Lecture 4

7. Sufficient statistics

Consider the "usual" statistical setup: the data is X and the paramter is Θ.

To gain information about the parameter we study various functions of the data X. For instance, if $X = (X_1, \dots, X_n)$ are IID Ber(θ) given $\Theta = \theta$, then we would use $T(X) = n^{-1}(X_1 + \cdots + X_n)$ to get information about the parameter. A function of the data is called a *statistic*.

Definition 7. Let $(\mathcal{T}, \mathcal{C})$ be a measurable space such that the σ-field $\mathcal{C}$ contains all singletons. A measurable mapping $T : \mathcal{X} \to \mathcal{T}$ is called a statistic.

As usual we think of a measurable space as a subspace of $\mathbb{R}^d$ and the σ-field as the corresponding sub-σ-field.

Although, formally a statistic is a mapping from the sample space $\mathcal{X}$ to some space $\mathcal{T}$, we can also think of the composition $T \circ X : S \to \mathcal{T}$ (recall that S is the underlying probability space). This is a random variable taking values in $\mathcal{T}$ and we often write T for this random quantity.

In the next sections we will look more closely at different classes of statistics. That is, functions of the data with certain interesting properties. The first such class is the class of *sufficient statistics*.

7.1. Sufficient statistics (classical).

The idea of sufficiency is to find a function T of the data X that summarizes the information about the parameter Θ. Above we mentioned the example of IID Ber(θ) random variables, $X_1, \dots, X_n$, where we know that we only need to know a function of the data, for instance $X_1 + \cdots + X_n$, in order to compute an estimate of θ. Similarly, we argued for the betting problem that decisions can be based entirely of knowing $X_1 + \cdots + X_n$ and not all the individual X_i's.

Elementary case: Let us first see what sufficient statistics is when we have densities. Suppose that the (conditional) distribution of X and $T = T(X)$ given $\Theta = \theta$ both have densities w.r.t. a measure ν (think Lebesgue measure or counting measure). Then we say that T is a sufficient statistic for Θ if $f_{X|T,\Theta}(x \mid t, \theta)$ does not depend on θ.

Note that, with $t = T(x)$,

$$f_{X|T,\Theta}(x \mid t, \theta) = \frac{f_{X,T|\Theta}(x, t \mid \theta)}{f_{T|\Theta}(t \mid \theta)} = \frac{f_{X|\Theta}(x \mid \theta)}{f_{T|\Theta}(t \mid \theta)}.$$

Hence T is sufficient if this ratio does not depend on θ.

To see how T captures the "information" about Θ we can write down the likelihood function as

$$f_{X|\Theta}(x \mid \theta) = f_{X|T,\Theta}(x \mid T(x), \theta) f_{T|\Theta}(T(x) \mid \theta)$$

If T is sufficient, then the first factor on the RHS does not depend on θ and the likelihood when observing $X = x$ is proportional (as a function of θ) to the likelihood when observing $T = T(x)$. That is, information about Θ comes only through the function T. If we, for example, want to maximize the likelihood we could maximize $f_{T|\Theta}(t \mid \theta)$ instead of maximizing $f_{X|\Theta}(x \mid \theta)$. In this sense, there is no need to know x itself, it is sufficient to know $t = T(x)$ to do inference.

General case: Formally, sufficient statistics are introduced as follows. First let $\mu_{T|\Theta}(\cdot \mid \theta)$ be the conditional distribution of T given $\Theta = \theta$. It is a probability measure on $\mathcal{C}$ given by $\mu_{T|\Theta}(C \mid \theta) = \mu_{X|\Theta}(T^{-1}C \mid \theta)$.

Definition 8. Suppose there exist versions of $\mu_{X|\Theta,T}(\cdot \mid \theta, t)$ and a function $r : \mathcal{B} \times \mathcal{T} \to [0,1]$ such that
(i) $r(\cdot, t)$ is a probability on $\mathcal{B}$ for each $t \in \mathcal{T}$,
(ii) $r(B, \cdot)$ is measurable $\mathcal{T}$ for each $B \in \mathcal{B}$,
and for each $\theta \in \Omega$ and $B \in \mathcal{B}$

$$\mu_{X|\Theta,T}(B \mid \theta, t) = r(B, t), \quad \text{for } \mu_{T|\Theta}(\cdot \mid \theta) - a.e.\ t.$$

Then T is called a *sufficient statistic for* Θ *(in the classical sense).*

Note that the function r satisfies the conditions of a conditional distribution and does not depend on θ. Hence, T is sufficient if $\mu_{X|\Theta,T}(\cdot \mid \theta, t)$ is a conditional distribution that does not depend on θ.

The simplest example (but not particularly useful) of a sufficient statistic is the data itself. That is $\mathcal{T} = \mathcal{X}$ and $T(x) = x$. Of course, $\mu_{X|\Theta,X}(B \mid \theta, x) = I_B(x)$ does not depend on θ so the statistic is sufficient. Using this statistic does not help you to summarize information about the parameter as it is as complicated as the data itself. Let's look at some simple cases where there exist simple sufficient statistics.

Example 13 (c.f. Example 6.2.3 in Casella & Berger)**.** Let $\{X_n\}$ be IID Ber(θ) given $\Theta = \theta$ and $X = (X_1, \ldots, X_n)$. Put $T(x) = x_1 + \cdots + x_n$. Let us show T is sufficient. Note that $T(X)$ is Bin(n, θ) given $\Theta = \theta$. For each $x = (x_1, \ldots, x_n)$ $x_i \in \{0, 1\}$ such that $t = T(x)$

$$f_{X|\Theta,T}(x \mid \theta, t) = \frac{f_{X,T|\Theta}(x, t)}{f_{T|\Theta}(t \mid \theta)} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.$$

Since this does not depend on θ, T is a sufficient statistic.

Example 14. Let $\{X_n\}$ be IID Exp(θ) given $\Theta = \theta$ and $X = (X_1, \ldots, X_n)$. Put $T(x) = x_1 + \cdots + x_n$. Let us show T is sufficient. Note that $T(X)$ is $\Gamma(n, \theta)$ given $\Theta = \theta$. For each $x_i > 0$ we have with $t = T(x)$

$$\frac{f_{X|\Theta}(x \mid \theta)}{f_{T|\Theta}(t \mid \theta)} = \frac{\prod_{i=1}^n \theta e^{-\theta x_i}}{\frac{\theta^n}{\Gamma(n)} t^{n-1} e^{-\theta t}} = \frac{(n-1)!}{t^{n-1}},$$

which does not depend on θ.

7.2. **Sufficient statistics (Bayesian).** In Bayesian statistics there is a slightly different notion of sufficiency, but it often coincides with the classical notion.

Definition 9. A statistic T is called a sufficient statistic for the parameter Θ (in the Bayesian sense) if, for every prior μ_Θ, there exist versions of the posterior distributions $\mu_{\Theta|X}$ and $\mu_{\Theta|T}$ such that, for every $A \in \tau$, $\mu_{\Theta|X}(B \mid x) = \mu_{\Theta|T}(B \mid T(x))$, μ_X-a.s., where μ_X is the marginal distribution of X.

Hence, no matter what prior one uses, one only has to consider the sufficient statistic for making inference, becuase the posterior distribution given $T = T(x)$ is the same as the posterior given the data $X = x$.

Let's see how this looks like with densities. If both $\mu_{\Theta|X}(\cdot \mid x)$ and $\mu_{\Theta|T}(\cdot \mid t)$ have densities w.r.t. the prior μ_Θ then

$$\mu_{\Theta|X}(A \mid x) = \int_A f_{\Theta|X}(\theta \mid x)\mu_\Theta(d\theta),$$
$$\mu_{\Theta|T}(A \mid t) = \int_A f_{\Theta|T}(\theta \mid t)\mu_\Theta(d\theta),$$

holds for any $A \in \tau$ and hence T is sufficient if and only if $f_{\Theta|X}(\theta \mid x) = f_{\Theta|T}(\theta \mid T(x))$ μ_X-a.s.

One way to check that T is sufficient in the Bayesian sense is to check that $\mu_{\Theta|X}(A \mid \cdot)$ is a function of $T(x)$. We have the following result.

Lemma 2. *Let T be a statistic and $\mathcal{B}_T$ the sub-σ-field of $\mathcal{B}$ generated by T. T is sufficient in the Bayesian sense if and only if, for every prior μ_Θ, there exists a version of $\mu_{\Theta|X}$ such that for each $A \in \tau$, $\mu_{\Theta|X}(A \mid \cdot)$ is measurable $\mathcal{B}_T$ (In other words, it is a function of $T(x)$).*

Proof. 'only if' part: If T sufficient in the Bayesian sense then for every prior and each $A \in \tau$, $\mu_{\Theta|X}(A \mid x) = \mu_{\Theta|T}(A \mid T(x))$ holds μ_X-a.e. Since $\mu_{\Theta|T}(A \mid T(\cdot))$ is measurable $\mathcal{B}_T$ it follows that so is $\mu_{\Theta|X}(A \mid \cdot)$.

'if' part: Suppose that for every prior and each $A \in \tau$, $\mu_{\Theta|X}(A \mid \cdot)$ is measurable $\mathcal{B}_T$. We want to show T sufficient in Bayesian sense. That is, that $\mu_{\Theta|X}(A \mid x) = \mu_{\Theta|T}(A \mid T(x))$ μ_X-a.s. We use the fact (e.g. Schervish, Proposition A.49 (4) p. 588) that for two functions functions f and g that are measurable w.r.t. a σ-field $\mathcal{F}$ and a measure μ

$$\int_B f d\mu = \int_B g d\mu \text{ for each } B \in \mathcal{F} \text{ implies } f(x) = g(x), \mu - a.e.$$

Hence, in our case it is sufficient to show that for each $B \in \mathcal{B}_T$

$$\int_B \mu_{\Theta|X}(A \mid x)\mu_X(dx) = \int_B \mu_{\Theta|T}(A \mid T(x))\mu_X(dx).$$

The LHS is $\Pr(\Theta \in A, X \in B)$. Since $B \in \mathcal{B}_T$ there is a set $C \in \mathcal{C}$ such that $B = T^{-1}C$. The RHS becomes

$$\begin{aligned}\int_B \mu_{\Theta|T}(A \mid T(x))\mu_X(dx) &= \{\text{change of variables}\} \\ &= \int_C \mu_{\Theta|T}(A \mid t)\mu_T(dt) \\ &= \Pr(\Theta \in A, T(X) \in C) \\ &= \Pr(\Theta \in A, X \in B).\end{aligned}$$

Hence, we have the desired equality and the proof is complete. □

Example 15. Let $\{X_n\}$ be conditionally IID Exp(θ) given $\Theta = \theta$ and $X = (X_1, \ldots, X_n)$. Put $T(x) = x_1 + \cdots + x_n$. Let us show T is sufficient in the Bayesian sense. Let μ_Θ be the prior (which is arbitrary). Then the posterior distribution has density (Bayes theorem)

$$f_{\Theta|X}(\theta \mid x) = \frac{\prod_{i=1}^n \theta e^{-\theta x_i}}{\int \prod_{i=1}^n \psi e^{-\psi x_i}\mu_\Theta(d\psi)} = \frac{\theta e^{-\theta \sum_{i=1}^n x_i}}{\int \psi^n e^{-\psi \sum_{i=1}^n x_i}\mu_\Theta(d\psi)} = .$$

Since $T(X)$ is $\Gamma(n,\theta)$ given $\Theta=\theta$ it follows that

$$f_{\Theta|T}(\theta\mid t)=\frac{\frac{\theta^n}{\Gamma(n)}t^{n-1}e^{-\theta t}}{\int\frac{\psi^n}{\Gamma(n)}t^{n-1}e^{-\psi t}\mu_\Theta(d\psi)}=\frac{\theta^n e^{-\theta t}}{\int\psi^n e^{-\psi t}\mu_\Theta(d\psi)}.$$

Hence $f_{\Theta|X}(\theta\mid x)=f_{\Theta|T}(\theta\mid T(x))$ so T is sufficient in the Bayesian sense.

It is satisfying to know that in most situations one may encounter the classical and Bayesian notion of sufficiency are the same.

Theorem 8. *Let $(\mathcal{T},\mathcal{C})$ be a measurable space and T a statistic. Suppose there exists a σ-finite measure ν such that $\mu_{X|\Theta}(\cdot\mid\theta)\ll\nu$ for all $\theta\in\Omega$. Then T is sufficient in the classical sense if and only if it is sufficient in the Bayesian sense.*

Density proof. Suppose all relevant densities exists.

Let μ_Θ be an arbitrary prior. If T sufficient in the classical sense, then $f_{X|\Theta,T}(x\mid\theta,t)=f_{X|T}(x\mid t)$. Hence the posterior density is (with $t=T(x)$)

$$\begin{aligned}\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta\mid x)&=\frac{f_{X|\Theta}(x\mid\theta)}{\int_\Omega f_{X|\Theta}(x\mid\theta)\mu_\Theta(d\theta)}\\&=\frac{f_{X|\Theta,T}(x\mid\theta,t)f_{T|\Theta}(t\mid\theta)}{\int_\Omega f_{X|\Theta,T}(x\mid\theta,t)f_{T|\Theta}(t\mid\theta)\mu_\Theta(d\theta)}\\&=\frac{f_{X|T}(x\mid t)f_{T|\Theta}(t\mid\theta)}{\int_\Omega f_{X|T}(x\mid t)f_{T|\Theta}(t\mid\theta)\mu_\Theta(d\theta)}\\&=\frac{f_{T|\Theta}(t\mid\theta)}{\int_\Omega f_{T|\Theta}(t\mid\theta)\mu_\Theta(d\theta)}\\&=\frac{d\mu_{\Theta|T}}{d\mu_\Theta}(\theta\mid t).\end{aligned}$$

For the converse suppose that T is sufficient in the Bayesian sense so that $\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta\mid x)=\frac{d\mu_{\Theta|T}}{d\mu_\Theta}(\theta\mid T(x))$. Then, with $t=T(x)$ we have

$$\begin{aligned}f_{X|T,\Theta}(x\mid t,\theta)&=\frac{f_{X|\Theta}(x\mid\theta)}{f_{T|\Theta}(t\mid\theta)}\\&=\frac{\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta\mid x)\int_\Omega f_{X|\Theta}(x\mid\theta)\mu_\Theta(d\theta)}{\frac{d\mu_{\Theta|T}}{d\mu_\Theta}(\theta\mid t)\int_\Omega f_{T|\Theta}(t\mid\theta)\mu_\Theta(d\theta)}\\&=\frac{f_X(x)}{f_T(t)}\end{aligned}$$

which does not depend on θ. Hence, T is sufficient in the classical sense. □

7.3. **How to find a sufficient statistic?** Suppose someone hands you a parametric family $\mathcal{P}_0=\{f_{X|\Theta}(\cdot\mid\theta),\theta\in\Omega\}$ of densities w.r.t. a measure ν. How do you come up with a sufficient statistic T? Further, if you have come up with a suggestion of a statistic T, how do you check if it is a sufficient statistic? The next theorem gives the answer.

Theorem 9 (Factorization Theorem, c.f. Theorem 6.2.6 in Casella & Berger)**.** *Suppose $\mathcal{P}_0=\{P_\theta:\theta\in\Omega\}$ is a parametric family and there exists a σ-finite ν such*

that $P_\theta \ll \nu$ for all $\theta \in \Omega$ with $dP_\theta/d\nu(x) = f_{X|\Theta}(x \mid \theta)$. Then $T(X)$ is sufficient for Θ (in either sense) if and only if there exist functions h and g such that

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

Density proof. Supposing all the relevant densities exist.

It is sufficient to check the equivalence in the Bayesian sense. If $f_{X|\Theta}(x \mid \theta) = h(x)g(\theta, T(x))$, the by Bayes' theorem

$$\begin{aligned}\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta \mid x) &= \frac{f_{X|\Theta}(x \mid \theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta)\mu_\Theta(d\theta)} \\ &= \frac{h(x)g(\theta, T(x))}{\int_\Omega h(x)g(\theta, T(x))\mu_\Theta(d\theta)} \\ &= \frac{g(\theta, T(x))}{\int_\Omega g(\theta, T(x))\mu_\Theta(d\theta)},\end{aligned}$$

which is a function of $T(x)$. Hence it is sufficient in the Bayesian sense (and also in the classical sense). Conversely, suppose $T(X)$ is sufficient in the Bayesian sense so that $f_{\Theta|X}(\theta \mid x) = f_{\Theta|T}(\theta \mid T(x))$. Then

$$f_{X|\Theta}(x \mid \theta) = f_{\Theta|X}(\theta \mid x)f_X(x) = \underbrace{f_X(x)}_{h(x)}\underbrace{f_{\Theta|T}(\theta|T(x))}_{g(\theta,T(x))}$$

□

Example 16 (Exponential families)**.** If we put $T(X) = (t_1(X), \ldots, t_k(X))$ then by the factorization theorem it follows that $T(X)$ is sufficient. Indeed,

$$f_{X|\Theta}(x \mid \theta) = \underbrace{h(x)}_{h(x)}\underbrace{c(\theta)\exp\{\sum_{i=1}^{k}\theta_i t_i(x)\}}_{g(\theta,T(x))}.$$

Hence, a sufficient statistic always exists. We can compute the density of the sufficient statistics.

8. A first glance at decision theory

Many statistical problems can be phrased in the language of decision theory. Suppose as usual that we have data X whose distribution depend on a parameter Θ. Based on observing $X = x$ we want to take some action. Let $\aleph$ be a set of possible actions. On $\aleph$ we take a σ-field α. The result of our action depend of course on the chosen action, but also on the parameter Θ. We say that every action induces a loss. A *loss function* is a function $L : \Omega \times \aleph \to \mathbb{R}$. We interpret $L(\theta, a)$ as the incurred loss if we took action a and $\Theta = \theta$.

One could let the loss depend on some other unobserved quantity V but we will not need this higher generality right now.

Definition 10. A deterministic decision rule is a measurable function δ from $\mathcal{X}$ to $\aleph$. We interpret $\delta(x)$ as the action to take if we observe $X = x$.

A randomized decision rule is a mapping from $\mathcal{X}$ to a probability measure on $(\aleph, \alpha)$ such that $x \mapsto \delta(A; x)$ is measurable for each $A \in \alpha$.

We think of executing a randomized decision rule as follows. Given $X = x$ we "throw a coin" according to $\delta(\cdot\,; x)$. This gives us an action $a \in \aleph$ which is the decision we take. A deterministic decision rule can be thought of as a special case of a randomized decision rule where all the probability mass is placed at a single action. In this case the action $\delta(x)$ is identified with the probability measure on $\aleph$ given by $\delta(A; x) = I_A(\delta(x))$.

If δ is a deterministic rule, then we associate the loss $L(\theta, \delta(x))$. If δ is a randomized decision rule we associate the loss $L(\theta, \delta(\cdot; x)) = \int_{\aleph} L(\theta, a)\delta(da; x)$. That is, the average loss when we draw the action from $\delta(\cdot; x)$.

In the Bayesian case, one introduces the *posterior risk function*

$$r(\delta \mid x) = \int_{\Omega} L(\theta, \delta(x))\mu_{\Theta|X}(d\theta \mid x).$$

That is, the average loss for decision rule δ given the observation $X = x$. One would like to find a decision rule that minimizes the posterior risk simulatenously for all $x \in \mathcal{X}$.

Definition 11. If δ_0 is a decision rule such that for all x, $r(\delta_0 \mid x) < \infty$ and for all x and all decision rules δ $r(\delta_0 \mid x) \leq r(\delta \mid x)$, then δ_0 is called a *formal Bayes rule*.

If δ_0 is a decision rule and there exists a subset $B \subset \mathcal{X}$ such that for all $x \in B$, $r(\delta_0 \mid x) < \infty$ and for all $x \in B$ and all decision rules δ, $r(\delta_0 \mid x) \leq r(\delta \mid x)$, then δ_0 is called a *partial Bayes rule*.

In classical decision theory we condition on $\Theta = \theta$ and introduce the *risk function*

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x))\mu_{X|\Theta}(dx \mid \theta).$$

That is, the conditional mean of the loss, given $\Theta = \theta$. Here we would like to find a rule δ that minimizes the risk function simultaneously for all values of θ.

8.1. **A coin tossing example.** Consider the following situation. You have an amount of m dollars to bet on the outcome of a Bernoulli random variable X_{n+1}. You observe $X = (X_1, \ldots, X_n)$. Suppose $X_1, \ldots, X_{n+1}$ are conditionally iid $\text{Ber}(\theta)$ random variables given $\Theta = \theta$. Based on the observations in X you have to make a decision whether to bet on $X_{n+1} = 0$ or $X_{n+1} = 1$. If you win, you win the amount m and otherwise you lose m.

Formulate this as a Bayesian decision problem. Write down the sample space $\mathcal{X}$, the parameter space Ω, and the action space $\aleph$. Choose an appropriate prior distribution and an appropriate loss function of your choice. Then find the best decision rule, i.e. the decision rule δ that minimizes the posterior risk simulatenously for all x.

8.2. **Sufficient statistics in decision theory.** If T is a sufficient statistic we would expect that we can base our decisions on T and do not need all the information in X since T contains all information about the unknown parameter Θ. In the Bayesian setting we have the following theorem that supports this.

Theorem 10. *If there is a formal Bayes rule and T is a sufficient statistics (in the Bayesian sense) then there is a formal Bayes rule which is a function of T.*

Proof. Let δ be a formal Bayes rule and take $x \in \mathcal{X}$. Since T is sufficient we have

$$r(\delta \mid x) = \int_\Omega L(\theta, \delta) \mu_{\Theta|X}(d\theta \mid x) = \int_\Omega L(\theta, \delta) \mu_{\Theta|T}(d\theta \mid T(x)).$$

We claim that for each y such that $T(x) = T(y)$ it follows that $r(\delta \mid x) = r(\delta \mid y)$. If not, suppose without loss of generality that $r(\delta \mid x) < r(\delta \mid y)$ for some $y \in \mathcal{X}$ with $T(y) = T(x)$. Let δ' be a decision rule such that $\delta'(z) = \delta(x)$ for all z such that $T(x) = T(z)$. Then it follows that

$$\begin{aligned} r(\delta' \mid y) &= \int_\Omega L(\theta, \delta') \mu_{\Theta|X}(d\theta \mid y) \\ &= \int_\Omega L(\theta, \delta') \mu_{\Theta|T}(d\theta \mid T(y)) \\ &= \int_\Omega L(\theta, \delta) \mu_{\Theta|T}(d\theta \mid T(x)) \\ r(\delta \mid x) &< r(\delta \mid y), \end{aligned}$$

which contradicts that δ is a formal Bayes rule. We conclude that the claim is true.

The decision rule δ' just defined is a function of $T(x)$ and satisfies $r(\delta' \mid T(x)) = r(\delta \mid x)$ for each x. Hence, it is a formal Bayes rule that is a function of T. □

Note that (in the proof above) the formal Bayes rule δ that we started with do not have to be a function of T. For instance, it may be the case that $T(x) = T(y)$, $\delta(x) \neq \delta(y)$, and $L(\theta, \delta(x)) = L(\theta, \delta(y))$ for each θ. Then $r(\delta \mid x) = r(\delta \mid y)$ although, $\delta(x) \neq \delta(y)$.

In the classical setting we have the following.

Theorem 11. *If δ_0 is a (randomized) decision rule and T is sufficient statistic (in classical sense), then there exists a decision rule δ_1 which is a function of T and $R(\theta, \delta_0) = R(\theta, \delta_1)$ for all θ.*

In the theorem, if δ_0 is deterministic we interpret it as the randomized rule δ by $\delta(A; x) = I_A(\delta_0(x))$. That is, the probability measure on $(\aleph, \alpha)$ that puts all its mass on $\delta_0(x)$.

Proof. Let $A \in \alpha$ and take

$$\delta_1(A; x) = E_\theta[\delta_0(A; X) \mid T = t].$$

Since T is sufficient the expectation does not depend on θ. We claim that for any δ_0-integrable function $h : \aleph \to \mathbb{R}$

$$E\Big[\int_\aleph h(a) \delta_0(da; X) \mid T = t\Big] = \int_\aleph h(a) \delta_1(da; t).$$

To see this, start with h as indicator, then simple function, and finally measurable function. Then we see that

$$\begin{aligned}
R(\theta,\delta_1) &= \int_{\mathcal{X}}\int_{\aleph} L(\theta,a)\delta_1(da;T(x))\mu_{X|\Theta}(dx\mid\theta)\\
&= \int_{\mathcal{X}} E\Big[\int_{\aleph} L(\theta,a)\delta_0(da;X)\mid T=T(x)\Big]\mu_{X|\Theta}(dx\mid\theta)\\
&= E_\theta\Big[E\Big[\int_{\aleph} L(\theta,a)\delta_0(da;X)\mid T\Big]\Big]\\
&= E_\theta\Big[\int_{\aleph} L(\theta,a)\delta_0(da;X)\Big]\\
&= \int_{\mathcal{X}}\int_{\aleph} L(\theta,a)\delta_0(da;x)\mu_{X|\Theta}(dx\mid\theta)\\
&= R(\theta,\delta_0).
\end{aligned}$$

□

One should note that even if δ_0 is a deterministic rule, the resulting rule $\delta_1(A;t) = E_\theta[\delta_0(A;X)\mid T=t]$ may be randomized.

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 12. 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 12. *If there exist version of $f_{X|\Theta}(x \mid \theta)$ for each θ 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 17. Let $\{X_n\}$ be IID Exp(θ) given $\Theta = \theta$ and $X = (X_1, \ldots, X_n)$. 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 θ 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 $h(x, y) = h(y)/h(x)$. □

The next concept is that of a complete statistic.

Definition 13. 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 18. Suppose that T has $\text{Bin}(n, \theta)$ 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 θ, 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 13. *If the natural parameter space Ω 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 ν_T' (see Section 4.2, Lecture 4). Let g be a measurable function such that $E_\theta[g(T)] = 0$ for all θ. 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 θ_0 in the interior of Ω. This is possible since Ω 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 $g^+(t) = g^-(t)$ ν_T'-a.e. and hence that $\mu_{T|\Theta}\{t : g(t) = 0 \mid \theta\} = 1$ for all θ. Hence, T is complete sufficient statistic. □

Completeness of a statistic is also related to minimal sufficiency.

Theorem 14 (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 $T = (T_1(X), \ldots, T_k(X))$ 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_θ-a.s. for all θ. 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_θ-a.s. in two steps.

First step: $S_i(T) = Y_i(T)$ P_θ-a.s. for all θ. 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 θ, $E_\theta[Y_i(T) - S_i(T)] = 0$ 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 θ.

Second step: $X_i(U) = Y_i(T)$ P_θ-a.s. for all θ. By step one we have $E_\theta[Y_i(T) \mid U] = X_i(U)$ P_θ-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_θ-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_θ-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 14. 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 θ.

Example 19. 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 θ.

Sometimes a statistic contains a coordinate that is ancillary.

Definition 15. 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 20. Let $X = (X_1, \ldots, X_n)$ 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 12 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 θ. 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 Θ.

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 21 (continued)**.** Consider the above example with $n = 2$ and consider finding a 50% confidence interval for Θ. 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 θ and probability $1/4$ that both are below θ.

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 Θ 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 Θ 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 15 (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 μ_Θ 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 θ and since T is boundedly complete $\mu_{T|\Theta}(\{t : g(t) = 0\} \mid \theta) = 1$ for all θ. 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 22. Let $X = (X_1, \dots, X_n)$ be conditionally IID Exp(θ) 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 $X \overset{\mathrm{d}}{=} \theta^{-1}Z$ 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 θ 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 15) $T(X)$ 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 $E_\theta[g(X)] = n^{-1}$.