Sufficiency and the Rao-Blackwell Theorem

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

Definition. ***Sufficiency*** *A statistic $T(\underline{X}) = T(x_1, \cdots, T_n)$ is sufficient for an unknown parameter θ if the conditional (joint) distribution of $X_1, \cdots, X_n$ given $T(\underline{X}) = t$ does not depend on θ for any given value of t.*

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

Remark. To prove a statistic is sufficient we compute $P(\underline{X} = \underline{x} | T(\underline{x}) = t)$ and check that it does not depend on θ. A statistic is sufficient with respect to a statistical model and its associated unknown parameter if no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter. Hence $T(\underline{X})$ contains all the information needed to compute any estimate of the parameter. Usually the sufficient statistic is a simple function of the data, i.e. the sum of all the data points.

Example. $T = \displaystyle\sum_{i=1}^{n} X_i$ is a sufficient statistic for $X_1, \cdots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$ We can verify this with

$$\begin{aligned} P(X_1 = x_1, \cdots, X_n = x_n | T = t) &= \frac{P(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P(T = t)} \\ &= \frac{p^t (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \end{aligned}$$

which does not involve θ at all hence T is sufficient

Theorem. ***The Fisher-Neyman Factorization Theorem*** *A statistics $T(\underline{X})$ is sufficient for θ if and only if for any value of θ, there exists nonnegative function g and h such that the joint probability density function can be written in the form*

$$\mathcal{L}(\theta) = f(x_1, \cdots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta) = g(T(x_1, \cdots, x_n), \theta) h(x_1, \cdots, x_n)$$

In other words, the joint density of f can be factored into a product such that one factor, h, does not depend on θ and the other factor, which does depend on θ, depends on x through $T(x)$.

Proof. (=>)

$$\begin{aligned}
L(\theta) &= P(\underline{X} = \underline{x}|\theta) \\
&= P(\underline{X} = \underline{x}, T(\underline{x}) = t|\theta) && (\underline{X} = \underline{x} \text{ implies sufficient statistics } T(\underline{x})) \\
&= P(\underline{X} = \underline{x}|T(\underline{x}) = t, \theta)P(T(\underline{x}) = t|\theta) && \text{(by conditional probability)} \\
&= h(\underline{x}) \cdot g(T(\underline{X}), \theta)
\end{aligned}$$

where $h(\underline{x}) = P(\underline{X} = \underline{x}|T(\underline{x}) = t, \theta) = P(\underline{X} = \underline{x}|T(\underline{x}) = t)$ (b/c $T(\underline{X})$ is sufficient) and $g(T(\underline{X}), \theta) = P(T(\underline{x}) = t|\theta)$, a function of $\underline{x}$ and parameter θ
(<=)
Given the factorization, we prove $T(\underline{X})$ is sufficient, that is

$$\begin{aligned}
P(\underline{X} = \underline{x}|T(\underline{x}) = t) &= \frac{P(\underline{X} = \underline{x})}{P(T(\underline{X}) = t)} && (\underline{X} = \underline{x} \text{ implies any statistics } T(\underline{X})) \\
&= \frac{P(\underline{X} = \underline{x})}{\sum_{T(\underline{x})=t} P(\underline{X} = \underline{x}|\theta)} \\
&= \frac{g(t, \theta)h(\underline{x})}{\sum_{T(\underline{x})=t} g(T(\underline{X}); \theta)h(\underline{X})} && \text{(by factorization of } \mathcal{L}(\theta) = P(\underline{X} = \underline{x}|\theta)) \\
&= \frac{g(t, \theta)h(\underline{x})}{g(t, \theta) \sum_{T(\underline{x})=t} h(\underline{X})} && (g(t, \theta) \text{ is a constant}) \\
&= \begin{cases} \dfrac{h(\underline{x})}{\sum_{T(\underline{x})=t} h(\underline{X})} & \text{T(\underline{x}) = t} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

which does not depend on θ. Hence sufficient □

Remark. This theorem offers a convenient way of identifying sufficient statistic T. For example, in the case of Binomial distribution,

$$\mathcal{L}(p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = g(p, \sum_{i=1}^n X_i) \cdot h(\underline{X})$$

where $h(\underline{X}) = 1$. $\mathcal{L}(\theta)$ can be factorized in into the form specified. Hence $T(\underline{X}) = \sum_{i=1}^n X_i$ is sufficient for p
There are distribution where sufficient statistics does not exist. For example in the case of Cauchy distribution with pdf

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

which cannot be factorized into the form.

Definition. ***Exponential family of distributions*** *provides a convenient characterization of a sufficient statistic. A distribution with cdf / pmf $f(x|\theta)$ is said to belong to a one parameter exponential family of distributions if*

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

where the support, A, does not depend on θ

Remark. Some example of distribution belonging to this family

1. Normal
2. Exponential
3. Gamma
4. Chi-squared
5. Bernoulli
$$f(x|p) = p^x(1-p)^x = exp\left\{x \log \frac{p}{1-p} + \log(1-p)\right\}$$
Hence Bernoulli is an exponential family of distribution with sufficient statistic
$$T = \sum_{i=1}^{n} X_i$$
6. Poisson
$$f(x_1|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = exp\{x_i \log \lambda - \lambda - \log x_i!\} = exp\{c(\lambda)T(x_i) + d(\lambda) + S(x_i)\}$$
Hence Poisson is an exponential family of distribution and hence
$$\sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$
is a sufficient statistic for λ

Note mixed distribution such as student's t, and family of uniform distribution where bounds are not fixed do not belong to exponential family of distribution. Uniform distribution is not exponential family of distribution because its support (a, b) depends on the parameter which violates condition specified above

Proposition. ***Exponential family of distribution is sufficient***

Theorem. ***The Rao-Blackwell Theorem*** *Let $\hat{\theta}$ is an estimator of θ with a finite variance. Suppose that T is sufficient for θ and let the Rao-Blackwell estimator be $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T]$. Then for all θ,*

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

where equality holds if and only if $\hat{\theta}^* = \hat{\theta}$. Also, the improved estimator $\hat{\theta}^*$ and the starting estimator $\hat{\theta}$ have the same bias (by law of total expectation). Hence, in particular, if $\hat{\theta}$ is unbiased, then so is $\hat{\theta}^*$

Proof. By law of total expectation we have

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

hence $\hat{\theta}^*$ and $\hat{\theta}$ have the same bias. So to compare MSE we compare variance. By law of total variance we have

$$Var(\hat{\theta}^*) = Var\left(\mathbb{E}[\hat{\theta}|T]\right) = Var(\hat{\theta}) - \mathbb{E}\left\{Var(\hat{\theta}|T)\right\} \leq Var(\hat{\theta})$$

Hence by $MSE(\hat{\theta}, \theta) = b^2(\hat{\theta}) + Var(\hat{\theta})$ we have $MSE(\hat{\theta}^*, \theta) \leq MSE(\hat{\theta}, \theta)$. Note $Var(\hat{\theta}^*) = Var(\hat{\theta})$ if and only if $Var(\hat{\theta}|T) = 0$, which implies that $\hat{\theta}$ is constant with respect to $\underline{X}$ when T is given

□

Note. The theorem characterizes the transformation of an arbitrarily crube estimator into an estimator that is optimal by the mean squared error criterion. Intuitively, the theorem states that if $g(X)$ is any kind of estimator of paramter θ, then the conditional expectation of $g(X)$ given $T(X)$ where T is a sufficient statistic, is typically a better estimator of θ, and is never worse. In summary, if an estimator is not a function of sufficient statistic, it can be improved by conditioning on the sufficient statistic

Remark. It is tempting to re-apply Rao-Blackwellization to resultant estimator $\hat{\theta}_{RB}$. It turns out that Rao-Blackwellization is an indempotent operation, i.e. using it to improve that already improved estimator does not obtain a further improvement, but merely returns as its output the same improved estimator

$$\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta}_0, T] = g(T)$$

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

Note that $\hat{\theta}_{RB}$ is always a function of sufficient statistic T.

Theorem. ***Lehmann-Scheffe Theorem*** *states that if T is both complete and sufficient and the starting estimator is unbiased, then the Rao-Blackwell estimator is the unique best unbiased estimator.*

Definition.

Conditional Expectation of** X **given the event $Y = y$ *is a function of* y *for* y *in the range of* Y

$$\mathbb{E}[X|Y=y] = \sum_{x\in\mathcal{X}} xP(X=x|Y=y) = \sum_{x\in\mathcal{X}} x\frac{P(X=x,Y=y)}{P(Y=y)}$$

if X, Y *are discrete. Or*

$$\mathbb{E}[X|Y=y] = \int_{\mathcal{X}} xf_X(x,y)dx = \int_{\mathcal{X}} x\frac{f_{X,Y}(x,y)}{P(Y=y)}dx$$

if continuous

Conditional Expectation with respect to a random variable *If* Y *is a discrete random variable with range y then we define on y the function*

$$g : y \mapsto \mathbb{E}(X|Y=y)$$

also called the conditional expectation of X *with respect to* Y*. Hence* $\mathbb{E}[X|Y]$ *is a random variable which assumes* $\mathbb{E}(X|Y=y)$ *at probability* $P(Y=y)$*. In fact*

$$E(X|Y) = g(Y) : \omega \mapsto \mathbb{E}[X|Y=Y(\omega)]$$

Proposition. *Some basic properties of conditional probability*

1. ***Law of total probability***

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

2. *The **conditional variance** is defined by*

$$Var[X|Y] = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y)$$

- ***Algebraic formula for variance***

$$Var[X|Y] = E[X^2|Y] - (\mathbb{E}[X|Y])^2$$

- ***Law of total variance***

$$Var(X) = \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y])$$

Example.

$$\begin{aligned}\mathbb{E}[temp] = \mathbb{E}[temp|month = jan]\frac{31}{365} + \mathbb{E}[temp|month = feb]\frac{28}{361} + \cdots +\\ \mathbb{E}[temp|month = dec]\frac{31}{361}\\ = \sum_i \mathbb{E}[temp|month = i]P(month = i)\\ = \mathbb{E}[\mathbb{E}[temp|month]]\end{aligned}$$

Example. What if instead of $\theta = e^{-\lambda}$ we wanted to take advantage of the RB theorem to improve $\hat{\lambda}_{MLE} = \overline{X}$

Solution. The sufficient estimator for λ is

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

So

$$\hat{\lambda} = \mathbb{E}[\hat{\lambda}_{MLE}|T] = \mathbb{E}[\overline{X}| \sum_{i=1}^{n} X_i] = \overline{X} = \hat{\lambda}_{MLE}$$

Note in the last step $\overline{X}$ is a function of $\displaystyle\sum_{i=1}^{n} X_i$ and therefore RB does not change the estimator. □