

MATHEMATICAL STATISTICS

1 Introduction

Wald decomposed the problem of statistical inference in three aspects: the statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, the decision space \mathcal{D} , and the loss function L . The statistical model is defined by the space of observations \mathcal{X} , a corresponding σ -algebra \mathcal{A} , and a family \mathcal{P} of probability laws over $(\mathcal{X}, \mathcal{A})$. In particular, a parametric model assumes $\mathcal{P} = \{P_{\underline{\theta}} \mid \underline{\theta} \in \Theta \subseteq \mathbb{R}^k\}$ whereas a nonparametric model cannot be indexed by any subset of \mathbb{R}^k for $k < \infty$.

There are two types of decision rules: pure and randomized. The former is just a function $\delta(x)$ mapping \mathcal{X} into \mathcal{D} . The latter associates a probability law over \mathcal{D} to possible decisions given the observation $x \in \mathcal{X}$. Therefore, for each possible value of $x \in \mathcal{X}$, there is a probability law P_{δ}^x over $(\mathcal{D}, \mathcal{A}_{\mathcal{D}})$. Note that if P_{δ}^x is degenerated, then it is equivalent to work with a pure decision rule.

Finally, the loss function $L_P(d)$ maps $(\mathcal{D}, \mathcal{P})$ into \mathbb{R}_+ , and its expectation is known as the risk function R_P^{δ} . For pure decision rules, $R_P^{\delta} = \int_{\mathcal{X}} L_P[\delta(x)] dP(x)$. For randomized decision rules, $R_P^{\delta} = \int_{\mathcal{X}, \mathcal{D}} L_P(d) dP_{\delta}^x(d) dP(x)$.

Definition 1. A decision rule δ_1 is uniformly better than another decision rule δ_2 if the risk associated with δ_1 is at most equal to the risk of adopting δ_2 , that is, $R_P^{\delta_1} \leq R_P^{\delta_2}$. Notation: $\delta_1 \succeq \delta_2$ or $P_{\delta_1}^x \succeq P_{\delta_2}^x$.

Definition 2. A decision rule δ_* is uniformly optimal in the class of decisions Δ if $\delta_* \succeq \delta$ for every $\delta \in \Delta$.

In general, \succeq is a partial pre-order, so that there is no optimal decision rule δ_* . Under this circumstance, it is natural to restrict ourselves to the following two definitions or to follow a Bayesian approach by imposing a prior over \mathcal{P} .

Definition 3. δ_* is minimax in Δ if it belongs to Δ and

$$\max_{P \in \mathcal{P}} R_P^{\delta_*} \leq \max_{P \in \mathcal{P}} R_P^\delta$$

for every $\delta \in \Delta$.

Definition 4. δ_* is uniformly most powerful in a sub-class of Δ induced by some statistical principle, e.g. unbiasedness, equivariance, invariance, and Neyman's principle.

Example 1. Consider a point estimation problem where the model at hand is represented by $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Then there is a functional g such that $P \in \mathcal{P}$ is transformed into $g(P)$. Denote by $\hat{g}(\underline{X})$ the random variable which estimates the functional, i.e. the estimator, and let $\hat{g}(\underline{x})$ denote the corresponding estimate. In the parametric case, for the sake of simplicity, we use the notation $g(\underline{\theta})$ instead of $g(P_{\underline{\theta}})$. The decision space \mathcal{D} consists in $g(\mathcal{P})$. Accordingly, in the parametric case, we write $\mathcal{D} = g(\Theta)$. Finally, the most common lost functions used for estimation purposes are the quadratic $L_2(d, P) = [d - g(P)]^2$ and the absolute value $L_\infty(d, P) = |d - g(P)|$.

Example 2. Consider a hypothesis testing problem where the statistical model is given by $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where the probability family has a partition of the form $\mathcal{P} = H_0 \oplus H_1$. Accordingly, the decision space is also decomposed in two components $\mathcal{D} = \{d_0, d_1\}$, where d_0 and d_1 indicate "acceptance" and rejection of the null hypothesis H_0 , respectively. The loss function can be based in a pure or randomized rule. Typically, a pure decision consists in

$$L(d_0, P) = \begin{cases} 0 & \text{if } P \in H_0 \\ a_0 & \text{if } P \in H_1 \end{cases} \quad \text{and} \quad L(d_1, P) = \begin{cases} a_1 & \text{if } P \in H_0 \\ 0 & \text{if } P \in H_1, \end{cases}$$

where a_0 and a_1 are positive constants standing for the error of type I and II, respectively. On the other hand, randomized decisions depend on a measurable application $\phi(\underline{X})$ on $(\mathcal{X}, \mathcal{A})$. For instance, this dependence is clear at the risk function level

$$R_P^\phi = \begin{cases} a_1 E_{H_0}[\phi(\underline{x})] & \text{if } P \in H_0, \\ a_0 E_{H_1}[1 - \phi(\underline{x})] & \text{if } P \in H_1. \end{cases}$$

Of course, the aim is at minimizing the risk R_P^ϕ for every $P \in \mathcal{P}$. Note that the constants a_0 and a_1 do not interfere in this minimization problem, so that we can take $a_0 = a_1 = 1$ without any loss of generality.

2 Sufficiency

2.1 Dominated families

Definition 1. A measure μ is σ -finite if there are sets A_1, A_2, \dots belonging to \mathcal{A} such that $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ and $\mu(A_i) < \infty$.

Definition 2. The measure ν is absolute continuous to (or dominated by) the σ -finite measure μ , i.e. $\nu \ll \mu$, if $\nu(A) = 0$, $\forall A \in \mathcal{A}$ such that $\mu(A) = 0$.

Radon-Nikodym's Theorem. $\nu \ll \mu$ if and only if there exists a function f which maps $(\mathcal{X}, \mathcal{A})$ into $(\mathbb{R}_+, \mathcal{B} \cap \mathbb{R}_+)$ such that $\nu(A) = \int_A d\nu = \int_A f d\mu$. The function f is called the Radon-Nikodym derivative of ν with respect to μ and it is essentially unique, that is, if there are f_1 and f_2 such that $\nu(A) = \int_A f_1 d\mu = \int_A f_2 d\mu$, then $f_1 = f_2$ μ -almost everywhere.

The set of all Radon-Nikodym derivatives is denoted by $\frac{d\nu}{d\mu}$. Note that if ν is a probability measure, then $f \in \frac{d\nu}{d\mu}$ is the probability density of ν with respect to μ .

Properties. Consider $P \ll Q \ll R$. If $f \in \frac{dP}{dQ}$ and $g \in \frac{dQ}{dR}$, then $fg = \frac{dP}{dR}$. Accordingly, if $f \in \frac{dP}{dR}$ and $g \in \frac{dQ}{dR}$, then $f/g = \frac{dP}{dQ} \in \frac{dP}{dQ}$.

Definitions. The family \mathcal{P} is dominated by the σ -finite measure μ , i.e. $\mathcal{P} \ll \mu$, if $P \ll \mu$, for every $P \in \mathcal{P}$. The family \mathcal{P} is dominated if there exists a σ -finite measure μ such that $\mathcal{P} \ll \mu$.

Halmos-Savage's Lemma. \mathcal{P} is a dominated family of probability measures if and only if there exists a countable subset $\{P_1, P_2, \dots\} \subseteq \mathcal{P}$ and a sequence of nonnegative real numbers $\{c_i\}_{i=1}^{\infty}$ such that $\sum_{i=1}^{\infty} c_i = 1$ and \mathcal{P} is dominated by the privileged measure $P_* \equiv \sum_{i=1}^{\infty} c_i P_i$.

2.2 Conditional expectation and probability

Definitions. A statistic T is a function that maps $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$. The σ -algebra generated by T is $\mathcal{A}_{\mathcal{T}} = T^{-1}(\mathcal{B}_{\mathcal{T}}) \subseteq \mathcal{A}$. The probability measure induced by T over $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ is defined by $P^T(B) = P(T^{-1}(B))$, $\forall B \in \mathcal{B}_{\mathcal{T}}$. Accordingly, the model induced by T for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ consists in $(\mathcal{T}, \mathcal{B}_{\mathcal{T}}, \mathcal{P}^T)$.

Transfer Theorem. A function $g : (\mathcal{X}, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$ is T -measurable if there exists a function $h : (\mathcal{T}, \mathcal{B}_{\mathcal{T}}) \rightarrow (\mathbb{R}, \mathcal{B})$ such that $g(x) = h(T(x))$, that is, $g = h \circ T$. In case g is P -integrable, then

$$\int_B h(t) dP^T(t) = \int_{T^{-1}(B)} h(T(x)) dP(x),$$

for all $B \in \mathcal{B}_{\mathcal{T}}$. Note that taking the integration over the whole space \mathcal{T} implies that

$$E_{P^T}[h(T)] = \int_{\mathcal{T}} h(t) dP^T(t) = \int_{\mathcal{X}} h(T(x)) dP(x) = E_P[h(T(X))].$$

Conditional Expectations. Consider the statistic T and a nonnegative P -measurable function g mapping $(\mathcal{X}, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B})$. If the measure ν such that $\nu(B) = \int_{T^{-1}(B)} g(x) dP(x)$ is absolute continuous to P^T , then the Radon-Nikodym derivative $d\nu/dP^T$ is T -measurable, denoted by $E_P[g(X) | T]$, and maps \mathcal{T} into \mathbb{R} such that

$$\nu(B) = \int_{T^{-1}(B)} g(x) dP(x) = \int_B E_P[g(X) | T = t] dP^T(t) = \int_B \frac{d\nu}{dP^T}(t) dP^T(t).$$

In other words, the conditional expectation is a Radon-Nikodym derivative.

Properties.

(1) *Law of Iterated Expectations.* Taking $B = \mathcal{T}$ yields

$$E_P[g(X)] = \int_{\mathcal{X}} g(x) dP(x) = \int_{\mathcal{T}} E_P[g(X) | T = t] dP^T(t) = E_{P^T}(E_P[g(X) | T]),$$

(2) *Linearity.* $E_P[\sum_k a_k g_k(X) | T] = \sum_k a_k E_P[g_k(X) | T]$,

(3) *T -measurability.* $E_P[h(T)g(X) | T] = h(T)E_P[g(X) | T]$.

Conditional Probabilities. For $A \in \mathcal{A}$, the conditional probability of A given T is given by $P(A | T) \equiv E_P[I_A(X) | T]$ which maps $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ into $(\mathbb{R}, \mathcal{B})$. Ac-

cordingly, the unconditional distribution is given by $P(A) = \int_{\mathcal{X}} I_A(x) dP(x) = E_P[I_A(X)]$.

Actually, $P(\cdot | T = t) : \mathcal{A} \rightarrow [0, 1]$ does not have, in general, all the properties of a probability measure (induced by $T = t$). However, if $(\mathcal{X}, \mathcal{A})$ is a metric space complete, separable, equipped with a Borel σ -field, and $\mathcal{B}_{\mathcal{T}}$ admits a countable generating subset, then there exists a set of versions of the conditional probability $P(A | T)$ such that, for every $t \in \mathcal{T}$ and $A \in \mathcal{A}$, $P_{X|T=t}(A) \equiv P(A | T = t)$ stands for a probability measure over $(\mathcal{X}, \mathcal{A})$. Moreover, $E_P[g(X) | T = t] = \int_{\mathcal{X}} g(x) dP_{X|T=t}(x)$, recovering the conventional definition.

Definition. A statistic $T : (\mathcal{X}, \mathcal{A}) \rightarrow (\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ if, $\forall A \in \mathcal{A}$, there exists a version of the conditional probability $P(A | T)$ which does not depend on $P \in \mathcal{P}$, i.e. $\forall P_1, P_2 \in \mathcal{P}, P_1(A | T = t) = P_2(A | T = t)$.

Halmos-Savage's Theorem. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a dominated model and P_* a privileged measure. A statistic T is sufficient if and only if $\forall P \in \mathcal{P}$, there exists a T -measurable version of $\frac{dP}{dP_*}$.

Proof. Consider T sufficient and $g_P(t)$ a version of $\frac{dP^T}{dP_*^T}(t)$, where $t \in \mathcal{T}$. Then, $\forall A \in \mathcal{A}$ and $P \in \mathcal{P}$,

$$\begin{aligned} P(A) &= \int_{\mathcal{T}} E_P[I_A(X) | T = t] dP^T(t) = \int_{\mathcal{T}} P(A | T = t) dP^T(t) \\ &= \int_{\mathcal{T}} P_*(A | T = t) dP^T(t) = \int_{\mathcal{T}} E_{P_*}[I_A(X) | T = t] dP^T(t), \end{aligned}$$

since, by sufficiency, there exists a version of the conditional probability that does not depend on the particular P taken. Hence, using Radon-Nikodym's Theorem yields

$$\begin{aligned} P(A) &= \int_{\mathcal{T}} E_{P_*}(I_A(X) | T = t) g_P(t) dP_*^T(t) \\ &= \int_{\mathcal{T}} E_{P_*}[I_A(X) g_P(T(X)) | T = t] dP_*^T(t) \\ &= \int_{\mathcal{X}} I_A(x) g_P[T(x)] dP_*(x) = \int_{\mathcal{A}} g_P[T(x)] dP_*(x). \end{aligned}$$

However, by definition, $P(A) = \int_{\mathcal{A}} dP(x)$ so that, by Radon-Nikodym's Theo-

rem, $g_P[T(x)]$, which is T -measurable, is a version of the Radon-Nikodym derivative $\frac{dP}{dP_*}(x)$, showing necessity. Consider now a T -measurable version $g_P[T(\cdot)]$ of $\frac{dP}{dP_*}(\cdot)$. For all $B \in \mathcal{B}_T$, $P^T(B) = P[T^{-1}(B)] = \int_{T^{-1}(B)} dP(x)$ by definition. Radon-Nikodym's Theorem thus implies that

$$\begin{aligned}
P^T(B) &= \int_{T^{-1}(B)} g_P[T(x)] dP_*(x) \\
&= \int_{\mathcal{T}} E_{P_*} [I_{T^{-1}(B)}(X) g_P(T(X)) | T = t] dP_*^T(t) \\
&= \int_{\mathcal{T}} E_{P_*} [I_{T^{-1}(B)}(X) | T = t] g_P(t) dP_*^T(t) \\
&= \int_B g_P(t) dP_*^T(t) = \int_B dP^T(t),
\end{aligned}$$

which means that $g_P(t) \in \frac{dP^T}{dP_*^T}(t)$ for every $t \in \mathcal{T}$. Then for every $A \in \mathcal{A}$, and $P \in \mathcal{P}$,

$$\begin{aligned}
\int_{\mathcal{T}} P[A | T = t] dP^T(t) &= \int_{\mathcal{T}} E_P [I_A(X) | T = t] dP^T(t) \\
&= \int_{\mathcal{X}} I_A(x) dP(x) \\
&= \int_{\mathcal{X}} I_A(x) g_P[T(x)] dP_*(x) \\
&= \int_{\mathcal{T}} E_{P_*} [I_A(X) g_P(T(X)) | T = t] dP_*^T(t) \\
&= \int_{\mathcal{T}} E_{P_*} [I_A(X) | T = t] g_P(t) dP_*^T(t) \\
&= \int_{\mathcal{T}} E_{P_*} [I_A(X) | T = t] dP^T(t) \\
&= \int_{\mathcal{T}} P_* [A | T = t] dP^T(t),
\end{aligned}$$

which implies that the conditional probability given T does not depend upon a particular measure $P \in \mathcal{P}$, i.e. T is sufficient. ■

2.3 Neyman-Fisher factorization criterion

Halmos-Savage's Theorem gives a necessary and sufficient condition for sufficiency. Although it gives a better intuition for the concept of sufficiency, it does not provide a straightforward way to assess whether a statistic is sufficient. The Neyman-Fisher factorization criterion gives an alternative necessary and suffi-

ciency condition for sufficiency which is more useful in practical considerations than Halmos and Savage's result.

Proposition. Consider a model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where \mathcal{P} is a family of measures dominated by a σ -finite measure μ . Denote the probability density $\frac{dP}{d\mu}$ by f_P . Then, the statistic T is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ if and only if, $\forall P \in \mathcal{P}$, there exists a version of $f_P \in \frac{dP}{d\mu}$ such that $f_P(x) = g_P[T(x)]h(x)$, where both g_P and h are measurable.

Proof. We need to show that the factorization criterion above is equivalent to the necessary and sufficient condition provided by Halmos-Savage's Theorem. Let T be sufficient, so that there exists a T -measurable version $g_P[T(\cdot)]$ of $\frac{dP}{d\mu}$. Then, it immediately follows that $f_P(x) = g_P[T(x)]h(x)$, where $f_P(x) \in dP/d\mu$, $g_P[T(x)] \in dP/dP_*$ and $h(x) \in dP_*/d\mu$. Consider now that $f_P(x) = g_P[T(x)]h(x)$, $\forall P \in \mathcal{P}$. Then,

$$f_{P_*}(x) = \frac{dP_*}{d\mu}(x) = \sum_i c_i \frac{dP_i}{d\mu}(x) = \sum_i c_i f_{P_i}(x) = \sum_i c_i g_{P_i}[T(x)]h(x).$$

However, $\forall P \in \mathcal{P}$,

$$\frac{dP}{dP_*}(x) = \frac{\frac{dP}{d\mu}(x)}{\frac{dP_*}{d\mu}(x)} = \frac{f_P(x)}{f_{P_*}(x)} = \frac{g_P[T(x)]h(x)}{\sum_i c_i g_{P_i}[T(x)]h(x)} = \frac{g_P[T(x)]}{\sum_i c_i g_{P_i}[T(x)]},$$

that is, $\frac{dP}{dP_*}$ is T -measurable. Therefore, by Halmos-Savage's Theorem, T is sufficient. ■

Example 1. Consider the random sample $\underline{X} = (X_1, \dots, X_n)$, where each X_i has distribution $P \in \mathcal{P}$ dominated by μ . Denote by $f_P \in \frac{dP}{d\mu}$ the probability density. Then, the joint probability density with respect to the product measure μ^n can be expressed by

$$f_P(x_1, \dots, x_n) = \prod_{i=1}^n f_P(x_i) = \prod_{i=1}^n f_P(x_{(i)}).$$

Thus, the order statistic $T = (X_{(1)}, \dots, X_{(n)})$ is sufficient.

Example 2. Consider the random sample $\underline{X} = (X_1, \dots, X_n)$, where each X_i has a Bernoulli distribution with parameter $p \in [0, 1]$. It is clear that \underline{X} takes

values in $\{0, 1\}^n$, but we can consider the whole \mathbb{R}^n space in order to use the Borel σ -algebra. It can be shown that the counting measure

$$\mu(B) = \#\{\underline{x} = (x_1, \dots, x_n) \in B \text{ such that } x_i \in \{0, 1\}, \forall i\}$$

is a dominating σ -finite measure for the model at hand. Then, the probability density f_p of \underline{X} with respect to μ , under the parameter value p , is

$$f_p(\underline{x}) = \frac{1}{\mu(\mathbb{R}^n)} \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = \frac{1}{\mu(\mathbb{R}^n)} p^{\sum_i x_i} (1-p)^{n-\sum_i x_i}.$$

Thus, the sum statistic $T = \sum_{i=1}^n X_i$ is sufficient since $f_p(\underline{x})$ can be factorized with $g_p[T(\underline{x})] = p^{T(\underline{x})} (1-p)^{n-T(\underline{x})}$ and $h(\underline{x}) = 1/\mu(\mathbb{R}^n)$.

2.4 Minimal sufficiency

It is a concept related to the maximal reduction of a problem. For instance, the observations \underline{X} clearly constitute a sufficient statistic, but they do not offer any reduction. More formally, consider two sufficient statistics T_1 and T_2 with induced σ -algebra \mathcal{A}_{T_1} and \mathcal{A}_{T_2} , respectively. If $\mathcal{A}_{T_2} \subseteq \mathcal{A}_{T_1}$, then T_2 is T_1 -measurable, which means that T_1 contains some irrelevant information.

Definition. Consider Υ the set of sufficient statistics. A statistic S is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ if S is T -measurable for every $T \in \Upsilon$, that is, $\mathcal{A}_S \subseteq \mathcal{A}_T$. Equivalently, S is a minimal sufficient statistic if $\mathcal{A}_S = \bigcap_{T \in \Upsilon} \mathcal{A}_T$.

If \mathcal{P} is a dominated family of measures, then the Radon-Nikodym derivative $\frac{dP}{dP_*}$, where $P \in \mathcal{P}$ and P_* is a privileged measure, is a minimal sufficient statistic. This result stems directly from Halmos-Savage's Theorem, but it is not so informative given that, in general, the set of Radon-Nikodym derivatives has an uncountable number of variables. However, for the particular case of exponential families, there is a finite number of random variables which are able to span the whole set $\left\{ \frac{dP}{dP_*} \mid P \in \mathcal{P} \right\}$.

Proposition 1. Let $\mathcal{P} = \{P_1, \dots, P_K\}$ and consider the privileged dominating measure $P_* = \frac{1}{K} \sum_{k=1}^K P_k$. Let f_k denote the density function $\frac{dP_k}{dP_*}$, $k = 1, \dots, K$. Then, $\underline{f} = (f_1, \dots, f_K)$ is minimal sufficient.

Proposition 2. Let $\mathcal{P}_0 \subset \mathcal{P}$. If \underline{S} is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ and is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, then \underline{S} is also minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

Proof. Consider \underline{T} a sufficient statistic for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, then \underline{T} is also sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$. However, \underline{S} is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$, thus it is by definition \underline{T} -measurable. Note that this is valid for every statistic T sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, hence \underline{S} is also minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. ■

Under these circumstances, the trick consists in constructing a sub-model with a finite family of measures and then uncover a minimal sufficient statistic that is also sufficient for the original model.

3 The exponential family

The model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is exponential if $\mathcal{P} = \{P_{\underline{\theta}} \mid \underline{\theta} \in \Theta \subseteq \mathbb{R}^K\}$ is dominated by a σ -finite measure μ with density

$$f_{\underline{\theta}}(\underline{x}) = C(\underline{\theta})h(\underline{x}) \exp\left(\sum_{k=1}^K \theta_k T_k(\underline{x})\right) \in \frac{dP_{\underline{\theta}}}{d\mu},$$

where $\underline{x} \in \mathcal{X}$, and $h(\cdot)$ and $C(\cdot)$ are nonnegative functions. The vector $\underline{\theta} = \{\theta_1, \dots, \theta_K\}$ is called the natural parameter and belongs to the parametric space Θ , which is a subset of the natural parameter space

$$\bar{\Theta} = \left\{ \underline{\theta} \in \mathbb{R}^K : \frac{1}{C(\underline{\theta})} = \int_{\mathcal{X}} h(\underline{x}) \exp\left(\sum_k \theta_k T_k(\underline{x})\right) d\mu(\underline{x}) < \infty \right\}.$$

Finally, the vector $\underline{T}(\underline{X})$ of real-valued statistics $(T_1(\underline{X}), \dots, T_K(\underline{X}))$ is called the privileged statistic.

Remark 1. The natural parameter space is convex, that is, for $\underline{\theta}, \underline{\theta}' \in \bar{\Theta}$ and $\alpha \in [0, 1]$, the convex combination $\underline{\theta}^* = \alpha \underline{\theta} + (1 - \alpha) \underline{\theta}'$ belongs to $\bar{\Theta}$.

$$\begin{aligned} \frac{1}{C(\underline{\theta}^*)} &= \int_{\mathcal{X}} h(\underline{x}) \exp\left(\sum_{k=1}^K \theta_k^* T_k(\underline{x})\right) d\mu(\underline{x}) \\ &= \int_{\mathcal{X}} h(\underline{x}) \exp\left(\alpha \sum_{k=1}^K \theta_k T_k(\underline{x}) + (1 - \alpha) \sum_{k=1}^K \theta'_k T_k(\underline{x})\right) d\mu(\underline{x}) \\ &= \int_{\mathcal{X}} h(\underline{x}) \exp\left(\alpha \sum_{k=1}^K \theta_k T_k(\underline{x})\right) \exp\left((1 - \alpha) \sum_{k=1}^K \theta'_k T_k(\underline{x})\right) d\mu(\underline{x}) \end{aligned}$$

$$\begin{aligned}
&\leq \left[\int_{\mathcal{X}} h \exp \left(\sum_{k=1}^K \theta_k T_k \right) d\mu \right]^\alpha \left[\int_{\mathcal{X}} h \exp \left(\sum_{k=1}^K \theta'_k T_k \right) d\mu \right]^{1-\alpha} \\
&= \left(\frac{1}{C(\underline{\theta})} \right)^\alpha \left(\frac{1}{C(\underline{\theta}')} \right)^{1-\alpha} < \infty,
\end{aligned}$$

where the third passage stems from the Holden inequality.

Remark 2. Let $\nu(A) = \int_A h(\underline{x}) d\mu(\underline{x})$. Then,

$$\frac{dP_{\underline{\theta}}}{d\nu} = C(\underline{\theta}) \exp \left(\sum_{k=1}^K \theta_k T_k(\underline{x}) \right),$$

which means that $h(\underline{x})$ plays no structural role. It can be set to one without loss of generality just by selecting the appropriate measure.

Remark 3. We tacitly assume that \underline{T} is of full rank K , which means that there are almost surely no linear constraints on \underline{T} . Otherwise, we can always set $T_K = \sum_{k=1}^{K-1} \alpha_k T_k$ and rearrange the exponential density in order to reduce the dimension of the parametric space. We also assume that the interior of Θ_0 is not empty with respect to \mathbb{R}^K . In case Θ_0 has empty interior with respect to \mathbb{R}^K , we shall say that the family \mathcal{P} is a curved exponential family, which typically leads to several problems.

Remark 4. If $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is an exponential model with natural parameter $\underline{\theta}$ and privileged statistic \underline{T} , then $(\mathbb{R}^K, \mathcal{B}^K, \mathcal{P}^T)$, where \mathcal{P}^T is the family of \underline{T} -induced measures, is also an exponential model with natural parameter $\underline{\theta}$ and privileged statistic \underline{T} .

$$\begin{aligned}
P_{\underline{\theta}}^T(B) &= \int_{T^{-1}(B)} C(\underline{\theta}) h(\underline{x}) \exp \left(\sum_{k=1}^K \theta_k T_k(\underline{x}) \right) d\mu(\underline{x}) \\
&= \int_{\mathcal{X}} C(\underline{\theta}) I_{T^{-1}(B)}(\underline{x}) \exp \left(\sum_{k=1}^K \theta_k T_k(\underline{x}) \right) d\nu(\underline{x}) \\
&= \int_{\mathcal{T}} C(\underline{\theta}) E \left[I_{T^{-1}(B)}(\underline{X}) \exp \left(\sum_{k=1}^K \theta_k T_k(\underline{X}) \right) \middle| \underline{T} = \underline{t} \right] d\nu^T(\underline{t}) \\
&= \int_B C(\underline{\theta}) \exp \left(\sum_{k=1}^K \theta_k t_k \right) d\nu^T(\underline{t}) \\
&= \int_B C(\underline{\theta}) h^T(\underline{t}) \exp \left(\sum_{k=1}^K \theta_k t_k \right) d\mu(\underline{t}), \quad \forall B \in \mathcal{B}_{\mathcal{T}},
\end{aligned}$$

which implies that $f_{\underline{\theta}}^T(\underline{t}) = C(\underline{\theta})h^T(\underline{t}) \exp\left(\sum_{k=1}^K \theta_k t_k\right)$.

Remark 5. Note that all $P_{\underline{\theta}}$'s belonging to \mathcal{P} are equivalent measures, that is, for $P_{\underline{\theta}}, P_{\underline{\theta}'} \in \mathcal{P}$, $P_{\underline{\theta}} \ll P_{\underline{\theta}'}$.

3.1 Sufficiency in exponential families

Proposition 1. The privileged statistic T is minimal sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ if \mathcal{P} is an exponential family of full rank.

Proof. It can be immediately seen that T is sufficient by the Neyman-Fisher factorization criterion. In addition, because \mathcal{P} is a probability family of full rank, there exist $\underline{\theta}^{(0)}, \dots, \underline{\theta}^{(K)} \in \bar{\Theta}$ forming a hypercube with nonzero volume, i.e.

$$|\mathbf{M}| = \begin{vmatrix} \theta_1^{(1)} - \theta_1^{(0)} & \theta_2^{(1)} - \theta_2^{(0)} & \dots & \theta_K^{(1)} - \theta_K^{(0)} \\ \theta_1^{(2)} - \theta_1^{(0)} & \theta_2^{(2)} - \theta_2^{(0)} & \dots & \theta_K^{(2)} - \theta_K^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^{(K)} - \theta_1^{(0)} & \theta_2^{(K)} - \theta_2^{(0)} & \dots & \theta_K^{(K)} - \theta_K^{(0)} \end{vmatrix} > 0.$$

Hence we can define $\mathcal{P}_0 = \{P_{\underline{\theta}^{(0)}}, \dots, P_{\underline{\theta}^{(K)}}\}$, where $P_{\underline{\theta}^{(0)}}$ can be taken as a privileged measure given that the $P_{\underline{\theta}^{(k)}}$'s are all equivalent measures. Then, the vector formed by the Radon-Nikodym derivatives $\left\{\frac{dP_{\underline{\theta}^{(1)}}}{dP_{\underline{\theta}^{(0)}}}, \dots, \frac{dP_{\underline{\theta}^{(K)}}}{dP_{\underline{\theta}^{(0)}}}\right\}$ is a minimal sufficient statistic for \mathcal{P}_0 . Now, for $k = 1, \dots, K$,

$$\begin{aligned} \frac{dP_{\underline{\theta}^{(k)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X}) &= \frac{f_{\underline{\theta}^{(k)}}}{f_{\underline{\theta}^{(0)}}}(\underline{X}) \\ &= \frac{C(\underline{\theta}^{(k)}) h(\underline{X}) \exp\left[\sum_j \theta_j^{(k)} T_j(\underline{X})\right]}{C(\underline{\theta}^{(0)}) h(\underline{X}) \exp\left[\sum_j \theta_j^{(0)} T_j(\underline{X})\right]} \\ &= \frac{C(\underline{\theta}^{(k)})}{C(\underline{\theta}^{(0)})} \exp\left[\sum_j (\theta_j^{(k)} - \theta_j^{(0)}) T_j(\underline{X})\right]. \end{aligned}$$

However, if $\left\{\frac{dP_{\underline{\theta}^{(1)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X}), \dots, \frac{dP_{\underline{\theta}^{(K)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X})\right\}$ is a minimal sufficient statistic, then $\left\{\log \frac{dP_{\underline{\theta}^{(1)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X}), \dots, \log \frac{dP_{\underline{\theta}^{(K)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X})\right\}$ is also minimal sufficient as well as the vector composed by

$$\log \left[\frac{C(\underline{\theta}^{(0)})}{C(\underline{\theta}^{(k)})} \frac{dP_{\underline{\theta}^{(k)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X}) \right] = \sum_{j=1}^K (\theta_j^{(k)} - \theta_j^{(0)}) T_j(\underline{X}) = \langle \theta_j^{(k)} - \theta_j^{(0)}, \underline{T}(\underline{X}) \rangle,$$

$k = 1, \dots, K$. In a matrix form, we have the following linear system

$$\underline{A}(\underline{X}) = \left[\log \left(\frac{C(\underline{\theta}^{(0)})}{C(\underline{\theta}^{(k)})} \frac{dP_{\underline{\theta}^{(k)}}}{dP_{\underline{\theta}^{(0)}}}(\underline{X}) \right) \right]_{K \times 1} = \mathbf{M} \underline{T}(\underline{X}),$$

where there is an inverse matrix \mathbf{M}^{-1} given that \mathbf{M} is nonsingular, i.e. $|\mathbf{M}| > 0$. Hence, as $\underline{A}(\underline{X})$ is a minimal sufficient statistic, $\mathbf{M}^{-1}\underline{A}(\underline{X}) = \underline{T}(\underline{X})$ is also minimal sufficient for \mathcal{P}_0 . Minimal sufficiency of $\underline{T}(\underline{X})$ with respect to \mathcal{P} follows then from sufficiency for \mathcal{P} and minimal sufficiency for \mathcal{P}_0 . ■

3.2 Sampling exponential models

Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ with exponential marginal distribution $P_{\underline{\theta}} \in \mathcal{P}$ with natural parameter $\underline{\theta}$ and privileged statistic $\underline{T}(X_i)$. Then, the joint density with respect to the product measure is given by

$$\begin{aligned} f_{\underline{\theta}}(\underline{x}) &= \prod_{i=1}^n f_{\underline{\theta}}(x_i) \\ &= \prod_{i=1}^n C(\underline{\theta}) h(x_i) \exp \left(\sum_{k=1}^K \theta_k T_k(x_i) \right) \\ &= C(\underline{\theta})^n \left[\prod_{i=1}^n h(x_i) \right] \exp \left(\sum_{k=1}^K \theta_k \sum_{i=1}^n T_k(x_i) \right), \end{aligned}$$

which is still an exponential model with natural parameter $\underline{\theta}$. Moreover, the privileged statistic is just the sum of the marginal privileged statistics. Thus, whatever sample size, the dimension of the problem remains always at K .

Example 1. Consider a random variable X with distribution of the Bernoulli family. Then, the density with respect to the counting measure of $\{0, 1\}$, $\mu(A) = \#\{A \cap \{0, 1\}\}$, is

$$\begin{aligned} f_p(x) &= p^x (1-p)^{1-x} = \exp \left[\log \left(p^x (1-p)^{1-x} \right) \right] \\ &= \exp [x \log p + (1-x) \log(1-p)] \\ &= \exp [x \log p + \log(1-p) - x \log(1-p)] \\ &= \exp \left[x \log \frac{p}{1-p} + \log(1-p) \right] = (1-p) \exp \left[x \log \frac{p}{1-p} \right], \end{aligned}$$

for $x \in \mathbb{R}$. Hence, in this case, $C(\theta) = 1 - p$, the natural parameter $\theta = \log \frac{p}{1-p}$, and the privileged statistic $T(X) = X$. In particular, a sample of n Bernoulli essays has an exponential density with natural parameter $\theta = \log \frac{p}{1-p}$ and privileged statistic $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$.

Example 2. Consider a random variable X that belongs to the Poisson family with density

$$f_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \frac{1}{x!} \exp(x \log \lambda).$$

Then, $C(\theta) = e^{-\lambda}$, $h(x) = 1/x!$, $\theta = \log \lambda$ is the natural parameter, and $T(X) = X$ stands for the privileged statistic.

Example 3. Consider a random variable X which belongs to the Gaussian family with density

$$\begin{aligned} f_{\mu, \sigma}(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (x^2 - 2x\mu + \mu^2) \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{\mu^2}{2\sigma^2} \right] \exp \left[\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 \right]. \end{aligned}$$

Then, $C(\underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{\mu^2}{2\sigma^2} \right]$, $\underline{\theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$ is the vector of natural parameters, and $\underline{T}(X) = (X, X^2)$ stands for the privileged statistic. Thus, a sample $\underline{X} = (X_1, \dots, X_n)$ of independent random variables normally distributed with natural parameter $\underline{\theta}$ has privileged statistic $\underline{T}(\underline{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$. In particular, this implies that the statistic formed by the sample mean and variance (\bar{X}, s^2) is minimal sufficient.

Example 4 (Behrens-Fisher's problem). Consider two random samples $\underline{X} = (X_1, \dots, X_n)$ and $\underline{Y} = (Y_1, \dots, Y_n)$, where $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_i \sim N(\mu_y, \sigma_y^2)$. Then, under the null hypothesis of $\mu = \mu_x = \mu_y$, the joint distribution of $(\underline{X}, \underline{Y})$ is a curved exponential distribution. If one writes the distribution as a function of three parameters $(\mu, \sigma_x, \sigma_y)$, then the property of exponential linearity is lost.

3.3 Moments of a privileged statistic

Proposition. For all $\underline{\theta} \in \Theta$, the moments $m_i = E_{\underline{\theta}}(T_k^i)$ exist and are finite.

In particular,

$$\begin{aligned} E_{\underline{\theta}}(\underline{T}) &= [\partial_{\theta_i} \psi(\underline{\theta})] = \text{grad}_{\underline{\theta}} \psi(\underline{\theta}) \\ \text{Var}_{\underline{\theta}}(\underline{T}) &= E_{\underline{\theta}} [(\underline{T} - E_{\underline{\theta}}(\underline{T}))(\underline{T} - E_{\underline{\theta}}(\underline{T}))'] = [\partial_{\theta_i \theta_j} \psi(\underline{\theta})] = \mathbf{I}(\underline{\theta}), \end{aligned}$$

where $\psi(\underline{\theta}) = -\log C(\underline{\theta})$ and $\mathbf{I}(\underline{\theta})$ is the Fisher information matrix, namely

$$\mathbf{I}(\underline{\theta}) = E_{\underline{\theta}} \left[(\text{grad}_{\underline{\theta}} \log f_{\underline{\theta}}(\underline{X})) (\text{grad}_{\underline{\theta}} \log f_{\underline{\theta}}(\underline{X}))' \right].$$

Proof. Recall that $C(\underline{\theta})$ is an integration constant which guarantees that the density integrates to one. Hence, by definition,

$$\int_{\mathcal{X}} h(\underline{x}) \exp \left[\sum_{k=1}^K \theta_k T_k(\underline{x}) \right] d\mu(\underline{x}) = \frac{1}{C(\underline{\theta})} = \exp[\psi(\underline{\theta})]$$

Differentiating both sides with respect to θ_i yields

$$\int_{\mathcal{X}} T_i(\underline{x}) h(\underline{x}) \exp \left[\sum_{k=1}^K \theta_k T_k(\underline{x}) \right] d\mu(\underline{x}) = \frac{\partial \psi(\underline{\theta})}{\partial \theta_i} \exp[\psi(\underline{\theta})],$$

which implies that $E_{\underline{\theta}}[T_i(\underline{x})] = \partial_{\theta_i} \psi(\underline{\theta})$ for $i = 1, \dots, K$. Thus $E_{\underline{\theta}}[\underline{T}(\underline{x})] = \text{grad}_{\underline{\theta}} \psi(\underline{\theta})$. Taking the second derivative with respect to θ_j gives

$$\begin{aligned} \int_{\mathcal{X}} T_i(\underline{x}) T_j(\underline{x}) h(\underline{x}) \exp \left[\sum_{k=1}^K \theta_k T_k(\underline{x}) \right] d\mu(\underline{x}) &= \frac{\partial^2 \psi(\underline{\theta})}{\partial \theta_i \partial \theta_j} \exp[\psi(\underline{\theta})] \\ &+ \frac{\partial \psi(\underline{\theta})}{\partial \theta_i} \frac{\partial \psi(\underline{\theta})}{\partial \theta_j} \exp[\psi(\underline{\theta})], \end{aligned}$$

which implies that

$$E_{\underline{\theta}}[T_i(\underline{x}), T_j(\underline{x})] = \frac{\partial^2 \psi(\underline{\theta})}{\partial \theta_i \partial \theta_j} + \frac{\partial \psi(\underline{\theta})}{\partial \theta_i} \frac{\partial \psi(\underline{\theta})}{\partial \theta_j} = \partial_{\theta_i \theta_j} \psi(\underline{\theta}) + \partial_{\theta_i} \psi(\underline{\theta}) \partial_{\theta_j} \psi(\underline{\theta}).$$

However, $E_{\underline{\theta}}[T_i(\underline{x})] = \partial_{\theta_i} \psi(\underline{\theta})$, so that $\text{Cov}_{\underline{\theta}}[T_i(\underline{x}), T_j(\underline{x})] = \partial_{\theta_i \theta_j} \psi(\underline{\theta})$ and $\text{Var}_{\underline{\theta}}[\underline{T}(\underline{x})] = [\partial_{\theta_i \theta_j} \psi(\underline{\theta})]$. In order to show that the covariance coincides with the Fisher information matrix, note that the latter is the variance of

$$\begin{aligned} \text{grad}_{\underline{\theta}} \log f_{\underline{\theta}}(\underline{x}) &= \text{grad}_{\underline{\theta}} \log C(\underline{\theta}) + \text{grad}_{\underline{\theta}} \sum_{k=1}^K \theta_k T_k(\underline{x}) \\ &= \text{grad}_{\underline{\theta}} \log C(\underline{\theta}) + \underline{T}(\underline{x}). \end{aligned}$$

But $\text{grad}_{\underline{\theta}} \log C(\underline{\theta})$ is constant, hence the information matrix reduces to $\mathbf{I}(\underline{\theta}) = \text{Var}_{\underline{\theta}}[\underline{T}(\underline{x})]$. ■

Lemma. Let g be a measurable and bounded function mapping $(\mathcal{X}, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B})$. Then,

$$E_{\underline{\theta}}[g(\underline{X})] = \int_{\mathcal{X}} g(\underline{x}) C(\underline{\theta}) h(\underline{x}) \exp \left[\sum_{k=1}^K \theta_k T_k(\underline{x}) \right] d\mu(\underline{x})$$

exists and is bounded for every $\underline{\theta} \in \Theta$.

4 Distribution free and complete statistics

4.1 Distribution freeness and ancillarity

Definition 1. A statistic \underline{S} is distribution free if, $\forall P_1, P_2 \in \mathcal{P}$, $P_1^{\underline{S}} = P_2^{\underline{S}}$, that is, \underline{S} does not depend on the probability measure $P \in \mathcal{P}$.

Example 1. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ with marginal density f with respect to the Lebesgue measure over $(\mathbb{R}, \mathcal{B})$. Denote the vector of ranks by $\underline{R} = (R_1, \dots, R_n)$, where R_i is the rank associated to X_i among X_1, \dots, X_n . Under a probability zero of occurring ties, \underline{R} takes value in the set of the $n!$ permutations of $(1, \dots, n)$. Hence, \underline{R} has uniform distribution over the $n!$ permutations set irrespective of the density function f of X_i .

Intuitively, a distribution free statistic does not carry any information about $P \in \mathcal{P}$, so that it sounds like the converse of a sufficient statistic, which provides all information available of $P \in \mathcal{P}$. However, these two concepts are not orthogonal in the sense that a sufficient statistic may possess some sort of impurity in the form of a distribution free component. This impurity is called ancillarity and it should be eliminated in order to avoid statistical problems.

Definition 2. A distribution free statistic \underline{S} which is measurable with respect to a minimal sufficient statistic is called an ancillary statistic.

Example 2. Consider the vector (N, X) , where $N - 1 \sim \text{Bin}(1, \frac{1}{2})$ and X has a conditional distribution given $N = n$ that is binomial with parameters (n, p) , $p \in (0, 1)$. Then, N is distribution free since it does not depend on

the parameter of interest p . The joint distribution of (N, X) takes values over $\{(1, 0), (1, 1), (2, 0), (2, 1), (2, 2)\}$, where the counting measure over $(\mathbb{R}^2, \mathcal{B}^2)$ can act as a dominating measure, and has density

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

Note that there is no way of getting rid off N by the factorization criterion, thereof (N, X) is a minimal sufficient statistic with ancillary statistic N .

Remark. Although the ancillary statistic does not provide any information on the parameters of interest, it is usually informative about the sort of information provided by the minimal sufficient statistic.

Example 3. Consider the linear model $Y = \underline{X}\beta + \epsilon$, where $\underline{X} \sim P^{\underline{X}}$ is independent of β and of the error $\epsilon \sim N(0, \sigma^2 \mathbf{I})$. Then, \underline{X} is an ancillary statistic.

Example 4. Consider a Cauchy model with density given by

$$f_\theta = \frac{1}{\pi [1 + (x - \theta)^2]},$$

where θ is a location parameter. Then, $X_{(i)} - X_{(j)} = (X_{(i)} - \theta) - (X_{(j)} - \theta)$ is distribution free and measurable with respect to the order statistic $\underline{X}_{(\cdot)}$ which is minimal sufficient. Thereof $X_{(i)} - X_{(j)}$ is ancillary.

Ancillarity Principle. This principle states that one should condition the minimal sufficient statistic upon the σ -algebra generated by the ancillary statistics. However, there are cases where this procedure cannot be implemented.

4.2 First-order ancillarity and completeness

Definition 1. Consider a minimal sufficient statistic \underline{T} for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ and a \underline{T} -measurable statistic \underline{S} which is also \mathcal{P} -integrable. Then, \underline{S} is first-order ancillary if, $E_{P_1}(\underline{S}) = E_{P_2}(\underline{S}), \forall P_1, P_2 \in \mathcal{P}$.

Trivial Example. A statistic which is \mathcal{P} -almost surely constant is first-order ancillary.

Definition 2. A model is complete if $E_P(\underline{S}) = \underline{c}, \forall P \in \mathcal{P}$, implies $\underline{S} = \underline{c}$ \mathcal{P} -almost surely, where \underline{c} is a constant vector and \underline{S} is an ancillary statistic.

Definition 3. A statistic $\underline{S}: (\mathcal{X}, \mathcal{A}) \mapsto (\mathcal{S}, \mathcal{B}_S)$ is complete if its induced model $(\mathcal{S}, \mathcal{B}_S, \mathcal{P}^S)$ is complete.

Note that completeness rules out the existence of nontrivial first-order ancillarity and there is no loss of generality in taking $\underline{c} = 0$. Accordingly, a minimal sufficient statistic \underline{T} is also complete if the ancillarity of its induced model $(\mathcal{T}, \mathcal{B}_T, \mathcal{P}^T)$ stems solely from a trivial first order ancillary statistic.

Proposition 1. In an exponential model of full rank, the privileged statistic \underline{T} is minimal sufficient and complete.

Proposition 2. Assume that \underline{T} is sufficient, complete, and \mathcal{P} -integrable. Then, \underline{T} is minimal sufficient.

Proof. Assume, for instance, that \underline{T} is not minimal sufficient. Then, there exists a sufficient statistic \underline{T}^* which is \underline{T} -measurable whereas \underline{T} is not \underline{T}^* -measurable. Let $\psi(\underline{X}) = \underline{T}(\underline{X}) - E[\underline{T}(\underline{X}) | \underline{T}^*]$. Since \underline{T}^* is sufficient, ψ does not depend on $P \in \mathcal{P}$ and so, it is a \underline{T} -measurable statistic.

$$\begin{aligned} E_P[\psi(\underline{X})] &= E_P\{\underline{T}(\underline{X}) - E[\underline{T}(\underline{X}) | \underline{T}^*]\} \\ &= E_P[\underline{T}(\underline{X})] - E_P[\underline{T}(\underline{X})] = 0, \quad \forall P \in \mathcal{P}, \end{aligned}$$

which implies that $\psi(\underline{X}) = 0$ \mathcal{P} -almost surely due to the fact that \underline{T} is complete. Hence, $\underline{T}(\underline{X}) = E[\underline{T}(\underline{X}) | \underline{T}^*]$ \mathcal{P} -almost surely, which means that \underline{T} is \underline{T}^* -measurable. Contradiction. ■

Basu's Theorem. Consider a sufficient and complete statistic \underline{T} and a distribution free statistic \underline{S} . Then, \underline{S} and \underline{T} are P -independent for all $P \in \mathcal{P}$.

Proof. The distribution freeness of \underline{S} implies that P^S does not depend on $P \in \mathcal{P}$. Hence, $P^S(B) = P[\underline{S}^{-1}(B)]$ is a constant over \mathcal{P} , $\forall B \in \mathcal{B}_S$. However, by the law of iterated expectations, $P^S(B) = E_P\{P[\underline{S}^{-1}(B) | T]\}$, which implies that

$$E_P\{P[\underline{S}^{-1}(B) | T] - P[\underline{S}^{-1}(B)]\} = 0, \quad \forall P \in \mathcal{P}.$$

But \underline{T} is sufficient, so that the above expectation does not depend on a particular choice of $P \in \mathcal{P}$, that is, it is a \underline{T} -measurable statistic. Since \underline{T} is also complete, $P[\underline{S}^{-1}(B) | \underline{T}] = P[\underline{S}^{-1}(B)]$ \mathcal{P} -almost surely, which is a form of stating \mathcal{P} -independence between the two statistics. ■

Corollary (Fisher's Lemma). Consider a random sample $\underline{X} = (X_1, \dots, X_n)$, $X_i \sim N(\mu, \sigma^2)$. Assume that σ^2 is fixed, so that the sample mean \bar{X} is a privileged statistic and thereof minimal sufficient and complete. Then, the distribution free statistic $S = ns^2/\sigma^2 \sim \chi_{n-1}^2$ is P_{μ, σ^2} -independent of \bar{X} for all $P_{\mu, \sigma^2} \in \mathcal{P}$.

Definition 4. Two measures $P_1, P_2 \in \mathcal{P}$ are disconnected if, for every N_1, N_2 such that $P_1(N_1) = P_2(N_2) = 0$, $N_1^c \cap N_2^c = \emptyset$.

Proposition 3. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be such that, $\forall P_1, P_2 \in \mathcal{P}$, P_1 and P_2 are not disconnected. Consider two \mathcal{P} -independent statistics \underline{S} and \underline{T} . Then, if \underline{T} is sufficient, \underline{S} is distribution free.

Proof. The \mathcal{P} -independence of \underline{S} and \underline{T} means that, for all $P \in \mathcal{P}$, $P(A | \underline{S}) = P(A)$ P -almost surely, for every $A \in \mathcal{A}_{\mathcal{T}} = \underline{T}^{-1}(\mathcal{B}_{\mathcal{T}})$ as well as $P(B | \underline{T}) = P(B)$ for every $B \in \mathcal{B}_{\mathcal{S}} = \underline{S}^{-1}(\mathcal{B}_{\mathcal{S}})$. However, \underline{T} is sufficient which implies that, for all $B \in \mathcal{B}_{\mathcal{S}}$ and $P_1, P_2 \in \mathcal{P}$, there is a version of the conditional probability such that $P_1[\underline{S}^{-1}(B) | \underline{T}] = P_2[\underline{S}^{-1}(B) | \underline{T}]$, P_1, P_2 -almost surely. Thus,

$$P_1[\underline{S}^{-1}(B)] \stackrel{P_1\text{-a.s.}}{=} P_1[\underline{S}^{-1}(B) | \underline{T} = \underline{t}] \stackrel{P_1, P_2\text{-a.s.}}{=} P_2[\underline{S}^{-1}(B) | \underline{T} = \underline{t}] \\ \stackrel{P_2\text{-a.s.}}{=} P_2[\underline{S}^{-1}(B)],$$

as long as $\underline{t} \in N_1^c \cap N_2^c$, for all N_i such that $P_i(N_i) = 0$, $i = 1, 2$. As P_1 and P_2 are not disconnected, $N_1^c \cap N_2^c \neq \emptyset$ and so, there exists $\underline{t} \in N_1^c \cap N_2^c$. Hence, \underline{S} is distribution free. ■

5 Point estimation

5.1 The role of sufficiency

Consider the problem of estimating a functional ℓ , which maps \mathcal{P} into \mathbb{R}^k . The loss function $L_P(\underline{t})$, $\underline{t} \in \mathbb{R}^k$, denotes the loss incurred if the true distribution is P and the estimate of $\ell(P)$ takes values \underline{t} . A loss function must satisfy three requirements:

- (i); $\underline{t} = \ell(P) \implies L_P(\underline{t}) = 0$
- (ii) $\underline{t} \neq \ell(P) \implies L_P(\underline{t}) > 0$
- (iii) $L_P(\underline{t})$ is a convex function over \mathbb{R}^k

For instance, the quadratic loss $L_P(\underline{t}) = [\underline{t} - \ell(P)]' A [\underline{t} - \ell(P)]$, $\underline{t} \in \mathbb{R}^k$, satisfies these assumptions as long as A is a positive definite matrix.

Denote by $\underline{\delta}$ an estimator of $\ell(P)$, which maps \mathcal{X} into \mathbb{R}^k , and consider a statistic \underline{T} which is sufficient for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Then, the Rao-Blackwellization of $\underline{\delta}$, $\underline{\delta}^T = E(\underline{\delta} | \underline{T})$, does not depend on a particular P , and so it stands for an alternative statistic for estimating $\ell(P)$.

Rao-Blackwell's Theorem. Assume convexity for the loss function $L_P(\cdot)$ irrespective to $P \in \mathcal{P}$. Then, $R_P^{\delta^T} \leq R_P^\delta$, $\forall P \in \mathcal{P}$, where R_P^δ and $R_P^{\delta^T}$ represent the risks associated with $\underline{\delta}$ and its Rao-Blackwellization $\underline{\delta}^T$, respectively.

Proof. By Jensen's inequality,

$$\begin{aligned} R_P^{\delta^T} &= E_P \left[L_P \left(\underline{\delta}^T \right) \right] = E_P \left(L_P \left[E \left(\underline{\delta} \mid \underline{T} \right) \right] \right) \\ &\leq E_P \left(E \left[L_P(\underline{\delta}) \mid \underline{T} \right] \right) = E_P \left[L_P(\underline{\delta}) \right] = R_P^\delta, \end{aligned}$$

completing the proof. ■

Lehmann-Scheffé's Theorem. Let $\underline{\delta}$ be an unbiased estimator of $\ell(P)$ and \underline{S} a sufficient and complete statistic for $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Then, the Rao-Blackwellization $\underline{\delta}^S = E(\underline{\delta} | \underline{S})$ is a uniformly minimal risk unbiased (UMRU) estimator for any convex loss function $L_P(\cdot)$, that is, for every $\underline{\delta}_*$ such that $E_P(\underline{\delta}_*) = \ell(P)$,

$\forall P \in \mathcal{P}$, $R_P^{\delta^S} \leq R_P^{\delta_*}$. Moreover, $\underline{\delta}^S$ is essentially unique in the sense that $\underline{\delta}^S = \underline{\delta}_*^S$ \mathcal{P} -almost surely for every unbiased estimators $\underline{\delta}$ and $\underline{\delta}_*$.

Proof. Take two unbiased estimators $\underline{\delta}$ and $\underline{\delta}_*$ of $\ell(P)$. Then, their Rao-Blackwellisations

$$\begin{aligned} E_P(\underline{\delta}^S) &= E_P[E(\underline{\delta} | \underline{S})] = E_P(\underline{\delta}) = \ell(P) \\ E_P(\underline{\delta}_*^S) &= E_P[E(\underline{\delta}_* | \underline{S})] = E_P(\underline{\delta}_*) = \ell(P) \end{aligned}$$

are also unbiased estimators for all $P \in \mathcal{P}$. Hence, $E_P(\underline{\delta}^S - \underline{\delta}_*^S) = E_P(\underline{\delta}^S) - E_P(\underline{\delta}_*^S) = 0$, $\forall P \in \mathcal{P}$, which implies, by completeness of \underline{S} , that $\underline{\delta}^S = \underline{\delta}_*^S$ \mathcal{P} -almost surely, i.e. the Rao-Blackwellization is essentially unique. Moreover, by Rao-Blackwell's Theorem, $R_P^{\delta^S} = R_P^{\delta_*^S} \leq R_P^{\delta_*}$, $\forall P \in \mathcal{P}$, which implies that $\underline{\delta}^S$ is UMRU. ■

Under these circumstances, given a known sufficient and complete statistic \underline{S} , there are two strategies for constructing UMRU estimators. The first consists in finding an unbiased estimator which is \underline{S} -measurable. The second is to consider a very simple unbiased estimator $\underline{\delta}$ and compute its Rao-Blackwellization $\underline{\delta}^S$.

Example 1. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ with each $X_i \sim N(\mu, \sigma^2)$. Then, it is known that the statistic $\underline{S} = (\bar{X}, s^2)$ is minimal sufficient and complete. Suppose we are looking for UMRU estimators for μ and σ^2 . Then, \bar{X} and $\tilde{s}^2 = \frac{n}{n-1}s^2$ are \underline{S} -measurable and unbiased, therefore UMRU estimators (first strategy). Alternatively, we can consider the trivial estimator X_i for μ , which is unbiased though not \underline{S} -measurable. Then,

$$E(X_1 | \bar{X}) = \dots = E(X_i | \bar{X}) = \dots = E(X_n | \bar{X}),$$

which implies that

$$nE(X_i | \bar{X}) = \sum_{i=1}^n E(X_i | \bar{X}) = E\left(\sum_{i=1}^n X_i | \bar{X}\right).$$

But $\sum_{i=1}^n X_i$ is \bar{X} -measurable, hence

$$E(X_i | \bar{X}) = \frac{1}{n}E\left(\sum_{i=1}^n X_i | \bar{X}\right) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X},$$

retrieving the sample mean as the UMRU estimator.

Example 2. Consider the same setup, but we are now interested on estimating μ^2 . The first estimator that comes in mind is \bar{X}^2 . However,

$$E(\bar{X}^2) = \text{Var}(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2,$$

hence it is biased. Nevertheless, it indicates that $\bar{X}^2 - \tilde{s}^2/n$ stands for an unbiased estimator. Note however that it is also \underline{S} -measurable, hence it is UMRU.

Example 3. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ with each $X_i \sim \text{Bin}(1, p)$, $p \in (0, 1)$. Suppose we want to estimate $1/p$. Then, let $\delta(\underline{X})$ be an unbiased estimator for $1/p$. However, it is well-known that $\sum_{i=1}^n X_i \sim \text{Bin}(n, p)$, hence

$$E_p[\delta(\underline{X})] = \sum_{t=0}^n \binom{n}{t} p^t (1-p)^{n-t} \delta(t) = \frac{1}{p}$$

for all $p \in (0, 1)$. In the case that $p \rightarrow 0$, the right-hand side goes to infinite, but the left-hand side converges to $\delta(0)$. But $\delta(0) \neq \infty$, otherwise the expectation of the estimator would be always infinity, which implies a contradiction. Thus there is no unbiased estimator for $1/p$ and, in this case, the two strategies fail.

Example 4. Consider the capture-recapture model, where we want to estimate the number of fishes, say N , dwelling in a lake. A possible experiment is to capture, mark, and release a random sample of k fishes, and then draw another sample of size n (with replacement) in order to count how many were marked. Denote by X_i the random variable that assume value one when the fish i was marked, and zero otherwise. Then, $\sum_{i=1}^n X_i \sim \text{Bin}(n, p = k/N)$. Given that k is known, in order to estimate N , it is clear that the parameter of interest is $1/p$. However, there is no unbiased estimator available. An alternative experiment consists in replacing the second stage of the first experiment in the following way. Instead of choosing a sample size n , we sample (with replacement) until the number of marked fishes reaches some arbitrary value m . This sampling scheme induces a negative binomial distribution with parameters m and k/N

for the total number Y of fishes captured in the second stage. The negative binomial belongs to the exponential family, so that its privileged statistic Y is minimal sufficient and complete. In this way, we only need to find a unbiased estimator $\delta(Y)$ for N . However, it can be shown that

$$E(Y) = m \frac{1 - k/N}{k/N} = m \frac{1 - p}{p},$$

which indicates that $\delta(Y) = (Y + m)/m$ is an UMRU estimator of $1/p$, since

$$\begin{aligned} E[\delta(Y)] &= E\left[\frac{Y + m}{m}\right] = E\left[1 + \frac{Y}{m}\right] = 1 + \frac{E(Y)}{m} = 1 + \frac{m(1 - p)}{mp} \\ &= 1 + \frac{1 - p}{p} = \frac{p + 1 - p}{p} = \frac{1}{p}. \end{aligned}$$

5.2 Semiparametric context

Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ and a nonparametric family of densities \mathcal{P} . A white noise model then corresponds to $\mathcal{P} = \mathcal{F}$, where \mathcal{F} is the set containing all continuous density functions with respect to the Lebesgue measure over $(\mathbb{R}, \mathcal{B})$. Accordingly, a symmetric white noise model corresponds to $\mathcal{P} = \mathcal{F}_+$, where \mathcal{F}_+ stands for the set of continuous density functions with respect to the Lebesgue measure over $(\mathbb{R}, \mathcal{B})$ which are symmetric around zero.

For the white noise model, the order statistic $\underline{X}_{(\cdot)}^{(n)} = (X_{(1)}, \dots, X_{(n)})$ is trivially sufficient by the factorization criterion. Moreover, after some boring algebra, it can be shown that $\underline{X}_{(\cdot)}^{(n)}$ is also a complete statistic. Analogously, for the symmetric white noise model, $|\underline{X}|_{(\cdot)}^{(n)} = (|X|_{(1)}, \dots, |X|_{(n)})$ is sufficient and complete. Under these circumstances, by Lehmann-Scheffé's Theorem, an UMRU estimator for any quantity $\psi(f)$, where $f \in \mathcal{F}$ is the joint-density function of \underline{X} , is an unbiased $\underline{X}_{(\cdot)}^{(n)}$ -measurable estimator. In the same way, if $f \in \mathcal{F}_+$, then an UMRU estimator for $\psi(f)$ should be unbiased and $|\underline{X}|_{(\cdot)}^{(n)}$ -measurable.

Example. Consider the family of measures \mathcal{F} containing every density function f with respect to the Lebesgue measure such that $\int x f(x) d\mu(x) < \infty$. Then, the sample mean $\bar{X} = \sum_{i=1}^n X_i$, due to its unbiasedness and $\underline{X}_{(\cdot)}^{(n)}$ -measurability, is an UMRU estimator for $\mu_f = \int x f(x) d\mu(x)$.

However, it is not always the case that we can immediately think of an UMRU estimator. In more complex and general situations, we may need a Rao-Blackwellization scheme, which boils down to the concept of U-statistics.

Definition. Consider a functional $\Psi : \mathcal{F} \rightarrow \mathbb{R}$, and the smallest sample size k such that $\Psi(f)$ can be estimated in an unbiased way, that is, there exists a function $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $E \left[\psi \left(\underline{X}^{(n)} \right) \right] = \Psi(f)$ for every $n \geq k$, where $\underline{X}^{(n)} = (X_1, \dots, X_n)$ stands for a random sample with $X_i \sim f \in \mathcal{F}$. Then the function ψ is called a kernel of order k for the estimation of Ψ .

For instance, for the estimation of the mean based on a random sample, only one observation ($k = 1$) is needed for constructing an unbiased estimator: $\psi(X_1) = X_1$. For the estimation of the variance, two observations ($k = 2$) are needed in order to have an unbiased estimation: $\psi(X_1, X_2) = \frac{1}{2}(X_2 - X_1)^2$.

Proposition. Consider a random sample X_1, \dots, X_n with marginal density $f \in \mathcal{F}$, and a kernel ψ of order k for estimating $\Psi(f)$. Then, the U-statistic

$$\frac{(n-k)!}{n!} \sum_{1 \leq i_1 \neq \dots \neq i_k \leq n} \psi(X_{i_1}, \dots, X_{i_k})$$

is an UMRU estimator for $\Psi(f)$.

Proof. By definition, $\psi(X_1, \dots, X_k)$ is an unbiased estimator for Ψ . Then, its Rao-Blackwellization $E \left[\psi(X_1, \dots, X_k) \mid X_{(1)}, \dots, X_{(n)} \right]$ is an UMRU estimator. However, the joint distribution of X_1, \dots, X_k given $X_{(1)}, \dots, X_{(n)}$ is equivalent, with probability one, to a discrete uniform distribution over the $n(n-1) \dots (n-k+1)$ possible distinct k -tuples of the form X_{i_1}, \dots, X_{i_k} . Hence, the U-statistic turns out to be the Rao-Blackwellization of ψ . ■

Example. Consider again the problem of estimating the variance based on a random sample X_1, \dots, X_n . For estimating $\Psi(f) = \sigma_f^2 = \int (x - \mu_f)^2 f(x) d\mu(x)$, the Rao-Blackwellization of the kernel $\psi(X_1, X_2) = \frac{1}{2}(X_2 - X_1)^2$ reads

$$\begin{aligned} \hat{\sigma}_f^2 &= \frac{(n-2)!}{n!} \sum_{1 \leq i \neq j \leq n} \frac{1}{2} (X_i - X_j)^2 = \frac{1}{2n(n-1)} \sum_{1 \leq i \neq j \leq n} (X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)} \sum_{1 \leq i \neq j \leq n} (X_i^2 - 2X_i X_j + X_j^2) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2n(n-1)} \left[(n-1) \sum_{i=1}^n X_i^2 + (n-1) \sum_{j=1}^n X_j^2 - 2 \sum_{1 \leq i \neq j \leq n} X_i X_j \right] \\
&= \frac{1}{2n(n-1)} \left[2(n-1) \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n \sum_{j=1}^n X_i X_j + 2 \sum_{i=1}^n X_i^2 \right] \\
&= \frac{1}{2n(n-1)} \left[2n \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i \sum_{j=1}^n X_j \right] \\
&= \frac{1}{2n(n-1)} \left[2n \sum_{i=1}^n X_i^2 - 2(n\bar{X})^2 \right] \\
&= \frac{1}{(n-1)} \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] = \frac{n}{n-1} s^2.
\end{aligned}$$

For the symmetric white noise model, the U-statistic consists in

$$\frac{1}{2^k} \frac{(n-k)!}{n!} \sum_{(s_1, \dots, s_k) \in \{-1, 1\}^k} \sum_{1 \leq i_1 \neq \dots \neq i_k \leq n} \psi(s_1 |X_{i_1}|, \dots, s_k |X_{i_k}|),$$

where the term 2^{-k} stems from the fact that s_1, \dots, s_k are uniformly distributed over the 2^k elements of $\{-1, 1\}^k$. Note that considering all permutations over $\{-1, 1\}^k$ implies $\psi(s_1 |X_{i_1}|, \dots, s_k |X_{i_k}|) = \psi(s_1 X_{i_1}, \dots, s_k X_{i_k})$.

5.3 Group invariance

Consider two probability measures P_1 and P_2 belonging to \mathcal{P} . The identifiability condition holds if $P_1 \neq P_2$. Let $\mathcal{G}, \cdot = \{g\}, \cdot$ be a group of measurable transformations of $(\mathcal{X}, \mathcal{A})$, i.e. g is a mapping from \mathcal{X} to \mathcal{X} .

Definition 1. The model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is invariant under the group of transformation \mathcal{G} if there exists a $Q \in \mathcal{P}$ such that $Q^{\underline{X}} = P^{g\underline{X}}$ for every $g \in \mathcal{G}$ and for all $P \in \mathcal{P}$. If the identifiability condition holds, then Q is (essentially) unique.

Notation. Henceforth, Q will be denoted by $\bar{g}P$. The transformation \bar{g} is a mapping from \mathcal{P} to \mathcal{P} , which belongs to the so-called induced group $\bar{\mathcal{G}}$. If the induced group has only one orbit in \mathcal{P} , i.e. \mathcal{P} itself, then \mathcal{G} and $\bar{\mathcal{G}}$ are called generating groups. In other words, if there exists a $\bar{g} \in \bar{\mathcal{G}}$ such that $P_2 = \bar{g}P_1$ for any $P_1, P_2 \in \mathcal{P}$, then \mathcal{G} is a generating group.

Consider the parametric case $\mathcal{P} = \{P_{\underline{\theta}} \mid \underline{\theta} \in \Theta\}$, for instance. Then, invari-

ance under \mathcal{G} means that, for every $\underline{\theta} \in \Theta$ and for all $g \in \mathcal{G}$, there exists a $\underline{\theta}' = \bar{g}\underline{\theta}$ such that $P_{\underline{\theta}}^{g\underline{X}} = P_{\underline{\theta}'}^{\underline{X}} = P_{\bar{g}\underline{\theta}}^{\underline{X}}$. The induced group $\bar{\mathcal{G}}$ acting on Θ is then characterized by $\bar{g} : \underline{\theta} \mapsto \underline{\theta}' = \bar{g}\underline{\theta}$ such that $P_{\bar{g}\underline{\theta}}(\underline{X} \in A) = P_{\underline{\theta}}(g\underline{X} \in A)$ for any $A \in \mathcal{A}$. Finally, as $\bar{\mathcal{G}}$ has only one orbit, Θ itself, it is a generating group.

Example 1. Consider the location family $\mathcal{P} = \{P_{\theta} \mid \theta \in \mathbb{R}\}$, P_{θ} on $(\mathbb{R}^n, \mathcal{B}^n)$ such that $P_{\theta}^{\underline{X}}(\underline{x}) = P_{\theta}^{\underline{X}}(x_1, \dots, x_n) = P_0^{\underline{X}}(x_1 - \theta, \dots, x_n - \theta)$, where P_0 is a fixed parent distribution. Then, the groups $\mathcal{G} = \{g_a \underline{x} = (x_1 + a, \dots, x_n + a) \mid a \in \mathbb{R}\}$ and $\bar{\mathcal{G}} = \{\bar{g}_a \theta = \theta + a \mid a \in \mathbb{R}\}$ are generating groups.

$$\begin{aligned} P_{\theta}^{g_a \underline{X}}(\underline{x}) &= P_{\theta}(g_a \underline{X} \leq \underline{x}) = P_{\theta}(X_1 + a \leq x_1, \dots, X_n + a \leq x_n) \\ &= P_0(X_1 + a \leq x_1 - \theta, \dots, X_n + a \leq x_n - \theta) \\ &= P_0(X_1 \leq x_1 - \theta - a, \dots, X_n \leq x_n - \theta - a) \\ &= P_{\theta+a}(X_1 \leq x_1, \dots, X_n \leq x_n) = P_{\theta+a}(\underline{X} \leq \underline{x}) \\ &= P_{\bar{g}_a \theta}^{\underline{X}}(\underline{x}). \end{aligned}$$

Example 2. Consider a scale-location family $\mathcal{P} = \{P_{\theta_1, \theta_2} \mid \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}_+\}$, P_{θ_1, θ_2} on $(\mathbb{R}^n, \mathcal{B}^n)$ such that

$$P_{\theta_1, \theta_2}^{\underline{X}}(x_1, \dots, x_n) = P_{0,1}^{\underline{X}}\left(\frac{x_1 - \theta_1}{\theta_2}, \dots, \frac{x_n - \theta_1}{\theta_2}\right),$$

where $P_{0,1}$ is a fixed parent distribution. Then,

$$\begin{aligned} \mathcal{G} &= \left\{g_{a,b}(x_1, \dots, x_n) = (a + bx_1, \dots, a + bx_n) \mid a \in \mathbb{R}, b \in \mathbb{R}_+\right\} \\ \bar{\mathcal{G}} &= \left\{\bar{g}_{a,b}(\theta_1, \theta_2) = (a + b\theta_1, b\theta_2) \mid a \in \mathbb{R}, b \in \mathbb{R}_+\right\} \end{aligned}$$

are generating groups.

$$\begin{aligned} P_{\theta_1, \theta_2}^{g_{a,b} \underline{X}}(\underline{x}) &= P_{\theta_1, \theta_2}(g_{a,b} \underline{X} \leq \underline{x}) \\ &= P_{\theta_1, \theta_2}(bX_1 + a \leq x_1, \dots, bX_n + a \leq x_n) \\ &= P_{\theta_1, \theta_2}^{\underline{X}}\left(\frac{x_1 - a}{b}, \dots, \frac{x_n - a}{b}\right) \\ &= P_{0,1}^{\underline{X}}\left(\frac{x_1 - (a + b\theta_1)}{b\theta_2}, \dots, \frac{x_n - (a + b\theta_1)}{b\theta_2}\right) \\ &= P_{a+b\theta_1, b\theta_2}^{\underline{X}}(\underline{x}) = P_{\bar{g}_{a,b} \underline{\theta}}^{\underline{X}}(\underline{x}). \end{aligned}$$

Example 3. Consider the general linear model $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{e}$, where $\underline{\beta} \in \mathbb{R}^k$ and the components of \underline{e} are independent and normally distributed with mean zero and known variance σ^2 . Then, for $\underline{v} = \mathbf{X}\underline{b}$ belonging to the space $\mathcal{V}(\mathbf{X})$ spanned by the columns of \mathbf{X} ,

$$\begin{aligned}\mathcal{G} &= \left\{ g_v \underline{y} = \underline{y} + \underline{v} \mid \underline{v} \in \mathcal{V}(\mathbf{X}) \right\} = \left\{ g_b \underline{y} = \underline{y} + \mathbf{X}\underline{b} \mid \underline{b} \in \mathbb{R}^k \right\} \\ \bar{\mathcal{G}} &= \left\{ \bar{g}_b \underline{\beta} = \underline{\beta} + \underline{b} \mid \underline{b} \in \mathbb{R}^k \right\}\end{aligned}$$

are generating groups. Trivially, $g_b \underline{Y} = \underline{Y} + \mathbf{X}\underline{b} = \mathbf{X}(\underline{\beta} + \underline{b}) + \underline{e} = \mathbf{X}\bar{g}_b \underline{\beta} + \underline{e}$.

Example 4. Consider again a nonparametric context where the random sample X_1, \dots, X_n has marginal distribution function F . Then, the order-preserving group

$$\mathcal{G} = \left\{ \underline{g}(\underline{x}) = (g(x_1), \dots, g(x_n)) \right\},$$

where g is an increasing monotonic transformation, is a generating group. Note that there exists a function g such that $g(X_1), \dots, g(X_n)$ are independent with marginal distribution G . In particular, X_1, \dots, X_n constitute a random sample with distribution F if and only if $F(X_1), \dots, F(X_n)$ are independent and uniformly distributed over the unity interval. In the same way, this is equivalent to say that $G^{-1}F(X_1), \dots, G^{-1}F(X_n)$ are independent with marginal distribution G .

Definition 2. Let $\mathcal{P} = \{P_\theta \mid \theta \in \mathbb{R}\}$ be generated by a translation group. Then, a location parameter ψ mapping \mathcal{P} into \mathbb{R} is such that

$$\psi \left(P_{\theta}^{\underline{X} + \iota a} \right) = \psi \left(P_{\theta+a}^{\underline{X}} \right) = \psi \left(P_{\theta}^{\underline{X}} \right) + a,$$

where $\iota = (1, \dots, 1)'$.

Definition 3. Let $\mathcal{P} = \{P_{\theta_1, \theta_2} \mid \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}_+\}$ be generated by a scale-translation group. Then, a scale parameter ψ mapping \mathcal{P} into \mathbb{R}_+ is such that

$$\psi \left(P_{\theta_1, \theta_2}^{b\underline{X} + \iota a} \right) = \psi \left(P_{a+b\theta_1, b\theta_2}^{\underline{X}} \right) = b\psi \left(P_{\theta_1, \theta_2}^{\underline{X}} \right).$$

5.4 Principle of equivariance

Assume a model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{P} = \{P_{\underline{\theta}} \mid \underline{\theta} \in \Theta\}$, which is invariant to a group \mathcal{G} of transformation. Suppose that we want to estimate the parameter vector $\underline{\theta}$ under an invariant loss function, that is to say, $\forall \underline{t}_1, \underline{t}_2 \in \Theta, L(\bar{g}\underline{t}_1, \bar{g}\underline{t}_2) = L(\underline{t}_1, \underline{t}_2)$. The principle of equivariance states that it is preferable to possess an equivariant estimator $\underline{T}(\underline{X})$, i.e. $\underline{T}(g\underline{X}) = \bar{g}\underline{T}(\underline{X})$.

Proposition. Consider an equivariant estimator $\hat{\underline{\theta}}$ and the corresponding risk function $R(\underline{\theta}, \hat{\underline{\theta}}) = E_{\underline{\theta}} [L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X}))]$, where L is an invariant loss function. Then, $R(\cdot, \hat{\underline{\theta}})$ is constant along the orbits of $\bar{\mathcal{G}}$.

Proof. For every $\bar{g} \in \bar{\mathcal{G}}$ and $\underline{\theta} \in \Theta$,

$$\begin{aligned} R(\bar{g}\underline{\theta}, \hat{\underline{\theta}}) &\equiv E_{\bar{g}\underline{\theta}} [L(\bar{g}\underline{\theta}, \hat{\underline{\theta}}(\underline{X}))] = E_{\underline{\theta}} [L(\bar{g}\underline{\theta}, \hat{\underline{\theta}}(g\underline{X}))] \\ &= E_{\underline{\theta}} [L(\bar{g}\underline{\theta}, \bar{g}\hat{\underline{\theta}}(\underline{X}))] = E_{\underline{\theta}} [L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X}))] \\ &= R(\underline{\theta}, \hat{\underline{\theta}}). \quad \blacksquare \end{aligned}$$

Corollary. Consider the same setting, but assume in addition that \mathcal{G} is a generating group. Then, $R(\underline{\theta}, \hat{\underline{\theta}}) = R(\hat{\underline{\theta}})$, i.e. the risk function is constant and does not depend on $\underline{\theta}$.

Proof. It is a trivial application of the proposition given that $\bar{\mathcal{G}}$, as a generating group, has only one orbit. \blacksquare

This is very important since we will be able to compare two estimators $\hat{\underline{\theta}}_1$ and $\hat{\underline{\theta}}_2$ by looking at their corresponding risk functions. Instead of using the not always reasonable concept of UMRU estimator, we will be more concerned with the uniformly minimum risk equivariant (UMRE) estimator.

5.5 Invariants and maximal invariants

Definition 1. Two elements $x, x' \in \mathcal{X}$ are equivalent, $x \sim x'$, if and only if there exists a transformation $g \in \mathcal{G}$ such that $x' = gx$. In other words, x is equivalent to x' if and only if they are in the same orbit.

Definition 2. Note that the orbits of \mathcal{G} in $(\mathcal{X}, \mathcal{A})$ constitute measurable partitions of \mathcal{X} . The smallest σ -algebra $\mathcal{A}_{\mathcal{G}} \subseteq \mathcal{A}$ which contains all orbits of \mathcal{G} is denominated the invariant σ -algebra.

Definition 3. A measurable function T defined on $(\mathcal{X}, \mathcal{A})$ is an invariant if $x \sim x'$ implies $T(x) = T(x')$. Putting in a different way, T is invariant if and only if it is $\mathcal{A}_{\mathcal{G}}$ -measurable. Note that an invariant is constant over each orbit.

Definition 4. A measurable function T defined on $(\mathcal{X}, \mathcal{A})$ is a maximal invariant if $T(x) = T(x')$ is a necessary and sufficient condition for $x \sim x'$, that is, $\mathcal{A}_{\mathcal{G}} = \mathcal{A}_T$.

Note that a maximal invariant is not only constant over each orbit, but it also associates a different value for each orbit. It works as if it were putting a different label in each orbit.

Trivial Consequence. If T is maximal invariant ($\mathcal{A}_T = \mathcal{A}_{\mathcal{G}}$), then a measurable function S is invariant if and only if it is T -measurable.

Proposition. If \mathcal{G} is a generating group, then invariance of S also implies distribution freeness.

Proof. For every orbit $A \in \mathcal{A}_{\mathcal{G}}$ and $\bar{g} \in \bar{\mathcal{G}}$,

$$P_{\underline{\theta}}^{\underline{X}}(A) = P_{\underline{\theta}}(\underline{X} \in A) = P_{\underline{\theta}}(\bar{g}\underline{X} \in A) = P_{\bar{g}\underline{\theta}}(\underline{X} \in A),$$

so that it does not depend on a particular $\underline{\theta}$. However, the invariance of S implies $\mathcal{A}_{\mathcal{G}}$ -measurability, hence $S^{-1}(B) \in \mathcal{A}_{\mathcal{G}}$ for $B \in \mathcal{B}_S$. Thus,

$$P_{\underline{\theta}}[S(\underline{X}) \in B] = P_{\underline{\theta}}[\underline{X} \in S^{-1}(B)] = P_{\bar{g}\underline{\theta}}[\underline{X} \in S^{-1}(B)] = P_{\bar{g}\underline{\theta}}[S(\underline{X}) \in B],$$

completing the proof. ■

Example 1. Consider a location model where the orbits are of the form $\{(x_1 + a, \dots, x_n + a) \mid a \in \mathbb{R}\}$ in \mathbb{R}^n . Then, $(X_2 - X_1, X_3 - X_1, \dots, X_n - X_1)$ and $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_{n-1} - \bar{X})$ are, for instance, maximal invariants.

5.6 Pitman estimators

Analogously to the Rao-Blackwellization procedure to produce UMRU estimators, there is a scheme for constructing UMRE estimators under a quadratic loss function. Recall that an UMRU estimator was achieved by taking the conditional expectation of an unbiased estimator given a sufficient statistic. In the present context, an UMRE estimator relates to the conditional bias of an equivariant estimator given a maximal invariant of the model. This estimator is denominated the Pitman estimator.

Consider, for instance, a location context where the random sample $\underline{X} = (X_1, \dots, X_n)$ has marginal density f_θ . One maximal invariant for the location model is, for instance, $T(\underline{X}) = (X_2 - X_1, X_3 - X_1, \dots, X_n - X_1)$. Consider now two equivariant estimators $\check{\theta}(\underline{X})$ and $\hat{\theta}(\underline{X})$. Then,

$$\begin{aligned} \check{\theta}(\underline{X}) - \hat{\theta}(\underline{X}) &= [\check{\theta}(\underline{X}) + a] - [\hat{\theta}(\underline{X}) + a] \\ &= \bar{g}\check{\theta}(\underline{X}) - \bar{g}\hat{\theta}(\underline{X}) \\ &= \check{\theta}(g\underline{X}) - \hat{\theta}(g\underline{X}), \end{aligned}$$

for every $g \in \mathcal{G}$, which implies that $\check{\theta}(\underline{X}) - \hat{\theta}(\underline{X})$ is invariant with respect to \mathcal{G} , and so it is measurable with respect to any maximal invariant. In particular, $\check{\theta}(\underline{X}) - \hat{\theta}(\underline{X})$ is measurable with respect to $T(\underline{X})$. Thus,

$$\check{\theta}(\underline{X}) = \hat{\theta}(\underline{X}) + \psi[T(\underline{X})],$$

where the function ψ can be chosen in order to minimize the risk function.

Under a quadratic loss function,

$$R(\check{\theta}) = E_\theta [\check{\theta}(\underline{X}) - \theta]^2 = E_\theta [\psi[T(\underline{X})] + \hat{\theta}(\underline{X}) - \theta]^2.$$

However, this is a well known problem with solution

$$\psi^*(\underline{X}) = \operatorname{argmin}_{\psi \in \mathcal{V}} E_\theta [\psi[T(\underline{X})] - \theta + \hat{\theta}(\underline{X})]^2 = -E_\theta [\hat{\theta}(\underline{X}) - \theta | T(\underline{X})],$$

where \mathcal{V} is the set of functions measurable with respect to $T(\underline{X})$. Note that $\psi^*(\underline{X})$ relates to the conditional bias of the estimator $\hat{\theta}(\underline{X})$ given the maximal

invariant $T(\underline{X})$. Then the UMRE estimator (under a quadratic loss function) is $\tilde{\theta}(\underline{X}) = \hat{\theta}(\underline{X}) - E_{\theta} [\hat{\theta}(\underline{X}) - \theta \mid T(\underline{X})]$. However, due to the fact that $\hat{\theta}(\underline{X})$ is invariant, the UMRE estimator reduces to

$$\tilde{\theta}(\underline{X}) = \hat{\theta}(\underline{X}) - E_0 [\hat{\theta}(\underline{X}) \mid T(\underline{X})].$$

Remark 1. The UMRE estimator is also unbiased.

$$\begin{aligned} E_{\theta} [\tilde{\theta}(\underline{X})] &= E_{\theta} [\hat{\theta}(\underline{X})] - E_{\theta} \left\{ E_{\theta} [\hat{\theta}(\underline{X}) - \theta \mid T(\underline{X})] \right\} \\ &= E_{\theta} [\hat{\theta}(\underline{X})] - E_{\theta} [\hat{\theta}(\underline{X}) - \theta] \\ &= E_{\theta} [\hat{\theta}(\underline{X})] - E_{\theta} [\hat{\theta}(\underline{X})] + \theta \\ &= \theta. \end{aligned}$$

Remark 2. Consider an UMVU estimator θ^* . If θ^* is equivariant, then $\theta^* = \tilde{\theta}$, that is, it is also UMRE. If θ^* is not equivariant, then $R(\theta, \theta^*) \leq R(\tilde{\theta})$. Note that $R(\tilde{\theta}) = \text{Var} [\tilde{\theta}(\underline{X})]$ does not depend on θ .

Remark 3. The Pitman form of an equivariant estimator, say $\hat{\theta} = X_1$, stands for the explicit form of the conditional expectation given a maximal invariant, say $T(\underline{X}) = (X_2 - X_1, \dots, X_n - X_1)$. The joint density g_0 of X_1 and $T(\underline{X})$ is

$$g_0(u, v_1, \dots, v_{n-1}) = f_0(u) \prod_{i=1}^{n-1} f_0(v_i + u),$$

where f_0 is the marginal density of X_i under $\theta = 0$. Accordingly, the joint density h_0 of $T(\underline{X})$ under $\theta = 0$ is

$$h_0(v_1, \dots, v_{n-1}) = \int_{-\infty}^{\infty} f_0(u) \prod_{i=1}^{n-1} f_0(v_i + u) du.$$

Therefore, the Pitman form of the equivariant estimator relates to

$$E_0 [X_1 \mid T(\underline{X})] = \frac{\int_{-\infty}^{\infty} u f_0(u) \prod_{i=1}^{n-1} f_0(v_i + u) du}{\int_{-\infty}^{\infty} f_0(u) \prod_{i=1}^{n-1} f_0(v_i + u) du}.$$

Remark 4. The UMRE estimator for the location model under an absolute deviation loss function is

$$\tilde{\theta}(\underline{X}) = \hat{\theta}(\underline{X}) - \text{med}_{\theta} [\hat{\theta}(\underline{X}) - \theta \mid T(\underline{X})] = \hat{\theta}(\underline{X}) - \text{med}_0 [\hat{\theta}(\underline{X}) \mid T(\underline{X})],$$

which is median unbiasedness. In this case, $\psi^*(\underline{X})$ relates to the conditional median of $\hat{\theta}(\underline{X}) - \theta$ given $T(\underline{X})$. Note, in particular, how the notion of unbiasedness is strongly related to the loss function at hand.

Definition. An estimator $\hat{\theta}$ for θ is risk unbiasedness if

$$R(\theta, \hat{\theta}) = E_{\theta} [L(\theta, \hat{\theta}(\underline{X}))] \leq E_{\theta} [L(\theta', \hat{\theta}(\underline{X}))] \neq R(\theta', \hat{\theta})$$

for every $\theta' \neq \theta$.

Remark 5. Gaussian location models are the worst case possible for the estimation of the location parameter! Consider a random sample $\underline{X} = (X_1, \dots, X_n)$, where $X_i - \theta \sim N(0, 1)$. Then, under a quadratic loss function, the sample mean \bar{X} is UMRE, UMRU, and UMVU for estimating θ . The corresponding risk is $R(\bar{X}) = \text{Var}[\bar{X}] = 1/n$. Consider now a nonnormal density f_0 belonging to the location family, and denote by $\tilde{\theta}$ the UMRE estimator. Then, by definition, the risk associated with $\tilde{\theta}$ is at most equal to the risk corresponding to the sample mean \bar{X} , i.e., $R(\tilde{\theta}) \leq R(\bar{X}) = 1/n$. In other words, the location parameter is more precisely estimated under nonnormality.

6 Hypothesis testing

Consider a model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{P} = H_0 \oplus H_1$. The set H_0 denotes the null hypothesis of the test, whereas H_1 denotes the alternative hypothesis. In the parametric case, $\mathcal{P} = \{P_{\underline{\theta}} \mid \underline{\theta} \in \Theta = H_0 \oplus H_1\}$. Testing is a decision problem completely characterized by the partition of \mathcal{P} . There are two possible decisions: reject the null hypothesis RH_0 or not. The decision function corresponds to a test ϕ which maps $(\mathcal{X}, \mathcal{A})$ into $([0, 1], \mathcal{B} \cap [0, 1])$ in such way that $\phi(\underline{X})$ represents the probability of rejecting the null hypothesis conditioned upon \underline{X} . Note that $\phi(\underline{X}) = P(RH_0 \mid \underline{X})$ does not depend on the parameter θ .

The expected probability of rejecting the null hypothesis when it is in fact correct, i.e. $E_{P \in H_0} \phi$, gives the size of the test. Accordingly, the expected

probability of rejecting the null when the alternative is correct, i.e. $E_{P \in H_1} \phi$, stands for the power of the test.

Neyman's Principle. One should attempt to control for the risk of Type I error $E_{P \in H_0} \phi \equiv \alpha(\phi, P)$ of the test ϕ . More precisely, one should consider only tests ϕ whose risks of Type I error are uniformly smaller than a given level $\alpha \in [0, 1]$, i.e., such that

$$\sup_{P \in H_0} \alpha(\phi, P) \leq \alpha.$$

The level α is called the significance level of these tests. Then, among these tests, one searches for a test, if it exists, that minimizes the risk of Type II error $\beta(\phi, P)$ for every $P \in H_1$, or equivalently, a test which maximizes the power function $E_{P \in H_1} \phi \equiv 1 - \beta(\phi, P)$.

Definition. A test ϕ^* is uniformly most powerful (UMP) within a class \mathcal{C}_α of α -level tests if $\phi^* \in \mathcal{C}_\alpha = \left\{ \phi : \sup_{P \in H_0} E_P(\phi) \leq \alpha \right\}$ is such that $E_P(\phi^*) \geq E_P(\phi)$ for every $P \in H_1$ and $\phi \in \mathcal{C}_\alpha$.

Note that uniformly most powerful tests seldom exist. Therefore it is often necessary to take into account additional principles, e.g. unbiasedness, invariance, asymptotic and likelihood principles.

6.1 Simple null versus simple alternative

Consider a family $\mathcal{P} = \{P_0, P_1\}$, where $H_0 = \{P_0\}$ and $H_1 = \{P_1\}$. Suppose that P_0 and P_1 are absolutely continuous. Then all feasible tests are represented in the region M of the power diagram illustrated in figure 1.

Elementary Properties. First, note that the point $(\alpha, \alpha) \in M$ for $\alpha \in [0, 1]$. In particular, $(0, 0)$ and $(1, 1)$ belong to M . Second, M is symmetric with respect to $(\frac{1}{2}, \frac{1}{2})$ since the fact that ϕ is a test implies that $1 - \phi$ is also a test. Third, M is compact (closed and bounded). Fourth, M is convex because if ϕ_1 and ϕ_2 are tests, then $\phi_\lambda = \lambda\phi_1 + (1 - \lambda)\phi_2$ is also a test for every $\lambda \in [0, 1]$.

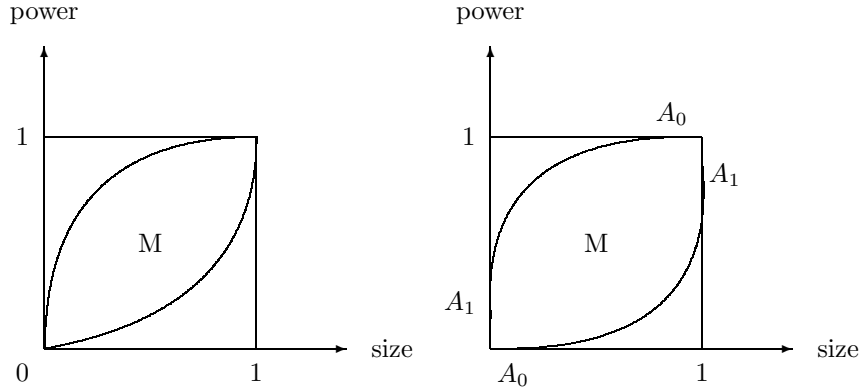


Figure 1

Figure 2

Consider now the general case where P_0 and P_1 are not necessarily absolutely continuous. Then, applying the Lebesgue decomposition to P_1 provides a partition (Q_1, S_1) such that Q_1 is absolutely continuous with respect to P_0 , whereas S_1 is singular with respect to P_0 , i.e. there exists $A \in \mathcal{A}$ such that $P_0(A) = 0$ and $S_1(A) > 0$. Then, for $\underline{X} \in A_1$ with $P_1(A_1) > 0$ and $P_0(A_1) = 0$, there is no size cost in rejecting the null hypothesis. Accordingly, for $\underline{X} \in A_0$ with $P_1(A_0) = 0$ and $P_0(A_0) > 0$, there is no power loss in not rejecting the null. This fact is illustrated in figure 2. By Neyman-Pearson's Theorem, there exists a UMP test under the probability level condition – a Neyman test – which is always placed at the upper border of M in order to maximize power. Note that the part of the upper border corresponding to the set A_0 contains inadmissible tests, since the size of these tests can be reduced without loss of power.

The points of M such that the northwest orthant with origin at $M(\phi) = (\alpha(\phi), 1 - \beta(\phi))$ intersects the risk diagram only at $M(\phi)$ stand for the set of admissible tests. Therefore, by changing the level of significance α , it is easy to see that the set of optimal tests in Neyman's sense belongs to the set of admissible tests. Accordingly, when P_0 and P_1 are absolutely continuous, these two sets are identical.

Given a significance level α , we now characterize the tests that are optimal in

Neyman's sense. These tests are those that minimize the risk of Type II error $\beta(\phi)$, i.e. that maximizes the power $E_1\phi$ subject to the level constraint $E_0\phi = \alpha(\phi) \leq \alpha$. Because there are only two probability distributions in the model, the model is always dominated by some measure μ — say, for instance, $P_0 + P_1$. Thus, P_0 and P_1 admit densities ℓ_0 and ℓ_1 with respect to μ , respectively.

Definition. A Neyman test is a test which satisfies

$$\phi(x) = \begin{cases} 1 & \text{if } \ell_1(x) > k\ell_0(x), \\ 0 & \text{if } \ell_1(x) < k\ell_0(x), \end{cases}$$

where k is a real positive number.

Note that the definition of Neyman tests does not impose any restriction when the equality $\ell_1(x) = k\ell_0(x)$ holds. In addition, Neyman tests have a natural interpretation. Namely, a hypothesis is accepted when it is sufficiently likely relative to the alternative hypothesis. Moreover the test function of a Neyman test can be written as

$$\phi(x) = \begin{cases} 1 & \text{if } \log \ell_1(x) - \log \ell_0(x) > \log k, \\ 0 & \text{if } \log \ell_1(x) - \log \ell_0(x) < \log k, \end{cases}$$

that is, it can be expressed in terms of log-likelihood ratios.

Neyman-Pearson's Theorem. (i) For any $\alpha \in [0, 1]$, there exists a Neyman test ϕ such that $E_0\phi = \alpha$. Moreover, for all x such that $\ell_1(x) = k\ell_0(x)$, where k is the real number defining ϕ , one can set ϕ to a constant γ irrespective of x . (ii) For any $\alpha \in [0, 1]$, a Neyman test ϕ satisfying $E_0\phi = \alpha$ is optimal at the significance level α . (iii) Conversely, for any given significance level $\alpha \in [0, 1]$, an optimal test is necessarily a Neyman test.

Proof. (i) Denote by ϕ a Neyman test corresponding to a real number k such that $\phi(x) = \gamma$ if $\ell_1(x) = k\ell_0(x)$. Then,

$$\begin{aligned} E_0\phi &= P_0(\ell_1(X) > k\ell_0(X)) + \gamma P_0(\ell_1(X) = k\ell_0(X)) \\ &= P_0\left(\frac{\ell_1(X)}{\ell_0(X)} > k\right) + \gamma P_0\left(\frac{\ell_1(X)}{\ell_0(X)} = k\right) \\ &= 1 - F(k) + \gamma[F(k) - F(k^-)], \end{aligned}$$

where F denotes the cumulative distribution of the likelihood ratio $\ell_1(X)/\ell_0(X)$ and $F(k^-) = \lim_{\kappa \uparrow k} F(\kappa)$. However F is right continuous, increasing, and such that $F(0^-) = 0$ and $\lim_{k \rightarrow \infty} F(k) = 1$, so that there exists a real number k_0 which satisfies $F(k_0^-) \leq 1 - \alpha \leq F(k_0)$. Then, two cases can be distinguished.

(a) If $F(k_0) = 1 - \alpha$, then a Neyman test ϕ corresponding to $k = k_0$ and $\gamma = 0$ satisfies $E_0\phi = \alpha$.

(b) If $F(k_0^-) \leq 1 - \alpha < F(k_0)$, then a Neyman test ϕ satisfies $E_0\phi = \alpha$ if associated with $k = k_0$ and

$$\gamma = \frac{F(k_0) - (1 - \alpha)}{F(k_0) - F(k_0^-)}.$$

(ii) Let ϕ denote a Neyman test which satisfies the size constraint $E_0\phi = \alpha$ for $k \in \mathbb{R}_+$. Consider now another test ϕ^* for which the level constraint $E_0\phi^* \leq \alpha$ holds. Then, the integral

$$\int [\phi(x) - \phi^*(x)] [\ell_1(x) - k\ell_0(x)] d\mu(x)$$

is nonnegative. This stems from $\phi = 1 \geq \phi^*$ if $\ell_1(x) > k\ell_0(x)$ and $\phi = 0 \leq \phi^*$ if $\ell_1(x) < k\ell_0(x)$. Thus,

$$E_1(\phi - \phi^*) \geq kE_0(\phi - \phi^*) = k(\alpha - E_0\phi^*).$$

Because the risk of Type I error associated with ϕ^* is at most α , it follows that $E_1\phi$ is greater than or equal to $E_1\phi^*$. Thus a Neyman test ϕ is optimal at the significance level α . (iii) Let ϕ' be an optimal test with significance level α . Let ϕ be a Neyman test such that $E_0\phi = \alpha$. From (ii), this Neyman test is also optimal at level α . Because both tests are optimal, their power must be equal, that is, $E_1\phi' = E_1\phi$. Consider now the integral

$$\int [\phi(x) - \phi'(x)] [\ell_1(x) - k\ell_0(x)] d\mu(x),$$

where k is an arbitrary positive real number. This integral is equal to

$$E_1(\phi - \phi') - kE_0(\phi - \phi') = k(E_0\phi' - \alpha) \leq 0.$$

Next choose k as the number corresponding to ϕ . From (ii), the above integral is known to be nonnegative. Hence, this integral must be equal to zero, which implies that the integrand is zero given that, from (ii), it is necessarily nonnegative. This implies that

$$\phi'(x) = \begin{cases} 1 & \text{if } \ell_1(x) > k\ell_0(x) \\ 0 & \text{if } \ell_1(x) < k\ell_0(x), \end{cases}$$

and so ϕ' is a Neyman test. ■

Note that if the probability distribution of the likelihood ratio $\ell_1(X)/\ell_0(X)$ is continuous under the null hypothesis H_0 , then $P_0(\ell_1(X) = k\ell_0(X)) = 0$ for every $k \in \mathbb{R}$. Thus, for any significance level α , there exists an optimal nonrandomized test whose critical region is $W = \{\ell_1(x) > k\ell_0(x)\}$, where k is defined by $P_0(\ell_1(X) > k\ell_0(X)) = \alpha$.

Suppose that there exists a sufficient statistic S . From the Neyman-Fisher factorization criterion, $\ell_0(x) = h(x)g_0(S(x))$ and $\ell_1(x) = h(x)g_1(S(x))$, so that the likelihood ratio is given by $\ell_1(x)/\ell_0(x) = g_1(S(x))/g_0(S(x))$. Therefore an optimal Neyman test at the significance level α depends on the observations solely through the sufficient statistic S .

Exponential Family. Consider a one-parameter exponential model with density $\ell(x; \theta) = C(\theta)h(x) \exp(Q(\theta)T(x))$, where the parameter θ can only assume two values, i.e. $\theta \in \Theta = \{\theta_0, \theta_1\}$. Assume that the model is identified and, without any loss of generality, that $\theta_1 > \theta_0$ and that Q is strictly increasing in θ , i.e. $Q(\theta_1) > Q(\theta_0)$. By Neyman-Pearson's Theorem, the optimal test for the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ is a Neyman test of the form

$$\phi(x) = \begin{cases} 1 & \text{if } \ell_1(x) > k\ell_0(x), \\ \gamma & \text{if } \ell_1(x) = k\ell_0(x), \\ 0 & \text{if } \ell_1(x) < k\ell_0(x). \end{cases}$$

However,

$$\frac{\ell_1(x)}{\ell_0(x)} = \frac{C(\theta_1)}{C(\theta_0)} \exp[(Q(\theta_1) - Q(\theta_0))T(x)],$$

where $Q(\theta_1) - Q(\theta_0) > 0$. Hence, a Neyman test can be written as

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > c, \\ \gamma & \text{if } T(x) = c, \\ 0 & \text{if } T(x) < c, \end{cases}$$

where the constants γ and c are determined by the size condition $E_0\phi = \alpha$. From this perspective, in the case of exponential models, the optimal Neyman test can be expressed as a simple function of the privileged statistic.

6.2 Composite null versus simple alternative

Consider a family $\mathcal{P} = H_0 \oplus H_1$, where $H_0 = \{P_1, \dots, P_m\}$ and $H_1 = \{P_{m+1}\}$. Denote by f_i the density of P_i with respect to the sum measure $\mu = \sum_{i=1}^{m+1} P_i$. For the sake of notation simplicity, let E_i represent the expectation under f_i . Then, the power diagrams $M_m \in \mathbb{R}^m$ and $M_{m+1} \in \mathbb{R}^{m+1}$ are associated with the sets $\{E_1\phi, \dots, E_m\phi \mid \phi \text{ is a test}\}$ and $\{E_1\phi, \dots, E_{m+1}\phi \mid \phi \text{ is a test}\}$, respectively.

Elementary Properties. The power diagrams conserve all the elementary properties described above. Namely, M_m and M_{m+1} are convex, compact, and symmetric with respect to the center $(1/2, \dots, 1/2)$ of the hypercube.

Proposition (Generalized Neyman-Pearson's Theorem). (i) For every $\underline{c} = (c_1, \dots, c_m) \in M_m$, there exists an (essentially) unique test ϕ maximizing the power $E_{m+1}\phi$ under the size constraint $E_i\phi = c_i, i = 1, \dots, m$. (ii) Consider a Neyman test ϕ which satisfies the size constraint, viz.

$$\phi(x) = \begin{cases} 1 & \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x), \\ \gamma(x) & \text{if } f_{m+1}(x) = \sum_{i=1}^m k_i f_i(x), \\ 0 & \text{if } f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x), \end{cases}$$

where $\gamma(x)$ stands for some randomization, such that $E_i\phi = c_i, i = 1, \dots, m$ for some constants k_1, \dots, k_m . Then, ϕ is equivalent to the maximizing test B^+ in figure 2. (iii) If ϕ is a Neyman test which satisfies the size constraint and the constants k_i are nonnegative for $i = 1, \dots, m$, then ϕ is the most powerful test under the level constraint $E_i\phi \leq c_i, i = 1, \dots, m$. (iv) If \underline{c} belongs to the interior of M_m , then there exists a Neyman test satisfying the size constraint.

Conversely, every test which maximizes power under the size constraint consists in a Neyman test.

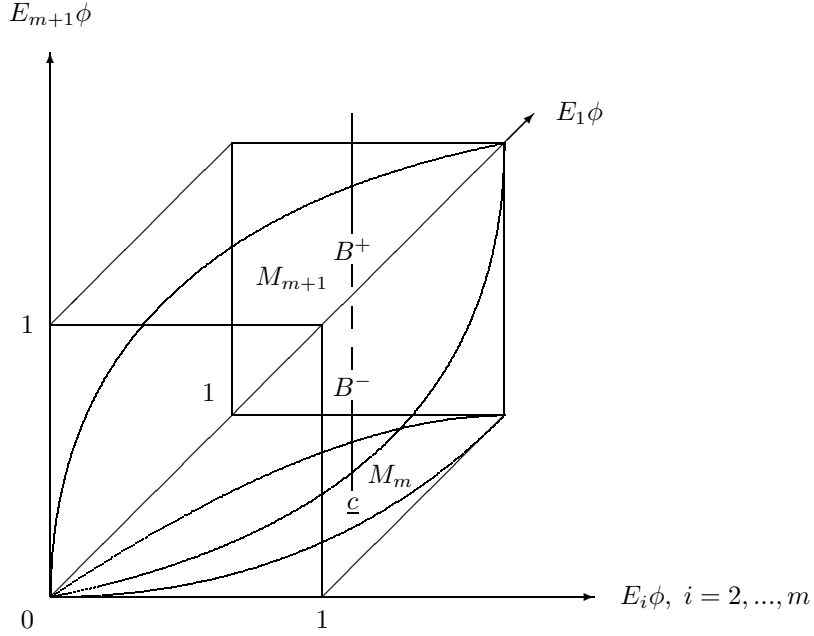


Figure 3

Proof. (i) By convexity and closeness of M_{m+1} , it is clear that there are at most two intersections with a vertical plane at \underline{c} , say B^- and B^+ (see figure 3). If $B^- = B^+$, then \underline{c} belongs to the boundary of M_m . If $B^- \neq B^+$, then B^+ corresponds to the most powerful test. (ii) Let ϕ' satisfy the size constraint. Then, for a Neyman test ϕ which also satisfies the size constraint, note that

$$\int (\phi - \phi') \left(f_{m+1} - \sum_{i=1}^m k_i f_i \right) d\mu \geq 0,$$

due to the nonnegativity of the integrand. Therefore,

$$E_{m+1}\phi - E_{m+1}\phi' - \sum_{i=1}^m k_i (E_i\phi - E_i\phi') \geq 0,$$

which implies that $E_{m+1}\phi \geq E_{m+1}\phi'$ since the last term of the left-hand side is zero given that both tests satisfy the size constraint. (iii) Consider a test ϕ'

such that $E_i\phi' \leq c_i$ for $i = 1, \dots, m$. Then, using the same rationale gives

$$E_{m+1}\phi \geq E_{m+1}\phi' + \sum_{i=1}^m k_i \underbrace{(E_i\phi - E_i\phi')}_{\geq 0},$$

which implies that a Neyman test ϕ satisfying the size constraint is the most powerful test only if the constants k_i are nonnegative. (iv) The convexity of the power diagram implies that there exists a hyperplane H separating B^+ and the interior of M_{m+1} . For a point \underline{y} belonging to this hyperplane, the following equation holds

$$\sum_{i=1}^{m+1} \tilde{k}_i y_i = \sum_{i=1}^m \tilde{k}_i c_i + \tilde{k}_{m+1} b,$$

where $\tilde{k}_{m+1} \neq 0$. Otherwise all points in the vertical plane cutting the power diagram would belong to the hyperplane. Taking $k_i = -\tilde{k}_i/\tilde{k}_{m+1}$ yields

$$y_{m+1} - \sum_{i=1}^m k_i y_i = b - \sum_{i=1}^m k_i c_i.$$

For $\underline{y} \in M_{m+1}$, it is clear that

$$y_{m+1} - \sum_{i=1}^m k_i y_i \leq b - \sum_{i=1}^m k_i c_i.$$

However, $y_i = E_i\phi$ given that it belongs to M_{m+1} , so that

$$E_{m+1}\phi - \sum_{i=1}^m k_i E_i\phi \leq b - \sum_{i=1}^m k_i c_i$$

for all ϕ . This means that $b - \sum_{i=1}^m k_i c_i = y_{m+1} - \sum_{i=1}^m k_i y_i$, for $\underline{y} \in H$ is the quantity that maximizes $E_{m+1}\phi - \sum_{i=1}^m k_i E_i\phi$. But B^+ is the intersection of the hyperplane with the power diagram, so that it corresponds to the most powerful test

$$E_{m+1}\phi - \sum_{i=1}^m k_i E_i\phi \leq E_{m+1}\phi_{B^+} - \sum_{i=1}^m k_i E_i\phi_{B^+},$$

for $\phi \in M_{m+1}$. Hence,

$$\int \phi \left(f_{m+1} - \sum_{i=1}^m k_i f_i \right) d\mu \leq \int \phi_{B^+} \left(f_{m+1} - \sum_{i=1}^m k_i f_i \right) d\mu,$$

where the equality holds for a Neyman test with constants k_i . However, B^+ also satisfies the size constraints, so that it corresponds to a Neyman test. The converse argument follows in the same rationale. ■

Remark. The most powerful test has the form of a classical Neyman test (with constant k) for the mixture $f_0 = \sum_{i=1}^m \frac{k_i}{k} f_i$ against f_{m+1} , i.e.

$$\phi(x) = \begin{cases} 1 & \text{if } f_{m+1}(x) > k \sum_{i=1}^m \frac{k_i}{k} f_i(x), \\ 0 & \text{if } f_{m+1}(x) < k \sum_{i=1}^m \frac{k_i}{k} f_i(x), \end{cases}$$

where $k_i \geq 0$ and $k = \sum_{i=1}^m k_i$. It is like associating the probability k_i/k to the density $f_i \in H_0$. For obvious reason, the vector $(k_1/k, \dots, k_m/k)$ is called the mixing distribution. If the problem of testing the composite null H_0 against the simple alternative H_1 has a solution (i.e. a most powerful test under a probability level constraint), then this solution belongs to the set of all possible Neyman tests for the simple null corresponding to the mixture of H_0 against the alternative H_1 .

6.3 Least favorable distributions

Consider a testing problem with composite null hypothesis $H_0 : \{f_{\underline{\theta}} \mid \underline{\theta} \in \Theta\}$ and simple alternative $H_1 : \{g\}$. As usual, the aim is at maximizing the power $E_g(\phi)$ subjected to a level constraint $E_{\underline{\theta}}(\phi) \leq \alpha$ for every $\underline{\theta} \in \Theta$. Consider now a simple hypothesis $H_\lambda : \{h_\lambda\}$, where λ is a probability measure and $h_\lambda(\underline{x}) \equiv \int_{\Theta} f_{\underline{\theta}}(\underline{x}) d\lambda(\underline{\theta})$. Then a Neyman test for H_λ against H_1 has the form

$$\phi_\lambda(x) = \begin{cases} 1 & g(x) > k_\lambda h_\lambda(x), \\ 0 & g(x) < k_\lambda h_\lambda(x). \end{cases}$$

Denote by $\Pi_\lambda = E_g(\phi_\lambda)$ the power of this Neyman test. The mixing distribution, say λ_0 , that we are looking for could well be the least favorable one, i.e. the mixture h_{λ_0} which is closest to g . In this way, we would be maximizing the power of the test in the most difficult situation for distinguishing the mixture of H_0 from the alternative H_1 .

Definition. The least favorable mixing distribution λ_0 is such that $\Pi_{\lambda_0} \leq \Pi_\lambda$ for every mixing distribution λ over Θ .

Proposition. Assume that the mixing distribution λ_0 is such that $E_{\underline{\theta}}(\phi_{\lambda_0}) \leq \alpha$ for all $\underline{\theta} \in \Theta$. Then, (i) ϕ_{λ_0} is most powerful for the initial problem H_0 against H_1 , and (ii) λ_0 is the least favorable mixing distribution.

Proof. Let ϕ' denote a test such that $E_{\underline{\theta}}(\phi') \leq \alpha$ for every $\underline{\theta} \in \Theta$. Then,

$$\begin{aligned} E_{h_\lambda}(\phi') &= \int_{\mathcal{X}} \phi'(x) h_\lambda(x) d\mu(x) = \int_{\mathcal{X}} \int_{\Theta} \phi'(x) f_{\underline{\theta}}(x) d\lambda(\underline{\theta}) d\mu(x) \\ &= \int_{\Theta} \int_{\mathcal{X}} \phi'(x) f_{\underline{\theta}}(x) d\mu(x) d\lambda(\underline{\theta}) \\ &\leq \int_{\Theta} \alpha d\lambda(\underline{\theta}) = \alpha \end{aligned}$$

for all mixing distributions λ over Θ . Hence, under the assumption that the probability level constraint holds, ϕ_{λ_0} has level α for any H_λ . However, ϕ_λ is a Neyman test for the null H_λ , thus $E_g \phi_{\lambda_0} = \Pi_{\lambda_0} \leq \Pi_\lambda$ for any λ over Θ . This means, by definition, that λ_0 is the least favorable mixing distribution. Moreover if ϕ' has level α under any H_λ , then ϕ' has also level α under H_{λ_0} , but it is less powerful than the correspondent Neyman test ϕ_{λ_0} . Therefore, ϕ_{λ_0} is the most powerful test for H_0 against H_1 . ■

Therefore, the strategy consists in identifying the least favorable distribution λ_0 which satisfies the α -level condition under the null hypothesis H_λ for every λ over Θ . If λ_0 really stands for the least favorable distribution for the null hypothesis H_0 , then the correspondent test ϕ_{λ_0} maximizes the power subject to the α -level constraint.

Example 1. Consider a parametric family of probabilities over $\Theta \subseteq \mathbb{R}$ with densities f_θ with respect to the Lebesgue measure satisfying the monotone likelihood ratio property. Let the null hypothesis correspond to $H_0 : \theta \leq \theta_0$ whereas $H_1 : \theta > \theta_0$ stand for the alternative hypothesis. It is easy to see that $\theta = \theta_0$ constitutes the case under the null hypothesis which is closest to the alternative hypothesis. Hence it seems plausible to take the mixing distribution degenerated at θ_0 as the least favorable mixing distribution. In addition, it is not difficult to show that a Neyman test of f_{θ_0} against f_{θ_1} has level α for every $\theta_1 \in H_1$, so that maximum power is achieved for testing the null H_0 against the alternative

H_1 (see subsection 6.4.1).

Example 2. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ with identical marginal density with respect to the Lebesgue measure over $(\mathbb{R}, \mathcal{B})$. Assume further that this density is positive almost everywhere. Denote by ξ_p the population quantile of order p , that is, $\xi_p = F^{-1}(p)$. Let the null hypothesis of interest be $H_0 : \xi_{p_0} \leq u$, whereas $H_1 : \xi_{p_0} > u$ stand for the alternative hypothesis. Denoting by $p = F(u)$ the true probability permits representing the hypotheses of interest by $H_0 : p \geq p_0$ and $H_1 : p < p_0$. Then, it is recommendable to use the sign statistic $M \equiv \#\{i | X_i \leq u\}$, which has exact distribution $\text{Bin}(n, p)$. Note that $\hat{p} = M/n$ stands for a consistent estimator of p . From this perspective, it seems natural to reject the null hypothesis when M is small. More precisely, the null is rejected if $F_{\text{Bin}(n,p)} \leq \alpha$ for a test ϕ such that $E_{p_0} \phi = \alpha$. In order to apply the concept of least favorable distribution, note that H_0 is a composite hypothesis and a cumulative distribution $F \in H_0$ can be characterized by (p, p^+, p^-) , where $p = F(u)$, p^+ corresponds to the conditional probability of X given that $X > u$, and p^- represents the conditional probability of X given that $X < u$. Then, the null hypothesis can be written as $H_0 = \{(p, p^+, p^-) | p \geq p_0\}$. Consider now an element (p_1, p_1^+, p_1^-) of the alternative hypothesis $H_1 = \{(p, p^+, p^-) | p < p_0\}$. A natural candidate for the least favorable distribution consists of (p_0, p_1^+, p_1^-) , so that we need to look at a Neyman test for (p_0, p_1^+, p_1^-) against (p_1, p_1^+, p_1^-) . As the two last elements of the hypotheses do not differ, a Neyman test will depend only in the discrepancy between the first element of each hypothesis. More formally,

$$\phi_N = \frac{p_1^M (1 - p_1)^{n-M}}{p_0^M (1 - p_0)^{n-M}} = \left(\frac{p_1}{p_0}\right)^M \left(\frac{1 - p_1}{1 - p_0}\right)^{n-M},$$

which is also a Neyman test for the null of $p = p_0$ against the alternative of $p = p_1 < p_0$. Note that a Neyman test rejects the null hypothesis H_0 in the case of a large ϕ_N . Moreover, this test belongs to the family of monotone likelihood ratio-based tests so that it is not only a most powerful test for the

problem under consideration, but also a uniformly most powerful test. Finally, two remarks are in place. First, as $E_{H_0}\phi_N = \alpha$ this test really considers the least favorable distribution. Second, under the null hypothesis, ϕ_N is large if and only if M is small, so that the sign test based on M is also uniformly most powerful.

6.4 Tests of one-sided hypotheses

Consider a one-parameter model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_\theta \mid \theta \in \Theta\})$, where Θ is an interval of \mathbb{R} , which is dominated by a measure μ . Let $\{\ell(x; \theta); \theta \in \Theta\}$ denote the family of strictly positive densities corresponding to the distributions P_θ , $\theta \in \Theta$. The null hypothesis H_0 of interest is defined by $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ and the alternative H_1 consists in $\Theta_1 = \{\theta : \theta > \theta_0\}$, where θ_0 is an element of Θ such that both Θ_0 and Θ_1 are nonempty.

6.4.1 Monotone likelihood ratio family

Definition. The family $\mathcal{P} = \{P_\theta \mid \theta \in \Theta \subseteq \mathbb{R}\}$ consists in a monotone likelihood ratio family if there exists a statistic $U(x)$ with values in \mathbb{R} such that, $\forall \theta' > \theta$, the ratio $\ell(x; \theta')/\ell(x; \theta)$ is a strictly increasing or decreasing function of U .

Note that by changing U to $-U$, there is no loss of generality in assuming that the likelihood ratios are strictly increasing functions of U . The next theorem states that there exists a UMP test under the conditions of the preceding definition.

Proposition. Suppose that $\{\ell(x; \theta); \theta \in \Theta\}$ belongs to a strictly increasing likelihood ratio family. Then, for any $\alpha \in [0, 1]$ there exists a UMP test with significance level α for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. This test is given by

$$\phi(x) = \begin{cases} 1 & \text{if } U(x) > c, \\ \gamma & \text{if } U(x) = c, \\ 0 & \text{if } U(x) < c, \end{cases}$$

where the constants γ and c are determined by the size condition $E_{\theta_0}\phi = \alpha$.

Proof. Consider the problem of testing the simple hypothesis $\tilde{H}_0 : \theta = \theta_0$

against the simple alternative $H_1 : \theta = \theta''$, where θ'' is a given value greater than θ_0 . From Neyman-Pearson's Theorem, an optimal test at level α is a Neyman test of form

$$\phi(x) = \begin{cases} 1 & \text{if } \ell(x; \theta'') > k\ell(x; \theta_0), \\ \gamma & \text{if } \ell(x; \theta'') = k\ell(x; \theta_0), \\ 0 & \text{if } \ell(x; \theta'') < k\ell(x; \theta_0). \end{cases}$$

From the monotone likelihood ratio property it follows that this test can be written as

$$\phi(x) = \begin{cases} 1 & \text{if } U(x) > c, \\ \gamma & \text{if } U(x) = c, \\ 0 & \text{if } U(x) < c, \end{cases}$$

where the constants γ and c are determined by the size condition $E_{\theta_0}\phi = \alpha$. Note that this test does not depend on the choice of θ'' ($\theta'' > \theta_0$). Therefore, the test is UMP at level α for testing $\tilde{H}_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta > \theta_0$. This means that ϕ is uniformly more powerful than any test $\tilde{\phi}$ for \tilde{H}_0 satisfying $E_{\theta_0}\tilde{\phi} \leq \alpha$. Hence, ϕ is more powerful than any test ϕ^* satisfying $\sup_{\theta \in \Theta_0} E_{\theta}\phi^* \leq \alpha$, i.e. than any test with level of significance α . In order to complete the proof, it is necessary to verify whether the level of the test ϕ is also less or equal to α . Let $\theta' < \theta_0$. From Neyman-Pearson's Theorem, it follows that the test ϕ is optimal at level $E_{\theta'}\phi$ for testing $H'_0 : \theta = \theta'$ against $H_1 : \theta = \theta_0$. However, $E_{\theta_0}\phi \geq E_{\theta'}\phi$, because the power of an optimal test is necessarily greater than or equal to its significance level. Hence $E_{\theta'}\phi \leq E_{\theta_0}\phi = \alpha$ for every $\theta' < \theta_0$. Thus the level of ϕ is α . ■

Exponential Family. Consider a one-parameter exponential model with density $\ell(x; \theta) = C(\theta)h(x)\exp(Q(\theta)T(x))$, where $Q(\theta)$ is strictly increasing. It is clear that such a family is a strictly increasing likelihood ratio family in its privileged statistic T . Therefore, it is not surprising that a Neyman test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$ can be written in terms of the privileged statistic (see subsection 6.1).

6.4.2 Locally most powerful tests and score tests

When the probability family does not satisfy the monotone likelihood ratio property, there may no longer exist a UMP test for one-sided hypotheses. In general, however, it is possible to construct a test which is locally more powerful for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Such a test is optimal in Neyman's sense only in a neighborhood of the boundary value θ_0 .

Definitions. Consider the problem of testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

1. A test ϕ is locally of level α if there exists a neighborhood \mathcal{V} of θ_0 such that

$$\sup_{\theta \leq \theta_0, \theta \in \mathcal{V}} E_\theta \phi = E_{\theta_0} \phi = \alpha.$$

2. A test ϕ is locally more powerful than a test ϕ^* if there exists a neighborhood \mathcal{V} of θ_0 such that $E_\theta \phi > E_\theta \phi^*$ for every $\theta > \theta_0, \theta \in \mathcal{V}$.
3. A test ϕ is locally uniformly most powerful (LUMP) at level α if it is locally at level α and if it is locally more powerful than any other tests that are locally of level α .

Proposition 1. Suppose one considers only tests ϕ^* such that $E_\theta \phi^*$ is differentiable at θ_0 . Then a test ϕ satisfying

$$\left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta=\theta_0} \neq 0$$

is LUMP at level α if and only if it is locally of level α and satisfies

$$\left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta=\theta_0} \geq \left. \frac{\partial E_\theta \phi^*}{\partial \theta} \right|_{\theta=\theta_0},$$

for any test ϕ^* which is locally of level α .

Proof. The result follows immediately from the Taylor expansion

$$\begin{aligned} E_\theta \phi - E_\theta \phi^* &= E_{\theta_0}(\phi - \phi^*) + (\theta - \theta_0) \left(\left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta=\theta_0} - \left. \frac{\partial E_\theta \phi^*}{\partial \theta} \right|_{\theta=\theta_0} \right) \\ &\quad + o(\theta - \theta_0) \\ &= (\theta - \theta_0) \left(\left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta=\theta_0} - \left. \frac{\partial E_\theta \phi^*}{\partial \theta} \right|_{\theta=\theta_0} \right) + o(\theta - \theta_0), \end{aligned}$$

given that $E_{\theta_0}\phi = E_{\theta_0}\phi^* = \alpha$ and, under the alternative hypothesis, $\theta \in \mathcal{V}$ such that $\theta > \theta_0$. ■

Proposition 2. Suppose that it is possible to differentiate under the integral sign. Then any LUMP test at level α is a score test, namely

$$\phi(x) = \begin{cases} 1 & \text{if } S_{\theta_0}(x) > k, \\ \gamma & \text{if } S_{\theta_0}(x) = k, \\ 0 & \text{if } S_{\theta_0}(x) < k. \end{cases}$$

where $S_{\theta_0}(X) = \frac{\partial}{\partial \theta} \log \ell(x; \theta)|_{\theta=\theta_0}$, and k and γ are constants determined by the condition $E_{\theta_0}\phi = \alpha$.

Proof. The proof is in the same line of the proof of Neyman-Pearson's Theorem. Note, however, that k is not necessarily positive. Nonetheless the proof still applies because the tests under consideration are such that $E_{\theta_0}\phi^* = \alpha$ instead of $E_{\theta_0}\phi^* \leq \alpha$. ■

It is worth to stress that the score test can be viewed as a limit case of a Neyman test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$, when θ_1 converges to θ_0 . Specifically, from Neyman-Pearson's Theorem, a Neyman test is of the form

$$\phi(x) = \begin{cases} 1 & \text{if } [\log \ell(x; \theta_1) - \log \ell(x; \theta_0)]/(\theta_1 - \theta_0) > k, \\ 0 & \text{if } [\log \ell(x; \theta_1) - \log \ell(x; \theta_0)]/(\theta_1 - \theta_0) < k. \end{cases}$$

Then it suffices to note that the ratio $[\log \ell(x; \theta_1) - \log \ell(x; \theta_0)]/(\theta_1 - \theta_0)$ converges to $S_{\theta_0}(x)$ as θ_1 approaches θ_0 . In the special case where the probability distribution of the score statistic is continuous under the null hypothesis H_0 , then there exists a nonrandomized LUMP test at level α . Its critical region is given by $W = \{S_{\theta_0}(X) \geq k\}$, where k is defined by the condition $P_{\theta_0}(W) = \alpha$ on the risk of Type I error.

The exact distribution of the score statistic $S_{\theta_0}(\underline{X}) = S_{\theta_0}(X_1, \dots, X_n)$ is not always easy to determine. However, it is quite straightforward to derive the asymptotic approximation to the critical region of the score test. In fact, under some regularity conditions, the score vector has an asymptotic normal distribution. Therefore, the score test is asymptotically nonrandomized with

critical region

$$W = \left\{ S_{\theta_0}(\underline{X}) > \sqrt{n\hat{I}(\theta_0)}z_{1-\alpha} \right\},$$

where $z_{1-\alpha}$ is the $(1 - \alpha)^{\text{th}}$ quantile of the standard normal distribution and $\hat{I}(\theta_0)$ stands for a consistent estimator of the Fisher information associated with one observation.

6.5 Unbiased tests

Definition 1. Consider a model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{H_0 \oplus H_1\})$. A test ϕ is unbiased at level α if $E_P\phi \leq \alpha$ for every $P \in H_0$ and $E_Q\phi \geq \alpha$ for all $Q \in H_1$. Put differently, a test is unbiased if its power is never inferior to its size.

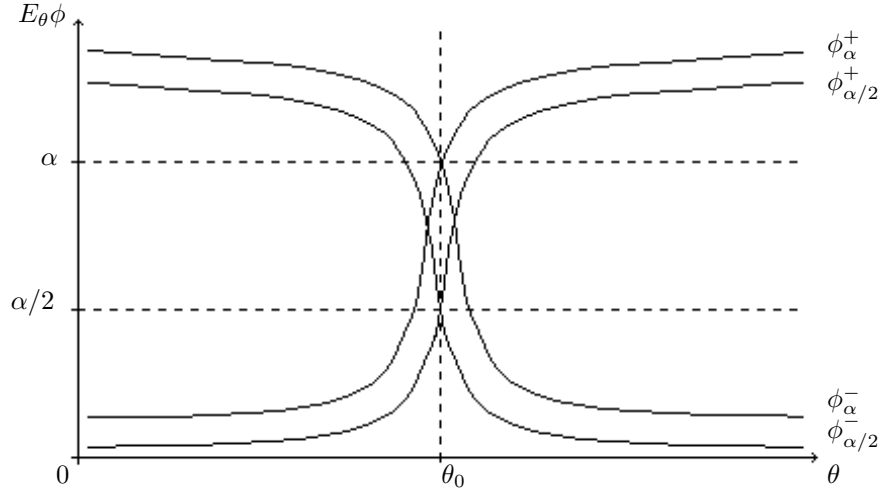


Figure 4

Example. Consider a family of monotone likelihood ratio-based tests for the null $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. Then a two-sided test $\phi_\alpha \equiv \phi_{\alpha/2}^+ + \phi_{\alpha/2}^-$ cannot have uniformly more power than the one-sided tests ϕ_α^+ and ϕ_α^- in figure 4. However, in contrast to ϕ_α , both ϕ_α^+ and ϕ_α^- are biased.

Definition 2. A test ϕ is uniformly most powerful and unbiased (UMPU) at level α if (i) it is unbiased at level α and (ii) it is more powerful than any other unbiased test of level α .

Note that the condition (i) can be replaced by the condition that ϕ is of level at most α , that is, the unbiasedness requirement is unnecessary. This stems from the fact that the constant test $\phi' = \alpha$ is trivially unbiased, which implies by condition (ii) that $E_P\phi \geq E_P\phi' = \alpha$ for any $P \in H_1$. Accordingly, a UMP test at level α is necessarily unbiased given that the power of such a test is at least equal to the constant test $\phi' = \alpha$. Therefore, any UMP test at level α is also a UMPU test at level α .

In certain circumstances it is natural to aim at constructing a uniformly most powerful test in the class of unbiased tests of level α , that is, a UMPU test. Thereby, it is useful to think in a topology for $\mathcal{P} = \{H_0 \oplus H_1\}$ such that $E_P\phi$ is continuous for every ϕ and $P \in \mathcal{P}$. Then, a necessary condition for α -unbiasedness of a test ϕ consists in belonging to the class of α -similar tests on $H^* = \overline{H_0} \cap \overline{H_1}$, that is, $E_P\phi = \alpha$ for every $P \in H^*$. Note also that a test is uniformly most powerful in the class of unbiased tests at level α if it is also uniformly most powerful in the class of α -similar tests.

Proposition. (i) Let ϕ^* be an unbiased test at level α . Suppose that the function $E_P\phi^*$ is continuous at any point of H^* . Then ϕ^* is an α -similar test on H^* . (ii) Suppose that the function $E_P\phi$ is continuous at any point of H_0 for every test ϕ . Then a test ϕ^* which is uniformly most powerful in the class of α -similar tests on H^* is necessarily uniformly most powerful and unbiased at level α .

Proof. (i) Every $P \in H^*$ is not only the limit of a sequence $\{P_{0n}\}$ of probability laws in H_0 , but also the limit of a sequence $\{P_{1n}\}$ of probability laws in H_1 . The continuity in P of the mapping $E_P\phi^*$ implies then that

$$E_P\phi^* = \lim_{n \rightarrow \infty} E_{P_{0n}}\phi^* \leq \alpha \quad \text{and} \quad E_P\phi^* = \lim_{n \rightarrow \infty} E_{P_{1n}}\phi^* \geq \alpha.$$

Thus $E_P\phi^* = \alpha$ for all $P \in H^*$, which implies that ϕ^* is an α -similar test on H^* . (ii) This is a direct consequence of (i), which implies that the class of α -similar tests contains the class of unbiased tests at level α . ■

6.5.1 UMPU tests for one-parameter exponential models

Consider a one-parameter exponential model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_\theta \mid \theta \in \Theta\})$, where Θ is an open interval of \mathbb{R} . The densities are $\ell(x; \theta) = C(\theta)h(x) \exp(Q(\theta)T(x))$, where $Q(\theta)$ is a strictly increasing and continuous function of θ . The problem of interest consists in testing the null hypothesis defined by $\Theta_0 = \{\theta : \theta_1 \leq \theta \leq \theta_2\}$ against the two-sided alternative hypothesis defined by $\Theta_1 = \Theta_0^c$, where $\theta_1, \theta_2 \in \Theta$. Albeit it is well known that there is no UMP test for this sort of problem, there is a UMPU test at level α as the next proposition states.

Proposition. There exists a UMPU test at level $\alpha \in [0, 1]$ for testing the null hypothesis $H_0 : \theta_1 \leq \theta \leq \theta_2$ against the two-sided alternative hypothesis $H_1 : \theta \in \Theta_0^c$. In particular, this test is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) < c_1 \text{ or } T(x) > c_2, \\ \gamma_1 & \text{if } T(x) = c_1, \\ \gamma_2 & \text{if } T(x) = c_2, \\ 0 & \text{if } c_1 < T(x) < c_2, \end{cases}$$

where the constants γ_i and c_i ($i = 1, 2$) are determined by the size conditions $E_{\theta_1}\phi = E_{\theta_2}\phi = \alpha$.

Proof. Even though there is no UMP test for the hypotheses at hand, Lehmann has shown that there exists a UMP test $\tilde{\phi}$ at level $1 - \alpha$ for testing $\tilde{H}_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ against $\tilde{H}_1 : \theta_1 < \theta < \theta_2$. In particular, this test $\tilde{\phi}$ is exactly identical to the test $1 - \phi$, where ϕ is the candidate for UMPU test at level α . Because it is optimal, the test $\tilde{\phi}$ must be preferred to the constant test equal to $1 - \alpha$. Thus, $E_\theta \tilde{\phi} = E_\theta(1 - \phi) \geq 1 - \alpha$ for every θ such that $\theta_1 < \theta < \theta_2$. It follows then that $E_\theta \phi \leq \alpha$ for all θ , $\theta_1 < \theta < \theta_2$, so that the level of ϕ is at most equal to α . To complete the proof, it suffices to verify that ϕ is optimal in the class of α -similar tests on $H^* = \{\theta_1, \theta_2\}$, i.e. to show that the test $\tilde{\phi}$ minimizes $E_\theta \phi^*$ subject to the constraints $E_{\theta_1} \phi^* = E_{\theta_2} \phi^* = 1 - \alpha$ for every value $\theta \in \Theta_1$. ■

Corollary. There exists a UMPU test at level $\alpha \in [0, 1]$ for testing $H_0 : \theta = \theta_0$

against $H_1 : \theta \neq \theta_0$, where θ_0 is an interior point of Θ . This test is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) < c_1 \text{ or } T(x) > c_2, \\ \gamma_1 & \text{if } T(x) = c_1, \\ \gamma_2 & \text{if } T(x) = c_2, \\ 0 & \text{if } c_1 < T(x) < c_2, \end{cases}$$

where the constants γ_i and c_i ($i = 1, 2$) are determined by the conditions $E_{\theta_0}\phi = \alpha$ and $E_{\theta_0}(\phi T) = \alpha E_{\theta_0}T$.

Proof. First, consider the previous result for θ_2 converging to θ_1 . Given that the form of the optimal test remains the same, we basically need to show how the conditions determining the constants c_i and γ_i ($i = 1, 2$) change when θ_2 approaches θ_1 . Note however that these conditions can be written as $E_{\theta_1}\phi = \alpha$ and $E_{\theta_2}\phi - E_{\theta_1}\phi = 0$. Let $\theta_1 = \theta_0$ and $\theta_2 = \theta_0 + \epsilon$, where ϵ converges to zero. Then, $E_{\theta_0}\phi = \alpha$ and $[E_{\theta_0+\epsilon}\phi - E_{\theta_0}\phi]/\epsilon = 0$. At the limit, the second constraint becomes $\frac{\partial}{\partial\theta}E_{\theta}\phi|_{\theta=\theta_0} = 0$. However,

$$\begin{aligned} \left. \frac{\partial E_{\theta}\phi}{\partial\theta} \right|_{\theta=\theta_0} &= \left. \frac{\partial C(\theta)}{\partial\theta} \right|_{\theta=\theta_0} \int_{\mathcal{X}} \phi(x)h(x) \exp(Q(\theta_0)T(x))d\mu(x) \\ &\quad + \left. \frac{\partial Q(\theta)}{\partial\theta} \right|_{\theta=\theta_0} \int_{\mathcal{X}} \phi(x)T(x)C(\theta_0)h(x) \exp(Q(\theta_0)T(x))d\mu(x) \\ &= \int_{\mathcal{X}} \left(\left. \frac{\partial \log C(\theta)}{\partial\theta} \right|_{\theta=\theta_0} + \left. \frac{\partial Q(\theta)}{\partial\theta} \right|_{\theta=\theta_0} T(x) \right) \phi(x)\ell(x; \theta_0)d\mu(x) \\ &= \left. \frac{\partial \log C(\theta)}{\partial\theta} \right|_{\theta=\theta_0} E_{\theta_0}\phi + \left. \frac{\partial Q(\theta)}{\partial\theta} \right|_{\theta=\theta_0} E_{\theta_0}(\phi T). \end{aligned}$$

But the score vector has zero mean, hence

$$E_{\theta} \frac{\partial \log \ell(x; \theta)}{\partial\theta} = \frac{\partial \log C(\theta)}{\partial\theta} + \frac{\partial Q(\theta)}{\partial\theta} E_{\theta}T = 0.$$

Then it follows that $E_{\theta_0}(\phi T) = E_{\theta_0}\phi E_{\theta_0}T = \alpha E_{\theta_0}T$ due to $\frac{\partial}{\partial\theta}E_{\theta}\phi|_{\theta=\theta_0} = 0$. Note that, in particular, these conditions determining the constants c_i and γ_i imply that the test ϕ is orthogonal to the minimal sufficient statistic T under the null hypothesis in the sense that $\text{Cov}_{\theta_0}(\phi, T) = 0$. ■

Example. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ drawn from a normal distribution $N(\theta, 1)$ with mean $\theta \in \mathbb{R}$. Consider a testing problem defined by

$H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The model is exponential with density

$$\ell(\underline{x}, \theta) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{n\theta^2}{2} - \frac{1}{2} \sum_{i=1}^n x_i^2\right) \exp(n\theta\bar{x}).$$

The function $Q(\theta) = n\theta$ is continuous and strictly increasing, thus there exists a UMPU test based on the privileged statistic $T(\underline{X}) = \bar{X}$. In addition, under the null, the distribution of the sample mean is normal $N(\theta_0, 1/n)$, which is symmetric about θ_0 . Thus the UMPU test at level α is a nonrandomized test with critical region $W = \{|\bar{X} - \theta_0| > c\}$, where the constant c is such that $P_{\theta_0}(W) = \alpha$. Note that this condition can be written as

$$P_{\theta_0}(\sqrt{n}|\bar{X} - \theta_0| > \sqrt{nc}) = \alpha,$$

which implies that \sqrt{nc} equals the $(1 - \alpha/2)$ -th quantile $z_{1-\alpha/2}$ of a standard normal distribution given that $\sqrt{n}(\bar{X} - \theta_0) \sim N(0, 1)$.

6.5.2 UMPU tests in the presence of nuisance parameters

Frequently one is interested in testing only a subset of the parameters of the model. For instance, the problem can be that of testing a one-parameter hypothesis in a multidimensional parameter model.

Definition. Consider a parametric model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_\theta \mid \theta \in \Theta\})$. Let S be a sufficient statistic in the submodel $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_\theta \mid \theta \in \Theta^*\})$, where $\Theta^* \subset \Theta$. A test ϕ is said to have a Neyman α -structure relative to S and Θ^* if $E_\theta(\phi \mid S) = \alpha$ for every $\theta \in \Theta^*$.

Proposition 0. A test ϕ with Neyman α -structure relative to S and Θ^* is α -similar on Θ^* . Further, if S is complete in the submodel $(\mathcal{X}, \mathcal{A}, \{P_\theta \mid \theta \in \Theta^*\})$, then every α -similar test on Θ^* has a Neyman α -structure relative to S and Θ^* .

Proof. Let ϕ be a test with Neyman α -structure relative to S and Θ^* . Then, $E_\theta(\phi \mid S) = \alpha$ for all $\theta \in \Theta^*$ and so, by the law of iterated expectations, $E_\theta\phi = E_\theta E_\theta(\phi \mid S) = \alpha, \forall \theta \in \Theta^*$. Therefore the test ϕ is α -similar on Θ^* . Conversely, if a test is α -similar on Θ^* , $E_\theta\phi = E_\theta E_\theta(\phi \mid S) = \alpha, \forall \theta \in \Theta^*$. Thus

$E_\theta [E_\theta (\phi | S) - \alpha] = 0, \forall \theta \in \Theta^*$. From the definition of completeness, it follows that $E_\theta (\phi | S) = \alpha, \forall \theta \in \Theta^*$, which means that ϕ has Neyman α -structure with regard to S and Θ^* . In addition, as S is sufficient for the submodel defined by Θ^* , note that the subscript θ can be omitted in this conditional expectation, i.e. $E(\phi | S) = \alpha, \forall \theta \in \Theta^*$. ■

The preceding definition and Proposition 0 are directly applicable to exponential models with many parameters. Hereafter, it is assumed that the chosen parameterization of the multi-parameter exponential model corresponds to its canonical parameterization so that densities can be written as

$$\ell(x; \underline{\theta}) = C(\underline{\theta}) \exp \left(\theta^* T^*(x) + \sum_{i=1}^{p-1} \lambda_i S_i(x) \right),$$

with $\underline{\theta} = (\theta^*, \lambda_1, \dots, \lambda_{p-1})'$. Denote by $\underline{S} = (S_1, \dots, S_{p-1})'$ and by $\underline{\lambda} = (\lambda_1, \dots, \lambda_{p-1})'$. The parameter θ^* is the scalar parameter of interest on which the hypothesis is bearing, and $\underline{\lambda}$ is the vector of nuisance parameters.

The privileged statistic $\underline{T} = (T^*, \underline{S}')'$ is minimal sufficient and complete for the exponential model under consideration. Sufficiency by itself ensures that there are no costs in working with the induced model. The probability distributions of the model induced by the privileged statistic \underline{T} have densities of the form

$$\ell(\underline{t}, \underline{\theta}) = \tilde{C}(\underline{\theta}) \exp \left(\theta^* t^* + \sum_{i=1}^{p-1} \lambda_i s_i \right),$$

where $\underline{t} = (t^*, \underline{s}')'$ and $\underline{s} = (s_1, \dots, s_{p-1})'$. It is then easy to verify that the conditional distribution of T^* given $\underline{S} = \underline{s}$ has a density of the form $\ell(t^* | \underline{s}; \theta^*) = C_s(\theta^*) \exp(\theta^* t^*)$. The importance of conditioning is clear: The conditional distribution does not depend on the nuisance parameter vector $\underline{\lambda}$. Therefore, the results on one-parameter models can be applied to the conditional model for T^* given \underline{S} .

Proposition 1. Consider a multi-parameter exponential model as described above. Then there exists a UMPU test at level α for testing the null hypothesis

$H_0 : \theta^* \leq \theta_0^*$ against the alternative $H_1 : \theta^* > \theta_0^*$. The test is given by

$$\phi(t^*, \underline{s}) = \begin{cases} 1 & \text{if } t^* > c(\underline{s}), \\ \gamma(\underline{s}) & \text{if } t^* = c(\underline{s}), \\ 0 & \text{if } t^* < c(\underline{s}), \end{cases}$$

where the functions $c(\underline{s})$ and $\gamma(\underline{s})$ are determined by conditions on the risk of Type I error, namely $E_{\theta_0^*} [\phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] = \alpha, \forall \underline{s}$.

Proof. The first step consists in showing that the test ϕ is unbiased at level α . Consider then the conditional model for T^* given $\underline{S} = \underline{s}$, where \underline{s} is arbitrary. This model is parameterized by a single parameter θ^* . Thus, it follows from previous results that the test $\phi(T^*, \underline{s})$ is UMP at level α for testing $H_0 : \theta^* \leq \theta_0^*$ against $H_1 : \theta^* > \theta_0^*$. Moreover, any UMP test at level α is also UMPU at level α because the power of such a test is greater than or equal to the constant test $\phi' = \alpha$. Therefore, the test $\phi(T^*, \underline{s})$ is conditionally unbiased at level α , i.e.

$$\begin{aligned} E_{\underline{\theta}} [\phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] &\leq \alpha, & \forall \underline{\theta} \in \Theta_0 = \{\underline{\theta} : \theta^* \leq \theta_0^*\}, \\ E_{\underline{\theta}} [\phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] &\geq \alpha, & \forall \underline{\theta} \in \Theta_1 = \{\underline{\theta} : \theta^* > \theta_0^*\}. \end{aligned}$$

Applying the law of iterated expectations yields that $\phi(T^*, \underline{S})$ is (unconditionally) unbiased at level α , as required. The second step of the proof establishes the optimality of the test ϕ . Note that, in the exponential model, the statistic \underline{S} is minimal sufficient and complete for the nuisance parameter $\underline{\lambda}$ when $\underline{\theta} \in \Theta_0^* = \{\underline{\theta} : \theta^* = \theta_0^*\}$. Under these circumstances, any unbiased test is α -similar, which in turn has a Neyman α -structure relative to \underline{S} and Θ_0^* . Thereof it follows that a test which is more powerful than any other test with a Neyman α -structure relative to \underline{S} and Θ_0^* must be UMPU. Now let $\psi(T^*, \underline{S})$ be a test with a Neyman α -structure relative to \underline{S} and Θ_0^* , i.e. such that $E_{\theta_0^*} (\psi(T^*, \underline{S}) \mid \underline{S} = \underline{s}) = \alpha, \forall \underline{s}$. The power of such a test is

$$E_{\underline{\theta}} \psi = \int \left(\int \psi(t^*, \underline{s}) dP_{\theta_0^*}^{T^* \mid S=s}(t^*) \right) dP_{\underline{\theta}}^S(\underline{s}), \quad \underline{\theta} \in \Theta_1.$$

However, the test ϕ is UMP at level α in the conditional model for any \underline{s} , hence it maximizes the quantity in parentheses. Integrating then yields $E_{\underline{\theta}} \phi \geq E_{\underline{\theta}} \psi$ for every $\underline{\theta} \in \Theta_1$. ■

However, it is not always easy to the functions $c(\underline{s})$ and $\gamma(\underline{s})$ corresponding to the UMPU test. The determination of such functions may become easier by invoking the next result, which replace the functions $c(\underline{s})$ and $\gamma(\underline{s})$ by two constants c and γ .

Proposition 2. Under the assumptions of Proposition 1, if $U = f(T^*, \underline{S})$ is an ancillary statistic for $\underline{\theta} \in \Theta_0^* = \{\underline{\theta} : \theta^* = \theta_0^*\}$ and if f is strictly increasing in t^* for any given \underline{s} , then the test defined by

$$\phi^*(u) = \begin{cases} 1 & \text{if } u > c, \\ \gamma & \text{if } u = c, \\ 0 & \text{if } u < c, \end{cases}$$

with $E_{\theta_0^*} \phi^* = \alpha$, where $u = f(t^*, \underline{s})$, is UMPU at level α for testing $H_0 : \theta^* \leq \theta_0^*$ against $H_1 : \theta^* > \theta_0^*$.

Proof. It is straightforward to show that the test ϕ^* exists by using a reasoning similar to the proof of (i) in Neyman-Pearson's Theorem. Then, given that the function f is strictly increasing in t^* , the test ϕ^* can be written as

$$\phi^*(f(t^*, \underline{s})) = \begin{cases} 1 & \text{if } t^* > c(\underline{s}), \\ \gamma(\underline{s}) & \text{if } t^* = c(\underline{s}), \\ 0 & \text{if } t^* < c(\underline{s}). \end{cases}$$

Note that this form corresponds to an optimal test according to Proposition 1, hence it remains to verify that the conditions of Proposition 1 which determines $c(\underline{s})$ and $\gamma(\underline{s})$ hold, i.e. that ϕ^* has a Neyman α -structure relative to \underline{S} and Θ_0^* . But \underline{S} is sufficient and complete for $\theta \in \Theta_0^*$ and U is ancillary for $\theta \in \Theta_0^*$. From Basu's Theorem, it follows that \underline{S} and U are independent when $\theta \in \Theta_0^*$. This implies, in turn, that $\alpha = E_{\theta_0^*} \phi^* = E_{\theta_0^*} [\phi^*(U) | \underline{S} = \underline{s}]$ because ϕ^* is U -measurable. ■

Similar results can be obtained when the null hypothesis is of the form $H_0 : \theta_1^* \leq \theta^* \leq \theta_2^*$ or $H_0' : \theta^* = \theta_0^*$, that is to say, when the alternative hypothesis is two-sided. In what follows, some propositions, which are completely analogous to those obtained in the one-sided case, are stated without any proof.

Proposition 3. Consider the multi-parameter exponential model as described above. Then, there exists a UMPU test at level α for testing the hypothesis

$H_0 : \theta_1^* \leq \theta^* \leq \theta_2^*$ against $H_1 : \theta^* < \theta_1^*$ or $\theta^* > \theta_2^*$. The test is given by

$$\phi(t^*, \underline{s}) = \begin{cases} 1 & \text{if } t^* < c_1(\underline{s}) \text{ or } t^* > c_2(\underline{s}), \\ \gamma_1(\underline{s}) & \text{if } t^* = c_1(\underline{s}), \\ \gamma_2(\underline{s}) & \text{if } t^* = c_2(\underline{s}), \\ 0 & \text{if } c_1(\underline{s}) < t^* < c_2(\underline{s}), \end{cases}$$

where the functions $c_i(\underline{s})$ and $\gamma_i(\underline{s})$ are determined by conditions on the risk of Type I error, namely $E_{\theta_i^*} [\phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] = \alpha$, $\forall \underline{s}$ ($i = 1, 2$).

Proposition 4. Consider the multi-parameter exponential model as described above. Then, there exists a UMPU test at level α for testing the hypothesis $H_0 : \theta^* = \theta_0^*$ against $H_1 : \theta^* \neq \theta_0^*$. The test reads

$$\phi(t^*, \underline{s}) = \begin{cases} 1 & \text{if } t^* < c_1(\underline{s}) \text{ or } t^* > c_2(\underline{s}), \\ \gamma_1(\underline{s}) & \text{if } t^* = c_1(\underline{s}), \\ \gamma_2(\underline{s}) & \text{if } t^* = c_2(\underline{s}), \\ 0 & \text{if } c_1(\underline{s}) < t^* < c_2(\underline{s}), \end{cases}$$

where the functions $c_i(\underline{s})$ and $\gamma_i(\underline{s})$ are determined by a condition on the risk of Type I error, namely $E_{\theta_0^*} [\phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] = \alpha$, $\forall \underline{s}$ and another of conditional independence between T^* and $\phi(T^*, \underline{S})$ given \underline{S} , viz.

$$E_{\theta_0^*} [T^* \phi(T^*, \underline{S}) \mid \underline{S} = \underline{s}] = \alpha E_{\theta_0^*} [T^* \mid \underline{S} = \underline{s}], \quad \forall \underline{s}.$$

As in the one-sided case, the functions $c_i(\underline{s})$ and $\gamma_i(\underline{s})$ may be replaced by constant functions.

Proposition 5. Consider the multi-parameter exponential model as described above and the problem of testing null hypotheses such as $H_0 : \theta_1^* \leq \theta^* \leq \theta_2^*$ or $H'_0 : \theta^* = \theta_0^*$. In the first case, it is assumed that there is a statistic $U = f(T^*, \underline{S})$ which is ancillary on $\Theta_1^* \equiv \{\underline{\theta} : \theta^* = \theta_1^*\}$ and $\Theta_2^* \equiv \{\underline{\theta} : \theta^* = \theta_2^*\}$ such that the function f is strictly increasing in t^* for every given \underline{s} . In the second case, it is assumed that the statistic $U = f(T^*, \underline{S})$ is ancillary on $\Theta_0^* \equiv \{\underline{\theta} : \theta^* = \theta_0^*\}$ and that f is linear and strictly increasing in t^* for every given \underline{s} . Then there exists a UMPU test at level α , which is given by

$$\phi^*(u) = \begin{cases} 1 & \text{if } u < c_1 \text{ or } u > c_2, \\ \gamma_1 & \text{if } u = c_1, \\ \gamma_2 & \text{if } u = c_2, \\ 0 & \text{if } c_1 < u < c_2, \end{cases}$$

where $u = f(t^*, \underline{s})$ and the constants c_i and γ_i are determined by $E_{\theta_i^*} \phi^* = \alpha$ ($i = 1, 2$), when testing H_0 , and by $E_{\theta_0^*} \phi^* = \alpha$ and $E_{\theta_0^*} (U \phi^*) = \alpha E_{\theta_0^*} U$, when testing H'_0 .

Proof. Given that f is an strictly increasing function, then the test ϕ^* is of the form considered in Propositions 3 and 4, namely

$$\phi^*(f(t^*, \underline{s})) = \begin{cases} 1 & \text{if } t^* < c_1(\underline{s}) \text{ or } t^* > c_2(\underline{s}), \\ \gamma_1(\underline{s}) & \text{if } t^* = c_1(\underline{s}), \\ \gamma_2(\underline{s}) & \text{if } t^* = c_2(\underline{s}), \\ 0 & \text{if } c_1(\underline{s}) < t^* < c_2(\underline{s}). \end{cases}$$

In the case of testing H_0 , U is ancillary on Θ_1^* and Θ_2^* . Thus, U is independent of \underline{S} on Θ_1^* and Θ_2^* . Hence, $E_{\theta_i^*} \phi^* = E_{\theta_i^*} (\phi^*(U) | \underline{S} = \underline{s}) = \alpha$, $\forall \underline{s}$, $i = 1, 2$. Therefore, ϕ^* satisfies the optimality conditions of Proposition 3 as required.

In the case of testing H'_0 , U is ancillary on Θ_0^* , and so it is independent of \underline{S} on Θ_0^* . Hence, $E_{\theta_0^*} \phi^* = E_{\theta_0^*} (\phi^*(U) | \underline{S} = \underline{s}) = \alpha$, $\forall \underline{s}$. Moreover, using $U = a(\underline{S}) + b(\underline{S})T^*$ and the same independence argument, the condition $E_{\theta_0^*} (U \phi^*) = \alpha E_{\theta_0^*} U$ can be written as

$$\begin{aligned} E_{\theta_0^*} [a(\underline{S})\phi^* + b(\underline{S})T^*\phi^*] &= \alpha E_{\theta_0^*} a(\underline{S}) + \alpha E_{\theta_0^*} [b(\underline{S})T^*] \\ \Rightarrow E_{\theta_0^*} [b(\underline{S})T^*\phi^*] &= \alpha E_{\theta_0^*} [b(\underline{S})T^*] \\ \Rightarrow E_{\theta_0^*} [b(\underline{S})E_{\theta_0^*} (T^*\phi^* - \alpha T^* | \underline{S})] &= 0. \end{aligned}$$

However, \underline{S} is complete, hence $b(\underline{s})E_{\theta_0^*} (T^*\phi^* - \alpha T^* | \underline{S} = \underline{s}) = 0$, $\forall \underline{s}$. Further, since $b(\underline{s})$ is strictly positive everywhere, then

$$E_{\theta_0^*} [T^*\phi(T^*, \underline{S}) | \underline{S} = \underline{s}] = \alpha E_{\theta_0^*} [T^* | \underline{S} = \underline{s}], \quad \forall \underline{s}.$$

Thus, ϕ^* satisfies the optimality conditions of Proposition 4 as required. \blacksquare

Example 1. Consider a random sample $\underline{X} = (X_1, \dots, X_n)$ drawn from a normal distribution $N(\mu, \sigma^2)$. The densities of the model are of the form

$$\ell(\underline{x}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i\right).$$

This family of densities is exponential with natural parameters $\theta^* = -1/(2\sigma^2)$ and $\lambda = \mu/\sigma^2$ and privileged statistics $T^* = \sum_{i=1}^n x_i^2$ and $S = \sum_{i=1}^n x_i$. Now

consider the null hypothesis $\sigma^2 \leq \sigma_0^2$. This hypothesis bears on one natural parameter since it can be written as $H_0 : \theta^* \leq \theta_0^*$. Thus there exists a UMPU test of this hypothesis at level α . To determine such a test, note that

$$U = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma_0^2} \sim \chi^2(n-1)$$

when $\sigma^2 = \sigma_0^2$. Hence U is distribution free when $\theta^* = \theta_0^*$. In addition, U is also ancillary due to its (T^*, S) -measurability, viz.

$$U = \frac{1}{\sigma_0^2} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \frac{1}{\sigma_0^2} (T^* - S^2/n).$$

Moreover, U is a linear function increasing in T^* , so that it follows from Proposition 2 that a UMPU test at level α is given by the critical region

$$W = \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 \geq \sigma_0^2 \chi_{1-\alpha}^2(n-1) \right\},$$

where $\chi_{1-\alpha}^2(n-1)$ is the $(1-\alpha)$ -th quantile of the chi-square distribution with $n-1$ degrees of freedom.

Example 2. Consider the problem of testing independence in a two-by-two contingency table. Suppose that the frequencies n_{ij} ($i, j = 0, 1$) follow a multinomial distribution $M(n; p_{ij})$. The hypothesis of independence then reads

$$H_0 : p_{00} p_{11} = p_{01} p_{10},$$

or equivalently as

$$H_0 : \theta^* = \log p_{00} + \log p_{11} - \log p_{01} - \log p_{10} = 0.$$

Thus, the problem boils down to testing the nullity of a linear combination of the logarithms of the cell probabilities. Given the structure of a multinomial density, there is a UMPU test at level α for testing independence. In order to find the form of such an optimal test, it is helpful to rewrite the density of the model so as to single out the parameter of interest θ^*

$$\begin{aligned} \ell(n; \underline{p}) &= \frac{n!}{n_{00}!n_{10}!n_{01}!n_{11}!} p_{00}^{n_{00}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{11}^{n_{11}} \\ &= \frac{n!}{n_{00}!n_{10}!n_{01}!n_{11}!} \exp(n \log p_{11}) \exp(\theta^* T^* + \lambda_1 S_1 + \lambda_2 S_2), \end{aligned}$$

where $n = n_{00} + n_{01} + n_{10} + n_{11}$, $\lambda_1 = \log p_{01} - \log p_{11}$, $\lambda_2 = \log p_{10} - \log p_{11}$, $T^* = n_{00}$, $S_1 = n_{00} + n_{01}$, and $S_2 = n_{00} + n_{10}$. From Proposition 4, it follows that a UMPU test can be constructed from the conditional distribution of T^* given $\underline{S} = (S_1, S_2)'$. However, it is well known that, conditionally on S_2 , the random variables n_{00} and n_{01} are independent and follow respectively the Binomial distributions

$$B\left(S_2, \frac{p_{00}}{p_{00} + p_{10}}\right) \quad \text{and} \quad B\left(n - S_2, \frac{p_{01}}{p_{01} + p_{11}}\right).$$

Now, under the null hypothesis of independence, the ratio $p_{00}/(p_{00} + p_{10})$ and $p_{01}/(p_{01} + p_{11})$, which can be interpreted as conditional probabilities, are equal. Then, it is readily seen that the conditional distribution of T^* given \underline{S} under H_0 is a hypergeometric distribution, i.e.

$$P_{H_0}(T^* = t^* \mid \underline{S} = \underline{s}) = \frac{\binom{s_2}{t^*} \binom{n - s_2}{s_1 - t^*}}{\binom{n}{s_1}}.$$

In principle, it is possible to use a table of hypergeometric distributions to determine the constants $c_i(\underline{s})$ and $\gamma(\underline{s})$ of Proposition 4 for all \underline{s} . Such derivation, however, is relatively complex. This is one reason for introducing asymptotic tests.

6.5.3 Locally uniformly most powerful and unbiased tests

In the last subsection, the problem was of testing a null hypothesis regarding one parameter of a multi-parameter exponential model. In what follows, the setting is somewhat different: the model has only one parameter θ , but it is not necessarily exponential. The null hypothesis is of the form $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$. Note that techniques for finding a locally most powerful test no longer apply, because these methods reduce here to find a test ϕ which maximizes simultaneously

$$\left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta = \theta_0} \quad \text{and} \quad - \left. \frac{\partial E_\theta \phi}{\partial \theta} \right|_{\theta = \theta_0}$$

in the class of tests satisfying $E_{\theta_0}\phi = \alpha$. A possible solution is to restrict this search to tests which are locally unbiased at level α .

Definitions. Consider the problem of testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$.

1. A test is locally unbiased at level α if $E_{\theta}\phi$ is twice differentiable in θ_0 and if $E_{\theta_0}\phi = \alpha$,

$$\left. \frac{\partial E_{\theta}\phi}{\partial \theta} \right|_{\theta=\theta_0} = 0 \quad \text{and} \quad \left. \frac{\partial^2 E_{\theta}\phi}{\partial \theta^2} \right|_{\theta=\theta_0} > 0.$$

2. A test is locally uniformly most powerful and unbiased (LUMPU) at level α if it is locally unbiased at level α and if

$$\left. \frac{\partial^2 E_{\theta}\phi}{\partial \theta^2} \right|_{\theta=\theta_0} \geq \left. \frac{\partial^2 E_{\theta}\phi^*}{\partial \theta^2} \right|_{\theta=\theta_0}$$

for any other locally unbiased test ϕ^* at level α .

Thus, in order to find a LUMPU test at level α , one needs to solve the following optimization problem

$$\begin{aligned} \max_{\phi} \quad & \left. \frac{\partial^2 E_{\theta}\phi}{\partial \theta^2} \right|_{\theta=\theta_0} \\ \text{subject to} \quad & E_{\theta_0}\phi = \alpha \\ & \left. \frac{\partial E_{\theta}\phi}{\partial \theta} \right|_{\theta=\theta_0} = 0. \end{aligned}$$

In the parametric case where P_{θ} has a density $\ell(x; \theta)$ with respect to some measure μ , this optimization problem can be written as

$$\begin{aligned} \max_{\phi} \quad & \int_{\mathcal{X}} \phi \left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} d\mu(x) \\ \text{subject to} \quad & \int_{\mathcal{X}} \phi \ell(x; \theta_0) d\mu(x) = \alpha \\ & \int_{\mathcal{X}} \phi \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} d\mu(x) = 0. \end{aligned}$$

A simple extension of Neyman-Pearson's Theorem resolves this problem.

Proposition. The test defined by

$$\phi(x) = \begin{cases} 1 & \text{if } \left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} > k_1 \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + k_2 \ell(x; \theta_0), \\ \gamma & \text{if } \left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} = k_1 \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + k_2 \ell(x; \theta_0), \\ 0 & \text{if } \left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} < k_1 \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + k_2 \ell(x; \theta_0), \end{cases}$$

with $E_{\theta_0} \phi = \alpha$ and

$$\left. \frac{\partial E_{\theta} \phi}{\partial \theta} \right|_{\theta=\theta_0} = 0$$

is LUMPU at level α for testing the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$.

Proof. Suppose that ϕ satisfies the stated conditions. Then for any test ϕ^* ,

$$\int_{\mathcal{X}} (\phi - \phi^*) \left(\left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} - k_1 \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} - k_2 \ell(x; \theta_0) \right) d\mu(x) \geq 0.$$

If ϕ^* is also locally unbiased at level α , then using the constraints on ϕ reduces the preceding inequality to

$$\int_{\mathcal{X}} (\phi - \phi^*) \left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} d\mu(x) \geq 0,$$

which means that

$$\left. \frac{\partial^2 E_{\theta} \phi}{\partial \theta^2} \right|_{\theta=\theta_0} \geq \left. \frac{\partial^2 E_{\theta} \phi^*}{\partial \theta^2} \right|_{\theta=\theta_0}.$$

Therefore, ϕ is uniformly more powerful than any other locally unbiased test at level α . ■

Note that the condition

$$\left. \frac{\partial^2 \ell(x; \theta)}{\partial \theta^2} \right|_{\theta=\theta_0} > k_1 \left. \frac{\partial \ell(x; \theta)}{\partial \theta} \right|_{\theta=\theta_0} + k_2 \ell(x; \theta_0)$$

can be rewritten in terms of the score vector $S_{\theta_0}(x) = \partial \log \ell(x; \theta_0) / \partial \theta$, i.e.

$$\left. \frac{\partial S_{\theta}(x)}{\partial \theta} \right|_{\theta=\theta_0} + S_{\theta_0}^2(x) > k_1 S_{\theta_0}(x) + k_2.$$

Therefore, given a sampling model where $\ell(\underline{x}; \theta) = \prod_{i=1}^n \ell(x_i; \theta)$ and under some appropriate regularity conditions, the random variable $n^{-1/2} S_{\theta_0}(\underline{X})$ is asymptotically distributed as a centered normal variable under the null hypothesis.

Moreover, $(1/n) \frac{\partial}{\partial \theta} S_{\theta}(\underline{X})|_{\theta=\theta_0}$ converges almost surely to a constant c . Therefore the critical region obtained above for a sample size of n can be asymptotically expressed as

$$\left\{ \left(\frac{S_{\theta_0}(\underline{X})}{\sqrt{n}} \right)^2 > \frac{k_1(n)}{\sqrt{n}} \frac{S_{\theta_0}(\underline{X})}{\sqrt{n}} + \frac{k_2(n)}{n} - c \right\},$$

is asymptotically equivalent to a critical region of the form

$$\left\{ \frac{S_{\theta_0}(\underline{X})}{\sqrt{n}} > c_1(n) \right\} \cup \left\{ \frac{S_{\theta_0}(\underline{X})}{\sqrt{n}} < c_2(n) \right\}.$$

However, using the conditions of the preceding proposition indicates that the LUMPU test is of the form $\{|S_{\theta_0}(\underline{X})| > \sqrt{n}c^*\}$. Then, a similar reasoning to the argument put forward in subsection 6.2.2 shows that, under suitable regularity conditions, this LUMPU test is asymptotically nonrandomized with critical region

$$\left\{ |S_{\theta_0}(\underline{X})| > \sqrt{n\hat{I}(\theta_0)}z_{1-\alpha/2} \right\},$$

where $z_{1-\alpha}$ is the $(1 - \alpha)^{\text{th}}$ quantile of the standard normal distribution and $\hat{I}(\theta_0)$ stands for a consistent estimator of the Fisher information associated with one observation. This test is called a two-sided score test.