# Lecture 5

## 9. Minimal sufficient and complete statistics

We introduced the notion of sufficient statistics in order to have a function of the data that contains all information about the parameter. However, a sufficient statistic does not have to be any simpler than the data itself. As we have seen, the identity function is a sufficient statistic so this choice does not simplify or summarize anything. A statistic is said to be minimal sufficient if it is as simple as possible in a certain sense. Here is a definition.

**Definition 11.** A sufficient statistic $T : \mathcal{X} \to \mathcal{T}$ is minimal sufficient if for any sufficient statistic $U : \mathcal{X} \to \mathcal{U}$ there is a measurable function $g : \mathcal{U} \to \mathcal{T}$ such that $T = g(U)$ $\mu_{X|\Theta}(\cdot \mid \theta)$-a.s. for all $\theta \in \Omega$.

How do we check if a statistic $T$ is minimal sufficient? It can be inconvenient to check the condition in the definition for all sufficient statistics $U$.

**Theorem 10.** *If there exist version of $f_{X|\Theta}(x \mid \theta)$ for each $\theta$ and a measurable function $T : \mathcal{X} \to \mathcal{T}$ such that $T(x) = T(y) \Leftrightarrow y \in \mathcal{D}(x)$, where*

$$\mathcal{D}(x) = \{y \in \mathcal{X} : f_{X|\Theta}(y \mid \theta) = f_{X|\Theta}(x \mid \theta)h(x, y), \forall\theta \text{ and some function } h(x, y) > 0\},$$

*then $T$ is a minimal sufficient statistic.*

**Example 14.** Let $\{X_n\}$ be IID Exp($\theta$) given $\Theta = \theta$ and $X = (X_1, \dots, X_n)$. Put $T(x) = x_1 + \cdots + x_n$. Let us show $T$ is minimal sufficient. The ratio

$$\frac{f_{X|\Theta}(x \mid \theta)}{f_{X|\Theta}(y \mid \theta)} = \frac{\theta^n e^{-\theta \sum_{i=1}^n x_i}}{\theta^n e^{-\theta \sum_{i=1}^n y_i}}$$

does not depend on $\theta$ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. In this case $h(x, y) = 1$, $\mathcal{D}(x) = \{y : \sum_{i=1}^n x_i = \sum_{i=1}^n y_i\}$, and $T$ is minimal sufficient.

*Proof.* Note first that the sets $\mathcal{D}(x)$ form a partition of $\mathcal{X}$. Indeed, by putting $h(y, x) = 1/h(x, y)$ we see that $y \in \mathcal{D}(x)$ implies $x \in \mathcal{D}(y)$. Similarly, taking $h(x, x) = 1$, we see that $x \in \mathcal{D}(x)$ and hence, the different $\mathcal{D}(x)$ form a partition. The condition says that the sets $\mathcal{D}(x)$ coincide with sets $T^{-1}\{T(x)\}$ and hence $\mathcal{D}(x) \in \mathcal{B}_T$ for each $x$. By Bayes theorem we have, for $y \in \mathcal{D}(x)$,

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta)\mu_\Theta(d\theta)} = \frac{h(x, y)f_{X|\Theta}(y \mid \theta)}{\int_\Omega h(x, y)f_{X|\Theta}(y \mid \theta)\mu_\Theta(d\theta)} = \frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta \mid y).$$

That is, the posterior density is constant on $\mathcal{D}(x)$. Hence, it is a function of $T(x)$ and by Lemma 1 $T$ is sufficient.

Let us check that $T$ is also minimal. Take $U : \mathcal{X} \to \mathcal{U}$ to be a sufficient statistic. If we show that $U(x) = U(y)$ implies $y \in \mathcal{D}(x)$, then it follows that $U(x) = U(y)$ implies $T(x) = T(y)$ and hence that $T$ is a function of $U(x)$. Then $T$ is minimal. By the factorization theorem (Theorem 2, Lecture 6)

$$f_{X|\Theta}(x \mid \theta) = h(x)g(\theta, U(x)).$$

We can assume that $h(x) > 0$ because $P_\theta(\{x : h(x) = 0\}) = 0$. Hence, $U(x) = U(y)$ implies

$$f_{X|\Theta}(y \mid \theta) = \frac{h(y)}{h(x)}g(\theta, U(x)).$$

That is, $y \in \mathcal{D}(x)$ with $h(x, y) = h(y)/h(x)$. $\square$

The next concept is that of a complete statistic.

**Definition 12.** Let $T : \mathcal{X} \to \mathcal{T}$ be a statistic and $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ the family of conditional distributions of $T(X)$ given $\Theta = \theta$. The family $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ is said to be complete if for each measurable function $g$, $E_\theta[g(T)] = 0, \forall\theta$ implies $P_\theta(g(T) = 0) = 1, \forall\theta$.

The family $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ is said to be boundedly complete if each bounded measurable function $g$, $E_\theta[g(T)] = 0, \forall\theta$ implies $P_\theta(g(T) = 0) = 1, \forall\theta$.

A statistic $T$ is said to be complete if the family $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ is complete.

A statistic $T$ is said to be boundedly complete if the family $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ is boundedly complete.

One should note that completeness is a statement about the entire family $\{\mu_{T|\Theta}(\cdot \mid \theta), \theta \in \Omega\}$ and not only about the individual conditional distributions $\mu_{T|\Theta}(\cdot \mid \theta)$.

**Example 15.** Suppose that $T$ has $\mathrm{Bin}(n, \theta)$ distribution with $\theta \in (0, 1)$ and $g$ is a function such that $E_\theta[g(T)] = 0 \ \forall\theta$. Then

$$0 = E_\theta[g(T)] = \sum_{k=0}^{n} g(k)\binom{n}{k}\theta^k(1-\theta)^{n-k} = (1-\theta)^n \sum_{k=0}^{n} g(k)\binom{n}{k}\Big(\frac{\theta}{1-\theta}\Big)^k.$$

If we put $r = \theta/(1 - \theta)$ we see that this equals

$$(1-\theta)^n \sum_{k=0}^{n} g(k)\binom{n}{k} r^k$$

which is a polynomial in $r$ of degree $n$. Since this is constant equal to 0 for all $r > 0$ it must be that $g(k)\binom{n}{k} = 0$ for each $k = 0, \dots, n$, i.e. $g(k) = 0$ for each $k = 0, \dots, n$. Since, for each $\theta$, $T$ is supported on $\{0, \dots, n\}$ it follows that $P_\theta(g(T) = 0) = 1 \ \forall\theta$ so $T$ is complete.

An important result for exponential families is the following.

**Theorem 11.** *If the natural parameter space $\Omega$ of an exponential family contains an open set in $\mathbb{R}^k$, then $T(X)$ is a complete sufficient statistic.*

*Proof.* We will give a proof for $k = 1$. For larger $k$ one can use induction. We know that the natural statistic $T$ has a density $c(\theta)e^{\theta t}$ with respect to $\nu'_T$ (see Section 4.2, Lecture 4). Let $g$ be a measurable function such that $E_\theta[g(T)] = 0$ for all $\theta$. That is,

$$\int_{\mathcal{T}} g(t)c(\theta)e^{\theta t}\nu_T(dt) = 0 \quad \forall\theta.$$

If we write $g^+$ and $g^-$ for the positive and negative part of $g$, respectively, then this says

$$\int_{\mathcal{T}} g^+(t)c(\theta)e^{\theta t}\nu_T(dt) = \int_{\mathcal{T}} g^-(t)c(\theta)e^{\theta t}\nu_T(dt) \quad \forall\theta. \tag{9.1}$$

Take a fixed value $\theta_0$ in the interior of $\Omega$. This is possible since $\Omega$ contains an open set. Put

$$Z_0 = \int_{\mathcal{T}} g^+(t)c(\theta_0)e^{\theta_0 t}\nu_T(dt) = \int_{\mathcal{T}} g^-(t)c(\theta_0)e^{\theta_0 t}\nu_T(dt)$$

and define the probability measures $P$ and $Q$ by

$$P(C) = Z_0^{-1} \int_C g^+(t) c(\theta_0) e^{\theta_0 t} \nu_T(dt)$$
$$Q(C) = Z_0^{-1} \int_C g^-(t) c(\theta_0) e^{\theta_0 t} \nu_T(dt).$$

Then, the equality (9.1) can be written

$$\int_{\mathcal{T}} \exp\{t(\theta - \theta_0)\} P(dt) = \int_{\mathcal{T}} \exp\{t(\theta - \theta_0\} Q(dt), \quad \forall \theta.$$

With $u = \theta - \theta_0$ we see that this implies that the moment generating function of $P$, $M_P(u)$, equals the mgf of $Q$, $M_Q(u)$ in a neighborhood of $u = 0$. Hence, by uniqueness of the moment generating function $P = Q$. It follows that $g^+(t) = g^-(t)$ $\nu_T'$-a.e. and hence that $\mu_{T|\Theta}\{t : g(t) = 0 \mid \theta\} = 1$ for all $\theta$. Hence, $T$ is complete sufficient statistic. □

Completeness of a statistic is also related to minimal sufficiency.

**Theorem 12** (Bahadur's theorem)**.** *If $T$ is a finite-dimensional boundedly complete sufficient statistic, then it is minimal sufficient.*

*Proof.* Let $U$ be an arbitrary sufficient statistic. We will show that $T$ is a function of $U$ by constructing the appropriate function. Put $T = (T_1(X), \dots, T_k(X))$ and $S_i(T) = [1 + e^{-T_i}]^{-1}$ so that $S_i$ is bounded and bijective. Let

$$X_i(u) = E_\theta[S_i(T) \mid U = u],$$
$$Y_i(t) = E_\theta[X_i(U) \mid T = t].$$

We want to show that $S_i(T) = X_i(U)$ $P_\theta$-a.s. for all $\theta$. Then, since $S_i$ is bijective we have $T_i = S_i^{-1}(X_i(U))$ and the claim follows. We show $S_i(T) = X_i(U)$ $P_\theta$-a.s. in two steps.

First step: $S_i(T) = Y_i(T)$ $P_\theta$-a.s. for all $\theta$. To see this note that

$$E_\theta[Y_i(T)] = E_\theta[E_\theta[X_i(U) \mid T]] = E_\theta[X_i(U)] = E_\theta[E_\theta[S_i(T) \mid U]] = E_\theta[S_i(T)].$$

Hence, for all $\theta$, $E_\theta[Y_i(T) - S_i(T)] = 0$ and since $S_i$ is bounded, so is $Y_i$ and bounded completeness implies $P_\theta(S_i(T) = Y_i(T)) = 1$ for all $\theta$.

Second step: $X_i(U) = Y_i(T)$ $P_\theta$-a.s. for all $\theta$. By step one we have $E_\theta[Y_i(T) \mid U] = X_i(U)$ $P_\theta$-a.s. So if we show that the conditional variance of $Y_i(T)$ given $U$ is zero we are done. That is, we need to show $\mathrm{Var}_\theta(Y_i(T) \mid U) = 0$ $P_\theta$-a.s. By the usual rule for conditional variance (Theorem B.78 p. 634)

$$\begin{aligned}\mathrm{Var}_\theta(Y_i(T)) &= E_\theta[\mathrm{Var}_\theta(Y_i(T) \mid U)] + \mathrm{Var}_\theta(X_i(U)) \\ &= E_\theta[\mathrm{Var}_\theta(Y_i(T) \mid U)] + E_\theta[\mathrm{Var}_\theta(X_i(U) \mid T)] + \mathrm{Var}_\theta(S_i(T)).\end{aligned}$$

By step one $\mathrm{Var}_\theta(Y_i(T)) = \mathrm{Var}_\theta(S_i(T))$ and $E_\theta[\mathrm{Var}_\theta(X_i(U) \mid T)] = 0$ since $X_i(U)$ is known if $T$ is known. Combining this we see that $\mathrm{Var}_\theta(Y_i(T) \mid U) = 0$ $P_\theta$-a.s. as we wanted. □

## 10. Ancillary statistics

As we have seen a sufficient statistic contains all the information about the parameter. The opposite is when a statistic does not contain any information about the parameter.

**Definition 13.** A statistic $U : \mathcal{X} \to \mathcal{U}$ is called ancillary if the conditional distribution of $U$ given $\Theta = \theta$ is the same for all $\theta$.

**Example 16.** Let $X_1$ and $X_2$ be conditionally independent $N(\theta, 1)$ distributed given $\Theta = \theta$. Then $U = X_2 - X_1$ is ancillary. Indeed, $U$ has $N(0, 2)$ distribution, which does not depend on $\theta$.

Sometimes a statistic contains a coordinate that is ancillary.

**Definition 14.** If $T = (T_1, T_2)$ is a sufficient statistic and $T_2$ is ancillary, then $T_1$ is called conditionally sufficient given $T_2$.

**Example 17.** Let $X = (X_1, \dots, X_n)$ be conditionally IID $U(\theta - 1/2, \theta + 1/2)$ given $\Theta = \theta$. Then

$$f_{X|\Theta}(x \mid \theta) = \prod_{i=1}^{n} I_{[\theta-1/2,\theta+1/2]}(x_i) = I_{[\theta-1/2,\infty)}(\min x_i) I_{(-\infty,\theta+1/2]}(\max x_i).$$

$T = (T_1, T_2) = (\max X_i, \max X_i - \min X_i)$ is minimal sufficient and $T_2$ is ancillary. Note that $f_{X|\theta}(y \mid \theta) = f_{X|\theta}(x \mid \theta) \Leftrightarrow \max x_i = \max y_i$ and $\min x_i = \min y_i$ $\Leftrightarrow T(x) = T(y)$. Hence, by Theorem 10 Lecture 7, $T$ is minimal sufficient. The conditional density of $(T_1, T_2)$ given $\Theta = \theta$ can be computed as (do this as an exercise)

$$f_{T_1,T_2|\Theta}(t_1, t_2 \mid \theta) = n(n-1)t_2^{n-2} I_{[0,1]}(t_2) I_{[\theta-1/2+t_2,\theta+1/2]}(t_1)$$

In particular, the marginal density of $T_2$ is

$$f_{T_2|\Theta}(t_2 \mid \theta) = n(n-1)t_2^{n-2}(1 - t_2)$$

and this does not depend on $\theta$. Hence $T_2$ is ancillary.

Note that the conditional distribution of $T_1$ given $T_2 = t_2$ and $\Theta = \theta$ is

$$f_{T_1|T_2,\Theta}(t_1 \mid t_2, \theta) = \frac{1}{(1 - t_2)} I_{[\theta-1/2+t_2,\theta+1/2]}(t_1).$$

That is, it is $U(\theta - 1/2 + t_2, \theta + 1/2)$. Hence, this distribution becomes more concentrated as $t_2$ becomes large. Although $T_2$ does not tell us something about the parameter, it tells us something about the conditional distribution of $T_1$ given $\Theta$.

The usual “rule” in classical statistics is to (whenever it is possible) perform inference conditional on an ancillary statistic.

In our example we can exemplify it.

**Example 18** (continued)**.** Consider the above example with $n = 2$ and consider finding a 50% confidence interval for $\Theta$. The naive way to do it is to consider the interval $I_1 = [\min X_i, \max X_i] = [T_1 - T_2, T_1]$. This interval satisfies $P_\theta(\Theta \in I_1) = 1/2$ since there is probability $1/4$ that both observations are above $\theta$ and probability $1/4$ that both are below $\theta$.

If one performs the inference conditional on the ancillary $T_2$ we get a very different result. We can compute

$$\begin{aligned} P_\theta(T_1 - T_2 \le \Theta \le T_1 \mid T_2) &= P_\theta(\Theta \le T_1 \le \Theta + T_2 \mid T_2 = t_2) \\ &= \frac{1}{1-t_2}\int_\theta^{\theta+t_2} I_{[\theta-1/2+t_2,\theta+1/2]}(t_1)dt_1 \\ &= \frac{t_2}{1-t_2} I_{[0,1/2)}(t_2). \end{aligned}$$

Hence, the level of confidence depends on $t_2$. In particular, we can construct an interval $I_2 = [T_1 - 1/4(1+T_2), T_1 + 1/4 - 3T_2/4]$ which has the property

$$P_\theta(\Theta \in I_2 \mid T_2 = t_2) = 1/2.$$

Indeed,

$$\begin{aligned} P_\theta(\Theta \in I_2 \mid T_2 = t_2) &= P_\theta(\Theta - 1/4 + 3T_2/4 \le T_1 \le \Theta + 1/4(1+T_2) \mid T_2 = t_2) \\ &= \int_{\theta-1/4+3t_2/4}^{\theta+(1+t_2)/4} I_{[\theta-1/2+t_2,\theta+1/2]}(t_1)dt_1 = 1/2. \end{aligned}$$

Since this probability does not depend on $t_2$ it follows that

$$P_\theta(\Theta \in I_2) = 1/2.$$

Let us compare the properties of $I_1$ and $I_2$. Suppose we observe $T_2$ small. This does not give us much information about $\Theta$ and this is reflected in $I_2$ being wide. On the contrary, $I_1$ is very small which is counterintuitive. Similarly, if we observe $T_2$ large, then we know more about $\Theta$ and $I_2$ is short. However, this time $I_1$ is wide!

Suppose $T$ is sufficient and $U$ is ancillary and they are conditionally independent given $\Theta = \theta$. Then there is no benefit of conditioning on $U$. Indeed, in this case

$$f_{T|U,\Theta}(t \mid u, \theta) = f_{T|\Theta}(t \mid \theta)$$

so conditioning on $U$ does not change anything. This situation appear when there is a boundedly complete sufficient statistic.

**Theorem 13** (Basu's theorem)**.** *If $T$ is boundedly complete sufficient statistic and $U$ is ancillary, then $T$ and $U$ are conditionally independent given $\Theta = \theta$. Furthermore, for every prior $\mu_\Theta$ they are independent (unconditionally).*

*Proof.* For the first claim (to show conditional independence) we want to show that for each measurable set $A \subset \mathcal{U}$

$$\mu_{U|\Theta}(A) = \mu_{U|T,\Theta}(A \mid t, \theta) \qquad \mu_{T|\Theta}(\cdot \mid \theta) - a.e.\ t,\ \forall \theta. \tag{10.1}$$

Since $U$ is ancillary $\mu_{U|\Theta}(A \mid \theta) = \mu_U(A)$, $\forall \theta$. We also have

$$\mu_{U|\Theta}(A \mid \theta) = \int_{\mathcal{T}} \mu_{U|T,\Theta}(A \mid t, \theta)\mu_{T|\Theta}(dt \mid \theta) = \int_{\mathcal{T}} \mu_{U|T}(A \mid t)\mu_{T|\Theta}(dt \mid \theta),$$

where the second equality follows since $T$ is sufficient. Indeed, $\mu_{X|T,\Theta}(B \mid t, \theta) = \mu_{X|T}(B \mid t)$ and since $U = U(X)$

$$\mu_{U|T,\Theta}(A \mid t, \theta) = \mu_{X|T,\Theta}(U^{-1}A \mid t, \theta) = \mu_{X|T}(U^{-1}A \mid t) = \mu_{U|T}(A \mid t).$$

Combining these two we get

$$\int_{\mathcal{T}} [\mu_U(A) - \mu_{U|T}(A \mid t)]\mu_{T|\Theta}(dt \mid \theta) = 0.$$

By considering the integrand as a function $g(t)$ we see that the above equation is the same as $E_\theta[g(T)] = 0$ for each $\theta$ and since $T$ is boundedly complete $\mu_{T|\Theta}(\{t : g(t) = 0\} \mid \theta) = 1$ for all $\theta$. That is (10.1) holds.

For the second claim we have by conditional independence that

$$\begin{aligned}
\mu_{U,T}(A \times B) &= \int_\Omega \int_B \mu_{U|T}(A \mid t)\mu_{T|\Theta}(dt \mid \theta)\mu_\Theta(d\theta) \\
&= \int_\Omega \mu_U(A)\mu_{T|\Theta}(B \mid \theta)\mu_\Theta(d\theta) \\
&= \mu_U(A)\mu_T(B)
\end{aligned}$$

so $T$ and $U$ are independent. □

Sometimes a combination of the recent results are useful for computing expected values in an unusual way:

**Example 19.** Let $X = (X_1, \dots, X_n)$ be conditionally IID Exp($\theta$) given $\Theta = \theta$. Consider computing the expected value of

$$g(X) = \frac{X_n}{X_1 + \cdots + X_n}.$$

To do this, note that $g(X)$ is an ancillary statistic. Indeed, if $Z = (Z_1, \dots, Z_n)$ are IID Exp(1) then $X \stackrel{d}{=} \theta^{-1}Z$ and we see that

$$\begin{aligned}
P_\theta(g(X) \le x) &= P_\theta\Big(\frac{1}{x} < \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1\Big) \\
&= P_\theta\Big(\frac{1}{x} < \frac{Z_1}{Z_n} + \cdots + \frac{Z_{n-1}}{Z_n} + 1\Big)
\end{aligned}$$

Since the distribution of $Z$ does not depend on $\theta$ we see that $g(X)$ is ancillary. The natural statistic $T(X) = X_1 + \cdots + X_n$ is complete (by the Theorem just proved) and minimal sufficient. By Basu's theorem (Theorem 13) $T(X)$ and $g(X)$ are independent. Hence,

$$\theta = E_\theta[X_n] = E_\theta[T(X)g(X)] = E_\theta[T(X)]E_\theta[g(X)] = n\theta E_\theta[g(X)]$$

and we see that $E_\theta[g(X)] = n^{-1}$.