Lecture Notes for STAT-F404, author: Marc Hallin[1]

2 Sufficiency

2.1 Dominated models

2.1.1 Measures

Denote by $(\mathcal{X}, \mathcal{A})$ a space equipped with a σ-field (a *measurable space*). Recall that a (positive) *measure* over $(\mathcal{X}, \mathcal{A})$ is a nonnegative set function $\mu : \mathcal{A} \longrightarrow \bar{\mathbb{R}}^+ = \mathbb{R}^+ \cup \{+\infty\}$ such that (σ-*additivity*)

$$\mu(A_1 \cup A_2 \cup \ldots) = \mu(A_1) + \mu(A_2) + \ldots$$

as soon as $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint. Note that this implies that $\mu(\emptyset) = 0$.

Familiar examples are

(i) *(Lebesgue measures)* the *Lebesgue measure* defined over $(\mathbb{R}^k, \mathcal{B}^k)$, where $\mathcal{B}^k$ is the Borel σ-field over $\mathbb{R}^k$, provides the Borel set's usual length for $k = 1$, area for $k = 2$, volume for $k = 3$, etc.

(ii) *(counting measures)* denoting by $\{a_i\}$ a finite or countable subset of $\mathcal{X}$, the measure μ defined over $(\mathcal{X}, \mathcal{A})$ by

$$\mu(A) := \#\{i : a_i \in A\}, \qquad A \in \mathcal{A}$$

(where $\#E$ stands for the possibly infinite cardinality of a set E) is called the *counting measure* associated with $\{a_i\}$. Examples are, over $(\mathbb{R}, \mathcal{B})$, the counting measures associated with $\{0, 1, \ldots, k\}$, with the set of integers $\mathbb{Z}$, with the set of natural numbers $\mathbb{N}$, or with the set of rationals $\mathbb{Q}$ (the latter yielding a rather weird measure under which all nonempty open intervals have measure ∞);

[1]With slight modifications by Davy Paindaveine and Thomas Verdebout.

(iii) *(probability measures)* a probability measure is a measure μ such that $\mu(\mathcal{X}) = 1$.

A measure over $(\mathcal{X}, \mathcal{A})$ is *σ-finite* if there exist $A_1, A_2, \ldots$ in $\mathcal{A}$ such that $\mu(A_i) < \infty$ and $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Examples are the Lebesgue measure over $(\mathbb{R}^k, \mathcal{B}^k)$, and the counting measures over $(\mathbb{R}, \mathcal{B})$ associated with $\mathbb{Z}$, $\mathbb{N}$, or $\mathbb{Q}$. A measure which is not σ-finite is μ defined over $(\mathcal{X}, \mathcal{A})$ by $\mu(\emptyset) = 0$, $\mu(A) = \infty$ for all $A \neq \emptyset$.

In the sequel, when a measurable space $(\mathcal{X}, \mathcal{A})$ is equipped with the measure μ, we tacitly assume that $\mathcal{A}$ has been *completed* for μ, that is, comprises all subsets of $\mathcal{X}$ that are included in a set with μ-measure zero; the μ-measure of such subsets is automatically zero[2].

2.1.2 Integrals

All integrals in the sequel are *Lebesgue integrals*. We will not attempt a rigorous definition of such integrals, for which we refer to measure theory or probability textbooks. Let f be a measurable function from $(\mathcal{X}, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$. The Lebesgue integral of f, when it exists, is denoted as

$$\int_{\mathcal{X}} f(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}).$$

Quite naturally, we let

$$\int_A f(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) := \int_{\mathcal{X}} I_A(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}),$$

where

$$I_A(\mathbf{x}) := \begin{cases} 1 & \mathbf{x} \in A \\ 0 & \mathbf{x} \notin A \end{cases}$$

is the indicator function of $A (\in \mathcal{A})$. For $f = 1$, we get $\int_A \mathrm{d}\mu = \mu(A)$.

[2]The Borel σ-field $\mathcal{B}$ for $\mathbb{R}$, for instance, is not complete for the Lebesgue measure μ. The σ-field $\mathcal{B}_0$ generated by $(\mathcal{B}, \mathcal{N}_\mu)$, where $\mathcal{N}_\mu$ is the collection of all subsets of Borel sets with Lebesgue measure zero, is called the *Lebesgue* σ-field. The elements B of $\mathcal{B}_0$ are of the form $A \cup C$, where $C \in \mathcal{N}$ and $A \cap C = \emptyset$; the Lebesgue measure μ then can be extended to $\mathcal{B}_0$ by putting $\mu(B) := \mu(A)$. The Lebesgue σ-field is complete for this extended Lebesgue measure.

(i) If μ is the Lebesgue measure over $(\mathbb{R}, \mathcal{B})$ and f is a bounded Riemann-integrable function, then its Lebesgue and Riemann integrals over intervals coincide:

$$\int_{[a,b]} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_{[a,b)} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_{(a,b]} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_{(a,b)} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_a^b f(\mathbf{x})\mathrm{d}\mathbf{x}$$

for all $a \leq b$, where the last integrable is the Riemann integral of f from a to b. Lebesgue-integrable functions, however, need not be Riemann-integrable. A classical counterexample is the indicator function $I_{\mathbb{Q}}$ of $\mathbb{Q}$ (since $\mathbb{Q}$ is a countable subset of $\mathbb{R}$, we have $\int_{[0,1]} I_{\mathbb{Q}}(x)\mathrm{d}\mu(x) = 0$, but the corresponding Riemann integral does not exist).

(ii) If μ is the counting measure of $\{a_i, \ldots, a_k\}$, then

$$\int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \sum_{i=1}^{k} f(a_i),$$

whereas if μ is the counting measure of $\{a_1, a_2, \ldots\}$, then

$$\int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \sum_{i=1}^{\infty} f(a_i).$$

(iii) If μ is a probability measure P, then the Lebesgue integral of f is nothing else than the expectation, under $\mathbf{X} \sim \mathrm{P}$, of $f(\mathbf{X})$:

$$\int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x})\mathrm{dP}(\mathbf{x}) = \mathrm{E}_{\mathrm{P}}[f(\mathbf{X})].$$

In particular, when μ is a discrete probability measure P, with atoms $x_1, x_2, \ldots$ and probability weights $p_1, p_2, \ldots$,

$$\int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \sum_{i=1}^{\infty} f(x_i)p_i$$

(obviously, this would just be a finite sum if P would have only finitely many atoms).

2.1.3 Radon-Nikodym derivatives

Let μ and ν be two measures defined over the same $(\mathcal{X}, \mathcal{A})$ space. We say that ν *is dominated by* μ or, equivalently, that ν *is absolutely continuous with respect to* μ (notation: $\nu << \mu$) if, for any $A \in \mathcal{A}$, $\mu(A) = 0$ implies $\nu(A) = 0$. When two (σ-finite) measures are mutually absolutely continuous, we say that they are *equivalent*. The following theorem then plays a central role in the definition of conditional expectations and conditional probabilities.

Theorem 1. (Radon-Nikodym) *Let μ and ν be two measures over $(\mathcal{X}, \mathcal{A})$, with μ being σ-finite. Then, $\nu << \mu$ if and only if there exists a function $f : \mathcal{X} \longrightarrow \mathbb{R}^+$ such that*

$$\Big(\nu(A) = \Big) \int_A \mathrm{d}\nu(\mathbf{x}) = \int_A f(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \tag{2.1}$$

for all $A \in \mathcal{A}$.

The function f in (2.1) is not uniquely defined; however it is *essentially unique*, in the sense that, if f_1 and f_2 are such that (2.1) holds, then

$$\mu\left(\{\mathbf{x} : f_1(\mathbf{x}) \neq f_2(\mathbf{x})\}\right) = 0,$$

that is, they coincide up to a set of μ-measure zero. The set of all μ-almost everywhere equal functions such that (2.1) holds is denoted as $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, and called the *Radon-Nikodym derivative* of ν with respect to μ. An arbitrary element (called a *version* of the Radon-Nikodym derivative) of $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, however, entirely characterizes the whole class; therefore, with a small abuse of notation, we also denote such a version by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, taking at $\mathbf{x} \in \mathcal{X}$ value $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\mathbf{x})$. The characteristic property (2.1) with that notation takes the form

$$\nu(A) = \int_A \mathrm{d}\nu(\mathbf{x}) = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \qquad \text{for all } A \in \mathcal{A}.$$

More generally, we have that, for any measurable function g,

$$\int_A g(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) = \int_A g(\mathbf{x}) \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \qquad \text{for all } A \in \mathcal{A}.$$

When $\mathrm{P} << \mu$, where μ is σ-finite and P is a probability measure, we say that $f_{\mathrm{P}} := \frac{\mathrm{dP}}{\mathrm{d}\mu}$ is the *probability density* of P with respect to μ, as (2.1) yields

$$\mathrm{P}[A] = \int f_{\mathrm{P}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \tag{2.2}$$

for all $A \in \mathcal{A}$. Probability densities, thus, are by essence defined up to a set of measure zero in the reference measure.

We now state two useful properties of Radon-Nikodym derivatives (we state these only for probability measures, although they extend to more general measures, which we will actually use in the sequel). Letting $\mathrm{P} << \mathrm{Q} << \mathrm{R}$ be probability measures over $(\mathcal{X}, \mathcal{A})$, we have the following:

(a) if $f \in \frac{\mathrm{dP}}{\mathrm{dQ}}$ and $g \in \frac{\mathrm{dQ}}{\mathrm{dR}}$, then $fg \in \frac{\mathrm{dP}}{\mathrm{dR}}$;

(b) if $f \in \frac{\mathrm{dP}}{\mathrm{dR}}$ and $g \in \frac{\mathrm{dQ}}{\mathrm{dR}}$, then $f/g \in \frac{\mathrm{dP}}{\mathrm{dQ}}$;

In (b), note that

$$\mathrm{Q}\left(\{\mathbf{x} : g(\mathbf{x}) = 0\}\right) = \int_{\{\mathbf{x}:g(\mathbf{x})=0\}} g(\mathbf{x})dR(\mathbf{x}) = 0,$$

so that $f(\mathbf{x})/g(\mathbf{x})$ is well-defined up to a set with Q-measure zero, hence can be given an arbitrary value at any $\mathbf{x}$ such that $g(\mathbf{x}) = 0$; "dividing by zero" thus is not a problem there.

Let us give a few examples of probability densities.

(i) The $\mathcal{N}(0, 1)$ probability measure over $(\mathbb{R}, \mathcal{B})$ has density

$$f(x) = (2\pi)^{-1/2} \exp\Big(-\frac{1}{2}x^2\Big), \qquad x \in \mathbb{R},$$

with respect to the Lebesgue measure. All probability distributions (over $\mathbb{R}$ or $\mathbb{R}^k$) called *absolutely continuous* in elementary textbooks, with density f defined as the derivative of a cumulative distribution function, actually are absolutely continuous with respect to the Lebesgue measure, and have density f (in the sense of (2.2)) with respect to the same (more precisely, f is a version of that density).

(ii) The Bernoulli $\text{Bin}(1, p)$ measure over $(\mathbb{R}, \mathcal{B})$, with $p \in (0, 1)$, is defined by

$$\mathrm{P}_p[A] = \begin{cases} 0 & \text{if } 0, 1 \notin A \\ p & \text{if } 0 \notin A \text{ and } 1 \in A \\ 1-p & \text{if } 0 \in A \text{ and } 1 \notin A \\ 1 & \text{if } 0, 1 \in A \end{cases}$$

for any $A \in \mathcal{B}$. That measure is absolutely continuous with respect to the counting measure associated with $\{0, 1\}$, with density

$$f_p(x) = p^x(1-p)^{1-x}, \qquad x \in \mathbb{R}.$$

Note that *any* other function f such that

$$f(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases}$$

is another *version* of the same density.

(iii) Similarly, the binomial $\text{Bin}(n, p)$ measure has density

$$f_{n,p}(x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x \in \mathbb{R},$$

with respect to the counting measure of $\{0, 1, \ldots, n\}$, the Poisson(λ) measure has density

$$f_\lambda(x) = \exp(-\lambda)\frac{\lambda^x}{x!}, \qquad x \in \mathbb{R},$$

with respect to the counting measure of $\mathbb{N}$, etc.

Denote by $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ a statistical model. That model is said to be *dominated* by the σ-finite measure μ if $\mathcal{P}$ is *dominated* by μ (notation: $\mathcal{P} << \mu$), namely, if for every $\mathrm{P} \in \mathcal{P}$, $\mathrm{P} << \mu$. Then, $\mathcal{P}$ can alternatively be described as a family of densities: $\{f_{\mathrm{P}} := \frac{\mathrm{dP}}{\mathrm{d}\mu} : \mathrm{P} \in \mathcal{P}\}$.

A model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ (a family $\mathcal{P}$) is called a *dominated model* (a *dominated family*) if there exists a σ-finite measure μ such that $\mathcal{P} << \mu$. Halmos and Savage (1949) proved the following lemma, showing that dominated families can be characterized without recurring to any "external" measure μ.

Lemma 1. *(Halmos and Savage, 1949) A family of probability measures $\mathcal{P}$ defined over the space $(\mathcal{X}, \mathcal{A})$ is a dominated family if and only if there exist a countable subset $\{\mathrm{P}_1, \mathrm{P}_2, \ldots\}$ of $\mathcal{P}$ and a sequence (c_i) of nonnegative real numbers satisfying $\sum_{i=1}^{\infty} c_i = 1$ such that*

$$\mathcal{P} << \mathrm{P}_* := \sum_{i=1}^{\infty} c_i \mathrm{P}_i. \tag{2.3}$$

The probability measure P_* is called a privileged *(dominating)* measure.

Note that (2.3) actually states that $\mathrm{P}_i[A] = 0$ for all i implies $\mathrm{P}[A] = 0$ for all $\mathrm{P} \in \mathcal{P}$, while the converse is trivially true. That fact could be described as $\mathcal{P}$ and the countable subfamily $\{\mathrm{P}_1, \mathrm{P}_2, \ldots\}$ being *mutually absolutely continuous* or *equivalent.* Lemma 1 then can be restated without mentioning any constants c_i nor any privileged P_*:

Lemma 1. *A family of probability measures $\mathcal{P}$ defined over the space $(\mathcal{X}, \mathcal{A})$ is a dominated family if and only if it is equivalent to one of its countable subsets.*

Privileged measures are indeterminate to a very large extent: if $\mathrm{P}_* = \sum_{i=1}^{\infty} c_i^* \mathrm{P}_i$ is a privileged measure, then any $\mathrm{P}_{**} = \sum_{i=1}^{\infty} c_i^{**} \mathrm{P}_i$ such that $c_i^{**} > 0$ if and only if $c_i^* > 0$ (with $\sum_{i=1}^{\infty} c_i^{**} = 1 = \sum_{i=1}^{\infty} c_i^*$) also is a privileged measure.

2.2 Conditional expectations

Denote by $\mathbf{T}$ a *statistic* defined over $(\mathcal{X}, \mathcal{A})$, with values in $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$, i.e. a function $\mathbf{T} : (\mathcal{X}, \mathcal{A}) \longrightarrow (\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ mapping $\mathbf{x} \in \mathcal{X}$ onto $\mathbf{T}(\mathbf{x}) \in \mathcal{T}$ and such that $\mathbf{T}^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}_{\mathcal{T}}$. Then $\mathcal{A}_{\mathbf{T}} := \mathbf{T}^{-1}(\mathcal{B}_{\mathcal{T}})$ is the smallest sub-σ-field of $\mathcal{A}$ with respect to which $\mathbf{T}$ is measurable. Call it the *σ-field generated by* $\mathbf{T}$. The statistic $\mathbf{T}$ maps each probability measure P defined over $(\mathcal{X}, \mathcal{A})$ onto a probability measure $\mathrm{P}^{\mathbf{T}}$ over $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$. Namely, for all $B \in \mathcal{B}_{\mathcal{T}}$, we have

$$\mathrm{P}^{\mathbf{T}}[B] := \mathrm{P}[\mathbf{T}^{-1}(B)].$$

That measure $\mathrm{P}^{\mathbf{T}}$ (the probability distribution of $\mathbf{T}(\mathbf{X})$ when $\mathbf{X} \sim \mathrm{P}$) is called an *induced* probability measure. Similarly, the family $\mathcal{P}^{\mathbf{T}} = \left\{\mathrm{P}^{\mathbf{T}} : \mathrm{P} \in \mathcal{P}\right\}$ is called an *induced* family and the statistical model $\left(\mathcal{T}, \mathcal{B}_{\mathcal{T}}, \mathcal{P}^{\mathbf{T}}\right)$ an *induced* model.

Such induced models typically are simpler than the original ones, sometimes much simpler, hence more convenient to work with. Intuitively, they cannot provide more information than the original ones: observing $\mathbf{T}(\mathbf{X})$ cannot be more informative than observing $\mathbf{X}$ itself. Very clearly, however, they can provide less, and even much less information. A question then naturally arises: is it possible to simplify, via a statistic $\mathbf{T}$, a model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ into a model $\left(\mathcal{T}, \mathcal{B}_{\mathcal{T}}, \mathcal{P}^{\mathbf{T}}\right)$ without losing any information on the data-generating process that generated $\mathbf{X}$? That question is the central one behind the concept of *sufficiency*: a statistic $\mathbf{T}$ will be called *sufficient* if $\mathbf{T}(\mathbf{X})$ carries as much information as $\mathbf{X}$ itself on the data-generating process that generated $\mathbf{X}$.

The mathematical translation of that simple idea will require the concepts of conditional expectation and conditional probability, which we now describe.

It can be shown that a measurable function $g : (\mathcal{X}, \mathcal{A}) \longmapsto (\mathbb{R}, \mathcal{B})$ is $\mathbf{T}$-measurable (equivalently, $\mathcal{A}_{\mathbf{T}}$-measurable) if there exists a measurable mapping $h : (\mathcal{T}, \mathcal{B}_{\mathcal{T}}) \longmapsto (\mathbb{R}, \mathcal{B})$ such that $g(\mathbf{x}) = h(\mathbf{T}(\mathbf{x}))$. Then, for all $B \in \mathcal{B}_{\mathcal{T}}$, we have

$$\int_B h(\mathbf{t}) \mathrm{d}\mathrm{P}^{\mathbf{T}}(\mathbf{t}) = \int_{\mathbf{T}^{-1}(B)} \underbrace{h(\mathbf{T}(\mathbf{x}))}_{=g(\mathbf{x})} \mathrm{d}\mathrm{P}(\mathbf{x}), \tag{2.4}$$

meaning that, as soon as one of those integrals exists, so does the other one, and they coincide (this is the *transfer property* of the Lebesgue integral).

In particular, for $B = \mathcal{T}$, hence $\mathbf{T}^{-1}(B) = \mathcal{X}$, with $\mathbf{X} \sim \mathrm{P}$, hence $\mathbf{T} \sim \mathrm{P}^{\mathbf{T}}$, adopting the expectation notation of the integral, (2.4) takes the familiar form

$$\mathrm{E}[h(\mathbf{T})] = \mathrm{E}[h(\mathbf{T}(\mathbf{X}))]. \tag{2.5}$$

Property (2.4) allows us to compute integrals of $\mathbf{T}$-measurable functions either in $\mathcal{T}$ or in $\mathcal{X}$, just as (2.5) tells us that expectations of $\mathbf{T}$ and $\mathbf{T}(\mathbf{X})$ are the same. Can that convenient property be extended also to functions g that are not $\mathbf{T}$-measurable? This is the purpose of *conditional expectations.*

Assume first that g is a *nonnegative* $\mathcal{A}$-measurable and P-integrable function. Can we define a $\mathbf{T}$-measurable function h such that

$$\int_B h(\mathbf{t})\mathrm{dP}^{\mathbf{T}}(\mathbf{t}) = \int_{\mathbf{T}^{-1}(B)} g(\mathbf{x})\mathrm{dP}(\mathbf{x}) \tag{2.6}$$

for all $B \in \mathcal{B}_{\mathcal{T}}$? Since g is nonnegative and P-integrable, the function ν_g from $\mathcal{B}_{\mathcal{T}}$ to $\mathbb{R}^+$ mapping B to

$$\nu_g(B) := \int_{\mathbf{T}^{-1}(B)} g(\mathbf{x})\mathrm{dP}(\mathbf{x})$$

is a finite measure over $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$. That measure ν_g is dominated by $\mathrm{P}^{\mathbf{T}}$, since $\mathrm{P}^{\mathbf{T}}[B] = 0$ implies $\mathrm{P}[\mathbf{T}^{-1}(B)] = 0$, hence $\nu_g(B) = 0$. The Radon-Nikodym theorem then guarantees the existence of an essentially unique function $h = \frac{\mathrm{d}\nu_g}{\mathrm{dP}^{\mathbf{T}}}$ such that

$$\nu_g(B) = \int_B h(\mathbf{t})\mathrm{dP}^{\mathbf{T}}(\mathbf{t}),$$

so that (2.6) holds. The class of functions $h = \frac{\mathrm{d}\nu_g}{\mathrm{dP}^{\mathbf{T}}}$ is called the *conditional expectation* of $g(\mathbf{X})$ given $\mathbf{T}$, and is denoted as $\mathrm{E}_{\mathrm{P}}[g(\mathbf{X})|\mathbf{T}]$. As usual, the same notation is used for any of the elements of that class, which are $\mathrm{P}^{\mathbf{T}}$-almost surely equal, $\mathbf{T}$-measurable, random variables; write $\mathrm{E}_{\mathrm{P}}[g(\mathbf{X})|\mathbf{T} = \mathbf{t}]$ for the value of $\mathrm{E}_{\mathrm{P}}[g(\mathbf{X})|\mathbf{T}]$ at $\mathbf{T} = \mathbf{t}$. With that notation, equation (2.6) takes the form

$$\int_B \mathrm{E}_{\mathrm{P}}[g(\mathbf{X})|\mathbf{T} = \mathbf{t}]\mathrm{dP}^{\mathbf{T}}(\mathbf{t}) = \int_{\mathbf{T}^{-1}(B)} g(\mathbf{x})\mathrm{dP}(\mathbf{x}) \qquad \text{for all } B \in \mathcal{B}_{\mathcal{T}}. \tag{2.7}$$

It remains to extend this construction to functions g that are not nonnegative: for an arbitrary $\mathcal{A}$-measurable, P-integrable, but not necessarily nonnegative g, we decompose g into $g^+ - g^-$, with

$$g^+(\mathbf{x}) := \begin{cases} |g(\mathbf{x})| & \text{if } g(\mathbf{x}) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad g^-(\mathbf{x}) := \begin{cases} |g(\mathbf{x})| & \text{if } g(\mathbf{x}) \leq 0 \\ 0 & \text{otherwise,} \end{cases}$$

and define $\mathrm{E_P}[g(\mathbf{X})|\mathbf{T}] = \mathrm{E_P}[g^+(\mathbf{X})|\mathbf{T}] - \mathrm{E_P}[g^-(\mathbf{X})|\mathbf{T}]$.

Since (2.7) involves the statistic $\mathbf{T}$ only through the σ-field $\mathcal{B}_\mathbf{T}$, the conditional expectation $\mathrm{E_P}[g(\mathbf{X})|\mathbf{T}]$ actually does only depend on $\mathbf{T}$ through $\mathcal{B}_\mathbf{T}$, so that the notation $\mathrm{E_P}[g(\mathbf{X})|\mathcal{A}_\mathbf{T}]$ is also used for $\mathrm{E_P}[g(\mathbf{X})|\mathbf{T}]$. This also implies that $\mathrm{E_P}[g(\mathbf{X})|\ell(\mathbf{T})] = \mathrm{E_P}[g(\mathbf{X})|\mathbf{T}]$ for any one-to-one mapping ℓ. In particular, for a real-valued T, the conditional expectations $\mathrm{E_P}[g(\mathbf{X})|T]$, $\mathrm{E_P}[g(\mathbf{X})|\exp(T)]$, and $\mathrm{E_P}[g(\mathbf{X})|T^3]$ always coincide.

Conditional expectations enjoy most of the elementary properties of expectations:

(a) *linearity*: for any constants c_i and real-valued measurable functions $g_i : \mathcal{X} \longrightarrow \mathbb{R}$,

$$\mathrm{E_P}\left[\sum_i c_i g_i(\mathbf{X})\middle|\mathbf{T}\right] = \sum_i c_i \mathrm{E_P}[g_i(\mathbf{X})|\mathbf{T}],$$

in the sense that if the right-hand side exists and is finite, so does the left-hand side;

(b) for any measurable function ℓ,

$$\mathrm{E_P}[\ell(\mathbf{T})g(\mathbf{X})|\mathbf{T}] = \ell(\mathbf{T})\mathrm{E_P}[g(\mathbf{X})|\mathbf{T}].$$

In particular, since it is easily checked that $\mathrm{E_P}[1|\mathbf{T}] = 1$, we always have that

$$\mathrm{E_P}[\ell(\mathbf{T})|\mathbf{T}] = \ell(\mathbf{T});$$

(c) $\mathrm{E}_{\mathrm{P}^\mathbf{T}}\big[\mathrm{E_P}[g(\mathbf{X})|\mathbf{T}]\big] = \mathrm{E_P}[g(\mathbf{X})]$ (this follows by taking $B = \mathcal{T}$ in (2.7)).

A simple and interesting geometric interpretation of conditional expectation is possible if we restrict to the L^2 space of square-integrable functions, namely the space of all real-valued measurable functions $\mathbf{x} \longmapsto f(\mathbf{x})$ such that $\int_\mathcal{X} f^2(\mathbf{x})\mathrm{dP}(\mathbf{x}) < \infty$, with scalar product

$$< f_1, f_2 >= \int_\mathcal{X} f_1(\mathbf{x})f_2(\mathbf{x})d\mathrm{P}(\mathbf{x}).$$

Let g and ψ belong to L^2, and let ψ be $\mathbf{T}$-measurable, hence of the form $\ell(\mathbf{T}(\mathbf{x}))$. Then, the squared L^2-distance between g and ψ is

$$\begin{aligned}
\mathrm{E}\big[\{g(\mathbf{X})-\psi(\mathbf{X})\}^2\big] &= \mathrm{E}\big[\{g(\mathbf{X})-\ell(\mathbf{T})\}^2\big] = \mathrm{E}\big[\{g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]+\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}^2\big] \\
&= \mathrm{E}\big[\{g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]\}^2\big] \\
&\qquad +2\mathrm{E}[\{g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]\}\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}] \\
&\qquad +\mathrm{E}\big[\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}^2\big]\,.
\end{aligned}$$

Quite obviously,

(a) the first term $\mathrm{E}\big[(g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}])^2\big]$ does not depend on $\ell(\cdot)$;

(b) By the properties of conditional expectations, the second term is zero: indeed,

$$\begin{aligned}
&\mathrm{E}[\{g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]\}\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}] \\
&\quad = \mathrm{E}\big[\mathrm{E}[\{g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]\}\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}\,|\mathbf{T}]\,\big] \\
&\quad = \mathrm{E}\big[\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}\,\mathrm{E}[g(\mathbf{X})-\mathrm{E}[g(\mathbf{X})|\mathbf{T}]|\mathbf{T}]\,\big] \\
&\quad = \mathrm{E}\big[\{\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T})\}\times 0\big] = 0;
\end{aligned}$$

(c) the minimal value of the third term $\mathrm{E}\big[(\mathrm{E}[g(\mathbf{X})|\mathbf{T}]-\ell(\mathbf{T}))^2\big]$ over all possible choices of $\psi(\mathbf{X})=\ell(\mathbf{T})$ is zero, a minimum which is reached at $\psi(\mathbf{X})=\ell(\mathbf{T})=\mathrm{E}[g(\mathbf{X})|\mathbf{T}]$.

It follows that the minimum, over all $\mathbf{T}$-measurable square-integrable functions ψ, of the squared L^2-distance $\mathrm{E}[\{g(\mathbf{X})-\psi(\mathbf{X})\}^2]$ is $\mathrm{E}[\{g(\mathbf{X})-E[g(\mathbf{X})|\mathbf{T}]\}^2]$; in other words, $E[g(\mathbf{X})|\mathbf{T}]$ is the L^2-projection of $g(\mathbf{X})$ onto the space of (square-integrable) $\mathbf{T}$-measurable variables.

2.3 Conditional probabilities

For any $A\in\mathcal{A}$, we have, with $\mathbf{X}\sim\mathrm{P}$,

$$\mathrm{P}[A] = \int_A \mathrm{dP} = \int_{\mathcal{X}} I_A(\mathbf{x})\mathrm{dP}(\mathbf{x}) = \mathrm{E}[I_A(\mathbf{X})]\,: \tag{2.8}$$

the probability of A is the expectation of the indicator of A. Therefore, it is natural to extend that characterization by defining the *conditional probability* $\mathrm{P}[A|\mathbf{T}]$ of A given $\mathbf{T}$ as the $\mathbf{T}$-measurable random variable

$$\mathrm{P}[A|\mathbf{T}] := \mathrm{E}_{\mathrm{P}}[I_A(\mathbf{X})|\mathbf{T}]. \tag{2.9}$$

While (2.8) is a property of expectations defined as integrals, (2.9) is the definition of a new concept: the conditional probability of A given $\mathbf{T}$. From the properties of conditional expectations, we have the following properties for conditional probabilities:

- $\mathrm{P}[A] = \mathrm{E}_{\mathrm{P}}[\mathrm{P}[A|\mathbf{T}]] = \int_{\mathcal{T}} \mathrm{P}[A|\mathbf{T} = \mathbf{t}]\mathrm{d}\mathrm{P}^{\mathbf{T}}(\mathbf{t})$
- $\mathrm{P}[A|\mathbf{T} = \mathbf{t}]$ is defined up to sets of $\mathrm{P}^{\mathbf{T}}$-measure zero.

Whereas for any *fixed* $A \in \mathcal{A}$, $\mathrm{P}[A|\mathbf{T}]$ is a class of $\mathbf{T}$-measurable random variables defined up to a set of $\mathrm{P}^{\mathbf{T}}$-measure zero, there is no guarantee that, for a given fixed value $\mathbf{t}$, there exists a collection of versions

$$\{\mathrm{P}[A|\mathbf{T} = \mathbf{t}] : A \in \mathcal{A}\}$$

constituting a *probability measure* over $(\mathcal{X}, \mathcal{A})$. If such a collection exists, it qualifies as being called the *conditional distribution* over $(\mathcal{X}, \mathcal{A})$ of $\mathbf{X}$, given $\mathbf{T}(\mathbf{X}) = \mathbf{t}$. However, it can be shown that, in "usual cases", such conditional distributions do exist.

Theorem 2. *Let $\mathcal{X}$ be a Borel set in a Euclidean space and $\mathcal{A}$ be the class of Borel subsets of $\mathcal{X}$. Then,*

(i) one can select, for each $A \in \mathcal{A}$, a version $\mathrm{P}^*[A|\mathbf{T}]$ of $\mathrm{P}[A|\mathbf{T}]$ in such a way that, for any fixed $\mathbf{t}$, $A \longmapsto \mathrm{P}^*[A|\mathbf{T} = \mathbf{t}]$, $A \in \mathcal{A}$ constitutes a probability measure over $(\mathcal{X}, \mathcal{A})$ (notation: $\mathrm{P}^{\mathbf{X}|\mathbf{T}=\mathbf{t}}$), and

(ii) $\mathbf{t} \longmapsto \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mathrm{P}^{\mathbf{X}|\mathbf{T}=\mathbf{t}}$ constitutes a version of $\mathrm{E}_{\mathrm{P}}[f|\mathbf{T}]$ (with f a P-integrable, possibly vector-valued random variable).

2.4 Sufficiency

We are now able to provide a precise definition of the concept of a *sufficient statistic*.

Definition 1. *A statistic* $\mathbf{T} : (\mathcal{X}, \mathcal{A}) \longrightarrow (\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ *is* sufficient *for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *if, for all* $A \in \mathcal{A}$*, there exists a version of* $\mathrm{P}[A|\mathbf{T}]$ *that does not depend on* P*, i.e. if, for all* $A \in \mathcal{A}$*,*

$$\bigcap_{\mathrm{P}\in\mathcal{P}} \mathrm{P}[A|\mathbf{T}] \neq \emptyset.$$

Intuitively, if a sufficient statistic $\mathbf{T}$ is known, then the (conditional) probability of any event $A \in \mathcal{A}$ does not depend on which particular $\mathrm{P} \in \mathcal{P}$ is generating the observation. Hence, once $\mathbf{T}$ is known, the observation $\mathbf{X}$ does not carry any additional information about P. *All information on* P *in* $\mathbf{X}$ *is contained in* $\mathbf{T}$, which justifies the terminology *sufficiency*.

Since $\mathrm{P}[A|\mathbf{T}]$ actually depends on $\mathbf{T}$ only through $\mathcal{A}_{\mathbf{T}}$, sufficiency is a property of $\mathcal{A}_{\mathbf{T}}$ rather than $\mathbf{T}$, which will allow us to sometimes write that $\mathcal{A}_{\mathbf{T}}$ itself is sufficient.

2.5 The Halmos-Savage theorem

The following theorem provides, in a *dominated model*, a necessary and sufficient condition for a statistic $\mathbf{T}$ being sufficient.

Theorem 3. *(Halmos and Savage, 1949) Let* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *be a dominated model. The following three statements are equivalent:*

(i) $\mathbf{T}$ *is sufficient;*

(ii) for any $\mathrm{P} \in \mathcal{P}$*, there exists a* $\mathbf{T}$*-measurable version of* $\frac{\mathrm{dP}}{\mathrm{dP}_*}$*, where* P_* *is a specific privileged probability measure;*

(iii) for any $\mathrm{P} \in \mathcal{P}$*, there exists a* $\mathbf{T}$*-measurable version of* $\frac{\mathrm{dP}}{\mathrm{dP}_*}$*, where* P_* *is an arbitrary privileged probability measure.*

Conditions (ii) and (iii) both are necessary and sufficient for sufficiency. Since (iii) obviously implies (ii), Condition (ii) is stronger than (iii) as a sufficient condition, and weaker as a necessary one.

Proof. (i) ⇒ (iii) Assume that $\mathbf{T}$ is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ and let $\mathrm{P}_* = \sum_i c_i \mathrm{P}_i$ be an arbitrary privileged measure. Then, for any $\mathrm{P} \in \mathcal{P}$, we have $\mathrm{P} << \mathrm{P}_*$, hence $\mathrm{P}^{\mathbf{T}} << \mathrm{P}_*^{\mathbf{T}}$. Thus, $\frac{\mathrm{dP}^{\mathbf{T}}}{\mathrm{dP}_*^{\mathbf{T}}}$ exists: arbitrarily pick one of its versions, and denote it as $\mathbf{t} \longmapsto g_{\mathrm{P}}(\mathbf{t})$. The proof then consists in showing that $\mathbf{x} \longmapsto g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))$ is a version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$. For any $A \in \mathcal{A}$, sufficiency of $\mathbf{T}$ implies that there exists a version of $\mathrm{P}[A|\mathbf{T}]$ that does not depend on P, hence is also a version of each of the $\mathrm{P}_i[A|\mathbf{T}]$'s and, therefore, a version of $\mathrm{P}_*[A|\mathbf{T}]$. Taking that fact into account and applying repeatedly the characteristic property of conditional expectations, we have, for any $A \in \mathcal{A}$,

$$
\begin{aligned}
\mathrm{P}[A] &= \int_{\mathcal{X}} \mathrm{I}_A(\mathbf{x}) \mathrm{dP}(\mathbf{x}) = \int_{\mathcal{T}} \mathrm{P}[A|\mathbf{T} = \mathbf{t}] \, \mathrm{dP}^{\mathbf{T}}(\mathbf{t}) \\
&= \int_{\mathcal{T}} \mathrm{P}_*[A|\mathbf{T} = \mathbf{t}] \, \mathrm{dP}^{\mathbf{T}}(\mathbf{t}) = \int_{\mathcal{T}} \mathrm{E}_{\mathrm{P}_*}[\mathrm{I}_A(\mathbf{X})|\mathbf{T} = \mathbf{t}] \, g_{\mathrm{P}}(\mathbf{t}) \mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}) \\
&= \int_{\mathcal{T}} \mathrm{E}_{\mathrm{P}_*}[g_{\mathrm{P}}(\mathbf{T}) \mathrm{I}_A(\mathbf{X})|\mathbf{T} = \mathbf{t}] \, \mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}) = \int_{\mathcal{X}} g_{\mathrm{P}}(\mathbf{T}(\mathbf{x})) \mathrm{I}_A(\mathbf{x}) \mathrm{dP}_*(\mathbf{x}) \\
&= \int_A g_{\mathrm{P}}(\mathbf{T}(\mathbf{x})) \mathrm{dP}_*(\mathbf{x}).
\end{aligned}
$$

This establishes that $\mathbf{x} \mapsto g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))$ is indeed a version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$. Since it is obviously $\mathbf{T}$-measurable, the result follows.

(iii) ⇒ (ii) Trivial.

(ii) ⇒ (i) Fix the privileged measure P_* mentioned in Condition (ii). For any P, let then $\mathbf{x} \longmapsto g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))$ be a $\mathbf{T}$-measurable version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$. First note that, for any $B \in \mathcal{B}_{\mathcal{T}}$,

$$
\begin{aligned}
\mathrm{P}^{\mathbf{T}}[B] = \mathrm{P}[\mathbf{T}^{-1}(B)] &= \int_{\mathbf{T}^{-1}(B)} g_{\mathrm{P}}(\mathbf{T}(\mathbf{x})) \mathrm{dP}_*(\mathbf{x}) \\
&= \int_B \mathrm{E}_{\mathrm{P}_*}[g_{\mathrm{P}}(\mathbf{T})|\mathbf{T} = \mathbf{t}] \, \mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}) = \int_B g_{\mathrm{P}}(\mathbf{t}) \mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}),
\end{aligned}
$$

which shows that $\mathbf{t} \longmapsto g_{\mathrm{P}}(\mathbf{t})$ is a version of $\frac{\mathrm{dP}^{\mathbf{T}}}{\mathrm{dP}_*^{\mathbf{T}}}$. Thus, for any $B \in \mathcal{B}_{\mathcal{T}}$, $\mathrm{P} \in \mathcal{P}$ and any real-valued measurable function ψ, we have

$$\begin{aligned}\int_{\mathbf{T}^{-1}(B)} \psi(\mathbf{x})\mathrm{dP}(\mathbf{x}) &= \int_{\mathbf{T}^{-1}(B)} \psi(\mathbf{x}) g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))\mathrm{dP}_*(\mathbf{x}) \\ &= \int_B \mathrm{E}_{\mathrm{P}_*}[\psi(\mathbf{X}) g_{\mathrm{P}}(\mathbf{T})|\mathbf{T}=\mathbf{t}]\, \mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}) \\ &= \int_B \mathrm{E}_{\mathrm{P}_*}[\psi(\mathbf{X})|\mathbf{T}=\mathbf{t}]\, g_{\mathrm{P}}(\mathbf{t})\mathrm{dP}_*^{\mathbf{T}}(\mathbf{t}) \\ &= \int_B \mathrm{E}_{\mathrm{P}_*}[\psi(\mathbf{X})|\mathbf{T}=\mathbf{t}]\, \mathrm{dP}^{\mathbf{T}}(\mathbf{t}). \end{aligned} \tag{2.10}$$

Thus, for any measurable real-valued function ψ, any version of $\mathrm{E}_{\mathrm{P}_*}[\psi(\mathbf{X})|\mathbf{T}]$ is a version of $\mathrm{E}_{\mathrm{P}}[\psi(\mathbf{X})|\mathbf{T}]$ that does not depend on P. Sufficiency of $\mathbf{T}$ follows by choosing $\psi = \mathrm{I}_A$. □

In view of (2.10), the definition of sufficiency could have been taken as the existence of a version of conditional *expectations* not depending on P, instead of that of a version of conditional *probabilities* not depending on P.

2.6 The Neyman-Fisher factorization criterion

In practice, the Halmos-Savage theorem is not convenient for checking sufficiency. Provided that a dominating measure is well identified, a much simpler method is based on the following result, which goes back to Neyman and Fisher.[3]

Proposition 1. *(The Neyman-Fisher factorization criterion) Let the model* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *be dominated by the* σ*-finite measure* μ*. A statistic* $\mathbf{T}$ *is sufficient for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *if and only if, for any* $\mathrm{P} \in \mathcal{P}$*, there exists a version of* $\frac{\mathrm{dP}}{\mathrm{d}\mu}$*,* f_{P} *say, factorizing* μ*-a.e. into*

$$f_{\mathrm{P}}(\mathbf{x}) = g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))h(\mathbf{x}),$$

where h *does not depend on* P*.*

[3]Neyman and Fisher, however, essentially took this result as a definition of sufficiency.

Proof. (⇒) Assume that $\mathbf{T}$ is sufficient. The Halmos-Savage theorem then guarantees existence, for any $\mathrm{P} \in \mathcal{P}$, of a $\mathbf{T}$-measurable version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$, where P_* is an arbitrary privileged measure; denote it as $\mathbf{x} \mapsto g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))$. Noting that, for any P, we have $\mathrm{P} << \mathrm{P}_* << \mu$, let h be an arbitrary version of $\frac{\mathrm{dP}_*}{\mathrm{d}\mu}$. The elementary properties of Radon-Nikodym derivatives then ensure that

$$f_{\mathrm{P}}(\mathbf{x}) := g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))h(\mathbf{x})$$

is a version of $\frac{\mathrm{dP}}{\mathrm{d}\mu}$, as was to be proved. (⇐) Assume that, for any $\mathrm{P} \in \mathcal{P}$, there exist some g_{P} and h (which, without loss of generality, can be assumed to be nonnegative) such that

$$f_{\mathrm{P}}(\mathbf{x}) = g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))h(\mathbf{x}) \quad \mu\text{-a.e.}$$

Fix then an arbitrary privileged measure $\mathrm{P}_* = \sum_{i=1}^{\infty} c_i \mathrm{P}_i$ and note that

$$f_{\mathrm{P}_*} := \sum_{i=1}^{\infty} c_i f_{\mathrm{P}_i} \in \frac{\mathrm{dP}_*}{\mathrm{d}\mu};$$

indeed, we have

$$\mathrm{P}_*[A] = \sum_{i=1}^{\infty} c_i \mathrm{P}_i[A] = \sum_{i=1}^{\infty} c_i \int_A f_{\mathrm{P}_i}(\mathbf{x}) d\mu(\mathbf{x}) = \int_A f_{\mathrm{P}_*}(\mathbf{x}) d\mu(\mathbf{x}).$$

Since $\mathrm{P} << \mathrm{P}_* << \mu$, the elementary properties of Radon-Nikodym derivatives ensure that a version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$ is given by

$$\frac{f_{\mathrm{P}}(\mathbf{x})}{f_{\mathrm{P}_*}(\mathbf{x})} = \frac{f_{\mathrm{P}}(\mathbf{x})}{\sum_{i=1}^{\infty} c_i f_{\mathrm{P}_i}(\mathbf{x})} = \frac{g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))h(\mathbf{x})}{\sum_{i=1}^{\infty} c_i g_{\mathrm{P}_i}(\mathbf{T}(\mathbf{x}))h(\mathbf{x})} = \frac{g_{\mathrm{P}}(\mathbf{T}(\mathbf{x}))}{\sum_{i=1}^{\infty} c_i g_{\mathrm{P}_i}(\mathbf{T}(\mathbf{x}))}.$$

Since this version of $\frac{\mathrm{dP}}{\mathrm{dP}_*}$ is $\mathbf{T}$-measurable, sufficiency of $\mathbf{T}$ follows from the Halmos-Savage theorem. □

As an example, let $\mathbf{X} = (X_1, \ldots, X_n)$ collect independently and identically distributed random variables that admit density f with respect to the Lebesgue measure on $\mathbb{R}$. This is thus a nonparametric model involving the family $\mathcal{P} = \{\mathrm{P}_f : f \in \mathcal{F}\}$, where $\mathcal{F}$ is the collection of all densities with respect to the Lebesgue measure on the real line. The density of $\mathbf{X}$ (in $\mathbb{R}^n$, with respect to the Lebesgue measure on $\mathbb{R}^n$), is, at $\mathbf{x} = (x_1, \ldots, x_n)$,

$$f^{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \underbrace{\left(\prod_{i=1}^{n} f(x_{(i)}) \right)}_{g_f(\mathbf{x}_{(.)})} \times \underbrace{1}_{h(\mathbf{x})},$$

where $\mathbf{x}_{(.)} = (x_{(1)}, \ldots, x_{(n)})$ is the order statistic. The factorization criterion thus entails that $\mathbf{x}_{(.)}$ is a sufficient statistic.

2.7 Minimal sufficiency (in dominated models)

Let $\mathbf{S}$ and $\mathbf{T}$ be two statistics, with values in $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$ and $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$, respectively. We say that $\mathbf{T}$ is $\mathbf{S}$-measurable if and only if $\mathbf{T}$ is $\mathcal{A}_{\mathbf{S}}$-measurable, in the sense that $\mathcal{A}_{\mathbf{T}} := \mathbf{T}^{-1}(\mathcal{B}_{\mathcal{T}}) \subseteq \mathcal{A}_{\mathbf{S}}$. It can be shown that this happens if and only if there exists a measurable function ℓ from $\mathcal{S}$ to $\mathcal{T}$ such that $T(\mathbf{x}) = \ell(\mathbf{S}(\mathbf{x}))$, or if and only if $S(\mathbf{x}) = S(\mathbf{y})$ implies that $T(\mathbf{x}) = T(\mathbf{y})$. Obviously, if $\mathbf{T}$ is $\mathbf{S}$-measurable and $\mathbf{S}$ is $\mathbf{T}$-measurable, then $\mathcal{A}_{\mathbf{S}} = \mathcal{A}_{\mathbf{T}}$, $T(\mathbf{x}) = \ell(\mathbf{S}(\mathbf{x}))$ for a *one-to-one* mapping ℓ, and $S(\mathbf{x}) = S(\mathbf{y})$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$; in this framework, both statistics provide the exact same reduction of information.

In the sequel, we assume that $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is a dominated model. If $\mathbf{T}$ is sufficient and $\mathbf{S}$-measurable (that is, if $\mathcal{A}_{\mathbf{T}} \subseteq \mathcal{A}_{\mathbf{S}}$), then $\mathbf{S}$ is also sufficient. Intuitively, if $\mathbf{T}$ is a function of $\mathbf{S}$, then all information carried by $\mathbf{T}$ is also carried by $\mathbf{S}$, whereas, mathematically, this readily follows from the Halmos-Savage theorem (since $\mathbf{T}$-measurability implies $\mathbf{S}$-measurability). Thus, many sufficient statistics may be available for a given model.

Suppose, for example, that $\mathbf{X} = (X_1, \ldots, X_n)$ collects independently and identically distributed $\mathcal{N}(0, \sigma^2)$ variables, and consider the resulting model parametrized by $\sigma^2 \in \mathbb{R}_0^+$. Then, the factorization criterion easily yields that the statistics

$\mathbf{T}_1(\mathbf{X}) = (X_1, \ldots, X_n)$

$\mathbf{T}_2(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$ (the order statistic)

$\mathbf{T}_3(\mathbf{X}) = (X_{(1)}^2, \ldots, X_{(n)}^2)$

$$\mathbf{T}_4(\mathbf{X}) = (X_{(1)}^2 + X_{(2)}^2, X_{(3)}^2 + \ldots + X_{(n)}^2)$$

$$\mathbf{T}_5(\mathbf{X}) = X_1^2 + \ldots + X_n^2$$

are all sufficient, with $\mathcal{A}_{\mathbf{T}_5} \subseteq \ldots \subseteq \mathcal{A}_{\mathbf{T}_1} \subseteq \mathcal{A}$. The smaller $\mathcal{A}_{\mathbf{T}}$, the larger the reduction associated with $\mathbf{T}$, and the simpler the model induced by $\mathbf{T}$: in this respect, $\mathbf{T}_5$ does a better job than $\mathbf{T}_4$, and a much better one than $\mathbf{T}_1$, which is trivially sufficient (no reduction at all). As we will be show later, $\mathbf{T}_5$ actually is *minimal sufficient*, in the sense that no further reduction is possible without losing sufficiency.

Definition 2. A statistic $\mathbf{T}$ is *minimal sufficient* (equivalently, the σ-field $\mathcal{A}_{\mathbf{T}}$ is *minimal sufficient*) if it is sufficient and if it is $\mathbf{S}$-measurable for any sufficient statistic $\mathbf{S}$ (equivalently, if $\mathcal{A}_{\mathbf{T}}$ is sufficient and if $\mathcal{A}_{\mathbf{T}} = \bigcap_{\mathbf{S} \text{ sufficient}} \mathcal{A}_{\mathbf{S}}$).

As an example, let $\mathbf{X} = (X_1, \ldots, X_n)$ collect independent and identically distributed random variables whose common distribution is the uniform distribution over the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. Denote by $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \mathbb{R}\}$ the family of joint distributions of such $\mathbf{X}$'s. Writing $\mathbb{I}[C]$ for the indicator function of Condition C (which takes value one if C is satisfied and value zero otherwise), the density of P_θ with respect to the Lebesgue measure in $\mathbb{R}^n$, at $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, is then

$$f_\theta(\mathbf{x}) = f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n \mathbb{I}\left[\theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2}\right]$$

$$= \mathbb{I}\left[\theta - \frac{1}{2} \leq x_{(1)}, x_{(n)} \leq \theta + \frac{1}{2}\right] = \mathbb{I}\left[x_{(n)} - \frac{1}{2} \leq \theta \leq x_{(1)} + \frac{1}{2}\right].$$

The factorization criterion thus implies that $\mathbf{T} := (X_{(1)}, X_{(n)})$ is sufficient. In order to establish minimal sufficiency, let $\mathbf{S}$ be sufficient. From the factorization criterion, we have that, for all $\theta \in \mathbb{R}$, the density f_θ factorizes into

$$f_\theta(\mathbf{x}) = g_\theta(\mathbf{S}(\mathbf{x}))h(\mathbf{x}) \qquad \mathrm{P}_\theta\text{-a.s.}$$

Now, note that $h(\mathbf{X}) > 0$ P_θ-a.s. for all $\theta \in \mathbb{R}$. Therefore, P_θ-a.s. for all $\theta \in \mathbb{R}$,

$$\begin{aligned} X_{(1)} &= \inf\big\{t \in \mathbb{R} : f_\theta(\mathbf{X}) = 0 \text{ for all } \theta \in (t, \infty)\big\} - \frac{1}{2} \\ &= \inf\big\{t \in \mathbb{R} : g_\theta(\mathbf{S}(\mathbf{X})) = 0 \text{ for all } \theta \in (t, \infty)\big\} - \frac{1}{2} \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} X_{(n)} &= \sup\big\{t \in \mathbb{R} : f_\theta(\mathbf{X}) = 0 \text{ for all } \theta \in (-\infty, t)\big\} + \frac{1}{2} \\ &= \sup\big\{t \in \mathbb{R} : g_\theta(\mathbf{S}(\mathbf{X})) = 0 \text{ for all } \theta \in (-\infty, t)\big\} + \frac{1}{2}\cdot \end{aligned} \tag{2.12}$$

It follows from (2.11)–(2.12) that $\mathbf{T}$ is $\mathbf{S}$-measurable, hence is minimal sufficient.

It remains rare that we can establish minimal sufficiency by using Definition 2 as we could do in the example above. We now present two results that together allow one to establish minimal sufficiency in many cases.

Proposition 2. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ and $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be two dominated models involving the same observation space $\mathcal{X}$, with $\mathcal{P}_0 \subset \mathcal{P}$. If $\mathbf{T}$ is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ and sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, then $\mathbf{T}$ is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.*

Proof. Let $\mathbf{S}$ be a sufficient statistic for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Then, $\mathbf{S}$ is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ (this follows, e.g., from the Halmos-Savage theorem). Since $\mathbf{T}$ is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$, we thus have, by definition, that $\mathbf{T}$ is $\mathbf{S}$-measurable, which was to be shown. □

Proposition 3. *Let $\mathcal{P} = \{\mathrm{P}_0, \mathrm{P}_1, \ldots, \mathrm{P}_K\}$ and assume that $\mathrm{P}_k << \mathrm{P}_0$ for $k = 1, \ldots, K$. Then, $\mathbf{T} := (T_1, \ldots, T_K)$, with $T_k := \frac{d\mathrm{P}_k}{d\mathrm{P}_0}$, is minimal sufficient.*

Proof. Obviously, the family $\mathcal{P}$ is dominated by P_0, and $\frac{d\mathrm{P}_0}{d\mathrm{P}_0} = 1$. It directly follows from the Halmos-Savage theorem (applied with $\mathrm{P}_* = \mathrm{P}_0$) that $\mathbf{T} := (T_1, \ldots, T_K)$ is sufficient. Let then $\mathbf{S}$ be an arbitrary sufficient statistic. From the Halmos-Savage theorem (still applied with $\mathrm{P}_* = \mathrm{P}_0$), there must exist, for any $k = 1, \ldots, K$, a function ℓ_k such that

$$\frac{\mathrm{dP}_k}{\mathrm{dP}_0} = \ell_k(\mathbf{S}).$$

This shows that $\mathbf{T}$ is $\mathbf{S}$-measurable, hence minimal sufficient. □

Let us provide some applications of Propositions 2–3.

Example 1: Let $\mathbf{X} = (X_1, \ldots, X_n)$ collect independent and identically distributed $\mathcal{N}(\mu, 1)$ random variables, with $\mu \in \mathbb{R}$. The density of $\mathbf{X}$ with respect to the Lebesgue measure on $\mathbb{R}^n$ is, at $\mathbf{x} = (x_1, \ldots, x_n)$,

$$\begin{aligned} f_\mu(x_1, \ldots, x_n) &= (2\pi)^{-n/2} \exp\Big(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \Big) \\ &= \underbrace{\exp\Big(\mu \sum_{i=1}^n x_i - \frac{n}{2}\mu^2 \Big)}_{g_\mu(\sum_{i=1}^n x_i)} \times \underbrace{(2\pi)^{-n/2} \exp\Big(-\frac{1}{2} \sum_{i=1}^n x_i^2 \Big)}_{h(\mathbf{x})}. \end{aligned}$$

The factorization criterion thus implies that $\sum_{i=1}^n X_i$ is a sufficient statistic. Now, denote as $\mathcal{P}$ the family of all $\mathcal{N}(\mu, 1)$ distributions associated with $\mu \in \mathbb{R}$, and by $\mathcal{P}_0$ a subfamily consisting of the $\mathcal{N}(\mu_0, 1)$ and $\mathcal{N}(\mu_1, 1)$ distributions associated with two arbitrary values $\mu_0 \neq \mu_1$. In view of Proposition 3,

$$\begin{aligned} T &:= \frac{f_{\mu_1}(x_1, \ldots, x_n)}{f_{\mu_0}(x_1, \ldots, x_n)} \\ &= \frac{\exp\big(\mu_1 \sum_{i=1}^n x_i - \frac{n}{2}\mu_1^2\big) (2\pi)^{-n/2} \exp\big(-\frac{1}{2}\sum_{i=1}^n x_i^2\big)}{\exp\big(\mu_0 \sum_{i=1}^n x_i - \frac{n}{2}\mu_0^2\big) (2\pi)^{-n/2} \exp\big(-\frac{1}{2}\sum_{i=1}^n x_i^2\big)} \\ &= \exp\Big((\mu_1 - \mu_0) \sum_{i=1}^n x_i + \frac{n}{2}(\mu_0^2 - \mu_1^2) \Big) \end{aligned}$$

is minimal sufficient for $\mathcal{P}_0$. Since $\sum_{i=1}^n X_i$ generate the same σ-field as T, it is also minimal sufficient for $\mathcal{P}_0$, hence (from Proposition 2) minimal sufficient for $\mathcal{P}$. Clearly,

$\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$, which generates the same σ-field as $\sum_{i=1}^n X_i$, is then also minimal sufficient for $\mathcal{P}$.

Example 2: Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an n-tuple of independent and identically distributed random variables, being logistic with location θ. More precisely, each X_i has density

$$f_\theta(x) = \frac{\exp\left(-(x-\theta)\right)}{\left\{1+\exp\left(-(x-\theta)\right)\right\}^2}, \qquad x \in \mathbb{R}.$$

Then, for the finite subfamily $\mathcal{P}_0$ corresponding to the $(K+1)$-tuple of pairwise distinct parameter values $\{\theta_0 = 0, \theta_1, \ldots, \theta_K\}$, a minimal sufficient statistic is, in view of Proposition 3,

$$\mathbf{T} = (T_1, \ldots, T_K), \quad \text{with } T_j := \exp(n\theta_j)\prod_{i=1}^n \left(\frac{1+\exp(-X_i)}{1+\exp(-X_i+\theta_j)}\right)^2.$$

Let us show that, for $K = n+1$, $\mathbf{T}(x_1, \ldots, x_n) = \mathbf{T}(y_1, \ldots, y_n)$ if and only if $(\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}) = (\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(n)})$. This would then imply that $\mathbf{T}$ generates the same σ-field as the order statistic $(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(n)})$ of $\mathbf{X}$, so that *the order statistic would then also be minimal sufficient for* $\mathcal{P}_0$. Since the factorization criterion implies that the order statistic is sufficient for the whole family $\mathcal{P}$ obtained for $\theta \in \mathbb{R}$, it would then be minimal sufficient for $\mathcal{P}$, too (from Proposition 2).

If $(\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}) = (\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(n)})$, then obviously $\mathbf{T}(x_1, \ldots, x_n) = \mathbf{T}(y_1, \ldots, y_n)$. Assume then that $\mathbf{T}(x_1, \ldots, x_n) = \mathbf{T}(y_1, \ldots, y_n)$. Let $\xi_j = \exp(\theta_j)$ and $u_i = \exp(-x_i)$, $v_i = \exp(-y_i)$. Since $\mathbf{T}(x_1, \ldots, x_n) = \mathbf{T}(y_1, \ldots, y_n)$, we have

$$\xi_j^n \prod_{i=1}^n \left(\frac{1+u_i}{1+\xi u_i}\right)^2 = \xi_j^n \prod_{i=1}^n \left(\frac{1+v_i}{1+\xi v_i}\right)^2 \quad \text{for } \xi = \xi_1, \ldots, \xi_{n+1}$$

(recall we took $K = n+1$), hence also

$$p(\xi) := \prod_{i=1}^n \frac{1+\xi u_i}{1+u_i} = \prod_{i=1}^n \frac{1+\xi v_i}{1+v_i} =: q(\xi) \quad \text{for } \xi = \xi_1, \ldots, \xi_{n+1}.$$

This last equation requires that two polynomials of degree n in ξ, namely $p(\xi)$ and $q(\xi)$, be equal at $n+1$ distinct values of ξ. This implies that these polynomials are identical, hence that they share the same roots. Since the roots of $p(\xi)$ are $-1/u_1, \ldots, -1/u_n$ and those of $q(\xi)$ are $-1/v_1, \ldots, -1/v_n$, it follows that $u_{(i)} = v_{(i)}$ for all $i = 1, \ldots, n$, hence that $x_{(i)} = y_{(i)}$ for all $i = 1, \ldots, n$, as was to be proved.

Example 3: Semiparametric location model: $X_1, \ldots, X_n$ are independently and identically distributed with density f_θ (with respect to the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$), with $f_\theta(x) = f_0(x-\theta)$ and $f_0 \in \mathcal{F}_0 := \left\{ f(x) : \int x f(x) \mathrm{d}\mu(x) = 0 \right\}$. That class $\mathcal{F}_0$ contains the centered logistic, so that the logistic family of Example 2 is a subfamily $\mathcal{P}_0$ of $\mathcal{P}$. Clearly, the order statistic is sufficient for $\mathcal{P}$ (this readily follows from the factorization criterion), while we have shown it is minimal sufficient for $\mathcal{P}_0$. Hence, the order statistic is minimal sufficient for $\mathcal{P}$.