ON HYPOTHESIS TESTING

FADOUA BALABDAOUI

Spring 2018

Notes: David Markwalder

Contents

Part 1. **Introduction and some fundamentals**

1. Posing the problem

Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\chi, \mathcal{B})$ be a random variable, $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space, $(\chi, \mathcal{B})$ a measurable space.
Result: X induces the probability measure P_X on $(\chi, \mathcal{B})$ given by $P_X(B) = \mathbb{P}(X \in B)$ for all $B \in \mathcal{B}$.

Example: Suppose $X \sim \mathcal{N}(\theta, 1)$ with $\theta \in \mathbb{R}$. Then

$$P_X(B) = \int_B \frac{1}{\sqrt{2\pi}} \exp\left(-1/2(x-\theta)^2\right) dx, \qquad \forall B \in \mathcal{B}.$$

We are going to assume that P_X belongs to some parametric family, that is, that there exists some parameter space Θ such that $P_X \in \{P_\theta : \theta \in \Theta\}$. Here, for all $\theta \in \Theta$, P_θ is a probability measure on $(\chi, \mathcal{B})$. In the previous example, $\Theta = \mathbb{R}$.

Example: $X \sim \text{Pois}(\theta)$, $\theta \in (0, +\infty)$. Then

$$P_X(B) = \sum_{x \in B} \frac{\exp(\theta)x}{x!}, \qquad \forall \theta \in 2^{\mathbb{N}}$$

the ensemble of all subsets of $\mathbb{N}$.

Problem: Let Θ_0 and Θ_1 be two subsets of Θ such that $\Theta_0 \cap \Theta_1 = \emptyset$.

Goal: We want, based on observed realisation of X_1, be able to decide between Θ_0 and Θ_1. This is a testing problem which can be formalized as follows:

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1,$$

where H_0 denotes the null- and H_1 denotes the alternative hypothesis.

Definition 1.1. ***critical function*** *We call a critical function any function Φ such that $\Phi(x) \in [0, 1]$ for all $x \in \chi$.*

Definition 1.2. ***test function*** *A test function is a critical function Φ such that for all $x \in \chi$ we either accept H_0 with probability $1 - \Phi(x)$ or we reject H_0 with probability $\Phi(x)$.*

Definition 1.3. ***type-I error, power, type-II error***

(i) for $\theta \in \Theta_0$, the function $\theta \mapsto \mathbb{E}_\theta [\Phi(X)]$ is called Type-I error.
(ii) for $\theta \in \Theta_1$, the same function is called power (usually denoted by $\beta(\theta)$)
(iii) $1 - \beta(\theta)$ is called type-II error.

Truth \Decision	Accept	Reject
Θ_0	✓	Type-I error
Θ_1	Type-II error	✓

The goal is to find a test function Φ such that $\begin{cases} \sup_{\theta \in \Theta_0} E_\theta(\Phi(X)) \leq \alpha \text{ for some given } \alpha \in (0,1) \\ \beta(\theta) \text{ is maximal } \forall \theta \in \Theta_1. \end{cases}$

Goal: Find a function Φ such that Type-I error is controlled if and only if $\sup_{\theta \in \Theta_0} E_\theta [\Phi(x)] \leq \alpha$ (for some given $\alpha \in (0, 1)$).

The power of Φ is the largest among all other testing functions $\Phi^\star(x)$ satisfying $\sup_{\theta \in \Theta_0} E_\theta [\Phi(x)] \leq \alpha$ if and only if for all $\theta \in \Theta_1$,$\beta(\theta) = E_\theta(\Phi(x)) \geq E_\theta(\Phi^\star(x)) = \beta^\star(\theta)$.

Definition 1.4. *We say that H_0 or H_1 is*

(i) simple if $\Theta_0 = \{\theta_0\}$ or $\Theta_1 = \{\theta_1\}$.
(ii) composite if $card(\Theta_0) > 1$ or $card(\Theta_1) > 1$.

Example: $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1$

$$\theta_0 \neq \theta_1$$

then we are testing a simple hypothesis against a simple hypothesis.

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta \geq \theta_1$$

2. The fundamental lemma on hypothesis testing

Definition 2.1. ***UMP*** *A test Φ is is said to be uniformly most powerful of level α (UMP of level α) if $\sup_{\theta \in \Theta_0} E_\theta [\Phi(X)] \leq \alpha$ and for any other test $\Phi^\star$ such that $\sup_{\theta \in \Theta_0} E_\theta [\Phi^\star(X)] \leq \alpha$ we have*

$$E_\theta \left[\Phi^\star(X)\right] \leq E_\theta [\Phi(X)]$$

for all $\theta \in \Theta_1$.

Theorem 2.2. ***Neyman-Pearson-Lemma*** *Let P_0 and P_1 be two probability measures on $(\chi, \mathcal{B})$ such that P_0 and P_1 admit densities p_0 and p_1 with respect to some $\sigma-$finite measure μ. Let $\alpha \in (0,1)$ and consider the problem $H_0 : p = p_0$ vs. $H_1 : p = p_1$.*

(i) There exists $k_\alpha \in (0,\infty)$ such that the test

$$\Phi(x) := \begin{cases} 1 & \text{if } p_1(x) > k_\alpha p_0(x) \\ 0 & \text{if } p_1(x) < k_\alpha p_0(x) \end{cases} \tag{1}$$

satisfies $E_{p_0}[\Phi(x)] = \alpha$ and Φ is UMP of level α (existence).

(ii) If Φ is a UMP test of level α (for the same problem), then it must be given by (1) μ-a.e. (uniqueness).

Lemma 2.3. *Let f be some measurable function on $(\chi, \mathcal{B})$ such that $f(x) > 0$ for all $x \in S$ (s is a set $\in \mathcal{B}$). Also let μ be some σ-finite measure on $(\chi, \mathcal{B})$. Then $\int_S f d\mu = 0 \Rightarrow \mu(S) = 0$.*

Proof. Define $S_n := \{x \in S : f(x) \geq 1/n\}$, $n > 0$. By definition of S ($f(x) > 0$ for all $x \in S$), we have $S \subset \cup_{n>0} S_n$. But, using the properties of measures we see that $\mu(S) \leq \sum_{n>0} \mu(S_n)$. But $\mu(S_n) \leq n \int_{S_n} f d\mu$ because $f \geq \frac{1}{n} m S_n$ which implies $\int_{S_n} f d\mu \geq \frac{1}{n}\mu(S)$. So

$$S_n \subset S \Rightarrow \int_{S_n} f d\mu \leq \int_S f d\mu = 0$$

by assumption. We conclude that $\mu(S) \leq 0$ if and only if $\mu(S) = 0$. □

Proof. We first show i) (existence) Consider the random variable $Y = \frac{p_1(x)}{p_0(x)}$ which, under H_0 is almost surely defined and we have $P_0(p_0(x) = 0) = \int_\chi \mathbb{1}_{\{p_0(x)=0\}} p_0(x) d\mu(x)$. Let F_0 be the cdf of Y under $H_0 : p = p_0$ and let $k_\alpha = \inf\{y : F_0(y) \geq 1 - \alpha\}$ be the $(1-\alpha)$ quantile of F_0. Let us consider the following test function

$$\Phi(x) := \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > k_\alpha \\ \gamma_\alpha & \text{if } \frac{p_1(x)}{p_0(x)} = k_\alpha \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < k_\alpha \end{cases}$$

such that γ_α satisfies $E_{p_0}[\Phi(x)] = \alpha$. This means that

$$1 \cdot P_{p_0}\left(\frac{p_1(x)}{p_0(x)} > k_\alpha\right) + \gamma_\alpha \cdot P_{p_0}\left(\frac{p_1(x)}{p_0(x)} = k_\alpha\right) + 0 \cdot P_{p_0}\left(\frac{p_1(x)}{p_0(x)} < k_\alpha\right) = \alpha$$

or equivalently

$$1 - F_0(k_\alpha) + \gamma_\alpha \left(F_0(k_\alpha) - F_0(k_\alpha-)\right) = \alpha.$$

Now define

$$\gamma_\alpha := \begin{cases} \frac{\alpha - (1 - F_0(k_\alpha))}{F_0(k_\alpha) - F_0(k_\alpha-)} & \text{if } F_0(k_\alpha) > F_0(k_\alpha-) \\ 0 & \text{if } F_0 \text{ is continuous in } k_\alpha. \end{cases}$$

Now we show that Φ is UMP among all tests of level α. Take another test $\Phi^\star$ such that $E_{p_0}[\Phi^\star(x)] \leq \alpha$. The goal is to show that $E_{p_1}[\Phi(x)] \geq E_{p_1}[\Phi^\star(x)]$.

$$\int_\chi \left(\Phi(x) - \Phi^\star(x)\right)(p_1(x) - k_\alpha p_0(x)) d\mu(x) =$$

$$= \int_L \left(\Phi(x) - \Phi^\star(x)\right)(p_1(x) - k_\alpha p_0(x)) d\mu(x) + \int_M \left(\Phi(x) - \Phi^\star(x)\right)(p_1(x) - k_\alpha p_0(x)) d\mu(x)$$

$$= \int_L \underbrace{\left(1 - \Phi^\star(x)\right)}_{\geq 0} \underbrace{(p_1(x) - k_\alpha p_0(x))}_{>0} d\mu(x) + \int_M \underbrace{\left(-\Phi^\star(x)\right)(p_1(x) - k_\alpha p_0(x))}_{\geq 0} d\mu(x) \geq 0,$$

where $L := \{x : p_1(x) > k_\alpha p_0(x)\}$ and $M := \{x : p_1(x) < k_\alpha p_0(x)\}$. Hence, $\int_\chi (\Phi(x) - \Phi^\star(x))(p_1(x) - k_\alpha p_0(x)) d\mu(x) \geq 0$ and thus we have

$$E_{p_1}[\Phi(x)] - E_{p_1}\left[\Phi^\star(x)\right] \geq k_\alpha \left(E_{p_0}[\Phi(x)] - E_{p_0}\left[\Phi^\star(x)\right]\right) = k_\alpha(\underbrace{\alpha - E_{p_0}\left[\Phi^\star(x)\right]}_{\geq 0}).$$

Therefore $E_{p_1}[\Phi(x)] \geq E_{p_1}[\Phi^\star(x)]$.
We now show ii) (uniqueness). Take another test $\Phi^\star$ of level α ($E_{p_0}[\Phi^\star(x)] \leq \alpha$) and such that $\Phi^\star$ is UMP among all

tests of level α. Let us consider the following set $S = \{x \in \chi : \Phi^\star(x) \neq \Phi(x)\} \cap \{x \in \chi : p_1(x) \neq k_\alpha p_0(x)\}$. We want to show that $\mu(S) = 0$. Assume $\mu(S) > 0$. Consider $f(x) = (\Phi(x) - \Phi^\star(x))(p_1(x) - k_\alpha p_0(x))$, $x \in \chi$. Note that $f(x) > 0$ for all $x \in S$. Using lemma we conclude that $\int_S f(x)d\mu(x) > 0$. Now,

$$\int_\chi f(x)d\mu(x) = \int_S f(x)d\mu(x) + \int_{S^c} f(x)d\mu(x)$$

where $f(x) = 0$ on S^c. This implies that

$$\begin{aligned} 0 < \int_\chi f(x)d\mu(x) &= \int_\chi \Big(\Phi(x) - \Phi^\star(x)\Big)(p_1(x) - k_\alpha p_0(x))d\mu(x) \\ &= \Big(E_{p_1}[\Phi(x)] - E_{p_1}[\Phi^\star(x)]\Big) - k_\alpha\Big(\alpha - E_{p_0}[\Phi^\star(x)]\Big) \end{aligned}$$

which means that $E_{p_1}[\Phi(x)] - E_{p_1}[\Phi^\star(x)] > k_\alpha(\alpha - E_{p_0}[\Phi^\star(x)) \geq 0$ It follows that $E_{p_1}[\Phi(x)] > E_{p_1}[\Phi^\star(x)]$ but this is impossible since by assumption $\Phi^\star$ is UMP. We conclude that $\mu(S) = 0$ and that $\mu-$a.e.

$$\Phi^\star(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > k_\alpha \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < k_\alpha. \end{cases}$$

□

Corollary 2.4. *Let $\alpha \in (0,1)$ and $\beta = E_{p_1}[\Phi(x)]$, the power of the Neyman-Pearson test of level α. Then $\alpha \leq \beta$ (we say that Φ is unbiased).*

Proof. Consider the constant test $\Phi^\star(x) = \alpha$ for all $x \in \chi$. $\Phi^\star$ is a test of level α and hence

$$\beta = E_{p_1}[\Phi(x)] \geq E_{p_1}[\Phi^\star(x)] = \alpha \Leftrightarrow \alpha \leq \beta.$$

□

Remark: We can even show that $\alpha < \beta$ (Φ is strictly unbiased).

Remark: The arguments used to prove the Neyman-Pearson lemma can be used to show that for any pair $(k, \gamma) \in (0, \infty) \times [0, 1]$, the test

$$\Phi(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > k \\ \gamma & \text{if } \frac{p_1(x)}{p_0(x)} = k \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < k \end{cases} \tag{2}$$

is UMP of level $E_{p_0}[\Phi(x)] = P_{p_0}\left(\frac{p_1(x)}{p_0(x)} > k\right) + \gamma P_{p_0}\left(\frac{p_1(x)}{p_0(x)} = k\right)$.

Example: (Quality control) We have a batch of items whose (unknown) proportion of defectiveness is $\theta \in (0, 1)$. To perform a quality control, n items are sampled from this batch to check whether they are defective or not. We want to test $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1$, $(\theta_1 > \theta_0)$ at some level $\alpha \in (0, 1)$. For $i \in \{1, \ldots, n\}$ define the random variable $X_i := \begin{cases} 1 & \text{if the i-th sampled item is defective} \\ 0 & \text{otherwise.} \end{cases}$

We have a random sample $(X_1, \ldots, X_n)$ of iid Ber(θ), i.e. $\chi = \{0, 1\}^n = \{0, 1\} \times \cdots \times \{0, 1\}$. We want to apply the Neyman-Pearson lemma to this testing problem. The joint density of $(X_1, \ldots, X_n)$ is

$$\begin{aligned} p_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

Under H_0 we have

$$\begin{aligned} p_{\theta_0}(x_1, \ldots, x_n) &= \theta_0^{\sum_{i=1}^n x_i}(1-\theta_0)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{\theta_0}{1-\theta_0}\right)^{\sum_{i=1}^n x_i}(1-\theta_0)^n, \end{aligned}$$

and under H_1 we have

$$\begin{aligned} p_{\theta_1}(x_1, \ldots, x_n) &= \theta_1^{\sum_{i=1}^n x_i}(1-\theta_1)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{\theta_1}{1-\theta_1}\right)^{\sum_{i=1}^n x_i}(1-\theta_1)^n. \end{aligned}$$

By applying the Neyman-Pearson lemma we know that the test Φ given by

$$\Phi(x_1, \ldots, x_n) := \begin{cases} 1 & \text{if } \left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]^{\sum_{i=1}^n x_i}\left(\frac{1-\theta_1}{1-\theta_0}\right)^n > k_\alpha \\ \gamma_\alpha & \text{if } \left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]^{\sum_{i=1}^n x_i}\left(\frac{1-\theta_1}{1-\theta_0}\right)^n = k_\alpha \\ 0 & \text{if } \left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]^{\sum_{i=1}^n x_i}\left(\frac{1-\theta_1}{1-\theta_0}\right)^n < k_\alpha. \end{cases}$$

Such that γ_α satisfies $E_{\theta_0}[\Phi(X_1, \ldots, X_n)] = \alpha$. Note that $\frac{\theta_1}{\theta_0} > 1$ implies $\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} > 1$ which means that the function $t \mapsto \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^t\left(\frac{1-\theta_0}{1-\theta_1}\right)^n$ is strictly increasing and continuous. Then the test Φ can also be rewritten as

$$\Phi(x_1, \ldots, x_n) := \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i > t_\alpha \\ \gamma_\alpha & \text{if } \sum_{i=1}^n x_i = t_\alpha \\ 0 & \text{if } \sum_{i=1}^n x_i < t_\alpha \end{cases}$$

where t_α is the $(1-\alpha)$-quantile of $\sum_{i=1}^n X_i$ under H_0 and γ_α satisfies $E_{\theta_0}[\Phi(x)] = \alpha$. Note that $\sum_{i=1}^n X_i \sim \text{Bin}(n, \theta_0)$ under H_0. Let F_{θ_0} be the cdf of $\text{Bin}(n, \theta_0)$:

$$F_{\theta_0}(y) := \begin{cases} 0 & \text{if } y < 0 \\ (1-\theta_0)^n & \text{if } 0 \le y < 1 \\ (1-\theta_0)^n + n\theta_0(1-\theta_0)^{n-1} & \text{if } 1 \le y < 2 \\ \vdots & \vdots \\ \sum_{j=0}^{n-1}\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j} & \text{if } n-1 \le y < n \\ 1 & \text{if } y \ge n. \end{cases}$$

$$\begin{aligned} \gamma_\alpha &= \frac{F_{\theta_0}(k_\alpha) - (1-\alpha)}{F_{\theta_0}(k_\alpha) - F_{\theta_0}(k_\alpha -)} \\ &= \frac{\sum_{j=0}^{k_\alpha}\binom{n}{j}\theta_0^j(1-\theta_0)^{n-j} - (1-\alpha)}{\binom{n}{k_\alpha}\theta_0^{k_\alpha}(1-\theta_0)^{n-k_\alpha}}. \end{aligned}$$

Graphical illustration:

A numerical illustration: $\theta_0 = 0.2$ and $\theta_1 = 0.4$

α	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.05	4	7	10	12	15
0.01	5	8	11	14	17

Values of t_α as a function of α and n.

$H_0 : \theta = 0.2$ vs. $H_1 : \theta = 0.4$

α	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
0.05	0.41	0.63	0.78	0.88	0.93
0.01	0.19	0.40	0.57	0.70	0.80

Power of Φ as a function of n and α. $E_{\theta_1}[\Phi(X_1,\dots,X_n)] = P_{\theta_1}\left(\sum_{i=1}^n X_i > t_\alpha\right) + \gamma_\alpha P_{\theta_1}\left(\sum_{i=1}^n X_i = t_\alpha\right)$.

3. Composite hypothese for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$

3.1. **Karlin-Rubin Theorem.** We will start this section with two examples.

Example 1: (Number of e-mails) The total number of e-mails that I received over a period of two weeks is

$$1, 0, 10, 11, 7, 8, 2, 0, 3, 7, 9, 13, 6, 5, 0.$$

Let X_i denote the number of daily e-mails received at day i, and denote by $\theta = E[X]$. Is it true that $\theta > 5$?

Example 2: (Airplane noise) The law requires that the noise caused by airplanes take-off should not exceed a certain threshold μ_0. From a sample of size n the noise intensity of airplanes was recorded. We want to test $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$, where μ is the true expectation of noise intensity.

Definition 3.1. *MLR Consider the parametric model $\{p_\theta : \theta \in \Theta\}$ and let $\Theta \subseteq \mathbb{R}$ be a parametric family of densities defined on $(\chi, \mathcal{B})$. This family is said to have a monotone likelihood ratio (MLR) if there exists a statistic T, and for any parameters $\theta_1 < \theta_2$ there exists a continuous and strictly increasing function g such that $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = g(T(x))$ for all $x \in \chi$ such that $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \in (0, +\infty)$.*

Remark: Note that g can depend on θ_1 or θ_2.

Example: (Quality Control with one sample) Let $X \sim \text{Bin}(n, \theta)$, $\theta \in \Theta = (0, 1)$. For $\theta_1 < \theta_2$, we have

$$\begin{aligned}\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} &= \frac{C_n^x \theta_2^x (1-\theta_2)^{n-x}}{C_n^x \theta_1^x (1-\theta_1)^{n-x}}\\ &= \left(\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right)^x \left(\frac{1-\theta_2}{1-\theta_1}\right)^n\end{aligned}$$

for $x \in \chi = \{1, \dots, n\}$. Put $T(x) = x$ and $g(t) = \left(\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right)^t \left(\frac{1-\theta_2}{1-\theta_1}\right)^n$. Note that $g(t)$ is continuous strictly increasing since $\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} > 1$.

Example: (Airplane noise with one sample) Suppose $X \sim \mathcal{N}(\mu, \sigma_0^2)$, σ_0^2 known and $\mu \in \Theta = \mathbb{R}$. We know that $p_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(x-\mu)^2\right)$. Let $\mu_1 \le \mu_2$:

$$\begin{aligned}\frac{p_{\mu_2}(x)}{p_{\mu_1}(x)} &= \exp\left\{-\frac{1}{2\sigma_0^2}\left((x-\mu_2)^2 - (x-\mu_1)^2\right)\right\}\\ &= \exp\left\{-\frac{1}{2\sigma_0^2}\left(x^2 - 2\mu_2 x + \mu_2^2 - x^2 + 2x\mu_1 - \mu_1^2\right)\right\}\\ &= \exp\left\{-\frac{1}{2\sigma_0^2}\left(2x(\mu_1-\mu_2) + \mu_2^2 - \mu_1^2\right)\right\}\\ &= \exp\left\{\frac{x(\mu_2-\mu_1)}{\sigma_0^2} - \frac{\mu_2^2-\mu_1^2}{2\sigma_0^2}\right\}\end{aligned}$$

Put $T(x) = x$ and $g(t) = \exp\left(\frac{t(\mu_2-\mu_1)}{\sigma_0^2} - \frac{\mu_2^2-\mu_1^2}{2\sigma_0^2}\right)$. Note that $g(t)$ is continuous and strictly increasing.

Theorem 3.2. ***Karlin-Rubin*** *Consider the testing problem $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ and fix $\alpha \in (0, 1)$. Suppose that $\{p_\theta : \theta \in \Theta\}$ admits the MLR property and let us denote by F_{θ_0} the cdf of $T(x)$ under $\theta = \theta_0$.*

(i) Then the test Φ given by $\Phi(x) = \begin{cases} 1 & \text{if } T(x) > t_\alpha \\ \gamma_\alpha & \text{if } T(x) = t_\alpha \\ 0 & \text{if } T(x) < t_\alpha, \end{cases}$
whereas t_α is the $(1-\alpha)-$ quantile of F_{θ_0} and γ_α satisfies

$$E_{\theta_0}[\Phi(X)] = P_{\theta_0}(T(X) > t_\alpha) + \gamma_\alpha P_{\theta_0}(T(X) = t_\alpha)) + 0P_{\theta_0}(T(X) < t_\alpha) = \alpha$$

is UMP of level α.

(ii) The function $\theta \mapsto E_\theta[\Phi(X)]$ is non-decreasing.

(iii) For all θ', the same test Φ is UMP for testing $H_0' : \theta \le \theta'$ versus $H_1' : \theta > \theta'$ at level $\alpha' = E_{\theta'}[\Phi(X)]$.

(iv) For any $\theta < \theta_0$, the same test Φ minimizes $E_\theta[\Phi(X)]$ among all tests $\Phi^\star$ satisfying $E_{\theta_0}[\Phi^\star(X)] = \alpha$.

Proof. *i*) and *ii*) Consider first the testing problem $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ with $\theta_1 > \theta_0$. By the Neyman-Pearson lemma, we know that the test

$$\Phi(x) := \begin{cases} 1 & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k_\alpha \\ \gamma_\alpha & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = k_\alpha \\ 0 & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k_\alpha, \end{cases}$$

where k_α is the $(1-\alpha)$ quantile of $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}$ under θ_0 and γ_α is such that $E_{\theta_0}[\Phi(X)] = \alpha$, is UMP of level α. But $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = g(T(x))$ is continuous and strictly increasing. Hence Φ can be rewritten as

$$\Phi(x) := \begin{cases} 1 & \text{if } T(x) > t_\alpha \\ \gamma_\alpha & \text{if } T(x) = t_\alpha \\ 0 & \text{if } T(x) < t_\alpha \end{cases}$$

with $t_\alpha = g^{-1}(k_\alpha)$, which is the $(1-\alpha)$–quantile of $T(x)$ under θ_0, and γ_α satisfies $E_{\theta_0}[\Phi(X)] = \alpha$. Since Φ does not involve θ_1, we conclude that Φ must be UMP of level α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.
Let us now show *ii*). Pick arbitrary θ' and θ'' such that $\theta' < \theta''$. The test Φ is the test you get for the hypothesis $H' : \theta = \theta'$ versus $H'' : \theta = \theta''$ by applying the Neyman-Pearson lemma and thus $\frac{p_{\theta''}(x)}{p_{\theta'}(x)} = \tilde{g}(T(x))$ where $\tilde{g}$ is continuous and strictly increasing (and may depend θ' and θ''). This implies that

$$\Phi(x) := \begin{cases} 1 & \text{if } \frac{p_{\theta''}(x)}{p_{\theta'}(x)} > k'_\alpha \\ \gamma_\alpha & \text{if } \frac{p_{\theta''}(x)}{p_{\theta'}(x)} = k'_\alpha \\ 0 & \text{if } \frac{p_{\theta''}(x)}{p_{\theta'}(x)} < k'_\alpha, \end{cases}$$

Furthermore, using the remark after the proof of the Neyman-Pearson lemma, we conclude that Φ must be UMP of level $\alpha' = E_{\theta'}[\Phi(X)]$. Using Corollary 2.1, we have that

$$\alpha' \le E_{\theta'}[\Phi(X)] \Leftrightarrow E_{\theta'}[\Phi(X)] \le E_{\theta''}[\Phi(X)]$$

(we say that Φ is unbiased). Since θ' and θ'' were chosen arbitrarily it follows that $\theta \mapsto E_\theta[\Phi(X)]$ is non-decreasing. This in turn implies that the supremum is admitted at θ_0 i.e. $\sup_{\theta \le \theta_0} E_\theta[\Phi(X)] = E_{\theta_0}[\Phi(X)] = \alpha$ (recall that the level of a test Φ for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ is $\sup_{\theta \in \Theta_0} E_\theta[\Phi(X)]$). This concludes the proof that Φ is UMP of level α for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta_0$.
iv) Fix $\theta < \theta_0$. By the MLR property, we know that there exists a strictly increasing and continuous function g such that $\frac{p_{\theta_0}(x)}{p_\theta(x)} = g(T(x))$. Thus the Karlin-Rubin test can be also given by

$$\Phi(x) := \begin{cases} 1 & \text{if } \frac{p_{\theta_0}(x)}{p_\theta(x)} > k_\alpha \\ \gamma_\alpha & \text{if } \frac{p_{\theta_0}(x)}{p_\theta(x)} = k_\alpha \\ 0 & \text{if } \frac{p_{\theta_0}(x)}{p_\theta(x)} < k_\alpha, \end{cases}$$

where k_α is linked to t_α through $k_\alpha = g(t_\alpha)$. Now

$$\int \Big(\Phi(x) - \Phi^\star(x)\Big) \left(p_{\theta_0}(x) - k_\alpha p_\theta(x)\right) d\mu(x) \ge 0$$

for any test $\Phi^\star$. Thus, $E_{\theta_0}(\Phi(X)) - E_{\theta_0}(\Phi^\star(X)) \ge k_\alpha \left(E_\theta(\Phi(X)) - E_\theta(\Phi^\star(X))\right)$ and $E_{\theta_0}(\Phi(X)) - E_{\theta_0}(\Phi^\star(X)) = 0$ if $E_{\theta_0}(\Phi^\star(X)) = 0$. Thus $E_\theta(\Phi(X)) \le E_\theta(\Phi^\star(X)))$. □

Corollary 3.3. ***application to exponential families*** *Suppose that* $p_\theta(x) = c(\theta)h(x)\exp(Q(\theta)T(x))$ *with* $\theta \in \Theta \subseteq \mathbb{R}$ *(one dimensional parameter space). If* $\theta \mapsto Q(\theta)$ *is continuous and strictly increasing, then* $\{p_\theta : \theta \in \Theta\}$ *admits the MLR property.*

We now go back to the introductory examples.

Example 1: (Number of e-mails) We want to test $H_0 : \theta \leq 5$ versus $H_1 : \theta > 5$. Here we assume that $X_1, \ldots, X_n \overset{iid}{\sim}$ Pois(θ) with $n = 15$. Hence we have density $p_\theta(x) = \frac{e^{-\theta}\theta^x}{x!}$, $x \in \{1, 2, \ldots\}$. The joint density of $(X_1, \ldots, X_n)$ is

$$\prod_{i=1}^{n} p_\theta(x_i) = \frac{e^{-n\theta}}{\prod_{i=1}^n x_i!}\theta^{\sum_{i=1}^n x_i} = \frac{e^{-n\theta}}{\prod_{i=1}^n x_i!}\exp\left(\log(\theta)\sum_{i=1}^{n} x_i\right) = c(\theta)h(x_1, \ldots, x_n)\exp(Q(\theta)T(x_1, \ldots x_n))$$

with $Q(\theta) = \log(\theta)$, $\theta \in \Theta$ and $T(x_1, \ldots x_n) = \sum_{i=1}^n x_i$. Hence at a given level α

$$\Phi(x) := \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i > t_\alpha \\ \gamma_\alpha & \text{if } \sum_{i=1}^n x_i = t_\alpha \\ 0 & \text{if } \sum_{i=1}^n x_i < t_\alpha, \end{cases}$$

with t_α being the $(1-\alpha)-$quantile of $\sum_{i=1}^n x_i$ under $\theta = \theta_0 = 5$ and γ_α such that $E_{\theta_0}[\Phi(x)] = \alpha$, is UMP at level α. We know that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\theta_0)$, then $\sum_{i=1}^n X_i \overset{\text{iid}}{\sim} \text{Pois}(n\theta_0)$. t_α is the $(1-\alpha)-$quantile of $\text{Pois}(n\theta_0) \overset{n=15,\theta_0=5,\alpha=0.05}{=} 90$. $\gamma_\alpha = \frac{F_{n\theta_0}(t_\alpha)-(1-\alpha)}{P_{n\theta_0}\left(\sum_{i=1}^{15} X_i = t_\alpha\right)} = \frac{0.960076-0.95}{0.0102} \approx 0.98$.

$$\Phi(x_1, \ldots, x_{15}) := \begin{cases} 1 & \text{if } \sum_{i=1}^{15} x_i > 90 \\ 0.98 & \text{if } \sum_{i=1}^{15} x_i = 90 \\ 0 & \text{if } \sum_{i=1}^{15} x_i < 90, \end{cases}$$

We have that $\sum_{i=1}^{15} X_i = 82$ and thus we accept $H_0 : \theta \leq 5$.

Example 2: (Take-off noise) If we assume that the noise intensity follows $\mathcal{N}(\mu, \sigma_0^2)$, $\sigma_0 > 0$ known, then

$$\begin{aligned} p_\mu(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2\sigma_0^2}(x_i-\mu)^2\right) \\ &= \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}\exp\left(-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right) \\ &= \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}\exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2} + \frac{\mu}{\sigma_0^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma_0^2}\right) \\ &= \underbrace{\frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}\exp\left(-\frac{n\mu^2}{2\sigma_0^2}\right)}_{c(\mu)}\underbrace{\exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right)}_{h(x_1,\ldots,x_n)}\exp(Q(\mu)T(x_1, \ldots, x_n)) \end{aligned}$$

with $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$, $Q(\mu) = \frac{\mu}{\sigma_0^2}$ continuous and strictly increasing. A UMP test of level α for testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ is given by

$$\Phi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i > t_\alpha \\ 0 & \text{if } \sum_{i=1}^n x_i \leq t_\alpha \end{cases}$$

with $E_{\mu_0}[\Phi(X_1, \ldots X_n)] = \alpha$ if and only if $P_{\mu_0}\left(\sum_{i=1}^n X_i > t_\alpha\right) = \alpha$.

$$
\begin{aligned}
P_{\mu_0}\left(\sum_{i=1}^{n} X_i > t_\alpha\right) = \alpha &\Leftrightarrow P_{\mu_0}\left(\overline{X_n} > t_\alpha/n\right) = \alpha \\
&\Leftrightarrow P_{\mu_0}\left(\overline{X_n} - \mu_0 > t_\alpha/n - \mu_0\right) = \alpha \\
&\Leftrightarrow P_{\mu_0}\left(\frac{\overline{X_n} - \mu_0}{\sqrt{\sigma_0^2/n}} > \frac{t_\alpha/n - \mu_0}{\sqrt{\sigma_0^2/n}}\right) = \alpha \\
&\Leftrightarrow P\left(Z > \frac{t_\alpha/n - \mu_0}{\sqrt{\sigma_0^2/n}}\right) = \alpha
\end{aligned}
$$

where $Z \sim \mathcal{N}(0,1)$. Hence $\frac{\sqrt{n}(t_\alpha/n-\mu_0)}{\sigma_0} = \zeta_\alpha$ the $(1-\alpha)-$quantile of $\mathcal{N}(0,1)$.

$$
\Phi(x_1,\ldots,x_n) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}(\overline{x}_n-\mu_0)}{\sigma_0} > \zeta_\alpha \\ 0 & \text{otherwise.} \end{cases}
$$

Now chose $\alpha = 0.05$ then (you can compute with software) $\zeta_\alpha \approx 1.64$. Let $n = 100, \sigma_0 = 7$ and $\mu_0 = 78$. Then, again using software, we compute $\mu_0 + \frac{\sigma_0}{\sqrt{n}}\zeta_\alpha \approx 79.15$. We observe $\overline{x}_n = 82 > 79.15$ and hence decide to reject H_0.

Remark:

As $n \to \infty$, the power of Φ increases to 1 for any fixed alternative. Indeed let $\mu \in \Theta_1 = (\mu_0, +\infty)$

$$
\begin{aligned}
\beta(\mu) &= E_\mu[\Phi(X_1,\ldots,X_n)] \\
&= P_\mu(\overline{X}_n > \mu_0 + \frac{\sigma_0}{\sqrt{n}}\zeta_\alpha) \\
&= P_\mu\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma_0} > \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} + \zeta_\alpha\right) \\
&= P\left(Z > \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} + \zeta_\alpha\right) \\
&= 1 - P\left(Z \leq \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} + \zeta_\alpha\right) \\
&= 1 - F_Z\left(-\frac{\sqrt{n}(\mu - \mu_0)}{\sigma_0} + \zeta_\alpha\right).
\end{aligned}
$$

But since $\lim_{n\to\infty} -\frac{\sqrt{n}(\mu-\mu_0)}{\sigma_0} + \zeta_\alpha = -\infty$ we conclude that $\lim_{n\to\infty} 1 - F_Z\left(-\frac{\sqrt{n}(\mu-\mu_0)}{\sigma_0} + \zeta_\alpha\right) = 1$. We say that the test Φ is consistent.

4. P-Values

Suppose we have an observation θ and want to make a decision whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. To do so we use a statistical procedure (a test) which we either accept or reject. Let us revisit Example 2 and suppose that we observed a mean $\overline{x}_n = 100$. This would not change our initial decision of rejecting H_0 but this somehow looks 'more convincing' or may seem like we have 'more' evidence against $H_0 : \mu \leq \mu_0$. This leads to the notion of p-values. Assume we are in a simple setting: $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$ (which may be composite but $\theta_0 \notin \Theta_1$). Consider a test function $\Phi(x) = \begin{cases} 1 & \text{if } T(x) > t_\alpha \\ 0 & \text{otherwise,} \end{cases}$ where t_α denotes the $(1-\alpha)$-quantile of $T(X)$ under $H_0 : \theta = \theta_0$. Assume that F_{θ_0}, the cdf of $T(X)$ under $\theta = \theta_0$, is continuous and strictly increasing, that is bijective.

Definition 4.1. ***p-value*** *Let $\mathcal{R}_\alpha = \{x' \in \chi : T(x') > t_\alpha\}$ be a rejection region for some fixed α. We define the p-value of an observation $x \in \chi$ with respect to Φ by $p_\Phi(x) = \inf\{\alpha : x \in \mathcal{R}_\alpha\}$.*

Lemma 4.2. *For the test Φ given above, it holds that $p_\Phi(x) = P_{\theta_0}(T(X) \geq T(x))$.*

Proof. Recall that $\Phi(x) = \begin{cases} 1 & \text{if } T(x) \geq t_\alpha \\ 0 & \text{otherwise} \end{cases}$ with $t_\alpha = F_{\theta_0}^{-1}(1-\alpha)$ (we have assumed that F_{θ_0} is bijective).

$$\begin{aligned} p_\Phi(x) &= \inf\{\alpha : x \in \mathcal{R}_\alpha\} \\ &= \inf\{\alpha : T(x) > F_{\theta_0}^{-1}(1-\alpha)\} \\ &= \inf\{\alpha : F_{\theta_0}(T(x)) > (1-\alpha)\} \\ &= \inf\{\alpha : \alpha > 1 - F_{\theta_0}(T(x))\} \\ &= \inf\{(1 - F_{\theta_0}(T(x)), +\infty)\} \\ &= 1 - F_{\theta_0}(T(x)) \\ &= P_{\theta_0}(T(X) > T(x)) \end{aligned}$$

whereas the last equality holds because F_{θ_0} is the cdf of $T(X)$ under $\theta = \theta_0$. □

Lemma 4.3. $p_\Phi(X) \sim \mathcal{U}([0,1])$ *under* $H_0 : \theta = \theta_0$.

Proof. We know that $p_\Phi(X) = 1 - F_{\theta_0}(T(X))$. Recall that if Y is some random variable with cdf equal to F, and F is bijective, then $U = F(Y) \sim \mathcal{U}([0,1])$. Indeed, since $F(Y) \leq u$ if and only if $Y \leq F^{-1}(u)$, we see that the cdf of U is $P(U \leq u) = \begin{cases} 0 & \text{if } u < 0 \\ u & \text{if } 0 \leq u < 1 \\ 1 & \text{if } u \geq 1 \end{cases}$, because $u = F(F^{-1}(u)) = P(Y \leq F^{-1}(u))$ and thus $F(Y) \sim \mathcal{U}([0,1])$. Thus $F_{\theta_0}(T(X)) \sim \mathcal{U}([0,1])$ and therefore $1 - F_{\theta_0}(T(X)) \sim \mathcal{U}([0,1])$. □

Recall that we have considered a simple setting. P-values can also be defined through the following definition

Definition 4.4. ***proper p-value*** *Consider testing* $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_1$ *such that* $\Theta_0 \cap \Theta_1 = \emptyset$. *A p-value* $p(X)$ *is said to be valid (or proper) if for all* $\theta \in \Theta_0$ *and for all* $t \in [0,1]$ *we have* $P_\theta(p(X) \leq t) \leq t$. *This means that* $p(X)$ *is a valid p-value if it is stochastically larger than* $U \sim \mathcal{U}([0,1])$ *under any* $\theta \in \Theta_0$.

Remark: Note that Definition (in the simple setting) gives a p-value that is stochastically equal to $U \sim \mathcal{U}([0,1])$.
Example: Let T be some statistic used for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. Define $p(x) = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$. We want to check that this defines a valid p-value. For that, we will need the following result.

Lemma 4.5. *Let Z be any random variable with distribution function F (not necessarily continuous or strictly increasing). Then* $U = F(Z)$ *satisfies* $P(U \leq u) \leq u$ *for all* $u \in [0,1]$.

Proof. We either have

$$F(\zeta) \leq u \Leftrightarrow \zeta \leq \zeta_u$$

or

$$F(\zeta) \leq u \Leftrightarrow \zeta < \zeta_u.$$

$$P(F(Z) \leq u) = \begin{cases} P(Z \leq \zeta_u) & \text{if } F(\zeta_u) = u \\ P(Z < \zeta_u) & \text{if } F(\zeta_u) > u \end{cases} = \begin{cases} F(\zeta_u) = u \\ F(\zeta_u -) \leq u \end{cases}$$

In any case we arrive at $P(F(Z) \leq u) = P(U \leq u) \leq u$. □

Remark: This is saying for any distribution function F, $F(z)$ is stochastically larger than $U \sim \mathcal{U}([0,1])$ with $Z \sim F$.
Now let us return to $p(x) = \sup_{\theta\in\Theta_0} P_\theta(T(X) \geq T(x))$. We will check that this defines a valid p-value.

Proof. Fix $\theta \in \Theta_0$ and denote by F_θ the cdf of $-T(X)$. Define

$$\begin{aligned} p_\theta(x) &= P_\theta\left(T(X) \geq T(x)\right) \\ &= P_\theta\left(-T(X) \leq -T(x)\right) = F_\theta(-T(x)). \end{aligned}$$

Using Lemma we know that $p_\theta(X)$ is stochastically larger than $\mathcal{U}([0,1])$.
For $\tilde{\theta} \in \Theta_0$:

$$\begin{aligned} P_{\tilde{\theta}}\left(p(X) \leq t\right) &= P_{\tilde{\theta}}\left(\sup_{\theta\in\Theta_0} F_\theta(-T(X)) \leq t\right) \\ &= P_{\tilde{\theta}}\left(\forall\theta \in \Theta_0\ F_\theta(-T(X)) \leq t\right) \\ &\leq P_{\tilde{\theta}}\left(F_{\tilde{\theta}}(-T(X)) \leq t\right) \\ &= P_{\tilde{\theta}}\left(p_{\tilde{\theta}}(X) \leq t\right) \leq t. \end{aligned}$$

In conclusion: $\forall t \in [0,1], \forall\tilde{\theta} \in \Theta_0$: $P_{\tilde{\theta}}\left(p(X) \leq t\right) \leq t \Leftrightarrow \sup_{\theta\in\Theta_0} P_\theta(p(X) \leq t) \leq t$ which means that $p(X)$ is indeed a valid p-value. □

What is the link between a valid p-value and testing? Given any valid p-value, we can construct the following test Φ at a given level α: $\Phi(x) = 1$ if and only if $p(x) \leq \alpha$.
Type-1 error $\sup_{\theta\in\Theta_0} E_\theta[\Phi(x)] = \sup_{\theta\in\Theta_0} P_\theta(\Phi(x) = 1) = \sup_{\theta\in\Theta_0} P_\theta(p(x) \leq \alpha) \leq \alpha$.

5. Brief look at multiple testing

Consider multiple hypothesis that we want to test at the same time. Call these (null) hypotheses $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(m)}$ for some integer $m \geq 2$. Suppose for all $i \in \{1,2,\ldots,m\}$ we have a test Φ_i for testing $H_0^{(i)}$ versus $H_1^{(i)}$ (some alternative). Consider the combined test Φ which rejects/accepts $H_0^{(i)}$ if Φ_i does. Let us suppose Φ_i has level α and that these tests are independent.

$$H_0 = H_0^{(1)} \cap H_0^{(2)} \cap \ldots \cap H_0^{(m)}$$

The Type-I error of

$$\begin{aligned} \Phi &= P_{H_0}(\text{rejecting at least one } H_0^{(i)} \text{ for some } i \in \{1,\ldots,m\}) \\ &= 1 - P_{H_0}(\text{accepting } H_0^{(1)} \text{ and } H_0^{(2)} \text{ and } \ldots \text{ and } H_0^{(m)}) \\ &= 1 - \prod_{i=1}^m P_{H_0}(\text{accepting } H_0) \\ &= 1 - \prod_{i=1}^m P_{H_0}(\Phi_i \text{accepts } H_0^{(i)}) \\ &= 1 - \prod_{i=1}^m P_{H_0^{(i)}}(\Phi_i \text{accepts } H_0^{(i)}) \\ &= 1 - (1-\alpha)^m \end{aligned}$$

Numerical illustration:

$$m = 10 \quad \alpha = 0.05 \quad \text{Type-I error} = 0.4$$

$$m = 50 \quad \alpha = 0.01 \quad \text{Type-I error} = 0.39$$

This means that we need to be more strict when choosing the levels of the individual tests.

5.1. **Bonferroni's correction.** gives a solution to this problem. Here we are not going to assume that tests Φ_i are independent.

$$\begin{aligned}
P_{H_0}\left(\text{rejecting at least } H_0^{(i)} \text{ for some } i \in \{1,\dots,m\}\right) &= P_{H_0}\left(\exists i \in \{1,\dots,m\} : \Phi \text{ rejects } H_0^{(i)}\right) \\
&= P_{H_0}\left(\cup_{1\le i\le m}\{\Phi \text{ rejects } H_0^{(i)}\}\right) \\
&\le \sum_{i=1}^{m} P_{H_0}\left(\Phi \text{ rejects } H_0^{(i)}\right) \\
&= \sum_{i=1}^{m} P_{H_0}\left(\Phi_i \text{ rejects } H_0^{(i)}\right) \\
&= \sum_{i=1}^{m} P_{H_0^{(i)}}\left(\Phi \text{ rejects } H_0^{(i)}\right)
\end{aligned}$$

If we chose the level of each test Φ_i to be $\frac{\alpha}{m}$, then the Type-I error of $\Phi \le m\frac{\alpha}{m} = \alpha$. Alternatively, we can require in this correction to have α_i (the level of Φ_i) satisfy $\sum_{i=1}^{m} \alpha_i \le \alpha$ (this will imply that the Type-I error of $\Phi \le \sum_{i=1}^{m} \alpha_i \le \alpha$).

Part 2. Further methods for constructing tests

1. Likelihood Ratio Tests

Definition 1.1. ***likelihood*** *Let $X_1,\dots X_n$ be iid random variables admitting a density assumed to belong to the parametric family $\{p_\theta, \theta \in \Theta\}$*

- *We call likelihood the function*

$$\Theta \to [0,\infty)$$

$$\theta \mapsto L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

- *We call log-likelihood the function*

$$\Theta \to \mathbb{R}$$

$$\theta \mapsto l_n(\theta) = \log\left(L_n(\theta)\right)$$

Definition 1.2. ***MLE*** *The maximum likelihood estimator (MLE) is any $\hat{\theta}_n$ satisfying $L_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} L_n(\theta)$ and since the logarithm is continuous and increasing $l_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} l_n(\theta)$*

Remarks:

- The MLE does not have to exist.
- If the MLE exists it is not necessarily unique.
- For any subset $\Theta' \subset \Theta$ we can define the restricted MLE which maximises $\theta \mapsto L_n(\theta)$ (or $\theta \mapsto l_n(\theta)$) over Θ'.

Definition 1.3. ***likelihood ratio statistic*** *Let Θ_0 and Θ_1 be two subsets of Θ such that $\Theta_0 \cap \Theta_1 = \emptyset$ ($\Theta_0 \cup \Theta_1 = \Theta$) and consider the testing problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ The likelihood ratio statistic is defined as $\Lambda_n = \frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}$.*

Definition 1.4. ***LRT*** *The likelihood ratio test for a given level α is given by*

$$\Phi(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \Lambda_n > \lambda_\alpha \\ \gamma_\alpha & \text{if } \Lambda_n = \lambda_\alpha \\ 0 & \text{if } \Lambda_n < \lambda_\alpha \end{cases}$$

where γ_α and λ_α are such that $\sup_{\theta\in\Theta} E_\theta[\Phi(X_1,\dots,X_n)] \le \alpha$.

Remark: The idea behind the definition of LRT is to reject $H_0 : \theta \in \Theta_0$ when $\frac{\sup_{\theta\in\Theta_1} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}$ is large. (see exercise)

2. Gaussian vectors and related distributions

2.1. Multivariate Gaussian distribution.

- Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$. We say that X is Gaussian if any linear combination of components, X_j $1 \le j \le d$, has a Gaussian distribution: For all $a_j \in \mathbb{R}$ for $j \in \{1, \dots, d\}$ $\sum_{i=1}^d a_j X_j$ is a normal random variable.
- Two Gaussian vectors $X = (X_1, \dots, X_d)$ and $Y = (Y_1, \dots, Y_m)$ are independent if and only if $\mathrm{Cov}(X_i, Y_j) = 0$ for all $(i, j) \in \{1, \dots, d\} \times \{1, \dots, m\}$.
- If $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathrm{Mat}(\mathbb{R}^d \times \mathbb{R}^d)$ then for any matrix $A \in \mathbb{R}^{m \times d}$ $(m \ge 1)$ we have $AX \sim \mathcal{N}(A\mu, A\Sigma A^\intercal)$
- If $X \sim \mathcal{N}(\mu, \Sigma)$ and Σ is invertible, then X admits density $f_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right)$.

2.2. Gamma-function. The gamma function is defined for all complex numbers except the non-positive integers. For complex numbers with a positive real part, it is defined via a convergent improper integral $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$. Note that if $n \in \mathbb{Z}_{>0}$ then $\Gamma(n) = (n-1)!$, $\Gamma(1) = 1$ and $n\Gamma(n) = \Gamma(n+1)$.

2.3. $\chi^2_{(k)}$: Chi-square distribution with k degrees of freedom. We say that $Y \sim \chi^2_{(k)}$ if we can find $X = (X_1, \dots, X_k) \sim \mathcal{N}(0, \mathbb{1}_k)$ such that $Y = \sum_{j=1}^k X_j^2 = \|X\|_2^2$ (the square of the euclidean norm of X). Y admits a density

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} \exp(-y/2)\, \mathbb{1}_{y>0}. \tag{3}$$

We recognize that $Y \sim \mathrm{Gamma}\left(\frac{k}{2}, \frac{1}{2}\right)$. Moreover if $X \sim \mathcal{N}(\mu, \Sigma)$ and Σ is invertible then $(x-\mu)^\intercal \Sigma^{-1}(x-\mu) \sim \chi^2_{(k)}$ (see exercise).

2.4. Distribution of Student(t-) of k degrees of freedom. We say that T follows a $t-$distribution with k degrees of freedom if we can find independent random variables X and Y with $X \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2_{(k)}$ such that $T = \frac{X}{\sqrt{Y/k}}$. We write $T \sim \mathcal{T}_{(k)}$. T admits density given by

$$f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k}\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{(k+1)/2}}, \quad t \in \mathbb{R}. \tag{4}$$

Note that $\mathcal{T}_{(1)}$ is the Cauchy distribution.

2.5. F-distribution. We say that Y admits an F-distribution with (p, q) degrees of freedom if we can find two random variables U and V such that U and V are independent, $U \sim \chi^2_{(p)}$, $V \sim \chi^2_{(q)}$ and $Y \sim \frac{U/p}{V/q}$. We will write $Y \sim \mathrm{F}_{p,q}$. Y admits density given by

$$f_Y(y) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(p/2)\Gamma(q/2)} p^{1/2} q^{1/2} \frac{y^{1/2-1}}{(q+py)^{(p+q)/2}} \mathbb{1}_{y>0}. \tag{5}$$

3. Example for LRT

3.1. Example a. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma_0^2)$, where $\theta \in \mathbb{R}$ and $\sigma_0 > 0$ is known. We want to test

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

Hence we have $\Theta_0 = \{\theta_0\}$ (a simple hypothesis) and $\Theta_1 = \mathbb{R} \setminus \{\theta_0\}$ (a composite hypothesis) such as $\Theta = \Theta_0 \cup \Theta_1 = \mathbb{R}$. Recall that $\Lambda_n = \frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} = \frac{\sup_{\mu \in \mathbb{R}} L_n(\theta)}{L_n(\theta_0)}$.

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n p_\theta(X_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(X-\theta)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X-\theta)^2\right). \end{aligned}$$

$$l_n(\theta) = \log(L_n(\theta)) = \text{ constant } - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \theta)^2$$

We want to show that $\operatorname{argmax}_{\theta\in\mathbb{R}} L_n(\theta) = \overline{X}_n$. Our goal is to maximize $\theta \mapsto \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta)^2\right)$ over $\mathbb{R}$ or equivalently maximize $-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta)^2$ over $\mathbb{R}$.

$$\frac{d}{d\theta}\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta)^2\right) = -2\sum_{i=1}^n (X_i-\theta) = 0 \Leftrightarrow \theta = \overline{X}_n \tag{6}$$

and

$$\frac{d^2}{d\theta^2}\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta)^2\right) = 2n > 0$$

which means that the function is convex on $\mathbb{R}$ and hence $\overline{X}_n$ gives the global maximum of L_n.

$$\begin{aligned}\Lambda_n &= \frac{L_n(\overline{X}_n)}{L_n(\theta_0)}\\ &= \frac{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2\right)}{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta_0)^2\right)}\\ &= \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2 + \frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta_0)^2\right)\end{aligned}$$

Recall that the event $\{\Lambda_n = \lambda_\alpha\}$ happens with probability equal to zero and hence the LRT is given by $\Phi(X_1,\dots,X_n) = \begin{cases}1 & \text{if } \Lambda_n > \lambda_\alpha\\ 0 & \text{if } \Lambda_n \le \lambda_\alpha\end{cases}$ almost surely and we are going to find λ_α such that $E_{\theta_0}(\Phi(X_1,\dots,X_n)) = \alpha$. Note that

$$\begin{aligned}\Lambda_n \text{ is 'large'} &\Leftrightarrow \sum_{i=1}^n (X_i-\theta_0)^2 - \sum_{i=1}^n (X_i-\overline{X}_n)^2 \text{ is 'large'}\\ &\Leftrightarrow \sum_{i=1}^n (X_i-\overline{X}_n+\overline{X}_n-\theta_0)^2 - \sum_{i=1}^n (X_i-\overline{X}_n)^2 \text{ is 'large'}\\ &\Leftrightarrow \sum_{i=1}^n (X_i-\overline{X}_n)^2 + 2\left(\sum_{i=1}^n (X_i-\overline{X}_n)\right)\cdot(\overline{X_i}-\theta_0) + n(\overline{X}_n-\theta_0)^2 - \sum_{i=1}^n (X_i-\overline{X}_n)^2 \text{ is 'large'}\\ &\Leftrightarrow n(\overline{X}_n-\theta_0)^2 \text{ is 'large'}\\ &\Leftrightarrow \frac{n(\overline{X}_n-\theta_0)^2}{\sigma_0^2} \text{ is 'large'}\\ &\Leftrightarrow \frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sigma_0} \text{ is 'large'}\end{aligned}$$

$\Phi(X_1,\dots,X_n) = \begin{cases}1 & \text{if } \frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sigma_0} > q_\alpha\\ 0 & \text{otherwise}\end{cases}$ such that $E_{\theta_0}(\Phi(X_1,\dots,X_n)) = P_{\theta_0}\left(\frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sigma_0} > q_\alpha\right) = \alpha$. We need to determine the quantile q_α. Recall $X_1,\dots,X_n \overset{iid}{\sim} \mathcal{N}(\theta_0,\sigma_0^2)$ under H_0 which means that $\overline{X}_n \sim \mathcal{N}(\theta_0,\sigma_0^2/n) \Leftrightarrow \frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma_0} \overset{d}{=} Z \sim \mathcal{N}(0,1)$.

$$\begin{aligned}P_{\theta_0}\left(\frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sigma_0} > q_\alpha\right) &= P(|Z| > q_\alpha)\\ &= P(Z > q_\alpha) + P(Z < -q_\alpha)\\ &= P(Z > q_\alpha) + P(-Z > q_\alpha)\\ &= 2P(Z > q_\alpha)\end{aligned}$$

by symmetry around zero of the Z distribution. Hence,

$$\begin{aligned}\alpha &= P_{\theta_0}(\Phi \text{ rejects } H_0)\\ &= 2P(Z > q_\alpha)\\ &\Leftrightarrow P(Z > q_\alpha) = \alpha/2\\ &\Leftrightarrow F_Z(q_\alpha) = 1 - \alpha/2\end{aligned}$$

therefore $\Phi(X_1,\dots,X_n) = \begin{cases}1 & \text{if } \frac{\sqrt{n}|\overline{X}_n - \theta_0|}{\sigma_0} > \zeta_{1-\alpha/2}\\ 0 & \text{otherwise}\end{cases}$ where $\zeta_{1-\alpha/2} = q_\alpha = (1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$ and $F_Z(\zeta) = \int_{-\infty}^{\zeta} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

3.2. **Cochrans Theorem.**

Theorem 3.1. ***Cochran*** *Let $(X_1,\dots,X_d) = X \sim \mathcal{N}_d(0,\mathbb{1})$ be a Gaussian vector. Let $A_1,\dots,A_J$ be $d\times d$ matricies such that $\sum_{i=1}^J rank(A_i) \le d$ and for all $i \in \{1,\dots,J\}$*

(i) A_i is symmetric and $A_i^2 = A_i$.
(ii) $A_iA_j = A_jA_i = 0$ for all $i \neq j$.

Then,

(i) $A_iX \sim \mathcal{N}(0,A_i)$ for all $i \in \{1,\dots J\}$ and $A_1X,\dots,A_JX$ are mutually independent.
(ii) The random variables $\|A_iX\|^2 \sim \chi^2_{rank(A_i)}$ and they are mutually independent.

Proof. i) We know that $X \sim \mathcal{N}(\mu,\Sigma)$ implies $AX \sim \mathcal{N}(A\mu, A\Sigma A^\intercal)$. Thus $A_iX \sim \mathcal{N}(0, A_iA_i^\intercal) \stackrel{d}{=} \mathcal{N}(0,A_i)$. Then, showing mutual independence of $A_iX,\dots A_JX$ is equivalent to showing $\mathrm{Cov}\left(A_iX, A_jX\right) = 0$ for all $i \neq j$. Let $E[X] = \mu$ and recall that

$$\begin{aligned}\mathrm{Cov}\,(AX,BX) &= E\left[A(X-\mu)(B(X-\mu))^\intercal\right]\\ &= E\left[A(X-\mu)(X-\mu)^\intercal B^\intercal\right]\\ &= AE\left[(X-\mu)(X-\mu)^\intercal\right]B^\intercal\\ &= A\Sigma B^\intercal.\end{aligned}$$

Hence in our case for $i \neq j \in \{1,\dots,J\}$ we have

$$\begin{aligned}\mathrm{Cov}(A_iX, A_jX) &= A_i\mathbb{1}A_j^\intercal\\ &= A_iA_j^\intercal\\ &= A_iA_j\\ &= 0\end{aligned}$$

by assumption.
ii) $A_1X,\dots,A_JX$ mutually independent implies $f(A_1X),\dots,f(A_JX)$ mutually independent for some measurable function f. In particular, this is true for $f(a) = \|a\|^2$ ($a \in \mathbb{R}^d$) continuous on $\mathbb{R}^d$ and hence measurable. We now show that $\|A_iX\|^2 \sim \chi^2_{(\mathrm{rank}(A_i))}$. A_i is symmetric. We can orthogonalize A_i in an orthonormal basis. There exists an orthogonal matrix P so that we can decompose $A_i = P^\intercal \begin{pmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \ddots & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \dots & 0 & \lambda_d\end{pmatrix} P$ where $\lambda_1,\dots,\lambda_d$ denote the eigenvalues of A_i. Using the assumption $A_i^2 = A_i$, we conclude that $\lambda_1,\dots,\lambda_d \in \{0,1\}$. Further we can decompose A_i^2 in the following

way

$$A_i^2 = P^\intercal \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_d \end{pmatrix} PP^\intercal \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_d \end{pmatrix} P$$
$$= P^\intercal \begin{pmatrix} \lambda_1^2 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_d^2 \end{pmatrix} P = A_i$$

which means that $\lambda_i^2 = \lambda_i$ for all $i \in \{1, \dots, d\}$ and hence there are only two solutions. We can also write $A_i = P^\intercal \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix} P$. Then $\mathbb{1}$ has size equal to the rank of A_i.

$$\begin{aligned} \|A_iX\|^2 &= (A_iX)^\intercal A_iX \\ &= X^\intercal A_i^\intercal A_iX \\ &= X^\intercal A_i^2 X \\ &= X^\intercal P^\intercal \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix} PX \\ &= (PX)^\intercal \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix} PX \\ &= Y^\intercal \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix} Y \\ &= \sum_{j=1}^{\mathrm{rank}(A_i)} Y_j^2. \end{aligned}$$

On the other hand, $Y = PX \sim \mathcal{N}(0, P\mathbb{1}P^\intercal)$. Hence $\|A_iX\|^2$ = the norm of a squared vector $\sim \mathcal{N}\left(0, \mathbb{1}_{\mathrm{rank}(A_i)}\right)$; in other words $Y_1, \dots, Y_{\mathrm{rank}(A_i)}$ are $\overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. □

3.3. **Example b.** Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\theta \in \mathbb{R}$ and $\sigma \in (0, \infty)$ both unknown. Here σ is acting as a nuisance parameter. We want to test

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

whereas $\Theta_0 = \{(\theta_0, \sigma) : \sigma \in (0, \infty)\} = \{\theta_0\} \times (0, \infty)$ and $\Theta = \{(\theta, \sigma) : \theta \in \mathbb{R} \text{ and } \sigma \in (0, \infty)\} = \mathbb{R} \times (0, \infty)$. Since σ is unknown, we have

$$\Lambda_n = \frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}$$

and

$$L_n(\theta, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2\right).$$

We need to maximize $(\theta, \sigma) \mapsto L_n(\theta, \sigma)$ over Θ. This is equivalent to maximizing

$$l_n(\theta, \sigma) = -n/2 \log(2\pi) - n \log(\sigma) - 1/(2\sigma^2) \sum_{i=1}^n (X_i - \theta)^2.$$

3.3.1. *Maximisation via profiling:* Let us fix $\sigma \in (0, \infty)$ and define the function $g_\sigma(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2$ which we are going to maximize over $\mathbb{R}$. Since $-\frac{1}{2\sigma^2}$ is a constant here. We can use previous calculations from example a). To show that the minimum is attained at $\theta = \overline{X}_n$. $\sup_{\theta\in\mathbb{R}} L_n(\theta,\sigma) = L_n(\overline{X}_n,\sigma)$ for any fixed $\sigma \in (0,\infty)$. Now, we go back to the log-likelihood and plug in $\overline{X}_n$: define the function

$$h(\sigma) = l_n(\overline{X}_n,\sigma) = -n/2\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_n - \overline{X}_n)^2$$

which we want to maximize over $(0,\infty)$.

$$\begin{aligned} h'(\sigma) &= -n/\sigma + 1/\sigma^3\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0 \\ &\Leftrightarrow \sigma^2 = 1/n\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\ &\Leftrightarrow \sigma = \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right]^{1/2} \end{aligned} \tag{7}$$

and

$$\begin{aligned} h''(\sigma) &= n/\sigma^2 - 3/\sigma^4\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\ &= n/\sigma^2 - 3/\sigma^4 n\hat{\sigma}^2 \\ &= n/\sigma^2 - \frac{3n\hat{\sigma}^2}{\sigma^4} \\ &= n/\sigma^4(\sigma^2 - 3\hat{\sigma}^2). \end{aligned}$$

The function h has a local maximum at (7). But, since h has a unique critical point, the function cannot go up to a larger value ($> h(\hat{\sigma})$) because otherwise h has to go down to reach another critical point. Therefore, (7) must be the global maximizer of h over $(0,\infty)$. We need to compute $\sup_{(\theta,\sigma)\in\Theta_0} L_n(\theta,\sigma) = \sup_{\sigma\in(0,\infty)} L_n(\theta_0,\sigma)$. Using similar arguments as for showing that (7) is the global maximizer of the function $\sigma \mapsto l_n(\overline{X}_n,\sigma)$ we can show that $\sup_{\sigma\in(0,\infty)} L_n(\theta_0,\sigma) = L_n(\theta_0,\hat{\sigma}_0)$ with

$$\hat{\sigma}_0 = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta_0)^2\right)^{1/2}. \tag{8}$$

$$\begin{aligned} \Lambda_n &= \frac{\sup_{(\theta,\sigma)\in\Theta} L_n(\theta,\sigma)}{\sup_{(\theta,\sigma)\in\Theta_0} L_n(\theta,\sigma)} \\ &= \frac{L_n(\overline{X}_n,\hat{\sigma})}{L_n(\theta_0,\hat{\sigma}_0)} \\ &= \frac{\frac{1}{(2\pi)^{n/2}}\frac{1}{\hat{\sigma}^n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)}{\frac{1}{(2\pi)^{n/2}}\frac{1}{\hat{\sigma}_0^n}\exp\left(-\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(X_i - \theta_0)^2\right)} \\ &= \frac{\frac{1}{\hat{\sigma}^n}\exp(-n/2)}{\frac{1}{\hat{\sigma}_0^n}\exp(-n/2)} \\ &= \left(\frac{\hat{\sigma}_0}{\hat{\sigma}}\right)^n = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{n/2}. \end{aligned}$$

We reject when Λ_n is 'large' but

$$\begin{aligned}
\Lambda_n \text{ is 'large'} &\Leftrightarrow \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \text{ is 'large'}\\
&\Leftrightarrow \frac{1/n\sum_{i=1}^n (X_i-\theta_0)^2}{1/n\sum_{i=1}^n (X_i-\overline{X}_n)^2} \text{ is 'large'}\\
&\Leftrightarrow \frac{\sum_{i=1}^n (X_i-\overline{X}_n)^2 + n(\overline{X}_n-\theta_0)^2}{\sum_{i=1}^n (X_i-\overline{X}_n)^2} \text{ is 'large'}\\
&\Leftrightarrow 1+\frac{n(\overline{X}_n-\theta_0)^2}{\sum_{i=1}^n (X_i-\overline{X}_n)^2} \text{ is 'large'}\\
&\Leftrightarrow \frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sqrt{\sum_{i=1}^n (X_i-\overline{X}_n)^2}} \text{ is 'large'}\\
&\Leftrightarrow \frac{\sqrt{n}|\overline{X}_n-\theta_0|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2}} \text{ is 'large'}.
\end{aligned}$$

We can find the distribution of $T_n := \frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2}}$ under $H_0: \theta=\theta_0$ using Cochrans theorem. If $(X_1,\ldots,X_n) = X \sim \mathcal{N}_n(\theta_0, \sigma^2\mathbb{1})$ then $(\frac{X_1-\theta_0}{\sigma_0},\ldots,\frac{X_n-\theta_0}{\sigma_0}) = Y \sim \mathcal{N}_n(0,\mathbb{1})$. Define $A_1 = \frac{1}{n}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix}$ and $A_2 = \mathbb{1} - A_1$. We have to check that A_1 and A_2 fulfil the assumptions of Cochrans theorem.

$$A_1^2 = \frac{1}{n^2}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix}\cdot\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix} = \frac{1}{n^2}\begin{pmatrix}n & \ldots & n\\ \vdots & & \vdots\\ n & \ldots & n\end{pmatrix} = \frac{1}{n}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix} = A_1$$

and $A_2 = \mathbb{1} - A_1$ $A_1(\mathbb{1}-A_1) = A_1 - A_1^2 = 0 = (\mathbb{1}-1)A_1$ $\operatorname{rank}(A_1) = 1$ and $\operatorname{rank}(A_2) = n-1$. Therefore, by Cochrans theorem, we know that A_1Y is independent of A_2Y and $\|A_2Y\|_2^2 \sim \chi^2_{(n-1)}$

$$A_1Y = \frac{1}{n}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix}\begin{pmatrix}\frac{X_1-\theta_0}{\sigma_0}\\ \vdots\\ \frac{X_n-\theta_0}{\sigma_0}\end{pmatrix} = \frac{\overline{X}_n-\theta_0}{\sigma_0}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

$$A_2Y = (\mathbb{1}-A_1)Y = Y - A_1Y = \begin{pmatrix}\frac{X_1-\theta_0}{\sigma_0}\\ \vdots\\ \frac{X_n-\theta_0}{\sigma_0}\end{pmatrix} - \frac{\overline{X}_n-\theta_0}{\sigma_0}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \begin{pmatrix}\frac{X_1-\overline{X}_n}{\sigma_0}\\ \vdots\\ \frac{X_n-\overline{X}_n}{\sigma_0}\end{pmatrix}$$

so that $\|A_2Y\|^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2$. .... Now $A_1Y \perp\!\!\!\perp A_2Y \Rightarrow A_1Y \perp\!\!\!\perp \|A_2Y\|^2 \Leftrightarrow \frac{\overline{X}_n-\theta_0}{\sigma} \perp\!\!\!\perp \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2$

$$\Rightarrow \underbrace{\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma}}_{\sim\mathcal{N}(0,1)} \perp\!\!\!\perp \underbrace{\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2}_{\sim\chi^2_{(n-1)}}$$

and using (4)

$$\Rightarrow \frac{\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2}} \sim \mathcal{T}_{(n-1)} \text{ under } H_0.$$

Note that the obtained statistic $T_n = \frac{\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n \frac{(X_i-\overline{X}_n)^2}{\sigma^2}}}$ Thus, the LRT is given by $\Phi(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } |T_n| > q_\alpha \\ 0 & \text{otherwise} \end{cases}$ where

$$\begin{aligned} P(|T_n| > q_\alpha) = \alpha &\Leftrightarrow 2P(T_n > q_\alpha) = \alpha \\ &\Leftrightarrow P(T_n > q_\alpha) = \alpha/2 \\ &\Leftrightarrow P(T_n \leq q_\alpha) = 1-\alpha/2 \end{aligned}$$

whereas $q_\alpha = t_{n-1,1-\alpha/2}$ the $(1-\alpha/2)$-quantile of $\mathcal{T}_{(n-1)}$.

3.4. **Example c.** Let $X_1,\dots,X_n \overset{iid}{\sim} \mathcal{N}(\theta_0,\sigma^2)$ with $\theta_0 \in \mathbb{R}$ known and $\sigma \in (0,\infty)$ unknown. We want to test

$$H_0 : \sigma = \sigma_0 \text{ versus } H_1 : \sigma \neq \sigma_0$$

whereas $\Theta_0 = \{\sigma_0\}$ and $\Theta = (0,+\infty)$.

$$\Lambda_n = \frac{\sup_{\sigma\in(0,\infty)} L_n(\theta_0,\sigma)}{L_n(\theta_0,\sigma_0)}$$

$L_n(\theta_0,\sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta_0)^2\right)$ then

$$l_n(\theta_0,\sigma) = -n/2\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta_0)^2.$$

$$\frac{d}{d\sigma}\left(l_n(\theta_0,\sigma)\right) = -n/\sigma + 1/\sigma^3\sum_{i=1}^n (X_i-\theta_0)^2 = 0 \Leftrightarrow \sigma^2 = 1/n\sum_{i=1}^n (X_i-\theta_0)^2$$

which implies that there exists a unique critical point

$$\hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^n (X_i-\theta_0)^2\right)^{1/2}$$

$$\frac{d^2}{d\sigma^2}(l_n(\theta_0,\sigma)) = n/\sigma^2 - 3/\sigma^4\sum_{i=1}^n (X_i-\theta_0)^2$$

and

$$\frac{d^2}{d\sigma^2}(l_n(\theta_0,\sigma))|_{\sigma=\hat{\sigma}} = n/\hat{\sigma} - \frac{3n\hat{\sigma}^2}{\hat{\sigma}^4} = \frac{2n}{\hat{\sigma}^2} < 0$$

which means that $\hat{\sigma}$ is a local maximizer and hence a global maximizer because otherwise the function $\sigma \mapsto l_n(\theta_0,\sigma)$ will have another critical point. Note that this obtained $\hat{\sigma}$ is equal to (**??**).

$$\begin{aligned} \Lambda_n &= \frac{L_n(\theta_0,\hat{\sigma})}{L_n(\theta_0,\sigma_0)} \\ &= \frac{\frac{1}{(2\pi)^{n/2}}\frac{1}{\hat{\sigma}^n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (X_i-\theta_0)^2\right)}{\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma_0^n}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\theta_0)^2\right)} \\ &= \frac{\frac{1}{\hat{\sigma}^n}\exp\left(-\frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2\right)}{\frac{1}{\sigma_0^n}\exp\left(-\frac{1}{2\sigma_0^2}n\hat{\sigma}^2\right)} \\ &= \frac{\sigma_0^n}{\hat{\sigma}^n}\exp\left(-n/2 + n/2\cdot\hat{\sigma}^2/\sigma_0^2\right) \\ &= \frac{1}{(\hat{\sigma}/\sigma_0)^n}\exp\left(-\frac{n}{2}\left[\left(\frac{\hat{\sigma}}{\sigma_0}\right)^2 - 1\right]\right) \\ &= g\left(\frac{\hat{\sigma}}{\sigma_0}\right) \end{aligned}$$

with $g(t) = 1/t^n\exp\left(n/2(t^2-1)\right)$ for $t \in (0,+\infty)$.

$$\begin{aligned} h(t) &= \log(g(t)) \\ &= -n\log(t) + n/2(t^2-1) \end{aligned}$$

$$h'(t) = -n/t + nt = n\frac{t^2-1}{t}$$

But we know that, by definition, $\Lambda_n \geq 1$ and hence $\Lambda_n = g\left(\frac{\hat{\sigma}}{\sigma_0}\right)$ which implies $\frac{\hat{\sigma}}{\sigma_0} \in [1,+\infty)$. Since g is strictly increasing on $[1,+\infty)$,

$$\begin{aligned}\Lambda_n \text{ is 'large'} &\Leftrightarrow \frac{\hat{\sigma}}{\sigma_0} \text{ is 'large'}\\ &\Leftrightarrow \frac{\hat{\sigma}^2}{\sigma_0^2} \text{ is 'large'}\\ &\Leftrightarrow \frac{1/n\sum_{i=1}^n (X_i-\theta_0)^2}{\sigma_0^2} \text{ is 'large'}\\ &\Leftrightarrow \sum_{i=1}^n \frac{(X_i-\theta_0)^2}{\sigma_0^2} \text{ is 'large'}.\end{aligned}$$

The LRT is given by $\Phi(X_1,\ldots,X_n) = \begin{cases}1 & \text{if } \sum_{i=1}^n \frac{(X_i-\theta_0)^2}{\sigma_0^2} > q_\alpha\\ 0 & \text{otherwise}\end{cases}$ with $P_{\sigma_0}\left(\sum_{i=1}^n \frac{(X_i-\theta_0)^2}{\sigma_0^2} > q_\alpha\right) = \alpha$.

$\frac{X_1-\theta_0}{\sigma_0},\ldots,\frac{X_n-\theta_0}{\sigma_0} \overset{iid}{\sim} \mathcal{N}(0,1)$ under $H_0 : \sigma = \sigma_0$ which implies $\sum_{i=1}^n \frac{(X_i-\theta_0)^2}{\sigma_0^2} \sim \chi^2_{(n)}$ and q_α the $(1-\alpha)$-quantile of $\chi^2_{(n)}$.

3.5. **Example d.** Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{N}(\theta,\sigma^2)$ with $\theta \in \mathbb{R}$ and $\sigma \in (0,\infty)$ both unknown. Here θ is acting as a nuisance parameter and we want to test

$$H_0 : \theta \text{ is something}, \sigma = \sigma_0 \text{ versus } H_1 : \theta \text{ is something}, \sigma \neq \sigma_0$$

whereas $\Theta_0 = \{(\theta,\sigma_0) : \theta \in \mathbb{R}\}$ and $\Theta = \mathbb{R} \times (0,+\infty)$.

$$\Lambda_n = \frac{\sup_{(\theta,\sigma)\in\Theta} L_n(\theta,\hat{\sigma})}{\sup_{\theta\in\mathbb{R}} L_n(\theta,\sigma_0)}$$

$$L_n(\theta,\sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta)^2\right)$$

We already know from example b that $\sup_{(\theta,\sigma)\in\Theta} = L_n(\overline{X}_n,\hat{\sigma})$ with $\hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2\right)^{1/2}$ and also

$$\begin{aligned}\Lambda_n &= \frac{\frac{1}{(2\pi)^{n/2}}\frac{1}{\hat{\sigma}^n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2\right)}{\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma_0^n}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2\right)}\\ &= \frac{1/\hat{\sigma}^n}{\hat{\sigma}_0^n}\exp\left(-n/2 + n/2\cdot\hat{\sigma}^2/\sigma_0^2\right)\end{aligned}$$

$\Lambda_n = g\left(\frac{\hat{\sigma}}{\sigma_0}\right)$ where g is the same function as before. Using similar arguments we show that Λ_n is 'large' if and only if $\sum_{i=1}^n \frac{(X_i-\overline{X}_n)^2}{\sigma_0^2}$ is 'large'. $\sum_{i=1}^n \frac{(X_i-\overline{X}_n)^2}{\sigma_0^2} \sim \chi^2_{(n-1)}$ as a result of Cochran's theorem. The LRT is given by $\Phi(X_1,\ldots,X_n) = \begin{cases}1 & \text{if } \sum_{i=1}^n \frac{(X_i-\overline{X}_n)^2}{\sigma_0^2} > q_\alpha\\ 0 & \text{otherwise}\end{cases}$ with $q_\alpha = (1-\alpha)$-quantile of $\chi^2_{(n-1)}$.

4. F-tests and application in linear regression

4.1. **Regression model.** A regression model aims at explaining the random behaviour of the response given the explanatory variables also called covariates/predictors. More specifically, a regression model assumes that $Y = f(\theta,x)+\epsilon$ whereas Y is the response, f and θ are unknown x are the covariate(s) and ϵ is the noise/error.

There are two settings:

(1) Random design: the covariate is random and the analysis is done conditionally on X but in the end randomness is taken into account.
(2) Fixed design: We observe a realisation x of X and we do the analysis conditionally on $X = x$.

In this course we will place ourselves in the fixed design.

4.2. **Linear Regression.** When $f(\theta, x) = \theta^\intercal x$ with $\theta, x \in \mathbb{R}^d$, then we talk about linear regression. The model is $Y = \theta^\intercal x + \epsilon$ with $E(\epsilon) = 0$. If $\theta_1, \ldots, \theta_d$ are the components of θ and $x_1, \ldots, x_d$ are the components of x then

$$Y = x_1\theta_1 + \ldots + \theta_d x_d + \epsilon.$$

The main goal is to estimate the unknown regression vector θ based on a random sample. We observe independent responses $Y_1, \ldots, Y_n$ and corresponding covariates $x_1, \ldots, x_n \in \mathbb{R}^d$. Let

$$Y_i = \theta^\intercal x_i + \epsilon_i$$

with $x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}$ for $i \in \{1, \ldots, n\}$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n$ and put $D = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ \vdots & & & \vdots \\ x_{i1} & \ldots & \ldots & x_{id} \\ \vdots & & & \vdots \\ x_{n1} & \ldots & \ldots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times d}$. The ith row of $D = x_i^\intercal = (x_{i1}, \ldots, x_{id})$. D is called the design-matrix. We can write the linear regression as

$$Y = D\theta + \epsilon. \tag{9}$$

4.3. **Least Squares Estimator.**

Definition 4.1. *LSE Consider the quadratic criterion*

$$Q_n(t) = \sum_{i=1}^{n} (Y_i - t^\intercal x_i)^2 \tag{10}$$

for $t \in \mathbb{R}^d$. $\hat{\theta}_n = argmin_{t \in \mathbb{R}^d} Q_n(t)$ is called (provided it exists) the least squares estimator if it minimizes Q_n over $\mathbb{R}^d$.

The rational behind $\hat{\theta}_n$ is that we can take some random variable Z with $\mu = E(Z) < \infty$ and $\sigma^2 = \mathrm{Var}(Z) < \infty$ then $\mu = \mathrm{argmin}_{a \in \mathbb{R}} E[(Z - a)^2]$. Indeed

$$\begin{aligned} E[(Z-a)^2] &= E[(Z - \mu + \mu - a)^2] \\ &= E[(Z-\mu)^2 + 2(Z-\mu)(\mu - a) + (\mu - a)^2] \\ &= \sigma^2 + 2(\mu - a)E[Z - \mu] + (\mu - a)^2 \\ &= \sigma^2 + (\mu - a)^2. \end{aligned}$$

Since $\mathrm{argmin}_a (\mu - a)^2 = \mu$ it follows that $\mu = \mathrm{argmin}_a E[(Z-a)^2]$. Let us go back to the regression problem and let us also assume that $\mathrm{Var}(Y_i) < \infty$ for $i \in \{1, \ldots, n\}$. Since $E(\epsilon_i) = 0$ for $i \in \{1, \ldots, n\}$, this means that $E(Y_i) = \theta^\intercal x_i = \mu_i$. We can also show as above that

$$(\mu_1, \ldots, \mu_n)^\intercal = \sum_{i=1}^{n} E[(Y_i - a_i)^2] \Rightarrow \theta = \mathrm{argmin}_{t \in \mathbb{R}^d} \sum_{i=1}^{n} E[(Y_i - t^\intercal x_i)^2].$$

Since we only observe $Y_1, \ldots, Y_n$ and $x_1, \ldots, x_n$ we replace this criterion by (10).

Proposition 4.2. *Assume that $D^\intercal D$ is invertible. Then, $\hat{\theta}_n$ exists and is unique. Furthermore*

$$\hat{\theta}_n = (D^\intercal D)^{-1} D^\intercal Y. \tag{11}$$

Proof. Recall that for $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ the euclidean norm is defined as $\| \sqrt{\sum_{i=1}^n v_i} \|$ and $\|v\|^2 = v^\intercal v$. Hence

$$\begin{aligned} Q_n(t) &= \sum_{i=1}^{n} (Y_i - t^\intercal x_i)^2 \\ &= \|Y - Dt\|^2 \\ &= (Y - Dt)^\intercal (Y - Dt) \\ &= Y^\intercal Y - Y^\intercal Dt - t^\intercal D^\intercal Y + t^\intercal D^\intercal Dt \\ &= Y^\intercal Y - 2t^\intercal D^\intercal Y + t^\intercal D^\intercal Dt \end{aligned}$$

We look now for a stationary point of $Q_n : \nabla Q_n(t) = -2D^\intercal Y + 2D^\intercal Dt$. Recall that for any differentiable function g defined on $\mathbb{R}^d$ we have

$$g(t + h) = g(t) + h^\intercal \nabla g(t) + o(\|h\|).$$

Therefore

$$\begin{aligned}\nabla Q_n(t) = 0 &\Leftrightarrow D^\intercal D t = D^\intercal Y \\ &\Leftrightarrow t = (D^\intercal D)^{-1} D^\intercal Y.\end{aligned}$$

The hessian of $Q_n(t)$ is $2D^\intercal D$, which is positive definite because for $a \in \mathbb{R}^d$

$$\begin{aligned}a^\intercal D^\intercal D a &= (Da)^\intercal Da \\ &= \|Da\|^2 \geq 0\end{aligned}$$

and

$$\begin{aligned}a^\intercal D^\intercal D a = 0 &\Leftrightarrow \|Da\|^2 = 0 \\ &\Leftrightarrow Da = 0 \\ &\Rightarrow D^\intercal D a = 0 \\ &\Rightarrow a = 0.\end{aligned}$$

It follows that $\hat{\theta}_n = (D^\intercal D)^{-1} D^\intercal Y$ is the unique minimizer of (the strictly convex function) Q_n. □

4.4. **Properties of the LSE.** In what follows we assume $E[\epsilon\epsilon^\intercal] = \sigma^2 \mathbb{1}_n$. In other words $E[\epsilon_i^2] = \mathrm{Var}(\epsilon_i) = \sigma^2$ for $i \in \{1, \ldots, n\}$ and $E[\epsilon_i \epsilon_j] = 0 \forall i \neq j \in \{1, \ldots, n\}$.

Proposition 4.3. *Assume that $D^\intercal D$ is invertible. Then,*

(i) $E[\hat{\theta}_n] = \theta$ *and*
(ii) $E[(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)^\intercal] = \sigma^2 (D^\intercal D)^{-1}$.

Proof. (i) Use (9) to see that

$$\begin{aligned}\hat{\theta}_n &= (D^\intercal D)^{-1} D^\intercal Y \\ &= (D^\intercal D)^{-1} D^\intercal (D\theta + \epsilon) \\ &= (D^\intercal D)^{-1} D^\intercal D\theta + (D^\intercal D)^{-1} D^\intercal \epsilon \\ &= \theta + (D^\intercal D)^{-1} D^\intercal \epsilon \end{aligned} \tag{12}$$

Since $E[\epsilon] = 0$ (i) follows.
(ii) Use (12) to see that

$$\begin{aligned}E\left[(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)^\intercal\right] &= E\left[(D^\intercal D)^{-1} D^\intercal \epsilon\epsilon^\intercal D (D^\intercal D)^{-1}\right] \\ &= (D^\intercal D)^{-1} D^\intercal E[\epsilon\epsilon^\intercal] D (D^\intercal D)^{-1} \\ &= (D^\intercal D)^{-1} D^\intercal \sigma^2 \mathbb{1}_n D (D^\intercal D)^{-1} \\ &= \sigma^2 (D^\intercal D)^{-1} D^\intercal D (D^\intercal D)^{-1} \\ &= \sigma^2 (D^\intercal D)^{-1}\end{aligned}$$

□

Proposition 4.4. *Let us assume that $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$. Then,*

(i) $\hat{\theta}_n \sim \mathcal{N}(\theta, \sigma^2 (D^\intercal D)^{-1})$.
(ii) $Y - D\hat{\theta}_n$ *and* $D(\hat{\theta}_n - \theta)$ *are independent Gaussian vectors.*
(iii) $\frac{\|Y - D\hat{\theta}_n\|^2}{\sigma^2} \sim \chi^2_{(n-d)}$ *and* $\frac{\|D(\hat{\theta}_n - \theta)\|^2}{\sigma^2} \sim \chi^2_{(d)}$.

Proof. (i) Recall that D is the design matrix and $Y = D\theta + \epsilon$. Then,

$$\begin{aligned}\hat{\theta}_n = (D^\intercal D)^{-1} D^\intercal Y &= (D^\intercal D)^{-1} D^\intercal (D\theta + \epsilon) \\ &= \theta + (D^\intercal D)^{-1} D^\intercal \epsilon\end{aligned}$$

whereas $(D^\intercal D)^{-1}D^\intercal$ is a matrix and ϵ is a gaussian vector. This means that $\hat{\theta}_n$ is also a gaussian vector with $E[\hat{\theta}_n] = \theta + 0 = \theta$ and covariance matrix $E[(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)^\intercal] = \sigma^2(D^\intercal D)^{-1}$ hence $\hat{\theta}_n \sim \mathcal{N}(\theta, \sigma^2(D^\intercal D)^{-1})$.
(ii) We want to show that $Y - D\hat{\theta}_n \perp\!\!\!\perp D(\hat{\theta}_n - \theta)$ whereas $Y - D\hat{\theta}_n$ denotes the estimated residuals.

$$\begin{aligned} D(\hat{\theta}_n - \theta) &= D\left((D^\intercal D)^{-1}D^\intercal Y - \theta)\right) \\ &= D\left((D^\intercal D)^{-1}D^\intercal(D\theta + \epsilon) - \theta\right) \\ &= A\epsilon \end{aligned}$$

Note that $A^\intercal = A$ and

$$\begin{aligned} A^2 &= D(D^\intercal D)^{-1}D^\intercal D(D^\intercal D)^{-1}D^\intercal \\ &= D(D^\intercal D)^{-1}D^\intercal \\ &= A \end{aligned}$$

On the other hand

$$\begin{aligned} Y - D\hat{\theta}_n &= D\theta + \epsilon - D(D^\intercal D)^{-1}D^\intercal(D\theta + \epsilon) \\ &= \epsilon - D(D^\intercal D)^{-1}D^\intercal\epsilon \\ &= (\mathbb{1} - A)\epsilon. \end{aligned}$$

$\mathbb{1} - A$ is symmetric and satisfies $(\mathbb{1} - A)^2 = (\mathbb{1} - A)(\mathbb{1} - A) = \mathbb{1} - A - A + A^2 = \mathbb{1} - A$. Furthermore, $(\mathbb{1} - A)A = A - A^2 = 0 = A(\mathbb{1} - A)$ and $\text{rank}(A) = d$ because $D^\intercal D$ is invertible (see in the notes on linear algebra) which implies that $\text{rank}(\mathbb{1} - A) = n - d$. Using Cochran's theorem, it follows that $Y - D\hat{\theta}_n \perp\!\!\!\perp D(\hat{\theta}_n - \theta)$ and

$$\begin{aligned} \frac{\|D(\hat{\theta}_n - \theta)\|^2}{\sigma^2} &= \left\|A\frac{\epsilon}{\sigma}\right\|^2 \sim \chi^2_{(\text{rank}(A))} \overset{d}{=} \chi^2_{(d)} \\ \frac{\|Y - D\hat{\theta}_n\|^2}{\sigma^2} &= \left\|(\mathbb{1}_n - A)\frac{\epsilon}{\sigma}\right\|^2 \sim \chi^2_{(n-d)}, \end{aligned}$$

which is also proof for (iii). □

Proposition 4.5. *Consider the linear regression model $Y = D\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbb{1}_n)$. Consider also the testing problem*

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0. \tag{13}$$

If $\sigma = \sigma_0$ is known then a test of level α for this problem is given by

$$\Phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \frac{\|D(\hat{\theta}_n - \theta_0)\|^2}{\sigma_0^2} > q_{d,1-\alpha} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $q_{d,1-\alpha}$ is the $(1 - \alpha)$ quantile of $\chi^2_{(d)}$.

Proof. Under H_0, we know from ((ii)) that $\frac{\|D(\hat{\theta}_n - \theta_0)\|^2}{\sigma_0^2} = \chi^2_{(d)}$ so that $P\left(\frac{\|D(\hat{\theta}_n - \theta_0)\|^2}{\sigma_0^2} > q_{d,1-\alpha}\right) = \alpha$. □

Proposition 4.6. *Let $Y = D\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbb{1}_n)$ and consider the problem* (13). *Suppose σ is known. Then a test of level α for this problem is given by*

$$\Phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \frac{\|D(\hat{\theta}_n - \theta_0)\|/d}{\|Y - D\hat{\theta}_n\|^2/(n-d)} > q_{d,n-d,1-\alpha} \\ 0 & \text{otherwise} \end{cases}$$

where $q_{d,n-d,1-\alpha}$ is the $(1 - \alpha)$ *quantile of the F-distribution* (5) *of d and $n - d$ degrees of freedom.*

Proof.

$$\frac{\frac{\|D(\hat{\theta}_n - \theta_0)\|^2}{\sigma^2}/d}{\frac{\|Y - D\hat{\theta}_n\|^2}{\sigma^2}/(n-d)} \sim F_{(d,n-d)}$$

under H_0 because$\|D(\hat{\theta}_n - \theta_0)\|^2 \perp\!\!\!\perp \|Y - D\hat{\theta}_n\|^2$, ((ii)) and ((iii)). □

4.5. χ^2**- and F-tests for variable selection.** The question we want to answer is: Which of the covariates are significant (have a non-trivial effect on the response). More formally, the question can be put in the context of testing. We want a test where θ is of the form $(\theta_1, \dots, \theta_{d-m}, 0, \dots, 0)^\intercal$. Even more formally, we want to test

$$H_0 : G\theta = 0 \quad \text{versus} \quad H_1 : G\theta \neq 0$$

where $G = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$ and $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$. Note that H_1 means that there exists $j \in \{d-m+1, \dots, d\}$ $\theta_j \neq 0$ and

$$G\theta = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & 0 & \ddots & \ddots & \vdots \\ & & & & & & \\ & & & & & & \\ \vdots & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{d-m} \\ \theta_{d-m+1} \\ \vdots \\ \theta_d \end{pmatrix} = \begin{pmatrix} \theta_{d-m+1} \\ \vdots \\ \\ \\ \vdots \\ \theta_d \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \\ \\ \vdots \\ 0 \end{pmatrix}.$$

4.5.1. *LRT for variable selection.* Let us assume that $\epsilon \sim \mathcal{N}(0, \sigma_0^2 \mathbb{1}_n)$ where σ_0^2 is known.

$$\Theta_0 = \{\theta \in \mathbb{R}^d : G\theta = 0\} = \{\theta \in \mathbb{R}^d \theta_{d-m+1} = \dots = \theta_d = 0\}$$

$$\Theta = \mathbb{R}^d$$

$$L_n(\theta) = \frac{1}{(2\pi)^{n/2}\,\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \theta^\intercal x_i)^2\right) = \frac{1}{(2\pi)^{n/2}\,\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}\|Y - D\theta\|^2\right)$$

$$l_n(\theta) = -n/2\log(2\pi) - n\log(\sigma_0) - 1/(2\sigma_0)\|Y - D\theta\|^2.$$

Maximizing $\theta \mapsto l_n(\theta)$ over $\mathbb{R}^d$ is equivalent to minimizing $\theta \mapsto \|Y - D\theta\|^2$ over $\mathbb{R}^d$. We know that the solution is the LSE (**??**). Hence $\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\theta\in\mathbb{R}^d} L_n(\theta) = L_n(\hat{\theta}_n)$.
Now, we need to maximize $\theta \mapsto l_n(\theta)$ over Θ_0. But this is equivalent to minimize $\theta \mapsto \|Y - D\theta\|^2$ over Θ_0. Under H_0 we have

$$\begin{aligned} D\theta &= \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \\ \vdots & & \vdots \\ x_{i1} & \dots & x_{id} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{d-m} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} x_{11} & \dots & x_{1(d-m)} \\ \vdots & & \vdots \\ x_{i1} & \dots & x_{i(d-m)} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{n(d-m)} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \vdots \\ \vdots \\ \theta_{d-m} \end{pmatrix} \\ &= \tilde{D}\tilde{\theta}. \end{aligned} \tag{15}$$

This problem is equivalent to minimizing $\tilde{\theta} \mapsto \|Y - \tilde{D}\tilde{\theta}\|^2$. We only need to check that $\tilde{D}^\intercal \tilde{D}$ is invertible. Note that $\tilde{D} = D\tilde{G}$ with $\tilde{G} = \begin{pmatrix} \mathbb{1}_{d-m} \\ 0_m \end{pmatrix}$. Let $a \in \mathbb{R}^{d-m}$. We want to show that $\tilde{D}^\intercal \tilde{D}a = 0$ implies $a = 0$.

$$\begin{aligned}
\tilde{D}^\intercal \tilde{D}a = 0 &\Rightarrow a^\intercal \tilde{D}^\intercal \tilde{D} = 0 \\
&\Leftrightarrow (\tilde{D}a)^\intercal \tilde{D}a = \|\tilde{D}a\|^2 = 0 \\
&\Leftrightarrow \tilde{D}a = 0 \\
&\Leftrightarrow D\tilde{G}a = 0 \\
&\Leftrightarrow Db = 0
\end{aligned}$$

because $D^\intercal D$ is invertible if and only if $\operatorname{rank}(D) = d$. Hence $\tilde{G}a = 0$ if and only if $a = 0$. $\tilde{D}^\intercal \tilde{D}$ is invertible and therefore we are in the same setting as in the least squares problem. Hence the minimizer of $\tilde{\theta} \mapsto \|Y - \tilde{D}\tilde{\theta}\|^2$ is given by $(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal Y$ if and only if the minimizer of $\theta \mapsto \|Y - D\theta\|^2$ under H_0 is given by $\hat{\theta}_n^0 = \begin{pmatrix} (\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal Y \\ 0_m \end{pmatrix}$. $\theta \mapsto l_n(\theta)$ is maximized by $\hat{\theta}_n^0$ under H_0 and

$$\begin{aligned}
\Lambda_n &= \frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)} \\
&= \frac{\frac{1}{(2\pi)^{n/2}\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}\|Y - D\hat{\theta}_n\|^2\right)}{\frac{1}{(2\pi)^{n/2}\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}\|Y - D\hat{\theta}_n^0\|^2\right)} \\
&= \exp\left[\frac{1}{2\sigma_0^2}\left(\|Y - D\hat{\theta}_n^0\|^2 - \|Y - D\hat{\theta}_n\|^2\right)\right].
\end{aligned}$$

We reject if Λ_n is 'large' which means that if $\|Y - D\hat{\theta}_n^0\|^2 - \|Y - D\hat{\theta}_n\|^2$ is large.

$$\begin{aligned}
\|Y - D\hat{\theta}_n^0\|^2 &= \|Y - D\hat{\theta}_n + D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2 \\
&= \|Y - D\hat{\theta}_n\|^2 + 2(Y - D\hat{\theta}_n)^\intercal D(\hat{\theta}_n - \hat{\theta}_n^0) + \|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2.
\end{aligned}$$

Now we show that $2(Y - D\hat{\theta}_n)^\intercal D(\hat{\theta}_n - \hat{\theta}_n^0) = 0$. We know that $\hat{\theta}_n$ is a zero of the gradient of the function $Q_n(t) = \|Y - Dt\|^2$, $t \in \mathbb{R}^d$. In other words

$$\begin{aligned}
D^\intercal D\hat{\theta}_n - D^\intercal Y = 0 &\Leftrightarrow D^\intercal(D\hat{\theta}_n - Y) = 0 \\
&\Leftrightarrow (Y - D\hat{\theta}_n)^\intercal D = 0 \\
&\Leftrightarrow (Y - D\hat{\theta}_n)^\intercal Dv = 0
\end{aligned}$$

for all $v \in \mathbb{R}^d$. In particular this holds true for $v = \hat{\theta}_n - \hat{\theta}_n^0$. Λ_n is 'large' if and only if $\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2$ is 'large'. What is the distribution of $\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2$ under H_0?

4.5.2. *The LRT for variable selection.* $\sigma = \sigma_0$ is known.

$$\begin{aligned}
\Lambda_n \text{ 'is large'} &\Leftrightarrow \|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2 \text{ 'is large'} \\
&\Leftrightarrow \frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2}{\sigma_0^2} \text{ 'is large'}
\end{aligned}$$

where $\hat{\theta}_n = (D^\intercal D)^{-1}D^\intercal Y$ and $\hat{\theta}_n^0 = (\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal Y$

Question: What is the distribution of $\frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2}{\sigma_0^2}$ under $H_0 : G\theta = 0$?

$$\begin{aligned}
D(\hat{\theta}_n - \hat{\theta}_n^0) &= D(\hat{\theta}_n - \theta) - D(\hat{\theta}_n^0 - \theta) \\
&= \Big(\underbrace{D(D^\intercal D)^{-1}D^\intercal}_{=:A} - \underbrace{\tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal}_{=:B}\Big)\epsilon
\end{aligned}$$

whereas $Y = D\theta + \epsilon = \tilde{D}\tilde{\theta} + \epsilon$ under H_0 and $\epsilon \sim \mathcal{N}(0, \sigma_0 \mathbb{1}_n)$. Recall 15 and observe that

$$\begin{aligned} AB &= D(D^\intercal D)^{-1} D^\intercal \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal \\ &= D(D^\intercal D)^{-1} D^\intercal D\tilde{G}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal \\ &= D\tilde{G}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal \\ &= \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal \\ &= B \end{aligned}$$

and

$$\begin{aligned} BA &= \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal D(D^\intercal D)^{-1} D^\intercal \\ &= \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{G}^\intercal D^\intercal D(D^\intercal D)^{-1} D^\intercal \\ &= \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{G}^\intercal D^\intercal \\ &= \tilde{D}(\tilde{D}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal. \\ &= B \end{aligned}$$

I.e. $BA = AB$ if and only if A and B commute ($A^\intercal = A$ and $B^\intercal = B$). Furthermore, the matrices are projections meaning $A^2 = A$ and $B^2 = B$. Hence, we can find an orthogonal matrix P such that

$$A = P^\intercal \begin{pmatrix} \mathbb{1}_d & 0 \\ 0 & 0 \end{pmatrix} P \text{ and } B = P^\intercal \begin{pmatrix} \mathbb{1}_{d-m} & 0 \\ 0 & 0 \end{pmatrix} P$$

because $\operatorname{rank}(A) = \operatorname{rank}(D^\intercal D) = d$ and $\operatorname{rank}(B) = \operatorname{rank}(\tilde{D}^\intercal \tilde{D})$ (see notes on linear algebra). Moreover

$$A - B = P^\intercal \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbb{1}_m & 0 \\ 0 & 0 & 0 \end{pmatrix} P$$

which implies $\operatorname{rank}(A - B) = m$. Hence we can write $\frac{\|D(\hat{\theta}_n - \theta_0)\|}{\sigma_0^2}^2 = \|(A - B)\frac{\epsilon}{\sigma_0}\|^2$ with $\frac{\epsilon}{\sigma_0} \sim \mathcal{N}(0, \mathbb{1}_n)$. Using Cochran's theorem, it follows that $\|(A - B)\frac{\epsilon}{\sigma_0}\|^2 \sim \chi^2_{\operatorname{rank}(A-B)}$, that is under H_0 $\frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|}{\sigma_0^2} \sim \chi^2_{(m)}$ with $\hat{\theta}_n^0 = \begin{pmatrix} (\tilde{\theta}^\intercal \tilde{D})^{-1}\tilde{D}^\intercal_{d-n} \\ 0_m \end{pmatrix}$ The LRT of level α can be given by

$$\Phi(Y_1, \ldots, Y_n) = \begin{cases} 1 & \text{if } \frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|}{\sigma_0^2} > q_{m,1-\alpha} \\ 0 & \text{otherwise} \end{cases}$$

with $q_{m,1-\alpha} = (1 - \alpha)$-quantile of $\chi^2_{(m)}$.

σ is unknown

The likelihood is

$$L_n = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|Y - D\theta\|^2\right)$$

with

$$\Theta = \{(\theta, \sigma) \in \mathbb{R}^d \times (0, +\infty)\} = \mathbb{R}^d \times (0, +\infty)$$

and

$$\begin{aligned} \Theta_0 &= \{(\theta, \sigma) : G\theta = 0 \text{ and } \sigma \in (0, +\infty)\} \\ &= \{\theta \in \mathbb{R}^d : \theta_{d-m+1} = \cdots = \theta_d = 0\} \times (0, +\infty). \end{aligned}$$

The log-likelihood is

$$l_n(\theta) = -n/2 \log(2\pi) - n \log(\sigma) - 1/(2\sigma^2)\|Y - D\theta\|^2.$$

To maximize $(\theta, \sigma) \mapsto l_n(\theta, \sigma)$ over Θ we can use the profiling approach:

- Fix $\sigma \in (0, +\infty)$ and maximize $\theta \mapsto l_n(\theta, \sigma)$ over $\mathbb{R}^d$. It is clear, for a fixed σ, the solution $\hat{\theta}_n$ is the one minimizing $\theta \mapsto \|Y - D\theta\|^2$ on $\mathbb{R}^d$, that is (11) the LSE.
- We plug the obtained solution $\hat{\theta}_n$ and maximize the function

$$\sigma \mapsto l_n(\hat{\theta}_n, \sigma) = -n/2 \log(2\pi) - n \log(\sigma) - 1/(2\sigma^2)\|Y - D\hat{\theta}\|^2$$

$$\begin{aligned}\frac{d}{d\sigma}l_n(\hat{\theta}_n,\sigma) &= -n/\sigma + 1/(\sigma^3)\|Y - D\hat{\theta}_n\|^2 = 0\\ &\Leftrightarrow \sigma^2 = 1/n\|Y - D\hat{\theta}_n\|^2\\ &\Leftrightarrow \sigma = 1/\sqrt{n}\|Y - D\hat{\theta}_n\|^2\end{aligned}$$

whereas σ is the unique critical point of $\sigma \mapsto l_n(\hat{\theta}_n,\sigma)$.

$$\begin{aligned}\frac{d^2}{d\sigma^2}l_n(\hat{\theta}_n,\sigma)|_{\sigma=\hat{\sigma}_n} &= -n/\hat{\sigma}^2 - 3/\hat{\sigma}^4\|Y - D\hat{\theta}_n\|^2\\ &= -n/\hat{\sigma}^2 - 3/\hat{\sigma}^4 n\hat{\sigma}_n^2\\ &= -\frac{2n}{\hat{\sigma}_n^2} < 0.\end{aligned}$$

Using the same arguments as for example b (for testing the mean of a Gaussian with unknown variance) we can show that $\hat{\sigma}_n$ gives the global maximum and also that

$$\sup_{(\theta,\sigma)\in\Theta} l_n(\theta,\sigma) = l_n(\hat{\theta}_n,\hat{\sigma}_n) \Leftrightarrow \sup_{(\theta,\sigma)\in\Theta} L_n(\theta,\sigma) = L_n(\hat{\theta}_n,\hat{\sigma}_n).$$

Now we need to find $\sup_{(\sigma,\theta)\in\Theta_0} L_n(\sigma,\theta)$. Similar arguments can be used to show that $\sup_{(\sigma,\theta)\in\Theta_0} L_n(\sigma,\theta) = L_n(\hat{\theta}_n^0,\hat{\sigma}_n^0)$ with $\sigma_n^0 = \begin{pmatrix}(\tilde{D}^\intercal\tilde{D})^{-1}\tilde{D}^\intercal Y\\ 0_m\end{pmatrix}$ and $\hat{\sigma}_n^0 = \frac{1}{\sqrt{n}}\|Y - D\hat{\theta}_n^0\|$.

$$\begin{aligned}\Lambda_n &= \frac{\sup_{(\theta,\sigma)\in\Theta} L_n(\theta,\sigma)}{\sup_{(\theta,\sigma)\in\Theta_0} L_n(\theta,\sigma)}\\ &= \frac{L_n(\hat{\theta}_n,\hat{\sigma}_n)}{L_n(\hat{\theta}_n^0,\hat{\sigma}_n^0)}\\ &= \frac{\frac{1}{(2\pi)^{n/2}\hat{\sigma}^n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\|Y - D\hat{\theta}_n\|\right)}{\frac{1}{(2\pi)^{n/2}}\frac{1}{(\hat{\sigma}_n^0)^n}\exp\left(-\frac{1}{2(\hat{\sigma}_n^0)}\|Y - D\hat{\theta}_n^0\|\right)}\\ &= \left(\frac{\hat{\sigma}_n^0}{\hat{\sigma}_n}\right)^n\\ &= \left(\frac{\left(\hat{\sigma}_n^0\right)^2}{\hat{\sigma}_n^2}\right)^{n/2}\end{aligned}$$

$$\begin{aligned}\Lambda_n \text{ 'is large' } &\Leftrightarrow \frac{\left(\hat{\sigma}_n^0\right)^2}{\hat{\sigma}_n^2}\text{ 'is large'}\\ &\Leftrightarrow \frac{1/n\|Y - D\hat{\theta}_n^0\|^2}{1/n\|Y - D\hat{\theta}_n\|^2}\text{ 'is large'.}\end{aligned}$$

$$\|Y - D\hat{\theta}_n^0\|^2 = \|Y - D\hat{\theta}_n\|^2 + 2\underbrace{(Y - D\hat{\theta}_n)^\intercal D(\hat{\theta}_n - \hat{\theta}_n^0)}_{=0} + \|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2$$

$$\begin{aligned}\Lambda_n \text{ 'is large' } &\Leftrightarrow 1 + \frac{\|Y - D\hat{\theta}_n^0\|^2}{\|Y - D\hat{\theta}_n\|^2}\text{ 'is large'}\\ &\Leftrightarrow \frac{\|Y - D\hat{\theta}_n^0\|^2}{\|Y - D\hat{\theta}_n\|^2}\text{ 'is large'.}\end{aligned}$$

We know that $D(\hat{\theta}_n - \hat{\theta}_n^0) = (A - B)\epsilon$. Also $Y - D\hat{\theta}_n = D\theta + \epsilon - D(D^\intercal D)^{-1}D^\intercal(D\theta + \epsilon) = (\mathbb{1}_n - A)\epsilon$.

$$\begin{aligned}(A - B)(\mathbb{1}_n - A) &= A - B - (A - B)A\\ &= A - B - (A - B) = 0\end{aligned}$$

and similarly $(\mathbb{1}_n - A)(A - B) = 0$. Also

$$\begin{aligned}(A-B)^2 &= (A-B)(A-B)\\ &= A^2 - AB - BA + B^2\\ &= A - B - B + B = A - B\end{aligned}$$

and

$$\begin{aligned}(\mathbb{1}_n - A)^2 &= (\mathbb{1}_n - A)(\mathbb{1}_n - A)\\ &= \mathbb{1}_n - A - A + A^2\\ &= \mathbb{1}_n - A.\end{aligned}$$

moreover we know $\operatorname{rank}(A-B) = m$ from previous calculations and $\operatorname{rank}(\mathbb{1} - A) = n - \operatorname{rank}(A) = n - d$. Using Cochran's theorem we have $D(\hat{\theta}_n - \hat{\theta}_n^0) \perp\!\!\!\perp Y - D\hat{\theta}_n$ and

$$\frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2}{\sigma^2} = \left\|(A-B)\frac{\epsilon}{\sigma}\right\|^2 \sim \chi^2_{(n)} \perp\!\!\!\perp \frac{\|Y - D\hat{\theta}_n\|^2}{\sigma^2} = \left\|(\mathbb{1}_n - A)\frac{\epsilon}{\sigma}\right\|^2 \sim \chi^2_{(n-d)}.$$

Hence, under H_0

$$\frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2}{\|Y - D\hat{\theta}_n\|^2} \sim F_{(m,n-d)}$$

with m and $n-d$ degrees of freedom. The LRT of level α is given by

$$\Phi(Y_1, \dots, Y_n) = \begin{cases} 1 & \text{if } \frac{\|D(\hat{\theta}_n - \hat{\theta}_n^0)\|^2}{\|Y - D\hat{\theta}_n\|^2} > q_{m,n-d,1-\alpha} \\ 0 & \text{otherwise,} \end{cases}$$

whereas $q_{m,n-d,1-\alpha}$ is the $(1-\alpha)$-quantile of $F_{(m,n-d)}$.

Email address: damark@math.uzh.ch