today: 3 things
HW4 is due tonight
OH today 2:30-4

06: SUFFICIENCY

Larsen & Marx 5.6

Prof Amanda Luby

1 Sufficient Estimators

So far, we've seen a few desirable properties for estimators: that they should be unbiased, that they should have minimum variance, and that they should converge to the parameter value with unlimited data. All of these properties are easy to motivate: they impose conditions on the probabilistic behavior of $\hat{\theta}$ that "make good sense". The next property we're going to introduce is not so intuitive, but has really important theoretical implications.

Assume we draw $X_1, .., X_n \sim f_x(\theta)$. Imagine two statisticians:

A 

$\left\{ X_1, X_2, \vdots, X_n \right. \Rightarrow \hat{\theta}_1 = g(X_i)$, $\hat{\theta}_2 = r(X_i)$, $\hat{\theta}_3 = X_1$, $\hat{\theta}_4 = X_{max}$, $\hat{\theta}_5 = h(X_i)$

$\hat{\theta}_1 = r(X)$, $\hat{\theta}_2 = g(r(X))$, $\hat{\theta}_3 = h(r(X))$, $\vdots$ $\Leftarrow r(X_i) \left. \right\}$  B

In general, A will be able to find a better estimator than B. BUT in some cases, B can just as well as statistician A. That's when we say $r(X_i)$ is 'sufficient' for θ

Whether or not an estimator is sufficient refers to the amount of "information" it contains about the unknown parameter. Estimates are calculated using values obtained from random samples (drawn from either p_x or f_x). If everything that we can possibly know from the data about θ is encapsulated in the estimate $\hat{\theta}$, then the corresponding estimator $\hat{\theta}$ is said to be sufficient.

Example of an estimator that is not sufficient: Let $Y_1, .., Y_n \sim f_y$, where $f_y = \frac{2y}{\theta^2}$ for $0 \leq y \leq \theta$. The MoM estimator for this distribution is $\hat{\theta}_{MoM} = \frac{3}{2}\bar{Y}$. Consider two random samples of size 3: $\{3, 4, 5\}$ and $\{1, 3, 8\}$.

sample 1: $\{3,4,5\}$ $\quad \hat{\theta}_{mom} = \frac{3}{2} \cdot 4 = 6$ ✓

sample 2: $\{1,3,8\}$: $\quad \hat{\theta}_{mom} = \frac{3}{2} \cdot 4 = 6$ X → we saw $X_3 = 8$, so we know that $\theta \geq 8$, so $\{X_1, X_2, X_3\}$ has more information about θ than $\frac{3}{2} \cdot \bar{Y}$. So $\hat{\theta}_{mom}$ is not sufficient.

In order to show, $P(W^n | \hat\theta = t) = \frac{P(W^n \cap \hat\theta = t)}{P(\hat\theta = t)}$

$\downarrow$ does not depend on θ

Sufficiency

Let $W_1, ..., W_n$ be a random sample from $f_w(w; \theta)$. The estimator $\hat{\theta} = h(W_1, ..., W_n)$ is said to be *sufficient* for θ if $P(W_1, ..., W_n | \hat{\theta} = t)$ does not depend on θ.

Factorization Criterion

Let $W_1, ..., W_n$ be a random sample from $f_w(w; \theta)$. The estimator $\hat{\theta} = h(W_1, ..., W_n)$ is sufficient for θ if and only if the likelihood function, $L(\theta)$, factors into the product of the pdf for $\hat{\theta}$ and a function of the sample that does not involve θ:

$$L(\theta) = \prod_{i=1}^{n} f_x(W_i; \theta) = f_{\hat\theta}(t; \theta) \cdot b(W_1, \ldots, W_n)$$

Example: Let $X_1, ..., X_n \sim Pois(\lambda)$. Show that $\hat{\lambda} = \sum X_i$ is a sufficient statistic for λ.

estimator $\Updownarrow$

$f_\lambda = \frac{e^{-\lambda}\lambda^k}{k!}$

① need pdf of $\hat\lambda = \sum X_i$. Sums of Poissons are Poissons

$\hat\lambda \sim Pois(n\lambda) \rightarrow f_{\hat\lambda} = \frac{e^{-n\lambda}(n\lambda)^{\sum X_i}}{(\sum X_i)!}$

② $$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = \frac{e^{-n\lambda}\lambda^{\sum X_i}}{\prod X_i!} \cdot \frac{n^{\sum X_i}}{n^{\sum X_i}} \cdot \frac{(\sum X_i)!}{(\sum X_i)!}$$

$$= \underbrace{\frac{e^{-n\lambda}(n\lambda)^{\sum X_i}}{(\sum X_i)!}}_{f_{\hat\lambda}} \cdot \underbrace{\frac{(\sum X_i)!}{n^{\sum X_i}\prod X_i!}}_{b(X_1, \ldots, X_n)}$$

By factorization criterion, $\hat\lambda = \sum X_i$ is sufficient for λ.

Factorization Criterion Round 2 Fisher-Neyman

Let $W_1, ..., W_n$ be a random sample from $f_w(w; \theta)$. The estimator $\hat{\theta} = h(W_1, ..., W_n)$ is sufficient for θ if if and only if the likelihood function, $L(\theta)$, factors into:

$$L(\theta) = \underbrace{g[h(W_1, \ldots, W_n); \theta]}_{\text{more relaxed}} \cdot b(W_1, \ldots, W_n)$$

since $g \neq f_{\hat\theta}(t)$

Example: Let $X_1, ..., X_n \sim Pois(\lambda)$. Show that $\hat{\lambda} = \sum X_i$ is a sufficient statistic for λ.

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$

$$= \underbrace{\left(e^{-n\lambda} \lambda^{\sum x_i}\right)}_{g(\sum x_i,\, \lambda)} \cdot \underbrace{\left(\frac{1}{\prod x_i!}\right)}_{b(x_1, \ldots, x_n)}$$

$\rightarrow$ by factorization criterion 2, $\hat{\lambda} = \sum x_i$ is sufficient for λ.

*Note: 1-1 functions of $\hat{\theta}_{suff}$ are also sufficient stats.

Proof:

Friday 10/6:
HW5 is posted → #1-3 today
#4 on Monday

* for proof, see end of document.
Taken from DeGroot & Schervish

$\mathbb{1}\{Y_i \le \theta\} = \begin{cases} 1 & Y_i \le \theta \\ 0 & \text{otherwise} \end{cases}$ $\quad f_y = \begin{cases} 2y/\theta^2 & 0 \le y \le \theta \\ 0 & y < 0 \text{ or } y > \theta \end{cases}$

$\mathbb{I}$

\mathbb

Example: Suppose $Y_1, \ldots, Y_n$ are drawn from $f_y(y;\theta) = \frac{2y}{\theta^2}$ where $0 \le y \le \theta$. The MLE for θ is $\hat{\theta} = Y_{max}$. Is Y_{max} sufficient for θ?

WTS: $L(\theta) = g(h(Y^n);\theta) \cdot b(Y^n)$
$= g(Y_{max};\theta) \cdot b(Y^n)$

$$L(\theta) = \prod_{i=1}^{n} \frac{2y_i}{\theta^2} \cdot \mathbb{1}\{Y_i \le \theta\}$$

$$= \prod_{i=1}^{n} 2y_i \cdot \left(\frac{1}{\theta^{2n}} \prod \mathbb{1}\{Y_i \le \theta\}\right) \leftarrow \mathbb{1}\{Y_1 \le \theta\}\mathbb{1}\{Y_2 \le \theta\} \cdots \mathbb{1}\{Y_n \le \theta\}$$

$= 1$ if all Y_i's $\le \theta$ } $\max Y_i \le \theta$

0 if any Y_i's $> \theta$ } $\max Y_i > \theta$

$$= \underbrace{\prod (2y_i)}_{b(Y^n)} \cdot \underbrace{\left(\frac{1}{\theta^{2n}} \cdot \mathbb{1}\{Y_{max} \le \theta\}\right)}_{g(Y_{max},\theta)}$$

By factorization theorem, Y_{max} is sufficient for θ

More notation: Estimator → Statistics

Estimator: $\hat{\theta}$ and they're always estimating θ parameter in pdf

Statistic: $T(X)$ can estimate anything (eg θ, $E(Y)$, $V(Y)$, $\frac{1}{\theta}$ ...)

Want to talk about specific values of estimator / statistic

$P(X \mid \hat{\theta} = \hat{\theta}_0)$

$P(X \mid T = t)$ ← more general, a little cleaner, more common in "advanced" materials

2 Jointly Sufficient Statistics

When a parameter θ is multidimensional, sufficient statistics will typically need to be multidimensional as well. Sometimes, no one-dimensional statistic is sufficient even when θ is one-dimensional. In either case, we need to extend the concept of sufficient statistic to deal with cases in which more than one statistic is needed in order to be sufficient.

Jointly Sufficient Statistics

Suppose that for each θ and each possible value of $(t_1, \ldots, t_k)$ of $(T_1, \ldots, T_k)$, where each $T_i = h_i(X_1, \ldots, X_n)$, the conditional joint distribution of $(X_1, \ldots, X_n)$ given $(T_1, \ldots, T_k) = (t_1, \ldots, t_k)$ does not depend on θ. Then $(T_1, \ldots, T_k)$ are called *jointly sufficient statistics* for θ

Factorization Theorem for Jointly Sufficient Statistics

Let $r_1, \ldots, r_k$ be functions. The statistics $T_i = r_i(X_1, \ldots, X_n)$ are jointly sufficient for θ if and only if the joint pdf $f(x_1, \ldots, x_n|\theta)$ can be factored into:

$L(\theta)$

$$L(\theta) = g[r_1(x^n), r_2(x^n), \ldots, r_k(x^n); \theta] \cdot b(x^n)$$

Example: Jointly sufficient statistics for the parameters of a normal distribution

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ $\quad f_x = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

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

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(x_i^2 - 2x_i\mu + \mu^2)}$$

$$= \underbrace{\left(\frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum x_i^2}\right)}_{r_1(x^n, \theta)} \underbrace{\left(e^{-\frac{\mu}{\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}}\right)}_{r_2(x^n, \theta)} \cdot \underbrace{1}_{b(x^n)}$$

$\rightarrow$ By joint factorization theorem,

$T_1 = \sum x_i$ and $T_2 = \sum x_i^2$ are sufficient for (μ, σ^2)

Round 2: $T_1' = \bar{x}$ $\quad T_2' = \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

If h + m are 1-1 functions, then if

$T_1' = h(T_1, T_2)$ and $T_2' = m(T_1, T_2)$ then

T_1' and T_2' are also sufficient

$T_1' = \frac{1}{n} T_1$

$T_1 = n T_1'$

$$\left.\begin{aligned} T_2' &= \frac{1}{n} T_2 - \frac{1}{n^2} T_1^2 \\ T_2 &= n(T_2' + T_1'^2) \end{aligned}\right\}$$ T_1' and T_2' are also sufficient

3 Rao-Blackwell Theorem

Mean Squared Error

$$\begin{aligned} MSE(\hat\theta, \theta) &= E_\theta[(\hat\theta - \theta)^2] \\ &= E[(\hat\theta - E(\hat\theta))^2] + (E(\hat\theta) - \theta)^2 \\ &= V(\hat\theta) + Bias(\theta, \hat\theta)^2 \end{aligned}$$

The following theorem says that if we want an estimator with small MSE we can confine our search to estimators which are functions of sufficient statistics.

Rao-Blackwell Theorem

Let $\hat\theta$ be an estimator of θ with $E(\hat\theta^2) < \infty$. Suppose that T is a sufficient estimator of θ, and let $\theta^* = E(\hat\theta|T)$. Then, for all θ,

$$E(\theta^* - \theta)^2 \leq E(\hat\theta - \theta)^2$$
$$MSE(\theta^*, \theta) \leq MSE(\hat\theta, \theta)$$

$\rightarrow$ inequality is strict unless $\hat\theta = f(T)$

Example: Let $X_1, ..., X_n \sim Pois(\lambda)$. We know that $\hat\lambda = \sum X_i$ is a sufficient statistic for λ. Let's "Rao-Blackwellize" the unbiased (but bad) estimator $\tilde\lambda = X_1$:

$E(\hat\lambda) = n\lambda$

$$\lambda^* = E[\tilde\lambda \mid \hat\lambda = t] = E[X_1 \mid \Sigma X_i = t]$$

Note: $\Sigma E(X_i \mid \Sigma X_i = t) = E(\Sigma X_i \mid \Sigma X_i = t) = t$

Since X_i's are iid, $E(X_i \mid \Sigma X_i = t)$ have to be equal $= c$

$$\Sigma c = t \rightarrow nc = t \rightarrow c = t/n = E[X_i \mid \Sigma X_i = t]$$
$$= E(X_1 \mid \Sigma X_i = t]$$
$$= \lambda^* = \frac{1}{n}\Sigma X_i = \bar{X}$$

$\hat\lambda$ | $E(\hat\lambda) = n\lambda$

$V(\hat\lambda) = n\lambda$

$MSE(\hat\lambda) = n\lambda + (n\lambda - \lambda)^2$

$\tilde\lambda$ | $E(\tilde\lambda) = \lambda$

$V(\tilde\lambda) = \lambda$

$MSE(\tilde\lambda) = \lambda + (\lambda - \lambda)^2$

$= \lambda$

λ^* | $E(\lambda^*) = \lambda$

$V(\lambda^*) = V(\frac{1}{n}\Sigma X_i)$

$= \frac{V(X_i)}{n}$

$= \frac{\lambda}{n}$

$MSE(\lambda^*) = \frac{\lambda}{n} + (\lambda - \lambda)^2$

$= \frac{\lambda}{n}$

Theorem 7.7.1 Factorization Criterion. Let $X_1, \ldots, X_n$ form a random sample from either a continuous distribution or a discrete distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of θ is unknown and belongs to a given parameter space Ω. A statistic $T = r(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \ldots, X_n$ can be factored as follows for all values of $\boldsymbol{x} = (x_1, \ldots, x_n) \in R^n$ and all values of $\theta \in \Omega$:

$$f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r(\boldsymbol{x}), \theta]. \tag{7.7.1}$$

Here, the functions u and v are nonnegative, the function u may depend on $\boldsymbol{x}$ but does not depend on θ, and the function v will depend on θ but depends on the observed value $\boldsymbol{x}$ only through the value of the statistic $r(\boldsymbol{x})$.

Proof We shall give the proof only when the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a discrete distribution, in which case

$$f_n(\boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta).$$

Suppose first that $f_n(\boldsymbol{x}|\theta)$ can be factored as in Eq. (7.7.1) for all values of $\boldsymbol{x} \in \boldsymbol{R}^n$ and $\theta \in \Omega$. For each possible value t of T, let $A(t)$ denote the set of all points $\boldsymbol{x} \in \boldsymbol{R}^n$ such that $r(\boldsymbol{x}) = t$. For each given value of $\theta \in \Omega$, we shall determine the conditional distribution of $\boldsymbol{X}$ given that $T = t$. For every point $\boldsymbol{x} \in A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x}|\theta)}{\Pr(T = t|\theta)} = \frac{f_n(\boldsymbol{x}|\theta)}{\sum_{\boldsymbol{y} \in A(t)} f_n(\boldsymbol{y}|\theta)}.$$

Since $r(\boldsymbol{y}) = t$ for every point $\boldsymbol{y} \in A(t)$, and since $\boldsymbol{x} \in A(t)$, it follows from Eq. (7.7.1) that

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{u(\boldsymbol{x})}{\sum_{\boldsymbol{y} \in A(t)} u(\boldsymbol{y})}. \tag{7.7.2}$$

Finally, for every point $\boldsymbol{x}$ that does not belong to $A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = 0. \tag{7.7.3}$$

It can be seen from Eqs. (7.7.2) and (7.7.3) that the conditional distribution of $\boldsymbol{X}$ does not depend on θ. Therefore, T is a sufficient statistic.

Conversely, suppose that T is a sufficient statistic. Then, for every given value t of T, every point $\boldsymbol{x} \in A(t)$, and every value of $\theta \in \Omega$, the conditional probability $\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta)$ will not depend on θ and will therefore have the form

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = u(\boldsymbol{x}).$$

If we let $v(t, \theta) = \Pr(T = t|\theta)$, it follows that

$$\begin{aligned} f_n(\boldsymbol{x}|\theta) &= \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) \Pr(T = t|\theta) \\ &= u(\boldsymbol{x})v(t, \theta). \end{aligned}$$

Hence, $f_n(\boldsymbol{x}|\theta)$ has been factored in the form specified in Eq. (7.7.1).

The proof for a random sample $X_1, \ldots, X_n$ from a continuous distribution requires somewhat different methods and will not be given here. ■

One way to read Theorem 7.7.1 is that $T = r(\boldsymbol{X})$ is sufficient if and only if the likelihood function is proportional (as a function of θ) to a function that depends on the data only through $r(\boldsymbol{x})$. That function would be $v[r(\boldsymbol{x}), \theta]$. When using the likelihood function for finding posterior distributions, we saw that any factor not depending on θ (such as $u(\boldsymbol{x})$ in Eq. (7.7.1)) can be removed from the likelihood without affecting