Theoretical Statistics - II

Bodhisattva Sen

April 19, 2021

Contents

1 Hypothesis Testing

We are given data $X \sim P_\theta$ ($X \in \mathcal{X}$) from a model that is parametrized by θ (e.g., say $X = (X_1, \ldots, X_n)$ where X_i's are i.i.d. from a parametric family with parameter θ). We consider a statistical problem involving θ whose value is unknown but must lie in a certain space Θ. We consider the testing problem

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1, \tag{1}$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$.

Here the hypothesis H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*. In hypothesis testing data are used to infer which of two competing hypotheses[1], H_0 or H_1, is correct. H_0 is *simple* if Θ_0 is a set with only one point; otherwise, H_0 is *composite*.

Example 1.2. Suppose that $X_1, \ldots, X_n$ are i.i.d $N(\theta, \sigma^2)$ where $\theta \in \mathbb{R}$ is unknown, and $\sigma > 0$ is assumed *known*. Suppose that we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Under the null hypothesis the X_i's are i.i.d $N(\theta_0, \sigma^2)$ and the sample mean $\overline{X} \sim N(\theta_0, \sigma^2/n)$.

Thus, a reasonable test can be to reject H_0 if $T := |\overline{X} - \mu_0| > c$, for some "large" constant c (as large deviations of the observed value of $\overline{X}$ from μ_0 would lead us to suspect that the null hypothesis might not be true).

But how large is large? We will discuss this soon...

A *nonrandomized test* of H_0 versus H_1 can be specified by a *critical region* $S \subset \mathcal{X}$ with the convention that we accept H_1 (or reject H_0) when $X \in S$ and accept H_0 when $X \notin S$. The performance of this test is described by its *power function* $\beta(\cdot)$, which gives the chance of rejecting H_0 as a function of $\theta \in \Theta$:

$$\beta(\theta) := \mathbb{P}_\theta(X \in S).$$

Ideally, we would want $\beta(\theta) = 0$ for $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$, but in practice this is generally impossible.

For technical reasons it is convenient to allow external randomization to "help" the researcher decide between H_0 and H_1. Randomized tests are characterized by a *test* or

[1]

Definition 1.1 (One-sided and two-sided hypotheses)**.** Let θ be a one-dimensional parameter.

- *one-sided hypotheses*
  - $H_0 : \theta \leq \theta_0$, and $H_1 : \theta > \theta_0$, or
  - $H_0 : \theta \geq \theta_0$, and $H_1 : \theta < \theta_0$
- *two-sided hypotheses* $H_0 : \theta = \theta_0$, and $H_1 : \theta \neq \theta_0$.

critical function ϕ with range a subset of $[0,1]$, i.e., $\phi : \mathcal{X} \to [0,1]$. Given $X = x$, $\phi(x)$ is the chance of rejecting H_0. The power function $\beta(\cdot)$ still gives the chance of rejecting H_0, and by smoothing,

$$\beta(\theta) = \mathbb{P}_\theta(\text{Reject } H_0) = \mathbb{E}_\theta[\mathbb{P}_\theta(\text{Reject } H_0|X)] = \mathbb{E}_\theta[\phi(X)].$$

A nonrandomized test with critical region S can be viewed as a randomized test with $\phi = 1_S$. Conversely, if $\phi(x)$ is always 0 or 1, then the randomized test with critical function ϕ can be considered a nonrandomized test with critical region $S = \{x \in \mathcal{X} : \phi(x) = 1\}$.

Goals: We would like the power function $\beta(\theta)$ to be *low* for values of $\theta \in \Theta_0$, and **high** for $\theta \in \Theta_1$. Hence, there is a need to strike an appropriate balance between the two goals of

$$\textit{low power in } \Theta_0 \qquad \text{and} \qquad \textit{high power in } \Theta_1.$$

The most popular method for striking a balance between the two goals is to choose a number $\alpha \in (0,1)$ and require that

$$\beta_\phi(\theta) \leq \alpha, \quad \text{for all} \quad \theta \in \Theta_0. \tag{2}$$

This α will usually be a small positive fraction (historically .05 or .01) and will be called the *level of significance* or simply *level*. Then, among all tests that satisfy (2), the statistician seeks a test whose power function is as high as can be obtained for $\theta \in \Theta_1$.

The *size* of a (randomized) test ϕ is defined as $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$.

1.1 Uniformly most powerful (UMP) tests

Definition 1.3. A test ϕ^* with level α is called uniformly most powerful (UMP) if

$$\mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)], \qquad \text{for all } \theta \in \Theta_1,$$

for all ϕ with level at most α.

Uniformly most powerful tests for composite hypotheses generally only arise when the parameter of interest is univariate, $\theta \in \Theta \subset \mathbb{R}$ and the hypotheses are of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a fixed constant[2]. In addition, the family of densities needs to have an appropriate structure.

1.2 Simple versus simple testing

A hypothesis is called *simple* if it completely specifies the distribution of the data, so $H_i : \theta \in \Theta_i$ is simple when Θ_i contains a single parameter value θ_i. When both hypotheses,

[2]Minor variants are possible here: H_0 could be $\theta = \theta_0, \theta < \theta_0, \theta \geq \theta_0$, etc.

H_0 and H_1 are simple, the Neyman-Pearson lemma (Theorem 1.4) provides a complete characterization of all reasonable tests. This result makes use of Lagrange multipliers, an important idea in optimization theory of independent interest.

Theorem 1.4 (Neyman-Pearson (NP) lemma)**.** Let P_{θ_0} and P_{θ_1} have densities p_0 and p_1 with respect to (w.r.t.) some dominating measure (recall that $\mu = P_{\theta_0} + P_{\theta_1}$ always works). Consider testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1.$$

Let $\alpha \in [0, 1]$. Then:

(i) There exists a constant k and a critical function ϕ_0 of the form

$$\phi_0(x) = \begin{cases} 1 & \text{when} \quad p_1(x) > kp_0(x), \\ 0 & \text{when} \quad p_1(x) < kp_0(x) \end{cases} \tag{3}$$

such that

$$\mathbb{E}_0[\phi_0(X)] = \alpha. \tag{4}$$

(ii) The test ϕ_0 in (3) satisfying (4) is a most powerful level α test of P_{θ_0} versus P_{θ_1}.

(iii) If ϕ is a most powerful level α test of P_{θ_0} versus P_{θ_1}, then it must be of the form (3) a.e. μ. It also satisfies (4) unless there is a test of size $< \alpha$ with power $= 1$.

1.3 Duality between testing and interval estimation

1.4 Generalized NP lemma

Theorem 1.5 (Generalized NP lemma)**.** Let $f_1, \ldots, f_{m+1}$ be real-valued, μ-integrable functions defined on a Euclidean space $\mathcal{X}$. Suppose that for given constants $c_1, \ldots, c_m$ there exists a critical function ϕ satisfying

$$\int \phi f_i d\mu = c_i, \qquad i = 1, \ldots, m. \tag{5}$$

Let $\mathcal{C}$ be the class of critical functions ϕ for which (5) holds.

(i) Among all members of $\mathcal{C}$ there exists one that maximizes $\int \phi f_{m+1} d\mu$.

(ii) A sufficient condition for a member ϕ_0 of $\mathcal{C}$ to maximize $\int \phi f_{m+1} d\mu$ (over $\mathcal{C}$) is the existence of constants $k_1, \ldots, k_m$ such that

$$\phi_0(x) = \begin{cases} 1 & \text{when} \quad f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x), \\ 0 & \text{when} \quad f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x). \end{cases} \tag{6}$$

(iii) If a member of $\mathcal{C}$ satisfies (6) with $k_1, \ldots, k_m \geq 0$, then it maximizes $\int \phi f_{m+1} d\mu$ among all critical functions satisfying $\int \phi f_i d\mu \leq c_i$, for $i = 1, \ldots, m$.

(iv) The set

$$M := \left\{ \left(\int \phi f_1 d\mu, \ldots, \int \phi f_m d\mu \right) : \phi \text{ is a critical function} \right\}$$

is convex and closed. If $(c_1, \ldots, c_m)$ is an interior point of M, then there exists constants $k_1, \ldots, k_m$ and a test ϕ_0 satisfying (5) and (6). And a necessary condition for a member of $\mathcal{C}$ to maximize $\int \phi f_{m+1} d\mu$ is that (6) holds a.e. μ.

Proof. We will only prove parts (ii) and (iii) here; for the proofs of the existence results see Lehmann and Romano [8, Theorem 3.6.1].

Proof of (ii): Take $\phi \in \mathcal{C}$. Note that $\int (\phi_0 - \phi)(f_{m+1} - \sum_{i=1}^m k_i f_i) d\mu \geq 0$ since the integrand is ≥ 0 (by the definition of ϕ_0). Hence,

$$\int (\phi_0 - \phi) f_{m+1} d\mu \geq \sum_{i=1}^m k_i \int (\phi_0 - \phi) f_i d\mu = 0 \quad \Rightarrow \quad \int \phi_0 f_{m+1} d\mu \geq \int \phi f_{m+1} d\mu.$$

This completes the proof of (i).

Proof of (iii): Suppose that $\phi_0 \in \mathcal{C}$ satisfies (6) with $k_1, \ldots, k_m \geq 0$. Take a critical function ϕ such that $\int \phi f_i d\mu \leq c_i$, for $i = 1, \ldots, m$. As in (i), $\int (\phi_0 - \phi)(f_{m+1} - \sum_{i=1}^m k_i f_i) d\mu \geq 0$, and thus,

$$\int (\phi_0 - \phi) f_{m+1} d\mu \geq \sum_{i=1}^m \int k_i (\phi_0 - \phi) f_i d\mu \geq 0 \quad \Rightarrow \quad \int \phi_0 f_{m+1} d\mu \geq \int \phi f_{m+1} d\mu,$$

as $\sum_{i=1}^m k_i \int \phi_0 f_i d\mu = \sum_{i=1}^m k_i c_i$ and $\sum_{i=1}^m k_i \int \phi f_i d\mu \leq \sum_{i=1}^m k_i c_i$. □

Example 1.6. Suppose that $X_1, \ldots, X_n$ are i.i.d. from the Cauchy location family $p_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$, for $x \in \mathbb{R}$ (let $\mathbf{X} = (X_1, \ldots, X_n)$). Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$[3]. Can we find a test ϕ of size α such that ϕ maximizes

$$\frac{d}{d\theta} \beta_\phi(\theta_0) = \frac{d}{d\theta} \mathbb{E}_\theta[\phi(\mathbf{X})]|_{\theta=\theta_0}? \tag{7}$$

For any test ϕ the power is given by

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(\mathbf{X})] = \int \phi(\mathbf{x}) p(\mathbf{x}; \theta) d\mathbf{x},$$

where $p(\mathbf{x}; \theta)$ is the joint density of the model. So, if the interchange of differentiation and integration is justifiable[4], then

$$\beta'_\phi(\theta) = \int \phi(\mathbf{x}) \frac{\partial}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x}.$$

[3] Exercise 1 (HW1): Show that here a UMP test for testing H_0 against H_1 does not exist when $n = 1$.

[4] Quite often, the dominated convergence theorem (DCT) can be used to justify the interchange.

Thus, by the generalized N-P lemma, a test of the form

$$\phi_0(x) = \begin{cases} 1 & \text{when} \quad \frac{\partial}{\partial\theta} p(\mathbf{x};\theta_0) > kp(\mathbf{x};\theta_0), \\ 0 & \text{when} \quad \frac{\partial}{\partial\theta} p(\mathbf{x};\theta_0) < kp(\mathbf{x};\theta_0). \end{cases}$$

maximizes $\beta'_\phi(\theta_0)$ among all ϕ with $\mathbb{E}_{\theta_0}\phi(\mathbf{X}) = \alpha$. This test is said to be *locally most powerful*[5] of size α; cf. Ferguson, Section 5.5, page 235. But

$$\frac{\partial}{\partial\theta} p(\mathbf{X};\theta_0) > kp(\mathbf{X};\theta_0) \quad \Leftrightarrow \quad \frac{\partial}{\partial\theta} \log p(\mathbf{X};\theta_0) > k \quad \Leftrightarrow \quad S_n(\theta_0) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) > k'.$$

Here $\ell_\theta(x) \equiv \log p_\theta(x)$ and $\dot{\ell}_{\theta_0}(x) = \frac{\partial}{\partial\theta}\ell_\theta(x)$. Hence for the Cauchy family (with $\theta_0 = 0$ without loss of generality), since $\dot{\ell}_{\theta_0}(x) = \frac{2(x-\theta)}{1+(x-\theta)^2}$, the locally most powerful test is given by

$$\phi(\mathbf{X}) = \begin{cases} 1 & \text{when} \quad n^{-1/2}\sum_{i=1}^n \frac{2X_i}{1+X_i^2} > k', \\ 0 & \text{when} \quad n^{-1/2}\sum_{i=1}^n \frac{2X_i}{1+X_i^2} < k'. \end{cases} \tag{8}$$

where k' is such that $\mathbb{E}_0[\phi(\mathbf{X})] = \alpha$. Although an exact value of k' above might be difficult to obtain, we can easily approximate k' as follows.

Under $H_0 : \theta = \theta_0 = 0$, with $Y_i = 2X_i/(1 + X_i^2)$,

$$\mathbb{E}_0 Y_i = 0 \qquad \text{and} \qquad \mathrm{Var}_0(Y_i) = \frac{1}{2}.$$

Hence, by the CLT, k' may be approximated by $2^{-1/2}z_\alpha$ where $\mathbb{P}(Z > z_\alpha) = \alpha$ with $Z \sim N(0, 1)$.

1.5 Unbiased tests

We know that in Example 1.2 a uniformly most powerful (UMP) test cannot exist[6]. One appealing constraint restricts attention to tests that are unbiased according to the following definition.

[5] When a UMP test does not exist, one may restrict the class of tests to, say, the class of unbiased tests (see Section 1.5), and then look for a UMP test in this smaller class. Alternatively, one may look for tests that have maximum power against alternatives in a subset of Θ_1. The case when the subset of alternatives is "close" to the null parameter values has received a good deal of attention, presumably because tests that have good power for "local alternatives", which are the hardest to detect, may also retain good power for "nonlocal" alternatives.

[6] Note that for testing

$$H_0 : \theta \le \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

a UMP test exists, since the family has *monotone likelihood ratio*; see e.g., Lehmann and Romano [8, Section 3.4]. Exercise 2 (HW1): Find a level α test has better power (for some θ's) than the usual two-sided z-test based on $\overline{X}$.

Definition 1.7 (Unbiased tests)**.** A test ϕ for $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ with level α is unbiased if its power $\beta_\phi(\theta) := \mathbb{E}_\theta[\phi(X)]$ satisfies

$$\beta_\phi(\theta) \leq \alpha, \quad \text{for all } \theta \in \Theta_0 \qquad \text{and} \qquad \beta_\phi(\theta) \geq \alpha \quad \text{for all } \theta \in \Theta_1.$$

If there is a UMP test ϕ^*, then it is automatically unbiased because $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$, for all $\theta \in \Theta_1$, where ϕ is the degenerate test, which equals α regardless of the observed data.

Definition 1.8 (UMP unbiased test)**.** A UMP unbiased (UMPU) level α test is a test ϕ_0 for which

$$\beta_{\phi_0}(\theta) \geq \beta_\phi(\theta) \quad \text{for all } \theta \in \Theta_1,$$

for all unbiased level α tests ϕ.

Indeed, we will see that a UMP unbiased test exists for Example 1.2. The following result, stated in the generality of a one-parameter exponential family, yields this desired result.

Theorem 1.9 (Application to one-parameter exponential family)**.** Consider i.i.d. data $X_1, \ldots, X_n$ from a one-parameter exponential family with density such that the joint density of the data can be expressed as

$$p(\mathbf{x}; \theta) = c(\theta) \exp(\theta T(\mathbf{x})) h(\mathbf{x}), \qquad \text{for } \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n,$$

for $\theta \in \Theta \subset \mathbb{R}$, w.r.t. a σ-finite measure on $\mathcal{X}$. For $\theta_0 \in \Theta$, consider testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

Then, for $\alpha \in (0, 1)$, the test ϕ_0 with

$$\mathbb{E}_{\theta_0}[\phi_0(T(\mathbf{X}))] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\theta_0}[T(\mathbf{X})\phi_0(T(\mathbf{X}))] = \alpha \mathbb{E}_{\theta_0}[T(\mathbf{X})]$$

and given by

$$\phi_0(T(\mathbf{x})) = \begin{cases} 1 & \text{if } \; T(\mathbf{x}) < c_1 \; \text{ or } \; T(\mathbf{x}) > c_2, \\ \gamma_i & \text{if } \; T(\mathbf{x}) = c_i, \\ 0 & \text{if } \; \text{otherwise.} \end{cases} \tag{9}$$

(for some $\gamma_1, \gamma_2 \in [0, 1]$ and $c_1 < c_2$) is a UMPU level α for H_0 versus H_1.

Furthermore, if T is symmetrically distributed about a under θ_0, then $\mathbb{E}_{\theta_0}[\phi_0(T(\mathbf{X}))] = \alpha$, $c_2 = 2a - c_1$ and $\gamma_1 = \gamma_2$ determine the constants.

Proof. We will restrict attention to tests of the form $\phi(\mathbf{x}) = \psi(T(\mathbf{x}))$ based on the sufficient statistic T[7], whose distribution is of the form $p_\theta(t) = c(\theta)e^{\theta t}$ (w.r.t. some σ-finite measure ν); see e.g., Lehmann and Romano [8, Lemma 2.7.2].

[7]Note that if $\phi(\mathbf{X})$ is any test then we can consider the test function $\tilde{\phi}(T) = \mathbb{E}[\phi(\mathbf{X})|T]$, which is valid test based on the the sufficient statistic T (as the distribution of $\mathbf{X}|T$ does not depend on θ), and has the exact same power function as ϕ, i.e., $\mathbb{E}_\theta[\phi(\mathbf{X})] = \mathbb{E}_\theta[\tilde{\phi}(T)]$ for all $\theta \in \Theta$.

Since all power functions are continuous in the case of an exponential family[8], it follows that any unbiased test ψ satisfies $\alpha = \beta_\psi(\theta_0) = \mathbb{E}_{\theta_0}[\psi(T)]$. Further, $\beta_\psi(\cdot)$ has a minimum at θ_0.

By Lehmann and Romano [8, Theorem 2.7.1], $\beta_\psi(\cdot)$ is differentiable, and can be differentiated under the integral sign. Hence

$$\begin{aligned}
\beta'_\psi(\theta) &= \frac{d}{d\theta}\int \psi(t)c(\theta)\exp(\theta t)d\nu(t) \\
&= \frac{c'(\theta)}{c(\theta)}\mathbb{E}_\theta[\psi(T)] + \mathbb{E}_\theta[T\psi(T)] \\
&= \{-\mathbb{E}_\theta[T]\}\mathbb{E}_\theta[\psi(T)] + \mathbb{E}_\theta[T\psi(T)]
\end{aligned}$$

since, with $\psi_0 \equiv \alpha$, $0 = \beta'_{\psi_0}(\theta) = \alpha\{c'(\theta)/c(\theta) + \mathbb{E}_\theta[T]\}$ (which implies that $c'(\theta)/c(\theta) = -\mathbb{E}_\theta[T]$). Thus,

$$0 = \beta'_\psi(\theta_0) = \mathbb{E}_{\theta_0}[T\psi(T)] - \alpha\mathbb{E}_{\theta_0}[T].$$

Thus any unbiased level α test $\psi(T)$ satisfies the two conditions:

$$\mathbb{E}_{\theta_0}[\psi(T(\mathbf{X}))] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\theta_0}[T(\mathbf{X})\psi(T(\mathbf{X}))] = \alpha\mathbb{E}_{\theta_0}[T(\mathbf{X})]. \tag{10}$$

We will apply the generalized NP lemma to show that ϕ_0 as given in (9) is UMPU.

Fix $\theta' \neq \theta \in \Theta$ and consider maximizing $\mathbb{E}_{\theta'}[\psi(T)]$ subject to the constraints in (10). By the generalized NP lemma Theorem 1.5-(iv)[9], there exist k_1, k_2 such that

$$\begin{aligned}
\psi_0(t) &= \begin{cases} 1 & \text{when} \quad c(\theta')e^{\theta' t} > c(\theta_0)(k_1 + k_2 t)e^{\theta_0 t}, \\ 0 & \text{when} \quad c(\theta')e^{\theta' t} < c(\theta_0)(k_1 + k_2 t)e^{\theta_0 t} \end{cases} \\
&= \begin{cases} 1 & \text{when} \quad e^{bt} > a_1 + a_2 t, \\ 0 & \text{when} \quad e^{bt} < a_1 + a_2 t, \end{cases}
\end{aligned} \tag{11}$$

maximizes $\mathbb{E}_{\theta'}[\psi(T)]$ subject to the constraints $\mathbb{E}_{\theta_0}[\psi(T)] = \alpha$ and $\mathbb{E}_{\theta_0}[T\psi(T)] = \alpha\mathbb{E}_{\theta_0}[T]$.

But the region described in (11) is either one-sided or the complement of an interval. But by Lehmann and Romano [8, Theorem 3.4.1] a one-sided test has a strictly monotone power function violating $\beta'_{\psi_0}(\theta_0) = 0$. Thus,

$$\psi_0(T) = \begin{cases} 1 & \text{when} \quad T < c_1 \ \text{ or } \ T > c_2, \\ 0 & \text{when} \quad c_1 < T < c_2. \end{cases}$$

[8] Exercise 3 (HW1): Show this (Hint: Apply Lehmann and Romano [8, Theorem 2.7.1] with $\phi \equiv 1$ to find that $c(\theta)$ is continuous; then apply it again with ϕ denoting an arbitrary critical function).

[9] Here we show that the assumption in Theorem 1.5-(iv) holds. Let

$$M \equiv \{(\mathbb{E}_{\theta_0}[\psi(T)], \mathbb{E}_{\theta_0}[T\psi(T)]) : \psi(T) \text{ is a critical function}\}.$$

Then M is convex and contains $\{(u, u\mathbb{E}_{\theta_0}T) : 0 < u < 1\}$. Also M contains points (α, v) with $v > \alpha\mathbb{E}_{\theta_0}T$; since, by Lehmann and Romano [8, Problem 18 of Chapter 3], there exist tests (UMP one-sided ones) having $\beta'(\theta_0) > 0$. Likewise M contains points (α, v) with $v < \alpha\mathbb{E}_{\theta_0}T$. Hence $(\alpha, \alpha\mathbb{E}_{\theta_0})$ is an interior point of M.

Since this test does not depend on $\theta' \neq \theta_0$, it is the UMP within the class of level α tests subject to (10). This test is unbiased, as is seen by comparing it with $\phi(x) \equiv \alpha$. It is then also UMP unbiased, since the class of tests satisfying (10) includes all unbiased tests. Hence ψ_0 is UMPU level α.

If T is distributed symmetrically about some point a under θ_0, then any test ψ symmetric about a that satisfies $\mathbb{E}_{\theta_0}[\psi(T)] = \alpha$ will also satisfy

$$\mathbb{E}_{\theta_0}[T\psi(T)] = \mathbb{E}_{\theta_0}[(T-a)\psi(T)] + a\mathbb{E}_{\theta_0}[\psi(T)] = 0 + a\alpha = \alpha\mathbb{E}_{\theta_0}[T],$$

automatically. □

1.6 UMPU tests in higher dimensions

Suppose that we have data $X \sim P_\theta$ where $\theta \in \Theta$ (here Θ is a subset of the Euclidean space). Consider testing

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1.$$

If the power function $\beta_\phi(\cdot)$ of an unbiased level α test ϕ is *continuous*, then $\beta_\phi(\theta) \leq \alpha$ for $\theta \in \overline{\Theta}_0$ (the closure of Θ_0) and $\beta_\phi(\theta) \geq \alpha$ for $\theta \in \overline{\Theta}_1$ (the closure of Θ_1). If we take $\Theta_B := \overline{\Theta}_0 \cap \overline{\Theta}_1$, the common boundary of Θ_0 and Θ_1, then

$$\beta_\phi(\theta) = \alpha \qquad \text{for } \theta \in \Theta_B. \tag{12}$$

Test functions ϕ satisfying (12) are called *similar on the boundary* (SOB).

Lemma 1.10. Suppose that the distributions $\{P_\theta\}_{\theta\in\Theta}$ are such that the power function of every test is continuous. Suppose that ϕ_0 is UMP among all tests satisfying (12) and is level α. Then ϕ_0 is a UMPU level α test.

Proof. The degenerate test that equals α regardless of the observed data is SOB level α. Since ϕ_0 has better power, $\beta_{\phi_0}(\theta) \geq \alpha$, for all $\theta \in \Theta_1$. As ϕ_0 is level α (i.e., $\beta_{\phi_0}(\theta) \leq \alpha$ for all $\theta \in \Theta_0$) ϕ_0 is unbiased.

Take a competing test ϕ which is level α and unbiased. Since its power function is continuous it is SOB level α. Then $\beta_\phi \leq \beta_{\phi_0}$ on Θ_1 because ϕ_0 is uniformly most powerful among all SOB tests. □

The tests we develop use conditioning to reduce to the univariate case. Part of why this works is that the tests have the structure in the following definition.

Definition 1.11. Suppose that T is sufficient for the subfamily $\mathcal{P}_B := \{P_\theta : \theta \in \Theta_B\}$. A test ϕ function is said to have *Neyman structure* w.r.t. T if[10]

$$\mathbb{E}_\theta[\phi(X)|T = t] = \alpha, \qquad \text{for a.e. } t\ (\mathcal{P}^T),\ \forall\, \theta \in \Theta_B. \tag{13}$$

[10] A statement is said to hold a.e. $\mathcal{P}$ if it holds except on a set N with $P(N) = 0$ for all $P \in \mathcal{P}$.

where $\mathcal{P}^{\mathcal{T}} := \{P_\theta^T : \theta \in \Theta_B\}$ and P_θ^T is the distribution of T under θ.

Remark 1.1. If ϕ has Neyman structure w.r.t. T, then ϕ is SOB. This easily follows from the fact that

$$\mathbb{E}_\theta[\phi(X)] = \mathbb{E}_\theta\big[\mathbb{E}[\phi(X)|T]\big] = \alpha, \qquad \forall\, \theta \in \Theta_B.$$

Lemma 1.12. If T is complete and sufficient for $\{P_\theta : \theta \in \Theta_B\}$, then every SOB test has Neyman structure.

Proof. Let ϕ be a SOB level α test and define $\psi(T) = \mathbb{E}[\phi(X)|T]$ (as T is sufficient, ψ does not dependent on $\theta \in \Theta_B$). Now

$$\mathbb{E}_\theta[\psi(T) - \alpha] = \mathbb{E}_\theta[\mathbb{E}[\phi(X)|T]] - \alpha = \mathbb{E}_\theta[\phi(X)] - \alpha = 0, \qquad \forall\, \theta \in \Theta_B,$$

and hence, by completeness $\psi(T) - \alpha = 0$ a.e., for all $\theta \in \Theta_B$. Hence ϕ has Neyman structure w.r.t. T. □

Remark 1.2. Suppose that:

(i) All critical functions have continuous power functions. Note that this is always true for exponential families.

(ii) T is complete sufficient for $\mathcal{P}_B = \{P_\theta : \theta \in \Theta_B\}$ (actually boundedly complete suffices; see Lehmann and Romano [8, Theorem 4.3.2]). Lehmann and Romano [8, Theorem 4.3.1] allows us to check (ii) for exponential families.

Then all unbiased tests are SOB and all SOB tests have Neyman structure (by Lemmas 1.10 and 1.12). Thus if we can find a UMP Neyman structure test ϕ_0 and we can show that ϕ_0 is unbiased, then ϕ_0 is UMPU.

Why is it easier to find UMP Neyman structure tests? Neyman structure tests are characterized by having conditional probability of rejection equal to α on each surface $T = t$. But the distribution on each such surface is independent of $\theta \in \Theta_B$ because T is sufficient for $\mathcal{P}_B$. Thus the problem has been reduced to testing a one parameter hypothesis for each fixed value of t; and in many problems we can easily find the most powerful test of this simple hypothesis (see e.g., Theorem 1.9).

1.6.1 Application to general exponential families

Suppose that X has distribution following an exponential family $\mathcal{P} = \{P_{\theta,\eta}\}_{(\theta,\eta)\in\Theta}$ with density

$$p_{\theta,\eta}(x) = c(\theta,\eta)\exp\left[\theta U(x) + \sum_{i=1}^{k} \eta_i T_i(x)\right] h(x) \tag{14}$$

w.r.t. a σ-finite dominating measure μ on some subset $\mathcal{X}$, where θ is univariate, $\eta = (\eta_1, \ldots, \eta_k)$ and $T = (T_1, \ldots, T_k)$, and the parameter space Θ is convex, has dimension $k+1$ and contains an interior point θ_0.

Goal: Find a UMPU test for

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0. \tag{15}$$

Idea: Consider the conditional distribution of U given T.

We know that the conditional distribution of U given T form a one-parameter exponential family with canonical parameter θ (independent of η)[11]; see e.g., Lehmann and Romano [8, Lemma 2.7.2]. Theorem 1.9 gives a UMPU conditional test of (15).

Theorem 1.14. If the exponential family (14) is of full rank[12] and Θ is open, then ϕ_0 given by

$$\phi_0(u,t) = \begin{cases} 1 & \text{if} \quad u < c_1(t) \ \text{ or } \ u > c_2(t), \\ \gamma_i(t) & \text{if} \quad u = c_i(t), \\ 0 & \text{if} \quad c_1(t) < u < c_2(t). \end{cases} \tag{17}$$

with the c_i's and γ_i's determined by (for a.e. t)

$$\mathbb{E}_{\theta_0,\eta}[\phi_0(U,T)|T=t] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\theta_0,\eta}[U\phi_0(U,T)|T=t] = \alpha\mathbb{E}_{\theta_0,\eta}[U|T=t] \tag{18}$$

is a UMPU test of (15).

Proof. We make the following observations:

(i) First note that the conditions on the exponential family ensure that the densities $\{P_{\theta_0,\eta}\}_\eta$ form a full rank exponential family with T as a complete[13] sufficient statistic.

(ii) When $T = t$ is given, U is the only remaining variable and the conditional distribution of U given $T = t$ is an exponential family with the form (16) (note that this conditional distribution does not depend on η).

[11]

Lemma 1.13 (Lemma 2.7.2 of Lehmann and Romano (2005))**.** Let X be distributed according to the exponential family (14). Then there exists a measure ν_t on $\mathbb{R}$ such that the conditional distribution of U given $T = t$ is an exponential family of the form

$$dP_{\theta\cdot}^{U|t}(u) = C_t(\theta)e^{\theta u}d\nu_t(u), \tag{16}$$

and hence in particular, is independent of η.

[12]The exponential family (14) is said to be of *full rank* if the interior of Θ is not empty and the sufficient statistics do not satisfy a linear constraint.

[13]In an exponential family of full rank, T is complete.

Claim 1: ϕ_0 is level α SOB (and unbiased).

From observation (ii), given $T = t$, by Theorem 1.9, ϕ_0 (as in (17)) is such that $\mathbb{E}_{\theta_0,\eta}[\phi_0|T = t] = \alpha$ for all t, so by smoothing $\mathbb{E}_{\theta_0,\eta}[\phi_0] = \alpha$ (thus ϕ_0 is level α SOB). Moreover, by Theorem 1.9, $\mathbb{E}_{\theta,\eta}[\phi_0] \geq \alpha$, by comparing with the degenerate level α test.

Take a competing test ϕ which is level α and unbiased. As before, we will restrict attention to tests of the form $\phi(x) \equiv \psi(U(x), T(x))$ based on the sufficient statistic $(U(X), T(X))$.

Claim 2: ϕ satisfies the two constraints:

$$\mathbb{E}_{\theta_0,\eta}[\phi|T] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\theta_0,\eta}[U(X)\phi(X)|T] = \alpha\mathbb{E}_{\theta_0,\eta}[U(X)|T]. \tag{19}$$

Since the power function of ϕ is continuous it is SOB level α. Thus, by Lemma 1.12, ϕ has Neyman structure, i.e., $\mathbb{E}_{\theta_0,\eta}[\phi|T = t] = \alpha$ for a.e. t (by observation (i) above).

The derivative $\frac{\partial}{\partial\theta}\beta_\phi(\theta,\eta)\Big|_{\theta=\theta_0}$ must be zero, and thus[14]

$$\mathbb{E}_{\theta_0,\eta}[U(X)\phi(X) - \alpha U(X)] = 0.$$

Conditioning on T, we have

$$0 = \mathbb{E}_{\theta_0,\eta}\Big[\mathbb{E}[U(X)\phi(X) - \alpha U(X)|T]\Big],$$

and since T is complete for the family of distributions with $\theta = \theta_0$, this implies that

$$\mathbb{E}[U(X)\phi(X) - \alpha U(X)|T] = 0 \quad \Rightarrow \quad \mathbb{E}_{\theta_0,\eta}[U(X)\phi(X)|T] = \alpha\mathbb{E}_{\theta_0,\eta}[U(X)|T].$$

Claim 3: ϕ_0 is UMP for testing (15) among all tests satisfying (19).

The power of a test $\phi(X) \equiv \psi(U(X), T(X))$ against an alternative (θ, η) is

$$\mathbb{E}_{\theta,\eta}[\psi(U,T)] = \mathbb{E}_{\theta,\eta}\left[\mathbb{E}_\theta[\psi(U,T)|T]\right].$$

As the conditional distribution of U given $T = t$ does not depend on η, one therefore maximizes the overall power by maximizing the power of the conditional test, $\mathbb{E}_\theta[\psi(U,T)|T = t]$, separately for each t.

[14] By (14), as $\frac{\partial}{\partial\theta}p_{\theta,\eta}(x) = \exp\left[\theta U(x) + \sum_{i=1}^k \eta_i T_i(x)\right] h(x)\left\{\frac{\partial}{\partial\theta}c(\theta,\eta) + U(x)c(\theta,\eta)\right\}$, we have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\beta_\phi(\theta,\eta) &= \frac{\partial}{\partial\theta}\int \phi(x)p_{\theta,\eta}(x)d\mu(x) \\
&= \frac{c'(\theta,\eta)}{c(\theta,\eta)}\mathbb{E}_{\theta,\eta}[\phi(X)] + \mathbb{E}[U(X)\phi(X)] \\
&= \{-\mathbb{E}_{\theta,\eta}[U(X)]\}\mathbb{E}_{\theta,\eta}[\phi(X)] + \mathbb{E}_{\theta,\eta}[U(X)\phi(X)]
\end{aligned}$$

since, with $\psi_0 \equiv \alpha$, $0 = \beta'_{\psi_0}(\theta,\eta) = \alpha\{c'(\theta,\eta)/c(\theta,\eta) + \mathbb{E}_{\theta,\eta}[U(X)]\}$ (which implies that $c'(\theta,\eta)/c(\theta,\eta) = -\mathbb{E}_{\theta,\eta}[U(X)]$). Thus,

$$0 = \frac{\partial}{\partial\theta}\beta_\phi(\theta,\eta)\Big|_{\theta=\theta_0} = \mathbb{E}_{\theta_0,\eta}[U(X)\phi(X)] - \alpha\mathbb{E}_{\theta_0,\eta}[U(X)].$$

Using observation (ii), given $T = t$, from the proof of Theorem 1.9 (see e.g., (10) and (11); a consequence of the generalized NP lemma), ϕ_0 (as in (17)) maximizes the conditional power against any $\theta \neq \theta_0$ subject to (19), and thus

$$\mathbb{E}_\theta[\phi_0(U,T)|T] \geq \mathbb{E}_\theta[\psi(U,T)|T] \quad \Rightarrow \quad \mathbb{E}_{\theta,\eta}[\phi_0(U,T)] \geq \mathbb{E}_{\theta,\eta}[\psi(U,T)] = \mathbb{E}_{\theta,\eta}[\phi(X)].$$

Thus ϕ_0 is uniformly most powerful unbiased. □

1.6.2 The t-test

Suppose that $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$. Let $X = (X_1, \ldots, X_n)$ and consider testing

$$H_0 : \mu \leq 0 \qquad \text{versus} \qquad H_1 : \mu > 0.$$

Letting $x = (x_1, \ldots, x_n)$, the joint density of the data is

$$\frac{1}{(\sqrt{2\pi})^n} \exp\left[\frac{\mu}{\sigma^2}U(x) - \frac{1}{2\sigma^2}T(x) - \frac{n\mu^2}{2\sigma^2} - n\log\sigma\right],$$

with $U(x) := \sum_{i=1}^n x_i$ and $T(x) = x_1^2 + \ldots + x_n^2$. This has form (14) with $\theta = \mu/\sigma^2$ and $\eta = -1/(2\sigma^2)$. Note that the hypotheses expressed using the canonical parameters are

$$H_0 : \theta \leq 0 \qquad \text{versus} \qquad H_1 : \theta > 0.$$

To proceed we need the conditional distribution of U given $T = t$ when $\mu = 0$[15]. As $U = \mathbf{1}^\top X$ (here $\mathbf{1}$ denotes a column of 1s) and $T = \|X\|^2$, we will study the distribution of $X|T$.

Note that the family of normal distributions with $\mu = 0$ is an exponential family with complete sufficient statistic T. Also, if we define $Z = X/\sigma$, so that $Z_1, \ldots, Z_n$ are i.i.d. standard normal, or $Z \sim N(0, I_n)$, then $W = X/\|X\| = Z/\|Z\|$ is ancillary. By Basu's theorem, T and W are independent. Because $\|X\| = \sqrt{T}, X = W\sqrt{T}$, and using independence between T and W, for any measurable function $h(\cdot)$,

$$\mathbb{E}[h(X)|T = t] = \mathbb{E}[h(W\sqrt{t})|T = t] = \mathbb{E}[h(W\sqrt{t})].$$

[15] Exercise 4 (HW1): Show that in the general exponential family setting of Section 1.6.1, if we consider testing

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0,$$

a UMP unbiased test is given by

$$\phi_0(u,t) = \begin{cases} 1 & \text{if} \quad u > c(t), \\ \gamma(t) & \text{if} \quad u = c(t), \\ 0 & \text{if} \quad u < c(t). \end{cases} \tag{20}$$

with the $c(\cdot)$ and $\gamma(\cdot)$ determined by (for a.e. t) $\mathbb{E}_{\theta_0,\eta}[\phi_0(U,T)|T = t] = \alpha$.

This shows that

$$X|T = t \sim W\sqrt{t}.$$

The vector W is said to be uniformly distributed on the unit sphere[16].

Using the above, since $U = \mathbf{1}^\top X$,

$$\mathbb{P}_{0,\sigma^2}[U > c(t)|T = t] = \mathbb{P}_{0,\sigma^2}\left[\mathbf{1}^\top W > \frac{c(t)}{\|X\|}\Big|T = t\right] = \mathbb{P}\left[\mathbf{1}^\top W > \frac{c(t)}{\sqrt{t}}\right].$$

This equals α if we take $c(t)/\sqrt{t} = q$, the upper α-th quantile for $\mathbf{1}^\top W$. Thus the uniformly most powerful unbiased test rejects H_0 if

$$\frac{U}{\sqrt{T}} > q.$$

Although it may not be apparent, this is equivalent to the usual test based on the t-statistic, as

$$t = \frac{\bar{X}}{s/\sqrt{n}} = \frac{U/\sqrt{n}}{\sqrt{(T - U^2/n)/(n-1)}} = \frac{\sqrt{n-1}\,U/\sqrt{T}}{\sqrt{n - U^2/T}} = g\left(\frac{U}{\sqrt{T}}\right),$$

where $\bar{X} = U/n$, $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1) = [T - U^2/n]/(n-1)$, and $g(v) = \frac{\sqrt{n-1}v}{\sqrt{n-v^2}}$.

The function $g(\cdot)$ here is strictly increasing (on $(-\sqrt{n}, \sqrt{n})$), and so

$$\frac{U}{\sqrt{T}} > q \qquad \text{if and only if} \qquad t > g(q).$$

When $\mu = 0$, t has the t-distribution on $n-1$ degrees of freedom, and so level α is achieved taking $g(q) = t_{\alpha,n-1}$, the upper α-th quantile of this distribution. So our test then rejects H_0 when

$$t > t_{\alpha,n-1}. \tag{21}$$

Example 1.15 (Two-sample t-test)**.** Exercise 5 (HW1): Suppose that we have data $X_1, \ldots, X_m$ i.i.d. $N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_n$ i.i.d. $N(\mu_Y, \sigma^2)$, where μ_X, μ_Y and $\sigma^2 > 0$ are unknown. Find UMP unbiased test for testing the hypothesis

$$H_0 : \mu_X = \mu_Y \qquad \text{versus} \qquad H_1 : \mu_Y > \mu_X.$$

This testing procedure naturally arises when comparing a treatment with a control to see if the treatment has an effect.

[16]Note that if A is an arbitrary orthogonal matrix (i.e., $AA^\top = I_n$), then $AZ \sim N(0, AA^\top) = N(0, I_n)$. Also $\|AZ\|^2 = (AZ)^\top(AZ) = Z^\top A^\top AZ = Z^\top Z = \|Z\|^2$. Thus Z and AZ have the same length and distribution. Then,

$$AW = \frac{AZ}{\|Z\|} = \frac{AZ}{\|AZ\|} \sim \frac{Z}{\|Z\|} = W.$$

So W and AW have the same distribution, which shows that the uniform distribution on the unit sphere (in $\mathbb{R}^n$) is invariant under orthogonal transformations. In fact, this is the only probability distribution on the unit sphere that is invariant under orthogonal transformations.

Example 1.16 (Comparing two Poisson distributions)**.** Exercise 6 (HW1): Suppose that we have independent data $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ where $\lambda, \mu > 0$ are unknown. It is natural to consider testing the hypothesis:

$$H_0 : \mu \leq a\lambda \qquad \text{versus} \qquad H_1 : \mu > a\lambda,$$

for some known a (e.g., $a = 1$). Using the framework of (14) (using the parameter of interest $\theta = \log(\mu/\lambda)$) find a UMP unbiased test for the above hypothesis. Describe the test and how to compute the critical value.

Example 1.17 (Fisher's exact test)**.** Exercise 7 (HW1): Fisher's exact test is a statistical significance test used in the analysis of contingency tables[17]. Suppose that we have two binary variables X and Y, each taking values 0 or 1. The goal is to test for the independence between X and Y. Thus, if

$$p_{ij} = \mathbb{P}(X = i, Y = j), \qquad \text{for } \ i, j = 0, 1,$$

given i.i.d. data $\{(X_k, Y_k)\}_{k=1}^n$ from the model, the goal is to test

$$H_0 : p_{00} = p_{0\cdot} p_{\cdot 0} \qquad \text{versus} \qquad H_1 : p_{00} \neq p_{0\cdot} p_{\cdot 0}, \tag{22}$$

where $p_{0\cdot} = \mathbb{P}(X = 0)$ and $p_{\cdot 0} = \mathbb{P}(Y = 0)$. The joint density of the data is

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_n = y_n) = \Pi_{i=0}^1 \Pi_{j=0}^1 p_{ij}^{n_{ij}},$$

where $n_{ij} = \#\{k : x_k = i, y_k = j\}$. If we take $N_{ij} = \#\{k : X_k = i, Y_k = j\}$ then $N = (N_{00}, N_{01}, N_{10}, N_{11})$ is a sufficient statistic. It is convenient to introduce new variables $U = N_{00}, T_1 = N_{00} + N_{01}$, and $T_2 = N_{00} + N_{10}$. There is a one-to-one relation between N and variables $T := (T_1, T_2)$ and U and thus, the distribution of (U, T_1, T_2) belongs to a full rank exponential family. Let $\theta = \log\left(\frac{p_{00}p_{11}}{p_{01}p_{10}}\right)$. Using Theorem 1.14 find the UMP unbiased test for testing (22) (which can be expressed as $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$). To describe the test in a more explicit fashion, we need the conditional distribution of U given $T = t$ when $\theta = 0$[18]. Show that U given $(T_1, T_2) = (t_1, t_2)$ follows a hypergeometric distribution, which arises in sampling theory.

1.6.3 Statistics independent of a sufficient statistic

A general expression for the UMP unbiased tests of the hypotheses $H_0 : \theta = \theta_0$ in the exponential family (14) was given in Theorem 1.14. However, this turns out to be inconvenient in the applications to normal and certain other families of continuous distributions, with

[17] It is named after its inventor, Ronald Fisher, and is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., p-value) can be calculated exactly.

[18] This distribution does not depend on η.

which we shall be concerned in the present chapter. In these applications, the tests can be given a more convenient form, in which they no longer appear as conditional tests in terms of U given $T = t$, but are expressed unconditionally in terms of a single test statistic. This is summarized in the following result.

Theorem 1.18. Suppose that the distribution of X is given by (14) and that $V = h(U, T)$ is independent of T when $\theta = \theta_0$. Then ϕ^* given by

$$\phi^*(v) = \begin{cases} 1 & \text{if} \quad v < C_1 \ \text{ or } \ u > C_2, \\ \gamma_i(t) & \text{if} \quad v = C_i, \\ 0 & \text{if} \quad C_1 < v < C_2. \end{cases} \tag{23}$$

with the C_i's and γ_i's determined by

$$\mathbb{E}_{\theta_0,\eta}[\phi^*(V)] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\theta_0}[V\phi^*(V)] = \alpha\mathbb{E}_{\theta_0}[V] \tag{24}$$

is UMP unbiased for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ provided

$$h(u, t) = a(t)u + b(t) \qquad \text{with} \quad a(t) > 0.$$

Proof. The test given in (17) is equivalent to (23) with constants C_i's and γ_i's determined by $\mathbb{E}_{\theta_0}[\phi_0(V, t)|t] = \alpha$ and

$$\mathbb{E}_{\theta_0}\left[\phi_0(V, t)\frac{V - b(t)}{a(t)}|t\right] = \alpha\mathbb{E}_{\theta_0}\left[\frac{V - b(t)}{a(t)}|t\right],$$

which reduces to $\mathbb{E}_{\theta_0}[V\phi_0(V, t)|t] = \alpha\mathbb{E}_{\theta_0}[V|t]$. Since V is independent of T for $\theta = \theta_0$, so are the C_i's and γ_i's as was to be proved. □

See Lehmann and Romano [8, Section 5.1] for the form of the UMP unbiased test when testing for one-sided alternatives as in Section 1.6.2.

1.7 Permutation tests

In Example 1.15 we compared two distributions, assuming normality of each population. For non-normal distributions however, the above method will not guarantee the level condition (for small m and n).

In the following we study a method that would yield an *exact* level α unbiased test when the two distributions have densities $f(\cdot)$ and $f(\cdot - \Delta)$, for unknown $f(\cdot)$ and Δ. The joint density of the data then has the form

$$p_\Delta(\mathbf{x}, \mathbf{y}) = f(x_1) \cdots f(x_m) f(y_1 - \Delta) \cdots f(y_n - \Delta), \qquad \text{where} \ \ f \in \mathcal{F} \tag{25}$$

for $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, and $\mathcal{F}$ can be taken to be the family of all probability densities that are continuous a.e. We consider testing

$$H_0 : \Delta = 0 \qquad \text{versus} \qquad H_1 : \Delta > 0.$$

Unbiasedness of a test $\phi(\cdot)$ in this case implies that, for all $f \in \mathcal{F}$,

$$\int_{\mathbb{R}^{m+n}} \phi(\mathbf{x}, \mathbf{y}) p_0(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \alpha, \tag{26}$$

where $p_0(\mathbf{x}, \mathbf{y})$ is defined as in (25) with $\Delta = 0$. In the following result we provide an easily verifiable equivalent condition for (26). Let $N = m + n$ and let $Z_i = X_i$ if $i = 1, \ldots, m$, and $Z_i = Y_{m-i}$ if $i = m + 1, \ldots, N$.

Theorem 1.19. If $\mathcal{F}$ is the family of all probability densities that are continuous a.e., then (26) holds for $f \in \mathcal{F}$ iff

$$\frac{1}{N!} \sum_{\mathbf{z}' \in S(\mathbf{z})} \phi(\mathbf{z}') = \alpha \qquad \text{a.e.}, \tag{27}$$

where $S(\mathbf{z})$ is the set of points obtained by permuting the coordinates of $\mathbf{z} = (z_1, \ldots, z_N) \in \mathbb{R}^N$ in all $N!$ possible ways.

Proof. Note that the set of order statistics $T(\mathbf{Z}) = (Z_{(1)}, \ldots, Z_{(N)})$ is a complete sufficient statistic for $\mathcal{F}$ (see e.g., Lehmann and Romano [8, Example 4.3.4]). Then, a necessary and sufficient condition for (26) is (see e.g., the proof of Lemma 1.12)

$$\mathbb{E}[\phi(\mathbf{Z}) | T(\mathbf{Z})] = \alpha \qquad \text{a.e.}$$

Note that $S(\mathbf{z}) = \{\mathbf{z}' : T(\mathbf{z}') = T(\mathbf{z})\}$. It follows that the conditional distribution of $\mathbf{Z}$ given $T(\mathbf{Z}) = T(\mathbf{z})$ assigns probability $1/N!$ to each of the $N!$ points of $S(\mathbf{z})$, thereby completing the proof. □

We shall now determine the test which, subject to (27), maximizes the power against a fixed alternative (25) or more generally against an alternative with arbitrary fixed alternative $h(\cdot)$. Thus, we want to test

$$H_0 : Z_1, \ldots, Z_N \text{ are i.i.d. } f \qquad \text{versus} \qquad H_1 : (Z_1, \ldots, Z_N) \sim h(\cdot).$$

The power of a test ϕ is $\int \phi(\mathbf{z}) h(\mathbf{z}) d\mathbf{z} = \mathbb{E}_h \left[\mathbb{E}_h[\phi(\mathbf{Z}) | T] \right]$. Since the conditional densities under the composite null hypothesis and under the simple alternative h are

$$p_0(\mathbf{z}|\mathbf{t}) = \frac{1}{N!} \qquad \text{and} \qquad p_1(\mathbf{z}|\mathbf{t}) = \frac{h(\mathbf{z})}{\sum_{\mathbf{z}' \in S(\mathbf{t})} h(\mathbf{z}')}, \qquad \text{for } \mathbf{z} \in S(\mathbf{t}),$$

the conditional power is $\mathbb{E}_h[\phi(\mathbf{Z}) | T = \mathbf{t}] = \sum_{\mathbf{z} \in S(\mathbf{t})} \phi(\mathbf{z}) p_1(\mathbf{z}|\mathbf{t})$, where $\mathbf{t} = (t_1, \ldots, t_N)$ such that $t_1 < t_2 < \ldots < t_N$. It is enough to maximize the conditional power for each $\mathbf{t}$ subject

to the constraint (27). By the NP lemma, this is achieved by rejecting $\mathbf{z} \in S(\mathbf{t})$ for large values of $\frac{p_1(\mathbf{z}|\mathbf{t})}{p_0(\mathbf{z}|\mathbf{t})}$. Thus the most powerful test is given by

$$\phi_0(\mathbf{z}) = \begin{cases} 1 & \text{if} \quad h(\mathbf{z}) > c(T(\mathbf{z})), \\ \gamma & \text{if} \quad h(\mathbf{z}) = c(T(\mathbf{z})), \\ 0 & \text{if} \quad h(\mathbf{z}) < c(T(\mathbf{z})). \end{cases} \tag{28}$$

Given $\mathbf{z}$ such that $T(\mathbf{z}) = \mathbf{t}$, to carry out the test, the $N!$ points of the set $S(\mathbf{t})$ are ordered according to the values of the density $h(\cdot)$. The hypothesis is rejected for the k largest values and with probability γ for the $(k+1)$'st value, where k and γ are defined by $k + \gamma = \alpha N!$ (assuming that the $k, (k+1)$'st and $(k+2)$'nd largest values are distinct).

Consider now in particular the alternatives (25). The most powerful permutation test is seen to depend on Δ and the f, and is therefore not UMP.

Example 1.20 (Permutation t-test (Exercise 8 (HW1))). Of special interest is the class of normal alternatives with common variance, i.e., $f \equiv N(\mu_X, \sigma^2)$ in (25). The most powerful permutation test, which turns out to be independent of μ_X, σ^2, and Δ[19], is appropriate when approximate normality is suspected but the assumption is not felt to be reliable. It may then be desirable to control the size of the test at level α regardless of the form of the densities f and to have the test unbiased against all alternatives (25). However, among the class of tests satisfying these broad restrictions it is natural to make the selection so as to maximize the power against the type of alternative one expects to encounter, that is, against the normal alternatives, i.e.,

$$h(\mathbf{z}) = (\sqrt{2\pi\sigma^2})^{-N} \exp\left[-\frac{1}{2\sigma^2} \Big\{ \sum_{i=1}^{m} (z_i - \mu_X)^2 + \sum_{i=m+1}^{N} (z_i - \mu_X - \Delta)^2 \Big\} \right].$$

1.8 Exercises

9. Lehmann and Romano [8, Problem 3.55].

10. Lehmann and Romano [8, Problem 3.62].

11. Suppose X is a random variable with density

$$p_\theta(x) := e^{\eta(\theta)T(x) - A(\theta)} h(x)$$

w.r.t. some dominating σ-finite measure μ on $\mathcal{X}$, and $\theta \in \Theta$, where Θ is an interval

[19] Show this. Also, show that the rejection region in this case has the form of the t-test in which the constant cutoff point $t_{\alpha,n-1}$ in (21) is replaced by a random one (under appropriate conditions it can be shown (you do not have to show this) that the difference between the random and the constant cut-off is small in an asymptotic sense). Further, show that this test is unbiased.

in $\mathbb{R}$. Assume that η is $\mathcal{C}^\infty$ and $\eta'(\theta) > 0$. Fix $\theta_0 \in \Theta$, and let

$$M := \left\{ \left(\mathbb{E}_{\theta_0} \psi(X), \mathbb{E}_{\theta_0} \psi(X) T(X) \right) : \psi : \mathcal{X} \mapsto [0,1] \right\} \subset \mathbb{R}^2.$$

Show that for any $\alpha \in (0,1)$, the point $(\alpha, \alpha \mathbb{E}_{\theta_0} T(X))$ is an interior point in M.

12. Lehmann and Romano [8, Problem 5.5].

13. Lehmann and Romano [8, Problem 5.11].

14. Lehmann and Romano [8, Problem 5.15].

2 Smooth parametric families

As seen in the last chapter, the finite sample theory of optimality for hypothesis testing applied only to rather special parametric families, primarily exponential families (and group families; see Lehmann and Romano [8, Chapter 6]). On the other hand, as we will see in this chapter, asymptotic optimality will apply more generally to parametric families satisfying smoothness conditions.

2.1 Local asymptotic normality and quadratic mean differentiability

Consider a parametric model $\{P_\theta : \theta \in \Theta\}$, where, throughout this section, Θ is assumed to be an open subset of $\mathbb{R}^k$ $(k \geq 1)$. The probability measures P_θ are defined on some measurable space $\mathcal{X}$. We assume that each P_θ is absolutely continuous w.r.t. a σ-finite measure μ, and set $p_\theta(x) = dP_\theta(x)/d\mu(x)$, for $x \in \mathcal{X}$.

We consider smooth parametric models. To motivate the smoothness condition given in Definition 2.3 below, consider the case of n i.i.d. random variables $X_1, \ldots, X_n$ and the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ (possibly dependent on n). The MP test rejects when the *loglikelihood ratio* statistic

$$\log[p_n(\theta_1)/p_n(\theta_0)]$$

is sufficiently large, where $p_n(\theta) := \Pi_{i=1}^n p_\theta(X_i)$ denotes the *likelihood function*. In the following two examples we illustrate the behavior of the loglikelihood ratio for two "simple" models.

Example 2.1 (Normal location model)**.** Suppose that P_θ is $N(\theta, \sigma^2)$, where σ^2 is known. Then,

$$\log[p_n(\theta_1)/p_n(\theta_0)] = \frac{n}{\sigma^2}[(\theta_1 - \theta_0)\bar{X}_n - \frac{1}{2}(\theta_1^2 - \theta_0^2)], \tag{29}$$

where $\bar{X}_n := \sum_{i=1}^n X_i/n$. By the weak law of large number (LLN), under $H_0 : \theta = \theta_0$,

$$(\theta_1 - \theta_0)\bar{X}_n - \frac{1}{2}(\theta_1^2 - \theta_0^2) \quad \xrightarrow{p} \quad (\theta_1 - \theta_0)\theta_0 - \frac{1}{2}(\theta_1^2 - \theta_0^2) = -\frac{1}{2}(\theta_1 - \theta_0)^2,$$

and so $\log[p_n(\theta_1)/p_n(\theta_0)] \xrightarrow{p} -\infty$[20].

A more useful result is obtained if θ_1 in (29) is replaced by $\theta_0 + hn^{-1/2}$ (local alternative); here $h \in \mathbb{R}$ is fixed. We then find,

$$\log[p_n(\theta_0 + hn^{-1/2})/p_n(\theta_0)] = \frac{h\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma^2} - \frac{h^2}{2\sigma^2} = hZ_n - \frac{h^2}{2\sigma^2}, \tag{30}$$

[20] It can also be shown that the power of the test (for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$) will converge to 1.

where $Z_n = \sqrt{n}(\bar{X}_n - \theta_0)/\sigma^2 \sim N(0, 1/\sigma^2)$. Notice that the expansion (30) is a linear function of Z_n and a simple quadratic function of h, with the coefficient of h^2 nonrandom. Furthermore, $\log[p_n(\theta_0 + hn^{-1/2})/p_n(\theta_0)]$ is distributed as $N(-h^2/(2\sigma^2), h^2/\sigma^2)$[21] under $H_0 : \theta = \theta_0$ for every n[22].

Example 2.2 (One-parameter exponential family)**.** Let $X_1, \ldots, X_n$ be i.i.d. having density $p_\theta(x) = \exp[\theta T(x) - A(\theta)]$ (for $x \in \mathcal{X}$) w.r.t. a σ-finite measure μ. Assume θ_0 lies in the interior of the natural parameter space $\Theta \subset \mathbb{R}$. Then[23], for a fixed $h \in \mathbb{R}$,

$$\log[p_n(\theta_0 + hn^{-1/2})/p_n(\theta_0)] = hZ_n - \frac{h^2}{2}A''(\theta_0) + o_p(1), \tag{31}$$

where, under $H_0 : \theta = \theta_0$,

$$Z_n := n^{-1/2} \sum_{i=1}^{n} \left\{T(X_i) - \mathbb{E}_{\theta_0}[T(X_i)]\right\} \xrightarrow{d} N(0, A''(\theta_0)).$$

Thus, the loglikelihood ratio (31) behaves asymptotically like the loglikelihood ratio (30) from a normal location model in Example 2.1 with $\sigma^2 = [A''(\theta_0)]^{-1}$. This is usually referred to as *local asymptotic normality*[24]. As we will see, such approximations allow one to deduce asymptotic optimality properties for the exponential model (or any model whose likelihood ratios satisfy an appropriate generalization of (31)) from optimality properties of the simple normal location model.

2.1.1 Local asymptotic normality: Heuristics

We would like to obtain an approximate result like (31) for more general families; in the following we give a heuristic proof sketch. Let $\ell_\theta(x) = \log p_\theta(x)$ be "twice differentiable" w.r.t. $\theta \in \Theta \subset \mathbb{R}^k$, and can be approximated by its second order Taylor series, i.e., for every fixed x,

$$\ell_{\theta+h}(x) = \ell_\theta(x) + h^\top \dot{\ell}_\theta(x) + \frac{1}{2}h^\top \ddot{\ell}_\theta(x)h + o_x(|h|^2).$$

Here $|\cdot|$ denotes the usual Euclidean norm and the subscript x in the remainder term is a reminder of the fact that this term depends on x as well as on h (and on θ). Then, using

[21] The relationship that the mean is the negative of half the variance will play a key role in the sequel.

[22] Moreover, it can be shown that the power of the test will converge to a number strictly between 0 and 1 (depending on h).

[23] Recall that $\mathbb{E}_{\theta_0}[T(X_i)] = A'(\theta_0)$, and $\mathrm{Var}_{\theta_0}(T(X_i)) = A''(\theta_0)$, and by a Taylor expansion,

$$n[A(\theta_0 + hn^{-1/2}) - A(\theta_0)] = hn^{1/2}A'(\theta_0) + \frac{h^2}{2}A''(\theta_0) + o(1).$$

[24] The notion of local asymptotic normality was introduced by Le Cam.

the above expansion,

$$\begin{aligned}\log[p_n(\theta+h_n)/p_n(\theta)] &= \sum_{i=1}^n [\log p_{\theta+h_n}(X_i) - \log p_\theta(X_i)] \\ &= h_n^\top \sum_{i=1}^n \dot{\ell}_\theta(X_i) + \frac{1}{2} h_n^\top \sum_{i=1}^n \ddot{\ell}_\theta(X_i) h_n + o(n|h_n|^2). \qquad (32)\end{aligned}$$

Note that $\dot{\ell}_\theta$ is called the *score function*. For $X \sim P_\theta$ (and for ℓ_θ satisfying regularity conditions), we have

1. The score function has mean zero: $P_\theta \dot{\ell}_\theta \equiv \mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$[25].
2. The mean curvature of the loglikelihood is the negative Fisher information: $P_\theta \ddot{\ell}_\theta = -I_\theta$, where $I_\theta := P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$.

Thus[26],

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n \dot{\ell}_\theta(X_i) &\overset{P_\theta}{\rightsquigarrow} N(0, I_\theta), \\ \frac{1}{n} \sum_{i=1}^n \ddot{\ell}_\theta(X_i) &\overset{P_\theta}{\to} -I_\theta. \end{aligned}$$

So, if $\sqrt{n} h_n \to h$, then using (32),

$$\begin{aligned}\log[p_n(\theta+h_n)/p_n(\theta)] &= (\sqrt{n}h_n)^\top \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i) + \frac{1}{2}(\sqrt{n}h_n)^\top \frac{1}{n} \sum_{i=1}^n \ddot{\ell}_\theta(X_i)(\sqrt{n}h_n) + o(1) \\ &\overset{P_\theta}{\rightsquigarrow} N(-\frac{1}{2} h^\top I_\theta h, h^\top I_\theta h).\end{aligned}$$

This behavior is known as *local asymptotic normality* (see Remark 2.2 below for a more detailed explanation).

What conditions make the above argument rigorous? As we will now see that a weaker condition than twice differentiability suffices; in fact, $\theta \mapsto \sqrt{p_\theta(x)}$ differentiable for most x suffices.

2.1.2 Quadratic mean differentiable

Classical smoothness conditions on the function $p_\theta(x)$ usually assume that, for fixed x, $p_\theta(x)$ is differentiable in θ at $\theta_0 \in \Theta \subset \mathbb{R}^k$; i.e., for some function $\dot{p}_\theta(x)$,

$$p_{\theta_0+h}(x) - p_{\theta_0}(x) - h^\top \dot{p}_\theta(x) = o(|h|)$$

[25] We can justified this if we are allowed to interchanging the order of integration and differentiation in the integral: $\int \dot{\ell}_{\theta_0}(x) p_{\theta_0}(x) d\mu(x) = \int \dot{p}_{\theta_0}(x) d\mu(x) = \frac{\partial}{\partial\theta}\left(\int p_\theta(x) d\mu(x)\right) = 0$.

[26] Here $\rightsquigarrow$ denotes weak convergence (i.e., convergence in distribution), and $\overset{P_\theta}{\rightsquigarrow}$ indicates that the true data distribution is P_θ.

as $|h| \to 0$. In addition, higher order differentiability is typically assumed with further assumptions on the remainder terms (e.g., twice continuous differentiability of $\log p_\theta(x)$ w.r.t θ along with a dominated second derivative).

In order to avoid such strong assumptions, it turns out to be useful to work with square roots of densities. For fixed x, differentiability of $p_\theta^{1/2}(x)$ at $\theta = \theta_0$ requires the existence of a function $\eta(x, \theta_0)$ such that

$$R(x, \theta_0, h) \equiv p_{\theta_0+h}^{1/2}(x) - p_{\theta_0}^{1/2}(x) - h^\top \eta(x, \theta_0) = o(|h|).$$

To obtain a weaker, more generally applicable condition, we will not require $R^2(x, \theta_0, h) = o(|h|^2)$ for every x, but we will impose the condition that $R^2(\cdot, \theta_0, h)$ averaged w.r.t. μ is $o(|h|^2)$. Let $L^2(\mu)$ denote the space of functions $g : \mathcal{X} \to \mathbb{R}$ such that $\|g\|^2_{L^2(\mu)} := \int g^2(x) d\mu(x) < \infty$. The convenience of working with square roots of densities is due to the fact that $p_\theta^{1/2}(\cdot) \in L^2(\mu)$ and, more importantly, it is an element with norm 1 (a fact first exploited by Le Cam).

Definition 2.3 (Quadratic mean differentiable)**.** The family $\mathcal{P} := \{P_\theta, \theta \in \Theta \subset \mathbb{R}^k\}$ is quadratic mean differentiable (abbreviated QMD) at θ_0 if there exists a vector of real-valued functions $\dot{\ell}_{\theta_0}(\cdot) = (\dot{\ell}_{\theta_0,1}(\cdot), \ldots, \dot{\ell}_{\theta_0,k}(\cdot))^\top$ such that

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2} h^\top \dot{\ell}_{\theta_0}(x) \sqrt{p_{\theta_0}(x)} \right]^2 d\mu(x) = o(|h|^2) \tag{33}$$

as $|h| \to 0$. Whenever (33) holds, we will call $\dot{\ell}_{\theta_0}$ as the *score function*.

In other words, if $\mathcal{P}$ satisfies QMD at θ_0, then we have: If $\{\theta_n\}$ is a sequence converging to θ_0 then

$$\sqrt{p_{\theta_n}(x)} = \sqrt{p_{\theta_0}(x)} + \frac{1}{2}(\theta_n - \theta_0)^\top \dot{\ell}_{\theta_0}(x) \sqrt{p_{\theta_0}(x)} + r_{\theta_n}(x) \tag{34}$$

for all $x \in \mathcal{X}$ and n with

$$\lim_{n\to\infty} \frac{\|r_{\theta_n}\|_{L^2(\mu)}}{|\theta_n - \theta_0|} = 0. \tag{35}$$

It is natural to ask why does the term $\frac{1}{2} h^\top \dot{\ell}_{\theta_0}(x) \sqrt{p_{\theta_0}(x)}$ arise in (33)? Note that if $p_\theta(x)$ is differentiable in θ at θ_0, then

$$\nabla_\theta \sqrt{p_\theta}\Big|_{\theta=\theta_0} = \frac{1}{2} \frac{\nabla_\theta p_\theta}{\sqrt{p_\theta}}\Big|_{\theta=\theta_0} = \frac{1}{2} \sqrt{p_{\theta_0}} \frac{\nabla_\theta p_{\theta_0}}{p_{\theta_0}} = \frac{1}{2} \sqrt{p_{\theta_0}}\, \nabla_\theta \ell_{\theta_0} = \frac{1}{2} \sqrt{p_{\theta_0}}\, \dot{\ell}_{\theta_0}.$$

From above it can be seen that, if smoothness conditions hold, $\dot{\ell}_{\theta_0}$ is indeed the usual score function, i.e., $\dot{\ell}_{\theta_0}(x) = \frac{\partial}{\partial\theta} \log p_\theta(x)\Big|_{\theta=\theta_0}$.

Le Cam showed that, under QMD, classical asymptotic results in statistics (such as the asymptotic normality of maximum likelihood estimators) can be proved without requiring the density $\theta \mapsto p_\theta(x)$ to be twice (or thrice) differentiable at θ_0. One appears to get the

benefit of the quadratic expansion without paying the twice-differentiability price usually demanded by such a Taylor expansion.

Example 2.4 (Normal distribution)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$, where $\theta \in \Theta = \mathbb{R}$. Show that this family is QMD. (Exercise 1 (HW2))

Example 2.5 (Double exponential)**.** For the model $p_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$ $(x, \theta \in \mathbb{R})$, differentiability fails at the point $\theta = x$, but this model satisfies QMD as (33) holds. (Exercise 2 (HW2))

Example 2.6 (Uniform distribution)**.** Suppose that $P_\theta = \text{Uniform}([0, \theta])$, for $\theta \in \Theta = (0, \infty)$. This model is not QMD as (for $\theta_0 > 0$)

$$\begin{aligned}
\int \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\right]^2 d\mu &\geq \int_{\theta_0}^{\theta_0+h} \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\right]^2 d\mu \\
&= \int_{\theta_0}^{\theta_0+h} \left[\frac{1}{\sqrt{\theta_0+h}} - 0 - \frac{1}{2}h \cdot \dot{\ell}_{\theta_0} \cdot 0\right]^2 d\mu \\
&= \frac{h}{\theta_0+h} = O(|h|).
\end{aligned}$$

Definition 2.7 (Fisher information)**.** For the QMD family $\mathcal{P}$ with score function $\dot{\ell}_{\theta_0}$, we define the Fisher Information matrix to be the matrix $I_\theta \in \mathbb{R}^{k\times k}$ with (i, j) entry $\int_{\mathcal{X}} \dot{\ell}_{\theta_0,i}(x)\dot{\ell}_{\theta_0,j}(x)p_{\theta_0}(x)\, d\mu(x)$. Thus,

$$I_{\theta_0} = \int (\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}})(\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}})^\top d\mu = \int_{\mathcal{X}} \dot{\ell}_{\theta_0}(x)\dot{\ell}_{\theta_0}(x)^\top p_{\theta_0}(x)\, d\mu(x).$$

A standard fact about the classical score function is that its expectation w.r.t. the probability measure P_{θ_0} equals zero. The classical proof for this involves interchanging the order of differentiation w.r.t. θ (see Section 2.1.1). The following lemma shows that the QMD assumption implies this fact directly.

Lemma 2.8. Suppose $\mathcal{P}$ satisfies QMD at θ_0 with score function $\dot{\ell}_{\theta_0}$. Then

$$\int \dot{\ell}_{\theta_0}(x)p_{\theta_0}(x)d\mu(x) = 0, \tag{36}$$

and the Fisher information matrix I_{θ_0} exists and is well-defined.

Proof. For $i = 1, \ldots, k$, we will first show that the i'th diagonal entry of the Fisher information matrix I_{θ_0} is finite, i.e., $\int \dot{\ell}^2_{\theta_0,i}(x)p_{\theta_0}(x)d\mu(x) < \infty$ (for $i = 1, \ldots, k$), which will show that I_{θ_0} is well-defined. This follows from the fact that for any $h \in \mathbb{R}^k$, taking $\theta_n = \theta_0 + hn^{-1/2}$ in (34), we get

$$\int \left[n^{1/2}\left\{\sqrt{p_{\theta_0+hn^{-1/2}}(x)} - \sqrt{p_{\theta_0}(x)}\right\} - \frac{1}{2}h^\top\dot{\ell}_{\theta_0}(x)\sqrt{p_{\theta_0}(x)}\right]^2 d\mu(x) \to 0,$$

as $n \to \infty$. Take $g_n := n^{1/2}(\sqrt{p_{\theta_0+hn^{-1/2}}} - \sqrt{p_{\theta_0}})$ and $g := \frac{1}{2}h^\top \dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}$. We will show that $g \in L^2(\mu)$. We are given that $\int (g_n - g)^2 d\mu \to 0$ as $n \to \infty$. Observe that $\{g_n\}_{n\geq 1}$ is Cauchy (as $\int (g_n - g_m)^2 d\mu = \int [(g_n - g) - (g_m - g)]^2 d\mu \leq 2\int (g_n - g)^2 d\mu + 2\int (g_m - g)^2 d\mu \to 0$ as $m, n \to \infty$). Hence, g_n has a limit q (say) in $L^2(\mu)$ (as $L^2(\mu)$ is a complete space). Hence $\int (g - q)^2 d\mu \leq 2\int (g - g_n)^2 d\mu + \int (g_n - q)^2 d\mu \to 0$ as $n \to \infty$, and thus, $\int (g - q)^2 d\mu = 0$. Therefore, $g = q$ in $L^2(\mu)$ and hence g belongs to $L^2(\mu)$.

By taking h to be the vector of zeros except for 1 in the i'th component shows that $\int \dot{\ell}^2_{\theta_0,i}(x) p_{\theta_0}(x) d\mu(x) < \infty$.

Let θ_n be a sequence converging to θ_0 as $n \to \infty$. By the QMD representation, we can write (34) with the remainder term r_{θ_n} satisfying (35). Note then that

$$1 = \int p_{\theta_n} d\mu = \int \left(\sqrt{p}_{\theta_0} + \frac{1}{2}(\theta_n - \theta_0)^\top \dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} + r_{\theta_n}\right)^2 d\mu.$$

We now expand the square in the right hand side above which will lead to six terms. One of the terms equals $\int p_{\theta_0} d\mu = 1$ which cancels with the left hand side. We thus obtain

$$\begin{aligned} 0 \;=\; & (\theta_n - \theta_0)^\top \int \dot{\ell}_{\theta_0} p_{\theta_0} d\mu + 2\int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \frac{1}{4}(\theta_n - \theta_0)^\top [\int \dot{\ell}_{\theta_0}\dot{\ell}^\top_{\theta_0} p_{\theta_0} d\mu]\,(\theta_n - \theta_0) \\ & + (\theta_n - \theta_0)^\top \int \dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \int r^2_{\theta_n} d\mu \end{aligned} \tag{37}$$

The first term in the right hand side above is clearly $O(|\theta_n - \theta_0|)$ in absolute value. The third term is $O(|\theta_n - \theta_0|^2)$. The final term (by (35)) equals $o(|\theta_n - \theta_0|^2)$. The remaining two terms (second and fourth) can be controlled via the Cauchy-Schwarz inequality as

$$2\int \sqrt{p_{\theta_0}}|r_{\theta_n}| d\mu \leq 2\sqrt{\int p_{\theta_0} d\mu}\sqrt{\int r^2_{\theta_n} d\mu} = 2\sqrt{\int r^2_{\theta_n} d\mu} = 2o(|\theta_n - \theta_0|)$$

by (35) and

$$\left|(\theta_n - \theta_0)^\top \int \dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}} r_{\theta_n} d\mu\right| \leq \sqrt{(\theta_n - \theta_0)^\top I_{\theta_0}(\theta_n - \theta_0)}\sqrt{\int r^2_{\theta_n} d\mu} = o(|\theta_n - \theta_0|^2)$$

again by (35) (and using the fact that $I_{\theta_0} = \int \dot{\ell}_{\theta_0}\dot{\ell}^\top_{\theta_0} p_{\theta_0} d\mu$). It is clear therefore that the leading term on the right hand side in (37) is the first term. By dividing the equation (37) through by $|\theta_n - \theta_0|$ and letting $n \to \infty$, we deduce (36). □

Remark 2.1 (Why work with square-roots of densities?)**.** The argument used in the proof of Lemma 2.8 above leads to an interesting and important fact involving $\dot{\ell}_{\theta_0}$ and the Fisher information. Because $\int \dot{\ell}_{\theta_0} p_{\theta_0} d\mu = 0$, we can plug this into (37) to obtain (also using the fact that the last two terms in (37) are $o(|\theta_n - \theta_0|^2)$):

$$2\int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu + \frac{1}{4}(\theta_n - \theta_0)^\top I_{\theta_0}(\theta_n - \theta_0) + o(|\theta_n - \theta_0|^2) = 0.$$

Thus, we obtain

$$2\int \sqrt{p_{\theta_0}} r_{\theta_n} d\mu = -\frac{1}{4}(\theta_n - \theta_0)^\top I_{\theta_0}(\theta_n - \theta_0) + o(|\theta_n - \theta_0|^2). \tag{38}$$

This fact is crucial for establishing that QMD implies local asymptotic normality (see Theorem 2.10 in Section 2.1.3). The interesting aspect about (38) is the following. The statement (35) implies that $\|r_{\theta_n}\|_{L^2(\mu)} = o(|\theta_n - \theta_0|)$. Therefore, if we use the Cauchy-Schwarz inequality on the left hand side in (38), we obtain that the left hand side is $o(|\theta_n - \theta_0|)$. But the equality above implies that the right hand side is $O(|\theta_n - \theta_0|^2)$ which is a much stronger conclusion that what can be derived from Cauchy-Schwarz inequality. Therefore $\int r_{\theta_n}\sqrt{p_{\theta_0}} d\mu$ is much smaller in comparison to the $L^2(\mu)$ norm of r_{θ_n}. Pollard [12] attributes this phenomenon to the fact that the functions $\sqrt{p_{\theta_n}}$ in $L^2(\mu)$ all have norm one (this is clear from the above proof of (38)) and argues that this is the main reason behind the magic of the QMD[27].

To establish QMD of specific models requires a convergence theorem for integrals. Usually one proceeds by showing differentiability of the map $\theta \mapsto p_\theta(x)$ for a.e. x plus μ-equi-integrability (e.g., domination). The following theorem takes care of many examples.

Theorem 2.9. Let Θ be an open subset of $\mathbb{R}^k$. Assume that the map $\theta \mapsto s_\theta(x) \equiv \sqrt{p_\theta(x)}$ is a continuously differentiable function of θ in some neighborhood of θ_0, for μ-a.e. x. If the elements of the matrix $I_\theta = \int (\dot{p}_\theta/p_\theta)(\dot{p}_\theta^\top/p_\theta)p_\theta\, d\mu$ are well-defined and continuous at θ_0, then the map $\theta \mapsto \sqrt{p_\theta}$ is QMD at θ_0 with $\dot{\ell}_{\theta_0}$ given by $\dot{p}_{\theta_0}/p_{\theta_0}$.

Proof. We will prove this when $\Theta \subset \mathbb{R}$ (i.e., $k = 1$).

By the chain rule, the map $\theta \mapsto p_\theta(x) = s_\theta^2(x)$ is differentiable in θ is some neighborhood of θ_0 with gradient $\dot{p}_\theta = 2s_\theta\dot{s}_\theta$, for μ-a.e. x. Because s_θ is nonnegative, its gradient $\dot{s}_\theta$ at a point at which $s_\theta = 0$ must be zero. Conclude that we can write $\dot{s}_\theta = \frac{1}{2}(\dot{p}_\theta/p_\theta)\sqrt{p_\theta}$, where the quotient $\dot{p}_\theta/p_\theta$ may be defined arbitrarily if $p_\theta = 0$. By assumption, the map $\theta \mapsto I_\theta = 4\int(\dot{s}_\theta)^2\, d\mu$ is continuous.

Because the map $\theta \mapsto s_\theta(x)$ is continuously differentiable around θ_0, the difference $s_{\theta_0+h}(x) - s_{\theta_0}(x)$ can be written as the integral $h\int_0^1 \dot{s}_{\theta_0+uh}(x)du$ of its derivative. Integrating over all x w.r.t. μ, and using Cauchy-Schwarz's inequality we have

$$\begin{aligned}\frac{1}{h^2}\int (s_{\theta_0+h}(x) - s_{\theta_0}(x))^2 d\mu(x) &= \int \left(\int_0^1 \dot{s}_{\theta_0+uh}(x)du\right)^2 d\mu(x) \\ &\le \int\int_0^1 (\dot{s}_{\theta_0+uh}(x))^2 du\, d\mu(x) = \frac{1}{4}\int_0^1 I_{\theta_0+uh}\, du,\end{aligned}$$

[27] A part of this subsection and the next is taken from Adityanand Guntuboyina's lecture notes (see https://www.stat.berkeley.edu/~aditya/resources/FullNotes210BSpring2018.pdf), which in turn is taken from Pollard [12].

where the last equality follows by Fubini's theorem and the definition of I_θ. For $h \to 0$ the right side converges to I_θ by the continuity of the map $\theta \mapsto I_\theta$ in a neighborhood of θ_0.

We will now use the following result: Suppose that f_n and f are arbitrary measurable functions such that $f_n \to f$ μ-a.e. (for some measure μ) and $\limsup \int |f_n|^2 d\mu \le \int |f|^2 d\mu < \infty$. Then $\int |f_n - f|^2 d\mu \to 0$ (see van der Vaart [15, Proposition 2.29]). This is known as Vitali's theorem.

Now consider a sequence $\{h_n\}_{n\ge1} \subset \mathbb{R}$ such that $h_n \to 0$. Let $f_n := (s_{\theta_0+h_n} - s_{\theta_0})/h_n$ and $f := \dot{s}_{\theta_0} = \dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}$. By the differentiability of the map $\theta \mapsto s_\theta$ at $\theta = \theta_0$, $f_n(x) \to f(x)$ for μ-a.e. x. Thus, by an application of Vitali's theorem yields,

$$\int \left[\frac{s_{\theta_0+h_n}(x) - s_{\theta_0}(x)}{h_n} - \dot{\ell}_{\theta_0}(x)\sqrt{p_{\theta_0}(x)}\right]^2 d\mu(x) \to 0,$$

thereby completing the proof. □

2.1.3 Local asymptotic normality

Theorem 2.10. Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$, where Θ is an open set, satisfies QMD at $\theta_0 \in \Theta$ with score function $\dot{\ell}_{\theta_0}$ and Fisher information matrix I_{θ_0}. Then for every fixed $h \in \mathbb{R}^k$, we have

$$\left|\sum_{i=1}^n \log\frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} - \frac{h^\top}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + \frac{1}{2}h^\top I_{\theta_0} h\right| \xrightarrow{P_{\theta_0}} 0 \qquad \text{as } n \to \infty.$$

Equivalently, the conclusion of the above theorem can be written

$$\sum_{i=1}^n \log\frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = \frac{h^\top}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2}h^\top I_{\theta_0} h + o_{P_{\theta_0}}(1) \qquad \text{as } n \to \infty. \tag{39}$$

We say that $\mathcal{P}$ satisfies the *local asymptotic normality* (LAN) property at θ_0 if the above holds for every $h \in \mathbb{R}^k$. To see this, note first that, by the CLT, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \xrightarrow{d} Z$$

where $Z \sim N(0, I_{\theta_0})$. Therefore, as a consequence of (39), we obtain that for every $h \in \mathbb{R}^k$,

$$\sum_{i=1}^n \log\frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \xrightarrow{d} h^\top Z - \frac{1}{2}h^\top I_{\theta_0} h \overset{d}{=} N\left(-\frac{1}{2}h^\top I_{\theta_0} h, h^\top I_{\theta_0} h\right).$$

Remark 2.2 (Why is this called local asymptotic normality?)**.** Now consider the estimation problem where we have one observation Y whose density belongs to the family $\{\mathcal{Q}_h : h \in \mathbb{R}^k\}$

were $\mathcal{Q}_h$ has the density q_h which is the density of the normal distribution with mean h and variance $I_{\theta_0}^{-1}$ (which we assume exist here). It is easy to see then that

$$\log \frac{q_h(Y)}{q_0(Y)} \sim h^\top I_{\theta_0} Y - \frac{1}{2} h^\top I_{\theta_0} h \qquad \text{under } Y \sim N(0, I_{\theta_0}^{-1}).$$

Therefore (39) effectively says that the likelihood ratios of $\{P_\theta, \theta \in \Theta\}$ (which can be arbitrary as long as $\mathcal{P}$ satisfies QMD) behave like the likelihood ratios of a normal experiment $\{Q_h : h \in \mathbb{R}^k\}$ where $Q_h = N(h, I_{\theta_0}^{-1})$. Hence asymptotically around θ_0 at the scale $n^{-1/2}$, the original statistical problem $\mathcal{P}$ becomes a normal mean estimation problem. This is why (39) is referred to as LAN.

We shall now prove Theorem 2.10.

Proof of Theorem 2.10. All expectations and probabilities in this proof are w.r.t. the probability measure P_{θ_0}. Write

$$L_n := \sum_{i=1}^n \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = 2 \sum_{i=1}^n \log \sqrt{\frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)}} = 2 \sum_{i=1}^n \log (1 + W_{ni})$$

where

$$W_{ni} := \sqrt{\frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)}} - 1.$$

We will use the fact that

$$\log(1+y) = y - \frac{y^2}{2} + \frac{1}{2} y^2 \beta(y) \qquad \text{where } \lim_{y \to 0} \beta(y) = 0$$

or equivalently, $\beta(y) = o(1)$ as $y \to 0$. This gives

$$L_n = 2 \sum_{i=1}^n W_{ni} - \sum_{i=1}^n W_{ni}^2 + \sum_{i=1}^n W_{ni}^2 \beta(W_{ni}).$$

Using the QMD representation (34), we can write

$$W_{ni} = \frac{\sqrt{p_{\theta_0 + hn^{-1/2}}(X_i)} - \sqrt{p_{\theta_0}(X_i)}}{\sqrt{p_{\theta_0}(X_i)}} = \frac{h^\top \dot{\ell}_{\theta_0}(X_i)}{2\sqrt{n}} + \frac{r_{\theta_0 + hn^{-1/2}}(X_i)}{\sqrt{p_{\theta_0}(X_i)}} = \frac{h^\top \dot{\ell}_{\theta_0}(X_i)}{2\sqrt{n}} + R_{ni} \tag{40}$$

where

$$R_{ni} = \frac{r_{\theta_0 + hn^{-1/2}}(X_i)}{\sqrt{p_{\theta_0}(X_i)}}.$$

We thus get

$$
\begin{aligned}
L_n &= \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}} + 2\sum_{i=1}^n R_{ni} - \sum_{i=1}^n \left(\frac{h^\top \dot{\ell}_{\theta_0}(X_i)}{2\sqrt{n}} + R_{ni} \right)^2 + \sum_{i=1}^n W_{ni}^2 \beta\left(W_{ni}\right) \\
&= \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}} + 2\sum_{i=1}^n R_{ni} - h^\top \frac{1}{4n} \left[\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \dot{\ell}_{\theta_0}(X_i)^\top \right] h \\
&\qquad - \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) R_{ni}}{\sqrt{n}} - \sum_{i=1}^n R_{ni}^2 + \sum_{i=1}^n W_{ni}^2 \beta\left(W_{ni}\right). \qquad (41)
\end{aligned}
$$

Observe now that by QMD, we know the following about the random variables R_{ni}:

$$
\mathbb{E}_{\theta_0}[R_{ni}^2] = \mathbb{E}_{\theta_0}\left[\frac{r_{\theta_0 + hn^{-1/2}}^2(X_i)}{p_{\theta_0}(X_i)} \right] = \left\| r_{\theta_0 + hn^{-1/2}}(\cdot) \right\|_{L^2(\mu)}^2 = o\left(\frac{|h|^2}{n} \right) = o(n^{-1}). \qquad (42)
$$

This gives that $\mathbb{E}_{\theta_0}\left[\sum_{i=1}^n R_{ni}^2\right] = o(1)$ and hence $\sum_{i=1}^n R_{ni}^2 \xrightarrow{p} 0$ (by Markov's inequality). Also, by the Cauchy-Schwarz inequality, we have

$$
\left| \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) R_{ni}}{\sqrt{n}} \right| \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (h^\top \dot{\ell}_{\theta_0}(X_i))^2} \sqrt{\sum_{i=1}^n R_{ni}^2} \xrightarrow{p} \sqrt{h^\top I_{\theta_0} h}\, \sqrt{0} = 0 \qquad (43)
$$

where we have used the weak law of large numbers. We thus have

$$
L_n = \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}} + 2\sum_{i=1}^n R_{ni} - h^\top \frac{1}{4n} \left[\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \dot{\ell}_{\theta_0}(X_i)^\top \right] h + \sum_{i=1}^n W_{ni}^2 \beta\left(W_{ni}\right) + o_{P_{\theta_0}}(1).
$$

We shall prove later that

$$
\sum_{i=1}^n W_{ni}^2 \beta\left(W_{ni}\right) = o_{P_{\theta_0}}(1), \qquad (44)
$$

so that we have

$$
L_n = \frac{h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}} + 2\sum_{i=1}^n R_{ni} - h^\top \frac{1}{4n} \left[\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \dot{\ell}_{\theta_0}(X_i)^\top \right] h + o_{P_{\theta_0}}(1).
$$

The third term in the right hand side above clearly converges to $-h^\top I_{\theta_0} h/4$ in probability (by the WLLN) so to complete the proof of Theorem 2.10, we only need to show that

$$
2\sum_{i=1}^n R_{ni} \xrightarrow{p} -\frac{1}{4} h^\top I(\theta_0) h. \qquad (45)
$$

For this, write

$$
2\sum_{i=1}^n R_{ni} = 2\sum_{i=1}^n \mathbb{E}_{\theta_0}[R_{ni}] + 2\sum_{i=1}^n (R_{ni} - \mathbb{E}_{\theta_0}[R_{ni}]).
$$

As, by (42),

$$\mathbb{E}\left(2\sum_{i=1}^{n}(R_{ni}-\mathbb{E}_{\theta_0}[R_{ni}])\right)^2 = 4\sum_{i=1}^{n}\mathrm{Var}(R_{ni}) \le 4\sum_{i=1}^{n}\mathbb{E}_{\theta_0}[R_{ni}^2] = o(1),$$

we get

$$2\sum_{i=1}^{n}R_{ni} = 2\sum_{i=1}^{n}\mathbb{E}_{\theta_0}[R_{ni}] + o_{P_{\theta_0}}(1) \qquad \text{as } n\to\infty. \tag{46}$$

Note that

$$2\sum_{i=1}^{n}\mathbb{E}_{\theta_0}[R_{ni}] = 2n\,\mathbb{E}_{\theta_0}[R_{n1}] = 2n\,\mathbb{E}_{\theta_0}\left[\frac{r_{\theta_0+hn^{-1/2}}(X_1)}{\sqrt{p_{\theta_0}(X_1)}}\right] = 2n\int r_{\theta_0+hn^{-1/2}}\sqrt{p_{\theta_0}}d\mu.$$

We shall now use the fact (38) which gives

$$2\int r_{\theta_0+hn^{-1/2}}\sqrt{p_{\theta_0}}d\mu = -\frac{1}{4n}h^\top I_{\theta_0}h + o(n^{-1})$$

so that

$$2\sum_{i=1}^{n}\mathbb{E}_{\theta_0}[R_{ni}] = -\frac{1}{4}h^\top I_{\theta_0}h + o(1).$$

Combining with (46), we obtain (45). To finish the proof of Theorem 2.10, we only need to verify (44). This is mainly a consequence of $\beta(y)=o(1)$ as $y\to 0$. Indeed,

$$\left|\sum_{i=1}^{n}W_{ni}^2\beta(W_{ni})\right| = \max_{1\le i\le n}|\beta(W_{ni})|\sum_{i=1}^{n}W_{ni}^2.$$

From the definition of W_{ni} in (40) (coupled with the expansion in (41) and (43) and the fact that $\sum_{i=1}^{n}R_{ni}^2 \xrightarrow{p} 0$), it follows that $\sum_{i=1}^{n}W_{ni}^2 \xrightarrow{p} \frac{1}{4}h^\top I_{\theta_0}h$. So, it suffices to show that $\max_{1\le i\le n}|\beta(W_{ni})| \xrightarrow{p} 0$ (under P_{θ_0}), which follows if we can show that (as $\lim_{y\to 0}\beta(y)=0$)

$$\max_{1\le i\le n}|W_{ni}| \xrightarrow{p} 0.$$

Using (40) it turns out that it is enough to show that

$$\max_{1\le i\le n}\left|\frac{h^\top\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right| = o_{P_{\theta_0}}(1) \quad \text{and} \quad \max_{1\le i\le n}|R_{ni}| = o_{P_{\theta_0}}(1). \tag{47}$$

We shall complete the proof now by proving the assertions in (47). For the first assertion in (47), write (for a fixed $\epsilon>0$),

$$\mathbb{P}_{\theta_0}\left(\max_{1\le i\le n}\left|\frac{h^\top\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right|>\epsilon\right) \le \sum_{i=1}^{n}\mathbb{P}_{\theta_0}\left(\left|\frac{h^\top\dot{\ell}_{\theta_0}(X_i)}{\sqrt{n}}\right|>\epsilon\right) = n\mathbb{P}_{\theta_0}\left(\left|\frac{h^\top\dot{\ell}_{\theta_0}(X_1)}{\sqrt{n}}\right|>\epsilon\right)$$
$$\le \frac{1}{\epsilon^2}\mathbb{E}_{\theta_0}\left[(h^\top\dot{\ell}_{\theta_0}(X_1))^2 I\left\{\left|\frac{h^\top\dot{\ell}_{\theta_0}(X_1)}{\sqrt{n}}\right|>\epsilon\right\}\right]$$

which converges to zero as $n \to \infty$ by the dominated convergence theorem. For the second assertion in (47), write

$$\mathbb{P}\left(\max_{1\le i\le n} |R_{ni}| > \epsilon\right) \le n\mathbb{P}\left(|R_{n1}| > \epsilon\right) \le \frac{n}{\epsilon^2}\mathbb{E}_{\theta_0}[R_{n1}^2] \to 0,$$

by (42). This completes the proof. □

2.2 Contiguity

The notion of contiguity is developed primarily as a technique for calculating the limiting distribution of a test statistic or power of a test function under an alternative sequence, especially when the limiting distribution under the null hypothesis is easy to obtain.

Contiguity is an "asymptotic" form of a probability measure Q being absolutely continuous w.r.t. another probability measure P. In order to motivate the concept, suppose P and Q are two probability measures on some measurable space $(\mathcal{X}, \mathcal{C})$. Assume that Q is *absolutely continuous* w.r.t. P. This means that $E \in \mathcal{C}$ and $P(E) = 0$ implies $Q(E) = 0$.

Suppose $T \equiv T(X)$ is a random vector from $\mathcal{X}$ to $\mathbb{R}^k$, such as an estimator, test statistic, or test function. How can one compute the distribution of T under Q if you know how to compute probabilities or expectations under P? Specifically, suppose it is required to compute $\mathbb{E}_Q[f(T)]$, where f is some measurable function from $\mathbb{R}^k$ to $\mathbb{R}$. Let p and q denote the densities of P and Q w.r.t. a common dominating measure μ. Then, assuming Q is absolutely continuous w.r.t. P, for any measurable f,

$$\mathbb{E}_Q[f(T(X))] = \int_{\mathcal{X}} f(T(x))dQ(x) = \int_{\mathcal{X}} f(T(x))\frac{q(x)}{p(x)}p(x)d\mu(x) = \mathbb{E}_P[f(T(X))L(X)],$$

where $L(X)$ is the usual *likelihood ratio statistic*, i.e., $L(x) = \frac{q(x)}{p(x)}$. Hence, the distribution of $T(X)$ under Q can be computed if the joint distribution of $(T(X), L(X))$ under P is known.

Contiguity is an asymptotic version of absolute continuity that permits an analogous asymptotic statement. Consider sequences of pairs of probabilities $\{P_n, Q_n\}$, where P_n and Q_n are probabilities on some measurable space $(\mathcal{X}_n, \mathcal{C}_n)$. Let $T_n : \mathcal{X}_n \to \mathbb{R}^k$ be some random vector. Suppose the asymptotic distribution of T_n under P_n is easily obtained, but the behavior of T_n under Q_n is also required. For example, if T_n represents a test function for testing P_n versus Q_n, the power of T_n is the expectation of T_n under Q_n. Contiguity provides a means of performing the required calculation. An example may help fix ideas.

Example 2.11 (Wilcoxon signed-rank statistic)**.** Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with unknown common density $f(\cdot - \theta)$, where $f(\cdot)$ is assumed symmetric about zero and $\theta \in \mathbb{R}$ in unknown. The problem is to test the null hypothesis $H_0 : \theta = 0$

against the alternative hypothesis $H_1 : \theta > 0$. Consider the Wilcoxon signed rank statistic defined by:

$$W_n = W_n(X_1, \ldots, X_n) = n^{-3/2} \sum_{i=1}^{n} R_{i,n}^{+} \text{sign}(X_i), \tag{48}$$

where $\text{sign}(X_i)$ is 1 if $X_i \geq 0$ and is -1 otherwise, and $R_{i,n}^{+}$ is the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$. We would reject H_0 when W_n is "large". Under the null hypothesis, the behavior of W_n is fairly easy to obtain. If $\theta = 0$, the variables $\text{sign}(X_i)$ are i.i.d., each 1 or -1 with probability 1/2, and are independent of the variables $R_{i,n}^{+}$[28]. In fact, the exact distribution of W_n is the *same* for all distributions with densities symmetric about 0. Thus, W_n is finite sample *distribution-free*. Thus, for finite n, critical values for W_n can be obtained exactly.

We can also study the asymptotic distribution of W_n under H_0, if we want to avoid simulations to find the critical value of the test. It is easy to see that $\mathbb{E}_{\theta=0}(W_n) = 0$. Define $\tilde{I}_k$ to be 1 if the k'th largest $|X_i|$ corresponds to a positive observation and -1 otherwise. Then, we have

$$\text{Var}_{\theta=0}(W_n) = \frac{1}{n^3} \text{Var}\left(\sum_{k=1}^{n} k \tilde{I}_k \right) = \frac{1}{n^3} \sum_{k=1}^{n} k^2 = \frac{1}{n^3} \frac{n(n+1)(2n+1)}{6} \to \frac{1}{3},$$

as $n \to \infty$. Not surprisingly, $W_n \xrightarrow{d} N(0, 1/3)$[29].

Thus, W_n is asymptotically normal with mean 0 and variance 1/3, and this is true whenever the underlying distribution has a symmetric density about 0. Hence, the test that rejects the null hypothesis if W_n exceeds $\frac{z_{1-\alpha}}{\sqrt{3}}$ has limiting level $1 - \alpha$ (here $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution).

Suppose now that we want to approximate the power of this test. The above argument does not generalize to even close alternatives since it heavily uses the fact that the variables are symmetric about zero. Contiguity provides a fairly simple means of attacking this problem, and we will reconsider this example later.

Definition 2.12 (Contiguity)**.** Let P_n and Q_n be probability distributions on $(\mathcal{X}_n, \mathcal{C}_n)$, for $n = 1, 2, \ldots$. We say that $\{Q_n\}$ is *contiguous* w.r.t. $\{P_n\}$, written $Q_n \lhd P_n$, if $P_n(A_n) \to 0$

[28] Exercise 3 (HW2): Find the joint distribution of $(\text{sign}(X_1), \ldots, \text{sign}(X_n), R_{1,n}^{+}, \ldots, R_{n,n}^{+})$ and deduce the independence between $(\text{sign}(X_1), \ldots, \text{sign}(X_n))$ and $(R_{1,n}^{+}, \ldots, R_{n,n}^{+})$ (when $\theta = 0$).

[29] To see why, note that we can show that (Exercise 4 (HW2): Show this.)

$$W_n - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \text{sign}(X_i) = o_p(1),$$

where $U_i = G(|X_i|)$ and G is the c.d.f. of $|X_i|$. Under the null hypothesis, U_i and $\text{sign}(X_i)$ are independent. Moreover, the random variables $U_i \text{sign}(X_i)$ are i.i.d., and so the CLT is applicable (see Lehmann and Romano [8, Problem 12.19]).

implies $Q_n(A_n) \to 0$ (as $n \to \infty$) for every sequence of sets $\{A_n\}_{n\geq 1}$ with $A_n \in \mathcal{C}_n$[30].

If $\{Q_n\}$ is *contiguous* to $\{P_n\}$, and $\{P_n\}$ is contiguous to $\{Q_n\}$, then we say the sequences $\{P_n\}$ and $\{Q_n\}$ are mutually contiguous, or just contiguous (and write $Q_n \lhd\rhd P_n$).

Example 2.13 (Normal location families)**.** Suppose that $P_n = N(0,1)$ and $Q_n = N(\theta_n, 1)$ with $\theta_n \to \theta \in \mathbb{R}$. Then, by using Theorem 2.15 (see below), we can show that $Q_n \lhd\rhd P_n$. Suppose now we assume that $\theta_n \to \infty$. Then taking $A_n = [\theta_n - 1, \theta_n + 1]$ shows that we do *not* have $Q_n \lhd P_n$. But notice that, regardless of the values of θ_n, Q_n is absolutely continuous w.r.t. P_n (i.e., $Q_n \ll P_n$) for all n.

Example 2.14 (Uniform location families)**.** Suppose that $P_n = \text{Uniform}([0,1])$, and $P_n = \text{Uniform}([0,\theta_n])$ where $\theta_n > 1$, and $\theta_n \to 1$ as $n \to \infty$. Then, using Theorem 2.15, we can show that $Q_n \lhd P_n$. But notice that Q_n is not absolutely continuous w.r.t. P_n.

We now would like a useful means of determining whether or not Q_n is contiguous to P_n. Suppose P_n and Q_n have densities p_n and q_n w.r.t. a dominating measure μ_n. For $x \in \mathcal{X}_n$, define the *likelihood ratio* of Q_n w.r.t. P_n by

$$L_n(x) = \begin{cases} \frac{q_n(x)}{p_n(x)} & \text{if } \; p_n(x) > 0 \\ +\infty & \text{if } \; p_n(x) = 0 < q_n(x) \\ 1 & \text{if } \; p_n(x) = q_n(x) = 0. \end{cases} \tag{49}$$

Note that L_n is regarded as an extended random variable, which means it is allowed to take on the value $+\infty$, at least under Q_n. Of course, under P_n, L_n is finite with probability one. Observe that

$$\mathbb{E}_{P_n}[L_n] = \int_{\mathcal{X}_n} L_n(x) p_n(x)\, d\mu_n(x) = \int\limits_{\{x \in \mathcal{X}_n : p_n(x) > 0\}} q_n(x)\, d\mu_n(x) = Q_n(\{x : p_n(x) > 0\}) \leq 1$$

with equality if and only if Q_n is absolutely continuous with respect to P_n[31]. Thus, the sequence of likelihood ratios L_n is uniformly tight[32] under P_n (this follows immediately from an application of Markov's inequality). By Prohorov's theorem[33], every subsequence has a further weakly converging subsequence. The next lemma (also known as Le Cam's first lemma) shows that the limit characterize contiguity[34].

[30] Compare this with *absolute continuity* of Q w.r.t. P (for probability distributions P and Q on $(\mathcal{X}, \mathcal{C})$) which means that $E \in \mathcal{C}$ and $P(E) = 0$ implies $Q(E) = 0$.

[31] Exercise 5 (HW2): Show this.

[32] Recall that a sequence of random variables $\{Y_n\}_{n\geq 1}$ is *uniformly tight* or just *tight* (i.e., $Y_n = O_p(1)$) if given any $\epsilon > 0$, $\exists\, M > 0$ such that $\mathbb{P}(|Y_n| \geq M) \leq \epsilon$ for all $n \geq 1$.

[33] Prohorov's theorem Lehmann and Romano [8, Theorem 11.2.15]: Suppose that $\{Y_n\}_{n\geq 1}$ is uniformly tight. Then $\exists$ a subsequence $\{n_j\}_{j\geq 1}$ and a distribution G such that $X_{n_j} \xrightarrow{d} G$.

[34] These are analogous to the characterizations we have for absolute continuity; see e.g., van der Vaart [15, Lemma 6.2].

Theorem 2.15 (Le Cam's first lemma)**.** Suppose that $\{P_n\}_{n\geq 1}$ and $\{Q_n\}_{n\geq 1}$ are sequences of probability distributions on $\{(\mathcal{X}_n, \mathcal{C}_n)\}_{n\geq 1}$ and let L_n be defined as in (49). Suppose $L_n \overset{P_n}{\rightsquigarrow} V$ (i.e., the distribution of L_n under P_n converges weakly to V). If $\mathbb{E}[V] = 1$ then Q_n is contiguous w.r.t. P_n.

Proof. Let G_n be the c.d.f. of L_n (under P_n) and G be the c.d.f. of V. Suppose that $P_n(E_n) = \alpha_n \to 0$. Let ϕ_n be a most powerful level α_n test of P_n versus Q_n. By the Neyman-Pearson Lemma, the test is of the form

$$\phi_n = \begin{cases} 1 & \text{if } L_n > k_n, \\ 0 & \text{if } L_n < k_n, \end{cases}$$

for some k_n chosen so the test is level α_n. Since ϕ_n is at least as powerful as the test that has rejection region E_n,

$$Q_n(E_n) \leq \int \phi_n dQ_n, \tag{50}$$

so it suffices to show the right side tends to zero. Now, for any $y < \infty$,

$$\begin{aligned} \int \phi_n dQ_n &= \int_{L_n \leq y} \phi_n dQ_n + \int_{L_n > y} \phi_n dQ_n \\ \leq \int_{L_n \leq y} \phi_n L_n dP_n + \int_{L_n > y} dQ_n &\leq y \int \phi_n dP_n + 1 - \int_{L_n \leq y} dQ_n \\ = y\alpha_n + 1 - \int_{L_n \leq y} L_n dP_n &= y\alpha_n + 1 - \int_0^y x dG_n(x). \end{aligned}$$

Fix any $\epsilon > 0$ and take y to be a continuity point of G with $\int_0^y x dG(x) > 1 - \epsilon/2$, which is possible since G has mean 1. As G_n converges weakly to G, $\int_0^y x dG_n(x) \to \int_0^y x dG(x)$[35]. Thus, for sufficiently large n, $1 - \int_0^y x dG_n(x) < \epsilon/2$, and $y\alpha_n < \epsilon/2$. Now, it follows that, for sufficiently large n, $\int \phi_n dQ_n < \epsilon$, which by (50) yields the desired result. □

The following result summarizes some equivalent characterizations of contiguity. The notation $\mathcal{L}(T|P)$ refers to the distribution (or law) of a random variable T under P.

Theorem 2.16. The following are equivalent characterizations of $\{Q_n\}$ being contiguous to $\{P_n\}$.

(i) For every sequence of real-valued random variables T_n such that $T_n \to 0$ in P_n-probability, it also follows that $T_n \to 0$ in Q_n-probability.

[35] To see why, construct $\tilde{L}_n \sim G_n$ and $\tilde{L} \sim G$ such that $\tilde{L}_n \overset{a.s}{\to} \tilde{L}$ (by the almost sure representation theorem) and then apply the DCT to $\tilde{L}_n \mathbf{1}_{[0,y]}(\tilde{L}_n)$ to conclude that

$$\mathbb{E}_{P_n}[L_n \mathbf{1}_{[0,y]}(L_n)] = \mathbb{E}[\tilde{L}_n \mathbf{1}_{[0,y]}(\tilde{L}_n)] \to \mathbb{E}[\tilde{L} \mathbf{1}_{[0,y]}(\tilde{L})] = \int_0^y x dG_n(x).$$

(ii) For every sequence T_n such that $\mathcal{L}(T_n|P_n)$ is tight, it also follows that $\mathcal{L}(T_n|Q_n)$ is tight.

(iii) If G is any limit point[36] of $\mathcal{L}(L_n|P_n)$, then G has mean 1.

Proof. See Lehmann and Romano [8, Theorem 12.3.2]. □

As will be seen in many important examples, loglikelihood ratios are typically asymptotically normally distributed, and the following corollary is useful.

Corollary 2.17 (Implication in LAN[37])**.** Suppose that P_n and Q_n are probability measures on arbitrary measurable spaces such that $\log L_n \overset{P_n}{\rightsquigarrow} N(\mu, \sigma^2)$. Then Q_n and P_n are mutually contiguous if and only if $\mu = -\sigma^2/2$.

Proof. To show $Q_n \lhd P_n$ let us apply Theorem 2.16-(iii). Let $e^Z \sim G$ where $Z \sim N(\mu, \sigma^2)$. Note that

$$\int x dG(x) = \mathbb{E}[e^Z] = e^{\mu+\sigma^2/2} = 1 \quad \Leftrightarrow \quad \mu = -\sigma^2/2.$$

Thus, $Q_n \lhd P_n$ if and only if $\mu = -\sigma^2/2$.

Exercise 6 (HW2): Now show that $P_n \lhd Q_n$. □

The following theorem solves the problem of obtaining a Q_n-limit law from a P_n-limit law that we posed at the start of this subsection.

Corollary 2.18 (Le Cam's third lemma)**.** Assume that, $(T_n, \log L_n) \overset{P_n}{\rightsquigarrow} (T, Z)$, where (T, Z) is bivariate normal with $\mathbb{E}(T) = \mu_1$, $\mathrm{Var}(T) = \sigma_1^2$, $\mathbb{E}(Z) = \mu_2$, $\mathrm{Var}(Z) = \sigma_2^2$ and $\mathrm{Cov}(T, Z) = \sigma_{1,2}$. Assume that $\mu_2 = -\sigma_2^2/2$, so that Q_n is contiguous to P_n. Then,

$$T_n \overset{Q_n}{\rightsquigarrow} N(\mu_1 + \sigma_{1,2}, \sigma_1^2).$$

Proof. See Lehmann and Romano [8, Corollary 12.3.2]. □

Example 2.19 (Asymptotically linear statistic)**.** Let $\{P_\theta : \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^k$ an open set, be QMD with corresponding densities $p_\theta(\cdot)$. By Corollary 2.17, in conjunction with Theorem 2.10, shows that $P^n_{\theta_0+hn^{-1/2}}$ and $P^n_{\theta_0}$ are mutually contiguous. In fact, the expansion (39) shows a lot more. For example, suppose an estimator (sequence) $\hat{\theta}_n$ is *asymptotically linear* in the following sense: under θ_0,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\theta_0}(X_i) + o_{P^n_{\theta_0}}(1), \tag{51}$$

[36] G is a limit point of a sequence G_n of distributions if G_{n_j} converges in distribution to G for some subsequence n_j.

[37] Recall that LAN for smooth parametric models implies that the loglikelihood ratio of local alternative to true parameter is asymptotically normal.

where $\mathbb{E}_{\theta_0}[\psi_{\theta_0}(X_1)] = 0$ and $\tau^2 := \mathrm{Var}_{\theta_0}(\psi_{\theta_0}(X_1)) < \infty$. Thus, under θ_0,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \tau^2). \tag{52}$$

Then, the joint behavior of $\hat{\theta}_n$ with the loglikelihood ratio satisfies

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \log L_n \end{bmatrix} = \begin{bmatrix} n^{-1/2} \sum_{i=1}^n \psi_{\theta_0}(X_i) \\ n^{-1/2} h^\top \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{2} h^\top I_{\theta_0} h \end{bmatrix} + o_{P^n_{\theta_0}}(1), \tag{53}$$

where $L_n = \frac{\Pi_{i=1}^n p_{\theta_0 + hn^{-1/2}}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)}$. By the bivariate CLT, this converges under θ_0 to a bivariate normal distribution with covariance

$$\sigma_{1,2} = \mathrm{Cov}_{\theta_0}(\psi_{\theta_0}(X_1), h^\top \dot{\ell}_{\theta_0}(X_1)). \tag{54}$$

Hence, under $P^n_{\theta_0 + hn^{-1/2}}$, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N(\sigma_{1,2}, \tau^2)$, by Corollary 2.18. It follows that, under $P^n_{\theta_0 + hn^{-1/2}}$,

$$\sqrt{n}[\hat{\theta}_n - (\theta_0 + hn^{-1/2})] \xrightarrow{d} N(\sigma_{1,2} - h, \tau^2).$$

Example 2.20 (Back to Example 2.11)**.** Recall the Wilcoxon signed-rank statistic W_n given by (48). Assume the underlying model is a location model $f(\cdot - \theta)$, where $f(\cdot)$ is assumed symmetric and unimodal[38] about zero. Assume $f'(x)$ exists for Lebesgue almost all x and

$$0 < I \equiv \int \frac{[f'(x)]^2}{f(x)} dx < \infty.$$

This model is QMD[39] and (39) holds with $\dot{\ell}_0(x) = -\frac{f'(x)}{f(x)}$. Under the null hypothesis $\theta = 0$, $W_n \xrightarrow{d} N(0, 1/3)$. Under the sequence of alternatives $\theta_n = hn^{-1/2}$, $W_n \xrightarrow{d} N(\sigma_{1,2}, 1/3)$, where $\sigma_{1,2}$ is given by (54) with $\theta_0 = 0$. In this case,

$$\sigma_{1,2} = \mathrm{Cov}_0 \left(U \mathrm{sign}(X), -h \frac{f'(X)}{f(X)} \right),$$

where $U = G(|X|)$ and G is the c.d.f. of $|X|$ when X has density $f(\cdot)$. So, $G(x) = 2F(x) - 1$, where F is the c.d.f. of X. By an integration by parts[40],

$$\sigma_{1,2} = -h\mathbb{E}_0 \left[G(|X|) \mathrm{sign}(X) \frac{f'(X)}{f(X)} \right] = 2h \int_{-\infty}^{\infty} f^2(x) dx.$$

Thus, under $\theta_n = hn^{-1/2}$,

$$W_n \xrightarrow{d} N\left(2h \int_{-\infty}^{\infty} f^2(x) dx, \frac{1}{3} \right).$$

[38] In statistics, a unimodal probability density is a probability density which has a single peak. Thus, here f is nonincreasing on either side of 0, as f is symmetric about 0.

[39] Exercise 7 (HW2): Show this.

[40] Exercise 8 (HW2): Show this.

Example 2.21 (Power of the Neyman-Pearson test)**.** Assume $\{P_\theta : \theta \in \Theta\}$ is QMD at θ_0, where Θ is an open subset of $\mathbb{R}^k$ and I_{θ_0} is nonsingular. Let p_θ be the corresponding density of P_θ. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_0 + hn^{-1/2}$ using the likelihood ratio statistic $L_n(h) \equiv \frac{\Pi_{i=1}^n p_{\theta_0+hn^{-1/2}}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)}$ based on n i.i.d. observations $X_1, \ldots, X_n$. Then, by Theorem 2.10, under $P_{\theta_0}^n$,

$$\log L_n(h) \xrightarrow{d} N\left(-\frac{\sigma_h^2}{2}, \sigma_h^2\right),$$

where $\sigma_h^2 = h^\top I_{\theta_0} h$.

Then, under $P_{\theta_0+hn^{-1/2}}^n$, $\log L_n(h)$ is asymptotically $N(\sigma_h^2/2, \sigma_h^2)$, by applying Corollary 2.18 with $T_n = \log L_n(h)$. Hence, the test that rejects when $\log L_n(h)$ exceeds $(-1/2)\sigma_h^2 + z_\alpha \sigma_h$ is asymptotically level α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_0 + hn^{-1/2}$, where z_α denotes the $1-\alpha$ quantile of $N(0,1)$. Then, the limiting power of this test sequence is $1 - \Phi(z_\alpha - \sigma_h)$.

2.3 Likelihood methods in parametric models

The above techniques will now be applied to classes of tests based on the likelihood function, namely the Wald, Rao, and likelihood ratio tests. This subsection is based on Lehmann and Romano [8, Chapter 12.4]; please read this book chapter carefully. I will only summarize some of main points below. Suppose that $X_1, \ldots, X_n$ are i.i.d. P_θ, taking values in $\mathcal{X}$ where $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ is a parametric family.

Theorem 2.22 (Asymptotic normality of the MLE)**.** Suppose that:

(i) $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ is QMD at θ_0 with nonsingular Fisher information matrix I_{θ_0}.

(ii) Let $X_1, \ldots, X_n$ are i.i.d. P_{θ_0}, taking values in $\mathcal{X}$.

(iii) Suppose further that there exists a measurable function $M : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_{\theta_0}[M^2(X_1)] < \infty$ such that, for ever θ_1 and θ_2 in a neighborhood of θ_0, we have

$$\left|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\right| \le M(x)|\theta_1 - \theta_2|.$$

(iv) Moreover, suppose that the MLE $\hat{\theta}_n$ is consistent for estimating θ_0.

Then,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_p(1). \tag{55}$$

As a consequence, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_k(0, I_{\theta_0}^{-1})$.

Proof. We will give a complete proof of this after we study the theory of M-estimation. □

2.3.1 Wald test

Given data $X_1, \ldots, X_n$ are i.i.d. P_θ suppose that we wish to test the hypothesis

$$H_0 : g(\theta) = 0 \qquad \text{versus} \qquad H_1 : g(\theta) > 0,$$

where $g : \Theta \to \mathbb{R}$ is assumed differentiable with gradient vector $\dot{g}(\cdot)$ (of dimension $k \times 1$). It is natural to test the above hypothesis with the test that rejects H_0 when

$$\sqrt{n} g(\hat{\theta}_n) > c_{n,\alpha}, \tag{56}$$

where $c_{n,\alpha}$ is the critical value such that the test has (approximate) level $\alpha \in (0, 1)$.

Question: How do we find $c_{n,\alpha}$? The following result helps us in that direction.

Theorem 2.23. Assume the setting of Theorem 2.22 with conditions (i)-(ii) holding. Suppose that $\hat{\theta}_n$ is an estimator of θ for which the expansion (55) holds when $\theta = \theta_0$. Let $\theta_n := \theta_0 + hn^{-1/2}$. Then (under $P^n_{\theta_n}$),

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \overset{P^n_{\theta_n}}{\rightsquigarrow} N_k(0, I^{-1}_{\theta_0}) \qquad \Leftrightarrow \qquad \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{P^n_{\theta_n}}{\rightsquigarrow} N_k(h, I^{-1}_{\theta_0}).$$

Furthermore, if $g : \Theta \to \mathbb{R}$ is a differentiable with nonzero gradient $\dot{g}(\theta_0)$ (of dimension $k \times 1$), then under $P^n_{\theta_n}$,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_n)) \overset{P^n_{\theta_n}}{\rightsquigarrow} N(0, \sigma^2_{\theta_0}),$$

where

$$\sigma^2_{\theta_0} = \dot{g}(\theta_0)^\top I^{-1}_{\theta_0} \dot{g}(\theta_0).$$

Proof. A complete proof was given in class; see Lehmann and Romano [8, Theorem 12.4.1] and its proof. □

Now, coming back to the test in (56), and assuming that $\dot{g}(\theta)$ and I_θ are continuous around $\theta = \theta_0$, the asymptotic variance of $g(\hat{\theta}_n)$ can be consistently estimated by

$$\hat{\sigma}^2_n := \dot{g}(\hat{\theta}_n)^\top I^{-1}_{\hat{\theta}_n} \dot{g}(\hat{\theta}_n).$$

Hence, we can take $c_{n,\alpha}$ in (56) to be $\hat{\sigma}_n z_\alpha$, where z_α is the upper α quantile of the standard normal distribution.

Exercise 9 (HW2): Find the power of the test in under the alternative $\theta_n = \theta_0 + hn^{-1/2}$. Also, how do we use the above strategy to construct a $1 - \alpha$ confidence set for $g(\theta)$.

Consider now the general problem of testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0, \tag{57}$$

where $\theta \in \Theta \subset \mathbb{R}^k$, under the assumptions of Theorem 2.23. We can show that, under H_0,

$$n(\hat{\theta}_n - \theta_0)^\top I_{\theta_0}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_k, \tag{58}$$

the chi-squared distribution with k degrees of freedom. Thus, for testing (139) the *Wald's test* rejects H_0 if

$$n(\hat{\theta}_n - \theta_0)^\top I_{\theta_0}(\hat{\theta}_n - \theta_0) > c_{k,1-\alpha}, \tag{59}$$

where $c_{k,1-\alpha}$ is the $1-\alpha$ quantile of χ^2_k. In the above, I_{θ_0} is often replaced by a consistent estimator, such as $I_{\hat{\theta}_n}$ (assuming I_θ is continuous). Under $\theta_n = \theta_0 + hn^{-1/2}$, the limiting distribution of the Wald statistic is $\chi^2_k(|I^{1/2}(\theta_0)h|^2)$[41], the noncentral chi squared distribution with k degrees of freedom and noncentrality parameter $|I^{1/2}(\theta_0)h|^2$.

A above ideas leads to an asymptotic level $1-\alpha$ confidence region for θ as

$$\{\theta \in \Theta : n(\hat{\theta}_n - \theta)^\top I_\theta(\hat{\theta}_n - \theta) \le c_{k,1-\alpha}\}.$$

When I_θ is replaced by $I_{\hat{\theta}_n}$ (assuming I_θ is continuous) the resulting confidence region is known as *Wald's confidence ellipsoid*. More generally, we can consider inference for $g(\theta)$, where $g = (g_1, \ldots, g_q) : \Theta \to \mathbb{R}^q$ for $q \ge 1$; see Lehmann and Romano [8, Equation (12.72)] and the related discussion.

2.3.2 Rao's score test

Instead of the Wald test, it is possible to construct tests based directly on the score function that have the advantage of not requiring computation of the MLE. Suppose that we are interested in testing (139). Letting

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i),$$

Rao's score test rejects when

$$Z_n^\top I_{\theta_0}^{-1} Z_n > c_{k,1-\alpha}$$

and is asymptotically level α. In this case, the Wald test (59) and the score test are asymptotically equivalent, in the sense that the probability that the two tests yield the same decision tends to one, both under the null hypothesis $\theta = \theta_0$ and under a sequence of alternatives $\theta_0 + hn^{-1/2}$. The equivalence follows from contiguity, the expansion (55), and the fact that $I_{\hat{\theta}_n} \to I_{\theta_0}$ in probability under θ_0 and under $\theta_0 + hn^{-1/2}$. Note that the two tests may differ greatly for alternatives far from θ_0.

[41] Exercise 10 (HW2). Show this.

2.4 Likelihood ratio test

Let $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ be a parametric family of distributions where P_θ has density p_θ w.r.t. a dominating measure μ. Suppose we observe $X_1, \ldots, X_n$ i.i.d. P_θ, and wish to test the hypothesis

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1. \tag{60}$$

If both the null and the alternative hypotheses consist of single points, then a MP test can be based on the loglikelihood ratio, by the Neyman-Pearson theory. Thus, if the two points are θ_0 and θ_1, respectively, then the optimal test statistic is given by

$$\log \frac{\Pi_{i=1}^n p_{\theta_1}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)}.$$

When Θ_0 and Θ_1 are composite, a sensible extension of the idea behind the Neyman-Pearson theory is to base a test on the loglikelihood ratio

$$\tilde{\Lambda}_n := \log \frac{\sup_{\theta \in \Theta_1} \Pi_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \Pi_{i=1}^n p_\theta(X_i)}.$$

Here the single points are replaced by maxima over the hypotheses. As before, the null hypothesis is rejected for large values of the statistic $\tilde{\Lambda}_n$. As the distributional properties of $\tilde{\Lambda}_n$ can be somewhat complicated, one usually replaces the supremum in the numerator by a supremum over the whole parameter set $\Theta := \Theta_0 \cup \Theta_1$[42]. This leads to the following statistic which is known as the (log) *likelihood ratio statistic* (LRS) for testing (60):

$$\Lambda_n := 2 \log \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} \Pi_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \Pi_{i=1}^n p_\theta(X_i)}. \tag{61}$$

The corresponding test for (60), which rejects for large values of Λ_n, is called the *likelihood ratio test* (LRT). Of course, the main question then becomes: "How do we find the critical value of the LRT?". This naturally leads to the study of the distribution of the LRS.

In this section we study the asymptotic properties of the LRS. The most important conclusion of this section is that, under H_0, the sequence Λ_n is asymptotically chi-squared. The main conditions needed for this conclusion are that the model is QMD at $\theta_0 \in \Theta_0$, and that Θ_0 and Θ are locally[43] equal to linear spaces. Then the test that rejects the null hypothesis if Λ_n exceeds the upper α-quantile of the chi-square distribution is asymptotically level α.

[42] This changes the test statistic only if $\tilde{\Lambda}_n \leq 0$, which is inessential, because in most cases the critical value will be positive (as Θ_1 is usually are "larger" set compared to Θ_0).

[43] The "local linearity" of the hypotheses is essential for the chi-square approximation (in which case the limiting distribution is χ^2_m where $m = \dim(\Theta) - \dim(\Theta_0)$), which fails already in a number of simple examples. An open set is certainly locally linear at each of its points, and so is a relatively open subset of an affine subspace. On the other hand, a half line or space, which arises, for instance, if testing a one-sided hypothesis $H_0 : \theta \leq 0$, is not locally linear at the boundary point $\theta = 0$. In that case the asymptotic null distribution of the LRS is not chi-square, but the distribution of a certain functional of a Gaussian vector.

Besides for testing, the LRS is often used for constructing confidence regions for a parameter $\psi(\theta)$. These can be constructed, as usual, as the values τ for which a null hypothesis $H_0 : \psi(\theta) = \tau$ is not rejected. Asymptotic confidence sets obtained by using the chi-square approximation are thought to have better coverage accuracy than those obtained by other asymptotic methods.

2.4.1 Deriving the asymptotic distribution of the LRS using LAN

An insightful derivation of the asymptotic distribution of the LRS is based on LAN. The approach applies also in the case that the (local) parameter spaces are not linear. Let us give a sketch of the argument below. Let $X_1, \ldots, X_n$ be i.i.d. P_{θ_0}, where $\theta_0 \in \Theta_0$. We can write Λ_n, as defined in (61), in terms of local likelihood ratios as:

$$\Lambda_n = 2 \sup_{h \in \mathcal{H}_n} \log \frac{\Pi_{i=1}^n p_{\theta_0 + hn^{-1/2}}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)} - 2 \sup_{h \in \mathcal{H}_{n,0}} \log \frac{\Pi_{i=1}^n p_{\theta_0 + hn^{-1/2}}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)} \tag{62}$$

where

$$\mathcal{H}_n := \sqrt{n}(\Theta - \theta_0) \qquad \text{and} \qquad \mathcal{H}_{n,0} := \sqrt{n}(\Theta_0 - \theta_0)$$

are the *local parameter spaces*. By LAN of $\{P_\theta\}$ at $\theta = \theta_0$, we have

$$\begin{aligned} 2 \sup_{h \in \mathcal{H}_n} \log \frac{\Pi_{i=1}^n p_{\theta_0 + hn^{-1/2}}(X_i)}{\Pi_{i=1}^n p_{\theta_0}(X_i)} &= 2 \sup_{h \in \mathcal{H}_n} \left[\frac{h^\top}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^\top I_{\theta_0} h + o_{P_{\theta_0}^n}(1) \right] \\ &= \sup_{h \in \mathcal{H}_n} \left[2h^\top I_{\theta_0} W_n - h^\top I_{\theta_0} h + o_{P_{\theta_0}^n}(1) \right] \end{aligned}$$

where

$$W_n := I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) \overset{P_{\theta_0}^n}{\rightsquigarrow} W \sim N_k(0, I_{\theta_0}^{-1}). \tag{63}$$

Using a similar expansion for the second term in (62), we get,

$$\begin{aligned} \Lambda_n &= \sup_{h \in \mathcal{H}_n} \left[2h^\top I_{\theta_0} W_n - h^\top I_{\theta_0} h + o_{P_{\theta_0}^n}(1) \right] - \sup_{h \in \mathcal{H}_{n,0}} \left[2h^\top I_{\theta_0} W_n - h^\top I_{\theta_0} h + o_{P_{\theta_0}^n}(1) \right] \\ &\approx -\inf_{h \in \mathcal{H}_n} \left[W_n^\top I_{\theta_0} W_n - 2h^\top I_{\theta_0} W_n + h^\top I_{\theta_0} h \right] + \inf_{h \in \mathcal{H}_{n,0}} \left[W_n^\top I_{\theta_0} W_n - 2h^\top I_{\theta_0} W_n + h^\top I_{\theta_0} h \right]. \end{aligned}$$

This suggests that, if the sets $\mathcal{H}_n$ and $\mathcal{H}_{n,0}$ converge in a suitable sense[44] to sets $\mathcal{H}$ and $\mathcal{H}_0$ respectively, the sequence Λ_n converges in distribution to the random variable Λ defined by

$$\begin{aligned} \Lambda &:= \inf_{h \in \mathcal{H}_0} (W - h)^\top I_{\theta_0} (W - h) - \inf_{h \in \mathcal{H}} (W - h)^\top I_{\theta_0} (W - h) \\ &= \inf_{h \in \mathcal{H}_0} |I_{\theta_0}^{1/2} W - I_{\theta_0}^{1/2} h|^2 - \inf_{h \in \mathcal{H}} |I_{\theta_0}^{1/2} W - I_{\theta_0}^{1/2} h|^2 \end{aligned} \tag{64}$$

[44] We use the following notion of convergence of sets. Write $\mathcal{H}_n \to \mathcal{H}$ where $\mathcal{H} := \{h \in \mathbb{R}^k : h = \lim_{j \to \infty} h_{n_j}$ for converging sequences $\{h_{n_j}\}_{j \geq 1}$ with $h_{n_j} \in \mathcal{H}_{n_j}$ for every $n_j\}$.

where W is defined in (63).

It can also be seen that the above argument also generalizes under contiguous alternatives, i.e., when the data is generated from $\theta_n := \theta_0 + gn^{-1/2} \in \Theta$, for $g \in \mathbb{R}^k$. Note that then $W_n \overset{P^n_{\theta_n}}{\rightsquigarrow} N_k(g, I^{-1}_{\theta_0})$, but the limiting distribution of the LRS is still Λ, as defined in (64).

2.4.2 Asymptotic distribution of the LRS

The following result, known as Wilk's theorem, gives the asymptotic distribution of the LRS (both under fixed and continuous alternatives).

Theorem 2.24 (Wilks' theorem)**.** Suppose that conditions (i)-(iv) hold in Theorem 2.22. Let $\hat{\theta}_{n,0} := \text{argmax}_{\theta \in \Theta_0} \Pi_{i=1}^n p_\theta(X_i)$ be the constrained MLE. Assume that the sets $\mathcal{H}_{n,0}$ and $\mathcal{H}_n$ converge to the sets $\mathcal{H}_0$ and $\mathcal{H}$ respectively. Let $\theta_n := \theta_0 + gn^{-1/2}$ for $g \in \mathbb{R}^k$. Then, the asymptotic distribution of the LRS Λ_n (defined in (61)) is given by

$$\Lambda_n \overset{P^n_{\theta_n}}{\rightsquigarrow} \Lambda$$

where the distribution of Λ is given in (64) with $W \sim N_k(g, I^{-1}_{\theta_0})$.

Proof. We will give a complete proof of this after we study the theory of M-estimation. □

Although (64) gives an explicit characterization of the limiting distribution of the LRS in a very general setting, it is not immediately clear if it is very useful. Observe that if θ_0 is an interior point of $\Theta \subset \mathbb{R}^k$, then $\mathcal{H}$ (the limit of the sequence of sets $\mathcal{H}_n$) is the whole of $\mathbb{R}^k$, and thus, Λ reduces to $\inf_{h \in \mathcal{H}_0} |I^{1/2}_{\theta_0} W - I^{1/2}_{\theta_0} h|^2$.

The following lemma shows that the distribution of the first term of Λ (in (61)), i.e., $\inf_{h \in \mathcal{H}_0} |I^{1/2}_{\theta_0} W - I^{1/2}_{\theta_0} h|^2$, indeed simplifies when $\mathcal{H}_0$ is an l-dimensional linear subspace of $\mathbb{R}^k$. First observe that, under H_0, $I^{1/2}_{\theta_0} W \sim N_k(0, I_{\theta_0})$.

Lemma 2.25. Let $Z \sim N_k(0, I_k)$ and let $\mathcal{S}$ be an l-dimensional linear subspace of $\mathbb{R}^k$. Then $\inf_{h \in \mathcal{S}} \|Z - h\|^2$ has a chi-square distribution with $k - l$ degrees of freedom.

Proof. For $z \in \mathbb{R}^k$, let $\text{Proj}_{\mathcal{S}}(z)$ denote the orthogonal projection of z onto the linear subspace $\mathcal{S}$. Note that $\text{Proj}_{\mathcal{S}}(z) = P_{\mathcal{S}} z$ for a $k \times k$ matrix $P_{\mathcal{S}}$ with rank l such that $P^2_{\mathcal{S}} = P_{\mathcal{S}} = P^\top_{\mathcal{S}}$. By Pythagoras' theorem, $z = \text{Proj}_{\mathcal{S}}(z) + \text{Proj}_{\mathcal{S}^\perp}(z)$ where $\langle \text{Proj}_{\mathcal{S}}(z), \text{Proj}_{\mathcal{S}^\perp}(z) \rangle = 0$. In fact, $Q := I_k - P_{\mathcal{S}}$ is the orthogonal projection matrix onto the linear space $\mathcal{S}^\perp$, the orthogonal complement of $\mathcal{S}$. Then,

$$\inf_{h \in \mathcal{S}} \|Z - h\|^2 \;=\; \inf_{h \in \mathcal{S}} \|P_{\mathcal{S}} Z - h\|^2 + \|QZ\|^2 \;=\; Z^\top Q^\top Q Z \;=\; Z^\top Q Z$$

where we have used the facts: (i) $\inf_{h \in \mathcal{S}} \|P_{\mathcal{S}} Z - h\|^2 = 0$, and (ii) $Q^2 = Q = Q^\top$. By eigendecomposition of the matrix Q, we have $Q = V^\top D V$ where V is a $k \times k$ orthogonal

matrix and D is a diagonal matrix with diagonal entries 0 or 1 (as the eigenvalues of a projection matrix are either 0 or 1). As Q has rank $k-l$, rank(D) is also $k-l$. Thus, without loss of generality, we can assume that $D_{1,1} = \cdots = D_{k-l,k-l} = 1$ and $D_{k-l+1,k-l+1} = \cdots = D_{k,k} = 0$, where $D := (D_{i,j})_{k\times k}$. As the standard normal distribution is invariant under orthogonal transformations, $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_k) := VZ \sim N_k(0, I_k)$. Thus,

$$Z^\top QZ = ZV^\top DVZ = \tilde{Z}D\tilde{Z} = \sum_{i>l} \tilde{Z}_i^2 \sim \chi^2_{k-l}.$$

Thus $\inf_{h\in\mathcal{S}} \|Z-h\|^2 = \sum_{i>l} \tilde{Z}_i^2$ has a chi-square distribution with $k-l$ degrees of freedom. □

Corollary 2.26. Consider the setting of Theorem 2.24. Let $\theta_n := \theta_0 + gn^{-1/2} \in \Theta$ for $g \in \mathbb{R}^k$. When $\theta_0 \in \Theta_0$ is an interior point of Θ, then

$$\Lambda_n \overset{P^n_{\theta_n}}{\rightsquigarrow} \inf_{h\in\mathcal{H}_0} \|I^{1/2}_{\theta_0} W - I^{1/2}_{\theta_0} h\|^2$$

where $W \sim N_k(g, I_\theta^{-1})$. Further, if $\mathcal{H}_0$ is a linear subspace of dimension l, then under the null (i.e., $g = 0$), $\Lambda_n \xrightarrow{d} \chi^2_{k-l}$.

Example 2.27 (Location-scale family)**.** Suppose we observe a random sample from the density $f((\cdot-\mu)/\sigma)/\sigma$ for a given probability density f, where the location-scale parameter $\theta = (\mu, \sigma)$ ranges over the set $\Theta = \mathbb{R} \times \mathbb{R}^+$. We consider two testing problems.

(i) Consider testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ which corresponds to the setting $\Theta_0 = \{0\}\times\mathbb{R}^+$. For a given point $\theta_0 = (0, \sigma) \in \Theta_0$, $\mathcal{H}_{n,0} := \sqrt{n}(\Theta_0 - \theta_0) = \{0\}\times(-\sqrt{n}\sigma, \infty)$ and converges to the linear space $\mathcal{H}_0 := \{0\} \times \mathbb{R}$. Also, note that here $\mathcal{H} = \mathbb{R}^2$. Under regularity conditions on f (e.g., QMD at 0), the sequence of LRSs is asymptotically chi-squared with 1 degree of freedom as, under H_0, Λ (defined in (64)) reduces to

$$\inf_{h\in\mathcal{H}_0'} \|Z-h\|^2 - \inf_{h\in\mathcal{H}} \|Z - I^{1/2}_{\theta_0} h\|^2 = \inf_{h\in\mathcal{H}_0'} \|Z-h\|^2 \sim \chi^2_1$$

where $Z \sim N_2(0, I_2)$ and $\mathcal{H}_0' := \{I^{1/2}_{\theta_0} h : h \in \mathcal{H}_0\}$ is a linear subspace of dimension 1.

(ii) Consider testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ which corresponds to the setting $\Theta_0 = (-\infty, 0] \times \mathbb{R}^+$. For a given point $\theta_0 = (0, \sigma)$ on the boundary of the null hypothesis, $\mathcal{H}_{n,0} := \sqrt{n}(\Theta_0 - \theta_0)$ converges to $\mathcal{H}_0 = (-\infty, 0] \times \mathbb{R}$. In this case, the limit distribution of the LRS is not chi-square but equals the distribution of the squared distance of a standard normal vector to the set $I^{1/2}_{\theta_0}\mathcal{H}_0$. The latter is a half-space with boundary line through the origin. Because a standard normal vector is rotationally symmetric, the distribution of its distance to a half-space of this type does not depend on the orientation of the half-space. Thus the limit distribution is equal to the distribution of the squared distance of a standard normal vector to the half-space $\{h = (h_1, h_2) \in \mathbb{R}^2 : h_2 \leq 0\}$, i.e., the distribution of $(Z\vee 0)^2$ for a standard normal variable Z. As $\mathbb{P}((Z \vee 0)^2 > c) = \frac{1}{2}\mathbb{P}(Z^2 > c)$ for every $c > 0$, we

must choose the critical value of the test equal to the upper 2α-quantile of the chi-square distribution with 1 degree of freedom. Then the asymptotic level of the test is α for every θ_0 on the boundary of the null hypothesis (provided $\alpha < 1/2$).

For a point θ_0 in the interior of the null hypothesis $H_0 : \mu \leq 0$ the sets $\sqrt{n}(\Theta_0 - \theta_0)$ converge to $\mathbb{R} \times \mathbb{R}$ and the LRS converges in distribution to the squared distance to the whole space, which is zero.

Example 2.28 (Power of the LRT)**.** We assume that the parameter θ_0 is an interior point of Θ and denote the true parameter by $\theta_n := \theta_0 + g/\sqrt{n}$. Under the conditions of Theorem 2.24, the LRS is asymptotically distributed as

$$\Lambda = \inf_{h \in \mathcal{H}_0} |Z + I_{\theta_0}^{1/2} g - I_{\theta_0}^{1/2} h|$$

for $Z \sim N_k(0, I_k)$. Suppose that the limiting local parameter set $\mathcal{H}_0$ is a linear subspace of dimension l, and that the null hypothesis is rejected for values of Λ_n exceeding the critical value $\chi^2_{k-l,\alpha}$. hen the local power functions of the resulting tests satisfy

$$\pi_n(\theta_n) = \mathbb{P}_{\theta_n}(\Lambda_n > \chi^2_{k-l,\alpha}) =: \pi(g).$$

The variable Λ is the squared distance of the vector Z to the affine subspace $-I_{\theta_0}^{1/2} g + I_{\theta_0}^{1/2} \mathcal{H}_0$. By the rotational invariance of the normal distribution, the distribution of Λ does not depend on the orientation of the affine subspace, but only on its codimension and its distance $\delta^2 := \inf_{h \in \mathcal{H}_0} |I_{\theta_0}^{1/2} g - I_{\theta_0}^{1/2} h|^2$ to the origin. This distribution is known as the *noncentral chi-square distribution*[45] with noncentrality parameter δ. Thus

$$\pi(g) = \mathbb{P}(\chi^2_{k-l}(\delta^2) > \chi^2_{k-l,\alpha}).$$

The noncentral chi-square distributions are stochastically increasing in the noncentrality parameter. It follows that the likelihood ratio test has good (local) power at g that yield a large value of the noncentrality parameter. Then $\mathcal{H}_0 = \{0\}$, and the noncentrality parameter reduces to the square root of $g^\top I_{\theta_0} g$.

2.5 Comparison of test functions

The aim of this section is to compare tests based on their power functions. Suppose that ϕ_n and $\tilde{\phi}_n$ are two *asymptotically level α tests*[46] with power functions π_n and $\tilde{\pi}_n$ respectively, for the testing problem:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

[45] Let $Y_1, \ldots, Y_r$ be independently normally distributed random variables with unit variance and means $\eta_1, \ldots, \eta_r$. Then $U = \sum_{i=1}^r Y_i^2$ is distributed according to the noncentral χ^2-distribution with r degrees of freedom and noncentrality parameter $\delta^2 = \sum_{i=1}^r \eta_i^2$.

[46] We say that a sequence of tests ϕ_n is asymptotically level α if $\limsup_{n\to\infty} \pi_n(\theta) \leq \alpha$ for every $\theta \in \Theta$.

when we have i.i.d. data $X_1, \ldots, X_n$ from a model $\{P_\theta : \theta \in \Theta\}$. We would say that ϕ_n is better than $\tilde{\phi}_n$ if

$$\pi_n(\theta) \geq \tilde{\pi}_n(\theta), \qquad \text{for all } \ \theta \in \Theta_1.$$

However, quite often it is not possible to find out, for every fixed n, which test is better. Hence, we may want to compare their asymptotic performance. A first idea is to consider their limiting power functions of the form

$$\pi(\theta) := \lim_{n\to\infty} \pi_n(\theta) \qquad \text{and} \qquad \tilde{\pi}(\theta) = \lim_{n\to\infty} \tilde{\pi}_n(\theta),$$

for $\theta \in \Theta_1$. If these limits exist for all θ, then the sequence ϕ_n[47] is better than the sequence $\tilde{\phi}_n$ if $\pi(\theta) \geq \tilde{\pi}(\theta)$, for all $\theta \in \Theta_1$. However, it turns out that this approach is too naive. The limiting power functions typically exist, but they are trivial and identical for all reasonable sequences of tests. Let us consider a simple example to demonstrate this.

Example 2.29 (Sign test)**.** Suppose that we have $X_1, \ldots, X_n$ i.i.d. from a density $f(\cdot - \theta)$, where $f(\cdot)$ is assumed symmetric about 0 (and unknown). Here $\theta \in \mathbb{R}$ is the unknown location parameter of interest and we want to test

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta > 0. \tag{65}$$

We assume further that f has a unique median (at 0). We can use the *sign statistic*

$$S_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i > 0\}$$

for the above testing problem, rejecting H_0 when S_n is 'large'. This is the *sign test*. Letting $F(\cdot - \theta)$ denote the c.d.f. of the underlying distribution, the mean and variance of S_n equals

$$\mu(\theta) := \mathbb{P}_\theta(X_1 > 0) = 1 - F(-\theta) \qquad \text{and} \qquad \frac{\sigma^2(\theta)}{n} := \frac{1}{n}[1 - F(-\theta)]F(-\theta).$$

In fact, $nS_n \sim \text{Bin}(n, 1-F(-\theta))$, and under $H_0 : \theta = 0$, $nS_n \sim \text{Bin}(n, 1/2)$ for any such F, as $F(0) = 1/2$ (as F has median 0). Thus, S_n is *distribution-free* under H_0 (cf. Example 2.11). Although the binomial critical values can be used to calibrate the test based on S_n, it is more convenient to use the normal approximation, as described below.

By the normal approximation to the binomial distribution, the sequence $\sqrt{n}(S_n - \mu(\theta))$ is asymptotically normal with mean 0 and variance $\sigma^2(\theta)$. Under H_0, the mean and variance are equal to $\mu(0) = 1/2$ and $\sigma^2(0) = 1/4$ respectively, so that $\sqrt{n}(S_n - \mu(0)) \xrightarrow{d} N(0, 1/4)$. The test that rejects H_0 if $\sqrt{n}(S_n - \mu(0))$ exceeds the critical value $z_\alpha/2$ has power function

$$\begin{aligned} \pi_n(\theta) &= \mathbb{P}_\theta\big(\sqrt{n}(S_n - \mu(\theta)) > \frac{1}{2}z_\alpha - \sqrt{n}(\mu(\theta) - \mu(0))\big) \\ &= 1 - \Phi\left(\frac{\frac{1}{2}z_\alpha - \sqrt{n}(F(0) - F(-\theta))}{\sigma(\theta)}\right) + o(1). \end{aligned}$$

[47] Typically, the tests corresponding to a sequence $\phi_1, \phi_2, \ldots$ are of the same type. For instance, they are all based on (i) sample averages or (ii) rank statistics, and only the number of observations changes with n.

Because $F(0) - F(-\theta) > 0$ for every $\theta > 0$ (as F has unique median 0), it follows that $\pi_n(\theta) \to 1$ if $\theta > 0$. Thus, the power at every fixed alternative converges to 1.

Definition 2.30 (Consistent test)**.** A sequence of tests with power functions $\theta \mapsto \pi_n(\theta)$ is asymptotically *consistent* at level α at (or against) the alternative θ if it is asymptotically level α and $\pi_n(\theta) \to 1$ as $n \to \infty$. If a family of sequences of tests contains for every level $\alpha \in (0, 1)$ a sequence that is consistent against every alternative, then the corresponding tests are simply called *consistent*.

Consistency is an optimality criterion for tests, but because most sequences of tests are consistent, it is too weak to be really useful. To make an informative comparison between sequences of (consistent) tests, we shall study the performance of the tests in problems that become harder as more observations become available. One way of making a testing problem harder is to choose null and alternative hypotheses closer to each other. In this section we fix the null hypothesis and consider the power at sequences of alternatives that converge to the null hypothesis.

Example 2.31 (Sign test, continued)**.** Consider the power of the sign test at sequences of alternatives $\theta_n \downarrow 0$. Extension of the argument of the preceding example yields[48]

$$\pi_n(\theta_n) = 1 - \Phi\left(\frac{\frac{1}{2}z_\alpha - \sqrt{n}(F(0) - F(-\theta_n))}{\sigma(\theta_n)}\right) + o(1).$$

The asymptotic power at θ_n depends on the rate at which $\theta_n \to 0$. If θ_n converges to zero fast enough to ensure that $\sqrt{n}(F(0) - F(-\theta_n)) \to 0$, then the power $\pi_n(\theta_n)$ converges to α: The sign test is not able to discriminate these alternatives from the null hypothesis. If θ_n converges to zero at a slow rate, then $\sqrt{n}(F(0) - (-\theta_n)) \to \infty$, and the asymptotic power is equal to 1: These alternatives are too easy. The intermediate rates, which yield a nontrivial asymptotic power, appear to be of most interest. Suppose that $f(0) > 0$. Then,

$$\sqrt{n}(F(0) - F(-\theta_n)) \to \sqrt{n}\theta_n f(0) + \sqrt{n}\, o(\theta_n).$$

This is bounded away from zero and infinity if θ_n converges to zero at rate $\theta_n = O(n^{-1/2})$. For such rates the power $\pi_n(\theta_n)$ is asymptotically strictly between α and 1. In particular, for every h,

$$\pi_n(hn^{-1/2}) \to 1 - \Phi(z_\alpha - 2hf(0)). \tag{66}$$

In the preceding example only alternatives θ_n that converge to the null hypothesis at rate $O(1/\sqrt{n})$ lead to a nontrivial asymptotic power. This is typical for parameters that depend "smoothly" on the underlying distribution. In this situation a reasonable method

[48] We can make this rigorous by using a Lindeberg-Feller CLT which allows the underlying data distribution to change with n.

for asymptotic comparison of two (or more) sequences of tests for (65) is to consider *local limiting power functions*, defined as

$$\pi(h) := \lim_{n\to\infty} \pi_n(hn^{-1/2}), \qquad h \geq 0.$$

These limits typically exist and can be derived by the same method as in the preceding example. A general scheme is as follows.

Let $\theta \in \Theta$ be a real parameter and let the test reject the null hypothesis $H_0 : \theta = \theta_0$ for large values of a test statistic T_n. Assume that T_n is asymptotically normal in the sense that, for all sequences of the form $\theta_n = \theta_0 + hn^{-1/2}$,[49]

$$\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \overset{P_{\theta_n}^n}{\rightsquigarrow} N(0,1). \tag{68}$$

Often $\mu(\theta)$ and $\sigma^2(\theta)/n$ can be taken to be the mean and the variance of T_n, but this is not necessary. The convergence (68) is sometimes referred to as "locally uniform" asymptotic normality. "Contiguity arguments" can reduce the derivation of asymptotic normality under $\theta_n = \theta_0 + hn^{-1/2}$ to derivation under $\theta = \theta_0$.

Assumption (68) includes that the sequence $\sqrt{n}(T_n - \mu(\theta_0))$ converges in distribution to $N(0, \sigma^2(\theta_0))$ under $\theta = \theta_0$. Thus, the tests that reject the null hypothesis $H_0 : \theta = \theta_0$ if $\sqrt{n}(T_n - \mu(\theta_0))$ exceeds $\sigma(\theta_0)z_\alpha$ are asymptotically of level α. The power functions of these tests can be written

$$\pi_n(\theta_n) = \mathbb{P}_{\theta_n}\Big(\sqrt{n}(T_n - \mu(\theta_n)) > \sigma(\theta_0)z_\alpha - \sqrt{n}(\mu(\theta_n) - \mu(\theta_0))\Big).$$

For $\theta_n = \theta_0 + hn^{-1/2}$, the sequence $\sqrt{n}(\mu(\theta_n) - \mu(\theta_0))$ converges to $h\mu'(\theta_0)$ if μ is differentiable at θ_0. If $\sigma(\theta_n) \to \sigma(\theta_0)$, then under (68),

$$\pi_n(\theta_0 + hn^{-1/2}) \to 1 - \Phi\left(z_\alpha - \frac{h\mu'(\theta_0)}{\sigma(\theta_0)}\right). \tag{69}$$

Thus, we have essentially proved the following result.

Theorem 2.32. Let μ and σ be functions of θ such that (68) holds for every sequence $\theta_n = \theta_0 + hn^{-1/2}$. Suppose that μ is differentiable and that σ is continuous at $\theta = \theta_0$. Then the power functions π_n of the tests that reject $H_0 : \theta = \theta_0$ for large values of T_n and are asymptotically of level α satisfy (69) for every h.

The limiting power function depends on the test statistics only through the quantity $\mu'(\theta_0)/\sigma(\theta_0)$. This is called the *slope* of the sequence of tests. Two sequences of tests

[49] As the convergence (68) is under a law indexed by θ_n that changes with n, the convergence is not implied by

$$\frac{\sqrt{n}(T_n - \mu(\theta))}{\sigma(\theta)} \overset{P_\theta^n}{\rightsquigarrow} N(0,1), \qquad \text{for every } \theta. \tag{67}$$

can be asymptotically compared by just comparing the sizes of their slopes[50]. The bigger the slope, the better the test. The size of the slope depends on the rate $\mu'(\theta_0)$ of change of the asymptotic mean of the test statistics relative to their asymptotic dispersion $\sigma(\theta_0)$.

Let us now compare the sign test with the t-test.

Example 2.33 (One-sample t-test)**.** The t-test for (65) rejects for large values of $\sqrt{n}\bar{X}_n/s_n$ where $\bar{X}_n$ and s_n^2 are the sample mean and (unbiased) sample variance. The sample variance s_n^2 converges in probability to the variance σ^2 of a single observation. By CLT and Slutsky's lemma we get, for $\theta_n = hn^{-1/2}$,

$$\sqrt{n}\left(\frac{\bar{X}_n}{s_n} - \frac{hn^{-1/2}}{\sigma}\right) = \frac{\bar{X}_n - hn^{-1/2}}{s_n} + h\left(\frac{1}{s_n} - \frac{1}{\sigma}\right) \overset{P_{\theta_n}^n}{\rightsquigarrow} N(0,1).$$

Thus Theorem 2.32 applies with $\mu(\theta) = \theta/\sigma$ and $\sigma(\theta) = 1$. Thus the slope of the t-test equals $1/\sigma$.

Example 2.34 (Sign test versus t-test)**.** We can now compare the sign test with the t-test. It suffices to compare the slopes of the two tests. By the preceding examples these are $2f(0)$ and $(\int x^2 f(x)dx)^{-1/2}$, respectively. Clearly the outcome of the comparison depends on the shape f. It is interesting that the two slopes depend on the underlying shape in an almost orthogonal manner. The slope of the sign test depends only on the height of f at zero; the slope of the t-test depends mainly on the tails of f. For the standard normal distribution the slopes are $\sqrt{2/\pi}$ and 1. The superiority of the t-test in this case is not surprising, because the t-test is uniformly most powerful for every n. For the Laplace distribution, the ordering is reversed: The slopes are 1 and $1/\sqrt{2}$. The superiority of the sign test has much to do with the "unsmooth" character of the Laplace density at its mode.

The simplicity of comparing slopes is attractive on the one hand, but indicates the potential weakness of asymptotics on the other. For instance, the slope of the sign test was seen to be $2f(0)$, but it is clear that this value alone cannot always give an accurate indication of the quality of the sign test. Consider a density that is basically a normal density, but a tiny proportion of $10^{-10}\%$ of its total mass is located under an extremely thin but enormously high peak at zero. The large value $f(0)$ would strongly favor the sign test. However, at moderate sample sizes the observations would not differ significantly from a sample from a normal distribution, so that the t-test is preferable. In this situation the asymptotics are only valid for unrealistically large sample sizes.

Even though asymptotic approximations should always be interpreted with care, in the

[50] If θ is the only unknown parameter in the problem, then the available tests can be ranked in asymptotic quality simply by the value of their slopes. In many problems there are also nuisance parameters (for instance the shape of a density), and the slope is a function of the nuisance parameter rather than a number. This complicates the comparison considerably. For every value of the nuisance parameter a different test may be best, and additional criteria are needed to choose a particular test.

present situation there is actually little to worry about. Even for $n = 20$, the comparison of slopes of the sign test and the t-test gives the right message for the standard distributions; see van der Vaart [15, Table 14.1].

2.5.1 Asymptotic relative efficiency

In the previous examples, the slopes of the sequence of tests helped us compare the sequences. A more general measure of comparing two tests is given below. This quantity is called the *asymptotic relative efficiency* (ARE). In the above problems this measure reduces to comparing the square of the quotient of two slopes.

Definition 2.35 (Asymptotic relative efficiency)**.** Consider a sequence of testing problems consisting of testing a null hypothesis

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_\nu.$$

We use the parameter ν to describe the asymptotics; thus $\nu \to \infty$. We require a priori that our tests attain asymptotically level α and power $\gamma \in (\alpha, 1)$. Usually we can meet this requirement by choosing an appropriate number of observations at "time" ν. A larger number of observations allows smaller level and higher power. If π_n is the power function of a test if n observations are available, then we define n_ν to be the *minimal number* of observations such that both

$$\pi_{n_\nu}(\theta_0) \leq \alpha, \qquad \text{and} \qquad \pi_{n_\nu}(\theta_\nu) \geq \gamma. \tag{70}$$

If two sequences of tests are available, then we prefer the sequence for which the numbers n_ν are smallest. Suppose that $n_{\nu,1}$ and $n_{\nu,2}$ observations are needed for two given sequences of tests. Then, if it exists, the limit

$$\lim_{\nu \to \infty} \frac{n_{\nu,2}}{n_{\nu,1}}$$

is called the *asymptotic relative efficiency* (ARE) or *Pitman efficiency* of the first with respect to the second sequence of tests.

A relative efficiency larger than 1 indicates that fewer observations are needed with the first sequence of tests, which may then be considered the better one.

In principle, the relative efficiency may depend on α, γ and the sequence of alternatives θ_ν. The concept is mostly of interest if the relative efficiency is the same for all possible choices of these parameters. This is often the case. In particular, in the situations considered previously, the relative efficiency turns out to be the square of the quotient of the slopes.

Suppose that in a given sequence of models $(\mathcal{X}_n, \mathcal{C}_n, P_{n,\theta} : \theta \in \Theta)$ it is desired to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternatives $H_1 : \theta = \theta_n$.

Theorem 2.36. Consider statistical models $(\mathcal{X}_n, \mathcal{C}_n, P_{n,\theta} : \theta \in \Theta)$ such that $\|P_{n,\theta} - P_{n,\theta_0}\| \to 0$[51] as $\theta \to \theta_0$, for every n. Let $T_{n,1}$ and $T_{n,2}$ be sequences of statistics that satisfy (68) for every sequence $\theta_n \downarrow \theta_0$ and functions μ_i and σ_i such that μ_i is differentiable at θ_0 and σ_i is continuous at θ_0, with $\mu_i'(\theta_0) > 0$ and $\sigma_i(\theta_0) > 0$. Then the ARE of the tests that reject the null hypothesis $H_0 : \theta = \theta_0$ for large values of $T_{n,i}$ is equal to

$$\left(\frac{\mu_1'(\theta_0)/\sigma_1(\theta_0)}{\mu_2'(\theta_0)/\sigma_2(\theta_0)}\right)^2$$

for every sequence of alternatives $\theta_n \downarrow 0$, irrespective of $\alpha > 0$ and $\gamma \in (\alpha, 1)$. If the power functions of the tests based on $T_{n,i}$ are nondecreasing for every n, then the assumption of asymptotic normality of $T_{n,i}$ can be relaxed to asymptotic normality under every sequence $\theta_n = O(n^{-1/2})$ only.

Proof. The condition that P_θ approaches P_{θ_0} in total variation distance as $\theta \to \theta_0$ implies that the minimal numbers $n_{\nu,i}$ must go to infinity as $\nu \to \infty$[52]. Further, the convergence to a continuous distribution implies that the asymptotic level and power attained for the minimal numbers of observations (minimal for obtaining at most level α and at least power γ) is exactly α and γ.

In order to obtain asymptotic level α the tests must reject H_0 if $\sqrt{n_{\nu,i}}(T_{n_{\nu,i},i} - \mu_i(\theta_0)) > \sigma_i(\theta_0) z_\alpha$. The powers of these tests are equal to

$$\begin{aligned}
\pi_{n_{\nu,i}}(\theta_\nu) &= \mathbb{P}_{\theta_\nu}\left(\sqrt{n_{\nu,i}}(T_{n_{\nu,i},i} - \mu_i(\theta_\nu)) > \sigma(\theta_0) z_\alpha - \sqrt{n_{\nu,i}}(\mu_i(\theta_\nu) - \mu_i(\theta_0))\right) \\
&= 1 - \Phi\left(z_\alpha + o(1) - \sqrt{n_{\nu,i}}\theta_\nu \frac{\mu_i'(\theta_0)}{\sigma_i(\theta_0)}(1 + o(1))\right) + o(1).
\end{aligned}$$

This sequence of powers tends to $\gamma < 1$ if and only if the argument of Φ tends to z_γ, i.e.,

$$z_\gamma \approx z_\alpha - \sqrt{n_{\nu,i}}\,\theta_\nu \frac{\mu_i'(\theta_0)}{\sigma_i(\theta_0)} \qquad \Leftrightarrow \qquad \sqrt{n_{\nu,i}}\,\theta_\nu \approx (z_\alpha - z_\gamma)\frac{\sigma_i(\theta_0)}{\mu_i'(\theta_0)}.$$

Thus the ARE of the two sequences of tests equals

$$\lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}} = \lim_{\nu\to\infty} \frac{n_{\nu,2}\theta_\nu^2}{n_{\nu,1}\theta_\nu^2} = \frac{\mu_1'(\theta_0)(z_\alpha - z_\gamma)^2\sigma_2(\theta_0)}{\mu_2'(\theta_0)(z_\alpha - z_\gamma)^2\sigma_1(\theta_0)} = \left(\frac{\mu_1'(\theta_0)/\sigma_1(\theta_0)}{\mu_2'(\theta_0)/\sigma_2(\theta_0)}\right)^2.$$

This proves the first assertion of the theorem.

The second assertion follows from the discussion in van der Vaart [15, Section 14.2]. □

[51] This refers to the L_1 distance between $P_{n,\theta}$ and P_{n,θ_0}; see Section 2.5.2.

[52] We show this using the concepts introduced in Section 2.5.2. The sum of the error probabilities of the first and second kind of any test, using Lemma 2.39, is

$$1 - \pi_n(\theta_\nu) + \pi_n(\theta_0) \geq 1 - \frac{1}{2}\|P_{n,\theta_\nu} - P_{n,\theta_0}\|.$$

By assumption, the right-hand side converges to 1 as $\nu \to \infty$ uniformly in every finite set of n. Thus, for every bounded sequence $n = n_\nu$ and any sequence of tests, the sum of the error probabilities is asymptotically bounded below by 1. But, from (70), we are given that the sum of the error probabilities is bounded above by $\alpha + 1 - \gamma < 1$, leading to a contradiction.

Example 2.37 (ARE of sign test versus t-test)**.** From the above result and Example 2.34 we can see that the ARE of the sign test versus the t-test is equal to

$$4f^2(0)\int x^2 f(x)dx.$$

For the uniform distribution, the relative efficiency of the sign test with respect to the t-test equals $1/3$[53]. It can be shown that this is the minimal possible value over all densities with mode zero. On the other hand, it is possible to construct distributions for which this relative efficiency is arbitrarily large, by shifting mass into the tails of the distribution. Thus, the sign test is "robust" against heavy tails, the t-test is not.

2.5.2 L_1-distance and power

Definition 2.38 (L_1-distance)**.** The L_1-distance between two distributions P and Q with densities $p = dP/d\mu$ and $q = dQ/d\mu$ is

$$\|P - Q\| = \int |p - q|\, d\mu. \tag{71}$$

Lemma 2.39. For a sequence of models P_n, with null hypothesis $H_0 : \theta = \theta_0$ and alternatives $H_1 : \theta = \theta_n$, the power function of any test satisfies

$$\pi_n(\theta_n) - \pi_n(\theta_0) \le \frac{1}{2}\|P_{n,\theta_n} - P_{n,\theta_0}\|. \tag{72}$$

Furthermore, there is a test for which equality holds.

Proof. If π_n is the power function of the test ϕ_n, then the difference on the left side in (72) can be written as $\int \phi_n(p_{n,\theta} - p_{n,\theta_0})d\mu_n$ (here we assume that $P_{n,\theta}$ is dominated by μ_n for all n). This expression is maximized for the test function $\phi_n = \mathbf{1}\{p_{n,\theta} > p_{n,\theta_0}\}$. Thus,

$$\pi_n(\theta_n) - \pi_n(\theta_0) = \int \phi_n(p_{n,\theta} - p_{n,\theta_0})d\mu_n \le \int_{p_{n,\theta} > p_{n,\theta_0}} (p_{n,\theta} - p_{n,\theta_0})d\mu_n = \frac{1}{2}\|P_{n,\theta_n} - P_{n,\theta_0}\|,$$

where in the last equality we have used the fact that: For any pair of probability densities p and q we have $\int_{q>p}(q-p)d\mu = \frac{1}{2}\int |p-q|d\mu$[54], since $\int (p-q)d\mu = 0$. □

2.6 Exercises

11. Lehmann and Romano [8, Problem 12.6].

[53] Exercise 19 (HW2): Show that $4f^2(0)\int x^2 f(x)dx \ge 1/3$ for every unimodal probability density f that has its mode at 0. (Hint: Use the invariance to scaling to reduce the problem to that of finding the minimum of $4\int y^2 f(y)dy$ over all probability densities f that are bounded by 1.)

[54] Note that $\int |p-q|d\mu = \int_{p\ge q}(p-q)d\mu + \int_{q>p}(q-p)d\mu = \left[\int (p-q)d\mu - \int_{q>p}(p-q)d\mu\right] + \int_{q>p}(q-p)d\mu = 0 + 2\int_{q>p}(q-p)d\mu$.

12. Lehmann and Romano [8, Problem 12.15].
13. Lehmann and Romano [8, Problem 12.46].
14. Lehmann and Romano [8, Problem 12.54].
15. Lehmann and Romano [8, Problem 12.64].
16. Lehmann and Romano [8, Problem 12.66].
17. Lehmann and Romano [8, Problem 13.20].
18. Lehmann and Romano [8, Problem 13.21].

3 Kernel density estimation

Let $X_1, \ldots, X_n$ be i.i.d. random variables having a probability density p (with respect to the Lebesgue measure on $\mathbb{R}$) and distribution function $F(x) := \int_{-\infty}^{x} p(t)dt$ (for $x \in \mathbb{R}$). Here F and p are unknown.

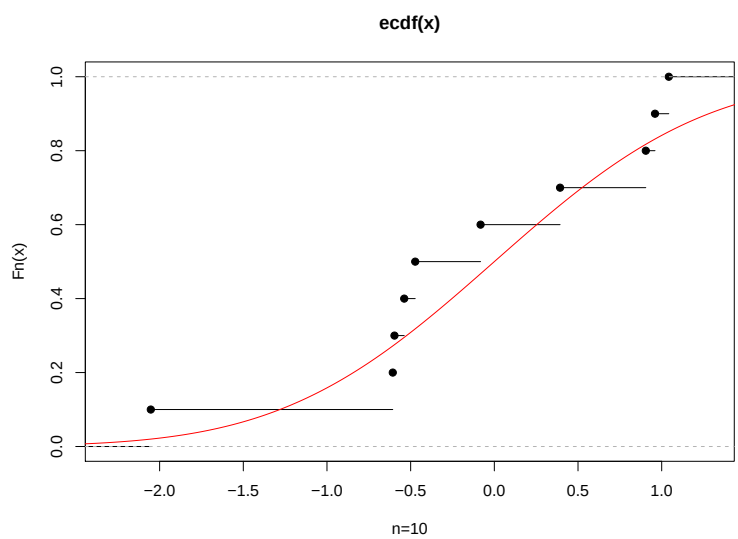
Question: Can we estimate F and p nonparametrically, making *minimal* assumptions?

Definition 3.1 (Empirical distribution function)**.** A natural estimator of F is the *empirical cumulative distribution function* (ECDF):

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \qquad \text{for } \ x \in \mathbb{R}, \tag{73}$$

where $I(\cdot)$ denotes the indicator function.

The ECDF is the distribution function of the *empirical distribution* $\mathbb{P}_n$, the probability distribution that puts mass $1/n$ at each of the data points X_i, i.e., $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where δ_x denotes the Dirac measure (i.e., $\delta_x(A) = 1$ if $x \in A$, and 0 otherwise). The following figure shows the ECDF obtained from $n = 10$ samples from a standard normal distribution overlaid with the true DF of $N(0, 1)$ (in red).



The Glivenko-Cantelli theorem shows that

$$\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| \overset{a.s.}{\to} 0, \qquad \text{as } n \to \infty^{55}.$$

Further we know that for every $x \in \mathbb{R}$,

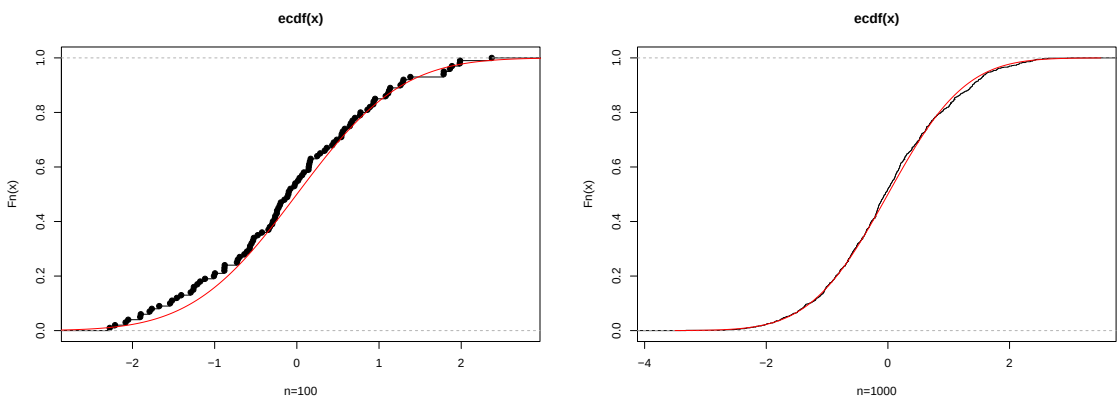
$$\sqrt{n}(\mathbb{F}_n(x) - F(x)) \overset{d}{\to} N(0, F(x)(1 - F(x))).$$

[55] Exercise 1 (HW3): Prove this.

Exercise 2 (HW3): Consider testing $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ where F_0 is a known continuous strictly increasing distribution function (e.g., standard normal) when we observe i.i.d. data $X_1, \ldots, X_n$ from F. The Kolmogorov-Smirnov (KS) test statistic is to consider

$$D_n := \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F_0(x)|,$$

and reject H_0 when $D_n > c_\alpha$, for a suitable $c_\alpha > 0$ (where α is the level of the test). Show that, under H_0, D_n is *distribution-free*, i.e., the distribution of D_n does not depend on F_0 (as long as it is continuous and strictly increasing). How would you compute (approximate/simulate) the critical value c_α, for every n.



As $F_n(x) \xrightarrow{p} F(x)$ for all $x \in \mathbb{R}$, we can also say that

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_n(x)(1 - F_n(x))}} \xrightarrow{d} N(0,1), \qquad \text{for every } x \in \mathbb{R}.$$

Thus, an asymptotic $(1 - \alpha)$ CI for $F(x)$ is

$$\left[F_n(x) - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{F_n(x)(1 - F_n(x))}, F_n(x) + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{F_n(x)(1 - F_n(x))}\right].$$

Likewise, we can also test the hypothesis $H_0 : F(x) = F_0(x)$ versus $H_1 : F(x) \neq F_0(x)$ for some known fixed c.d.f F_0, and $x \in \mathbb{R}$.

Let us come back to the estimation of p. As p is the derivative of F, for small $h > 0$, we can write the approximation

$$p(x) \approx \frac{F(x + h) - F(x - h)}{2h}.$$

As $\mathbb{F}_n$ is a natural estimator of F, it is intuitive to define the following (Rosenblatt) estimator of p:

$$\hat{p}_n^R(x) = \frac{\mathbb{F}_n(x + h) - \mathbb{F}_n(x - h)}{2h}.$$

We can rewrite $\hat{p}_n^R$ as

$$\hat{p}_n^R(x) = \frac{1}{2nh}\sum_{i=1}^n I(x-h < X_i \le x+h) = \frac{1}{nh}\sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right),$$

where $K_0(u) = \frac{1}{2}I_{(-1,1]}(u)$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{p}_n(x) := \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \tag{74}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)du = 1$. Such a function K is called a *kernel* and the parameter h is called the *bandwidth* of the estimator (74). The function $\hat{p}_n$ is called the *kernel density estimator* (KDE) or the Parzen-Rosenblatt estimator. Some classical examples of kernels are the following:

$$\begin{aligned}
K(u) &= \tfrac{1}{2}I(|u| \le 1) && \text{(the rectangular kernel)}\\
K(u) &= \frac{1}{\sqrt{2\pi}}\exp(-u^2/2) && \text{(the Gaussian kernel)}\\
K(u) &= \tfrac{3}{4}(1-u^2)I(|u| \le 1) && \text{(the Epanechnikov kernel).}
\end{aligned}$$

Note that if the kernel K takes only nonnegative values and if $X_1, \ldots, X_n$ are fixed, then $\hat{p}_n$ is a probability density.

The Parzen-Rosenblatt estimator can be generalized to the multidimensional case easily. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with (joint) density $p(\cdot, \cdot)$. A kernel estimator of p is then given by

$$\hat{p}_n(x, y) := \frac{1}{nh^2}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right), \tag{75}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel defined as above and $h > 0$ is the bandwidth.

3.1 The choice of the bandwidth and the kernel

It turns out that the choice of the bandwidth h is far more crucial for the quality of $\hat{p}_n$ as an estimator of p than the choice of the kernel K. We can view the KDE (for unimodal, nonnegative kernels) as the sum of n small "mountains" given by the functions

$$x \mapsto \frac{1}{nh}K\left(\frac{X_i - x}{h}\right).$$

Every small mountain is centered around an observation X_i and has area $1/n$ under it, for any bandwidth h. For a small bandwidth the mountain is very concentrated (peaked), while for a large bandwidth the mountain is low and fat. If the bandwidth is small, then the mountains remain separated and their sum is peaky. On the other hand, if the bandwidth

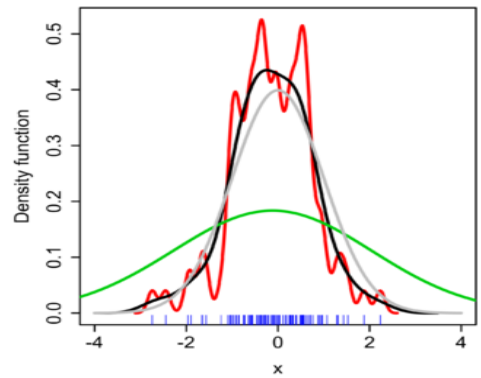


Figure 1: KDE with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with h=0.05. Black: KDE with h=0.337. Green: KDE with h=2.

is large, then the sum of the individual mountains is too flat. Intermediate values of the bandwidth should give the best results.

For a fixed h, the KDE $\hat{p}_n(x_0)$ is not consistent in estimating $p(x_0)$, where $x_0 \in \mathbb{R}$. However, if the bandwidth decreases with sample size at an appropriate rate, then it is, regardless of which kernel is used.

Exercise 3 (HW3): Suppose that p is continuous at x_0, that $h_n \to 0$, and that $nh_n \to \infty$ as $n \to \infty$. Then, $\hat{p}_n(x_0) \to p(x_0)$ in probability under the following assumptions on the kernel K: (i) $\int K(u)du = 1$, (ii) $\sup_{u\in\mathbb{R}} |K(u)| < \infty$, (iii) $\int |K(u)|du < \infty$, and (iv) $\lim_{u\to\infty} |uK(u)| = 0$. (Hint: Study the bias and variance of the estimator separately; see Parzen [9])].

3.2 Mean squared error of kernel density estimator (KDE)

A basic measure of the accuracy of $\hat{p}_n$ is its *mean squared risk* (or *mean squared error*) at an arbitrary fixed point $x_0 \in \mathbb{R}$:

$$\text{MSE} = \text{MSE}(x_0) := \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big].$$

Here $\mathbb{E}_p$ denotes the expectation with respect to the distribution of $(X_1, \ldots, X_n)$:

$$\text{MSE}(x_0) := \int \cdots \int \Big(\hat{p}_n(x_0; z_1, \ldots, z_n) - p(x_0)\Big)^2 \Big[\prod_{i=1}^n p(z_i)\Big] dz_1 \ldots dz_n.$$

Of course,

$$\text{MSE}(x_0) = b^2(x_0) + \sigma^2(x_0)$$

where

$$b(x_0) := \mathbb{E}_p[\hat{p}_n(x_0)] - p(x_0), \qquad \text{(bias)}$$

and

$$\sigma^2(x_0) := \mathbb{E}_p\Big[\Big(\hat{p}_n(x_0) - \mathbb{E}_p[\hat{p}_n(x_0)]\Big)^2\Big] \qquad \text{(variance)}.$$

To evaluate the mean squared risk of $\hat{p}_n$ we will analyze separately its variance and bias.

3.2.1 Variance of KDE

Proposition 3.2 (Variance of $\hat{p}_n$)**.** Suppose that the density p satisfies $p(x) \le p_{\max} < \infty$ for all $x \in \mathbb{R}$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function such that

$$\int K^2(u)du < \infty.$$

Then for any $x_0 \in \mathbb{R}$, $h > 0$, and $n \ge 1$ we have

$$\sigma^2(x_0) \le \frac{C_1}{nh},$$

where $C_1 = p_{\max} \int K^2(u)du$.

Proof. Observe that $\hat{p}_n(x_0)$ is an average of n i.i.d. random variables and so

$$\sigma^2(x_0) = \mathrm{Var}(\hat{p}_n(x_0)) = \frac{1}{n}\mathrm{Var}\left(\frac{1}{h}K\left(\frac{X_1 - x_0}{h}\right)\right) \le \frac{1}{nh^2}\mathbb{E}_p\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right]$$

Now, observe that

$$\mathbb{E}_p\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right] \;=\; \int K^2\left(\frac{z - x_0}{h}\right)p(z)dz \le p_{\max}\, h \int K^2(u)du.$$

Combining the above two displays we get the desired result. □

Thus, we conclude that if the bandwidth $h \equiv h_n$ is such that $nh \to \infty$ as $n \to \infty$, then the variance of $\sigma^2(x_0)$ goes to 0 as $n \to \infty$.

3.2.2 Bias of KDE

To analyze the bias of the KDE (as a function of h) we need certain conditions on the density p and on the kernel K. In what follows, for $\beta > 0$ we let $\lfloor \beta \rfloor$ denote the greatest integer *strictly less* than β.

Definition 3.3. Let T be an interval in $\mathbb{R}$ and let β and L be two positive numbers. The *Hölder class* $\Sigma(\beta, L)$ on T is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose derivative $f^{(\ell)}$ satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \le L|x - x'|^{\beta - \ell}, \qquad \text{for all } x, x' \in T.$$

The special case $\beta = 1$ is sometimes called the Lipschitz space[56]. If $\beta = 2$ then we have

$$|f'(x) - f'(x')| \leq L|x - x'|, \qquad \text{for all } x, x' \in T.$$

Roughly speaking, this means that the functions have bounded "second" derivative (f has second derivative a.e.).

Definition 3.4. Let $\ell \geq 1$ be an integer. We say that $K : \mathbb{R} \to \mathbb{R}$ is a *kernel of order* ℓ if the functions $u \mapsto u^j K(u)$, $j = 0, 1, \ldots, \ell$, are integrable and satisfy

$$\int K(u)du = 1, \qquad \int u^j K(u)du = 0, \quad j = 1, \ldots, \ell.$$

Does bounded kernels of order ℓ exist? See Tsybakov [14, Section 1.2.2] for constructing such kernels.

Observe that when $\ell \geq 2$ then the kernel has to take negative values which may lead to negative values of $\hat{p}_n$. This is sometimes mentioned as a drawback of using higher order kernels ($\ell \geq 2$). However, observe that we can always define the estimator

$$\hat{p}_n^+(x) = \max\{0, \hat{p}_n(x)\}$$

whose risk is smaller than or equal to the risk of $\hat{p}_n(x)$:

$$\mathbb{E}_p\Big[(\hat{p}_n^+(x_0) - p(x_0))^2\Big] \leq \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big], \qquad \forall x \in \mathbb{R}.$$

Suppose now that p belong to a class of densities $\mathcal{P} = \mathcal{P}(\beta, L)$ defined as follows:

$$\mathcal{P}(\beta, L) := \left\{ p : p \geq 0, \int p(x)dx = 1, \text{and } p \in \Sigma(\beta, L) \text{ on } \mathbb{R} \right\}.$$

Proposition 3.5 (Bias of $\hat{p}_n$)**.** Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)|du < \infty.$$

Then for any $x_0 \in \mathbb{R}$, $h > 0$, and $n \geq 1$ we have

$$|b(x_0)| \leq C_2 h^\beta, \tag{76}$$

where $C_2 = \frac{L}{\ell!} \int |u|^\beta |K(u)|du$.

[56] A Lipschitz function $g : \mathbb{R} \to \mathbb{R}$ is absolutely continuous and therefore is differentiable a.e., that is, differentiable at every point outside a set of Lebesgue measure zero. Its derivative is essentially bounded in magnitude by the Lipschitz constant, and for $a < b$, the difference $g(b)?g(a)$ is equal to the integral of the derivative g' on the interval $[a, b]$.

Proof. We have

$$\begin{aligned} b(x_0) &= \frac{1}{h}\int K\left(\frac{z-x}{h}\right)p(z)dz - p(x_0) \\ &= \int K(u)\Big[p(x_0+uh)-p(x_0)\Big]du. \end{aligned}$$

Next, using Taylor's theorem[57], we get

$$p(x_0+uh) = p(x_0) + p'(x_0)uh + \ldots + \frac{(uh)^\ell}{\ell!}p^{(\ell)}(x_0+\tau uh),$$

where $0 \le \tau \le 1$. Since K has order $\ell = \lfloor \beta \rfloor$, we obtain

$$\begin{aligned} b(x_0) &= \int K(u)\frac{(uh)^\ell}{\ell!}p^{(\ell)}(x_0+\tau uh)du \\ &= \int K(u)\frac{(uh)^\ell}{\ell!}(p^{(\ell)}(x_0+\tau uh) - p^{(\ell)}(x_0))du, \end{aligned}$$

and

$$\begin{aligned} |b(x_0)| &\le \int |K(u)|\frac{|uh|^\ell}{\ell!}\Big|p^{(\ell)}(x_0+\tau uh) - p^{(\ell)}(x_0)\Big|du \\ &\le L\int |K(u)|\frac{|uh|^\ell}{\ell!}|\tau uh|^{\beta-\ell}du \le C_2 h^\beta. \end{aligned}$$

□

From Propositions 3.2 and 3.5, we see that the upper bounds on the bias and variance behave in opposite ways as the bandwidth h varies. The variance decreases as h grows, whereas the bound on the bias increases. The choice of a small h corresponds to a large variance and leads to *undersmoothing*. Alternatively, with a large h the bias cannot be reasonably controlled, which leads to *oversmoothing*. An optimal value of h that balances

[57] **Taylor's theorem**: Let $k \ge 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $R_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + R_k(x),$$

where $R_k(x) = o(|x-a|^k)$ as $x \to a$.

Mean-value forms of the remainder: Let $f : \mathbb{R} \to \mathbb{R}$ be $k+1$ times differentiable on the open interval with $f^{(k)}$ continuous on the closed interval between a and x. Then

$$R_k(x) = \frac{f^{(k+1)}(\xi_L)}{(k+1)!}(x-a)^{k+1}$$

for some real number ξ_L between a and x. This is the *Lagrange form* of the remainder.

Integral form of the remainder: Let $f^{(k)}$ be absolutely continuous on the closed interval between a and x. Then

$$R_k(x) = \int_a^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k\,dt. \tag{77}$$

Due to absolute continuity of $f^{(k)}$, on the closed interval between a and x, $f^{(k+1)}$ exists a.e.

bias and variance is located between these two extremes. To get an insight into the optimal choice of h, we can minimize in h the upper bound on the MSE obtained from the above results.

If p and K satisfy the assumptions of Propositions 3.2 and 3.5, we obtain

$$\text{MSE} \leq C_2^2 h^{2\beta} + \frac{C_1}{nh}. \tag{78}$$

The minimum with respect to h of the right hand side of the above display is attained at

$$h_n^* = \left(\frac{C_1}{2\beta C_2^2}\right)^{1/(2\beta+1)} n^{-1/(2\beta+1)}.$$

Therefore, the choice $h = h_n^*$ gives

$$\text{MSE}(x_0) = O\Big(n^{-\frac{2\beta}{2\beta+1}}\Big), \quad \text{as } n \to \infty,$$

uniformly in x_0. Thus, we have the following result.

Theorem 3.6. Assume that the conditions of Proposition 3.5 hold and that $\int K^2(u)du < \infty$. Fix $\alpha > 0$ and take $h = \alpha n^{-1/(2\beta+1)}$. Then for $n \geq 1$, the KDE $\hat{p}_n$ satisfies

$$\sup_{x_0 \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where $C > 0$ is a constant depending only on β, L, α and on the kernel K.

Proof. We apply (78) to derive the result. To justify the application of Proposition 3.2, it remains to prove that there exists a constant $p_{\max} < \infty$ satisfying

$$\sup_{x \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} p(x) \leq p_{\max}. \tag{79}$$

To show that (79) holds, consider K^* which is a bounded kernel of order ℓ (not necessarily equal to K). Applying Proposition 3.5 with $h = 1$ we get that, for any $x \in \mathbb{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$\Big| \int K(z - x)\, p(z) dz - p(x) \Big| \leq C_2^* := \frac{L}{\ell!} \int |u|^\beta |K^*(u)| du.$$

Therefore, for any $x \in \mathbb{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$p(x) \leq C_2^* + \int |K^*(z - x)|\, p(z) dz \leq C_2^* + K_{\max}^*,$$

where $K_{\max}^* = \sup_{u \in \mathbb{R}} |K^*(u)|$. Thus, we get (79) with $p_{\max} = C_2^* + K_{\max}^*$. □

Under the assumptions of Theorem 3.6, the rate of convergence of the estimator $\hat{p}_n(x_0)$ is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$, which means that for a finite constant C and for all $n \geq 1$ we have

$$\sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big] \leq C \psi_n^2.$$

Now the following two questions arise. Can we improve the rate ψ_n by using other density estimators? What is the best possible rate of convergence? To answer these questions it is useful to consider the minimax risk R_n^* associated to the class $\mathcal{P}(\beta, L)$:

$$R_n^*(\mathcal{P}(\beta, L)) = \inf_{T_n} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p\Big[(T_n(x_0) - p(x_0))^2\Big],$$

where the infimum is over all estimators. One can prove a lower bound on the minimax risk of the form $R_n^*(\mathcal{P}(\beta, L)) \geq C'\psi_n^2 = C'n^{-\frac{2\beta}{2\beta+1}}$ with some constant $C' > 0$. This implies that under the assumptions of Theorem 3.6 the KDE attains the optimal rate of convergence $n^{-\frac{\beta}{2\beta+1}}$ associated with the class of densities $\mathcal{P}(\beta, L)$. Exact definitions and discussions of the notion of optimal rate of convergence will be given later.

Remark 3.1. Quite often in practice it is assumed that $\beta = 2$ and that p'' is continuous at x_0. Also, the kernel is taken to be of order one and symmetric around 0. Then it can be shown that (Exercise 4 (HW3))

$$\mathrm{MSE}(x_0) = \frac{1}{nh}\int K^2(u)du p(x_0) + \frac{1}{4}h^4\left(\int u^2 K(u)du\right)^2 p''(x_0)^2 + o((nh)^{-1} + h^4).$$

You may assume that p is bounded from above.

Remark 3.2. Since $2\beta/(2\beta+1)$ approaches 1 as β becomes large, Theorem 3.6 implies that, for sufficiently smooth densities, the convergence rate can be made arbitrarily close to the parametric n^{-1} convergence rate. The fact that higher-order kernels can achieve improved rates of convergence means that they will eventually dominate first-order kernel estimators for large n. However, this does not mean that a higher-order kernel will necessarily improve the error for sample sizes usually encountered in practice, and in many cases, unless the sample size is very large there may actually be an increase in the error due to using a higher-order kernel.

3.3 Pointwise asymptotic distribution

Whereas the results from the previous subsection have shown us that $\hat{p}_n(x_0)$ converges to $p(x_0)$ in probability under certain assumptions, we cannot straightforwardly use this for statistical inference. Ideally, if we want to estimate $p(x_0)$ at the point x_0, we would like to have exact confidence statements of the form

$$\mathbb{P}\left(p(x_0) \in [\hat{p}_n(x_0) - c(n, \alpha, x_0, K), \hat{p}_n(x_0) - c(n, \alpha, x_0, K)]\right) \geq 1 - \alpha,$$

where α is the significance level and $c(n, \alpha, x_0, K)$ sequence of constants that one would like to be as small as possible (given α).

Theorem 3.7. Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)| du < \infty.$$

Suppose that p also satisfies $p(x) \leq p_{\max} < \infty$ for all $x \in \mathbb{R}$. Let K further satisfy (a) $\|K\|_2^2 := \int K^2(u)du < \infty$, (b) $\|K\|_\infty := \sup_{u \in \mathbb{R}} K(u) < \infty$. Suppose that the sequence of bandwidths $\{h_n\}_{n=1}^\infty$ satisfy $h_n \to 0$, $nh_n \to \infty$, and $n^{1/2} h_n^{\beta+1/2} \to 0$ as $n \to \infty$. Then, as $n \to \infty$,

$$\sqrt{nh_n}\left(\hat{p}_n(x_0) - p(x_0)\right) \xrightarrow{d} N\Big(0, p(x_0)\|K\|_2^2\Big).$$

Proof. We first find the limit for the 'variance term'. We use the Lindeberg-Feller central limit theorem for triangular arrays of independent random variables[58] with

$$Y_{ni} := \sqrt{nh}\frac{1}{nh}K\left(\frac{X_i - x_0}{h}\right) = \sqrt{\frac{1}{nh}}K\left(\frac{X_i - x_0}{h}\right), \qquad i = 1, \ldots, n,$$

so that $Y_{n1}, \ldots, Y_{nn}$ are i.i.d. and we have

$$\sqrt{nh}\left(\hat{p}_n(x_0) - \mathbb{E}_p[\hat{p}_n(x_0)]\right) = \sum_{i=1}^n (Y_{ni} - \mathbb{E}(Y_{ni})).$$

Thus, we only need to show that the two conditions in the Lindeberg-Feller CLT hold. Clearly,

$$\begin{aligned} n\mathbb{E}(Y_{ni}^2) &= \frac{1}{h}\int K^2\left(\frac{z - x_0}{h}\right)p(z)dz \\ &= \int K^2(u)\, p(x_0 + uh)du \to p(x_0)\int K^2(u)du, \qquad \text{as } n \to \infty, \end{aligned}$$

by the dominated convergence theorem (DCT), since $p(\cdot)$ is continuous at x_0 and bounded on $\mathbb{R}$. Now,

$$\begin{aligned} n\mathbb{E}(Y_{ni})^2 &= \frac{1}{h}\left(\int K\left(\frac{z - x_0}{h}\right)p(z)dz\right)^2 = h\left(\int K(u)p(x_0 + uh)du\right)^2 \\ &\leq h\|K\|_2^2\, p_{\max} \to 0, \qquad \text{as } n \to \infty, \end{aligned}$$

which shows that $\sum_{i=1}^n \mathbb{E}[(Y_{ni} - \mathbb{E}(Y_{ni}))^2] \to p(x_0)\int K^2(u)du$. Furthermore,

$$|Y_{ni}| \leq \frac{1}{\sqrt{nh}}\|K\|_\infty \to 0, \qquad \text{as } n \to \infty,$$

[58] **Lindeberg-Feller CLT** (see e.g., van der Vaart [15, p.20]): For each n let $Y_{n1}, \ldots, Y_{nn}$ be independent random variables with finite variances. If, as $n \to \infty$, (i) $\sum_{i=1}^n \mathbb{E}[Y_{ni}^2 I(|Y_{ni}| > \epsilon)] \to 0$, for every $\epsilon > 0$, and (ii) $\sum_{i=1}^n \mathbb{E}[(Y_{ni} - \mathbb{E}(Y_{ni}))^2] \to \sigma^2$, then

$$\sum_{i=1}^n (Y_{ni} - \mathbb{E}(Y_{ni})) \xrightarrow{d} N(0, \sigma^2), \qquad \text{as } n \to \infty.$$

by the assumption on the sequence of bandwidths. Thus, $I(|Y_{ni}| > \epsilon) \to 0$, for every $\epsilon > 0$ and by the DCT

$$\sum_{i=1}^{n} \mathbb{E}[Y_{ni}^2 I(|Y_{ni}| > \epsilon)] = \mathbb{E}[nY_{n1}^2 I(|Y_{n1}| > \epsilon)] \to 0.$$

By (76) we see that the bias term can be bounded above as

$$\sqrt{nh}|b(x_0)| \leq \sqrt{nh}h^{\beta} \to 0, \qquad \text{as } n \to \infty.$$

Therefore, we have the desired result. □

Exercise 5 (HW3): Suppose that you are given an i.i.d. sample from a bounded density p with bounded derivatives at x_0. Suppose that $c(\alpha, x_0)$ is such that $\mathbb{P}(-c(\alpha, x_0) \leq Z \leq c(\alpha, x_0)) = 1 - \alpha$ where $Z \sim N(0, p(x_0))$. Use a kernel density estimator (with a suitable kernel) to obtain a 95 percent confidence interval (CI) for $p(x_0)$ in such a way that the size of the interval shrinks at rate $1/\sqrt{nh_n}$ as $n \to \infty$, and that h_n can be chosen so that this rate is 'almost' (say, up to a $\log n$ term) of order $n^{-1/3}$.

Exercise 6 (HW3): Under the setup of Remark 3.1 and the assumption that $h = \alpha n^{-1/5}$, where $\alpha > 0$, find the asymptotic distribution of $\sqrt{nh}(\hat{p}_n(x_0) - p(x_0))$. Can this be used to construct a CI for $p(x_0)$? What are the advantages/disadvantages of using this result versus the setup of Theorem 6.16 with $\beta = 2$ to construct a CI for $p(x_0)$?

3.4 Introduction to kernel regression

Regression models are used to study how a dependent or response variable depends on a predictor variable. Let (X, Y) be a pair of real-valued jointly distributed random variables such that $\mathbb{E}|Y| < \infty$. The *regression function* $f : \mathbb{R} \to \mathbb{R}$ of Y on X is defined as

$$f(x) := \mathbb{E}[Y|X = x].$$

Suppose that we have a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n i.i.d. pairs of random variables having the same distribution as (X, Y).

Question: How to estimate f *nonparametrically* from the data?

Classically, the regression function f is assumed to lie in a class of functions specified by a finite number of parameters (e.g., linear regression). The nonparametric approach only assumes that $f \in \mathcal{F}$, where $\mathcal{F}$ is a given nonparametric class of functions. The set of values $\{X_1, \ldots, X_n\}$ is called the *design*. Here the design is random.

The *conditional residual* $\xi := Y - \mathbb{E}[Y|X]$ has mean zero, i.e., $\mathbb{E}(\xi) = 0$ (by definition), and we may write

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{80}$$

where ξ_i are i.i.d. random variables with the same distribution as ξ. In particular, $\mathbb{E}(\xi_i) = 0$ for all $i = 1, \ldots, n$. The variables ξ_i can therefore be interpreted as "errors".

Idea: The key idea we use in estimating f nonparametrically in this section is called *"local averaging"*. Given a kernel K and a bandwidth h, one can construct kernel estimators for nonparametric regression. There exist different types of kernel estimators of the regression function f. The most celebrated one is the Nadaraya-Watson estimator defined as follows:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, \qquad \text{if } \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0,$$

and $f_n^{NW}(x) = 0$ otherwise. This estimator was proposed separately in two papers by Nadaraya and Watson in the year 1964.

Example: If we choose $K(u) = \frac{1}{2} I(|u| \leq 1)$, then $f_n^{NW}(x)$ is the average of Y_i such that $X_i \in [x - h, x + h]$. Thus, for estimating $f(x)$ we define the "local" neighborhood as $[x - h, x + h]$ and consider the average of the observations in that neighborhood. For fixed n, the two extreme cases for the bandwidth are:

(i) $h \to \infty$. Then $f_n^{NW}(x)$ tends to $n^{-1} \sum_{i=1}^n Y_i$ which is a constant independent of x. The systematic error (bias) can be too large. This is a situation of *oversmoothing*.

(ii) $h \to 0$. Then $f_n^{NW}(X_i) = Y_i$ whenever $h < \min_{i,j} |X_i - X_j|$ and $\lim_{h \to 0} f_n^{NW}(x) = 0$, if $x \neq X_i$. The estimator f_n^{NW} is therefore too oscillating: it reproduces the data Y_i at the points X_i and vanishes elsewhere. This makes the stochastic error (variance) too large. In other words, *undersmoothing* occurs.

Thus, the bandwidth h defines the "width" of the local neighborhood and the kernel K defines the "weights" used in averaging the response values in the local neighborhood. As we saw in density estimation, an appropriate choice of the bandwidth h is more important than the choice of the kernel K.

The Nadaraya-Watson estimator can be represented as a weighted sum of the Y_i:

$$f_n^{NW}(x) = \sum_{i=1}^n Y_i W_i^{NW}(x)$$

where the weights are

$$W_i^{NW}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} \, I\left(\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right) \neq 0\right).$$

Definition 3.8. An estimator $\hat{f}_n(x)$ of $f(x)$ is called a *linear nonparametric regression estimator* if it can be written in the form

$$\hat{f}_n(x) = \sum_{i=1}^n Y_i W_{ni}(x)$$

where the weights $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ depend only on n, i, x and the values $X_1, \ldots, X_n$.

Typically, the weights $W_{ni}(x)$ of linear regression estimators satisfy the equality $\sum_{i=1}^n W_{ni}(x) = 1$ for all x (or for almost all x with respect to the Lebesgue measure).

Another intuitive motivation of f_n^{NW} is given below. Suppose that the distribution of (X, Y) has density $p(x, y)$ with respect to the Lebesgue measure and $p^X(x) = \int p(x, y)dy > 0$. Then,

$$f(x) = \mathbb{E}[Y|X = x] = \frac{\int yp(x, y)dy}{p^X(x)}.$$

If we replace here $p(x, y)$ by the KDE $\hat{p}_n(x, y)$ of the density of (X, Y) and use the corresponding KDE $\hat{p}_n^X(x)$ to estimate $p^X(x)$, we obtain $\hat{f}_n^{NW}$ in view of the following result.

Exercise 7 (HW3): Let $\hat{p}_n^X(x)$ and $\hat{p}_n(x, y)$ be the KDEs of p^X and p respectively (as defined in the previous lecture) with a kernel K of order 1. Then

$$f_n^{NW}(x) = \frac{\int y\hat{p}_n(x, y)dy}{\hat{p}_n^X(x)}$$

if $\hat{p}_n^X(x) \neq 0$.

4 U-statistics

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution P on $\mathbb{R}^d$. Given a known function $h : \mathbb{R} \to \mathbb{R}$, consider estimation of the "parameter"

$$\theta \equiv \theta(P) = \mathbb{E}_P[h(X_1, \ldots, X_r)] \tag{81}$$

(here we assume that $n \geq r$). We will assume that the function h is *permutation symmetric/invariant* in its r arguments[59], i.e., $h(\pi x) = h(x)$ for every $x \in \mathbb{R}^r$ and $\pi \in \Pi_r$, the set of all permutations of $\{1, \ldots, r\}$[60] (here $\pi x = (x_{\pi(1)}, \ldots, x_{\pi(r)})$ for $x = (x_1, \ldots, x_r)$).

Question: Why should we study estimation of θ as in (81)? The following examples show that many interesting parameters arising in statistics can be expressed in the form (81). In fact, a parameter $\theta(P)$ admits an unbiased estimator if and only if for some r there is an h such that (81) holds.

Example 4.1 (Population variance)**.** If $X_1, \ldots, X_n$ are i.i.d. random variables with mean $\mathbb{E}(X_1) = \mu$ then the population variance is defined as

$$\begin{aligned} \mathrm{Var}(X_1) = \mathbb{E}[(X_1 - \mu)^2] &= \frac{1}{2}\mathbb{E}\left[(X_1 - \mu)^2 + (X_2 - \mu)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\{(X_1 - \mu) - (X_2 - \mu)\}^2\right] = \mathbb{E}\left[\frac{1}{2}(X_1 - X_2)^2\right] \end{aligned}$$

Thus, $\theta = \mathrm{Var}(X_1)$ can be expressed in the form (81) with $\theta = \mathbb{E}[h(X_1, X_2)]$ where $h(x_1, x_2) := \frac{1}{2}(x_1 - x_2)^2$.

Example 4.2 (Gini mean difference)**.** When $h(x_1, x_2) = |x_1 - x_2|$, $\theta = \mathbb{E}[|X_1 - X_2|]$ is the mean pairwise deviation or Gini mean difference.

Question: What is a natural estimator of θ in (81)?

A natural unbiased estimator of θ in (81) is $h(X_1, \ldots, X_r)$. Since n observations (with $n \geq r$) are available, this simple estimator can be improved: By Rao-Blackwell theorem, the new unbiased estimator formed by computing the conditional expectation given a sufficient statistic has smaller variance. Here, for X_i's with values in $\mathbb{R}$, the vector of order statistics $(X_{(1)}, \ldots, X_{(n)})$ is sufficient; and for i.i.d. X_i's more generally, the empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is sufficient (see e.g., Dudley [2, Theorem 5.1.9, page 177]).

Definition 4.3 (U-statistics)**.** A U-statistic of order r with kernel h is defined as

$$U_n := \frac{1}{\binom{n}{r}} \sum_{(i_1, \ldots, i_r)} h(X_{i_1}, \ldots, X_{i_r}) \tag{82}$$

[59] A given h could always be replaced by a symmetric one as $\theta = \frac{1}{\binom{n}{r}} \sum_{(i_1, \ldots, i_r)} \mathbb{E}[h(X_{i_1}, \ldots, X_{i_r})]$.

[60] Thus, h is permutation invariant if the value of $h(x)$ does not change if we permute the components of x, i.e., for instance, when $r = 3$: $h((x_1, x_2, x_3)) = h((x_2, x_1, x_3)) = h((x_3, x_1, x_2)) = h((x_1, x_3, x_1)) = h((x_2, x_3, x_1)) = h((x_3, x_2, x_1))$.

where $(i_1, \ldots, i_r)$ denotes one of the $\binom{n}{r}$ collections of unordered subsets of r distinct integers chosen from $\{1, \ldots, n\}$.

Here "U" stands for "unbiased". The theory of U-statistics was introduced by Wassily Hoeffding in the 1940s.

Exercise 8 (HW3): Show that if $X_{(1)}, \ldots, X_{(n)}$ denote the values $X_1, \ldots, X_n$ stripped from their order (the order statistics in the case of real-valued variables), then $U_n = \mathbb{E}[h(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(n)}]$. Now, using the Rao-Blackwell theorem show that U_n has smaller variance than $h(X_1, \ldots, X_r)$.

Example 4.4 (Sample variance)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables and consider the (unbiased) sample variance:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Observe that

$$U_n := \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{1}{2}(X_i - X_j)^2 \quad = \quad \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2}(X_i - X_j)^2 = s_n^2.$$

by a simplification as in Example 4.1[61].

We will study many properties of U-statistics: Unbiasedness, lower variance, asymptotic variance, asymptotic distribution, etc.

Example 4.5 (U-statistic of degree $r = 1$)**.** U-statistic of degree $r = 1$ is a mean $U = n^{-1} \sum_{i=1}^{n} h(X_i)$. The asymptotic normality of U is then just a consequence of the CLT.

4.1 Projection

Although the asymptotic distribution of a U-statistic of degree $r = 1$ can be easily obtained by the CLT, i is not immediate how we can handle U-statistics of degree $r \geq 1$. The idea of a *projection*[62] of a random variables becomes important in this regard.

More generally, a common method to derive the limit distribution of a sequence of statistics T_n is to show that it is asymptotically equivalent to a sequence S_n of which the limit

[61] Note that $U = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(X_i - \bar{X}_n) - (X_j - \bar{X}_n)\right]^2 = \frac{1}{n(n-1)}[n \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \bar{X}_n)(X_j - \bar{X}_n)]$ which yields the desired result.

[62] In mathematics, the **Hilbert projection theorem** is a famous result of convex analysis that says that for every vector x in a Hilbert space H and every nonempty closed convex $C \subset H$, there exists a unique vector $y \in C$ for which $\|x - z\|$ is minimized over the vectors $z \in C$. This is, in particular, true for any closed subspace M of H. In that case, a necessary and sufficient condition for y is that the vector $x - y$ be orthogonal to M.

behavior is known. The basis of this method is Slutsky's lemma, which shows that the sequence $T_n = T_n - S_n + S_n$ converges in distribution to S if both $T_n - S_n \to 0$ and $S_n \xrightarrow{d} S$.

Question: How do we find a suitable sequence S_n?

First, the variables S_n must be of a simple form, because the limit properties of the sequence S_n must be known. Second, S_n must be close enough. One solution is to search for the closest S_n of a certain predetermined form. In this chapter, "closest" is taken as closest in square expectation.

Example 4.6 (Hájek projection)**.** Suppose that $X_1, \ldots, X_n$ are independent random vectors on $\mathbb{R}^d$, and let $\mathcal{S}$ denote the set of all variables of the form $\sum_{i=1}^n g_i(X_i)$, for for arbitrary measurable functions $g_i : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}[g_i^2(X_i)] < \infty$ for $i \in \{1, \ldots, n\}$. Let $T \equiv T(X_1, \ldots, X_n)$ be a statistic with finite second moment. The projection of T onto $\mathcal{S}$ is called the Hájek projection of T onto $\mathcal{S}$. What is the form the Hájek projection? We will answer this question below.

Definition 4.7 (Projection)**.** Let T and $\{S : S \in \mathcal{S}\}$ be random variables (defined on the same probability space) with finite second-moments. A random variable $\hat{S}$ is called a *projection* of T onto $\mathcal{S}$ (or L_2-projection) if $\hat{S} \in \mathcal{S}$ and minimizes

$$S \mapsto \mathbb{E}[(T - S)^2], \qquad \text{over} \quad S \in \mathcal{S}.$$

Often $\mathcal{S}$ is a linear space in the sense that $\alpha_1 S_1 + \alpha_2 S_2 \in \mathcal{S}$ for every $\alpha_1, \alpha_2 \in \mathbb{R}$, whenever $S_1, S_2 \in \mathcal{S}$. In this case $\hat{S}$ is the projection of T if and only if $T - \hat{S}$ is orthogonal to S for the inner product $\langle S_1, S_2 \rangle = \mathbb{E}[S_1 S_2]$. This is the content of the following theorem.

Theorem 4.8 (Projection onto a linear space)**.** Let $\mathcal{S}$ be a linear space of random variables with finite second moments. Then $\hat{S}$ is the projection of T onto $\mathcal{S}$ if and only if $\hat{S} \in \mathcal{S}$ and

$$\mathbb{E}[(T - \hat{S})S] = 0, \qquad \text{for every} \quad S \in \mathcal{S}. \tag{83}$$

Every two projections of T onto $\mathcal{S}$ are almost surely (a.s.) equal. If the linear space $\mathcal{S}$ contains the constant variables, then $\mathbb{E}[T] = \mathbb{E}[\hat{S}]$ and $\mathrm{Cov}(T - \hat{S}, S) = 0$ for every $S \in \mathcal{S}$.

Proof. For any S and $\hat{S} \in \mathcal{S}$,

$$\mathbb{E}[(T - S)^2] = \mathbb{E}[(T - \hat{S})^2] + 2\mathbb{E}[(T - \hat{S})(\hat{S} - S)] + \mathbb{E}[(\hat{S} - S)^2].$$

If $\hat{S}$ satisfies the orthogonality condition, then the middle term is zero (by (83) as $\hat{S} - S \in \mathcal{S}$), and we conclude that $\mathbb{E}[(T - S)^2] \geq \mathbb{E}[(T - \hat{S})^2]$, with strict inequality unless $\mathbb{E}[(\hat{S} - S)^2] = 0$ (i.e., $S = \hat{S}$ a.s.). Thus, the orthogonality condition implies that $\hat{S}$ is a projection, and also that it is unique a.s.

Conversely, for any number $\alpha \in \mathbb{R}$ and $S \in \mathcal{S}$,

$$\mathbb{E}[(T - \hat{S} - \alpha S)^2] - \mathbb{E}[(T - \hat{S})^2] = -2\alpha \mathbb{E}[(T - \hat{S})S] + \alpha^2 \mathbb{E}[S^2].$$

If $\hat{S}$ is a projection, then this expression is nonnegative for every α. But the parabola $\alpha \mapsto -2\alpha\mathbb{E}[(T-\hat{S})S] + \alpha^2\mathbb{E}[S^2]$ is nonnegative if and only if the orthogonality condition $\mathbb{E}[(T-\hat{S})S] = 0$ is satisfied.

If the constants are in $\mathcal{S}$, then the orthogonality condition implies $\mathbb{E}[(T-\hat{S})1] = 0$, whence the last assertions of the theorem follow. □

The orthogonality of $T - \hat{S}$ and $\hat{S}$ yields the Pythagorean rule

$$\mathbb{E}[T^2] = \mathbb{E}[(T-\hat{S})^2] + \mathbb{E}[\hat{S}^2]. \tag{84}$$

Now suppose a sequence of statistics T_n and linear spaces $\mathcal{S}_n$ is given. For each n, let $\hat{S}_n$ be the projection of T_n on $\mathcal{S}_n$. Then, as shown by the following result, the limiting behavior of the sequence T_n follows from that of $\hat{S}_n$ and vice versa, provided the quotient $\mathrm{Var}(T_n)/\mathrm{Var}(\hat{S}_n)$ converges to 1.

Theorem 4.9 (Asymptotic equivalence)**.** Let $\mathcal{S}_n$ be linear spaces of random variables with finite second moments that contain the constants. Let T_n be random variables with projections $\hat{S}_n$ onto $\mathcal{S}_n$. If $\mathrm{Var}(T_n)/\mathrm{Var}(\hat{S}_n) \to 1$[63], then

$$R_n := \frac{T_n - \mathbb{E}[T_n]}{\sqrt{\mathrm{Var}(T_n)}} - \frac{\hat{S}_n - \mathbb{E}[\hat{S}_n]}{\sqrt{\mathrm{Var}(\hat{S}_n)}} \xrightarrow{p} 0. \tag{85}$$

Proof. First note that $\mathbb{E}(R_n) = 0$. We will show that $\mathrm{Var}(R_n) \to 0$ as $n \to \infty$, which will complete the proof.

As $\mathcal{S}_n$ contain all constants, applying (83) with the constant random variable 1 we get $\mathbb{E}[(T_n - \hat{S}_n)1] = 0$, which implies $\mathbb{E}[T_n - \hat{S}_n] = 0$. Thus, $\mathrm{Cov}(T_n - \hat{S}_n, \hat{S}_n) = \mathbb{E}[(T_n - \hat{S}_n)\hat{S}_n] = 0$ where the last equality follows from (83). Thus,

$$\mathrm{Cov}(T_n, \hat{S}_n) = \mathrm{Cov}(T_n - \hat{S}_n, \hat{S}_n) + \mathrm{Var}(\hat{S}_n) = \mathrm{Var}(\hat{S}_n).$$

Therefore,

$$\mathrm{Var}(R_n) = 2 - \frac{2\mathrm{Cov}(T_n, \hat{S}_n)}{\sqrt{\mathrm{Var}(T_n)\mathrm{Var}(S_n)}} = 2 - \frac{2\mathrm{Var}(\hat{S}_n)}{\sqrt{\mathrm{Var}(T_n)\mathrm{Var}(S_n)}} = 2\left(1 - \sqrt{\frac{\mathrm{Var}(\hat{S}_n)}{\mathrm{Var}(T_n)}}\right) \to 0,$$

as $n \to \infty$, which completes the proof. □

In the preceding theorem it is essential that the $\hat{S}_n$ are the projections of the variables T_n, because the condition $\mathrm{Var}(T_n)/\mathrm{Var}(\hat{S}_n) \to 1$ for general sequences $\hat{S}_n$ and T_n, does not imply anything.

[63] The condition $\mathrm{Var}(T_n)/\mathrm{Var}(\hat{S}_n) \to 1$ in the theorem implies that the projections $\hat{S}_n$ are asymptotically of the same size as the original T_n.

4.1.1 The Hájek projection

Recall the setting of Example 4.6, in particular the definition of T and $\mathcal{S}$. The following result gives the form of the of T onto $\mathcal{S}$.

Lemma 4.10. The projection of an arbitrary random variable $T = T(X_1, \ldots, X_n)$ with finite second moment onto $\hat{S}$ is given by

$$\hat{S} := \sum_{i=1}^{n} \mathbb{E}[T|X_i] - (n-1)\mathbb{E}[T]. \tag{86}$$

$\hat{S}$ is called the *Hájek projection* of T onto $\mathcal{S}$.

Proof. Note that $\hat{S} \in \mathcal{S}$. Thus it suffices to show that (83) holds. But,

$$\begin{aligned}
\mathbb{E}[(T-\hat{S})S] &= \mathbb{E}[(T-\hat{S})\sum_{i=1}^{n} g_i(X_i)] = \sum_{i=1}^{n} \mathbb{E}[(T-\hat{S})g_i(X_i)] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}[(T-\hat{S})g_i(X_i)|X_i]\right] = \sum_{i=1}^{n} \mathbb{E}\left[g_i(X_i)\mathbb{E}[T-\hat{S}|X_i]\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[g_i(X_i)\left(\mathbb{E}[T|X_i] - \mathbb{E}[\hat{S}|X_i]\right)\right]
\end{aligned}$$

where for each $i \in \{1, \ldots, n\}$, by (86),

$$\begin{aligned}
\mathbb{E}[\hat{S}|X_i] &= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{E}(T|X_j)|X_i\right] - (n-1)\mathbb{E}(T) \\
&= (n-1)\mathbb{E}(T) + \mathbb{E}[T|X_i] - (n-1)\mathbb{E}(T) = \mathbb{E}[T|X_i]
\end{aligned}$$

This completes the proof. □

Note that if $X_1, \ldots, X_n$ are i.i.d. and $T \equiv T(X_1, \ldots, X_n)$ is permutation symmetric, then

$$\mathbb{E}[T|X_i = x] = \mathbb{E}[T|X_1 = x] = \mathbb{E}[T(x, X_2, \ldots, X_n)] \tag{87}$$

for all $i \in \{1, \ldots, n\}$ which does not depend on i (as T is permutation invariant).

Exercise 9 (HW3): Show that in this case $\hat{S}$ is also the projection of T onto the smaller set $\mathcal{S}_0$ consisting of all variables of the form $\sum_{i=1}^{n} g(X_i)$ for an arbitrary measurable function g (with finite second moment).

4.2 U-statistics and Hájek's projection

Recall the estimation problem (81) and our estimator (82) — the U-statistic U_n. The Hájek projection of U_n is $\hat{U}_n$ (see (86)) which leads to

$$\hat{U}_n - \theta = \sum_{i=1}^{n} \mathbb{E}[U_n|X_i] - n\theta = \frac{r}{n}\sum_{i=1}^{n} h_1(X_i) \tag{88}$$

where

$$h_1(x) := \mathbb{E}[h(x, X_2, \ldots, X_r)] - \theta.$$

The first equality in (88) is the consequence of the Hájek projection principle. The second equality is established in the proof below. The sequence $\sqrt{n}(\hat{U}_n - \theta)$ is asymptotically normal by the CLT, provided $\mathbb{E}[h_1^2(X_1)] < \infty$. The difference between U_n and $\hat{U}_n$ is asymptotically negligible.

Theorem 4.11 (Asymptotic normality of U-statistics)**.** If $\mathbb{E}[h^2(X_1, \ldots, X_r)] < \infty$ and $\xi_1 := \mathrm{Var}(h_1(X_1)) > 0$, then

$$\sqrt{n}(U_n - \hat{U}_n) \xrightarrow{p} 0. \tag{89}$$

Hence

$$\sqrt{n}(U_n - \theta) = \sqrt{n}(\hat{U}_n - \theta) + o_p(1) = \frac{r}{\sqrt{n}} \sum_{i=1}^{n} h_1(X_i) + o_p(1) \xrightarrow{d} N\left(0, r^2 \mathrm{Var}(h_1(X_1))\right)$$

where $\xi_1 = \mathrm{Var}(h_1(X_1)) = \mathrm{Cov}(h(X_1, X_2, \ldots, X_r), h(X_1, X_2', \ldots, X_r'))$. Here $X_2', \ldots, X_r'$ are i.i.d. having the same distribution of X_1 and independent of $X_1, \ldots, X_n$.

Proof. The first task is to prove that the form of the Hájek projection $\hat{U}_n$ is as claimed in (88). Since the X_i's are independent and h is permutation symmetric,

$$\mathbb{E}\left[h(X_{i_1}, \ldots, X_{i_r}) - \theta | X_i = x\right] = \begin{cases} h_1(x), & \text{if } i \in \{i_1, \ldots, i_r\} \\ 0, & \text{if } i \notin \{i_1, \ldots, i_r\}. \end{cases}$$

Thus,

$$\mathbb{E}[U_n - \theta | X_i] = \frac{1}{\binom{n}{r}} \sum_{(i_1, \ldots, i_r)} h_1(X_i) \mathbf{1}_{\{i_1, \ldots, i_r\}}(i) = \frac{\binom{n-1}{r-1}}{\binom{n}{r}} h_1(X_i) = \frac{r}{n} h_1(X_i).$$

Now summing these over i yields (88). We can calculate $\mathrm{Var}(\hat{U}_n)$ easily as

$$\mathrm{Var}(\hat{U}_n) = \frac{r^2}{n^2} \sum_{i=1}^{n} \mathrm{Var}(h_1(X_i)) = \frac{r^2}{n} \xi_1. \tag{90}$$

Since the random variables $Y_i := h_1(X_i)$ are i.i.d. with mean zero and finite variance, by the CLT, we have

$$\sqrt{n}(\hat{U}_n - \theta) \xrightarrow{d} N(0, r^2 \xi_1).$$

To show (89) we will apply Theorem 4.9, and thus it is enough to show[64]

$$\frac{\mathrm{Var}(\hat{U}_n)}{\mathrm{Var}(U_n)} \to 1 \tag{91}$$

[64] Exercise 10 (HW3): Show that (91) indeed yields (89).

Calculating $\mathrm{Var}(U_n)$ is a bit more involved which we do now. Note that

$$\mathrm{Var}(U_n) = \frac{1}{\binom{n}{r}^2} \sum_{(i_1,\ldots,i_r)} \sum_{(i'_1,\ldots,i'_r)} \mathrm{Cov}\left(h(X_{i_1}, X_{i_2}, \ldots, X_{i_r}), h(X_{i'_1}, X_{i'_2}, \ldots, X_{i'_r})\right).$$

Let k be the number of indices in common between $\{i_1, \ldots, i_r\}$ and $\{i'_1, \ldots, i'_r\}$. Then, since there are $\binom{n}{r}$ ways of choosing $\{i_1, \ldots, i_r\}$, $\binom{r}{k}$ ways of choosing k common indices from the choice of $\{i_1, \ldots, i_r\}$, and then $\binom{n-r}{r-k}$ ways of choosing the rest of the second block $\{i'_1, \ldots, i'_r\}$, we find that

$$\mathrm{Var}(U_n) \;=\; \frac{1}{\binom{n}{r}^2} \sum_{k=1}^{r} \binom{n}{r}\binom{r}{k}\binom{n-r}{r-k}\xi_k \;=\; \sum_{k=1}^{r} \frac{\binom{r}{k}\binom{n-r}{r-k}}{\binom{n}{r}}\xi_k \tag{92}$$

where[65] for $k \in \{1, \ldots, r\}$,

$$\xi_k := \mathrm{Cov}\left(h(X_1, \ldots, X_k, X_{k+1}, \ldots, X_r), h(X_1, \ldots, X_k, X'_{k+1}, \ldots, X'_r)\right).$$

The expression of $\mathrm{Var}(U_n)$ in (92) can be rewritten as

$$\begin{aligned} &\sum_{k=1}^{r} \frac{r!}{k!\,(r-k)!} \cdot \frac{(n-r)!}{(r-k)!\,(n-2r+k)!} \cdot \frac{(n-r)!\,r!}{n!}\,\xi_k \\ =\; &\sum_{k=1}^{r} \frac{r!^2}{k!\,(r-k)!^2} \frac{(n-r)\cdots(n-2r+k+1)}{n(n-1)\cdots(n-r+1)}\xi_k \;\sim\; r^2\xi_1\frac{1}{n} \qquad \text{if } \xi_1 > 0, \end{aligned}$$

as in the above sum the first term is $O(1/n)$, the second term is $O(1/n^2)$, and so forth. Putting this together with (92) yields (91) and completes the proof. □

Example 4.12 (Asymptotic normality of sample variance)**.** Recall the setting in Example 4.4. Here the parameter of interest is $\theta = \sigma^2 = \mathrm{Var}(X_1)$ and the U-statistic is the (unbiased) sample variance $U_n \equiv s_n^2$. Let us find the asymptotic distribution of s_n^2 (where $r = 2$). In this case $h_1(x) = \mathbb{E}[\frac{1}{2}(x - X_2)^2] - \sigma^2 = \frac{1}{2}(x^2 + \mathbb{E}[X_2^2]) - x\mathbb{E}[X_2] - \sigma^2$. Thus,

$$\begin{aligned} 2^2\xi_1 &= 4\mathrm{Var}(h_1(X_1)) = \mathrm{Var}(X_1^2 - 2X_1\mu) = \mathbb{E}[(X_1^2 - 2X_1\mu)^2] - (\mathbb{E}[X_1^2] - 2\mu^2)^2 \\ &= \mathbb{E}[X_1^4 - 4X_1^3\mu + 4X_1^2\mu^2] - (\mathbb{E}[X_1^2])^2 - 4\mu^4 + 4\mathbb{E}[X_1^2]\mu^2 \\ &= \mathbb{E}[X_1^4] - 4\mu\mathbb{E}[X_1^3] + 4\mathbb{E}[X_1^2]\mu^2 - (\sigma^2 + \mu^2)^2 - 4\mu^4 + 4(\sigma^2 + \mu^2)\mu^2 \\ &= \mathbb{E}[X_1^4] - 4\mu\mathbb{E}[X_1^3] - \sigma^4 + 3\mu^4 + 6\sigma^2\mu^2 \;=\; \mu_4 - \sigma^4. \end{aligned}$$

where $\mu_4 = \mathbb{E}[(X_1 - \mu)^4]$ is the 4th central moment. So $n\mathrm{Var}(U_n) \to \mu_4 - \sigma^4$. Thus, $\sqrt{n}(U_n - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4)$.

[65] Note that from the right hand term in (92) can be expressed as $\mathbb{E}[\xi_K]$ where $\xi_0 = 0$ and $K \sim \mathrm{Hypergeometric}(n, r)$ (sampling without replacement from an urn containing n balls, r of which are red and $n - r$ of which are blue), i.e.,

$$\mathbb{P}(K = k) = \frac{\binom{r}{k}\binom{n-r}{r-k}}{\binom{n}{r}} \qquad \text{for } k \in \{1, \ldots, r\}.$$

If $\xi_1 = 0$, we say that U_n is *degenerate*[66].

4.3 Exercises

12. Suppose we have bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The *Kendall's* τ-statistic is

$$\tau := \frac{4}{n(n-1)} \sum_{i<j} \mathbf{1}\left\{(Y_j - Y_i)(X_j - X_i) > 0\right\} - 1.$$

This statistic is a measure of dependence between X and Y and counts the number of concordant pairs (X_i, Y_i) and (X_j, Y_j) in the observations. Two pairs are *concordant* if the indicator in the definition of τ is equal to 1. Large values of τ indicate positive dependence (or concordance), whereas small values indicate negative dependence. Under independence of X and Y and continuity of their distributions, the distribution of τ is centered about zero, and in the extreme cases that all or none of the pairs are concordant τ is identically 1 or -1, respectively.

Show that $\tau + 1$ is a U-statistic of order 2, and find it's asymptotic distribution. Use this to develop an asymptotic level α test for "independence" (describe the test explicitly).

[66] Exercise 11 (HW3): Suppose that $X_1, \ldots, X_n$ are i.i.d. P, an unknown distribution on $\mathbb{R}$, with mean $\mathbb{E}(X_1) = \mu$ and variance $\sigma^2 = \mathrm{Var}(X_1) > 0$. How would one estimate the square of the mean, $\theta(P) = \mu^2$? Find the limiting distribution of the estimator. Suppose next that $\mu = \mathbb{E}(X_1) = 0$. Find a nondegenerate limiting distribution of the estimator in this case. [Hint: As $\mathbb{E}(X_1 X_2) = \mu^2$, it is an estimable parameter with degree at most 2.]

5 General linear model

The general linear model incorporates many of the most popular and useful models that arise in applied statistics, including models for multiple regression and the analysis of variance. The basic model can be written succinctly in matrix form as

$$Y = X\beta + \varepsilon \tag{93}$$

where $Y \in \mathbb{R}^n$ is the observed response vector, X is an $n \times p$ matrix of known constants (also known as the design matrix) consisting of the predictor variables, $\beta \in \mathbb{R}^p$ is an unknown parameter, and ε is a random vector in $\mathbb{R}^n$ of unobserved errors. We will first assume (the less stringent assumption) that the error vector $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)$ satisfies

$$\mathbb{E}[\epsilon_i] = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma^2 \ \text{ (for all } \ i) \qquad \text{and} \qquad \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0 \ \text{ for all } \ i \neq j. \tag{94}$$

Example 5.1 (Quadratic regression)**.** In quadratic regression, the response variable is modeled as a quadratic function of some explanatory variable plus a random error. Specifically, we model the observed data $\{(x_i, Y_i)\}_{i=1}^n$ as

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i, \tag{95}$$

Here the explanatory variables $x_1, \ldots, x_n$ are taken to be known constants, β_1, β_2, and β_3 are the unknown parameters, and ϵ_i's are unobserved errors satisfying (94). We can succinctly express (95) in the form (93) with

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

It is often more convenient to view the unknown mean of Y, namely,

$$\mu := \mathbb{E}[Y] = X\beta \tag{96}$$

in $\mathbb{R}^n$ as the unknown parameter. If $c_1, \ldots, c_p$ are the columns of X, then

$$\mu = X\beta = \beta_1 c_1 + \ldots + \beta_p c_p,$$

which shows that μ must be a linear combination of the columns of X. So μ must lie in the vector (sub)-space

$$\omega := \mathrm{span}\{c_1, \ldots, c_p\} = \{X\beta : \beta \in \mathbb{R}^p\}. \tag{97}$$

Using μ instead of β, the vector of unknown parameter is $\theta = (\mu, \sigma)$ taking values in $\Theta = \omega \times (0, \infty)$.

5.1 Estimation

To estimate $\beta \in \mathbb{R}^p$ it is natural to *project* the response vector $Y \in \mathbb{R}^n$ onto the column space of X and define

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \tag{98}$$

Here $\hat{\beta}$ is called the *least squares estimator* (LSE) of β. Of course, when the rank r of X is less than p, $\hat{\beta}$ will not be unique[67]. The following fundamental result from convex analysis can be used to characterize the LSE $\hat{\beta}$.

Theorem 5.2 (Hilbert projection theorem)**.** For every vector x in a Hilbert space $(H, \langle \cdot, \cdot \rangle)$[68] and every nonempty closed convex $C \subset H$, there exists a unique vector $y \in C$ such that

$$y = \operatorname*{argmin}_{z \in C} \|x - z\|^2. \tag{99}$$

This is, in particular, true for any closed subspace $M \subset H$. In that case, a necessary and sufficient condition for y to satisfy (99) is that the vector $x - y$ be orthogonal to M, i.e., $\langle x - y, z \rangle = 0$ for all $z \in M$. This closest point y (to x) is called the projection of x onto C.

Applying Theorem 5.2 with $H = \mathbb{R}^n$ (with the usual inner product) and $M = \omega$, we see that

$$\hat{Y} := \arg\min_{z \in \omega} \|Y - z\|^2 \;=\; \arg\min_{z = X\beta : \beta \in \mathbb{R}^p} \|Y - z\|^2$$

is the projection of Y onto ω. The mapping $Y \mapsto \hat{Y}$ is linear and can be represented by an $n \times n$ matrix P, i.e.,

$$\hat{Y} = PY,$$

with P called the (orthogonal) *projection matrix*[69] onto ω. Since $\hat{Y} \in \omega$, $P\hat{Y} = \hat{Y}$, and so $P^2 Y = P(PY) = P\hat{Y} = \hat{Y} = PY$. Because Y can take arbitrary values in $\mathbb{R}^n$, this shows that $P^2 = P$ (and thus the eigenvalues of P are either 0 or 1; why?). Matrices that satisfy the equation $P^2 = P$ are called *idempotent*.

Once we obtain $\hat{Y}$, we may find $\hat{\beta} \in \mathbb{R}^p$ such that $\hat{Y} = X\hat{\beta}$. If X has full column rank p, then $\hat{\beta}$ is also unique.

[67] Since Y has mean μ, it is fairly intuitive that our data must provide some information distinguishing between any two values for μ, since the distributions for Y under two different values for μ must be different. Whether this also holds for β depends on the rank r of X. Since X has p columns, this rank r is at most p. If the rank of X equals p then the mapping $\beta \mapsto X\beta$ is one-to-one, and each value $\mu \in \omega$ is the image of a unique value $\beta \in \mathbb{R}^p$. But if the columns of X are linearly dependent, then a nontrivial linear combination of the columns of X will equal zero, so $Xv = 0$ for some $v \neq 0$. But then $X(\beta + v) = X\beta + Xv = X\beta$, and parameters β and $\beta^* := \beta + v$ both give the same mean μ. Here our data Y provides no information to distinguish between parameter values β and β^*.

[68] See Appendix A.1 for a brief review and examples of Hilbert spaces.

[69] Tukey coined the term "hat matrix" for P because it puts the hat on Y.

A convenient way to calculate $\hat{\beta}$ and then $\hat{Y}$ using calculus is to realize that all partial derivatives of the least squares criterion $\|Y - X\beta\|^2$ must vanish at $\beta = \hat{\beta}$[70]. Another approach to characterizing $\hat{\beta}$ and $\hat{Y}$ proceeds via geometric considerations which we describe below. Since the columns c_i, for $i = 1, \ldots, p$, of X lie in ω, the *residual* vector

$$e := Y - \hat{Y} \tag{100}$$

must be orthogonal to every element in ω, i.e., we must have $c_i^\top e = 0$, for all $i = 1, \ldots, p$, which implies

$$X^\top e = 0.$$

Since $Y = \hat{Y} + e$,

$$X^\top Y = X^\top(\hat{Y} + e) = X^\top \hat{Y} + X^\top e = X^\top \hat{Y} = X^\top X \hat{\beta}.$$

If $X^\top X$ is invertible[71] then this gives

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

In this case

$$PY = \hat{Y} = X\hat{\beta} = X(X^\top X)^{-1} X^\top Y,$$

and so the projection matrix P onto ω can be written as

$$P = X(X^\top X)^{-1} X^\top.$$

Thus P is symmetric.

5.2 Gauss-Markov theorem

Recall the setting (93) where ε satisfies (94). One of the most famous results in statistics asserts that the LSE of β have the smallest variance among *all linear unbiased* estimates. We will make this precise here[72]. We focus on estimation of any linear combination of the parameter β, e.g.,

$$\xi := b^\top \beta$$

[70] Exercise 13 (HW3): Using this approach find expressions for $\hat{\beta}$ and then $\hat{Y}$.

[71] The matrix $X^\top X$ is invertible iff X has full column rank, i.e., $r = p$ (here $\text{rank}(X) = r$). In fact, $X^\top X$ is positive definite in this case. To see this, let $v \in \mathbb{R}^p$ be an eigenvector of $X^\top X$ with $\|v\| = 1$ and eigenvalue λ. Then

$$\|Xv\|^2 = v^\top X^\top X v = \lambda v^\top v = \lambda,$$

which must be strictly positive since $Xv = c_1 v_1 + \ldots + c_p v_p$ cannot be zero if X has full column rank (as $v \neq 0$).

[72] We should also make clear that the restriction to unbiased estimates is not necessarily a wise one. This observation will lead to considering biased estimates of β such as ridge regression.

with $b \in \mathbb{R}^p$ being a fixed vector. For example, when b is the j'th unit vector (i.e., $b = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^p$ with 1 being in the j'th place) we are interested in estimating β_j, for $j \in \{1, \ldots, p\}$.

It is natural to estimate ξ by an estimator of the form $a^\top Y = a_1 Y_1 + \ldots + a_n Y_n$, for some $a \in \mathbb{R}^n$. Such an estimator is called a *linear estimate* (as the estimator is a linear function of Y).

Question: Among all linear unbiased estimators of ξ which one has the smallest variance?

The following result answers this questions.

Theorem 5.3 (Gauss-Markov theorem)**.** Suppose that $X^\top X$ is invertible, and

$$\mathbb{E}[Y] = X\beta \qquad \text{and} \qquad \mathrm{Var}(Y) = \sigma^2 I_n.$$

Then $b^\top \hat{\beta} = b^\top (X^\top X)^{-1} X^\top Y$ is unbiased and has uniformly minimum variance among all linear unbiased estimators of η.

Proof. Let $\delta = a^\top Y$ be a competing linear unbiased estimator. Observe that if $\delta = a^\top Y$ is an unbiased estimator of η, then we must have

$$b^\top \beta = \eta = \mathbb{E}[a^\top Y] = a^\top X\beta \qquad \text{for all} \quad \beta \in \mathbb{R}^p. \tag{101}$$

Therefore, $a^\top X = b$. Note that

$$\begin{aligned}
\mathrm{Var}(a^\top Y) &= \mathrm{Var}(b^\top \hat{\beta} + (a^\top Y - b^\top \hat{\beta})) \\
&= \mathrm{Var}(b^\top \hat{\beta}) + \mathrm{Var}(a^\top Y - b^\top \hat{\beta}) + 2\mathrm{Cov}(b^\top \hat{\beta}, a^\top Y - b^\top \hat{\beta}).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{Cov}(b^\top \hat{\beta}, a^\top Y - b^\top \hat{\beta}) &= \mathrm{Cov}\left(b^\top (X^\top X)^{-1} X^\top Y \ , \ [a^\top - b^\top (X^\top X)^{-1} X^\top] Y\right) \\
&= [b^\top (X^\top X)^{-1} X^\top] \, (\sigma^2 I_n) \, [a - X(X^\top X)^{-1} b] \\
&= \sigma^2 b^\top (X^\top X)^{-1} \left[X^\top a - X^\top X (X^\top X)^{-1} b\right] \\
&= \sigma^2 b^\top (X^\top X)^{-1} \left[X^\top a - b\right] = 0
\end{aligned}$$

as $a^\top X = b$ (from (101)). Therefore,

$$\mathrm{Var}(a^\top Y) = \mathrm{Var}(b^\top \hat{\beta}) + \mathrm{Var}(a^\top Y - b^\top \hat{\beta}) \geq \mathrm{Var}(b^\top \hat{\beta}),$$

which completes the proof. □

In fact, the above idea generalizes to estimating $w^\top \mu$ for any fixed $w \in \mathbb{R}^n$ (recall that $\mu = \mathbb{E}[Y] = X\beta$). A natural linear unbiased estimator of $w^\top \mu$ is $w^\top \hat{Y}$, as

$$\mathbb{E}[w^\top \hat{Y}] = w^\top \mathbb{E}[PY] = w^\top P(X\beta) = w^\top X\beta = w^\top \mu.$$

Exercise 14 (HW3): Consider estimating $w^\top \mu$, where $w \in \mathbb{R}^n$ is a known vector. Show that $w^\top \hat{Y}$ has the smallest variance among all linear unbiased estimators of $w^\top \mu$. As a consequence of the above (re)-derive Theorem 5.3.

Although $w^\top \hat{Y}$ is the "best" linear unbiased estimate (BLUE), in nonlinear estimates can be more precise at times. The following exercise demonstrates this.

Exercise 15 (HW3): Suppose that $Y_i = \beta + \epsilon_i$ for $i = 1, \ldots, n$ where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with common density $f(x) = \frac{e^{-\sqrt{2}|x|/\sigma}}{\sigma\sqrt{2}}$, for $x \in \mathbb{R}$. Shows that the conditions of the Gauss-Markov theorem are satisfied. Further, show that the sample median has roughly (for large n) half the variance of $\hat{\beta}$.

5.3 Normal linear model

In the last subsection we saw that just assuming uncorrelatedness and homoscedasticity of the errors (i.e., (94)) was enough to make $\hat{\beta}$ the BLUE of β. However, for exact inference (i.e., hypothesis testing and confidence intervals) we need to make distributional assumptions on ε. We usually assume that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma^2)$, with $\sigma > 0$ an unknown parameter, so that

$$\varepsilon \sim N(0, \sigma^2 I_n).$$

This leads to the *normal linear model*

$$Y \sim N(X\beta, \sigma^2 I_n). \tag{102}$$

5.3.1 Canonical form

Many results about testing (and estimation) in the general linear model follow easily once the data are expressed in a canonical form. Let $v_1, \ldots, v_n$ be an orthonormal basis for $\mathbb{R}^n$, chosen so that $v_1, \ldots, v_r$ span ω (as defined in (97)). Then, the data vector Y can be expressed as

$$Y = Z_1 v_1 + \ldots + Z_n v_n. \tag{103}$$

Algebraically, $Z = (Z_1, \ldots, Z_n)$ can be found introducing an $n \times n$ orthogonal matrix A with columns $v_1, \ldots, v_n$. Then A is an orthogonal matrix, i.e., $A^\top A = AA^\top = I_n$, and Y and Z are related by

$$Y = AZ \qquad \text{or} \qquad Z = A^\top Y.$$

Since $Y = \mu + \varepsilon$, $Z = A^\top(\mu + \varepsilon) = A^\top \mu + A^\top \varepsilon$. If we define

$$\eta := A^\top \mu \qquad \text{and} \qquad \varepsilon^* := A^\top \varepsilon,$$

then

$$Z = \eta + \varepsilon^* \qquad \Rightarrow \qquad Z \sim N(\eta, \sigma^2 I_n),$$

as $\varepsilon^* \sim N(0, \sigma^2 I_n)$. Next, recalling that $c_1, \ldots, c_p$ are the columns of the design matrix X, we have $\mu = X\beta = \sum_{i=1}^p \beta_i c_i$ and

$$\eta = A^\top \mu = \begin{pmatrix} v_1^\top \\ \vdots \\ v_n^\top \end{pmatrix} \sum_{i=1}^p \beta_i c_i = \begin{pmatrix} \sum_{i=1}^p \beta_i v_1^\top c_i \\ \vdots \\ \sum_{i=1}^p \beta_i v_n^\top c_i \end{pmatrix}.$$

Since $c_1, \ldots, c_p$ all lie in ω, and $v_{r+1}, \ldots, v_n$ all lie in $\omega^\perp$, we have $v_k^\top c_i = 0$ for $k > r$ and $i = 1, \ldots, p$, and thus

$$\eta_{r+1} = \cdots = \eta_n = 0.$$

Now, using $\eta = A^\top \mu$,

$$\mu = A\eta = [v_1 \cdots v_n][\eta_1, \ldots, \eta_r, 0 \ldots 0]^\top = \sum_{i=1}^r \eta_i v_i. \tag{104}$$

This establishes a one-to-one relation between points $\mu \in \omega$ and $(\eta_1, \ldots, \eta_r) \in \mathbb{R}^r$. Since $Z \sim N(\eta, \sigma^2 I_n)$, the variables $Z_1, \ldots, Z_n$ are independent with $Z_i \sim N(\eta_i, \sigma^2)$. The density of Z, taking advantage of the fact that $\eta_{r+1} = \ldots = \eta_n = 0$, is

$$f_Z(z_1, \ldots, z_n) = \frac{1}{(2\pi\sigma^2)^2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^r (z_i - \eta_i)^2 - \frac{1}{2\sigma^2} \sum_{i=r+1}^n z_i^2 \right] \tag{105}$$

$$= \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n z_i^2 + \frac{1}{\sigma^2} \sum_{i=1}^r z_i \eta_i - \frac{1}{2\sigma^2} \sum_{i=1}^r \eta_i^2 - \frac{n}{2} \log(2\pi\sigma^2) \right]. \tag{106}$$

These densities form a full rank $(r+1)$-parameter exponential family with complete sufficient statistic

$$\left(Z_1, \ldots, Z_r, \sum_{i=r+1}^n Z_i^2 \right).$$

Exploiting the canonical form, many parameters are easy to estimate. Because $\mathbb{E}[Z_i] = \eta_i, i = 1, \ldots, r$, Z_i is the UMVU estimator of η_i, for $i = 1, \ldots, r$.

Exercise 16 (HW3): Find the UMVU estimator of $w^\top \mu$, where $w \in \mathbb{R}^n$ is a fixed vector. In particular, find the UMVU estimator of $b^\top \beta$, where $b \in \mathbb{R}^p$ is a fixed vector (assuming X has full column rank).

5.3.2 Estimating σ^2

From the above discussion, $Z_{r+1}, \ldots, Z_n$ are i.i.d. from $N(0, \sigma^2)$. Thus $\mathbb{E}[Z_i^2] = \sigma^2$, for $i = r+1, \ldots, n$, and the average of these variables,

$$S^2 := \frac{1}{n-r} \sum_{i=r+1}^n Z_i^2 \tag{107}$$

is an unbiased estimator of σ^2. But S^2 is a function of the complete sufficient statistic $(Z_1, \ldots, Z_r, \sum_{i=1}^n Z_i^2)$, and so S^2 is the UMVU estimator of σ^2. The estimator S^2 can be computed from the length of the residual vector $e = Y - \hat{Y}$ in (100). To see this, first observe that, from (103)

$$\hat{Y} = PY = PAZ = [Pv_1 \vdots \cdots \vdots Pv_n]Z = [v_1 \vdots \cdots \vdots v_r \vdots 0 \vdots \cdots \vdots 0]Z = \sum_{i=1}^r Z_i v_i \tag{108}$$

and from (104), we have

$$\|e\|^2 = e^\top e = \left(\sum_{i=r+1}^n Z_i v_i^\top\right)\left(\sum_{i=r+1}^n Z_i v_i\right) = \sum_{i=r+1}^n \sum_{i=r+1}^n Z_i Z_j v_i^\top v_j = \sum_{i=r+1}^n Z_i^2 \tag{109}$$

as $v_1, \ldots, v_n$ is an orthonormal basis, $v_i^\top v_j = \delta_{ij}$ (where δ_{ij} equals 0if $i \neq j$ and equals 1 when $i = j$).

As $\hat{\mu} \equiv \hat{Y}$ is a function of $Z_1, \ldots, Z_r$, and the residual e is a function of $Z_{r+1}, \ldots, Z_n$, S^2 and $\hat{\mu}$ are independent. Also, using the above, and the definition of the chi-square distribution,

$$\frac{(n-r)S^2}{\sigma^2} = \sum_{i=r+1}^n (Z_i/\sigma)^2 \sim \chi^2_{n-r},$$

since $Z_i/\sigma \sim N(0,1)$, for $i = r+1, \ldots, n$.

The distribution theory just presented can be used to set confidence intervals for linear estimators. If a is a constant vector in $\mathbb{R}^n$, then the variance of (unbiased) LSE $a^\top \hat{\mu}$ of $a^\top \mu$ is $\sigma^2 \|Pa\|^2$, which is naturally estimated as $\hat{\sigma}_{a^\top \hat{\mu}} := S\|Pa\|$.

Exercise 17 (HW3): Show that in the general linear model with $Y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \omega$, and $\sigma^2 > 0$,

$$\left(a^\top \hat{\mu} - \hat{\sigma}_{a^\top \hat{\mu}} t_{\alpha/2, n-r}, a^\top \hat{\mu} + \hat{\sigma}_{a^\top \hat{\mu}} t_{\alpha/2, n-r}\right)$$

is a $1-\alpha$ confidence interval for $a^\top \mu$. In particular, when X has full column rank, find a $1-\alpha$ confidence interval for β_i, for $i = 1, \ldots, p$.

5.3.3 Noncentral F and chi-square distributions

Distribution theory for testing in the general linear model relies on noncentral F and chi-square distributions.

Definition 5.4 (Noncentral chi-square distribution)**.** If $Z_1, \ldots, Z_k$ are independent and $\delta \geq 0$ with

$$Z_1 \sim N(\delta, 1) \qquad \text{and} \qquad Z_j \sim N(0,1), \;\; j = 2, \ldots, k,$$

then $W := \sum_{i=1}^k Z_i^2$ has the *noncentral chi-square* distribution with noncentrality parameter δ^2 and k degrees of freedom, denoted by

$$W \sim \chi^2_k(\delta^2).$$

Lemma 5.5. If $Z \sim N(\gamma, I_k)$, then $Z^\top Z \sim \chi_k^2(\|\gamma\|^2)$.

Proof. Let B be an $k \times k$ orthogonal matrix where the first row is $\gamma^\top/\|\gamma\|$, so that $\tilde{\gamma} := B\gamma = (\|\gamma\|, 0, \ldots, 0) \in \mathbb{R}^k$. Then, $\tilde{Z} := BZ \sim N(\tilde{\gamma}, I_k)$. As $\tilde{Z}^\top \tilde{Z} = \sum_{i=1}^k \tilde{Z}_i^2 \sim \chi_k^2(\|\gamma\|^2)$, the lemma follows as $\tilde{Z}^\top \tilde{Z} = Z^\top B^\top B Z = Z^\top Z$. □

The next lemma shows that certain quadratic forms for multivariate normal vectors have noncentral chi-square distributions.

Lemma 5.6. If Σ is a $k \times k$ positive definite matrix and if $Z \sim N(\gamma, \Sigma)$, then

$$Z^\top \Sigma^{-1} Z \sim \chi_k^2(\gamma^\top \Sigma^{-1} \gamma).$$

Proof. Let $A := \Sigma^{-1/2}$, the symmetric square root of Σ^{-1} [This can be found by writing $\Sigma = VDV^\top$ where V is an orthogonal matrix (so that $VV^\top = I$) and D is diagonal, and defining $\Sigma^{1/2} = VD^{1/2}V^\top$, where $D^{1/2}$ is diagonal with entries the square roots of the diagonal entries of D. Then $\Sigma^{1/2}$ is symmetric and $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.].

Then $AZ \sim N(A\gamma, I_k)$, and so

$$Z^\top \Sigma^{-1} Z = (AZ)^\top (AZ) \sim \chi_k^2(\|A\gamma\|^2).$$

The lemma now follows as $\|A\gamma\|^2 = (A\gamma)^\top (A\gamma) = \gamma^\top A A \gamma = \gamma^\top \Sigma^{-1} \gamma$. □

Definition 5.7 (Noncentral F-distribution)**.** If V and W are independent variables with $V \sim \chi_k^2(\delta^2)$ and $W \sim \chi_m^2$, then

$$\frac{V/k}{W/m} \sim F_{k,m}(\delta^2),$$

the noncentral F-distribution with degrees of freedom k and m and noncentrality parameter δ^2. When $\delta^2 = 0$ this distribution is simply called the F-distribution, $F_{k,m}$.

5.4 Testing in the general linear model

In the general linear model, $Y \sim N(\mu, \sigma^2 I_n)$ with the mean $\mu = X\beta = \sum_{i=1}^p \beta_i c_i$ in a linear subspace ω with dimension r. In this subsection we consider testing

$$H_0 : \mu \in \omega_0 \qquad \text{versus} \qquad H_1 : \mu \in \omega \setminus \omega_0$$

with ω_0 being a q-dimensional linear subspace of ω, with $0 \le q < r$. Null hypotheses of this form arise when β satisfies linear constraints. For instance we might want to test $H_0 : \beta_1 = \beta_2$ or $H_0 : \beta_1 = 0$. Note that $H_0 : \beta_1 = \beta_2$ is equivalent to $\mu \in \text{span}\{c_1 + c_2, c_3, \ldots, c_p\} =: \omega_0 \subset \omega$ and $H_0 : \beta_1 = 0$ is equivalent to $\mu \in \text{span}\{c_2, c_3, \ldots, c_p\}$.

Let $\hat{\mu}$ and $\hat{\mu}_0$ denote LSEs for μ under the full model and under H_0. Specifically,

$$\hat{\mu} \equiv \hat{Y} = PY \qquad \text{and} \qquad \hat{\mu}_0 = P_0 Y,$$

where P and P_0 are the projection matrices onto ω and ω_0. The test statistic of interest is based on $\|Y - \hat{\mu}\|$, the distance between Y and ω, and $\|Y - \hat{\mu}_0\|$, the distance between Y and ω_0. Because $\omega_0 \subset \omega$, the former distance must be smaller, but if the distances are comparable, then at least qualitatively H_0 may seem adequate. The test statistic is

$$T = \frac{n-r}{r-q} \frac{\|Y - \hat{\mu}_0\|^2 - \|Y - \hat{\mu}\|^2}{\|Y - \hat{\mu}\|^2}, \tag{110}$$

and H_0 will be rejected if T exceeds a suitable constant. Noting that $Y - \hat{\mu} \in \omega^\perp$ and $\hat{\mu} - \hat{\mu}_0 \in \omega$, the vectors $Y - \hat{\mu}$ and $\hat{\mu} - \hat{\mu}_0$ are orthogonal, and the squared length of their sum, by the Pythagorean theorem, is

$$\|Y - \hat{\mu}_0\|^2 = \|Y - \hat{\mu}\|^2 + \|\hat{\mu} - \hat{\mu}_0\|^2.$$

Using this, the formula for T can be rewritten as

$$T = \frac{n-r}{r-q} \frac{\|\hat{\mu} - \hat{\mu}_0\|^2}{\|Y - \hat{\mu}\|^2} = \frac{\|\hat{\mu} - \hat{\mu}_0\|^2}{(r-q)S^2},$$

where we have used (107) and (109)[73]. For level and power calculations we need the distribution of T which is given in the next result.

Theorem 5.8. Under the normal linear model,

$$T \sim F_{r-q,n-r}(\delta^2),$$

where

$$\delta^2 := \|\mu - P_0\mu\|^2/\sigma^2. \tag{111}$$

Proof. Write $Y = \sum_{i=1}^n Z_i v_i$, where $v_1, \ldots, v_n$ is an orthonormal basis chosen so that $v_1, \ldots, v_q$ span ω_0 and $v_1, \ldots, v_r$ span ω. Then, as in (108),

$$\hat{\mu}_0 = \sum_{i=1}^q Z_i v_i \qquad \text{and} \qquad \hat{\mu} = \sum_{i=1}^r Z_i v_i.$$

We know that $Z \sim N(\eta, \sigma^2 I_n)$ with $\eta_{r+1} = \cdots = \eta_n = 0$. Since $v_i^\top v_j$ is zero when $i \neq j$ and one when $i = j$,

$$\|Y - \hat{\mu}\|^2 = \Big\| \sum_{i=r+1}^n Z_i v_i \Big\|^2 = \sum_{i=r+1}^n Z_i^2.$$

[73]Exercise 18 (HW3): Show that this test statistic is equivalent to the generalized likelihood ratio test statistic. Show that when $r - q = 1$ the test is UMPU.

Similarly,

$$\|Y - \hat{\mu}_0\|^2 = \Big\| \sum_{i=q+1}^{n} Z_i v_i \Big\|^2 = \sum_{i=q+1}^{n} Z_i^2,$$

and so,

$$T = \frac{\frac{1}{r-q}\sum_{i=q+1}^{r}(Z_i/\sigma)^2}{\frac{1}{n-r}\sum_{i=r+1}^{n}(Z_i/\sigma)^2}.$$

The variables Z_i are independent, and so the numerator and denominator in this formula for T are independent. Because $Z_i/\sigma \sim N(\eta_i/\sigma, 1)$, by Lemma 5.5, $\sum_{i=q+1}^{r}(Z_i/\sigma)^2 \sim \chi^2_{r-q}(\delta^2)$ where

$$\delta^2 := \sum_{i=q+1}^{r} \eta_i^2/\sigma^2. \tag{112}$$

Also, since $\eta_i = 0$ for $i = r+1, \ldots, n$, $Z_i/\sigma \sim N(0,1), i = r+1, \ldots, n$, and so $\sum_{i=r+1}^{n}(Z_i/\sigma)^2 \sim \chi^2_{n-r}$. So by Definition 5.7 for the noncentral F-distribution, $T \sim F_{r-q,n-r}(\delta^2)$. To complete the proof we must show that (111) and (112) agree, or that $\sum_{i=q+1}^{r}\eta_i^2 = \|\mu - P_0\mu\|^2$. Since $\mu = \mathbb{E}[\hat{Y}] = \sum_{i=1}^{r}\eta_i v_i$, and $P_0\mu = \mathbb{E}[P_0 Y] = \mathbb{E}[\hat{\mu}_0] = \sum_{i=1}^{q}\eta_i v_i$, we have

$$\mu - P_0\mu = \sum_{i=q+1}^{r}\eta_i v_i.$$

The result now follows as $\|\mu - P_0\mu\|^2 = \sum_{i=q+1}^{r}\eta_i^2$. □

5.5 Exercises

19. Consider a general linear model $Y \sim N(\mu, \sigma^2 I_n), \mu \in \omega, \sigma^2 > 0$ with $\dim(\omega) = r$. Define $\psi = A\mu \in \mathbb{R}^q$ where $q < r$, and assume $A = AP$ where P is the projection onto ω, so that $\hat{\psi} := A\hat{\mu} = AY$, and that A has full rank q. The F-test derived in Section 5.4 allows us to test $\psi = 0$ versus $\psi \neq 0$. Modify that theory and give a level α test of $H_0 : \psi = \psi_0$ versus $H_1 : \psi \neq \psi_0$ with ψ_0 some constant vector in $\mathbb{R}^q$. [Hint: Let $Y^* := Y - \mu_0$ with $\mu_0 \in \omega$ and $A\mu_0 = \psi_0$. Then the null hypothesis will be $H_0 : A\mu^* = 0$ where $\mu^* = \mu - \mu_0$.]

6 M-estimation (or empirical risk minimization)

Suppose that we are interested in a parameter (or "functional") θ attached to the distribution of the observations $X_1, \ldots, X_n$ i.i.d. P taking values in some space $\mathcal{X}$ (e.g., a metric space). A popular method (in statistics and machine learning) for finding an estimator $\hat{\theta}_n \equiv \hat{\theta}(X_1, \ldots, X_n)$ is to *maximize a criterion function* of the type

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \mathbb{M}_n(\theta) \qquad \text{where} \quad \mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta] = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i). \tag{113}$$

where $\mathbb{P}_n$ denotes the empirical measure[74]. Here Θ denotes the parameter space and, for each $\theta \in \Theta$, m_θ denotes the a real-valued function on $\mathcal{X}$ (usually $-m_\theta$ is thought of as a "loss" function). Such a quantity $\hat{\theta}_n$ is called an M*-estimator* as it is obtained by maximizing (or minimizing) an objective function. The map

$$\theta \mapsto \mathbb{P}_n[-m_\theta] = \frac{1}{n}\sum_{i=1}^n [-m_\theta(X_i)]$$

can be thought of as the "empirical risk" and $\hat{\theta}_n$ denotes the *empirical risk minimizer* over $\theta \in \Theta$. Here are some examples:

1. **Maximum likelihood estimators**: These correspond to $m_\theta(x) = \log p_\theta(x)$.

2. **Location estimators**:

   (a) **Median**: corresponds to $m_\theta(x) = |x - \theta|$.

   (b) **Mode**: may correspond to $m_\theta(x) = \mathbf{1}\{|x - \theta| \leq 1\}$.

3. **Nonparametric maximum likelihood**: Suppose $X_1, \ldots, X_n$ are i.i.d. from a density $\theta(\cdot)$ on $[0, \infty)$ that is known to be nonincreasing. Then take Θ to be the collection of all nonincreasing densities on $[0, \infty)$ and $m_\theta(x) = \log \theta(x)$. The corresponding M-estimator is the MLE over all non-increasing densities. It can be shown that $\hat{\theta}_n$ exists and is unique; $\hat{\theta}_n$ is usually known as the Grenander estimator.

4. **Regression estimators**: Let $\{X_i = (Z_i, Y_i)\}_{i=1}^n$ denote i.i.d. from a regression model and let

$$m_\theta(x) = m_\theta(z, y) := -(y - \theta(z))^2,$$

[74] Suppose now that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. Then the *empirical measure* $\mathbb{P}_n$ is defined by $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x denotes the Dirac measure at x. For each $n \geq 1$, $\mathbb{P}_n$ denotes the random discrete probability measure which puts mass $1/n$ at each of the n data points $X_1, \ldots, X_n$. Thus, for any Borel set $A \subset \mathcal{X}$, $\mathbb{P}_n(A) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_A(X_i) = \frac{\#\{i \leq n : X_i \in A\}}{n}$. For a real-valued function f on $\mathcal{X}$, we write

$$\mathbb{P}_n(f) := \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

for a class $\theta \in \Theta$ of real-valued functions from the domain of Z[75]. This gives the usual least squares estimator over the class Θ. The choice $m_\theta(z, y) = -|y - \theta(z)|$ gives the least absolute deviation estimator over Θ.

In these problems, the (true value of the) parameter of interest is

$$\theta_0 := \arg\max_{\theta \in \Theta} M(\theta) \qquad \text{where} \quad M(\theta) := P[m_\theta] = \mathbb{E}_{X \sim P}[m_\theta(X)]. \tag{114}$$

Perhaps the simplest general way to see that $\hat{\theta}_n$ (in (113)) is a reasonable estimator of θ_0 is to reason as follows. By the law of large numbers, we can approximate the 'risk' for a fixed parameter θ by the empirical risk which depends only on the data, i.e.,

$$\mathbb{P}_n[m_\theta] \approx P[m_\theta], \qquad \text{i.e.,} \;\; \mathbb{M}_n(\theta) \approx M(\theta).$$

However, the pointwise convergence $\mathbb{M}_n(\theta) \xrightarrow{p} M(\theta)$, for all $\theta \in \Theta$, is too weak to ensure the convergence of their maximizers (see e.g., Figure 2).

If $\mathbb{M}_n(\theta)$ and $M_n(\theta)$ are *uniformly* close, then maybe their argmax's $\hat{\theta}_n$ and θ_0 are close. This naturally leads to the investigation of quantities such as the uniform deviation

$$\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| = \sup_{\theta \in \Theta} |(\mathbb{P}_n - P)[m_\theta]|. \tag{115}$$

The study of such deviations, uniformly over finite/infinite dimensional parameter spaces Θ, is the main topic of the subject *empirical process theory*, a full study of which is beyond the scope of the current course. However, in this chapter we will develop some basic tools and work under (slightly) restrictive assumptions (on the functions $\mathbb{M}_n, M$) to try to provide some intuition for the main ideas. Here is the motivation example.

Example 6.1 (Sample median)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. observations from a continuous density f with median θ_0 and variance 1 (such that $f(\theta_0) > 0$). Let $\hat{\theta}_n$ be a sample median based on $X_1, \ldots, X_n$ that is defined as any minimizer of

$$\mathbb{M}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta| = \mathbb{P}_n[|\cdot - \theta|] \tag{116}$$

over $\theta \in \mathbb{R}$. Here are a few questions about $\hat{\theta}_n$ that we would like to answer in this chapter: (i) Is $\hat{\theta}_n$ consistent for estimating θ_0? (ii) What is the asymptotic distribution of $\hat{\theta}_n$? (iii) Can we develop a theory to study such M-estimators where the objective function $\mathbb{M}_n(\cdot)$ is not differentiable everywhere?

Remark 6.1 (Z-estimators)**.** Often the maximizing value of $\mathbb{M}_n$ is sought by setting a derivative (or the set of partial derivatives in the multidimensional case) equal to zero, i.e.,

$$\Psi_n(\theta) := \mathbb{P}_n[\psi_\theta] = 0, \tag{117}$$

[75] In the simplest setting we could parametrize $\theta(\cdot)$ as $\theta_\beta(z) := \beta^\top z$, for $\beta \in \mathbb{R}^d$, in which case $\Theta = \{\theta_\beta(\cdot) : \beta \in \mathbb{R}^d\}$.

where $\psi_\theta(x) := \nabla_\theta m_\theta(x)$. These are estimating equations (and need not correspond to a maximization problem) and the corresponding estimator is called a Z-estimator (as it solves the 'zero' of an equation). We can also develop a theory of Z-estimators, although we will be more concerned with M-estimators in this chapter.

6.1 Consistency of M-estimators

If $\hat{\theta}_n$ (as defined in (113)) is used to estimate the parameter θ_0 (in (139)), then it is certainly desirable that the sequence $\hat{\theta}_n$ converges in probability to θ_0, i.e., $\hat{\theta}_n$ be consistent.

Here we assume that Θ is a metric space with the metric $d(\cdot,\cdot)$. We want to show that

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0 \tag{118}$$

As we have seen in (115), to study the consistency of $\hat{\theta}_n$ we have to deal will the uniform convergence of the empirical measure for the class of functions $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$, or a subset thereof.

6.1.1 Glivenko-Cantelli (GC) classes of functions

Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables taking values in the space $\mathcal{X}$ with probability measure P. Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. The main object of study in this section is to obtain probability estimates of the random quantity

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf|.$$

The law of large numbers says that $\mathbb{P}_n f \to Pf$ almost surely, as soon as the expectation Pf exists. A class of functions is called *Glivenko-Cantelli* if this convergence is uniform in the functions belonging to the class.

Definition 6.2. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $P|f| < \infty$ for every $f \in \mathcal{F}$ is called *Glivenko-Cantelli*[76] (GC) if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to 0, \qquad \text{in probability.}$$

However, uniform convergence of $\mathbb{M}_n$ to M is not enough to guarantee (118). We will further need to assume that θ_0 is a *well separated* maximizer[77] of $M(\cdot)$ (see Figure 2), i.e.,

[76] As the Glivenko-Cantelli property depends on the distribution P of the observations, we also say, more precisely, P-Glivenko-Cantelli.

[77] This is an identifiability condition which says that approximately maximizing $M(\theta)$ unambiguously specifies θ_0. This condition holds if $M(\cdot)$ has a unique maximizer, Θ is compact, and $M(\cdot)$ is continuous.

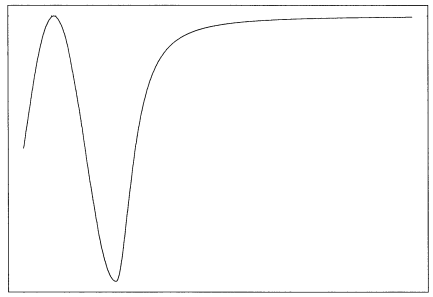


Figure 2: An example of a function whose point of maximum is not well separated.

for every $\delta > 0$, $M(\theta_0) > \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta)$. Fix $\delta > 0$ and let

$$\psi(\delta) := M(\theta_0) - \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} \mathbb{M}(\theta) > 0.$$

Observe that,

$$\begin{aligned} \{d(\hat{\theta}_n, \theta_0) \geq \delta\} \quad &\Rightarrow \quad M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta) \\ &\Leftrightarrow \quad M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\ &\Rightarrow \quad M(\hat{\theta}_n) - M(\theta_0) + (\mathbb{M}_n(\theta_0) - \mathbb{M}_n(\hat{\theta}_n)) \leq -\psi(\delta) \\ &\Rightarrow \quad 2 \sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta). \end{aligned} \tag{119}$$

Therefore,

$$\mathbb{P}\left(d(\hat{\theta}_n, \theta_0) \geq \delta\right) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta)/2\right) \to 0$$

if $\mathcal{F}$ is P-Glivenko Cantelli.

However, as we will see later, showing that $\mathcal{F}$ is GC is not always easy when Θ is not a compact set[78]. When Θ is not a compact set, the following lemma may be used to reduce the problem to showing GC property only over a compact subset of Θ, by using a *concavity* assumption.

Lemma 6.3 (When m_θ is concave)**.** Suppose that Θ is a convex subset of $\mathbb{R}^k$, and that $\theta \mapsto m_\theta(x)$, is concave[79], for all $x \in \mathcal{X}$. We assume that θ_0 (as defined in (139)) exists and is unique. Suppose that for some $\epsilon > 0$,

$$\sup_{\theta \in \Theta : \|\theta - \theta_0\| \leq \epsilon} |\mathbb{M}_n(\theta) - M(\theta)| \xrightarrow{p} 0. \tag{120}$$

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

[78] We can get around the problem by restricting to a compact set where most of the mass of P lies, and showing that this does not affect the asymptotics. However, this needs techniques that are problem specific.

[79] A function $g : \Theta \to \mathbb{R}$ is concave if for any $u, v \in \Theta$ and $\alpha \in [0, 1]$,

$$g(\alpha u + (1 - \alpha)v) \geq \alpha g(u) + (1 - \alpha) g(v).$$

Proof. Let us define

$$\alpha := \frac{\epsilon}{\epsilon + \|\hat{\theta}_n - \theta_0\|} \in (0,1] \qquad \text{and} \qquad \tilde{\theta}_n := \alpha\hat{\theta}_n + (1-\alpha)\theta_0. \tag{121}$$

The idea of the proof is to compare $\mathbb{M}_n(\tilde{\theta}_n)$ (instead of $\mathbb{M}_n(\hat{\theta}_n)$) with $\mathbb{M}_n(\theta_0)$.

Note that by (121), $\tilde{\theta}_n - \theta_0 = \alpha(\hat{\theta}_n - \theta_0)$, and thus, for any $\delta > 0$, using arguments similar to in (119), we have

$$\begin{aligned}
\{\|\tilde{\theta}_n - \theta_0\| \geq \delta\} \quad &\Rightarrow \quad M(\tilde{\theta}_n) \leq \sup_{\theta\in\Theta:\|\theta-\theta_0\|\geq\delta} M(\theta) \\
&\Leftrightarrow \quad M(\tilde{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\
&\Rightarrow \quad M(\tilde{\theta}_n) - M(\theta_0) + (\mathbb{M}_n(\theta_0) - \mathbb{M}_n(\tilde{\theta}_n)) \leq -\psi(\delta) \\
&\Rightarrow \quad 2\sup_{\theta\in\Theta:\|\theta-\theta_0\|\leq\epsilon} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta)
\end{aligned}$$

where we have used the facts: (i) $\mathbb{M}_n(\theta_0) - \mathbb{M}_n(\tilde{\theta}_n) \leq 0$ and (ii) $\|\tilde{\theta}_n - \theta_0\| \leq \epsilon$. Note that (i) follows from the observation that $\mathbb{M}_n(\cdot)$ is a concave function (as a sum of concave functions is also concave), and thus,

$$\mathbb{M}_n(\tilde{\theta}_n) = \mathbb{M}_n\big(\alpha\hat{\theta}_n + (1-\alpha)\theta_0\big) \geq \alpha\mathbb{M}_n(\hat{\theta}_n) + (1-\alpha)\mathbb{M}_n(\theta_0)$$

which implies that

$$\mathbb{M}_n(\tilde{\theta}_n) - \mathbb{M}_n(\theta_0) \geq \alpha\left(\mathbb{M}_n(\hat{\theta}_n) - \mathbb{M}_n(\theta_0)\right) \geq 0,$$

as $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\cdot)$.

To show (ii), observe that, if $\|\hat{\theta}_n - \theta_0\| \neq 0$, then

$$\|\tilde{\theta}_n - \theta_0\| = \alpha\|\hat{\theta}_n - \theta_0\| \leq \frac{\epsilon}{\|\hat{\theta}_n - \theta_0\|}\|\hat{\theta}_n - \theta_0\| \leq \epsilon.$$

Therefore,

$$\mathbb{P}\left(\|\tilde{\theta}_n - \theta_0\| \geq \delta\right) \quad \leq \quad \mathbb{P}\left(\sup_{\theta\in\Theta:\|\theta-\theta_0\|\leq\epsilon} |\mathbb{M}_n(\theta) - M(\theta)| \geq \frac{\psi(\delta)}{2}\right) \to 0$$

by the assumption in the lemma. Thus, $\tilde{\theta}_n \xrightarrow{p} \theta_0$. But then also,

$$\|\hat{\theta}_n - \theta_0\| = \frac{\epsilon\|\tilde{\theta}_n - \theta_0\|}{\epsilon - \|\tilde{\theta}_n - \theta_0\|} \xrightarrow{p} 0.$$

□

6.1.2 Bracketing numbers

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$. We are mostly thinking of $L_r(Q)$-spaces for probability measures Q, i.e., the $L_r(Q)$-norm is $\|f\|_{Q,r} = \left(\int |f|^r dQ\right)^{1/r}$.

Recall that a *cover* of a set $\mathcal{F}$ is a collection of sets whose union includes $\mathcal{F}$ as a subset. Formally, if $C = \{U_\alpha : \alpha \in A\}$ is an indexed family of sets U_α, then C is a cover of $\mathcal{F}$ if $\mathcal{F} \subseteq \bigcup_{\alpha \in A} U_\alpha$.

Definition 6.4 (ε-bracket)**.** Given two functions $l(\cdot)$ and $u(\cdot)$, the *bracket* $[l, u]$ is the set of all functions $f \in \mathcal{F}$ with $l(x) \le f(x) \le u(x)$, for all $x \in \mathcal{X}$. An *ε-bracket* is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$.

Definition 6.5 (Bracketing numbers)**.** The *bracketing number* $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.

In the definition of the bracketing number, the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.

Example 6.6. (Distribution function). Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with c.d.f. F. When $\mathcal{F}$ is equal to the collection of all indicator functions of the form $f_t(\cdot) = \mathbf{1}_{(-\infty,t]}(\cdot)$, with t ranging over $\mathbb{R}$, then

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| = \sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)|$$

where $\mathbb{F}_n$ is the empirical distribution function.

Consider brackets of the form $[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i)}]$ for grid points $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with the property $F(t_i-) - F(t_{i-1}) < \varepsilon$ for each i (points at which F jumps more than ε are points of the partition). These brackets have $L_1(P)$-size ε. Their total number k can be chosen smaller than $3/\varepsilon$ (for ε small)[80].

6.1.3 GC by bracketing

Theorem 6.7. Let $\mathcal{F}$ be a class of measurable functions such that $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is Glivenko-Cantelli.

Proof. Fix $\epsilon > 0$. Choose finitely many ϵ-brackets $[l_i, u_i]$ whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) < \epsilon$, for every i. Then, for every $f \in \mathcal{F}$, there is a bracket such that

$$(\mathbb{P}_n - P)f \le (\mathbb{P}_n - P)u_i + P(u_i - f) \le (\mathbb{P}_n - P)u_i + \epsilon.$$

[80] Exercise 1 (HW4): Show this.

Consequently,

$$\sup_{f\in\mathcal{F}}(\mathbb{P}_n - P)f \le \max_i(\mathbb{P}_n - P)u_i + \epsilon.$$

The right side converges almost surely (a.s.) to ϵ by the strong law of large numbers for real variables. Thus, $\limsup_{n\to\infty}\sup_{f\in\mathcal{F}}(\mathbb{P}_n - P)f \le \epsilon$ a.s. A similar argument also yields

$$\begin{aligned} &(\mathbb{P}_n - P)f \ge (\mathbb{P}_n - P)l_i + P(l_i - f) \ge (\mathbb{P}_n - P)l_i - \epsilon \\ \Rightarrow\quad &\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f \ge \min_i(\mathbb{P}_n - P)l_i - \epsilon. \end{aligned}$$

Thus, by SLLN we can show that $\limsup_{n\to\infty}[-\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f] \le \epsilon$ a.s. As,

$$\sup_{f\in\mathcal{F}}|(\mathbb{P}_n - P)f| = \max\left\{\sup_{f\in\mathcal{F}}(\mathbb{P}_n - P)f, -\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f\right\},$$

we see that $\limsup\|\mathbb{P}_n - P\|_{\mathcal{F}} \le \epsilon$ a.s., for every $\epsilon > 0$. This is true for every ϵ and hence the limit superior is zero, yielding the desired result. □

Example 6.8. (Distribution function). The previous proof generalizes a well-known proof of the classical GC theorem for the empirical distribution function on the real line. Indeed, the set of indicator functions of cells $(-\infty, c]$ possesses finite bracketing numbers for any underlying distribution; simply use the brackets $\left[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i)}\right]$ for a grid of points $-\infty = t_0 < t_1 < \ldots < t_k = +\infty$ with the property $P(t_{i-1}, t_i) < \epsilon$ for each $i = 1, \ldots, k$.

Example 6.9 (Pointwise compact class)**.** Let $\mathcal{F} = \{m_\theta(\cdot) : \theta \in \Theta\}$ be a collection of measurable functions with integrable *envelope*[81] function F indexed by a compact metric space Θ such that the map $\theta \mapsto m_\theta(x)$ is continuous for every x. Then the bracketing numbers of $\mathcal{F}$ are finite and hence $\mathcal{F}$ is Glivenko-Cantelli.

We can construct the brackets in the obvious way in the form $[m_B, m^B]$, where $B \subset \Theta$ is an open ball and m_B and m^B are the infimum and supremum of m_θ for $\theta \in B$, respectively (i.e., $m_B(x) = \inf_{\theta\in B} m_\theta(x)$, and $m^B(x) = \sup_{\theta\in B} m_\theta(x)$).

Given a sequence of balls B_k with common center a given $\theta \in \Theta$ and radii decreasing to 0, we have $m^{B_k} - m_{B_k} \downarrow m_\theta - m_\theta = 0$ by the continuity, pointwise in x, and hence also in L_1 by the dominated convergence theorem and the integrability of the envelope. Thus, given $\epsilon > 0$, for every $\theta \in \Theta$ there exists a ball B around θ such that the bracket $[m_B, m^B]$ has size at most ϵ. By the compactness of Θ[82], the collection of balls constructed in this way has a finite subcover. The corresponding brackets cover $\mathcal{F}$. This construction shows that the bracketing numbers are finite, but it gives no control on their sizes.

An example of such a class would be the log-likelihood function of a parametric model $\{p_\theta(x) : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is assumed to be *compact* and $\theta \mapsto p_\theta(x)$ is assumed to be continuous in θ, for every x.

[81] An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \le F(x)$, $\forall\, x \in \mathcal{X}$, $f \in \mathcal{F}$.

[82] A topological space Θ is said to be compact if every open cover has a finite subcover.

Example 6.10 (Back to Example 6.1)**.** We can now use Lemma 6.3, coupled with Example 6.9 and Theorem 6.7, to show that the sample median $\hat{\theta}_n$, defined as the minimizer of (116), converges in probability to the population median $\theta_0 := \operatorname{argmin}_{\theta\in\mathbb{R}} \mathbb{E}[|X-\theta|]$ (which is assumed to be unique).

To use Lemma 6.3 to show $\hat{\theta}_n \xrightarrow{p} \theta_0$ we only need to verify (120). Here, we can take $m_\theta(x) := -(|x-\theta| - |x|)$, for $\theta \in [\theta_0 - 1, \theta_0 + 1]$. Note that $\hat{\theta}_n = \operatorname{argmax}_{\theta\in\mathbb{R}} \mathbb{P}_n[m_\theta]$ and $\theta_0 = \operatorname{argmax}_{\theta\in\mathbb{R}} P[m_\theta]$. As $\mathcal{F} = \{m_\theta : |\theta - \theta_0| \leq 1\}$ is a collection of functions with integrable envelope $G(x) := \sup_{|\theta-\theta_0|\leq 1} |m_\theta(x)| \leq \theta_0 + 1$ such that $\theta \mapsto m_\theta(x)$ is continuous for every x, by Example 6.9, $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$, and thus Theorem 6.7 yields (120).

6.2 Asymptotic normality of Z-estimators

Suppose that $\hat{\theta}_n$ is consistent for θ_0. The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta_0$ converges to zero. The answer depends on the specific situation, but for estimators based on n replications of an experiment the order is often $n^{-1/2}$. Then multiplication with the inverse of this rate creates a proper balance, and the sequence $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution, most often to a normal distribution. This is interesting from a theoretical point of view. It also makes it possible to obtain approximate confidence sets.

In this section we derive the asymptotic normality of Z-estimators. We can use a characterization of M-estimators either by maximization or by solving estimating equations (as in Remark 6.1). Consider the second possibility first. For $\theta \in \Theta$, let

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i) = \mathbb{P}_n[\psi_\theta], \qquad \text{and} \qquad \Psi(\theta) := P[\psi_\theta],$$

where $\psi_\theta(x) := \nabla_\theta m_\theta(x)$.

6.2.1 Heuristic proof of asymptotic normality of Z-estimators

Assume that $\hat{\theta}_n$ is a zero of Ψ_n and converges in probability to θ_0, a zero of Ψ. As $\hat{\theta}_n \xrightarrow{p} \theta_0$, it makes sense to expand $\Psi_n(\hat{\theta}_n)$ in a Taylor series around θ_0. Assume for simplicity that $\Theta \subset \mathbb{R}$, i.e., θ is one-dimensional. Then,

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\ddot{\Psi}_n(\tilde{\theta}_n),$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and θ_0. This can be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)}.$$

If $P[\psi_{\theta_0}^2]$ is finite, then the numerator

$$-\sqrt{n}\Psi_n(\theta_0) = -n^{-1/2}\sum_{i=1}^n \psi_{\theta_0}(X_i) \ \xrightarrow{d}\ N(0, P[\psi_{\theta_0}^2])$$

by the CLT (as $P[\psi_{\theta_0}] = \Psi(\theta_0) = 0$). Next consider the denominator. The first term $\dot{\Psi}_n(\theta_0)$ is an average and can be analyzed by the law of large numbers: $\dot{\Psi}_n(\theta_0) \xrightarrow{p} P[\dot{\psi}_{\theta_0}]$, provided the expectation exists. The second term in the denominator is a product of $\hat{\theta}_n - \theta_0 = o_p(1)$ and $\ddot{\Psi}_n(\tilde{\theta}_n)$ and converges in probability to zero under the reasonable condition that $\ddot{\Psi}_n(\tilde{\theta}_n)$ (which is also an average) is $O_p(1)$. Together with Slutsky's lemma, these observations yield

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, P[\psi_{\theta_0}^2](P[\dot{\psi}_{\theta_0}])^{-2}\right). \tag{122}$$

The preceding derivation can be made rigorous by imposing appropriate conditions; one challenge is to show that $\ddot{\Psi}_n(\tilde{\theta}_n) = O_p(1)$.

The derivation can be extended to higher-dimensional parameters. For a k-dimensional parameter, we use k estimating equations. Then the criterion functions are maps $\Psi_n : \mathbb{R}^k \to \mathbb{R}^k$ and the derivatives $\dot{\Psi}_n(\theta_0)$ are $k \times k$-matrices that converge to the $k \times k$ matrix $P[\dot{\psi}_{\theta_0}]$ with entries $P[(\partial/\partial\theta_j)\psi_{\theta_0,i}]$. The final statement becomes

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_k\left(0, (P[\dot{\psi}_{\theta_0}])^{-1}P[\psi_{\theta_0}\psi_{\theta_0}^\top](P[\dot{\psi}_{\theta_0}])^{-1}\right). \tag{123}$$

See van der Vaart [15, Theorem 5.41] for a formal statement and a proof of the above result.

In the preceding derivation it is implicitly understood that the function $\theta \mapsto \psi_\theta(x)$ possesses two continuous derivatives with respect to θ, for every x. This is true in many examples but fails, for instance in Example 6.1, where $\psi_\theta(x) = \text{sign}(x - \theta)$, which yields the median. Nevertheless, the median is asymptotically normal. That such a simple, but important, example cannot be treated by the preceding approach has motivated much effort to derive the asymptotic normality of Z/M-estimators by more refined methods.

The following result (which we state without proof), taken from van der Vaart [15, Chapter 5], gives some sufficient (less stringent) conditions for a Z-estimator to be asymptotically normal. Compare this with Theorem 2.22 where we gave a similar result for the asymptotic normality of MLEs.

Theorem 6.11 (Asymptotic normality of Z-estimators)**.** For each θ in an open subset of Euclidean space, let $x \mapsto \psi_\theta(x)$ be a measurable vector-valued function such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $M(\cdot)$ with $P[M^2] < \infty$,

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \le M(x)\|\theta_1 - \theta_2\|.$$

Assume that $P\|\psi_{\theta_0}\|^2 < \infty$ and that the map $\theta \mapsto P[\psi_\theta]$ is differentiable at a zero θ_0, with nonsingular derivative matrix V_{θ_0}. If $\mathbb{P}_n[\psi_{\hat{\theta}_n}] = o_p(n^{-1/2})$, and $\hat{\theta}_n \xrightarrow{p} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta_0}(X_i) + o_p(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} P[\psi_{\theta_0}\psi_{\theta_0}^\top](V_{\theta_0}^{-1})^\top$.

Example 6.12 (Location estimators)**.** Let $X_1, \ldots, X_n$ be a random sample of real-valued observations and suppose we want to estimate the location of their distribution. "Location" is a vague term; it could be made precise by defining it as the mean or median, or the center of symmetry of the distribution if this happens to be symmetric. Two examples of location estimators are the sample mean and the sample median. Both can be thought of as Z-estimators, because they solve the equations

$$\sum_{i=1}^n (X_i - \theta) = 0, \qquad \text{and} \qquad \sum_{i=1}^n \operatorname{sign}(X_i - \theta) = 0,$$

respectively[83]. It seems reasonable to study estimators that solve a general equation of the type

$$\sum_{i=1}^n \psi(X_i - \theta) = 0.$$

Popular examples are the Huber estimators corresponding to the functions

$$\psi_k(x) \equiv \psi(x) = \begin{cases} -k & \text{if } x \le -k \\ x & \text{if } |x| \le k \\ k & \text{if } x \ge k. \end{cases}$$

Exercise 2 (HW4): Find the asymptotic distribution of the Huber's location estimator $\hat{\theta}_n$, using ψ_k, when we have i.i.d. data $X_1, \ldots, X_n$ from P on $\mathbb{R}$ (state clearly the assumptions you need on the model P). Notice that $\hat{\theta}_n$ is asymptotically normal regardless of whether P has a finite variance or not. This is an attractive property of $\hat{\theta}_n$.

Example 6.13 (Robust regression)**.** Suppose that $(X, Y) \in \mathbb{R}^k \times \mathbb{R}$ satisfy the following linear regression model:

$$Y = \theta_0^\top X + \varepsilon \tag{124}$$

where ε, the unobserved error, is assumed to be independent of X. Suppose that we have i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (124). The classical estimator for the regression parameter θ is the LSE, which minimizes $\sum_{i=1}^n (Y_i - \theta^\top X_i)^2$. Outlying values of X_i ("leverage points") or extreme values of (X_i, Y_i) jointly ("influence points") can have an arbitrarily large influence on the value of the LSE, which therefore is nonrobust.

As in the case of location estimators, a more robust estimator for θ can be obtained by replacing the square by a function $m(x)$ that grows less rapidly as $x \to \infty$, for instance

[83] The *sign-function* is defined as $\operatorname{sign}(x) = -1, 0, 1$ if $x < 0$, $x = 0$ or $x > 0$, respectively. For the median we assume that there are no tied observations.

$m(x) = |x|$ or $m(x)$ equal to the primitive function of Huber's ψ. Usually, minimizing an expression of the type $\sum_{i=1}^n m(Y_i - \theta^\top X_i)$ is equivalent to solving a system of equations

$$\sum_{i=1}^n \psi(Y_i - \theta^\top X_i) X_i = 0. \tag{125}$$

As $\mathbb{E}[\psi(Y - \theta_0^\top X)X] = \mathbb{E}[\psi(\varepsilon)]\mathbb{E}[X]$, we can expect the resulting estimator to be consistent provided $\mathbb{E}[\psi(\varepsilon)] = 0$. Furthermore, we should expect that, for $V_{\theta_0} = \mathbb{E}[\psi'(\varepsilon)XX^\top]$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} V_{\theta_0}^{-1} \sum_{i=1}^n \psi(Y_i - \theta^\top X_i) X_i + o_p(1).$$

6.3 Asymptotic normality of M-estimators

In this section we present one result that gives the asymptotic distribution of M-estimators for the case of i.i.d. observations. To motivate our high-level approach to deriving the asymptotic distribution of M-estimators, it is probably instructive to study MLEs in parametric models.

Example 6.14 (Parametric maximum likelihood estimators)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. from an unknown density p_{θ_0} belonging to a known class $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. Let $\hat{\theta}_n$ denote the MLE of θ_0. A classical result is that, under some smoothness assumptions, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N_k(0, I^{-1}(\theta_0))$ where $I(\theta_0)$ denotes the Fisher information matrix.

The first step is to observe that if $\theta \mapsto p_\theta(x)$ is sufficiently smooth at θ_0 (e.g., if P_θ is QMD at $\theta = \theta_0$; see (33)), then, for any $h \in \mathbb{R}^k$,

$$\sum_{i=1}^n \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^\top \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^\top I(\theta_0) h + o_{P_{\theta_0}}(1) \tag{126}$$

where $\dot{\ell}_{\theta_0}(x) := \nabla_\theta \log p_\theta(x)$ denotes the score function. Condition (126) is known as the LAN (local asymptotic normality) condition. Let us now define the "local" processes:

$$\tilde{M}_n(h) := \sum_{i=1}^n \log \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \qquad \text{for } \ h \in \mathbb{R}^k \tag{127}$$

and

$$\tilde{M}(h) := h^\top \Delta - \frac{1}{2} h^\top I(\theta_0) h, \qquad \text{for } \ h \in \mathbb{R}^k, \tag{128}$$

where $\Delta \sim N(0, I(\theta_0))$. By LAN we know that (see e.g., Theorem 2.10)

$$\tilde{M}_n(h) \xrightarrow{d} \tilde{M}(h)$$

for every $h \in \mathbb{R}^k$. Observe that

$$\hat{h}_n := \operatorname*{argmax}_{h \in \mathbb{R}^k} \tilde{M}_n(h) = \sqrt{n}(\hat{\theta}_n - \theta_0)$$

is the maximizer of the process (127), and if $\hat{h}$ is the maximizer of $M(\cdot)$, over $\mathbb{R}^k$, then

$$\hat{h} = I^{-1}(\theta_0)\Delta \sim N_k(0, I^{-1}(\theta_0)).$$

We shall prove the asymptotic normality of $\hat{h}_n$ assuming the marginal convergence of (126) (for every fixed h) can be suitably strengthened to a process level result, for every $K \subset \mathbb{R}^k$ compact[84].

The next big idea is: "As $\tilde{M}_n(\cdot) \approx \tilde{M}(\cdot)$ it is reasonable to expect that their maximizers are also close, i.e., $\hat{h}_n \approx \hat{h}$, i.e., the argmax functional is "continuous"[85]. This is formalized by the a class of results that go by the name of the *argmax continuous mapping theorem*[86]. The argmax theorem (Theorem 6.15) will then imply

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{h}_n \xrightarrow{d} \hat{h} \sim N_k(0, I^{-1}(\theta_0)),$$

provided the conditions of the argmax theorem hold. The main condition is tightness of $\{\hat{h}_n\}$ which means that the rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{-1/2}$.

The above idea can be easily extended to derive the asymptotic distributions of other $\sqrt{n}$-consistent estimators, e.g., non-linear regression, robust regression, etc. (see van der Vaart [15, Chapter 5] for more details).

Theorem 6.16 (Asymptotic normality of M-estimators)**.** Suppose that $x \mapsto m_\theta(x)$ is a measurable function for each $\theta \in \Theta \subset \mathbb{R}^k$ for an open set Θ, that $\theta \mapsto m_\theta(x)$ is differentiable at $\theta_0 \in \Theta$ for P-almost every x with derivative $\dot{m}_{\theta_0}(x)$, and that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq F(x)\|\theta_1 - \theta_2\| \tag{130}$$

[84] Formalizing this is a bit technical and we will not get into this in the course.

[85] Note that here both $\hat{h}_n$ and $\hat{h}$ are random, as opposed to the situation when proving the consistency of M-estimators.

[86] In the following result, by $\ell^\infty(K)$ we mean the space of all bounded functions on K. Also, recall the definition of upper semicontinuity: f is upper semicontinuous (u.s.c.) at x_0 if $\limsup_{n\to\infty} f(x_n) \leq f(x_0)$ whenever $x_n \to x_0$ as $n \to \infty$.

Theorem 6.15 (Argmax continuous mapping theorem)**.** Let $\{\mathbb{M}_n(h) : h \in \mathbb{R}^k\}$ and $\{M(h) : h \in \mathbb{R}^k\}$ be stochastic processes indexed by $\mathbb{R}^k$. Suppose that the following conditions hold:

1. $\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(K)$ for every compact subset $K \subset \mathbb{R}^k$ (think of this assumption as being a strengthening of pointwise weak convergence $\mathbb{M}_n(h) \xrightarrow{d} M(h)$ for every $h \in \mathbb{R}^k$).
2. Almost all sample paths $h \mapsto M(h)$ are u.s.c. and possess a unique maximum at a random point $\hat{h}$.
3. For each n, let $\hat{h}_n$ be a random element of $\mathbb{R}^k$ such that $\mathbb{M}_n(\hat{h}_n) \geq \sup_{h\in\mathbb{R}^k} \mathbb{M}_n(h) - o_p(1)$.
4. The following *tightness* condition holds: For every $\epsilon > 0$, there exists a compact set $K_\epsilon \subseteq \mathbb{R}^k$ such that
$$\limsup_{n\to\infty} \mathbb{P}\Big(\hat{h}_n \notin K_\epsilon\Big) \leq \epsilon \qquad \text{and} \qquad \mathbb{P}\Big(\hat{h} \notin K_\epsilon\Big) \leq \epsilon. \tag{129}$$

Then $\hat{h}_n \xrightarrow{d} \hat{h}$ in $\mathbb{R}^k$.

holds for all θ_1, θ_2 in a neighborhood of θ_0, where $F \in L_2(P)$. Also suppose that $M(\theta) = P[m_\theta]$ has a second order Taylor expansion

$$P[m_\theta] - P[m_{\theta_0}] = \frac{1}{2}(\theta - \theta_0)^\top V(\theta - \theta_0) + o(\|\theta - \theta_0\|^2)$$

where θ_0 is a point of maximum of M and V is symmetric and nonsingular (negative definite since M is a maximum at θ_0). If $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_p(n^{-1})$ and $\hat{\theta}_n \xrightarrow{p} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{m}_{\theta_0}(X_i) + o_p(1) \xrightarrow{d} N(0, V^{-1}P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top]V^{-1}).$$

Proof. (Sketch) We can show that, for $\Delta \sim N_k(0, P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top])$,

$$\begin{aligned} \tilde{M}_n(h) &:= n[\mathbb{M}_n(\theta_0 - hn^{-1/2}) - \mathbb{M}_n(\theta_0)] = n\mathbb{P}_n[m_{\theta_0 + hn^{-1/2}} - m_{\theta_0}] \\ &\xrightarrow{d} h^\top\Delta + \frac{1}{2}h^\top V h =: \tilde{M}(h) \end{aligned} \tag{131}$$

as a stochastic "process"[87]. Here $\tilde{M}_n(\cdot)$ is called the *localized, centered* and *rescaled* stochastic process. Then the conclusion will follow from the argmax continuous mapping theorem (see Theorem 6.15) upon noticing that

$$\hat{h}_n := \operatorname*{argmax}_{h \in \mathbb{R}^k} \tilde{M}_n(h) = \sqrt{n}(\hat{\theta}_n - \theta_0),$$

and

$$\hat{h} := \operatorname*{argmax}_{h \in \mathbb{R}^k} \tilde{M}(h) = -V^{-1}\Delta \sim N_k(0, V^{-1}P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top]V^{-1}).$$

One may wonder how does (131) follow. Observe that

$$n\mathbb{P}_n[m_{\theta_0 + hn^{-1/2}} - m_{\theta_0}] = \sqrt{n}(\mathbb{P}_n - P)[\sqrt{n}(m_{\theta_0 + hn^{-1/2}} - m_{\theta_0})] + nP[m_{\theta_0 + hn^{-1/2}} - m_{\theta_0}].$$

By the second order Taylor expansion of $M(\theta) := P[m_\theta]$ about θ_0, the second term of the right side of the last display converges to $(1/2)h^\top V h$ (uniformly on compacta, i.e., for $\|h\| \leq K$, for any $K > 0$). To handle the first term in the above display we need the notion of weak convergence of stochastic processes which we will avoid. We will just show that for every fixed $h \in \mathbb{R}^k$,

$$\tilde{M}_n(h) \xrightarrow{d} \tilde{M}(h),$$

which can be strengthened to show weak convergence of the stochastic processes on compacta. Define $f_{n,h} : \mathcal{X} \to \mathbb{R}$ as

$$f_{n,h}(x) := \sqrt{n}(m_{\theta_0 + hn^{-1/2}} - m_{\theta_0})(x).$$

Then

$$\tilde{\mathbb{M}}_n(h) := \sqrt{n}(\mathbb{P}_n - P)[\sqrt{n}(m_{\theta_0 + hn^{-1/2}} - m_{\theta_0})] = \sqrt{n}(\mathbb{P}_n - P)[f_{n,h}] \xrightarrow{d} N\left(0, h^\top P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top]h\right).$$

by the Lindeberg-Feller CLT (see the next subsections for an complete illustration of these ideas). We conclude that $\tilde{\mathbb{M}}_n(h)$ converges weakly to $h^\top\Delta$, and the desired result holds (here we have also assumed that the tightness condition in Theorem 6.15 holds). □

[87] Note the connection with LAN given in Example 6.14 when $m_\theta(x) = \log p_\theta(x)$.

6.4 Limiting distribution of the sample median

Suppose $X_1, \ldots, X_n$ are i.i.d. observations from a distribution P on $\mathbb{R}$. Assume that P has distribution function F which is differentiable at its median θ_0 (i.e., $F(\theta_0) = 1/2$) with positive derivative $f(\theta_0)$. Let $\hat{\theta}_n$ denote a sample median based on $X_1, \ldots, X_n$ defined as any maximizer of

$$\mathbb{M}_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} |X_i - \theta|$$

over $\theta \in \mathbb{R}$. Also let $M(\theta) := -\mathbb{E}|X_1 - \theta|$ and note that θ_0 uniquely maximizes[88] $M(\theta)$ over $\theta \in \mathbb{R}$. As seen in Example 6.10 we can also work with $M(\theta) = -\mathbb{E}[|X - \theta| - |X|]$ which avoids moment assumptions.

We have already shown that $\hat{\theta}_n$ converges to θ_0 in probability, i.e., $\hat{\theta}_n$ is a consistent estimator of θ_0. Here we will assume that $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$, i.e., the rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{-1/2}$.

We shall now address the question of finding the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. There are many approaches for finding this limiting distribution but we shall follow the standard empirical processes approach which easily generalizes to other M-estimators. This approach also highlights the need to study convergence of stochastic processes.

Our approach for finding the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is based on the following *localized, centered* and *rescaled* stochastic process:

$$\tilde{M}_n(h) := n\left(\mathbb{M}_n(\theta_0 + n^{-1/2}h) - \mathbb{M}_n(\theta_0)\right) \qquad \text{for } h \in \mathbb{R}.$$

This is a stochastic process that is indexed by $h \in \mathbb{R}$. Its important property (easy to see) is that $\hat{h}_n := \sqrt{n}(\hat{\theta}_n - \theta_0)$ maximizes $\tilde{M}_n(h), h \in \mathbb{R}$, i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{h}_n := \operatorname*{argmax}_{h \in \mathbb{R}} \tilde{M}_n(h).$$

This suggests the following approach to find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We study the process $\tilde{M}_n(h), h \in \mathbb{R}$, and argue that it converges as $n \to \infty$ to some limit process $\tilde{M}(h), h \in \mathbb{R}$, in an appropriate sense. If this process convergence is 'strong' enough, then

[88] Indeed, first write

$$-M(\theta) = \mathbb{E}|X_1 - \theta| = \int_{-\infty}^{\theta} (\theta - x)f(x)dx + \int_{\theta}^{\infty} (x - \theta)f(x)dx = \theta\,(2F(\theta) - 1) + \int_{\theta}^{\infty} xf(x)dx - \int_{-\infty}^{\theta} xf(x)dx.$$

This gives

$$-M'(\theta) = 2\theta f(\theta) + 2(F(\theta) - 1) - 2\theta f(\theta) = 2(F(\theta) - 1)$$

and $M''(\theta) = -2f(\theta)$. Note that $M'(\theta) = 0$ implies $F(\theta) = 1/2$ which shows that θ_0 is the unique maximizer of $M(\cdot)$ (as $M''(\theta_0) < 0$ and M is concave on $\mathbb{R}$).

we can hopefully argue that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \operatorname*{argmax}_{h \in \mathbb{R}} \tilde{M}_n(h) \xrightarrow{d} \operatorname*{argmax}_{h \in \mathbb{R}} \tilde{M}(h).$$

It is actually not too hard to understand the behavior of $\tilde{M}_n(h)$ as $n \to \infty$ for each fixed $h \in \mathbb{R}$. For this, we can write

$$\begin{aligned}
\tilde{M}_n(h) &= n\left(\mathbb{M}_n(\theta_0 + n^{-1/2}h) - \mathbb{M}_n(\theta_0)\right) \\
&= n\left((\mathbb{M}_n - M)(\theta_0 + n^{-1/2}h) - (\mathbb{M}_n - M)(\theta_0)\right) + n\left(M(\theta_0 + n^{-1/2}h) - M(\theta_0)\right) \\
&= \sqrt{n}(\mathbb{P}_n - P)[\sqrt{n}(m_{\theta_0 + hn^{-1/2}} - m_{\theta_0})] + nP[m_{\theta_0 + hn^{-1/2}} - m_{\theta_0}] \qquad (132) \\
&=: A_n(h) + B_n(h). \qquad (133)
\end{aligned}$$

where

$$m_{\theta_0 + hn^{-1/2}}(x) - m_{\theta_0}(x) = -\left(|x - \theta_0 - hn^{-1/2}| - |x - \theta_0|\right).$$

Let us now analyze A_n and B_n separately. Clearly, $B_n(h)$ is a deterministic sequence (or every h). To understand this, we shall use a second order Taylor expansion for $M(\theta_0 + n^{-1/2}h)$ around θ_0. Note that $M(\theta) := -\mathbb{E}|X_1 - \theta|$ is a smooth function. Also note that $M'(\theta_0) = 0$ because θ_0 maximizes $M(\theta), \theta \in \mathbb{R}$. We thus get

$$B_n(h) = n\left(M(\theta_0 + n^{-1/2}h) - M(\theta_0)\right) = \frac{1}{2}M''(\theta_0)h^2 + o(1)$$

so that

$$B_n(h) \to \frac{1}{2}M''(\theta_0)h^2 \qquad \text{as } n \to \infty.$$

Let us now come to the mean zero random variable $A_n(h)$. To understand it, let us first compute its variance:

$$\begin{aligned}
\operatorname{Var}(A_n(h)) &= n^2 \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n \left\{|X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\right\}\right) \\
&= n \operatorname{Var}\left(|X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\right) \\
&\approx n \operatorname{Var}\left(I\{X_1 < \theta_0\}n^{-1/2}h - I\{X_1 > \theta_0\}n^{-1/2}h\right)
\end{aligned}$$

where I have ignored the contribution from X_1 lying between θ_0 and $\theta_0 + n^{-1/2}h$ (should not matter for large n; verify this). This gives

$$\operatorname{Var}(A_n(h)) \approx h^2 \operatorname{Var}\Big(I\{X_1 < \theta_0\} - I\{X_1 > \theta_0\}\Big).$$

Now because $\mathbb{P}(X_1 < \theta_0) = \mathbb{P}(X_1 > \theta_0) = 1/2$ (as θ_0 is a population median), it is easy to check that the variance of $I\{X_1 < \theta_0\} - I\{X_1 > \theta_0\}$ appearing above equals 1. We have therefore obtained

$$\operatorname{Var}(A_n(h)) \to h^2 \qquad \text{as } n \to \infty.$$

It is actually possible to prove that

$$A_n(h) \xrightarrow{d} N(0, h^2) = hN(0,1) \qquad \text{as } n \to \infty. \tag{134}$$

For this, we can use the Lindeberg-Feller CLT (stated next).

6.4.1 Lindeberg-Feller Central Limit Theorem

Theorem 6.17 (Lindeberg-Feller CLT)**.** For each n, let $Y_{n1}, \dots, Y_{nk_n}$ be k_n independent random vectors with $\mathbb{E}\|Y_{ni}\|^2 < \infty$ for each $i = 1, \dots, k_n$. Suppose the following two conditions hold:

$$\sum_{i=1}^{k_n} \text{Var}(Y_{ni}) \to \Sigma \qquad \text{as } n \to \infty \tag{135}$$

where $\text{Var}(Y_{ni})$ denotes the covariance matrix of the random vector Y_{ni} and

$$\sum_{i=1}^{k_n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) \to 0 \qquad \text{as } n \to \infty \text{ for every } \epsilon > 0. \tag{136}$$

Then

$$\sum_{i=1}^{k_n} (Y_{ni} - \mathbb{E}Y_{ni}) \xrightarrow{d} N(0, \Sigma) \qquad \text{as } n \to \infty. \tag{137}$$

For a proof of this result, see, for example, van der Vaart [15, Proposition 2.27]. It is easy to see that this result generalizes the usual CLT. Indeed, the usual CLT states that for i.i.d random variables $X_1, X_2, \dots$ with $\mathbb{E}X_i = \mu$, $\mathbb{E}\|X_i\|^2 < \infty$ and $\text{Var}(X_i) = \Sigma$, we have

$$\sum_{i=1}^{n} \left(\frac{X_i}{\sqrt{n}} - \frac{\mu}{\sqrt{n}}\right) \xrightarrow{d} N(0, \Sigma) \qquad \text{as } n \to \infty.$$

Indeed this can be proved by applying Theorem 6.17 to

$$Y_{ni} = \frac{X_i}{\sqrt{n}}.$$

The condition (135) is obvious while for (136) note that

$$\sum_{i=1}^{n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\|X_i\|^2 I\{\|X_i\| > \sqrt{n}\epsilon\}\right) = \mathbb{E}\left(\|X_1\|^2 I\{\|X_1\| > \sqrt{n}\epsilon\}\right)$$

which clearly converges to zero by the DCT (under the assumption $\mathbb{E}\|X_1\|^2 < \infty$).

6.4.2 Back to the limiting distribution of the sample median

Recall the random variables A_n from (133). The Lindeberg-Feller CLT can be used to show (134). Note first that

$$A_n(h) = \sum_{i=1}^{n} (Y_{ni} - \mathbb{E}Y_{ni}) \qquad \text{where} \qquad Y_{ni} := |X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\,.$$

We have already checked that $\text{Var}(A_n(h)) = \sum_{i=1}^{n} \text{Var}(Y_{ni}) \to h^2$ as $n \to \infty$. To check (136), note that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\Big(|X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\Big)^2 I\left\{\big||X_i - \theta_0 - n^{-1/2}h| - |X_i - \theta_0|\big| > \epsilon\right\}\right] \\
&= n\mathbb{E}\left[\Big(|X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\Big)^2 I\left\{\big||X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\big| > \epsilon\right\}\right]
\end{aligned}$$

Using the trivial inequality

$$\Big||X_1 - \theta_0 - n^{-1/2}h| - |X_1 - \theta_0|\Big| \le n^{-1/2}|h|,$$

and the fact that the function $y \mapsto y^2 I\{y > \epsilon\}$ is nondecreasing, we obtain

$$\sum_{i=1}^{n} \mathbb{E}\left(\|Y_{ni}\|^2 I\{\|Y_{ni}\| > \epsilon\}\right) \le h^2 I\{n^{-1/2}|h| > \epsilon\} \to 0 \qquad \text{as } n \to \infty.$$

The conditions of Theorem 6.17 therefore hold and we obtain (134). Thus if we define

$$\tilde{M}(h) := hZ + \frac{1}{2}h^2 M''(\theta_0), \qquad \text{for } h \in \mathbb{R}$$

where $Z \sim N(0,1)$, then we have shown that

$$\tilde{M}_n(h) \xrightarrow{d} \tilde{M}(h) \qquad\qquad \text{for every } h \in \mathbb{R}.$$

It turns out that the process $\tilde{M}_n$ converges to $\tilde{M}$ in a stronger sense than convergence in distribution for each fixed $h \in \mathbb{R}$. We shall see this later. This stronger convergence allows us to deduce that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{h}_n := \operatorname*{argmax}_{h\in\mathbb{R}} \tilde{M}_n(h) \xrightarrow{d} \operatorname*{argmax}_{h\in\mathbb{R}} \tilde{M}(h) = \frac{-Z}{M''(\theta_0)} \sim N\left(0, \frac{1}{(M''(\theta_0))^2}\right).$$

We can simplify this slightly by writing $M''(\theta_0)$ in terms of $f(\theta_0)$. We thus have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\theta_0)}\right)$$

To make this argument rigorous, we have to prove that the stochastic process $\tilde{M}_n$ converges to $\tilde{M}$ in a strong enough sense so that their argmaxs also converge.

Example 6.18 (Asymptotic normality of the sample median using Theorem 6.16)**.** Recall that the sample median maximizes the criterion function $\theta \mapsto -\sum_{i=1}^n |X_i - \theta|$. We take $m_\theta(x) := -(|x-\theta| - |x|)$ (as in Example 6.10). We can apply Theorem 6.16 to find the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We can show that (130) holds with $F(x) = 1$. Furthermore, the map $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 except if $x = \theta_0$, with $\dot{m}_{\theta_0}(x) = -\text{sign}(x - \theta_0)$. Under the minimal condition that F is differentiable at θ_0, $P[m_\theta]$ has a two-term Taylor expansion (around θ_0) as $P[m_\theta] = P[m_{\theta_0}] - \frac{1}{2}(\theta-\theta_0)^2 2f(\theta_0) + o((\theta-\theta_0)^2)$, so that we can set $V_{\theta_0} = -2f(\theta_0)$. As $P[\dot{m}_{\theta_0}^2] = \mathbb{E}[1] = 1$, the asymptotic variance of the median is $1/(2f(\theta_0))^2$.

Example 6.19 (Nonlinear least squares)**.** Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of a vector (X, Y) that follows the regression model

$$Y = f_{\theta_0}(X) + \epsilon, \qquad \mathbb{E}(\epsilon|X) = 0.$$

Here f_θ is a parametric family of regression functions, for instance $f_\theta(x) = \theta_1 + \theta_2 e^{\theta_3 x}$, and we aim at estimating the unknown vector θ. The least squares estimator that minimizes

$$\theta \mapsto \sum_{i=1}^n (Y_i - f_\theta(X_i))^2$$

is an M-estimator for $m_\theta(x, y) = (y - f_\theta(x))^2$. It should be expected to converge to the minimizer of the limit criterion function

$$\theta \mapsto P[m_\theta] = P[(f_{\theta_0} - f_\theta)^2] + \mathbb{E}[\epsilon^2].$$

Thus the LSE should be consistent if θ_0 is identifiable from the model, in the sense that $\theta \neq \theta_0$ implies that $f_\theta(X) \neq f_{\theta_0}(X)$ with positive probability. For sufficiently regular regression models, we have

$$P[m_\theta] \approx P\left[\{(\theta - \theta_0)^\top \dot{f}_{\theta_0}\}^2\right] + \mathbb{E}[\epsilon^2].$$

This suggests that the conditions of Theorem 6.16 are satisfied with $V_{\theta_0} = 2P[\dot{f}_{\theta_0}\dot{f}_{\theta_0}^\top]$ and $\dot{m}_{\theta_0}(x, y) = -2(y - f_{\theta_0}(x))f_{\theta_0}(x)$. If ϵ and X are independent, then this leads to the asymptotic covariance matrix $V_{\theta_0}^{-1} 2\mathbb{E}[\epsilon^2]$.

Exercise 3 (HW4): Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. from a joint distribution P on $[0, 1] \times \mathbb{R}$. Suppose that we postulate a model of the form

$$Y_i = \alpha_0(X_i - \beta_0)_+ + \epsilon_i,$$

where ϵ_i's are i.i.d. with mean 0 and unknown constant variance σ^2 (here $x_+ = \max(0, x)$). Let $(\hat{\alpha}_n, \hat{\beta}_n)$ be the least squares estimator of the unknown parameter (α_0, β_0). Under appropriate conditions (state clearly the conditions you need on the distribution P) and assuming the consistency of $(\hat{\alpha}_n, \hat{\beta}_n)$, find its asymptotic distribution.

6.5 Asymptotics for minimizers of convex processes

To study the asymptotic for M-estimators (here we consider minimizers instead of maximizers) it is more convenient to first consider minimizers of *convex* processes. In this case the proofs are simpler and the main ideas become more transparent (also a comparison can be drawn easily with the approach of QDM, as in Section 2.1.2)[89].

6.5.1 Preliminaries

= Our main result will crucially use the fact that $\mathbb{M}_n(\theta)$ (as in (113)) is a convex function of θ and is reliant on the following convexity lemma (we will not be proving the lemma here; you may see Pollard [11, Section 6] if you are interested in its proof).

Lemma 6.20 (Uniform convergence on compacta)**.** Let $\{\lambda_n(h) : h \in \mathcal{S} \subset \mathbb{R}^k\}$ be a sequence of random convex functions defined on a convex, open subset $\mathcal{S}$ of $\mathbb{R}^k$. Suppose $\lambda(\cdot)$ is a real-valued function on $\mathcal{S}$ for which $\lambda_n(h) \xrightarrow{p} \lambda(h)$ as $n \to \infty$, for each $h \in \mathcal{S}$. Then for every compact subset K of $\mathcal{S}$, we have

$$\sup_{h \in K} |\lambda_n(h) - \lambda(h)| \xrightarrow{p} 0.$$

Qualitatively, the convexity lemma states that, under the assumption of convexity, pointwise convergence can be turned into uniform convergence on compacta.

Lemma 6.21 (Nearness of argmins)**.** Suppose $\lambda_n(\cdot)$ is convex as in Lemma 6.20 and is approximated by $\tilde{\lambda}_n(\cdot)$. Let α_n be the argmin of $\lambda_n(\cdot)$[90] and let β_n be the unique minimizer of $\tilde{\lambda}_n(\cdot)$ over $\mathcal{S}$. Then, for each $\delta > 0$,

$$\mathbb{P}\left(|\alpha_n - \beta_n| \geq \delta\right) \leq \mathbb{P}\left(\Delta_n(\delta) \geq \frac{1}{2} v_n(\delta)\right)$$

where

$$\Delta_n(\delta) := \sup_{h:|h-\beta_n| \leq \delta} \left|\lambda_n(h) - \tilde{\lambda}_n(h)\right| \qquad \text{and} \qquad v_n(\delta) := \inf_{h:|h-\beta_n|=\delta} \tilde{\lambda}_n(h) - \tilde{\lambda}_n(\beta_n).$$

Proof. The lemma as stated has nothing to do with convergence or indeed with the 'n' subscript at all; but it is stated in that form so that it can be useful for later purposes. To prove it, let h be an arbitrary point outside the ball around β_n with radius δ, say $h = \beta_n + lu$ for a unit vector $u \in \mathbb{R}^k$, where $l \geq \delta$. Convexity of λ_n implies

$$\lambda_n(\beta_n + \delta u) = \lambda_n\left((1 - \delta/l)\beta_n + (\delta/l)h\right) \leq (1 - \delta/l)\lambda_n(\beta_n) + (\delta/l)\lambda_n(h).$$

[89] Moreover, this approach avoids relying on weak convergence of processes and the argmin continuous mapping theorems; see e.g. van der Vaart and Wellner [16, Chapter 3.2].

[90] A convex function is continuous and attains it minimum on compact sets, but it can be flat at its bottom and have several minima. For simplicity we speak about 'the argmin' when referring to any of the possible minimizers.

Writing for convenience $r_n(h) := \lambda_n(h) - \tilde{\lambda}_n(h)$ (for all $h \in \mathcal{S}$), we deduce that

$$\begin{aligned}(\delta/l)\big\{\lambda_n(h) - \lambda_n(\beta_n)\big\} &\geq \lambda_n(\beta_n + \delta u) - \lambda_n(\beta_n) \\ &= \big[\tilde{\lambda}_n(\beta_n + \delta u) - \tilde{\lambda}_n(\beta_n)\big] + r_n(\beta_n + \delta u) - r_n(\beta_n) \\ &\geq v_n(\delta) - 2\Delta_n(\delta). \end{aligned} \tag{138}$$

Thus, if $|\alpha_n - \beta_n| \geq \delta$, then by taking $h = \alpha_n$ we have $\lambda_n(h) - \lambda_n(\beta_n) \leq 0$ (as α_n is an argmin of $\lambda_n(\cdot)$), which implies that (by (138)) $v_n(\delta) - 2\Delta_n(\delta) \leq 0$, thereby yielding the desired result. □

The above results gives a probabilistic bound on how far α_n can be from β_n.

We record a couple of useful implications of Lemma 6.21. If $\lambda_n - \tilde{\lambda}_n$ goes to zero uniformly on bounded sets in probability and β_n is stochastically bounded, then $\Delta_n(\delta) \xrightarrow{p} 0$ by a simple argument[91]. It follows that $\alpha_n - \beta_n \xrightarrow{p} 0$ provided only that $1/v_n(\delta)$ is stochastically bounded for each fixed δ. This last requirement says that $\tilde{\lambda}_n$ shouldn't flatten out around its minimum as n increases.

Corollary 6.22. Suppose $M_n(\cdot)$ is convex (random) function such that

$$M_n(h) = \frac{1}{2}h^\top V h + U_n^\top h + C_n + r_n(h), \qquad \text{for } \ h \in \mathcal{S},$$

where V is a symmetric and positive definite matrix, U_n is stochastically bounded, C_n is arbitrary, and $r_n(h)$ goes to zero in probability for each $h \in \mathcal{S}$. Then α_n, the argmin of λ_n, is only $o_p(1)$ away from $\beta_n = -V^{-1}U_n$, the argmin of $\tilde{M}_n(\cdot)$, where

$$\tilde{M}_n(h) := \frac{1}{2}h^\top V h + U_n^\top h + C_n.$$

If also $U_n \xrightarrow{d} U$ then $\alpha_n \xrightarrow{d} -V^{-1}U$.

Proof. The function $\lambda_n(h) := M_n(h) - U_n^\top h - C_n$ is convex and goes to $(1/2)h^\top V h$ in probability for each h. By Lemma 6.20 the convergence is uniform on bounded sets. Let $\Delta_n(\delta)$ be the supremum of $|r_n(h)|$ over $\{h \in \mathcal{S} : |h - \beta_n| \leq \delta\}$. Then, by Lemma 6.21,

$$\alpha_n = -V^{-1}U_n + \epsilon_n, \qquad \text{where} \quad \mathbb{P}(|\epsilon_n| \geq \delta) \leq \mathbb{P}\left(\Delta_n(\delta) \geq \frac{1}{2}c_{min}\,\delta^2\right) \to 0.$$

where c_{min} is the smallest eigenvalue of V, and $\Delta_n(\delta) \xrightarrow{p} 0$, by the arguments used above. □

A useful slight extension of this is when $M_n(h) = (1/2)h^\top V_n h + U_n^\top h + C_n + r_n(h)$ is convex, with a nonnegative definite symmetric matrix V_n that converges in probability to a positive definite V. Writing $V_n = V + \eta_n$ the remainder η_n can be absorbed into $r_n(h)$ and the result above holds.

[91] Exercise 4 (HW4): Show this.

6.5.2 Asymptotic normality of M-estimators for convex processes

Recall the M-estimation setup: Given $X_1, \ldots, X_n$ i.i.d. P taking values in some space $\mathcal{X}$ (e.g., a metric space) the M-estimator of interest is

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} \mathbb{M}_n(\theta) \qquad \text{where} \quad \mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta] = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i),$$

for $\Theta \subset \mathbb{R}^k$, and $m_\theta : \mathcal{X} \to \mathbb{R}$ is a "loss" function. Here we assume that $m_\theta(x)$ is a *convex function in* θ, for every $x \in \mathcal{X}$. The (true value of the) parameter of interest is

$$\theta_0 := \arg\min_{\theta \in \Theta} M(\theta) \qquad \text{where} \quad M(\theta) := P[m_\theta] = \mathbb{E}_{X \sim P}[m_\theta(X)]. \tag{139}$$

We can obtain the consistency of $\hat{\theta}_n$, the argmin of $\mathbb{M}_n(\theta)$ (over $\theta \in \Theta$) by appealing to Lemma 6.21 with $\tilde{\lambda}(\theta) = M(\theta)$ and $\beta_n = \theta_0$. Compare this with Lemma 6.3.

To study the asymptotic normality of $\hat{\theta}_n$ we need some 'weak' expansion of $m_\theta(x)$ around the value θ_0 is needed, but we avoid explicitly requiring pointwise derivatives to exist. With this in mind, write

$$m_{\theta_0+h}(x) - m_{\theta_0}(x) = \dot{m}_{\theta_0}(x)^\top h + R_h(x), \tag{140}$$

for a function $\dot{m}_{\theta_0} : \mathcal{X} \to \mathbb{R}$ with mean zero under P. If $\mathbb{E}[R_h(X)^2]$ is of order $o(|h|^2)$ as $h \to 0$, as we will usually require, then $\dot{m}_{\theta_0}(x)$ is nothing but the derivative in quadratic mean of the function $m_{\theta_0+h}(x)$ at $h = 0$.

Theorem 6.23. Suppose that $m_\theta(x)$ is convex in θ (for every x) and that (140) holds with

$$M(\theta_0 + h) - M(\theta_0) = \mathbb{E}[R_h(X)] = \frac{1}{2}h^\top V h + o(|h|^2) \tag{141}$$

where θ_0 is a point of minimum of M and V is symmetric and nonsingular (positive definite since M is a maximum at θ_0). Suppose also that $\mathrm{Var}(R_h(X)) = o(|h|^2)$, and that $\dot{m}_{\theta_0}(X)$ has a finite covariance matrix $J := P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top]$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{P}_n[\dot{m}_{\theta_0}] + o_p(1) \xrightarrow{d} N\left(0, V^{-1}JV^{-1}\right). \tag{142}$$

Proof. Consider the convex function

$$\lambda_n(h) := n\left[\mathbb{M}_n(\theta_0 + hn^{-1/2}) - \mathbb{M}_n(\theta_0)\right] = \sum_{i=1}^n \left[m_{\theta_0+hn^{-1/2}}(X_i) - m_{\theta_0}(X_i)\right].$$

This random process is minimized at $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Also, note that by taking $\theta = \theta_0 + hn^{-1/2}$ in (141), we get

$$n\left[M(\theta_0 + hn^{-1/2}) - M(\theta_0)\right] = h^\top V h + q_n(h),$$

where $q_n(h) = no(|h|^2/n) \to 0$, for fixed h. Accordingly, by (140),

$$\begin{aligned}
\lambda_n(h) &:= \sum_{i=1}^n \left[\dot{m}_{\theta_0}(X_i)^\top h n^{-1/2} + R_{hn^{-1/2}}(X_i)\right] \\
&= \sum_{i=1}^n \left[\dot{m}_{\theta_0}(X_i)^\top h n^{-1/2} + R_{hn^{-1/2}}(X_i) - \mathbb{E}[R_{hn^{-1/2}}(X_i)]\right] + n\mathbb{E}[R_{hn^{-1/2}}(X_i)] \\
&= U_n^\top h + r_n(h) + \frac{1}{2} h^\top V h + q_n(h)
\end{aligned}$$

where

$$U_n = n^{-1/2} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i), \qquad \text{and} \qquad r_n(h) := \sum_{i=1}^n \left\{R_{hn^{-1/2}}(X_i) - \mathbb{E}[R_{hn^{-1/2}}(X_i)]\right\}.$$

Now $r_n(h)$ tends to zero in probability for each h, since its mean is zero and its variance is $\sum_{i=1}^n \mathrm{Var}(R_{hn^{-1/2}}(X_i)) = no(1/n)$. This, together with Corollary 6.22, proves (142) and the limit distribution result, since U_n goes to a $N(0, J)$ by the CLT. Note that both consistency and asymptotic normality followed from the same approximation argument. □

This convexity based argument can be used to prove asymptotics of many other M-estimators that are based on convex optimization (see Hjort and Pollard [7]).

Exercise 5 (HW4): (Asymptotic distribution of sample median) Recall Example 6.1. The goal of this problem is to find the asymptotic distribution of the sample median (properly normalized). For $h \in \mathbb{R}$, let

$$\lambda_n(h) := n\left(\mathbb{M}_n(\theta_0 + n^{-1/2}h) - \mathbb{M}_n(\theta_0)\right).$$

Also let

$$\tilde{\lambda}_n(h) := h^2 f(\theta_0) + \frac{h}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i) \qquad \text{where} \quad \dot{m}_{\theta_0}(X_i) := I\{X_i \le \theta_0\} - I\{X_i > \theta_0\}.$$

1. For every $h \in \mathbb{R}$, prove that $\lambda_n(h) - \tilde{\lambda}_n(h)$ converges in probability to zero as $n \to \infty$.
2. Use Lemma 6.20 to prove that, for every $L > 0$, $\sup_{h:|h|\le L} \left|\lambda_n(h) - \tilde{\lambda}_n(h)\right| \xrightarrow{p} 0$.
3. Note that $\alpha_n := \sqrt{n}(\hat{\theta}_n - \theta_0)$ minimizes $\lambda_n(h)$ over $h \in \mathbb{R}$. Combine the results of the two parts above to argue that $\alpha_n - \beta_n$ converges to zero in probability, where β_n (uniquely) minimizes $\tilde{\lambda}_n(h)$ over $h \in \mathbb{R}$.
4. Deduce the asymptotic distribution of $\alpha_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$.

7 Bootstrap methods

Suppose that we have data $\mathbf{X} \sim P$, and $\theta \equiv \theta(P)$ is a parameter of interest. Think of $\mathbf{X} = (X_1, \ldots, X_n)$, where n is the sample size. Let $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ be an estimator of θ. Suppose that we would want to construct a level-$(1-2\alpha)$ *confidence interval* (CI) for θ, i.e., find κ_α and $\kappa_{1-\alpha}$ such that

$$\mathbb{P}(\hat{\theta} - \kappa_\alpha \leq \theta \leq \hat{\theta} + \kappa_{1-\alpha}) = 1 - 2\alpha. \tag{143}$$

Question: How do we find (estimate) κ_α and $\kappa_{1-\alpha}$ in such a general setting?

Problem: To solve the above problem we need to find (estimate) the distribution of $\hat{\theta} - \theta$. However, the distribution of $\hat{\theta} - \theta$ depends on P and might be *unknown*. Even if we know its asymptotic distribution (e.g., $\hat{\theta} - \theta$ is asymptotically normal), we may want more accurate quantiles for a fixed sample size. In some situations, the asymptotic limiting distribution can depend on *nuisance* parameters that can be hard to estimate.

Example 7.1. Let $X_1, \ldots, X_n$ be a random sample from a univariate distribution with distribution function F having a density f continuous and positive in a neighborhood of it median $\theta_0 := \inf\{t : F(t) \geq 1/2\}$. The sample median may be defined as

$$\hat{\theta}_n := \inf\{t : \mathbb{F}_n(t) \geq 1/2\} \tag{144}$$

or, using Example 6.1 it is any minimizer of (116) (Note that $\mathbb{F}_n(t) := \frac{1}{n}\sum_{i=1}^n I(X_i \leq t)$, for $t \in \mathbb{R}$ is the *empirical distribution function* of the data). Suppose the goal is now to construct a confidence interval for the population median θ_0. Under the above conditions, it can be shown that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \tau^2), \qquad \text{as } n \to \infty,$$

where $\tau^2 = 1/(4f^2(\theta_0))$. Although, the above limit distribution can be used to construct a CI for θ_0, it would involve the estimation of $f(\theta_0)$, which is tricky to estimate. A natural question that arises now: Can we find a CI for θ_0 that avoids the estimation of the nuisance parameter $f(\theta_0)$?

Example 7.2. Let $X_1, \ldots, X_n$ be a random sample from a univariate distribution with finite first and second moments. Suppose that the goal is to construct a CI for $\mu = \mathbb{E}(X_1)$. Of course, we can use the CLT directly to construct a CI for μ:

$$\left[\bar{X}_n - z_\alpha \frac{s}{\sqrt{n}}, \bar{X}_n + z_\alpha \frac{s}{\sqrt{n}}\right],$$

where z_α is the upper α-th quantile of the standard normal distribution, $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, and $s^2 := \sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)$ is the (unbiased) sample variance. However, we may ask the following question: Can we get a more accurate CI?

We will see in the following that *bootstrap* can be a useful technique in these two problems and beyond.

7.1 Bootstrap: Introduction

To motivate the bootstrap method, let us consider the following simple scenario. Suppose that we model our data $\mathbf{X} = (X_1, \ldots, X_n)$ as a random sample from some distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is a class of probability distributions. Let $\eta(\mathbf{X}, P)$ be a *root*, i.e., a random variable that possibly depends on both the distribution P and the sample $\mathbf{X}$ drawn from P (e.g., think of $\eta(\mathbf{X}, P)$ as $\sqrt{n}(\bar{X}_n - \mu)$, where $\bar{X}_n = \sum_{i=1}^n X_i/n$ and $\mu = \mathbb{E}(X_1)$). In fact, $\hat{\theta} - \theta$ (as described above) is a root.

In general, we may wish to estimate the mean or a quantile or some other probabilistic feature or the entire *distribution* of $\eta(\mathbf{X}, P)$. As mentioned above, the distribution of $\hat{\theta} - \theta$ depends on P and is thus unknown. Let $H_n(x, P)$ denote the c.d.f. of $\eta(\mathbf{X}, P)$, i.e.,

$$H_n(x, P) := \mathbb{P}_{\mathbf{X} \sim P}(\eta(\mathbf{X}, P) \leq x). \tag{145}$$

Of course, if we can estimate $H_n(\cdot, P)$ then we can use this to construct CIs, test hypotheses; e.g., if $\eta(\mathbf{X}, P) = \hat{\theta} - \theta$ then being able to estimate $H_n(\cdot, P)$ immediately yields estimates of κ_α and $\kappa_{1-\alpha}$ as defined in (143).

Question: What if we knew P and could draw unlimited samples from P?

In that case we could approximate $H_n(x, P)$ as follows: Draw repeated samples (of size n) from P resulting in a series of values for the root $\eta(\mathbf{X}, P)$, then we could form an estimate of $H_n(x, P)$ by counting how many of the $\eta(\mathbf{X}, P)$'s are $\leq x$.

But, of course, we do not know P. However we can *estimate* P by $\hat{P}_n$ and use the above idea. This is the notion of bootstrap.

Definition 7.3 (Bootstrap)**.** The bootstrap is a method of replacing (plugging in) the unknown distribution P with $\hat{P}_n$ (estimated from the data) in probability/expectation calculations.

The bootstrap approximation of $H_n(\cdot, P)$ is $H_n(\cdot, \hat{P}_n)$, where $\hat{P}_n$ is an estimator of P obtained from the observed data (that we think is close to P), i.e.,

$$\hat{H}_n(x) \equiv H_n(x, \hat{P}_n) \;\; := \;\; \mathbb{P}_{\mathbf{X}^* \sim \hat{P}_n}\left(\eta(\mathbf{X}^*, \hat{P}_n) \leq x | \mathbf{X}\right) \equiv \mathbb{P}^*\left(\eta(\mathbf{X}^*, \hat{P}_n) \leq x | \mathbf{X}\right). \tag{146}$$

where $\mathbb{P}^*_{\mathbf{X}^* \sim \hat{P}_n}(\cdot|\mathbf{X})$ is the conditional probability given the observed data $\mathbf{X}$ (under the estimated $\hat{P}_n$). Thus, bootstrap estimates the distribution of $\eta(\mathbf{X}, P)$ by that of $\eta(\mathbf{X}^*, \hat{P}_n)$, where $\mathbf{X}^*$ is a random sample drawn from the distribution $\hat{P}_n$ (conditional on the data). The idea is that

$$\text{if } \hat{P}_n \approx P, \qquad \text{then} \qquad H_n(\cdot, \hat{P}_n) \approx H_n(\cdot, P).$$

Question: How do we find $H_n(\cdot, \hat{P}_n)$, the distribution of $\eta(\mathbf{X}^*, \hat{P}_n)$?

Answer: In most cases, the distribution of $\eta(\mathbf{X}^*, \hat{P}_n)$ is difficult to analytically compute, but it can *always* be approximated easily by Monte Carlo *simulations*.

Thus, the bootstrap can thus be broken down in the following simple steps:

- Find a "good" estimator $\hat{P}_n$ of P.
- Draw a large number (say, B) of random samples $\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(B)}$ from the distribution $\hat{P}_n$ and then compute $T^{*(j)} := \eta(\mathbf{X}^{*(j)}, \hat{P}_n)$, for $j = 1, \ldots, B$.
- Finally, compute the desired feature of $\eta(\mathbf{X}^*, \hat{P}_n)$ using the empirical c.d.f. $\tilde{H}_n^B(\cdot, \hat{P}_n)$ of the values $T^{*(1)}, \ldots, T^{*(B)}$, i.e.,

$$\tilde{H}_n^B(x, \hat{P}_n) := \frac{1}{B} \sum_{j=1}^{B} I(T^{*(j)} \leq x), \qquad \text{for } \ x \in \mathbb{R}. \tag{147}$$

Intuitively,

$$\tilde{H}_n^B(\cdot, \hat{P}_n) \ \approx \ H_n(\cdot, \hat{P}_n) \ \approx \ H_n(\cdot, P),$$

where the first approximation is from Monte Carlo error (and can be as small as we would like, by taking B as large as we want) and the second approximation is due to the bootstrap method. If $\hat{P}_n$ is a good approximation of P, then the bootstrap can yield a very useful approximation of $H_n(\cdot, P)$.

Example 7.4 (Bootstrapping the sample mean)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. F and that $\sigma^2 := \mathrm{Var}(X_1) < \infty$. Let $\eta(\mathbf{X}, F) := \sqrt{n}(\bar{X}_n - \mu)$, where $\mu := \mathbb{E}(X_1)$ and $\bar{X}_n := \sum_{i=1}^n X_i/n$. A natural estimator of F is $\mathbb{F}_n$, the e.d.f. of the data. Thus, we approximate the distribution of $\eta(\mathbf{X}, F)$ by that of $\eta(\mathbf{X}^*, \mathbb{F}_n)$ where $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ are drawn i.i.d. from $\mathbb{F}_n$. Note that

$$\eta(\mathbf{X}^*, \mathbb{F}_n) = \sqrt{n}(\bar{X}_n^* - \bar{X}_n), \qquad \text{where} \qquad \bar{X}_n^* := \frac{1}{n} \sum_{i=1}^{n} X_i^*.$$

We may approximate the distribution of $\eta(\mathbf{X}^*, \mathbb{F}_n)$ by drawing many Monte Carlo samples from $\mathbb{F}_n$ and computing $\eta(\mathbf{X}^*, \mathbb{F}_n)$ again and again, as in (147). Question: How does one draw Monte Carlo samples from $\mathbb{F}_n$?

7.2 Parametric bootstrap

In parametric models it is more natural to take $\hat{P}_n$ as the fitted parametric model.

Example 7.5 (Estimating the standard deviation of a statistic)**.** Suppose that $X_1, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$. Suppose that we are interested in the parameter

$$\theta = \mathbb{P}(X \leq c) = \Phi\left(\frac{c - \mu}{\sigma}\right),$$

where c is a given known constant. A natural estimator of θ is its MLE $\hat{\theta}$:

$$\hat{\theta} = \Phi\left(\frac{c - \bar{X}_n}{\hat{\sigma}}\right),$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Question: How do we estimate the standard deviation of $\hat{\theta}$? There is no easy closed form expression for this.

Solution: We can bootstrap! Draw many (say B) bootstrap samples of size n from

$$N(\bar{X}, \hat{\sigma}^2) \equiv \hat{P}_n.$$

For the j-th *bootstrap sample* we compute a sample average $\bar{X}^{*(j)}$, a sample standard deviation $\hat{\sigma}^{*(j)}$. Finally, we compute

$$\hat{\theta}^{*(j)} = \Phi\left(\frac{c - \bar{X}^{*(j)}}{\hat{\sigma}^{*(j)}}\right).$$

We can estimate the mean of $\hat{\theta}$ by $\bar{\theta}^* = \frac{1}{B}\sum_{j=1}^B \hat{\theta}^{*(j)}$. The standard deviation of $\hat{\theta}$ can then be estimated by the bootstrap standard deviation of the $\hat{\theta}^{*(j)}$ values, i.e.,

$$\left[\frac{1}{B}\sum_{j=1}^B (\hat{\theta}^{*(j)} - \bar{\theta}^*)^2\right]^{1/2}.$$

Example 7.6 (Comparing means when variances are unequal)**.** Suppose that we have two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal populations. Suppose that

$$X_1, \ldots, X_m \text{ are i.i.d. } N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y_1, \ldots, Y_n \text{ are i.i.d. } N(\mu_2, \sigma_2^2).$$

Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_2.$$

We can use the test statistic

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}},$$

where $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_X^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2$ and $S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Note that as $\sigma_1^2 \neq \sigma_2^2$, U does not necessarily follow a t-distribution.

Question: How do we find the critical value of this test?

The parametric bootstrap can proceed as follows:

First choose a large number B, and for $j = 1, \ldots, B$, simulate $(\bar{X}_m^{*(j)}, \bar{Y}_n^{*(j)}, S_X^{2*(j)}, S_Y^{2*(j)})$, where all four random variables are independent with the following distributions:

- $\bar{X}_m^{*(j)} \sim N(0, \hat{\sigma}_X^2/m)$,
- $\bar{Y}_n^{*(j)} \sim N(0, \hat{\sigma}_Y^2/n)$,
- $S_X^{2*(j)} \sim \hat{\sigma}_X^2 \, \chi_{m-1}^2$,
- $S_Y^{2*(j)} \sim \hat{\sigma}_Y^2 \, \chi_{n-1}^2$,

where $\hat{\sigma}_X^2 = S_X^2/(m-1)$ and $\hat{\sigma}_Y^2 = S_Y^2/(n-1)$. Then we compute

$$U^{*(j)} = \frac{(m+n-2)^{1/2}(\bar{X}_m^{*(j)} - \bar{Y}_n^{*(j)})}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^{2*(j)} + S_Y^{2*(j)})^{1/2}}$$

for each j. We approximate the null distribution of U by the empirical distribution of the $\{U^{*(j)}\}_{j=1}^B$. Let c_n^* be the $\left(1-\frac{\alpha}{2}\right)$-quantile of the empirical distribution of $\{U^{*(j)}\}_{j=1}^B$. Then, we can reject H_0 if

$$|U| > c_n^*.$$

7.3 The nonparametric bootstrap

In problems where the distribution P is not indexed by a parametric family, a natural estimator of P is the empirical distribution $\hat{P}_n$ given by the distribution that puts $1/n$-mass at each of the observed data points.

Example 7.7. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample from a distribution F on $\mathbb{R}$. Suppose that we want a CI for the median θ_0 of F. We can base a CI on the sample median $\hat{\theta}_n$ (see (144)).

We want to estimate the distribution of $\hat{\theta}_n - \theta_0$. Let $\eta(\mathbf{X}, F) := \hat{\theta}_n - \theta_0$. We may choose $\hat{F} = \mathbb{F}_n$, the empirical distribution function of the observed data. Thus, our method can be broken in the following steps:

- Choose a large number B and simulate many samples $\mathbf{X}^{*(j)}$, for $j = 1, \ldots, B$, (conditionally i.i.d. given the data) from $\mathbb{F}_n$. This reduces to drawing *with replacement sampling* from $\mathbf{X}$.
- For each bootstrap sample we compute the sample median $\hat{\theta}_n^{*(j)}$ and then find the appropriate sample quantiles of $\{\hat{\theta}_n^{*(j)} - \hat{\theta}_n\}_{i=1}^B$. Observe that $\eta(\mathbf{X}^*, \hat{P}_n) = \hat{\theta}_n^* - \hat{\theta}_n$.

7.4 Consistency of the bootstrap

Suppose that $\hat{F}_n$ and F are the corresponding c.d.f.'s for $\hat{P}_n$ and P respectively. Suppose that $\hat{P}_n$ is a consistent estimator of P. This means that at each x in the support of X_1

where $F(x)$ is continuous, $\hat{F}_n(x) \to F(x)$ in probability or a.s. as $n \to \infty$[92]. If, in addition, $H_n(x, P)$, considered as a functional of P, is *continuous* in an appropriate sense, it can be expected that $H_n(x, \hat{P}_n)$ will be close to $H_n(x, P)$, when n is large.

Observe that $\hat{H}_n(x) \equiv H_n(x, \hat{P}_n)$ is a *random distribution function* (as it depends on the observed data). Let ρ be any notion of distance between two probability distributions that metrizes weak convergence, i.e., for any sequence of c.d.f.'s $\{G_n\}_{n\geq 1}$, we have

$$G_n \xrightarrow{d} G \qquad \text{if and only if} \qquad \rho(G_n, G) \to 0 \ \text{ as } \ n \to \infty.$$

In particular, we can take ρ to be the Levy metric[93]. For simplicity, we can also use the uniform distance (Kolmogorov metric) between G_n and G (which metrizes weak convergence if G is a continuous c.d.f.).

Definition 7.8. We say that the bootstrap is *weakly consistent* under ρ for $\eta(\mathbf{X}_n, P)$ if

$$\rho(H_n, \hat{H}_n) \xrightarrow{p} 0 \qquad \text{as } \ n \to \infty,$$

where H_n and $\hat{H}_n$ are defined in (145) and (146) respectively. We say that the bootstrap is *strongly consistent* under ρ for $\eta(\mathbf{X}_n, P)$ if

$$\rho(H_n, \hat{H}_n) \xrightarrow{a.s.} 0 \qquad \text{as } \ n \to \infty.$$

In many problems, it can be shown that $H_n(\cdot, P)$ converges in distribution to a limit $H(\cdot, P)$. In such situations, it is much easier to prove that the bootstrap is consistent by showing that

$$\rho(\hat{H}_n, H) \xrightarrow{a.s./p} 0 \qquad \text{as } \ n \to \infty.$$

In applications, e.g., for construction of CIs, we are quite often interested in approximating the quantiles of H_n by that of $\hat{H}_n$ (as opposed to the actual c.d.f.). The following simple result shows that weak convergence, under some mild conditions, implies the convergence of the quantiles.

Exercise 6 (HW4): Let $\{G_n\}_{n\geq 1}$ be a sequence of distribution functions on the real line converging weakly to a distribution function G, i.e., $G_n(x) \to G(x)$ at all continuity points x of G. Assume that G is continuous and strictly increasing at $y = G^{-1}(1-\alpha)$. Then,

$$G_n^{-1}(1-\alpha) := \inf\{x \in \mathbb{R} : G_n(x) \geq 1-\alpha\} \to y = G^{-1}(1-\alpha).$$

[92] If F is a continuous c.d.f., then it follows from Polya's theorem that $\hat{F}_n \to F$ in probability or a.s. uniformly over x. Thus, $\hat{F}_n$ and F are uniformly close to one another if n is large.

[93] Let $F, G : \mathbb{R} \to [0,1]$ be two cumulative distribution functions. Define the Lévy distance between them to be

$$L(F, G) := \inf\{\varepsilon > 0 | F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \,\text{for all}\, x \in \mathbb{R}\}.$$

7.4.1 Bootstrapping the sample mean

Theorem 7.9 (Bootstrapping the sample mean)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. F and that $\sigma^2 := \mathrm{Var}(X_1) < \infty$. Let $\eta(\mathbf{X}, F) := \sqrt{n}(\bar{X}_n - \mu)$, where $\mu := \mathbb{E}(X_1)$ and $\bar{X}_n := \sum_{i=1}^n X_i/n$. Then,

$$\rho(\hat{H}_n, H_n) := \sup_{x\in\mathbb{R}} |H_n(x) - \hat{H}_n(x)| \stackrel{a.s.}{\to} 0 \qquad \text{as } \ n \to \infty,$$

where $\hat{H}_n(x) \equiv H_n(x, \mathbb{F}_n)$ and $\mathbb{F}_n$ is the empirical c.d.f. of the sample $X_1, X_2, \ldots, X_n$.

Proof. For a fixed sequence $X_1, X_2, \ldots$, the variable $\bar{X}_n^*$ is the average of n observations $X_1^*, \ldots, X_n^*$ sampled from the empirical distribution. The (conditional) mean and variance of these observations are

$$\begin{aligned}
\mathbb{E}(X_i^*|\mathbb{F}_n) &= \sum_{i=1}^n \frac{1}{n} X_i = \bar{X}_n \\
\mathrm{Var}(X_i^*|\mathbb{F}_n) &= \sum_{i=1}^n \frac{1}{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2.
\end{aligned}$$

By the strong law of large numbers, the conditional variance converges to σ^2 for almost every sequence $X_1, X_2, \ldots$.

The asymptotic distribution of $\bar{X}_n^*$ can be established by the Lindeberg-Feller CLT (see Theorem 6.17). Note that as the observations $X_1^*, \ldots, X_n^*$ are sampled from a different distribution $\mathbb{F}_n$ for every n, a CLT for a triangular array is necessary. In this setup we take $k_n = n$, $Y_{n,i} = X_i^*/\sqrt{n}$. It suffices to show that, for every $\epsilon > 0$,

$$\mathbb{E}_{\mathbb{F}_n}[|X_i^*|^2 I(|X_i^*| > \epsilon\sqrt{n})] = \frac{1}{n}\sum_{i=1}^n X_i^2 I(|X_i| > \epsilon\sqrt{n}) \stackrel{a.s.}{\to} 0.$$

The left side is smaller than $n^{-1}\sum_{i=1}^n X_i^2 I(|X_i| > M)$ as soon as $\epsilon\sqrt{n} \geq M$. By the strong law of large numbers, the latter average converges to $\mathbb{E}[|X_i|^2 I(|X_i| > M)]$ for almost every sequence $X_1, X_2, \ldots$. For sufficiently large M, this expression is arbitrarily small. Conclude that the limit superior of the left side of the preceding display is smaller than any number $\delta > 0$ almost surely and hence the left side converges to zero for almost every sequence $X_1, X_2, \ldots$.

Exercise 7 (HW4): Complete the proof now. □

Remark 7.1. The proof of Theorem 7.10 shows that to prove the consistency of the bootstrap it is enough to try to understand the limiting behavior of $H_n(\cdot, P_n)$, where P_n is any sequence of distributions "converging" (in some appropriate sense) to P. Thus, quite often, showing the consistency of the bootstrap boils down to showing the weak convergence of

$\eta(\mathbf{X}_n, P_n)$ under a triangular array setup, as $\mathbf{X}_n$ is now an i.i.d. sample from P_n. For example, if the CLT plays a crucial role in proving that $H_n(\cdot, P)$ converges weakly to a limit $H(\cdot, P)$, the Lindeberg-Feller CLT theorem can be used to show that $H_n(\cdot, P_n)$ converges weakly to $H(\cdot, P)$.

Exercise 8 (HW4): What do you think would be the limiting behavior of $\sqrt{n}(\bar{X}_n^* - \mu)$, conditional on the data $\mathbf{X}$? [Hint: You may use the law of the iterated logarithm]

7.5 Second-order accuracy of the bootstrap

One philosophical question about the use of the bootstrap is whether the bootstrap has any advantages at all when a CLT is already available. To be specific, suppose that $\eta(\mathbf{X}, F) = \sqrt{n}(\bar{X}_n - \mu)$. If $\sigma^2 := \mathrm{Var}(X_1) < \infty$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \qquad \text{and} \qquad \rho(\hat{H}_n, H_n) \xrightarrow{p} 0 \qquad \text{as } n \to \infty.$$

So two competitive approximations to $H_n(x)$ are $\Phi(x/\hat{\sigma}_n)$ (where $\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$) and $\hat{H}_n \equiv H_n(x, \mathbb{F}_n)$. It turns out that, for certain types of statistics, the bootstrap approximation is (theoretically) more *accurate* than the approximation provided by the CLT. Because any normal distribution is symmetric, the CLT cannot capture information about the skewness in the finite sample distribution of $\eta(\mathbf{X}, F)$. The bootstrap approximation does so. So the bootstrap succeeds in correcting for skewness, just as an Edgeworth expansion[94] would do. This is called Edgeworth correction by the bootstrap, and the property is called *second-order accuracy of the bootstrap.*

Theorem 7.10 (Second-order accuracy)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. F and that $\sigma^2 := \mathrm{Var}(X_1) < \infty$. Let $\eta(\mathbf{X}, F) := \sqrt{n}(\bar{X}_n - \mu)/\sigma$, where $\mu := \mathbb{E}(X_1)$ and $\bar{X}_n := \sum_{i=1}^n X_i/n$. If $\mathbb{E}[|X_1|^3] < \infty$ and F is continuous, then,

$$\rho(H_n, \hat{H}_n) = o_p(n^{-1/2}) \qquad \text{as } n \to \infty, \tag{148}$$

where $\hat{H}_n(x) \equiv H_n(x; \mathbb{F}_n)$ is the c.d.f. of $\eta(\mathbf{X}^*, \mathbb{F}_n) := \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma}$ ($\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$) and $\mathbb{F}_n$ is the empirical c.d.f. of the sample $X_1, X_2, \ldots, X_n$.

Compare (148) with the usual CLT approximation of H_n which states that $\rho(H_n, \Phi) = O(n^{-1/2})$, where $\Phi(\cdot)$ is the standard normal c.d.f. Thus, the bootstrap approximation is more accurate than that by the CLT. In fact, under certain assumptions it can be shown that $\rho(H_n, \hat{H}_n) = O_p(n^{-1})$; see e.g., Hall [5].

[94] We note that $T := \sqrt{n}(\bar{X}_n - \mu)/\sigma$ admits the following Edgeworth expansion:

$$\mathbb{P}(T \le x) = \Phi(x) + \frac{p_1(x|F)}{\sqrt{n}}\phi(x) + \frac{p_2(x|F)}{n}\phi(x) + \text{ smaller order terms},$$

where $p_1(x|F)$ and $p_2(x|F)$ are polynomials in x with coefficients depending on F (here $\phi(\cdot)$ is the standard normal density function).

Remark 7.2 (Rule of thumb)**.** Let $X_1, X_2, \ldots, X_n$ are i.i.d. F and $\eta(\mathbf{X}, F)$ be a root. If $\eta(\mathbf{X}, F) \xrightarrow{d} N(0, \tau^2)$, where τ does not dependent of F, then second-order accuracy is likely. Proving it will depend on the availability of an Edgeworth expansion for $\eta(\mathbf{X}, F)$. If τ depends on F (i.e., $\tau = \tau(F)$), then the bootstrap should be just first-order accurate.

7.6 Bootstrapping regression models

Regression models are among the key ones that differ from the i.i.d. setup and are also among the most widely used. Bootstrap for regression cannot be model-free; the particular choice of the bootstrap scheme depends on whether the errors are i.i.d. or not. We will only talk about the linear model with deterministic x's and i.i.d. errors. Additional moment conditions will be necessary depending on the specific problem to which the bootstrap will be applied; see e.g., Freedman [4]. First let us introduce some notation.

We consider the model

$$y_i = \beta^\top x_i + \epsilon_i,$$

where β is a $p \times 1$ ($p < n$) vector and so is x_i, and ϵ_i's are i.i.d. F with mean 0 and variance $\sigma^2 < \infty$. Let X be the $n \times p$ design matrix with the i'th row equal to x_i and let $Y := (y_1, \ldots, y_n) \in \mathbb{R}^n$. The least squares estimator of β is defined as

$$\hat{\beta}_n := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 = (X^\top X)^{-1} X^\top Y,$$

where we assume that $(X^\top X)^{-1}$ is nonsingular. We may be interested in the sampling distribution of

$$(X^\top X)^{-1}(\hat{\beta}_n - \beta) \sim H_n(\cdot, F).$$

First observe that H_n only depends on F. The *residual bootstrap* scheme is described below. Compute the *residual* vector

$$\hat{\epsilon} = (e_1, \ldots, e_n)^\top := Y - X\hat{\beta}_n.$$

We consider the centered residuals:

$$\tilde{\epsilon}_i = y_i - x_i^\top \hat{\beta}_n - \frac{1}{n}\sum_{j=1}^n e_j, \qquad \text{for} \quad i = 1, \ldots, n.$$

The bootstrap estimator of the distribution $H_n(\cdot, F)$ is $H_n(\cdot, \tilde{F}_n)$, where $\tilde{F}_n$ is the empirical c.d.f. of $\tilde{e}_1, \ldots, \tilde{e}_n$.

We can show that, under appropriate conditions[95], by an application of the Lindeberg-Feller CLT, the above bootstrap scheme is consistent.

[95] We may assume that: (i) p is fixed (as n grows); (ii) $\frac{1}{n} X_n^\top X_n \to \Sigma$, where Σ is positive definite; (iii) $\frac{1}{\sqrt{n}}|x_{ij,n}| \to 0$ as $n \to \infty$, where $X \equiv X_n = (x_{ij,n})$.

7.7 Failure of the bootstrap

In spite of the many consistency theorems in the previous sections, there are instances where the usual (nonparametric) bootstrap based on sampling with replacement from the original data actually does not work. Typically, these are instances where the root $\eta(\mathbf{X}, F)$ fails to admit a CLT. Before seeing a few examples, we list a few situations where the ordinary bootstrap fails to estimate the c.d.f. of $\eta(\mathbf{X}, F)$ consistently:

(a) $\eta(\mathbf{X}, F) = \sqrt{n}(\bar{X}_n - \mu)$ when $\mathrm{Var}_F(X_1) = \infty$.

(b) $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$ and $\nabla g(\mu) = 0$.

(c) $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$ and g is not differentiable at μ.

(d) The underlying population F_θ is indexed by a parameter θ, and the support of F_θ depends on the value of θ.

(e) The underlying population F_θ is indexed by a parameter θ, and the true value θ_0 belongs to the *boundary* of the parameter space Θ.

Exercise 9 (HW4): Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an i.i.d. sample from F and $\sigma^2 = \mathrm{Var}_F(X_1) = 1$. Let $g(x) = |x|$ and let $\eta(\mathbf{X}, F) = \sqrt{n}(g(\bar{X}_n) - g(\mu))$. If the true value of μ is 0, then by the CLT for $\bar{X}_n$ and the continuous mapping theorem, $\eta(\mathbf{X}, F) \xrightarrow{d} |Z|$ with $Z \sim N(0, \sigma^2)$. Simulate the distribution of $\eta(\mathbf{X}^*, \mathbb{F}_n)$ and empirically compare it with that of $\eta(\mathbf{X}, F)$ (you may take F to be standard normal). Does the usual bootstrap work in this case?

Exercise 10 (HW4): Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample of size n from the uniform distribution on $[0, \theta]$, $\theta > 0$ unknown. Let $X_{(n)}$ be the maximum of the observations, and suppose that the goal is to estimate the distribution of $n(X_{(n)} - \theta)$. Let $X^*_{(n)}$ be the maximum of a sample of size n from the empirical distribution of the sample. Show that $\mathbb{P}(X^*_{(n)} = X_{(n)} | \mathbf{X}) \to 1 - e^{-1}$ as $n \to \infty$. What does this mean regarding the consistency of the empirical bootstrap estimator of the distribution of the maximum?

Remark 7.3 (Subsampling)**.** A remedy when the usual bootstrap fails is to use *subsampling*. The basic idea of *subsampling* is to approximate the sampling distribution of a statistic based on the values of the statistic computed over *smaller subsets* of the data. For example, in the case where the data are n observations that are i.i.d., a statistic is computed based on the entire data set and is recomputed over all $\binom{n}{b}$ data sets of size b ($b \leq n$). These recomputed values of the statistic are suitably normalized to approximate the true sampling distribution. Typically, if $b/n \to 0$ and $b \to \infty$ as $n \to \infty$, and there exists a limiting non-degenerate c.d.f. $H(\cdot, P)$ such that $H_n(\cdot, P)$ converges weakly to $H(\cdot, P)$ (as $n \to \infty$) the resulting subsampling approximation is consistent; see Politis et al. [10] for a detailed study of subsampling and its various applications in statistics.

8 Multiple hypothesis testing

8.1 Motivation

In the *multiple hypothesis testing*[96] problem we wish to test many hypotheses simultaneously. The null hypotheses are denoted by $H_{0,i}, i = 1, \ldots, n$, where n denotes the total number of hypotheses. Consider the following example.

Example 8.1 (Prostate cancer study)**.** DNA microarrays measure expression levels of tens of thousands of genes. The data consist of levels of mRNA, which are thought to measure how much of a protein the gene produces. A larger number implies a more active gene.

Suppose that we have n genes and data on the expression levels for each gene among healthy individuals and those with prostate cancer. In the example considered in [3], $n = 6033$ genes were measured on 50 control patients and 52 patients with prostate cancer. The data obtained are (X_{ij}) where

$$X_{ij} = \text{ gene expression level on gene } i \text{ for the } j\text{'th individual.}$$

We want to test the effect of the i'th gene, i.e.,

$$H_{0,i} : i\text{'th gene has same expression level for both control and cancer patients,}$$

whereas the alternative for the i'th hypothesis is that the gene expression for the cancer patients is more than that of the control patients. For the i'th gene, we use the following test statistic:

$$\frac{\bar{X}_{i\cdot}^P - \bar{X}_{i\cdot}^C}{\text{sd}(\ldots)} \sim t_{100}, \qquad \text{under } \; H_{0,i},$$

where $\bar{X}_{i\cdot}^P$ denotes the average expression level for the i'th gene for the 52 cancer patients and $\bar{X}_{i\cdot}^C$ denotes the corresponding value for the control patients and $\text{sd}(\ldots)$ denotes the standard error of the difference. We reject the null $H_{0,i}$ for gene i if the test statistic exceeds the critical value $t_{100}^{-1}(1-\alpha)$, for $\alpha \in (0,1)$.

Consider the above prototypical example where we test $n \approx 6000$ null hypotheses at level 0.05 (say). In general, the task is how do we detect the true non-null effects (hypotheses where the null is not true) when a majority of the null hypotheses are true? The problem is that even when none of the genes have a significant effect (i.e., all the null hypotheses are true), even then on an average we expect $6000 \times 0.05 = 300$ rejections, which is a lot of rejection for scientists to do follow-up studies.

[96] Much of the material here is taken from the lecture notes by Emmanuel Candes (Stanford U; see https://statweb.stanford.edu/~candes/teaching/stats300c/) and Rina Foygel Barber's (U Chicago) lectures on 'Topics in Selective Inference' (see https://sites.google.com/view/topics-in-selective-inference/).

Thus, a natural first question is: "Are there any significant genes"? If we reject this *global testing* problem, we may focus on each hypothesis and try to decide about accepting/rejecting the i'th hypothesis. The two main questions that we will address here are:

- *Global testing.* Here our primary interest is *not* on the individual hypotheses $H_{0,i}$, but instead on the global hypothesis $H_0 : \cap_{i=1}^n H_{0,i}$, the intersection of $H_{0,i}$'s.
- *Multiple testing.* In this scenario we are concerned with the individual hypotheses $H_{0,i}$ and want to say something about each hypothesis.

8.2 Global testing

Consider the following prototypical (Gaussian sequence model) example:

$$y_i = \mu_i + \epsilon_i, \quad \text{for } \ i = 1, \ldots, n, \tag{149}$$

where ϵ_i's are i.i.d. $N(0,1)$, the μ_i's are unknown constants and we only observe the y_i's. We want to test

$$H_{0,i} : \mu_i = 0 \qquad \text{versus} \qquad H_{1,i} : \mu_i \neq 0 \text{ (or } \mu_i > 0).$$

In global testing, the goal is to test the hypothesis:

$$H_0 : \mu_i = 0, \text{ for all } i \quad \text{(no signal)}, \qquad \text{versus} \qquad H_1 : \text{ at least one } \mu_i \text{ is non-zero.}$$

The complication is that if we do each of these tests $H_{0,i}$ at level α, and then want to combine them, the global null hypothesis H_0 might not have level α. This is the first hurdle.

Data: $p_1, p_2, \ldots, p_n$: p-values for the n hypotheses.

We will assume that under $H_{0,i}$, $p_i \sim \text{Unif}(0,1)$. (we are not assuming independence among the p_i's yet.)

8.2.1 Bonferroni procedure

Suppose that $\alpha \in (0,1)$ is given. The *Bonferroni procedure* can be described as:

- Test $H_{0,i}$ at level α/n, for all $i = 1, \ldots, n$.
- Reject the global null hypothesis H_0 if we reject $H_{0,i}$ for some i.

This can be succinctly expressed as looking at the minimum of the p-values, i.e.,

$$\text{Reject } H_0 \qquad \text{if} \qquad \min_{i=1,\ldots,n} p_i \leq \frac{\alpha}{n}.$$

Question: Is this a valid level-α test, i.e., is $P_{H_0}(\text{Type I error}) \overset{?}{\leq} \alpha$?

Answer: Yes. Observe that

$$\begin{aligned}
\mathbb{P}_{H_0}(\text{Rejecting } H_0) &= \mathbb{P}_{H_0}\left(\min_{i=1,\ldots,n} p_i \leq \alpha/n\right) \\
&= \mathbb{P}_{H_0}(\cup_{i=1}^n \{p_i \leq \alpha/n\}) \\
&\leq \sum_{i=1}^n \mathbb{P}_{H_{0,i}}(p_i \leq \alpha/n), \quad \text{(crude upper bound)} \\
&= n \cdot \alpha/n, \quad \text{since } p_i \sim \text{Unif}([0,1]) \text{ under null} \\
&= \alpha.
\end{aligned}$$

So this is a valid level-α test, whatever the p_i's are (the p_i's could be dependent).

Question: Are we being too conservative (the above is an upper bound)? As we are testing each hypothesis using a very small level α/n most of the p-values would fail to be significant. The feeling is that we need a *very strong signal* for some i to detect the global null using the Bonferroni method.

Answer: We are not doing something very crude, if all the p-values are independent. This can be seen by directly calculating the exact level of the test. If the p_i's are independent, then observe that

$$\begin{aligned}
\mathbb{P}_{H_0}\left(\min_i p_i \leq \alpha/n\right) &= 1 - \mathbb{P}_{H_0}\left(\cap_{i=1}^n \{p_i > \alpha/n\}\right) \\
&= 1 - \prod_{i=1}^n \mathbb{P}_{H_{0,i}}(p_i > \alpha/n) \quad \text{(using independence)} \\
&= 1 - \left(1 - \frac{\alpha}{n}\right)^n \xrightarrow{\text{as } n\to\infty} 1 - e^{-\alpha} \quad \approx \alpha \qquad \text{(for } \alpha \text{ small)}.
\end{aligned}$$

Thus, the Bonferroni approach is not a bad thing to do, especially when we have independent p-values[97]. Note that, under independence to obtain an exact level α test we can use the Šidák correction, i.e., instead of α/n (in the above) we can take $\tilde{\alpha} := [1 - (1-\alpha)^{1/n}]$.

8.2.2 Power of the Bonferroni procedure

Let us now focus on the power of the Bonferroni method. To discuss power we need a model for the alternative.

Question: Consider the example of the Gaussian sequence model mentioned previously. Under what scenario for the μ_i's do we expect the Bonferroni test to do well?

[97] On the other extreme, if the p-values are very dependent the Bonferroni method can be quite conservative. The Type I error of the Bonferroni method when $p_1 = \ldots = p_n$ is α/n.

Answer: If we have (a few) strong signals, then the Bonferroni procedure is good. We will try to formalize this now.

In the Gaussian sequence model the Bonferroni procedure reduces to: Reject $H_{0,i}$ ($H_{0,i}$: $\mu_i = 0$ vs. $H_{1,i} : \mu_i > 0$) if

$$y_i > z_{\alpha/n},$$

where $z_{\alpha/n}$ is the $(1-\alpha/n)$'th quantile of the standard normal distribution.

Question: How does $z_{\alpha/n}$ behave? Do we know its order (when α is fixed and n is large)?

Answer: As first approximation, $z_{\alpha/n}$ is like $\sqrt{2\log n}$ (an important number for Gaussian random variables)[98].

Fact 1. Here is a fact from extreme value theory about the order of the maximum of the ϵ_i's, i.e., $\max_{i=1,\ldots,n} \epsilon_i$:

$$\frac{\max_{i=1,\ldots,n} \epsilon_i}{\sqrt{2\log n}} \xrightarrow{\text{a.s.}} 1,$$

i.e., if we have a bunch of n independent standard normals, the maximum is like $\sqrt{2\log n}$.

The mean of $\max_{i=1,\ldots,n} \epsilon_i$ is like $\sqrt{2\log n}$ and the fluctuations around the mean is of order $O_p(1)$.

To study the power of the Bonferroni procedure, we consider the following stylistic regimes (in the following the superscript (n) is to allow the variables to vary with n):

(i) $\mu_1^{(n)} = (1+\eta)\sqrt{2\log n}$ and $\mu_2 = \ldots = \mu_n = 0$,

(ii) $\mu_1^{(n)} = (1-\eta)\sqrt{2\log n}$ and $\mu_2 = \ldots = \mu_n = 0$,

where $\eta > 0$. So, in both settings, we have one strong signal, and everything else is 0.

In case (i), the signal is slightly stronger than $\sqrt{2\log n}$; and in case (ii), the signal is slightly weaker than $\sqrt{2\log n}$. We will show that Bonferroni actually works for case (i) (by that we mean the power of the test actually goes to 1). Meanwhile, the Bonferroni procedure fails for case (ii) — the power of the test converges to α.

[98] We can bound $1-\Phi(t)$ as:

$$\frac{\phi(t)}{t}\left(1-\frac{1}{t^2}\right) \le 1-\Phi(t) \le \frac{\phi(t)}{t} \qquad \Rightarrow \qquad 1-\Phi(t) \approx \frac{\phi(t)}{t} \qquad \text{for } t \text{ large.}$$

Here is a heuristics proof of the fact that $z_{\alpha/n} \approx \sqrt{2\log n}$:

$$\begin{aligned}
1-\Phi(t) \;\approx\; \frac{\phi(t)}{t} = \frac{\alpha}{n} \quad &\Leftrightarrow \quad \frac{e^{-t^2/2}}{\sqrt{2\pi}t} = \frac{\alpha}{n} \\
&\Leftrightarrow \quad -\frac{t^2}{2} = \cancel{\log(\sqrt{2\pi}t)} + \log(\alpha/n) \quad (\text{as } \log(\sqrt{2\pi}t) \text{ is a smaller order term}) \\
&\approx \quad t^2 \;=\; -2\log(\alpha/n) = 2\log n - 2\log\alpha \;\approx\; \sqrt{2\log n}.
\end{aligned}$$

This is not only a problem with the Bonferroni procedure — it can be shown that no test can detect the signal in case (ii).

Case (i):

$$\begin{aligned}\mathbb{P}(\max y_i > z_{\alpha/n}) &= \mathbb{P}\left(\{y_1 > z_{\alpha/n}\} \cup \left\{\max_{i=2,\ldots,n} y_i > z_{\alpha/n}\right\}\right) \\ &\geq \mathbb{P}(\{y_1 > z_{\alpha/n}\}) \\ &\approx \mathbb{P}\left(\epsilon_1 > \sqrt{2\log n} - (1+\eta)\sqrt{2\log n}\right) \to 1.\end{aligned}$$

In this regime, just by looking at y_1, we will be able to detect that H_0 is not true.

Case (ii):

$$\mathbb{P}(\max y_i > z_{\alpha/n}) \leq \mathbb{P}(y_1 > z_{\alpha/n}) + \mathbb{P}\left(\max_{i=2,\ldots,n} y_i > z_{\alpha/n}\right).$$

Note that the first term is equal to $\mathbb{P}(\epsilon_1 > \eta\sqrt{2\log n}) \to 0$ as $n \to \infty$; whereas the second term converges to $1 - e^{-\alpha}$. Hence, we have shown that in this case the power of the test is less than or equal to the level of the test. So the test does as well as just plain guesswork.

This shows the dichotomy in the Bonferroni procedure; that by just changing the signal strength you can always recover or you can fail $(1-\alpha)$ of the time.

Whenever we have a hypothesis testing procedure, there has to be an effort in trying to understand the power of the procedure. And it is quite often the case that different tests (using different test statistics) are usually geared towards detecting different kinds of departures from the null. Here, the Bonferroni procedure is geared towards detecting sparse, strong signals.

8.2.3 Chi-squared test

Consider the Gaussian sequence model described in (149) and suppose that we want to test the global null hypothesis:

$$H_0 : \mu_i = 0, \text{ for all } i, \text{(no signal)} \qquad \text{versus} \qquad H_1 : \text{ at least one } \mu_i \text{ is non-zero.}$$

Letting $Y = (y_1, \ldots, y_n)$, the *chi-squared test* can be expressed as:

$$\text{Reject } H_0 \text{ if} \quad T := \|Y\|^2 > \chi^2_n(1-\alpha).$$

Note that under H_0,

$$T \sim \chi^2_n,$$

and under H_1,

$$T \sim \chi^2_n(\|\mu\|^2),$$

where $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ and $\chi^2_n(\|\mu\|^2)$ denotes the non-central χ^2_n distribution with non-centrality parameter $\|\mu\|^2$.

This test is going to have high power when $\|\mu\|^2$ is large. So, this test would have high power when there are many weak signals (even if each μ_i is slightly different from zero as we square it and add these up we can get a substantially large $\|\mu\|^2$). The Bonferroni procedure may not be able to detect a scenario like this — given α/n to each hypothesis if the signal strengths are weak all of the p-values (for the different hypotheses) might be considerably large.

Remark 8.1 (Fisher's combination test)**.** Suppose that $p_1, \ldots, p_n$ are the n p-values obtained from the n hypotheses tests. We assume that the p_i's are independent. The *Fisher's combination test* rejects the global null hypothesis if

$$T := \sum_{i=1}^{n} -2 \log p_i$$

is large. Observe that, under H_0, $T := -2\sum_{i=1}^{n} \log p_i \sim \chi^2_{2n}$. This follows from the fact that under $H_{0,i}$,

$$-\log p_i \sim \text{Exp}(1) \equiv \text{Gamma}(1, 1).$$

Again, as this test is aggregating the p-values, it will hopefully be able to detect the presence of many weak signals.

8.3 Simultaneous inference

In simultaneous inference, we have a long list of hypotheses we are interested in testing (it could be a finite list or even an infinite list) and our requirement is that with probability at least $1 - \alpha$ all of these hypotheses tests have to be simultaneously correct (i.e., the probability of making any error is at most α). We can alternatively be in the setting where our goal is to construct a CI for each parameter in a list of parameters and we would like to ensure that all the CIs cover the respective parameters simultaneously with probability at least $1 - \alpha$.

Let us illustrate this using the Gaussian sequence setting (149)[99]. We may be interested in the following questions:

(a) We could ask questions about each μ_i, e.g., can we test the hypothesis $H_{0,i} : \mu_i = 0$ simultaneously for all i. Correspondingly, we can ask if we can form confidence intervals for the μ_i's that have simultaneous coverage, i.e.,

$$\mathbb{P}\left(\mu_i \in \text{CI}_i \ \ \forall i\right) \geq 1 - \alpha$$

where CI_i is a confidence interval for μ_i, for all i.

[99] Sometimes also called as the one-way layout with Gaussian noise.

(b) We can also ask if μ_i and μ_j are the same, i.e., test $H_{0,i,j} : \mu_i = \mu_j$ simultaneously for all $i, j \in \{1, \ldots, n\}$? What can we do to obtain simultaneously valid inference for these $\binom{n}{2}$ hypotheses? Similarly, we can ask, if we can construct CIs for each of these $\binom{n}{2}$ parameters $\mu_i - \mu_j$ that are simultaneously valid.

To solve (a) we can use the Šidák correction, i.e., by taking $\tilde{\alpha} := [1 - (1 - \alpha)^{1/n}]$, we can ensure that

$$\mathbb{P}\left(\mu_i \in [y_i - z_{\tilde{\alpha}/2}, y_i + z_{\tilde{\alpha}/2}] \ \ \forall i\right) = \mathbb{P}\left(|\epsilon_i| \leq z_{\tilde{\alpha}/2} \ \ \forall i\right) = (1 - \tilde{\alpha})^n = 1 - \alpha.$$

Now, to solve (b), observe that the above display implies that

$$\mathbb{P}\left(\mu_i - \mu_j \in [y_i - y_j - 2z_{\tilde{\alpha}/2}, y_i - y_j + 2z_{\tilde{\alpha}/2}] \ \ \forall i\right) \geq 1 - \alpha.$$

This is our option (i) for solving (a). However the above approach could be conservative.

In option (ii), we can apply Bonferroni directly to the following $\binom{n}{2}$ data points:

$$(y_i - y_j) = (\mu_i - \mu_j) + (\epsilon_i - \epsilon_j),$$

where $\epsilon_i - \epsilon_j \sim N(0, 2)$. This will lead to the following CIs which are simultaneously valid with probability at least $1 - \alpha$:

$$(y_i - y_j) \pm \sqrt{2} z_{\alpha/(2\binom{n}{2})}.$$

Again, this approach is likely to be a bit conservative because of the dependence between the variables.

Yet another approach is the *Tukey's honest significant difference test* (HSD) which avoids the union bound and the overly conservative nature of the Bonferroni procedure. The idea is to look at $q_{n,\alpha}$, the $(1 - \alpha)$'th quantile of the distribution of

$$\max_{i=1,\ldots,n} \epsilon_i - \min_{i=1,\ldots,n} \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, 1)$ (the quantiles of this distribution can be easily simulated).

We can now show that

$$\mathrm{CI}_{ij} := (y_i - y_j) \pm q_{m,\alpha}$$

is a valid level α simultaneously valid confidence interval for $(\mu_i - \mu_j)$ (and unlike Bonferroni, Tukey's HSD is not conservative):

$$\begin{aligned}
\mathbb{P}\left(\mu_i - \mu_j \in \mathrm{CI}_{ij} \ \ \forall i \neq j\right) &= \mathbb{P}\left(|(\mu_i - \mu_j) - (y_i - y_j)| \leq q_{n,\alpha} \ \ \forall i \neq j\right) \\
= \ \mathbb{P}\big(|\epsilon_i - \epsilon_j| \leq q_{n,\alpha} \ \ \forall i \neq j\big) &= \mathbb{P}\Big(\max_{i,j} |\epsilon_i - \epsilon_j| \leq q_{n,\alpha}\Big) \\
= \ \mathbb{P}\Big(\max_i \epsilon_i - \min_i \epsilon_i \leq q_{n,\alpha}\Big) &= 1 - \alpha.
\end{aligned}$$

8.4 Multiple testing/comparison problem: False discovery rate

Until now, we have been considering tests of the global null $H_0 = \cap_i H_{0,i}$ or simultaneous inference. For some testing problems, however, our goal is to accept or reject each individual $H_{0,i}$, for $i \in \{1, \ldots, n\} =: [n]$. Let $\mathcal{H}_0 \subset [n]$ (here $[n]$ denotes all the hypotheses being tested) denote the true nulls with $|\mathcal{H}_0| = n_0$ and the remaining hypotheses are non-null. Let $\mathcal{R} \subset \{1, \ldots, n\}$ denote the rejections by any multiple testing procedure.

We have four types of outcomes for a multiple testing proecure:

	Accept $H_{0,i}$	Reject $H_{0,i}$	
$H_{0,i}$ true	U	V	n_0
$H_{0,i}$ false	T	S	$n - n_0$
	$n - R$	R	n

where $R =$ number of rejections (an observed random variable) and U, V, S, T are unobserved random variables. Note that

$$V = \text{ number of false discoveries.}$$

8.4.1 Family-wise error rate

Ideally, we would not like to make *false discoveries* (i.e., reject the null when the null is true). But if you are not willing to make any false discoveries, which basically translates to our threshold/cutoff being really large for each test, then we will not be able make any discoveries at all.

Traditionally, statisticians want to control the *family-wise error rate* (FWER):

$$\text{FWER } := \mathbb{P}(V \geq 1).$$

It is very easy to design a test whose FWER is controlled by a predetermined level α: reject or accept each hypothesis $H_{0,i}$ according to a test whose type I error is at most α/n. Indeed, this is the Bonferroni method. By the union bound, one then has

$$\text{FWER} = \mathbb{P}\left(\cup_{i \in \mathcal{H}_0} \{\text{Reject } H_{0,i}\}\right) \leq \sum_{i \in \mathcal{H}_0} \mathbb{P}\left(\text{Reject } H_{0,i}\right) \leq \frac{\alpha n_0}{n} \leq \alpha.$$

In modern theory of hypothesis testing, control of the FWER is considered too stringent mainly because it leads to tests that fail to reject many non-null hypotheses as well.

8.4.2 False discovery rate

The *false discovery rate* (FDR) is an error control criterion developed in the 1990s as an alternative to the FWER. When the number of tests is in the tens of thousands or even

higher, FWER control is so stringent a criterion that individual departures from the null have little chance of being detected. In such cases, it may be unreasonable to control the probability of having any false rejections. Attempting to do so would leave us with virtually no power to reject individual non-nulls.

A new point of view advanced by Benjamini and Hochberg [1] proposes controlling the (expected) proportion of errors among the rejected hypotheses. The *false discovery (or rejection) proportion* (FDP) is defined as

$$\text{FDP} := \frac{\#\text{ false discoveries}}{\#\text{ discoveries}} = \frac{|\mathcal{H}_0 \cap \mathcal{R}|}{|\mathcal{R}|} = \frac{V}{R} \qquad (\text{here } \tfrac{0}{0} \text{ is taken to be } 0). \tag{150}$$

Intuitively, FDP reflects the "cost" of following up on discoveries, i.e., if FDP ≤ 0.1, then approximately 90% of discoveries are expected to be "real". It seems perfectly reasonable to try to develop multiple testing procedures with small FDP.

But FDP is an unobserved random variable, so the criterion we propose to control is its expectation, which we refer to as the *false discovery rate* (FDR), i.e.,

$$\text{FDR} := \mathbb{E}(\text{FDP}). \tag{151}$$

Question: What if we reject all $H_{0,i}$ with $p_i \leq \alpha$ (for some fixed $\alpha \in (0,1)$)?

Answer: In this case we expect around $\alpha \cdot |\mathcal{H}_0|$ many false positives (discoveries/rejections); this will be the numerator in (150). When the number of true signals (i.e., hypotheses where the null is false) is (extremely) small, then most of our discoveries are going to be false discoveries, and thus FDP ≈ 1, which is not necessarily good for follow-up studies. However, if the number of true signals is large, then $\alpha \cdot |\mathcal{H}_0|$ many false positives is quite a reasonable number of false discoveries in the context a very large number of true discoveries (we expect R to be large here).

So until we know how many signals there are we will not know whether we have a high or low FDP. Thus, we would want to choose a p-value threshold *adaptively* to the data which may ensure that FDR is controlled.

8.4.3 Benjamini-Hochberg procedure

The Benjamini-Hochberg (BH) procedure controls FDR at any desired level (e.g., suppose we take $q = 0.2$), i.e., FDR $\leq q = 0.2$; thus out of all of the rejections we make we are willing to have 20% of them be false, on an average.

The BH procedure can be described as: suppose that $p_1, \ldots, p_n$ are the p-values from the n hypotheses tests. Let

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$$

be the sorted p-values. Let

$$i_0 := \max\left\{i \le n : p_{(i)} \le q\frac{i}{n}\right\}, \qquad 0 < q < 1.$$

We reject all hypotheses $H_{0,(i)}$ for $1 \le i \le i_0$ (reject those hypotheses with p-values from $p_{(1)}$ to $p_{(i_0)}$). Pictorially this can be easily expressed as: draw the line with slope q passing through the origin and plot the ordered p-values $\{(\frac{i}{n}, p_{(i)})\}_{i=1}^n$, and accept all the hypotheses whose p-values lie above the line after the last time it was below the line.

Another way to view the BH procedure is via the following sequential description: start with $\{i = n\}$ and keep accepting the hypothesis corresponding to $p_{(i)}$ as long as $p_{(i)} > qi/n$. As soon as $p_{(i)} \le iq/n$, stop and reject all the hypotheses corresponding to $p_{(j)}$ for $j \le i$.

We can also think of the BH procedure as finding the maximum k such that at least k many p_i's are $\le qk/n$ and then rejecting all p_i's such that $p_i \le qk/n$.

Theorem 8.2. Suppose that the p-values $p_1, \ldots, p_n$ are independent. Suppose further that the BH procedure (at level q) is used to accept/reject the hypotheses. Then

$$\text{FDR } = \mathbb{E}\left(\frac{V}{\max(R,1)}\right) = q\frac{n_0}{n}.$$

Remark 8.2. Note that the above result states that the BH procedure controls FDR for all configurations of $\{H_{0,i}\}_{i=1}^n$.

Proof. This proof can be found in the recent paper by Heesen and Janssen [6][100]. Although there are many ways to prove this result, we will use the so-called *leave-one-out* technique[101]. Another useful technique uses martingale ideas; see Storey et al. [13].

We may assume that $\mathcal{H}_0$ is nonempty for otherwise $V \equiv 0$ and there will be nothing to prove. Let $\mathbf{p} := (p_1, \ldots, p_n)$ and let $R \equiv R(\mathbf{p})$ denote the number of rejections made by the BH procedure. From the description, it should be clear that $R(\mathbf{p})$ is exactly equal to i_0. We can therefore write the FDP as

$$\text{FDP} = \frac{V}{R(\mathbf{p}) \vee 1} = \sum_{j \in \mathcal{H}_0} \frac{I\{p_j \le qR(\mathbf{p})/n\}}{R(\mathbf{p}) \vee 1}.$$

We now fix $j \in \mathcal{H}_0$ and let $\tilde{\mathbf{p}} := (p_1, \ldots, p_{j-1}, 0, p_{j+1}, \ldots, p_n)$, i.e., the j'th p-value is replaced by 0 and the rest of the p-values are unchanged. Let $R(\tilde{\mathbf{p}})$ denote the number of rejections

[100] I came to know of this proof from the blog https://statpacking.wordpress.com/

[101] The intuition behind the *leave-one-out* proof technique is as follows:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R}\right] = \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\frac{I\{p_j \le qR/n\}}{R}\right] \approx \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\frac{qR/n}{R}\right] = \sum_{j \in \mathcal{H}_0} \frac{q}{n} = q\frac{n_0}{n},$$

where the $\approx$ step may hold if R were independent of p_j (which is not true).

of the BH procedure for $\tilde{\mathbf{p}}$. It should be noted that $R(\tilde{\mathbf{p}}) \geq 1$ because of the presence of a zero p-value in $\tilde{\mathbf{p}}$. The key observation now is

$$\frac{I\left\{p_j \leq qR(\mathbf{p})/n\right\}}{R(\mathbf{p}) \vee 1} = \frac{I\left\{p_j \leq qR(\tilde{\mathbf{p}})/n\right\}}{R(\tilde{\mathbf{p}})}. \tag{152}$$

To see this, first observe that when $p_j \leq qR(\mathbf{p})/n$, then the j'th hypothesis is rejected and $R(\mathbf{p}) = R(\tilde{\mathbf{p}})$, and thus $I\left\{p_j \leq qR(\mathbf{p})/n\right\} = I\left\{p_j \leq qR(\tilde{\mathbf{p}})/n\right\}$. Suppose that the left indicator is zero and that p_j equals $p_{(k)}$ (i.e., p_j equals the k'th order statistic). Then the fact that the left indicator is zero implies that $p_j \equiv p_{(k)} > qk/n$ (because otherwise, we would reject the j'th hypothesis). By definition of the BH procedure, we must then have $p_{(i)} > qi/n$ for every $i > k$. Let us denote $\tilde{\mathbf{p}}$ by $(\tilde{p}_1, \ldots, \tilde{p}_n)$. As $\tilde{p}_{(i)} = p_{(i)}$ for $i > k$, we also have $\tilde{p}_{(i)} > qi/n$ for every $i > k$ and this implies that $R(\tilde{\mathbf{p}}) \leq k$. We therefore have $p_j = p_{(k)} > qk/n \geq qR(\tilde{\mathbf{p}})/n$ which means that the right hand side of (152) is also zero.

Using (152), we can write

$$\text{FDR} = \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\frac{I\left\{p_j \leq qR(\mathbf{p})/n\right\}}{R(\mathbf{p}) \vee 1}\right] = \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\frac{I\left\{p_j \leq qR(\tilde{\mathbf{p}})/n\right\}}{R(\tilde{\mathbf{p}})}\right].$$

The independence assumption of $p_1, \ldots, p_n$ now implies that p_j and $R(\tilde{\mathbf{p}})$ are independent. Also because p_j is uniformly distributed on $[0, 1]$ as $j \in \mathcal{H}_0$, we deduce that

$$\text{FDR} = \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{I\left\{p_j \leq qR(\tilde{\mathbf{p}})/n\right\}}{R(\tilde{\mathbf{p}})}\Big|\tilde{\mathbf{p}}\right]\right] = \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\frac{q}{n}\right] = q\frac{n_0}{n}$$

and this completes the proof. $\square$

A Appendix

A.1 Hilbert spaces

A vector space in $\mathbb{R}^n$ can be spanned by a finite set of vectors. A Hilbert space is a generalization of the notion of a Euclidean space and admit expansions like that as in a finite dimensional vector space.

Definition A.1 (Hilbert space)**.** Let $\mathcal{H}$ be a (real) vector space together with a function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ (the inner product) for which

$$\begin{aligned} \langle x, y \rangle &= \langle y, x \rangle, \qquad \forall\, x, y \in \mathcal{H} \text{ (symmetric)}, \\ \langle x, ay + bz \rangle &= a\langle x, y \rangle + b\langle x, z \rangle, \qquad \forall\, x, y, z \in \mathcal{H},\ \alpha, \beta \in \mathbb{R} \text{ (bilinear)}, \\ \langle x, x \rangle &\geq 0, \qquad x \in \mathcal{H}, \ \text{ with equality if and only if } x = 0. \end{aligned}$$

Suppose that the norm in $\mathcal{H}$ is defined by

$$\|x\| := \sqrt{\langle x, x \rangle}$$

and $\mathcal{H}$ is complete[102] in the metric $d(x, y) := \|x - y\|$. Then $\mathcal{H}$ forms a Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$.

Example A.2 (Euclidean space)**.** Let $\mathcal{H} = \mathbb{R}^m$ and $\langle x, y \rangle := \sum_{i=1}^m x_i y_i$ (where $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$); or more generally $\langle x, y \rangle = x^\top A y$ where A is a symmetric positive definite matrix.

Example A.3 (Euclidean matrices)**.** Let $\mathcal{H} = \mathbb{R}^{m \times m}$ be the set of all $m \times m$ matrices. Define $\langle x, y \rangle := \mathrm{tr}(xy^\top)$. Then $\langle \cdot, \cdot \rangle$ defines a Hilbert space over $m \times m$ matrices.

Example A.4 (L_2 space)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $L_2(\Omega, \mathcal{A}, \mu)$ be the set (of equivalence classes) of all square integrable functions with

$$\langle f, g \rangle := \int f g \, d\mu.$$

Example A.5 (Sobolev space)**.** The Sobolev space $W_m[0, 1]$ is the collection of all functions $f : [0, 1] \to \mathbb{R}$ with $m-1$ continuous derivatives, $f^{(m-1)}$ absolutely continuous, and $\|f^{(m)}\| < \infty$. With an inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle f, g \rangle := \sum_{k=0}^{m-1} f^{(k)}(0) g^{(k)}(0) + \int_0^1 f^{(m)}(x) g^{(m)}(x) dx, \qquad f, g \in W_m[0, 1], \tag{153}$$

$W_m[0, 1]$ is a Hilbert space.

Here are some properties of any Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$:

[102] A metric space $\mathcal{H}$ is said to be *complete* if every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$.

- The *Cauchy-Schwarz* inequality holds:

$$|\langle x, y\rangle| \leq \|x\|\|y\|, \qquad \forall\, x, y \in \mathcal{H}.$$

- The *Parallelogram laws* assert that

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \;\text{ and }\; \|x+y\|^2 - \|x-y\|^2 = 4\langle x, y\rangle \quad \forall\, x, y \in \mathcal{H}.$$

- (Linear functional) A function $\varphi : \mathcal{H} \to \mathbb{R}$ is said to be a *linear functional* if $\varphi(\alpha x + \beta y) = \alpha\varphi(x) + \beta\varphi(y)$ whenever $x, y \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$. For example, for a fixed $y \in \mathcal{H}$,

$$\varphi_y(x) := \langle x, y\rangle, \qquad \forall\, x \in \mathcal{H}, \tag{154}$$

  defines a continuous linear functional, a linear functional that is continuous with respect to the metric induced by the inner product.

- (Dual space) The *dual space* $\mathcal{H}^*$ (of $\mathcal{H}$) is the space of all continuous linear functions from $\mathcal{H}$ into $\mathbb{R}$. It carries a natural norm[103], defined by

$$\|\varphi\|_{\mathcal{H}^*} = \sup_{\|x\|=1, x\in\mathcal{H}} |\varphi(x)|, \qquad \varphi \in \mathcal{H}^*.$$

  This norm satisfies the parallelogram laws.

  The following result, known as the *Riesz representation theorem* ,gives a convenient description of the dual.

  Theorem A.6 (Riesz representation theorem)**.** Any continuous linear functional can be represented in the form (154) for some $y \in \mathcal{H}$ depending on the linear functional.

  Thus to every element φ of the dual $\mathcal{H}^*$ there exists one and only one $u_\varphi \in \mathcal{H}$ such that $\langle x, u_\varphi\rangle = \varphi(x)$, for all $x \in \mathcal{H}$. The inner product on the dual space $\mathcal{H}^*$ satisfies

$$\langle \varphi, \psi\rangle_{\mathcal{H}^*} := \langle u_\psi, u_\varphi\rangle_{\mathcal{H}}.$$

  So the dual space is also an inner product space. The dual space is also complete, and so it is a Hilbert space in its own right.

[103] Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces over $\mathbb{R}$. A function $T : \mathcal{X} \to \mathcal{Y}$ is called a linear operator if

$$T(cx_1 + x_2) = cT(x_1) + T(x_2), \qquad \forall\, x_1, x_2 \in \mathcal{X}, c \in \mathbb{R}.$$

The *operator norm* (or spectral norm) of T is defined as $\|T\| := \sup\{\|T(x)\| : \|x\| \leq 1\}$, and T is called *bounded* if $\|T\| < \infty$.

(a) Show that a bounded operator T is continuous: If $\|x_n - x\| \to 0$, then $\|T(x_n) - T(x)\| \to 0$.

(b) Show that a continuous linear operator T is bounded.

(c) Let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, with the usual Euclidean norms. Let A be an $n \times m$ matrix, and define a linear operator T by $T(x) = Ax$. Relate the operator norm $\|T\|$ to the eigenvalues of $A^\top A$.

- (Convex sets) Recall that a subset $\mathcal{H}_0 \subset \mathcal{H}$ is called a *linear subspace* if it is closed under addition and scalar multiplication; i.e., $\alpha x + \beta y \in \mathcal{H}_0$ whenever $x, y \in \mathcal{H}_0$ and $\alpha, \beta \in \mathbb{R}$.

  A subset $C \subset \mathcal{H}$ is said to be *convex* if it contains the line joining any two of its elements, i.e., $\alpha x + (1-\alpha)y \in C$ whenever $x, y \in C$ and $0 \le \alpha \le 1$.

  A set $C \subset \mathcal{H}$ is said to be a *cone* if $\alpha x \in C$ whenever $x \in C$ and $\alpha \ge 0$. Thus, C is a convex cone if $\alpha x + \beta y \in C$ whenever $x, y \in C$ and $0 \le \alpha, \beta < \infty$. Any linear subspace is, by definition, also a convex cone. Any ball, $B = \{x \in \mathcal{H} : \|x\| \le c\}$, $c > 0$, is a convex set, but not a convex cone.

- (Projection theorem) If $C \subset \mathcal{H}$ is a closed convex set and $z \in \mathcal{H}$, then there is a unique $x \in C$ for which
$$\|x - z\| = \inf_{z \in C} \|y - z\|.$$
  In fact, $x \in C$ satisfies the condition
$$\langle z - x, y - x \rangle \le 0, \qquad \forall\, y \in C. \tag{155}$$
  The element $x \in C$ is called the *projection* of z onto C and denoted by $\Pi_C(z)$. Prove the projection theorem.

  In particular, if C is a convex cone, setting $y = x/2$ and $y = 2x$ in (155) shows that $\langle z - x, x \rangle = 0$. Thus, x is the unique element of C for which
$$\langle z - x, x \rangle = 0 \qquad \text{and} \qquad \langle z - x, y \rangle \le 0 \quad \forall\, y \in C.$$

  If C is a linear subspace, then $z - x$ is *orthogonal* to C, i.e.,
$$\langle z - x, y \rangle = 0 \quad \forall\, y \in C.$$

- (Orthogonal complement) Suppose that $\mathcal{H}_0 \subset \mathcal{H}$. The *orthogonal complement* of $\mathcal{H}_0$ is
$$\mathcal{H}_0^\perp := \{x \in \mathcal{H} : \langle x, y \rangle = 0,\ \forall\, y \in \mathcal{H}_0\}.$$

  Result: The orthogonal complement of a subset of a Hilbert space is a closed linear subspace.

  The projection theorem states that if $C \subset \mathcal{H}$ is a closed subspace, then any $z \in C$ may be uniquely represented as $z = x + y$, where $x \in C$ is the best approximation to z, and $y \in C^\perp$.

  Result: If $C \subset \mathcal{H}$ is a closed subspace, then $\mathcal{H} = C \oplus C^\perp$, where
$$A \oplus B := \{x + y : x \in A, y \in B\}.$$
  Thus, every closed subspace C of $\mathcal{H}$ has a closed complementary subspace $C^\perp$.

- (Orthonormal basis) A collection $\{e_t : t \in T\} \subset \mathcal{H}$ (where T is any index set) is said to be *orthonormal* if $e_s \perp e_t$ (i.e., $\langle e_s, e_t \rangle = 0$) for all $s \neq t$ and $\|e_t\| = 1$, for all $t \in T$.

  As in the finite-dimensional case, we would like to represent elements in our Hilbert space as linear combinations of elements in an orthonormal collection, but extra care is necessary because some infinite linear combinations may not make sense.

  The *linear span* of $S \subset \mathcal{H}$, denoted $\text{span}(S)$, is the collection of all finite linear combinations $\alpha_1 x_1 + \cdots + \alpha_n x_n$ with $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in S$. The closure of this set is denoted by $\overline{\text{span}}(S)$.

  An orthonormal collection $\{e_t, t \in T\}$, is called an *orthonormal basis* for the Hilbert space $\mathcal{H}$ if $\langle e_t, x \rangle \neq 0$ for some $t \in T$, for every nonzero $x \in \mathcal{H}$.

  Result: *Every Hilbert space has an orthonormal basis.*

  When $\mathcal{H}$ is *separable*[104], a basis can be found by applying the Gram-Schmidt algorithm to a countable dense set, and in this case the basis will be countable.

  Result: If $\{e_n\}_{n \geq 1}$, is an orthonormal basis of $\mathcal{H}$, then each $x \in \mathcal{H}$ may be written as $x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$. Show this.

References

[1] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

[2] Dudley, R. M. (1999). *Uniform central limit theorems*, Volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

[3] Efron, B. (2010). *Large-scale inference*, Volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge. Empirical Bayes methods for estimation, testing, and prediction.

[4] Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist. 9*(6), 1218–1228.

[5] Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York.

[6] Heesen, P. and A. Janssen (2015). Inequalities for the false discovery rate (FDR) under dependence. *Electron. J. Stat. 9*(1), 679–716.

[104] A topological space is called *separable* if it contains a countable, dense subset; i.e., there exists a sequence $\{x_n\}_{n=1}^{\infty}$ of elements of the space such that every nonempty open subset of the space contains at least one element of the sequence.

[7] Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

[8] Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. Springer, New York.

[9] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist. 33*, 1065–1076.

[10] Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.

[11] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*(2), 186–199.

[12] Pollard, D. (1997). Another look at differentiability in quadratic mean. In *Festschrift for Lucien Le Cam*, pp. 305–314. Springer, New York.

[13] Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(1), 187–205.

[14] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[15] van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[16] van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.