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

1 Statistical Inference: a decision theoretic approach

1.1 Anatomy of a statistical inference problem

Statistical inference is about making decisions, based on observed random variables, whose distributions are not fully specified. What do we need in order to describe and then, hopefully, solve such decision problems? Basically, three constituents are required—four, if a Bayesian approach is to be considered:

(i) a statistical model, describing the behavior of some observable random variable $\mathbf{X}$, called the *observation*;

(ii) a decision space, specifying a set of feasible decisions;

(iii) a loss function providing an evaluation of the various decisions, and

(iv) (Bayesian approach) a *prior distribution*.

1.1.1 Statistical models

A *statistical model* is a triple $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{X}$ is a space equipped with a σ-field $\mathcal{A}$ (the *observation space*) and $\mathcal{P} = \{\mathrm{P}\}$ is a collection of probability measures P over $(\mathcal{X}, \mathcal{A})$. This triple characterizes the possible behaviors or the possible data-generating processes for an observable random variable $\mathbf{X}$ called the *observation*. That random variable $\mathbf{X}$, with values in $(\mathcal{X}, \mathcal{A})$, has probability distribution P, where P is an unspecified element of $\mathcal{P}$.

When $\mathcal{P}$ is of the form $\mathcal{P} = \{\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}$, we call it a *parametric family*, yielding a *parametric model*, with *parameter* $\boldsymbol{\theta}$ ranging over the *parameter space* Θ. A family $\mathcal{P}$ that cannot be indexed by a finite-dimensional parameter is called a *nonparametric family*, yielding a *nonparametric model*.

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

The notation $\mathbf{x}$ is used for a point in $\mathcal{X}$ (a possible value of $\mathbf{X}$). As a random variable, $\mathbf{X}$ is defined over some measurable space $(\Omega, \mathcal{A}_\Omega)$, with elements ω. That space will seldom appear here. One always can think of $(\Omega, \mathcal{A}_\Omega) = (\mathcal{X}, \mathcal{A})$, with (for $\omega = \mathbf{x}$) $\mathbf{X}(\omega) = \mathbf{X}(\mathbf{x}) := \mathbf{x}$ (the identity mapping).

1.1.2 Decision spaces and decision rules

The *decision space* $\mathcal{D}$ is the collection of all possible actions d, equipped with some σ-field $\mathcal{B}_\mathcal{D}$. Combining the decision space and the statistical model just described, one can define *decision rules* or *statistical procedures* as follows.

A *pure* (*non-randomized*) *decision rule* is a measurable mapping

$$\delta : \mathbf{x} \in \mathcal{X} \longmapsto \delta(\mathbf{x}) \in \mathcal{D}$$

from the observation space $(\mathcal{X}, \mathcal{A})$ to the decision space $(\mathcal{D}, \mathcal{B}_\mathcal{D})$.

A *randomized decision rule* is a collection $\left\{\mathrm{P}^\delta_\mathbf{x} | \mathbf{x} \in \mathcal{X}\right\}$ of probability measures over $(\mathcal{D}, \mathcal{B}_\mathcal{D})$ such that for all $D \in \mathcal{B}_\mathcal{D}$ the mapping $\mathbf{x} \longmapsto \mathrm{P}^\delta_\mathbf{x}(D)$ be measurable (from $(\mathcal{X}, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$). For any given $D \in \mathcal{B}_\mathcal{D}$, thus, $\mathrm{P}^\delta_\mathbf{X}(D)$ is a correctly defined random variable, and the probability, when $\mathbf{X}$ has distribution P, of the final decision, δ, say, falling into D is $\mathrm{P}[\delta \in D] = \mathrm{E}_\mathrm{P}[\mathrm{P}^\delta_\mathbf{X}(D)] = \int_\mathcal{X} \mathrm{P}^\delta_\mathbf{x}(D)\mathrm{dP}$. In practice, under the randomized decision rule $\left\{\mathrm{P}^\delta_\mathbf{x}\right\}$, if the observation $\mathbf{X}$ takes the value $\mathbf{x}$, a decision is randomly selected according to the distribution $\mathrm{P}^\delta_\mathbf{x}$, that is, in such a way that the probability (conditional on $\mathbf{X} = \mathbf{x}$) that this decision falls into any $D \in \mathcal{B}_\mathcal{D}$ is $\mathrm{P}^\delta_\mathbf{x}(D)$. Conditionally on $\mathbf{X} = \mathbf{x}$, δ therefore can be considered as a random variable taking values in $\mathcal{D}$ and distribution function $\mathrm{P}^\delta_\mathbf{x}$. In the sequel, with a slight abuse of notation, we will write δ for any decision rule, whether it be pure or randomized.

Note that a pure decision rule is the particular case of a randomized decision rule for which the probability measures in $\left\{\mathrm{P}^\delta_\mathbf{x}\right\}$ are all degenerate—namely, such that, for any $\mathbf{x} \in \mathcal{X}$, there exists a $\delta(x) \in \mathcal{D}$ for which $\mathrm{P}^\delta_\mathbf{x}[\delta = \delta(x)] = 1$.

1.1.3 Loss functions

A *loss function* is a function mapping $\mathcal{D} \times \mathcal{P}$ onto the nonnegative real line:

$$(d, \mathrm{P}) \in \mathcal{D} \times \mathcal{P} \longmapsto L_{\mathrm{P}}(d) \in \mathbb{R}^+,$$

such that for all P, the mapping $d \longmapsto L_{\mathrm{P}}(d)$ is measurable. If the decision d is taken and the underlying distribution (the one that actually generated $\mathbf{X}$) is P, a loss $L_{\mathrm{P}}(d)$ is incurred: the loss function thus provides an evaluation of each possible decision d under each possible P.

In the parametric case with parameter $\boldsymbol{\theta}$, we will write $L_{\boldsymbol{\theta}}(d)$ instead of $L_{\mathrm{P}_{\boldsymbol{\theta}}}(d)$.

1.2 Risk functions and uniformly optimal decision rules

1.2.1 Risk functions

Combining constituents (i), (ii), and (iii) allows us to define, for any decision rule, a *risk function*. The risk function associated with a decision rule is the corresponding expected loss, considered as a function of $\mathrm{P} \in \mathcal{P}$. The risk function of a pure decision rule δ thus takes the form

$$\mathrm{P} \longmapsto R^{\delta}_{\mathrm{P}} := \mathrm{E}_{\mathrm{P}}\left[L_{\mathrm{P}}(\delta(\mathbf{X}))\right] = \int_{\mathcal{X}} L_{\mathrm{P}}(\delta(\mathbf{x}))\, \mathrm{dP}(\mathbf{x}), \quad \mathrm{P} \in \mathcal{P},$$

whereas, for a randomized decision rule $\left\{\mathrm{P}^{\delta}_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\right\}$, we have (for simplicity, we adopt the same notation R^{δ}_{P})

$$\mathrm{P} \longmapsto R^{\delta}_{\mathrm{P}} := \mathrm{E}_{\mathrm{P}}\left[\int_{\mathcal{D}} L_{\mathrm{P}}(d)\, \mathrm{dP}^{\delta}_{\mathbf{X}}\right] = \int_{\mathcal{X}} \int_{\mathcal{D}} L_{\mathrm{P}}(d)\, \mathrm{dP}^{\delta}_{\mathbf{x}}\, \mathrm{dP}(\mathbf{x}), \quad \mathrm{P} \in \mathcal{P}.$$

In the parametric case, we write R^{δ}_{θ} instead of $R^{\delta}_{\mathrm{P}_{\boldsymbol{\theta}}}$.

Risk functions are crucial in the comparison of decision rules. Actually, in order to compare two decision rules, we compare their risk functions. Whether pure or randomized, we say that a decision rule δ_1 *uniformly dominates* (weakly) a decision rule δ_2 (notation: $\delta_1 \succeq \delta_2$) if

$$R^{\delta_1}_{\mathrm{P}} \leq R^{\delta_2}_{\mathrm{P}} \quad \text{for all } \mathrm{P} \in \mathcal{P}.$$

We say that a decision rule δ^* is *uniformly optimal* or *uniformly minimum risk* (UMR) within a class $\mathcal{C}$ of decision rules if

(i) $\delta^* \in \mathcal{C}$, and

(ii) $\delta^* \succeq \delta$ for all $\delta \in \mathcal{C}$.

The problem is that, given two decision rules δ_1 and δ_2, in general, we neither have $\delta_1 \succeq \delta_2$ nor $\delta_2 \succeq \delta_1$: risk functions are not well-ordered for the binary relation $\succeq$. As a consequence, for given $\mathcal{C}$, a uniformly optimal decision rule δ^* in general does not exist. As a rule, the problem of finding an "optimal" statistical procedure δ is an ill-posed problem.

Faced with this ill-posed problem, the attitude of "classical" (also called "frequentist") and Bayesian statisticians differ. While classical statisticians will try to turn the problem into a well-posed one by imposing sensible restrictions on the class $\mathcal{C}$ via *statistical principles*, the Bayesian solution relies on an additional fourth constituent: the prior, the use of which we now briefly describe.

1.2.2 The Bayesian approach

Bayesians, in addition to (i) a statistical model, (ii) a decision space, and (iii) a loss function also consider a fourth constituent, (iv) a *prior distribution*.

A prior distribution or, simply, a prior, is a probability distribution Π over $\mathcal{P}$, equipped with an adequate σ-field—for simplicity, let us only consider the parametric case, where the prior can be defined over the parameter space $(\Theta, \mathcal{B}_\Theta)$, $\mathcal{B}_\Theta$ denoting the intersection of the Borel σ-field on $\mathbb{R}^k$ and Θ. Such a prior associates with each decision function δ a nonnegative real number—the *Bayesian risk*

$$R^\delta := \int_\Theta R^\delta(\boldsymbol{\theta}) \, \mathrm{d}\Pi(\boldsymbol{\theta}),$$

that is the expectation, under Π, of the risk $R^\delta(\boldsymbol{\theta})$ considered as a measurable function of $\boldsymbol{\theta}$.

Contrary to ordinary risk functions, Bayesian risks, which are real numbers, are well-ordered: for any δ_1 and δ_2, either $\mathrm{R}^{\delta_1} \geq \mathrm{R}^{\delta_2}$ ($\delta_2 \succeq_\Pi \delta_1$), or $\mathrm{R}^{\delta_1} \leq \mathrm{R}^{\delta_2}$ ($\delta_1 \succeq_\Pi \delta_2$), and optimal solutions, minimizing the Bayesian risk, typically exist. The resulting (weak) ordering $\succeq_\Pi$,

however, depends on the choice of the prior Π. In a sense, the prior reduces the family $\mathcal{P}$ of distributions to a unique element, the mixture $\int_\Theta \mathrm{P}_{\boldsymbol{\theta}}\,\mathrm{d}\Pi(\boldsymbol{\theta})$ (with mixing distribution Π) of the elements $\mathrm{P}_{\boldsymbol{\theta}}$ of $\mathcal{P}$.

We will not pursue along this way.

1.2.3 Statistical principles

If no optimal decision rule exists within a class $\mathcal{C}$ of decision rules, another solution consists in reducing $\mathcal{C}$ by means of some arbitrary but well-accepted conditions: *statistical principles*, such as the principles of *sufficiency* or *unbiasedness* (estimation and testing), the principles of *equivariance* (estimation) or *invariance* (testing), the *Neyman principle* (testing), the principle of *ancillarity* (treatment of *nuisances*), etc. Many of those principles will be considered in the subsequent chapters.

1.2.4 Other optimality concepts

Uniform dominance in some cases can be considered too strong a requirement. Other weaker optimality concepts, such as *stringency* or *minimaxity*, still over some class $\mathcal{C}$ of decision rules, can be considered. For instance, a decision rule δ^* is called *minimax* over $\mathcal{C}$ if

(i) $\delta^* \in \mathcal{C}$, and

(ii) $\max_{\mathrm{P}\in\mathcal{P}} R_{\mathrm{P}}^{\delta^*} \leq \max_{\mathrm{P}\in\mathcal{P}} R_{\mathrm{P}}^{\delta}$ for all $\delta \in \mathcal{C}$.

Minimaxity thus corresponds to a most "cautious" attitude consisting in evaluating each decision rule via its worst performance.

1.3 Examples

Let us conclude with a brief description of the two most usual problems of statistical inference: point estimation and hypothesis testing.

1.3.1 Point estimation

Consider a parametric model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\})$, and the problem of estimating $\boldsymbol{\theta}$. The decision space here is $\mathcal{D} = \Theta$; the pure decision rules, the *estimators* $\delta : \mathcal{X} \longrightarrow \Theta$ (i.e., for any $\mathbf{x} \in \mathcal{X}, \delta(\mathbf{x}) \in \Theta$). One also may be interested in estimating the value $g(\boldsymbol{\theta})$ of some given function g at $\boldsymbol{\theta}$, in which case $\mathcal{D} = g(\Theta)$. For simplicity, however, we concentrate on the problem of estimating $\boldsymbol{\theta}$ itself, and restrict to the scalar case $\theta \in \Theta \subseteq \mathbb{R}$.

Loss functions, in this context, typically involve bowl-shaped functions ℓ of the difference $(\delta(\mathbf{x}) - \theta)$ satisfying $\ell(0) = 0$. Examples are

– the *quadratic loss function* is $L_\theta\left(\delta\left(\mathbf{x}\right)\right) = \left(\delta(\mathbf{x}) - \theta\right)^2$;

– the *absolute deviation loss* is $L_\theta\left(\delta\left(\mathbf{x}\right)\right) = |\delta(\mathbf{x}) - \theta|$;

– "robust" loss functions, of the form (Huber 1964)

$$L_\theta\left(\delta\left(\mathbf{x}\right)\right) = \begin{cases} \ell(\delta(\mathbf{x}) - \theta) & \text{if} \quad |\ \delta(\mathbf{x}) - \theta| \leq c \\ \ell(c) + \ell'(c)(\delta(\mathbf{x}) - \theta - c) & \text{if} \quad \delta(\mathbf{x}) - \theta > c \\ \ell(-c) + \ell'(-c)(\delta(\mathbf{x}) - \theta + c) & \text{if} \quad \delta(\mathbf{x}) - \theta < -c \end{cases}$$

(where ℓ' stands for the derivative of ℓ), or

– bounded loss functions, as

$$L_\theta\left(\delta\left(\mathbf{x}\right)\right) = \begin{cases} \ell(\delta(\mathbf{x}) - \theta) & \text{if} \quad \ell(\delta(\mathbf{x}) - \theta) \leq m \\ m & \text{if} \quad \ell(\delta(\mathbf{x}) - \theta) > m. \end{cases} \tag{1}$$

Most sensible loss functions, in this context, are *convex* functions of $(\delta(\mathbf{x}) - \theta)$, a property bounded functions such as (1) cannot possess. The corresponding risks are

– the *quadratic* or L_2 *risk* $\mathrm{R}_\theta^\delta = \mathrm{E}_\theta\left[\left(\delta(\mathbf{x}) - \theta\right)^2\right]$

– the *expected absolute deviation* or L_1 *risk* $\mathrm{R}_\theta^\delta = \mathrm{E}_\theta[|\delta(\mathbf{x}) - \theta|]$, etc.

As already mentioned, a uniformly optimal estimator in general does not exist unless we put restrictions on the class $\mathcal{C}$ of estimators under consideration. Typically, an estimator

having uniformly minimal (quadratic, absolute deviation,...) risk within the class $\mathcal{C}$ of *all* estimators of θ does not exist. Indeed, no δ can "beat" the degenerate estimator

$$\delta(\mathbf{x}) := \theta_0 \quad \text{for all } \mathbf{x} \in \mathcal{X} \tag{2}$$

uniformly over Θ: its L_2 and L_1 risks reduce to $(\theta - \theta_0)^2$ and $|\theta - \theta_0|$, respectively, taking value zero at θ_0. An estimator which has risk zero at $\theta = \theta_0$ clearly is unbeatable at θ_0, but is extremely bad away from θ_0, and does not exploit any of the information contained in $\mathbf{X}$. If we impose $\mathcal{C} = \mathcal{C}_0$, the class of *unbiased estimators* or $\mathcal{C} = \mathcal{C}_{\mathcal{G}}$, the class of estimators that are *shift-equivariant* (see a later chapter), such degenerate estimators as (2) are ruled out, and uniformly optimal solutions may exist.

For instance, let $\mathbf{X}$ denote an i.i.d. sample $X_1, \ldots, X_n$, with $\theta = \mathrm{E}[X_1]$. It follows from the Lehmann-Scheffé theorem (see a later chapter again) that the sample mean $\overline{X} := \frac{1}{n}\sum X_i$ has uniformly minimum risk within the class of unbiased estimators (we say that it is *uniformly minimum risk unbiased* (UMRU)), irrespective of the (convex) loss function ... This at first sight is a very strong property. The condition of unbiasedness (under any distribution for which $\mathrm{E}[X_1]$ exists and is finite), however, puts a very severe restriction on the class $\mathcal{C}$ of $\overline{X}$'s competitors.

Another classical example is the optimality of the ordinary least squares (OLS) estimator in the general linear model. The observation here, traditionally denoted as $\mathbf{Y}$, is a $n \times 1$ vector satisfying

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{X}$ is a $n \times k$ matrix of constants with maximal rank $k \leq n$ and $\mathbf{e}$ is an unobserved $n \times 1$ random error with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$. The Gauss-Markov theorem tells us that the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the *best linear unbiased estimator* (BLUE), meaning that it achieves uniformly minimum quadratic risk within the class of linear (in $\mathbf{Y}$) unbiased estimators. Again, this optimality property perhaps is not as strong as it may look, since the class of linear unbiased estimators essentially reduces to weighted means of $\mathbf{Y}$.

1.3.2 Hypothesis testing

Hypothesis testing problems, in a sense, are the simplest of all statistical inference problems, as the decision space only contains two elements: *rejection* and *non-rejection*.

Consider a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, and assume that $\mathcal{P}$ has been divided into two nonoverlapping parts, which we denote as $\mathcal{P} = \mathcal{H}_0 \oplus \mathcal{H}_1$: $\mathcal{H}_0$ is called the *null hypothesis*, $\mathcal{H}_1$ the *alternative*. The decision space $\mathcal{D} = \{\mathrm{R}\mathcal{H}_0, \not{\mathrm{R}}\mathcal{H}_0\} = \{1, 0\}$ contains two points: $\mathrm{R}\mathcal{H}_0$ (rejecting $\mathcal{H}_0$) and $\not{\mathrm{R}}\mathcal{H}_0$ (not rejecting $\mathcal{H}_0$), which we conveniently will code as 1 and 0. In the parametric case, the same partition of $\mathcal{P}$ into a null and an alternative is generally represented as the corresponding partition of the parameter space: $\Theta = \mathcal{H}_0 \oplus \mathcal{H}_1$. Whether $\mathcal{H}_0$ and $\mathcal{H}_1$ are subsets of $\mathcal{P}$ or Θ in general is clear from the context, and we will use both acceptions.

Decision rules in this context are denoted by ϕ instead of δ. By definition, pure decision rules are measurable functions from $(\mathcal{X}, \mathcal{A})$ to $\{0, 1\}$: call them (pure or nonrandomized) *tests*. A test ϕ is thus a statistic with values in $\{0, 1\}$, hence the *indicator* of the region $\{\mathbf{x} : \phi(\mathbf{x}) = 1\} \in \mathcal{A}$ of all observation values leading to rejection, also called the *critical region* of ϕ.

We also will consider randomized decision rules. Since a distribution over the two-point set $\mathcal{D} = \{0, 1\}$ is entirely characterized by the probability it puts on the value 1, a randomized decision rule or *randomized test* associates such a probability will each point $\mathbf{x}$ in the observation space $\mathcal{X}$: the probability of rejecting $\mathcal{H}_0$ when $\mathbf{x}$ has been observed. The same notation ϕ is conveniently used for that mapping, and we henceforth define a randomized test as a measurable mapping ϕ from $(\mathcal{X}, \mathcal{A})$ to the interval $[0, 1]$. This is coherent, as it redefines a pure test as the special case of a randomized test taking only the values $\phi(\mathbf{x}) = 0$ (rejection with probability zero) and $\phi(\mathbf{x}) = 1$ (rejection with probability 1).

In practice, thus, a test ϕ, conditional on $\mathbf{X} = \mathbf{x}$, rejects $\mathcal{H}_0$ with probability $\phi(\mathbf{x})$. Then, $\mathrm{E}_\mathrm{P}[\phi(\mathbf{X})]$ is the probability that ϕ leads to rejection of the null hypothesis under P.

Loss functions, since $\mathcal{D}$ has only two points, 0 and 1, are fully characterized by

$$
\begin{aligned}
L_{\mathrm{P}}(0) &= \begin{cases} 0 & \text{if} \quad \mathrm{P} \in \mathcal{H}_0 \quad \text{(the cost of not rejecting } \mathcal{H}_0 \text{ when it is true is zero)} \\ a > 0 & \text{if} \quad \mathrm{P} \in \mathcal{H}_1 \quad \text{(the cost of not rejecting } \mathcal{H}_0 \text{ when it is false is } a\text{)} \end{cases} \\
L_{\mathrm{P}}(1) &= \begin{cases} b > 0 & \text{if} \quad \mathrm{P} \in \mathcal{H}_0 \quad \text{(the cost of rejecting } \mathcal{H}_0 \text{ when it is true is } b\text{)} \\ 0 & \text{if} \quad \mathrm{P} \in \mathcal{H}_1 \quad \text{(the cost of rejecting } \mathcal{H}_0 \text{ when it is false is zero),} \end{cases}
\end{aligned}
$$

with $a > 0$ and $b > 0$. The resulting risk (for a test ϕ, be it pure or randomized) is

$$
R_{\mathrm{P}}^{\phi} = \begin{cases} b\mathrm{E}_{\mathrm{P}}[\phi(\mathbf{X})] & \text{if} \quad \mathrm{P} \in \mathcal{H}_0 \quad \text{(called the } \textit{type I risk}\text{)} \\ a\mathrm{E}_{\mathrm{P}}[1-\phi(\mathbf{X})] & \text{if} \quad \mathrm{P} \in \mathcal{H}_1 \quad \text{(called the } \textit{type II risk}\text{).} \end{cases}
$$

Minimizing type I risk implies minimizing the *size* $\mathrm{E}_{\mathrm{P}}[\phi(\mathbf{X})]$ of ϕ under $\mathrm{P} \in \mathcal{H}_0$, whereas minimizing type II risk implies maximizing the *power* $\mathrm{E}_{\mathrm{P}}[\phi(\mathbf{X})]$ of ϕ under $\mathrm{P} \in \mathcal{H}_1$.

A test ϕ^* is uniformly optimal within the class $\mathcal{C}$ of all tests if its risk is uniformly minimal, that is, if, for all $\phi \in \mathcal{C}$,

$$
\begin{cases} b\mathrm{E}_{\mathrm{P}}[\phi^*(\mathbf{X})] \leq b\mathrm{E}_{\mathrm{P}}[\phi(\mathbf{X})] & \text{for all } \mathrm{P} \in \mathcal{H}_0 \\ a\mathrm{E}_{\mathrm{P}}[1-\phi^*(\mathbf{X})] \leq a\mathrm{E}_{\mathrm{P}}[1-\phi(\mathbf{X})] & \text{for all } \mathrm{P} \in \mathcal{H}_1, \end{cases}
$$

or, equivalently, since the constants a and b obviously play no role,

$$
\begin{cases} \mathrm{E}_{\mathrm{P}}[\phi^*] \leq \mathrm{E}_{\mathrm{P}}[\phi] & \text{for all } \mathrm{P} \in \mathcal{H}_0 \qquad (3a) \\ \mathrm{E}_{\mathrm{P}}[\phi^*] \geq \mathrm{E}_{\mathrm{P}}[\phi] & \text{for all } \mathrm{P} \in \mathcal{H}_1. \qquad (3b) \end{cases}
$$

In particular, (3a) should hold for the test $\phi(\mathbf{x}) = 0$ (the test that rejects with probability zero, irrespective of $\mathbf{x}$), and (3b) for the test $\phi(\mathbf{x}) = 1$ (the test that rejects with probability one, irrespective of $\mathbf{x}$), so that ϕ^* should satisfy $\mathrm{E}_{\mathrm{P}}[\phi^*] = 0$ for all $\mathrm{P} \in \mathcal{H}_0$ and $\mathrm{E}_{\mathrm{P}}[\phi^*] = 1$ for all $\mathrm{P} \in \mathcal{H}_1$, that is, achieve risk zero uniformly. Such a test clearly does not exist, unless the problem is degenerate (for instance, distributions in $\mathcal{H}_0$ and $\mathcal{H}_1$ have disjoint supports).

If an optimal solution is to be found, the class $\mathcal{C}$ thus should be restricted via some statistical principle. The most classical one is the *Neyman principle*, under which an upper bound α is imposed on the size. More precisely, consider the class $\mathcal{C}_\alpha$ of α-*level tests*, namely,

the tests satisfying the probability level condition

$$\mathrm{E}_{\mathrm{P}}[\phi^*] \leq \alpha \text{ for all } \mathrm{P} \in \mathcal{H}_0, \tag{3}$$

where the predefined level $\alpha \in (0, 1)$ takes conventional values of $1\%, 5\%, 10\%$, etc. Along with this restriction to $\mathcal{C}_\alpha$, the cost a of type I error is put to zero, and the problem (without loss of generality, one can take $b = 1$) consists in maximizing the power within $\mathcal{C}_\alpha$. A *uniformly most powerful* (UMP) α-level test ϕ^* then is such that

(i) $\phi^* \in \mathcal{C}_\alpha$, that is, $\mathrm{E}_{\mathrm{P}}[\phi^*] \leq \alpha$ for all $\mathrm{P} \in \mathcal{H}_0$, and

(ii) $\mathrm{E}_{\mathrm{P}}[\phi^*] \geq \mathrm{E}_{\mathrm{P}}[\phi]$ for all $\phi \in \mathcal{C}_\alpha$ and all $\mathrm{P} \in \mathcal{H}_1$.

Under such an approach, optimal tests can be found for simple problems; further restrictions of the class $\mathcal{C}_\alpha$, however, may be needed, based on the principles of *unbiasedness*, *invariance*, etc.