Statistical Inference.

Q. What is the average height of a BSDS student?

parameter of interest

population.

We observe the heights of the current batch of students.

sample.

Goal: To make inference about the parameter based on the sample at hand.

To connect the gap between the sample and the population, we make use of statistical models.

Assume: The height of a BSDS student is H. $H \sim N(\mu, \sigma^2)$.

The observations we have

$H_1, H_2, \cdots, H_{62} \overset{iid}{\sim} H$

$H_1, H_2, \cdots, H_{62} \overset{iid}{\sim} N(\mu, \sigma^2)$

Statistical modeling

Average height of a BSDS student

$\mathbb{E}(H) = \mu$

We want to estimate μ based on $H_1, \cdots, H_{62}$.

Eg. What is the average lifetime of a Phillips LED bulb?

parameters of interest — population

If Phillips gives a one-year warranty on its bulbs, what proportion of bulbs need to be replaced?

parameters of interest

observations:

Fix a time period T (eg. 1000 hours).

Let 500 bulbs run for T hours.

If a bulb is fused, we know when it got fused.

If it is not fused, then we only know that it lasted for T hours.

The Lifetime of a Phillips LED bulb is L

$$L \sim \text{Exponential}(\text{mean} = \lambda).$$

We observe

$$\tilde{L}_1, \ldots, \tilde{L}_{500}$$

where $\tilde{L}_i = \begin{cases} L_i & \text{if } L_i \leq T \\ T & \text{if } L_i \geq T \end{cases}$ Truncated at T.

$\tilde{L}_1, \ldots, \tilde{L}_{500}$ are iid.

They have the same distribution as L truncated at T.

The distribution of $\tilde{L}$ involves λ and T

From this, we can get to λ.

In statistical inference,

we have a population. (an infinite one).

we have a sample from the population (finite).

We postulate certain (statistical) model on the population.

Eg. $H \sim N(\mu, \sigma^2)$, $L \sim Exp(\text{mean} = \lambda)$

We also make certain assumptions on the sample collection mechanism.

Eg. $H_1, \ldots, H_{50} \overset{i.i.d}{\sim} H$. $\tilde{L}_1, \ldots, \tilde{L}_{500} \overset{i.i.d}{\sim} \tilde{L} = \begin{cases} L \text{ if } L \le T \\ T \text{ o.w.} \end{cases}$

Our task is to make inference about the unknown parameters of the population based on the sample.

Parametric model: When we assume the form of the population distribution to be known (except for the unknown parameters).

Nonparametric model: We do not assume the form of the distribution to be known.

Eg. $X \sim F$ where F is continuous.

$X \sim f$, and $E(X)$ exists

$E(X^4)$ exists.

We will focus on parametric models.

Throughout the course, we will assume we have observations

$$X_1, \dots, X_n \overset{iid}{\sim} f_\theta$$

where f_θ is a known parametric form.

The form of f_θ is known upto the parameter θ, we know f_θ completely as soon as we know θ.

Eg. $f_\theta = N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$$f_{(\mu,\sigma^2)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$

Eg. f_θ = Exponential (mean $= \lambda$), $\theta = \lambda$

$$f_\lambda = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, \quad x > 0.$$

Eg. f_θ = Binomial (n, p) $\quad \theta = p$

$$f_p = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n$$

~~We have a parametric model~~

~~$X_1, \cdots, X_n \overset{iid}{\sim} f_\theta$, θ is the unknown parameter.~~

We have a parametric model

$$X_1, \dots, X_n \overset{iid}{\sim} f_\theta .$$

θ is the unknown parameter.

We want to make inference about θ.

- **Estimation**: Estimate the value of θ based on $X_1, \dots X_n$

  - want a single value (point estimation)
  - want an interval where θ lies with high probability (interval estimation)

- **Testing**: We want to check whether θ presents to some preconceived values that we have in mind.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Binomial}(5, p)$

We may want to estimate p. $\frac{\bar{X}}{5}$.

$$\left(\frac{\bar{X}}{5} - c, \frac{\bar{X}}{5} + c\right)$$

We may want to test whether the value of p equals 0.1

$$H_0: p = 0.1 \quad \text{vs} \quad H_1: p \neq 0.1$$

$$H_0: p < 0.1 \quad \text{vs.} \quad H_1: p \geq 0.1$$

$$>$$
$$<$$

In our statistical model, the values of the parameters can vary within certain sets. These are known as the parameter space.

Eg. $f_\theta = N(\mu, \sigma^2)$. $\mu \in \mathbb{R}$, $\sigma > 0$

Eg. $F_\theta = Exp(\text{mean}=\lambda)$, $\lambda > 0$.

Eg. $f_\theta = Bin(5, p)$, $p \in [0,1]$

In this course, we will focus on optimal inferential procedures.

Note: Optimality of the procedure comes as a result of our modeling assumption.

The same procedure which is optimal under one setup may not be good under a different setup.

Grading

Mid-term	—	30
End-term	—	50
2x class tests	—	20
		100

References:

- Statistical Inference - Casella & Berger
- Introduction to the Theory of Statistics - Mood, Graybill & Boes
- Introduction to Mathematical Statistics - Hogg, McKean, Craig
- Theory of Point Estimation - Lehmann & Casella
- Testing Statistical Hypotheses - Lehmann & Romano

$X_1, \dots, X_n \overset{iid}{\sim} f_\theta$, θ unknown parameter

$\theta \in \Theta$ ← parameter space.

We want to do inference on θ.

(a) To estimate θ

(b) Testing hypotheses involving θ.

X_i's take values in $\mathcal{X}$ ← sample space.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \in [0,1]$ (Θ)

$X_i \in \{0,1\} \leftarrow \mathcal{X}$

To estimate p.

To test. $p = \frac{1}{2}$. vs. $p \neq \frac{1}{2}$.

$p > \frac{1}{2}$ / $p < \frac{1}{2}$.

Eg. $X_1, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$

$\underset{\sim}{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$

$(-\infty, \infty) \times (0, \infty)$

$X_i \in \mathbb{R}$

$X_i \in (-\infty, \infty) \leftarrow \mathcal{X}$.

(H)

To estimate $\underset{\sim}{\theta} : \mu, \sigma^2$.

To test $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0 \left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix} \right)$ v/s. $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}_0$

To test $\mu = \mu_0$ vs $\mu \neq \mu_0$

In this case, σ^2 is a nuisance parameter.

$X_1, \dots, X_n \overset{iid}{\sim} F_\theta$, $\theta \in (H)$

The first step is to summarize the data.

data reduction.

Eg. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ← sample mean

A reduction from $\mathbb{R}^n$ to $\mathbb{R}$.

Eg. $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ ← sample variance

A reduction from $\mathbb{R}^n$ to $\mathbb{R}_+$

Eg. The minimum or the maximum values.

$X_{(1)} = \min_{i=1,\dots,n} X_i$, $X_{(n)} = \max_{i=1,\dots,n} X_i$

Eg. The middlemost value / median

$$\begin{cases} \dfrac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ even} \\ X_{(\frac{n+1}{2})} & \text{if } n \text{ odd.} \end{cases}$$

Eg. Skewness, Kurtosis. . — — .

How to choose the summary?

Statistic: A statistic $T = T(\underline{X})$ is a measurable function of the sample.

— Something that can be computed based on the sample (it should only be based on the sample)

Measurable function: A random function $T(\underline{X})$ is measurable if we can assign probabilities to it.

In particular, $\mathbb{P}_\theta(T(\underline{X}) \in O)$ is defined for every open set O

We can also define $\mathbb{P}_\theta(T(X) \in (a,b))$

$[a,b]$, $(a,b]$, $[a,b)$
$(-\infty, b]$, $(-\infty, b)$
(a, ∞), $[a, \infty)$

$\mathbb{P}_\theta(T(\underline{X}) \in (-\infty, t])$ is well defined

$= \mathbb{P}_\theta(T(\underline{X}) \leq t)$ ← cdf of $T(\underline{X})$.

- Usually, most functions are measurable. But not all.
- All the functions that we will encounter in this course are measurable

$X_1, \ldots, X_n$: sample

$\downarrow$

$T = T(\underline{X}) = T(X_1, \ldots, X_n)$: summary statistic.
(can be a vector)

Starting from the sample we can compute the statistic

But, usually, not the other way round.

Which summary to use?

Depends on the modelling assumption.

Modelling assumption: $X_1, \dots X_n \overset{iid}{\sim} f_\theta$

We write $p_\theta(\underline{x}) = p_\theta(x_1 \dots x_n)$: joint dist of the sample

$$= \prod_{i=1}^{n} f_\theta(x_i)$$

Idea: Choose the summary statistic in such a way that no information on θ is lost in the process.

- The distribution of the sample given the value of the summary statistic does not involve θ.
- The conditional distribution of $\underset{\sim}{X}$ given $T(\underline{X})$ is free of θ.

Def: We say that a statistic $T(\underline{X})$ is **sufficient** for θ if the conditional distribution of $\underline{X}$ given $T(\underline{X})$ is free of θ.

- Knowing the statistic is **sufficient** to get all the information about θ.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \in [0,1]$

$$f_\theta(x_i) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases} = p^{x_i}(1-p)^{1-x_i}$$

$$f_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i)$$

$$= \prod_{i=1}^{n} \left\{ p^{x_i}(1-p)^{1-x_i} \right\}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

$$T(\underline{X}) = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$$

$$\mathbb{P}_\theta\left(\underline{X} = \underline{x} \mid T(\underline{X}) = t\right) = \frac{\mathbb{P}_\theta\left(\{\underline{X} = \underline{x}\} \cap \{T(\underline{X}) = t\}\right)}{\mathbb{P}_\theta\left(T(\underline{X}) = t\right)}$$

$$\mathbb{P}_\theta\left(T(\underline{X}) = t\right) = \begin{cases} \binom{n}{t} p^t (1-p)^{n-t}, & t = 0, 1, \cdots, n. \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbb{P}_\theta\left(\{\underline{X} = \underline{x}\} \cap \{T(\underline{X}) = t\}\right) = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \mathbb{P}_\theta(\underline{X} = \underline{x}) & \text{if } T(\underline{x}) = t \end{cases}$$

$$\mathbb{P}_\theta(\underline{X} = \underline{x}) = p_\theta(\underline{x}) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

$$\text{Numerator} = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ p^{\left(\sum_{i=1}^n x_i\right)^t} (1-p)^{n - \left(\sum_{i=1}^n x_i\right)^t} & \text{if } T(\underline{x}) = t \end{cases}$$

$$T(\underline{x}) = \sum_{i=1}^n x_i$$

$$\text{Numerator} = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ p^t (1-p)^{n-t} & \text{if } T(\underline{x}) = t \end{cases}$$

$$\mathbb{P}_\theta\left(\underline{X} = \underline{x} \mid T(\underline{X}) = t\right) = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \dfrac{p^t (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} & \text{if } T(\underline{x}) = t \end{cases}$$

$$= \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \dfrac{1}{\binom{n}{t}} & \text{if } T(\underline{x}) = t \end{cases}$$

So, $T(\underline{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for free of p.

p. Knowing the total no. of successes is enough for p.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$.

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}, \quad x \in \mathbb{R}$$

$$p_\theta(\underset{\sim}{x}) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2}$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i^2+\mu^2-2\mu x_i)}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{-\frac{n\mu^2}{2}} e^{\mu \sum_{i=1}^n x_i}$$

$$T(\underset{\sim}{X}) = \sum_{i=1}^n X_i \sim N(n\mu, n)$$

pdf of $T(\underline{X})$

$$q_\theta(t) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{1}{2n}(t-n\mu)^2} \qquad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi n}} e^{-\frac{1}{2n}(t^2 + n^2\mu^2 - 2n\mu t)}$$

$$= \frac{1}{\sqrt{2\pi n}} e^{-\frac{t^2}{2n}} e^{-\frac{n\mu^2}{2}} e^{\mu t}$$

The pdf of $\underline{X}$ given $T(\underline{x})$

$$\tilde{p}_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t) = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \dfrac{p_\theta(\underline{x})}{q_\theta(t)} & \text{if } T(\underline{x}) = t \end{cases}$$

$$= \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \dfrac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{-\frac{n\mu^2}{2}} e^{\mu \sum_{i=1}^n x_i}}{\frac{1}{\sqrt{2\pi n}} e^{-\frac{t^2}{2n}} e^{-\frac{n\mu^2}{2}} e^{\mu t}} & \text{if } T(\underline{x}) = t \end{cases}$$

$T(\underline{x}) = \sum_{i=1}^n x_i$

free of μ

$T(\underline{X}) = \sum_{i=1}^n X_i$ is sufficient for μ

The same conclusion holds if the variance σ^2 is known (not necessarily 1)

Result: A statistic $T = T(\underline{X})$ is sufficient for θ if the ratio $\dfrac{p_\theta(\underline{x})}{q_\theta(T(\underline{x}))}$ is free of θ

where $p_\theta(\underline{x})$ is the "joint pdf/pmf of the sample $\underline{X}$ and $q_\theta(t)$ is the pdf/pmf of the statistic $T(\underline{X})$.

We need to check this for all $\underline{x} \in \mathcal{X}$.

Result: (Factorization Theorem) Let $p_\theta(\underline{x})$ be the joint pdf/pmf of the sample $\underline{X}$. A statistic $T = T(\underline{X})$ is sufficient for θ if and only if we can factorize $p_\theta(\cdot)$ as

$$p_\theta(\underline{x}) = \underbrace{g_\theta(T(\underline{x}))}_{\text{involves } \theta,\ \text{only depends on the statistic } T} \underbrace{h(\underline{x})}_{\text{free of } \theta} \quad \forall \underline{x} \in \mathcal{X}.$$

Eg. $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$

$$p_\theta(\underline{x}) = p^{\sum x_i} (1-p)^{n-\sum x_i} = \left(\frac{p}{1-p}\right)^{\sum x_i} (1-p)^n$$

$$p_\theta(\underline{x}) = \begin{cases} p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}, & x_i \in \{0,1\}, i = 1, \cdots n \\ 0, & \text{otherwise} \end{cases}$$

$$= g_\theta\left(\sum_{i=1}^n x_i\right) h(\underline{x})$$

where $g_\theta(t) = \begin{cases} \left(\frac{p}{1-p}\right)^t (1-p)^n, & t = 0, 1, \cdots, n \\ 0, & \text{o.w.} \end{cases}$

$$h(\underline{x}) = \begin{cases} 1 & \text{if } x_i \in \{0,1\}, i = 1, \cdots n \\ 0 & \text{o.w.} \end{cases}$$

$T(\underline{X}) = \sum_{i=1}^n X_i$ is sufficient for p.

$X_1, \dots X_n$ ← sample

$p_\theta(\underline{x})$, $\theta \in \Theta$ ← statistical model

$T = T(\underline{X})$ is sufficient for θ if the conditional dist. of $\underline{X}$ given $T(\underline{X})$ is free of θ.

Result: (Factorization Theorem) A statistic $T = T(\underline{X})$ is sufficient for θ if and only if we can factorize

$$p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x}) \qquad \forall\, \underline{x} \in \mathcal{X},\ \theta \in \Theta$$

Proof: **If part**: We have to show:

if $p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x}) \quad \forall \underline{x} \in \mathcal{X}$

Then $T(\underline{X})$ is sufficient for θ.

We have to show: $\mathbb{P}_\theta\left(\underline{X} = \underline{x} \mid T(\underline{X}) = t\right)$ is free of θ

$$\mathbb{P}_\theta\left(\underline{X} = \underline{x} \mid T(\underline{X}) = t\right) = \begin{cases} 0 & \text{if } T(\underline{x}) \neq t \\ \dfrac{\mathbb{P}_\theta(\underline{X} = \underline{x})}{\mathbb{P}_\theta(T(\underline{X}) = t)} & \text{if } T(\underline{x}) = t \end{cases}$$

$$\frac{\mathbb{P}_\theta(\underline{X} = \underline{x})}{\mathbb{P}_\theta(T(\underline{X}) = t)} = \frac{p_\theta(\underline{x})}{\sum_{\underline{y} \in \mathcal{X} : T(\underline{y}) = t} p_\theta(\underline{y})}$$

$$\frac{\mathbb{P}_\theta(\underline{X}=\underline{x})}{\mathbb{P}_\theta(T(\underline{X})=T(\underline{x}))} = \frac{p_\theta(\underline{x})}{\sum_{\underline{y}\in\mathcal{X}:T(\underline{y})=T(\underline{x})} p_\theta(\underline{y})}$$

$$= \frac{g_\theta(T(\underline{x}))\, h(\underline{x})}{\sum_{\underline{y}\in\mathcal{X}:T(\underline{y})=T(\underline{x})} g_\theta(T(\underline{y}))\, h(\underline{y})}$$

$$= \frac{g_\theta(T(\underline{x}))\, h(\underline{x})}{g_\theta(T(\underline{x})) \sum_{\underline{y}\in\mathcal{X}:T(\underline{y})=T(\underline{x})} h(\underline{y})} \quad \leftarrow \text{free of } \theta$$

$\therefore T(\underline{X})$ is sufficient for θ.

Note: The derivation shows us

$$P_\theta(T(\underline{X}) = t) = \sum_{\underline{y} \in \mathcal{X} : T(\underline{y}) = t} p_\theta(\underline{y})$$

$$= \sum_{\underline{y} \in \mathcal{X} : T(\underline{y}) = t} g_\theta(T(\underline{y}))\, h(\underline{y})$$

$$= g_\theta(t) \sum_{\underline{y} \in \mathcal{X} : T(\underline{y}) = t} h(\underline{y}).$$

Only if part: $T(\underline{X})$ is sufficient for θ

To show: $p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x}) \quad \forall\, \underline{x} \in \mathcal{X}$

$T(\underline{X})$ is sufficient $\Rightarrow$ $\mathbb{P}_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t)$ is free of θ

$$p_\theta(\underline{x}) = \mathbb{P}_\theta(\underline{X} = \underline{x})$$

$$= \sum_{t : t = T(x) \text{ for some } x \in \mathcal{X}} \mathbb{P}_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t) \cdot \mathbb{P}_\theta(T(\underline{X}) = t)$$

i.e. $t \in$ range of T

$$= \sum_{t \text{ in range of } T} \mathbb{P}_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t) \quad \mathbb{P}_\theta(T(\underline{X}) = t)$$

$$\mathbb{P}_\theta\left(\underline{X}=\underline{x} \mid T(\underline{X})=t\right) = 0 \quad \text{if } T(\underline{x}) \neq t$$

$$\sum_{t \text{ in range } T} \mathbb{P}_\theta\left(\underline{X}=\underline{x} \mid T(\underline{X})=t\right)\ \mathbb{P}_\theta\left(T(\underline{X})=t\right)$$

$$= \sum_{t\,:\,t=T(\underline{x})} \mathbb{P}_\theta\left(\underline{X}=\underline{x} \mid T(\underline{X})=t\right)\ \mathbb{P}_\theta\left(T(\underline{X})=t\right).$$

$$= \underbrace{\mathbb{P}_\theta\left(T(\underline{X})=T(\underline{x})\right)}_{\text{a function of } \theta \text{ and } \underline{x} \text{ only via } T(\underline{x})} \underbrace{\mathbb{P}_\theta\left(\underline{X}=\underline{x} \mid T(\underline{X})=T(\underline{x})\right)}_{\text{is free of } \theta \quad h(\underline{x})}$$

$$g_\theta(T(\underline{x})) \Rightarrow p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x}).$$

Some observations

- Suppose we have a statistic $T = T(\underline{X})$

$$T : \mathcal{X} \longrightarrow T(\mathcal{X})$$

For $t \in T(\mathcal{X})$, define $A_t = \left\{ \underline{x} \in \mathcal{X} : T(\underline{x}) = t \right\}$

$\sim$ is an equivalence relation if

- $x \sim x$, $x \sim y$, $y \sim x$, $x \sim y$, $y \sim z$, $z \sim x$
- $T(\underline{x}) = T(\underline{x})$
- $T(\underline{x}) = T(\underline{y}) \iff T(\underline{y}) = T(\underline{x})$
- $T(\underline{x}) = T(\underline{y})$, $T(\underline{y}) = T(\underline{z}) \Rightarrow T(\underline{x}) = T(\underline{z})$

$x \sim y$ if $T(\underline{x}) = T(\underline{y})$

So, $\sim$ defined as $x \sim y$ if $T(\underline{x}) = T(\underline{y})$ is an equivalent relation.

$$A_t = \{ \underline{x} : T(\underline{x}) = t \}$$

$\underline{x} \in A_{T(\underline{x})} \Rightarrow$ every $\underline{x} \in A_t$ for some $t \in T(\mathcal{X})$

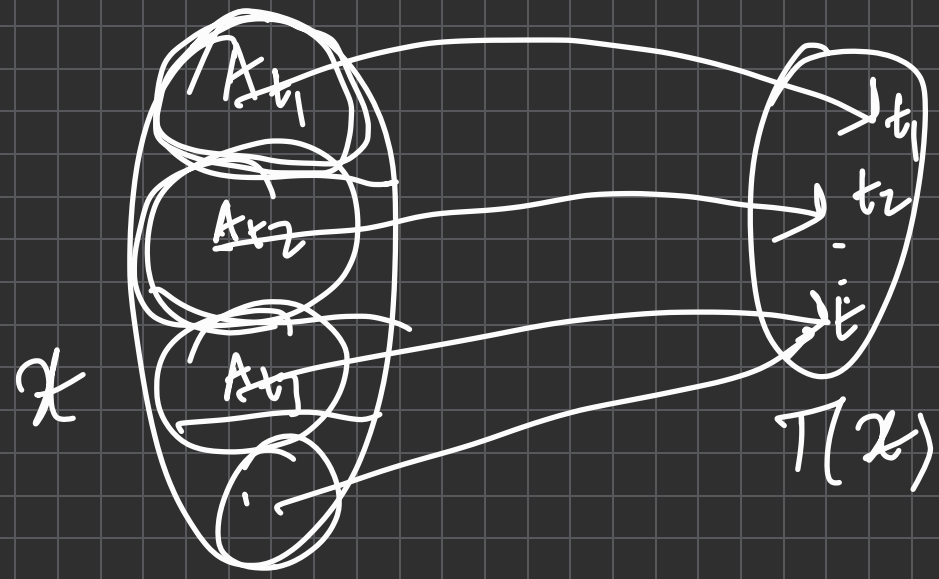
$\underline{x} \sim \underline{y}$ and $\underline{x} \in A_t$, then $\underline{y} \in A_t$

Every $\underline{x}$ can belong to only one A_t. for some $t \in T(\mathcal{X})$

$\underline{x} \in A_t, A_s \quad t \neq s.$

The space $\mathcal{X}$ is partitioned by the sets A_t, $t \in T(\mathcal{X})$.

These are known as equivalence classes.



$X_1, \dots, X_n \longrightarrow X_1 + \cdots + X_n$

A statistic $T(\underline{X})$ is sufficient if it contains all the information about θ.

A sufficient statistic is also called a sufficient partition.

- $\mathbb{P}_\theta\left(\underline{X}=\underline{x} \,\middle|\, T(\underline{X})=T(\underline{x})\right)$ is free of θ.

  this is known to us.

  we can generate observations from this distribution

Eg.

For the Bernoulli case

$$\mathbb{P}_\theta\left(\underline{X}=\underline{x} \mid T(\underline{X})=t\right) = \begin{cases} 0 & \text{if } \sum x_i \neq t \\ \dfrac{1}{\binom{n}{\sum x_i}} & \text{if } \sum x_i = t \end{cases}$$

$$\mathbb{P}_\theta\left(\underline{X}=\underline{x} \,\middle|\, T(\underline{X})=T(\underline{x})\right) = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}, \quad x_i \in \{0,1\}, \; i=1,\dots,n$$

Generate $\underset{\sim}{X}' \sim \mathbb{P}_\theta\left(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = T(\underset{\sim}{x})\right)$

$$\mathbb{P}_\theta\left(\underset{\sim}{X}' = \underset{\sim}{x}\right) = \sum_{T(\underset{\sim}{x})} \mathbb{P}_\theta\left(\underset{\sim}{X}' = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = T(\underset{\sim}{x})\right) \mathbb{P}_\theta\left(T(\underset{\sim}{X}) = T(\underset{\sim}{x})\right)$$

$$= \sum_{T(\underset{\sim}{x})} \mathbb{P}_\theta\left(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = T(\underset{\sim}{x})\right) \mathbb{P}_\theta\left(T(\underset{\sim}{X}) = T(\underset{\sim}{x})\right)$$

$$= \mathbb{P}_\theta\left(\underset{\sim}{X} = \underset{\sim}{x}\right)$$

In the Bernoulli example

$$\mathbb{P}_\theta\left(\underline{X}'=\underline{x}' \,\middle|\, T(\underline{X})=t\right) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} x_i' \neq t \\ \dfrac{1}{\binom{n}{\sum_{i=1}^{n} x_i'}}, & \sum_{i=1}^{n} x_i' = t \end{cases}$$

From $\underline{x}$, we set $T(\underline{x}) = t_0$

$$\frac{1}{\binom{n}{\sum x_i'}}, \text{ if } \sum x_i' = t_0$$

$\underset{\sim}{X}$ ← Person 1 has access to it
can compute

Person 2 only has access to $T(\underset{\sim}{X})$

Once the sample is observed. $\underset{\sim}{X} = \underset{\sim}{x} \longrightarrow T(\underset{\sim}{x})$

Person 1 has access to entire $\underset{\sim}{x}$

Person 2 only gets to know $T(\underset{\sim}{x})$

We have a statistical model $F_\theta(\cdot)$

$X_1, \dots, X_n \overset{iid}{\sim} F_\theta(\cdot)$

$\mathbb{P}_\theta(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t)$ is free of θ
← completely known

Once we know t, $\mathbb{P}_\theta\left(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t\right)$ is a distribution of $\underset{\sim}{X}$. and it is completely known.

So, we can generate a random sample $\underset{\sim}{Y}$ from this distribution.

That is, $\mathbb{P}_t\left(\underset{\sim}{Y} = \underset{\sim}{y}\right) = \mathbb{P}_\theta\left(\underset{\sim}{X} = \underset{\sim}{y} \mid T(\underset{\sim}{X}) = t\right)$

↑ conditional distribution depending on the value of t.

Unconditional distribution of $\underset{\sim}{Y}$.

$$\mathbb{P}\left(\underline{Y}=\underline{y}\right) = \sum_t \underbrace{\mathbb{P}_t\left(\underline{Y}=\underline{y}\right)}\ \mathbb{P}_\theta(t)$$

$$= \sum_t \mathbb{P}_\theta\left(\underline{X}=\underline{y} \mid T(\underline{X})=t\right)\ \mathbb{P}_\theta\left(T(\underline{X})=t\right)$$

$$= \mathbb{P}_\theta\left(\underline{X}=\underline{y}\right).$$

So, the unconditional dist. of $\underline{y}$ is same as the unconditional dist. of $\underline{X}$.

Eg. $X_1 \dots X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

We have seen $T(\underline{X}) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

$\uparrow$ # of 1's in the sample

$$\mathbb{P}_\theta\left(\underline{X} = \underline{y} \mid T(\underline{X}) = t\right) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} y_i \neq t \\ \dfrac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^{n} y_i = t \end{cases}$$

Once we observe $\underline{x} = (x_1 \dots x_n)$

we can compute $t = \sum_{i=1}^{n} x_i$.

We know $\mathbb{P}_\theta\left(\underline{X} = \underline{x} \mid T(\underline{X}) = t\right)$

Generate $\underset{\sim}{Y} = (Y_1 \dots Y_n)$ from this distribution.

$$\mathbb{P}_t(\underline{Y} = \underline{y}) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} y_i \neq t \\ \dfrac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^{n} y_i = t \end{cases}$$

Choose t places at random (out of the n places)

Put 1 in those t places.

Put 0 in the rest

This gives us $Y_1 \dots Y_n$

In a particular problem, there are multiple sufficient statistics.

Eg. $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

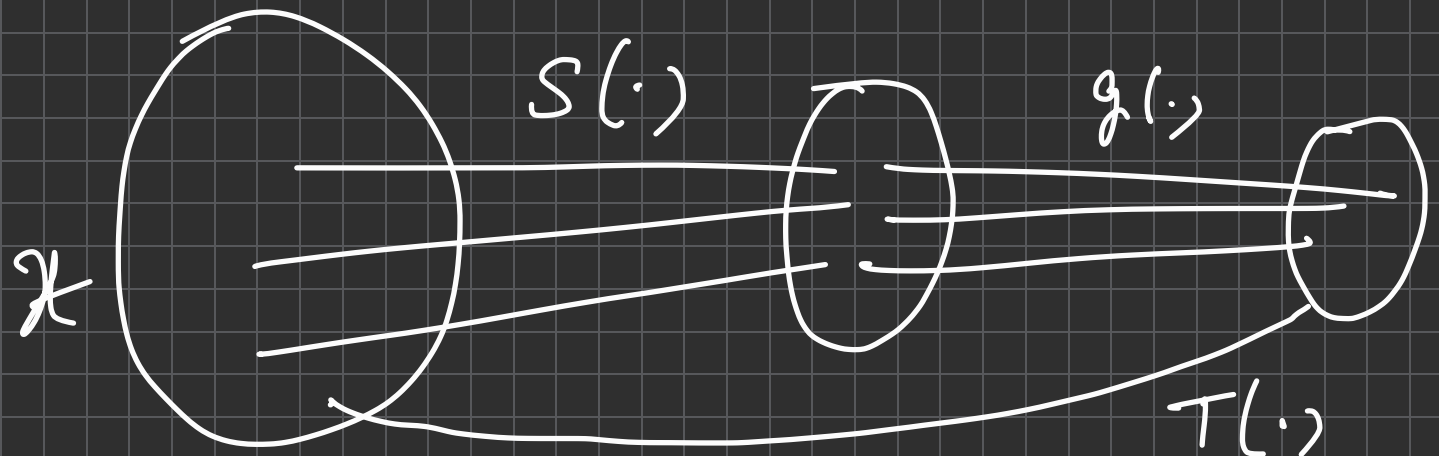
$(X_1, \ldots, X_n)$ is sufficient
$(X_{(1)}, \ldots, X_{(n)})$ is sufficient
$\sum_{i=1}^{n} X_i$ is sufficient.

} which one to choose?

We should go for the one which gives us the most possible reduction.

Def: A statistic $T(\underline{X})$ is said to be minimal sufficient for a parameter θ if it is sufficient and for any sufficient statistic $S(\underline{X})$ of θ, $T(\underline{X})$ can be written as a function of $S(\underline{X})$.

- $T(\underline{X})$ is sufficient.
- If $S(\underline{X})$ is sufficient, then there exists g such that $T(\underline{x}) = g(S(\underline{x}))$. $\forall\ \underline{x} \in \mathcal{X}$.



A minimal sufficient statistic is a possible reduction to every sufficient statistic.

Minimal sufficient statistics is not unique.

Results
Two statistics $T(\underline{X})$ and $S(\underline{X})$ are both minimal sufficient for θ if and only if there exist functions $g(\cdot)$ and $h(\cdot)$ such that

$T(\underline{x}) = g(S(\underline{x}))$ and $S(\underline{x}) = h(T(\underline{x}))$.

Result: Let $p_\theta(\underline{x})$ be the joint pdf/pmf of the sample. If $\dfrac{p_\theta(\underline{x})}{p_\theta(\underline{y})}$ is free of θ if and only if $T(\underline{x}) = T(\underline{y})$, then $T(\underline{X})$ is a minimal sufficient statistic for θ.

Eg. $X_1 \cdots X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

$$p_\theta(\underline{x}) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$$

$$\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}}{\theta^{\sum_{i=1}^n y_i} (1-\theta)^{n - \sum_{i=1}^n y_i}} = \theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i} (1-\theta)^{\sum_{i=1}^n y_i - \sum_{i=1}^n x_i}$$

is free of θ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

So, $T(\underline{X}) = \sum_{i=1}^{n} X_i$ is minimal sufficient for θ.

Eg. $X_1 \cdots X_n \overset{iid}{\sim} N(\mu, 1)$

$$p_\theta(\underline{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}, \quad x_i \in \mathbb{R}, \ i=1,\cdots n$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\left\{\sum_{i=1}^{n} x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n} x_i\right\}}$$

$$\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = e^{-\frac{1}{2}\left\{\sum_{i=1}^{n} x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i^2 - n\mu^2 + 2\mu\sum_{i=1}^{n} y_i\right\}}$$

$$= e^{-\frac{1}{2}\left\{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2 - 2\mu\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right)\right\}}$$

This is free of μ if and only if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

So, $\sum_{i=1}^{n} X_i$ is a minimal sufficient statistic for μ

Eg. $X_1, \dots, X_n \overset{iid}{\sim} U(0,\theta)$, $\theta > 0$

$$p_\theta(\underline{x}) = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 < x_i < \theta \ \forall i \\ 0 & \text{o.w.} \end{cases}$$

$$= \begin{cases} \frac{1}{\theta^n} & \text{if } x_{(1)} > 0 \text{ and } x_{(n)} < \theta \\ 0 & \text{o.w.} \end{cases}$$

$$= \frac{1}{\theta^n} \mathbb{1}\{x_{(1)} > 0\} \mathbb{1}\{x_{(n)} < \theta\}$$

$$p_\theta(\underline{y}) = \frac{1}{\theta^n} \mathbb{1}\{y_{(1)} > 0\} \mathbb{1}\{y_{(n)} < \theta\}$$

$\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})}$ is free of θ if and only if $\mathbb{1}\{x_{(n)} < \theta\} = \mathbb{1}\{y_{(n)} < \theta\}$

This is true iff $x_{(n)} = y_{(n)}$. So, $X_{(n)}$ is minimal suff. for θ.

Sufficiency: Gives us a reduction of the data without sacrificing any information on the relevant parameter.

Minimal sufficiency: Gives us the "best" possible sufficient reduction

<u>Result:</u> If $\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})}$ is free of θ if and only if $T(\underline{x}) = T(\underline{y})$, then $T(\underline{X})$ is a minimal sufficient statistic for θ.

Proof: We are given $\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})}$ is free of θ $\iff$ $T(\underline{x}) = T(\underline{y})$

We want to show: $T(\underline{X})$ is a minimal suff. statistic for θ.

- $T(\underline{X})$ is sufficient
- For any suff. statistic $S(\underline{X})$, we can find a fn. $g_s(\cdot)$ s.t. $T(\underline{x}) = g_s(S(\underline{x}))$ $\forall\, \underline{x} \in \mathcal{X}$.
- $T(\underline{X})$ is sufficient iff $p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x})$ $\forall \underline{x} \in \mathcal{X}$

Define $\mathcal{T} = \{ t : t = T(\underline{x}) \text{ for some } \underline{x} \in \mathcal{X} \}$

$\leftarrow$ image of $\mathcal{X}$ under $T(\cdot)$

$A_t = \{ \underline{x} \in \mathcal{X} : T(\underline{x}) = t \}$ $\leftarrow$ level set.

Now, for a fixed $t \in \mathcal{T}$, fix an element $x_t \in A_t$.

Take $\underline{x} \in \mathcal{X}$. If $\underbrace{T(\underline{x}) = t}$, then $x_{T(\underline{x})} = x_t \in A_t$

$\Rightarrow x \in A_t$.

$x_{T(\underline{x})}$ is the fixed element in A_t

x and $x_{T(\underline{x})}$ belong to the same set A_t.

$T(\underline{x}) = T(x_{T(\underline{x})})$.

By the condition of the theorem $\dfrac{p_\theta(\underline{x})}{p_\theta(\underline{x}_{T(\underline{x})})} = h(\underline{x})$

is free of θ.

$p_\theta(\underline{x}) = p_\theta(\underline{x}_{T(\underline{x})})\, h(\underline{x}) \quad \forall\, \underline{x}$

Define $g_\theta(t) = p_\theta(x_t)$, $t \in \mathcal{T}$.

$g_\theta(\cdot)$ is well-defined since x_t is a fixed element for every $t \in \mathcal{T}$

With this, we set $p_\theta(\underline{x}) = g_\theta(T(\underline{x}))\, h(\underline{x})$ $\forall \underline{x} \in \mathcal{X}$

$\Rightarrow$ $T(\underline{X})$ is sufficient for θ.

Next, we will show that $T(\underline{X})$ is minimal.

Let $S(\underline{X})$ be a sufficient statistic.

$\Rightarrow$ we can find $\tilde{g}$ and $\tilde{h}$ s.t.

$$p_\theta(\underline{x}) = \tilde{g}_\theta(S(\underline{x}))\, \tilde{h}(\underline{x}) \quad \forall \underline{x} \in \mathcal{X}.$$

$$\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = \frac{\tilde{g}_\theta(S(\underline{x}))}{\tilde{g}_\theta(S(\underline{y}))} \frac{\tilde{h}(\underline{x})}{\tilde{h}(\underline{y})}$$

If $S(\underline{x}) = S(\underline{y})$, then $\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = \frac{\tilde{h}(\underline{x})}{\tilde{h}(\underline{y})}$ ← free of θ

By the condition of the theorem, $T(\underline{x}) = T(\underline{y})$

So, $S(\underline{x}) = S(\underline{y}) \Rightarrow T(\underline{x}) = T(\underline{y})$.

Thus, there must exist a function $g(\cdot)$ s.t.

$T(\underline{x}) = g(S(\underline{x}))$ $\forall$ $\underline{x} \in \mathcal{X}$.

So, $T(\cdot)$ is minimal.

Result: Let f and g be two functions such that $f(x) = f(y) \Rightarrow g(x) = g(y)$. Then, there exists a function h such that $g(x) = h(f(x))\ \forall x$

Proof: Take $z = f(x)$.

Define $h(z) = g(x)$ | $h(f(x)) = g(x)$

Suppose $z = f(x)$ and $z = f(y) \Rightarrow f(x) = f(y)$

$\Rightarrow g(x) = g(y)$.

$\Rightarrow$ The value of the function h does not depend on the choice of x

So, h is a valid function.

And, by definition, $g(x) = h(f(x))$. (Proved).

One-parameter exponential family

A pmf / pdf $f_\theta(\cdot)$ belongs to the (one-parameter) exponential family if

$$f_\theta(x) = e^{a(\theta)\, T(x) - b(\theta)}\, c(x) \qquad \forall\, x \in \mathcal{Y},\ \theta \in \Theta$$

$\mathcal{Y}$ is the support of the distribution (pmf / pdf)

Θ is the parameter space.

The support $\mathcal{Y}$ must not involve the parameters θ.

Eg. Uniform $(0, \theta)$, $f_\theta(x) = \begin{cases} \frac{1}{\theta}, & \text{if } x \in (0, \theta) \\ 0 & \text{o.w.} \end{cases}$

$\mathcal{Y} = (0, \theta)$ ← depends on θ. NOT in exponential family.

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i)$$

$$= \prod_{i=1}^{n} \left\{ e^{a(\theta)\tau(x_i) - b(\theta)} c(x_i) \right\}$$

$$= e^{a(\theta)\sum_{i=1}^{n}\tau(x_i) - nb(\theta)} \prod_{i=1}^{n} c(x_i)$$

$$= g_\theta(T(\underline{x}))\ h(\underline{x})$$

$$g_\theta(t) = e^{a(\theta)t - nb(\theta)}, \quad h(\underline{x}) = \prod_{i=1}^{n} c(x_i)$$

$$T(\underline{x}) = \sum_{i=1}^{n} \tau(x_i)$$

$$\Rightarrow T(\underline{X}) = \sum_{i=1}^{n} \tau(X_i) \text{ is sufficient for } \theta.$$

In fact, $\dfrac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = e^{a(\theta)\left\{\sum_{i=1}^{n} \tau(x_i) - \sum_{i=1}^{n} \tau(y_i)\right\}} \dfrac{\prod_{i=1}^{n} c(x_i)}{\prod_{i=1}^{n} c(y_i)}$

This is free of θ iff $\sum_{i=1}^{n} \tau(x_i) = \sum_{i=1}^{n} \tau(y_i)$

$\Updownarrow$

$T(\underline{x}) = T(\underline{y})$

So, $T(\underline{X}) = \sum_{i=1}^{n} \tau(X_i)$ is a minimal sufficient statistic for θ.

Multiparameter exponential family.

$\underline{\theta} = (\theta_1, \cdots, \theta_r) \leftarrow r$ parameters.

A pdf/pmf belongs to the multiparameter exponential family if

$$f_\theta(x) = e^{\sum_{j=1}^{k} a_j(\theta)\, \tau_j(x) - b(\theta)}\, c(x), \quad x \in \mathcal{X}, \quad \underline{\theta} \in \Theta.$$

Eg. $N(\mu, \sigma^2)$. $\underline{\theta} = (\mu, \sigma^2)$, $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R} \times (0, \infty)$

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R}$$

$$= \frac{1}{\sqrt{2\pi}}\, \frac{1}{\sigma}\, e^{-\frac{1}{2\sigma^2}(x^2 + \mu^2 - 2\mu x)}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}}$$

$$-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \log\sigma$$

$$= e^{a_1(\underline{\theta})\tau_1(x) + a_2(\theta)\tau_2(x) - b(\theta)} \cdot \frac{1}{\sqrt{2\pi}} \cdot c(x)$$

$$= e^{\quad}$$

$$a_1(\underline{\theta}) = -\frac{1}{2\sigma^2}, \quad a_2(\theta) = \frac{\mu}{\sigma^2}$$

$$\tau_1(x) = x^2, \quad \tau_2(x) = x$$

$$b(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$$

$$c(x) = \frac{1}{\sqrt{2\pi}} \quad \forall\ x \in \mathbb{R}$$

f_{μ,σ^2} belongs to the multiparameter exponential family

$$f_\theta(\underline{x}) = \prod_{i=1}^{n} \left\{ e^{\sum_{j=1}^{k} a_j(\theta)\, \tau_j(x_i) - b(\theta)}\, c(x_i) \right\}$$

$$= e^{\sum_{i=1}^{n} \left\{ \sum_{j=1}^{k} a_j(\theta)\, \tau_j(x_i) - b(\theta) \right\}} \prod_{i=1}^{n} c(x_i)$$

$$= e^{\sum_{j=1}^{k} \left\{ a_j(\theta) \sum_{i=1}^{n} \tau_j(x_i) \right\} - n b(\theta)} \prod_{i=1}^{n} c(x_i)$$

$$\Rightarrow T(\underline{X}) = \left(\sum_{i=1}^{n} \tau_1(x_i), \ldots, \sum_{i=1}^{n} \tau_k(x_i) \right) \text{ is sufficient for } \underline{\theta}$$

$$\frac{p_\theta(x_{\underline{\cdot}})}{p_\theta(y_{\underline{\cdot}})} = e^{\sum_{j=1}^{K}\{a_j(\theta)\sum_{i=1}^{n}\tau_j(x_i)\}-\sum_{j=1}^{K}\{a_j(\theta)\sum_{i=1}^{n}\tau_j(y_i)\}} \left\{\frac{\prod_{i=1}^{n}c(x_i)}{\prod_{i=1}^{n}c(y_i)}\right\}$$

$$= e^{\sum_{j=1}^{K} a_j(\theta)\left\{\sum_{i=1}^{n}\tau_j(x_i)-\sum_{i=1}^{n}\tau_j(y_i)\right\}} \prod_{i=1}^{n}\frac{c(x_i)}{c(y_i)}$$

This ratio is free of θ if and only if

$$\sum_{i=1}^{n}\tau_j(x_i) = \sum_{i=1}^{n}\tau_j(y_i) \quad \forall\, j=1,\dots,K$$

$$\Rightarrow T(\underline{X}) = \left(\sum_{i=1}^{n}\tau_1(x_i),\dots,\sum_{i=1}^{n}\tau_K(x_i)\right) \text{ is a minimal sufficient statistic for } \underline{\theta}.$$

Eg. $N(\mu, \sigma^2)$

$$\tau_1(x) = x^2, \quad \tau_2(x) = x$$

$\left(\sum_{i=1}^n \tau_1(x_i), \sum_{i=1}^n \tau_2(x_i) \right)$ is minimal suff. for (μ, σ^2)

$$= \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i \right)$$

In many situations, the order statistics is the best possible reduction that we can get.

Order statistics

For a random sample $X_1, \dots, X_n$, the order statistics are $X_{(1)} \le \dots \le X_{(n)}$

$X_1 \dots X_n \overset{iid}{\sim} f_\theta$.

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} f_\theta(x_{(i)})$$

So, $(X_{(1)}, \dots, X_{(n)})$ is sufficient for θ.

Order statistics are sufficient even under the nonparametric regime.

Suppose that $X_1, \dots, X_n \overset{iid}{\sim} f$, where f is a pdf.

So, X_i's are continuous random variables.

Therefore, $\mathbb{P}(X_i = X_j) = 0$

So, the order statistics are distinct (with prob. 1)

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

$$\mathbb{P}\left(\underline{X} = \underline{x} \mid X_{(1)} = y_1, \dots, X_{(n)} = y_n\right) = \begin{cases} 0 & \text{if the ordered values of } \underline{x} \text{ don't match } y_1, \dots, y_n \\ \frac{1}{n!} & \text{otherwise.} \end{cases}$$

$\Rightarrow (X_{(1)}, \dots, X_{(n)})$ is sufficient for f

Order statistics

Data: $X_1 \dots X_n$

Ordered values: $X_{(1)} < X_{(2)} < \dots < X_{(n)}$

If X_i's are continuous, then the probability of two random variables taking the same value is 0

In that case $\mathbb{P}(X_i\text{'s are all distinct}) = 1$.

Order statistics are sufficient if the underlying distribution is continuous.

The joint distribution / pdf of the order statistics

$$f_{X_{(1)}\cdots X_{(n)}}(y_1,\cdots y_n) = \begin{cases} n!\, f(y_1)\cdots f(y_n), \\ \qquad\qquad \text{if } y_1<y_2<\cdots<y_n \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

Marginal distribution of the order statistics:

Pdf of $X_{(j)}$.

$U(0,\theta)$ — $X_{(n)}$ is minimal suff. for θ.

$U(\theta,1)$ — $X_{(1)}$ is minimal suff. for θ

If n is odd, then $X_{(\frac{n+1}{2})}$ is the median

$X_{(\frac{3n}{4})} - X_{(\frac{n}{4})}$ = inter-quartile range.

Pdf of $X_{(1)}$:

The cdf of $X_{(1)}$

$$F_{(1)}(y) = \mathbb{P}(X_{(1)} \le y) = 1 - \mathbb{P}(X_{(1)} > y)$$

$$\{X_{(1)} > y\} \iff \{X_1 > y, X_2 > y, \dots, X_n > y\}$$

$$\therefore \mathbb{P}(X_{(1)} > y) = \mathbb{P}(X_1 > y, X_2 > y, \dots, X_n > y)$$

($X_1 \dots X_n$ are indep) $= \mathbb{P}(X_1 > y)\,\mathbb{P}(X_2 > y) \cdots \mathbb{P}(X_n > y)$

($X_1 \dots X_n$ are identically distributed) $= \{\mathbb{P}(X_1 > y)\}^n$

$= \{1 - F_X(y)\}^n$ $\left(F_X(y) = \mathbb{P}(X_1 \le y)\right)$

$$F_{(1)}(y) = \mathbb{P}(X_{(1)} \le y) = 1 - \mathbb{P}(X_{(1)} > y)$$

$$= 1 - \{1 - \mathbb{P}(X_1 \le y)\}^n$$

$$= 1 - \{1 - F_X(y)\}^n$$

If $F_X(\cdot)$ is differentiable, then $F_{(1)}(\cdot)$ is differentiable.

In that case, the pdf of $X_{(1)}$ is

$$f_{(1)}(y) = -n\{1 - F_X(y)\}^{n-1}(-f_X(y))$$

$$= n\{1 - F_X(y)\}^{n-1} f_X(y).$$

$X_{(1)}$ has the same range as X_1.

The cdf of $X_{(n)}$

$$F_{(n)}(y) = \mathbb{P}(X_{(n)} \le y)$$

$$\{X_{(n)} \le y\} \Leftrightarrow \{X_1 \le y, \cdots, X_n \le y\}$$

$$\therefore \mathbb{P}(X_{(n)} \le y) = \mathbb{P}(X_1 \le y, \cdots, X_n \le y)$$

$$= \mathbb{P}(X_1 \le y) \cdots \mathbb{P}(X_n \le y) \quad (\text{indep}).$$

$$= \{\mathbb{P}(X_1 \le y)\}^n$$

$$= \{F_X(y)\}^n$$

$$\therefore F_{(n)}(y) = \{F_X(y)\}^n$$

The pdf of $X_{(n)}$

$$f_{(n)}(y) = n\left\{F_X(y)\right\}^{n-1} f_X(y)$$

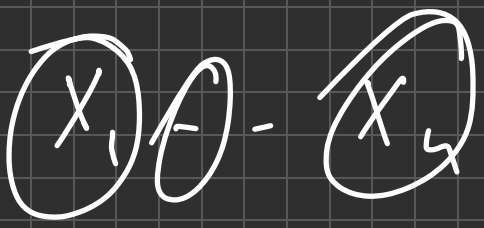
The pdfs of $X_{(1)}$ & $X_{(n)}$ are

$$f_{(1)}(y) = n\left\{1 - F_X(y)\right\}^{n-1} f_X(y)$$

$$f_{(n)}(y) = n\left\{F_X(y)\right\}^{n-1} f_X(y)$$

Both $X_{(1)}$ and $X_{(n)}$ have the same range as X_i (marginally).

$X_{(j)}$ For general j :



The cdf of $X_{(j)}$

$$F_{(j)}(y) = \mathbb{P}\left(X_{(j)} \le y\right).$$

Define $W_i = \begin{cases} 1 & \text{if } X_i \le y \\ 0 & \text{if } X_i > y \end{cases} = \mathbb{1}\{X_i \le y\}$

$\{X_{(j)} \le y\} \iff$ at least j many of the W_i's are 1.

$$\sum_{i=1}^{n} W_i \ge j$$

$$W_i = \mathbb{1}\{X_i \le y\}$$

$$\{X_{(j)} \le y\} \iff \sum_{i=1}^{n} W_i \ge j$$

W_i's are Bernoulli

W_i's are independent.

$$\mathbb{P}(W_i = 1) = \mathbb{P}(X_i \le y) = F_X(y) \leftarrow \text{free of } i$$

($\because$ X_i's are identical)

$$\therefore W_1, \dots, W_n \overset{iid}{\sim} \text{Bernoulli}(F_X(y))$$

$$\Rightarrow \sum_{i=1}^{n} W_i \sim \text{Binomial}(n, F_X(y))$$

$$\{X_{(j)} \le y\} \iff \sum_{i=1}^{n} W_i \ge j$$

$$\sum_{i=1}^{n} W_i \sim \text{Binomial}(n, F_X(y))$$

$$\therefore \mathbb{P}\left(X_{(j)} \le y\right) = \mathbb{P}\left(\text{Bin}(n, F_X(y)) \ge j\right)$$

$$= \sum_{k=j}^{n} \mathbb{P}\left(\text{Bin}(n, F_X(y)) = k\right)$$

$$= \sum_{k=j}^{n} \binom{n}{k} \{F_X(y)\}^k \{1 - F_X(y)\}^{n-k}$$

$$=: F_{(j)}(y).$$

If $F_X(\cdot)$ is differentiable, then the pdf of $X_{(j)}$ is

$$\sum_{k=j}^{n} \binom{n}{k} \Big[\{F_X(y)\}^k (n-k) \{1-F_X(y)\}^{n-k-1} \{-f_X(y)\} + k \{F_X(y)\}^{k-1} f_X(y) \{1-F_X(y)\}^{n-k} \Big]$$

$$= \sum_{k=j}^{n} \binom{n}{k} \{F_X(y)\}^{k-1} \{1-F_X(y)\}^{n-k-1} f_X(y) \Big[-(n-k) F_X(y) + k \{1-F_X(y)\} \Big]$$

$$= \sum_{k=j}^{n} \{F_X(y)\}^{k-1} \{1-F_X(y)\}^{n-k-1} f_X(y) \Big[\binom{n}{k} k \{1-F_X(y) - \binom{n}{k} (n-k) F_X(y) \Big]$$

$$\sum_{k=j}^{n} \{F_X(y)\}^{k-1} \{1-F_X(y)\}^{n-k-1} f_X(y)$$

$$\left[\binom{n}{k} k \{1-F_X(y) - \binom{n}{k} (n-k) F_X(y)\right]$$

$$\binom{n}{k} k = \frac{n!}{k!(n-k)!} k = \frac{n!}{(k-1)!(n-k)!} = \frac{n!\,(n-k+1)}{(k-1)!(n-k+1)!}$$

$$= (n-k+1)\binom{n}{k-1}$$

$$= f_X(y) \sum_{k=j}^{n} \left\{ F_X(y)^{k-1} \{1-F_X(y)\}^{n-k} \binom{n}{k-1} (n-k+1) \right.$$

$$\left. - F_X(y)^k \{1-F_X(y)\}^{n-k-1} \binom{n}{k} (n-k) \right\}$$

$$a_k = F_X(y)^{k-1} \{1-F_X(y)\}^{n-k} \binom{n}{k-1} (n-k+1)$$

$$= f_X(y) \sum_{k=j}^{n} (a_k - a_{k+1})$$

$$\sum_{k=j}^{n} (a_k - a_{k+1}) = a_j - \cancel{a_{j+1}} + \cancel{a_{j+1}} - \cancel{a_{j+2}} + \cdots + \cancel{a_n} - a_{n+1}$$

$$= a_j - a_{n+1}$$

$$a_j - a_{n+1} = \{F_X(y)\}^{j-1} \{1 - F_X(y)\}^{n-j} \binom{n}{j-1} (n-j+1)$$

$$- \{F_X(y)\}^{n} \{1 - F_X(y)\}^{-1} \binom{n}{n} (\underbrace{n - n - 1 + 1}_{0})$$

$$= \{F_X(y)\}^{j-1} \{1 - F_X(y)\}^{n-j} \frac{n!}{(j-1)!\,(n-j)!}$$

$\therefore$ The pdf of $X_{(j)}$ is

$$f_{(j)}(y) = \frac{n!}{(j-1)!\,(n-j)!} \{F_X(y)\}^{j-1} \{1 - F_X(y)\}^{n-j} f_X(y)$$

So, The pdf of $X_{(j)}$ is

$$f_{(j)}(y) = \frac{n!}{(j-1)!\,(n-j)!} \{F_X(y)\}^{j-1} \{1-F_X(y)\}^{n-j} f_X(y).$$

Eg. Let $X_1 \cdots X_n \overset{i.i.d}{\sim} U(0,1)$

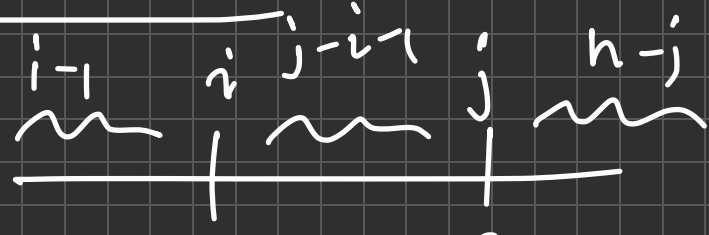
Then, $f_X(y) = \begin{cases} 1 & \text{if } 0<y<1 \\ 0 & \text{o.w.} \end{cases}$

$$F_X(y) = \begin{cases} 0 & \text{if } y<0 \\ y & \text{if } 0 \le y < 1 \\ 1 & \text{if } y \ge 1 \end{cases}$$

$$\therefore f_{(j)}(y) = \frac{n!}{(j-1)!\,(n-j)!} y^{j-1}(1-y)^{n-j}, \quad 0<y<1$$

Joint distribution of two order statistics

The jt. pdf of $X_{(i)}$ and $X_{(j)}$ where $i<j$



$$f_{(i,j)}(y,z) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \{F_X(y)\}^{i-1} \{F_X(z)-F_X(y)\}^{j-i-1} \{1-F_X(z)\}^{n-j}, \quad y<z$$

$$f_X(y)\, f_X(z)$$

$$f_{(i,j)}(y,z) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} f_X(y)\, f_X(z) \{F_X(y)\}^{i-1} \{F_X(z)-F_X(y)\}^{j-i-1} \{1-F_X(z)\}^{n-j}, \quad y<z$$

Ancillary statistic

A statistic $S = S(\underline{X})$ is said to be ancillary if the distribution of S is free of the unknown parameter θ.

Eg. $X_1, X_2 \overset{iid}{\sim} \text{Normal}(\theta, 1)$, $\theta \in \mathbb{R}$.

— $X_1 - X_2 \sim \text{Normal}(0, 2)$ ← free of θ.

$X_1 - X_2$ is ancillary.

Eg. $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$

— $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$

$\bar{X} \sim N(\theta, \frac{1}{n})$, $(n-1)S^2 \sim \chi^2_{n-1}$, $\bar{X}, S^2$ indep.

↑ free of θ

S^2 is ancillary.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$, $\theta > 0$.

$X_{(n)}$ is minimal sufficient.

$\frac{X_{(1)}}{X_{(n)}}$ has distribution free of θ.

$\frac{X_{(1)}}{X_{(n)}}$ is ancillary.

$\frac{X_{(i)}}{X_{(j)}}$, $1 \le i < j \le n$ is ancillary.

Eg Location family of distributions

$X_1, \dots, X_n \overset{iid}{\sim} f_\theta$, $F_\theta(x) = F(x - \theta)$ $\forall x$ $\forall \theta$

where F is a known pdf (does not involve any unknown parameter)

In this case, the cdf satisfies

$$F_\theta(x) = \int_{-\infty}^{x} f_\theta(t)\,dt = \int_{-\infty}^{x} f(t-\theta)\,dt$$

$$= \int_{-\infty}^{x-\theta} f(z)\,dz = F(x-\theta)$$

↑ cdf corr. to f.

$F_\theta(x) = F(x-\theta)$, F is completely known

Eg. Normal $(\theta, 1)$

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}, \quad x \in \mathbb{R}$$

$$= f(x-\theta), \text{ where } f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}$$

Eg. Cauchy $(\theta, 1)$

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}, \quad x \in \mathbb{R}$$

$$= f(x-\theta), \qquad f(z) = \frac{1}{\pi} \frac{1}{1+z^2}, \quad z \in \mathbb{R}$$

Eg. Laplace $(\theta, 1)$ is a member of the location family.

If X has pdf $f_\theta(x) = f(x-\theta)$

then $X-\theta$ has pdf $f(x)$

$\Rightarrow X-\theta = Y$ has a known distribution

$X = \theta + Y$

For location family:

$$X_1, \cdots X_n \overset{i.i.d}{\sim} f_\theta(x) = f(x-\theta)$$

$$\underbrace{X_1-\theta, \cdots, X_n-\theta}_{\text{random variables, but not observable.}} \overset{iid}{\sim} f(x) \leftarrow \text{completely known.}$$

$$X_{(n)} - X_{(1)} = \max_i X_i - \min_i X_i$$

$$= \max_i (X_i-\theta) + \theta - \left\{\min_i (X_i-\theta) + \theta\right\}$$

$$= \max_i (X_i-\theta) + \cancel{\theta} - \min_i (X_i-\theta) - \cancel{\theta}$$

$$= \max_i Y_i - \min_i Y_i, \quad Y_i = X_i-\theta$$

$$= Y_{(n)} - Y_{(1)}$$

$Y_1, \cdots Y_n \overset{i.i.d}{\sim} f(\cdot)$ ← completely known $\Rightarrow Y_{(n)} - Y_{(1)}$ has dist. free of θ.

$\therefore$ $X_{(n)} - X_{(1)} = Y_{(n)} - Y_{(1)}$ is ancillary.

The range $R = X_{(n)} - X_{(1)}$ for a location family $f(x-\theta)$ is ancillary.

Scale family $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$, $\theta > 0$.

f_θ is said to belong to the scale family if $f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$, $\forall x \; \forall \theta$, where $f(\cdot)$ is completely known

Ancillary statistic A statistic $S = S(\underline{X})$ is ancillary if its distribution is free of the unknown parameter θ.

Sufficient statistics The part of the sample that contains all the information about the unknown parameter θ.

Minimal sufficient statistics Give us the best possible reduction.

Maybe ancillary statistics & minimal suff. stat. are unrelated $\Leftarrow$ NOT True.

Eg. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(\theta, \theta+1)$, $\theta \in \mathbb{R}$.

A minimal suff. statistic for θ is $(X_{(1)}, X_{(n)})$.

$\Rightarrow (X_{(n)} - X_{(1)},\ X_{(1)} + X_{(n)})$ is also minimal sufficient.

$\text{Uniform}(\theta, \theta+1) = \text{Uniform}(0,1) + \theta$.

$$f_\theta(y) = \begin{cases} 1 & \text{if } \theta < y < \theta+1 \\ 0 & \text{o.w.} \end{cases} \qquad f(y) = \begin{cases} 1 & \text{if } 0<y<1 \\ 0 & \text{o.w.} \end{cases}$$

$$= \begin{cases} 1 & \text{if } 0 < y-\theta < 1 \\ 0 & \text{o.w.} \end{cases}$$

$= f(y-\theta)$ ← a distribution from the location family.

$\therefore X_{(n)} - X_{(1)}$ is ancillary.

The distribution of $X_{(n)} - X_{(1)}$ can be computed directly.

$$f_R(r) = n(n-1)\, r^{n-2}, \quad 0 < r < 1$$ } free of θ.

$R = X_{(n)} - X_{(1)}$.

$X_{(n)} - X_{(1)}$ is ancillary.

$(X_{(n)} - X_{(1)},\ X_{(1)} + X_{(n)})$ is minimal sufficient.

They are not independent.

When is it true that a minimal sufficient statistic is unrelated to an ancillary statistic?

Complete statistic

Def 4: A statistic $T = T(\underline{X})$ is said to be complete if

$$\mathbb{E}_\theta \{ g(T) \} = 0 \quad \forall \theta \implies \mathbb{P}_\theta \big(g(T) = 0 \big) = 1 \quad \forall \theta$$

$g(T) = 0$ with pr. 1

$g(\cdot)$ is a real-valued function

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $0 < \theta < 1$

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

Take any function $g(\cdot)$

$$\mathbb{E}_\theta\{g(T)\} = \sum_{t=0}^{n} g(t)\, \mathbb{P}_\theta(T = t)$$

$$= \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t}$$

$\mathbb{E}_\theta\{g(T)\} = 0 \quad \forall \theta \in (0,1)$

$$\Leftrightarrow \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = 0 \quad \forall \theta \in (0,1)$$

$$\Leftrightarrow (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0 \quad \forall \ \theta \in (0,1)$$

$$\Leftrightarrow \sum_{t=0}^{n} g(t) \binom{n}{t} \eta^t = 0 \quad \forall \ \eta \in (0,\infty) \quad \left[\begin{array}{l} \eta = \frac{\theta}{1-\theta} \\ \theta \in (0,1) \Leftrightarrow \eta \in (0,\infty) \end{array}\right.$$

↑ a polynomial in η of degree n

↖ does not involve η.

$$\Leftrightarrow g(t) \binom{n}{t} = 0 \quad \forall \ t = 0, 1, \cdots, n$$

$$\Leftrightarrow g(t) = 0 \quad \forall \ t = 0, 1, \cdots, n \quad \left[\because \binom{n}{t} \neq 0\right]$$

[since the LHS is a polynomial of degree n which is identically 0]

$\therefore \ E_\theta\{g(T)\} = 0 \Rightarrow g(T) = 0$ wp 1.

$\therefore \ T = \sum_{i=1}^{n} X_i$ is complete.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$, $\theta > 0$ $[\theta \in (0, \infty)]$

$T = X_{(n)}$ is minimal sufficient for θ.

$$f_{(n)}(y) = n \left(\frac{y}{\theta}\right)^{n-1} \frac{1}{\theta}, \quad 0 < y < \theta$$

$$= n \frac{y^{n-1}}{\theta^n}, \quad 0 < y < \theta$$

Take a function g.

$$\mathbb{E}_\theta\{g(T)\} = \int_0^\theta g(t) f_T(t)\, dt$$

$$= \int_0^\theta g(t)\, n \frac{t^{n-1}}{\theta^n}\, dt$$

$$= \frac{n}{\theta^n} \int_0^\theta g(t)\, t^{n-1}\, dt$$

$$\mathbb{E}_\theta[g(T)] = 0 \quad \forall \theta \in (0,\infty)$$

$$\Leftrightarrow \int_0^\theta g(t)\, t^{n-1}\, dt = 0 \quad \forall \theta \in (0,\infty)$$

The LHS as a function of θ is differentiable and its derivative is

$$g(\theta)\,\theta^{n-1}$$

The derivative of the RHS w.r.t. θ is 0

$$\therefore \quad g(\theta)\,\theta^{n-1} = 0 \quad \forall \theta \in (0,\infty) \quad \left[\text{since the two derivatives should match}\right]$$

$$\Rightarrow g(\theta) = 0 \quad \forall \theta \in (0,\infty)$$

$$\Rightarrow \mathbb{P}_\theta[g(T) = 0] = 1 \quad \therefore T = X_{(n)} \text{ is complete.}$$

Def: A complete statistic that is sufficient is called a complete sufficient statistic.

Result: A complete sufficient statistic is minimal sufficient.

Pf: Let $T = T(\underline{X})$ be complete & sufficient.

Let $S = S(\underline{X})$ be a sufficient statistic

We will show that $T = g(S)$ for some $g(\cdot)$.

Define $\psi(T) = T - \underbrace{\mathbb{E}_\theta(T|S)}_{\text{does not depend on } \theta \ (\because S \text{ is sufficient})}$

S is sufficient. $\underline{X}|S$ is free of θ. $T|S$ is free of θ.

$$\mathbb{E}_\theta[\psi(T)] = \mathbb{E}_\theta\left[T - \mathbb{E}_\theta(T|S)\right]$$

$$= \mathbb{E}_\theta(T) - \mathbb{E}_\theta\,\mathbb{E}_\theta(T|S)$$

$$= \mathbb{E}_\theta(T) - \mathbb{E}_\theta(T) = 0 \quad \forall\,\theta$$

T is complete $\Rightarrow \psi(T) = 0$ wp 1.

$\Rightarrow T - \mathbb{E}_\theta(T|S) = 0$ wp 1

$\Rightarrow T = \mathbb{E}_\theta(T|S)$ wp 1

$$\mathbb{P}_\theta\left[T = \underbrace{\mathbb{E}_\theta(T|S)}\right] = 1$$

a function of S and does not involve θ.

$\Rightarrow \mathbb{P}_\theta[T = g(S)] = 1$

$\therefore$ T can be written as a function of S.

$\Rightarrow$ T is a minimal sufficient statistic.

If T is complete and S is sufficient, then we can always write T as a function of S. But, it is not always true that T is sufficient. However, if T is sufficient, then it must be minimal sufficient.

Result: A complete sufficient statistic is independent of any ancillary statistic. (Basu's Theorem)

Pf: Let T be complete sufficient.

S be ancillary

We will show:

$$\mathbb{P}_\theta[S \in A \mid T] = \mathbb{P}_\theta(S \in A) \quad \text{for every } A.$$

$\mathbb{P}_\theta(S \in A)$ ← is free of θ

$\mathbb{P}_\theta(S \in A \mid T)$ ← free of θ (since T is suff.)

Define $p_A = \mathbb{P}_\theta(S \in A)$, $q_A(T) = \mathbb{P}_\theta(S \in A \mid T)$

$$\mathbb{P}(A) = \int_A f(x)\,dx = \int_{-\infty}^{\infty} \underbrace{\mathbb{1}\{x \in A\}}\, f(x)\,dx$$

$$= \mathbb{E}\{\mathbb{1}(X \in A)\}$$

$$= \mathbb{E}\{\mathbb{1}_A\}$$

$$\mathbb{P}(A) = \underset{X}{\mathbb{E}}(\mathbb{1}_A)$$

$$\mathbb{P}(A|B) = \mathbb{E}(\mathbb{1}_A | B)$$

$$\mathbb{E}\,\mathbb{P}(A|B) = \mathbb{E}\,\mathbb{E}(\mathbb{1}_A | B) = \mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$$

$$\therefore \quad p_A = \mathbb{P}_\theta(S \in A) = \mathbb{E}_\theta\, \mathbb{P}_\theta(S \in A \mid T)$$

$$= \mathbb{E}_\theta\left[q_A(T)\right] \qquad \forall \theta$$

$$\therefore \quad \mathbb{E}_\theta\left\{ q_A(T) - p_A \right\} = 0 \qquad \forall \theta$$

T is complete $\Rightarrow$ $\mathbb{P}_\theta\left\{ q_A(T) = p_A \right\} = 1 \quad \forall \theta$

$\therefore$ w.p. 1

$$\mathbb{P}_\theta(S \in A \mid T) = \mathbb{P}_\theta(S \in A) \qquad \forall A$$

The conditional dist of S given T equals the unconditional dist of S

$\Rightarrow$ T & S are independent.

Completeness: T is said to be complete if

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

Result: If T is complete and sufficient, then T is minimal sufficient.

Result: (Basu's Theorem) If T is complete sufficient and S is ancillary, then T and S are independent.

Eg. Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is minimal sufficient for θ.

$\bar{X}$ is also complete

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

$$S^2 = \frac{1}{2n(n-1)} \sum_{1 \le i \ne j \le n} (x_i - x_j)^2$$

$$\frac{1}{2n(n-1)} \sum_{1 \le i \ne j \le n} \underbrace{(x_i - \bar{x} - x_j + \bar{x})^2}$$

$$(x_i - \bar{x})^2 + (x_j - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x})$$

$$\sum_{i,j=1}^{n} (x_i - \bar{x})^2 = n \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\sum_{i,j=1}^{n} (x_j - \bar{x})^2 = n \sum_{j=1}^{n} (x_j - \bar{x})^2 \quad \text{equal.}$$

$$\sum_{i,j=1}^{n} (x_i - \bar{x})(x_j - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) \sum_{j=1}^{n} (x_j - \bar{x}) = 0$$

$$\frac{1}{2n(n-1)} \sum_{1 \le i \ne j \le n} (X_i - X_j)^2 = \frac{1}{2n(n-1)} \left\{ 2n \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}$$

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

$X_1, \cdots, X_n \overset{iid}{\sim} N(\theta, 1)$

$\underbrace{X_1 - \theta}_{Y_1}, \cdots, \underbrace{X_n - \theta}_{Y_n} \overset{iid}{\sim} N(0,1)$

$Y_i = X_i - \theta$

$X_i = Y_i + \theta$

$$X_i - X_j = Y_i + \theta - (Y_j + \theta) = Y_i - Y_j$$

$$S^2 = \frac{1}{2n(n-1)} \sum_{i \ne j} (X_i - X_j)^2 = \frac{1}{2n(n-1)} \sum_{i \ne j} (Y_i - Y_j)^2$$

$\therefore$ S^2 is ancillary

$\bar{X}$ is complete sufficient.

S^2 is ancillary.

So, by Basu's theorem $\bar{X}$ and S^2 are independent.

Eg. Let $X_1 \cdots X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, $\theta > 0$

$X_{(n)}$ is complete sufficient.

$\dfrac{X_{(1)}}{X_{(n)}}$ is ancillary

$\therefore$ Basu's Theorem tells us that $\dfrac{X_{(1)}}{X_{(n)}}$ & $X_{(n)}$ are independent

Let $X_1, \dots X_n \overset{iid}{\sim} N(\theta, 1)$, $T = \bar{X} \sim N(\theta, \frac{1}{n})$

$$\mathbb{E}_\theta\{g(T)\} = \frac{1}{\sqrt{2\pi\frac{1}{n}}} \int_{-\infty}^{\infty} g(t)\, e^{-\frac{1}{2\frac{1}{n}}(t-\theta)^2} dt$$

$$= \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} g(t)\, e^{-\frac{n(t-\theta)^2}{2}} dt = 0 \quad \forall_{\theta \in \mathbb{R}}$$

Integral transforms: A mapping between function classes, defined in terms of integrals.

Let f be a continuous function.

Define

$$\psi(t) = \int_0^\infty e^{-tx} f(x)\, dx, \quad t \in \mathbb{R}$$

← one-sided Laplace transform.

$$\psi(t) = \int_{-\infty}^\infty e^{-tx} f(x)\, dx, \quad t \in \mathbb{R}$$

← two-sided Laplace transform

$$\psi(t) = \int_0^\infty x^{t-1} f(x)\, dx, \quad t > 0$$

← Mellin's transform

$$\psi(t) = \int_0^\infty \frac{1}{t+x} f(x)\, dx, \quad t \geq 0$$

← Stieltjes transform.

$$\psi(t) = \int_{-\infty}^\infty e^{itx} f(x)\, dx, \quad t \in \mathbb{R}$$

← Fourier transform

All these transforms have the property that

if $\psi_f(t) = \psi_g(t) \quad \forall t$ then $f(x) = g(x) \ \forall x$

$f \longmapsto \psi_f$, $g \longmapsto \psi_g$

$\psi_f \equiv \psi_g \Rightarrow f \equiv g$

In particular, if $\psi_f \equiv 0$ $(\psi_f(t) = 0 \ \forall t)$

then $f \equiv 0$ $(f(x) = 0 \ \forall x)$

$$\mathbb{E}_\theta\{g(T)\} = \int_{-\infty}^{\infty} g(t) f(t)\, dt \longleftarrow c\, \psi_g(\theta)$$

$\psi_g(\theta) = 0 \ \forall \theta \Rightarrow g(t) = 0 \ \forall t.$

Back to the normal example:

$$\mathbb{E}_\theta\{g(T)\} = \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} g(t)\, e^{-\frac{n(t-\theta)^2}{2}}\, dt$$

$$= \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} g(t)\, e^{-\frac{nt^2}{2}} e^{-\frac{n\theta^2}{2}} e^{nt\theta}\, dt$$

$$= \sqrt{\frac{n}{2\pi}}\, e^{-\frac{n\theta^2}{2}} \int_{-\infty}^{\infty} g(t)\, e^{-\frac{nt^2}{2}} e^{nt\theta}\, dt$$

$$\mathbb{E}_\theta\{g(T)\} = 0 \quad \forall \theta$$

$$\Rightarrow \int_{-\infty}^{\infty} g(t)\, e^{-\frac{nt^2}{2}} e^{nt\theta}\, dt = 0 \quad \forall\, \theta$$

By uniqueness of integral transforms,

$$g(t)\, e^{-\frac{nt^2}{2}} = 0 \quad \forall\, t \in \mathbb{R}$$

$$e^{-\frac{nt^2}{2}} > 0 \quad \forall t \Rightarrow g(t) = 0 \quad \forall\, t \in \mathbb{R}.$$

$$\therefore \quad \mathbb{P}_\theta \{ g(T) = 0 \} = 1 \quad \forall\, \theta.$$

$$\therefore \quad T = \bar{X} \text{ is complete.}$$

Complete sufficient statistic for exponential family

The pdf or pmf is given by

$$e^{\sum_{j=1}^{k} a_j(\theta)\, \tau_j(x) - b(\theta)}\, c(x), \quad x \in \mathcal{J}$$

($\mathcal{J}$: free of θ)

If $X_1, \ldots, X_n$ are iid with the above pdf/pmf,

then $\left(\sum_{i=1}^{n} \tau_1(X_i), \ldots, \sum_{i=1}^{n} \tau_k(X_i) \right)$ is minimal sufficient for θ.

Result: Let $X_1, \ldots, X_n$ be iid with pdf/pmf

$$e^{\sum_{j=1}^{k} a_j(\theta)\, \tau_j(x) - b(\theta)}\, c(x), \quad x \in \mathcal{J}$$

Then, $T = \left(\sum_{i=1}^{n} \tau_1(X_i), \ldots, \sum_{i=1}^{n} \tau_k(X_i) \right)$ is complete sufficient for θ if $\{ (a_1(\theta), \ldots, a_k(\theta)) : \theta \in \Theta \}$ contains a k-dimensional open set.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\mu \in \mathbb{R}$, $\sigma > 0$

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

$a_1(\mu, \sigma^2) = \mu$, $a_2(\mu, \sigma^2) = \frac{1}{\sigma^2}$

$$\{ (a_1(\mu, \sigma^2), a_2(\mu, \sigma^2)) : (\mu, \sigma^2) \in \Theta \}$$

$$= \left\{ \left(\mu, \frac{1}{\sigma^2}\right) : \mu \in \mathbb{R}, \sigma > 0 \right\}$$

$$= \{ (\nu, \kappa) : \nu \in \mathbb{R}, \kappa > 0 \}, \quad \nu = \mu, \ \kappa = \frac{1}{\sigma^2}$$

$= \mathbb{R} \times (0, \infty)$ contains a (two-dimensional) open subset of $\mathbb{R}^2$.

$\therefore \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$ is complete sufficient for (μ, σ^2)

Eg. $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \theta)$, $\theta > 0$

$\sum_{i=1}^n X_i^2$ is minimal sufficient.

$\therefore \left(\sum X_i, \sum X_i^2\right)$ is not minimal sufficient.

$\therefore \left(\sum X_i, \sum X_i^2\right)$ is not complete.

Result: Let T be a complete sufficient statistic.

Then, any 1-1 function of T is also complete sufficient.

Pf: Suppose S is a 1-1 function of T.

$\exists\ h(\cdot)$ s.t. $S = h(T)$

$\mathbb{E}_\theta\{g(S)\} = 0 \quad \forall \theta \Rightarrow \mathbb{P}_\theta\{g(S) = 0\} = 1 \quad \forall \theta$

$$\mathbb{E}_\theta\{g(S)\} = \mathbb{E}_\theta\{g(h(T))\} = 0 \quad \forall\theta$$

$$\Leftrightarrow \mathbb{E}_\theta\{g\circ h(T)\} = 0 \quad \forall\theta$$

$$\Rightarrow \mathbb{P}_\theta\{g\circ h(T) = 0\} = 1 \quad \forall\theta$$

$$\Leftrightarrow \mathbb{P}_\theta\{g(S) = 0\} = 1 \quad \forall\theta$$

$\Rightarrow$ S is also complete sufficient.

Sufficiency principle: In a statistical inference problem, if there is a (minimal) sufficient statistic, then our inference should depend on the sample only via the (minimal) sufficient statistic.

That is, if $T = T(\underline{X})$ is (minimal) sufficient.

For two sets of observations $\underline{x}$ and $\underline{y}$, we have $T(\underline{x}) = T(\underline{y})$. Then, our inference should be the same irrespective of whether we observe $\underline{X} = \underline{x}$ or $\underline{X} = \underline{y}$.

Likelihood.

$$X_1 \cdots X_n \overset{iid}{\sim} f_\theta$$

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) = g_\theta(T(x))\, h(x)$$

$$p_\theta(\underline{y}) = g_\theta(T(\underline{y}))\, h(\underline{y})$$

$$\frac{p_\theta(\underline{x})}{p_\theta(\underline{y})} = \frac{h(x)}{h(y)} \quad \frac{g_\theta(T(x))}{g_\theta(T(y))}$$

Def: Let $X_1 \dots X_n$ have joint dist (pdf/pmf) $p_\theta(\cdot)$ which is indexed by $\theta \in \Theta$. The likelihood function $L(\theta)$ is defined as

$$L(\theta) = L(\theta | \underline{x}) = p_\theta(\underline{x}) \quad \forall \ \theta \in \Theta$$

Eg. $X_1 \dots X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

$$p_\theta(\underline{x}) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}.$$

$\Theta = \left\{ \frac{1}{4}, \frac{1}{2}, \frac{3}{4} \right\}$. We observe 5 successes out of 20 draws.

$$n = 20, \quad \sum_{i=1}^n x_i = 5$$

If $\theta = \frac{1}{4}$, then $L(\theta) = \left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^{15} = \frac{3^{15}}{4^{20}}$.

$\theta = \frac{1}{2}$, $L(\theta) = \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{15} = \frac{2^5 \cdot 2^{15}}{4^{20}} = \frac{2^{20}}{4^{20}}$.

$\theta = \frac{3}{4}$, $L(\theta) = \left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right)^{15} = \frac{3^5}{4^{20}}$.

θ	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
	3^{15}	2^{20}	3^5
	1,43,48,907 >>	10,48,576 >>	243

The likelihood: Let $X_1, \dots X_n$ be a random sample with jt. pdf / pmf $p_\theta(\underline{x})$, $\theta \in \Theta$. Then the likelihood of θ is defined as

$$L(\theta) = L(\theta | \underline{x}) = p_\theta(\underline{x}) \quad \forall \theta \in \Theta$$
$$\forall \underline{x} \in \mathcal{X}.$$

↑ viewed as a function of θ.

$L(\theta | \underline{x})$ tells us how likely it is that the true parameter is θ when we observe $\underline{x}$ as our sample.

Likelihood principle: Let $L(\theta|\underline{x})$ be the likelihood function. Suppose for two sets of observations $\underline{x}$ and $\underline{y}$, we have $L(\theta|\underline{x}) = C(\underline{x},\underline{y})\, L(\theta|\underline{y})$. Then, the inference on θ based on whether we observe $\underline{x}$ or $\underline{y}$ should be the same.

Eg. $X_1 \dots X_n \overset{i.i.d}{\sim} \text{Bernoulli}(\theta)$ $\quad 0<\theta<1$

$$L(\theta|\underline{x}) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}, \quad 0 \le \sum_{i=1}^n x_i \le n$$

Let $\underline{y}$ be another observation.

$$L(\theta|\underline{y}) = \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n-\sum_{i=1}^n y_i}$$

$$\frac{L(\theta|\underline{x})}{L(\theta|\underline{y})} = \theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i} (1-\theta)^{\sum_{i=1}^n y_i - \sum_{i=1}^n x_i}$$

$\underline{1}$ free of θ if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$

$L(\theta|\underline{x}) = L(\theta|\underline{y})$ if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$;

So, inference on θ based on two samples $\underline{x}$ and $\underline{y}$ should be the same if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$.

. For two samples if the no. of successes are the same, then our inference about θ should also match.

$$L(\theta|\underline{x}) = p_\theta(\underline{x})$$

$$\frac{L(\theta|\underline{x})}{L(\theta|\underline{y})} = \frac{p_\theta(\underline{x})}{p_\theta(\underline{y})}$$ ← is free of θ iff $T(\underline{x}) = T(\underline{y})$

$\Leftrightarrow$ $T(\underline{X})$ is min. suff.

$\therefore$ If $T(\underline{X})$ is minimal sufficient for θ and $T(\underline{x}) = T(\underline{y})$, then

$$L(\theta|\underline{x}) = C(\underline{x},\underline{y})\, L(\theta|\underline{y})$$

So, inference on θ based on $\underline{x}$ and that based on $\underline{y}$ should be the same.

Therefore, inference on θ based on $\underline{x}$ should so via $T(\underline{x})$.

That is, our inference based on the sample only depends on the minimal sufficient statistic.

Point Estimation

$X_1, \dots, X_n \overset{iid}{\sim} f_\theta$, $\theta \in \Theta$

We want to estimate θ.

A point estimator of θ is any statistic $T(\underline{X})$ that takes values in Θ.

With the observed sample $\underline{x}$, we get the estimate $T(\underline{x})$.

- Which T to use? How to estimate?
- What can we say about the qualities of T?

How to estimate? How do we construct estimators?

We will discuss about :
- Method of moments
- Maximum likelihood estimator

Method of moments estimators

$X_1, \dots, X_n \overset{iid}{\sim} f_\theta$. $\theta = (\theta_1, \dots, \theta_K)$

$\mu_j' = \mathbb{E}_\theta(X^j)$, $m_j' = \frac{1}{n}\sum_{i=1}^{n} X_i^j$

$\mu_j' = a_j(\theta)$: depends on θ ; $m_j' = b_j$: free of θ can be observed.

MoM estimators are obtained by solving :

$$\left.\begin{aligned} a_1(\theta) &= b_1 \\ a_2(\theta) &= b_2 \\ &\vdots \\ a_K(\theta) &= b_K \end{aligned}\right\}$$

Another version:

$$\mu_1 = \mathbb{E}_\theta(X) \quad , \quad m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \leftarrow b_1$$

$\uparrow a_1(\theta)$

$$\mu_j = \mathbb{E}_\theta\left((X-\mu_1)^j\right) \quad , \quad m_j = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^j \qquad j \geq 2$$

$\uparrow \tilde{a}_j(\theta) \qquad j \geq 2 \qquad \uparrow \tilde{b}_j$

The MoM estimators of $\theta_1 - \theta_K$ are obtained by solving:

$$\left.\begin{aligned} a_1(\theta) &= b_1 \\ \tilde{a}_2(\theta) &= \tilde{b}_2 \\ &\vdots \\ \tilde{a}_K(\theta) &= \tilde{b}_K \end{aligned}\right\}$$

$$\mu_j' = \mathbb{E}_\theta\left(X^j\right) = \mathbb{E}_\theta\left((\underbrace{X-\mu_1}_{Y}+\mu_1)^j\right)$$

$$= \mathbb{E}_\theta\left\{\sum_{k=0}^{j}\binom{j}{k}(X-\mu_1)^{j-k}\mu_1^{\,k}\right\}$$

$$= \sum_{k=0}^{j}\binom{j}{k}\mu_1^{\,k}\,\underbrace{\mathbb{E}_\theta\left((X_1-\mu_1)^{j-k}\right)}_{\mu_{j-k}}$$

$$= \sum_{k=0}^{j}\binom{j}{k}\mu_1^{\,k}\,\mu_{j-k}$$

$$\mu_j = \mathbb{E}_\theta\left((X-\mu_1)^j\right) = \mathbb{E}_\theta\left\{\sum_{k=0}^{j}\binom{j}{k}X^k(-\mu_1)^{j-k}\right\}$$

$$= \sum_{k=0}^{j}(-1)^{j-k}\binom{j}{k}\mu_1^{(j-k)}\,\mu_k'$$

- We can get the central moments from the raw moments by using linear combinations.

  We can also get the raw moments from the central moments using linear combinations.

- This is true for both the population moments and the sample moments.

Therefore, the MOM estimators based on the two approaches are the same.

We use one or the other based on which one is more convenient to us.

Maximum likelihood estimator:

$$X_1, \ldots, X_n \overset{iid}{\sim} f_\theta, \qquad \theta \in \Theta$$

$$L(\theta | \underline{x}) = \prod_{i=1}^{n} f_\theta(x_i), \qquad \theta \in \Theta$$

The maximum likelihood estimator (MLE) of θ is the maximizer of the likelihood.

$$\hat{\theta} = \hat{\theta}(\underline{x}) = \underset{\theta \in \Theta}{\operatorname{argmax}} \; L(\theta | \underline{x}) \qquad \underline{x} \in \mathcal{X}.$$

$\hat{\theta}(\underline{X})$ is the MLE.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $0 \le \theta \le 1$.

$$L(\theta | \underline{x}) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}, \quad 0 \le \theta \le 1$$

log-likelihood:

$$\ell(\theta | \underline{x}) = \log L(\theta | \underline{x}) = \sum_{i=1}^n x_i \log \theta + \left(n - \sum_{i=1}^n x_i\right) \log(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ell(\theta | \underline{x}) = \sum_{i=1}^n x_i \frac{1}{\theta} - \left(n - \sum x_i\right) \frac{1}{1-\theta}$$

$$\frac{\partial^2}{\partial \theta^2} \ell(\theta | \underline{x}) = -\frac{\sum x_i}{\theta^2} - \frac{(n - \sum x_i)}{(1-\theta)^2} < 0 \quad \forall \theta \in (0,1)$$

$$\frac{\partial}{\partial \theta} \ell(\theta | \underline{x}) = 0 \iff \frac{\sum x_i}{\theta} = \frac{n - \sum x_i}{1-\theta} \iff \frac{1-\theta}{\theta} = \frac{n - \sum x_i}{\sum x_i}$$

$$\frac{1}{\theta} = \frac{n}{\sum x_i} \therefore \theta = \frac{\sum x_i}{n} = \bar{x}. \quad \hat{\theta}_{MLE} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Eg. $X_1 \cdots X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$, $\theta > 0$

$$L(\theta | \underline{x}) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & 0 \le x_1 \cdots x_n \le \theta \\ 0, & \text{o.w.} \end{cases}$$

$$= \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} \le \theta\} \mathbb{1}\{x_{(1)} \ge 0\}$$

$\theta \mapsto \frac{1}{\theta^n}$ is decreasing.

$\therefore L(\theta | \underline{x})$ is max. at the smallest value of θ which is $x_{(n)}$.

$\therefore \hat{\theta}_{MLE} = X_{(n)}$.

Eg. $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $0 \le \theta \le 1$

θ : prob. of success in one trial.

θ^2 : prob of getting two successes in two trials

$\hat{\theta}_{MLE} = \bar{X}$

What is the MLE of θ^2? $\bar{X}^2$. Since MLEs are invariant.

$X_1, \ldots, X_n \sim \underbrace{p_\theta(\cdot)}_{\text{jt. dist (pdf/pmf) of the random sample}}, \quad \theta \in \Theta$

likelihood function is

$$L(\theta) = L(\theta | \underline{x}) = p_\theta(\underline{x}), \quad \theta \in \Theta, \ \underline{x} \in \mathcal{X}$$

Maximum likelihood estimator (MLE) is

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \ L(\theta | \underline{X})$$

$$\hat{\theta}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}} \ L(\theta | \underline{x}) \qquad \forall \ \underline{x} \in \mathcal{X}$$

$$\hat{\theta}_{MLE} = \hat{\theta}(\underline{X}).$$

log-likelihood

$$\ell(\theta) = \ell(\theta|\underline{x}) = \log L(\theta|\underline{x}), \quad \theta \in \Theta, \quad \underline{x} \in \mathcal{X}.$$

Since $t \mapsto \log(t)$ is strictly increasing

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\ L(\theta|\underline{x}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta|\underline{x})$$

If $\ell(\theta) = \ell(\theta|\underline{x})$ is twice differentiable,

Then $\hat{\theta}_{MLE}$ is found by solving

$$\ell'(\theta) = 0 \longleftarrow \text{likelihood equation.}$$

and verifying $\ell''(\hat{\theta}) < 0$

$\hat{\theta}$ is obtained as a solution to the likelihood equation.

The parameter space Θ plays an important role in the determination of MLE.

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

$\hat{\theta}_{MLE} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (True if $\theta \in [0,1]$)

Suppose $\Theta = \left(\frac{1}{4}, \frac{1}{2}\right)$ $\quad \frac{1}{4} < \theta < \frac{1}{2}$

$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}}\ L(\theta | \underline{x})$

Eg. $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$ $\quad \Theta = \mathbb{R}$

$\hat{\mu}_{MLE} = \bar{X}$

If $\mu > 0$. $\quad \hat{\theta}_{MLE} = \begin{cases} \bar{X} & \text{if } \bar{X} > 0 \\ 0 & \text{if } \bar{X} \le 0 \end{cases}$ (Technically undefined)

MLE of different parameterization

Eg. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $\theta \in [0,1]$

$\theta \leftarrow$ prob. of success in a single trial

$\theta^2 \leftarrow$ prob. of two successes in two trials

We are interested in θ^2.

What is the MLE of θ^2? $\left(\bar{X}^2\right)$

$$L(\theta) = L(\theta \mid \underline{x}) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}, \quad \theta \in [0,1]$$

$$\left(t = \sum_{i=1}^n x_i\right) \qquad = \theta^t (1-\theta)^{n-t}, \quad \theta \in [0,1]$$

$\eta = \theta^2 \iff \theta = \sqrt{\eta}$. $\quad \theta \in [0,1] \iff \eta \in [0,1]$

↑ parameter of interest.

$$\tilde{L}(\eta) = [L(\theta)]_{\theta=\sqrt{\eta}} = [\theta^t (1-\theta)^{n-t}]_{\theta=\sqrt{\eta}}$$

$$= (\sqrt{\eta})^t (1-\sqrt{\eta})^{n-t}, \quad \eta \in [0,1]$$

$$\tilde{\ell}(\eta) = \frac{t}{2} \log \eta + (n-t) \log (1-\sqrt{\eta})$$

$$\tilde{\ell}'(\eta) = \frac{t}{2\eta} + (n-t) \frac{-\frac{1}{2\sqrt{\eta}}}{1-\sqrt{\eta}}$$

$$= \frac{t}{2\eta} - \frac{n-t}{2\sqrt{\eta}(1-\sqrt{\eta})}$$

$$\tilde{\ell}''(\eta) = -\frac{t}{2\eta^2} + \frac{(n-t)\frac{1}{2\sqrt{\eta}} - 1}{2\eta(1-\sqrt{\eta})^2}$$

$\tilde{\ell}''(\bar{x}) < 0$ (check this).

$\therefore \hat{\theta}_{MLE} = \bar{X}^2$

$\tilde{\ell}'(\eta) = 0$

$\Leftrightarrow \frac{t}{2\eta} = \frac{n-t}{2\sqrt{\eta}(1-\sqrt{\eta})}$

$\Leftrightarrow \frac{t}{n-t} = \frac{\sqrt{\eta}}{1-\sqrt{\eta}}$

$\Leftrightarrow \frac{t}{n} = \sqrt{\eta}$

$\Leftrightarrow \eta = \left(\frac{t}{n}\right)^2 = \bar{x}^2$

Now, take an arbitrary likelihood $L(\theta) = L(\theta | \underline{x})$, $\theta \in \Theta$

Suppose g is a 1-1 function

We are interested in $\eta = g(\theta)$, $\eta \in g(\Theta) = \{g(\theta) : \theta \in \Theta\}$

$\eta = g(\theta)$, g is 1-1 $\Leftrightarrow$ $\theta = g^{-1}(\eta)$

Now, $\tilde{L}(\eta) = [L(\theta)]_{\theta = g^{-1}(\eta)} = L(g^{-1}(\eta))$, $\eta \in g(\Theta)$

$$\hat{\eta}(\underline{x}) = \underset{\eta \in g(\Theta)}{\operatorname{argmax}} \tilde{L}(\eta | \underline{x})$$

$$= \underset{\eta \in g(\Theta)}{\operatorname{argmax}} L(g^{-1}(\eta))$$

$L(\theta)$ is max. if $\theta = \hat{\theta}(\underline{x}) \Rightarrow \tilde{L}(\eta)$ is max if $g^{-1}(\eta) = \hat{\theta}(\underline{x})$

$$\hat{\eta}(x) = \underset{\eta \in g(\Theta)}{\operatorname{argmax}} \; \tilde{L}(\eta | \underline{x}) = \underset{\eta \in g(\Theta)}{\operatorname{argmax}} \; L(g^{-1}(\eta) | \underline{x})$$

$L(\theta | \underline{x})$ is max. if $\theta = \hat{\theta}(\underline{x}) \Rightarrow \tilde{L}(\eta | \underline{x})$ is max if

$$g^{-1}(\eta) = \hat{\theta}(\underline{x})$$

$$\Leftrightarrow \tilde{L}(\eta | \underline{x}) \text{ is max. if}$$

$$\eta = g(\hat{\theta}(\underline{x}))$$

$$\therefore \hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})$$

This property holds even if g is not 1-1.

Result: Suppose that in a particular statistical problem $\hat{\theta}_{MLE}$ is the MLE of a parameter θ. If $\eta = g(\theta)$ is another parameter, then the MLE of η is $\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})$.

In many cases, closed-form expression for MLE does not exist.

Eg. Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ Cauchy $(\theta, 1)$. $\theta \in \mathbb{R}$

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}, \quad x \in \mathbb{R}$$

$$L(\theta) = L(\theta | \underline{x}) = \prod_{i=1}^n f_\theta(x_i) = \frac{1}{\pi^n} \prod_{i=1}^n \frac{1}{1+(x_i-\theta)^2}$$

We know that a unique maximizer of $L(\theta)$ exists.
A closed-form expression for the MLE does not exist.

In such cases, we use some kind of optimization technique to find the MLE.

Suppose $\ell(\theta|\underline{x}) = \log L(\theta|\underline{x})$, $\theta \in \Theta$

Let ℓ' and ℓ'' exist

So, $\hat{\theta}_{MLE}$ is obtained by solving $\ell'(\theta) = 0$.

However, if cannot be solved analytically.

Newton - Raphson method : To find the root of $f(x) = 0$

Start with x_0 and define $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$\vdots$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

till convergence.

The root is taken to be

The convergent point.

In our case, $f = \ell'$, $f' = \ell''$.

So, our iterative procedure would look like:

- start with $\theta_0 \leftarrow$ initial guess
- Define $\theta_{n+1} = \theta_n - \dfrac{\ell'(\theta_n)}{\ell''(\theta_n)}$, $n = 0, 1, 2, \ldots$
- continue till convergence.
- $\hat{\theta}_{MLE}$ is the convergent point

$$\ell(\theta) = \ell(\theta|x) = \log L(\theta|x) = \log\left(\prod_{i=1}^{n} f_\theta(x_i)\right)$$

$$= \sum_{i=1}^{n} \log f_\theta(x_i)$$

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(x_i)$$

$$\ell''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(x_i)$$

If often holds:

$$E_\theta\left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right] = 0$$

$$E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right)^2\right]$$

$$= -E_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(X)\right]$$

$I(\theta)$ = Fisher's information

If, instead of using $\ell''(\theta)$, we use $\mathbb{E}_\theta[\ell''(\theta|\underline{X})]$ in Newton-Raphson iterations, then the resulting algorithm is known as Fisher's scoring.

So, the iterates become

$$\theta_{n+1} = \theta_n - \frac{\ell'(\theta_n)}{\mathbb{E}_\theta[\ell''(\theta_n|\underline{X})]}$$

$$= \theta_n + \frac{\ell'(\theta_n)}{n\, I(\theta_n)}, \text{ where } I(\theta) \text{ is the Fisher information.}$$

MLE with multiple parameters:

$$L(\underline{\theta}) = L(\underline{\theta} \mid \underline{x}) = p_{\underline{\theta}}(\underline{x}) \quad , \theta \in \Theta \quad , \forall \underline{x} \in \mathcal{X}$$

$$\hat{\underline{\theta}}(\underline{x}) = \underset{\underline{\theta} \in \Theta}{\operatorname{argmax}} \; L(\underline{\theta} \mid \underline{x})$$

$$\hat{\theta}_{MLE} = \hat{\underline{\theta}}(\underline{X})$$

$$\ell(\underline{\theta} \mid \underline{x}) = \log L(\underline{\theta} \mid \underline{x}) \, , \; \underline{\theta} \in \Theta$$

If $\ell(\underline{\theta})$ is twice-differentiable (all second order partial derivatives exist)

Then we solve

$\nabla \ell(\hat{\underline{\theta}}) = 0$ ← likelihood equations

$\nabla^2 \ell(\hat{\underline{\theta}})$ is negative-definite.

In case a closed form does not exist, we can use a similar algorithm to Newton-Raphson.

$\underline{\theta}_0$

$$\underline{\theta}_{n+1} = \underline{\theta}_n - \left(\nabla^2 \ell(\underline{\theta}_n)\right)^{-1} \nabla \ell(\theta_n), \quad n = 0, 1, \ldots$$

till convergence.

Fisher scoring:

$$\underline{\theta}_{n+1} = \underline{\theta}_n - \left(n\, I(\theta_n)\right)^{-1} \nabla \ell\,(\underline{\theta}_n)$$

$I(\theta)$ is the Fisher's information matrix.

How to judge an estimator?

We have a model $X_1, \dots, X_n \sim f_\theta$, $\theta \in \Theta$

θ is the parameter of interest.

$T = T(\underline{X})$ is our estimator.

$|T - \theta|^2$ ← gives the departure of our estimator from the truth

↑ a random variable

Take expectation:

$MSE_\theta(T) = \mathbb{E}_\theta (T-\theta)^2$ ← mean squared error.

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

$$= \mathbb{E}_\theta \{ (T - \mathbb{E}_\theta(T))^2 \} + \mathbb{E}_\theta \{ (\mathbb{E}_\theta(T) - \theta)^2 \} + 2\,\mathbb{E}_\theta \{ (T - \mathbb{E}_\theta(T))(\mathbb{E}_\theta(T) - \theta) \}$$

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

$$= Var_\theta(T) + Bias_\theta^2(T)$$

$Bias_\theta(T)$ or $B_\theta(T)$

↑ bias variance decomposition

We would like to have the estimator which has the minimum MSE.

More precisely, we would like to have T such that

$$MSE_\theta(T) \leq MSE_\theta(S) \quad \forall\, \theta \in \Theta \text{ and all estimator } S.$$

No such estimator exists.

Take any $\theta_0 \in \Theta$

Define $S = \theta_0$ with prob 1. ← The statistic S always takes the value θ_0 irrespective of the sample

Then, $MSE_{\theta_0}(S) = 0$

and $MSE_{\theta_0}(T) > 0$ unless $T = \theta_0$ with prob 1.

So, setting an estimator which has lower MSE than any other estimator for all values of the parameter is not possible.

To set something meaningful, we need to restrict our class of estimators.

In particular, we focus on estimators which are unbiased. That is, $Bias_\theta(T) = 0 \quad \forall \theta$

Def: An estimator T of θ is said to be unbiased if $\mathbb{E}_\theta(T) = \theta \quad \forall \theta \in \Theta$.

Def: An estimator T is said to be best unbiased estimator or uniformly minimum variance unbiased estimator (UMVUE) of θ if it has the least variance among all unbiased estimators of θ.

That is: • $\mathbb{E}_\theta(T) = \theta \quad \forall \theta \in \Theta$

• $\forall S$ s.t. $\mathbb{E}_\theta(S) = \theta \quad \forall \theta \in \Theta$

$$\text{var}_\theta(T) \leq \text{var}_\theta(S) \quad \forall \theta \in \Theta$$

This also entails that $MSE_\theta(T) \leq MSE_\theta(S) \quad \forall \theta \in \Theta$.

$p_\theta(\underline{x})$ ⟵ joint pmf / pdf of the n random sample $(X_1 \cdots X_n)$

$X_1, \cdots X_n$ ⟵ random sample

$p_\theta(\underline{x}) = \mathbb{P}_\theta\left[X_1 = x_1, \cdots, X_n = x_n\right]$ ⟵ $X_1 \cdots X_n$ discrete

$=$ jt pdf of $(X_1 \cdots X_n)$ evaluated at $(x_1 \cdots x_n)$

Take any estimator $T = T(\underline{X})$. Then,

$$\mathbb{E}_\theta\left[T(\underline{X})\right] = \int_{\mathcal{X}} T(\underline{x})\, p_\theta(\underline{x})\, d\underline{x}$$

, $\mathcal{X}$ is the support of $p_\theta(\cdot)$

$\mathcal{X}$ can depend on θ.

If the LHS is differentiable w.r.t. θ, then

$$\frac{d}{d\theta}\,\mathbb{E}_\theta\left[T(\underline{X})\right] = \frac{d}{d\theta}\int_{\mathcal{X}} T(\underline{x})\, p_\theta(\underline{x})\, d\underline{x}$$

$$\frac{d}{d\theta}\,\mathbb{E}_\theta[T(\underline{x})] = \frac{d}{d\theta}\int_{\mathcal{X}} T(\underline{x})\; p_\theta(\underline{x})\; d\underline{x}$$

$$= \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left\{T(\underline{x})\, p_\theta(\underline{x})\right] d\underline{x}$$

(assuming this change of differentiation and integral is possible)

$$= \int_{\mathcal{X}} T(\underline{x})\; \frac{\partial}{\partial\theta}\, p_\theta(\underline{x})\; d\underline{x}$$

$$= \int_{\mathcal{X}} T(\underline{x})\; \frac{\frac{\partial}{\partial\theta} p_\theta(\underline{x})}{p_\theta(\underline{x})}\; p_\theta(\underline{x})\; d\underline{x}$$

$$= \int_{\mathcal{X}} T(\underline{x})\; \frac{\partial}{\partial\theta} \log p_\theta(\underline{x})\; p_\theta(\underline{x})\; d\underline{x}$$

$$= \mathbb{E}_\theta\left[T(\underline{X})\, \frac{\partial}{\partial\theta} \log p_\theta(\underline{X})\right]$$

For any estimator $T = T(\underline{X})$,

$$\frac{d}{d\theta}\mathbb{E}_\theta[T(\underline{X})] = \mathbb{E}_\theta\left[T(\underline{X})\frac{\partial}{\partial\theta}\log p_\theta(\underline{X})\right]$$

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(\underline{X})\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, p_\theta(\underline{x})\, d\underline{x}$$

$$= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta}p_\theta(\underline{x})}{p_\theta(\underline{x})}\, p_\theta(\underline{x})\, d\underline{x}$$

$$= \int_{\mathcal{X}} \frac{\partial}{\partial\theta} p_\theta(\underline{x})\, d\underline{x}$$

$$= \frac{d}{d\theta}\int_{\mathcal{X}} p_\theta(\underline{x})\, d\underline{x} \quad \left[\begin{array}{l}\text{assuming the} \\ \text{interchange is possible}\end{array}\right]$$

$$= \frac{d}{d\theta}1 = 0$$

Now,

$$\mathrm{Cov}_\theta\left(T(\underline{X}), \frac{\partial}{\partial\theta}\log f_\theta(\underline{x})\right) = \mathbb{E}_\theta\left[T(\underline{X})\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right] - \mathbb{E}_\theta[T(\underline{X})]\,\underbrace{\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right]}_{=0}$$

$$= \mathbb{E}_\theta\left[T(\underline{X})\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right]$$

$$= \frac{d}{d\theta}\mathbb{E}_\theta[T(\underline{X})]$$

For two random variables Y and Z

$$\{\mathrm{cov}(Y,Z)\}^2 \le \mathrm{var}(Y)\,\mathrm{var}(Z)$$ Cauchy-Schwarz inequality.

$$\Rightarrow \mathrm{var}(Y) \ge \frac{\{\mathrm{cov}(Y,Z)\}^2}{\mathrm{var}(Z)}$$

$$\therefore \quad \mathrm{var}_\theta\left(T(\underline{X})\right) \geq \frac{\left\{\mathrm{cov}_\theta\left(T(\underline{X}), \frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)\right\}^2}{\mathrm{var}_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right)}$$

$$= \frac{\left\{\frac{d}{d\theta}\,\mathbb{E}_\theta\left(T(\underline{X})\right)\right\}^2}{\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right\}^2\right]}, \quad \left[\because \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right] = 0\right]$$

So, the variance of any estimator must satisfy this bound. This bound is known as the Cramér-Rao lower bound (for the variance of an estimator).

If T is an unbiased estimator of θ, then

$$\mathrm{var}_\theta\left(T(\underline{X})\right) \geq \frac{1}{\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(\underline{X})\right\}^2\right]}.$$

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(\underline{X})\right] = 0 \qquad \forall\theta$$

$$\Leftrightarrow \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, p_\theta(\underline{x})\, d\underline{x} = 0 \qquad \forall\theta$$

$$\Rightarrow \frac{d}{d\theta}\int_{\mathcal{X}} \frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, p_\theta(\underline{x})\, d\underline{x} = 0$$

$$\Rightarrow \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, p_\theta(\underline{x})\right\} d\underline{x} = 0$$

$$\Rightarrow \int_{\mathcal{X}} \left\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(\underline{x})\, p_\theta(\underline{x}) + \frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, \underbrace{\frac{\partial}{\partial\theta} p_\theta(\underline{x})}\right\} d\underline{x} = 0$$

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\, p_\theta(\underline{x})$$

$$\Rightarrow \int_{\mathcal{X}} \left\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(\underline{x})\, p_\theta(\underline{x}) + \left[\frac{\partial}{\partial\theta}\log p_\theta(\underline{x})\right]^2 p_\theta(\underline{x})\right\} d\underline{x} = 0$$

$$\int_{\underline{x}} \frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{x}) \; p_\theta(\underline{x}) \, d\underline{x} + \int_{\underline{x}} \left\{ \frac{\partial}{\partial \theta} \log p_\theta(\underline{x}) \right\}^2 p_\theta(\underline{x}) \, d\underline{x} = 0$$

$$\Rightarrow \int_{\underline{x}} \left\{ \frac{\partial}{\partial \theta} \log p_\theta(\underline{x}) \right\}^2 p_\theta(\underline{x}) \, d\underline{x} = - \int_{\underline{x}} \frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{x}) \, p_\theta(\underline{x}) \, d\underline{x}$$

$$\Rightarrow \mathbb{E}_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log p_\theta(\underline{X}) \right\}^2 \right] = - \mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{X}) \right]$$

Then,

$$\mathrm{Var}_\theta\left(T(\underline{X})\right) \geq \frac{1}{- \mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{X}) \right]} .$$

When $X_1, \cdots, X_n \overset{iid}{\sim} f_\theta$

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i)$$

$$\log p_\theta(\underline{x}) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

$$\frac{\partial}{\partial\theta} \log p_\theta(\underline{x}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(x_i)$$

$$\frac{\partial^2}{\partial\theta^2} \log p_\theta(\underline{x}) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(x_i)$$

$$\therefore -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log p_\theta(\underline{x})\right] = -\sum_{i=1}^{n} \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(x_i)\right]$$

$$\therefore -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log p_\theta(\underline{x})\right] = -n\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(x_1)\right]$$

$$= n\,I(\theta).$$

$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_1)\right]$ is the Fisher information.

We can also start with

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{x}) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(x_i)$$ and show that

$$\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\log p_\theta(\underline{X})\right\}^2\right] = n\,\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right\}^2\right]$$

$$= n\,I(\theta).$$

$$\therefore \operatorname{var}_\theta\left(T(\underline{X})\right) \geq \frac{\left[\frac{d}{d\theta}\mathbb{E}_\theta\left(T(\underline{X})\right)\right]^2}{n\,I(\theta)}$$

If T is unbiased for θ, then $\frac{d}{d\theta}\mathbb{E}_\theta(T(\underline{X})) = 1$

If T is unbiased for $a(\theta)$, then $\frac{d}{d\theta}\mathbb{E}_\theta(T(\underline{X})) = a'(\theta)$.