Advanced Statistical Inference

STAT 561 - Advanced Statistical Inference

Dr. Bing Li

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2013

Contents

Part One

1. Basic Ideas in Bayesian Anaylysis

Mathematical Preparation

Monday January 9

1. Product σ-Field

   $(\Omega_1, \mathscr{F}_1, \mu_1), (\Omega_2, \mathscr{F}_2, \mu_2)$ are two measure spaces. The goal is to construct a σ-field on $\Omega_1 x \Omega_2$.

   Let $\mathscr{A} = \{AxB : A \in \mathscr{F}_1, B \in \mathscr{F}_2\}$.

   The σ-field generated $\mathscr{A}$ is called the product σ-field, written as $\mathscr{F}_1 x \mathscr{F}_2$, that is $\sigma(\mathscr{A})$. This is NOT a cartesian product, which would be $\{(A,B) : A \in \mathscr{F}_1, B \in \mathscr{F}_2\}$.
2. Proctuct Measure

   Let $E \in \mathscr{F}_1 x \mathscr{F}_2$. Let $E_2(\omega_1) = \{\omega_2 : (\omega_1, \omega_2) \in E\}$ and similarly, $E_1(\omega_2) = \{\omega_1 : (\omega_1, \omega_2) \in E\}$.

   It is true (in Billingsly) that

   Theorem 1.0.1 — Number Unknown. If $E \in \mathscr{F}_1 x \mathscr{F}_2$ then $E_1(\omega_2) \in \mathscr{F}_1$ for all $\omega_2 \in \Omega_2$. Similarly, $E_2(\omega_1) \in \mathscr{F}_2$ for all $\omega_1 \in \Omega_1$.

   If $f : \Omega_1 x \Omega_2 \to \mathbb{R}$ measurable $\mathscr{F}_1 x \mathscr{F}_2 \setminus \mathscr{R}$. Then for each $\omega_1 \in \Omega_1$,

   $$f(\omega_1, \cdot) ⓜ \mathscr{F}_2 \setminus \mathscr{R} \text{ for each } \omega_2 \in \Omega_2.$$

$$f(\cdot,\omega_2)\text{ⓜ}\mathscr{F}_1 \setminus \mathscr{R}$$

Now, for each $E \in \mathscr{F}_1 x \mathscr{F}_2$ consider

$$f_{1,E} : \Omega_1 \to \mathscr{R}, \omega_1 \mapsto \mu_2(E_2,(\omega_2))$$

It can be shown that $f_{1,E}$ is uniformly measurable $\mathscr{F}_1 \setminus \mathscr{R}$ for all E.

Proof. Outline.

- Show that if $\mathscr{L} = \{E : f_{1,E}\text{ⓜ}\mathscr{F}_1 \setminus \mathscr{R}\}$ then $\mathscr{L}$ is a λ-system.
- Let $\mathscr{P} = \{AxB : A \in \mathscr{F}_1, B \in \mathscr{F}_2\}$ then it is a π-system.
  Furthermore, if $E = AxB$,

$$E_2(\omega_1) = \begin{cases} B & \omega_1 \in A \\ \emptyset & \omega_1 \notin A \end{cases}$$

  So, $\mu_2(E_2(\omega_1)) = \begin{cases} \mu_2(B) & \omega_1 \in A \\ \emptyset & \omega_1 \notin A \end{cases} = I_A(\omega_1)\mu(B) = f_{1,E}$
  So, $f_{1,E}\text{ⓜ}\mathscr{F}_1$.
  Thus $\mathscr{P} \subseteq \mathscr{L}$.
- By $\pi - \lambda$ Theorem, $\mathscr{F}_1 x \mathscr{F}_2 \subseteq \mathscr{L}$.

Similarly, $f_{2,E}\text{ⓜ}\mathscr{F}_2 \setminus \mathscr{R}$.
We can now define two set functions,

$$\pi'(E) = \int f_{1,E} d\mu_1$$

$$\pi''(E) = \int f_{2,E} d\mu_2$$

Again using $\pi - \lambda$ Theorem, it can be shown that, π', π'' are both measure and if μ_1, μ_2 are σ-finite, then

$$\pi' = \pi'' \text{ on } \mathscr{F}_1 x \mathscr{F}_2$$

Note that here, $\mathscr{P}$ equals $\mathscr{A}$ used at begining of notes.
We did not have a measure in $\mathscr{F}_1 x \mathscr{F}_2$. Now we have π', π'' both measures on $\mathscr{F}_1 x \mathscr{F}_2$, they are the same. We call this measure the product meaure, written as $\mu_1 x \mu_2$.
Note that $(\Omega_1 x \Omega_2, \mathscr{F}_1 x \mathscr{F}_2, \mu_1 x \mu_2)$ is called product measure space. ■

3. Tonelli's Theorem

   $(\Omega_1, \mathscr{F}_1, \mu_1), (\Omega_2, \mathscr{F}_2, \mu_2)$ are two σ-finite measure spaces.

   $(\Omega_1 x \Omega_2, \mathscr{F}_1 x \mathscr{F}_2, \mu_1 x \mu_2)$ is the product measure space.

   Suppose we have $f : \Omega_1 x \Omega_2 \to \mathbb{R}\text{ⓜ}\mathscr{F}_1 x \mathscr{F}_2 \setminus \mathscr{R}$. Where $f \geq 0$ and

$$\int f d(\mu_1 x \mu_2) = \int \left[\int (f(\cdot,\omega_2) d\mu_1) \right] d\mu_2$$

4. Fubini's Theorem

The conclusion of Tonelli's Theorem still holds if f is NOT nonnegative, but if f is integrable μ_2. (integrable - integral of absolute value of function is finite)

Wednesday January 11

5. Conditional Probability

This is a special application of Radon- Nikodgm Theorem. We know that

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

We may define $P(A|\mathscr{G})$ when $\mathscr{G} \subseteq \mathscr{F}$ as sub-σ-field. We defined this intuitively in elementary probability course (definition above), but we ar enot going to define it generally.

Now let $A \in \mathscr{F}$ and $\mathscr{G} \subset \mathscr{F}$ be a σ-field. Consider the set function

$$\nu : \mathscr{G} \to \mathbb{R}, G \mapsto P(AG)$$

It can be easily shown that ν is a measure on $\mathscr{G}$. Consider another set function,

$$\mu : \mathscr{G} \to \mathbb{R}, G \mapsto P(G)$$

So μ is nothting but P restricted on $\mathscr{G}$.

It's easy to show that $\nu << \mu$.

$$\mu(G) = 0 \Rightarrow P(G) = 0 \Rightarrow P(AG) = 0 \Rightarrow \nu(G) = 0$$

By Radon-Nikodgm Theorem, there exists a δ such that

$$\nu(G) = \int_G \delta d\mu \quad \forall G \in \mathscr{G}$$

δ is called R-N Derivative, written as

$$\delta = \frac{d\nu}{d\mu}$$

and is similar in to $\frac{P(AG)}{P(G)}$, but it's more general.

δ is called the conditional probability of A given $\mathscr{G}$. To distinguish it form P(A|B), where B is a set, we use $P(A||\mathscr{G})$, where $\mathscr{G}$ is a σ-field. By construction,

(a) δ is measurable $\mathscr{G}$
(b) $\int_G \delta dp = P(AG) \quad \forall G \in \mathscr{G}$

Note that, by RNT, δ is unique with probability 1. Any δ' satisfying (a) and (b) has $\delta' = \delta a.e.P$. So, we say that δ is a version of conditional probability.

So, δ is a version of $P(A||\mathscr{G})$ if and only if (a) and (b) are satisfied. We may define $P(A||\mathscr{G})$ either by RNT or (a) and (b).

Properties of Conditional Probability

It behaves like probability, but since it is a function, unique up to a.e. P, these properties have to be qualified by a.s. P.

(a) $P(\emptyset||\mathscr{G}) = 0, P(\Omega||\mathscr{G}) = 1$ a.s. P
(b) $0 \leq P(A||\mathscr{G}) \leq 1$ a.s. P
(c) If $A_1, A_2, \ldots$ are disjoint members of $\mathscr{F}$ then $P(\bigcup_n A_n||\mathscr{G}) = \sum_n P(A_n||\mathscr{G})$ a.s. P

Let's consider the special case where $\mathscr{G}$ is a σ-field generated by some random element, T (i.e. $\mathscr{G} = \sigma(T)$). More specifically, for some measurable space $(\Omega_T, \mathscr{F}_T)$ where

$$T : \Omega \to \Omega_T \text{ⓜ} \mathscr{F} \setminus \mathscr{F}_T \quad \mathscr{G} = T^{-1}(\mathscr{F}_T)$$

Here, we write

$$\begin{aligned} P(A||\mathscr{G}) &= P(A||\sigma(T)) \\ &= P(A||T^{-1}(\mathscr{F}_T)) \\ &= P(A||T) \end{aligned}$$

The following theorem makes checking that something is a conditional probability easier. In principle, we have to check $\int_G \delta dp = P(AG) \quad \forall G \in \mathscr{G}$.

Theorem 1.0.2 — 33.1 in Billingsly. Let $\mathscr{P}$ be a *pi*-system generating $\mathscr{G}$ and suppose that Ω is a countable union of sets in $\mathscr{P}$. An integrable function, f, is a version of $P(A||\mathscr{G})$ if

(a) f is measurable $\mathscr{G}$
(b) $\int_G f dp = P(AG) \quad \forall G \in \mathscr{P}$

6. Conditional Distribution

Let there be probability space $(\Omega, \mathscr{F}, P)$, measurable space $(\Omega_X, \mathscr{F}_X)$, and a random element, $X : \Omega \to \Omega_X \text{ⓜ} \mathscr{F} \setminus \mathscr{F}_X$. Also, let $\mathscr{G} \subseteq \mathscr{F}$ be a sub σ-field.

We are going to define conditional distribution of X given G. Under very mild conditions there is a function

$$f : \mathscr{F}_X x \Omega \to \mathbb{R}$$

such that for each $A \in \mathscr{F}_X$, $f(A, \cdot)$ is a version of

$$P(X \in A||\mathscr{G}) = P(X^{-1}(A)||\mathscr{G})$$

and, for each $\omega \in \Omega$, $f(\cdot, \omega)$ is a probability measure on $(\Omega_X, \mathscr{F}_X)$.

The only condition for this existance is $(\Omega_X, \mathscr{F}_X)$ must be a Borel Space, that is $\mathscr{F}_X$ is Borel σ-field. This should always be the case for our purposes.

7. Conditional Expectation

Let us have the same probability space, measurable space, random element, and sub σ-field as defined before, but here with $\bar{\mathbb{R}}$.

We want to define conditional expectaiton of X given $\mathscr{G}$.

First, assume $X \geq 0$. Consider a set function,

$$\nu : \mathcal{G} \to \mathbb{R}, G \mapsto \int_G X dP$$

It can be easily shown that ν is a measure.
Let μ again be $\mathcal{G} \to \mathbb{R}, G \mapsto P(G)$. Then $\nu << \mu$. By RNT, $\delta = \frac{d\nu}{d\mu}$ is well defined. This is defined to be conditional expectation of X given $\mathcal{G}$, written as

$$E(X||\mathcal{G})$$

Suppose $X \ngeq 0$, but integrable P. Recall that $X = X^+ - X^-$. Since $X^+, X^- \geq 0$, then both $E(X^+||\mathcal{G}), E(X^-||\mathcal{G})$ are defined by RNT. We define,

$$E(X||\mathcal{G}) = E(X^+||\mathcal{G}) - E(X^-||\mathcal{G})$$

Friday January 13
As in the case of $P(A||\mathcal{G})$, the equivalent conditions for $d : \Omega \to \mathbb{R}$ is a version of $E(X||\mathcal{G})$.
(a) δ measurable $\mathcal{G}$
(b) $\int_G \delta dP = \int X dP \quad \forall G \in \mathcal{G}$
INSERT PHOTO FROM BOARD - "Mesh"
The value of δ in each thick outlined cell is the average (with respect to P measure) of $X(\omega)$ over the subcells (thin outlined) in thick cells.
We see from this definition that if $A \in \mathcal{F}$, $X = I_A$ then the second condition becomes

$$\int_G \delta dP = \int_G I_A dP = P(A \bigcap G)$$

So, $E(I_A||\mathcal{G}) = P(A||\mathcal{G})$.
Properties of Conditional Expectations

Theorem 1.0.3 34.2 in Billingsly Suppose that X, Y, X_n are integrable P.

If X = a a.e. P, then $E(X||\mathcal{G})$ a.s. P

(b) If $a, b \in \mathbb{R}$ then

$$E(aX + bY||\mathcal{G}) = a(E(X||\mathcal{G})) + b(E(X||\mathcal{G}))a.s.P$$

(c) If $X \leq Y$ a.s. P then

$$E(X||\mathcal{G}) \leq E(Y||\mathcal{G})$$

(d) $|E(X||\mathcal{G})| \leq E(|X|||\mathcal{G})$ a.s. P (in fact this is true for all convex functions).
(e) If $X_n \to X$ a.s. P, $|X_n| \leq Y$, and Y integrable P, then

$$E(X_n||\mathcal{G}) \to E(X||\mathcal{G})a.s.P$$

Proof. Found in Billingsly. ■

Theorem 1.0.4 — 34.4 in Billingsly. If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subset \mathcal{F}$ and X integrable P, then

$$E(E(X||\mathcal{G}_2)||\mathcal{G}_1) = E(X||\mathcal{G}_1)$$

This is called the Law of Iterative Conditional Expectation.

Theorem 1.0.5 — 34.3 in Billingsly. If X measurable $\mathscr{G}$, Y ⓜ $\mathscr{F}$, then

$$E(XY||\mathscr{G}) = XE(Y||\mathscr{G}) a.s. P$$

Other Properties

(a) X, Y are random elements such that XY integrable P.

(b) If $\mathscr{G} \subseteq \mathscr{F}$ is the sub σ-field, then

$$E(XE(Y||\mathscr{G})) = E(E(X||\mathscr{G})Y) = E(E(X||\mathscr{G})E(Y||\mathscr{G}))$$

Conditional expectation is a self-adjoint operation.

Proof. "Wire Theorem"

$$\begin{aligned} E(XE(Y||\mathscr{G})) &= E(E(XE(Y||\mathscr{G})||\mathscr{G})) \\ &= E(E(Y||\mathscr{G})E(X||\mathscr{G})) \\ &= E(E(E(X||\mathscr{G})Y||\mathscr{G})) \\ &= E(E(X||\mathscr{G})Y) \end{aligned}$$

■

8. Conditional Distribution of a Random Element Given Another Random Element

Here we have the typical probability space, measurable spaces for X and Y.
Let there be a function,
$h : \mathscr{F}_X x \Omega_Y \to \mathbb{R}$
This function is called the conditional distribution of X given Y if

$$\tilde{h}(A, \omega) = h(A, Y(\omega))$$

We say that $\tilde{h} : \mathscr{F}_X x \Omega \to \mathbb{R}$ is the condiitional distribution of X given $\mathscr{G} = Y^{-1}(\mathscr{F}_Y$.
That is,

(a) For each $A \in \mathscr{F}_X$

$$\tilde{h}(A, Y(\cdot)) = P(X^{-1}(A)||Y^{-1}(\mathscr{F}_Y))$$

(b) For each $\omega \in \Omega$

$$\tilde{h}(\cdot, Y(\omega)) = P_{X|Y}(A|y)$$

9. Conditional Density of One Random Element Given Another Random Element

Suppose probability space and σ-finite measure spaces for X and Y.
Here our relevant function is

$$g : \Omega_X x \Omega_Y$$

which is the conditional density of X given Y if for all $A \in \mathscr{F}_X$,

$$\int_A g(x,y) d\mu_X(x) = P_{X|Y}(A|y)$$

In the following special case, g ahs an explicit formula.
$(\Omega, \mathscr{F}, P)$
$(\Omega_X, \mathscr{F}_X, \mu_X)$
$(\Omega_Y, \mathscr{F}_Y, \mu_Y)$
$(\Omega_X x \Omega_Y, \mathscr{F}_X x \mathscr{F}_Y, \mu_X x \mu_Y)$
$(X,Y) : \Omega \to \Omega_X x \Omega_Y$ ⓜ $\mathscr{F} \setminus \mathscr{F}_X x \mathscr{F}_Y$

Let $P_X = PX^{-1}, P_Y = PY^{-1}, P_{XY} = P(XY)^{-1}$.
Assume $P_X << \mu_X, P_Y << \mu_Y, P_{XY} << \mu_X x \mu_Y$.

$$f_X = \frac{dP_X}{d\mu_X}$$

$$f_Y = \frac{dP_Y}{d\mu_Y}$$

$$f_{XY} = \frac{dP_{XY}}{d(\mu_X x \mu_Y)}$$

Let

$$f_{X|Y} = \begin{array}{ll} \frac{f_{XY}}{f_Y} & if f_Y \neq 0 \\ 0 & f_Y = 0 \end{array}$$

$$f_{Y|X} = \begin{array}{ll} \frac{f_{XY}}{f_X} & if f_X \neq 0 \\ 0 & f_X = 0 \end{array}$$

Then it is easy to show that each is indeed the conditional density of their respective elements (first given second).
Wednesday January 18

Claim: $g(x,y)$ is the conditional density.

Proof. Want to show that for all $A \in \mathscr{F}_X$,

$$\int_A g(x,y) d\mu_x(x) = P_{X|Y}(A|y)$$

Which means that

$$\int_A g(x, y(\omega)) d\mu_x(x) = P_{X|Y}(X^{-1}(A)|\sigma(y))$$

This is true if for all $G' \in \sigma(y)$

$$\int_{G'} \int_A g(x, y(\omega)) d\mu_X(x) dP(\omega) = P(X^{-1}(A) \bigcap G')$$

But note that

$$\begin{array}{l} G' \in \sigma(y) \\ \Leftrightarrow G' \in Y^{-1}(\mathscr{F}_Y) \\ G' = Y^{-1}(G) \text{ for some } G \in \mathscr{F}_Y \end{array}$$

So we want to check that

$$\int_{Y^{-1}(G)} \int_A g(x,y(\omega)) d\mu_X(x) dP(\omega) = P(X^{-1}(A) \bigcap Y^{-1}(G))$$

$$\begin{aligned}
\int_{Y^{-1}(G)} \int_A g(x,y(\omega)) d\mu_X(x) dP(\omega) &= \int_G \int_A g(x,y) d\mu_X(x) dP_Y(y) \\
&= \int_G \int_A \frac{f_{XY}(x,y)}{f_Y(y)} d\mu_X(x) [f_Y(y)] d\mu_Y(y) \\
&= \int_G \int_A f_{XY}(x,y) d\mu_X(x) d\mu_Y(y) \\
&= \int_{GxA} f_{XY}(x,y) d(\mu_X x \mu_Y)(x,y) \\
&= P_{XY}(GXA) \\
&= P \circ (X,Y)^{-1}(AxG) \\
&= P(X \in A, Y \in G) \\
&= P(\omega : \omega \in X^{-1}(A) \& \omega \in Y^{-1}(G)) \\
&= P(X^{-1}(A) \bigcap Y^{-1}(G))
\end{aligned}$$

■

1.1 Frequentist & Bayesian Settings

We have our probability space $(\Omega, \mathscr{F}, P)$. We also have some data,

$$(\Omega_X, \mathscr{F}_X, \mu_X)$$

$$X : \Omega \to \Omega_X \text{ⓜ} \mathscr{F} / \mathscr{F}_X$$

Here, usually Ω_X is a $\mathbb{R}^m$.

Typically we have

$$X = (X_1, \ldots, X_n)$$

and possibly,

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$$

We could say that these data are independent and identically distributed (iid) random vectors of dimension p. In this case m = np.

The goal of statical inference is to estimate.

$$P_X = PX^{-1} = P_0$$

The ?? distribution of X.

There are two schools of thought

1. Frequentist Approach - assume a family of distributions, $\mathscr{P}$, where $\mathscr{P} << \mu_x$. Usually we assume that $\mathscr{P}$ is a parametric family, $\mathscr{P} = \{P_\theta : \theta \in \Omega_\theta \subseteq \mathbb{R}^p\}$. We assume that $P_0 \in \mathscr{P}$, that is there exisgts $\theta_0 \in \Omega_\theta$ such that $P_\theta = P_0$. The goal is to estimate P_0.
2. Bayesian Approach - here we assume the data is generated by the conditional distribtuion $P_{X|\theta}$. We observe X, then determine what is the best estimate of the random θ.

1.2 Prior Posterior & Likelihood

Here let there be probability space $(\Omega, \mathscr{F}, P)$; σ-finite measurable spaces $(\Omega_X, \mathscr{F}_X, \mu_X)$, $(\Omega_\theta, \mathscr{F}_\theta, \mu_\theta)$. Together,

$$(\Omega_X x \Omega_\theta, \mathscr{F}_X x \mathscr{F}_\theta, \mu_X x \mu_\theta)$$

Also, a random element,

$$(X, \theta) : \Omega \to \Omega_X x \Omega_\theta \text{ⓜ} \mathscr{F} \setminus \mathscr{F}_X x \mathscr{F}_\theta$$

$P_X = P \circ X^{-1} \leftarrow$ marginal distribution of X
$P_\theta = P \circ \theta^{-1} \leftarrow$ prior distribution
$P_{X,\theta} = P \circ (X, \theta)^{-1} \leftarrow$ joint distribution of X and θ
$P_{X|\theta}(A|\theta) : \mathscr{F}_X x \Omega_\theta \to \mathbb{R}$. Likeliehood distribution
$P_{\theta|X}(G|x) : \mathscr{F}_\theta x \Omega_X \to \mathbb{R}$. Posterior distribution

Note in the following the first inequalities are **assumed**.

$P_X << \mu_X \Rightarrow f_X = \frac{dP_X}{d\mu_X}$ Marginal Density
$P_\theta << \mu_\theta \Rightarrow \pi_\theta = \frac{dP_\theta}{d\mu_\theta}$ Prior Density
$P_(X, \theta) << \mu_X x \mu_\theta \Rightarrow f_{X,\theta}(x, \theta) = \frac{dP_{X,\theta}}{d(\mu_x x \mu_\theta)}$ Joint Density

FINISH FROM PHOTO
One way to estimate θ is by maximizing $\pi_{\theta|X}(\theta|x)$. Want to do so with value that is most likely to happen (given the data).

$$\pi_{\theta|X}(\theta|x) = P_{\theta|X}(\theta = \theta|x)$$

By construction,

$$\begin{aligned} \pi_{\theta|X} &= \frac{f_X \theta}{f_X} \\ &= \frac{f_{X|\theta} \pi_\theta}{\int_{\Omega_\theta} f_{X|\theta} \pi_\theta d\mu_t heta} \end{aligned}$$

1.3 Conditional Independence and Frquentist/Bayesian Sufficiency

Independence

Two random elements are said to be independent if for all $A' \in \sigma(X), G' \in \sigma(\theta)$ we have

$$P(A' \bigcap G') = P(A')P(G')$$

This statement can also be expressed in $(\Omega_X x \Omega_\theta, \mathscr{F}_X x \mathscr{F}_\theta, P_X x P_\theta)$ as follows.

Since $A' \in \sigma(X) = X^{-1}(\mathscr{F}_X), A' = X^{-1}(A)$ for some $A \in \mathscr{F}_X$. So $G' = \Theta^{-1}(G).G \in \mathscr{F}_\Theta$.

$$\begin{aligned}
P(A' \bigcap G') &= P(X^{-1}(A) \bigcap \Theta^{-1}(G)) \\
&= P(\{\omega : \omega \in X^{-1}(A) \bigcap \Theta^{-1}(G)\}) \\
&= P(\{\omega : \omega \in X^{-1}(A) \& \omega \in \Theta^{-1}(G)\}) \\
&= P(\{\omega : X(\omega) \in A, \Theta(\omega) \in G\}) \\
&= P(\{\omega : (X(\omega), \Theta(\omega)) \in AxG\}) \\
&= P(\{\omega : (X, \Theta)(\omega) \in AxG\}) \\
&= P(\{\omega : \omega \in (X, \Theta)^{-1} AxG\}) \\
&= [P \circ (X, \Theta)^{-1}](AxG) \\
&= P_{X,\Theta}(AxG)
\end{aligned}$$

Also note that

$$P(A') = P(X^{-1}(A)) = P_X(A)$$

$$P(G') = P_\Theta(G)$$

So with independence, (and for $A \in \mathscr{F}_X, G \in \mathscr{F}_\Theta$)

$$P_{X,\Theta}(AxG) = P_X(A) P_\Theta(G)$$

But we know that this implies that $P_{X,\Theta}$ is the product measure $P_X x P_\Theta$.

Conditional Independence

Now, given sub σ-field $\mathscr{G} \in \mathscr{F}$ we want to define $X \& \Theta$ conditionally independent given $\mathscr{G}$.

Definition 1.3.1 We say that $X \& \Theta$ are conditionally independent given $\mathscr{G}$ (i.e. $X \perp\!\!\!\perp \Theta | \mathscr{G}$) if for all $A' \in \sigma(X), G' \in \sigma(\Theta)$ we have

$$P[A' \cap G' || \mathscr{G}] = P[A' || \mathscr{G}] P[G' || \mathscr{G}] a.s.P$$

Equivalently for all $A \in \mathscr{F}_X, G \in \mathscr{F}_\Theta$,

$$P[X^{-1}(A) \cap \Theta^{-1}(G) || \mathscr{G}] = P[X^{-1}(A) || \mathscr{G}] P[\Theta^{-1}(G) || \mathscr{G}]$$

Equivalently,

$$P_{X,\Theta|\mathscr{G}}(AxG|\mathscr{G}) = P_{X|\mathscr{G}}(A|\mathscr{G}) P_{\Theta|\mathscr{G}}(G|\mathscr{G})$$

Equivalent Condition for Conditional Independence

Theorem 1.3.1 — 1.1 in Notes. The following statements are equivalent.

1. $X \perp\!\!\!\perp \Theta | \mathscr{G}$
2. $P(X^{-1}(A) || \Theta, \mathscr{G}) = P(X^{-1}(A) | \mathscr{G}) a.s.P \quad \forall A \in \sigma(X)$
3. $P(\Theta^{-1}(G) || X, \mathscr{G}) = P(\Theta^{-1}(G) || \mathscr{G}) a.s.P \quad \forall G \in \sigma(\Theta)$

Proof. It suffies to proof that $1 \Leftrightarrow 2$.

$1 \Rightarrow 2$. We know that for all $A \in \mathscr{F}_X, G \in \mathscr{F}_\Theta$ that

$$P[X^{-1}(A) \cap \Theta^{-1}(G)||\mathscr{G}] = P[X^{-1}(A)||\mathscr{G}]P[\Theta^{-1}(G)||\mathscr{G}]$$

Want that for all $A \in \mathscr{F}_X$ that $P(X^{-1}(A)||\Theta,\mathscr{G}) = P(X^{-1}(A)|\mathscr{G})$.

$$\begin{aligned} P(X^{-1}(A)||\Theta,\mathscr{G}) &\equiv P\left(X^{-1}(A)||\sigma(\sigma(\Theta) \cup \mathscr{G})\right) \\ &= P(\ldots||\sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G})) \end{aligned}$$

So it suffices to show that

$$P(X^{-1}(A)||\sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G})) = P(X^{-1}(A)||\mathscr{G})$$

From the definition given we want to show that the above statement is true. which is so that the for all $B \in \sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G})$,

$$\int_B P(X^{-1}(A)||\mathscr{G})dP = P(X^{-1}(A) \cap B)$$

But this is very hard because B is hard to characterize. But we have theorem that says you only have to check (*) for all B in a π-system generating $\sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G})$.

$$\mathscr{P} = \{\Theta^{-1}(G) \cap F : G \in \mathscr{F}_\Theta, F \in \mathscr{G}\}$$

It is trivial to show that $\mathscr{P}$ is a π-system.
MORE IN PHOTO
Meanwhile,

$$\mathscr{P} \subseteq \sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G})$$

Therefore,

$$\sigma(\Theta^{-1}(\mathscr{F}_\Theta) \cup \mathscr{G}) = \sigma(\mathscr{P})$$

So, sufficent to check (*) $\forall B \in \mathscr{P}'$

$$B \in \mathscr{P} \Rightarrow B = \Theta^{-1}(G) \cap F, G \in \mathscr{F}_\Theta, F \in \mathscr{G}$$

So, we want

$$\int_{\Theta^{-1}(G)\cap F} P(X^{-1}(A)||\mathscr{G})dP = P(\Theta^{-1}(G) \cap F \cap X^{-1}(A))$$

$$\begin{aligned}
\int_{\Theta^{-1}(G)\cap F} P(X^{-1}(A)||\mathscr{G})dP &= \int_{\Theta^{-1}(G)\cap F} E\left(I_{X^{-1}(A)}||\mathscr{G}\right)dP \\
&= E\left(I_{\Theta^{-1}(G)}I_F E(I_{X^{-1}(A)}||\mathscr{G})\right) \\
&= E\left(E(I_{\Theta^{-1}(G)}I_F||\mathscr{G})E(I_{X^{-1}(A)}||\mathscr{G})\right) \\
&= E\left(I_F E(I_{\Theta^{-1}(G)}||\mathscr{G})E(I_{X^{-1}(A)}||\mathscr{G})\right) \\
&= E\left(I_F E(I_{\Theta^{-1}(G)}I_{X^{-1}(A)}||\mathscr{G})\right) \\
&= E\left(E(I_F I_{\Theta^{-1}(G)}I_{X^{-1}(A)}||\mathscr{G})\right) \\
&= E\left(I_F I_{\Theta^{-1}(G)}I_{X^{-1}(A)}\right) \\
&= P(F\cap\Theta^{-1}(G)\cap X^{-1}(A))
\end{aligned}$$

Monday January 23

$2 \Rightarrow 1$. We want to show that

$$P(X^{-1}(A)||\mathscr{G})P(\Theta^{-1}(G)||\mathscr{G})$$

is conditional probability of

$$P(X^{-1}(A)\cap\Theta^{-1}(G)||\mathscr{G})$$

for all $F \in \mathscr{G}$.

$$\begin{aligned}
\int_F P(X^{-1}(A)||\mathscr{G})P(\Theta^{-1}(G)||\mathscr{G})dP &= E\left[I_F E\left(I_{X^{-1}(A)}||\mathscr{G}\right)E\left(I_{\Theta^{-1}(G)}||\mathscr{G}\right)\right] \\
&= E\left[E\left(I_{X^{-1}(A)}||\mathscr{G}\right)E\left(I_F I_{\Theta^{-1}(G)}||\mathscr{G}\right)\right] \\
&= E\left[E\left(I_{X^{-1}(A)}||\mathscr{G}\right)I_F I_{\Theta^{-1}(G)}\right] \\
&= E\left[E\left(I_{X^{-1}(A)}I_F I_{\Theta^{-1}(G)}||\Theta,\mathscr{G}\right)\right] \\
&= E\left[I_{X^{-1}(A)}I_F I_{\Theta^{-1}(G)}\right] \\
&= P(X^{-1}(A)\cap\Theta^{-1}(G)\cap F)
\end{aligned}$$

■

1.4 Equivalence of Frequentist & Bayesian Sufficiency

Here we have,

$$(\Omega_\Theta,\mathscr{F}_\Theta,\mu_\Theta),(\Omega_X,\mathscr{F}_X,\mu_X),(\Omega_T,\mathscr{F}_T)$$

Where

$$T:\Omega_X\to\Omega_T \text{ⓜ} \mathscr{F}_X/\mathscr{F}_T$$

is called a statistic.

$$T = T(X) \text{ or } T \circ X = T(X(\omega))$$

In fewquentist setting, we say that T is **suffienct** if $P_{X|T,\Theta}$ does not depend on Θ. It can be easily verified (see Homework) that $P_{X|T,\Theta}$ doesn't depend on Θ implies that

$$P_{X|T,\Theta} = P_{X|T} \text{ a.s. P}$$

This is "nearly" fequentist. Above is exchangeable with "$X \perp\!\!\!\perp \Theta | T$", but can't say this in frequentist setting.

$$P_{\Theta|T,X} = P_{\Theta|T} \Leftrightarrow P_{\Theta|X} = P_{\Theta|T}$$

That is to say that a statistic, T, is sufficient for Θ if and only iff the posterir distribution of $\Theta|X$ is the same as the posterior distribution of $\Theta|T$. This would be used in a Bayesian setting.

Definition 1.4.1 — Bayesian Sufficient. We say that $T \circ X$ is **Bayesian sufficient** if

$$P_{\Theta|X} = P_{\Theta|T} \text{ a.s. P}$$

Lemma 1.1 (HW 2) Suppose that $f(\theta)$ is a p.d.f such that

$$f(\theta) \propto exp\{-a\theta^2 + b\theta\}, \quad a > 0$$

Then,

1. $\theta \sim N(\frac{b}{2a}, \frac{1}{2a})$
2. $\int exp\{-a\theta^2 + b\theta\} d\theta = \sqrt{\frac{\pi}{a}} exp\{\frac{b^2}{4a}\}$

■ **Example 1.1** Suppose that

$$X|\Theta \sim N(\Theta, \sigma^2)$$
$$\Theta \sim N(\mu, \tau^2)$$

Find $\pi_{\Theta|X}(\theta|x), f_X(x)$.
Solution:

$$\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\} * \frac{1}{\sqrt{2\pi\tau^2}} exp\{-\frac{1}{2}\frac{(\theta-\mu)^2}{\tau^2}\} \\
&\propto exp\{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\} * exp\{-\frac{1}{2}\frac{(\theta-\mu)^2}{\tau^2}\} \\
&= exp\{-(\frac{1}{2\sigma^2} + \frac{1}{2\tau^2})\theta^2 + (\frac{x}{\sigma^2} + \frac{\theta}{\tau^2})\theta\}
\end{aligned}$$

Using Lemma 1.1,

$$\theta|X \sim N\left(\frac{\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}}{1/2(2\sigma^{-2} + 1/2\tau^{-1})}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

How about $f_X(x)$?

$$
\begin{aligned}
f_X(x) &= \int f(x|\theta)\pi(\theta)d\theta \\
&\vdots \\
&= \frac{1}{2\pi\sigma\tau} * exp\{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2}\} \int exp\{-(\frac{1}{2\sigma^2} + \frac{1}{2\tau^2})\theta^2 + (\frac{x}{\sigma^2} + \frac{\mu}{\tau^2})\theta\} \\
&= \cdots * \sqrt{\frac{\pi}{\frac{1}{2\sigma^2} + \frac{1}{2\tau^2}}} exp\{\frac{(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2})^2}{4(\frac{1}{2\sigma^2} + \frac{1}{2\tau^2})}\}
\end{aligned}
$$

We want to identify this as a p.d.f of x, so we can treat anything that is not x as a constant. Using elementary algebra we get...

$$\propto exp\{-(\frac{x^2}{2(\tau^2+\sigma^2)} + \frac{x\mu}{(\sigma^2+\tau^2)})\}$$

Applying Lemma 1.1 for x and simplifying,

$$X \sim N(\mu, \tau^2 + \sigma^2)$$

■

This can be extended to multivariate setting, 2-sample setting, ANOVA setting, regression setting, etc. It is essential to all aspects of linear models.

Wednesday January 25

■ **Example 1.2** Suppose

$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$$
$$\theta \sim N(\mu, \tau^2)$$

$\pi(\theta|X_1, \ldots, X_n) =?$

By Example 1.1,

$$\theta|\bar{X} \sim N\left(\frac{\frac{\bar{X}}{\sigma^2/n} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2/n} + \frac{1}{\tau^2})}, \frac{1}{\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}}\right)$$

Note that the sample size will effect how much weight each of the pervious means is given. As $n \to \infty$, depending on which coeffecient goes to 1,

$$E(\theta|\underline{X}) \to \bar{X} \text{ or } \mu$$

Similarly with variance, as $n \to \infty$,

$$Var(\theta|\underline{X}) \to 0$$

We can generalize/approximate this phenomenon as follows,

$$\theta|X_1, \ldots, X_n \sim N(\hat{\theta}_1, I^{-1}(\hat{\theta}))$$

where $\hat{\theta}$ is the MLE.

In our special case,

$$\theta|X_1,\ldots,X_n \sim N(\bar{X}, \frac{\sigma^2}{n})$$

■

1.5 Conjugate Priors

1.5.1 Introduction

In general, computing posterior distributions or posterior means is a hard problem involving high-dimentional numerical integration. This was a big hurdle for Bayesian methods before computers. Now we can do methods such as Monte Carlo inegration (MCMC).

In special cases, such as with the Exponential Family and mixture of distributions, posterior can be expressed explicitly through the use of conjugate families.

Definition 1.5.1 — Conjugate Family. A family of distributions, $\mathscr{P}$, on $(\Omega_\theta, \mathscr{F}_\theta)$ is a **conjugate family** if

$$P_\Theta \in \mathscr{P} \Rightarrow \mathscr{P}_{\Theta|X} \in \mathscr{P}$$

Not unique, for example if you let $\mathscr{P}$ be the colleciton of all distributions on $(\Omega_\theta, \mathscr{F}_\theta)$, then then it is always conjugate. Usually there is a suitable conjugate family.

1.5.2 Exponential Family

■ **Example 1.3** $X_1,\ldots,X_n|\theta \sim Pois(\theta)$

$$\begin{aligned} f(x_1,\ldots,x_n|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta} \\ &= \frac{\theta^{\Sigma x_i} e^{-n\theta}}{\prod(x_i!)} \\ &\propto \theta^{c_1} e^{-c_2\theta} \end{aligned}$$

Note that we are treating θ as the variable of interest here, not x. Recall that if $\theta \sim Gamma(\alpha,\beta)$ then we have a distribution approaching the form above, that is,

$$\pi(\theta) \propto \theta^{\alpha-1} e^{-\theta/\beta}$$

If we use this form, then we may find $pi(\theta|\underline{X})$ using the following:

$$pi(\theta|\underline{X}) \propto \theta^{c_1+\alpha-1} e^{-(c_2+1/\beta)\theta}$$

Which gives us that

$$\theta|X_1,\ldots,X_n \sim Gamma(c_1+\alpha, (c_2+1/\beta)^{-1})$$

This is generally true for all exponential family distributions.

■

We say that X has exponential family distrubiton if is has p.d.f in the form of

$$\frac{e^{\theta^T t(x)}}{\int e^{\theta^T t(x) d\mu(x)}}$$

where μ is a σ-finite measure on $(\Omega_X, \mathscr{F}_X)$.

Essentially, p.d.f. with $\mu(x) \propto e^{\theta^T t(x)}$.

More generally, suppose that $\phi : \Theta \mapsto \Phi$ bijection (one-to-one onto). Then X has Exponential Family distribution if and only if the p.d.f of X with μ is

$$\frac{e^{\phi^T t(x)}}{\int e^{\phi^T t(x) d\mu(x)}}$$

In this case, $X \sim Exp(\phi, t, \mu)$, when ϕ is identity, this is called the canonical form of exponential family.

Theorem 1.5.1 — 1.3 from class. If

$$P_\Theta \sim Ep(\xi, \phi, \nu)$$

$$P_{X|\Theta} \sim Ex(\phi, t, \mu)$$

then

$$P_{\Theta|X} \in Ep(\xi_X, \phi, \nu_x)$$

where

$$\xi_X(\alpha) = \xi(\alpha) + t(x)$$

$$d\nu_X(\theta) = \frac{d\nu(\theta)}{\int e^{\phi^T(\theta) t(x)} d\mu(x)}$$

Proof. $P_\Theta \in Ep(\xi, \phi, nu) \Rightarrow \pi(\theta) = \frac{e^{\xi^T(\alpha) t(\theta)}}{\int e^{\xi^T(\alpha) t(\theta) d\nu(x)}}$

$$f(x|\theta) = \frac{e^{\phi^T t(x)}}{\int e^{\phi^T t(x) d\mu(x)}}$$
$$\pi(\theta|x) = \text{PHOTO}$$
$$=$$

GIANT fraciton of fractions = PHOTO
So, with respect to the new measure,

$$d\nu_X(\theta) = \frac{d\nu(\theta)}{\int e^{\phi^T(\theta) t(x)} d\mu(x)}$$

where the pdf is

So, $\theta|X \sim Ep(\xi(\alpha)+t(x), \phi(\theta), \frac{d\nu(\theta)}{\int e^{\phi^T(\theta)t(x)d\mu(x)}})$

■

Friday January 27

■ **Example 1.4 — One Parameter Normal.**

$$X|\theta \sim N(\theta, \sigma^2)$$

Want to assign conjugate prior for θ.

$$f(x|\theta) \propto exp\{(\frac{-1}{2\sigma^2})\theta^2 + (\frac{\mu}{\sigma^2})\theta\}$$

Recall Lemma 1.1.

$$f(\theta) \propto exp\{(\frac{-1}{2\sigma^2})\theta^2 + (\frac{\mu}{\sigma^2})\theta\}$$

$$\Rightarrow \theta \sim N(\mu, \sigma^2)$$

So, consider the following family,

$$\mathscr{F} = \{f(\theta) \propto exp\left\{\frac{-1}{2\alpha_2^2}\theta^2 + \frac{\alpha_1}{\alpha_2^2}\theta\}\right\}$$

Suppose that $\pi \in \mathscr{F}$. Then we have that

$$\begin{aligned}
\pi(\theta|X) &\propto f(x|\theta)\pi(\theta) \\
&\propto exp\{(\frac{-1}{2\sigma^2})\theta^2 + (\frac{\mu}{\sigma^2})\theta\} * exp\left\{\frac{-1}{2\alpha_2^2}\theta^2 + \frac{\alpha_1}{\alpha_2^2}\theta\}\right\} \\
&= exp\left\{(\frac{-1}{2\alpha_2^2} - \frac{1}{2\alpha_2})\theta^2 + (\frac{x}{\sigma^2} + \frac{\alpha_1}{\alpha_2^2})\theta\}\right\} \\
&= -\frac{1}{2}(\frac{1}{\sigma^2} + \frac{1}{\alpha^2})
\end{aligned}$$

So we have that

$$\theta|X \sim N(\frac{\frac{\alpha_1}{\alpha_2} + \frac{x}{\sigma_2}}{(\frac{1}{\sigma^2} + \frac{1}{\alpha^2})}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\alpha^2}})$$

■

1.5.3 Convex Hull of Conjugate Family

R

Mathematically, if $A \subseteq \mathbb{R}^k$,

$$Conv(A) = \{\alpha_1 s_1 + \cdots + \alpha_k s_k : \alpha_1 \geq 0, \ldots, \alpha_k \geq 0; \alpha_1 + \cdots + \alpha_k = 1; s_1 \in A, \ldots, s_k \in A\}$$

R By the way, a convex combination of a set of probability measures, say $P_1, \ldots, P_k$, is called the **mixture** of $P_1, \ldots, P_k$. So $Conv(\mathscr{P})$ is simply the colleciton of all mixture distributions derived from $\mathscr{P}$.

Theorem 1.5.2 If $\mathscr{P}$ is conjugate to $P_{X|\theta}$, then $Conv(\mathscr{P})$ is also conjugate.

Proof. Suppose that $P_\Theta \in Conv(\mathscr{P})$. Want to show that $P_{\Theta|X} \in Conv(\mathscr{P})$.

Sicne $P_\Theta \in Conv(\mathscr{P})$ there exists

$$\alpha_1, \ldots, \alpha_k \in \mathbb{R}, \sum^k \alpha_i = 1, \alpha_i \geq 0$$

$$P_\Theta^{(1)} \ldots P_\Theta^{(k)} \in \mathscr{P}$$

such that

$$P_\Theta = \sum^k \alpha_i P_\Theta^{(i)}$$

$$\begin{aligned}
f_X(x) &= \int f(x|\theta) dP_\Theta \\
&= \int f(x|\theta) d(\sum^k \alpha_i P_\Theta^{(i)}) \\
&= \sum^k \alpha_i \int f(x|\theta) dP_\Theta^{(i)} \\
&= \sum^k \alpha_i \int m_i(x) dP_\Theta^{(i)} \\
dP_\Theta(\cdot|x) &= \frac{f(x|\theta)}{f_X(x)} dP_\Theta \\
&= \frac{f(x|\theta)}{\sum^k \alpha_i m_i(x)} \sum^k \alpha_k dP_\Theta^{(i)} \\
&= \sum^k \frac{\alpha_i m_i(x)}{\sum^k \alpha_j m_j(x)} \frac{f(x|\theta) dP_\Theta^{(i)}}{m_i(x)}
\end{aligned}$$

Note that the first term is great than zero and sums to 1. It's our new α^*. In the second term note that $P_\Theta^{(i)} \in \mathscr{P}$ and $P_{\Theta|X}^{(i)}(\cdot|x) \in \mathscr{P}$.

Thus this is a convex combination of members of $\mathscr{P}$ in $Conv(\mathscr{P})$.

■

This is a nice way to construct conjugate families of mixtures of exponential family.

1.6 Two-Parameter, Normal Family

Definition 1.6.1 — Inverse χ^2 Disribution. A random variable T is said to have an inverse χ^2 distribution of $\chi_{(k)}^{-2}$ if and only if,

$$\frac{1}{T} \sim \chi_{(k)}^2$$

If, for some $\tau < 0$, $\frac{T}{\tau} \sim \chi_{(k)}^{-2}$ then we write that

$$T \sim \tau\chi^{-2}_{(k)}$$

By Jacobian theorem,

$$T = g(X)$$

$$f_T(t) = f_X(g^{-1}(t))|det(\frac{\partial g^{-1}(t)}{\partial t})|$$

Above will be shown in HW.

Theorem 1.6.1 If $T \sim \chi^{-2}_{(\nu)}$ then,

$$f(t) = \frac{1}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}} t^{-\frac{\nu}{2}-1} e^{\frac{-1}{t}}, t > 0$$

Proof. Shown in HW 2. ■

Now suppose that $X_1, \ldots, X_n|\phi, \lambda \overset{iid}{\sim} N(\lambda, \phi)$ where both λ and ϕ are random. How do we assign conjugate prior of (λ, ϕ)?

From classical statistics (Stat 514) we know that the sufficience statistic for λ, ϕ is $(\sum X_i, \sum X_i^2)$ or $(T = \bar{X}, S = \sum_{i=1}^n (X_i - \bar{X})^2$.
So $\pi(\lambda, \phi|X_1, \ldots, X_n) = \pi(\lambda, \phi|T, S)$. Moreover, form Normal theory,

$$T \perp\!\!\!\perp X|\lambda, \phi$$
$$T|\lambda, \phi \sim N(\lambda, \phi/n)$$
$$S|\lambda, \phi \sim \phi\chi^2_{(n-1)}$$

Monday January 30

So, we need only to look at [GET NOTES FROM BUDDY]

$$\begin{aligned} f(T, S|\lambda, \phi) &= f(S|\lambda, \phi) f(T|\lambda, \phi) \\ &= f(S|\phi) f(T|\lambda, \phi) \\ &= [S|\phi][T|\lambda, \phi] \\ &= \left[\phi^{-(\frac{n-1}{2}-1)} exp\left\{(\frac{-s}{2\phi})(\frac{1}{\phi})\right\}\right] * \left[\frac{1}{\sqrt{2\pi}(\phi/n)^{1/2}} exp\left\{\frac{-1}{2(\phi/n)}\lambda^2 + \frac{t}{\phi/n}t + (\frac{-1}{2(\phi/n)})t^2\right\}\right] \end{aligned}$$

So we have $N(t, \frac{\phi}{n}) s\chi^{-2}_{(n-3)})$.
More generally, the distribution,

$$N(a, \frac{\phi}{m}) * \tau\chi^{-2}_{(k)}$$

is referred to (by Bing Li) as NICH (Normal-Inverse CHi-square). Here,

$$NICH(a, m, \tau, k)$$

We may say that $(\lambda, \phi) \mapsto [S|\phi][T|\lambda, \phi]) \sim NICH(t, n, s, n-3)$. The likelihood is NICH.

Lemma 1.2

$$NICH(a_1,m_1,\tau_1,k_1) * NICH(a_2,m_2,\tau_2,k_2) = NICH(a_3,m_3,\tau_3,k_3)$$

Where we have that

$a_3 = \frac{m_1a_1+m_2a_2}{m_1+m_2}$
$m_3 = m_1+m_2$
$\tau_3 = \tau_1+\tau_2+m_1a_1^2+m_2a_2^2-m_3a_3^2$
$k_3 = k_1+k_2-3$

Before proof, let's rewrite likelihood as

$$NICH(a,m,\tau,k) \propto \phi^{-1/2}exp\{\frac{-1}{2(\phi/m)}\lambda^2+\frac{a}{\phi/m}\lambda\}\phi^{-(k/2)-1}exp\{-\frac{\tau+ma^2}{2\phi}\}$$

Proof. By definition

$$\begin{aligned} NICH_1 * NICH_2 &\propto \phi^{-1/2}exp\{\frac{-1}{2(\phi/m_1)}\lambda^2+\frac{a_1}{\phi/m_1}\lambda\}\phi^{-(k_1/2)-1}exp\{-\frac{\tau_1+m_1a_1^2}{2\phi}\} * \phi^{-1/2}exp\{\frac{-1}{2(\phi/m_2)}\lambda^2+ \\ &= \phi^{-1/2}exp\{(\frac{-1}{2(\phi/m_1)}-\frac{1}{2(\phi/m_2)})\lambda^2+(\frac{a_1}{\phi/m_1}+\frac{a_2}{\phi/m_2})\lambda\}\phi^{-(k_1/2)-1}\phi^{-(k_2/2)-1}exp\{-\frac{\tau_1+ \\ &= \phi^{-1/2}exp\{(\frac{-1}{2(\phi/m_3)})\lambda^2+(\frac{a_3}{\phi/m_3})\lambda\}\phi^{-(k_3/2)-1}exp\{-\frac{\tau_3+m_3a_3^2}{2\phi}\} \end{aligned}$$

■

Recall we have that $(\lambda,\phi) \mapsto [S|\phi][T|\lambda,\phi]) \sim NICH(t,n,s,n-3)$. By Lemma 1.2, we assign prior

$$[\lambda|\phi][\phi] \sim NICH(a,m,\tau,k)$$

This gives us that

$$\pi(\lambda,\phi|T,S) \sim NICH(a^*,m^*,\tau^*,k^*)$$

where we have that

$a^* = \frac{ma+nT}{m+n}$
$m^* = m+n$
$\tau^* = \tau+S+ma^2+nt^2-(m+n)(\frac{ma+nt}{m+n})^2$
$k^* = k+n-3+3$

So, $[\lambda|\phi,X] \sim N(a^*,\frac{\phi}{m^*})$, and $[\phi|X] \sim \tau^*\chi^{-2}_{k^*}$.
Note that we can make inference about λ,ϕ using

$$\frac{\lambda-a^*}{\sqrt{\phi/m^*}}|\phi,X \sim N(0,1)$$

since the RHS doesn't depend on ϕ we have that it is independence form $\phi|X$.

$$\frac{\lambda - a^*}{\sqrt{\phi/m^*}} | X \sim N(0,1)$$

Wednesday February 1

We want to make inference about

$$\frac{\frac{\lambda - a^*}{\sqrt{\phi/m^*}}}{\sqrt{\frac{\tau^*}{\phi}/m^*}} | X \sim t_{m^*}$$

We may reorganize this to be

$$\frac{m^*(\lambda - a^*)}{\sqrt{\tau^*/m^*}} \sim t_{m^*}$$

Similarly, for making inference about ϕ,

$$\phi | X \sim \tau^* \chi^{-2}_{m^*}$$

$$\frac{\tau^*}{\phi} | X \sim \chi^2_{m^*}$$

1.7 Noninformative Prior

1.7.1 General Concept

Even when you don't have a prior distribution it's still beneficial to use Bayesian setting. For example, in dealing with nuisance parameters we deal with high dimension prior.

So here, try to use Bayesian to solve "no prior information" proglem. We want to use a flat prior, e.g. the Lebesgue measure. But, in what is this flat? Supppose we impose the Lebesgue measure on θ. Then the prior for monotone transformation of θ, say θ^3, is not Lebesgue anymore.

Definition 1.7.1 — Improper Prior. An infinite, but σ-finite measure on Ω_Θ is called an **improper prior**.

■ **Example 1.5**

$$X | \theta \sim N(\theta, \phi)$$

Note that here ϕ is known.

$$\pi(\theta) \equiv 1$$

$$f(X|\theta) = \sqrt{2\pi}^{-1} e^{-1/2*(x-\theta)^2}$$

$$\pi(\theta|X) = \sqrt{2\pi}^{-1} e^{-1/2*(x-\theta)^2} * 1 = N(X,1)$$

$$f_X(x) = \frac{f(x|\theta)\pi(\theta)}{\pi(\theta|X)} = \frac{f(x|\theta)}{\pi(\theta|x)} = 1$$

The marginal is Lebesgue improper! But, posterior is important ■

■ **Example 1.6**

$$X_1,\ldots,X_n|\lambda,\phi \sim N(\lambda,\phi)$$

Here both parameters are random.
We have sufficient statistics, $(S=\bar{X}, T=\sum(X_i-\bar{X})^2)$.

$$[(\lambda,\phi)\mapsto f(t,s|\lambda,\phi)] \sim NICH(t,n,s,n-3)$$
$$\propto \phi^{-1/2} exp\left\{\frac{-1}{2(\phi/n)}\lambda^2+\frac{t}{\phi/n}\lambda\right\}\phi^{-((n-3)/2)-1}exp\{-\frac{s+nt^2}{2\phi}\}$$

$$[\lambda|\phi]=1$$
$$[\phi]=\frac{1}{\phi}$$

When we multiply the p.d.f of our NICH by $\frac{1}{\phi}$ we get NICH(t, n, s, n-1) following the same argument in Seciton 1.5 (?) (proper case).

We can show that

$$\frac{\sqrt{n}(\lambda-t(x))}{\sqrt{s(x)/(n-1)}}|X \sim t_{(n-1)}$$
$$\frac{s(x)}{\phi}|X \sim \chi^2_{(n-1)}$$

Exactly the same as frequentist sample distribution, except what random has changed.

■

1.7.2 Invariant Prior

What is flat? What is a natural generalization of Lebesgue Measure?

Lebesgue measure is invariant under translation.

$$\theta \mapsto \theta + c$$

If λ is Lebesgue and T is translation,

$$\lambda \cdot T^{-1} = \lambda$$

This is natural generalization of flatness. Change T to be some other transformation that some how resembles translation = group of transformation.

R Review definition of group of transformations.
Ω (set)

$\mathcal{G}$ is a set of bijections (one-to-one on to) on Ω such that

1. for all $g_1,g_2 \in \mathcal{G}, g_1 \circ g_2 \in \mathcal{G}$
2. for all $e \in \mathcal{G}$ such that $e \circ g = g \circ e = g$ for all $g \in \mathcal{G}$

3. $g \in \mathcal{G} \Rightarrow g^{-1} \in \mathcal{G}$

With our own problem,

$$\Omega_\Theta \subseteq \mathbb{R}^P$$

Consider a parametric group,

$$\{g_t : t \in \Omega_\theta\}$$

Consider two types of transformations,

$$\theta \mapsto g_t(\theta) = L_t$$
$$\theta \mapsto g_\theta(t) = R_t$$

We may refer to these transformations as Left and Right transformations, respectively.

Two types of invariant priors or two generalizations of Lebesgue measure, 2-generalization of flatness.

Definition 1.7.2 A measure, Π, on Ω_Θ is the left Haar measure.

$$\Pi = \Pi \circ L_t^{-1} \forall t \in \Omega_\Theta$$

A measure, Π, is the right Harr measure if

$$\Pi = \Pi \circ R_t^{-1} \forall t \in \Omega_\Theta$$

■ **Example 1.7**

$$\Omega_\Theta = \mathbb{R}$$

Translation group;

$$\mathcal{G} = \{(\theta \mapsto \theta + c = g_c(\theta)) : c \in \mathbb{R}\}$$

Suppose that the improper prior density of θ is $\pi(\theta)$.

$$L_t(\theta) = g_c(\theta) = \theta + c = \eta$$

then the density of η is

$$\phi(g_c^{-1}(\eta)) |\frac{\partial g_c^{-1}(\eta)}{\partial \eta}|$$

We want to find

$$\pi(\cdot) = \pi(\cdot - c)$$
$$\pi(\theta) = \pi(\theta - c) \quad \forall \theta \in \mathbb{R}$$

If we take $\theta = 0$,

$$\pi(0) = \pi(-c)$$

If we take $-c = \theta$,

$$\pi(\theta) = \pi(0) \propto \text{ Lebesgue}$$

Left Haat measure IS the Lebesgue measure.

Right transformation,

$$\begin{aligned} R_c(\theta) &= g_\theta(c) \\ &= c+\theta \\ &= \theta + c \\ &= g_c(\theta) \\ &= L_c(\theta) \end{aligned}$$

The right Haar measure is also proportional to the Lebesgue. ■

Friday February 3

■ **Example 1.8** Haan measures for this group:

$$\Omega_\Theta = (0,\infty)$$

$$\mathscr{G} = \{(\theta \mapsto a\theta) : a > 0\}$$

We can show that a goup used for distributions like N(0, θ)
Left tansformation:

$$L_a = a\theta = \eta$$

if density for θ is $\pi(\theta)$.

$$\begin{aligned} \pi_\eta(\eta) &= \pi(\frac{\eta}{a})|\frac{d\eta/a}{d\eta}| \\ &= \pi\left(\frac{\eta}{a}\right)\frac{1}{a} \\ \pi_\eta(1) &= \pi\left(\frac{1}{a}\right)\frac{1}{a} \\ \pi(\frac{1}{a}) &\propto a \\ \pi(\theta = \frac{1}{a}) &\propto \frac{1}{\theta} \end{aligned}$$

Right transformation:

$$R_a(\theta) = g_\theta(a) = \theta a = L_a(\theta)$$

Right Haar density is also

$$\pi(\theta) = \frac{1}{\theta}$$

■

■ **Example 1.9** $N(\mu,\sigma^2), \Omega_\Theta = \{(\mu,\sigma) : \mu \in \mathbb{R}, \sigma > 0\}$

Consider group,

$$\mathscr{G} = \{[(\mu,\sigma) \mapsto (c\mu + b, c\sigma) = (g_{b,c}(\mu,\sigma))] : b \in \mathbb{R}, c > 0\}$$

Left transformation:

$$L_{b,c}(\mu,\sigma) = g_{b,c}(\mu,\sigma) = (c\mu + b, c\sigma) = (\tilde{\mu},\tilde{\sigma})$$

$$\begin{aligned}
\pi_\eta(\tilde{\mu},\tilde{\sigma}) &= \pi(\frac{\tilde{\mu}-b}{c},\frac{\tilde{\sigma}}{c}) \begin{vmatrix} \frac{\partial\mu}{\partial\tilde{\mu}} & \frac{\partial\mu}{\partial\tilde{\sigma}} \\ \frac{\partial\sigma}{\partial\tilde{\mu}} & \frac{\partial\sigma}{\partial\tilde{\sigma}} \end{vmatrix} \\
&= \pi(\frac{\tilde{\mu}-b}{c},\frac{\tilde{\sigma}}{c}) \begin{vmatrix} \frac{1}{c} & 0 \\ 0 & \frac{1}{c} \end{vmatrix} \\
&= \pi(\frac{\tilde{\mu}-b}{c},\frac{\tilde{\sigma}}{c})\frac{1}{c^2} \\
\pi_\eta(0,1) &= \pi(\frac{-b}{c},\frac{1}{c})\frac{1}{c^2}
\end{aligned}$$

So if we consider $(\frac{-b}{c},\frac{1}{c}) = (\mu,\sigma)$ we get that

$$\pi(\mu,\sigma) \propto \frac{1}{\sigma^2}$$

which is the Left Haan measure.
Right transformation:

$$R_{b,c}(\mu,\sigma) = g_{\mu,\sigma}(b,c) = (\mu + \sigma b, c\sigma) = (\tilde{\mu},\tilde{\sigma}) \neq L_{b,c}$$

$$\begin{aligned}
\sigma b + \mu &= \tilde{\mu} \\
\sigma c &= \tilde{\sigma} \\
\sigma &= \frac{\tilde{\sigma}}{c} \\
\mu &= \tilde{\mu} - \sigma b = \tilde{\mu} - \tilde{\sigma}\frac{b}{c}
\end{aligned}$$

$$\begin{aligned}
\pi_\eta(\tilde{\mu},\tilde{\sigma}) &= \pi(\tilde{\mu} - tilde\sigma(\frac{b}{c}),\frac{\tilde{\sigma}}{c}) \begin{vmatrix} \frac{\partial\mu}{\partial\tilde{\mu}} & \frac{\partial\mu}{\partial\tilde{\sigma}} \\ \frac{\partial\sigma}{\partial\tilde{\mu}} & \frac{\partial\sigma}{\partial\tilde{\sigma}} \end{vmatrix} \\
&= \pi(\tilde{\mu} - \tilde{\sigma}(\frac{b}{c}),\frac{\tilde{\sigma}}{c}) \begin{vmatrix} 1 & \frac{-b}{c} \\ 0 & \frac{1}{c} \end{vmatrix} \\
&= \pi(\tilde{\mu} - \tilde{\sigma}(\frac{b}{c}),\frac{\tilde{\sigma}}{c}) \begin{vmatrix} 1 & \frac{-b}{c} \\ 0 & \frac{1}{c} \end{vmatrix} \\
&= \pi(\tilde{\mu} - \tilde{\sigma}(\frac{b}{c}),\frac{\tilde{\sigma}}{c})\frac{1}{c}
\end{aligned}$$

If we replace $(\tilde{\mu},\tilde{\sigma}) = (0,1)$ we see that

$$\pi(\mu,\sigma) \propto \frac{1}{\sigma}$$

Thus the Right and Left Harr are not the same. ∎

1.7.3 Jeffreys' Prior

Haar requires natural group of transformations which is not always available in particular appliations. An easily available prior, Jeffrey's prior is constructed using the folloiwng principle.

If we assign θ a measure, Π and $\eta = T(\theta)$ is one-to-one transformation, then the prior assigned to η should satisfy $\Pi \circ T^{-1}$

Jefferys' Prior

Let $f(X|\theta)$ be the likelihood. Let $I(\theta)$ be the Fisher Information

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(X|\theta)|\theta\right]$$

The Jeffreys' Prior is defined as

$$\Pi(\theta) \propto \sqrt{detI(\theta)}$$

Theorem 1.7.1 — 1.16. Let π_Θ be the Jeffreys' prior density of θ. $\Pi_\Phi(\phi)$ is the Jeffreys' prior density for ϕ where $\phi = h(\theta)$. Here, h is one-to-one. Then we have that

$$\pi_\Phi(\phi) = \Phi_\Theta(h^{-1}(\phi))\left|det\left(\frac{\partial h^{-1}(\phi)}{\partial\phi}\right)\right|$$

In terms of measure, $\Pi_\Phi = \Pi_\Theta \circ h^{-1}$.

Proof. Let $f_{X|\Phi}(x|\phi)$ represent the likelihood of ϕ. Because Φ and Θ are one-to-one,

$$f_{X|\Phi}(x|\phi) = f_{X|\theta}(x|\theta) = f_{X|\Theta}(X|h^{-1}(\phi))$$

Conditional distributions only depends on σ-field and two one-to-one variables generage same σ-field.

GET MORE FROM PHOTOS ∎

Monday February 6

▪ Example 1.10

$$X_1,\ldots,X_n \sim N(\theta,\sigma^2)$$

where both $\theta\&\sigma$ are unknown.

Find Jeffreys' prior for θ,σ.

$$f(x_i|\theta,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_i-\theta)^2}{2\sigma^2}\}$$

Table 1.1: Paper Rock Scissor Loss Table

	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

$$f(x|\theta,\sigma) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\{-\frac{\sum_i(x_i-\theta)^2}{2\sigma^2}\}$$

$$\log(f(x_i|\theta,\sigma)) = -n\sqrt{2\pi\sigma^2} - \frac{(x_i-\theta)^2}{2\sigma^2}$$

$$\frac{\partial^2}{\partial\theta^2}\log(f(x_i|\theta,\sigma)) = -\frac{n}{\theta}$$

$$\frac{\partial^2}{\partial\theta\partial\sigma^2}\log(f(x_i|\theta,\sigma)) = -(\sigma^2)^{-2}\sum(x_i-\theta)$$

$$\frac{\partial^2}{\partial\sigma^2}\log(f(x_i|\theta,\sigma)) = \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}\sum(x_i-\theta)^2$$

Want to take negative expecation of each of theses.

$$I(\theta,\sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2}(\sigma^2)^{-2} \end{pmatrix}$$

Want to find the determinent of I. Here it's proportional to σ^{-6}.

$$\sqrt{det(I)} = \sigma^{-3} = \phi^{\frac{3}{2}}$$

Which gives us our Jeffrey's prior for (θ,σ^2).

■

1.8 Statistical Decision Theory

Rock, Paper, Scissor.

Player One Action Space: $\Omega_{A,1} = \{R,P,S\}$

Player Two Action Space: $\Omega_{A,2} = \{R,P,S\}$

Loss Function - $L : \Omega_{A,1} x \Omega_{A,2} \to \mathbb{R}$

Without loss of generality, we may assume the lose to be player one's loss.

Winner gains one dollar, loser loses a dollar, and in the case of a tie, no money is exchanged. So we have $(\Omega_{A,1},\Omega_{A,2},L)$. In statistics, we may refer to player one as the "statistician" and player two as "nature".

Player one's action is called the action space, Ω_A. Player two's action space is called the parameter space, Ω_Θ.

But where does our data come from in this scenario? We are dealing with a statistical decision problem.

Statistical Decision Problem

The statisticain (player one) wants to know nature's action, Θ, but he doesn't. He observes a random variable, whose distribution depends on θ. This is defused information about θ. The random element, X, takes value on

$$(\Omega_X, \mathscr{F}_X)$$

where $X \sim P_\theta$, $\theta \in \Omega_\Theta$.

So we have

$$(\Omega_X, \mathscr{F}_X, \{P_\theta : \theta \in \Omega_\Theta\})$$

on top of $(\Omega_A, \Omega_\Theta, L$.
Decision rule:

$$d : \Omega_X \to \Omega_A$$

and the collection of all decision rule is $\mathscr{D}$. This incurs a loss $L(\theta, d(x))$.

Before we take an action, our anticipation of loss is

$$E(L(\theta, d(x))|\theta)$$

which we may refer to as **risk**.

Game: $(\Omega_A, \Omega_\Theta, L)$
Data: $(\Omega_X, \mathscr{F}_X)$
Model: $\{P_\theta : \theta \in \Omega_\Theta\}$
Decision Rule: $\mathscr{D}$
Risk: R

Bayesian Statistical Decision Problem

Here, the game, data, model, and decision rule remain the same. The difference is in how we calculate the risk and the use of a prior.

Game: $(\Omega_A, \Omega_\Theta, L)$
Data: $(\Omega_X, \mathscr{F}_X)$
Model: $\{P_\theta : \theta \in \Omega_\Theta\}$
Decision Rule: $\mathscr{D}$
Prior: P_Θ
Risk: r

$$r(d) = E_{\Theta,X}[L(\Theta, d(x))] = \int_{\Omega_\Theta x \Omega_X} L(\theta, d(x)) dP_{\Theta,X}$$

Wednesday February 8

Bayes Rule: Optimal estimation, optimal test. This is (mathematically) easier than frequentist optimal procedures becuase we have a metric in Θ. Don't have to make uniform statements about θ, such as UMVUE, UMP, UMPU, ...

Definition 1.8.1 — Bayes' Rule. The Bayes rule is

$$d_B = \arg\min\{r(d) : d \in \mathscr{D}\}$$

If P_Θ is improper then d_B is called **generalized Bayes' Rule**.

$$r(d) = \int_{\Omega_\Theta x \Omega_X} L(\theta, d(x)) f_{X,\Theta}(x,\theta) d(\mu_x x \mu_\Theta)(x,\theta)$$

Note that usually $L(\theta, d(x)) \geq 0$ so we have that

$$f_{x,\Theta} = \begin{cases} f(x|\theta)\pi(\theta) \\ \pi(\theta|x)f(x) \end{cases}$$

First way:

$$\begin{aligned} r(d) &= \int_{\Omega_\Theta} \left[\int_{\Omega_X} L(\theta, d(x)) f_{X|\Theta}(x|\theta) d(\mu_x)(x) \right] \pi(\theta) d\mu_\Theta \\ &= \int_{\Omega_\Theta} R(\theta, d)\pi(\theta) d\mu_\Theta \end{aligned}$$

Other way would swap Θ, X. (??)

$$\begin{aligned} r(d) &= \int_{\Omega_X} \left[\int_{\Omega_T heta} L(\theta, d(x)) \pi_{\Theta|X}(x|\theta) d(\mu_\Theta)(\theta) \right] d\mu_X \\ &= \int_{\Omega_X} \rho(x, d(x)) d\mu_X(x) \end{aligned}$$

Where we have the posterior expected loss,

$$\rho(x,a) = E(L(\theta,a)|X)$$

INSERT PHOTO
How to calculate Bayes's rule, not by definition.

Definition 1.8.2

$$\arg\min(r(d) : d \in \mathscr{D})$$

is the minimum over a set of functions, $\mathscr{D}$. A member of $\mathscr{D}$ is

$$d : \Omega_X \to \Omega_A$$

But this can be converted into minimization over numbers or vectors by the following theorem.

Theorem 1.8.1 — 1.7 in Notes. Suppose $L(\theta,a) \geq C < -\infty$ for all $\theta \in \Omega_\Theta, a \in \Omega_A$. Then the decision rule,

$$d_B : \Omega_X \to \Omega_A, x \mapsto \arg\min\{\rho(x,a) : a \in \Omega_A\}$$

is the Bayes' Rule.

Ω_A is usually $\mathbb{R}, \mathbb{R}^P$ or subsets thereof. So we are not minimizing over functional spaces.

Proof. By Tonelli's theorem, (because we can assume WLoG that $L(\theta,a) \geq 0, \quad \forall \theta, a$)

$$r(d) = \int_{\Omega_X} \rho(x, d(x)) f_X(x) d\mu_X(x)$$

So, for any x,

$$\rho(x, d_B(x)) \leq \rho(x, d(x))$$

Thus,

$$\int_{\Omega_X} \rho(x, d_B(x)) f_X(x) d\mu_X(x) \leq \int_{\Omega_X} \rho(x, d(x)) f_X(x) d\mu_X(x)$$

And we have that $r(d_B) \leq r(d) \forall d \in \mathscr{D}$. ■

In frequentist theory, optimality have to be stated uniformly.

"Commonly used optimal criteria" in frequentist decision theory
1. admissibility
2. minimax

Definition 1.8.3 A decision rule is **inadmissible** if there exists $d' \in \mathscr{D}$ such that
1. $R(\theta,d') \leq R(\theta,d) \forall \theta \in \Omega_\Theta$
2. $R(\theta,d') < R(\theta,d)$ for some $\theta \in \Omega_\Theta$

A decision rule that is not inadmissible is admissible.

Definition 1.8.4 A decision rule $d \in \mathscr{D}$ is a **minimax rule** if for all $d' \in \mathscr{D}$,

$$\sup_{\theta \in \Omega_\Theta} R(\theta,d) \leq \sup_{\theta \in \Omega_\Theta} R(\theta,d')$$

The relation between Bayes' rule and admissible rule, generally "all Bayes rule are adimissible".

Theorem 1.8.2 — 1.8 in Notes. Suppose
1. for each $d \in \mathscr{D}$, $R(\theta,d)$ is integrable with respect to P_Θ
2. for any $d_1, d_2 \in \mathscr{D}$,

$$R(\theta,d_1) - R(\theta,d_2) < 0$$

for some $\theta \in \Omega_\Theta$ which implies that

$$P(R(\theta,d_1) - R(\theta,d_2) < 0) > 0$$

Then any Bayes or generalized Bayes rule is admissiable.

Proof. Suppose $d_1 \in \mathscr{D}$ is Bayes and inadmissible. Then there exists $d_2 \in \mathscr{D}$ such that

1. $R(\theta,d_2) \leq R(\theta,d_1) \forall \theta \in \Omega_\Theta$
2. $R(\theta,d_2) < R(\theta,d_1)$ for some $\theta \in \Omega_\Theta$

But this gives us that

$$R(\theta,d_2) - R(\theta,d_1) \geq 0 \forall \theta$$

and

$$P(R(\theta,d_2) - R(\theta,d_1) < 0) > 0$$

Recall from STAT 517 that if $f \geq 0, \mu(f > 0) > 0$ then

$$\int f d\mu > 0$$

or that if $f \leq 0, \mu(f < 0) > 0$ then

$$\int f d\mu < 0$$

This implies that

$$\begin{aligned} &\Rightarrow \int R(\theta,d_2) - R(\theta,d_1) dP_\Theta < 0 \\ \Rightarrow &\int R(\theta,d_2) dP_\Theta < \int R(\theta,d_1) dP_\Theta \\ &\Rightarrow r(d_2) < r(d_1) \\ &\Rightarrow d_1 \text{ not Bayes'} \end{aligned}$$

■

Friday February 10

Theorem 1.8 condition 2 is a mild condition which is satifried by the two important cases

1. $R(\theta,d)$ is continuous function for all $d \in \mathscr{D}$, and $P(\Theta \in G) > 0$ for any empty open set. For example, if P « λ, λ « P ($P \equiv \lambda$), then $P(\Theta \in G) < 0$ fro all open $G \neq 0$. To see that the above statement is sufficient for the second condition of Theorem 1.8,
   FINISH FROM PHOTO
2. Ω_Θ is countable (σ-finite)

   $$\Omega_\Theta = \{\theta_1, \theta_2, \dots\}$$

   and

   $$P(\Theta = \theta_i) > 0, \forall i = 1,2,\dots$$

   If $R(\theta,d_2) - R(\theta,d_1) > 0$ for some $\theta \in \Omega_\Theta$ then this implies that

   $$R(\theta_i,d_2) - R(\theta_i,d_1) > 0 \quad i \in \{1,2,\dots\}$$

   But we know that $P(\Theta = \theta_i) > 0$ so

   $$P(R(\theta_i,d_2) - R(\theta_i,d_1) < 0) > 0$$

2. Estimation & Inference

Typically for esitmation,

$$\Omega_\Theta = \mathbb{R}^P, \Theta \subseteq \mathbb{R}^P, \Omega_A = \Omega_\Theta$$

so for testing problems,

$$\Omega_\Theta = \{0,1\} = \Omega_A$$

but for clarification problesm,

$$\Omega_\Theta = \{a_1, \ldots, a_k\} = \Omega_A$$

2.1 Estimation

$$\Omega_\Theta = \Omega_A = \mathbb{R}^P$$

One of the most commonly used loss is the squared loss or L_2 loss.

Definition 2.1.1 The L_2 loss is

$$L(\theta, a) = (\theta - a)^T w(\theta)(\theta - a)$$

where $w(\theta > 0), \forall \theta \in \Omega_\Theta$

In this case, the Bayes' rule is explicit.

Theorem 2.1.1 Suppose that

1. $E[\theta^T w(\theta)\theta|X] < \infty$ a.s.
2. $E[w(\theta)|X] > 0$ a.s

Then,

$$d_B(x) = (E[w(\theta)|X])^{-1} E[w(\theta)\theta|X]$$

is the Bayes' rule.

Note that if $w(\theta) = I_P$, then Bayes' rule for loss, $||\theta - a||^2$, is simply $E(\theta|X)$.

Proof. Because $L(\theta, a) \geq 0$, we need to minimize

$$\rho(x,a) = E[(\theta - a)^T w(\theta)(\theta - a)|X]$$

where a can depend on x.

$$\begin{aligned}
&E[(\theta - d_B(x) + d_B(x) - a)^T w(\theta)(\theta - d_B(x) + d_B(x) - a)|X] \\
&\quad = E[(d_B(x) - a)^T w(\theta)(d_B(x) - a)|X] + E[(d_B(x) - a)^T w(\theta)(\theta - d_B(x))|X] + E[(\theta - d_B(x))^T w(\theta \\
&\quad = (d_B - a)^T E[w(\theta - d_B)|X] \\
&\quad = (d_B - a)^T \left(E[w\theta|X] - E(wd_B|X)\right) \\
&\quad = 0
\end{aligned}$$

Overview:

$$\rho(x,a) = \text{nonneg } + 0 + 0 + \rho(x, d_B) \Rightarrow \rho(x,a) \geq \rho(x, d_B)$$

So we have that d_B is the Bayes' rule. ■

Another commonly used loss is the L_1 loss.

Definition 2.1.2 If $p = 1$ then the L_1 loss is

$$L(\theta, a) = |\theta - a|$$

What is the Bayes rules here?

Definition 2.1.3 — Median. Let U be a random varaible with c.d.f. F. Then any n umber m that satisfies

$$F(m-) \leq \frac{1}{2} \leq F(m)$$

is called the **median**. Note that this definition does not require F to be monotone.

Theorem 2.1.2 If U is integrable and m is a median of U, then

$$\int |U - m| dP \leq \int |U - a| dP \quad \forall a \in \Omega_U$$

Proof. Case 1. m < a

$$\begin{aligned}
\int_{\Omega_U} |U - m| - |U - a| dP &= \int_{U \leq m} + \int_{m < U \leq a} + \int_{U < a} \\
&= \int_{U \leq m} m - a dP + \int_{m < U \leq a} 2U - m - a dP + \int_{U < a} a - m dP \\
&= (m - a)P(U \leq m) + (**) + (a - m)P(U > a)
\end{aligned}$$

For (**),

$$\int_{m<U\le a} 2U-m-adP \le (2a-m-U)P(m<U\le a) \tag{2.1}$$
$$= (m-a)[P(U\le m)-P(U>a)-P(m<U\le a)] \tag{2.2}$$
$$= (m-a)[P(U\le m)-P(U>m)] \tag{2.3}$$
$$= (m-a)[P(U\le m)-(1-P(U\le m))] \qquad = (m-a)[2F(m)-1] \tag{2.4}$$

So when a > m,

$$E|U-m|-E|U-a| \le (2F(m)-1)(m-a)$$

Case 2. $a < m$

By the same argument,

$$E|U-m|-E|U-a| \le (1-2F(m-))(a-m)$$

But, because m is a median,

$$F(m-) \le \frac{1}{2} \le F(m)$$

FINISH FROM PHOTO ■

Monday February 13

Corollary 2.1

The posterior median $M(\theta|X)$ is a Bayes rule with respect to

$$L(\theta,a) = |\theta-a|$$

More generally,

$$L(\theta,a) = \begin{cases} \alpha_1(\theta-a) & \theta \ge a \\ \alpha_2(a-\theta) & \theta < a \end{cases}$$

Here, Bayes rule is quantile.
Besides Bayes estimations, another populat estimation is the generalized MLE. This is simply,

$$\hat{\theta} = \arg\max\{\pi(\theta|X) : \theta \in \Omega_\Theta\}$$

the posterior mode. MLE is a special case where $\pi(\theta)$ is a constant.

■ **Example 2.1** Bayes estimation under

$$L(\theta,a) = (\theta-a)^2 = |\theta-a|$$

Also generalized MLE.

We have a linear regression,

$$Y_i = \theta x_i + \varepsilon_i \quad i = 1,\ldots,n$$

$$\varepsilon_1, \ldots, \varepsilon_n | \theta \overset{iid}{\sim} N(\theta, \sigma^2)$$

We have prior knowledge that $\theta \geq 0$ (e.g. growth rate). how do we incorporate their information for estimation.

$$\Pi(\theta) = 1 \quad \theta \geq 0 = I(\theta \geq 0)$$

Here we find the imporoper prior.

$$f(x|\theta) \propto \exp\{-\frac{x^T x}{2\sigma^2}\theta^2 + \frac{x^T y}{\sigma^2}\theta\}$$

where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

By Lemma 1.1, if $f(\theta) \propto \exp\{-a\theta^2 + b\theta\}$ then

$$\theta \sim N(\frac{b}{2a}, \frac{1}{2a})$$

So, as a function of θ, $\theta \mapsto f(x|\theta)$ has the form

$$N(\frac{x^T y}{x^T x}, \frac{\sigma^2}{x^T x})$$

and we can then see that the posterior density would be

$$\pi(\theta|x) = \frac{N(\mu(x,y), \tau^2(x,y))I(\theta \geq 0)}{\int_0^\infty N(\mu, \tau^2) d\mu}$$

Note that this integral integrated with repsect to θ becomes

$$\int_0^\infty N(\mu, \tau^2) d\theta = \int_{\mu/\tau}^\infty N(0,1) d\theta = \Phi(\mu/\tau)$$

Now with the conjugate Bayes rule and $L(\theta, a) = (\theta - a)^2$ we may let $\gamma = \frac{\theta - \mu}{\tau}$ and get that

$$\pi(\gamma|x) = \frac{N(0,1)}{\Phi(\frac{\mu}{\tau})} I(\gamma \geq -\frac{\mu}{\tau})$$

$$\begin{aligned} E(\gamma|X) &= \frac{1}{\Phi(\mu/\tau)} \int_{-\mu/\tau}^\infty \gamma N(0,1) d\gamma \\ &= \frac{\exp\{-\mu^2/(2\tau^2)\}}{\sqrt{2\pi}\Phi(\mu/\tau)} \end{aligned}$$

$$E(\theta|X) = \mu + \tau * \frac{\exp\{-\mu^2/(2\tau^2)\}}{\sqrt{2\pi}\Phi(\mu/\tau)}$$

where, as above, $\mu = \frac{x^T y}{x^T x}$, and $\tau^2 = \frac{\sigma^2}{x^T x}$.

■

Bayes rule under $L(\theta,a) = |\theta - a|$ is the median. So the solution would be

$$\int_0^m N(\mu,\tau^2)d\theta = \frac{1}{2}\int_0^\infty N(\mu,\tau^2)d\theta$$

So with the change of variables,

$$\int_{-\mu/\tau}^{(m-\mu)/\tau} N(0,1)d\gamma = \frac{1}{2}\int_{-\mu/\tau}^{\infty} N(0,1)d\gamma$$

and after some algebra we get that

$$m = \tau\Phi^{-1}(1-\frac{1}{2}\Phi(\mu/\tau)) + \mu$$

and so here the generalized MLE is the MLE.

This is simply maximizing $N(\mu,\tau^2)$ over $\theta \geq 0$.
So the generalized MLE over $\theta \geq 0$ is

$$\max(0, \frac{x^T y}{x^T x})$$

2.2 Bayes Rule and Unbiasedness

Another relation between Bayesian and Frequentist criteria. Whereas bayes rule and admissibility are somewhat consistant, unibased is a fundamentally unBayesian idea.

Recall that d(x) is unbiased for θ if and only if

$$E(d(x)|\theta) = \theta \forall \theta$$

Theorem 2.2.1 — 2.3 in notes. Suppose $d_B(x)$ is Bayes rule with respect to

$$L(\theta,a) = (\theta - a)^T W(\theta)(\theta - a)$$

then $d_B(x)$ is biased unless

$$d(x) = \theta a.s. P_{\theta,X}$$

Proof. Recall that

$$d_B(x) = [E(W(\theta)|X)]^{-1} E(W(\theta)\theta|X)$$

In the follwoig, write $W(\theta) = W$, $d_B(x) = d_B$. We want to show that if d_B is unbiased then

$$r(d_B) = 0 \Leftrightarrow \theta = d_B(x) a.s. P$$

$$\begin{aligned} r(d_B) &= E\left[(\theta - E(W|X)^{-1}E(W\theta|X)^T W(\dots)\right] \\ &= E\left[E(\theta^T W\theta|X) - E(W\theta|X)^T E(W|X)^{-1}E(W\theta|X)\right] \\ &= E\left[E(\theta^T W\theta|X) - d_B^T E(W\theta|X)\right] \\ &= E\left[E(\theta^T W\theta - d_B^T W\theta|X)\right] \\ &= E(\theta^T W\theta - d_B^T W\theta) \\ &= E(\theta^T W\theta - E(d_B|\theta)^T W\theta) \\ &= 0 \end{aligned}$$

■

Wednesday February 15

2.3 Error Assessment in Bayesian Setting

In frequentist setting, X is nothing but one realizeation of a random variable that potentially takes many values. So we may use

$$\mathrm{Var}_\theta(x) = \mathrm{Var}(x|\theta)$$

or

$$\mathrm{MSE}(X|\theta) = \mathrm{E}(X - \mathrm{E}(X|\theta))^2$$

In the Bayesian context, X is fixed (i.e. it's being conditional on). Once you condition on something it's out of the probability picture.

$$\mathrm{Var}(\theta|X) = \mathrm{E}[(\theta - \mathrm{E}(\theta|X))(\theta - \mathrm{E}(\theta|X))^T]$$
$$\mathrm{MSE}_d(\theta|X) = \mathrm{E}(\theta - d(x)|x)^2 = \mathrm{E}[(\theta - d(x))(\theta - d(x))^T]$$

Note that $\mathrm{Var}(\theta|X)$ is a special case of $\mathrm{MSE}_d(\theta|X)$ when $d = d_B$ under squared loss. It's easy to show the identity

$$\mathrm{MSE}_d(\theta|X) = \mathrm{Var}(\theta|X) + [\mathrm{E}(\theta|X) - d(x)][\mathrm{E}(\theta|X) - d(x)]^T$$

■ **Example 2.2** Suppose that

$$X_1, \dots, X)n|\phi \overset{iid}{\sim} N(0,\phi), \phi \sim \tau\chi^{-2}_{(\nu)}$$

Then as in your HW #2, you can show that

$$\phi|X_1, \dots, X_n \sim (s+\tau)\chi^{-2}_{(\nu+n)}$$

where $s = \sum_{i=1}^n X_i^2$.
Also, as shown in the homework,

$$\mathrm{E}(\phi|X_1, \dots, X_n) = \frac{s+\tau}{\nu+n-2}$$

$$\mathrm{Var}(\phi|X_1, \dots, X_n) = \frac{2(s+\tau)^2}{(\nu+n-2)^2(\nu+n-4)}$$

Suppose that $d_1(x) = \frac{1}{n}S$, unbiased estimator ($\mu = 0$). Also that $d_2(x) = \frac{s+\tau}{\nu+n+2}$ is the GMLE (HW 2).

$$\text{MSE}_{d_1}(\phi|X_1,\ldots,X_n) = \frac{2(s+\tau)^2}{(\nu+n-2)^2(\nu+n-4)} + \left(\frac{s+\tau}{\nu+n-2} - \frac{1}{n}S\right)^2$$

$$\text{MSE}_{d_2}(\phi|X_1,\ldots,X_n) = \frac{2(s+\tau)^2}{(\nu+n-2)^2(\nu+n-4)} + \left(\frac{s+\tau}{\nu+n-2} - \frac{s+\tau}{\nu+n+2}S\right)^2$$

So you can present $d \pm \sqrt{\text{MSE}_d(\theta|X_1,\ldots,X_n)}$. ∎

2.4 Credible Set (or Interval)

In the frequentist setting, CI is a random interval. C(X):

$$P_\Theta(\theta \in C(X)) = 1-\alpha$$

In the Bayesian setting, C(X) is fixed, we want:

$$P(\theta \in C(X)|X) = 1-\alpha$$

Definition 2.4.1 A $(1-\alpha)$-credible set is any $C \in \mathscr{F}_\theta$ such that

$$P(\Theta^{-1}(C)||X) \geq 1-\alpha \quad a.s.$$

Shortest interval is preferred. HPD: Highest Posterior Density Credible Set.

Definition 2.4.2 The $(1-\alpha)$-highest posterior density for $\theta$is a $C \in \mathscr{F}_\theta$ such that

$$C = \{\theta \in \Omega_\theta | \pi(\theta|X) \geq k_\alpha\}$$

where

$$k_\alpha = \sup\{k|P(\pi(\theta|X) \geq k) \geq 1-\alpha\}$$

Intuitively,
PHOTO
Continuous $\pi(\theta|X)$ then in this case (PHOTO)

$$P(\pi(\theta|X) \geq k_\alpha) = 1-\alpha$$

$$P(\{\theta|pi(\theta|X) \geq K_\alpha\}||X)$$

As we move K up, $P(\pi(\theta|x) \geq K)$ takes only three values. What if $1-\alpha$ is not one of the three?
PHOTO
HPD is the shortest credible set as proved in next theorem.

Theorem 2.4.1 — 2.4. Suppose that $\pi(0) > 0$, for all $\theta \in \Omega_\Theta$. Let C^*_α be a $(1-\alpha)$ HPD credible set and C_α be any $(1-\alpha)$ credible set.

Furthermore, assume that $P_{\Theta|X}(C^*_\alpha|X) = 1-\alpha$. Then,

$$\mu_\Theta(C^*_\alpha) \leq \mu_\Theta(C_\alpha)$$

Proof. Want to show:
$\mu_\Theta(c) < \mu_\Theta(C^*_\alpha) \Rightarrow P_{\Theta|X}(C|X) < P_{\Theta|X}(C^*_\alpha|X) = 1-\alpha$
Let $C \subseteq \Omega_\Theta, \mu_\Theta(C) < \mu_\Theta(C^*_\alpha) \Rightarrow \mu(C \setminus C^*_\alpha) < \mu(C^*_\alpha \setminus C)$
Friday February 17

FINSIH PROOF
We know that
$\pi(\theta|X) \geq K_\alpha$ on $C^*_\alpha \Rightarrow$ on $C^*_\alpha \setminus C$.

$\pi(\theta|X) \leq K_\alpha$ on $C(^*_\alpha)^C \Rightarrow$ on $C \setminus C^*_\alpha$.

$$\begin{aligned} P_{\Theta|X}(C^*_\alpha \setminus C|X) &= \int_{C^*_\alpha \setminus C} \pi(\theta|x) d\mu_\Theta(\theta) \\ &\geq K_\alpha \mu_\Theta(C^*_\alpha \setminus C) \\ &> K_\alpha \mu_\Theta(C \setminus C^*_\alpha) \\ &\geq \int_{C \setminus C^*_\alpha} \pi(\theta|X) d\mu_\Theta \\ &= P_{\Theta|X)}(C \setminus C^*_\alpha) \end{aligned}$$

■

INCLUDE PICTURE

■ **Example 2.3**

$$X_1, \ldots, X_n|\lambda, \phi \overset{iid}{\sim} N(\lambda, \phi)$$

$$\lambda|\phi \sim N(a, \frac{\phi}{m})$$

$$\phi \sim \tau\chi^2_{(k)}$$

So we know that

$$\frac{\sqrt{n+m}(\lambda - a(x))}{\sqrt{\tau(x) \setminus (n+k)}} \mid X \sim t_{(n_k)}$$

where,
$a(x) = (n\bar{x} + ma) \setminus (n+m)$
$\tau(x) = \sum(X_i - \bar{X})^2 + \tau + (\bar{X} - a)^2(m^{-1} + n^{-1})$

INSERT PHOTO
So the $(1-\alpha) * 100\%$ credible set for λ is

$$\{\lambda : -t_{(n+k)}(\frac{\alpha}{2}) < \frac{\sqrt{n+m}(\lambda - a(x))}{\sqrt{\tau(x) \setminus (n+k)}} < t_{(n+k)}(\frac{\alpha}{2})\}$$

Here we get,

$$a(x) \pm t_{(n+k)}(\frac{\alpha}{2}\sqrt{\tau(x)[(n+k)(n+m)]})$$

Recall that we also know that

$$\frac{\phi}{\tau(x)}|x \sim \chi^{-2}_{(n+k)}$$

PHOTO
C_1, C_2 are solution to

$$h(C_1) = h(C_2)$$
$$\int_{C_1}^{C_2} h(t)dt = 1-\alpha$$

where h is pdf of χ^{-2}_{k+m}
HDD set of $\{\phi : C_1 < \frac{\phi}{\tau(x)} < C_2\}$

■

2.5 Hypothesis Test

PHOTO
Common we use the following losses

- 0-1 <- if wrong lose 1
- 0 - C_1 - C_2 <- more nuanced. If wrong one way, C_1. If wrong the other way, C_2.

$$L(\theta,a) = \begin{cases} 0 & (\theta,a) \in (\Omega_\Theta^{(0)} x \{a_0\}) \cup (\Omega_\Theta^{(1)} x \{a_1\}) \\ C_1 & (\theta,a) \in (\Omega_\Theta^{(0)} x \{a_1\}) \\ C_2 & (\theta,a) \in (\Omega_\Theta^{(1)} x \{a_0\}) \end{cases}$$

Theorem 2.5.1 Suppose that

$$0 < P_{\Theta|X}(\Omega_\Theta^{(0)}|X) < 1$$

then the Bayes' rule for 0 - C_1 - C_2 Loss is

$$d_B(x) = \begin{cases} a_0 & C_1 P_{\Theta|X}(\Omega_\Theta^{(1)}|X) \le C_0 P_{\Theta|X}(\Omega_\Theta^{(0)}|X) \\ a_1 & C_1 P_{\Theta|X}(\Omega_\Theta^{(1)}|X) > C_0 P_{\Theta|X}(\Omega_\Theta^{(0)}|X) \end{cases}$$

Proof. Since the loss is bad, need to minimize

$$\rho(x,a)$$

over (a_0, a_1).

$$\begin{aligned} \rho(x,a) &= E(L(\theta,a)|X) \\ &= \int_{\Omega_\Theta^{(0)}} L(\theta,a) dP_{\Theta|X} + \int_{\Omega_\Theta^{(1)}} L(\theta,a) dP_{\Theta|X} \\ \rho(x,a_0) &= 0 + C_1 P_{\Theta|X}(\Omega_\Theta^{(1)}) \\ \rho(x,a_1) &= C_2 P_{\Theta|X}(\Omega_\Theta^{(0)}) + 0 \end{aligned}$$

■

There is a problem in the frequentist setting (e.g. $H_0 : \theta = \theta_0$)

$$\Omega_\Theta^{(0)} = \{\theta_0\}$$

So, if we put an absolute continous prior, then there is no mass assigned on $\Omega_\Theta^{(0)}$. So, the above Bayes rule does not work.

Definition 2.5.1 Let $(\Omega, \mathscr{F})$ be a measureable space. Then the Divas measure for $a \in \Omega$ is the set function

$$\delta_a(B) = \begin{cases} 1 & a \in B \\ 0 & a \notin B \end{cases}$$

This is a measure (proved in 517).

Note that is also can be shown that for any function $f : \Omega \to \mathbb{R}, B \in \mathscr{F}$

$$\int_B f(\omega)\delta_a(d\omega) = f(a)\delta_a(B)$$

Monday February 20

Lemma 2.1 Suppose that $(\Omega, \mathscr{F})$ is a measureable space. We have that $a \in \Omega, B \in \mathscr{F}, \delta_a(B)$ is the Dirac measure about a. Suppose that for any a, $\{a\} \in \mathscr{F}$, (true for Borel). Suppose that

$$f : \Omega \to \mathbb{R}, \textcircled{m}\mathscr{F} \setminus \mathbb{R}$$

then we have that

$$\int_B f(\omega)d\delta_a(\omega) = f(a)\delta_a(B)$$

Proof.

$$\int_B f d\delta_a = \sup_{\{A_i\} \in \mathscr{P}} \sum_{i=1}^{k} [\inf_{\omega < A_i} f(\omega)]\delta_a(A_i)$$

If $a \notin B$, then the above summation part of the equality is 0 for all $\{A_i\} \in \mathscr{P}$.

Now suppose that $a \in B$, then there exists a unique $A_i \in \{A_i\}$ such that$a \in A_i$.

So for any $\{A_i\} \in \mathscr{P}$

$$\sum_{j=1}^{k} [\inf_{\omega < A_j} f(\omega)]\delta_a(A_j) = [\inf_{\omega < A_i} f(\omega)]\delta_a(A_i), \quad a \in A_i$$

We know that this term is less than f(a), so this implies that

$$\int_B f d\delta_a \leq f(a)$$

From here we want to take the special partition

$$(\{a\}, B \setminus \{a\}) \in \mathscr{P}$$

$$\int f d\delta_a \geq [\inf_{\omega \in \{a\}} f(\omega)]\delta_a(\{a\}) + [\inf_{\omega \in B \setminus \{a\}} f(\omega)]\delta_a(B \setminus \{a\})$$
$$= f(a) * 1 + 0$$
$$= f(a)$$

So we have that $\int f d\delta_a = f(a)$ and thus we may conclude that

$$\int f d\delta_a = \begin{cases} 0 & a \notin B \\ f(a) & a \in B \end{cases} \Rightarrow \int f d\delta_a = f(a)\delta_a(B)$$

■

Suppose that Q_Θ is a measure on $(\Omega_\Theta, \mathscr{F}_\Theta)$ such that

$$Q_\Theta << \nu_\Theta << \lambda$$

Let $P_\Theta = (1-\varepsilon)Q_\Theta + \varepsilon\delta_{\Theta_0}$ and $\Pi_\Theta(\theta) = \frac{dQ_\Theta}{d\nu_\Theta}$.

Theorem 2.5.2 — 2.6. Suppose that P_Θ is defined as above. Then, for $a \in A_i$,

$$P_{\Theta|X}(\{\theta_0\}|x) = \frac{\varepsilon f(x|\theta_0)}{(1-\varepsilon)\int_{\Omega_\Theta} f(x|\theta)\pi_\Theta(\theta)d\nu_\Theta(\theta) + \varepsilon f(x|\theta_0)}$$

Proof. Let $\mu_\Theta = (1-\varepsilon)\nu_\Theta + \varepsilon\delta_{\Theta_0}$. We want to show that $P_\Theta << \mu_\Theta$ and find $\frac{dP_\Theta}{d\mu_\Theta}$. But in order to find this derivative we'll have to guess and check!

$$\tau_\Theta(\theta) = \begin{cases} \pi_\Theta(\theta) & \theta \neq \theta_0 \\ 1 & \theta = \theta_0 \end{cases}$$

We want to check $\tau_\Theta = \frac{dP_\Theta}{d\mu_\Theta}$.
Let $B \in \mathscr{F}_\Theta$,

$$\begin{aligned}
\int_B \tau_\Theta(\theta) d\mu_\Theta(\theta) &= \int_B \tau_\Theta(\theta) d((1-\varepsilon)\nu_\Theta + \varepsilon\delta_{\theta_0}) \\
&= (1-\varepsilon)\int_B \tau_\Theta d\nu_\Theta + \varepsilon\int \tau_\Theta d\delta_{\theta_0} \\
&= \cdots \int_{B\setminus\{\theta_0\}} \tau_\Theta d\nu_\Theta + \ldots \\
&= \cdots \int_{B\setminus\{\theta_0\}} \Pi_\Theta d\nu_\Theta + \ldots \\
&= \cdots \int_B \Pi_\Theta d\nu_\Theta + \ldots \\
&= (1-\varepsilon)\int_B \Pi_\Theta d\nu_\Theta + \varepsilon\int \tau_\Theta d\delta_{\theta_0} \\
&= (1-\varepsilon)\int_B \Pi_\Theta d\nu_\Theta + \varepsilon\delta_{\theta_0}(B) \\
&= (1-\varepsilon)Q_\Theta(B) + \varepsilon\delta_{\theta_0}(B) \\
&= P_\Theta(B)
\end{aligned}$$

Therefore by RN Theorem

$$P_\Theta << \mu_\Theta, \frac{dP_\Theta}{d\mu_\Theta}$$

the posterior density from $\tau_\Theta(\theta)$ (the prior density) and the likelihood gives us,

$$\tau_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)\tau_\Theta(\theta)}{\int_{\Omega_\Theta} f_{X|\Theta}(x|\theta)\tau_\Theta d\mu_\Theta(\theta)}$$

So we have that

$$\begin{aligned} P_{\Theta|X}(\{\theta_0\}|x) = \int_{\{\theta_0\}} \tau_{\Theta|x}(\theta|x)d\mu_\Theta(\theta) \\ = \frac{\int_{\{\theta_0\}} f(x|\theta)\tau_\Theta(\theta)d\mu_\Theta(\theta)}{\int_{\Omega_\Theta} f_{X|\Theta}(x|\theta)\tau_\Theta d\mu_\Theta(\theta)} \end{aligned}$$

By Lemma 2.1,

$$\begin{aligned} \int_{\{\theta_0\}} f(x|\theta)\tau_\Theta(\theta)d\mu_\Theta(\theta) &= (1-\varepsilon)\int_{\{\theta_0\}} \ldots d\nu_\Theta(\theta) + \varepsilon \int_{\{\theta_0\}} \ldots d\delta_{\theta_0} \\ &= 0 + \varepsilon f(x|\theta_0) \end{aligned}$$

$$\int_{\Omega_\Theta} f_{X|\Theta}(x|\theta)\tau_\Theta d\mu_\Theta(\theta) = (1-\varepsilon)\int f(x|\theta)\pi(\theta)d\mu(\theta) + \varepsilon \int f(x|\theta)\pi(\theta)d\delta_{\theta_0}(\theta)$$

Therefore,

$$P_{\Theta|X}(\{\theta_0\}|x) = \frac{\varepsilon f(x|\theta_0)}{(1-\varepsilon)\int f(x|\theta)\pi(\theta)d\nu(\theta) + \varepsilon f(x|\theta_0)}$$

■

Corollary 2.2 Consider setting

$$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$$

and action $\mathscr{A} = \{a_0, a_1\}$ where we accept and fail to reject H_0, respectively.
Suppose the loss function is

$$L(\theta, a) = \begin{cases} 0 & (\theta, a) \in (\{\theta_0\} x \{a_0\}) \bigcup (\{\theta_0\}^C x \{a_1\}) \\ C_0 & (\theta, a) \in (\{\theta_0\} x \{a_1\}) \\ C_1 & (\theta, a) \in (\{\theta_0\}^C x \{a_0\}) \end{cases}$$

and the prior is

$$P_\Theta = (1-\varepsilon)Q_\Theta + \varepsilon\delta_{\theta_0}$$

the Bayes' rule is

$$\frac{\varepsilon f(x|\theta_0)}{(1-\varepsilon)\int f(x|\theta)\pi(\theta)d\nu(\theta) + \varepsilon f(x|\theta_0)} < \frac{C_1}{C_0 + C_1}$$

■ **Example 2.4** Suppose that

$$X_1,\ldots,X_n|\theta \overset{iid}{\sim} Exp(\theta)$$

Then the likelihood

$$f(x|\theta) = \theta^{-n}e^{-t(x)/\theta}, \quad \theta > 0$$

$$t(x) = \sum_{i=1}^{n} X_i$$

Suppose that $\theta \sim \tau\chi^{-2}_{(m)}$. Then (you may check)

$$\theta|\underline{X} \sim (2t(x)+\tau)\chi^{-2}_{(m+2n)}$$

Suppose we want to test

$$H_0 : \theta \leq a, H_1 : \theta > a$$

The Bayes' rule: reject if

$$P(\Omega^{(1)}_\Theta|x) > \frac{C_0}{C_0+C_1}$$

Note that

$$\begin{aligned} P(\Omega^{(1)}_\Theta|x) = P(\theta > a|x) & \\ &= 1 - P(\theta \leq a|x) \\ &= 1 - P\left(\frac{\theta}{2t(x)+\tau} \leq \frac{a}{2t(x)+\tau}|x\right) \\ &= 1 - F(\frac{a}{2t(x)+\tau}) \end{aligned}$$

where F is the c.d.f. of $\chi^{-2}_{(2n+m)}$.
We want to solve

$$a = (2t(x)+\tau)F^{-1}(\frac{C_1}{C_0+C_1})$$

■

Wednesday February 22

Suppose we want to test

$$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$$

Under the $0-C_0-C_1$ loss and the slab & spike prior.
The Bayes' rule: reject if

$$\frac{\varepsilon f(x|\theta_0)}{(1-\varepsilon)\int_{\Omega_\Theta} f(x|\theta)\pi(\theta)d\nu(\theta) + \varepsilon f(x|\theta_0)} < \frac{C_1}{C_0+C_1}$$

$$f(x|\theta_0) = \theta_0^{-1}e^{-t(x)/\theta_0}$$

$$\pi(\theta) = \tau\chi^{-2}_{(\nu)} \quad \text{slab}$$

$$\pi(\theta) = \frac{1}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}(\frac{\theta}{\tau})^{-\frac{\nu}{2}-1}e^{-\frac{\tau}{2\theta}}\tau^{-1}$$

Take the integral from the rejection denominator, and using the information above

$$\int_{-\infty}^{\infty}\theta^{-n}e^{-t(x)/\theta}\frac{1}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}(\frac{\theta}{\tau})^{-\frac{\nu}{2}-1}e^{-\frac{\tau}{2\theta}}\tau^{-1}d\theta = \text{Algebra to isolate new pdf (integrates to 1)}$$
$$= \frac{\Gamma(\frac{2n+2}{2})}{\Gamma(\frac{\nu}{2})}\tau^{\frac{\nu}{2}}(2t(x)+\tau)\frac{2n+\nu}{2}2^n$$

2.6 Classificaiton

Ω_Θ is a finite set true lables of classes.

$$\Omega_\Theta = \{1,\ldots,k\}$$
$$\Omega_A == \{1,\ldots,k\}$$

PHOTO
The loss function,
$L : \{1,\ldots,k\}x\{1,\ldots,k\}$
PHOTO
The $0-C_0-C_1$ is a sepcial case in this.
PHOTO
The 0-1 loss is also a further special case where $C_{01} = C_{10} = 1$

Theorem 2.6.1 The Bayes' Rule for $(\Omega_\Theta, \Omega_A, L)$ is

$$d_B(x) - \arg\min\left\{\sum_{\theta=1}^{k} C_{\theta a}f_{x|\theta}(x|\theta)\pi_\Theta(\theta) : a = 1,\ldots,k\right\}$$

Proof. Because the los if inte we need to minimize,

$$\rho(x,a) : a = 1,\ldots,k$$

Recall that,

$$\begin{aligned}\rho(x,a) &= \mathrm{E}(L(\theta,a)|X)\\ &= \sum_{\theta=1}^{k} L(\theta,a)\pi(\theta|x)\\ &= \sum_{\theta=1}^{k} C_{\theta a}\pi(\theta|x)\\ &= \sum_{\theta=1}^{k} C_{\theta a}f(x|\theta)\pi(\theta)\end{aligned}$$

Above, note that

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

■

usually, there is a training set and a testing test. In the training set, we know the 'labels', whereas in the testing set we do not.

We have for label θ,

$$X_{\theta 1}, \ldots, X_{\theta n_\theta} \overset{iid}{\sim} P_{X|\theta}$$

where $\theta = 1, \ldots, k$. An example would be handwritting for letters/numbers.

Generally, there are extra parameters Ψ_θ that determines the shape of clan θ.

$$X_{\theta 1}, \ldots, X_{\theta n_\theta} \sim P_{X|\theta,\Psi_\theta}$$

Most commonly used models for $P_{X|\theta,\Psi_\theta}$ is

1. LDA - $N(\mu_\theta, \Sigma)$
2. QDA - $N(\mu_\theta, \Sigma_\theta)$

The parameter, Φ_θ is estimated from training sets, whereas $\pi(\theta)$ is estimated by

$$\frac{n_\theta}{\sum_{\theta=1}^k n_\theta}$$

Plug these into the Bayes then you have LDA, QDA. So, under first model if we use UMVUE to estimate μ_θ, Σ, then

$$\hat{\mu_\theta} = \frac{1}{n_\theta} \sum_{i=1}^{n_\theta} X_{\theta i}$$

$$\hat{\Sigma} = \frac{1}{\sum_{\theta=1}^k n_\theta - k} \sum_{\theta=1}^{k} \sum_{i=1}^{n_\theta} (x_{\theta i} - \hat{\mu}_\theta)(x_{\theta i} - \hat{\mu}_\theta)^T$$

Plug in Bayes rule and we get LDA.

For the second model,

$$\hat{\mu_\theta} = \frac{1}{n_\theta} \sum_{i=1}^{n_\theta} X_{\theta i}$$

$$\hat{\Sigma} = \frac{1}{n_\theta - 1} \sum_{\theta=1}^{k} \sum_{i=1}^{n_\theta} (x_{\theta i} - \hat{\mu}_\theta)(x_{\theta i} - \hat{\mu}_\theta)^T$$

Plug this into Bayes rule and get QDA.

Suppose we make a plot using QDA,

PHOTO

2.7 Stein's Estimate

Stein's Estimate

Stein (1956) shows that if the dimension is greater than 3 and sample size is 1, then MLE is not admissible.

Friday February 24

Lemma 2.2 Suppose that

$$X \sim N(\mu, \sigma^2), g : \mathbb{R} \to \mathbb{R}$$

where g is differentiable,

$$\mathrm{E}|g(x)| < \infty$$

Then we have that

$$\mathrm{Cov}(X, g(X)) = \mathrm{Var}(x)\mathrm{E}g^{\cdot}(x)$$

Proof. First we assume that

$$X \sim N(0,1)$$

Let ϕ be the pdf of $N(0,1)$. Recall that

$$\phi(x) = -x\phi(x)$$

$$\mathrm{E}[g^{\cdot}(x)] = \int_{-\infty}^{\infty} g^{\cdot}(x)\phi(x)dx$$

Let $a \in \mathbb{R}$, then this integral above becomes

$$\begin{aligned}
\int_{-\infty}^{a} g^{\cdot}(x)\phi(x)dx + \int_{a}^{\infty} g^{\cdot}(x)\phi(x)dx &= \int_{-\infty}^{a} g^{\cdot}(x)(\int_{-\infty}^{x} \phi^{\cdot}(z)dz)dx + \int_{a}^{\infty} g^{\cdot}(x)(\int_{x}^{\infty} \phi^{\cdot}(z)dz)dx \\
&= \int_{-\infty}^{a}\int_{-\infty}^{x} g^{\cdot}(x)\phi^{\cdot}(z)dzdx + \int_{a}^{\infty}\int_{x}^{\infty} g^{\cdot}(x)\phi^{\cdot}(z)dzdx \\
&= \int_{-\infty}^{a}\int_{z}^{a} g^{\cdot}(x)\phi^{\cdot}(z)dxdz + \int_{a}^{\infty}\int_{z}^{a} g^{\cdot}(x)\phi^{\cdot}(z)dxdz \\
&= \int_{-\infty}^{a} \phi^{\cdot}(z)(g(a) - g(z))dz + \int_{a}^{\infty} \phi^{\cdot}(z)((g(z) - g(z))dz) \\
&= \int_{-\infty}^{\infty} \phi^{\cdot}(z)(g(a) - g(z))dz \\
&= g(a)\int_{-\infty}^{\infty} \phi^{\cdot}(z)dz - \int_{-\infty}^{\infty} \phi^{\cdot}(z)g(z)dz \\
&= \int zg^{\cdot}(z)\phi(z)dz \\
&= \mathrm{E}(Xg(x)) \\
&= \mathrm{Cov}(X, g^{\cdot}(x)) \\
&= E(g^{\cdot}(x))
\end{aligned}$$

More generally, if $X \sim N(\mu, \sigma^2)$ then $X = \sigma Z + \mu$ where $Z \sim N(0,1)$.

$$\mathrm{Cov}(X, g(X)) = \sigma\mathrm{Cov}(Z, g_1(z))$$

Note that

$$
\begin{aligned}
g_1(z) &= g(\sigma z+\mu) && (2.5)\\
&= \sigma \mathrm{E}(g_1^{\cdot}(z)) && (2.6)\\
&= \sigma \mathrm{E}(\sigma g_1^{\cdot}(\sigma Z+\mu)) && (2.7)\\
&= \sigma^2 \mathrm{E}(g^{\cdot}(\sigma Z+\mu)) && (2.8)\\
&= \sigma^2 \mathrm{E}(g^{\cdot}(x)) && (2.9)\\
&= \mathrm{Var}(x)\mathrm{E}g^{\cdot}(x) && (2.10)
\end{aligned}
$$

■

Lemma 2.3 (HW #4). If

$$X \sim N(\mu,\Sigma), g : \mathbb{R}^P \to \mathbb{R}^p$$

where g is differentiable such that $\frac{\partial g}{\partial x^T}$ has integrable components. Then

$$\mathrm{Cov}(X, g(X)) = \Sigma \mathrm{E}\left(\frac{\partial g^T(X)}{\partial X}\right)$$

Theorem 2.7.1 — 2.8. Suppose that

$$X \sim N(\theta, I_P)$$

where $P \geq 3$.
Let

$$L(\theta, a) = ||\theta - a||^2$$

and d(x) = X. Then we have that d is inadmissable. ("Very weird")

Proof. Let $d_1(x) = (1 - g(X))X$ where $h : \mathbb{R}^P \to \mathbb{R}$ is to be specified later.

$$
\begin{aligned}
R(\theta, d_1) &= \mathrm{E}||\theta - d_1(X)||^2\\
&= \mathrm{E}||X - \theta - h(x)X||^2\\
&= \mathrm{E}||X-\theta||^2 - 2\mathrm{E}[(X-\theta)^T h(X)X] + \mathrm{E}[h^2(X)||X|]\\
&= \mathrm{E}||X-\theta||^2 - 2tr[\mathrm{Cov}(X, h(X)X)] + \mathrm{E}[h^2(X)||X|]\\
&= \mathrm{E}||X-\theta||^2 - 2tr[I_P \mathrm{E}\frac{\partial h(X)}{\partial X}] + \mathrm{E}[h^2(X)||X|]\\
&= \mathrm{E}||X-\theta||^2 - 2\mathrm{E}(X^T(\frac{\partial h}{\partial X} + ph(X))) + \mathrm{E}[h^2(X)||X|]\\
&= R(\theta, d) - 2\mathrm{E}(X^T(\frac{\partial h}{\partial X} + ph(X))) + \mathrm{E}[h^2(x)||X||^2]
\end{aligned}
$$

Let $h(x) = \frac{\alpha}{||x||^2}$. Then

$$\frac{\partial h}{\partial X} = \alpha \frac{\partial}{\partial X}(||X||^2)^{-1} \tag{2.11}$$

$$= \alpha(-1)(||X||^2)^{-2}\frac{dX^T X}{dX} \qquad = -\alpha||X||^{-4}2x \tag{2.12}$$

$$= -\alpha 2X||X||^{-2} \tag{2.13}$$

So if we plug this into equations above we get

$$||X||^{-2}\alpha(4-2P+\alpha)$$

So we have that

$$R(\theta,d_1) - R(\theta,d) = \alpha(4-2P+\alpha)\mathrm{E}||X||^2$$

If $4-2P+\alpha < 0$ then d is inadmissible.
Any $0 < \alpha < 2P-4$ makes this happen. Note we need $P \geq 3$. ■