Sufficient Statistics

February 21, 2008

For observations $X = (X_1, \ldots, X_n)$ and statistic $T(X)$, the conditional probability

$$\mathbf{f}_{X|T(X)}(\mathbf{x}|t, \theta) = P_\theta\{X_1 = x_1, \ldots, X_n = x_n | T(X) = t\} \tag{1}$$

is, typically, a function of both t *and* θ.

However, consider the case $X = (X_1, \ldots, X_n)$, a sequence of n Bernoulli trials with success probability parameter θ and the statistic $T(X) = X_1 + \cdots + X_n$ the total number of successes. Then

$$\begin{aligned} \mathbf{f}_X(\mathbf{x}|\theta) &= P_\theta\{X_1 = x_1, \ldots, X_n = x_n\} = \theta^{x_1}(1-\theta)^{(1-x_1)} \cdots \theta^{x_n}(1-\theta)^{(1-x_n)} \\ &= \theta^{x_1+\cdots+x_n}(1-\theta)^{(n-(x_1\cdots+x_n))} \end{aligned}$$

and $T(X)$ is a $Bin(n, \theta)$ random variable.

Referring to equation (1), if $\sum_{i=1}^n x_i \neq t$, then the value of the statistic is incompatible with the observations. In this case, equation (1) equals 0. On the other hand, if $\sum_{i=1}^n x_i = t$, then, we have,

$$\mathbf{f}_{X|T(X)}(\mathbf{x}|t, \theta) = \frac{\mathbf{f}_X(\mathbf{x}|\theta)}{f_{T(X)}(t|\theta)} = \frac{P_\theta\{X_1 = x_1, \ldots, X_n = x_n\}}{P_\theta\{T(X) = t\}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1},$$

an answer that does not depend on the parameter θ.

Definition 1. *For observations $X_1, \ldots, X_n$, the statistic T, is called a* **sufficient statistic** *if equation* (1) *is a function of the values, t, of the statistic and does not depend on the value of the parameter θ.*

Thus, by the law of total probability

$$P_\theta\{X_1 = x_1, \ldots, X_n = x_n\} = P\{X_1 = x_1, \ldots, X_n = x_n | T(X) = T(\mathbf{x})\} P_\theta\{T(X) = T(\mathbf{x})\}.$$

and once we know the value of the sufficient statistic, we cannot obtain any additional information about the value of θ from knowing the observed values $X_1 \ldots, X_n$.

How we find sufficient statistics is given by the Neyman-Fisher factorization theorem.

1 Neyman-Fisher Factorization Theorem

Theorem 2. *The statistic T is sufficient for θ if and only if functions g and $\mathbf{h}$ can be found such that*

$$\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x}) g(\theta, T(\mathbf{x})) \tag{2}$$

The central idea in proving this theorem can be found in the case of discrete random variables.

Proof. Because T is a function of $\mathbf{x}$,

$$\mathbf{f}_X(\mathbf{x}|\theta) = f_{X,T(X)}(\mathbf{x}, T(\mathbf{x})|\theta) = \mathbf{f}_{X|T(X)}(\mathbf{x}|T(\mathbf{x}), \theta) f_{T(X)}(T(\mathbf{x})|\theta).$$

If we assume that T is sufficient, then $\mathbf{f}_{X|T(X)}(\mathbf{x}|T(\mathbf{x}), \theta)$ is not a function of θ and we can set it to be $\mathbf{h}(\mathbf{x})$. The second term is a function of $T(\mathbf{x})$ and θ. We will write it $g(\theta, T(\mathbf{x}))$.

If we assume the factorization in equation (3), then, by the definition of conditional expectation,

$$P_\theta\{X = \mathbf{x}|T(X) = t\} = \frac{P_\theta\{X = \mathbf{x}, T(X) = t\}}{P_\theta\{T(X) = t\}}.$$

or,

$$\mathbf{f}_{X|T(X)}(\mathbf{x}|t, \theta) = \frac{\mathbf{f}_{X,T(X)}(\mathbf{x}, t|\theta)}{f_{T(X)}(t|\theta)}.$$

The numerator is 0 if $T(\mathbf{x}) \neq t$ and is

$$\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x}) g(\theta, t)$$

otherwise.

The denominator

$$f_{T(X)}(t|\theta) = \sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{f}_X(\tilde{\mathbf{x}}|\theta) = \sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{h}(\tilde{\mathbf{x}}) g(\theta, t).$$

The ratio

$$\mathbf{f}_{X|T(X)}(\mathbf{x}|t, \theta) = \frac{\mathbf{h}(\mathbf{x}) g(\theta, t)}{\sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{h}(\tilde{\mathbf{x}}) g(\theta, t)} = \frac{\mathbf{h}(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{h}(\tilde{\mathbf{x}})},$$

which is independent of θ and, therefore, T is sufficient. □

2 Maximum Likelihood Estimation

Looking at the likelihood in the case of a sufficient statistic, we have that

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{h}(\mathbf{x}) g(\theta, T(\mathbf{x})).$$

Thus, maximizing the likelihood is equivalent to maximizing $g(\theta, T(\mathbf{x}))$ and the maximum likelihood estimator

$$\hat{\theta}(T(\mathbf{x}))$$

is a function of the sufficient statistic.

3 Unbiased Estimation

We shall learn something about the value of sufficient statistics for unbiased estimators after we review a couple of facts about conditional expectation. Write

$$\phi(u) = E[Y|U = u] = \sum_y y f_{Y|U}(y|u)$$

In words, $\phi(u)$ is the average of Y on the set $\{U = u\}$. Thus, by the law of total probability

$$E\phi(U) = \sum_u E[Y|U=u]f_U(u) = \sum_u \sum_y y f_{Y|U}(y|u) f_Y(u) = \sum_y y \sum_u f(y,u) = \sum_y y f_Y(y) = \mu_Y. \qquad (4)$$

Also,

$$\begin{aligned}\sigma_Y^2 = E[(Y-\mu_Y)^2] &= E[(Y-\phi(U)) + (\phi(U)-\mu_y))^2] \\ &= E[(Y-\phi(U))^2] + 2E[(Y-\phi(U))(\phi(U)-\mu_Y)] + E[(\phi(U)-\mu_Y)^2]\end{aligned}$$

The second term

$$\begin{aligned}E[(Y-\phi(U))(\phi(U)-\mu_Y)] &= \sum_u \sum_y (y-\phi(u))(\phi(u)-\mu_Y) f(y,u) \\ &= \sum_u (\phi(u)-\mu_Y)\left(\sum_y (y-\phi(u)) f_{Y|U}(y|u)\right) f_Y(u) = 0.\end{aligned}$$

The sum in parenthesis is 0 because $\phi(u)$ is the mean of the conditional density function $f_{Y|U}(y|u)$, i.e.,

$$\sum_y (y-\phi(u)) f_{Y|U}(y|u) = \sum_y y f_{Y|U}(y|u) - \phi(u) \sum_y f_{Y|U}(y|u) = \sum_y y f_{Y|U}(y|u) - \phi(u) = 0.$$

Consequently, from equation (3)

$$\sigma_Y^2 = E[(Y-\phi(U))^2] + \sigma^2_{\phi(U)}.$$

and

$$\sigma_Y^2 \geq \sigma^2_{\phi(U)}. \qquad (5)$$

with equality if and only if $Y = \phi(U)$.

If $d(X)$ is an estimator and $T(X)$ is a sufficient statistic, then $E_\theta[d(X)|T(X)]$ does not depend on θ and thus, it is also a statistic. Let's call it $\phi(T(X))$.

By equation (4),

$$g(\theta) = E_\theta d(X) = E_\theta \phi(T(X)).$$

Thus, if $d(X)$ is an unbiased estimator, then so is $\phi(T(X))$. In addition, be equation (5),

$$\mathrm{Var}_\theta(\phi(T(X)) \leq \mathrm{Var}_\theta(d(X)). \qquad (6)$$

with equality if and only if $d(X) = \phi(T(X))$ and the estimator is a function of the sufficient statistic.

Equation (6) is called the **Rao-Blackwell theorem**.

4 Examples

Example 3 (Uniform random variables)**.** *Let $X_1, \cdots, X_n$ be $U(0,\theta)$ random variables. Then, the joint density function*

$$\mathbf{f}(\mathbf{x}|\theta) = \begin{cases} 1/\theta^n & \text{if, for all } i,\ 0 \leq x_i \leq \theta, \\ 0 & \text{otherwise.}\end{cases}$$

If we rewrite this using indicator function notation, then

$$\mathbf{f}_X(\mathbf{x}|\theta) = \frac{1}{\theta^n} I_{[0,\theta]}(\max_{1\le i\le n} x_i).$$

Thus, $T(\mathbf{x}) = \max_{1\le i\le n} x_i$ *is a sufficient statistic with the factorization*

$$\mathbf{h}(\mathbf{x}) = 1 \text{ and } g(\theta,t) = I_{[0,\theta]}(t)/\theta^n.$$

Example 4 (Exponential families)**.** *Recall that an exponential family of random variables has its density of the form*

$$f_X(x|\theta) = c(\theta)h(x)\exp(\nu(\theta)T(x)).$$

Thus by the factorization theorem, T *is a sufficient statistic.* ν *is called the natural parameter.*

Example 5 (Bernoulli observations)**.** *The density is*

$$\mathbf{f}_X(\mathbf{x}|\theta) = \theta^{x_1+\cdots+x_n}(1-\theta)^{(n-(x_1\cdots+x_n))} = (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{(x_1+\cdots+x_n)}$$

Thus, the sufficient statistic is sum of the observations $T(\mathbf{x}) = x_1 + \cdots + x_n$ *and the natural parameter* $\nu(\theta) = \ln(\theta/(1-\theta))$, *the* **log-odds**,

Example 6 (Gamma random variables)**.** *For a multidimensional parameter space, the exponential family is defined with the product in the exponential replaced by the inner product.*

$$f_X(x|\theta) = c(\theta)h(x)\exp\langle\nu(\theta), T(x)\rangle.$$

For a gamma random variable, we have the density,

$$f_X(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x}.$$

Thus for n *independent* $\Gamma(\alpha,\beta)$ *random variables*

$$\begin{aligned}\mathbf{f}_X(\mathbf{x}|\alpha,\beta) &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}(x_1\cdots x_n)^{\alpha-1}\exp(-\beta(x_1+\cdots+x_n)) \\ &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}(x_1\cdots x_n)^{-1}\exp(\alpha(\ln x_1\cdots+\ln x_n) - \beta(x_1+\cdots+x_n)).\end{aligned}$$

Thus, the sufficient statistic

$$T(\mathbf{x}) = (\ln x_1 + \cdots + \ln x_n, x_1 + \cdots + x_n).$$

and the natural parameters

$$\nu(\alpha,\beta) = (\alpha, -\beta).$$