Notes on Sufficient Statistics and the Rao-Blackwell Theorem

• Let $f(x, \theta)$ be the pdf of the random vector $X$ with parameter $\theta$. The statistic $S = \phi(X)$ is a sufficient statistic for $\theta$ if the conditional distribution of $X$ given $S = s$ does not depend on $\theta$ for all values of $s$.

• A sufficient statistic is best understood as a partition of the sample space. When the dimension of $S$ is smaller than the dimension of $X$, then many $x$ will lead to the same $s = \phi(x)$. A sufficient statistic, just like any statistic, can be thought of inducing a partition (that is, a collection of mutually disjoint sets) of the possible realizations of $X$, where those realizations $x$ that lead to the same $s = \phi(x)$ are lumped together in one set. Now what is nice about the partition induced by the sufficient statistic is that knowing to which of those sets $X$ belongs contains already all information about $\theta$—it is a "sufficient" description of $X$ for all purposes that relate to questions about $\theta$.

• Example: Suppose $X = (X_1, X_2, X_3)'$ is a sample of iid Bernoulli random variables with parameter $p$. Let $S = \sum_{i=1}^3 X_i$, and denote by $\{X|S = s\}$ the set of possible realizations of $X$ if $S = s$. Then the set of all possible realizations of $X$ is partitioned by $S$ in the following four subsets:

$$\{\{X|S = 0\}, \{X|S = 1\}, \{X|S = 2\}, \{X|S = 3\}\}$$

$$= \{(0, 0, 0)', (0, 0, 1)', (0, 1, 0)', (0, 1, 1)', (1, 0, 0)', (1, 0, 1)', (1, 1, 0)', (1, 1, 1)\}$$

Now $S = \sum_{i=1}^3 X_i$ is a sufficient statistic, because the probability of $X$ being exactly equal to any one member of the four sets $\{X|S = 0\}, \{X|S = 1\}, \{X|S = 2\}, \{X|S = 3\}$ is independent of $p$: It is 1 if $S = 0$ or $S = 1$ (since there is only a single realization in these two sets) and $1/3$ if $S = 1$ or $S = 2$. But neither 1 nor $1/3$ depends on $p$, so $S$ satisfies the definition of a sufficient statistic.
Factorization Theorem: Let \( f(x, \theta) \) be the pdf of the random vector \( X \) with parameter \( \theta \). A statistic \( S = \phi(X) \) is a sufficient statistic for \( \theta \) if and only if \( f(x, \theta) \) factors as
\[
    f(x, \theta) = h(x)g(\phi(x), \theta) = h(x)g(s, \theta)
\]
where the function \( h(\cdot) \) does not depend on \( \theta \) and both \( h(\cdot) \) and \( g(\cdot, \cdot) \) are nonnegative.

Proof of "if" in the discrete case: Let \( q(s, \theta) \) be the pdf of \( S \). Then
\[
    q(s, \theta) = P(S = s) = \sum_{u \in Q(s)} f(u, \theta),
\]
where \( Q(s) = \{u : \phi(u) = s\} \) are the partitions induced by \( S \).

We find
\[
    P(X = x | S = s) = \frac{f(x, \theta)}{\sum_{u \in Q(\phi(x))} f(u, \theta)} = \frac{h(x)g(\phi(x), \theta)}{\sum_{u \in Q(\phi(x))} h(u)g(\phi(u), \theta)} = \frac{h(x)}{g(\phi(x), \theta) \sum_{u \in Q(\phi(x))} h(u)} = \frac{h(x)}{g(\phi(x), \theta) \sum_{u \in Q(\phi(x))} h(u)}
\]
so that \( P(X = x | S = s) \) does not depend on \( \theta \), as was to be shown.

Example: Let \( X = (X_1, \cdots, X_n)' \) with \( X_i \sim iidU[0, \theta] \). Then
\[
    f_X(x) = \prod_{i=1}^n (\theta^{-1} 1[0 \leq x_i \leq \theta]) = (1[\min x_i \geq 0])(1[\max x_i \leq \theta] \theta^{-n})
\]
so that by the Factorization Theorem, \( \max X_i \) is sufficient for \( \theta \).

Rao-Blackwell Theorem: Let \( X \) be a random vector with pdf \( f(x, \theta) \) and sufficient statistic \( S = \phi(X) \) for \( \theta \). The minimum variance unbiased estimator of \( \theta \), if it exists, depends on \( X \) only through \( S \).

Proof: Take any unbiased estimator \( g(X) \) of \( \theta \), i.e. \( E[g(X)] = \theta \). Let \( h(S) = E[g(X) | S] \). By the Law of Iterated Expectations,
\[
    E[h(S)] = E[E[g(X) | S]] = \theta
\]
$E[g(X)] = \theta$, so that $h(S)$ is also an unbiased estimator. (It is an estimator, because $E[g(X)|S = s]$ does not depend on $\theta$, so is a function of the sample only). But

$$Var[g(X)] = E[Var[g(X)|S]] + Var[E[g(X)|S]]$$

so that $Var[h(S)] \leq Var[g(X)]$.

- The Rao-Blackwell Theorem is constructive in the sense that it says how a given unbiased estimator can be improved upon. But this improvement is often hard to do, because one would need to figure out the conditional distribution of $g(X)|S = s$, and then carry out the integration. Rather, the Theorem is useful, because it tells us that we can restrict attention to functions of a sufficient statistic in a search of a minimum variance unbiased estimator.

- Sufficient statistics can be more than one dimensional, and the Factorization Theorem goes through with $S$ and $\theta$ of any (finite) dimension. Example: Let $X = (X_1, \cdots, X_n)'$, where $X_j \sim iid \mathcal{N}(\mu, \sigma^2)$. Then, with $\theta = (\mu, \sigma^2)'$,

$$f(x, \theta) = (2\pi \sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2] = (2\pi \sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2 - \frac{1}{2} \frac{1}{\sigma^2} n(\bar{x} - \mu)^2]$$

so that $(\bar{x}, \sum (x_i - \bar{x})^2)$ is a two-dimensional sufficient statistic for $\theta$.

- In fact, one can always add another statistic to a given sufficient statistic to obtain a higher dimensional sufficient statistic: In the Binomial example, also the two-dimensional statistic $S_2 = (\sum_{i=1}^3 X_i, X_2) = (S, X_2)$ is a sufficient statistic (check!). This is undesirable, because this limits the dimensionality reduction obtained by sufficiency. A statistic $S = \phi(X)$ is called minimal sufficient if it induces the coarsest possible partition of the sample space.