



Advanced Statistical Inference

STAT 561 - Advanced Statistical Inference

Dr. Bing Li

Copyright © 2013 John Smith

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2013

Contents

I	Part One	
1	Basic Ideas in Bayesian Analysis	7
1.1	Frequentist & Bayesian Settings	14
1.2	Prior Posterior & Likelihood	15
1.3	Conditional Independence and Frequentist/Bayesian Sufficiency	15
1.4	Equivalence of Frequentist & Bayesian Sufficiency	18
1.5	Conjugate Priors	21
1.5.1	Introduction	21
1.5.2	Exponential Family	21
1.5.3	Convex Hull of Conjugate Family	23
1.6	Two-Parameter, Normal Family	24
1.7	Noninformative Prior	27
1.7.1	General Concept	27
1.7.2	Invariant Prior	28
1.7.3	Jeffreys' Prior	32
1.8	Statistical Decision Theory	33



Part One

1	Basic Ideas in Bayesian Analysis	7
1.1	Frequentist & Bayesian Settings	
1.2	Prior Posterior & Likelihood	
1.3	Conditional Independence and Frequentist/Bayesian Sufficiency	
1.4	Equivalence of Frequentist & Bayesian Sufficiency	
1.5	Conjugate Priors	
1.6	Two-Parameter, Normal Family	
1.7	Noninformative Prior	
1.8	Statistical Decision Theory	

1. Basic Ideas in Bayesian Analysis

Mathematical Preparation

Monday January 9

1. Product σ -Field

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

Let $\mathcal{A} = \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$.

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

2. Product Measure

Let $E \in \mathcal{F}_1 \times \mathcal{F}_2$. 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 $E \in \mathcal{F}_1 \times \mathcal{F}_2$ then $E_1(\omega_2) \in \mathcal{F}_1$ for all $\omega_2 \in \Omega_2$. Similarly, $E_2(\omega_1) \in \mathcal{F}_2$ for all $\omega_1 \in \Omega_1$.

If $f : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ measurable $\mathcal{F}_1 \times \mathcal{F}_2 \setminus \mathcal{R}$. Then for each $\omega_1 \in \Omega_1$,

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

$$f(\cdot, \omega_2) \otimes \mathcal{F}_1 \setminus \mathcal{R}$$

Now, for each $E \in \mathcal{F}_1 \times \mathcal{F}_2$ consider

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

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

Proof. Outline.

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

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

$$\text{So, } \mu_2(E_2(\omega_1)) = \begin{cases} \mu_2(B) & \omega_1 \in A \\ 0 & \omega_1 \notin A \end{cases} = I_A(\omega_1)\mu(B) = f_{1,E}$$

So, $f_{1,E} \otimes \mathcal{F}_1$.

Thus $\mathcal{P} \subseteq \mathcal{L}$.

- By $\pi - \lambda$ Theorem, $\mathcal{F}_1 \times \mathcal{F}_2 \subseteq \mathcal{L}$.

Similarly, $f_{2,E} \otimes \mathcal{F}_2 \setminus \mathcal{R}$.

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

$$\pi' = \pi'' \text{ on } \mathcal{F}_1 \times \mathcal{F}_2$$

Note that here, \mathcal{P} equals \mathcal{A} used at beginning of notes.

We did not have a measure in $\mathcal{F}_1 \times \mathcal{F}_2$. Now we have π', π'' both measures on $\mathcal{F}_1 \times \mathcal{F}_2$, they are the same. We call this measure the product measure, written as $\mu_1 \times \mu_2$.

Note that $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2)$ is called product measure space. ■

3. Tonelli's Theorem

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

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

Suppose we have $f : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R} \otimes \mathcal{F}_1 \times \mathcal{F}_2 \setminus \mathcal{R}$. Where $f \geq 0$ and

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

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 $P(A|\mathcal{G})$ when $\mathcal{G} \subseteq \mathcal{F}$ as sub- σ -field. We defined this intuitively in elementary probability course (definition above), but we are not going to define it generally.

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

$$\nu : \mathcal{G} \rightarrow \mathbb{R}, G \mapsto P(AG)$$

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

$$\mu : \mathcal{G} \rightarrow \mathbb{R}, G \mapsto P(G)$$

So μ is nothing but P restricted on \mathcal{G} .

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

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

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

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

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 \mathcal{G} . To distinguish it from $P(A|B)$, where B is a set, we use $P(A|\mathcal{G})$, where \mathcal{G} is a σ -field. By construction,

- (a) δ is measurable \mathcal{G}
- (b) $\int_G \delta d\mu = P(AG) \quad \forall G \in \mathcal{G}$

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 $P(A|\mathcal{G})$ if and only if (a) and (b) are satisfied. We may define $P(A|\mathcal{G})$ 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) $P(\emptyset|\mathcal{G}) = 0, P(\Omega|\mathcal{G}) = 1$ a.s. P

(b) $0 \leq P(A|\mathcal{G}) \leq 1$ a.s. P

(c) If A_1, A_2, \dots are disjoint members of \mathcal{F} then $P(\bigcup_n A_n|\mathcal{G}) = \sum_n P(A_n|\mathcal{G})$ a.s. P

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

$$T : \Omega \rightarrow \Omega_T \text{ in } \mathcal{F} \setminus \mathcal{F}_T \quad \mathcal{G} = T^{-1}(\mathcal{F}_T)$$

Here, we write

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

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

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

(a) f is measurable \mathcal{G}

(b) $\int_G f dp = P(AG) \quad \forall G \in \mathcal{P}$

6. Conditional Distribution

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

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

$$f : \mathcal{F}_X \times \Omega \rightarrow \mathbb{R}$$

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

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

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

The only condition for this existence is $(\Omega_X, \mathcal{F}_X)$ must be a Borel Space, that is \mathcal{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 \mathbb{R} .

We want to define conditional expectation of X given \mathcal{G} .

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

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

It can be easily shown that ν is a measure.

Let μ again be $\mathcal{G} \rightarrow \mathbb{R}, G \mapsto P(G)$. Then $\nu \ll \mu$. 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 $X \not\geq 0$, 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 $d : \Omega \rightarrow \mathbb{R}$ is a version of $E(X|\mathcal{G})$.

- (a) δ measurable \mathcal{G}
- (b) $\int_G \delta dP = \int_G X dP \quad \forall G \in \mathcal{G}$

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

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

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

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

- (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 $X_n \rightarrow X$ a.s. P , $|X_n| \leq Y$, and Y integrable P , then

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

Proof. Found in Billingsly. ■

Theorem 1.0.4 — 34.4 in Billingsly. If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ 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 \mathcal{G} , $Y \in \mathcal{F}$, then

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

Other Properties

- (a) X, Y are random elements such that XY integrable P .
- (b) If $\mathcal{G} \subseteq \mathcal{F}$ is the sub σ -field, then

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

Conditional expectation is a self-adjoint operation.

Proof. "Wire Theorem"

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

■

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,

$$h : \mathcal{F}_X \times \Omega_Y \rightarrow \mathbb{R}$$

This function is called the conditional distribution of X given Y if

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

We say that $\tilde{h} : \mathcal{F}_X \times \Omega_Y \rightarrow \mathbb{R}$ is the conditional distribution of X given $\mathcal{G} = Y^{-1}(\mathcal{F}_Y)$.

That is,

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

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

- (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

$$g : \Omega_X \times \Omega_Y$$

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

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

In the following special case, g has an explicit formula.

$$\begin{aligned}
&(\Omega, \mathcal{F}, P) \\
&(\Omega_X, \mathcal{F}_X, \mu_X) \\
&(\Omega_Y, \mathcal{F}_Y, \mu_Y) \\
&(\Omega_X \times \Omega_Y, \mathcal{F}_X \times \mathcal{F}_Y, \mu_X \times \mu_Y) \\
&(X, Y) : \Omega \rightarrow \Omega_X \times \Omega_Y \subseteq \mathcal{F} \setminus \mathcal{F}_X \times \mathcal{F}_Y
\end{aligned}$$

Let $P_X = PX^{-1}, P_Y = PY^{-1}, P_{XY} = P(XY)^{-1}$.

Assume $P_X \ll \mu_X, P_Y \ll \mu_Y, P_{XY} \ll \mu_X \times \mu_Y$.

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

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

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

Let

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

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

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 $A \in \mathcal{F}_X$,

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

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 $G' \in \sigma(y)$

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

But note that

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

So we want to check that

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

■

1.1 Frequentist & Bayesian Settings

We have our probability space (Ω, \mathcal{F}, P) . We also have some data,

$$\begin{aligned}
&(\Omega_X, \mathcal{F}_X, \mu_X) \\
X : \Omega &\rightarrow \Omega_X \text{ (mod } \mathcal{F} / \mathcal{F}_X)
\end{aligned}$$

Here, usually Ω_X is a \mathbb{R}^m .

Typically we have

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

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 statistical 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, \mathcal{P} , where $\mathcal{P} \ll \mu_X$. Usually we assume that \mathcal{P} is a parametric family, $\mathcal{P} = \{P_\theta : \theta \in \Omega_\theta \subseteq \mathbb{R}^p\}$. We assume that $P_0 \in \mathcal{P}$, that is there exists $\theta_0 \in \Omega_\theta$ 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 distribution $P_{X|\theta}$. 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 (Ω, \mathcal{F}, P) ; σ -finite measurable spaces $(\Omega_X, \mathcal{F}_X, \mu_X)$, $(\Omega_\theta, \mathcal{F}_\theta, \mu_\theta)$. Together,

$$(\Omega_X \times \Omega_\theta, \mathcal{F}_X \times \mathcal{F}_\theta, \mu_X \times \mu_\theta)$$

Also, a random element,

$$(X, \theta) : \Omega \rightarrow \Omega_X \times \Omega_\theta \subseteq \mathcal{F} \setminus \mathcal{F}_X \times \mathcal{F}_\theta$$

$P_X = P \circ X^{-1} \leftarrow$ 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
 $P_{X|\theta}(A|\theta) : \mathcal{F}_X \times \Omega_\theta \rightarrow \mathbb{R}$. Likelihood distribution
 $P_{\theta|X}(G|x) : \mathcal{F}_\theta \times \Omega_X \rightarrow \mathbb{R}$. Posterior distribution

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

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

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

1.3 Conditional Independence and Frequentist/Bayesian Sufficiency

Independence

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

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

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

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

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

Also note that

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

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

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

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

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

Conditional Independence

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

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

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

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

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

Equivalently,

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

Equivalent Condition for Conditional Independence

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

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

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

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

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

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

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

So it suffices to show that

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

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

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

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 $\sigma(\Theta^{-1}(\mathcal{F}_\Theta) \cup \mathcal{G})$.

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

It is trivial to show that \mathcal{P} is a π -system.

MORE IN PHOTO

Meanwhile,

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

Therefore,

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

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

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

So, we want

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

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

Monday January 23

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

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

is conditional probability of

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

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

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

■

1.4 Equivalence of Frequentist & Bayesian Sufficiency

Here we have,

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

Where

$$T : \Omega_X \rightarrow \Omega_T \text{ } \mathbb{M} \mathcal{F}_X / \mathcal{F}_T$$

is called a statistic.

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

In frequentist setting, we say that T is **sufficient** 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" frequentist. 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 posterior 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

$$\begin{aligned} X|\Theta &\sim N(\Theta, \sigma^2) \\ \Theta &\sim N(\mu, \tau^2) \end{aligned}$$

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

Solution:

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

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 $f_X(x)$?

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

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...

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

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

$$\begin{aligned}
X_1, \dots, X_n | \theta &\sim^{iid} N(\theta, \sigma^2) \\
\theta &\sim N(\mu, \tau^2)
\end{aligned}$$

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

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 $n \rightarrow \infty$, depending on which coefferient goes to 1,

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

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

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

We can generalize/approximate this phenomenon as follows,

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

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

In our special case,

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

■

1.5 Conjugate Priors

1.5.1 Introduction

In general, computing posterior distributions or posterior means is a hard problem involving high-dimensional numerical integration. This was a big hurdle for Bayesian methods before computers. Now we can do methods such as Monte Carlo integration (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, \mathcal{P} , on $(\Omega_\theta, \mathcal{F}_\theta)$ is a **conjugate family** if

$$P_\theta \in \mathcal{P} \Rightarrow \mathcal{P}_{\theta|X} \in \mathcal{P}$$

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

1.5.2 Exponential Family

■ **Example 1.3** $X_1, \dots, X_n | \theta \sim \text{Pois}(\theta)$

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

Note that we are treating θ as the variable of interest here, not x . Recall that if $\theta \sim \text{Gamma}(\alpha, \beta)$ 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

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

This is generally true for all exponential family distributions.

■

We say that X has exponential family distribution if it 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 $(\Omega_X, \mathcal{F}_X)$.

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

Theorem 1.5.1 — 1.3 from class. If

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

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

then

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

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

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

$$\pi(\theta|x) = \text{PHOTO}$$

$$=$$

GIANT fraction 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, $\theta|X \sim \text{Exp}(\xi(\alpha) + t(x), \phi(\theta), \frac{dv(\theta)}{\int e^{\phi^T(\theta)t(x)d\mu(x)})}$

■

Friday January 27

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

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

Want to assign conjugate prior for θ .

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

Recall Lemma 1.1.

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

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

So, consider the following family,

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

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

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

So we have that

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

■

1.5.3 Convex Hull of Conjugate Family



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

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

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

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

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

Since $P_\Theta \in \text{Conv}(\mathcal{P})$ there exists

$$\alpha_1, \dots, \alpha_k \in \mathbb{R}, \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0$$

$$P_\Theta^{(1)} \dots P_\Theta^{(k)} \in \mathcal{P}$$

such that

$$P_\Theta = \sum_{i=1}^k \alpha_i P_\Theta^{(i)}$$

$$\begin{aligned} f_X(x) &= \int f(x|\theta) dP_\Theta \\ &= \int f(x|\theta) d\left(\sum_{i=1}^k \alpha_i P_\Theta^{(i)}\right) \\ &= \sum_{i=1}^k \alpha_i \int f(x|\theta) dP_\Theta^{(i)} \\ &= \sum_{i=1}^k \alpha_i \int m_i(x) dP_\Theta^{(i)} \\ dP_\Theta(\cdot|x) &= \frac{f(x|\theta)}{f_X(x)} dP_\Theta \\ &= \frac{f(x|\theta)}{\sum_{i=1}^k \alpha_i m_i(x)} \sum_{i=1}^k \alpha_i dP_\Theta^{(i)} \\ &= \sum_{i=1}^k \frac{\alpha_i m_i(x)}{\sum_{j=1}^k \alpha_j m_j(x)} \frac{f(x|\theta) dP_\Theta^{(i)}}{m_i(x)} \end{aligned}$$

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

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

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 Distribution. 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

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

By Jacobian theorem,

$$T = g(X)$$

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

Above will be shown in HW.

Theorem 1.6.1 If $T \sim \chi_{(v)}^{-2}$ then,

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

Proof. Shown in HW 2. ■

Now suppose that $X_1, \dots, X_n | \phi, \lambda \sim^{iid} N(\lambda, \phi)$ 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 sufficiency 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 $\pi(\lambda, \phi | X_1, \dots, X_n) = \pi(\lambda, \phi | T, S)$. Moreover, from Normal theory,

$$T \perp\!\!\!\perp X | \lambda, \phi$$

$$T | \lambda, \phi \sim N(\lambda, \phi/n)$$

$$S | \lambda, \phi \sim \phi \chi_{(n-1)}^2$$

Monday January 30

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

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

So we have $N(t, \frac{\phi}{n}) s \chi_{(n-3)}^{-2}$.

More generally, the distribution,

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

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

$$\begin{aligned} a_3 &= \frac{m_1 a_1 + m_2 a_2}{m_1 + m_2} \\ m_3 &= m_1 + m_2 \\ \tau_3 &= \tau_1 + \tau_2 + m_1 a_1^2 + m_2 a_2^2 - m_3 a_3^2 \\ k_3 &= k_1 + k_2 - 3 \end{aligned}$$

Before proof, let's rewrite likelihood as

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

Proof. By definition

$$\begin{aligned} NICH_1 * NICH_2 &\propto \phi^{-1/2} \exp\left\{\frac{-1}{2(\phi/m_1)} \lambda^2 + \frac{a_1}{\phi/m_1} \lambda\right\} \phi^{-(k_1/2)-1} \exp\left\{-\frac{\tau_1 + m_1 a_1^2}{2\phi}\right\} * \phi^{-1/2} \exp\left\{\frac{-1}{2(\phi/m_2)} \lambda^2 + \frac{a_2}{\phi/m_2} \lambda\right\} \phi^{-(k_2/2)-1} \exp\left\{-\frac{\tau_2 + m_2 a_2^2}{2\phi}\right\} \\ &= \phi^{-1/2} \exp\left\{\left(\frac{-1}{2(\phi/m_1)} - \frac{1}{2(\phi/m_2)}\right) \lambda^2 + \left(\frac{a_1}{\phi/m_1} + \frac{a_2}{\phi/m_2}\right) \lambda\right\} \phi^{-(k_1/2)-1} \phi^{-(k_2/2)-1} \exp\left\{-\frac{\tau_1 + \tau_2 + m_1 a_1^2 + m_2 a_2^2 - (m_1 + m_2) a_3^2}{2\phi}\right\} \\ &= \phi^{-1/2} \exp\left\{\left(\frac{-1}{2(\phi/m_3)}\right) \lambda^2 + \left(\frac{a_3}{\phi/m_3}\right) \lambda\right\} \phi^{-(k_3/2)-1} \exp\left\{-\frac{\tau_3 + m_3 a_3^2}{2\phi}\right\} \end{aligned}$$

■

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

$$\begin{aligned} a^* &= \frac{ma + nT}{m+n} \\ m^* &= m + n \\ \tau^* &= \tau + S + ma^2 + nt^2 - (m+n)\left(\frac{ma+nt}{m+n}\right)^2 \\ k^* &= k + n - 3 + 3 \end{aligned}$$

So, $[\lambda|\phi, X] \sim N(a^*, \frac{\phi}{m^*})$, and $[\phi|X] \sim \tau^* \chi_{k^*}^{-2}$.

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 ,

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

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

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" problem. We want to use a flat prior, e.g. the Lebesgue measure. But, in what is this flat? Suppose 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

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

Here both parameters are random.

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

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

$$[\lambda | \phi] = 1$$

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

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 Section 1.5 (?) (proper case).

We can show that

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

$$\frac{s(x)}{\phi} | X \sim \chi_{(n-1)}^2$$

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 somehow 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

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

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:

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

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

We can show that a group used for distributions like $N(0, \theta)$

Left transformation:

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

if density for θ is $\pi(\theta)$.

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

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,

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

Left transformation:

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

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

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:

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

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

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

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 applications. An easily available prior, Jeffreys' prior is constructed using the following 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 $\Pi \circ T^{-1}$

Jeffreys' 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

$$\Pi(\theta) \propto \sqrt{\det I(\theta)}$$

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

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

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 generate same σ -field.

GET MORE FROM PHOTOS ■

Monday February 6

■ Example 1.10

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

where both θ & σ are unknown.

Find Jeffreys' prior for θ, σ .

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

Table 1.1: Paper Rock Scissor Loss Table

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

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

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

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

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

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

Want to take negative expectation 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 determinant 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 - $L : \Omega_{A,1} \times \Omega_{A,2} \rightarrow \mathbb{R}$

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 statistician (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

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

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

So we have

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

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

Decision rule:

$$d : \Omega_X \rightarrow \Omega_A$$

and the collection of all decision rule is \mathcal{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: $(\Omega_X, \mathcal{F}_X)$

Model: $\{P_\theta : \theta \in \Omega_\Theta\}$

Decision Rule: \mathcal{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: $(\Omega_X, \mathcal{F}_X)$

Model: $\{P_\theta : \theta \in \Omega_\Theta\}$

Decision Rule: \mathcal{D}

Prior: P_Θ

Risk: r

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

Wednesday February 8

Bayes Rule: Optimal estimation, optimal test. This is (mathematically) easier than frequentist optimal procedures because 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

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

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

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

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 . (??)

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

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

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

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

$$d : \Omega_X \rightarrow \Omega_A$$

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,

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

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 $L(\theta, a) \geq 0, \forall \theta, a$)

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

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 $r(d_B) \leq r(d) \forall d \in \mathcal{D}$. ■

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 $d' \in \mathcal{D}$ 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 $d \in \mathcal{D}$ is a **minimax rule** if for all $d' \in \mathcal{D}$,

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

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

Theorem 1.8.2 — 1.8 in Notes. Suppose

1. for each $d \in \mathcal{D}$, $R(\theta, d)$ is integrable with respect to P_Θ
2. for any $d_1, d_2 \in \mathcal{D}$,

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

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 admissible.

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

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

■

