Graduate Mathematical Statistics Notes

Fangyuan Lin

June 23, 2025

Contents

0.1 Introduction

0.1.1 Topics of the Course

1. Statistical Models: $(P_\theta : \theta \in \Theta)$, a parametrized model. We have n data points

$$X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$$

   (a) Sufficiency and Exponential Family.

      i. Factorization

      ii. Minimal Sufficiency: is it possible to keep information while compressing the data.

      iii. Ancillary Statistic

      iv. Completeness

      v. Rao-Blackwell Theorem: a consequence of sufficiency. If you use an estimator not based on a sufficient statistic, it can always be improved.

2. Decision Theory: Compare the performance of different estimators.

   (a) Loss function: $l(\hat{\theta}, \theta)$, the distance between the estimated parameter and the true parameter. It is itself a random variable.

   (b) Risk: $\mathbb{E}l(\hat{\theta}, \theta)$

   (c) Bayes and Minimax Optimality

   (d) Admissibility

   (e) James-Stein Estimator: considered the most interesting topic in this course. Application in optimal adaptive non-parametric estimators.

   (f) Neyman-Pearson Lemma

   (g) Minimax Lower Bound: used to argue that estimatotion error is at least something: Le Cam two-point method. Estimation is always going to be harder than testing - a lower bound for the testing problem implies a lower bound for the estimation problem.

3. Estimation under Constraints

   (a) Unbiasedness assumption: UMVUE, Lehmann-Scheffe

   (b) Invariance: location family, Pitman Estimator

4. Likelihood and Asymptotics

   (a) Consistency of MLE

   (b) Fisher info and score.

   (c) LAN and DQM

   (d) Cramer Rao Lower bound: (People use this to justify asymptotic optimality of MLE but it's not true?)

   (e) Hodges estimator

   (f) Convolution Theorem and Local Asymptotic Minimaxity

   (g) Bernstein-von Mises theorem

0.1.2 Recommended Textbooks

1. E. Lehmann and G. Casella, *Theory of Point Estimation*: Covers section 1, 2 and part of section 3.
2. E. Lehmann and J. Romano, *Testing Statistical Hypotheses*: Will only use some pages.
3. I. Johnstone, *Gaussian Sequence Model*: Very important and relevant to current research.
4. A. van der Vaart Asymptotic Statistics: the book the instructor uses everyday in his research - should read very carefully every page of it.

0.2 Statistical Model/Experiment

Statistical Model/Experiment

A statistical model/experiment is a collection of probability distributions

$$P_\theta : \theta \in \Theta$$

Also we have data/observations

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$$

We usually assume i.i.d. observations.

Statistic

A statistic or estimator is a function of data

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

We should think of statistic as a summary of the data, or a way to compress the data.

A natural requirement is that we don't want to throw away some of the data, e.g. the statistic only uses the first observation. The idea of sufficiency gives a rigorous way to characterize no-information-loss.

Sufficient Statistic

T is sufficient iff and the conditional distribution of $X|T$ does not depend on θ.

- Why is this a good definition and how do we interpret it?
- Image that we have two statisticians Alice and Bob. We give Alice the raw data $X_1, \ldots, X_n$ but we give Bob a summary/function of the data $T = T(X_1, \ldots, X_n)$. Now who has more information? Well, the information Alice has is not less than the information Bob has. However, if T is

sufficient, then Bob has no less information.

- Bob's strategy: sample $\tilde{X}_1, \ldots, \tilde{X}_n$ from the conditional distribution $X|T$. The marginal joint distribution of the new data $(\tilde{X}_1, \ldots, \tilde{X}_n)$ is the same as $(X_1, \ldots, X_n)$.

Gaussian Example

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, 1), \quad T(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is sufficient.

$$\left.\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}\right| \bar{X} \sim N\begin{pmatrix} \begin{pmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix}, I_n - \frac{1}{n}1_n 1_n^T) \end{pmatrix}$$

$$I_n - \frac{1}{n}1_n 1_n^T = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & & \\ -\frac{1}{n} & 1-\frac{1}{n} & & \\ & & \ddots & \\ & & & 1-\frac{1}{n} \end{bmatrix}$$

Note that to see $\mathbb{E}[X_1|\bar{X}] = \bar{X}$, write

$$\mathbb{E}(\bar{X}|\bar{X}) = \bar{X} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By symmetry, the conditional expectation of X_i given $\bar{X}$ are all the same, and their average is equal to $\bar{X}$, so they are all equal to $\bar{X}$.
The covariance matrix is related to Schur formula.
Bob can sample

$$\begin{pmatrix} \tilde{X} \\ \vdots \\ \tilde{X} \end{pmatrix} \sim N\begin{pmatrix} \begin{pmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix}, I_n - \frac{1}{n}1_n 1_n^T) \end{pmatrix}$$

which has the same distribution as $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$. We can check manually this by seeing that

$$\mathbb{E}\tilde{X}_1 = \mathbb{E}[\mathbb{E}[\tilde{X}_1|\bar{X}]] = \mathbb{E}\bar{X} = \theta$$

For the second moment, note that it's equal to mean squared plus variance:

$$\mathbb{E}[\tilde{X}_1{}^2] = \mathbb{E}[\mathbb{E}[\tilde{X}_1{}^2|\bar{X}]] = \mathbb{E}[1 - \frac{1}{n} + \bar{X}^2] = 1 - \frac{1}{n} + \frac{1}{n} + \theta^2 = 1 + \theta^2$$

$$Var(\tilde{X}) = \mathbb{E}\tilde{X}_1{}^2 - (\mathbb{E}\tilde{X}_1)^2 = 1 + \theta^2 - \theta^2 = 1$$

We next compute the cross moment $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$:

$$\mathbb{E}[\tilde{X}_1\tilde{X}_2] = \mathbb{E}[\mathbb{E}[\tilde{X}_1\tilde{X}_2|\bar{X}]] = \mathbb{E}(-\frac{1}{n} + \bar{X}^2) = \mathbb{E}(-\frac{1}{n} + \frac{1}{n} + \theta^2) = \theta^2$$

Therefore,

$$\mathrm{Cov}(\tilde{X}_1, \tilde{X}_2) = \theta^2 - \theta^2 = 0$$

Therefore, we see that $\tilde{X}$ follows the same distribution as X. (Mean and Covariance are all we need to characterize Gaussian.)

Bernoulli Example

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta), \quad T(X) = \sum_{i=1}^{n} X_i$$

is sufficient. We consider the following quantity.

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)}$$

$$\begin{aligned} P(X = x, T = t) &= \begin{cases} P(X = x) & \sum_{i=1}^{n} X_i = t \\ 0 & \sum X_i \neq t \end{cases} \\ &= 1_{\sum_{i=1}^{n} X_i = t} P(X = x) \\ &= 1_{\sum X_i = t} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= 1_{\sum X_i = t} \theta^t (1-\theta)^{n-t} \end{aligned}$$

$$P(T = t) = \binom{n}{t} \theta^T (1-\theta)^{n-t}$$

Therefore,

$$P(X = x|T = t) = 1_{\sum X_i = t} \frac{1}{\binom{n}{t}}$$

which does not depend on θ.

Arbitrary Distribution Example

Consider observations from an arbitrary probability distribution and the *order statistic*

$$X_1, \quad , X_n \overset{i.i.d.}{\sim} P_\theta, \quad T = (X_{(1)}, \dots, X_{(n)}),$$

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Well this is a function of the data. Some information is lost since if we are given the order statistic, we cannot get back to the original data. The question is: even

if we lose information, do we lose information relevant to θ? The answer is no and we can show that the order statistic is always **sufficient**.
The verification is very easy. All we need to do is to consider

$$X_1, \ldots, X_n | X(1), X_{(2)}, \ldots, X_{(n)}$$

Given the order statistic, $(X_1, \ldots, X_n)$ has $n!$ possibilities since they must be a permutation of the order statistic and by symmetry, each permutation has equal probability. Therefore, $X_1, \ldots, X_n | X(1), X_{(2)}, \ldots, X_{(n)}$ is a uniform distribution over all the $n!$ permutations. If Bob is given the order statistic, he can just shuffle the order statistic and get $\tilde{X}$ that has the same distribution as the raw data.
If the data are not independently sampled, the order statistic is no longer sufficient.

Uniform Example

Consider observations from a uniform distribution on the interval $(0, \theta)$:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} Uniform(0, \theta), \quad T(X_1, \ldots, X_n) = \max_{1 \le i \le n} X_i = X_{(n)}$$

is actually sufficient.
We can argue that by consider the order statistic, and note that

$$X_{(1)}, \ldots, X_{(n-1)} | X_{(n)} = t$$

is an order statistic from $n-1$ i.i.d. samples from $Uniform(0, t)$.
Bob can sample the remaining $n-1$ data from Uniform distribution on $(0, t)$.

Discussion question: Should we always use sufficient statistic and throw away the data?

- Infomration-Theoretic perspective: Yes
- Computation perspective: No, you need to sampling artificial data from $X|T$ and sampling can be NP hard. (Montanari 2015, Bresler, Gramamik and Shah 2014)

0.3 Review: Sufficiency

Recall the definition of sufficient statistics: Suppose we have a distribution parametrized by θ:

$$(P_\theta, \theta \in \Theta), \quad X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$$

$T = T(X_1, \ldots, X_n)$ is called suffient iff $X|T$ does not dependent on θ.

An Alternative Bayesian Definition of Sufficiency

T is sufficient if and only if

$$\theta \to T \to X$$

forms a Markov chain, i.e.

$$\theta \perp X | T$$

A useless remark: Note that $\theta \to X \to T$ is always a Markov chain.

The following theorem is very easy to use in practice.

Factorization Theorem

Suppose $(P_\theta : \theta \in \Theta)$ is continuous or discrete (has pdf or pmf), then T is sufficient if and only if

$$p(X|\theta) = g_\theta(T(X))h(X)$$

for some function g_θ and h.

- If given T, the value of g_θ is deterministic.

Proof. We present the proof for the discrete case. Assume that the factorization condition holds, i.e.

$$P(X|\theta) = g_\theta(T(X))h(X).$$

Let's check T is sufficient:

$$\begin{aligned}
P(X = x|T = t) &= \frac{P(X = x, T = t)}{P(T = t)} \\
P(X = x, T = t) &= \begin{cases} P(X = x) & T(x) = t \\ 0 \quad T(x) \neq t \end{cases} \quad = \mathbf{1}_{T(x)=t}P(X = x) \\
&= \mathbf{1}_{T(x)=t}g_\theta(T(X))h(X) \\
&= \mathbf{1}_{T(x)=t}g_\theta(t)h(X)
\end{aligned}$$

Let's now look at the denominator and we use the law of total probability.

$$\begin{aligned}
P(T = t) &= \sum_{x':T(x')=t} p(x'|\theta) \\
&= \sum_{x':T(x')=t} g_\theta(T(x'))h(X) \\
&= \sum_{x':T(x')=t} g_\theta(t)h(x') \\
&= g_\theta(t) \sum_{x':T(x')=t} h(x')
\end{aligned}$$

The radio (conditional probability) is independent of θ because $g_\theta(t)$ gets cancelled out.

$$P(X = x|T = t) = \frac{\mathbf{1}_{T(x)=t}h(x)}{\sum_{x':T(x')=t} h(x')}$$

does not dependent on θ, so T is sufficient.
Now suppose that T is sufficient.

$$P(x|\theta) = P_\theta(X = x)$$

Note that it is equal to

$$P_\theta(X = x) = P_\theta(X = x, T(X) = T(x))$$

Now we can factorize this joint distribution into conditional distribution and the marginal distribution.

$$\begin{aligned} &P_\theta(X = x|T(X) = T(x))P_\theta(T(X) = T(x)) \\ &= h(x)g_\theta(T(x)) \end{aligned}$$

This first factor does not depend on θ by the sufficiency of T. □

Factorization Theorem on i.i.d. Normal

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. Then

$$\begin{aligned} P(X|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i^2)-\frac{1}{2}n\theta^2+\theta\sum_{i=1}^{n}X_i} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i^2)} e^{-\frac{1}{2}n\theta^2+\theta\bar{X}} \end{aligned}$$

Therefore, $\bar{X}$ is sufficient.

Factorization Theorem on i.i.d. Uniform Distribution

Let X_i be iid uniform distribution on the interval $(, \theta)$. Then

$$\begin{aligned} p(x|\theta) &= \prod_{i=1}^{n} (\frac{1}{\theta}\mathbf{1}_{0<X_i<\theta}) \\ &= \theta^{-n}\prod_{i=1}^{n}\mathbf{1}_{0<x_i<\theta} \\ &= \theta^{-n}\mathbf{1}_{0<\min_i x_i, \max_i x_i<\theta} \\ &= \theta^{-n}\mathbf{1}_{0<\min_i x_i}\mathbf{1}_{\max_i x_i<\theta} \end{aligned}$$

Therefore, $\max_i x_i$ is sufficient.

0.4 Exponential Family

Exponential Family

A distribution p (pmf or pdf) is in the exponential family if

$$p(x|\theta) = \exp\left(\sum_{j=1}^{d}\eta_j(\theta)T_j(x) - B(\theta)\right)h(x)$$

where η is called natural parameter, a function of the underlying parameter θ. T_j is a sufficient statistics. $B(\theta)$ is a normalizing factor, i.e.

$$B(\theta) = \log \int e^{\sum_{j=1}^{d} \eta_j(\theta) T_j(x)} h(x) d\mu(x).$$

$h(x)$ is called the base measure.

Exponential Family and Exponential Distribution

The exponential distribution $\exp(\theta)$ belongs to the exponential function.

$$\begin{aligned} p(x|\theta) &= \theta e^{-x\theta} \mathbf{1}_{x \geq 0} \\ &= \exp(-\theta x + \log \theta) \mathbf{1}_{x \geq 0} \end{aligned}$$

Here θ is the natural parameter. x is the sufficient statistic. $\log(\theta)$ is the log-partition function. The indicator is the base measure.

Exponential Family and Gaussian Distribution

Consider $N(\mu, \sigma^2)$ where

$$\begin{aligned} p(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \\ &= \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log\left(2\pi\sigma^2\right)\right) \end{aligned}$$

Most common distributions are in the exponential family. The exponential family is a convenient concept when we consider i.i.d. observations, where the joint likelihood is

$$p(x_1, \ldots, x_n|\theta) = \exp\left(\sum_{j=1}^{d} \eta_j(\theta) \left(\sum_{i=1}^{n} T_j(x_i)\right)\right) \prod_{i=1}^{n} h(x)$$

Note that this is still an exponential family where the sufficient statistic is the sum

$$T = \left(\sum_{i=1}^{n} T_1(x_i), \ldots, \sum_{i=1}^{n} T_d(x_i)\right)$$

The sufficient statistic is still d dimensional, so you can always compress your data into d dimension.

Canonical Form of Exponential Family

An exponential family distribution p is of the canonical form if

$$p(x|\eta) = \exp\left(\sum_{j=1}^{d} \eta_j T_j(x) - A(\eta)\right) h(x)$$

where the natural parameter $\eta = \theta$ is the identity function. $A(\eta)$ is the normalizing function:

$$\log \int e^{\sum_{j=1}^{d} \eta_j T_j(x)} h(x) d\mu(x)$$

0.4.1 Minimal Exponential Family

We should make sure d is minimized and if so, the exponential family is called minimal.

Minimal Exponential Family (Informal)

An exponential family $(P_\eta : \eta \in H)$ (of canonical form) is minimal if its dimension cannot be reduced.
(This is not a formal definition)

A non-minimal example

Let

$$\begin{aligned} p(x|\eta) &= \exp(\eta_1 T(x) + \eta_2(3T(x) + 2) - A(\eta)) \\ &= \exp((\eta_1 + 3\eta_2)T(x) + 2\eta_2 - A(\eta)) \end{aligned}$$

In this example, we reduced the dimension of the exponential family from 2 to 1. This happened because the sufficient statistics are linearly dependent.
Now if the natural parameters are linearly dependent, then we can also reduce dimension:

$$p(x|\eta) = \exp(\eta T_1(x) + (4 - 5\eta)T_2(x) - A(\eta)) \tag{1}$$
$$= \exp(\eta(T_1(x) - 5T_2(x)) - A(\eta)) \exp(4T_2(x)) \tag{2}$$

0.4.2 Canonical Form

Now we present the formal definition of canonical form.

Formal Definition of Canonical Form

An exponential family $(P_\eta, \eta \in H)$ (of canonical form) is minimal if its sufficient statistics are linearly independent and natural parameters are linearly independent.

There are two types of minimal exponential families.

1. Full rank: the parameter space H contains an open d-dimensional rectangle.

2. Curved: The natural parameters $\eta_1, \ldots, \eta_d$ are related in non-linear ways.

For example:

Normal Distribution Example

$$p(x|\mu,\sigma^2) = \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right) \tag{3}$$

Let

1. $$T_1(x) = -x^2, \quad T_2(x) = x$$

2. $$\eta_1 = \frac{1}{2\sigma^2}, \quad \eta_2 = \frac{\mu}{\sigma^2}$$

   Let's consider a weird Poisson-like example, $N(\sigma^2, \sigma^2)$. We get that

   $$\eta_2 = 1$$

   and the expression becomes non-minimal and $N(\sigma^2, \sigma^2)$ is a one-dimensional exponential family.
   Now let's consider $\mu = \sqrt{\sigma^2}$. Then

   $$\eta_1 = \frac{1}{2\sigma^2}, \quad \eta_2 = \frac{1}{\sqrt{\sigma^2}}$$

   The natural parameters are related in a non-linear way, so we cannot reduce the dimension further. $N(\sqrt{\sigma^2}, \sigma^2)$ a 2-dimensional curved exponential family.
   Now if there is no constraint on μ and σ^2, then the exponential family is minimal and full rank.

   $$H = (0, \infty) \times \mathbb{R}$$

To summarize, non-minimal exponential families are over-parameterized.

0.4.3 Minimal Sufficiency

Minimally Sufficient

S is minimally sufficient if and only if for every sufficient T, S is a function of T.

Example of minimally sufficient statistic

X_i i.i.d. $N(\theta, 1)$

1. $T_1 = (X_1, \ldots, X_n)$

2. $$T_2 = (X_1 + X_2, X_3 + X_4, \ldots, X_{n-1} + X_n)$$

3.
$$T_3 = (\sum_{i \leq n/2} X_i, \sum_{i ? n/2} X_i)$$

4.
$$T_4 = \sum_i X_i$$

They are all sufficient statistics. We see that T_4 is a function of T_1, T_2 and T_3, but not vice versa. We will later show that T_4 is minimal statistic.

0.4.4 Finding minimally sufficient statistic

Sub-Family Method

Lemma

Suppose $\Theta_0 \subset \Theta$, S is minimally sufficient for the small family $(P_\theta : \theta \in \Theta_0)$ and sufficient for the big family $(P_\theta : \theta \in \Theta)$, then it is minimally sufficient for the big family.

- To check minimal sufficiency, you only need to find a convenient sub-family and check minimal sufficiency for that small family.

Proof. The proof directly uses the definition of minimal sufficiency. Suppose T is an arbitrary sufficient statistic. Then $S = f(T)$ since S is minimally sufficient on the small family $(P_\theta : \theta \in \Theta_0)$. □

Theorem: Minimal sufficiency of likelihood ratios

Assume $(P_\theta : \theta \in \theta_0, \theta_1, \ldots, \theta_d)$ share common support, then

$$T(X) = \left(\frac{P_{\theta_1}(X)}{P_{\theta_0}(X)}, \ldots, \frac{P_{\theta_d}(X)}{P_{\theta_0}(X)} \right)$$

is minimally sufficient.

- Note that the assumption is not true for uniform distribution on $(0, \theta)$ since the support does depend on θ, but the assumption is true for Gaussian, binomial, exponential family etc.
- If $d = 1$, i.e. we only have θ_0 and θ_1, then the likelihood ratio of the distributions itself is a 1-dimensional minimally sufficient statistic.

Proof. The proof is actually easy.

1. We need to review the factorization theorem. T is sufficient if and only if the distribution of X can be factored into two parts. The first part only depends on θ through the statistic $T(X)$. The second part is function of X.
2. We can always factorize the likelihoods using the following algorithm:

(a)

$$P_{\theta_0}(X) = P_{\theta_0}(X)$$

(b)

$$P_{\theta_j}(X) = T_j(X)P_{\theta_0}(X), \quad j = 1, \dots, d$$

This is immediate from the definition of T.

3. Now define

$$g_{\theta_j}(T(x)) = \begin{cases} 1 & j = 0, \\ T_j(x) & j = 1, \dots, k \end{cases}$$

$$h(x) = P_{\theta_0}(x)$$

θ_0 can be an arbitrary element in the parameter space so we have a valid h because it does not depend on knowledge of θ.

4. Note that if a statistic T is sufficient, then

$$\frac{P(x|\theta_1)}{P(x|\theta_0)} = \frac{g_{\theta_1}(T(x))}{g_{\theta_0}(T(x))}$$

$h(x)$ gets cancelled out. The likelihood ratio only depends on x through $T(x)$.

5. Now suppose T' is an arbitrary sufficient statistic, by the above conclusion, the likelihood ratio is a function of $T'(x)$.

6. Since T is a function of likelihood ratio, T is a function of T', meaning that T is a minimally sufficient by definition.

$\square$

Bernoulli Likelihood Ratio Example

Let X_i be i.i.d. Bernoulli(θ). $\theta \in [0, 1]$.

$$\sum_{i=1}^{n} X_i$$

is a sufficient statistic.
We will now show it's minimally sufficient using the subfamily method.
Consider the subfamily $\theta_0 = 0.5, \theta_1 = 0.6$. The likelihood ratio is going to be our minimally sufficient statistic:

$$\frac{p(x|\theta_1)}{p(x|\theta_0)} = \frac{\theta_1^{\sum_{i=1}^{n} x_i}(1-\theta_1)^{n-\sum_i x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum_i x_i}}$$

It's equal to

$$\left(\frac{\theta_1}{\theta_0}\right)^{\sum x_i} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-\sum x_i} = \left(\frac{\theta_1}{\theta_0}\frac{1-\theta_1}{1-\theta_0}\right)^{\sum x_i} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n}$$

Which is equal to

$$(\frac{3}{2})^{\sum x_i}(\frac{4}{5})^n$$

This guy is minimally sufficient for the subfamily $\{0.5, 0.6\}$. Therefore it's always minimally sufficient for the original family $[0, 1]$. However, note that this is a monotonic function of the sum statistic $\sum x_i$, so it's equivalent/bijective to the sums $\sum x_i$. Therefore, $\sum x_i$ is also minimally sufficient.

Recall that $T = T(X)$ is sufficient iff $X|T$ is independent of $\theta \in \Theta$. S is minimally sufficient iff S is sufficient and for every sufficient T, S is a function T, i.e. we can compute S from T.

1. Sub-family method:
   Lemma: Suppose $\Theta_0 \subset \Theta_1$, S is minimally sufficient on Θ_0 and sufficient on Θ_1, it is also minimal sufficient on Θ_1.
   Theorem: For $(P_\theta) : \theta \in \{\theta_0, \theta_1, \ldots, \theta_d\}$ with common support.

$$T(X) = \left(\frac{P_{\theta_1}}{P_{\theta_0}}(X), \ldots, \frac{P_{\theta_d}}{P_{\theta_0}}(X)\right)$$

   is minimally sufficient.

A minimal exponential family is defined such that the dimension cannot be reduced.

Minimal Exponential Family

A minimal exponential family $\exp(\langle \eta, T(X)\rangle - A(\eta))h(X)$.

$$\eta \in H \subset \mathbb{R}^d$$

is minimal if the natural parameters η_j are not linearly dependent and the sufficient statistics $T_j(X)$ are not linearly dependent.

- Note that we used $\langle \eta, T(X)\rangle$ to represent $\sum_j \eta_j T_j(X)$.

0.5 Minimal Exponential Family and Minimal Sufficient Statistic

Theorem: Minimal exponential family and minimal sufficient statistic

The minimal exponential family $\exp(\langle \eta, T(x)\rangle - A(\eta))h(x)$.

$$\eta \in H \subset \mathbb{R}^d,$$

then

$$T(x) = (T_1(x), \ldots, T_d(x))$$

is minimally sufficient.

Proof. 1. Since the exponential family is minimal, we can find $\eta_0, \eta_1, \ldots, \eta_d \in H$

such that

$$\begin{bmatrix} (\eta_1 - \eta_0)^T \\ (\eta_2 - \eta_0)^T \\ \vdots \\ (\eta_d - \eta_0)^T \end{bmatrix} \in \mathbb{R}^{d \times d}$$

has full rank. (Note that this is a consequence of minimal exponential family.)

Illustration with d equal to 2

Consider two situations

(a) Full rank exponential family:

Full rank exponential family

An exponential family is of full rank if the following equivalent conditions are true:

i. The statistics T_i are linearly independent as functions.

ii. The parameter space H is an open set.

In this case, you can find a rectangle inside H and let η_i be the vertices. Then their differences are linearly independent.

(b) Curved exponential family: in this case, the parameters η_i are related in a non-linear way. Because of the curvature of the parameter space, we can find η_0, η_1, η_2 such that $\eta_2 - \eta_1$ and $\eta_1 - \eta_0$ are linearly independent, i.e. the two by two matrix has full rank.

2. If you have a non-minimal exponential family: This means that H is a *linear* subspace of $\mathbb{R}^d$ because the η_i are related in a linear way. Their differences are always parallel.

3. Now consider the subfamily $\{\eta_0, \eta_1, \dots, \eta_d\} \subset H$ and the minimal sufficient statistic $\frac{P(X|\eta_j)}{P(X|\eta_0)}, j = 1, \dots, d$.

$$\begin{aligned} \frac{P(X|\eta_j)}{P(X|\eta_0)} &= \frac{\exp(\langle \eta_j, T(x) \rangle - A(\eta_j))}{\exp(\langle \eta_0, T(x) \rangle - A(\eta_0))} \\ &= \exp(\langle \eta_j - \eta_0, T(x) \rangle - A(\eta_j) + A(\eta_0)) \end{aligned}$$

This is equivalent to $\langle \eta_j - \eta_0, T(x), \quad j = 1, \dots, d$. We can turn them into a column vector:

$$\begin{bmatrix} (\eta_1 - \eta_0, T(x))^T \\ \vdots \\ (\eta_d - \eta_0, T(x))^T \end{bmatrix} = \begin{bmatrix} \langle \eta_1 - \eta_0 \rangle \\ \vdots \\ \langle \eta_d - \eta_0 \rangle \end{bmatrix} T(x)$$

which, since the matrix is of full rank, is equivalent to

$$T(x) = \begin{bmatrix} T_1(x) \\ \vdots \\ T_d(x) \end{bmatrix},$$

which is minimally sufficient on the subspace $\{\eta_0, \dots, \eta_d\}$ and therefore on H.

□

With the above theorem, we can derive minimally sufficient statistic for Bernoulli, Poisson, Gaussian, etc.

0.6 Completeness

The idea of the completeness method is to remove all ancillary information.

Example

Suppose we have $X_1, X_2 \sim N(\theta, 1)$.

$$T = (X_1, X_2)$$

is sufficient but not minimal. T is a *trivial* sufficient statistic. We can use the previous theorem to show that a minimally sufficient statistic is the sum of the data, but T is not a function of the sum.
Now note that T is equivalent to $(X_1 - X_2, X_1 + X_2)$. The distribution of $X_1 - X_2$ is $N(0, 2)$, which does not dependent on θ, so it's useless when estimating θ. Therefore it's said to be **ancillary**.

Ancillary Statistic

$A = A(X)$ is *ancillary* iff its distribution does not dependent on $\theta \in \Theta$.
It is said to be *first-order ancillary* iff its expectation $\mathbb{E}_\theta A(X)$ does not dependent on $\theta \in \Theta$. (This is a weaker version).

Complete Statistic

$T = T(X)$ is complete iff

$$\mathbb{E}_\theta f(T(X)) = 0$$

implies that for any function f

$$f(T(X)) = 0 \quad a.s. \quad \forall \theta \in \Theta,$$

i.e. $P_\theta(f(T(X)) = 0) = 1$, i.e. the zero function is the only possible f.

- This means that there is no non-constant function of T is first-order ancillary.

Theorem: Bahadur

If T is sufficient and complete, then T is minimally sufficient.

Proof. Assume a minimal sufficient statistic $U = U(X)$ exists. Then by definition of minimal sufficiency, U is a function of T, $U = h(T)$. It suffices to show that T is also a function of U.
Let's now construct such function h.

1. Define
$$g(u) = \mathbb{E}_\theta(T|U = u),$$
which is a function independent of θ since it is a function of a sufficient statistic U.

2. Then,
$$\begin{aligned}\mathbb{E}_\theta g(h(T)) &= \mathbb{E}_\theta g(U) = \mathbb{E}_\theta(\mathbb{E}_\theta(T|U)) = \mathbb{E}_\theta(T) \\ &\implies \mathbb{E}_\theta(g(h(T)) - T) = 0 \quad \forall\theta \in \Theta\end{aligned}$$

3. By completeness of T,
$$g(h(T)) = T \quad a.s. \implies g(U) = T \quad a.s.$$

□

Bernoulli Example

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

$$T = \sum_{i=1}^{n} X_i \sim Binomial(n, \theta)$$

Suppose $\mathbb{E}_\theta f(T(X)) = 0$, then

$$\begin{aligned}&\sum_{i=1}^{n} f(i)\binom{n}{i}\theta^i(1-\theta)^{n-i} \\ &= \sum_{i=1}^{n} f(i)\binom{n}{i}(\frac{\theta}{1-\theta})^i(1-\theta)^n = 0 \quad \forall\theta \in (0,1) \\ &\implies \sum_{i=1}^{n} f(i)\binom{n}{i}(\frac{\theta}{1-\theta})^i = 0\end{aligned}$$

Set

$$\beta = \frac{\theta}{1-\theta}$$
$$\sum_{i=1}^{n} f(i)\binom{n}{i}\beta^i = 0 \quad \forall\beta > 0$$

This is a polynomial of degree n. It has at most n roots. But the above equation says the equation has an infinitely amount of solutions. This means that the coefficients of the polynomial must ALL be zero! This shows that T is complete.

Uniform Distribution Example

Consider $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. This is NOT an exponential family since each distribution has different support. Let

$$T = \max_i X_i$$

Let's find the distribution:

$$P(T \leq t) = \prod_{I=1}^{n} P(X_i \leq t) = \left(\frac{t}{\theta}\right)^n, \quad t \in (0, \theta)$$
$$p(t|\theta) = \frac{\mathrm{d}}{\mathrm{d}t} P(T \leq t) = -\theta^n \cdot n \cdot t^{n-1}, \quad t \in (0, \theta)$$

Suppose

$$\mathbb{E}_\theta f(T(x)) = 0, \forall \theta > 0$$

Then

$$\int_0^\theta f(t)\theta^{-n} \cdot n \cdot t^{n-1} \mathrm{d}t = 0$$
$$\implies \int_0^\theta t^{n-1} f(t) \mathrm{d}t = 0, \quad \forall \theta > 0 \quad \text{(want to show)}$$

To show that T is complete, we want to show that the function f is the zero function. We need a trick from *real analysis*. The trick is **Positive Part and Negative Part**!

$$f^+ = \max(f, 0), \quad f^- = \max(-f, 0)$$

Then f can always be decomposed into difference of positive and negative parts:

$$f = f^+ - f^-$$

Then we have that

$$\int_0^\theta t^{n-1} f^+(t) \mathrm{d}t = \int_0^\theta t^{n-1} f^-(t) \mathrm{d}t \quad \forall \theta > 0.$$
$$\implies \int_{\theta_1}^{\theta_2} t^{n-1} f^+(t) \mathrm{d}t = \int_{\theta_1}^{\theta_2} t^{n-1} f^-(t) \mathrm{d}t \quad \forall 0 < \theta_1 < \theta_2$$
$$\implies \int_A t^{n-1} f^+(t) \mathrm{d}t = \int_A t^{n-1} f^-(t) \mathrm{d}t \quad \forall \text{Borell set } A$$
$$\implies t^{n-1} f^+(t) = t^{n-1} f^-(t) \quad \text{can also derive from line 2 if have not taken measure theory}$$
$$\implies f(t) = 0 \quad a.s.$$
$$\implies T \text{ is complete.}$$

Intuitively, the above measure theoretic argument is true because both t^{n-1} and $f^+ t^{n-1} f^-$ are positive, so the integrals cannot be equal by coincidental cancellations.

Normal Distribution Example

Let $X_1, \dots X_m \overset{iid}{\sim} N(\theta, 1)$.

$$T = \frac{1}{\sqrt{n}} \sum_i X_i \sim N(\theta, 1)$$

Suppose that

$$\mathbb{E}_\theta f(T(x)) = 0 \quad \forall \theta \in \mathbb{R}.$$

which implies that

$$\begin{aligned} &\int f(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \mathrm{d}x = 0 \\ \implies &\int f(x) e^{-\frac{1}{2}x^2 + x\theta} \mathrm{d}x = 0 \quad \forall \theta \in \mathbb{R} \\ \implies &\int f^+(x) e^{-\frac{1}{2}x^2 + x\theta} \mathrm{d}x = \int f^-(x) e^{-\frac{1}{2}x^2 + x\theta} \mathrm{d}x \quad \forall \theta \in \mathbb{R} \end{aligned}$$

Take $\theta = 0$, we get that

$$\int f^+(x) e^{-\frac{1}{2}x^2} \mathrm{d}x = \int f^-(x) e^{-\frac{1}{2}x^2} \mathrm{d}x$$

which implies that

$$\frac{\int f^+(x) e^{-\frac{1}{2}x^2} e^{\theta x} \mathrm{d}x}{\int f^+(x) e^{-\frac{1}{2}x^2} \mathrm{d}x} = \frac{\int f^-(x) e^{-\frac{1}{2}x^2} e^{\theta x} \mathrm{d}x}{\int f^-(x) e^{-\frac{1}{2}x^2} \mathrm{d}x}$$

Note very importantly that these are **moment generating functions**! Same MGF implies same density, so

$$f^+ = f^- \quad a.e. \implies f = 0 \quad a.e.$$

Full Rank Exponential Family

$$e^{\sum_{j=1}^d \eta_j T_j(x) - A(\eta)} h(x), \quad \eta \in H$$

Then

$$T = (T_1(x), \dots, T_d(x))$$

is complete.

- The proof is similar to that of the normal distribution. You apply the moment generating function argument.

Completeness means that we have moved all the first order ancillary information.

Basu's Theorem

Suppose T is complete and sufficient and A is ancillary, then T and A are independent.

Proof. We want to show that

$$P_\theta(A \in B | T = t) = P_\theta(A \in B) \quad \forall t$$

Let

$$C = P_\theta(A \in B),$$

which does not dependent on θ because A is ancillary.
Let

$$g(t) = P_\theta(A \in B | T = t),$$

which does not dependent on θ either, since T is sufficient.
Now

$$\begin{aligned} \mathbb{E}_\theta(g(T) - c) &= \mathbb{E}_\theta[P_\theta(A \in B) - P_\theta(A \in B)] \\ &= P_\theta(A \in B) - P_\theta(A \in B) = 0 \quad \forall \theta \in \Theta \end{aligned}$$

Then by completeness, $g(t) = c$ a.s.. □

Power of Basu's theorem

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$. Then

$$\bar{X} \perp \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Proof. The proof of this depends on linear algebra (from undergraduate mathematical statistics which I never took, anyways. I just read the proof and I think I understood it lol). However, we know that $\bar{X}$ is sufficient and complete. The sum of difference of squares follows a χ^2_{n-1} distribution which does not dependent on θ. Therefore, Basu's theorem tells us that they are independent. □

0.7 Decision Theory

Today we are going to start a new topic: decision theory!

- Abraham Wald from Columbia established this theory. As we know, he unfortunately died in a plane crash.

Suppose we have the family $(P_\theta : \theta \in \Theta)$ and we have data $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$. Suppose we want to estimate θ with $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. Suppose we also have a loss function $L(\hat{\theta}, \theta)$, e.g. $\|\hat{\theta} - \theta\|^2$.

- Note that $L(\hat{\theta}, \theta)$ is a randomly variable. We can get rid of the randomness by taking expectation:

$$\mathbb{E}_\theta L(\hat{\theta}, \theta) = \int L(\hat{\theta}(x), \theta) p_\theta(x) \mathrm{d}x$$

This is called the *risk function* and we denote it with $R(\hat{\theta}, \theta)$.

0.7.1 Rao-Blackwell Theorem

Theorem: Rao-Blackwell

Assume $L(\hat{\theta}, \theta)$ is convex in $\hat{\theta}$, for any $\hat{\theta}$ and any sufficient statistic T, defined

$$\tilde{\theta} = \mathbb{E}_\theta(\hat{\theta}|T)$$

Then

$$R(\tilde{\theta}, \theta) \leq R(\hat{\theta}, \theta)$$

- Unless your estimator is already a function of the sufficient statistic T, this will be a strict inequality.

Proof.

1. $$\begin{aligned} L(\tilde{\theta}, \theta) &= L(\mathbb{E}_\theta(\hat{\theta}|T), \theta) \\ &\leq \mathbb{E}_\theta[L(\hat{\theta}, \theta)|T] \end{aligned}$$

   by Jensen's inequality and the convexity of the loss function L.

2. Finally, the proof is done by taking expectation of both sides.

□

- Taking conditional expectation is called Rao-Blackwellization.
- Note that T is required to be sufficient since otherwise, we are not able to compute the conditional expectation as it depends on θ.

0.8 Bayes Estimator and Minimax Estimator

Comparing Two Estimators

Suppose we have two estimators $\hat{\theta}$ and $\tilde{\theta}$. Let

$$r_1(\theta) = R(\hat{\theta}, \theta) \quad r_2(\theta) = R(\tilde{\theta}, \theta)$$

- However, we do not know the true location of θ. So how can we compare two estimators?
- One idea is to compute the **average risk**:

  $$\int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta.$$

  Note that we need an piror distribution on θ.

- Another idea is to compute the **maximum risk**:

$$\sup_{\theta\in\Theta} R(\hat{\theta}, \theta)$$

Bayes' Estimator and Minimax Estimator

$\hat{\theta}$ is called a **Bayes' estimator** if

$$\hat{\theta} = \text{argmin}_{\tilde{\theta}} \int R(\tilde{\theta}, \theta)\pi(\theta)\mathrm{d}\theta$$

which is equivalent to

$$\forall\tilde{\theta} \quad \int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta \le \int R(\tilde{\theta}, \theta)\pi(\theta)\mathrm{d}\theta.$$

$\hat{\theta}$ is called a **minimax estimator** if

$$\hat{\theta} = \text{argmin}_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\tilde{\theta}, \theta).$$

which is equivalent to

$$\forall\tilde{\theta} \quad \sup_{\theta\in\Theta} R(\hat{\theta}, \theta) \le \sup_{\theta\in\Theta} R(\hat{\theta}, \theta)$$

Let's take a look at the Baye's estimator

$$\int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta = \int\int L(\hat{\theta}(x), \theta)P_\theta(x)\pi(\theta)\mathrm{d}x\mathrm{d}\theta.$$

Note that $P_\theta(x)\pi(\theta) = P(x|\theta)\pi(\theta)$ is the joint distribution of (x, θ). It's also equal to $\pi(\theta|x)m(x)$ where $\pi(\theta|x)$ is the posterior distribution of θ and $m(x)$ is the marginal of x. Then the average risk can be represented as:

$$\int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta = \int\int L(\hat{\theta}(x), \theta)\pi(\theta|x)\mathrm{d}\theta m(x)\mathrm{d}x$$

Note that we exchanged the order of integration with a non-rigorous application of the Fubini's theorem. (Most functions in this course are nice functions so we usually just apply Fubini's theorem without any check.)

- Now note that
$$\int L(\hat{(}x), \theta)\pi(\theta|x)\mathrm{d}\theta)$$
is a function of x.

- We can find a number $\hat{\theta}_\pi(x)$ that minimizes this function:
$$\hat{\theta}_\pi(x) = \text{argmin}_a \int L(a, \theta)\pi(\theta|x)\mathrm{d}\theta.$$

- Claim: this estimator is Bayes.

- Note that in
$$\text{argmin}_{\hat{\theta}} \int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta,$$
it's a minimization over all functions. But now we are dealing a minimization over all numbers.

Proof. We want to show that for any $\hat{\theta}$
$$\int R(\hat{\theta}_\pi, \theta)\pi(\theta)\mathrm{d}\theta \le \int R(\hat{\theta},\ theta)\pi(\theta)\mathrm{d}\theta$$

$$\begin{aligned}\int R(\hat{\theta}_\pi,\ theta)\pi(\theta)\mathrm{d}\theta &= \int\int L(\hat{\theta}_\pi(x), \theta)\pi(\theta|x)\mathrm{d}\theta m(x)\mathrm{d}x \\ &\le \int\int L(\hat{\theta}(x), \theta)\pi(\theta|x)\mathrm{d}\theta m(x)\mathrm{d}x \\ &= \int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta\end{aligned}$$

by the definition of $\hat{\theta}_\pi$ □

An important example

Consider $\Theta \subset \mathbb{R}$.
$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

$$\begin{aligned}\hat{\theta}_\pi(x) &= \text{argmin}_a \int (a-\theta)^2\pi(\theta|x)\mathrm{d}\theta \\ &= \text{argmin}_a \mathbb{E}((a-\theta)^2|x)\end{aligned}$$

The solution of this minization problem is apparently
$$\mathbb{E}[\theta|x]$$

- (some useful remark: think of expectation as projection.)
- Very important fact to remember: Suppose we have a random variable $Y \in \mathbb{R}$.
$$\mathbb{E}[Y-\mu]^2 = \text{Var}(Y) + (\mathbb{E}Y - \mu)^2$$
The mean square error is the sum of variance and bias squared. We just used the conditional version of this fact.

Bernoulli Example

Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), and consider
$$L(\hat{p}, p) = (\hat{p} - p)^2$$
Consider the beta prior
$$\pi = Beta(\alpha, \beta), \quad \pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

Then

$$p|X_1, \dots, X_n \sim Beta(\sum_{X_i} +\alpha, \sum_{1-X_i} +\beta)$$

Then the Bayes estimator is

$$\hat{p} = \mathbb{E}(p|X_1, \dots, X_n) = \frac{\sum X_i + \alpha}{n + \alpha + \beta}$$

Let's now compute the risk.

$$\begin{aligned} R(\hat{p}, p) &= \mathbb{E}_p(\hat{p} - \ p)^2 \\ &= \mathrm{Var}(\hat{p}) + (\mathbb{E}_p(\hat{p} - p))^2 \\ &= (\frac{n}{n+\alpha+\beta})^2 \frac{p(1-p)}{n} + (\frac{\alpha+\beta}{\alpha+\beta+n})^2 (\frac{\alpha}{\alpha+\beta} - p)^2 \end{aligned}$$

Now let's find the minimax estimator:

$$\hat{\theta}_{minimax} = \mathrm{argmin}_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$$

- Note that this is analogous to the equilibrium of a game in game theory.
- Prof. Chao remarked that the minimax estimator is harder to find than the average estimator.

Theorem: Bayes and Minimax Estimator

Suppose for some prior distribution π, $\hat{\theta}$ satisfies that

$$\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \inf_{\tilde{\theta}} \int R(\tilde{\theta}, \theta) \pi(\theta) \mathrm{d}\theta$$

then $\hat{\theta}$ is minimax.

- To find the minimax estimator, we are actually looking for a Bayes' estimator such that the average risk is minimized for some prior distribution on θ.

  Proof. First of all,

  $$\forall \tilde{\theta}, \sup_{\theta \in \Theta} R(\tilde{\theta}, \theta) \geq \int R(\tilde{\theta}, \theta) \pi(\theta) \mathrm{d}\theta$$

  This is apparently true because this is just saying that "largest is greater than average." This inequality is always used.

  $$\begin{aligned} \sup_{\theta \in \Theta} R(\tilde{\theta}, \theta) &\geq \int R(\tilde{\theta}, \theta) \pi(\theta) \mathrm{d}\theta \\ &\geq \inf_{\tilde{\theta}} \int R(\tilde{\theta}, \theta) \pi(\theta) \mathrm{d}\theta \\ &= \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) \end{aligned}$$

  ∎

The theorem might be hard to use but it has a nice corollary and it is an important tool for finding minimax estimator.

Corollary

If $\hat{\theta} = \hat{\theta}_\pi$ for some π and $R(\hat{\theta}_\pi, \theta)$ is constant over $\theta \in \Theta$, then $\hat{\theta}$ is minimax.

Proof. Let's check the condition of the above theorem.

1. First, the worst risk is equal to the average risk because the risk is constant

$$\begin{aligned}\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) &= \int R(\hat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta \\ &= \inf_{\tilde{\theta}} \int R(\tilde{\theta}, \theta)\pi(\theta)\mathrm{d}\theta\end{aligned}$$

2. The second equality is because $\hat{\theta}$ is a Bayes estimator.

   Now the condition of theorem is satisfied and $\hat{\theta}$ is minimax.

□

Bernoulli Minimax Example

Let $X_1, \dots, X_n$ be iid Bernoulli(p), and lost function $L(\hat{p}, p) = (\hat{p} - p)^2$.

- The Bayes estimator as we have found is

$$\hat{p} = \mathbb{E}(p|X_1, \dots, X_n) = \frac{\sum X_i + \alpha}{\alpha + \beta + n}$$

-

$$\begin{aligned}R(\hat{p}, p) &= \mathbb{E}_\theta(\hat{p} - p)^2 \\ &= (\frac{n}{n+\alpha+\beta})^2\frac{p(1-p)}{n} + (\frac{\alpha+\beta}{n+\alpha+\beta})^2(\frac{\alpha}{\alpha+\beta} - p)^2\end{aligned}$$

- This is a quadratic function of p:

$$\begin{aligned}&(\frac{n}{n+\alpha+\beta})^2\frac{p(1-p)}{n} + (\frac{\alpha+\beta}{n+\alpha+\beta})^2(\frac{\alpha}{\alpha+\beta} - p)^2 \\ &= \left[\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2 - \frac{1}{n}\left(\frac{n}{n+\alpha+\beta}\right)^2\right]p^2 \\ &+ \left[\frac{1}{n}\left(\frac{n}{n+\alpha+\beta}\right)^2 - \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2\frac{2\alpha}{\alpha+\beta}\right]p \\ &+ \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2\left(\frac{\alpha}{\alpha+\beta}\right)^2\end{aligned}$$

- To make the Bayes estimator constant, we need both the quadratic and linear terms to be zero.

- $$\begin{cases} (\alpha+\beta)^2 = n \\ 2\alpha(\alpha+\beta) = n \end{cases}$$

  Let's solve these equations. We can take the square of the second equation to get that

  $$n = 4\alpha^2, \quad \alpha = \frac{\sqrt{n}}{2}, \quad b = \frac{\sqrt{n}}{2}$$

  Therefore,

  $$\hat{p}_{minimax} = \frac{\sum_i X_i + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$$

  Note that the MLE is $\hat{p}_{MLE} = \bar{X}$.

  - Let's compare the minimax and the MLE estimators:

    $$R(\hat{p}_{MLE}, p) = \mathbb{E}_p(\hat{p}_{MLE}, p) = \mathbb{E}_p(\hat{p} - p)^2 = \frac{p(1-p)}{n}$$

    $$R(\hat{p}_{minimax}, p) = \frac{1}{4(\sqrt{n}+1)^2}$$

    Note that $\max_p R(\hat{p}_{MLE}, p) = \frac{1}{4n}$. Therefore, $\hat{p}_{minimax}$ is doing a little better than the MLE in terms of maximum risk.

Bernoulli Example with a Different Loss Function

Let's normalize the loss function with the Fisher information. Let $X_1, \ldots, X_n$ be iid Bernoulli(p) with loss function

$$L(\hat{p}, p) = \frac{(\hat{p} - p)^2}{p(1-p)}$$

Let the prior be the uniform prior: $\pi(p) = 1$.

$$\hat{p}(x) = \text{argmin}_a \int \frac{(a-p)^2}{p(1-p)} \pi(p|x) \text{d}p$$

1. Note that

   $$\text{argmin}_a \int \frac{(a-p)^2}{p(1-p)} \pi(p|x) \text{d}p = \text{argmin}_a \int (a-p)^2 \frac{\pi(p|x)}{p(1-p)} \text{d}p$$

   $$\frac{\pi(p|x)}{p(1-p)} \propto p^{\sum_{i=1}^n X_i - 1}(1-p)^{\sum_i (1-X_i) - 1} = Beta(\sum_{i=1}^n X_i, \sum_i (1 - X_i))$$

2. Therefore,

   $$\hat{p}(x) = \frac{\sum X_i}{\sum X_i + \sum (1 - X_i)} = \bar{X} = \hat{p}_{MLE}$$

3. Now let's look at the risk:

   $$R(\hat{p}, p) = \mathbb{E}_p \frac{(\bar{X} - p)^2}{(p(1-p)} = \frac{1}{n}$$

which is a constant. Therefore, $\hat{p} = \bar{X}$ is minimax.

4. Note that the result is very different from the last example, but we only normalized the loss function this time.

Normal Distribution Example

Let $X_1, \dots, X_n$ iid $N(\mu, \sigma^2), \quad \mu \in \mathbb{R}^2$. The loss function is

$$(\hat{\mu} - \mu)^2$$

Note that there's no question that the square error is the most natural choice of loss function for normal distribution, since the Fisher information is a constant.

- Question: Is $\bar{X}$ minimax?

$$R(\bar{X}, \mu) = \mathbb{E}_\mu (\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$$

  Note that this is not Bayes estimator since it's unbiased and ordinary squared-error loss, Bayes estimator must be biased.

- Our current tool box is not enough to prove this minimax. We need new tools to show that this is minimax.

0.9

In last session, we talked about decision theory where we have data $X \sim P_\theta$ and parameter space $(P_\theta : \theta \in \Theta)$. We have a loss function to quantify the error of an estimator: $L(\hat{\theta}, \theta)$ and the risk function:

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta L(\hat{\theta}, \theta) = \int L(\hat{\theta}(x), \theta) \mathrm{d}P_\theta(x)$$

Normal Distribution Example

Consider $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad \mu \in \mathbb{R}$ with loss function $(\hat{\mu} - \mu)^2$.

- Question: Is $\bar{X}$ minimax?
- Well $R(\bar{X}, \mu) = \mathbb{E}_\theta (\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$.
- Let's first ask: Is $\bar{X}$ Bayes? Consider $\pi = N(0, \iota^2)$.

$$\pi(\mu | X) \propto \pi(\mu) \prod_{i=1}^n p(X_i | \mu) \propto e^{-\frac{\mu^2}{2\iota^2} - \frac{1}{2} \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}}$$

  Define

$$f(\mu) = \frac{\mu^2}{\iota^2} + \sum_i \frac{(X_i - \mu)^2}{\sigma^2}$$

$$f'(\mu) = \frac{2\mu}{\iota^2} + \frac{1}{\sigma^2}\sum_i 2(\mu - X_i) = 0$$

$$\iff \frac{\mu}{\iota^2} + n\frac{\mu}{\sigma^2} = \frac{1}{\sigma^2}\sum X_i$$

- This implies that our Bayes estimator is

$$\mathbb{E}[\mu|X] = \frac{\frac{1}{\sigma^2}\sum X_i}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}} = \frac{\frac{n}{\sigma^2}}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}\bar{X}$$

- Observe that no matter what ι we choose, $\mathbb{E}[\mu|X]$ is not equal to $\bar{X}$; therefore, $\bar{X}$ is never a Bayes estimator.

- Let's look at the the risk:

$$\begin{aligned} R(\hat{\mu}, \mu) &= \text{Var}(\hat{\mu}) + (\mathbb{E}[\hat{\mu}] - \mu^2) \\ &= \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}\right)^2 \frac{\sigma^2}{n} + \left(\frac{\frac{1}{\iota^2}}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}\right)^2 \mu^2 \end{aligned}$$

- To find the average risk, we integrate the risk:

$$\int R(\hat{\mu}, \mu)\pi(\mu)\mathrm{d}\mu = \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}\right)^2 \frac{\sigma^2}{n} + \left(\frac{\frac{1}{\iota^2}}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}\right)^2 \iota^2 = \frac{1}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}$$

- To prove $\bar{X}$ is minimax, Note that that $\sup_{\mu\in\mathbb{R}} R(\bar{X}, \mu) = \frac{\sigma^2}{n}$.

-
$$\begin{aligned} \forall \hat{\mu}, \sup_{\mu\in\mathbb{R}} R(\hat{\mu}, \mu) &\geq \int R(\hat{\mu}, \mu)\pi(\mu)\mathrm{d}\mu \\ &\geq \inf_{\hat{\mu}} \int R(\hat{\mu}, \mu)\pi(\mu)\mathrm{d}\mu \\ &= \frac{1}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}} \end{aligned}$$

- Letting $\iota^2 \to \infty$ on both sides, we get that

$$\lim_{\iota^2\to\infty} \sup_{\mu\in\mathbb{R}} R(\hat{\mu}, \mu) = \sup_{\mu\in\mathbb{R}} R(\hat{\mu}, \mu) \geq \lim_{\iota^2\to\infty} \frac{1}{\frac{1}{\iota^2} + \frac{n}{\sigma^2}}$$

  Therefore

$$\sup_{\mu\in\mathbb{R}} R(\hat{\mu}, \mu) \geq \frac{\sigma^2}{n} = R(\bar{,}\mu)$$

The idea of the above example leads to the following theorem:

Theorem

If there exist prior distributions $\{\pi_m\}$ such that

$$\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \lim_{m \to \infty} \inf_{\hat{\theta}} \int R(\hat{\theta}, \theta) \pi_m(\theta) \mathrm{d}\theta$$

then $\hat{\theta}$ is minimax.