Stat Inference Review

Yuling Max Chen

May 2021

Abstract

This note is generated from the hand-written lecture notes of Prof. Julien Berestycki(http://www.stats.ox.ac.uk/~berestyc/), who taught the course "*Foundation of Statistical Inference*" in the Michaelmas Term 2020. Instead of simply copying, this note also contains some clarifications of the materials and completes some proofs by padding some skipped steps, as well as comments with some personal thoughts. No individual or party should use this note for any purpose other than personal study. Should there be any typo, please contact me at yuling.chen@mansfield.ox.ac.uk. All criticisms and/or comments are very much welcome.

Contents

1 Exponential Families

Def 1.1: A family $\{f(x;\theta), \theta = (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k\}$ of pdf/pmf's indexed by θ is *k-parameter exponential family* if the pdf/pmf's $f(x;\theta)$ takes the form:

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

$$\stackrel{or}{=} \exp\left[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\eta)\right] h(x), \text{ the canonical form}$$

$$\because \forall \theta \in \Theta, 1 = \int_{\mathcal{X}} f(x;\theta)dx = \exp(-B(\theta)) \int_{\mathcal{X}} \exp\left[\sum_{i=1}^{k} \eta_i(\theta) T_i(x)\right] h(x)dx$$

$$\implies B(\theta) \text{ only depends on } \eta(\theta)$$

Where,
- $\eta_i()$ are the natural/canonical parameters
- $T_i(x)$ are real valued statistics(functions of x), called natural/canonical obs
- θ, x can be multidimensional

- The support of an exponential family does not depend on θ. - The Cauchy family, $f(x;\mu) = \frac{1}{\pi(1+(x-\mu)^2)}, x \in \mathbb{R}$, is not exponential as it cannot be transformed into the exponential form, despite the support does not depend on μ.

1.1 1-Parameter Case

Poisson:

$$pdf_X(x) = f(x;\theta) = e^{-\theta}\frac{\theta^x}{x!}$$

$$= \underbrace{\frac{1}{x!}}_{h(x)} \exp(\underbrace{x}_{T(x)} \underbrace{\log\theta}_{\eta(\theta)} - \underbrace{\theta}_{B(\theta)}), x = 0, 1, 2 \cdots$$

Binomial(n,p), where n is known:

$$pdf_X(x) = f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \underbrace{\binom{n}{x}}_{h(x)} \exp\left[\underbrace{x}_{T(x)} \underbrace{\left(\log(\frac{p}{1-p})\right)}_{\eta(p)} + \underbrace{n\log(1-p)}_{-B(p)}\right], x = 0, 1, 2, \cdots, n$$

Normal($\mu, \sigma^2 = 1$) with known variance:

$$pdf_X(x) = f(x;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

$$= \underbrace{\frac{\exp(-x^2/2)}{\sqrt{2\pi}}}_{h(x)} \cdot \exp\left(\underbrace{\mu}_{\eta(\mu)} \underbrace{x}_{T(x)} - \underbrace{\frac{\mu^2}{2}}_{B(\mu)}\right)$$

1.2 n-parameter Case

Gamma(α, β):

$$pdf_X(x) = f(x;\theta) = \frac{\beta^\alpha x^{\alpha-1}\exp(-\beta x)}{\Gamma(\alpha)}\mathbb{I}_{x\geq 0}$$
$$= \exp\left[\underbrace{(\alpha-1)}_{\eta_1(\theta)}\underbrace{\log x}_{T_1(x)} - \underbrace{\beta}_{\eta_2(\theta)}\underbrace{x}_{T_2(x)} - \underbrace{(\log(\Gamma(\alpha)) - \alpha\log\beta)}_{B(\theta)}\right]\cdot\underbrace{\mathbb{I}_{x\geq 0}}_{h(x)}$$

$N(\mu,\sigma^2)$:

$$pdf_X(x) = f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$= \exp\left[\underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\theta)}\underbrace{x^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{\eta_2(x)}\underbrace{x}_{T_2(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{B(\theta)}\right]$$

1.3 The Parameter Space

Def 1.2:

$$\Theta = \left\{\theta : \int_{\mathcal{X}} h(x)\exp\left[\sum_{i=1}^n \eta_i(\theta)T_i(x)\right]dx < \infty\right\}$$
$$\Xi = \left\{\theta : \int_{\mathcal{X}} h(x)\exp\left[\sum_{i=1}^n \eta_i T_i(x)\right]dx < \infty\right\}, \text{ the natural parameter space}$$

- Note that $\eta(\Theta) \subseteq \Xi$ and $\eta(\Theta) \neq \Xi$.

Prop 1.1: Ξ is convex. Proof of Prop 1.1: Take $\eta, \eta' \in \Xi$ and let $\alpha \in (0,1)$. Define $B(\eta) = \log(\int \exp[\sum_i \eta_i T_i(x)]h(x)dx$. ($\because 1 = \int_{\mathcal{X}} f(x;\theta)dx = \exp(-B(\theta))\int_{\mathcal{X}}\exp\left[\sum_{i=1}^k \eta_i(\theta)T_i(x)\right]h(x)dx$) Then,

$$B(\alpha\eta + (1-\alpha)\eta')$$
$$= \log\int \exp(\alpha\sum_i \eta_i T_i(x) + (1-\alpha)\sum_i \eta_i' T_i(x))h(x)dx$$
$$= \log\int \left[\exp(\sum_i \eta_i T_i(x))h(x)\right]^\alpha\left[\exp(\sum_i \eta_i' T_i(x))h(x)\right]^{(1-\alpha)}dx$$
($\because h = h^\alpha h^{(1-\alpha)}$)
$$\leq \log\left[\int \exp(\sum_i \eta_i T_i(x))h(x)dx\right]^\alpha\left[\int \exp(\sum_i \eta_i' T_i(x))h(x)dx\right]^{(1-\alpha)}$$
$\because$ By Holder's Ineq, $E(XY) \leq E(|X|^p)^{1/p}E(|Y|^q)^{1/q}$ for $\frac{1}{p}+\frac{1}{q}=1$
$$\implies \int |fg|dx \leq \left(\int |f|^p\right)^{1/p} + \left(\int |g|^q\right)^{1/q}$$
$$\leq \alpha B(\eta) + (1-\alpha)B(\eta') < \infty$$

■

Def 1.3: A family s.t Ξ is open and non-empty is called **regular**. **Def 1.4:** The function $T_1, \cdots T_n$ are called $\mathcal{P}$ **-affine independent**, if $\forall c_j \in \mathbb{R}, c_0 \in \mathbb{R}$,

$$\sum_{j=1}^n c_j T_j(x) = c_0, \forall x \in \mathcal{A} \implies c_j = 0, j = 0, 1\cdots n$$

Prop 1.2: The functions T_i are $\mathcal{P}$-affine independent, if $Cov_\eta T$ is positive definite $\forall \eta \in \Xi$.
Def 1.5: The family is **strictly k-dimensional** if the functions $\eta_i(\theta)$ are linearly indep and the $T_i's$ are $\mathcal{P}$-affine independent.
Ex 1.1: Omitted, see P13.
Thm 1.1: The natural parameter space Ξ of a strictly k-parameter exponential family is convex and contains a non-empty k-dim interval.
Thm 1.2: Let $\mathcal{P}$ be a strictly k-parameter exponential family and natural parameter space Ξ. $\forall \eta \in Int(\Xi)$, we have: (a) All moments of T w.r.t $f(x;\eta)$ exist;
(b) $\mathbb{E}_\eta[T_i(x)] = \frac{\partial}{\partial \eta_i} B(\eta)$;
(c) $Cov(T_i, T_j) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} B(\eta)$.

Proof Thm 1.2:
(a):

$$\begin{aligned} MGF_{T(x)}(s) = E_\eta(e^{s\cdot T(x)}) &= E_\eta\left[\exp\left(\sum_{i=1}^k s_i T_i(x)\right)\right] \\ &= \int \exp\left(\sum_{i=1}^k s_i T_i(x)\right) h(x)dx \\ &= \int \exp\left(\sum_{i=1}^k (s_i+\eta_i)T_i(x) - \eta_i T_i(x)\right) h(x)dx \\ &= \exp(B(\eta+s) - B(\eta)) \\ &\because \exp(B(\eta)) = \int \exp\left(\sum_{i=1}^k \eta_i T_i(x)\right) h(x)dx \\ \implies \log(MGF_{T(x)}(s)) &= B(\eta+s) - B(\eta) \end{aligned}$$

$\because \eta \in Int(\Xi)$ and $\Xi = \{z : B(z) < \infty\} \implies \exists \delta > 0 : \forall |s| < \delta, MGF_{T(x)}(s) < +\infty. \implies E_\eta[|T(x)|^k] < \infty, \forall k$, all moments are finite. ■

(b): Differentiate both sides of $\exp(B(\eta)) = \int \exp\left(\sum_{i=1}^k \eta_i T_i(x)\right) h(x)dx$ w.r.t η_i:

$$\begin{aligned} \exp(B(\eta))\frac{\partial}{\partial \eta_i} B(\eta) &= \int T_i(x) \exp\left[\sum \eta_i T_i(x)\right] h(x)dx \\ \implies \frac{\partial}{\partial \eta_i} B(\eta) &= \int T_i(x) \exp\left[\sum \eta_i T_i(x) - B(\eta)\right] h(x)dx \\ &= \int T_i(x) pdf_X(x)dx = E_\eta(T_i(x)) \end{aligned}$$

■

(c): omitted

2 Sufficient and Minimal Sufficiency

2.1 General Case

Def 2.1: Suppose $X \sim f(x;\theta)$. A **statistic** $T(x)$ is a function of the data that does not depend on θ.
$T(x)$ is **sufficient** for θ if $f(x|T=t,\theta) = f(x|T=t)$.

Ex 2.1: omitted, see P18.

Thm 2.1 (Factorization Theorem):
$T(x)$ is a sufficient statistic for θ iff $\exists g, h > 0 : f(x;\theta) = g(T(x);\theta)h(x)$.

Proof of Thm 2.1:

($\Longrightarrow$) : Suppose T sufficient and $T(x) = t$.

$$f(x;\theta) = P_\theta(X = x) = P_\theta(X = x, T = t) = \underbrace{P_\theta(X = x|T = t)}_{\perp\theta, \because T \text{ is suff}} \underbrace{P_\theta(T = t)}_{g(t,\theta)}$$

($\Longleftarrow$) : Suppose $f(x;\theta) = g(t;\theta)h(x)$

$$\Longrightarrow P_\theta(T = t) = \sum_{x:T(x)=t} P_\theta(X = x) = \sum_{x:T(x)=t} f(x;\theta) = g(t,\theta) \sum_{x:T(x)=t} h(x)$$

$$\Longrightarrow P_\theta(X = x|T = t) = \frac{P_\theta(X = x, T = t)}{P_\theta(T = t)} = \frac{P_\theta(X = x)}{P_\theta(T = t)} = \frac{f(x;\theta)}{P_\theta(T = t)} = \underbrace{\frac{h(x)}{\sum_{x:T(x)=t} h(x)}}_{\perp\theta}$$

■

Def 2.2: A statistic is **minimal sufficient** if it can be expressed as a function of any other sufficient statistics.

Thm 2.2 (Lehmon-Scheffe Theorem): A statistic T is minimal sufficient iff $T(x) = T(y) \iff \frac{f(y;\theta)}{f(x;\theta)} = k(x, y) \perp \theta$.

Proof of Thm 2.2:
($\Longleftarrow$) : Suppose T is a statistic s.t. $T(x) = T(y) \iff \frac{f(y;\theta)}{f(x;\theta)} = k(x, y) \perp \theta$. Sufficiency:

$$f(x|t,\theta) = P_\theta(X = x|T = t) = \frac{P_\theta(X = t)}{P_\theta(T = t)} = \frac{f(x,\theta)}{\sum_{y\in\tau} f(y,\theta)}, \tau = \{y : T(y) = t\}$$

$$= \frac{f(x,\theta)}{\sum_{y\in\tau} f(x,\theta)k(x,y)} = [\sum_{y\in\tau} k(x,y)]^{-1} \perp \theta \Longrightarrow \text{Sufficient}$$

Minimality: Suppose U is another sufficient statistic and that $U(x) = U(y)$ for some x, y. Hence,

$$\frac{f(y,\theta)}{f(x,\theta)} = \frac{g(U(y),\theta)h(y)}{g(U(x),\theta)h(x)} = \frac{h(y)}{h(x)} \perp \theta$$
$$\Longrightarrow T(x) = T(y), T(x) \text{ can be expressed by } U(x).$$

($\Longrightarrow$) : Suppose T is minimal sufficient, then,

$$T(x) = T(y) \Longrightarrow \frac{f(y,\theta)}{f(x,\theta)} \overset{suff.}{=} \frac{g(T(y),\theta)h(y)}{g(T(x),\theta)h(x)} = \frac{h(y)}{h(x)} \perp \theta$$

Let $D(x) = \{y : f(x,\theta) = k(x,y)f(y,\theta)\}, \forall\theta$ and $D_0 = \{x : f(x,\theta)\}$. For each class $D(x)$, choose a representation $\bar{x}$ and define $G : y \in D(x) \to \bar{x}$. G is a statistic constant on classes of the partition, i.e. $\forall x_1, x_2 \in D(x), G(x_1) = G(x_2)$. Also, G is sufficient because $f(x,\theta) = k(x,\bar{x})f(\bar{x},\theta) = \underbrace{k(x, G(y))}_{\perp\theta} f(G(y),\theta)$, by the factorization thm. Since T is minimal, it is a function of G. Hence, $T(x_1) = g(G(x_1)) = g(G(x_2)) = T(x_2)$. ■

2.2 Case of Exponential Families

Thm 2.3: Suppose $f(x,\theta) = \exp\left[\sum_{i=1}^k \eta_i(\theta)T_i(x) - B(\theta)\right] h(x)$ form a strictly k-parameter exponential family. Let $(X_1, \cdots, X_n) \overset{iid}{\sim} f(x,\theta)$. Then,
(a) $T_{(n)}(x) = (\sum_{i=1}^n T_1(x_i), \cdots, \sum_{i=1}^n T_k(x_i))$ is minimal sufficient.
(b) The distribution of $T_{(n)}(x)$ belongs to a k-parameter family.

Proof of Thm 2.3(a):

$$\frac{f((x_1,\cdots,x_n),\theta)}{f((y_1,\cdots,y_n),\theta)} = \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(y_i)} \exp\left[\sum_{j=1}^k \eta_i(\theta)\left(\sum_{i=1}^n T_j(x_i) - \sum_{i=1}^n T_j(y_i)\right)\right] \perp \theta$$

$$\iff \sum_{i=1}^n T_j(x_i) = \sum_{i=1}^n T_j(y_i), \forall j = 1,\cdots,k$$

Ex 2.2: Find the minimal sufficient statistic for θ. partially omitted, see P26-27. (2) Let $X_1,\cdots,X_n \overset{iid}{\sim} Unif(a,b)$. Note,

$$f((x_1,\cdots,x_n),\theta) = \prod_{i=1}^n \frac{1}{b-a}\mathbb{I}_{[a,b]}(x_i) = (b-a)^{-1}\cdot\mathbb{I}(x_{(1)}\geq a)\cdot\mathbb{I}(x_{(n)}\leq b)$$

By factorization thm, the sufficient statistic is $T(x) = (x_{(1)}, x_{(n)})$. Then,

$$\frac{f((x_1,\cdots,x_n),\theta)}{f((y_1,\cdots,y_n),\theta)} = \frac{\mathbb{I}(X_{(1)}\geq a)\cdot\mathbb{I}(X_{(n)}\leq b)}{\mathbb{I}(Y_{(1)}\geq a)\cdot\mathbb{I}(Y_{(n)}\leq b)} \perp (a,b) \implies \text{sufficient}$$

$$\not\iff T(x) = T(y) \implies \text{but not minimal}$$

(3) Let $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Then, for the parameter $\theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}^+$,

$$\frac{f(x,\theta)}{f(y,\theta)} = \frac{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2)}{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 - 2\mu\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right)\right)$$

$$\perp\theta \iff \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \text{ and } \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$$

$$\implies T(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right) \text{ is minimal sufficient}$$

Also note, $T'(x) = (\bar{x}, s^2) = \left(\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2\right) = \left(\frac{1}{n}T_1(x), \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - (\frac{1}{n}\sum_{i=1}^n x_i)^2\right)\right)$

$$= \left(\frac{1}{n}T_1(x), \frac{1}{n-1}\left(T_2(x) - \frac{1}{n}T_1(x)^2\right)\right)$$

is 1-1 corresponded to $T(x)$, hence is also minimal sufficient.

3 Fisher Information & Point Estimation

3.1 Regularity Assumptions

Consider the likelihood function $L(\theta,x) = f(x;\theta)$. (1) The support of $L(\theta,x)$, $\mathcal{A} = \{x : f(x;\theta) > 0\} \perp \theta$; - Exponential familiy satisfies this. (2) $\Theta \subseteq \mathbb{R}^k$ is open; if k = 1, the interval can be either finite or infinite. (3) $\forall x\in\mathcal{A}, \theta\in\Theta, \left|\frac{\partial f(x;\theta)}{\partial\theta_i}\right| < +\infty$. • Under Regularity Assumptions, $\frac{\partial}{\partial\theta}\int_{\mathcal{A}} f(x,\theta)dx = \int_{\mathcal{A}}\frac{\partial}{\partial\theta}f(x,\theta)dx$. (4) $\forall x\in\mathcal{A}, \theta\in\Theta, l''$ exists and

$$\frac{\partial^2}{\partial\theta^2}\int_{\mathcal{A}} f(x,\theta)dx = \int_{\mathcal{A}}\frac{\partial^2}{\partial\theta^2}f(x,\theta)dx$$

$$\frac{\partial^2}{\partial\theta^2}\sum_{x\in\mathcal{A}} f(x,\theta) = \sum_{x\in\mathcal{A}}\frac{\partial^2}{\partial\theta^2}f(x,\theta)$$

3.2 Score function

Def 3.1 (Score function):

$$\text{1-dim: } S(\theta,x) = l'(\theta,x) = \frac{\partial}{\partial\theta}\log L(\theta,x)$$
$$\text{k-dim: } S(\theta,x) = \nabla_\theta l(\theta,x) = \left(\frac{\partial}{\partial\theta_1}\log L(\theta,x), \cdots, \frac{\partial}{\partial\theta_k}\log L(\theta,x)\right)^T$$

Thm 3.1: Under Reg 1-3:

$$\mathbb{E}_\theta S(\theta,x) = 0, \forall\theta\in\Theta$$

Proof of Thm 3.1:

$$\mathbb{E}_\theta S(\theta,x) = \int_{\mathcal{A}} \underbrace{l'(\theta,x)}_{(\log L)'} f(x,\theta)dx = \int_{\mathcal{A}} \frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)} f(x,\theta)dx = \frac{\partial}{\partial\theta}\int_{\mathcal{A}} f(x,\theta)dx = \frac{\partial}{\partial\theta}1 = 0$$

■

3.3 Fisher Information

Def 3.2 (Fisher Information):

$$\text{1-dim: } I_X(\theta) = Var_\theta[S(\theta,x)] = \mathbb{E}_\theta[l'(\theta,x)^2]$$
$$\text{k-dim: } I_X(\theta) = Cov_\theta(S(\theta,x))$$
$$\text{where, } I_X(\theta)_{jr} = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j}l(\theta,x)\frac{\partial}{\partial\theta_r}l(\theta,x)\right]$$

- $\because \mathbb{E}_\theta S(\theta,x) = 0$.

Thm 3.2: Under Reg 1-4

$$\text{1-dim: } I_X(\theta) = -\mathbb{E}_\theta[l''(\theta,x)] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log P(x,\theta)\right]$$
$$\text{k-dim: } I_X(\theta) = \mathbb{E}_\theta[J(\theta,x)],$$
$$\text{where, the observee Fisher Info Matrix } J(\theta,x) = -\frac{\partial^2 l(\theta,x)}{\partial\theta_j\partial_r}, \forall j,r = 1,\cdots,k$$

Proof of Thm 3.2: (1-dim)

$$l''(\theta,x) = \frac{\partial^2}{\partial\theta^2}\log f(x,\theta) = \frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta}f}{f} = \frac{(\frac{\partial^2}{\partial\theta^2}f)f - (\frac{\partial}{\partial\theta}f)^2}{f^2} = \frac{\frac{\partial^2}{\partial\theta^2}f}{f} - \left(\frac{\frac{\partial}{\partial\theta}f}{f}\right)^2$$
$$\text{By Reg 4, } \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f}{f}\right] = \int_{\mathcal{A}}\frac{\frac{\partial^2}{\partial\theta^2}f}{f}\cdot f dx = \int_{\mathcal{A}}\frac{\partial^2}{\partial\theta^2}f dx = \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{A}} f dx = 0$$
$$\implies -\mathbb{E}_\theta[l''] = \mathbb{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}f}{f}\right)^2\right] = \mathbb{E}_\theta[l'^2] = I_X(\theta)$$

■

Properties of Fisher Information: (1) $X\perp Y \implies I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$;
(2) Reparametrization: If $\theta = h(\xi)$, where $h()$ differentiable, then the Fisher info of $X\sim P_{h(\xi)}$ is $I_X^*(\xi) = I_X(h(\xi))[h(\xi)]^2$.

3.4 Point Estimation

Def 3.3: A function $T : \mathcal{X} \to \Gamma$ is an **estimator** of $\gamma = g(\theta)$, for $g : \Theta \to \Gamma$. The value $T(x)$ is called the **estimate** of $g(\theta)$.
Def 3.4: The **bias** of T for $\gamma = g(\theta)$ is, $bias(T, \theta) = E_\theta[T] - g(\theta)$. T is **unbiased** for $g(\theta)$ if $E_\theta[T] = g(\theta)$.

3.5 Method of Moments (MoM)

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} P_\theta$, then $m_\gamma = E_\theta[X^r]$ depends on θ, for $\gamma = 1, 2, \cdots$. Let $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n x_i^k$, for $k = 1, 2, \cdots, \gamma$. Then the **moment estimator** for γ is:

$$\hat{\gamma}_{MoM} = h(\hat{m}_1, \cdots, \hat{m}_\gamma)$$

Ex 3.1: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} Poisson(\lambda), \lambda > 0$. Note: $m_1 = E[X_i] = \lambda$ and $Var(X_i) = \lambda = m_2 - m_1^2$. Then we have:

$$\hat{\lambda}_{MME} = \hat{m}_1 = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\tilde{\lambda}_{MME} = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

3.6 Maximum Likelihood Estimator (MLE)

Def 3.5: $T(x)$ is the **MLE** of θ if

$$L(T(x); x) = \max_{\theta \in \Theta} L(\theta, x), \forall x \in \mathcal{X}$$

Thm 3.3: If $\gamma = g(\theta)$ and g is bijective, then $\hat{\theta}$ is a MLE iff $\hat{\gamma} = g(\hat{\theta})$ is a MLE for γ.
- If g is not bijective, define $\hat{\gamma}_{MLE} = g(\hat{\theta}_{MLE})$.
Thm 3.4: If $L(\theta, x)$ is differntiable w.r.t θ and has a unique max in $Int(\Theta)$, then $\hat{\theta}_{MLE}$ is the unique solution of $\frac{\partial}{\partial\theta} L(\theta, x) = 0$.

3.7 Variance and Mean Square Error (MSE)

Def 3.6: The **MSE** (also called the **quadratic loss function**) of T for $g(\theta)$ is:

$$MSE_\theta(T) = \mathbb{E}_\theta[(T - g(\theta)^2]$$

Prop 3.1:

$$MSE_\theta(T) = Var_\theta(T) + \underbrace{(E_\theta(T) - g(\theta))}_{bias^2}^2$$

- MSE = Var, for unbiased estimators.
Ex 3.2: $X = (X_1, \cdots, X_n) \overset{iid}{\sim} Unif(0, \theta)$. (1) MLE:

$$cdf_{(X_1,\cdots,X_n)}(x) = P(X_1 \le x, \cdots, X_n \le x) = P(X_{\max} \le x) = \prod_{i=1}^n P(X_i \le x) = \left(\frac{x}{\theta}\right)^n \mathbb{I}_{0 \le x \le \theta}$$

$$\implies pdf_{(X_1,\cdots,X_n)}(x) = P(X_{\max} = x) = \frac{\partial}{\partial x} P(X_{\max} \le x) = n\frac{x^{n-1}}{\theta^n}\mathbb{I}_{0 \le x \le \theta} = L(\theta, x)$$

$$\implies \hat{\theta}_{MLE} = \hat{\theta} = X_{\max}, \because \theta \ge x_i, \forall i$$

(2) MSE:

$$E_{\hat{\theta}}(X_{\max}) = \int_0^\theta n\frac{x^{n-1}}{\theta^n}\cdot x dx = \frac{n}{\theta^n}\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta$$

$$E_{\hat{\theta}}(X_{\max}^2) = \int_0^\theta n\frac{x^{n-1}}{\theta^n}\cdot x^2 dx = \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2$$

$$\implies Var_{\hat{\theta}}(X_{\max}) = E_{\hat{\theta}}(X_{\max}^2) - [E_{\hat{\theta}}(X_{\max})]^2 = \frac{n\theta^2}{(n+1)^2(n+2)}$$

$$Bias_{\hat{\theta}}(X_{\max}) = E_{\hat{\theta}}(X_{\max}) - \theta = -\frac{\theta}{n+1}$$

$$\implies MSE_{\hat{\theta}}(X_{\max}) = Var_{\hat{\theta}}(X_{\max}) + (Bias_{\hat{\theta}}(X_{\max}))^2$$

$$= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}$$

(3) An unbiased estimator: $\tilde{\theta} = \frac{n+1}{n}X_{\max}$

$$MSE_{\tilde{\theta}} = Var_{\tilde{\theta}}(X_{\max}) + bias_{\tilde{\theta}}(X_{\max})^2$$

$$= \frac{(n+1)^2}{n^2}Var_{\hat{\theta}}(X_{\max}) + \left(\frac{n+1}{n}E_{\hat{\theta}}(X_{\max}) - \theta\right)^2$$

$$= \frac{\theta^2}{n(n+2)} + 0 < \frac{2\theta^2}{(n+1)(n+2)} = MSE_{\hat{\theta}}$$

4 Minimum Variance Unbiased Estimators (MVUE) and the Cramer Rao's Lower Bound (CRLB)

Def 5.1: T_1 is a **better estimator than** T_2 **in quadratic mean** if

$$\forall\theta \in \Theta, MSE_\theta(T_1) \leq MSE_\theta(T_2)$$

- If $\hat{\theta} = \theta_0$ (the true parameter), then $MSE_{\theta_0}(\hat{\theta}) = 0$ and hence no other estimator can be uniformly better.

4.1 Single Parameter ($\Theta \in \mathbb{R}$) Case

Def 5.2: $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ is the **minimum variance unbiased estimator (MVUE)** for θ or $g(\theta)$ if: (1) $E_\theta[\hat{\theta}] = \theta, \forall\theta$ (Unbiased), and; (2) $\forall\tilde{\theta}$ unbiased, $Var_\theta(\tilde{\theta}) \geq Var_\theta(\hat{\theta})$ (Minimum variance).

Thm 5.1 (CRLB): Given the Regularity Conditions, and that $0 < I_X(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}l(\theta, x)\right] < \infty$. Let $\gamma = g(\theta)$ (continuously differentiable and real-valued with $g' \neq 0$). Let T be an unbiased estimator of γ, then:

$$Var_\theta(T) \geq \frac{|g'(\theta)|^2}{I_X(\theta)} = CRLB$$

The equality holds iff:

$$T(x) - g(\theta) = \frac{g'(\theta)S(\theta, x)}{I_X(\theta)}, \forall x \in \mathcal{A}, \theta \in \Theta$$

Remark of Thm 5.1: • If T attains the CRLB, then it is a MVUE. (But the CRLB is not guaranteed to exist.) • If $g(\theta) = \theta$, $Var_\theta(T) \geq \frac{1}{I_X(\theta)} = CRLB$ and T attains CRLB iff $S(\theta, x) = I_X(\theta)(T(x) - \theta)$ (or $T(x) = \theta + \frac{S(\theta,x)}{I_x(\theta)}$).

Proof of Thm 5.1:

$$
\begin{aligned}
Cov_\theta(T, S(\theta,x)) &= E_\theta(T\cdot S(\theta,x)) + E_\theta(T)\cdot E_\theta(S(\theta,x)) = E_\theta(T\cdot S(\theta,x)), \because E_\theta(S(\theta,x)) = 0 \\
&= \int_{\mathcal{X}} T(x)\frac{\partial}{\partial\theta}\ln P(x,\theta)\cdot P(x,\theta)dx = \int_{\mathcal{X}} T(x)\frac{\partial}{\partial\theta}P(x,\theta)dx \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} T(x)P(x,\theta)dx, \text{ by Reg Cond.} \\
&= \frac{\partial}{\partial\theta}\underbrace{E_\theta[T]}_{g(\theta)} = g'(\theta) - -(*)
\end{aligned}
$$

Setting $c(\theta) = \frac{g'(\theta)}{I_X(\theta)}$, then:

$$
\begin{aligned}
0 \le Var_\theta(T - c(\theta)S(\theta,x)) &= Var_\theta(T) + c^2(\theta)Var_\theta(S(\theta,x)) - 2c(\theta)Cov_\theta(T, S(\theta,x)) \\
&= Var_\theta(T) + c^2(\theta)I_X(\theta) - 2c(\theta)g'(\theta), \text{ by Def 3.2 and } (*) \\
&= Var_\theta(T) - \frac{|g'(\theta)|^2}{I_X(\theta)}
\end{aligned}
$$

The equality holds when $T - c(\theta)S(\theta,x)$ is almost surely constant, which must equal to its expectation $E_\theta[T - c(\theta)S(\theta,x)] = E_\theta[T] - E_\theta[c(\theta)S(\theta,x)] = g(\theta)$. ($\because T$ is unbiased estimator of $\gamma = g(\theta)$ and $E[S] = 0$.) That is:

$$
T(x) - c(\theta)S(\theta,x) = g(\theta) \iff T(x) - g(\theta) = \frac{g'(\theta)S(\theta,x)}{I_X(\theta)}
$$

■

Def 5.3: A statistic T is **efficient** if its variance attains the CRLB.
Ex 5.1: Consider $X \sim Bin(n,\theta)$. Let $\gamma = g(\theta) = \theta(1-\theta)$ (hence, $g'(\theta) = 1 - 2\theta$). Then,

$$
\begin{aligned}
l(\theta,x) &= \ln\binom{n}{x} + (n-x)\ln(1-\theta) + x\ln\theta \\
S(\theta,x) &= -\frac{n-x}{1-\theta} + \frac{x}{\theta} \\
I_X(\theta) &= -E_\theta\left[\frac{\partial}{\partial\theta}S(\theta,x)\right] = \frac{n - \overbrace{E_\theta[x]}^{n\theta}}{1-\theta} + \frac{E_\theta[x]}{\theta} = \frac{n}{(1-\theta)\theta}
\end{aligned}
$$

Given an statistic $T(x) = \frac{1}{n-1}x(1-\frac{x}{n})$. Observe that it is unbiased for γ and its variance is greater than CRLB:

$$
\begin{aligned}
E_\theta(T(x)) &= \frac{1}{n-1}(\overbrace{E_\theta[x]}^{n\theta} - \frac{\overbrace{E_\theta[x^2]}^{n\theta(1-\theta)+n^2\theta^2}}{n}) = \frac{1}{n-1}(n\theta - \theta(1-\theta) - n\theta^2) = \theta, \text{ Unbiased} \\
Var_\theta(T(x)) &= E_\theta(T^2(x)) - [E_\theta(T(x))]^2 \overset{skipped}{=} \frac{\theta}{n} - \frac{\theta^2(5n-7) - 4\theta^2(2n-3) + \theta^4(4n-6)}{n(n-1)} \\
CRLB &= \frac{|g'(\theta)|^2}{I_X(\theta)} = \frac{(1-2\theta)^2\theta(1-\theta)}{n} < Var_\theta(T(x))
\end{aligned}
$$

4.2 Multiple Parameters ($\Theta \in \mathbb{R}^k$) Case

Suppose $\gamma = g(\theta) = \in \mathbb{R}^m$.
Def 5.4: Let T, T^* unbiased estimators for γ. Then T^* has a **smaller covariance matrix than** T at $\theta \in \Theta$ ($Cov_\theta T^* \preceq Cov_\theta T$) if:

$$
u^T(Cov_\theta T^* - Cov_\theta T)u \le 0, \forall u \in \mathbb{R}^m
$$

Thm 5.2: Given Regularity Conditions and that $T_X(\theta)$ is not singular, the CRLB is:

$$Cov_\theta T \geq (\Delta_\theta g)(\theta) I_X^{-1}(\theta)(\Delta_\theta g)^T(\theta), \forall \theta \in \Theta$$

Where,

$$\bullet \Delta_\theta g = \frac{\partial}{\partial \theta_l} g_i(\theta), i = 1, \cdots, m; l = 1, \cdots, k$$
$$\bullet I_X(\theta) = Cov_\theta(S(\theta, x))$$
$$\bullet S(\theta, x) = \frac{\partial}{\partial \theta_i} l(\theta, x), i = 1, \cdots, k$$

Ex 5.2: Consider $X = (X_1, \cdots, X_n) \overset{iid}{\sim} N(\mu, \sigma^2)$. Then $I_X(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$. The estimators $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are independent and:

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n}$$
$$Var(s^2) = \frac{2\sigma^4}{n-1} > (I_X(\theta)_{22})^{-1}$$
$$\because \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \implies Var\left(\chi^2_{n-1}\right) = 2(n-1) = Var\left(\frac{(n-1)s^2}{\sigma^2}\right)$$

4.3 Other Sepcial Cases

Thm 5.3 (for MLE's): Under Reg Conds, if $\hat{\theta}_{MLE}$ exists and there exists an unbiased estimator $\tilde{\theta}$ that attains the CRLB, then $\tilde{\theta} \overset{a.s}{=} \hat{\theta}_{MLE}$.
Thm 5.4 (for Exp Family): Suppose $X = (X_1, \cdots, X_n)$ is a one-parameter Exponential family with η and T. Then the sufficiency of a statistic T implies efficiency, if $E_\theta[T] = g(\theta) = \gamma$.
Proof of Thm 5.4: The normal form of the exponential family gives:

$$P(x, \theta) = h(x)\exp[T(x)\eta(\theta) - B(\theta)], \text{ by Sufficiency of } T$$
$$\implies S(\theta, x) = \frac{\partial}{\partial \theta} l(\theta, x) = \eta'(\theta)T(x) - B'(\theta)$$
$$\implies T(x) \text{ and } S(\theta, x) \text{ are linearly related.}$$
$$\implies Cov_\theta(S(\theta, x), T(x)) = \eta' Var_\theta(T) = \frac{1}{\eta'} Vat_\theta(S)$$
$$\implies \underbrace{Cov^2_\theta(S(\theta, x), T(x))}_{(g'(\theta))^2, \text{ by Pf of Thm 5.1}} = Var_\theta(T(x)) \underbrace{Var_\theta(S(\theta, x))}_{I_X(\theta)}$$
$$\implies Var_\theta(T) = \frac{|g'(\theta)|^2}{I_X(\theta)} = CRLB$$
$$\implies \text{Efficient, by Def 5.3}$$

■

5 Rao-Blackwell Theorem and Completeness

5.1 Rao-Blackwell Theorem

Thm 6.1 (Rao-Blackwell Theorem): Let RV $X \sim P_\theta$, statistic T sufficient, and $\hat{\gamma}$ unbiased estimator for $\gamma = g(\theta)$. Define $\hat{\gamma} = \mathbb{E}_\theta[\hat{\gamma} \mid T]$, then:
(a) $\hat{\gamma} \perp \theta$,
(b) $\forall \theta, \mathbb{E}_\theta[\hat{\gamma}_T] = \gamma$ (unbiased),

(c) For $\gamma \in \mathbb{R}^k$, $Var_\theta(\hat{\gamma}_T) \leq Var_\theta(\hat{\gamma})$ $(k=1)$ or $Cov_\theta(\hat{\gamma}_T) \preceq Cov_\theta(\hat{\gamma})$ $(k>1)$. (smaller variance)

- If $trace(Cov_\theta(\hat{\gamma})) < \infty$, then $Cov_\theta(\hat{\gamma}_T) = Cov_\theta(\hat{\gamma})$ iff $P(\hat{\gamma} = \gamma) = 1$.
- Any unbiased estimator can be improved by a sufficient statistic.

Proof of Thm 6.1:
(a) T is sufficient $\implies$ $f(x|\theta, T) \perp \theta$, by Def 2.1. Then,

$$\hat{\gamma}_T = \mathbb{E}_\theta[\hat{\gamma} \mid T=t] = \int_{\mathcal{X}} \hat{\gamma}(x) f(x|\theta, T=t)dx = \int_{\mathcal{X}} \hat{\gamma}(x) f(x|T=t)dx \perp \theta$$

(b)

$$\mathbb{E}[\hat{\gamma}_T] = \mathbb{E}[\mathbb{E}_\theta[\hat{\gamma} \mid T=t]] = \mathbb{E}[\hat{\gamma}] = \gamma, \because \hat{\gamma} \text{ is unbiased}$$

(c) $k=1$ case, $k>1$ case can be proved similarly:

$$\begin{aligned} Var(\hat{\gamma}) &= E[(\hat{\gamma}-\gamma)^2] = E[(\hat{\gamma}-\hat{\gamma}_T+\hat{\gamma}_T-\gamma)^2] = E[E[(\hat{\gamma}-\hat{\gamma}_T+\hat{\gamma}_T-\gamma)^2|T]] \\ &= E[E[(\hat{\gamma}-\hat{\gamma}_T)^2|T]] + E[E[(\hat{\gamma}_T-\gamma)^2|T]] + E[E[2(\hat{\gamma}-\hat{\gamma}_T)(\hat{\gamma}_T-\gamma)|T]] \\ &= \underbrace{E[E[(\hat{\gamma}|T - E[\hat{\gamma}|T])^2]]}_{\mathbb{E}[Var(\hat{\gamma}|T)]\geq 0} + \underbrace{E[(\hat{\gamma}_T - E[\hat{\gamma}_T])^2]}_{Var(\hat{\gamma}_T)} + \underbrace{2E[E[(\hat{\gamma}-\hat{\gamma}_T)|T]\cdot(\hat{\gamma}_T-\gamma)]}_{0} \\ &\geq Var(\hat{\gamma}_T) \end{aligned}$$

■

Ex 6.1: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} Bern(\theta)$. $\hat{\theta} = X_1$ is unbiased. $T = \sum_{i=1}^n X_i$ is sufficient. Then,

$$\begin{aligned} \hat{\theta}_T &= E_\theta[X_1|T=t] = P_\theta(X_1=1|T=t) = \frac{P(X_1=1, \sum_{i=1}^n X_1 = t-1)}{P(T=t)} \\ &= \frac{\theta \cdot \binom{n-1}{t-1}\theta^{t-1}(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{t}{n} \end{aligned}$$

5.2 Completeness

Def 6.1: A statistical model $\{P_\theta, \theta \in \Theta\}$ is **complete** if $\forall h: \mathcal{X} \to \mathbb{R}, \mathbb{E}_\theta[h(X)] = 0 \implies \mathbb{P}_\theta(h(x)=0) = 1, \forall \theta \in \Theta$.

Lemma 6.1: A statistic T is **complete** if the model $\{P_\theta, \theta \in \Theta\}$ is **complete**, i.e.

$$\mathbb{E}_\theta[h(T)] = 0 \implies \mathbb{P}_\theta(h(T)=0) = 1, \forall \theta \in \Theta$$

Ex 6.2: Consider a model that consists of the 2 distributions: $N(1,2)$ and $N(0,1)$. This model is not complete. Consider $h(x) = (x-1)^2 - 2$. Then, both $E(h(x)) = 0$ but $h(x) \neq 0, \forall x \neq 1 \pm \sqrt{2}$.

Ex 6.3: (P60) Model $\{Unif[0,\theta], \theta \in \mathbb{R}_+\}$ is complete. Because:

$$\begin{aligned} &\text{Suppose } E_\theta[h(x)] = \int_0^\theta \frac{1}{\theta} h(x)dx = 0, \forall \theta > 0 \\ &\implies \frac{\partial}{\partial \theta}\int_0^\theta h(x)dx = 0 \overset{a.e.}{=} h(\theta) \end{aligned}$$

If $(X_1 \cdots X_n) \overset{iid}{\sim} U[0,\theta]$, then $T(X) = X_{\max}$ is a complete statistic.

$$\begin{aligned} E_\theta[h(X_{\max})] &= \int_{-\infty}^{\infty} h(t) f_\theta(t)dt = \int_0^\theta h(t)\frac{n}{\theta^n}t^{n-1}dt = 0, \forall \theta \\ \implies \int_0^\theta h^-(t)t^{n-1}dt &= \int_0^\theta h^+(t)t^{n-1}dt, \forall \theta \\ \implies h^-(t) = h^+(t) &\implies h(t) = 0 \end{aligned}$$

Thm 6.2: Suppose $\mathcal{P}$ is a k-parameter exponential family with natural parameter $\eta = (\eta_1, \cdots, \eta_k)$ and that the natural parameter space Ξ contains a non-empty k-dimensional interval. Then, $T(x) = (T_1(x), \cdots, T_k(x))$ is both **sufficient** and **complete**.

Corallary 6.1: If P_θ belongs to a strictly k-parameter exponential family, then T(x) is sufficient and complete.

Thm 6.3 (Lehman-Scheffe's Theorem): Let T be a sufficient and complete statistic for the statistical model $\mathcal{P}$ and $\hat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$. Then, $\hat{\gamma} = E[\hat{\gamma}|T]$ is MVUE for γ.
• If T is a sufficient and complete statistic, the such unbiased estiamtor $\hat{\gamma}$ is unique.
Proof of Thm 6.3: By contradition. Suppose $\exists\tilde{\gamma}$ unbiased with $Cov_{\theta_0}\tilde{\gamma} \prec Cov_{\theta_0}\hat{\gamma}_T$ for some θ_0. By Rao-Blackwell's Thm, for $\tilde{\gamma}_T = E_\theta[\tilde{\gamma}|T]$,

$$Cov_{\theta_0}\tilde{\gamma}_T \preceq Cov_{\theta_0}\tilde{\gamma} \prec Cov_{\theta_0}\hat{\gamma}_T - - - (*)$$

But note that both $\hat{\gamma}_T$ and $\tilde{\gamma}_T$ are unbiased, hence

$$\begin{aligned} E(\hat{\gamma}_T) = E(\tilde{\gamma}_T) = \gamma &\implies E(\hat{\gamma}_T - \tilde{\gamma}_T) = E(h(T)) = 0 \\ &\implies P(h(T) = 0) = 1 = P(\hat{\gamma}_T = \tilde{\gamma}_T) \text{ , by completeness of } T \\ &\implies Cov_{\theta_0}\tilde{\gamma}_T = Cov_{\theta_0}\hat{\gamma}_T \text{ , contradicts } (*) \end{aligned}$$

■

Ex 6.4: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} Unif[0, \theta]$. Recall that $E_\theta[X_{\max}] = \frac{n}{n+1}\theta$ (by Ex 3.2), and $X_{\max}$ is complete and sufficient (by Ex 2.2 and 6.3). So, $\hat{\theta} = \frac{n+1}{n}X_{\max}$ is the MVUE, despite the CRLB does not apply here ($\because$ support $[0, \theta]$ depends on θ).

Ex 6.5: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} N(\mu, \sigma^2)$, this is a strictly 2-parameter exponential family. Also, from Ex 2.2, $T(x) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient. By Lemma 6.1, $T(x)$ is also complete, since the model is itself complete as a strictly 2-parameter exponential family. Consider $(\bar{X}, s^2)$. $E(\bar{X}) = \mu$ unbiased, easy to proof. $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is also unbiased because:

$$\begin{aligned} E\left[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2\right] &= \frac{1}{n-1}E\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] \\ &= \frac{n}{n-1}[(Var(X_1) + E(X_1)^2) - (Var(\bar{X}) + E(\bar{X})^2)] \\ &= \frac{n}{n-1}\left(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right) = \sigma^2 \end{aligned}$$

Also, $(\bar{X}, s^2)$ is a function of T, hence it is the MVUE by Thm 6.3, despite that s^2 does not attain the CRLB by Ex 5.2.

Ex 6.6: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} Poisson(\lambda)$. Recall $\hat{\lambda}_{MME} = \frac{1}{n}\sum_{i=1}^n X_i, \tilde{\lambda}_{MME} = \frac{1}{n}\sum_{i=1}^n (X_1 - \bar{X})^2$ from Ex 3.1. Poisson family is a strictly 1-parametric exponential family with $h(x) = \frac{1}{x!}, B(\lambda) = \lambda, T(x) = x, \eta(\lambda) = \log\lambda$. Hence, $\bar{X}$ is sufficient and complete, Corallary 6.1. Then, by Thm 6.3, $\hat{\lambda}_{MME}$ is MVUE.

(a) Get the Cramer-Rao bound. we have: $S(x, \lambda) = \frac{\sum_{i=1}^n x}{\lambda} - 1$ and $I_x(\lambda) = n\lambda^{-1}$. Hence, $CRLB = (I_x(\lambda))^{-1} = \frac{\lambda}{n} = Var(\bar{X})$. So, $\bar{X}$ is efficient. **(b)** Conditional distribution of X_i given $T(X) = \sum Xi$.

$$
\begin{aligned}
P_\lambda\left(X_i = m | \sum_{j=1}^n X_j = k\right) &= \frac{P_\lambda(X_i = m, \sum_{j=1}^n X_j = k)}{P_\lambda(\sum_{j=1}^n X_j = k)} \\
&= \frac{P_\lambda(X_i = m) P_\lambda(\sum_{j\neq i} X_j = k - m)}{\exp(-n\lambda)(n\lambda)^k/k!}, \because \text{ sun of indep Poisson is Poisson} \\
&= \frac{\lambda^m((n-1)\lambda)^{k-m}}{(n\lambda)^k} \exp(n\lambda - \lambda - (n-1)\lambda) \frac{k!}{(k-m)!m!} \\
&= \binom{k}{m}\left(1 - \frac{1}{n}\right)^{k-m}\left(\frac{1}{n}\right)^m \\
\implies X_i | \sum_{j=1}^k X_j = k \sim Bin\left(k, \frac{1}{n}\right) &\implies \mu = \frac{k}{n}; \sigma^2 = \frac{k}{n}(1 - \frac{1}{n}) - - - - - (*)
\end{aligned}
$$

Hence for $S^2 = \frac{n}{n-1}\tilde{\lambda}$, we get:

$$
\begin{aligned}
S_T^2 = E_\lambda(S^2|T) &= E_\lambda\left[\frac{1}{n-1}\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] | \sum_{j=1}^n X_j = k\right] \\
&= \frac{n}{n-1}\left[E_\lambda\left(X_1^2 | \sum_{j=1}^n X_j = k\right) - \frac{k^2}{n^2}\right] \\
&= \frac{n}{n-1}\left[\left(\frac{k}{n}\left(1 - \frac{1}{n}\right) + \frac{k^2}{n^2}\right) - \frac{k^2}{n^2}\right], \text{ by } (*) \\
&= \frac{k}{n} = \bar{X}
\end{aligned}
$$

So, starting from S^2 as an unbiased estimator for λ, we arrive at $\bar{X}$ by Thm 6.1 (Rao-Blackwell).

6 Bayesian Inference

In Bayesian inference, both the observation and the parameter are treated as random variables, i.e. full probabilistic model.
The posterior distribution is:

$$
\pi(\theta|x) = \frac{L(\theta, x)\pi(\theta)}{\underbrace{\int L(\theta, x)\pi(\theta)d\theta}_{\text{marginal likelihood}}} \implies \pi(\theta|x) \propto L(\theta, x)\pi(\theta)
$$

6.1 Conjugacy

Def 7.1: Consider a model $(L(\theta, x), \theta \in \Theta, x \in \mathcal{X})$. Then a family of prior distribution $(\pi_\gamma, \gamma \in \Gamma)$ is **conjugate** if:

$$
\forall \gamma \in \Gamma, x \in \mathcal{X}, \exists \gamma(x) : \pi_\gamma(\cdot|x) = \pi_{\gamma(x)}(\cdot)
$$

- i.e. posterior takes the same form as prior.

Ex 7.1: Omiited, see P72-73.

6.2 Priors

Proposition 7.1: Suppose $L(\theta, x) = h(x)\exp\{\sum_{i=1}^{k}\eta_i(\theta)T_i(x) - B(\theta)\}$ is a k-parameter exponential family. Then $\pi_\gamma(\theta) \propto \exp\{\gamma_0 B(\theta) + \sum_{i=1}^{k}\gamma_i\eta_i(\theta)\}$ are a **conjugate prior family**, where $\gamma = (\gamma_0, \cdots, \gamma_k)$ are parameters.

Ex 7.2: Consider $(X_1, \cdots, X_n) \overset{iid}{\sim} Poisson(\theta)$. Then $L(\theta, x) \propto \exp(-n\theta)\theta^{t(x)} = \exp(t(x)\log\theta - n\theta)$, with $t(x) = \sum_{i=1}^{n} x_i, \theta > 0$. The natual conjugate prior is: $\pi(\theta) \propto \exp\{\gamma_0 n\theta + \gamma_1 \log\theta\}$. Note that, this is normalisable if $\gamma_0 < 0$ and $\gamma_1 > -1$. Take $\gamma_0 = -\beta/n$ and $\gamma_1 = \alpha - 1$, then: $\pi(\theta) \propto \theta^{\alpha-1}\exp(-\beta\theta) \sim \Gamma(\alpha, \beta)$. Hence, the posterior: $P(\theta|x) \sim \Gamma(\alpha + \sum_{i=1}^{n} x, \beta + n)$.

Def 7.1: A pdf/pmf π is an **improper prior** if it has infinite mass $\int_\Theta \pi(\theta)d\theta = \infty, \pi(\theta) \geq 0$.
But a posterior $\pi(\theta|x)$ can be defined as long as $\int_\Theta f(x,\theta)\pi(\theta)d\theta < +\infty$ almost surely in x.

Ex 7.3: $X|p \sim Bin(n,p), \pi(p) = (p(1-p))^{-1}$, and $\pi(p|x) \propto p^{x-1}(1-p)^{n-x-1}$ improper if $x = 0$ or $x = n$. So, the posterior is not always well defined.

Ex 7.4: If X is discrete and can take only finitely many values $\{x_1, \cdots, x_n\} = \mathcal{X}$, then show that one cannot use an improper prior:
Note that $\pi(\theta|x) = \frac{\pi(\theta)f(x,\theta)}{m(x)}$, where $m(x) = \int f(x,\theta)\pi(\theta)d\theta$. Hence, $\sum_{i=1}^{n} m(x_i) = \sum_{i=1}^{n}\int f(x_i,\theta)\pi(\theta)d\theta = \int \underbrace{\sum_{i=1}^{n} f(x_i,\theta)}_{=1}\pi(\theta)d\theta = \infty$, because $\pi(\theta)$ is improper. Then, $m(x_i)$ cannot be finite for all $x_i \in \mathcal{X}$, because a finite sum gives an infinite value.

6.3 Non-Informative Priors

Subjective Prior (non-informative prior): a distribution representing prior knowledge about the parameter before any data is collected; can try different priors representing different "points of view".
Objective Prior: several approaches offer the promise of an "automatic" and even "objective" prior. They can be used when little or non-reliable info is available.

6.3.1 Uniform Priors

Def 7.2: Naive representation of lack of information: $\pi(\theta) =$ constant $\propto 1$. - Hence, the posterior: $\pi(\theta|x) = \frac{L(\theta,x)}{\int_\Theta L(\theta,x)d\theta}$, which exists if $\int_\Theta L(\theta,x)d\theta < \infty$ almost surely in x.

Ex 7.5: $X \sim Exp(\theta)$. Consider the uniform prior: $\pi(\theta) = 1$. The posterior is well defined because: $\int_{-\infty}^{\infty}\exp(-\theta x)\theta d\theta < \infty, \forall x > 0$. Then, $\theta|x \sim Exp(x)$. But, does it have good properties?
Reparametrizing the prior: let $\eta = \log\theta$. Then, $\tilde{\pi}(\theta) = \pi(\theta(\eta))\frac{d\theta}{d\eta} = \frac{d\theta}{d\eta} = e^\eta \neq 1$. As a prior in η, $\tilde{\pi}$ is very informative (put a lot of wright for large η).

6.3.2 Jeffery's Prior

Jeffery:"If we have a rule for constructing a prior, it should not depend on parametrization."

Def 7.3: Jeffery's Prior is: $\pi(\theta) \propto \sqrt{I_\theta}$, where $I_\theta = E_\theta\left[\frac{\partial^2}{\partial\theta^2}L(\theta,x)\right]$ is the fisher information.
- In k dimension $\Theta \subset \mathbb{R}^k$, under Reg Cond 1-4, the **k-dim Jeffery's Prior** is: $\pi(\theta) \propto \det(I_\theta)^{1/2}$, where $(I_\theta)_{ij} = -E_\theta\left[\frac{\partial^2 l(\theta,x)}{\partial\theta_i\partial\theta_j}\right]$.

• Reparametrization does not change the prior: Consider $g(\psi) = \theta$, for some continuous differentiable function g. Then, $\tilde{\pi}(\psi) \propto \pi(g(\psi))|g'(\psi)| = \sqrt{I_\theta}|g'(\psi)| = \sqrt{I_\theta \cdot g'(\psi)^2} = \sqrt{I_\psi}$.

Ex 7.6: Suppose $X \sim Poisson(\lambda), f(x,\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, \cdots$. Then the Jeffery's Prior is:

$$\pi(\lambda) \propto \sqrt{I_x(\lambda)} = \left(E\left[\left(\frac{\partial}{\partial\lambda}l(\lambda,x)\right)^2\right]\right)^{1/2} = \left(E\left[\left(\frac{x}{\lambda}-1\right)^2\right]\right)^{1/2}$$
$$= \left(\sum_{x=0}^{\infty} f(x,\lambda)\left(\frac{n-\lambda}{\lambda}\right)^2\right)^{1/2} = \left(\frac{1}{\lambda^2}\underbrace{E\left[(x-\lambda)^2\right]}_{Var(x-\lambda)=Var(x)=\lambda}\right)^2 = \lambda^{-1/2} \text{ (improper)}$$

6.3.3 Maximum Entropy Prior

Def 7.4: The **Entropy** of a pdf/pmf is defined as $Ent(\pi) = -\int_\Theta \pi(\theta)\log\pi(\theta)d\theta$. (Large entropy = space well explored at all scales.)

Def 7.5: The **maximum entropy prior**is: $\pi(\theta) = \arg\max_\pi Ent(\pi)$.

Thm 7.1: The density $\pi(\theta)$ that maximizes $Ent(\pi)$ subject to $E[T_j(\theta)] = t_j, j = 1, \cdots, p$ takes the p-parameter exponential family form $\pi(\theta) \propto \exp\left\{\sum_{i=1}^p \lambda_i T_i(\theta)\right\}, \forall\theta \in \Theta$, where $\lambda_1, \cdots, \lambda_p$ are determined by the constraints.

Ex 7.7: Find π which maximize $Ent(\pi)$ on $\Theta \in \mathbb{R}$ subject to $\int_0^\infty \pi(\theta)d\theta = 1; E(\underbrace{\theta}_{T_1(\theta)}) = \int_0^\infty \theta\pi(\theta)d\theta = \mu; E(\underbrace{(\theta-\mu)^2}_{T_2(\theta)}) = \int_0^\infty(\theta-\mu)^2\pi(\theta)d\theta = \sigma^2$. Then, $\pi(\theta) \propto \exp(\lambda_1\theta + \lambda_2(\theta-\mu)^2) = \exp(-\frac{(\theta-\mu)^2}{2\sigma^2})$, by the constraints.

6.4 Predictive Distribution

Def 7.6: If $(x_i)_{i=1}^{n+1} \overset{iid}{\sim} f(x,\theta)$ with prior π, the posterior predictive distribution is:

$$f(x_{n+1}|x) = \int f(x_{n+1},\theta)\pi(\theta|x)d\theta, x = (x_1, \cdots, x_n)$$

Ex 7.8: Suppose $Y \sim Poisson(\lambda), \pi(\lambda) = \Gamma(\alpha,\beta)$.

$$P(y) = \frac{P(y|\lambda)\pi(\lambda)}{\underbrace{P(\lambda|y)}_{\sim\Gamma(\alpha+y,\beta+1) \text{ by Ex7.2}}}, \because P(\lambda|y) = \frac{P(y|\lambda)\pi(\lambda)}{P(y)}$$
$$= \frac{\frac{\exp(-\lambda)\lambda^y}{y!}\frac{\beta^\alpha\exp(-\beta\lambda)\lambda^{\alpha-1}}{\Gamma(\alpha)}}{\frac{(\beta+1)^{\alpha+y}\lambda^{\alpha+y-1}\exp(-(\beta+1)\lambda)}{\Gamma(\alpha+y)}}$$
$$= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y \sim NegBinom(\alpha,\beta)$$
$$\implies NegBin(y|\alpha,\beta) = \int_\lambda Poisson(y|\lambda)\Gamma(\lambda|\alpha,\beta)d\lambda$$
$$\implies P(y_{n+1}|y) = \int Poisson(y_{n+1}|\lambda)\Gamma\left(\lambda|\alpha+\sum_{i=1}^n y_i, \beta+n\right) \sim NegBin\left(y_{n+1}|\alpha+\sum_{i=1}^n y_i, \beta+n\right)$$

Ex 7.9: Consider $(X_i)_{i=1}^n \overset{iid}{\sim} N(\theta, \sigma^2)$ (σ known) and $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$\begin{aligned}
\pi(\theta|x) \propto \pi(\theta)p(x|\theta) &\propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right) \\
&\propto \exp\left(\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu)^2 - \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\theta)^2\right]\right) \\
\text{Completing squares...} &\propto \exp\left(-\frac{1}{2\sigma_n^2}(\theta-\mu_n)^2\right), \text{ where } \mu_n = \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^n x_i/\sigma^2}{\sigma_0^{-2}+n\sigma^{-2}}; \sigma_n^{-2} = \sigma_0^{-2} + n\sigma^{-2} \\
&\implies \theta|X \sim N(\mu_n, \sigma_n^2) \text{ and } X_{n+1} \sim \theta + N(0,\sigma^2) \\
&\implies X_{n+1}|X \sim \theta|X + N(0,\sigma^2) \sim N(\mu_n, \sigma_n^2+\sigma^2)
\end{aligned}$$

- If $\sigma_0 < \sigma$, then we put more weight on the prior, and counts more on the single observation. If $\sigma_0 = \sigma$, then prior has same weight as a single extra observation. Then,

7 Hierarchical Models

7.1 Hierarchical Structures

Consider a model $Y_i \sim Binom(n_i, \theta_i)$. Then, $(\theta_1, \cdots, \theta_k)$ has the following structures:
Def 8.1: (a) **Identical parameters:** $\theta_1 = \cdots = \theta_k = \theta_*$ (pooled data), ignoring the structure of the problem;
(b) **Independent parameters:** $\theta_i \perp \theta_j, \forall i \neq j$ (unpooled data), result from each unit can be analysed independently;
(c) **Exchangeable/Symmetric parameters:** $\theta_i' s$ are assumed to be similar, i.e. θ_i is not a prior better than θ_j.
- For any permutation $\sigma, (\theta_1, \cdots, \theta_I) \overset{d}{=} (\theta_{\sigma(1)}, \cdots, \theta_{\sigma(I)})$

Ex 8.1: The Binomial example under different approaches:
(a) Non-Bayesian + all θ_i equal: $Y_i \sim Binom(n_i, \theta) \implies \hat{\theta}^{MLE} = \frac{\sum y_i}{\sum n_i}$;
(b) Non-Bayesian + indep θ_i: $Y_i \sim Binom(n_i, \theta_i) \implies \hat{\theta}_i^{MLE} = \frac{y_i}{n_i}$; (c) Bayesian + all θ_i equal: $Y_i \sim Binom(n_i, \theta), \theta \sim Beta(\alpha, \beta) \implies \hat{\theta}^{MLE} = \frac{\sum y_i + \alpha}{\sum n_i + \alpha + \beta}$;
(d) Bayesian + indep θ_i: $Y_i \sim Binom(n_i, \theta_i), \theta_i \overset{iid}{\sim} Beta(\alpha, \beta) \implies \hat{\theta}_i^{MLE} = \frac{y_i+\alpha}{n_i+\alpha+\beta}$.

7.1.1 Approximate Empirical Bayesian Approach

Def 8.2: The **Approximate Empirical Bayesian Approach**is: $\theta_i \overset{iid}{\sim} Beta(\alpha, \beta)$, want to choose (α, β),
(a) calculate $r_i = \frac{y_i}{n_i}$, then find the sample mean μ_r and variance σ_r^2 for the $r_i' s$;
(b) solve for $\hat{\alpha}, \hat{\beta}$ s.t. $Beta(\hat{\alpha}, \hat{\beta})$ has mean $= \mu_r$ and variance $= \sigma_r^2$;
(c) use $\theta_i \overset{iid}{\sim} Beta(\hat{\alpha}, \hat{\beta})$ to obtain posterior $P(\theta_i|\hat{\alpha}, \hat{\beta}, y_i)$.

• problem: (i) over-confidence on the chosen $\hat{\alpha}, \hat{\beta}$, because we used the data twice at step (b) and (c); (ii) ignoring uncertainty because we only consider one choice of $\hat{\alpha}, \hat{\beta}$.

7.1.2 Hierarchical Model: Go Bayesian twice

Def 8.3: Assume joint probability model for (θ, ϕ), where ϕ is the hyperprior
- *Level1:* $Y_i|\theta_i \overset{iid}{\sim} Likelihhod(Y_i, \theta_i)$;
- *Level2:* $\theta_i|\phi \overset{iid}{\sim} Prior(\theta, \phi)$ ($\theta_i' s$ are not indep, but they are cond indep given ϕ);

- *Level3:* $\phi \sim HyperPrior(\phi)$. $\implies$ **Joint piror distribution:** $p(\theta, \phi) = p(\theta|\phi)p(\phi)$;
$\implies$ **Joint posterior distribution:** $p(\theta, \phi|y) \propto p(y|\theta, \phi)p(\phi, \theta) = p(y|\theta)p(\theta|\phi)p(\phi)$.

Ex 8.2: $(\alpha, \beta) \sim P(\alpha, \beta), \theta_i|\alpha, \beta \overset{iid}{\sim} Beta(\alpha, \beta), Y_i|\theta_i \overset{iid}{\sim} Bin(n_i, \theta_i)$

$$
\begin{aligned}
p(\theta, \alpha, \beta|y) &\propto p(y|\theta)p(\theta|\alpha, \beta)p(\alpha, \beta) \\
&= \left(\prod_{i=1}^{I} p(y_i|\theta_i)\right)\left(\prod_{i=1}^{I} p(\theta_i|\alpha, \beta)\right)p(\alpha, \beta) \\
&\propto \left(\prod_{i=1}^{I} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}\right)\left(\prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\right)p(\alpha, \beta) \\
\implies p(\theta_i|\alpha, \beta, y_i) &\propto \left(\prod_{i=1}^{I} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}\right) \sim Beta(\alpha+y_i, \beta+n_i-y_i) \\
p(\alpha, \beta|y) &\propto p(\alpha, \beta)p(y|\alpha, \beta) \\
&\propto p(\alpha, \beta)\int_{\Theta} p(y|\theta)p(\theta|\alpha, \beta)d\theta \\
&= p(\alpha, \beta)\int_{\Theta} \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}d\theta \\
&= p(\alpha, \beta)\prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\beta+n_i-y_i)\Gamma(\alpha+y_i)}{\Gamma(\alpha+\beta+n_i)}
\end{aligned}
$$

Q: How to draw fro $p(\theta, \alpha, \beta|y)$:
(a) draw $\phi \sim p(\phi|y)$;
(b) draw $\theta \sim p(\theta|\phi, y)$;
(c) Draw predictive values $\tilde{y}$ from $p(y|\theta)$.

Thm 8.1: (De Finetti's) A large (infinite) sequence of Bernoulli RVs is exchangeable iff it is a mixture of iid Bernoulli RVs.
- $p(\theta) = \int[\prod_{i=1}^{T} \pi(\theta_i|\phi)]g(\phi)d\phi$.

7.2 Normal Data Example

Omitted, See P100-114
Model:
(a) Level1: $(X_{i,1}, \cdots, X_{i,n_i}) \overset{iid}{\sim} N(\theta_i, \sigma^2)$, for $i = 1, \cdots, J$ and *sigma* known. - $\theta = (\theta_1, \cdots, \theta_J)$ is the vector of means; - $\bar{X}_{j,\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$; $\bar{X}_{\cdot\cdot} = \frac{1}{\sum_{j=1}^{J} n_j}\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ij}$ (b) Level2: $\theta_i|\phi, \tau^2 \overset{iid}{\sim} N(\phi, \tau^2)$;
(c) Level3: $\psi = (\phi, \tau^2) \sim g(\psi)$.

7.2.1 Step1: Joint Prior $p(\theta, \psi|X)$

$p(\theta, \psi|x) \propto g(\psi)p(\theta|\psi)p(x|\theta)$, where $\psi = (\phi, \tau^2)$

$$\propto g(\psi)\prod_{j=1}^{J} \underbrace{N(\theta_j|\phi, \tau^2)}_{\theta_j \sim N(\phi,\tau^2)} \prod_{j=1}^{J} p(\vec{x_j}|\theta_j), \text{ where } \vec{x_j} = (x_{j,1}, \cdots, x_{j,n_j})$$

$$\propto g(\psi)\prod_{j=1}^{J} N(\theta_j|\phi, \tau^2)\prod_{j=1}^{J} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}(x_{ji} - \theta_j)^2\right)$$

$$\propto g(\psi)\prod_{j=1}^{J} N(\theta_j|\phi, \tau^2)\prod_{j=1}^{J} \exp\left(-\frac{1}{2\sigma^2}n_j(\theta_j^2 - 2\theta_j\bar{x}_{j\cdot})\right), \text{ only leave the term with } \theta_j$$

$$\propto g(\psi)\prod_{j=1}^{J} N(\theta_j|\phi, \tau^2)\prod_{j=1}^{J} \exp\left(-\frac{1}{2\sigma^2/n_j}(\theta_j - \bar{x}_{j\cdot})^2\right)$$

$$\propto g(\psi)\prod_{j=1}^{J} N(\theta_j|\phi, \tau^2)\prod_{j=1}^{J} N(\bar{x}_{j\cdot}|\theta_j, \sigma_j^2) \text{ , where } \sigma_j^2 = \frac{\sigma^2}{n_j}$$

7.2.2 Step2: Conditional Posterior (of θ)

$p(\theta|\psi, X)$ Condition on ψ, the θ_j are iid, $p(\theta|\psi, x) = \frac{p(\theta,\psi|x)}{g(\psi|x)} \propto \frac{p(\theta,\psi|x)}{g(\psi)}$

$$p(\theta_j|\psi, \vec{x_j}) \propto N(\theta_j|\phi, \tau^2)N(\bar{x}_{j\cdot}|\theta_j, \sigma_j^2)$$

$$\propto \exp\left(-\frac{1}{2\tau^2}(\theta_j - \phi)^2 - \frac{1}{2\sigma_j^2}(\bar{x}_{j\cdot} - \theta_j)^2\right)$$

$$\propto \exp\left(-\frac{1}{2V_j}(\theta_j - \hat{\theta}_j)^2\right), \text{ where } V_j = \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1} \text{ and } \hat{\theta}_j = V_j\left(\frac{\bar{x}_{j\cdot}}{\sigma_j^2} + \frac{\phi}{\tau^2}\right)$$

$$\implies \theta_j|\psi, \vec{x_j} \sim N(\hat{\theta}_j, V_j)$$

7.2.3 Step3: Marginal Posterior (of ψ) $p(\psi|X)$

Note $p(\psi|x) \propto g(\psi)p(x|\psi)$. This does not have a closed form. So, we assume ϕ is uniform condition on τ^2, i.e. $g(\psi) = p(\phi|\tau^2)p(\tau^2) \propto p(\tau^2)$.
Now, need to calculate the marginal posterior $p(x|\psi)$: (since we want $p(\psi|x)$ upto proportionality constant such that all terms that do not depend on ψ are ignored, we want $p(x|\psi)$ up to terms with no ψ)

$$\because \psi \text{ is fixed, the variables } \left((\theta_j, \vec{x}_j)_{j=1}^J\right) \perp$$

$$p(x|\psi) = \prod_{j=1}^{J} p(\vec{x}_j|\psi) = \prod_{j=1}^{J} \int p(\theta_j|\psi)p(\vec{x}_j|\theta_j)d\theta_j$$

Recall the identity $\sum_{i=1}^{n_j}(x_{ji} - \theta_j)^2 = n_j(\bar{x}_{j\cdot} - \theta_j)^2 + n_j s_j^2$, where $s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}\bar{x}_{j\cdot})^2$, then:

$$p(\vec{x}_j|\theta_j) \propto \exp\left(-\frac{1}{2\sigma_j^2}[(\bar{x}_{j\cdot} - \theta_j)^2 + s_j^2]\right) \propto \exp\left(-\frac{1}{2\sigma_j^2}(\bar{x}_{j\cdot} - \theta_j)^2\right)$$

$$\propto p(\bar{x}_{j\cdot}|\psi), \text{ up to terms indep. of } \psi$$

Since $\theta_j \sim N(\phi, \tau^2)$ and $\bar{x}_{j\cdot}|\theta_j \overset{d}{=} \theta_j + \sigma_j Z$ ($Z \sim N(0, 1)$), we have that $\bar{x}_{j\cdot}|\psi \sim N(\phi, \sigma_j^2 + \tau^2) - - - (*)$ (see proof below).
Hence we deduce that: $p(\vec{x}_j|\theta_j) \propto \exp\left(-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{x}_{j\cdot} - \phi)^2\right) \sim N(\bar{x}_{j\cdot}|\phi, \sigma_j^2 + \tau^2)$.

Proof of (*): consider the following model: $X|\theta \sim N(\theta, \sigma^2), \theta \sim N(\phi, \tau^2)$, want to show that $X \sim N(\phi, \tau^2 + \sigma^2)$.

$$
\begin{aligned}
p(x) &= \int p(x|\theta)\pi(\theta)d\theta \\
&\propto \int \frac{1}{\tau\sigma} \exp\left\{-\frac{1}{2}[\sigma^{-2}(x-\theta)^2 + \tau^{-2}(\theta-\phi)^2]\right\} d\theta \\
&\propto \int \exp\left\{-\frac{1}{2}\left[\theta^2(\sigma^{-2}+\tau^{-2}) - 2\theta(\sigma^{-2}x + \tau^{-2}\phi) + \frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{\sigma^{-2}+\tau^{-2}} - \frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{\sigma^{-2}+\tau^{-2}} + \sigma^{-2}x^2 + \tau^{-2}\phi^2\right]\right\} d\theta \\
&= \int \exp\left\{-\frac{1}{2}\left[(\sigma^{-2}+\tau^{-2})\left(\theta^2 - 2\theta\frac{(\sigma^{-2}x+\tau^{-2}\phi)}{\sigma^{-2}+\tau^{-2}} + \frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{(\sigma^{-2}+\tau^{-2})^2}\right)\right]\right\} d\theta \\
&\quad \times \exp\left\{-\frac{1}{2}\left[-\frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{\sigma^{-2}+\tau^{-2}} + \sigma^{-2}x^2 + \tau^{-2}\phi^2\right]\right\} \\
&= \int \exp\left\{-\frac{1}{2}\left[(\sigma^{-2}+\tau^{-2})\left(\theta^2 - 2\theta\frac{(\sigma^{-2}x+\tau^{-2}\phi)}{\sigma^{-2}+\tau^{-2}} + \frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{(\sigma^{-2}+\tau^{-2})^2}\right)\right]\right\} d\theta \\
&\quad \times \exp\left\{-\frac{1}{2}\left[x^2\left(\sigma^{-2} - \frac{\sigma^{-4}}{(\sigma^{-2}+\tau^{-2})^2}\right) - x\left(\frac{2\sigma^{-2}\tau^{-2}\phi}{\sigma^{-2}+\tau^{-2}}\right) + \tau^{-2}\phi^2\right]\right\} \\
&= \int \exp\left\{-\frac{1}{2}\left[(\sigma^{-2}+\tau^{-2})\left(\theta^2 - 2\theta\frac{(\sigma^{-2}x+\tau^{-2}\phi)}{\sigma^{-2}+\tau^{-2}} + \frac{(\sigma^{-2}x+\tau^{-2}\phi)^2}{(\sigma^{-2}+\tau^{-2})^2}\right)\right]\right\} d\theta \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\frac{\sigma^{-2}\tau^{-2}}{\sigma^{-2}+\tau^{-2}}(x^2 - 2x\phi) + \tau^{-2}\phi^2\right]\right\} \\
&= \int \exp\left\{-\frac{1}{2}\left[(\sigma^{-2}+\tau^{-2})\left(\theta - \frac{(\sigma^{-2}x+\tau^{-2}\phi)}{\sigma^{-2}+\tau^{-2}}\right)^2\right]\right\} d\theta \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\frac{\sigma^{-2}\tau^{-2}}{\sigma^{-2}+\tau^{-2}}[(x-\phi)^2 - \phi^2] + \tau^{-2}\phi^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\frac{\sigma^{-2}\tau^{-2}}{\sigma^{-2}+\tau^{-2}}(x-\phi)^2\right\} \cdot \int \exp\left\{(\sigma^{-2}+\tau^{-2})\left(\theta - \frac{(\sigma^{-2}x+\tau^{-2}\phi)}{\sigma^{-2}+\tau^{-2}}\right)^2\right\} d\theta \\
&\propto \exp\left\{-\frac{1}{2}\frac{\sigma^{-2}\tau^{-2}}{\sigma^{-2}+\tau^{-2}}(x-\phi)^2\right\}
\end{aligned}
$$

7.2.4 Extra Step: Compute $p(\theta|X)$

Assuming $p(\tau) \propto \tau^{-a} (a \geq 0)$,
There are 2 approaches: (1) $p(\theta|x) = \int \underbrace{p(\theta|x,\psi)}_{step2}\underbrace{p(\psi|x)}_{step3} d\psi$; (2) $p(\theta|x) = \int \underbrace{p(\theta,\psi|x)}_{step1} d\psi$.

8 Decision Theory

Framework:
• Parameter space: Θ;
• Model: $X|\theta \sim f(x,\theta)$ for some parametric family $\{f(x,\theta); \theta \in \Theta\}$
• Action/Decison space: $\mathcal{A}$, e.g. selecting a hypothesis $\mathcal{A} = \{H_0, H_1\}$, or estimating a function $g(\theta), \mathcal{A} = g(\Theta)$;
• Loss function: $L : \Theta \times \mathcal{A} \to \mathbb{R}+$ (given action $a \in \mathcal{A}$, if true parameter is θ, we incur loss $L(\theta, a)$);
• Set of Decision Rule: $\mathcal{D} \subseteq \{\delta : \mathcal{X} \to \mathcal{A}\}$ (rule δ specify which action we take given observation $x \in \mathcal{X}$).

8.1 Frequentiest Risk & Admissibility

Def 9.1: For given rule $\delta \in \mathcal{D}$ and $\theta \in \Theta$, the **Frequentist Risk**is the expected loss assuming the true parameter is θ:

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x, \theta) dx$$

Ex 9.1: Consider $\delta(x)$ is an estimation of $\theta \in \mathbb{R}^k$ and $L(\theta, a) = \|a - \theta\|^2$, then $R(\theta, \delta) = E_\theta[\|\delta(x) - \theta\|^2]$. Testing $\theta \in H_0$ again $\theta \in H_1$, i.e. $\mathcal{A} = \{0, 1\}$. Hence the loss gives:

$$L(\theta, a) = \begin{cases} 1, \theta \in H_0, a = 1 \\ 1, \theta \in H_1, a = 0 \\ 0, \text{otherwise} \end{cases} \implies R(\theta, \delta(x)) = \begin{cases} P_\theta(\delta(x) = 0) \text{ if } \theta \in H_1 (\text{type I error}) \\ P_\theta(\delta(x) = 1) \text{ if } \theta \in H_0 (\text{type II error}) \end{cases}$$

Def 9.2: δ_2 **strctly dominates**δ_1 if $R(\theta, \delta_1) \geq R(\theta, \delta_2), \forall \theta \in \Theta$, with $R(\theta, \delta_1) > R(\theta, \delta_2)$ for at least some θ.

Def 9.3: A procedure δ_1 is inadmissible if $\exists \delta_2$ such that δ_2 strictly dominates δ_1.

Ex 9.2: $X \sim Unif(0, \theta)$. $\mathcal{D}$ = estimators of the form $\hat{\theta}(x) = ax$, family indexed by a. Show that $a = \frac{3}{2}$ is a necessary condition for $\hat{\theta}$ to be admissible for quadratic loss:

$$R(\theta, \hat{\theta}) = \int_0^\theta (ax - \theta)^2 \frac{1}{\theta} dx = (\frac{a^2}{3} - a + 1)\theta^2 \implies \min \text{ at } a = \frac{3}{2}$$

So, all $\hat{\theta} = ax$ are inadmissible if $a \neq \frac{3}{2}$, although $a = \frac{3}{2}$ does not guarantee the admissibility of $\hat{\theta}$.

8.2 Minimax Rule and Bayes Rule

Def 9.4: δ^* is a **minimax rule**if $\sup_\theta R(\theta, \delta^*) \leq \sup_\theta R(\theta, \delta), \forall \delta \in \mathcal{D}$.
- It minimizes the maximum risk $\delta^* = \arg\min_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$

Def 9.5: The **Bayes integrated risk**is $r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta) d\theta$, where priot $\pi(\theta)$ is given.

Def 9.6: δ is a **Bayes rule**wrt π if it minimizes the Bayes risk: $r(\pi, \delta) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta') =: m_\pi$.

Def 9.7: If the infimum is not attained, for $\epsilon > 0$, if δ_ϵ is such that $r(\pi, \delta_\epsilon) < m_\pi + \epsilon$, then δ_ϵ is said $\epsilon-$Bayes wrt π.

Def 9.8: δ is **Exxtended Bayes** if $\forall \epsilon > 0, \exists \pi : \delta$ is $\epsilon-$Bayes wrt pi.

Def 9.9: The **Expected posterior loss** of a rule δ wrt π is: $\Lambda(x) = \int_\Theta L(\theta, \delta(x))\pi(\theta|x) d\theta$.

Prop 9.1: A Bayes rule minimizes the expected postrior loss.
Proof of Prop 9.1:

$$\begin{aligned} r(\pi, \delta) &= \int R(\theta, \delta)\pi(\theta) d\theta = \int \int L(\theta, \delta(x)) f(\theta, x)\pi(\theta) dx d\theta \\ &= \int \int L(\theta, \delta(x))\pi(\theta|x) \underbrace{h(x)}_{\text{normalization constant}} dx d\theta \\ &= \int h(x) \underbrace{\left(\int L(\theta, \delta(x))\pi(\theta|x) d\theta \right)}_{\Lambda(x)} dx \end{aligned}$$

$$\implies \delta^* = \arg\min_{\delta \in \mathcal{D}} r(\pi, \delta) = \arg\min_{\delta \in \mathcal{D}} \Lambda(x)$$

Def 9.10: δ^* is $\pi-$**admissible**iff $\forall\delta : R(\theta,\delta) \leq R(\theta,\delta^*), \forall\theta \in \Theta$ and $\pi(\{\theta : R(\theta,\delta) < R(\theta,\delta^*)\}) = 0$, i.e. $\pi(\{\theta : R(\theta,\delta) = R(\theta,\delta^*)\}) = 1$ ($R(\theta,\delta)$ and $R(\theta,\delta^*)$ are also everywhere equal wrt π).

Thm 9.1: A Bayes rule wrt π is $\pi-$admissible.
Proof of Thm 9.1: By contradiction. Suppose Bayes rule δ^* is not $\pi-$admissible, $\exists\delta : \pi(\underbrace{\{\theta : R(\theta,\delta) < R(\theta,\delta^*)\}}_{\mathcal{A}_\delta \subseteq \Theta}) >$ 0. Hence:

$$r(\pi,\delta) - r(\pi,\delta^*) = \int_{\mathcal{A}_\delta} \underbrace{[R(\theta,\delta) - R(\theta,\delta^*)]}_{<0}\pi(\theta)d\theta + \int_{\mathcal{A}_\delta^C} \underbrace{[R(\theta,\delta) - R(\theta,\delta^*)]}_{=0 \text{ on } \mathcal{A}_\delta^C}\pi(\theta)d\theta < 0$$

$$\implies \text{Contradicts } \delta^* \text{ is Bayes, because } r(\pi,\delta^*) = \inf_{\delta\in\mathcal{D}} r(\pi,\delta)$$

Prop 9.2: (Bayes Rules and Admissibility) Let δ^π be a Bayes rule wrt π, then:
(a) If δ^π is unique and $r(\pi,\delta^\pi) < \infty$, then is it admissible;
(b) If $\forall\delta, \theta \to R(\theta,\delta)$ is continuous, $r(\pi,\delta^\pi) < \infty$ and π has a positive density wrt Lebesgue, then δ^π is admissible.
Proof of Prop 9.2:
(a): By contrapositive. If δ^π is not admissible, then $\exists\delta : R(\theta,\delta) \leq R(\theta,\delta^\pi), \forall\theta \in \Theta$. This implies, $r(\pi,\delta) \leq r(\pi,\delta^\pi)$. Note that $\delta^\pi = \arg\min_{\delta\in\mathcal{D}} r(\pi,\delta)$. Then no longer unique.
(b) By contradiction. If δ^π is not admissible, then $\exists\delta : R(\theta,\delta) \leq R(\theta,\delta^\pi), \forall\theta \in \Theta$. Since $\theta \to R(\theta,\delta) - R(\theta,\delta^\pi)$ is continuous, $\mathcal{A}_\delta \neq \emptyset$ and it is an open set. Then, $\pi(\mathcal{A}_\delta) > 0$, a contradiction to def 9.10.

8.3 Finite Decision Problem

Def 9.11: A decision problem is **finite**when $\Theta = (\theta_1, \cdots, \theta_k)$ is finite.

Def 9.12: The **risk set**$S \subseteq \mathbb{R}^k$ is the set of points $\{(R(\theta_1,\delta), \cdots, R(\theta_k,\delta)), \delta \in \mathcal{D}\}$.

Lemma 9.1: S is a convex set.
Proof of Lemma 9.1: Let $\delta_1, \delta_2 \in \mathcal{D}$ be 2 rules, take $\alpha \in (0,1)$. Then define a randomized rule $\delta'(x) = \begin{cases}\delta_1(x) \text{ with prob } \alpha \\ \delta_2(x) \text{ with prob } 1-\alpha\end{cases} \implies R(\theta,\delta') = \alpha R(\theta,\delta_1) + (1-\alpha)R(\theta,\delta_2)$.

Ex 9.2: some graph examples omitted, see P128-130.

8.4 Bayes & Minimax

Thm 9.2: If δ is a Bayes rule wrt π with $r(\pi,\delta) = c$ and δ_0 is a rule s.t. $\max_\theta R(\theta,\delta_0) = c$, then δ_0 is minimax. Proof pf Thm 9.2: By contradiction. If for some other rule $\delta', \max_\theta R(\theta,\delta') = c - \epsilon, \forall\epsilon > 0$. (So δ_0 is not minimax.)

$$\begin{aligned} r(\pi,\delta') &= \int R(\theta,\delta')\pi(\theta)d\theta \\ &\leq \int (c-\epsilon)\pi(\theta)d\theta \\ &= (c-\epsilon) < r(\pi,\delta) \implies \delta \text{ is not Bayes for } \pi, \text{ contradiction!} \end{aligned}$$

■

Thm 9.3: If δ is Bayes for π with the property that $R(\theta,\delta) \perp \theta$, then δ is minimax.

Proof pf Thm 9.3: By contradiction. Let $R(\theta,\delta) = C, \forall\theta$, then $r(\pi,\delta) = \int R(\theta,\delta)\pi(\theta)d\theta = C$. If $\exists\delta'$ with $\max_\theta R(\theta,\delta') = c - \epsilon, \forall\epsilon > 0$, then $r(\pi,\delta') \leq c - \epsilon < r(\pi,\delta)$. Again, δ is not Bayes, contradiction!

Lemma 9.2: The Bayes estimator with constant risk is minimax.

Ex 9.3: Minimax estimator for quadratic loss. Consider $X \sim Bin(n, \theta), \pi(\theta) \sim Beta(\alpha, \beta)$. Bayes estimator is $\hat{\theta} = \frac{\alpha+x}{\alpha+\beta+n}$.

$$\begin{aligned} R(\hat{\theta}, \theta) &= E_\theta[(\hat{\theta} - \theta)^2] = MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta}) \\ &= \left[\theta - E_{\hat{\theta}}\left(\frac{\alpha + x}{\alpha + \beta + n}\right)\right]^2 + Var\left(\frac{\alpha + x}{\alpha + \beta + n}\right) \\ &= \left[\theta - \left(\frac{\alpha + n\theta}{\alpha + \beta + n}\right)\right]^2 + \frac{n\theta(1-\theta)}{(\alpha + \beta + n)^2} \\ &= \frac{[(\alpha + \beta)\theta - \alpha]^2 + n\theta(1-\theta)}{[\alpha + \beta + n]^2} \end{aligned}$$

If $\alpha = \beta = \frac{\sqrt{n}}{2}$, this is constant in θ. Then the minimax estimator for quadratic loss is: $\frac{x+\sqrt{n}/2}{n+\sqrt{n}}$.

8.5 Point Estimation

Def 9.13: The **Zero-One Loss**is: $L(\hat{\theta}, \theta) = \begin{cases} a, & \text{if } |\hat{\theta} - \theta| > b \\ 0, & \text{else} \end{cases}$.

• Expected posterior loss:

$$\begin{aligned} \Lambda(x) &= \int L(\hat{\theta}, \theta)\pi(\theta|x)d\theta \\ &= a\int_{\hat{\theta}+b}^{\infty} \pi(\theta|x)d\theta + a\int_{-\infty}^{\hat{\theta}-b} \pi(\theta|x)d\theta \\ &\propto 1 - \int_{\hat{\theta}-b}^{\hat{\theta}+b} \pi(\theta|x)d\theta \end{aligned}$$

- Then the Bayes rule is: $\hat{\theta}(x) = \arg\max_\theta \int_{\hat{\theta}-b}^{\hat{\theta}+b} \pi(\theta|x)d\theta$. (when $b \to 0$, $\hat{\theta}$ is the posterior mode.)

Def 9.14: The **Absolute-Error Loss**is: $L(\hat{\theta}, \theta) = k|\hat{\theta} - \theta|$.

• Expected posterior loss:

$$\begin{aligned} \Lambda(x) &= \int L(\hat{\theta}, \theta)\pi(\theta|x)d\theta \\ &\propto \int |\hat{\theta} - \theta|\pi(\theta|x)d\theta = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)\pi(\theta|x)d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})\pi(\theta|x)d\theta \\ \implies \frac{\partial}{\partial\hat{\theta}}\Lambda(x) &= \int_{-\infty}^{\hat{\theta}} \pi(\theta|x)d\theta - \int_{\hat{\theta}}^{\infty} \pi(\theta|x)d\theta = 0 \implies \hat{\theta} = \text{median of } \pi(\theta|x) \end{aligned}$$

Def 9.15: The **Quadratic Loss**is: $L(\hat{\theta}, \theta) = k(\hat{\theta} - \theta)^2$.

• Expected posterior loss:

$$\begin{aligned} \Lambda(x) &= E_\theta[(\hat{\theta} - \theta)^2] = E_\theta[(\hat{\theta} - \mu_x + \mu_x - \theta)^2] \\ &= (\hat{\theta} - \mu_x)^2 + 2(\hat{\theta} - \mu_x)\underbrace{E_{\theta|x}[\mu_x - \theta]}_{=0} + E_{\theta|x}[(\mu_x - \theta)^2] \\ &= (\hat{\theta} - \mu_x)^2 + Var_{\theta|x}(\theta) \\ \implies \Lambda(x) &\text{ is minimized when } \hat{\theta} = \mu_x, \text{ the posterior mean} \end{aligned}$$

9 Bayesian Hypothesis Tests

9.1 Testing Simple Hypothesis with Loss Functions

Suppose $(X_i)_{i=1}^n \overset{iid}{\sim} f(x;\theta)$ and want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

The decision rule is: $\delta_C(x) = \begin{cases} H_1, x \in C \\ H_0, x \notin C \end{cases}$, where C is the critical region.

The loss function is: $L_s(\theta, \delta_C(x)) = \begin{cases} a1_{x\in C}, \text{ if } \theta = \theta_0 \\ b1_{x\notin C}, \text{ if } \theta = \theta_1 \end{cases}$.

The Tpye I error is: $\alpha = P(\text{reject } H_0|H_0)$; The Tpye II error is: $\beta = P(\text{accept } H_0|H_1)$; power $= 1 - \beta$.
The Risk function for δ_C is:

$$\begin{cases} R(\theta_0, \delta_C) & = \int L_s(\theta_0, \delta_C(x)) f(x;\theta_0)dx = \int a1_{x\in C} f(x;\theta_0)dx = a\alpha \\ R(\theta_1, \delta_C) & = \int L_s(\theta_1, \delta_C(x)) f(x;\theta_1)dx = \int b1_{x\notin C} f(x;\theta_1)dx = b\beta \end{cases}$$

The Bayes Risk is:

$$r(\pi, \delta_C) = \sum_{\theta\in\{\theta_0,\theta_1\}} R(\theta;\delta_C)\pi(\theta), \text{ where } \pi(\theta) = \begin{cases} p_0, \theta = \theta_0 \\ p_1, \theta = \theta_1 \end{cases}$$

9.2 Bayes Test

The Bayes test chooses the critical region C to minimize the Bayes risk.

Lemma 12.1: (Neyman-Pearson lemma) The best test of size α of H_0 vs H_1 is a likelihood ratio test with critical region $C_{NP} = \left\{x : \frac{L(\theta_1;x)}{L(\theta_0;x)} \geq A\right\}$, for some constant $A > 0$ chosen s.t. $P(X \in C_{NP}|H_0) = \alpha$.
- The best Bayes test has the highest power = 1 - β.

Thm 12.1: The critical region for the Bayes test is the critical region for a LRT with $A = \frac{p_0 a}{p_1 b}$.
- Every LRT is a Bayes test for some p_0, p_1.
Proof of Thm 12.1: The Bayes test minimizes:

$$\begin{aligned} r(\pi, \delta_C) &= p_0 a\alpha + p_1 b\beta = p_0 aP(X \in C|H_0) + p_1 bP(X \notin C|H_1) \\ &= p_0 a \int_C L(\theta_0;x)dx + p_1 b \int_{C^c} L(\theta_1;x)dx \\ &= p_0 a \int_C L(\theta_0;x)dx + p_1 b \left(1 - \int_C L(\theta_1;x)dx\right) \\ &= p_1 b + \int_C \left[p_0 aL(\theta_0;x) - p_1 bL(\theta_1;x)\right] dx \end{aligned}$$

Choose C s.t. $x \in C$ iff $p_0 aL(\theta_0;x) - p_1 bL(\theta_1;x) \leq 0$, hence $\frac{L(\theta_1;x)}{L(\theta_0;x)} \geq \frac{p_0 a}{p_1 b} = A$. ■

Ex 12.1: Consider $(X_i)_{i=1}^n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 known. Want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$ using the classical test and the Bayes test, where $\mu_1 > \mu_0$. Then the critical region for the LRT is:

$$\begin{aligned}
\frac{L(\mu_1; x)}{L(\mu_0; x)} \geq A &\iff \log L(\mu_1; x) - \log L(\mu_0; x) \geq \log A \\
LHS &= -\frac{n(\bar{x} - \mu_1)^2}{2\sigma^2} - \left(-\frac{n(\bar{x} - \mu_0)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2\sigma^2}(\bar{x}^2 - 2\bar{x}\mu_1 + \mu_1^2 - \bar{x}^2 + 2\bar{x}\mu_0 - \mu_0^2) \\
&= \frac{n}{\sigma^2}(\bar{x} - \frac{1}{2}(\mu_1 + \mu_0))(\mu_1 - \mu_0) \geq \log A \\
\implies \bar{x} &\geq \frac{\sigma^2 \log A}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_1 + \mu_0)
\end{aligned}$$

Suppose $\mu_0 = 0, \mu_1 = 1, \sigma^2 = 1, n = 4$. In a classical test, $\alpha = 0.05$ is pre-determined. Then,

$$\beta = P\left(\frac{\bar{x} - \mu_0}{sd(\bar{x})} < 1.96 \mid \mu = \mu_1, \frac{\sigma^2}{n} = \frac{1}{4}\right) = P\left(\bar{x} < 1.96 \times \sqrt{\frac{1}{4}} \mid \mu = \mu_1, \frac{\sigma^2}{n} = \frac{1}{4}\right) \approx 0.484$$

For Bayes test, $A = \frac{p_0 a}{p_1 b}$, where $p_0 = \frac{1}{4}, p_1 = \frac{3}{4}, a = 2, b = 1$. Then,

$$\begin{aligned}
\bar{x} &\geq \frac{1}{4}\log(\frac{2}{3}) + \frac{1}{2} = 0.399 \\
\alpha &= P(\bar{x} \geq 0.399 | \mu = \mu_0 = 0, \frac{\sigma^2}{n} = \frac{1}{4}) \approx 0.212 \\
\beta &= P(\bar{x} < 0.399 | \mu = \mu_1 = 1, \frac{\sigma^2}{n} = \frac{1}{4}) \approx 0.363
\end{aligned}$$

Compared to the classical test, the Bayes test has higher Type I error but lower Type II error, hence higher power.

Def 12.1: The **Maximum A Posteriori (MAP)**test chooses the hypothesis with the highest posterior probability $P(X_i|y)$.

Lemma 12.2: The MAP estimator is the Bayes estimator under the zero-one loss function: $L(\theta, \delta) = \begin{cases} 1, \text{ if } 1_{\theta \in \Theta_1} \neq \delta \\ 0, \text{ o/w} \end{cases}$, where $\delta = \{0, 1\}$.

Proof of Lemma 12.2: The risk under zero-one loss is:

$$\begin{aligned}
r(\pi, \delta) &= \int L(\theta_0, \delta)\pi(\theta_0|x) + L(\theta_1, \delta)\pi(\theta_1|x)dx \\
&= \int L(\theta_0, \delta)\frac{f(x; \theta_0)\pi(\theta_0)}{f_X(x)} + L(\theta_1, \delta)\frac{f(x; \theta_1)\pi(\theta_1)}{f_X(x)}dx \\
&= \int L(\theta_0, \delta)\frac{f(x|H_0)p(H_0)}{f_X(x)} + L(\theta_1, \delta)\frac{f(x|H_1)p(H_1)}{f_X(x)}dx \\
&= \int L(\theta_0, \delta)P(H_0|x) + L(\theta_1, \delta)P(H_1|x)dx
\end{aligned}$$

Suppose $P(H_0|x) > P(H_1|x)$, then we choose δ' s.t. $L(\theta_0, \delta') = 0$ and $L(\theta_1, \delta') = 1$. So, $\delta' = 0$, which has the highest posterior probability.

9.3 Composite Hypothesis

Suppose model is $X|\theta \sim f_\theta(\cdot), \theta \in \Theta \sim \Pi$. Consider the testing problem: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1, \Theta_0 \cap \Theta_1 = \emptyset$.

Def 12.2: A hypothesis $H_j : \theta \in \Theta_j$ is **simple**iff Θ_j is a singleton. A hypothesis $H_j : \theta \in \Theta_j$ is **composite**iff Θ_j is NOT a singleton.
- If H_j is simple, we cannot use $p_0 = \pi(\theta \in \Theta_0) = 0$ if Π has a density.

Def 12.3: Computing the posteriors: $\Pi(\Theta_0|X) = \frac{p_0 m_0(X)}{p_0 m_0(X)+(1-p_0)m_1(X)}$, where $m_0(X) = \int_{\Theta_0} f(X;\theta)\Pi_0(d\theta)$ and $m_1(X) = \int_{\Theta_1} f(X;\theta)\Pi_1(d\theta)$ are the marginal likelihoods under H_0, H_1 respectively.

Def 12.4: The **Bayes Factor**of H_0 over H_1 is: $B_{0/1}(X) = \frac{m_0(X)}{m_1(X)}$, i.e. the ratio of marginal likelihoods.

- For $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, the Bayes factor is $B_{0/1}(X) = \frac{f_{\theta_0}(X)}{\int_{\Theta_1} f_\theta(X)\pi(\theta)d\theta}$.
- The Bayes test associated to te zero-one loss function verifies: $\delta(X) = 1 \iff B_{0/1}(X) < \frac{1-p_0}{p_0}$.
- Interpretation: by Adrian Raftey

$P(H_0\|x)$	B	$2\log B$	evidence for H_0
< 0.5	< 1	< 0	negative (support H_1)
0.5-0.75	1-3	0-2	barely worth mentioning
0.75-0.92	3-12	2-5	positive
0.92-0.99	12-150	5-10	strong
> 0.99	> 150	> 10	very strong

- $2\log B$ is reported because it is on the same scale as the familiar deviance and LRT statistic.

9.3.1 Test in the case of a simple hypothesis

Construct prior as a mixture between a prior on $\Theta_0 = \{\theta_0\}$ and a prior on Θ_1.

$$\pi(d\theta) = p_0\delta_{\theta_0}(d\theta) + (1-p_0)\pi_1(d\theta), \text{ where } \pi_1 \text{ is a prob dist. on } \Theta_1$$

Then the posterior is:

$$\pi(\{\theta_0\}|X) = \frac{p_0 f(X,\theta_0)}{p_0 f(X,\theta_0) + (1-p_0)\int_{\Theta_1} f(X,\theta)\pi_1(\theta)d\theta}$$

The Bayes Test rejects H_0 when:

$$\begin{aligned}\delta(X) = 1 &\iff p_0 f(X,\theta_0) < (1-p_0)\int_{\Theta_1} f(X,\theta)\pi_1(\theta)d\theta \\ &\iff \underbrace{\frac{f(X,\theta_0)}{\int_{\Theta_1} f(X,\theta)\pi_1(\theta)d\theta}}_{\text{Bayes factor}} < \frac{1-p_0}{p_0}\end{aligned}$$

Ex 12.2: Consider $X \sim Binom(n, \theta)$, want to test $H_0 : \{\theta = \frac{1}{2}\}$ vs $H_1 : \{\theta \neq \frac{1}{2}\}$. The priors are $p_0 = \pi(H_0) = \frac{1}{2}$ and $\pi_1 \sim Unif(0, 1)$ ($\pi_1(1/2) = 0$). Compute the posterior density:

$$\begin{aligned}
\pi(H_0|x) = \pi(\{1/2\}|x) &= \frac{p_0 f(x; \theta = \frac{1}{2})}{p_0 f(x; \theta = \frac{1}{2}) + (1 - p_0) \int_0^1 f_\theta(x) d\theta} \\
&= \frac{\binom{n}{x} 2^{-n}}{\binom{n}{x} 2^{-n} + \frac{1}{n+1}} \\
\text{where, } \int_0^1 f_\theta(x) d\theta &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta = \int_0^1 \frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} d\theta \\
&= \frac{1}{n+1} \int_0^1 \underbrace{\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^x (1-\theta)^{n-x}}_{\sim Gamma(x+1, n-x)} d\theta = \frac{1}{n+1}
\end{aligned}$$

9.3.2 Point Composite Hypothesis

Consider $(X_i)_{i=1}^n \sim N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$. Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0, \Theta_0 = \{0\} \times \mathbb{R}_+$.
Same approach as for simple hypothesis, but $\pi_0 = \delta_0 \otimes \Pi_\sigma$, where Π_σ is the prior distribution on σ with density π_σ. π_1 has density on $\mathbb{R} \times \mathbb{R}_+$, e.g.:

$$\pi_1(\mu, \sigma^2) \propto \frac{\phi\left(\frac{\mu - \mu_0}{\sigma\tau}\right)}{\sigma\tau} \times (\sigma^2)^{-a-1} e^{-b/\sigma^2} \frac{b^a}{\Gamma(a)} \implies \mu|\sigma \sim N(\mu_0, \sigma^2\tau^2), \sigma^2 \sim IGamma(a, b)$$

- this is a hierarchical prior under H_1.

• Computing the posterior: $\pi(\Theta_0|x) = \frac{p_0 m_0(x)}{p_0 m_0(x) + (1-p_0) m_1(x)}$, where $m_0(X) = \int_0^\infty f(x|\mu = 0, \sigma^2) \pi_\sigma(\sigma) d\sigma$ and $m_1(x) = \int_{\mathbb{R}} \int_0^\infty f(x|\mu, \sigma^2) \pi(\mu|\sigma^2) \pi_\sigma(\sigma) d\sigma$ are the marginal likelihood under H_0, H_1 respectively.

9.3.3 Framework for Bayesian Model Selection

Models/Hypothesis for data $x : \mathcal{M}_1, \cdots, \mathcal{M}_k$, under model $\mathcal{M}_i$;
- $X \sim f_i(x; \theta_i)$ where θ_i unknown parameter;
- Prior for θ_i is $\pi_i(\theta_i)$, and prior probability is $P(\mathcal{M}_i)$ ($= \frac{1}{k}$ in the uniform prior case);
- Marginal density of X is: $m_i(x) = m(x|\mathcal{M}_i) = \int f_i(x; \theta_i) \pi_i(\theta_i) d\theta_i$.

(1) Posterior density: $\pi_i(\theta_i|x) = \frac{f_i(x|\theta_i)\pi_i(\theta_i)}{m(x|\mathcal{M}_i)}$;
(2) Bayes factor of $\mathcal{M}_j$ to $\mathcal{M}_i$ is: $B_{ji} = \frac{m(x|\mathcal{M}_j)}{m(x|\mathcal{M}_i)}$;
(3) Posterior: $\Pi(\mathcal{M}_i|x) = \frac{\Pi(\mathcal{M}_i) m(x|\mathcal{M}_i)}{\sum_j \Pi(\mathcal{M}_j) m(x|\mathcal{M}_j)} = \left[\sum_j \frac{\Pi(\mathcal{M}_j)}{\Pi(\mathcal{M}_i)} B_{ji}\right]^{-1}$.

Ex 12.3: Model $X \sim Binom(n, \theta), \pi(\theta) = 30\theta(1-\theta)^4$, want to test $H_0 : \theta \leq 0.2$ vs $H_1 : \theta > 0.2$.
Given the prior and the tests, we have the prior probabilities: $p_0 = p(H_0) = \pi(\theta \in \Theta_0) = \int_0^{0.2} 30\theta(1-\theta)^4 d\theta \approx 0.345$. Then $p_1 = p(H_1) \approx 1 - 0.345$. Thus, the prior distribution for H_0 and H_1 are $\pi(\theta|H_0) = \frac{30\theta(1-\theta)^4}{p(H_0)}, 0 < \theta \leq 0.2$ and $\pi(\theta|H_1) = \frac{30\theta(1-\theta)^4}{p(H_1)}, 0.2 < \theta < 1$. The marginal likelihoods for H_0 and H_1 are: (For $n = 5, X = 0$).

$$
\begin{aligned}
m_0(x) &= \int_{\Theta_0} f(x;\theta)\pi(\theta|H_0)d\theta = \binom{n}{x}\int_0^{0.2}\theta^x(1-\theta)^{n-x}\frac{30\theta(1-\theta)^4}{p(H_0)}d\theta \\
&= \binom{5}{0}\int_0^{0.2}\frac{30\theta(1-\theta)^9}{p(H_0)}d\theta \approx 0.185/0.345 = 0.536 \\
m_1(x) &= \int_{\Theta_0} f(x;\theta)\pi(\theta|H_1)d\theta = \binom{n}{x}\int_{0.2}^{1}\theta^x(1-\theta)^{n-x}\frac{30\theta(1-\theta)^4}{p(H_1)}d\theta \\
&= \binom{5}{0}\int_{0.2}^{1}\frac{30\theta(1-\theta)^9}{p(H_1)}d\theta \approx 0.134
\end{aligned}
$$

Hence, the Bayes factor $B = \frac{m_0(x)}{m_1(x)} \approx \frac{0.536}{0.134} = 4$.
The posteriors are:

$$
\begin{aligned}
\Pi(H_0|x) &= \frac{m_0(x)p_0}{m_0(x)p_0 + m_1(x)(1-p_0)} \approx \frac{0.536\times 0.345}{0.536\times 0.345 + 0.134\times(1-0.345)} \approx 0.678 \\
\Pi(H_1|x) &= \frac{m_1(x)(1-p_0)}{m_0(x)p_0 + m_1(x)(1-p_0)} = 1 - \Pi(H_0|x) \approx 0.322
\end{aligned}
$$

Ex 12.4: Model $(X_i)_{i=1}^n \overset{iid}{\sim} N(\theta,\sigma^2)$, with σ^2 known. Want to test $H_0 : \theta = 0$ vs $H_1 : \theta|H_1 \sim N(\mu,\tau^2)$. The marginalized likelihood for H_0 and H_1 are:

$$
\begin{aligned}
m_0 &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right) \\
m_1 &= (2\pi\sigma^2)^{-n/2}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right)\times(2\pi\tau^2)^{-1/2}\exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right)d\theta \\
&= (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2}\left[\frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\bar{x})^2\right]\right] \\
\implies B &= \frac{m_1}{m_0} = \left(1+\frac{n\tau^2}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\frac{n\bar{x}^2}{\sigma^2} - \frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2\right]\right\} \\
&= \left(1+\frac{1}{\rho^2}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\frac{(t-\rho\eta)^2}{1+\rho^2} - \eta^2\right]\right\}, \text{ where } t = \frac{\sqrt{n}\bar{x}}{\sigma^2}, \eta = \frac{-\mu}{\tau}, \rho = \frac{\sigma}{\tau\sqrt{n}}
\end{aligned}
$$

- A problem of choosing the prior: if we take a diffuse prior under H_1, i.e. $\tau \to \infty$, then $B \to \infty$ whatever x, giving overwhelming support for H_0.
• Bayes factor and Bayes tests (under zero-one loss) are not defined when the priors are improper.