Testing Statistical Hypotheses
Lecture Notes

Vladimir Koltchinskii

October 19, 2020

1 Preliminaries on Probability and Measure Theory

In probability theory, we deal with random variables $X : \Omega \mapsto S$ defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{A})$. It is always assumed that the mapping X is measurable, which means that for all $A \in \mathcal{A}$

$$X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A)\} \in \Sigma.$$

This allows one to define probabilities of all events $\{X \in A\}, A \in \mathcal{A}$ and to define the distribution of random variable X as a probability measure $P = \mathbb{P} \circ X^{-1}$,

$$P(A) := (\mathbb{P} \circ X^{-1})(A) = \mathbb{P}(\{X \in A\}) = \mathbb{P}(\{\omega : X(\omega) \in A\}), A \in \mathcal{A}.$$

If X has distribution P, we will often write $X \sim P$ (meaning X has been sampled from the distribution P).

The distributions of random variables (r.v.) X are most often described by their densities. If μ is a given fixed measure (finite or σ-finite) on $(S, \mathcal{A})$, we say that P is absolutely continuous with respect to μ, $P \ll \mu$, if, for all $A \in \mathcal{A}$, $\mu(A) = 0$ implies $P(A) = 0$. By Radon-Nikodym Theorem, if $P \ll \mu$, then there exists a non-negative μ-integrable function $p : S \mapsto \mathbb{R}_+$ such that

$$P(A) = \int_A p(x)\mu(dx), A \in \mathcal{A}.$$

Moreover, such a function is unique in the sense that, for any two functions p_1, p_2 satisfying these properties, $p_1 = p_2$ μ a.s. Such a function p is called the density of measure P with respect to measure μ (or also Radon-Nikodym derivative of P w.r.t. μ, $p = \frac{dP}{d\mu}$). For a probability measure P, the density is integrable to 1 : $\int_S p d\mu = 1$.

In the case when $S = \mathbb{R}^d$ equipped with Borel σ-algebra of sets, measure μ is typically the Lebesgue measure on $\mathbb{R}^d$ and the integrals are written in the form

$$P(A) = \int_A p(x)dx.$$

Another common case is when S is discrete (finite or countable) equipped with σ-algebra of all its subsets. In this case, μ is typically the counting measure $\mu(A) := \text{card}(A), A \subset S$. In this case, the integral becomes a sum

$$P(A) = \int_A p d\mu = \sum_{x \in A} p(x)$$

and the density p is called the probability mass function of distribution P.

Two probability distributions P and Q on $(S, \mathcal{A})$ are called mutually singular ($P \perp Q$) iff there exists a set $A \in \mathcal{A}$ such that $P(A) = 0$ and $Q(S \setminus A) = 0$. In som sense, these

two distributions have disjoint supports. If P and Q have densities p and q with respect to μ this would be the case $pq = 0$ μ a.s.

In statistical problems, it is assumed that observed data is sampled at random from an unknown distribution P in some space S (often called the sample space of statistical experiment). The data could be represented by a random variable X defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in S with distribution P. Any measurable function $T(X)$ of data X is called a statistic.

The set $\mathcal{P}$ of all possible distributions in a particular problem is called a statistical model. It could range from all the distributions on measurable space $(S, \mathcal{A})$ (if nothing is known about the distribution P of the data X) to relatively small sets of distributions (such as all normal distributions in the real line). We will often write $X \sim P, P \in \mathcal{P}$ to say that $\mathcal{P}$ is used as a model for data X.

In many cases, there is a measure μ on $(S, \mathcal{A})$ (finite or σ-finite) such that, for all $P \in \mathcal{P}$, $P \ll \mu$ implying that P has a density $p = \frac{dP}{d\mu}$ with respect to measure μ. In such cases, we say that model $\mathcal{P}$ is dominated by μ. It is convenient to describe dominated models in terms of densities. Say, if $\tilde{\mathcal{P}} := \{\frac{dP}{d\mu} : P \in \mathcal{P}\}$, we can write $X \sim p, p \in \tilde{\mathcal{P}}$. If $\mathcal{P} := \{P_1, \dots, P_N\}$ is a finite set of distributions (which is often the case in hypotheses testing problems), one can always set $\mu := P_1 + \dots + P_N$ to define a dominating measure for our model. Similarly, if $\mathcal{P} := \{P_1, P_2, \dots\}$ is countable, one can set $\mu := \sum_{n \geq 1} 2^{-n} P_n$. More generally, it is known that model $\mathcal{P}$ is dominated by a σ-finite measure μ if and only if there exists $\mathcal{P}_0 \subset \mathcal{P}$ such that $\mathcal{P}_0$ is countable and, for all $A \in \mathcal{A}$, the condition $P(A) = 0, P \in \mathcal{P}_0$ implies that $P(A) = 0, P \in \mathcal{P}$ (see, e.g., Lehmann, Theorem A.4.2).

It is common in statistics to parameterize statistical models by some parameter θ from a parameter space Θ. In this case, the model is the family of distributions $\{P_\theta : \theta \in \Theta\}$, or their densities w.r.t. μ $\{p_\theta : \theta \in \Theta\}$. We will also write in such cases $X \sim P_\theta, \theta \in \Theta$ or $X \sim p_\theta, \theta \in \Theta$, where θ is an unknown parameter of the model. Typical examples are: normal model $X \sim N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$ in the real line, Poisson model $X \sim \mathcal{P}(\theta), \theta > 0$, uniform model $X_1, \dots, X_n$ i.i.d. $\sim U[0, \theta], \theta > 0$, binomial model $X \sim B(n, \theta), \theta \in [0, 1]$, multivariate normal models, Gamma distributions, Beta distributions, etc. One can also consider models with infinite-dimensional parameter (called in statistics nonparametric models) such as $X_1, \dots, X_n$ i.i.d. $\sim \theta$, where θ is an unknown density of i.i.d. observations $X_1, \dots, X_n$ in $\mathbb{R}$ that belongs to some set Θ of densities in the real line (for instance, densities of certain degree of smoothness). Note that in this case the notation $X_1, \dots, X_n$ i.i.d. $\sim \theta$ is not quite precise. One should rather define the joint density

$$p_\theta(x_1, \dots, x_n) := \theta(x_1) \dots \theta(x_n)$$

of $(X_1, \dots, X_n)$ and write $(X_1, \dots, X_n) \sim p_\theta, \theta \in \Theta$.

It will be said that model $\{P_\theta : \theta \in \Theta\}$ is identifiable (or, more precisely, its parameterization is identifiable) iff $\theta_1 \neq \theta_2, \theta_1, \theta_2 \in \Theta$ implies that $P_{\theta_1} \neq P_{\theta_2}$ (in terms of densities p_θ, it means that p_{θ_1} and p_{θ_2} do not coincide μ a.s.).

Note that since there is a set $\mathcal{P}$ of possible distributions of random variable X, there should be also a set of probability measures $\mathbb{P}_P : P \in \mathcal{P}$ on (Ω, Σ) that would generate these distributions. The simplest approach is just to use $(S, \mathcal{A})$ as the underlying probability space and to equip it with measure $\mathbb{P}_P = P$. If the data consists of i.i.d. observations $X_1, x_2, \cdots \sim P$ in space $(S, \mathcal{A})$, one can set $\Omega = S^\infty = S \times S \times \ldots$, equip with σ-algebra $\Sigma = \mathcal{A}^\infty = \mathcal{A} \times \mathcal{A} \times \ldots$ and with probability measure $\mathbb{P}_P := P^\infty := P \times P \times \ldots$. Usually, it is not important how exactly the underlying probability spaces $(\Omega, \Sigma, \mathbb{P}_P)$ is constructed. But the subscript P indicating what is the underlying distribution of the data is commonly used for both probabilities and expectations. Say, if $T(X)$ is a real valued statistic for the model $X \sim P, P \in \mathcal{P}$, we would write $\mathbb{P}_P\{T(X) \leq t\}$, $\mathbb{E}_P T(X)$ to indicate that $X \sim P$ when the probability or the expectation is computed. In the case when the model is described in terms of density p, we could write $\mathbb{P}_p\{T(X) \leq t\}$ and $\mathbb{E}_p T(X)$, and, in the case when the model is parameterized with parameter θ, say, $X \sim P_\theta, \theta \in \Theta$, we would write $\mathbb{P}_\theta\{T(X) \leq t\}$ and $\mathbb{E}_\theta T(X)$.

2 Simple Hypotheses Testing: Neyman-Pearson Lemma

Let X be a random variable with values in a measurable space $(S, \mathcal{A})$ defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. It is known that X has one of two possible distributions, P or Q. Our goal is to test the hypothesis $H_0 : X \sim P$ against the alternative $H_a : X \sim Q$ based on an observation of random variable X. Such a problem will be called *a simple hypotheses testing.* In what follows, H_0 is called the null hypotheses and H_a is called the alternative.

Definition 2.1 *A (randomized) test for the null hypotheses H_0 against the alternative H_a is a measurable function $\phi : S \mapsto [0, 1]$. We will denote the set of all such functions by Φ.*

Test ϕ will be used as follows: given that $X = x$, a coin is tossed with probability of getting a head equal to $\phi(x)$ and probability of getting a tail $1 - \phi(x)$. We reject the null hypothesis H_0 if the toss results in a head and we do not reject it if it results in a tail. Such a randomized procedure is an example of what is called in statistical decision theory a decision rule. A deterministic test is a test ϕ that takes only the values 0 and 1.

Note also that the set Φ of all tests is convex: if $\phi_1, \phi_2 \in \Phi$ and $\lambda \in [0, 1]$, then $\lambda\phi_1 + (1 - \lambda)\phi_2 \in \Phi$. The test $\lambda\phi_1 + (1 - \lambda)\phi_2$ could be viewed as a mixture of tests ϕ_1 and ϕ_2 : it uses ϕ_1 with probability λ and ϕ_2 with probability $1 - \lambda$.

Each test ϕ could make two possible errors:

- to reject H_0 when it is true (type I error);
- not to reject H_0 when H_a is true (type II error).

Definition 2.2 *Let*

$$\alpha_\phi := \mathbb{E}_P \phi(X) = \int_S \phi dP$$

and

$$\beta_\phi := \mathbb{E}_Q \phi(X) = \int_S \phi dQ.$$

α_ϕ the significance level or the size of test ϕ and β_ϕ will be called the power of test ϕ.

Note that α_ϕ is the probability of type I error and $1 - \beta_\phi$ is the probability of type II error. Ideally, one would like to find a test with as large as possible power and as small as possible significance level. A natural question to ask is:

• is there a test $\phi^\star$ for which $\alpha_\phi = 0$ and $\beta_\phi = 1$?

Let us call such a test $\phi^\star$ "the best test".

Proposition 2.1 *The best test $\phi^\star$ exists if and only if the distributions P and Q are mutually singular, $P \perp Q$ (in other words, if there exists a set $A \in \mathcal{A}$ such that $P(A) = 0$ and $Q(A^c) = 0$).*

The condition that $P \perp Q$ means that distributions P and Q have disjoint supports. This makes testing problem rather trivial and, for such problems, one could indeed find the best test that makes errors with probability zero. In more interesting statistical problems, the distributions could have overlapping supports (such as, for instance, two normal distributions in real line), and testing becomes less trivial. One way to pose a simple hypotheses testing problems as an optimization problem is to try to find a test that maximizes the power β_ϕ subject to a constraint that the significance level is not too large, say, $\alpha_\phi \leq \alpha$ for some level $\alpha \in [0, 1]$.

Definition 2.3 *For $\alpha \in [0, 1]$, let us say that a test ϕ is of size α (more precisely, of size at most α) iff $\alpha_\phi \leq \alpha$. A test ϕ^* is a most powerful test of size α iff it is of size α and*

$$\beta_{\phi^*} \geq \beta_\phi$$

for all tests ϕ of size α.

In other words, most powerful tests of size α are solutions of the following optimization problem:

$$\phi^* \in \text{Argmin}\{\beta_\phi : \alpha_\phi \leq \alpha, \phi : S \mapsto [0, 1]\}.$$

Since $\phi \mapsto \alpha_\phi = \int_S \phi dP$ and $\phi \mapsto \beta_\phi = \int_S \phi dQ$ are linear functionals, this optimization problem is an example of an infinite-dimensional linear program. This problem has a very simple solution described by *Neyman-Pearson Lemma.*

Without loss of generality, we can assume that distributions P and Q are absolutely continuous with respect to a σ-finite measure μ on $(S, \mathcal{A})$ with densities p and q (if there is no natural reference measure λ on $(S, \mathcal{A})$, one could always set $\lambda := P + Q$).

Definition 2.4 *Let*

$$L(X) := \frac{q(X)}{p(X)}.$$

$L(X)$ will be called the likelihood ratio statistic. For $c \in \mathbb{R}_+, \gamma \in [0,1]$, define

$$\phi_{c,\gamma}(X) := \begin{cases} 1 & \text{if } L(X) > c \\ \gamma & \text{if } L(X) = c \\ 0 & \text{if } L(X) < c. \end{cases}$$

Tests $\phi_{c,\gamma}, c \in \mathbb{R}_+, \gamma \in [0,1]$ will be called Neyman-Pearson tests.

We are now ready to state Neyman-Pearson Lemma.

Theorem 2.1 *Consider a simple hypotheses testing problem $H_0 : X \sim P$ against $H_a : X \sim Q$. Let $\alpha \in (0,1)$ be a significance level. The following statements hold:*

1. *There exist $c \in \mathbb{R}_+, \gamma \in [0,1]$ such that $\alpha_{\phi_{c,\gamma}} = \alpha$.*
2. *For such a choice of c, γ, Neyman-Pearson test $\phi_{c,\gamma}$ is most powerful of size α.*
3. *If ϕ^* is a most powerful test of size α, then $\phi^*(x) = \phi_{c,\gamma}(x)$ μ a.s. on the set $\{x : L(x) \neq c\}$. Moreover, $\alpha_{\phi^*} = \alpha$ unless $\beta_{\phi^*} = 1$.*

Proof. *Claim 1.* Let

$$F(c) := \mathbb{P}_p\{L(X) \leq c\}, c \geq 0,$$

be the distribution function of likelihood ratio statistics $L(X)$ under the null hypothesis. Then

$$\begin{aligned} \alpha_{\phi_{c,\gamma}} = \mathbb{E}\phi_{c,\gamma}(X) &= \mathbb{P}_p\{L(X) > c\} + \gamma \mathbb{P}_p\{L(X) = c\} \\ &= 1 - F(c) + \gamma(F(c) - F(c-)). \end{aligned}$$

The condition $\alpha_{\phi_{c,\gamma}} = \alpha$ is then equivalent to

$$F(c) - \gamma(F(c) - F(c-)) = 1 - \alpha.$$

By standard properties of distribution function, either there exists a value of c such that $F(c) = 1 - \alpha$, or there exists a value of c such that $F(c-) \leq 1 - \alpha < F(c)$. In the first case, set $\gamma := 0$ and in the second case set

$$\gamma := \frac{F(c) - (1 - \alpha)}{F(c) - F(c-)}.$$

With such choices of c and γ, we do have $\alpha_{\phi_{c,\gamma}} = \alpha$, which proves the first claim of the theorem.

Claim 2. To prove the second claim, let ϕ be a test with $\alpha_\phi \le \alpha$ By the definition of $\phi_{c,\gamma}$,

$$(\phi_{c,\gamma}(x) - \phi(x))(q(x) - cp(x)) \ge 0, x \in S$$

(check it!). Therefore,

$$\int_S (\phi_{c,\gamma} - \phi)(q - cp)d\mu \ge 0,$$

which implies

$$\int_S \phi_{c,\gamma} q d\mu - \int_S \phi q d\mu \ge c\Big(\int_S \phi_{c,\gamma} p d\mu - \int_S \phi p d\mu\Big),$$

or,

$$\beta_{\phi_{c,\gamma}} - \beta_\phi \ge c(\alpha_{\phi_{c,\gamma}} - \alpha_\phi). \tag{2.1}$$

Since $\alpha_{\phi_{c,\gamma}} = \alpha$, $\alpha_\phi \le \alpha$ and $c \ge 0$, we can conclude that $\beta_{\phi_{c,\gamma}} \ge \beta_\phi$, implying that $\phi_{c,\gamma}$ is a most powerful test of size α,

Claim 3. Suppose now ϕ^* is most powerful of size α. Then $\alpha_{\phi^*} \le \alpha$ and $\beta_{\phi^*} = \beta_{\phi_{c,\gamma}}$. It follows from (2.1) that

$$0 = \beta_{\phi_{c,\gamma}} - \beta_{\phi^*} \ge c(\alpha_{\phi_{c,\gamma}} - \alpha_{\phi^*}) = c(\alpha - \alpha_{\phi^*}). \tag{2.2}$$

Since $c(\alpha - \alpha_{\phi^*}) \ge 0$, we have $c(\alpha - \alpha_{\phi^*}) = 0$ and

$$0 = \beta_{\phi_{c,\gamma}} - \beta_{\phi^*} - c(\alpha_{\phi_{c,\gamma}} - \alpha_{\phi^*}) = \int_S (\phi_{c,\gamma} - \phi^*)(q - cp)d\mu \ge 0.$$

This implies that

$$\int_S (\phi_{c,\gamma} - \phi^*)(q - cp)d\mu = 0$$

and, since the function under the last integral is nonnegative, we have $(\phi_{c,\gamma} - \phi^*)(q - cp) = 0$ μ a.s. Thus, on the set $\{x : L(x) \ne c\} = \{x : q(x) \ne cp(x)\}$, we have $\phi^*(x) = \phi_{c,\gamma}(x)$ μ a.s.

Finally, since $c(\alpha - \alpha_{\phi^*}) = 0$, we either have $\alpha_{\phi^*} = \alpha$, or $c = 0$. In the last case,

$$\begin{aligned}\beta_{\phi^*} = \beta_{\phi_{c,\gamma}} &\ge \mathbb{P}_q\{L(X) > 0\} \ge \mathbb{P}_q\{0 < q(X) < +\infty, p(X) < +\infty\} \\ &\ge \int_{\{q>0\}} q d\mu - \int_{\{p=+\infty\}} q d\mu = \int_S q d\mu = 1.\end{aligned}$$

■

Example 2.1 Let $X \sim N(a, I_d)$ be a normal r.v. in $\mathbb{R}^d$. Consider the following simple hypotheses testing problem:

$$\begin{cases} H_0 : a = 0 \\ H_a : a = u \end{cases}$$

for some $u \in \mathbb{R}^d, u \neq 0$. In this case,

$$p(x) = \frac{1}{(2\pi)^{d/2}} \exp\Big\{-\frac{|x|^2}{2}\Big\} \text{ and } q(x) = \frac{1}{(2\pi)^{d/2}} \exp\Big\{-\frac{|x-u|^2}{2}\Big\}, x \in \mathbb{R}^d.$$

We have

$$\begin{aligned} \log L(X) = \log \frac{q(X)}{p(X)} &= -\frac{|X-u|^2}{2} + \frac{|X|^2}{2} \\ &= \langle u, X\rangle - \frac{|u|^2}{2}. \end{aligned}$$

Therefore, $L(X) > c$ can be rewritten as $\langle u, X\rangle > c'$ for some c', and we can write Neyman-Pearson test in the form

$$\phi_c(X) = \begin{cases} 1 & \text{if } \langle u, X\rangle \geq c \\ 0 & \text{otherwise} \end{cases}$$

If we find a value of c such that $\mathbb{E}_p\phi_c(X) = \alpha$, then ϕ_c is a most powerful test of size α. To this end, it is enough to observe that

$$\mathbb{E}_p\phi_c(X) = \mathbb{P}_p\{\langle u, X\rangle \geq c\}$$

and that, under the null hypotheses X is a standard normal vector in $\mathbb{R}^d$. Therefore, $\langle u, X\rangle \sim N(0; |u|^2)$ and

$$\mathbb{P}_p\{\langle u, X\rangle \geq c\} = \mathbb{P}\Big\{Z \geq \frac{c}{|u|}\Big\}.$$

If z_α denotes $(1-\alpha)$-quantile of Z, it is enough to set $c := z_\alpha|u|$ to satisfy the size α constraint. Thus, the test ϕ_c with this choice of c is most powerful of size α. The rejection region of this test is a half-space $\{x : \langle u, x\rangle \geq c\}$ (on one side of the hyperplane $\{x : \langle u, x\rangle = c\}$). Note that this test is not randomized (which is typical when the underlying distributions are continuous and the likelihood ratio $L(X)$ is a continuous r.v.).

Example 2.2 Consider the problem of testing hypothesis $H_0 : \theta = \theta_1$ against the alternative $H_a : \theta = \theta_2$ based on an observation $X \sim B(n; \theta), \theta \in [0, 1]$. Assume that $0 \leq \theta_0 < \theta_1 \leq 1$. In this case,

$$p(x) = \binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}, \quad q(x) = \binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}, x = 0, 1, \ldots, n$$

and

$$\log L(X) = \log \frac{q(X)}{p(X)} = X \log\Big(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\Big) + n \log\Big(\frac{1-\theta_1}{1-\theta_0}\Big).$$

Note that $\log\Big(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\Big) > 0$ and the condition $L(X) > c$ is equivalent to $X > c'$ for some c'. Thus, Neyman-Pearson tests could be written as

$$\phi_{c,\gamma}(X) := \begin{cases} 1 & \text{if } X > c \\ \gamma & \text{if } X = c \\ 0 & \text{if } X < c. \end{cases}$$

If, for some $k = 0, \ldots, n-1$

$$1 - \alpha = \sum_{j=0}^{k} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j},$$

then the test $\phi_{k,0}$ is most powerful of size α. Otherwise, if

$$\sum_{j=0}^{k-1} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} < 1 - \alpha < \sum_{j=0}^{k} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j},$$

then the test $\phi_{k,\gamma}$, where

$$\gamma := \frac{\sum_{j=0}^{k} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} - (1-\alpha)}{\binom{n}{k} \theta_0^k (1-\theta_0)^{n-k}}.$$

is most powerful of size α.

2.1 Problems

1. Prove Proposition 2.1 and find the best test in the case of mutually singular hypotheses.

2. Let $g : S \mapsto \mathbb{R}$ be a μ-integrable function. Show that if $\int_S \phi g d\mu = 0$ for all $\phi \in \Phi$, then $g = 0$ μ a.s.

3. Let $g : S \mapsto \mathbb{R}$ be a μ-integrable function. Show that the test $\phi^* \in \Phi$ that maximizes $\int_S \phi g d\mu$ over the set Φ if and only if the following conditions hold μ a.s: $\phi^*(x) = 1$ on the set $\{x : g(x) > 0\}$ and $\phi^*(x) = 0$ on the set $\{x : g(x) < 0\}$.

4. Describe *least powerful* tests of size α in simple hypotheses testing problem $H_0 : X \sim p$ against $H_a : X \sim q$.

5. Let α be the significance level and let β be the power of the most powerful test of size α in a simple hypotheses testing problem $H_0 : X \sim p$ against $H_a : X \sim q$. Show that $\beta \geq \alpha$ and, moreover, $\beta > \alpha$ unless $p = q$ a.s.

6. Consider the following testing problem:

$$H_0 : X_1, \ldots, X_n \text{ i.i.d. } \sim N(\mu_1, \Sigma) \text{ in } \mathbb{R}^d$$
$$H_a : X_1, \ldots, X_n \text{ i.i.d. } \sim N(\mu_2, \Sigma) \text{ in } \mathbb{R}^d,$$

where $\mu_1, \mu_2 \in \mathbb{R}^d$ are known means and Σ is a known non-singular covariance matrix. Find the most powerful test of size α for testing H_0 against H_a and determine its power.

7. Let

$$D := \left\{ \left(\int_S \phi dP, \int_S \phi dQ \right) : \phi \in \Phi \right\} \subset [0,1]^2.$$

Show that D is a convex subset of the square $[0,1]^2$, that it contains the diagonal of the square and it is symmetric about the center of the square $(1/2, 1/2)$. Also show that the "upper boundary" of set D (the part of the boundary above the diagonal) is the graph of a concave function representing the power of most powerful tests of size $\alpha, \alpha \in [0,1]$. Similarly, the "lower boundary" of D (the part of the boundary below the diagonal) is the graph of a convex function representing the power of least powerful tests of size $\alpha, \alpha \in [0,1]$. More formally, let

$$\Phi_\alpha := \{\phi \in \Phi : \int_S \phi dP = \alpha\}, \alpha \in [0,1].$$

Then prove that the function

$$\psi^+(\alpha) := \sup_{\phi \in \Phi_\alpha} \int_S \phi dQ, \alpha \in [0,1]$$

is concave and the function

$$\psi^-(\alpha) := \inf_{\phi \in \Phi_\alpha} \int_S \phi dQ, \alpha \in [0,1]$$

is convex.

3 Hellinger Distance: Consistency of Neyman-Pearson Tests and Minimax Lower Bounds

Let p, q be two densities with respect to a σ-finite measure μ on $(S, \mathcal{A})$ and let P, Q be the corresponding probability distributions. Define

$$H^2(p,q) := \int_S (\sqrt{p} - \sqrt{q})^2 d\mu.$$

Definition 3.1 *$H(p,q)$ will be called the Hellinger distance between p and q.*

Note that, for any probability density p w.r.t. μ, $\sqrt{p} \in L_2(\mu)$ and, moreover, $\|\sqrt{p}\|_{L_2(\mu)} = 1$ (soo, $\sqrt{p}$ belongs to the unit sphere of the space $L_2(\mu)$). We also have

$$H(p,q) = \|\sqrt{p} - \sqrt{q}\|_{L_2(\mu)}.$$

Definition 3.2 *Let*

$$A(p,q) := \int_S \sqrt{p}\sqrt{q} d\mu.$$

$A(p,q)$ will be called the Hellinger affinity between p and q.

Clearly,

$$A(p,q) = \langle \sqrt{p}, \sqrt{q} \rangle_{L_2(\mu)},$$

which is the cosine of the angle between $\sqrt{p}$ and $\sqrt{q}$ in the Hilbert space $L_2(\mu)$.

The following properties are obvious:

- $H^2(p,q) = 2(1 - A(p,q))$;
- $0 \le H^2(p,q) \le 2$;
- $0 \le A(p,q) \le 1$;
- $H^2(p,q) = 0$ if and only if $A(p,q) = 1$ if and only if $p = q$ μ a.s.;
- $H^2(p,q) = 2$ if and only if $A(p,q) = 0$ if and only if $pq = 0$ a.s. if and only if the distributions P and Q are mutually singular.

We will use Hellinger distance and Hellinger affinity to bound errors in hypotheses testing. In particular, the following simple fact holds.

Proposition 3.1 *Consider simple hypotheses testing problem*

$$\begin{cases} H_0 : X \sim p \\ H_a : X \sim q \end{cases}$$

for two densities p and q with respect to μ. For likelihood ratio statistic $L(X) := \frac{q(X)}{p(X)}$, define a Neyman-Pearson test

$$\phi_c(X) := \begin{cases} 1 & \text{if } L(X) \ge c \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} \max(\alpha_{\phi_c}, 1 - \beta_{\phi_c}) &\le \max(c^{1/2}, c^{-1/2}) A(p,q). \\ &= \max(c^{1/2}, c^{-1/2}) \Big(1 - \frac{H^2(p,q)}{2}\Big). \end{aligned}$$

Proof. Note that

$$\begin{aligned}\alpha_{\phi_c} &= \mathbb{P}_p\{L(X) \geq c\} \leq \mathbb{E}_p \frac{\sqrt{L(X)}}{\sqrt{c}} \\ &= c^{-1/2} \int_S \sqrt{\frac{q(x)}{p(x)}} p(x)\mu(dx) = c^{-1/2} \int_S \sqrt{p(x)}\sqrt{q(x)}\mu(dx) \\ &= c^{-1/2} A(p,q).\end{aligned}$$

Similarly, one can prove that

$$1 - \beta_\phi \leq c^{1/2} A(p,q),$$

implying the claim. ∎

On the other hand, we will prove the following minimax lower bound.

Proposition 3.2 *Consider simple hypotheses testing problem*

$$\begin{cases} H_0 : X \sim p \\ H_a : X \sim q \end{cases}$$

for two densities p and q with respect to μ. Then

$$\begin{aligned}\inf_{\phi \in \Phi} \max(\alpha_\phi, 1 - \beta_\phi) &\geq \frac{1}{2} - \frac{1}{4} \int_S |p - q| d\mu \\ &\geq \frac{1}{2}\left(1 - H(p,q)\sqrt{1 - \frac{H^2(p,q)}{4}}\right).\end{aligned}$$

where Φ is the set of all tests based on an observation X.

Proof. We write

$$\begin{aligned}\max(\alpha_\phi, 1 - \beta_\phi) &\geq \frac{\alpha_\phi + 1 - \beta_\phi}{2} = \frac{1}{2} + \frac{1}{2}(\alpha_\phi - \beta_\phi) \\ &= \frac{1}{2} + \frac{1}{2} \int_S \phi(p - q) d\mu.\end{aligned}$$

Clearly, the right hand side of the last bound is minimized for the test

$$\phi^*(X) := \begin{cases} 1 & \text{if } q(X) \geq p(X) \\ 0 & \text{otherwise.} \end{cases}$$

For this test ϕ^*,

$$\max(\alpha_\phi, 1 - \beta_\phi) \geq \frac{\alpha_{\phi^*} + 1 - \beta_{\phi^*}}{2} = \frac{1}{2} + \frac{1}{2} \int_S \phi^*(p - q) d\mu = \frac{1}{2} - \frac{1}{2} \int_{\{q \geq p\}} (q - p) d\mu.$$

By symmetry of p and q, we also have

$$\max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2} - \frac{1}{2}\int_{\{p\geq q\}} (p-q)d\mu, \phi \in \Phi.$$

Adding these two inequalities, we get

$$\max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2} - \frac{1}{4}\int_S |p-q|d\mu.$$

To complete the proof, it is enough to use the following lemma.

Lemma 3.1 *For all densities p, q w.r.t. μ,*

$$\frac{1}{2}\int_S |p-q|d\mu \leq H(p,q)\sqrt{1-\frac{H^2(p,q)}{4}}.$$

Proof. Note that, by Cauchy-Schwarz inequality,

$$\begin{aligned} A(p,q)^2 &= \Big(\int_S \sqrt{p}\sqrt{q}d\mu\Big)^2 = \Big(\int_S \sqrt{\min(p,q)}\sqrt{\max(p,q)}d\mu\Big)^2 \\ &\leq \int_S \min(p,q)d\mu \int_S \max(p,q)d\mu \end{aligned}$$

Since

$$\min(p,q) = \frac{p+q-|p-q|}{2},\ \max(p,q) = \frac{p+q+|p-q|}{2},$$

we get

$$\begin{aligned} A(p,q)^2 &\leq \int_S \frac{p+q-|p-q|}{2}d\mu \int_S \frac{p+q+|p-q|}{2}d\mu \\ &= \Big(1-\frac{1}{2}\int_S |p-q|d\mu\Big)\Big(1+\frac{1}{2}\int_S |p-q|d\mu\Big) \\ &= 1-\frac{1}{4}\Big(\int_S |p-q|d\mu\Big)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} &\frac{1}{2}\int_S |p-q|d\mu \leq \sqrt{1-A(p,q)^2} = \sqrt{1-A(p,q)}\sqrt{1+A(p,q)} \\ &\sqrt{\frac{H^2(p,q)}{2}}\sqrt{2-\frac{H^2(p,q)}{2}} = H(p,q)\sqrt{1-\frac{H^2(p,q)}{4}}. \end{aligned}$$

■

■

For $n \geq 1$, let $\mu^{(n)} := \mu \times \cdots \times \mu$ be the product measure on the space $S^n := S \times \cdots \times S$ of n copies of measure μ. Similarly, let $P^{(n)} := P \times \cdots \times P$ and $Q^{(n)} := Q \times \cdots \times Q$. Clearly, $P^{(n)}$ and $Q^{(n)}$ are absolutely continuous with respect to $\mu^{(n)}$ with densities

$$p^{(n)}(x_1, \ldots, x_n) = p(x_1) \ldots p(x_n) \text{ and } q^{(n)}(x_1, \ldots, x_n) = q(x_1) \ldots q(x_n), \ (x_1, \ldots, x_n) \in S^n.$$

The following proposition is straightforward.

Proposition 3.3 *For all densities p, q with respect to μ and all $n \geq 1$,*

$$A(p^{(n)}, q^{(n)}) = A(p, q)^n.$$

Proof. Indeed,

$$\begin{aligned} A(p^{(n)}, q^{(n)}) &= \int_{S^n} \sqrt{p^{(n)}} \sqrt{q^{(n)}} d\mu^{(n)} \\ &= \int_S \cdots \int_S \sqrt{p(x_1) \ldots p(x_n)} \sqrt{q(x_1) \ldots q(x_n)} \mu(dx_1) \ldots \mu(dx_n) \\ &= \int_S \sqrt{p(x_1) q(x_1)} \mu(dx_1) \cdots \int_S \sqrt{p(x_n) q(x_n)} \mu(dx_n) = A(p, q)^n. \end{aligned}$$

■

An important consequence of this simple fact is that, if p and q are two different densities, then $A(p, q) < 1$ and $A(p^{(n)}, q^{(n)}) \to 0$ as $n \to \infty$, implying that $H^2(p^{(n)}, q^{(n)}) \to 2$ as $n \to \infty$. This means that distributions $P^{(n)}$ and $Q^{(n)}$ are "asymptotically singular" when $n \to \infty$.

Consider now the following simple hypotheses testing problem:

$$\begin{cases} H_0^{(n)} : X_1, \ldots, X_n \text{ i.i.d. } \sim p \\ H_a^{(n)} : X_1, \ldots, X_n \text{ i.i.d. } \sim q \end{cases} \iff \begin{cases} H_0^{(n)} : X^{(n)} = (X_1, \ldots, X_n) \sim P^{(n)} \\ H_a^{(n)} : X^{(n)} = (X_1, \ldots, X_n) \sim Q^{(n)} \end{cases} \tag{3.1}$$

Recall that, for two singular distributions P and Q in a simple hypotheses testing problem, there exists a test ϕ for which $\alpha_\phi = 0$ and $\beta_\phi = 1$ (thus, ϕ makes an error with zero probability). It is natural to expect that the same property holds asymptotically for two distributions $P^{(n)}$ and $Q^{(n)}$ that are "asymptotically singular".

Definition 3.3 *A sequence of tests $\phi^{(n)}$ for the null hypothesis $H_0^{(n)}$ against the alternative $H_a^{(n)}$ is called consistent iff, for all densities $p \neq q$,*

$$\alpha_{\phi^{(n)}} \to 0 \text{ and } \beta_{\phi^{(n)}} \to 1 \text{ as } n \to \infty.$$

We will show that a sequence of Neyman-Pearson tests $\phi_{c,1}^{(n)}$ with an arbitrary fixed threshold $c \in (0, +\infty)$ is consistent for simple hypotheses testing problem (3.1). It easily follows from the next proposition.

Proposition 3.4 *Let*

$$L_n(X_1, \dots, X_n) := \frac{q(X_1) \dots q(X_n)}{p(X_1) \dots p(X_n)}$$

be the likelihood ratio statistics for testing problem (3.1). *For* $c \in (0, +\infty), n \geq 1$, *denote*

$$\phi_c^{(n)}(X_1, \dots, X_n) := \begin{cases} 1 & \text{if } L_n(X_1, \dots, X_n) \geq c \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\max(\alpha_{\phi_c^{(n)}}, 1 - \beta_{\phi_c^{(n)}}) \leq \max(c^{1/2}, c^{-1/2}) A(p, q)^n. \tag{3.2}$$

As a consequence, the sequence of Neyman-Pearson tests $\{\phi_c^{(n)}\}$ *is consistent for* $H_0^{(n)}$ *against* $H_a^{(n)}$.

Proof. Bound (3.2) immediately follows from Lemma 3.1 and Proposition 3.3. Since, for $p \neq q$, we have $A(p, q) < 1$, consistency of the sequence of tests $\{\phi_c^{(n)}\}$ follows from bound (3.2). ■

It is of interest to look at a similar testing problem in the case when the densities p and q are allowed to be close to each other when the sample size n is large. Namely, consider the problem

$$\begin{cases} H_0^{(n)} : X_1, \dots, X_n \text{ i.i.d. } \sim p_n \\ H_a^{(n)} : X_1, \dots, X_n \text{ i.i.d. } \sim q_n \end{cases}$$

for two sequences of densities $\{p_n\}$ and $\{q_n\}$. Let

$$\delta_n := H(p_n, q_n) \to 0 \text{ as } n \to \infty.$$

Thus, the null hypotheses and the alternative are getting close when $n \to \infty$. The question is how fast δ_n could go to zero so that still there exists a consistent sequence of tests ϕ_n for $H_0^{(n)}$ against $H_a^{(n)}$. The following theorem provides an answer.

Theorem 3.1 *If*

$$\sqrt{n}\delta_n \to \infty \text{ as } n \to \infty, \tag{3.3}$$

then there exists a sequence of tests $\{\phi_n\}$ *such that*

$$\max(\alpha_{\phi_n}, 1 - \beta_{\phi_n}) \to 0 \text{ as } n \to \infty.$$

On the other hand, if

$$\limsup_{n\to\infty} \sqrt{n}\delta_n < \infty, \tag{3.4}$$

then, for some constant $b > 0$,

$$\liminf_{n\to\infty} \inf_{\phi\in\Phi_n} \max(\alpha_\phi, 1-\beta_\phi) \geq b,$$

where Φ_n *is the set of all tests based on the observations* $X_1, \ldots, X_n$. *Moreover, if*

$$\sqrt{n}\delta_n \to 0 \text{ as } n \to \infty, \tag{3.5}$$

then

$$\liminf_{n\to\infty} \inf_{\phi\in\Phi_n} \max(\alpha_\phi, 1-\beta_\phi) = \frac{1}{2}.$$

Proof. To prove the first claim, consider a Neyman-Pearson test

$$\phi_n(X_1, \ldots, X_n) := \begin{cases} 1 & \text{if } \frac{q_n(X_1)\ldots q_n(X_n)}{p_n(X_1)\ldots p_n(X_n)} \geq c \\ 0 & \text{otherwise} \end{cases}$$

and use bound (3.2) to get

$$\max(\alpha_{\phi_n}, 1-\beta_{\phi_n}) \leq \max(c^{1/2}, c^{-1/2}) A(p_n, q_n)^n.$$

Note that

$$A(p_n, q_n) = 1 - \frac{H^2(p_n, q_n)}{2} = 1 - \frac{\delta_n^2}{2}.$$

Therefore,

$$\max(\alpha_{\phi_n}, 1-\beta_{\phi_n}) \leq \max(c^{1/2}, c^{-1/2}) \Big(1 - \frac{\delta_n^2}{2}\Big)^n \to 0$$

as $n \to \infty$ under assumption (3.3).

To prove the remaining claims, we use the minimax lower bound of Proposition 3.2. It yields

$$\inf_{\phi\in\Phi_n} \max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2}\Big(1 - H(p^{(n)}, q^{(n)})\sqrt{1 - \frac{H^2(p^{(n)}, q^{(n)})}{4}}\Big).$$

Using Proposition 3.3, we have

$$\begin{aligned} H^2(p^{(n)}, q^{(n)}) &= 2(1 - A(p^{(n)}, q^{(n)})) = 2(1 - A(p_n, q_n)^n) \\ &= 2\Big(1 - \Big(1 - \frac{H^2(p_n, q_n)}{2}\Big)^n\Big) = 2\Big(1 - \Big(1 - \frac{\delta_n^2}{2}\Big)^n\Big). \end{aligned}$$

Therefore, under condition (3.4),

$$B := \limsup_{n\to\infty} H^2(p^{(n)}, q^{(n)}) < 2,$$

implying

$$\liminf_{n\to\infty} \inf_{\phi\in\Phi_n} \max(\alpha_\phi, 1-\beta_\phi) \geq b := \frac{1}{2}\Big(1 - B\sqrt{1-\frac{B^2}{4}}\Big) > 0.$$

Under condition (3.5),

$$H^2(p^{(n)}, q^{(n)}) \to 0 \text{ as } n \to \infty,$$

implying

$$\liminf_{n\to\infty} \inf_{\phi\in\Phi_n} \max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2},$$

which completes the proof. ■

3.1 Problems

1. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$ random variables. Consider the problem of testing $H_0^{(n)} : \theta = 0$ against $H_a^{(n)} : \theta = \varepsilon_n$. Determine necessary and sufficient conditions on the "separation rate" ε_n such that consistent testing is possible.

2. The same question for the model $X_1, \ldots, X_n$ i.i.d. $U[0, \theta], \theta > 0$ and for the testing problem $H_0^{(n)} : \theta = 1$ against $H_a^{(n)} : \theta = 1 + \varepsilon_n$.

4 Bayes optimal tests

Consider hypotheses testing problem

$$\begin{cases} H_0 : X \sim p_0 \\ H_1 : X \sim p_1, \end{cases}$$

where p_0, p_1 are densities on $(S, \mathcal{A})$ with respect to μ. Let $\Theta = \{0, 1\}$ be a parameter space with $\theta = 0$ corresponding to hypothesis H_0 and $\theta = 1$ corresponding to hypothesis H_1. We will assign prior probability $\pi_0 \in [0, 1]$ to $\theta = 0$ and $\pi_1 = 1 - \pi_0$ to $\theta = 1$ (in other words, π_0 is the probability that H_0 is true and π_1 is the probability that H_1 is true). This defines a prior distribution π on the parameter space Θ. We will also assign "costs" $W_0, W_1 \geq 0$ to the two errors in our testing problem: W_0 will be the cost of

rejecting H_0 when it is true and W_1 will be the cost of rejecting H_1 when it is true. As before, for a test ϕ, denote

$$\alpha_\phi := \mathbb{E}_0\phi(X) = \int_S \phi p_0 d\mu, \quad \beta_\phi := \mathbb{E}_1\phi(X) = \int_S \phi p_1 d\mu.$$

With these notations, we can write the average risk of test ϕ with respect to the prior π (or the expected "cost" of making an error) as

$$R_\pi(\phi) = W_0\pi_0\alpha_\phi + W_1\pi_1(1-\beta_\phi).$$

Definition 4.1 *A test ϕ_π that minimizes the average risk $R_\pi(\phi)$ over the set Φ of all randomized tests is called the Bayes optimal test with respect to the prior π.*

In other words, Bayes optimal tests are the solutions of the following optimization problem:

$$\phi_\pi \in \mathrm{Argmin}_{\phi\in\Phi} R_\pi(\phi).$$

Note that the average risk $R_\pi(\phi)$ can be rewritten as follows:

$$R_\pi(\phi) = \int_S \phi(W_0\pi_0 p_0 - W_1\pi_1 p_1)d\mu + W_1\pi_1.$$

To find Bayes optimal tests, it is enough to minimize over the set Φ the integral

$$\int_S \phi(W_0\pi_0 p_0 - W_1\pi_1 p_1)d\mu.$$

Clearly, a solution of this problem is the following deterministic test:

$$\phi_\pi(x) := \begin{cases} 1 & \text{if } W_1\pi_1 p_1(x) \geq W_0\pi_0 p_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

Note also that any other solution must coincide with ϕ_π on the set

$$\{x : W_1\pi_1 p_1(x) \neq W_0\pi_0 p_0(x)\}$$

μ a.s. and it could take arbitrary values in $[0,1]$ on the set

$$\{x : W_1\pi_1 p_1(x) = W_0\pi_0 p_0(x)\}.$$

Essentially, it means that to minimize the average risk R_π it is enough to deal with deterministic tests with rejection region

$$L(X) \geq \frac{W_0\pi_0}{W_1\pi_1},$$

where $L(X) := \frac{p_1(X)}{p_0(X)}$ is the likelihood ratio.

This Bayesian approach to testing could be easily extended to more general *composite hypotheses testing problem*. Namely, assume that $X \sim P_\theta, \theta \in \Theta$, where $\{P_\theta : \theta \in \Theta\}$ is a family of probability distributions on $(S, \mathcal{A})$ parametrized by unknown parameter $\theta \in \Theta$. We will suppose that $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ with probability densities $p_\theta = \frac{dP_\theta}{d\mu}$. Let Θ_0, Θ_1 be a partition of parameter space Θ into two disjoint sets: $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0, \Theta_1 \neq \emptyset$. We will be interested in the following hypotheses testing problem:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases} \quad X \sim P_\theta, \theta \in \Theta.$$

Assume now that Θ is equipped with a σ-algebra $\mathcal{B}$ and with a probability measure Π on $(\Theta, \mathcal{B})$. We will call Π a *prior* distribution on Θ. In most of the examples, prior Π will have density π with respect to some given σ-finite measure ν on $(\Theta, \mathcal{B})$: $\Pi(d\theta) = \pi(\theta)\nu)(d\theta)$. Introducing prior Π means that parameter θ is now viewed as a random variable sampled from the distribution Π. Therefore, it becomes natural to view density p_θ as a conditional density of X given θ and to use the following notation:

$$p(x|\theta) := p_\theta(x), x \in S, \theta \in \Theta.$$

To allow the integration with respect to two variables (x, θ), we assume that the function $(x, \theta) \mapsto p(x|\theta)$ is measurable. We can also define *posterior distribution* of θ given $X = x$ (conditional distribution of θ given $X = x$) using Bayes formula:

$$\Pi(d\theta|x) = \frac{p(x|\theta)\Pi(d\theta)}{p(x)},$$

where

$$p(x) := \int_\Theta p(x|\theta)\Pi(d\theta)$$

is the mixture of densities $p(\cdot|\theta), \theta \in \Theta$ with respect to the prior Π (in other words, the unconditional density of X). Note that, by Fubini theorem,

$$\int_S p(x)\mu(dx) = \int_S \int_\Theta p(x|\theta)\Pi(d\theta)\mu(dx) = \int_\Theta \int_S p(x|\theta)\mu(dx)\Pi(d\theta) = 1,$$

so, $p(x) < +\infty$ μ a.s. and $\Pi(\cdot|x)$ is well defined.

In the case when the prior Π has density π with respect to ν, $\Pi(d\theta) = \pi(\theta)\nu(d\theta)$, the corresponding posterior density

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}$$

is also well defined.

As in the case of simple hypotheses testing, we introduce the costs W_0 and W_1 of making an error (rejecting H_0 or H_1 when it is true). We also define *the power function* of a test $\phi \in \Phi$.

Definition 4.2 *Let $\phi \in \Phi$ be a randomized test. The function*

$$\beta_\phi(\theta) := \mathbb{E}_\theta \phi(X), \theta \in \Theta$$

will be called the power function of ϕ.

Note that, for $\theta \in \Theta_0$ (when hypotheses H_0 is true), $\beta_\phi(\theta)$ is the probability that test ϕ makes an error; on the other hand, for $\theta \in \Theta_1$ (when hypothesis H_1 is true), the probability that ϕ makes an error is $1 - \beta_\phi(\theta)$. This allows us to write the risk function of test ϕ (its expected loss) as

$$R(\theta, \phi) = W_0 \beta_\phi(\theta) I_{\Theta_0}(\theta) + W_1(1 - \beta_\phi(\theta)) I_{\Theta_1}(\theta), \theta \in \Theta.$$

The average risk of ϕ with respect to the prior Π is then

$$\begin{aligned} R_\Pi(\phi) &:= \int_\Theta R(\theta, \phi) \Pi(d\theta) \\ &= \int_\Theta [W_0 \beta_\phi(\theta) I_{\Theta_0}(\theta) + W_1(1 - \beta_\phi(\theta)) I_{\Theta_1}(\theta)] \Pi(d\theta). \end{aligned}$$

Definition 4.3 *A test ϕ_Π that minimizes the average risk $R_\Pi(\phi)$ with respect to Π over the set Φ of all randomized tests is called the Bayes optimal test with respect to the prior Π.*

Theorem 4.1 *Any test of the following form*

$$\phi_{\Pi,\gamma}(x) := \begin{cases} 1 & \text{if } W_1 \Pi(\Theta_1|x) > W_0 \Pi(\Theta_0|x) \\ \gamma(x) & \text{if } W_1 \Pi(\Theta_1|x) = W_0 \Pi(\Theta_0|x) \\ 0 & \text{if } W_1 \Pi(\Theta_1|x) < W_0 \Pi(\Theta_0|x), \end{cases}$$

where $\gamma : S \mapsto [0, 1]$ is a measurable function, is Bayes optimal with respect to Π. If ϕ is Bayes optimal with respect to Π, it coincides μ a.s. with $\phi_{\Pi,\gamma}$ for some γ.

Proof. To find Bayes optimal tests ϕ_Π, rewrite the average risk $R_\Pi(\phi)$ as follows:

$$R_\Pi(\phi) = \int_\Theta \beta_\phi(\theta) [W_0 I_{\Theta_0} - W_1 I_{\Theta_1}(\theta)] \Pi(d\theta) + W_1 \Pi(\Theta_1).$$

Note also that, using the definition of posterior distribution and Fubini theorem, we have

$$\begin{aligned}
&\int_\Theta \beta_\phi(\theta)[W_0 I_{\Theta_0} - W_1 I_{\Theta_1}(\theta)]\Pi(d\theta) \\
&= \int_\Theta \int_S \phi(x) p(x|\theta)\mu(dx)[W_0 I_{\Theta_0}(\theta) - W_1 I_{\Theta_1}(\theta)]\Pi(d\theta) \\
&= \int_\Theta \int_S \phi(x)[W_0 I_{\Theta_0}(\theta) - W_1 I_{\Theta_1}(\theta)]p(x|\theta)\mu(dx)\Pi(d\theta) \\
&= \int_S \int_\Theta \phi(x)[[W_0 I_{\Theta_0}(\theta) - W_1 I_{\Theta_1}(\theta)]\Pi(d\theta|x)p(x)\mu(dx) \\
&= \int_S \phi(x)[W_0\Pi(\Theta_0|x) - W_1\Pi(\Theta_1|x)]p(x)\mu(dx).
\end{aligned}$$

It is clear that to minimize $R_\Pi(\phi)$ over $\phi \in \Phi$, one has to minimize the integral

$$\int_S \phi(x)[W_0\Pi(\Theta_0|x) - W_1\Pi(\Theta_1|x)]p(x)\mu(dx),$$

and tests $\phi_{\Pi,\gamma}$ are solutions of this problem for any measurable function $\gamma : S \mapsto [0,1]$. Moreover, any other minimizer $\phi \in \Phi$ of $R_\Pi(\phi)$ should satisfy the conditions $\phi(x) = 1$ μ a.s. on the set

$$\{x : W_1\Pi(\Theta_1|x) > W_0\Pi(\Theta_0|x)\}$$

and $\phi(x) = 0$ μ a.s. on the set

$$\{x : W_1\Pi(\Theta_1|x) < W_0\Pi(\Theta_0|x)\}.$$

Since $\phi : S \mapsto [0,1]$ is a measurable function and ϕ could take arbitrary values in $[0,1]$ on the set

$$\{x : W_1\Pi(\Theta_1|x) = W_0\Pi(\Theta_0|x)\},$$

it coincides with $\phi_{\Pi,\gamma}$ μ a.s. for some measurable function $\gamma : S \mapsto [0,1]$. ■

Remark. Note that $\Pi(\Theta_0|X)$, $\Pi(\Theta_1|X) = 1 - \Pi(\Theta_0|X)$ are posterior probabilities of sets Θ_0, Θ_1. Let

$$T(X) := \frac{\Pi(\Theta_1|X)}{\Pi(\Theta_0|X)}.$$

and let $c := \frac{W_0}{W_1}$. Then deterministic test that rejects H_0 when $T(X) \geq c$ and accepts H_0 otherwise is Bayes optimal with respect to Π.

Proposition 4.1 *Let $\pi_0 := \Pi(\Theta_0), \pi_1 := \Pi(\Theta_1)$ and assume that $\pi_0, \pi_1 > 0$. Denote*

$$\Pi_0(A) := \frac{\Pi(A \cap \Theta_0)}{\pi_0},\ \Pi_1(A) := \frac{\Pi(A \cap \Theta_1)}{\pi_1}, A \in \mathcal{B}.$$

Then $\Pi = \pi_0\Pi_0 + \pi_1\Pi_1$. *Let*

$$p_0(x) = \int_{\Theta_0} p_\theta(x)\Pi_0(d\theta),\ p_1(x) = \int_{\Theta_1} p_\theta(x)\Pi_1(d\theta)$$

be the mixtures of densities p_θ *w.r.t. distributions* Π_0, Π_1. *Then Bayes optimal tests for simple hypotheses testing problem*

$$\begin{cases} H_0 : X \sim p_0 \\ H_1 : X \sim p_1 \end{cases}$$

with respect to prior $\pi = \{\pi_0, \pi_1\}$ *coincides with Bayes optimal tests for composite hypotheses testing problem*

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases} \quad X \sim P_\theta, \theta \in \Theta.$$

with respect to prior Π.

4.1 Problems

1. Prove Proposition 4.1.

2. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1), \theta \in \mathbb{R}$ r.v. Find the Bayes optimal test for testing $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with respect to the prior $N(\mu, \tau^2)$ on the parameter space $\Theta = \mathbb{R}$.

5 Sequential Testing: Bayes Optimality

We will consider the following simple hypotheses testing problem

$$\begin{cases} H_0 : X_1, \ldots, X_n, \ldots \text{ i.i.d. } \sim p \\ H_1 : X_1, \ldots, X_n, \ldots \text{ i.i.d. } \sim q, \end{cases}$$

where p, q are densities with respect to a σ-finite measure μ on $(S, \mathcal{A})$ and $p \neq q$. The testing will be performed sequentially. At moment of time $t = 0$, a sequential test should decide whether to accept H_0, to accept H_1, or to request an observation X_1. In the last case, at moment of time $t = 1$, the test again decides (based on observation X_1) whether to stop and accept H_0 or H_1, or to continue and request an additional observation X_2, an so on. Of course, if the observations have no cost, the sequential test could continue for an arbitrary long time N. If, at the end, it makes a decision using a Neyman-Pearson test, it could achieve arbitrary small probabilities of the errors (in view of consistency of Neyman-Pearson tests).

5.1 Stopping times and sequential tests

In what follows, we assume that an observation has some cost $a > 0$ and there are also costs $W_0, W_1 > 0$ of making the errors of two types (rejecting H_0 when it is true and rejecting H_1 when it is true). Of course, the moment of time τ when a sequential test stops should possess the following property: for all $n \geq 1$, the decision that $\tau = n$ should be bases only on the observations $X_1, \dots, X_n$ available by moment of time n. This leads to the following definition.

Definition 5.1 *Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and let*

$$\mathcal{F}_n := \sigma(X_1, \dots, X_n), n \geq 1$$

be the σ-algebra generated by r.v. $X_1, \dots, X_n$. A nondecreasing sequence $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ of σ-algebras is called a filtration.

Note that, for all $n \geq 1$, $E \in \mathcal{F}_n$ if and only if $E = \{(X_1, \dots, X_n) \in B\}$ for some set $B \subset S^n = S \times \dots \times S, B \in \mathcal{A}^n = \mathcal{A} \times \dots \times \mathcal{A}$ (thus, $E \in \mathcal{F}_n$ means that it is possible to decide whether event E occurs or not based on observations $X_1, \dots, X_n$).

Definition 5.2 *A random variable τ with values in $\{0, 1, 2, \dots\}$ is called a stopping time with respect to the filtration $\{\mathcal{F}_n\}$ iff $\{\tau = n\} \in \mathcal{F}_n$ for all $n \geq 0$.*

For a stopping time τ and for $n \geq 1$, the event $\{\tau = n\}$ can be represented as

$$\{\tau = n\} = \{(X_1, \dots, X_n) \in A_n\},$$

where $A_n \in \mathcal{A}^n$.

Example 5.1 *Let $Y_1, \dots, Y_n, \dots$ be a sequence of random variables and $\mathcal{F}_n := \sigma(Y_1, \dots, Y_n)$, $n \geq 1$. Then*

$$\tau := \inf\{n : Y_n \geq 1\}$$

is a stopping time with respect to $\{\mathcal{F}_n\}$, but

$$\tilde{\tau} := \sup\{n : Y_n \geq 1\}$$

is not.

Definition 5.3 *A sequential test is a couple $\delta = \delta(X_1, X_2, \dots) = (\tau, \{\phi_n : n \geq 0\})$, where $\tau = \tau(X_1, X_2, \dots)$ is a stopping time and $\{\phi_n : n \geq 0\}$ is a sequence of tests $\phi_n : A_n \mapsto [0, 1]$ (where sets $A_n \in \mathcal{A}^n$ are such that $\{\tau = n\} = \{(X_1, \dots, X_n) \in A_n\}$).*

The test $\delta = (\tau, \{\phi_n : n \geq 0\})$ is used as follows: if $\tau = n$ (which is decided based on whether $(X_1, \ldots, X_n) \in A_n$), the test stops at moment of time n and applies ϕ_n to make a decision. Given $(X_1, \ldots, X_n) \in A_n$, with probability $\phi_n(X_1, \ldots, X_n)$, the test accepts H_1 (rejects H_0); with probability $1 - \phi_n(X_1, \ldots, X_n)$ it accepts H_0 (rejects H_1). Of course, for $n = 0$, ϕ_0 does not use the data to make a decision (H_1 is accepted with probability ϕ_0 and H_0 is accepted with probability $1 - \phi_0$).

In other words, sequential test $\delta = (\tau, \{\phi_n : n \geq 0\})$ outputs a random sequence of decisions $\nu := \{\nu_j : j \geq 0\}$, where $\nu_j \in \{c, 0, 1, *\}$. If $\tau = n$, then $\nu_j = c, j < n$ (meaning that, at moment of time j, δ decides to continue and request an additional observation), at moment of time n, $\nu_n \in \{0, 1\}$ (meaning that δ decides to stop and accept either H_0 or H_1) and, at moment of time $j > n$, $\nu_j = *$ (meaning that the test has already stopped). Using the terminology of statistical decision theory, we call ν an action, and, given the data $X_1, X_2, \ldots$, $\delta(X_1, X_2, \ldots)$ generates a probability distribution $D_{\{X_1, X_2, \ldots\}}$ on actions, so, it is a decision rule.

5.2 Bayes optimal sequential tests

Let $\Theta := \{0, 1\}$ and let ϑ denote a random variable in space Θ. Let

$$\pi := \mathbb{P}\{\vartheta = 1\}, \quad 1 - \pi = \mathbb{P}\{\vartheta = 0\}$$

be a prior distribution on Θ. We will assume that the value θ of r.v. ϑ determines whether H_0 or H_1 is true.

If sequential test δ stops at moment of time τ and chooses an action ν, it suffers a loss

$$L(\theta, \nu) := a\tau + W_\theta \nu_\tau^{1-\theta}(1 - \nu_\tau)^\theta$$

(we set $0^0 = 1$). The expected loss of δ with respect to a random choice of action for given sequence of observations $X_1, X_2, \ldots$ is equal to

$$L(\theta, \delta) = \mathbb{E}_{\nu \sim D_{\{X_1, X_2, \ldots\}}} L(\theta, \nu) = a\tau + W_\theta \phi_\tau(X_1, \ldots, X_\tau)^{1-\theta}(1 - \phi_\tau(X_1, \ldots, X_\tau))^\theta.$$

The risk function of the test δ is

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta L(\theta, \delta) = \mathbb{E}(L(\theta, \delta) | \vartheta = \theta) \\ &= \mathbb{E}_\theta \Big[a\tau + W_\theta \phi_\tau(X_1, \ldots, X_\tau)^{1-\theta}(1 - \phi_\tau(X_1, \ldots, X_\tau))^\theta \Big], \theta \in \Theta, \end{aligned}$$

and the average risk with respect to the prior is

$$R_\pi(\delta) = \mathbb{E}L(\vartheta, \delta) = \mathbb{E}\mathbb{E}(L(\vartheta, \delta) | \vartheta) = (1 - \pi)\mathbb{E}_0 L(0, \delta) + \pi \mathbb{E}_1 L(1, \delta).$$

Definition 5.4 *A Bayes optimal sequential test for prior π is a test δ_π that minimizes the average risk $R_\pi(\delta)$ over all sequential tests δ.*

To find Bayes optimal tests δ_π, consider first only sequential tests δ for which $\tau = \tau_\delta = 0$. Such tests do not use any data and to make a decision they should use an arbitrary test $\phi_\alpha = \alpha$ that rejects H_0 with probability α and accepts it with probability $1-\alpha$. Let us denote such sequential test by δ_α. Obviously,

$$R_\pi(\delta_\alpha) = (1-\pi)\alpha W_0 + \pi(1-\alpha)W_1.$$

The minimum of $R_\pi(\delta_\alpha)$ with respect to α is attained at $\alpha = 1$ if $(1-\pi)W_0 \le \pi W_1$ and at $\alpha = 0$ otherwise. Thus,

$$\psi_0(\pi) := \min_{\alpha\in[0,1]} R_\pi(\delta_\alpha) = \min_{\delta:\tau_\delta=0} R_\pi(\delta) = \min(R_\pi(\delta_0), R_\pi(\delta_1)) = \min(\pi W_1, (1-\pi)W_0).$$

Clearly, $\psi_0(\pi) = 0$ for $\pi = 0$ and $\pi = 1$, and the maximum of $\psi(\pi)$ is attained at $\pi = \frac{W_0}{W_0+W_1}$ and it is equal to $\frac{W_0W_1}{W_0+W_1}$. Moreover, the function ψ_0 is linear between 0 and $\frac{W_0}{W_0+W_1}$ and between $\frac{W_0}{W_0+W_1}$ and 1. The function ψ_0 describes the optimal average risk attainable with no data. Note that, for $\pi \le \frac{W_0}{W_0+W_1}$, the test δ_1 optimizes the average risk and, for $\pi > \frac{W_0}{W_0+W_1}$, the test δ_0 does.

Consider now the sequential tests δ for which $\tau_\delta \ge 1$ (that utilize at least one observation). Let

$$\begin{aligned}\psi_{\ge1}(\pi) &:= \inf_{\delta:\tau_\delta\ge1} R_\pi(\delta)\\ &= \inf_{\delta:\tau_\delta\ge1}[(1-\pi)\mathbb{E}_0 L(0,\delta) + \pi\mathbb{E}_1 L(1,\delta)].\end{aligned}$$

Note that $\psi_{\ge1}(\pi) \ge a, \pi \in [0,1]$ (since at least one observation with cost a is being used) and $\psi_{\ge1}$ is a concave function on $[0,1]$ (as an infimum of linear functions). Using concavity of $\psi_{\ge1}$ it is now easy to check that the graph of function $\psi_{\ge1}$ intersects the graph of function ψ_0 at exactly two points if

$$\frac{W_0W_1}{W_0+W_1} > \psi_{\ge1}\Big(\frac{W_0}{W_0+W_1}\Big),$$

at one point if

$$\frac{W_0W_1}{W_0+W_1} = \psi_{\ge1}\Big(\frac{W_0}{W_0+W_1}\Big)$$

and the graphs do not intersect if

$$\frac{W_0W_1}{W_0+W_1} < \psi_{\ge1}\Big(\frac{W_0}{W_0+W_1}\Big)$$

Let $0 < \gamma_1 \le \gamma_2 < 1$ be such that

$$\psi_0(\gamma_1) = \psi_{\ge1}(\gamma_1),\ \psi_0(\gamma_2) = \psi_{\ge1}(\gamma_2),$$

provided that the graphs do intersect. It will be also convenient to set $\gamma_1 = \gamma_2 = \frac{W_0}{W_0+W_1}$ if they do not. Clearly, $\psi_{\geq 1}(\pi) < \psi_0(\pi)$ for $\pi \in (\gamma_1, \gamma_2)$ and $\psi_{\geq 1}(\pi) \geq \psi_0(\pi)$ otherwise.

Note also that

$$\psi(\pi) := \inf_{\delta} R_\pi(\delta) = \min(\psi_0(\pi), \psi_{\geq 1}(\pi))$$

is the minimal average risk with respect to the prior π attainable in our testing problem.

An important consequence of the above considerations is the following proposition.

Proposition 5.1 *A Bayes optimal sequential test δ_π (if it exists) must at moment of time $t = 0$:*

- *stop and accept H_0, if $\pi \leq \gamma_1$;*
- *stop and accept H_1, if $\pi \geq \gamma_2$;*
- *continue and request an observation X_1, if $\pi \in (\gamma_1, \gamma_2)$.*

Any test δ that does not follow this strategy at moment $t = 0$ is not Bayes optimal, so, there exists a sequential test δ' such that $R_\pi(\delta') < R_\pi(\delta)$.

It will be said in what follows that two sequential tests $\delta = (\tau, \{\phi_n : n \geq 0\})$ and $\delta' = (\tau', \{\phi'_n : n \geq 0\})$ *agree* $\mathbb{P}_0, \mathbb{P}_1$ a.s. iff $\tau = \tau'$ $\mathbb{P}_0, \mathbb{P}_1$ a.s. and, for all $n \geq 1$, $\phi_n(X_1, \ldots, X_n) = \phi'_n(X_1, \ldots, X_n)$ $\mathbb{P}_0, \mathbb{P}_1$ a.s. on the event $\{\tau = \tau' = n\}$ (we mean here a.s. property with respect to both measures $\mathbb{P}_0$ and $\mathbb{P}_1$).

Note now that, by Bayes fromula, the posterior distribution on $\Theta = \{0, 1\}$ at moment $t = n$ given observations $X_1, \ldots, X_n$ is

$$\pi_n := \mathbb{P}\{\vartheta = 1 | X_1, \ldots, X_n\} = \frac{\pi q(X_1) \ldots q(X_n)}{(1-\pi)p(X_1) \ldots p(X_n) + \pi q(X_1) \ldots q(X_n)},$$
$$1 - \pi_n := \mathbb{P}\{\vartheta = 0 | X_1, \ldots, X_n\} = \frac{(1-\pi)p(X_1) \ldots p(X_n)}{(1-\pi)p(X_1) \ldots p(X_n) + \pi q(X_1) \ldots q(X_n)}, n \geq 1.$$

We will also set $\pi_0 = \pi$. Recall that the sequence of observations $X_1, \ldots, X_n, \ldots$ is i.i.d., the cost of observations is a constant a and the costs of making errors W_0, W_1 are also constant. Therefore, given the event $\{\tau \geq n\}$, Bayes optimal sequential tests should behave at moment $t = n$ similarly to its behavior at moment $t = 0$, the only difference is that probabilities of ϑ being 1 or 0 are now π_n and $1 - \pi_n$ instead of π and $1 - \pi$. This makes it plausible to conjecture that the following statement holds.

Theorem 5.1 *Let δ^*_π be a sequential test with stopping time*

$$\tau^*_\pi := \inf\{n : \pi_n \notin (\gamma_1, \gamma_2)\}.$$

For all $n \geq 0$, given that $\tau^*_\pi = n$,

- *the tests accepts H_0 if $\pi_n \le \gamma_1$;*
- *the test accepts H_1 if $\pi_n \ge \gamma_2$.*

Then δ_π^* is the unique Bayes optimal test (in the sense that any other Bayes optimal test agrees with δ_π^* $\mathbb{P}_0, \mathbb{P}_1$ a.s.).

5.3 Problems

1. **Wald's identity**. Let $Y_1, Y_2, \dots$ be i.i.d. random variables with $\mathbb{E}|Y_1| < +\infty$ and let τ be a stopping time with values in $\{1, 2, \dots\}$ and with $\mathbb{E}\tau < \infty$. Prove that

$$\mathbb{E}(Y_1 + \dots + Y_\tau) = \mathbb{E}Y_1 \mathbb{E}\tau.$$

Hints. First show that

$$\mathbb{E}(Y_1 + \dots + Y_\tau) = \sum_{k=1}^{\infty} \sum_{j=1}^{k} \mathbb{E}Y_j I(\tau = k).$$

Change the order of summation in the above formula. You also have to know that

$$\mathbb{E}\tau = \sum_{j=1}^{\infty} \mathbb{P}\{\tau \ge j\}.$$

If you want to be rigorous, prove the identity first for non-negative random variables (in this case, there is no problem with changing the order of summation). In the general case, apply it to the absolute values of r.v. Y_j to justify the absolute convergence of the series and the change of the order of summation.

6 Likelihood Ratios, Wald's Sequential Probability Ratio Test and the Proof of Bayes Optimality

We will now discuss another way to describe Bayes optimal tests (in terms of likelihood ratios). Let

$$L_n := L_n(X_1, \dots, X_n) = \frac{q(X_1) \dots q(X_n)}{p(X_1) \dots p(X_n)}, \ n \ge 1$$

be the sequence of likelihood ratio statistics for our testing problem. Also set $L_0 := 1$. Note that condition $\pi_n \le \gamma_1$ is equivalent to $L_n \le \frac{1-\pi}{\pi} \frac{\gamma_1}{1-\gamma_1} =: \Gamma_1$ and condition $\pi_n \ge \gamma_2$ is equivalent to $L_n \ge \frac{1-\pi}{\pi} \frac{\gamma_2}{1-\gamma_2} =: \Gamma_2$. Thus, we could equivalently define the test δ_π^* as a sequential likelihood ratios test with stopping time

$$\tau_\pi^* = \inf\{n : L_n \notin (\Gamma_1, \Gamma_2)\}.$$

For all $n \ge 0$, given that $\tau_\pi^* = n$,

- the tests accepts H_0 if $L_n \le \Gamma_1$;
- the test accepts H_1 if $L_n \ge \Gamma_2$.

The following simple lemma provides some important information about stopping times of such tests.

Lemma 6.1 *Let $Y_1, \dots, Y_n$ be i.i.d. random variables and let $S_n := Y_1 + \cdots + Y_n$. For $n = 0$, set $S_n := 0$. For $c_1 \le c_2$, let*

$$\tau := \inf\{n \ge 0 : S_n \notin (c_1, c_2)\}.$$

If

$$\mathbb{P}\{Y_1 = 0\} < 1, \tag{6.1}$$

then there exists $\varepsilon > 0$ such that

$$\mathbb{P}\{\tau > n\} \le (1-\varepsilon)^n, n \ge 1.$$

Proof. If $c_1 < c_2 < 0$, or $0 < c_1 < c_2$, or $c_1 = c_2$, then $\tau = 0$ and the claim is trivial. In what follows, we assume that $c_1 < c_2$ and $c_1 c_2 \le 0$. It follows from condition (6.1) that, for some $t > 0, \lambda > 0$ we either have $\mathbb{P}\{Y_1 \ge t\} \ge \lambda$, or $\mathbb{P}\{Y_1 \le -t\} \ge \lambda$. To be specific, assume that the first inequality holds. Choose $m \ge 1$ such that $c_2 - c_1 < mt$. Then

$$\mathbb{P}\{S_m > c_2 - c_1\} \ge \mathbb{P}\{S_m \ge mt\} \ge \mathbb{P}\{Y_1 \ge t, \dots, Y_m \ge t\} \ge \lambda^m.$$

Assume that $n = mk$ for some $k \ge 1$. If $\tau > mk$, then $S_{jm} \in (c_1, c_2)$ for all $j = 0, \dots, k$. Therefore, we also have

$$S_{jm} - S_{(j-1)m} \le c_2 - c_1, j = 1, \dots, k.$$

This implies that

$$\begin{aligned}
&\mathbb{P}\{\tau > km\} \le \mathbb{P}\{S_{jm} - S_{(j-1)m} \le c_2 - c_1, j = 1, \dots, k\} \\
&= \prod_{j=1}^{k} \mathbb{P}\{S_{jm} - S_{(j-1)m} \le c_2 - c_1\} = \mathbb{P}\{S_m \le c_2 - c_1\}^k \le (1 - \lambda^m)^k.
\end{aligned}$$

For arbitrary n, choose k such that $km \le n < (k+1)m$. Then

$$\mathbb{P}\{\tau > n\} \le \mathbb{P}\{\tau > km\} \le (1-\lambda^m)^k \le (1-\lambda^m)^{\frac{n}{m}-1},$$

which easily implies the claim. ∎

The following proposition immediately follows.

Proposition 6.1 *If $p \neq q$, then for some $\varepsilon \in (0,1)$,*

$$\mathbb{P}_\theta\{\tau_\pi^* > n\} \leq (1-\varepsilon)^n, n \geq 1, \theta = 0, 1.$$

In particular, it follows that

$$\mathbb{E}_\theta(\tau_\pi^*)^k < \infty, k \geq 1, \theta = 0, 1.$$

Proof. Apply the previous lemma to r.v. $Y_j := \log \frac{q(X_j)}{p(X_j)}, j \geq 1$. ■

Let $\delta = (\tau, \{\phi_t : t \geq 0\})$ be a sequential test. Given $n \geq 1$, suppose event $\{\tau \geq n\}$ occurs, so, the test δ does not stop before moment n. Denote by $\delta_{X_1,\dots,X_n}$ the sequential test evolving in time $t' = t - n$ with stopping time $\tau' = \tau - n$ and the sequence of tests $\{\phi_{t';X_1,\dots,X_n} : t' = 1, 2, \dots\}$

$$\phi_{t';X_1,\dots,X_n}(X_1', \dots, X_{t'}') = \phi_{t'+n}(X_1, \dots, X_n, X_1', \dots, X_{t'}').$$

If $X_1' = X_{n+1}, \dots, X_k' = X_{n+k}, \dots$, the test $\delta_{X_1,\dots,X_n}$ makes the same decisions as δ (provided that δ does not stop before moment n), but it "resets the clock" $t' = t-n, \tau' = \tau - n$ at moment of time n.

We will need the following proposition that provides a formula for average posterior risk of sequential test δ given $X_1, \dots, X_n$.

Proposition 6.2 *For all $n \geq 1$, on the event $\{\tau \geq n\}$,*

$$\mathbb{E}(L(\vartheta, \delta)|X_1, \dots, X_n) = an + R_{\pi_n}(\delta_{X_1,\dots,X_n}) \text{ a.s.}$$

Proof. Indeed, on the event $\{\tau \geq n\}$, test δ has already suffered the loss an. After "resetting the clock", we can write the expected loss with respect to the action ν in terms of the loss of test $\delta_{X_1,\dots,X_n}$ as

$$L(\vartheta, \delta(X_1, X_2, \dots)) = an + L(\vartheta, \delta_{X_1,\dots,X_n}(X_1', X_2', \dots))$$

with $X_k' = X_{n+k}, k = 1, 2, \dots$. Since $(X_1', X_2', \dots)$ and $X_1, \dots, X_n$ are independent, and $(X_1', X_2', \dots) \stackrel{d}{=} (X_1, X_2, \dots)$, we have

$$\begin{aligned}
\mathbb{E}(L(\vartheta, \delta)|X_1, \dots, X_n) &= an + \mathbb{E}(L(\vartheta, \delta_{X_1,\dots,X_n}(X_1', X_2', \dots))|X_1, \dots, X_n) \\
&= an + \mathbb{P}\{\vartheta = 0|X_1, \dots, X_n\}\mathbb{E}(L(0, \delta_{X_1,\dots,X_n}(X_1', X_2', \dots))|\vartheta = 0, X_1, \dots, X_n) \\
&+ \mathbb{P}\{\vartheta = 1|X_1, \dots, X_n\}\mathbb{E}(L(1, \delta_{X_1,\dots,X_n}(X_1', X_2', \dots))|\vartheta = 1, X_1, \dots, X_n) \\
&= an + (1 - \pi_n)\mathbb{E}_0 L(0, \delta_{X_1,\dots,X_n}) + \pi_n \mathbb{E}_1 L(1, \delta_{X_1,\dots,X_n}) \\
&= an + R_{\pi_n}(\delta_{X_1,\dots,X_n}),
\end{aligned}$$

which completes the proof.

■

It will be said in what follows that two sequential tests $\delta = (\tau, \{\phi_n : n \geq 0\})$ and $\delta' = (\tau', \{\phi'_n : n \geq 0\})$ agree $\mathbb{P}_0, \mathbb{P}_1$ a.s. up to moment of time n iff, for all $j = 0, \dots, n$, events $\{\tau = j\}$ and $\{\tau' = j\}$ are equal $\mathbb{P}_0, \mathbb{P}_1$ a.s. and $\phi_j(X_1, \dots, X_j) = \phi'_j(X_1, \dots, X_j)$ $\mathbb{P}_0, \mathbb{P}_1$ a.s. on the event $\{\tau = \tau' = j\}$. This property means that $(\nu_j : j = 0, \dots, n) \stackrel{d}{=} (\nu'_j : j = 0, \dots, n)$ conditionally on $X_1, \dots, X_n$ $\mathbb{P}_0, \mathbb{P}_1$ a.s., where $\nu = \{\nu_j : j \geq 0\}$ and $\nu' = \{\nu'_j : j \geq 0\}$ are actions taken by δ and δ'. It will be also said that δ and δ' agree $\mathbb{P}_0, \mathbb{P}_1$ a.s. up to moment of time $n-1$, but disagree at moment n, iff they agree $\mathbb{P}_0, \mathbb{P}_1$ a.s. up to moment $n-1$ and either $\mathbb{P}_0(E) > 0$, or $\mathbb{P}_1(E) > 0$ for the following event $E \in \mathcal{F}_n$:

$$E := E_{n,\delta,\delta'} := (\{\tau = n\} \setminus \{\tau' = n\}) \cup (\{\tau' = n\} \setminus \{\tau = n\}) \cup (\{\tau = \tau' = n\} \cap \{\phi_n(X_1, \dots, X_n) \neq \phi'_n(X_1, \dots, X_n)\}). \tag{6.2}$$

Note that event E occurs when either one of the tests δ and δ' stops at moment n and another one does not, or when both tests stop at moment n, but they are choosing their actions with different probabilities.

The following proposition is crucial in the proof of Theorem 5.1.

Proposition 6.3 *Let δ be a sequential test.*

1. *If δ does not agree with δ^*_π at moment $n = 0$ with a positive probability, then there exists a sequential test $\bar{\delta}$ that agrees with δ^*_π at moment $n = 0$ with probability 1 and such that $R_\pi(\bar{\delta}) < R_\pi(\delta)$.*
2. *Similarly, for all $n \geq 1$, if δ agrees with δ^*_π up to moment $n-1$ $\mathbb{P}_0, \mathbb{P}_1$ a.s., but disagrees at moment n with a positive probability, then there exists a sequential test $\bar{\delta}$ that agrees with δ^*_π up to moment n and such that $R_\pi(\bar{\delta}) < R_\pi(\delta)$.*

Proof. The first claim follows from Proposition 5.1. To prove the claim for $n \geq 1$, assume that sequential test δ with stopping time τ agrees with δ^*_π up to moment $n-1$ with probability 1. This implies that the events $\{\tau \leq n-1\}$ and $\{\tau^*_\pi \leq n-1\}$ coincide $\mathbb{P}_0, \mathbb{P}_1$ a.s. Let $E = E_{n,\delta,\delta^*_\pi} \subset \{\tau \geq n\}, E \in \mathcal{F}_n$ be the event that δ and δ^*_π disagree at moment n (see (6.2)). Assume that $\mathbb{P}(E) > 0$ (which is equivalent to either $\mathbb{P}_0(E) > 0$ or $\mathbb{P}_1(E) > 0$). Using Proposition 6.2, the average risk of test δ could be written as follows:

$$\begin{aligned}
R_\pi(\delta) &= \mathbb{E}L(\vartheta, \delta) = \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + \mathbb{E}L(\vartheta, \delta)I(\tau \geq n) \\
&= \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + \mathbb{E}\mathbb{E}(L(\vartheta, \delta)|X_1, \dots, X_n)I(\tau \geq n) \\
&= \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + an\mathbb{P}\{\tau \geq n\} + \mathbb{E}R_{\pi_n}(\delta_{X_1,\dots,X_n})I(\tau \geq n) \\
&= \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + an\mathbb{P}\{\tau \geq n\} + \mathbb{E}R_{\pi_n}(\delta_{X_1,\dots,X_n})I_E + \mathbb{E}R_{\pi_n}(\delta_{X_1,\dots,X_n})I_{\{\tau \geq n\} \setminus E}.
\end{aligned}$$

Note that, on the event E where δ disagrees with δ^*_π at moment $t = n$, the test $\delta_{X_1,\dots,X_n}(X'_1, X'_2, \dots)$ disagrees with $(\delta^*_\pi)_{X_1,\dots,X_n}(X'_1, X'_2, \dots)$ at moment $t' = 0$. By Proposition 5.1, it follows that on this event

$$R_{\pi_n}(\delta_{X_1,\dots,X_n}) > \psi(\pi_n).$$

Therefore, there exists a sequential test $\tilde{\delta}_{X_1,\ldots,X_n} = (\tilde{\tau}, \{\tilde{\phi}_{t';X_1,\ldots,X_n} : t' \geq 0\})$ such that

$$R_{\pi_n}(\delta_{X_1,\ldots,X_n}) > R_{\pi_n}(\tilde{\delta}_{X_1,\ldots,X_n}).$$

We will now define a sequential test $\bar{\delta} = (\bar{\tau}, \{\bar{\phi}_t : t \geq 0\})$ that agrees with δ up to moment of time $n-1$ and, on the event $\{\tau \geq n\} \setminus E$ continues as $\delta_{X_1,\ldots,X_n}$, whereas, on the event E, it continues as $\tilde{\delta}_{X_1,\ldots,X_n}$. To this end, we set $\bar{\tau} := \tau$ on the event $\{\tau \leq n-1\}$. Suppose also that, for $j = 0, \ldots, n-1$, $\bar{\phi}_j(X_1, \ldots, X_j) = \phi_j(X_1, \ldots, X_j)$ on the event $\{\tau = j\} = \{\bar{\tau} = j\}$. Let $\bar{\tau} := \tau$ on the event $\{\tau \geq n\} \setminus E$. For $j = 0, 1, \ldots,$ on each of the events $\{\tau = n + j\} \setminus E$, define

$$\bar{\phi}_{n+j}(X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+j}) = \phi_{n+j}(X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+j}).$$

Finally, on the event E, define $\bar{\tau} := \tilde{\tau} + n$ and, on each of the events $\{\bar{\tau} = n + j\} \cap E$, set

$$\bar{\phi}_{n+j}(X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+j}) = \tilde{\phi}_{j,X_1,\ldots,X_n}(X_{n+1}, \ldots, X_{n+j}).$$

For test $\bar{\delta}$,

$$\begin{aligned}
&R_\pi(\bar{\delta}) \\
&= \mathbb{E}L(\vartheta, \bar{\delta})I(\bar{\tau} \leq n-1) + an\mathbb{P}\{\bar{\tau} \geq n\} \\
&+ \mathbb{E}R_{\pi_n}(\bar{\delta}_{X_1,\ldots,X_n})I_E + \mathbb{E}R_{\pi_n}(\bar{\delta}_{X_1,\ldots,X_n})I_{\{\bar{\tau}\geq n\}\setminus E} \\
&< \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + an\mathbb{P}\{\tau \geq n\} \\
&+ \mathbb{E}R_{\pi_n}(\delta_{X_1,\ldots,X_n})I_E + \mathbb{E}R_{\pi_n}(\delta_{X_1,\ldots,X_n})I_{\{\tau\geq n\}\setminus E} \\
&< R_\pi(\delta),
\end{aligned}$$

where we use the facts that

$$\mathbb{E}L(\vartheta, \bar{\delta})I(\bar{\tau} \leq n-1) + an\mathbb{P}\{\bar{\tau} \geq n\} = \mathbb{E}L(\vartheta, \delta)I(\tau \leq n-1) + an\mathbb{P}\{\tau \geq n\},$$

$$\mathbb{E}R_{\pi_n}(\bar{\delta}_{X_1,\ldots,X_n})I_{\{\bar{\tau}\geq n\}\setminus E} = \mathbb{E}R_{\pi_n}(\delta_{X_1,\ldots,X_n})I_{\{\tau\geq n\}\setminus E}$$

and

$$\mathbb{E}R_{\pi_n}(\bar{\delta}_{X_1,\ldots,X_n})I_E < \mathbb{E}R_{\pi_n}(\delta_{X_1,\ldots,X_n})I_E$$

since $R_{\pi_n}(\bar{\delta}_{X_1,\ldots,X_n} < R_{\pi_n}(\delta_{X_1,\ldots,X_n})$ on E and $\mathbb{P}(E) > 0$. ■

We can now complete the proof of Theorem 5.1.

Proof. Assume there exists a sequential test δ with stopping time τ such that $R_\pi(\delta) < R_\pi(\delta^*_\pi)$. This test could not agree with δ^*_π for all $n \geq 0$. Thus, either δ disagrees with δ^*_π with a positive probability at moment $n = 0$, or there exists an $n \geq 1$ such that δ agrees with δ^*_π up to moment of time $n-1$ $\mathbb{P}_0, \mathbb{P}_1$ a.s. and disagrees with δ^*_π at moment n with a positive probability. By induction, we can deduce from Proposition 6.3 that, for any

$n \geq 0$, there exists a sequential test $\delta^{(n)}$ such that $\delta^{(n)}$ agrees with δ_π^* up to moment of time n with probability 1 and

$$R_\pi(\delta_\pi^*) > R_\pi(\delta) \geq \cdots \geq R_\pi(\delta^{(n)}) \geq R_\pi(\delta^{(n+1)}) \geq \ldots$$

This implies that

$$\begin{aligned}
&0 < \varepsilon := R_\pi(\delta_\pi^*) - R_\pi(\delta) \leq R_\pi(\delta_\pi^*) - R_\pi(\delta^{(n)}) \\
&= \mathbb{E}L(\vartheta, \delta_\pi^*) I(\tau_\pi^* \leq n) + \mathbb{E}L(\vartheta, \delta_\pi^*) I(\tau_\pi^* \geq n+1) - \mathbb{E}L(\vartheta, \delta^{(n)}) I(\tau^{(n)} \leq n) - \mathbb{E}L(\vartheta, \delta^{(n)}) I(\tau^{(n)} \geq n+1) \\
&\leq \mathbb{E}L(\vartheta, \delta_\pi^*) I(\tau_\pi^* \geq n+1),
\end{aligned}$$

where we used the fact that tests δ_π^* and $\delta^{(n)}$ agree up to moment $n-1$ with probability 1 and, as a consequence,

$$\mathbb{E}L(\vartheta, \delta_\pi^*) I(\tau_\pi^* \leq n) = \mathbb{E}L(\vartheta, \delta^{(n)}) I(\tau^{(n)} \leq n).$$

Therefore,

$$\begin{aligned}
\varepsilon &\leq \mathbb{E}L(\vartheta, \delta_\pi^*) I(\tau_\pi^* \geq n+1) \\
&\leq a(1-\pi)\mathbb{E}_0 \tau_\pi^* I(\tau_\pi^* \geq n+1) + a\pi \mathbb{E}_1 \tau_\pi^* I(\tau_\pi^* \geq n+1) \\
&+ (1-\pi) W_0 \mathbb{P}_0\{\tau_\pi^* \geq n+1\} + \pi W_1 \mathbb{P}_1\{\tau_\pi^* \geq n+1\}.
\end{aligned}$$

Using Proposition 6.1, it is easy to see that the right hand side of the last bound tends to zero as $n \to \infty$, which contradicts the fact that $\varepsilon > 0$.

We can conclude that, for any sequential test δ, $R_\pi(\delta_\pi^*) \leq R_\pi(\delta)$. Hence δ_π^* is Bayes optimal.

Suppose there exists a sequential test δ such that $R_\pi(\delta) = R_\pi(\delta_\pi^*)$. Then δ must agree with δ_π^* at moment $n = 0$. If we assume that δ agrees with δ_π^* up to moment $n-1$ with probability 1, but they disagree at moment n with a positive probability, then there exists a test $\bar{\delta}$ for which $R_\pi(\bar{\delta}) < R_\pi(\delta) = R_\pi(\delta_\pi^*)$. This contradicts the fact that δ_π^* is Bayes optimal. Thus, any Bayes optimal test δ must agree with δ_π^* for all $n \geq 0$ with probability 1.

■

We now turn again to the definition of sequential test in terms of likelihood ratios

$$L_n := L_n(X_1, \ldots, X_n) = \frac{q(X_1) \ldots q(X_n)}{p(X_1) \ldots p(X_n)},\ n \geq 1.$$

Suppose $\Gamma_1 < 1 < \Gamma_2$ and let $\delta_{\Gamma_1, \Gamma_2}$ be the sequential test with stopping time

$$\tau(\Gamma_1, \Gamma_2) := \inf\{n \geq 1 : L_n \notin (\Gamma_1, \Gamma_2)\}$$

that, for $\tau(\Gamma_1, \Gamma_2) = n$, accepts H_1 if $L_n \geq \Gamma_2$ and accepts H_0 if $L_n \leq \Gamma_1$.

For a sequential test δ, denote by $\alpha_0(\delta)$ the probability that δ rejects H_0 when it is true and by $\alpha_1(\delta)$ the probability to reject H_1 when it is true. Note that, for a sequential test delta with stopping time τ,

$$R_\pi(\delta) = (1-\pi)[a\mathbb{E}_0\tau + W_0\alpha_0(\delta)] + \pi[a\mathbb{E}_0\tau + W_1\alpha_1(\delta)].$$

Our main goal is to prove the following theorem due to A. Wald.

Theorem 6.1 *Denote*

$$\alpha_0 := \alpha_0(\delta_{\Gamma_1,\Gamma_2}),\ \alpha_1 := \alpha_1(\delta_{\Gamma_1,\Gamma_2}).$$

For any sequential test δ with stopping time τ and with $\alpha_0(\delta) \le \alpha_0, \alpha_1(\delta) \le \alpha_1$, the following inequalities hold:

$$\mathbb{E}_0\tau \ge \mathbb{E}_0\tau(\Gamma_1,\Gamma_2),\ \mathbb{E}_1\tau \ge \mathbb{E}_1\tau(\Gamma_1,\Gamma_2).$$

Wald's Theorem shows that among all the sequential tests achieving certain probabilities of the errors α_0, α_1, the test $\delta_{\Gamma_1,\Gamma_2}$ with properly chosen thresholds Γ_1, Γ_2 does it, on average, as fast as possible. We will prove this theorem by showing that $\delta_{\Gamma_1,\Gamma_2}$ could be viewed as a Bayes optimal test for proper costs a, W_0, W_1.

Proof. Let $a = 1$ and $W_0, W_1 > 0$. It was proved in the previous section that there exists numbers $\gamma_1 = \gamma_1(W_0, W_1)$, $\gamma_2 = \gamma_2(W_0, W_1)$ such that $\gamma_1 \le \gamma_2$ and, for $\pi \in (0,1)$, sequential test δ^*_π with thresholds γ_1, γ_2 defined in Theorem 5.1 is Bayes optimal. Moreover, if we set

$$\Gamma_1 := \frac{1-\pi}{\pi}\frac{\gamma_1}{1-\gamma_1}\quad \Gamma_2 := \frac{1-\pi}{\pi}\frac{\gamma_2}{1-\gamma_2}, \tag{6.3}$$

then $\delta^*_\pi = \delta_{\Gamma_1,\Gamma_2}$. Thus, for given Γ_1, Γ_2 and π, it is enough to find γ_1, γ_2 from the equations (6.3), and then to solve the equations

$$\begin{cases} \gamma_1(W_0, W_1) = \gamma_1 \\ \gamma_2(W_0, W_1) = \gamma_2 \end{cases} \tag{6.4}$$

to determine the costs W_0, W_1. It could be proved (based on the properties of the functions $\gamma_1(W_0, W_1)$, $\gamma_2(W_0, W_1)$) that the solution of equations (6.4) does exist (but we will skip it here). Since $\delta_{\Gamma_1,\Gamma_2}$ is Bayes optimal for prior π, we have $R_\pi(\delta_{\Gamma_1,\Gamma_2}) \le R_\pi(\delta)$ for all sequential tests δ. Therefore, for all $\pi \in (0,1)$,

$$\begin{aligned}&(1-\pi)[\mathbb{E}_0\tau(\Gamma_1,\Gamma_2) + W_0\alpha_0(\delta_{\Gamma_1,\Gamma_2})] + \pi[\mathbb{E}_0\tau(\Gamma_1,\Gamma_2) + W_1\alpha_1(\delta_{\Gamma_1,\Gamma_2})] \\ &\le (1-\pi)[\mathbb{E}_0\tau + W_0\alpha_0(\delta)] + \pi[\mathbb{E}_0\tau + W_1\alpha_1(\delta)].\end{aligned}$$

Since

$$\alpha_0(\delta) \le \alpha_0 = \alpha_0(\delta_{\Gamma_1,\Gamma_2}), \alpha_1(\delta) \le \alpha_1 = \alpha_1(\delta_{\Gamma_1,\Gamma_2}),$$

we can conclude that

$$\mathbb{E}_0\tau \ge \mathbb{E}_0\tau(\Gamma_1,\Gamma_2),\ \mathbb{E}_1\tau \ge \mathbb{E}_1\tau(\Gamma_1,\Gamma_2).$$

■

To find the thresholds Γ_1, Γ_2 of Wald's test for the desired probabilities of the error α_0, α_1, on needs to solve the equations $\alpha_0(\delta_{\Gamma_1,\Gamma_2}) = \alpha_0$, $\alpha_1(\Gamma_1, \Gamma_2) = \alpha_1$, which is not an easy problem. The following simple proposition could be useful to solve this problem approximately.

Proposition 6.4 *The following inequalities hold:*

$$\alpha_0(\delta_{\Gamma_1,\Gamma_2}) \le \frac{1-\alpha_1(\delta_{\Gamma_1,\Gamma_2})}{\Gamma_2},\ \alpha_1(\delta_{\Gamma_1,\Gamma_2}) \le \Gamma_1(1-\alpha_0(\delta_{\Gamma_1,\Gamma_2})).$$

Proof. Let

$$C_n := \{(x_1,\dots,x_n) : L_k(x_1,\dots,x_k) \in (\Gamma_1,\Gamma_2), k = 1,\dots,n-1,\ L_n(x_1,\dots,x_n) \ge \Gamma_2\}.$$

Then

$$\begin{aligned}\alpha_0(\delta_{\Gamma_1,\Gamma_2}) &= \sum_{n=1}^{\infty} \int_{C_n} p(x_1)\dots p(x_n)\mu(dx_1)\dots\mu(dx_n)\\ &\le \frac{1}{\Gamma_2}\sum_{n=1}^{\infty}\int_{C_n} q(x_1)\dots q(x_n)\mu(dx_1)\dots\mu(dx_n)\\ &= \frac{1-\alpha_1(\delta_{\Gamma_1,\Gamma_2})}{\Gamma_2}.\end{aligned}$$

The proof of the second inequality is similar. ■

If now Γ_1, Γ_2 are such that $\alpha_0(\delta_{\Gamma_1,\Gamma_2}) = \alpha_0$, $\alpha_1(\Gamma_1,\Gamma_2) = \alpha_1$, we have

$$\alpha_0 \le \frac{1-\alpha_1}{\Gamma_2},\ \alpha_1 \le \Gamma_1(1-\alpha_0),$$

implying that

$$\Gamma_1 \ge \Gamma_1' := \frac{\alpha_1}{1-\alpha_0},\ \ \Gamma_2 \le \frac{1-\alpha_1}{\alpha_0} =: \Gamma_2'.$$

Using again the same inequalities for the test $\delta_{\Gamma_1',\Gamma_2'}$ yields

$$\alpha_0(\delta_{\Gamma_1',\Gamma_2'}) \le \frac{\alpha_0}{1-\alpha_1},\ \ \alpha_1(\delta_{\Gamma_1',\Gamma_2'}) \le \frac{\alpha_1}{1-\alpha_0}.$$

If α_0, α_1 are small numbers, the test $\delta_{\Gamma_1',\Gamma_2'}$ does achieve the desired probabilities of the errors approximately.

6.1 Problems

1. Prove that

$$\alpha_1(\delta_{\Gamma_1,\Gamma_2}) \leq \Gamma_1(1 - \alpha_0(\delta_{\Gamma_1,\Gamma_2})).$$

2. Let

$$K(p||q) := \mathbb{E}_p \log \frac{p(X)}{q(X)}$$

be the Kullback-Leibler divergence between densities p and q (using Jensen's inequality, it is easy to prove that $K(p||q) \geq 0$ with $K(p||q) = 0$ if and only if $p = q$ μ a.s.). Let $\tau := \tau(\Gamma_1, \Gamma_2)$. Use Wald's identity to show that

$$\mathbb{E}_0 \log L_\tau(X_1, \ldots, X_\tau) = -\mathbb{E}_0 \tau K(p||q), \quad \mathbb{E}_1 \log L_\tau(X_1, \ldots, X_\tau) = \mathbb{E}_1 \tau K(q||p). \tag{6.5}$$

Assuming that, for some constant $C > 0$, $|\log \frac{q(X)}{p(X)}| \leq C$ P, Q-a.s., use (6.5) to derive upper bounds on the expectations $\mathbb{E}_0\tau, \mathbb{E}_1\tau$.

7 Composite Hypotheses Testing: Uniformly Most Powerful Tests

Let $X \sim p_\theta, \theta \in \Theta$, where p_θ are densities w.r.t. a σ-finite measure μ on $(S, \mathcal{A})$. Suppose that Θ_0, Θ_1 is a partition of parameter space $\Theta : \Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0, \Theta_1 \neq \emptyset$. Consider the following composite hypotheses testing problem:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1. \end{cases}$$

Let $\phi : S \mapsto [0, 1]$ be a test for the null hypotheses H_0 against the alternative H_a.

Definition 7.1 *Let*

$$\beta_\phi(\theta) := \mathbb{E}_\theta \phi(X) = \int_S \phi p_\theta d\mu, \theta \in \Theta.$$

We will call $\beta_\phi(\theta), \theta \in \Theta$ the power function of test ϕ.

Clearly, $\beta_\phi(\theta)$ is the probability that test ϕ rejects H_0 when H_0 is true provided that $X \sim p_\theta$. For a good test ϕ, β_ϕ should be small on set Θ_0 (where H_0 is true) and large on set Θ_1 (where H_a is true).

Definition 7.2 *It will be said that ϕ is a test of size α for $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$ iff $\beta_\phi(\theta) \leq \alpha, \theta \in \Theta_0$.*

Definition 7.3 *Let ϕ^* be a test of size α. It will be called a uniformly most powerful (UMP) test of size α iff for all tests ϕ of size α*

$$\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta), \theta \in \Theta_1.$$

In other words, UMP tests of size α are the solutions of the following optimization problem:

$$\begin{cases} \beta_\phi(\theta) \longrightarrow \max, \theta \in \Theta_1 \\ \beta_\phi(\theta) \leq \alpha, \theta \in \Theta_0. \end{cases}$$

This is a multicriteria optimization problem (simultaneous maximization of powers for all alternatives $\theta \in \Theta_1$ subject to the size constraints on the null set Θ_0).

Note that if there exists a UMP test ϕ^* of size α for $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$, then ϕ^* is most powerful of size α for all hypotheses testing problems $H_0 : \theta \in \Theta_0$ against $H_{a,\theta_1} : \theta = \theta_1$, $\theta_1 \in \Theta_1$. All such tests should "agree" for all choices of $\theta_1 \in \Theta_1$, which is a restrictive constraint on the model $\{p_\theta : \theta \in \Theta\}$. Thus, UMP tests exist only for very special statistical models and for very special hypotheses.

Definition 7.4 *Let $X \sim p_\theta$, $\theta \in \Theta \subset \mathbb{R}$. The model $\{p_\theta : \theta \in \Theta\}$ will be called a monotone likelihood ratios family (MLR) iff there exists a measurable function $T : S \mapsto \mathbb{R}$ and, for all $\theta', \theta'' \in \Theta, \theta' < \theta''$, there exists a function $\psi_{\theta',\theta''}$ such that*

$$\frac{p_{\theta''}(x)}{p_{\theta'}(x)} = \psi_{\theta',\theta''}(T(x)),$$

and, moreover, all the functions $\psi_{\theta',\theta''}$ are either non-decreasing, or non-increasing. In what follows, we will call $\psi_{\theta',\theta''}$ link functions of MLR family.

Clearly, by Neyman-Fisher factorization, $T(X)$ is a sufficient statistic for such an MLR family.

Example 7.1 Let

$$p_\theta(x) := \frac{1}{Z(\theta)} \exp\{A(\theta)T(x)\}h(x), x \in S, \theta \in \Theta \subset \mathbb{R}$$

be a one parameter exponential family. Here $h : S \mapsto \mathbb{R}_+$ is a non-negative measurable function and

$$Z(\theta) := \int_S \exp\{A(\theta)T(x)\}h(x)\mu(dx) < \infty, \theta \in \Theta.$$

If $A : \Theta \mapsto \mathbb{R}$ is a monotone (non-decreasing or non-increasin function), then $\{p_\theta : \theta \in \Theta\}$ is an MLR family w.r.t. T.

Many classical models in statistics are one parameter exponential families (and, hence, also MLR families), in particular:

- normal model with unknown mean $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, 1), \theta \in \mathbb{R}$;
- normal model with unknown variance $X_1, \ldots, X_n$ i.i.d. $\sim N(0, \sigma^2), \sigma^2 > 0$;
- Poisson model $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{P}(\theta), \theta > 0$;
- Binomial model $X \sim B(n; \theta), \theta \in [0, 1]$.

Example 7.2 Consider the following Cauchy model: $X \sim$ Cauchy$(\theta), \theta \in \mathbb{R}$, where Cauchy$(\theta)$ is the distribution in $\mathbb{R}$ with density:

$$p_\theta(x) := \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, x \in \mathbb{R}$$

It is easy to check that this model is not an MLR family.

Example 7.3 A more general example is a location family $p_\theta(x) := p(x - \theta), x, \theta \in \mathbb{R}$ generated by a density p. Such a family is MLR if and only if p is log-concave (which means that $\log p$ is a concave function on $\mathbb{R}$, see A. Saumard and J. Wellner, *Statistics Surveys*, 2014, 8, 45–114, Proposition 2.3 (b).

Indeed, MLR assumption for p_θ means that, for all $\theta' < \theta'', x' < x''$

$$\frac{p(x' - \theta'')}{p(x' - \theta')} \leq \frac{p(x'' - \theta'')}{p(x'' - \theta')}.$$

If $g := \log p$, we can rewrite the last inequality as

$$g(x' - \theta'') + g(x'' - \theta') \leq g(x' - \theta') + g(x'' - \theta"). \tag{7.1}$$

Denote $a := x' - \theta''$ and $b := x'' - \theta'$. Then $b - a = (x'' - x') + (\theta'' - \theta') > 0$. Note that $x' - \theta' = a + (\theta'' - \theta')$ and $x'' - \theta'' = b - (\theta'' - \theta')$. It easily follows that

$$x' - \theta' = (1 - \lambda)a + \lambda b, \quad x'' - \theta'' = \lambda a + (1 - \lambda)b$$

with $\lambda := \frac{\theta'' - \theta'}{b - a}$. Assuming that p is log-concave (or g is concave), we get

$$g(x' - \theta') = g((1 - \lambda)a + \lambda b) \geq (1 - \lambda)g(a) + \lambda g(b)$$

and

$$g(x'' - \theta'') = g(\lambda a + (1 - \lambda)b) \geq \lambda g(a) + (1 - \lambda)g(b).$$

Adding up the last two inequalities yields (7.1).

On the other hand, assume that (7.1) holds. For arbitrary $a < b$, choose $\theta' < \theta''$ such that $\theta'' - \theta' = \frac{b-a}{2}$. Let $x' := \theta' + \frac{a+b}{2}$ and $x'' := \theta'' + \frac{a+b}{2}$. Then $x' - \theta' = x'' - \theta'' = \frac{a+b}{2}$ and it follows from (7.1) that $g(\frac{a+b}{2}) \geq \frac{g(a)+g(b)}{2}$, so g is midpoint concave. Since g is also a Borel measurable function, its concavity follows from Sierpinski's theorem.

The following simple fact is called Karlin-Rubin Theorem.

Theorem 7.1 *Let $\{p_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$ be an MLR family with respect to statistic T and with (to be specific) non-decreasing link functions. Let $X \sim p_\theta, \theta \in \Theta$. For $\theta_0 \in \Theta$, consider a composite hypotheses testing problem*

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0. \end{cases}$$

Then

1. *There exist $c \in \mathbb{R}, \gamma \in [0,1]$ such that*

$$\beta_{\phi^*_{c,\gamma}}(\theta_0) = \mathbb{E}_{\theta_0} \phi^*_{c,\gamma}(X) = \alpha$$

   for the test of the following form:

$$\phi^*_{c,\gamma}(X) := \begin{cases} 1 & \text{if } T(X) > c \\ \gamma & \text{if } T(X) = c \\ 0 & \text{if } T(X) < c. \end{cases}$$

2. *For this choice of c, γ, $\phi^* := \phi^*_{c,\gamma}$ is UMP of size α.*

3. *The power function $\beta_{\phi^*}(\theta), \theta \in \Theta$ is non-decreasing. Moreover, if the model $\{p_\theta : \theta \in \Theta\}$ is identifiable ($p_{\theta'} \neq p_{\theta''}$ for $\theta' \neq \theta''$), then $\beta_{\phi^*}(\theta)$ is strictly increasing on the set $\{\theta \in \Theta : \beta_{\phi^*}(\theta) \in (0,1)\}$.*

Proof. Let $\theta' < \theta'', \theta', \theta'' \in \Theta$. Consider the following simple hypotheses testing problem:

$$\begin{cases} H'_0 : \theta = \theta' \\ H'_a : \theta = \theta'' \end{cases}$$

Let

$$\phi^*(X) := \begin{cases} 1 & \text{if } T(X) > c \\ \gamma & \text{if } T(X) = c \\ 0 & \text{if } T(X) < c. \end{cases}$$

It easily follows from MLR property that

$$\frac{p_{\theta''}(X)}{p_{\theta'}(X)} > \psi_{\theta',\theta''}(c) \text{ implies that } T(X) < c \text{ implies that } \phi^*(X) = 1$$
$$\frac{p_{\theta''}(X)}{p_{\theta'}(X)} < \psi_{\theta',\theta''}(c) \text{ implies that } T(X) < c \text{ implies that } \phi^*(X) = 0.$$

This means that ϕ^* is a Neyman-Pearson type test of size $\alpha^{'} = \beta' = \beta_{\phi^*}(\theta')$, and, as a consequence of Neyman-Pearson Lemma, it is most powerful of size α' for H_0' against H_a'.

Since $\beta^{''} := \beta_{\phi^*}(\theta'')$ is the power of this test, we have $\beta^{''} \geq \alpha^{'} = \beta^{'}$, implying that β_{ϕ^*} is non-decreasing. Moreover, for $\alpha' \in (0,1)$, $\beta^{''} > \alpha^{'} = \beta^{'}$ unless $p_{\theta'} = p_{\theta''}$ μ a.s. Therefore, β_{ϕ^*} is a strictly increasing function on the set $\{\theta \in \Theta : \beta_{\phi^*}(\theta) \in (0,1)\}$, provided that $\{p_\theta : \theta \in \Theta\}$ is an identifiable model.

One can show (similarly to the proof of the statement 1 of Neyman-Pearson Lemma) that there exist $c \in \mathbb{R}, \gamma \in [0,1]$ such that $\beta_{\phi^*}(\theta_0) = \mathbb{E}_{\theta_0}\phi^*(X) = \alpha$. For $\theta' = \theta_0$ and $\theta'' > \theta_0$, we can conclude that ϕ^* is most powerful test of size α for hypothesis $\theta = \theta_0$ against any alternative $\theta'' > \theta_0$, which implies that ϕ^* is UMP test of size α for $\theta = \theta_0$ against $\theta > \theta_0$. Since β_{ϕ^*} is non-decreasing, we also have $\beta_{\phi^*}(\theta) \leq \beta_{\phi^*}(\theta_0) = \alpha, \theta \leq \theta_0$, implying that ϕ^* is UMP of size α for $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$. ■

We will conclude this section with a curious example.

Example 7.4 Let $X_1, \ldots, X_n$ be i.i.d. $\sim$ Cauchy$(\theta), \theta \in \mathbb{R}$. The following statement will be proved:

Proposition 7.1 *For all $\alpha \in (0, 1/2)$, there is no UMP test of size α for hypothesis $H_0 : \theta \leq 0$ against the alternative $H_a : \theta > 0$.*

Here is the plan of our proof. We will first show that there exists a unique test ϕ^* for this testing problem satisfying a weaker optimality property, namely, it is *locally most powerful*.

Consider a testing problem

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_a : \theta > \theta_0 \end{cases} \quad X \sim p_\theta, \theta \in \mathbb{R}.$$

Suppose, for all $\phi \in \Phi$, β_ϕ is a differentiable function in $\mathbb{R}$. A test ϕ^* is said to be locally most powerful (LMP) of size α for H_0 against H_a iff $\beta_{\phi^*}(\theta_0) = \alpha$, and, for all tests ϕ with $\beta_\phi(\theta_0) = \alpha$, we have $\beta'_{\phi^*}(\theta_0) \geq \beta'_{\phi^*}(\theta_0)$. Thus, locally most powerful tests maximize the slope of the power function at point θ_0 subject to the size constraint. It is very easy to check that if ϕ^* is uniformly most powerful of size α for H_0 against H_a then it is also locally most powerful at point θ_0. Therefore, to prove that there are no UMP tests of

size α it would be enough to show that there is a unique LMP test of size α, but it is not a UMP test.

Assuming that $\theta \mapsto p_\theta(\cdot)$ is sufficiently regular, we have for all tests ϕ

$$\beta'_\phi(\theta) = \int_S \phi \frac{\partial p_\theta}{\partial \theta} d\mu$$

The proof of the following proposition is a straightforward modification of the proof of Neyman-Pearson Lemma.

Proposition 7.2 *Consider the following test*

$$\phi_c^*(X) := \begin{cases} 1 & \text{if } \frac{\partial p_\theta(X)}{\partial \theta}_{|\theta=\theta_0} \geq c p_{\theta_0}(X) \\ 0 & \text{otherwise.} \end{cases}$$

If there exists $c \in \mathbb{R}$ such that $\beta_{\phi_c^*}(\theta_0) = \alpha$, then ϕ_c^* is LMP of size α.

In other words, this LMP test is based on the test statistic $\frac{\partial}{\partial \theta} \log p_\theta(X)_{|\theta=\theta_0}$ (often called the score function) whose values are to be compared with threshold c to make a decision. It is also easy to see that in the case when the distribution of the score function is continuous, ϕ_c^* is the unique LMP test a.s..

Proof. We have

$$p_\theta(X_1, \ldots, X_n) = \pi^{-n} \prod_{j=1}^n \frac{1}{1 + (X_j - \theta)^2}$$

and

$$\begin{aligned} \frac{\partial}{\partial \theta} \log p_\theta(X_1, \ldots, X_n) &= -\sum_{j=1}^n \frac{\partial}{\partial \theta} \log(1 + (X_j - \theta)^2) \\ &= 2 \sum_{j=1}^n \frac{X_j - \theta}{1 + (X_j - \theta)^2}. \end{aligned}$$

For $\theta = 0$, we get the following test statistic:

$$T_n(X_1, \ldots, X_n) := 2 \sum_{j=1}^n \frac{X_j}{1 + X_j^2}.$$

The rejection region of LMP test of size α is $T_n \geq c$ for c satisfying $\mathbb{P}_0\{T_n \geq c\} = \alpha$. Note that, under H_0, T_n is a symmetric r.v. and $\mathbb{P}_0\{T_n \geq 0\} = 1/2$. Thus, for $\alpha \in (0, 1/2)$, we have $c > 0$. Since $\frac{x}{1+x^2} \to 0$ as $x \to \infty$, the rejection region $\{(x_1, \ldots, x_n) : T_n(x_1, \ldots, x_n) \geq c\}$ satisfies the condition

$$\{(x_1, \ldots, x_n) : T_n(x_1, \ldots, x_n) \geq c\} \subset \{(x_1, \ldots, x_n) : \min_{1 \leq j \leq n} |x_j| \leq a\}$$

for some $a > 0$. Therefore,

$$\beta_{\phi_c^*}(\theta) = \mathbb{P}_\theta\{T_n(X_1, \ldots, X_n) \geq c\} \leq \mathbb{P}_\theta\{\min_{1 \leq j \leq n} |X_j| \leq a\}$$
$$\leq n\mathbb{P}_\theta\{|X_1| \leq a\} \to 0$$

as $|\theta| \to \infty$. This shows that the test ϕ_c^* is not UMP (since for UMP test we would have $\beta_\phi(\theta) > \alpha$ for $\theta > 0$.)

■

7.1 Problems

1. Check the claims of examples 7.1 and 7.2.
2. State and prove a version of Karlin-Rubin Theorem in each of the following cases:

   (a) when the link functions are non-decreasing and you want to test $H_0 : \theta \geq \theta_0$ against $H_a : \theta < \theta_0$;

   (b) when the link functions are non-increasing and you want to test $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$;

   (c) when the link functions are non-increasing and you want to test $H_0 : \theta \geq \theta_0$ against $H_a : \theta < \theta_0$.
3. Show that the test described in Theorem 7.1 is least powerful of size α for $H_0 : \theta \geq \theta_0$ against $H_a : \theta < \theta_0$.
4. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta; 1)$. Find most the powerful test of size α for $H_0 : \theta \geq 1$ against $H_a : \theta \leq 0$.
5. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta; 1), \theta \in \mathbb{R}$. Show that there is no most powerful test of size α for $H_0 : \theta \in [0, 1]$ against $H_a : \theta \notin [0, 1]$.
6. Prove that UMP test of size α is aso LMP of size α (under proper assumptions).
7. Prove Proposition 7.2.

8 Generalized Neyman-Pearson Lemma

In this section, we consider a well known generalization of Neyman-Pearson Lemma due to Dantzig and Wald (see G. Dantzig and A. Wald, On the Fundamental Lemma of Neyman and Pearson, *Ann. Math. Statis.*, 1951, 1, 87–93). It describes the solutions of

the following optimization problem

$$\int_S \phi f_{m+1} d\mu \longrightarrow \max$$
$$s.t. \int_S \phi f_j d\mu = \alpha_j, j = 1, \ldots, m, \ \phi : S \mapsto [0,1]$$

for given μ-integrable functions $f_1, \ldots, f_{m+1}$ and numbers $\alpha_1, \ldots, \alpha_j$, and it is widely used to develop tests with optimal properties in a variety of problems in which Neyman-Pearson Lemma itself is not sufficient.

Theorem 8.1 *Let $f_1, \ldots, f_{m+1} : S \mapsto \mathbb{R}$ be μ-integrable functions on $(S, \mathcal{A})$. Suppose, for given numbers $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, there exists a measurable function $\phi : S \mapsto [0,1]$ such that*

$$\int_S \phi f_j d\mu = \alpha_j, j = 1, \ldots, m. \tag{8.1}$$

Then the following statements hold:

1. *There exists a measurable function $\phi^* : S \mapsto [0,1]$ that maximizes $\int_S \phi f_{m+1} d\mu$ subject to constraints* (8.1).

2. *If*

$$\phi^*(x) := \begin{cases} 1 & \text{for } f_{m+1}(x) > \sum_{j=1}^m u_j f_j(x) \\ 0 & \text{for } f_{m+1}(x) < \sum_{j=1}^m u_j f_j(x) \end{cases} \tag{8.2}$$

   with some $u_1, \ldots, u_m \in \mathbb{R}$, and ϕ^* satisfies the constraints (8.1)*, then ϕ^* maximizes $\int_S \phi f_{m+1} d\mu$ subject to these constraints.*

3. *If ϕ^* is defined by* (8.2) *with $u_j \geq 0, j = 1, \ldots, m$ and satisfies the constraints* (8.1)*, then ϕ^* maximizes $\int_S \phi f_{m+1} d\mu$ subject to the constraints*

$$\int_S \phi f_j d\mu \leq \alpha_j, j = 1, \ldots, m. \tag{8.3}$$

4. *Finally, the set*

$$C := \left\{ \left(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_m d\mu \right) : \phi : S \mapsto [0,1] \right\} \subset \mathbb{R}^m$$

   is closed and convex. If $(\alpha_1, \ldots, \alpha_m)$ belongs to relative interior of set C,[1] *then there exist constants $u_1, \ldots, u_m \in \mathbb{R}$ and a test ϕ^* satisfying* (8.2) *and* (8.1) *and maximizing $\int_S \phi f_{m+1} d\mu$ subject to* (8.1)*. Moreover, it is necessary for ϕ^* to satisfy* (8.2) *μ a.s. in order to maximize $\int_S \phi f_{m+1} d\mu$ subject to* (8.1).

[1] Let C be a convex subset of $\mathbb{R}^m$ and let $x_0 \in C$. Let $L := \text{l.s.}(\{x - x_0 : x \in C\})$. Clearly, C is a convex subset of the affine space $\text{aff}(C) := x_0 + L$. The relative interior $\text{ri}(C)$ of set C is the set of all interior points of C in the usual topology of $\text{aff}(C)$. The relative boundary $\partial_{\text{ri}} C := \text{cl}(C) \setminus \text{ri}(C)$, where $\text{cl}(C)$ denotes the closure of set C.

Proof. *Claim 1.* Let $\Phi_\alpha, \alpha = (\alpha_1, \dots, \alpha_m)$ be the set of tests $\phi \in \Phi$ satisfying the constraints (8.1). Note that

$$\sup_{\phi \in \Phi_\alpha} \int_S \phi f_{m+1} d\mu \le \int_S |f_{m+1}| d\mu < \infty$$

and there exists a sequence $\{\phi_n\} \subset \Phi_\alpha$ of tests such that

$$\int_S \phi_n f_{m+1} d\mu \to \sup_{\phi \in \Phi_\alpha} \int_S \phi f_{m+1} d\mu \text{ as } n \to \infty.$$

We will use the following fact: there exists a subsequence $\{\phi_{n_k}\}$ of sequence $\{\phi_n\}$ and a test $\phi^* \in \Phi_\alpha$ such that for all $f \in L_1(\mu)$

$$\int_S \phi_{n_k} f d\mu \to \int_S \phi^* f d\mu$$

as $k \to \infty$.[2] Therefore, applying this fact to $f = f_{m+1}$, we get

$$\int_S \phi^* f_{m+1} d\mu = \sup_{\phi \in \Phi_\alpha} \int_S \phi f_{m+1} d\mu,$$

implying Claim 1 of the theorem.

Claim 2. Let $\phi^* \in \Phi_\alpha$ be a test such that (8.2) holds. Then, for any test $\phi \in \Phi_\alpha$, we have

$$\int_S (\phi^* - \phi)(f_{m+1} - \sum_{j=1}^m u_j f_j) d\mu \ge 0$$

since the expression under the integral is non-negative. This implies that

$$\int_S \phi^* f_{m+1} d\mu - \int_S \phi f_{m+1} d\mu \ge \sum_{j=1}^m u_j \Big(\int_S \phi^* f_j d\mu - \int_S \phi f_j d\mu \Big) = 0 \tag{8.4}$$

since $\phi, \phi^* \in \Phi_\alpha$. This implies Claim 2.

Claim 3. It again follows from (8.2) since $\phi^* \in \Phi_\alpha$, ϕ satisfies (8.3) and $u_j \ge 0$.

Claim 4. The proof of the last claim is more involved and it relies on some facts of convex geometry.

[2]Indeed, note that $L_\infty(\mu)$ is the dual space of $L_1(\mu)$. If $L_1(\mu)$ is separable, which is the case, for instance if σ-algebra $\mathcal{A}$ of S is countably generated, then, by a sequential version of Banach-Alaouglu Theorem, the unit ball of $L_\infty(\mu)$ is sequentially compact with respect to weak* topology. Since Φ_α is a closed subset of the unit ball of $L_\infty(\mu)$, this implies the claim in the case of countably generated σ-algebra $\mathcal{A}$. In the general case, one can replace $\mathcal{A}$ by the σ-algebra generated by sequence $\{\phi_n\}$ (which is countably generated) to complete the proof.

First note that set C is indeed convex as an image of convex set of tests $\Phi = \{\phi : \phi : S \mapsto [0,1], \phi \text{ is } \mathcal{A} - \text{measurable}\}$ under a linear mapping

$$\phi \mapsto \Big(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_m d\mu\Big) \in \mathbb{R}^m.$$

Set C is also closed: for a limit point c of this set, there exists a sequence $\{\phi_n\}$ of tests such that

$$\Big(\int_S \phi_n f_1 d\mu, \ldots, \int_S \phi_n f_m d\mu\Big) \to c \text{ as } n \to \infty.$$

Extracting from $\{\phi_n\}$ a subsequence $\{\phi_{n_k}\}$ such that $\int_S \phi_{n_k} f d\mu \to \int_S \phi d\mu$ as $k \to \infty$ for all $f \in L_1(\mu)$ and some $\phi \in \Phi$ (based on sequential compactness argument already used in the proof of Claim 1) allows us to conclude that $c = \Big(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_m d\mu\Big) \in C$.

Denote now

$$D := \Big\{\Big(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_m d\mu, \int_S \phi f_{m+1} d\mu\Big) : \phi \in \Phi\Big\} \subset \mathbb{R}^{m+1}.$$

Clearly, D is a closed convex subset of $\mathbb{R}^{m+1}$.

For $\alpha := (\alpha_1, \ldots, \alpha_m) \in C$, define

$$c^+_{m+1}(\alpha) := \sup\Big\{\int_S \phi f_{m+1} d\mu : \phi \in \Phi_\alpha\Big\}$$
$$c^-_{m+1}(\alpha) := \inf\Big\{\int_S \phi f_{m+1} d\mu : \phi \in \Phi_\alpha\Big\}.$$

Since both the supremum and the infimum in the definition of c^+_{m+1}, c^-_{m+1} are attained (for the supremum, this is Claim 1 of the theorem and for the infimum it is similar), we have

$$c^+(\alpha) := (\alpha, c^+_{m+1}(\alpha)) \in D,\ c^-(\alpha) := (\alpha, c^-_{m+1}(\alpha)) \in D.$$

Since D is convex, it also contains the segment between these two points. We will need the following simple lemma.

Lemma 8.1 *The function $C \ni \alpha \mapsto c^+_{m+1}(\alpha)$ is concave and the function $C \ni \alpha \mapsto c^-_{m+1}(\alpha)$ is convex (as a consequence, both of them are continuous). The set D can be represented as follows:*

$$D = \bigcup_{\alpha \in C} \{\alpha\} \times [c^-_{m+1}(\alpha), c^+_{m+1}(\alpha)].$$

Thus, convex set $D \subset \mathbb{R}^{m+1}$ has convex set $C \subset \mathbb{R}^m$ as its "base". The boundary of D includes the graph of concave function $c^+_{m+1}(\alpha), \alpha \in C$ (on the top) and concave function $c^-_{m+1}(\alpha), \alpha \in C$ (at the bottom).

The following lemma will be also useful.

Lemma 8.2 *If $c^+_{m+1}(\alpha) = c^-_{m+1}(\alpha)$ for some $\alpha \in \mathrm{ri}(C)$, then $c^+_{m+1}(y) = c^-_{m+1}(y)$ for all $y \in C$. In this case, $c_{m+1}(y) := c^+_{m+1}(y) = c^-_{m+1}(y), y \in C$ is a linear function $c_{m+1}(y) = \langle u, y\rangle$ for some $u \in \mathbb{R}^m$ and set D belongs to a hyperplane in $\mathbb{R}^{m+1}$ passing through the point $(0, \ldots, 0) \in \mathbb{R}^{m+1}$.*

Proof. Suppose $c^+_{m+1}(\alpha) = c^-_{m+1}(\alpha) =: c_{m+1}(\alpha)$ for some $\alpha \in \mathrm{ri}(C)$, but there is a point $\alpha' \in C$ such that $c^+_{m+1}(\alpha') > c^-_{m+1}(\alpha')$. Denote $c := (\alpha, c_{m+1}(\alpha)) \in D$, $c^+ := (\alpha', c^+_{m+1}(\alpha')) \in D$ and $c^- := (\alpha', c^+_{m+1}(\alpha')) \in D$. Since $\alpha \in \mathrm{ri}(C)$, there exists also a point $a := (\alpha'', a_{m+1}) \in D$ with $\alpha'' \in C, a_{m+1} \in \mathbb{R}$ and such that α is an interior point of the straight line segment between α' and α''. Consider the following set

$$T := \mathrm{conv}(\{c^+, c^-, a\}).$$

Clearly, T is a triangle and $T \subset D$. Note that $\alpha = \lambda\alpha' + (1-\lambda)\alpha''$ for some $\lambda \in (0, 1)$. Let

$$\tilde{c}^+ := \lambda c^+ + (1-\lambda)a,\ \tilde{c}^- := \lambda c^- + (1-\lambda)a.$$

Clearly, $\tilde{c}^+, \tilde{c}^- \in D$ and

$$\tilde{c}^+ = (\alpha, \tilde{c}^+_{m+1}), \tilde{c}^- = (\alpha, \tilde{c}^-_{m+1})$$

with

$$c^+_{m+1}(\alpha) \geq \tilde{c}^+_{m+1} > \tilde{c}^-_{m+1} \geq c^-_{m+1}(\alpha).$$

This contradicts the assumption that $c^+_{m+1}(\alpha) = c^-_{m+1}(\alpha)$. Thus, we have $c^+_{m+1}(y) = c^-_{m+1}(y)$ for all $y \in C$.

Since $c^+_{m+1}(y), y \in C$ is concave and $c^-_{m+1}(y), y \in C$ is convex, the function $c_{m+1}(y) := c^+_{m+1}(y) = c^-_{m+1}(y), y \in C$ is both concave and convex. It is well known that convexity of c_{m+1} implies that

$$c_{m+1}(y) \geq c_{m+1}(\alpha) + \langle u, y - \alpha\rangle$$

and concavity of c_{m+1} implies that

$$c_{m+1}(y) \leq c_{m+1}(\alpha) + \langle u, y - \alpha\rangle,$$

where $u \in \partial c_{m+1}(\alpha)$ is a subgradient of c_{m+1} at point α. Thus, we have

$$c_{m+1}(y) = c_{m+1}(\alpha) + \langle u, y - \alpha\rangle.$$

Since point $(0, \ldots, 0) \in D$ (take $\phi \equiv 0$), we have $0 = c_{m+1}(0) = c_{m+1}(\alpha) - \langle u, \alpha\rangle$, implying that $c_{m+1}(y) = \langle u, y\rangle$, and the remaining claims follow. ∎

Let us fix $\alpha \in \mathrm{ri}(C)$. Recall that $c^+(\alpha) = (\alpha, c^+_{m+1}) \in D, c^-(\alpha) = (\alpha, c^-_{m+1}) \in D$. Moreover, it is easy to see that $c^+(\alpha), c^-(\alpha) \in \partial_{\mathrm{ri}}(D)$. Assume first that $c^-_{m+1}(\alpha) < c^+_{m+1}(\alpha)$ implying that $c^+(\alpha) \neq c^-(\alpha)$. We will need the following geometric fact: there exists a hyperplane in the space $\mathbb{R}^{m+1}$ passing through the point $c^+(\alpha)$ and such that convex set D is located on one side of this hyperplane. Moreover, we could assume that the hyperplane does not contain points of $\mathrm{ri}(D)$. In other words, there exists a vector $w = (u, \lambda) \in \mathbb{R}^{m+1}$ with $u \in \mathbb{R}^m, \lambda \in \mathbb{R}$ such that the hyperplane is

$$H := \{x \in \mathbb{R}^{m+1} : \langle x, w\rangle = \langle c^+(\alpha), w\rangle\}$$

and

$$\begin{aligned} &\langle x, w\rangle \geq \langle c^+(\alpha), w\rangle, x \in D, \\ &\langle x, w\rangle > \langle c^+(\alpha), w\rangle, x \in \mathrm{ri}(D). \end{aligned} \tag{8.5}$$

We will now prove that $\lambda \neq 0$. Indeed, for $\lambda = 0$, we would have that all the points of the segment $\{(\alpha, x_{m+1}) : x_{m+1} \in [c^-_{m+1}(\alpha), c^+_{m+1}(\alpha)]\}$ belongs to the hyperplane H. However, one can show that (α, x_{m+1}) is a point of $\mathrm{ri}(D)$ for $x_{m+1} \in (c^-_{m+1}(\alpha), c^+_{m+1}(\alpha))$, so, it could not be a point of H. To check this, let U be a small cube in $\mathrm{aff}(C)$ (of the same dimension as $\mathrm{aff}(C)$) with center α such that $U \subset C$. Since $c^+_{m+1}(\alpha) > c^-_{m+1}(\alpha)$ and functions $c^+_{m+1}(y), c^-_{m+1}(y), y \in C$ are continuous, one can choose U small enough so that, for some numbers $\gamma_1 < x_{m+1} < \gamma_2$ we have $c^+_{m+1}(y) > \gamma_2, y \in U$ and $c^-_{m+1}(y) < \gamma_1, y \in U$. This means that the set $U \times [\gamma_1, \gamma_2] \subset D$ and (α, x_{m+1}) is an interior point of $U \times [\gamma_1, \gamma_2]$, implying $(\alpha, x_{m+1}) \in \mathrm{ri}(D)$. This also proves that $\lambda \neq 0$. By rescaling vector $w = (u, \lambda)$ if needed, we can set $\lambda = -1$ and rewrite (8.5) for $x = (y, x_{m+1})$ as follows:

$$\begin{aligned} &\langle y, u\rangle - x_{m+1} \geq \langle \alpha, u\rangle - c^+_{m+1}(\alpha), (y, x_{m+1}) \in D, \\ &\langle y, u\rangle - x_{m+1} > \langle \alpha, u\rangle - c^+_{m+1}(\alpha), (y, x_{m+1}) \in \mathrm{ri}(D). \end{aligned} \tag{8.6}$$

If ϕ^* is a test maximizing $\int_S \phi f_{m+1} d\mu$ subject to (8.1), we have $c^+_{m+1}(\alpha) = \int_S \phi^* f_{m+1} d\mu$. For

$$(y, x_{m+1}) = \Big(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_{m+1} d\mu\Big) \in D$$

with $\phi \in \Phi$ we can further rewrite (8.6) as

$$\int_S \phi^*(f_{m+1} - \sum_{j=1}^m u_j f_j) d\mu \geq \int_S \phi(f_{m+1} - \sum_{j=1}^m u_j f_j) d\mu, \phi \in \Phi.$$

Thus, ϕ^* maximizes the integral $\int_S \phi(f_{m+1} - \sum_{j=1}^m u_j f_j) d\mu$ over all tests $\phi \in \Phi$. For this, it is necessary and sufficient that ϕ^* is equal to 1 μ a.s. on the set $\{f_{m+1} > \sum_{j=1}^m u_j f_j\}$ and is equals to 0 μ a.s. on the set $\{f_{m+1} < \sum_{j=1}^m u_j f_j\}$. This completes the proof in the case when $c^-_{m+1}(\alpha) < c^+_{m+1}(\alpha)$.

In the case when $c^-_{m+1}(\alpha) = c^+_{m+1}(\alpha)$, we have, by Lemma 8.2, that, for $x = (y, x_{m+1}) \in D$, $x_{m+1} = \langle y, u\rangle$ with some $u \in \mathbb{R}^m$. For

$$(y, x_{m+1}) = \Big(\int_S \phi f_1 d\mu, \ldots, \int_S \phi f_{m+1} d\mu\Big), \phi \in \Phi,$$

this yields

$$\int_S \phi\Big(f_{m+1} - \sum_{j=1}^m u_j f_j\Big) d\mu = 0$$

for all $\phi \in \Phi$. It follows that $f_{m+1} = \sum_{j=1}^m u_j f_j$ μ a.s. and condition (8.2) is trivially satisfied μ a.s. for all tests ϕ.

■

8.1 Problems

1. Prove Lemma 8.1.
2. Suppose f_1, f_2, f_3 are densities w.r.t. μ and, for all $u_1, u_2 \in \mathbb{R}$, f_3 does not coincide with the linear combination $u_1 f_1 + u_2 f_2$ μ a.s. Then, for all $\alpha \in (0, 1)$, there exists a test $\phi \in \Phi$ such that $\int_S \phi f_1 d\mu = \int_S \phi f_2 d\mu = \alpha$ and $\int_S \phi f_3 d\mu > \alpha$.
3. Let $X \sim N(\theta, 1), \theta \in \mathbb{R}$. Find the most powerful test of size α for the null hypothesis $H_0 : \theta = -1$ or $\theta = 1$ against the alternative $\theta = 0$.

9 Applications of Generalized Neyman-Pearson Lemma: Uniformly Most Powerful Unbiased Tests in One-Parameter Exponential Families

Let $X \sim p_\theta : \theta \in \Theta \subset \mathbb{R}$ be an observation in space $(S, \mathcal{A})$ with density p_θ w.r.t. measure μ given by the following expression:

$$p_\theta(x) := \frac{1}{Z(\theta)} \exp\{\theta T(x)\} h(x), x \in S,$$

where $T : S \mapsto \mathbb{R}, h : S \mapsto \mathbb{R}_+$ are measurable functions and

$$Z(\theta) := \int_S \exp\{\theta T(x)\} h(x) \mu(dx) < \infty, \theta \in \Theta.$$

Note that $Z(\theta)$ is a convex function and the set $\{\theta : Z(\theta) < \infty\}$ is an interval of $\mathbb{R}$. Thus, without loss of generality, we could assume that Θ is an open interval. We will also assume that $T(x)$ is not a constant (to avoid the trivial case). Such a model is called

a one-parameter exponential family with sufficient statistic $T(X)$. Note that this model is a monotone likelihood ratio family w.r.t. to $T(X)$. Thus, there exist UMP tests of size α for such hypotheses as $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$, or $H_0 : \theta \geq \theta_0$ against $H_a : \theta < \theta_0$. These tests are given by Karlin-Rubin Theorem.

However, *there are no* UMP tests of size α for the hypothesis $H_0 : \theta \in [\theta_1, \theta_2]$ against *two sided* alternative $H_a : \theta < \theta_1$ or $\theta > \theta_2$, where $\theta_1, \theta_2 \in \Theta, \theta_1 < \theta_2$. This is due to the fact that such a test must be UMP for $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_a : \theta > \theta_2$. By Karlin-Rubin Theorem (and its proof), its power function would be strictly increasing on the whole set Θ. On the other hand, it must be also UMP for $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_a : \theta < \theta_1$, and, again by Karlin-Rubin Theorem, its power function would be strictly decreasing. This contradiction shows that such tests do not exist.

For such hypotheses with two sided alternatives, there is another notion of optimality in testing, *uniformly most powerful unbiased* tests.

Definition 9.1 *Let $X \sim p_\theta, \theta \in \Theta$ and let $\{\Theta_0, \Theta_1\}$ be a partition of parameter space Θ into two disjoint nonempty subsets. Consider hypotheses testing problem $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$. A test $\phi : S \mapsto [0, 1]$ is called an unbiased test of size α iff $\beta_\phi(\theta) \leq \alpha, \theta \in \Theta_0$ and $\beta_\phi(\theta) \geq \alpha, \theta \in \Theta_1$.*

Definition 9.2 *A test $\phi^* : S \mapsto [0, 1]$ is called a uniformly most powerful unbiased (UMPU) test of size α for H_0 against H_a iff it is unbiased of size α and, for any unbiased test ϕ of size α, $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta), \theta \in \Theta_1$.*

Clearly, if there exists a UMP test of size α for H_0 against H_a, then it is UMPU test of size α. But UMPU tests could exist also when UMP tests do not.

We will show the following proposition.

Proposition 9.1 *Let $X \sim p_\theta, \theta \in \Theta$, where $\Theta \subset \mathbb{R}$ is an open interval and $\{p_\theta : \theta \in \Theta\}$ is a one-parameter exponential family. Let $\alpha \in (0, 1)$. Then, there exists a UMPU test of size α for the hypothesis $H_0 : \theta \in [\theta_1, \theta_2]$ against the alternative $H_a : \theta \notin [\theta_1, \theta_2]$, where $\theta_1 < \theta_2, \theta_1, \theta_2 \in \Theta$.*

Proof. It is easy to see that, in the case of exponential family, the power function $\beta_\phi(\theta), \theta \in \Theta$ is continuous for all tests ϕ. Therefore, if ϕ is an unbiased test for H_0 against H_a, we must have $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$. Let $\theta' > \theta_2$ and consider the following optimization problem:

$$\begin{cases} \beta_\phi(\theta') = \int_S \phi p_{\theta'} d\mu \to \max \\ \beta_\phi(\theta_1) = \int_S \phi p_{\theta_1} d\mu = \alpha, \\ \beta_\phi(\theta_2) = \int_S \phi p_{\theta_2} d\mu = \alpha, \\ \phi \in \Phi. \end{cases} \tag{9.1}$$

We will use the generalized Neyman-Pearson Lemma to show that this problem has a solution ϕ^* of the following form:

$$\phi^*(x) := \begin{cases} 1 & \text{if } p_{\theta'}(x) > u_1 p_{\theta_1}(x) + u_2 p_{\theta_2}(x) \\ 0 & \text{if } p_{\theta'}(x) < u_1 p_{\theta_1}(x) + u_2 p_{\theta_2}(x) \end{cases}$$

for some constants u_1, u_2. To this end, we need to check that (α, α) is an interior point of the convex set

$$C := \Big\{ \Big(\int_S \phi p_{\theta_1} d\mu, \int_S \phi p_{\theta_2} d\mu \Big) : \phi \in \Phi \Big\}.$$

Note that since $p_{\theta_1} \neq p_{\theta_2}$, the power $\beta_{\max}(\alpha)$ of the most powerful test of size α for the hypothesis $\theta = \theta_1$ against the alternative $\theta = \theta_2$ is strictly larger that α and the power $\beta_{\min}(\alpha)$ of the least powerful test of size α is strictly smaller that α. Thus, we have $\alpha \in (\beta_{\min}(\alpha), \beta_{\max}(\alpha))$ for all $\alpha \in (0, 1)$. Since also $\beta_{\max}(\alpha)$ is a concave function on $(0, 1)$ and $\beta_{\min}(\alpha)$ is convex, these two functions are continuous and it is easy to conclude that (α, α) is an interior point of C (see the proof of generalized Neyman-Pearson Lemma for a similar argument). Thus, constants u_1, u_2 for which ϕ^* solves problem (9.1) do exist.

We can assume without loss of generality that $h(x) > 0$ for all $x \in S$ (otherwise, one can replace S with the set $S' := \{x : h(x) > 0\}$). Then, the condition $p_{\theta'}(x) > u_1 p_{\theta_1}(x) + u_2 p_{\theta_2}(x)$ is equivalent to the condition

$$\frac{1}{Z(\theta')} e^{\theta' T(x)} > \frac{u_1}{Z(\theta_1)} e^{\theta_1 T(x)} + \frac{u_2}{Z(\theta_2)} e^{\theta_2 T(x)}. \tag{9.2}$$

Note that, if $u_1 \leq 0, u_2 \leq 0$, then condition (9.2) holds for all x, which means that the test ϕ^* does not satisfy the constraints of problem (9.1). If $u_1 > 0$ and $u_2 \leq 0$, we can rewrite (9.2) as

$$\frac{-u_2}{Z(\theta_2)} e^{(\theta_2 - \theta_1) T(x)} + \frac{1}{Z(\theta')} e^{(\theta' - \theta_1) T(x)} > \frac{u_1}{Z(\theta_1)}.$$

Since we have an increasing function of $T(x)$ in the left hand side of the last equation, the rejection region of test ϕ^* would be of the form $\{x : T(x) > c\}$. However, we know (from Karlin-Rubin Theorem) that the power of such a test in our MLR family is a strictly increasing function of θ and such a test could not be a solution of problem (9.1). Similarly, one could rule out the cases $u_1 \geq 0$ and $u_2 \geq 0$. Thus, we must have $u_1 < 0, u_2 > 0$. In this case, it is again easy to check that the rejection region of the test ϕ^* is of the form $\{x : T(x) \notin [c_1, c_2]\}$ for some constants $c_1 < c_2$. In other words, ϕ^* must satisfy

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [c_1, c_2] \\ 0 & \text{if } T(x) \in (c_1, c_2) \end{cases}$$

and there is a choice of constants c_1, c_2 for which such a test solves problem (9.1). Moreover, test ϕ^* solves problem (9.1) simultaneously for all $\theta' > \theta_2$. It is also easy to check by a similar analysis that the following two claims hold:

1. test ϕ^* solves problem (9.1) for all $\theta' < \theta_1$;

2. test ϕ^* *minimizes* the power $\beta_\phi(\theta')$ subject to constraints $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$ for all $\theta' \in (\theta_1, \theta_2)$.

It immediately follows from these facts that $\beta_{\phi^*}(\theta) \leq \alpha$ for all $\theta \in [\theta_1, \theta_2]$ and $\beta_{\phi^*}(\theta) \geq \alpha$ for all $\theta \notin [\theta_1, \theta_2]$. Thus, ϕ^* is an unbiased test of size α for H_0 against H_a, and it maximizes the power $\beta_\phi(\theta)$ for all $\theta \notin [\theta_1, \theta_2]$ among all the unbiased tests of size α (since all of them satisfy the constraints $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$). This shows that ϕ^* is UMPU test of size α. ■

Remark. It could be proved (and it is sometimes done in the literature) that you could find test of the form

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [c_1, c_2] \\ \gamma_1 & \text{if } T(x) = c_1 \\ \gamma_2 & \text{if } T(x) = c_2 \\ 0 & \text{if } T(x) \in (c_1, c_2) \end{cases}$$

for some $c_1 < c_2, \gamma_1, \gamma_2 \in [0, 1]$ that satisfies the constraints $\beta_{\phi^*}(\theta_1) = \beta_{\phi^*}(\theta_2) = \alpha$ and is UMPU of size α.

Another common testing problem for one-parameter exponential family $\{p_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$ is to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta \neq \theta_0$ for a given $\theta_0 \in \Theta$. As in the previous example, UMP tests of any size $\alpha \in (0, 1)$ do not exist for this problem, but it is possible to develop UMPU test of size α. Note that this problem could be viewed as the limit case of the previous one when $\theta_1 \uparrow \theta_0$ and $\theta_2 \downarrow \theta_0$. The size constraint $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$ becomes in the limit $\beta_\phi(\theta_0) = \alpha$ and $\beta'_\phi(\theta_0) = 0$.

Note also that if ϕ is an unbiased test of size α for $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$, then $\beta_\phi(\theta_0) = \alpha$ and the power function $\beta_\phi(\theta), \theta \in \Theta$ has a minimum at point θ_0. Therefore, we must have $\beta'_\phi(\theta_0) = 0$ (provided that β_ϕ is differentiable at θ_0).

It is not hard to check that, in the case of one-parameter exponential family, the power function $\beta_\phi(\theta), \theta \in \Theta$ is differentiable for all tests $\phi \in \Phi$. Moreover, the following simple proposition holds:

Proposition 9.2 *For all* $\theta \in \Theta$,

$$\beta'_\phi(\theta) = \mathrm{cov}_\theta(\phi(X), T(X)).$$

Thus, the condition $\beta'_\phi(\theta_0) = 0$ means that $\phi(X)$ and $T(X)$ are uncorrelated random variables for $X \sim p_{\theta_0}$. Since

$$\mathrm{cov}_{\theta_0}(\phi(X), T(X)) = \mathbb{E}_{\theta_0}\phi(X)T(X) - \mathbb{E}_{\theta_0}\phi(X)\mathbb{E}_{\theta_0}T(X)$$

and we also have $\beta_\phi(\theta_0) = \alpha$, the condition $\beta'_\phi(\theta_0) = 0$ becomes

$$\mathbb{E}_{\theta_0}\phi(X)T(X) = \alpha\mathbb{E}_{\theta_0}T(X).$$

or

$$\int_S \phi T p_{\theta_0} d\mu = \alpha \int_S T p_{\theta_0} d\mu.$$

Thus, we can try to find UMPU tests of size α by solving the following optimization problem: for $\theta \neq \theta_0$,

$$\begin{cases} \int_S \phi p_\theta d\mu \to \max \\ \int_S \phi p_{\theta_0} d\mu = \alpha \\ \int_S \phi T p_{\theta_0} d\mu = \alpha \int_S T p_{\theta_0} d\mu \\ \phi \in \Phi. \end{cases} \tag{9.3}$$

The analysis of this optimization problem is based on the Generalized Neyman-Pearson Lemma and is similar to what we did in the proof of Proposition 9.1. It turns out that it results in the same test

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [c_1, c_2] \\ 0 & \text{if } T(x) \in (c_1, c_2) \end{cases}$$

that for some $c_1 < c_2$ satisfies the constraints and solves problem (9.3). Moreover, this test does not depend on the choice of the alternative $\theta \neq \theta_0$ and it is unbiased. Hence, it is UMPU test of size α.

It would be a very good exercise to try to do all the remaining details of the proof on your own. The following fact could be useful to show that one-sided tests do not satisfy the constraints of the problem.

Proposition 9.3 *Let T be a random variable in $\mathbb{R}$ and let ψ be a monotone function in $\mathbb{R}$. Then $\operatorname{cov}(\psi(T), T) \geq 0$ if ψ is non-decreasing and $\operatorname{cov}(\psi(T), T) \leq 0$ if ψ is non-increasing. Moreover, these inequalities are strict unless r.v. $\psi(T)$ is constant a.s.*

Note that one-sided tests based on statistic $T(X)$ could be viewed as $\psi(T), T = T(X)$ for such functions ψ as

$$\psi(T) := \begin{cases} 1 & \text{for } T > c \\ \gamma & \text{for } T = c \\ 0 & \text{for } T < c \end{cases} \quad \text{or} \quad \psi(T) := \begin{cases} 1 & \text{for } T < c \\ \gamma & \text{for } T = c \\ 0 & \text{for } T > c, \end{cases}$$

which allows us to use the above proposition to check that $\operatorname{cov}(\psi(T), T) \neq 0$ for one-sided tests.

The conclusion is that the following proposition holds:

Proposition 9.4 *Let $X \sim p_\theta, \theta \in \Theta$, where $\Theta \subset \mathbb{R}$ is an open interval and $\{p_\theta : \theta \in \Theta\}$ is a one-parameter exponential family. Let $\alpha \in (0,1)$. Then, there exists a UMPU test of size α for the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta \neq \theta_0$, where $\theta_0 \in \Theta$.*

Note that UMPU test of size α for $H_0 : \theta = \theta_0$ $H_a : \theta \neq \theta_0$ could be written in the form

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [c_1, c_2] \\ \gamma_1 & \text{if } T(x) = c_1 \\ \gamma_2 & \text{if } T(x) = c_2 \\ 0 & \text{if } T(x) \in (c_1, c_2) \end{cases}$$

for some $c_1 < c_2, \gamma_1, \gamma_2 \in [0,1]$ such that the constraints of problem (9.3) are satisfied.

9.1 Problems

1. Check claims 1, 2 in the proof of Proposition 9.1.
2. Prove Proposition 9.2.
3. Prove Proposition 9.3.
4. Try to fill out the details of the proof of Proposition 9.4. If it does not work, read it in Lehmann's book.
5. Use the Generalized Neyman-Pearson Lemma to show that there exists a UMP test of size α for the hypothesis $\theta \notin (\theta_1, \theta_2)$ against the alternative $\theta \in (\theta_1, \theta_2)$ (in the case of one-parameter exponential family).

10 Likelihood Ratio Tests

Let $X_1, \ldots, X_n$ be i.i.d. observations sampled from the distribution $P_\theta, \theta \in \Theta$ in $(S, \mathcal{A})$, where Θ is a parameter space. Suppose that μ is a measure on $(S, \mathcal{A})$ (finite or σ-finite) and, for all $\theta \in \Theta$, P_θ is absolutely continuous with respect to μ with density p_θ. The likelihood function for this model is defined as

$$L_n(\vartheta) := L_n(\vartheta; X_1, \ldots, X_n) := \prod_{j=1}^{n} p_\vartheta(X_j), \vartheta \in \Theta.$$

Maximum likelihood estimators (MLE) are the maximizers of the likelihood function:

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) \in \text{Argmin}_{\vartheta \in \Theta} L_n(\vartheta).$$

In what follows, it will be assumed that maximum likelihood estimators exist.

Let Θ_0, Θ_1 be a partition of Θ :

$$\Theta_0, \Theta_1 \subset \Theta, \Theta_0, \Theta_1 \neq \emptyset, \Theta_0 \cup \Theta_1 = \Theta, \Theta_0 \cap \Theta_1 = \emptyset.$$

Consider the following testing problem:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_a : \theta \in \Theta_1. \end{cases}$$

The likelihood ratio (LR) statistics for this problem is defined as

$$\frac{\sup_{\vartheta \in \Theta_1} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)}.$$

If $\Theta := \{\theta_0, \theta_1\}$ and $\Theta_0 := \{\theta_0\}, \Theta_1 := \{\theta_1\}$, our testing problem becomes a simple hypotheses testing. In this case,

$$\frac{\sup_{\vartheta \in \Theta_1} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)} = \frac{p_{\theta_1}(X_1) \dots p_{\theta_1}(X_n)}{p_{\theta_0}(X_1) \dots p_{\theta_0}(X_n)}$$

is the usual test statistic of Neyman-Pearson test for $H_0 : \theta = \theta_0$ against $H_a : \theta = \theta_1$. In general, if

$$\hat{\theta}_0 \in \operatorname{Argmin}_{\vartheta \in \Theta_0} L_n(\vartheta), \ \hat{\theta}_1 \in \operatorname{Argmin}_{\vartheta \in \Theta_1} L_n(\vartheta)$$

are MLEs under H_0 and under H_1, then

$$\frac{\sup_{\vartheta \in \Theta_1} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)} = \frac{L_n(\hat{\theta}_1)}{L_n(\hat{\theta}_0)} = \frac{p_{\hat{\theta}_1}(X_1) \dots p_{\hat{\theta}_1}(X_n)}{p_{\hat{\theta}_0}(X_1) \dots p_{\hat{\theta}_0}(X_n)}$$

could be viewed as an estimator of this statistic.

In what follows, it will be convenient to use log-likelihood ratio (log-LR) statistic instead of LR-statistic:

$$\tilde{\Lambda}_n := \log \frac{\sup_{\vartheta \in \Theta_1} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)} = \log \frac{L_n(\hat{\theta}_1)}{L_n(\hat{\theta}_0)}.$$

The likelihood ratio test rejects H_0 for sufficiently large values of $\tilde{\Lambda}_n$. Moreover, often $\tilde{\Lambda}_n$ is replaced by its modified version

$$\Lambda_n := 2 \log \frac{\sup_{\vartheta \in \Theta} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)} = 2(\tilde{\Lambda}_n \vee 0).$$

Usually, the threshold for which $\tilde{\Lambda}_n$ is large is a positive number and "$\tilde{\Lambda}_n$ is large" is equivalent to "Λ_n is large". Note that if $\hat{\theta}$ is the MLE for the whole model, then

$$\Lambda_n = 2 \log \frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)}.$$

A difficulty with implementing LR-test is related to the fact that we need to find a critical value (a threshold) c for statistic Λ_n such that the test is of size α, that is

$$\mathbb{P}_\theta\{\Lambda_n \geq c\} \leq \alpha, \theta \in \Theta_0.$$

This would be hard to accomplish if the above probability depended on $\theta \in \Theta_0$. A surprising property of LR-tests is that for regular statistical models and for reasonable null hypotheses, the asymptotic distribution of log likelihood statistic Λ_n as $n \to \infty$ does not depend on $\theta \in \Theta_0$, which allows one to overcome the above difficulty.

10.1 Example: Gaussian shift model

Let $X \sim N(\theta, I_d), \theta \in \Theta$, $\Theta \subset \mathbb{R}^d$. In other words, $X = \theta + Z, \theta \in \Theta \subset \mathbb{R}^d$ with $Z \sim N(0, I_d)$ being the standard normal noise. This model will be called a Gaussian shift model. The likelihood function is given by

$$L(\vartheta) = L(\vartheta; X) := \frac{1}{(2\pi)^{d/2}} \exp\Big\{-\frac{1}{2}|X - \vartheta|^2\Big\}, \theta \in \Theta$$

and the log likelihood function is

$$\log L(\vartheta) := -\frac{d}{2}\log(2\pi) - \frac{1}{2}|X - \vartheta|^2.$$

Clearly, the MLE $\hat{\theta}$ coincides with the "projection" $P_\Theta X$ of X onto Θ :

$$\hat{\theta} = \hat{\theta}(X) \in \mathrm{Argmax}_{\vartheta \in \Theta} L(\vartheta; X) = \mathrm{Argmin}_{\vartheta \in \Theta} |X - \vartheta|^2 =: P_\Theta X.$$

Note that $P_\Theta X$ exists and is unique if Θ is a closed convex set (in particular, if $\Theta = L$ is a subspace of $\mathbb{R}^d$). We also have

$$\log L(\hat{\theta}) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}|X - \hat{\theta}|^2 = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\mathrm{dist}^2(X, \Theta).$$

Similarly, if $\Theta_0 \subset \Theta$ and $\hat{\theta}_0$ is the MLE under H_0, we have

$$\log L(\hat{\theta}_0) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}|X - \hat{\theta}_0|^2 = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\mathrm{dist}^2(X, \Theta_0).$$

Therefore,

$$\Lambda = 2\log\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} = \mathrm{dist}^2(X, \Theta_0) - \mathrm{dist}^2(X, \Theta).$$

Consider now the case when $\Theta := \mathbb{R}^d$ and $\Theta_0 := L \subset \mathbb{R}^d$ is a subspace of $\mathbb{R}^d$ and $\dim(L) < d$. In this case,

$$\Lambda = \mathrm{dist}^2(X, L) = |P_{L^\perp} X|^2.$$

Under the null hypothesis $\theta \in L$, we have

$$\Lambda = |P_{L^\perp} X|^2 = |P_{L^\perp}(\theta + Z)|^2 = |P_{L^\perp} Z|^2 \sim \chi^2_{d-\dim(L)}$$

since $P_{L^\perp}\theta = 0, \theta \in L$.

This proves the following proposition:

Proposition 10.1 *Let $X \sim N(\theta, I_d), \theta \in \mathbb{R}^d$ and let $L \subset \mathbb{R}^d$ be a subspace of $\dim(L) < d$. Consider the following testing problem:*

$$\begin{cases} H_0 : \theta \in L \\ H_a : \theta \notin L. \end{cases}$$

The log likelihood ratio statistic Λ has the distribution $\chi^2_{d-\dim(L)}$ for $X \sim N(\theta, I_d)$ and for all $\theta \in L$.

Given $\alpha \in (0,1)$, we can choose c such that $\mathbb{P}\{\chi^2_{d-\dim(L)} \geq c\} = \alpha$. Then, the LR-test that rejects H_0 when $\Lambda \geq c$ is of size α.

10.2 Example: Multinomial model

Let $X = (X_1, \ldots, X_k) \sim \mathcal{M}(n; \theta_1, \ldots, \theta_k), \theta_j \geq 0, \sum_{j=1}^k \theta_j = 1$ be a multinomial r.v. with the number of trials n and unknown probabilities of outcomes $\theta_1, \ldots, \theta_k$.[3] The parameter space Θ is the simplex of all probability distributions on $\{1, \ldots, k\}$:

$$\Theta := \Big\{(\theta_1, \ldots, \theta_k) : \theta_j \geq 0, j = 1, \ldots, k, \sum_{j=1}^k \theta_j = 1\Big\}.$$

The likelihood function is

$$L_n(\theta) = L_n(\theta; X) := \frac{n!}{X_1! \ldots X_k!} \theta_1^{X_1} \ldots \theta_k^{X_k}.$$

It is easy to find the MLE for this model:

$$\begin{aligned}
\hat{\theta} &:= \operatorname{argmax}_{\vartheta \in \Theta} L_n(\vartheta) = \operatorname{argmax}_{\vartheta \in \Theta} \log L_n(\vartheta) = \\
&= \operatorname{argmax}_{\vartheta \in \Theta} \Big[\log \frac{n!}{X_1! \ldots X_k!} + X_1 \log \vartheta_1 + \cdots + X_k \log \vartheta_k\Big] \\
&= \operatorname{argmax}_{\vartheta \in \Theta} \Big[\frac{X_1}{n} \log \vartheta_1 + \cdots + \frac{X_k}{n} \log \vartheta_k\Big] \\
&= \operatorname{argmim}_{\vartheta \in \Theta} \Big[-\frac{X_1}{n} \log \vartheta_1 - \cdots - \frac{X_k}{n} \log \vartheta_k\Big] \\
&= \operatorname{argmim}_{\vartheta \in \Theta} \Big[\sum_{j=1}^k \frac{X_j}{n} \log \frac{X_j/n}{\vartheta_j}\Big].
\end{aligned}$$

[3] Recall the definition of multinomial r.v. Suppose n independent trials are performed with k possible outcomes $1, \ldots, k$ of each trial. Probabilities of these outcomes are $\theta_1, \ldots, \theta_k$ and X_j is the number of trials resulting in outcome $j, j = 1, \ldots, k$. Then $X = (X_1, \ldots, X_k) \sim \mathcal{M}(n; \theta_1, \ldots, \theta_k)$.

Denote by

$$K(\theta'||\theta'') := \sum_{j=1}^{k} \theta'_j \log \frac{\theta'_j}{\theta''_j}$$

the Kullback-Leibler (KL) divergence between distributions $\theta', \theta'' \in \Theta$. Then, we can write

$$\hat{\theta} = \text{argmin}_{\vartheta \in \Theta} K(\tilde{\theta}||\vartheta),$$

where $\tilde{\theta} := \left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right) \in \Theta$ is the vector of frequencies. Using Jensen's inequality, it is easy to check that

$$K(\tilde{\theta}||\tilde{\theta}) = 0 \leq K(\tilde{\theta}||\vartheta), \vartheta \in \Theta, \quad K(\tilde{\theta}||\vartheta) > 0, \vartheta \neq \tilde{\theta},$$

implying that $\hat{\theta} = \tilde{\theta}$ is the unique MLE.

If $\Theta_0 \subset \Theta$ and it is known that $\theta \in \Theta_0$, then the MLE under hypothesis $\theta \in \Theta_0$ is

$$\hat{\theta}_0 := \text{argmax}_{\vartheta \in \Theta_0} L_n(\vartheta) = \text{argmin}_{\vartheta \in \Theta_0} K(\hat{\theta}||\vartheta).$$

Consider the following testing problem:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_a : \theta \notin \Theta_0 \end{cases}$$

for a subset $\Theta_0 \subset \Theta$. Then, the log LR-statistic for this problem could be written in the form

$$\Lambda_n := 2 \log \frac{\sup_{\vartheta \in \Theta} L_n(\vartheta)}{\sup_{\vartheta \in \Theta_0} L_n(\vartheta)} = 2 \log \frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)} = 2nK(\hat{\theta}||\hat{\theta}_0) = 2n \inf_{\vartheta \in \Theta_0} K(\hat{\theta}||\vartheta).$$

We will study the asymptotic distribution of Λ_n as $n \to \infty$ in the case when $\Theta := \{\theta^{(0)}\}$ and we want to test the hypotheses $H_0 : \theta = \theta^{(0)}$ against the alternative $H_a : \theta \neq \theta^{(0)}$ for some $\theta^{(0)} \in \Theta$ with $\theta_j^{(0)} > 0, j = 1, \ldots, k$. In this case,

$$\Lambda_n = 2nK(\hat{\theta}||\theta^{(0)}) = 2n \sum_{j=1}^{k} \hat{\theta}_j \log \frac{\hat{\theta}_j}{\theta_j^{(0)}} = 2n \sum_{j=1}^{k} \hat{\theta}_j \log\Big(1 + \frac{\hat{\theta}_j - \theta_j^{(0)}}{\theta_j^{(0)}}\Big).$$

Using Taylor expansion, write

$$\log(1+x) = x - \frac{x^2}{2} + x^2 r(x),$$

where $r(x) \to 0$ as $x \to 0$. Since, under the null hypothesis H_0,

$$\hat{\theta}_j - \theta_j^{(0)} = O_{\mathbb{P}}(n^{-1/2}),$$

we get

$$\log\Big(1 + \frac{\hat{\theta}_j - \theta_j^{(0)}}{\theta_j^{(0)}}\Big) = \frac{\hat{\theta}_j - \theta_j^{(0)}}{\theta_j^{(0)}} - \frac{1}{2}\frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{(\theta_j^{(0)})^2} + o_{\mathbb{P}}(n^{-1})$$

as $n \to \infty$. Therefore, using the facts that $\sum_{j=1}^k \hat{\theta}_j = \sum_{j=1}^k \theta_j^{(0)} = 1$ and $(\hat{\theta}_j - \theta_j^{(0)})^3 = O_{\mathbb{P}}(n^{-3/2})$, we get

$$\begin{aligned}
\Lambda_n &= 2n\sum_{j=1}^k \hat{\theta}_j \frac{\hat{\theta}_j - \theta_j^{(0)}}{\theta_j^{(0)}} - n\sum_{j=1}^k \hat{\theta}_j \frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{(\theta_j^{(0)})^2} + 2no_{\mathbb{P}}(n^{-1}) \\
&= 2n\sum_{j=1}^k \theta_j^{(0)} \frac{\hat{\theta}_j - \theta_j^{(0)}}{\theta_j^{(0)}} + 2n\sum_{j=1}^k \frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{\theta_j^{(0)}} - n\sum_{j=1}^k \theta_j^{(0)} \frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{(\theta_j^{(0)})^2} - n\sum_{j=1}^k \frac{(\hat{\theta}_j - \theta_j^{(0)})^3}{(\theta_j^{(0)})^2} + o_{\mathbb{P}}(1) \\
&= n\sum_{j=1}^k \frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{\theta_j^{(0)}} + o_{\mathbb{P}}(1).
\end{aligned}$$

The statistic

$$\hat{\chi}^2(n) := n\sum_{j=1}^k \frac{(\hat{\theta}_j - \theta_j^{(0)})^2}{\theta_j^{(0)}}$$

is the well known Pearson's chi-square statistic. Its asymptotic distribution as $n \to \infty$ is chi-square with $k-1$ degrees of freedom. We will sketch the proof of this result. Consider random vectors

$$Y_n := \Big(\frac{\sqrt{n}(\hat{\theta}_j - \theta_j^{(0)})}{\sqrt{\theta_j^{(0)}}} : j = 1, \dots, k\Big).$$

By the Central Limit Theorem, Y_n converges in distribution as $n \to \infty$ to a normal vector Y with mean zero and covariance $\Sigma = (\sigma_{ij})_{i,j=1,\dots,k}$, where

$$\sigma_{ij} := \begin{cases} 1 - \theta_i^{(0)} & i = j \\ -\sqrt{\theta_i^{(0)}}\sqrt{\theta_j^{(0)}} & i \neq j. \end{cases}$$

Denoting $v = \Big(\sqrt{\theta_i^{(0)}} : i = 1, \dots, k\Big)$, we get $\Sigma = I_k - P_v$, where $P_v = vv^T$ is the orthogonal projection onto the one dimensional space spanned on v. Therefore, Y has the same distribution as $Z - \langle Z, v\rangle v, Z \sim N(0, I_k)$. This implies that

$$\hat{\chi}^2(n) = |Y_n|^2 \xrightarrow{d} |Z - \langle Z, v\rangle v|^2 \sim \chi^2_{k-1}.$$

Thus, we proved the following statement:

Proposition 10.2 *Let Λ_n be log-LR statistic for hypothesis $H_0 : \theta = \theta^{(0)}$ against the alternative $H_a : \theta \neq \theta^{(0)}$ for some $\theta^{(0)} \in \Theta$ with $\theta_j^{(0)} > 0, j = 1, \ldots, k$. Then*

$$\Lambda_n \xrightarrow{d} \chi^2_{k-1} \text{ as } n \to \infty.$$

In the next sections, we will show that chi-square limit distribution of log LR-statistic Λ_n under null hypothesis is a very common phenomenon for regular statistical models.

10.3 Maximum Likelihood Estimators: Asymptotic Normality

Recall that the maximum likelihood estimators (MLE) could be equivalently defined as maximizers of the log-likelihood function:

$$\hat{\theta}_n \in \text{Argmax}_{\vartheta \in \Theta} \log L_n(\vartheta) = \text{Argmax}_{\vartheta \in \Theta} n^{-1} \sum_{j=1}^{n} \log p_\vartheta(X_j).$$

If $\mathbb{E}_\theta |\log p_\vartheta(X)| < \infty, \vartheta, \theta \in \Theta$, then, by the Strong Law of Large Numbers, for all $\vartheta \in \Theta$,

$$n^{-1} \sum_{j=1}^{n} \log p_\vartheta(X_j) \to \mathbb{E}_\theta \log p_\vartheta(X) \text{ as } n \to \infty \text{ a.s.}$$

provided that $X_1, \ldots, X_n, \ldots$ are i.i.d. $\sim p_\theta, \theta \in \Theta$.

Note also that, if the model $\{p_\theta : \theta \in \Theta\}$ is identifiable, then the function $\Theta \ni \vartheta \mapsto \mathbb{E}_\theta \log p_\vartheta(X)$ has unique maximum at the point $\vartheta = \theta$. Indeed, for two densities p, q with respect to μ, the Kullbac-Leibler divergence between p and q is defined as

$$K(p||q) := \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

The following proposition is well known:

Proposition 10.3 *For all densities p, q,*

$$K(p||q) \geq H^2(p, q).$$

Proof. We will use the following elementary bound:

$$\log x \leq 2(\sqrt{x} - 1), x \geq 0.$$

It implies that

$$\log \frac{q(X)}{p(X)} \leq 2\Big(\sqrt{\frac{q(X)}{p(X)}} - 1\Big).$$

Therefore,

$$-K(p||q) = \mathbb{E}_p \log \frac{q(X)}{p(X)} \le 2\Big(\mathbb{E}_p \sqrt{\frac{q(X)}{p(X)}} - 1\Big)$$
$$= 2\Big(\int_S \sqrt{p(x)}\sqrt{q(x)}\mu(dx) - 1\Big) = -H^2(p,q),$$

implying the result. ∎

We can write

$$\mathbb{E}_\theta \log p_\theta(X) - \mathbb{E}_\theta \log p_\vartheta(X) = K(p_\theta||p_\vartheta) \ge H^2(p_\theta, p_\vartheta).$$

If the model is identifiable, $H^2(p_\theta, p_\vartheta) > 0$ for all $\vartheta \ne \theta$, implying that θ is the unique maximal point of the function $\Theta \ni \vartheta \mapsto \mathbb{E}_\theta \log p_\vartheta(X)$.

Since $n^{-1}\sum_{j=1}^n \log p_\vartheta(X_j) \approx \mathbb{E}_\theta \log p_\vartheta(X)$ for large n, one could expect and it could be proved that the MLE $\hat\theta_n$ converges to θ as $n \to \infty$ (in probability and a.s.) under some regularity of the model (consistency of MLE).

Another important property of maximum likelihood estimators is asymptotic normality. We will describe below in some detail the regularity assumptions needed for this property.

10.3.1 Quadratic mean differentiability

Let $X \sim p_\theta, \theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^d$ and p_θ are densities w.r.t. μ on $(S, \mathcal{A})$.

Definition 10.1 *It will be said that $\{p_\vartheta : \vartheta \in \Theta\}$ is quadratic mean differentiable (QMD) at point $\theta \in \Theta$ iff there exists a measurable function $\psi_\theta : S \mapsto \mathbb{R}^d$ such that $\int_S |\psi_\theta(x)|^2 \mu(dx) < \infty$ and*

$$\sqrt{p_{\theta+h}} - \sqrt{p_\theta} = \langle \psi_\theta, h\rangle + o(|h|) \text{ as } |h| \to 0$$

in $L_2(\mu)$.

In other words, quadratic mean differentiability means that

$$\Big\| \sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \langle \psi_\theta, h\rangle \Big\|_{L_2(\mu)} = o(|h|) \text{ as } |h| \to 0.$$

For $d = 1$, it is equivalent to

$$\lim_{h\to 0} \frac{\sqrt{p_{\theta+h}} - \sqrt{p_\theta}}{h} = \psi_\theta \text{ in } L_2(\mu).$$

Informally, we can write

$$\frac{\partial}{\partial\theta}\sqrt{p_\theta} = \psi_\theta,$$

and, by the Chain Rule,

$$\frac{\partial}{\partial\theta}p_\theta = 2\sqrt{p_\theta}\psi_\theta,$$
$$\frac{\partial}{\partial\theta}\log p_\theta = \frac{2\psi_\theta}{\sqrt{p_\theta}}.$$

Recall that $\frac{\partial}{\partial\theta}\log p_\theta(X)$ is the score function of our model. With these conventions, it is easy to check that QMD implies that

$$\mathbb{E}_\theta\frac{\partial}{\partial\theta}\log p_\theta(X) = 0.$$

Recall also that the Fisher information for regular statistical model is usually defined as

$$I(\theta) := \mathbb{E}_\theta\frac{\partial}{\partial\theta}\log p_\theta(X)\Big(\frac{\partial}{\partial\theta}\log p_\theta(X)\Big)^T = \operatorname{cov}\Big(\frac{\partial}{\partial\theta}\log p_\theta(X)\Big)$$

(the covariance matrix of the score function). This leads to the following definition.

Definition 10.2 *Under quadratic mean differentiability, the Fisher information matrix is defined as*

$$I(\theta) := 4\int_S \psi_\theta\psi_\theta^T\,d\mu.$$

10.3.2 Local Asymptotic Normality (LAN)

Let $X_1, \dots, X_n$ be i.i.d. $\sim p_\theta$, where $\theta \in \Theta$, $\Theta \subset \mathbb{R}^d$ is an open set. Define

$$Z_n(\theta_0; u) := \log\frac{\prod_{j=1}^n p_{\theta+n^{-1/2}u}(X_j)}{\prod_{j=1}^n p_\theta(X_j)} = \sum_{j=1}^n\Big[\log p_{\theta+n^{-1/2}u}(X_j) - \log p_\theta(X_j)\Big],$$

where $u \in \mathbb{R}^d, \theta + n^{-1/2}u \in \Theta$. Stochastic process $Z_n(\theta; u)$ will be called the *log likelihood ratio process*. Clearly, for any $u \in \mathbb{R}^d$, $\theta + n^{-1/2}u \in \Theta$ for large enough n, implying that $Z_n(\theta; u)$ is well defined for large enough n. Note also that

$$\operatorname{Argmax}_{u\in\mathbb{R}^d,\theta+n^{-1/2}u\in\Theta} Z_n(\theta; u) = \sqrt{n}\Big(\operatorname{Argmax}_{\vartheta\in\Theta} L_n(\vartheta) - \theta\Big),$$

implying that any

$$\hat{u}_n \in \operatorname{Argmax}_{u\in\mathbb{R}^d,\theta+n^{-1/2}u\in\Theta} Z_n(\theta; u) = \sqrt{n}(\hat{\theta}_n - \theta),$$

where $\hat{\theta}$ is an MLE.

The following important result is due to Le Cam (see van der Vaart, *Asymptotic Statistics*, pp. 94–95).

Theorem 10.1 *Suppose $\{p_\vartheta : \vartheta \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^d$, is quadratic mean differentiable at point $\theta \in \Theta$. Let $X_1, \dots, X_n$ be i.i.d. $\sim p_\theta$. Then, for all $u \in \mathbb{R}^d$,*

$$Z_n(\theta; u) = \langle Y_n(\theta), u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle + o_{\mathbb{P}}(1) \text{ as } n \to \infty, \tag{10.1}$$

where

$$Y_n(\theta) := \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log p_\theta(X_j). \tag{10.2}$$

Representation (10.1) of log-likelihood ratio process $Z_n(\theta_0; u)$ is called *local asymptotic normality (LAN)*. It shows that the limit of $Z_n(\theta; u)$ is a relatively simple quadratic function w.r.t. $u \in \mathbb{R}^d$. Recall also that, under quadratic mean differentiablity,

$$\mathbb{E}_\theta \frac{\partial}{\partial\theta} \log p_\theta(X) = 0, \ \operatorname{cov}\Big(\frac{\partial}{\partial\theta} \log p_\theta(X)\Big) = I(\theta).$$

By the central limit theorem, this implies that

$$Y_n(\theta) \xrightarrow{d} N(0, I(\theta)) \text{ as } n \to \infty.$$

Thus, for all $u \in \mathbb{R}^d$,

$$Z_n(\theta; u) \xrightarrow{d} \langle Y(\theta), u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle$$

where $Y(\theta) = I(\theta)^{1/2}Z, Z \sim N(0; I_d)$. This allows one to reduce the asymptotic analysis of regular statistical models (experiments) to the analysis of limit Gaussian model (experiment). In particular, the local asymptotic normality easily implies the following heuristic derivation of asymptotic normality of MLE. Note that

$$\begin{aligned}
\sqrt{n}(\hat\theta_n - \theta) = \hat u_n &= \operatorname{argmax}_{\theta + n^{-1/2}u \in \Theta, u \in \mathbb{R}^d} Z_n(\theta; u) \\
&\approx \operatorname{argmax}_{u\in\mathbb{R}^d} \Big[\langle Y_n(\theta), u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle\Big] \\
&\overset{d}{\approx} \operatorname{argmax}_{u\in\mathbb{R}^d} \Big[\langle Y(\theta), u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle\Big] \\
&= \operatorname{argmax}_{u\in\mathbb{R}^d} \Big[\langle I(\theta)^{1/2}Z, u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle\Big] \\
&= \operatorname{argmax}_{u\in\mathbb{R}^d} \Big[\langle Z, I(\theta)^{1/2}u\rangle - \frac{1}{2}|I(\theta)^{1/2}u|^2\Big] \\
&= I(\theta)^{-1/2}\operatorname{argmax}_{v\in\mathbb{R}^d} \Big[\langle Z, v\rangle - \frac{1}{2}|v|^2\Big] \\
&= I(\theta)^{-1/2}Z \sim N(0, I(\theta)^{-1}),
\end{aligned}$$

provided that $I(\theta)$ is non-singular. Thus,

$$\sqrt{n}(\hat\theta_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}).$$

To justify this heuristic, additional assumptions are needed in order to make the remainder of LAN representation (10.1) *uniformly* small in bounded subsets of $\mathbb{R}^d$. Namely, the following lemma holds (see van der Vaart, *Asymptotic Statistics* for the proof).

Lemma 10.1 *Let*

$$Q_n(\theta; u) := \langle Y_n(\theta), u\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle, u \in \mathbb{R}^d$$

and

$$R_n(\theta; u) := Z_n(\theta; u) - Q_n(\theta; u), u \in \mathbb{R}^d, \theta + n^{-1/2}u \in \Theta.$$

Suppose $\{p_\vartheta : \vartheta \in \Theta\}$, *where* Θ *is an open subset of* $\mathbb{R}^d$, *is quadratic mean differentiable at point* $\theta \in \Theta$. *In addition, suppose that there exists a neighborhood* $U \subset \Theta$ *of point* θ *such that*

$$|\log p_{\vartheta'}(X) - \log p_{\vartheta''}(X)| \le L_\theta(X)|\vartheta' - \vartheta''|, \vartheta', \vartheta'' \in U,$$

with $L_\theta(X)$ *satisfying the condition* $\mathbb{E}_\theta L_\theta^2(X) < \infty$. *Let* $X_1, \dots, X_n$ *be i.i.d.* $\sim p_\theta$. *Then, for all* $R > 0$,

$$\sup_{|u|\le R} |R_n(\theta; u)| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty.$$

It is not hard to check that the result of the lemma also implies that there exists a sequence $R_n \to \infty$ such that

$$\sup_{|u|\le R_n} |R_n(\theta; u)| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty. \tag{10.3}$$

The following theorem provides sufficient conditions for asymptotic normality of MLE. Its proof relies on Lemma 10.1 to justify the heuristic argument based on LAN.

Theorem 10.2 *Let* $\{p_\theta : \theta \in \Theta\}$, *where* $\Theta \subset \mathbb{R}^d$ *is an open subset, be quadratic mean differentiable at point* $\theta \in \Theta$ *with non-singular Fisher information* $I(\theta)$. *Suppose that there exists a neighborhood* $U \subset \Theta$ *of point* θ *such that*

$$|\log p_{\vartheta'}(X) - \log p_{\vartheta''}(X)| \le L_\theta(X)|\vartheta' - \vartheta''|, \vartheta', \vartheta'' \in U,$$

where $L_\theta(X)$ *satisfies the condition* $\mathbb{E}_\theta L_\theta^2(X) < \infty$. *Let* $X_1, \dots, X_n$ *be i.i.d.* $\sim p_\theta$. *Let* $\hat{\theta}_n$ *be a maximum likelihood estimator and suppose that* $\hat{\theta}_n$ *is consistent:*

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta \text{ as } n \to \infty.$$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}).$$

10.4 Wilks' Theorem

Let $X_1, \dots, X_n$ be i.i.d. $\sim p_\theta, \theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^d$. Let $L \subset \mathbb{R}^d$ be a subspace of $\dim(L) < d$. Consider the following testing problem:

$$\begin{cases} H_0 : \theta \in \Theta \cap L \\ H_a : \theta \in \Theta \setminus L. \end{cases}$$

Let $\hat{\theta}_n$ be MLE for the whole model and let $\hat{\theta}_{n,0}$ be MLE under H_0.

Theorem 10.3 *Suppose $\{p_\theta : \theta \in \Theta\}$ is quadratic mean differentiable at point $\theta \in \Theta$ with non-singular Fisher information $I(\theta)$. Suppose that there exists a neighborhood $U \subset \Theta$ of point θ such that*

$$|\log p_{\vartheta'}(X) - \log p_{\vartheta''}(X)| \le L_\theta(X)|\vartheta' - \vartheta''|, \vartheta', \vartheta'' \in U,$$

where $L_\theta(X)$ satisfies the condition $\mathbb{E}_\theta L_\theta^2(X) < \infty$. Let $X_1, \dots, X_n$ be i.i.d. $\sim p_\theta$. Suppose $\hat{\theta}_n$ is consistent and $\hat{\theta}_{n,0}$ is consistent under H_0. Then, for all $\theta \in \Theta \cap L$,

$$\Lambda_n := 2 \log \frac{\sup_{\vartheta \in \Theta} L_n(\vartheta)}{\sup_{\vartheta \in \Theta \cap L} L_n(\vartheta)} \xrightarrow{d} \chi^2_{d - \dim(L)} \text{ as } n \to \infty.$$

Proof. Let $\theta \in L$ and let

$$\hat{u}_n := \sqrt{n}(\hat{\theta}_n - \theta) = \text{argmin}_{u \in \mathbb{R}^d, \theta + n^{-1/2}u \in \Theta} Z_n(\theta; u)$$

and

$$\hat{u}_{n,0} := \sqrt{n}(\hat{\theta}_{n,0} - \theta) = \text{argmin}_{u \in L, \theta + n^{-1/2}u \in \Theta} Z_n(\theta; u).$$

Then

$$\begin{aligned} \Lambda_n &= 2 \log \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_{n,0})} = 2 \log \frac{L_n(\hat{\theta}_n)}{L_n(\theta)} - 2 \log \frac{L_n(\hat{\theta}_{n,0})}{L_n(\theta)} \\ &= 2[Z_n(\theta; \hat{u}_n) - Z_n(\theta, \hat{u}_{n,0})]. \end{aligned}$$

By Theorem 10.2, estimator $\hat{\theta}_n$ is asymptotically normal and so is estimator $\hat{\theta}_{n,0}$ under H_0. Thus, for $\theta \in L$, both sequences $\{\hat{u}_n\}$ and $\{\hat{u}_{n,0}\}$ converge in distribution to normal r.v. and are stochastically bounded. This implies that, for all $\theta \in L$,

$$\mathbb{P}_\theta\{|\hat{u}_n| \ge R_n\} \to 0 \text{ and } \mathbb{P}_\theta\{|\hat{u}_{n,0}| \ge R_n\} \to 0 \text{ as } n \to \infty$$

for any sequence $R_n \to \infty$.

Let

$$\tilde{u}_n := \text{argmax}_{u \in \mathbb{R}^d} Q_n(\theta_0; u) = I(\theta)^{-1} Y_n(\theta)$$

and, for $\theta \in L$, let

$$\tilde{u}_{n,0} := \text{argmax}_{u \in L} Q_n(\theta_0; u) = (P_L I(\theta) P_L)^{-1} P_L Y_n(\theta).$$

Similarly, since $Y_n(\theta)$ converges in distribution (to a normal r.v.), it is stochastically bounded and we have

$$\mathbb{P}_\theta\{|\tilde{u}_n| \geq R_n\} \to 0 \text{ and } \mathbb{P}_\theta\{|\tilde{u}_{n,0}| \geq R_n\} \to 0 \text{ as } n \to \infty$$

for any sequence $R_n \to \infty$.

In what follows, we will fix $\theta \in L$ and write $\mathbb{P} = \mathbb{P}_\theta$. Recall that, by Lemma 10.1,

$$Z_n(\theta; u) = Q_n(\theta; u) + R_n(\theta; u), u \in \mathbb{R}^d, \theta + n^{-1/2} u \in \Theta,$$

where

$$\sup_{|u| \leq R_n} |R_n(\theta; u)| = o_{\mathbb{P}}(1)$$

for some sequence $R_n \to \infty$. Therefore, we have

$$\begin{aligned}
Z_n(\theta; \hat{u}_n) &= Z_n(\theta; \hat{u}_n) I(|\hat{u}_n| \leq R_n) + Z_n(\theta; \hat{u}_n) I(|\hat{u}_n| > R_n) \\
&= \sup_{|u| \leq R_n} Z_n(\theta; u) I(|\hat{u}_n| \leq R_n) + o_{\mathbb{P}}(1) \\
&= \sup_{|u| \leq R_n} Z_n(\theta; u) - \sup_{|u| \leq R_n} Z_n(\theta; u) I(|\hat{u}_n| > R_n) + o_{\mathbb{P}}(1) \\
&= \sup_{|u| \leq R_n} Z_n(\theta; u) + o_{\mathbb{P}}(1) \\
&= \sup_{|u| \leq R_n} Q_n(\theta; u) + o_{\mathbb{P}}(1) \\
&= Q_n(\theta; \tilde{u}_n) I(|\tilde{u}_n| \leq R_n) + \sup_{|u| \leq R_n} Q_n(\theta; u) I(|\tilde{u}_n| > R_n) + o_{\mathbb{P}}(1) \\
&= \sup_{u \in \mathbb{R}^d} Q_n(\theta; u) - \sup_{u \in \mathbb{R}^d} Q_n(\theta; u) I(|\tilde{u}_n| > R_n) + o_{\mathbb{P}}(1) \\
&= \sup_{u \in \mathbb{R}^d} Q_n(\theta; u) + o_{\mathbb{P}}(1)
\end{aligned}$$

Similarly, for all $\theta \in L$,

$$Z_n(\theta; \hat{u}_{n,0}) = \sup_{u \in L} Q_n(\theta; u) + o_{\mathbb{P}}(1).$$

Therefore,

$$\begin{aligned}
\Lambda_n &= 2[Z_n(\theta; \hat{u}_n) - Z_n(\theta, \hat{u}_{n,0})] \\
&= 2 \sup_{u \in \mathbb{R}^d} Q_n(\theta; u) - 2 \sup_{u \in L} Q_n(\theta; u) + o_{\mathbb{P}}(1).
\end{aligned}$$

Clearly,

$$\begin{aligned}2\sup_{u\in\mathbb{R}^d} Q_n(\theta;u) &= \sup_{u\in\mathbb{R}^d}\Big[-|I(\theta)^{-1/2}Y_n(\theta) - I(\theta)^{1/2}u|^2 + |I(\theta)^{-1/2}Y_n(\theta)|^2\Big]\\ &= |I(\theta)^{-1/2}Y_n(\theta)|^2 - \inf_{u\in\mathbb{R}^d}|I(\theta)^{-1/2}Y_n(\theta) - I(\theta)^{1/2}u|^2\\ &= |I(\theta)^{-1/2}Y_n(\theta)|^2\end{aligned}$$

and

$$\begin{aligned}2\sup_{u\in L} Q_n(\theta;u) &= \sup_{u\in L}\Big[-|I(\theta)^{-1/2}Y_n(\theta) - I(\theta)^{1/2}u|^2 + |I(\theta)^{-1/2}Y_n(\theta)|^2\Big]\\ &= |I(\theta)^{-1/2}Y_n(\theta)|^2 - \inf_{u\in L}|I(\theta)^{-1/2}Y_n(\theta) - I(\theta)^{1/2}u|^2,\end{aligned}$$

implying that

$$\begin{aligned}\Lambda_n &= \inf_{u\in L}|I(\theta)^{-1/2}Y_n(\theta) - I(\theta)^{1/2}u|^2 + o_{\mathbb{P}}(1)\\ &= \inf_{v\in I(\theta)^{1/2}L}|I(\theta)^{-1/2}Y_n(\theta) - v|^2 + o_{\mathbb{P}}(1)\\ &= \mathrm{dist}^2(I(\theta)^{-1/2}Y_n(\theta), \tilde{L}) + o_{\mathbb{P}}(1),\end{aligned}$$

where $\tilde{L} = I(\theta)^{1/2}L$ and $\dim(\tilde{L}) = \dim(L)$. Note also that

$$I(\theta)^{-1/2}Y_n(\theta) \xrightarrow{d} Z, Z \sim N(0; I_d).$$

Therefore,

$$\Lambda_n \xrightarrow{d} \mathrm{dist}^2(Z, \tilde{L}) = |P_{\tilde{L}^\perp}Z|^2 \sim \chi^2_{d-\dim(L)}.$$

■

10.5 Bahadur's Efficiency of Likelihood Ratio Tests

Let $X_1, \ldots, X_n$ be i.i.d. observations in $(S, \mathcal{A})$ with distribution P. Suppose that P is absolutely continuous w.r.t. measure μ on $(S, \mathcal{A})$ (finite or σ-finite) with density p. Consider the following hypotheses testing problem:

$$\begin{cases}H_0 : P \in \mathcal{P}_0\\ H_a : P \in \mathcal{P}_1,\end{cases}$$

where $\mathcal{P}_0, \mathcal{P}_1$ are two sets of distributions P. For simplicity, assume that $\mathcal{P}_0, \mathcal{P}_1$ are finite sets.

Let $\{T_n\}$ be a sequence of test statistics $T_n(X_1, \ldots, X_n)$ and suppose that H_0 has to be rejected for large values of T_n. Denote

$$\alpha_n(P; t) := \mathbb{P}_P\{T_n \geq t\}, t \in \mathbb{R}$$

and

$$\alpha_n(\mathcal{P}_0; t) := \sup_{P \in \mathcal{P}_0} \alpha_n(P; t), t \in \mathbb{R}.$$

Clearly, $\alpha_n(\mathcal{P}_0; t)$ is a nonincreasing left continuous function of t. Given $\alpha \in (0, 1)$, let

$$C_n(\alpha) := \{t : \alpha_n(\mathcal{P}_0; t) \leq \alpha\}.$$

It is easy to see that $C_n(\alpha) = [c_{n,\alpha}, +\infty)$ provided that there exists $c_{n,\alpha}$ such that $\alpha_n(\mathcal{P}_0; c_{n,\alpha}) = \alpha$, or $C_n(\alpha) = (c_{n,\alpha}, +\infty)$ if there exists $c_{n,\alpha}$ such that

$$\alpha \in (\alpha_n(\mathcal{P}_0; c_{n,\alpha}+), \alpha_n(\mathcal{P}_0; c_{n,\alpha})).$$

Note also that, for all $P \in \mathcal{P}_0$,

$$\mathbb{P}_P\{T_n \in C_n(\alpha)\} \leq \alpha.$$

Definition 10.3 *Denote*

$$L_n := \alpha_n(\mathcal{P}_0; T_n).$$

Random variable L_n will be called the observed significance level (p-value) of T_n.

Note that T_n is "large" if and only if L_n is "small". Note also that, for all $P \in \mathcal{P}_0$,

$$\mathbb{P}_P\{L_n \leq \alpha\} = \mathbb{P}_P\{\alpha_n(\mathcal{P}_0; T_n) \leq \alpha\} = \mathbb{P}_P\{T_n \in C_n(\alpha)\} \leq \alpha.$$

Therefore, the test that rejects H_0 when $L_n \leq \alpha$ and does not reject otherwise is of size α.

Finally, note that, under alternative H_a, it is desirable that H_0 is rejected which happens when L_n is small. Thus, one can characterize the quality of the tests based on $\{T_n\}$ by the rate of decay of sequence L_n as $n \to \infty$ for $Q \in \mathcal{P}_1$ (the faster the decay is, the better).

Definition 10.4 *For $Q \in \mathcal{P}_1$, $X_1, \ldots, X_n$ i.i.d. $\sim Q$, define*

$$B_Q(\{T_n\}) := \lim_{n \to \infty} \Big[-\frac{2}{n} \log L_n\Big],$$

provided that the limit "in probability" exists. $B_Q(\{T_n\})$ is called the Bahadur's slope of $\{T_n\}$.

Roughly speaking, $L_n \approx e^{-\frac{B_Q n}{2}}$, where $B_Q = B_Q(\{T_n\})$. The larger B_Q is, the faster L_n converges to 0. In what follows, we show that the sequence of likelihood ratio test statistics for H_0 against H_a has the largest Bahadur's slope (is Bahadur efficient).

Recall that the LR-statistic is defined as

$$\tilde{\Lambda}_n := \log \frac{\sup_{P\in\mathcal{P}_1} \prod_{j=1}^n p(X_j)}{\sup_{P\in\mathcal{P}_0} \prod_{j=1}^n p(X_j)}.$$

Recall also that the Kullback-Leibler (KL) divergence between distributions P and Q with densities p and q is defined as

$$K(P||Q) = K(p||q) = \mathbb{E}_P \log \frac{p(X)}{q(X)}.$$

The following result will be proved.

Theorem 10.4 *Suppose $\{T_n\}$ is a sequence of statistics. Then, for all $Q \in \mathcal{P}_1$,*

$$B_Q(\{T_n\}) \le 2 \min_{P\in\mathcal{P}_0} K(Q||P) \text{ a.s.}$$

On the other hand, for all $Q \in \mathcal{P}_1$,

$$B_Q(\{\tilde{\Lambda}_n\}) = 2 \min_{P\in\mathcal{P}_0} K(Q||P) \text{ a.s.}$$

Proof. To prove the first claim, it is enough to consider the case $\mathcal{P}_0 = \{P\}$ (otherwise, note that if $P \in \mathcal{P}_0$, then the Bahadur slope for the smaller null hypothesis $\{P\}$ is at least as large as the Bahadur slope for $\mathcal{P}_0$).

Let $B > A > K(Q||P)$. Denote

$$\Lambda_n := \log \frac{q(X_1)\dots q(X_n)}{p(X_1)\dots p(X_n)}.$$

For all Q, by the strong law of large numbers,

$$n^{-1}\Lambda_n = n^{-1}\sum_{j=1}^n \log \frac{q(X_j)}{p(X_j)} \to K(Q||P) \text{ as } n\to\infty \ \mathbb{P}_Q - \text{a.s.}$$

Therefore, $\Lambda_n \le nA$ for all n large enough $\mathbb{P}_Q$-a.s. Note also that

$$d(Q\times\dots\times Q) = e^{\Lambda_n} d(P\times\dots\times P)$$

and we have

$$\begin{aligned}\mathbb{P}_Q\{L_n \le e^{-nB}, \Lambda_n \le nA\} &= \mathbb{E}_P I(L_n \le e^{-nB}, \Lambda_n \le nA) e^{\Lambda_n} \\ &\le e^{nA}\mathbb{P}_P\{L_n \le e^{-nB}\} \le e^{n(A-B)}.\end{aligned}$$

For $B > A$, $\sum_n e^{n(A-B)} < \infty$ implying, by Borel-Cantelli Lemma, that

$$\mathbb{P}_Q\{L_n \le e^{-nB}, \Lambda_n \le nA \text{ i.o.}\} = 0.$$

Therefore, $\mathbb{P}_Q$ a.s. either $L_n > e^{-nB}$, or $\Lambda_n > nA$ for all n large enough. As we proved before, the last inequality does not hold for large n $\mathbb{P}_Q$ a.s. Therefore, $L_n > e^{-nB}$ for all large n $\mathbb{P}_Q$ a.s. and we have $-\frac{2}{n}\log L_n < 2B$, implying that

$$\limsup_n \Big[-\frac{2}{n}\log L_n\Big] \le 2B \ \mathbb{P}_Q - \text{a.s.}$$

Since it holds for all $B > K(Q||P)$, we get

$$\limsup_n \Big[-\frac{2}{n}\log L_n\Big] \le 2K(Q||P) \ \mathbb{P}_Q - \text{a.s.},$$

implying that $B_Q(\{T_n\}) \le 2K(Q||P)$, which proves the first claim.

To prove the second claim, note that for statistics $\tilde{\Lambda}_n$,

$$\alpha_n(\mathcal{P}_0; t) = \sup_{P\in\mathcal{P}_0} \mathbb{P}_P\{\tilde{\Lambda}_n \ge t\}$$

We have, for all $P \in \mathcal{P}_0$,

$$\mathbb{P}_P\{\tilde{\Lambda}_n \ge t\} = \mathbb{P}_P\left\{\log \frac{\sup_{P\in\mathcal{P}_1}\prod_{j=1}^n p(X_j)}{\sup_{P\in\mathcal{P}_0}\prod_{j=1}^n p(X_j)} \ge t\right\} \le \mathbb{P}_P\left\{\sup_{Q\in\mathcal{P}_1}\log\frac{\prod_{j=1}^n q(X_j)}{\prod_{j=1}^n p(X_j)} \ge t\right\}$$

$$\le \text{card}(\mathcal{P}_1)\max_{Q\in\mathcal{P}_1}\mathbb{P}_P\left\{\log\frac{\prod_{j=1}^n q(X_j)}{\prod_{j=1}^n p(X_j)} \ge t\right\} \le \text{card}(\mathcal{P}_1)\max_{Q\in\mathcal{P}_1}\frac{\mathbb{E}_P\frac{\prod_{j=1}^n q(X_j)}{\prod_{j=1}^n p(X_j)}}{e^t}$$

$$= \text{card}(\mathcal{P}_1)\max_{Q\in\mathcal{P}_1}\frac{\int_S\cdots\int_S \frac{q(x_1)\ldots q(x_n)}{p(x_1)\ldots p(x_n)}p(x_1)\ldots p(x_n)\mu(dx_1)\ldots\mu(dx_n)}{e^t} \le \text{card}(\mathcal{P}_1)e^{-t}.$$

Therefore,

$$L_n = \alpha_n(\mathcal{P}_0; \tilde{\Lambda}_n) \le \text{card}(\mathcal{P}_1)e^{-\tilde{\Lambda}_n},$$

implying that

$$-\frac{2}{n}\log L_n \le -\frac{\log\text{card}(\mathcal{P}_1)}{n} + \frac{\tilde{\Lambda}_n}{n}.$$

It follows from the strong law of large numbers that, for all $Q \in \mathcal{P}_1$,

$$\frac{\tilde{\Lambda}_n}{n} = \min_{P\in\mathcal{P}_0}\max_{Q\in\mathcal{P}_1}\frac{1}{n}\sum_{j=1}^n\log\frac{q(X_j)}{p(x_j)} \ge \min_{P\in\mathcal{P}_0}\frac{1}{n}\sum_{j=1}^n\log\frac{q(X_j)}{p(x_j)}$$

$$\to \min_{P\in\mathcal{P}_0} K(Q||P) \text{ as } n\to\infty \ \mathbb{P}_Q - \text{a.s.}$$

Thus,

$$\liminf_n \Big[-\frac{2}{n}\log L_n\Big] \geq 2 \min_{P\in\mathcal{P}_0} K(Q||P)\ \mathbb{P}_Q - \text{a.s.}$$

and $B_Q(\{\tilde{\Lambda}_n\}) \geq 2\min_{P\in\mathcal{P}_0} K(Q||P)$, which allows us to complete the proof.

■

10.6 Problems

1. Check that the quadratic mean differentiability condition holds for the normal model $N(\theta,1), \theta\in\mathbb{R}$ and does not hold for the uniform model $U[0,\theta], \theta>0$.

2. Despite the fact that the proof of local asymptotic normality under quadratic mean differentiablity assumption is rather delicate, it is not hard to prove this property under stronger assumptions using Taylor formula for $\log p_{\theta+n^{-1/2}u}(X_j) - \log p_\theta(X_j)$. Try to write down such a proof.

3. Deduce (10.3) from the result of Lemma 10.1.

4. If A_n are events such that $\mathbb{P}(A_n)\to 0$ as $n\to\infty$ and ξ_n is an arbitrary sequence of r.v. in $\mathbb{R}$, then

$$\xi_n I_{A_n} = o_{\mathbb{P}}(1).$$

   Prove it.

5. Similarly to the proof of Theorem 10.3, try to write down the proof of Theorem 10.2 based on Lemma 10.1.

6. Suppose $\{p_\theta : \theta\in\mathbb{R}\}$ is a parametric family of densities satisfying quadratic mean differentiabilty with Fisher information $I(\theta)$. Let $(X_1,\dots,X_n)$ be i.i.d. with density p_θ, where θ is an unknown parameter. Consider a sequence of hypotheses testing problems $H_0 : \theta=\theta_0$ against $H_n : \theta=\theta_n$, where $\theta_n := \theta_0 + n^{-1/2}h$. Let ϕ_n be a sequence of Neyman-Pearson tests with $\phi_n(X_1,\dots,X_n)=1$ when $p_{\theta_n}(X_1)\dots p_{\theta_n}(X_n) \geq p_{\theta_0}(X_1)\dots p_{\theta_0}(X_n)$ and $\phi_n(X_1,\dots,X_n)=0$ otherwise. Denote α_n and α_n' the probabilities of the test making an error under H_0 and under H_n, respectively. Prove that

$$\limsup_{n\to\infty} \alpha_n \leq \exp\{-I(\theta_0)h^2/8\}$$

   and

$$\limsup_{n\to\infty} \alpha_n' \leq \exp\{-I(\theta_0)h^2/8\}.$$

7. Find the likelihood ratio test for the hypothesis that a die is fair (i.e., the probability of any number $1,2,\dots,6$ is equal to $1/6$) based on n independent rolls of the die.

8. Let $X = (X_1, X_2, X_3, X_4) \sim N(\theta; I_4), \theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4$, where I_4 is 4×4 identity matrix. Describe likelihood ratio test of level α for hypothesis $H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4$ against the alternative H_a : for some $1 \leq i \neq j \leq 4$, $\theta_i \neq \theta_j$.

9. Suppose $X_1, \ldots, X_n$ are i.i.d. with density p_θ, $\theta \in \mathbf{R}^2$. Consider the following hypotheses testing problem: $H_0 : \theta \in S$ against $H_a : \theta \notin S$, where $S \subset \mathbb{R}^2$ is an equilateral triangle in the plane. Assuming that the model $\{p_\theta\}$ satisfies standard regularity assumptions (in particular, that a proper version of Local Asymptotic Normality (LAN) holds and that the Fisher information matrix $I(\theta)$ is nonsingular), find the asymptotic distribution of the loglikelihood ratio statistics as $n \to \infty$. Determine it, in particular, when $I(\theta)$ is the identity matrix. In this case, describe precisely how the asymptotic distribution depends on $\theta \in S$.

10. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta; 1), \theta \in (-\infty, \infty)$. Consider the problem of testing $H_0 : \theta \leq 0$ against $H_a : \theta > 0$. Find the Bahadur slope of the test statistic $\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}$ for an alternative $\theta > 0$.

    Hint: you might want to use something like the asymptotic relationship $\mathbf{P}\{Z \geq x\} \sim \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}$ as $x \to \infty$, where Z is a standard normal r.v.