Lecture 5: Sufficiency & the Rao-Blackwell Theorem

STA261 – Probability & Statistics II

Ofir Harari

Department of Statistical Sciences

University of Toronto

Outline

Sufficient statistics

- We have discussed the likelihood function extensively (and will continue doing so)
- It drives parameter estimation via the Maximum Likelihood principle
- We have studied the many good large-sample properties of MLEs (asymptotic normality & efficiency)
- We shall see that it also drives inference on parameters (i.e. hypothesis testing & confidence intervals)
- All in all, the likelihood is the single most important function in statistics, as summarized in the *Likelihood Principle*:

> In the inference about θ, after $\underline{x} = (x_1, \ldots, x_n)$ is observed, all relevant experimental information is contained in the likelihood function for the observed $\underline{x}$.

Sufficient statistics (cont.)

- So, all information about θ is encoded in the likelihood
- But what if the likelihood itself depends on the data through a mere summary (statistic)?
- Consider, for example, a sequence of Bernoulli trials,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Binom}(1, p),$$

with the likelihood function

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

- If we know the total number of "successful" trials, nothing more can be learned on p from knowing the detailed observations (i.e. the sequences $(1,0,1,0,1)$ and $(1,1,1,0,0)$ are the equivalent from a likelihood standpoint).
- Data compression!

Sufficient statistics (cont.)

Definition

A statistic $T(\underline{X}) = T(X_1, \ldots, X_n)$ is *sufficient* for an unknown parameter θ if the conditional (joint) distribution of $X_1, \ldots, X_n$ given $T(\underline{X})$ does not depend on θ.

- In other words: $T(\underline{X})$ teaches us all we need to know about θ.
- To continue with the Bernoulli trials example, let us now verify that $\sum_{i=1}^n X_i$ is indeed sufficient for p.
- Note that $\sum_{i=1}^n X_i \sim \text{Binom}(n, p)$, thus

$$\mathbb{P}\left(\underline{X} = \underline{x} \,\middle|\, \sum_{i=1}^n X_i = t\right) = \frac{\mathbb{P}\left(\underline{X} = \underline{x} \,,\, \sum_{i=1}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)}$$

$$= \begin{cases} \dfrac{\mathbb{P}\left(\underline{X} = \underline{x}\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} & , \quad \sum_{i=1}^n X_i = t \\ 0 & , \quad \text{otherwise} \end{cases}$$

Sufficient statistics (cont.)

$$\mathbb{P}\left(\underline{X}=\underline{x}\,\middle|\,\sum_{i=1}^{n}X_i=t\right)=\begin{cases}\dfrac{\mathbb{P}\left(\underline{X}=\underline{x}\right)}{\mathbb{P}\left(\sum_{i=1}^{n}X_i=t\right)} & , \quad \sum_{i=1}^{n}X_i=t\\ 0 & , \quad \text{otherwise}\end{cases}$$

$$=\begin{cases}\dfrac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} & , \quad \sum_{i=1}^{n}X_i=t\\ 0 & , \quad \text{otherwise}\end{cases}$$

$$=\begin{cases}\dfrac{1}{\binom{n}{t}} & , \quad \sum_{i=1}^{n}X_i=t\\ 0 & , \quad \text{otherwise}\end{cases}.$$

The above does not depend on p, hence $\sum_{i=1}^{n}X_i$ is sufficient for p.

The Fisher–Neyman Factorization Theorem

- You may have noticed that a direct verification of sufficiency can be messy
- Intuitively, the likelihood depends on the data via $\sum_{i=1}^{n} X_i$, ergo, it is sufficient
- Our intuition is right this time!

Theorem

A statistic $T(\underline{X})$ is sufficient for $\theta \iff$ for any value of θ we can write

$$\mathcal{L}(\theta) = g\left(T(\underline{x}), \theta\right) h(\underline{x}).$$

- Note that in the binary case

$$\mathcal{L}(p) = \underbrace{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}_{g\left(p, \sum_{i=1}^n x_i\right)},$$

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

Fisher–Neyman Factorization Theorem (cont.)

Jerzy Neyman, 1894-1981

Source: statistics.berkeley.edu

Fisher–Neyman Factorization Theorem (cont.)

Proof for the discrete case:

$(\Longrightarrow)$

Suppose that $T(\underline{X})$ is sufficient for θ, and let $\underline{x}$ be our sample, such that $T(\underline{x}) = t$.

$$\mathcal{L}(\theta) = \mathbb{P}\left(\underline{X} = \underline{x} \middle| \theta\right) = \mathbb{P}\left(\underline{X} = \underline{x}, T(\underline{X}) = t \middle| \theta\right)$$

$$= \mathbb{P}\left(\underline{X} = \underline{x} \middle| T(\underline{X}) = t, \theta\right) \mathbb{P}\left(T(\underline{X}) = t \middle| \theta\right)$$

- $\mathbb{P}\left(\underline{X} = \underline{x} \middle| T(\underline{X}) = t, \theta\right)$ does not depend on θ (why?) – call it $h(\underline{x})$
- Denote $g(t, \theta) := \mathbb{P}\left(T(\underline{X}) = t \middle| \theta\right)$ – and we're done.

Fisher–Neyman Factorization Theorem (cont.)

Proof (cont.):

($\Longleftarrow$)

Suppose now that the likelihood can be factorized as

$$\mathcal{L}(\theta) := \mathbb{P}(\underline{X} = \underline{x} \big| \theta) = g\left(T(\underline{x}), \theta\right) h(\underline{x}).$$

Note that

$$\mathbb{P}\left(T(\underline{X}) = t\right) = \sum_{T(\underline{x})=t} \mathbb{P}\left(\underline{X} = \underline{x} \big| \theta\right) = g(t, \theta) \sum_{T(\underline{x})=t} h(\underline{x}),$$

hence

$$\mathbb{P}(\underline{X} = \underline{x} \big| T(\underline{X}) = t) = \begin{cases} \dfrac{\mathbb{P}\left(\underline{X} = \underline{x}, T(\underline{X}) = t \big| \theta\right)}{\mathbb{P}\left(T(\underline{X}) = t\right)} & , \quad T(\underline{x}) = t \\ 0 & , \quad \text{otherwise} \end{cases}$$

Fisher–Neyman Factorization Theorem (cont.)

$$\mathbb{P}(\underline{X} = \underline{x} \big| T(\underline{X}) = t) = \begin{cases} \dfrac{\mathbb{P}\left(\underline{X} = \underline{x}, T(\underline{X}) = t \big| \theta\right)}{\mathbb{P}\left(T(\underline{X}) = t\right)} & , \quad T(\underline{x}) = t \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{\mathbb{P}\left(\underline{X} = \underline{x} \big| \theta\right)}{g(t,\theta) \sum_{T(\underline{x})=t} h(\underline{x})} & , \quad T(\underline{x}) = t \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{g(t,\theta) h(\underline{x})}{g(t,\theta) \sum_{T(\underline{x})=t} h(\underline{x})} & , \quad T(\underline{x}) = t \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{h(\underline{x})}{\sum_{T(\underline{x})=t} h(\underline{x})} & , \quad T(\underline{x}) = t \\ 0 & , \quad \text{otherwise} \end{cases} ,$$

that does not depend on θ.

Example: Poisson distribution

Example

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$. Find a sufficient statistic for λ.

Solution:

- Here $\mathcal{L}(\lambda) = \mathrm{e}^{-n\lambda} \dfrac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$
- Write
$$\mathcal{L}(\lambda) = g\left(\sum_{i=1}^n x_i, \lambda\right) h(\underline{x})$$
where $g\left(\sum_{i=1}^n x_i, \lambda\right) = \mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}$ and $h(\underline{x}) = \dfrac{1}{\prod_{i=1}^n x_i!}$
- Then, from the factorization Theorem, $\displaystyle\sum_{i=1}^n X_i$ is sufficient for λ.

Example: the Cauchy distribution

Example

Let $X_1, \ldots, X_n$ be a random sample from the Cauchy distribution, with pdf

$$f(x|\theta) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]}.$$

Does a sufficient statistics for θ exist?

Solution:

Here

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\pi^n \prod_{i=1}^{n}\left[1 + (x_i-\theta)^2\right]}$$

$$= \frac{1}{\pi^n} \exp\left\{-\sum_{i=1}^{n} \log\left[1 + (x_i-\theta)^2\right]\right\} = \ldots$$

- ⋆ Slice and dice it all you like, the x_i's and θ cannot be separated
- ⋆ No sufficient statistic, keep all the data

Example: Uniform distribution

Example

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{U}[0,\theta]$ (continuous uniform). Find a sufficient statistic for θ.

Solution:

Recall that

$$f(x|\theta) = \begin{cases} \dfrac{1}{\theta} & , \quad 0 \leq x \leq \theta \\ 0 & , \quad \text{otherwise} \end{cases} = \frac{1}{\theta} \cdot I\left\{0 \leq x \leq \theta\right\},$$

for

$$I\left\{0 \leq x \leq \theta\right\} = \begin{cases} 1 & , \quad x \in [0,\theta], \\ 0 & , \quad \text{otherwise.} \end{cases}$$

In light of this,

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I\left\{0 \leq x_i \leq \theta\right\} = \frac{1}{\theta^n} I\left\{x_{\max} \leq \theta\right\},$$

where $x_{\max} = \max(x_1, \ldots, x_n)$.

Example: Uniform distribution (cont.)

Solution (cont):

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} I\left\{x_{\max} \leq \theta\right\}$$

- We can write
$$\mathcal{L}(\theta) = g(x_{\max}, \theta) h(\underline{x})$$
for $g(x_{\max}, \theta) = \dfrac{1}{\theta^n} I\left\{x_{\max} \leq \theta\right\}$ and $h(\underline{x}) = 1$.
- Then by the factorization Theorem, $X_{\max} = \max\left(X_1, \ldots, X_n\right)$ is sufficient for θ.

The Exponential family of distributions

The following definition covers a surprisingly large subset of the distributions we have familiarized ourselves with.

Definition

A distribution with cdf/pmf $f(x|\theta)$ is said to belong to a *one parameter exponential family of distributions* if

$$f(x|\theta) = \begin{cases} \exp\{c(\theta)T(x) + d(\theta) + S(x)\} & , \quad x \in A \\ 0 & , \quad \text{otherwise} \end{cases}$$

where A does not depend on θ.

- Fantastic! What is it good for?
- Sufficiency, among other things

Sufficiency in Exponential families

- Suppose for a moment that we have a random sample from an exponential family, with
$$f(x\big|\theta) = \exp\left\{c(\theta)\,T(x) + d(\theta) + S(x)\right\}.$$

- Then, the likelihood is given by
$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{n} \exp\left\{c(\theta)\,T(x_i) + d(\theta) + S(x_i)\right\} \\
&= \exp\left\{c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta) + \sum_{i=1}^{n} S(x_i)\right\} \\
&= \underbrace{\exp\left\{c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta)\right\}}_{g\left(\sum_{i=1}^{n} T(x_i),\theta\right)} \underbrace{\exp\left\{\sum_{i=1}^{n} S(x_i)\right\}}_{h(\underline{x})}
\end{aligned}$$

- Evidently, $\sum_{i=1}^{n} T(X_i)$ is sufficient for θ.

Example: Poisson distribution

- Here

$$f(x_i \big| \lambda) = \frac{\mathrm{e}^{-\lambda} \lambda_i^x}{x_i!} = \exp\left\{-\lambda + x_i \log \lambda - \log x_i!\right\}$$

- Denoting $c(\lambda) = \log \lambda$, $T(x_i) = x_i$, $d(\lambda) = -\lambda$ and $S(x_i) = -\log x_i!$, we have

$$f(x_i \big| \lambda) = \exp\left\{c(\lambda)\, T(x_i) + d(\lambda) + S(x_i)\right\}.$$

- The Poisson family is an exponential family of distributions then
- And indeed, we have shown that $\sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$ is sufficient for λ

Example: Gamma distribution (λ known)

- Here $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda)$, i.e.

$$f(x_i|\alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\lambda x} = \exp\left\{\alpha \log \lambda - \log \Gamma(\alpha) + (\alpha - 1)\log x_i - \lambda x_i\right\}$$

$$= \exp\left\{\alpha \log x_i + \alpha \log \lambda - \log \Gamma(\alpha) - \log x_i - \lambda x_i\right\}, \; x \geq 0$$

- Set $c(\alpha) = \alpha, \quad T(x_i) = \log x_i, \quad d(\alpha) = \alpha \log \lambda - \log \Gamma(\alpha)$ and $S(x_i) = -\log x_i - \lambda x_i$, we can write

$$f(x_i|\alpha) = \exp\left\{c(\alpha)\, T(x_i) + d(\alpha) + S(x_i)\right\},$$

hence the Gamma family is an exponential family of distributions.

- Moreover, $\displaystyle\sum_{i=1}^n T(X_i) = \sum_{i=1}^n \log X_i$ is sufficient for α.

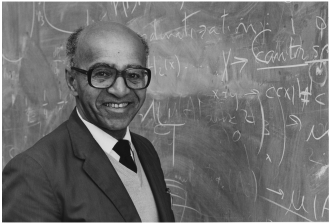
The Rao–Blackwell Theorem

Theorem

Let $\widehat{\theta}$ be an estimator of θ with a finite variance. Suppose that T is sufficient for θ, and let $\widehat{\theta}^* = \mathbb{E}\left[\widehat{\theta}\middle|T\right]$. Then for all θ

$$\mathrm{MSE}(\hat{\theta}^*, \theta) \leq \mathrm{MSE}(\hat{\theta}, \theta),$$

where equality holds $\iff \widehat{\theta}^* = \widehat{\theta}$.



David Blackwell, 1919-2010

Source: nationalmedals.org

The Rao–Blackwell Theorem (cont.)

Proof: Recall that from the law of total expectation

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}\left\{\mathbb{E}[\hat{\theta}\big| T]\right\} = \mathbb{E}[\hat{\theta}],$$

thus $\widehat{\theta}^*$ and $\widehat{\theta}^*$ have the same bias, and to compare their MSEs we need to compare their variances.

⋆ **In particular, if $\widehat{\theta}$ is unbiased then so is $\widehat{\theta}^*$.**

Now, applying the law of total variance we have –

$$\mathrm{Var}[\hat{\theta}^*] = \mathrm{Var}\left\{\mathbb{E}[\hat{\theta}\big| T]\right\} + \mathbb{E}\left\{\mathrm{Var}[\hat{\theta}\big| T]\right\} = \mathrm{Var}[\hat{\theta}] + \mathbb{E}\left\{\mathrm{Var}[\hat{\theta}\big| T]\right\} \geq \mathrm{Var}[\hat{\theta}],$$

where

$$\begin{array}{c}\text{equality}\\ \text{holds}\end{array} \Longrightarrow \mathrm{Var}[\hat{\theta}\big| T] = 0 \Longrightarrow \begin{array}{c}\widehat{\theta} \text{ is a constant}\\ \text{w.r.t. } \underline{X} \text{ when}\\ T \text{ is given}\end{array} \Longrightarrow \begin{array}{c}\widehat{\theta} \text{ is a}\\ \text{function of } T,\\ \text{say, } \widehat{\theta} = g(T)\end{array}$$

$$\Longrightarrow \widehat{\theta}^* = \mathbb{E}[\widehat{\theta}\big| T] = \mathbb{E}[g(T)\big| T] = g(T) = \widehat{\theta}$$

Comments on the Rao–Blackwell Theorem

- Where in the proof did we use the fact that $T(\underline{X})$ was sufficient for θ?
  - ⋆ Implicitly: we called $\widehat{\theta}^* = \mathbb{E}\big[\hat{\theta}\big|T\big]$ an "estimator", but that is only true because the distribution of $\widehat{\theta}^*$ does not depend on θ.

- It is tempting to re-apply Rao–Blackwellization to the resultant estimator – could it be further improved?
  - ⋆ Remember: $\widehat{\theta}_{\mathrm{RB}} = \mathbb{E}[\hat{\theta}_0\big|T] = g(T)$ (a function of T), therefore

$$\mathbb{E}[\hat{\theta}_{\mathrm{RB}}\big|T] = \mathbb{E}[g(T)\big|T] = g(T) = \hat{\theta}_{\mathrm{RB}},$$

    suggesting that the process stops after one stage.

- A follow-up result, the **Lehmann–Scheffé Theorem**:
  if, in addition to sufficiency, T has a property called *completeness*, Rao–Blackwellizing an unbiased estimator would yield the unique optimal unbiased estimator.

“Rao–Blackwellization” example

Example

Suppose that the annual number of earthquakes in a certain seismic region follows a Pois(λ) distribution, where different years are assumed to be independent. We wish to estimate the probability that there will be no earthquakes next year, based on a sample $X_1, \ldots, X_n$. “Rao–Blackwellize” the naive estimator

$$\widehat{\theta}_0 = \begin{cases} 1 & , \quad X_1 = 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

to obtain an improved estimator.

Solution:

- First, note that the parameter we wish to estimate is $\theta = \mathbb{P}(X = 0) = \mathrm{e}^{-\lambda}$
- Also note that

$$\mathbb{E}[\hat{\theta}_0] = \mathbb{P}(X_1 = 0) = \mathrm{e}^{-\lambda} = \theta,$$

hence $\widehat{\theta}_0$ (naive as it may be) is unbiased.

“Rao–Blackwellization” example (cont.)

Solution (cont.):

- We have verified that $T = \sum_{i=1}^{n} X_i$ is sufficient for λ, thus it is also sufficient for $\theta = \mathrm{e}^{-\lambda}$ (or any other monotonic transformation of λ)
- Keep in mind that $\sum_{i=1}^{n} X_i \sim \text{Pois}(n\lambda)$. Likewise, $\sum_{i=2}^{n} X_i \sim \text{Pois}\left((n-1)\lambda\right)$.
- Just like $\widehat{\theta}_0$, $\widehat{\theta}_0 \big| T$ is binary (returns either 0 or 1), hence

$$\widehat{\theta}_{\mathrm{RB}} := \mathbb{E}\left[\widehat{\theta}_0 \middle| T\right] = \mathbb{P}\left(\widehat{\theta}_0 = 1 \middle| \sum_{i=1}^{n} X_i = T\right) = \mathbb{P}\left(X_1 = 0 \middle| \sum_{i=1}^{n} X_i = T\right)$$

$$= \frac{\mathbb{P}\left(X_1 = 0\,,\, \sum_{i=1}^{n} X_i = T\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = T\right)} = \frac{\mathbb{P}\left(X_1 = 0\,,\, \sum_{i=2}^{n} X_i = T\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = T\right)}$$

"Rao–Blackwellization" example (cont.)

$$\widehat{\theta}_{\text{RB}} = \frac{\mathbb{P}\left(X_1 = 0\,,\, \sum_{i=2}^{n} X_i = T\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = T\right)} = \frac{\mathbb{P}\left(X_1 = 0\right)\mathbb{P}\left(\sum_{i=2}^{n} X_i = T\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = T\right)}$$

$$= \frac{\mathrm{e}^{-\lambda}\cdot \mathrm{e}^{-(n-1)\lambda}\frac{[(n-1)\lambda]^T}{T!}}{\mathrm{e}^{-n\lambda}\frac{[n\lambda]^T}{T!}} = \left(1-\frac{1}{n}\right)^{T} = \left(1-\frac{1}{n}\right)^{\sum_{i=1}^{n} X_i}$$

- Remember: $\widehat{\theta}_0$ was unbiased– thus so is $\widehat{\theta}_{\text{RB}}$
- It can be shown that it is the <u>best unbiased estimator</u> of $\theta = \mathrm{e}^{-\lambda}$, for all λ
- Not the best estimator of θ overall, though
- For large n

$$\widehat{\theta}_{\text{RB}} = \left(1-\frac{1}{n}\right)^{\sum_{i=1}^{n} X_i} = \left(1-\frac{1}{n}\right)^{n\overline{X}} \approx \mathrm{e}^{-\overline{X}} = \mathrm{e}^{-\widehat{\lambda}_{\text{MLE}}} = \widehat{\theta}_{\text{MLE}}$$

Comparison of $\widehat{\theta}_{\mathrm{RB}}$ and $\widehat{\theta}_{\mathrm{MLE}}$

```
#Calculating MSE by Monte-Carlo Simulation
MSEs <- function(lambda, n){
        theta <- exp(-lambda)
        samp <- matrix(rpois(n*1e5, lambda),
                       ncol=n)
        xBar <- apply(samp, 1, mean)
        thetaHat1 <- exp(-xBar)
        thetaHat2 <- (1-1/n)^(n*xBar)
        MSE1 <- mean((thetaHat1-theta)^2)
        MSE2 <- mean((thetaHat2-theta)^2)
        return(c(MSE1,MSE2))
}
>
> n <- 5
> Vals1 <- t(sapply(.1*c(1:40), MSEs, n=n))
> plot(.1*c(1:40), Vals1[,1], type='l')
> lines(.1*c(1:40), Vals1[,2], lty=2, col=2)
>
```