Graduate Mathematical Statistics Notes

Fangyuan Lin

June 11, 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

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

We usually assume i.i.d. observations.

Statistic

A statistic or estimator is a function of data

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

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 $X_1, \dots, X_n$ but we give Bob a summary/function of the data $T = T(X_1, \dots, X_n)$. 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.

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

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

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

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

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

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

is sufficient. We consider the following quantity.

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

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

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

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*

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

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

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 $(0, \theta)$:

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

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:

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

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

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

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

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

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

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

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.

□

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 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 $\sum_j \eta_j T_j(X)$.

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.

$\square$