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

3 Sufficiency and point estimation

3.1 The Rao-Blackwell theorem

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model, and let $\mathbf{T} = \mathbf{T}(\mathbf{X})$ be a sufficient statistic for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Consider a decision space $(\mathcal{D}, \mathcal{B}_{\mathcal{D}})$, where $\mathcal{D}$ *is a convex Borel set of* $\mathbb{R}^k$, equipped with the Borel sub-σ-field $\mathcal{B}_{\mathcal{D}} = \mathcal{B}^k \bigcap \mathcal{D}$. In this context, a pure decision rule $\delta : (\mathcal{X}, \mathcal{A}) \longrightarrow (\mathcal{D}, \mathcal{B}_{\mathcal{D}})$ is a statistic with values in $(\mathcal{D}, \mathcal{B}_{\mathcal{D}})$. The typical example is point estimation for the value (at $\mathrm{P} \in \mathcal{P}$) of a real-valued function $\varphi(\mathrm{P})$: then, the decision space is $\mathcal{D} = \varphi(\mathcal{P})$. We henceforth use the terminology associated with point estimation.

An estimator δ is called $\mathcal{P}$*-integrable* if $\mathrm{E}_{\mathrm{P}}[\delta(\mathbf{X})]$ exists and is finite for all $\mathrm{P} \in \mathcal{P}$ (convexity of $\mathcal{D}$ implies that $\mathrm{E}_{\mathrm{P}}[\delta(\mathbf{X})] \in \mathcal{D}$ for all P). If $\mathrm{E}_{\mathrm{P}}[\delta(\mathbf{X})] = \varphi(\mathrm{P})$ for all P, then we say that δ is an *unbiased* estimator of $\varphi(\mathrm{P})$. Define the *Rao-Blackwellization of* δ with respect to $\mathbf{T}$ as

$$\delta^{\mathbf{T}} := \mathrm{E}_{\not{\mathrm{P}}}\left[\delta(\mathbf{X})|\mathbf{T}\right],$$

where independence on P follows from sufficiency of $\mathbf{T}$ (sufficiency indeed implies that there exists a version of this conditional expectation that does not depend on P[2]). Thus, $\delta^{\mathbf{T}}$ is still a statistic, that, from convexity, takes its values in $\mathcal{D}$. Note that if δ is an unbiased estimator of $\varphi(\mathrm{P})$, then $\delta^{\mathbf{T}}$ is an unbiased estimator of $\varphi(\mathrm{P})$, too, as

$$\mathrm{E}_{\mathrm{P}}\left[\delta^{\mathbf{T}}(\mathbf{X})\right] = \mathrm{E}_{\mathrm{P}}[\mathrm{E}_{\mathrm{P}}\left[\delta(\mathbf{X})|\mathbf{T}\right]] = \mathrm{E}_{\mathrm{P}}[\delta(\mathbf{X})] = \psi(\mathrm{P})$$

for all $\mathrm{P} \in \mathcal{P}$. Now, denote by $\mathrm{R}^{\delta}_{\mathrm{P}}$ and $\mathrm{R}^{\delta^{\mathbf{T}}}_{\mathrm{P}}$ the risks at P associated with δ and $\delta^{\mathbf{T}}$, respectively, for a given loss function $(\mathrm{P}, d) \longmapsto \mathrm{L}_{\mathrm{P}}(d)$.

Theorem 1. (Rao-Blackwell). *Let* $\mathbf{T}$ *be sufficient for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. *Consider a loss function* $(\mathrm{P}, d) \longmapsto \mathrm{L}_{\mathrm{P}}(d)$ *such that* $d \longmapsto \mathrm{L}_{\mathrm{P}}(d)$ *is strictly* convex *for all* $\mathrm{P} \in \mathcal{P}$. *Then, for any* $\mathrm{P} \in \mathcal{P}$ *at which the risk* $\mathrm{R}^{\delta}_{\mathrm{P}}$ *exists and is finite,* $\mathrm{R}^{\delta}_{\mathrm{P}} > \mathrm{R}^{\delta^{\mathbf{T}}}_{\mathrm{P}}$ *unless* $\delta = \delta^{\mathbf{T}}$ P*-almost surely (in which case one obviously has* $\mathrm{R}^{\delta}_{\mathrm{P}} = \mathrm{R}^{\delta^{\mathbf{T}}}_{\mathrm{P}}$*).*

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

[2] This was actually established in the proof of the Halmos-Savage Theorem.

Proof. The proof directly follows from an application of *Jensen's inequality.* Recall that, for any real-valued convex function g, any $\mathcal{A}$-measurable random vector $\boldsymbol{\xi}$ and any sub-σ-field $\mathcal{A}_0$ of $\mathcal{A}$,

$$\mathrm{E}[g(\boldsymbol{\xi})|\mathcal{A}_0] \geq g(\mathrm{E}[\boldsymbol{\xi}|\mathcal{A}_0]) \tag{3.1}$$

provided that $\mathrm{E}[g(\boldsymbol{\xi})]$ exists and is finite; if the function g is strictly convex and $\boldsymbol{\xi}$ is not $\mathcal{A}_0$-measurable, then inequality is strict. Here, for all $\mathrm{P} \in \mathcal{P}$, Jensen's inequality yields

$$\begin{aligned} \mathrm{R}_{\mathrm{P}}^{\delta^{\mathbf{T}}} &= \mathrm{E}_{\mathrm{P}}\left[\mathrm{L}_{\mathrm{P}}(\delta^{\mathbf{T}})\right] = \mathrm{E}_{\mathrm{P}}\left[\mathrm{L}_{\mathrm{P}}\left(\mathrm{E}_{\not{\mathrm{P}}}[\delta(\mathbf{X})|\mathbf{T}]\right)\right] \\ &\leq \mathrm{E}_{\mathrm{P}}\left[\mathrm{E}_{\not{\mathrm{P}}}[\mathrm{L}_{\mathrm{P}}(\delta(\mathbf{X}))\,|\mathbf{T}]\right] = \mathrm{E}_{\mathrm{P}}[\mathrm{L}_{\mathrm{P}}(\delta(\mathbf{X}))] = \mathrm{R}_{\mathrm{P}}^{\delta}, \end{aligned} \tag{3.2}$$

where the inequality in (3.2) follows from (3.1), and is strict unless δ is $\mathbf{T}$-measurable. □

Conditioning with respect to a sufficient statistic thus uniformly improves any estimator δ which is not $\mathbf{T}$-measurable. If the loss function in convex, but not strictly convex, then the inequality may become weak, $\mathrm{R}_{\mathrm{P}}^{\delta} \geq \mathrm{R}_{\mathrm{P}}^{\delta^{\mathbf{T}}}$, even for a non $\mathbf{T}$-measurable δ. In all cases, thus, the $\mathbf{T}$-measurable estimator $\delta^{\mathbf{T}}$ is uniformly preferable to δ, irrespective of the convex loss function considered.

Example 1: Let $X_1, \ldots, X_n$ be a sample of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu \in \mathbb{R}$ but otherwise unspecified Lebesgue density f. The weighted mean $\bar{X}_{\mathbf{w}} := \sum_{i=1}^n w_i X_i$, with $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = 1$, is an unbiased estimator of μ. The order statistic $\mathbf{X}_{(\cdot)} := \left(X_{(1)}, \ldots, X_{(n)}\right)$ is sufficient. In view of the linearity properties of conditional expectations, the Rao-Blackwellization $\bar{X}_{\mathbf{w}}^{\mathbf{X}_{(\cdot)}}$ of $\bar{X}_{\mathbf{w}}$ is $\bar{X}_{\mathbf{w}}^{\mathbf{X}_{(\cdot)}} = \sum_{i=1}^n w_i \mathrm{E}[X_i|\mathbf{X}_{(\cdot)}]$. The distribution of $\mathbf{X}$ conditional on $\mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})$ is uniform over the $n!$ permutations of $\mathbf{x}_{(\cdot)}$. Hence, the distribution of X_i conditional on $\mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}$ is uniform over $x_{(1)}, \ldots, x_{(n)}$, so that

$$\mathrm{E}[X_i|\mathbf{X}_{(\cdot)}] = \frac{1}{n}\sum_{i=1}^n X_{(i)} = \frac{1}{n}\sum_{i=1}^n X_i =: \bar{X}.$$

Therefore, Rao-Blackwellizing $\bar{X}_{\mathbf{w}}$ yields

$$\bar{X}_{\mathbf{w}}^{\mathbf{X}_{(\cdot)}} = \sum_{i=1}^n w_i \mathrm{E}[X_i|\mathbf{X}_{(\cdot)}] = \sum_{i=1}^n w_i \bar{X} = \bar{X}.$$

Unweighted means, thus, in this respect are preferable to weighted means (which is quite plausible in view of the symmetry of the problem). Note that, in particular, Rao-Blackwellization of X_1 provides $\bar{X}$ (this is the particular case associated with $\mathbf{w} = (1, 0, \ldots, 0)$).

Example 2: 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 $[0, \theta]$. Denote by $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \mathbb{R}\}$ the family of joint distributions of such $\mathbf{X}$'s. We consider the problem of estimating $\varphi(P) = \theta$. 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

$$f_\theta(\mathbf{x}) = \prod_{i=1}^{n} \mathbb{I}\big[0 \leq x_i \leq \theta\big] = \mathbb{I}\big[0 \leq x_{(1)}, x_{(n)} \leq \theta\big] = \mathbb{I}\big[0 \leq x_{(1)}\big]\mathbb{I}\big[x_{(n)} \leq \theta\big].$$

The factorization criterion thus implies that $\mathbf{T} := X_{(n)}$ is sufficient. An unbiased estimator for θ is $\delta = 2X_1$. Its Rao-Blackwellization using $\mathbf{T} = X_{(n)}$ is given by

$$\delta^{X_{(n)}} = 2\mathrm{E}[X_1 | X_{(n)}] = 2\bigg\{ X_{(n)} \times \frac{1}{n} + \mathrm{E}[Z | X_{(n)}] \times \bigg(1 - \frac{1}{n}\bigg)\bigg\},$$

where Z, conditional on $X_{(n)}$, is uniformly distributed over $[0, X_{(n)}]$. Therefore,

$$\delta^{X_{(n)}} = 2\bigg\{ \frac{X_{(n)}}{n} + \frac{X_{(n)}}{2}\bigg(1 - \frac{1}{n}\bigg)\bigg\} = \frac{n+1}{n} X_{(n)}.$$

It is easy to check explicitly that this is indeed an unbiased estimator of θ.

3.2 Distribution-freeness and ancillarity

We now turn to the concept of *distribution-freeness*, which, in a sense, is exactly the opposite of sufficiency: whereas a sufficient statistic carries all the available information, a distribution-free statistic does not carry any information at all. As we shall see, however, things are more subtle.

As usual, let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model.

Definition 1. *A statistic* $\mathbf{S}$ *is* distribution-free *(under* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *or under* $\mathcal{P}$*) if its distribution is the same under any* $\mathrm{P} \in \mathcal{P}$*, that is, if* $\mathrm{P}_1^{\mathbf{S}} = \mathrm{P}_2^{\mathbf{S}}$ *for all* $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{P}$.

Note that distribution-freeness, like sufficiency, is a property of σ-fields: if $\mathbf{S}$ is distribution-free, so is any $\mathcal{A}_{\mathbf{S}}$-measurable statistic. Hence, we may also speak of distribution-free σ-fields in the sequel.

Definition 2. *A distribution-free statistic* $\mathbf{S}$ *measurable with respect to* any *sufficient statistic is called* ancillary *for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

Example 3: Let $\mathbf{X} = (N, Y)$, where $N - 1 \sim \text{Bin}(1, \frac{1}{2})$ and, conditionally on $N = n$, $Y \sim \text{Bin}(n, p)$, with $p \in (0, 1)$, which characterizes the distribution $\mathrm{P}_p^{\mathbf{X}}$ of $\mathbf{X}$. Denoting as μ the counting measure of $\{(1,0),\ (1,1),\ (2,0),\ (2,1),\ (2,2)\}$, the family $\mathcal{P} = \{\mathrm{P}_p : p \in (0,1)\}$ is a one-parameter family dominated by μ; it is easy to check that the density of P_p with respect to μ is, at $\mathbf{x} = (n, y)$,

$$f_p(\mathbf{x}) = \frac{1}{2}\binom{n}{y} p^y (1-p)^{n-y}.$$

One can show that $\mathbf{X} = (N, Y)$ is minimal sufficient, while $N \sim \text{Bin}(1, \frac{1}{2})$ is obviously distribution-free, and $\mathbf{X}$-measurable. Hence, N is ancillary. Although N does not carry any information on p, it is needed in interpreting the information contained in Y.

Example 4: In the logistic location family (Example 2 of Chapter 2), the order statistic $\mathbf{X}_{(\cdot)}$ is minimal sufficient. Since the *spacings* $X_{(i+1)} - X_{(i)}$ are distribution-free (indeed, $X_{(i+1)} - X_{(i)} = (X_{(i+1)} - \theta) - (X_{(i)} - \theta)$) and $\mathbf{X}_{(\cdot)}$-measurable, each of them is *ancillary*.

The *principle of ancillarity* consists in getting rid of ancillary statistics/σ-fields. This principle, in the presence of $\mathbf{S}$, distribution-free and measurable with respect to a minimal sufficient statistic $\mathbf{T}$, consists in reducing the original model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ by conditioning on $\mathbf{S}$, yielding $\left(\mathcal{T}, \mathcal{B}_{\mathcal{T}}, \mathcal{P}^{\mathbf{T}|\mathbf{S}=\mathbf{s}}\right)$, where $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ is $\mathbf{T}$'s observation space, and $\mathcal{P}^{\mathbf{T}|\mathbf{S}=\mathbf{s}}$ is the collection of conditional distributions of $\mathbf{T}$ conditional on $\mathbf{S} = \mathbf{s}$ (provided that such conditional distributions exist).

Example 3 (continued): Observe $N = n$, then treat it as a constant, with a model which is either binomial with exponent 1, or binomial with exponent 2.

3.3 Completeness and the Lehmann-Scheffé theorem

Reduction of the data via sufficiency is most effective when there is no ancillary statistic except for the trivial case—the almost sure constants. Characterizing such an absence is difficult, and an even stronger requirement, that of the absence of *first-order ancillary* statistics, is considered, leading to the concept of *completeness*.

Definition 3. *A statistic* $\mathbf{S}$ *is* first-order distribution-free *under* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *if and only if (i) it is*

$\mathcal{P}$-integrable (that is, $\mathrm{E}_{\mathrm{P}}[\mathbf{S}]$ *exists and is finite for any* $\mathrm{P} \in \mathcal{P}$*) and (ii)* $\mathrm{E}_{\mathrm{P}}[\mathbf{S}] = \mathbf{c}_{\not{\mathrm{P}}}$*, a constant that does not depend on* $\mathrm{P} \in \mathcal{P}$*.*

A statistic $\mathbf{S}$ is said to be *a $\mathcal{P}$-almost sure constant* if there exists a constant $\mathbf{c}$ and, for all $\mathrm{P} \in \mathcal{P}$, a set $N_{\mathrm{P}} \in \mathcal{A}$ with P-probability zero such that $\mathbf{S}(\mathbf{x}) = \mathbf{c}$ for all $\mathbf{x} \notin N_{\mathrm{P}}$. The $\mathcal{P}$-almost sure constants are trivially first-order distribution-free. A statistic $\mathbf{T}$ is called *complete* if the only $\mathbf{T}$-measurable first-order distribution-free statistics are those trivial ones.

Definition 4. *A statistic* $\mathbf{T}$ *(or the corresponding σ-field $\mathcal{A}_{\mathbf{T}}$) is* complete *for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *if and only if the fact that* $\mathrm{E}_{\mathrm{P}}[\ell(\mathbf{T})] = \mathbf{0}$ *for all* $\mathrm{P} \in \mathcal{P}$ *implies that* $\ell(\mathbf{T}) = \mathbf{0}$ *P-almost surely for all* $\mathrm{P} \in \mathcal{P}$*.*

It is easy to show that a statistic $\mathbf{T}$ is complete if and only if, indeed, the $\mathcal{P}$-almost sure constants are the only $\mathbf{T}$-measurable first-order distribution-free statistics, in the sense that the fact that $\mathrm{E}_{\mathrm{P}}[\ell(\mathbf{T})] = \mathbf{c}_{\not{\mathrm{P}}}$ for all $\mathrm{P} \in \mathcal{P}$ implies that $\ell(\mathbf{T}) = \mathbf{c}_{\not{\mathrm{P}}}$ P-almost surely for all $\mathrm{P} \in \mathcal{P}$.

<u>Example 2 (continued)</u>: 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 $[0, \theta]$. Denote by $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \mathbb{R}\}$ the family of joint distributions of such $\mathbf{X}$'s. For $n \geq 2$, the statistic $\mathbf{T} = \mathbf{X}$ is not complete because

$$\ell(\mathbf{T}) = X_2 - X_1$$

is $\mathbf{T}$-measurable, is not a $\mathcal{P}$-almost sure constant, yet provides $\mathrm{E}_\theta[\ell(\mathbf{T})] = 0$ for all $\theta > 0$ (we write E_θ instead of $\mathrm{E}_{\mathrm{P}_\theta}$). Similarly, for $n \geq 2$, the order statistic $\mathbf{T} = \mathbf{X}_{(\cdot)}$ is not complete because

$$\ell(\mathbf{T}) = \frac{n+1}{n} X_{(n)} - (n+1) X_{(1)}$$

is $\mathbf{T}$-measurable, is not a $\mathcal{P}$-almost sure constant, yet provides $\mathrm{E}_\theta[\ell(\mathbf{T})] = 0$ for all $\theta > 0$. For the same reason, the statistic $\mathbf{T} = (X_{(1)}, X_{(n)})$ is not complete either for $n \geq 2$. In contrast, the statistic $T = X_{(n)}$ is complete, as we now show. Assume that there exists ℓ such that $\mathrm{E}_\theta[\ell(T)] = 0$ for all $\theta > 0$, that is, such that, for all $\theta > 0$,

$$0 = \mathrm{E}_\theta[\ell(T)] = \mathrm{E}_\theta[\ell(X_{(n)})] = \int_{-\infty}^{\infty} \ell(z) f_\theta^{X_{(n)}}(z)\, dz = \frac{n}{\theta^n} \int_0^\theta \ell(z) z^{n-1}\, dz.$$

Thus,

$$\int_0^\theta \ell(z) z^{n-1}\, dz = 0$$

for all $\theta > 0$, so that we must have that $\ell(z)z^{n-1} = 0$ μ_+-almost everywhere (where μ_+ is the Lebesgue measure on $\mathbb{R}^+$), hence also that $\ell(z) = 0$ μ_+-almost everywhere. Since $X_{(n)}$ is nonnegative P_θ-almost surely for any $\theta > 0$, it follows that $\ell(X_{(n)}) = 0$ P_θ-almost surely for any $\theta > 0$. This establishes that $T = X_{(n)}$ is complete.

As we shall see, completeness is a property that nicely complements sufficiency. The example above suggests that a sufficient statistic may be complete only if it provides performs a large/maximal reduction of $\mathbf{X}$ still ensuring sufficiency. The next result supports this.

Theorem 2. *Let* $\mathbf{T}$ *be sufficient and complete for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. *Then, provided that a minimal sufficient statistic exists,* $\mathbf{T}$ *is minimal sufficient.*

Proof. Let $\mathbf{T}_*$ be a minimal sufficient statistic. It is enough to prove that $\mathbf{T}$ is $\mathbf{T}_*$-measurable (since the sufficient statistic $\mathbf{T}$ is then measurable with respect to any sufficient statistic, hence is a minimal sufficient statistic). To this end, consider the statistic

$$\mathbf{V} := \mathbf{T} - \mathrm{E}_{\mathrm{P}}[\mathbf{T}|\mathbf{T}_*]$$

(sufficiency of $\mathbf{T}_*$ implies that this is indeed a statistic). Clearly, it has expectation zero under any $\mathrm{P} \in \mathcal{P}$. Also, $\mathbf{V}$ is $\mathbf{T}$-measurable (since $\mathbf{T}_*$ is minimal sufficient, the $\mathbf{T}_*$-measurable statistic $\mathrm{E}_{\mathrm{P}}[\mathbf{T}|\mathbf{T}_*]$ is also $\mathbf{T}$-measurable). Completeness of $\mathbf{T}$ thus entails that $\mathbf{V} = 0$ P-almost surely under any $\mathrm{P} \in \mathcal{P}$, hence that

$$\mathbf{T} = \mathrm{E}_{\mathrm{P}}[\mathbf{T}|\mathbf{T}_*]$$

P-almost surely under any $\mathrm{P} \in \mathcal{P}$. It follows that $\mathbf{T}$ is $\mathbf{T}_*$-measurable, which establishes the result. □

Two remarks are in order. First, it is not always so that minimal sufficient statistic exists (although existence actually holds under extremely mild assumptions). Second, a minimal sufficient statistic may fail to be complete; for instance, we saw in Chapter 2 that when one observes a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ collecting i.i.d. realizations from the $\mathrm{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ distribution, then $\mathbf{T} = (X_{(1)}, X_{(n)})$ is minimal sufficient; however, it is not complete since $\mathrm{E}_\theta[X_{(n)} - X_{(1)} - \frac{n-1}{n+1}] = 0$ for any $\theta \in \mathbb{R}$.

The following result provides another interesting property of sufficient and complete statistics. Recall that a statistic $\mathbf{S}$, with values in $(\mathcal{S}, \mathcal{B}_\mathcal{S})$, and a statistic $\mathbf{T}$, with values in $(\mathcal{T}, \mathcal{B}_\mathcal{T})$, are P-independent if

$$\text{for all } B \in \mathcal{B}_\mathcal{S}, \quad \mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}] = \mathrm{P}[\mathbf{S}^{-1}(B)] \quad \text{P-almost surely,}$$

or, equivalently, if for all $B \in \mathcal{B}_\mathcal{T}$, $\mathrm{P}[\mathbf{T}^{-1}(B)|\mathbf{S}] = \mathrm{P}[\mathbf{T}^{-1}(B)]$ P-almost surely.

Theorem 3. (Basu, 1955) *If (i)* $\mathbf{T}$ *is sufficient and complete for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *and (ii)* $\mathbf{S}$ *is distribution-free, then* $\mathbf{S}$ *and* $\mathbf{T}$ *are* P*-independent for any* $\mathrm{P} \in \mathcal{P}$*.*

Proof. Fix $B \in \mathcal{B}_\mathcal{S}$ arbitrarily. Since

$$\mathrm{E}_\mathrm{P}\big[\mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}]\big] = \mathrm{P}[\mathbf{S}^{-1}(B)] \quad \text{for all } \mathrm{P} \in \mathcal{P},$$

one has

$$\mathrm{E}_\mathrm{P}\big[\mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}] - \mathrm{P}[\mathbf{S}^{-1}(B)]\big] = 0 \quad \text{for all } \mathrm{P} \in \mathcal{P}. \tag{3.3}$$

Note that $\mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}]$ does not depend on P (since $\mathbf{T}$ is sufficient) and that this is also the case for $\mathrm{P}[\mathbf{S}^{-1}(B)]$ (since $\mathrm{P}[\mathbf{S}^{-1}(B)] = \mathrm{P}^\mathbf{S}[B]$ and $\mathbf{S}$ is distribution-free). Thus,

$$\mathrm{P}\left[\mathbf{S}^{-1}(B)|\mathbf{T}\right] - \mathrm{P}[\mathbf{S}^{-1}(B)]$$

is a $\mathbf{T}$-measurable *statistic*. In view of (3.3), that $\mathbf{T}$-measurable statistic has expectation zero for all $\mathrm{P} \in \mathcal{P}$. Since $\mathbf{T}$ is complete, we must then have

$$\mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}] - \mathrm{P}[\mathbf{S}^{-1}(B)] = 0 \quad \text{P-almost surely}$$

for any $\mathrm{P} \in \mathcal{P}$. Since $B \in \mathcal{B}_\mathcal{S}$ was fixed arbitrarily, we thus proved that

$$\mathrm{P}[\mathbf{S}^{-1}(B)|\mathbf{T}] = \mathrm{P}[\mathbf{S}^{-1}(B)] \quad \text{P-almost surely}$$

for any $\mathrm{P} \in \mathcal{P}$ and any $B \in \mathcal{B}_\mathcal{S}$, which establishes the result. □

Example 2 (continued): In the framework of this example, we have seen that $T = X_{(n)}$ is sufficient and complete. The statistic $\mathbf{S} = (X_2/X_1, X_3/X_1, \ldots, X_{n-1}/X_1)$ is distribution-free (this follows by noting that $\mathbf{S} = (Z_2/Z_1, \ldots, Z_{n-1}/Z_1)$, where $(Z_1, \ldots, Z_n) := (X_1/\theta, \ldots, X_n/\theta)$ is distribution-free). The Basu theorem thus implies that $X_{(n)}$ and $(X_2/X_1, X_3/X_1, \dots, X_{n-1}/X_1)$ are independent under P_θ for any $\theta > 0$.

Example 5: the Basu theorem can be seen as an extension of the classical Fisher Lemma on the independence of $\bar{X}$ and s^2 in Gaussian samples. Let indeed $\mathbf{X} = (X_1, \dots, X_n)$, where the X_i's are independent and identically distributed $\mathcal{N}(\mu, \sigma^2)$; take $\mu \in \mathbb{R}$ as the parameter of this model, and consider σ^2 as fixed. Then,

(a) from the factorization criterion, $\bar{X}$ is sufficient for this one-parameter family;

(b) it can be shown (a property of *exponential families*, to be covered in Chapter 4) that $\bar{X}$ is complete for the same family;

(c) from classical results, we know that $ns^2/\sigma^2 \sim \chi^2_{n-1}$, irrespective of μ; hence, s^2 is distribution-free.

The Basu theorem thus implies that $\bar{X}$ and s^2 are independent. Now, such independence holds for any σ^2, and thus extends to the two-parameter family indexed by μ and σ^2. For the same reason, the following pairs also are mutually independent: $\bar{X}$ and $X_{\max} - X_{\min}$; $\bar{X}$ and the vector of *spacings* $(X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)})$; $\bar{X}$ and the vector of *ranks* $(R_1^{(n)}, \dots, R_n^{(n)})$.[3]

In point estimation, completeness essentially complements sufficiency in the Rao-Blackwell theorem, yielding the *Lehmann-Scheffé theorem.*

Theorem 4. (Lehmann-Scheffé) *Let* $\mathbf{T}$ *be sufficient and complete for* $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, *and let* $\mathbf{S}$ *be an unbiased estimator of* $\varphi(\mathrm{P})$. *Then,*

(i) $\mathbf{S}^{\mathbf{T}} := \mathrm{E}_{\not{\mathrm{P}}}[\mathbf{S}|\mathbf{T}]$ *is* essentially unique (*in the sense that if* $\mathbf{S}_1$ *and* $\mathbf{S}_2$ *are two unbiased estimators of* $\varphi(P)$, *then* $\mathbf{S}_1^{\mathbf{T}} = \mathbf{S}_2^{\mathbf{T}}$ P-*almost surely for all* $\mathrm{P} \in \mathcal{P}$);

(ii) irrespective of the convex *loss function* $\mathbf{d} \longmapsto \mathrm{L}_{\mathrm{P}}(\mathbf{d})$, $\mathbf{S}^{\mathbf{T}}$ *has uniformly minimum risk in the class of unbiased estimators of* $\varphi(\mathrm{P})$ (*"*$\mathbf{S}^{\mathbf{T}}$ *is* UMRU *for* $\varphi(\mathrm{P})$*"*)[4].

---

[3] The rank of X_i is defined as $R_i^{(n)} := \#\{\, j = 1, \dots, n : X_j \leq X_i\}$.

[4] UMRU stands for *Uniformly Minimum Risk Unbiased.*

Proof. (i) Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be unbiased estimators of $\varphi(\mathrm{P})$, and denote as $\mathbf{S}_1^{\mathbf{T}}$ and $\mathbf{S}_2^{\mathbf{T}}$ their Rao-Blackwellized versions. Then, for all $\mathrm{P} \in \mathcal{P}$,

$$\mathrm{E}_{\mathrm{P}}\big[\mathbf{S}_i^{\mathbf{T}}\big] := \mathrm{E}_{\mathrm{P}}\big[\mathrm{E}_{\cancel{\mathrm{P}}}[\mathbf{S}_i|\mathbf{T}]\big] = \mathrm{E}_{\mathrm{P}}[\mathbf{S}_i] = \varphi(\mathrm{P}),$$

so that $\mathbf{S}_1^{\mathbf{T}}$ and $\mathbf{S}_2^{\mathbf{T}}$ still are unbiased estimators of $\varphi(\mathrm{P})$. Hence,

$$\mathrm{E}_{\mathrm{P}}\big[\mathbf{S}_1^{\mathbf{T}} - \mathbf{S}_2^{\mathbf{T}}\big] = 0 \qquad \text{for all } \mathrm{P} \in \mathcal{P}.$$

But $\mathbf{S}_1^{\mathbf{T}} - \mathbf{S}_2^{\mathbf{T}}$ is a $\mathbf{T}$-measurable statistic. Completeness of $\mathbf{T}$ thus entails that

$$\mathbf{S}_1^{\mathbf{T}} - \mathbf{S}_2^{\mathbf{T}} = 0 \qquad \text{P-almost surely for all } \mathrm{P} \in \mathcal{P},$$

which establishes the result. (ii) Pick an arbitrary unbiased estimator $\mathbf{V}$ of $\varphi(P)$. Applying Part (i) of the result, then the Rao-Blackwell theorem, we obtain

$$\mathrm{R}_{\mathrm{P}}^{\mathbf{S}^{\mathbf{T}}} = \mathrm{R}_{\mathrm{P}}^{\mathbf{V}^{\mathbf{T}}} \leq \mathrm{R}_{\mathrm{P}}^{\mathbf{V}} \qquad \text{for all } \mathrm{P} \in \mathcal{P}.$$

This establishes that $\mathbf{S}^{\mathbf{T}}$ is UMRU for $\varphi(\mathrm{P})$. □

Exponential families are the main domain of application for the Lehmann-Scheffé theorem. Those families are the subject of Chapter 4. Here, we treat an example that does not belong to exponential families.

<u>Example 2 (continued)</u>: In the framework of this example, we have seen that $T = X_{(n)}$ is sufficient and complete. We also showed that the Rao-Blackwellized version of the unbiased estimator $S = 2X_1$ of θ is

$$S^T = \frac{n+1}{n} X_{(n)}.$$

From the Lehmann-Scheffé theorem, S^T is UMRU for θ. Irrespective of the convex loss function considered. For the L_2 loss function, this shows in particular that this estimator is UMVU (Uniformly Minimum Variance Unbiased) for θ.

3.4 A more involved application: U-statistics

Consider the nonparametric model under which the observation is an independent and identically distributed n-tuple $\mathbf{X} = (X_1, \ldots, X_n)$, where X_i has unspecified density $f \in \mathcal{F}$, the family of all probability densities (with respect to the Lebesgue measure μ) over $(\mathbb{R}, \mathcal{B})$. As we have seen in Chapter 2, the order statistic $\mathbf{X}_{(\cdot)} := \big(X_{(1)}, \ldots, X_{(n)}\big)$ is then sufficient, and it can be shown that it is also complete[5].

That property of the order statistic is not affected if moment restrictions are applied to the densities $f \in \mathcal{F}$ (that is, if $\mathcal{F}$ is replaced with the family $\mathcal{F}_0$ of those densities for which moments, or the moments of some functions, exist up to some order). The Lehmann-Scheffé theorem thus applies to such nonparametric families.

Assume that $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}$ is such that

$$\varphi(\mathrm{P}) := \mathrm{E}_{\mathrm{P}}[\psi(X_1, \ldots, X_k)]$$

exists and is finite for all P such that $f = \frac{\mathrm{dP}}{\mathrm{d}\mu} \in \mathcal{F}_0$ and consider the problem of estimating $\varphi(P)$ on the basis of $\mathbf{X}$. Obviously, $\mathbf{S} := \psi(X_1, \ldots, X_k)$ is an unbiased estimator of $\varphi(P)$. The Rao-Blackwellized version $\mathbf{S}^{\mathbf{X}_{(\cdot)}}$ of this estimator is then UMRU for $\varphi(P)$. Now, conditional on $\mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}$, where $\mathbf{x}_{(\cdot)} = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ has strictly increasing entries (ties have Lebesgue measure zero, hence probability zero, so that they safely can be neglected), $(X_1, \ldots, X_k)$ is uniformly distributed over the $n(n-1)\ldots(n-k+1)$ ordered k-tuples $(x_{i_1}, \ldots, x_{i_k})$ with entries in $\{x_{(1)}, \ldots, x_{(n)}\}$, or equivalently in $\{x_1, \ldots, x_n\}$. It follows that

$$\begin{aligned}\mathbf{S}^{\mathbf{X}_{(\cdot)}} &= \mathrm{E}_{\mathrm{P}}\big[\psi(X_1, \ldots, X_k)|\mathbf{X}_{(\cdot)}\big] \\ &= \frac{1}{n(n-1)\ldots(n-k+1)} \sum_{\substack{i_1,\ldots,i_k=1 \\ i_j \text{ pairwise} \neq}}^{n} \psi(X_{i_1}, \ldots, X_{i_k}),\end{aligned}$$

where the sum is over all distinct ordered k-tuples of integers in $\{1, \ldots, n\}$. This leads to the definition of a U*-statistic* (Hoeffding 1948).

Definition 5. *Let* $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}$ *be such that*

[5] For a proof, one may consider the exponential subfamily $\mathcal{F}_{\mathbf{T}_n}$ associated with the priviliged statistic $\mathbf{T}_n := \big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \ldots, \sum_{i=1}^n X_i^n\big)$. It can be shown that $\mathbf{X}_{(\cdot)}$ generates the same σ-field as $\mathbf{T}_n$, hence is complete for the exponential subfamily (see Chapter 4), which in turn implies that it is complete for the broader family $\mathcal{F}$.

(i) $\varphi(\mathrm{P}) := \mathrm{E}_{\mathrm{P}}[\psi(X_1, \ldots, X_k)]$ *for all* P *with density* $f \in \mathcal{F}_0$ *(see above), and*

(ii) k *is the smallest integer for which such a* ψ *exists (that is,* k *is the smallest number of observations required for unbiased estimation of* $\varphi(\mathrm{P})$*).*

Then, ψ *is called a* kernel *(of order* k*) for* Ψ*, and the statistic*

$$U_\psi = U_\psi(X_1, \ldots, X_n) := \frac{1}{n(n-1)\ldots(n-k+1)} \sum_{\substack{i_1,\ldots,i_k=1 \\ i_j \text{ pairwise } \neq}}^{n} \psi(X_{i_1}, \ldots, X_{i_k})$$

is called a U-statistic with kernel ψ.

<u>Remark 1</u>: If the kernel ψ is symmetric in its arguments (meaning that $\psi(X_{\pi(1)}, \ldots, X_{\pi(k)}) = \psi(X_1, \ldots, X_k)$ for any permutation π of $\{1, \ldots, n\}$), then U_ψ takes the form

$$U_\psi = \frac{k!}{n(n-1)\ldots(n-k+1)} \sum_{\substack{i_1,\ldots,i_k=1 \\ i_1<\ldots<i_k}}^{n} \psi(X_{i_1}, \ldots, X_{i_k}).$$

<u>Remark 2</u>: If $\mathcal{F}$ is further restricted to the subfamily of symmetric (with respect to 0) densities, then the order statistic loses its minimal sufficiency and completeness properties to the order statistic of absolute values. A U-statistic can be defined with appropriate changes in the form of conditional expectations.

It follows from the Lehmann-Scheffé theorem that when f is unspecified in $\mathcal{F}_0 := \{f \in \mathcal{F} : \varphi(\mathrm{P}) \text{ exists and is finite}\}$, the U-statistic U_ψ is, for convex loss functions, a UMRU estimator for $\varphi(\mathrm{P})$, and is essentially unique (in the sense that if ψ_1 and ψ_2 are two kernels for $\varphi(\mathrm{P})$, then $U_{\psi_1} = U_{\psi_2}$ P-almost surely for any $f \in \mathcal{F}_0$).

<u>Example 6</u>: With the above notation, let $\mathcal{F}_0$ be the family of densities $f = \frac{\mathrm{dP}}{\mathrm{d}\mu}$ for which the mean $m_{\mathrm{P}} := \int x f(x) \mathrm{d}\mu(x)$ exists and is finite. Clearly, $\psi(X_1) = X_1$ is a kernel of order one for $\varphi(\mathrm{P}) = m_{\mathrm{P}}$. It follows that

$$U_\psi = \frac{1}{n} \sum_{i=1}^{n} X_i =: \bar{X}$$

is UMRU for m_{P} in the model with unspecified density $f \in \mathcal{F}_0$.

<u>Example 7</u>: Let now $\mathcal{F}_0$ be the family of densities for which the variances $\sigma^2_{\mathrm{P}} := \mathrm{Var}_{\mathrm{P}}[X_i]$ are

finite. Since

$$\mathrm{E_P}[(X_1 - X_2)^2] = \mathrm{E_P}[X_1^2] + \mathrm{E_P}[X_2^2] - 2\mathrm{E_P}[X_1 X_2] = 2\{\mathrm{E_P}[X_1^2] - (\mathrm{E_P}[X_1])^2\} = 2\sigma_{\mathrm{P}}^2$$

for all P such that $f = \frac{\mathrm{dP}}{\mathrm{d}\mu} \in \mathcal{F}_0$, a kernel of order two for $\varphi(\mathrm{P}) = \sigma_{\mathrm{P}}^2$ is

$$\psi(X_1, X_2) = \frac{1}{2}\left(X_1 - X_2\right)^2.$$

The corresponding U-statistic is

$$\begin{aligned}
U_\psi &= \frac{1}{2n(n-1)} \sum_{\substack{i,j=1\\ i\neq j}}^{n} (X_i - X_j)^2 = \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} (X_i - X_j)^2 \\
&= \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} (X_i^2 + X_j^2 - 2X_i X_j) = \frac{1}{2n(n-1)} \left\{ 2n \sum_{i=1}^{n} X_i^2 - 2\left(\sum_{i=1}^{n} X_i \right)^2 \right\} \\
&= \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right\} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 =: S^2,
\end{aligned}$$

which is the traditional unbiased estimator of σ^2. From the Lehmann-Scheffé theorem, S^2 is thus UMRU for σ_{P}^2 in the model with unspecified density $f \in \mathcal{F}_0$.

Those UMRU results look impressively strong, as the corresponding families of distributions are quite big. Yet one should not forget that the bigger $\mathcal{F}$, the more severe the unbiasedness constraint (the fact that $\bar{X}$ and S^2 are UMRU estimators is largely due to the fact that they do not have many competitors).