Hypothesis Testing

Lecture Notes for MATH 6263 at Georgia Tech

Cheng Mao

School of Mathematics, Georgia Tech

December 8, 2022

This set of notes is based on the books [LR06, Was04, Ros20, JN20, PW19, Efr12], lecture notes by Vladimir Koltchinskii, lecture notes by Emmanuel Candès, and other sources. It is provided to the students in the course MATH 6263 at Georgia Tech as a complement to the lectures. It is not meant to be a complete introduction to the subject.

Contents

Chapter 1

Fundamentals of hypothesis testing

1.1 Background and setup

1.1.1 Review of probability

Sample space Consider a sample space $\mathcal{X}$ containing all possible outcomes of an experiment. Let μ be the reference (or natural) measure on $\mathcal{X}$. We primarily consider the following spaces:

- A finite or countable set $\mathcal{X}$ equipped with the counting measure μ. For example, when we roll a die, the outcome lies in $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Moreover, if we consider the number of times we flip a coin before a "heads" is observed, then this number lies in $\mathcal{X} = \{1, 2, 3, \dots\}$.
- $\mathcal{X} = \mathbb{R}^d$ equipped with the Lebesgue measure μ. For example, tomorrow's temperature is in $\mathcal{X} = \mathbb{R}$, while tomorrow's temperature and humidity jointly lie in $\mathcal{X} = \mathbb{R}^2$.

Random variable and distribution A random variable X is an experiment taking values in $\mathcal{X}$. We write $X \sim \mathcal{P}$ if X follows a distribution $\mathcal{P}$. There are several ways to describe a random variable or a distribution:

- If X is discrete, i.e., $\mathcal{X}$ is finite or countable, we can specify the probability mass function (PMF) f_X of X. For example, for the uniform random variable $X \sim \mathsf{Unif}([n])$ where $[n] := \{1, \dots, n\}$, we have $f_X(i) = \mathbb{P}\{X = i\} = 1/n$ for $i = 1, \dots, n$.
- If X is continuous, e.g., $X = \mathbb{R}$ or $\mathbb{R}^d$, we can specify the probability density function (PDF or density) f_X of X. For example, for the standard Gaussian random variable $X \sim \mathsf{N}(0, 1)$, we have $f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ for $t \in \mathbb{R}$.
- The cumulative distribution function (CDF) of a random variable X on $\mathbb{R}$ is $F_X(t) = \mathbb{P}\{X \le t\}$. We have $F_X'(t) = f_X(t)$ and $\int_{-\infty}^{t} f_X(s)\, ds = F_X(t)$. The CDF of $X = (X_1, \dots, X_d)$ on $\mathbb{R}^d$ is $F_X(t_1, \dots, t_d) = \mathbb{P}\{X_1 \le t_1, \dots, X_d \le t_d\}$.

Event, probability, and expectation In general, for a subset $E \subset \mathcal{X}$, the probability of the event $\{X \in E\}$ is $\mathbb{P}\{X \in E\} = \int_E f_X \, d\mu$. Examples:

- Roll a die; the outcome is $X \sim \mathsf{Unif}([6])$. The probability of seeing 2 or 3 is $\mathbb{P}\{X \in \{2, 3\}\} = \sum_{i=2}^{3} 1/6 = 1/3$.

- Consider $X \sim \mathsf{N}(0,1)$. The probability that X is positive is $\mathbb{P}\{X > 0\} = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 1/2$.

The expectation of X is $\mathbb{E}[X] = \int_{\mathcal{X}} t\, f_X(t)\, d\mu(t)$. Given a function $g : \mathcal{X} \to \mathbb{R}$, the expectation of $g(X)$ is $\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(t) f_X(t)\, d\mu(t)$. Examples:

- For $X \sim \mathsf{Unif}([6])$, $\mathbb{E}[X] = \sum_{i=1}^6 i \cdot \frac{1}{6} = 3.5$.
- For $X \sim \mathsf{N}(0,1)$, the variance of X is $\mathbb{E}[(X-0)^2] = \int_{-\infty}^\infty t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 1$.

1.1.2 Setup of statistical hypothesis testing

Statistics is in some sense the reverse engineering of probability. Observing a set of data $X = (X_1, X_2, \ldots, X_n)$, we aim to say something about the underlying distribution that generates the data. Let us describe the basic setup of hypothesis testing using a biased coin flip as a running example. Consider a biased coin for which we see 1 (heads) with probability $\theta \in [0,1]$ and see 0 (tails) with probability $1-\theta$. In other words, the observation follows the $\mathsf{Ber}(\theta)$ distribution. The following is a list of basic concepts in (parametric) statistics:

- Parameter: θ, which is typically a real number. E.g., $\theta = 0.3, 0.5$, or 0.8.
- Parameter space: the set Θ of parameters. E.g., $\Theta = [0,1]$.
- Probability distribution: $\mathcal{P}_\theta$. E.g., $\mathcal{P}_\theta = \mathsf{Ber}(\theta)$.
- Observation: $X = (X_1, \ldots, X_n)$, where we have i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}_\theta$. E.g., $X_1, \ldots, X_n$ are the binary outcomes of n independent coin flips.
- Family of distributions: the set $\mathscr{P}$ containing all $\mathcal{P}_\theta$. E.g., $\mathscr{P} = \{\mathsf{Ber}(\theta) : \theta \in [0,1]\}$.

Observing the data $X \sim \mathcal{P}_\theta^{\otimes n}$ (or simply $X \sim \mathcal{P}_\theta$), statistical hypothesis testing consists in testing between two hypotheses:

- Null hypothesis, H_0: $\theta \in \Theta_0$ for a subset $\Theta_0 \subset \Theta$;
- Alternative hypothesis, H_1: $\theta \in \Theta_1$ where $\Theta_1 := \Theta \setminus \Theta_0$.

For coin flips, an example is $\Theta_0 = [0, 0.5]$.

1.1.3 Non-randomized and randomized tests

A non-randomized test $\psi = \psi(X)$ is a function of X taking values in $\{0,1\}$, where

- $\psi(X) = 0$ means that we accept the null hypothesis H_0;
- $\psi(X) = 1$ means that we reject the null hypothesis H_0.

Since X is a random variable, $\psi(X)$ is also a random variable, and, clearly, it has to be a Bernoulli random variable. For coin flips, an example is $\psi(X) = \mathbb{1}\{\bar{X} > 0.5\}$ where $\bar{X} := (X_1 + \cdots X_n)/n$.

Defining a non-randomized test is equivalent to specifying the following two regions:

- Region of acceptance, S_0: $\psi(X) = 0$ if and only if $X \in S_0$;

- Region of rejection or critical region, S_1: $\psi(X) = 1$ if and only if $X \in S_1$.

For coin flips, an example is $S_0 = \{x \in \{0,1\}^n : \bar{x} \le 0.5\}$ where $\bar{x} = (x_1 + \cdots + x_n)/n$.

A randomized test can be defined through a function $\phi = \phi(X)$ of X taking values in $[0,1]$: Given X, we

- accept the null hypothesis H_0 with probability $1 - \phi(X)$;
- reject the null hypothesis H_0 with probability $\phi(X)$.

Therefore, a randomized test is a generalization of the non-randomized version, and the test is a $\mathsf{Ber}(\phi(X))$ random variable conditional on X.

1.1.4 Level of significance, power, and two types of errors

When designing and analyzing a (non-randomized) test ψ, we usually select a number $\alpha \in (0,1)$, called the level of significance (also known as significance, statistical significance, or significance level), such that

$$\mathbb{P}_\theta\{\psi(X) = 1\} = \mathbb{P}_\theta\{X \in S_1\} \le \alpha \quad \text{for all } \theta \in \Theta_0. \tag{1.1}$$

In other words, α is an upper bound on the size of the critical region under the probability $\mathbb{P}_\theta$ for any $\theta \in \Theta_0$. For example, when $\alpha = 0.05$, the above condition says that, whenever H_0 holds, the test ψ gives 1 with probability at most 0.05, i.e., ψ gives 0 with probability at least 0.95.

Furthermore, for any $\theta \in \Theta_1 = \Theta \setminus \Theta_0$, the power (function) of the test ψ against the alternative θ is defined as

$$\beta(\theta) := \mathbb{P}_\theta\{\psi(X) = 1\} = \mathbb{P}_\theta\{X \in S_1\}.$$

We typically would like to maximize the power $\beta(\theta)$ over all alternatives $\theta \in \Theta_1$ subject to the level of significance (1.1).

Note that a test ψ can make two types of errors:

- Type I error (false positive): reject H_0 when H_0 is true;
- Type II error (false negative): accept H_0 when H_1 is true.

In terms of the two types of errors, we are interested in minimizing the type II error $1 - \beta(\theta)$ subject to the constraint that the type I error is no larger than α.

The definitions of significance and power can be generalized to a randomized test ϕ by replacing $\mathbb{P}_\theta\{\psi(X) = 1\}$ with $\mathbb{E}_\theta[\phi(X)]$. That is, we would like to maximize the power

$$\beta(\theta) := \mathbb{E}_\theta[\phi(X)] \quad \text{for } \theta \in \Theta_1$$

subject to the level of significance

$$\mathbb{E}_\theta[\phi(X)] \le \alpha \quad \text{for } \theta \in \Theta_0.$$

1.2 Neyman–Pearson lemma

1.2.1 Simple hypothesis testing

Let us consider the simplest case where $\Theta = \{0, 1\}$, $\Theta_0 = \{0\}$, and $\Theta_1 = \{1\}$. That is, we are testing between two simple hypotheses

- $H_0 : X \sim \mathcal{P}_0$;
- $H_1 : X \sim \mathcal{P}_1$.

Here a simple hypothesis means that there is a single distribution associated to it. On the other hand, if the associated set of distributions contains multiple elements, then it is called a composite hypothesis, which we will study later.

Recall that designing a non-randomized test ψ is equivalent to specifying the critical region S_1 (i.e., the region of rejection), which is a subset of the sample space $\mathcal{X}$. Let us revisit our goal: to maximize the power of the test subject to a certain level of significance, i.e.,

$$\max_{S_1 \subset \mathcal{X}} \mathbb{P}_1\{X \in S_1\} \quad \text{s.t. } \mathbb{P}_0\{X \in S_1\} \leq \alpha.$$

This formulation offers the geometric intuition that we are maximizing the size of S_1 under the probability distribution $\mathbb{P}_1$ under the constraint that its size is no larger than α under $\mathbb{P}_0$.

For theory, we consider directly the more general case of a (possibly) randomized test ϕ and study the problem

$$\max_{\phi} \mathbb{E}_1[\phi(X)] \quad \text{s.t. } \mathbb{E}_0[\phi(X)] \leq \alpha.$$

What is the most powerful test ϕ^* that solves the above optimization problem?

1.2.2 Likelihood-ratio test

Let p_0 and p_1 denote the densities of $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively, with respect to a reference measure μ. The likelihood ratio (statistic) is defined as

$$L(X) := \frac{p_1(X)}{p_0(X)}.$$

For $c > 0$ and $\gamma \in [0, 1]$, define

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

The randomized test $\phi_{c,\gamma}$ is called the likelihood-ratio test or the Neyman–Pearson test. When X is a continuous variable, we usually have $L(X) = c$ with probability zero, so the test becomes non-randomized. The following theorem shows that, if c and γ are chosen appropriately, then the test $\phi_{c,\gamma}$ is the most powerful test that we are aiming for.

Theorem 1.1 (Neyman–Pearson lemma)**.** *For any level of significance $\alpha \in (0, 1)$, the following statements hold:*

1. *There exists $c > 0$ and $\gamma \in [0, 1]$ such that $\mathbb{E}_0[\phi_{c,\gamma}(X)] = \alpha$.*

2. *For such a choice of c and γ, the test $\phi_{c,\gamma}$ maximizes the power $\beta_\phi := \mathbb{E}_1[\phi(X)]$ among all tests ϕ such that $\mathbb{E}_0[\phi(X)] \le \alpha$. In other words, $\phi_{c,\gamma}$ is the most powerful test at significance level α.*

3. *If ϕ^* is a most powerful test at significance level α, then we have $\phi^*(x) = \phi_{c,\gamma}(x)$ on the set $\{x : L(x) \neq c\}$ μ-almost everywhere. Moreover, we have $\mathbb{E}_0[\phi^*(X)] = \alpha$ unless $\mathbb{E}_1[\phi^*(X)] = 1$.*

Proof. 1. Let $F(t) := \mathbb{P}_0\{L(X) \le t\}$ be the CDF of the likelihood ratio $L(X)$ under the null hypothesis. Then

$$\mathbb{E}_0[\phi_{c,\gamma}(X)] = \mathbb{P}_0\{L(X) > c\} + \gamma\, \mathbb{P}_0\{L(X) = c\} = 1 - F(c) + \gamma(F(c) - F(c-))$$

where $F(c-) := \lim_{t \nearrow c} F(t)$. Consider two cases: (i) If there exists c such that $F(c) = 1 - \alpha$, then we set $\gamma = 0$. It follows immediately that $\mathbb{E}_0[\phi_{c,\gamma}(X)] = \alpha$. (ii) Otherwise, there must exist c such that $F(c-) \le 1-\alpha < F(c)$. Then we set $\gamma = \frac{F(c)-(1-\alpha)}{F(c)-F(c-)}$. It follows that $\mathbb{E}_0[\phi_{c,\gamma}(X)] = \alpha$.

2. Let ϕ be a test such that $\mathbb{E}_0[\phi(X)] \le \alpha$. By the definition of $\phi_{c,\gamma}$, it is easy to check that

$$\big(\phi_{c,\gamma}(x) - \phi(x)\big)\big(p_1(x) - c\, p_0(x)\big) \ge 0$$

for any $x \in \mathcal{X}$. Therefore,

$$\int_{\mathcal{X}} (\phi_{c,\gamma} - \phi)(p_1 - c\, p_0)\, d\mu \ge 0 \quad \Longleftrightarrow \quad \int_{\mathcal{X}} (\phi_{c,\gamma} - \phi) p_1\, d\mu \ge c \int_{\mathcal{X}} (\phi_{c,\gamma} - \phi) p_0\, d\mu.$$

In other words, we have

$$\mathbb{E}_1[\phi_{c,\gamma}(X)] - \mathbb{E}_1[\phi(X)] \ge c\big(\mathbb{E}_0[\phi_{c,\gamma}(X)] - \mathbb{E}_0[\phi(X)]\big).$$

Since $\mathbb{E}_0[\phi(X)] \le \alpha = \mathbb{E}_0[\phi_{c,\gamma}(X)]$, it follows that $\beta_{\phi_{c,\gamma}} - \beta_\phi \ge 0$.

3. Suppose that ϕ^* is the most powerful test at significance level α. Replacing ϕ with ϕ^* in the previous display yields

$$0 = \mathbb{E}_1[\phi_{c,\gamma}(X)] - \mathbb{E}_1[\phi^*(X)] \ge c\big(\mathbb{E}_0[\phi_{c,\gamma}(X)] - \mathbb{E}_0[\phi^*(X)]\big) \ge 0,$$

so the equality holds. As a result, all the inequalities in the previous part are in fact equalities, and we must have

$$\big(\phi_{c,\gamma}(x) - \phi^*(x)\big)\big(p_1(x) - c\, p_0(x)\big) = 0$$

μ-almost everywhere. Therefore, on the set $\{x : L(x) \neq c\} = \{x : p_1(x) \neq c\, p_0(x)\}$, we have $\phi^*(x) = \phi_{c,\gamma}(x)$ μ-almost everywhere.

Finally, since $c\big(\mathbb{E}_0[\phi_{c,\gamma}(X)] - \mathbb{E}_0[\phi^*(X)]\big) = 0$, we have either $\mathbb{E}_0[\phi^*(X)] = \alpha$ or $c = 0$. In the latter case,

$$\mathbb{E}_1[\phi^*(X)] = \mathbb{E}_1[\phi_{c,\gamma}(X)] \ge \mathbb{P}_1\{L(X) > 0\} \ge \mathbb{P}_1\{p_1(X) > 0\} = 1.$$

□

1.2.3 Examples

Gaussian Fix a nonzero vector $\mu \in \mathbb{R}^d$. Consider testing between simple hypotheses:

- $H_0 : X \sim \mathcal{N}(0, I_d)$;
- $H_1 : X \sim \mathcal{N}(\mu, I_d)$.

We have

$$p_0(x) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{\|x\|^2}{2}\Big), \qquad p_1(x) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{\|x-\mu\|^2}{2}\Big)$$

for $x \in \mathbb{R}^d$, so

$$\log L(x) = -\frac{\|x-\mu\|^2}{2} + \frac{\|x\|^2}{2} = \langle x, \mu\rangle - \frac{\|\mu\|^2}{2}.$$

Therefore, the condition $L(x) > c$ can be written equivalently as $\langle x, \mu\rangle > \tau$ for some τ depending on c. The likelihood-ratio test takes the form

$$\phi_\tau(X) = \begin{cases} 1 & \text{if } \langle X, \mu\rangle > \tau, \\ 0 & \text{if } \langle X, \mu\rangle \le \tau. \end{cases}$$

It remains to find τ such that $\mathbb{E}_0[\phi_\tau(X)] = \mathbb{P}_0\{\langle X, \mu\rangle > \tau\} = \alpha$. Then the above general theory guarantees that ϕ_τ is the most powerful test at significance level α. Towards this end, note that $\langle X, \mu\rangle \sim \mathcal{N}(0, \|u\|^2)$ and so

$$\mathbb{P}_0\{\langle X, \mu\rangle > \tau\} = \mathbb{P}\{Z > \tau/\|u\|\}$$

where Z is a standard Gaussian. Therefore, if z_α denotes the $(1-\alpha)$-quantile of Z, then $\tau = z_\alpha\|u\|$ satisfies that $\mathbb{P}_0\{\langle X, \mu\rangle > \tau\} = \alpha$.

Binomial Fix $\theta_0, \theta_1 \in [0, 1]$ such that $\theta_0 < \theta_1$. Consider testing between simple hypotheses:

- $H_0 : X \sim \mathsf{Bin}(n, \theta_0)$;
- $H_1 : X \sim \mathsf{Bin}(n, \theta_1)$.

We have

$$p_0(x) = \binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}, \qquad p_1(x) = \binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}$$

for $x = 0, 1, \ldots, n$, so

$$\log L(x) = x\log\frac{\theta_1}{\theta_0} + (n-x)\log\frac{1-\theta_1}{1-\theta_0} = x\log\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} + n\log\frac{1-\theta_1}{1-\theta_0}.$$

Since $\log\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} > 0$, the condition $L(x) > c$ can be written equivalently as $x > \tau$ for some τ depending on c. The likelihood-ratio test takes the form

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

If it holds that

$$\mathbb{P}_0\{X \leq k\} = \sum_{j=0}^{k} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} = 1 - \alpha$$

for some $k = 0, 1, \ldots, n$, then we can set $\tau = k$ and $\gamma = 0$ so that $\phi_{k,0}$ is the most powerful test at significance level α according to the above general theory. Otherwise, if

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

then we can set $\tau = k$ and

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

so that $\phi_{k,\gamma}$ is the most powerful test at significance level α.

1.2.4 Geometric intuition

For a simple hypothesis testing problem, we can gain some geometric intuition about the set of points (α, β) for which there exists a test ϕ such that $\mathbb{E}_0[\phi(X)] = \alpha$ and $\mathbb{E}_1[\phi(X)] = \beta$. See Figure 3.1 of [LR06]. In short, this convex subset of $[0,1]^2$ is symmetric around the center $(1/2, 1/2)$, and the most powerful tests correspond to the points on the upper boundary of the set.

1.3 UMP and unbiasedness

1.3.1 One-sided testing and uniformly most powerful tests

The Neyman–Pearson lemma shows that the most powerful test for a simple hypothesis testing problem is the likelihood-ratio test. We now turn to composite hypothesis testing between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$ given $X \sim \mathcal{P}_\theta$ for $\theta \in \Theta$. Subject to a level of significance α, if there is a test ϕ that maximizes the power for all alternatives $\theta \in \Theta_1$, then we call ϕ a uniformly most powerful (UMP) test. It is often difficult to find a UMP test, but the likelihood-ratio test works in the following special case.

Suppose that the parameter θ is real-valued. Given $X \sim \mathcal{P}_\theta$, for a fixed $\theta_0 \in \mathbb{R}$, we test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. We say that the densities $p_\theta(x)$ have monotone likelihood ratios if there exists a real-valued function $T(x)$ such that for any $\theta < \theta'$, we have $\mathcal{P}_\theta \neq \mathcal{P}_{\theta'}$ and $p_{\theta'}(x)/p_\theta(x)$ is an increasing function of $T(x)$. This is the case for the Gaussian and binomial examples above, where $T(x) = \langle x, \mu \rangle$ and $T(x) = x$ respectively. More generally, any exponential family with densities of the following form has monotone likelihood ratios:

$$p_\theta(x) = C(\theta) e^{Q(\theta)\, T(x)} h(x),$$

where Q is strictly monotone.

Theorem 1.2. *Consider the setting described above. For $\tau > 0$ and $\gamma \in [0,1]$, define the likelihood-ratio test*

$$\phi_{\tau,\gamma}(X) := \begin{cases} 1 & \text{if } T(X) > \tau, \\ \gamma & \text{if } T(X) = \tau, \\ 0 & \text{if } T(X) < \tau. \end{cases} \tag{1.2}$$

Then the following statements hold:

1. *The likelihood-ratio test $\phi_{\tau,\gamma}$ is UMP, where the constants τ and γ are determined by the condition $\mathbb{E}_{\theta_0}[\phi_{\tau,\gamma}(X)] = \alpha$.*
2. *The power function $\beta(\theta) = \mathbb{E}_\theta[\phi_{\tau,\gamma}(X)]$ is strictly increasing at all θ such that $\beta(\theta) \in (0,1)$.*
3. *For any $\theta < \theta_0$, the test $\phi_{\tau,\gamma}$ minimizes $\beta(\theta)$ among all tests ϕ such that $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.*

Proof. 1. Fix any $\theta_1 > \theta_0$. We can modify the proof of part 1 of Theorem 1.1 to obtain τ and γ. For example, analogous to case (i) in that proof, τ can be defined so that

$$1 - \alpha = \mathbb{P}_0\{T(X) \le \tau\},$$

and γ is set to zero. In case (ii), the test can be also defined in a similar way. Since $p_{\theta_1}(x)/p_{\theta_0}(x)$ is an increasing function of $T(x)$, the above quantity is equal to $F(c) = \mathbb{P}_0\{p_{\theta_1}(X)/p_{\theta_0}(X) \le c\}$ that appears in the proof of Theorem 1.1, where c depends on τ. As a result, the test $\phi_{\tau,\gamma}$ is the most powerful for the alternative θ_1. Furthermore, crucially, τ and γ are defined through $T(x)$ only and does not depend on the choice θ_1. It follows that $\phi_{\tau,\gamma}$ is UMP over $\theta > \theta_0$.

The final condition we need to check is that $\mathbb{E}_\theta[\phi_{\tau,\gamma}(X)] \le \alpha$ for $\theta < \theta_0$, but this is guaranteed by the next part.

2. For any $\theta' < \theta''$, we have in fact showed (again, in parts 1 and 2 of Theorem 1.1) that the test $\phi_{\tau,\gamma}$ is the most powerful for testing $H_0 : \theta = \theta'$ against $\theta = \theta''$ at significance level $\beta(\theta') = \mathbb{E}_{\theta'}[\phi_{\tau,\gamma}(X)]$. By definition, $\phi_{\tau,\gamma}$ is no less powerful than the constant test $\phi = \beta(\theta')$, so we obtain that $\beta(\theta'') \ge \beta(\theta')$. If $\beta(\theta'') = \beta(\theta')$, then by part 3 of Theorem 1.1, the likelihood ratio is constant μ-almost everywhere, so we must have $\mathcal{P}_{\theta'} = \mathcal{P}_{\theta''}$ and $\theta = \theta'$.

3. This follows from symmetry: We can consider testing $H_0 : \theta \ge \theta_0$ against $H_1 : \theta < \theta_0$ and reverse all the inequalities in the above proofs.

□

1.3.2 Two-sided testing and unbiased tests

Suppose that we observe $X \sim \mathcal{P}_\theta$ from an exponential family with densities

$$p_\theta(x) = C(\theta)e^{Q(\theta)T(x)}h(x)$$

where Q is strictly monotone, so that the densities have monotone likelihood ratios with respect to $T(x)$. For simplicity, let us focus on the case where the likelihood ratios are continuous. For $\theta_0 < \theta_1$, consider testing $H_0 : \theta \notin (\theta_0, \theta_1)$ against $H_1 : \theta \in (\theta_0, \theta_1)$. Although this is the two-sided case which has not been discussed, it is natural to consider the likelihood-ratio test

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) \in (\tau_0, \tau_1), \\ 0 & \text{if } T(x) \notin (\tau_0, \tau_1), \end{cases}$$

where τ_0 and τ_1 are determined by

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{E}_{\theta_1}[\phi(X)] = \alpha. \tag{1.3}$$

It is known that (Theorem 3.7.1 of [LR06]):

- The test $\phi(x)$ is UMP;
- The test minimizes the power function $\beta_\phi(\theta) := \mathbb{E}_\theta[\phi(X)]$ for all $\theta \notin [\theta_0, \theta_1]$ subject to (1.3);
- The power function $\beta_\phi(\theta)$ has a maximum at some $\theta' \in [\theta_0, \theta_1]$ and decreases strictly as θ goes away from θ' in either direction.

However, if we test $H_0 : \theta \in (\theta_0, \theta_1)$ against $H_1 : \theta \notin (\theta_0, \theta_1)$, then there exists no UMP in general. In this case, we can consider a weaker goal as follows.

When testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, we say that a test ϕ is unbiased if $\beta_\phi(\theta) \leq \alpha$ for $\theta \in \Theta_0$ and $\beta_\phi(\theta) \geq \alpha$ for $\theta \in \Theta_1$, so that $\beta_\phi(\theta) = \alpha$ on the boundary between Θ_0 and Θ_1. It is not hard to see that any UMP test is unbiased.

For testing $H_0 : \theta \in (\theta_0, \theta_1)$ against $H_1 : \theta \notin (\theta_0, \theta_1)$, a test ϕ is unbiased if $\beta_\phi(\theta) \leq \alpha$ for $\theta \in [\theta_0, \theta_1]$ and $\beta_\phi(\theta) \geq \alpha$ for $\theta \notin (\theta_0, \theta_1)$. Consider the likelihood-ratio test

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) \notin (\tau_0, \tau_1), \\ 0 & \text{if } T(x) \in (\tau_0, \tau_1), \end{cases}$$

where τ_0 and τ_1 are determined by (1.3). Then ϕ is unbiased by monotonicity.

1.4 *p*-values

1.4.1 Definition

Consider a family of tests ϕ_α, each at significance level $\alpha \in (0, 1)$. Suppose that the regions of rejection $S_1(\alpha) := \{x \in \mathcal{X} : \phi_\alpha(x) = 1\}$ are nested as α varies in $(0, 1)$, in the sense that

$$S_1(\alpha) \subset S_1(\alpha') \quad \text{if } \alpha < \alpha'.$$

(This is typically true except in some corner cases.)

A widely used notion in hypothesis testing is the p-value, which is defined to be the number

$$\hat{p} = \hat{p}(X) = \inf\{\alpha : X \in S_1(\alpha)\}.$$

In other words, given the data X, the p-value is the smallest significance level at which the null is rejected. It indicates how strongly the data contradicts the null hypothesis: the smaller $\hat{p}$ is, the stronger the contradiction is. At a (fixed) significance level α', we reject the null if $\hat{p} < \alpha'$.

1.4.2 Examples

Gaussian Recall the testing problem between $H_0 : X \sim \mathcal{N}(0, I_d)$ and $H_1 : X \sim \mathcal{N}(\mu, I_d)$. The likelihood-ratio test rejects the null if $\langle X, \mu \rangle > z_\alpha \|u\|$, where z_α denotes the $(1-\alpha)$-quantile of Z. Hence, we have

$$S_1(\alpha) = \{x \in \mathbb{R}^d : \langle x, \mu \rangle > z_\alpha \|u\|\}$$

and

$$\hat{p}(X) = \inf\{\alpha : X \in S_1(\alpha)\} = \inf\{\alpha : \langle X, \mu \rangle > z_\alpha \|u\|\}.$$

It follows that $z_{\hat{p}} = \langle X, \mu/\|\mu\| \rangle$ and

$$\hat{p}(X) = 1 - \Phi\big(\langle X, \mu/\|\mu\| \rangle\big) = \mathbb{P}\left\{Z > \langle X, \mu/\|\mu\| \rangle\right\},$$

where Z denotes a standard Gaussian and Φ denotes its CDF.

Multinomial Suppose that X takes values in $\{1, 2, \ldots, 10\}$. Consider testing between hypotheses

- $H_0 : p_0(j) = 1/10$ for $j = 1, \ldots, 10$;
- $H_1 : p_1(j) = j/55$ for $j = 1, \ldots, 10$.

Consider a test with

$$S_1(\alpha) = \{x : x \geq 11 - 10\alpha\}.$$

The p-value is

$$\hat{p}(X) = \inf\{\alpha : X \geq 11 - 10\alpha\} = (11 - X)/10.$$

1.5 Confidence regions

Hypothesis testing is closely related to the inference problem: Given the data $X \sim \mathcal{P}_\theta$, we would like to have a subset $\mathcal{C}(X) \subset \Theta$ such that $\theta \in \mathcal{C}(X)$ with high probability. To be more precise, we say that $\mathcal{C}(X) \subset \Theta$ is a confidence region (or confidence set) at confidence level $1 - \alpha$ if

$$\mathbb{P}_\theta\{\theta \in \mathcal{C}(X)\} \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

The relation of confidence regions to hypothesis testing is explained by the following theorem.

Theorem 1.3. *For any $\theta_0 \in \Theta$, let $S_0(\theta_0)$ be the region of acceptance of a level-α test for testing $H_0 : \theta = \theta_0$ against an alternative given $X \sim \mathcal{P}_\theta$. Define $\mathcal{C}(X) := \{\theta \in \Theta : X \in S_0(\theta)\}$. Then $\mathcal{C}(X)$ is a confidence region at confidence level $1 - \alpha$.*

Proof. By definition, we have $\theta \in \mathcal{C}(X)$ if and only if $X \in S_0(\theta)$, so

$$\mathbb{P}_\theta\{\theta \in \mathcal{C}(X)\} = \mathbb{P}_\theta\{X \in S_0(\theta)\} \geq 1 - \alpha.$$

□

The above theorem is extremely general and (therefore) almost vacuous. To further explore the connection between confidence regions and hypothesis testing, we now consider a special case where the confidence region $\mathcal{C}(X)$ takes the form $[\theta_\ell(X), \infty)$ for a real number $\theta_\ell(X)$. In other words, the parameters are real-valued, and we aim to find a lower confidence bound $\theta_\ell(X)$ such that

$$\mathbb{P}_\theta\{\theta_\ell(X) \le \theta\} \ge 1 - \alpha.$$

Subject to this constraint, we would like $\theta_\ell(X)$ to minimize $\mathbb{P}_\theta\{\theta_\ell(X) \le \theta'\}$ for all $\theta' < \theta$. If this is indeed the case, we say that $\theta_\ell(X)$ is a uniformly most accurate lower confidence bound for θ at confidence level $1 - \alpha$.

Theorem 1.4. *For $\Theta \subset \mathbb{R}$, let $\{p_\theta(x) : \theta \in \Theta\}$ be a family of densities with continuous likelihood ratios that are monotone with respect to $T(x)$. Suppose that for $\theta \in \Theta$ and $X \sim \mathcal{P}_\theta$, the CDF $F_\theta(t)$ of $T(X)$ is continuous in θ and t. Then the following statements hold:*

1. *There exists a uniformly most accurate confidence bound $\theta_\ell(X)$ for θ at confidence level $1 - \alpha$ for each $\alpha \in (0, 1)$.*
2. *Moreover, if the equation $F_\theta(T(x)) = 1 - \alpha$ has a solution $\theta \in \Theta$, then the solution is unique and is equal to $\theta_\ell(x)$.*

Proof. 1. Consider the likelihood-ratio test (1.2) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Recall that the region of acceptance is $S_0(\theta_0) = \{x : T(x) \le \tau(\theta_0)\}$, where $\tau(\theta_0)$ is defined so that $\mathbb{P}_{\theta_0}\{T(X) \le \tau(\theta_0)\} = 1 - \alpha$ for any $\theta_0 \in \Theta$. Recall that the power $\mathbb{P}_{\theta_1}\{T(X) > \tau(\theta_0)\}$ is larger than α for $\theta_1 > \theta_0$. As a result, we have $\tau(\theta_0) < \tau(\theta_1)$.

Define $\mathcal{C}(X) := \{\theta \in \Theta : T(X) \le \tau(\theta)\}$. It follows from the monotonicity of τ that $\mathcal{C}(X)$ is of the form $[\theta_\ell(X), \infty)$ where

$$\theta_\ell(X) := \inf\{\theta \in \Theta : T(X) \le \tau(\theta)\}.$$

By the previous theorem, $[\theta_\ell(X), \infty)$ is a confidence region for θ at level $1 - \alpha$. Furthermore, for any $\theta' < \theta$, the probability

$$\mathbb{P}_\theta\{\theta_\ell(X) > \theta'\} = \mathbb{P}_\theta\{\theta' \notin \mathcal{C}(X)\} = \mathbb{P}_\theta\{T(X) > \tau(\theta')\}$$

is the power of the likelihood-ratio test for testing $H_0 : \theta \le \theta'$ against $\theta > \theta'$. Since the likelihood-ratio test is UMP, the above quantity is maximized, or, equivalently, the probability $\mathbb{P}_\theta\{\theta_\ell(X) \le \theta'\}$ is minimized. Hence $\theta_\ell(X)$ is a uniformly most accurate lower confidence bound.

2. The reasoning is again similar to the previous part. For $\theta < \theta'$, since $F_{\theta'}(t)$ is the power of the likelihood-ratio test at level $F_\theta(t)$, we have $F_\theta(t) < F_{\theta'}(t)$. That is, $F_\theta(t)$ is strictly increasing in θ. It follows that the equation $F_\theta(T(x)) = 1 - \alpha$ has at most one solution θ.

Suppose that the solution exists. By the definition of $\theta_\ell(x)$, we have $\tau(\theta_\ell(x)) = T(x)$ and thus

$$\mathbb{P}_{\theta_\ell(x)}\{T(X) \le T(x)\} = \mathbb{P}_{\theta_\ell(x)}\{T(X) \le \tau(\theta_\ell(x))\} = 1 - \alpha.$$

The claim follows from the uniqueness of the solution.

□

Suppose that $\theta_\ell(X)$ and $\theta_u(X)$ are lower and upper confidence bounds for θ at confidence levels $1-\alpha_1$ and $1-\alpha_2$ respectively. Then $[\theta_\ell(X), \theta_u(X)]$ is a confidence interval for θ at confidence level $1-\alpha$ where $\alpha = \alpha_1 + \alpha_2$:

$$\mathbb{P}_\theta\{\theta \in [\theta_\ell(X), \theta_u(X)]\} \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

If $\theta_\ell(X)$ and $\theta_u(X)$ are uniformly most accurate, then they together minimize

$$\mathbb{P}_\theta\{\theta_\ell(X) \leq \theta'\} + \mathbb{P}_\theta\{\theta'' \leq \theta_u(X)\}$$

for any $\theta' < \theta$ and $\theta'' > \theta$. However, this is a somewhat unnatural measure of accuracy when we consider confidence intervals. A more widely used measure of accuracy is simply the length $\theta_u(X) - \theta_\ell(X)$ of the confidence interval.

Chapter 2

Examples of statistical tests

2.1 Hypothesis testing for linear models

2.1.1 Setup

Consider the linear regression model

$$Y_i = \tilde{\alpha} + \tilde{\beta}\tilde{x}_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are constants, and ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Let us standardize $\tilde{x}_i$ by setting

$$x_i := \frac{\tilde{x}_i - \bar{x}}{\sqrt{\sum_{j=1}^n (\tilde{x}_j - \bar{x})^2}}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^n \tilde{x}_i.$$

Then $\sum_{i=1}^n x_i = 0$ and $\sum_{i=1}^n x_i^2 = 1$. Moreover, if we set

$$\beta^* := \tilde{\beta}\sqrt{\sum_{i=1}^n (\tilde{x}_i - \bar{x})^2}, \qquad \alpha^* := \tilde{\alpha} + \beta^* \frac{\bar{x}}{\sqrt{\sum_{i=1}^n (\tilde{x}_i - \bar{x})^2}},$$

then the model can be written in the following standardized form

$$Y_i = \tilde{\alpha} + \tilde{\beta}\tilde{x}_i + \varepsilon_i = \alpha^* + \beta^* x_i + \varepsilon_i.$$

Suppose that we are interested in a linear function of $\tilde{\alpha}$ and $\tilde{\beta}$ (or, equivalently, of α^* and β^*). In other words, let c and d be constants, and consider the unknown quantity

$$\rho = c\alpha^* + d\beta^*.$$

Suppose that we would like to test $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$; or, we would like to obtain a confidence interval for ρ.

2.1.2 *z*-test

Towards this end, let us first consider the least squares estimator $(\hat{\alpha}, \hat{\beta})$ of (α^*, β^*). Solving

$$\min_{\alpha', \beta'} \sum_{i=1}^n (Y_i - \alpha' - \beta' x_i)^2$$

yields

$$\hat{\alpha} = \bar{Y}, \qquad \hat{\beta} = \sum_{i=1}^{n} x_i Y_i = x^\top Y.$$

Therefore, we can estimate ρ by

$$\hat{\rho} = c\hat{\alpha} + d\hat{\beta} = c\bar{Y} + dx^\top Y = \sum_{i=1}^{n} \Big(\frac{c}{n} + dx_i\Big) Y_i.$$

At $\rho = \rho_0$, the estimator $\hat{\rho}$ is Gaussian with mean

$$\sum_{i=1}^{n} \Big(\frac{c}{n} + dx_i\Big)(\alpha^* + \beta^* x_i) = c\alpha^* + d\beta^* = \rho_0$$

and variance

$$\sigma^2 \sum_{i=1}^{n} \Big(\frac{c}{n} + dx_i\Big)^2 = \sigma^2 \Big(\frac{c^2}{n} + d^2\Big).$$

As a result,

$$\frac{\hat{\rho} - \rho_0}{\sigma\sqrt{c^2/n + d^2}} \sim \mathcal{N}(0, 1).$$

If σ is known, we can derive a test at significance level $\delta \in (0, 1)$ from the above fact. Namely, we accept H_0 if

$$\left| \frac{\hat{\rho} - \rho_0}{\sigma\sqrt{c^2/n + d^2}} \right| \le z_{\delta/2}$$

where $z_{\delta/2}$ is the $(1 - \delta/2)$-quantile of $\mathcal{N}(0, 1)$. Moreover, a confidence interval for ρ at confidence level $1 - \delta$ is

$$\Big[\hat{\rho} - z_{\delta/2} \cdot \sigma\sqrt{c^2/n + d^2}, \hat{\rho} + z_{\delta/2} \cdot \sigma\sqrt{c^2/n + d^2}\Big].$$

2.1.3 *t*-test

However, since σ is typically unknown, we need a few more steps. A natural idea is to replace σ in the above formula with an estimate. The sample variance is

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \\
&= \frac{1}{n-2} \|Y - \bar{Y}\mathbf{1} - x^\top Y x\|^2 \\
&= \frac{1}{n-2} \Big\| \Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\Big) Y \Big\|^2 \\
&= \frac{1}{n-2} \Big\| \Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\Big) (\alpha^* \mathbf{1} + \beta^* x + \varepsilon) \Big\|^2 \\
&= \frac{1}{n-2} \Big\| \Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\Big) \varepsilon \Big\|^2,
\end{aligned}$$

and so

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} = \Big\| \Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\Big) \frac{\varepsilon}{\sigma} \Big\|^2.$$

Based on this formula, we have two observations:

- Since x is a unit vector orthogonal to $\mathbf{1}$, the matrix $I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top$ is the orthogonal projection onto the $(n-2)$-dimensional subspace S of $\mathbb{R}^n$ orthogonal to the span of $\mathbf{1}$ and x. This implies that $\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\right)\frac{\varepsilon}{\sigma}$ is a standard Gaussian vector in S, so $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, i.e., it is a chi-squared random variable with $n-2$ degrees of freedom.
- Since $\hat{\rho} = \left(\frac{c}{n}\mathbf{1} + dx\right)^\top Y = \left(\frac{c}{n}\mathbf{1} + dx\right)^\top(\alpha^*\mathbf{1} + \beta^* x + \varepsilon)$, we see that $\hat{\rho}$ only depends on the randomness of the projection of ε on the span of $\mathbf{1}$ and x. Therefore, $\hat{\rho}$ is independent of $\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n} - xx^\top\right)\frac{\varepsilon}{\sigma}$ and thus of $(n-2)\frac{\hat{\sigma}^2}{\sigma^2}$.

By a standard characterization of the t-distribution with $n-2$ degrees of freedom, denoted by t_{n-2}, it follows that

$$\frac{\hat{\rho} - \rho_0}{\hat{\sigma}\sqrt{c^2/n + d^2}} = \frac{\hat{\rho} - \rho_0}{\sigma\sqrt{c^2/n + d^2}} \Bigg/ \sqrt{\frac{(n-2)\hat{\sigma}^2/\sigma^2}{n-2}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} = t_{n-2}.$$

Similar to the case of a Z-test, here we accept $H_0 : \rho = \rho_0$ if

$$\left|\frac{\hat{\rho} - \rho_0}{\hat{\sigma}\sqrt{c^2/n + d^2}}\right| \le \tau_{\delta/2}$$

where $\tau_{\delta/2}$ is the $(1-\delta/2)$-quantile of t_{n-2}. Moreover, a confidence interval for ρ at confidence level $1-\delta$ is

$$\left[\hat{\rho} - \tau_{\delta/2}\cdot\hat{\sigma}\sqrt{c^2/n + d^2}, \hat{\rho} + \tau_{\delta/2}\cdot\hat{\sigma}\sqrt{c^2/n + d^2}\right].$$

2.2 Analysis of variance

Consider m independent samples, each of which consists of i.i.d. observations

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad j = 1, \dots n_i,\ i = 1, \dots, m,$$

where σ is unknown. We are interested in testing $H_0 : \mu_1 = \cdots = \mu_m$ against $H_1 :$ otherwise. The idea is to construct two estimators of the variance σ^2 and compare them. The first estimator is good regardless of whether H_0 or H_1 is true, while the second is good only under H_0. As a result, comparing the two will serve as the test.

- Let $Y_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ and $X_i = (X_{i1}, \dots, X_{in_i})^\top$. Then

$$\begin{aligned}\sum_{j=1}^{n_i}(X_{ij} - Y_i)^2 &= \left\|X_i - \frac{\mathbf{1}_{n_i}^\top X_i}{n_i}\mathbf{1}_{n_i}\right\|_2^2 \\ &= \left\|\left(I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i}\right)X_i\right\|_2^2 = \left\|\left(I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i}\right)(X_i - \mu_i\mathbf{1}_{n_i})\right\|_2^2,\end{aligned}$$

  where $\mathbf{1}_{n_i}$ is the all-ones vector in $\mathbb{R}^{n_i}$. Since $I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i}$ is the orthogonal projection onto the orthogonal complement of the all-ones vector, and $\frac{1}{\sigma}(X_i - \mu_i\mathbf{1}_{n_i})$ is a standard Gaussian vector, we see that $\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(X_{ij} - Y_i)^2 \sim \chi^2_{n_i-1}$. It follows that

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - Y_i)^2 \sim \chi^2_{n-m},$$

  where $n := \sum_{i=1}^m n_i$. Then the quantity $\frac{1}{n-m}\sum_{i=1}^m\sum_{j=1}^{n_i}(X_{ij} - Y_i)^2$ is an estimator of σ^2.

- Let $\bar{Y} = \frac{1}{m}\sum_{i=1}^m Y_i$ and $Y = (Y_1, \ldots, Y_m)^\top$. Under H_0, let $\mu = \mu_1$ and we have i.i.d. $\sqrt{n_i}(Y_i - \mu) \sim \mathcal{N}(0, \sigma^2)$. Similar to the above case, we can derive

$$\sum_{i=1}^m n_i (Y_i - \bar{Y})^2 = \left\| \left(I_m - \frac{\mathbf{1}_m \mathbf{1}_m^\top}{m} \right) \sqrt{n_i} (Y - \mu \mathbf{1}_m) \right\|_2^2.$$

  It follows that

$$\frac{1}{\sigma^2} \sum_{i=1}^m n_i (Y_i - \bar{Y})^2 \sim \chi^2_{m-1}.$$

  Then the quantity $\frac{1}{m-1}\sum_{i=1}^m n_i (Y_i - \bar{Y})^2$ is an estimator of σ^2.

Furthermore, we claim that the two estimators $\frac{1}{n-m}\sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - Y_i)^2$ and $\frac{1}{m-1}\sum_{i=1}^m n_i (Y_i - \bar{Y})^2$ are independent. To this end, let us compute the covariance matrix

$$\begin{aligned}
&\mathbb{E}\left[\left(\left(I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i} \right)(X_i - \mu \mathbf{1}_{n_i}) \right) \left(\left(I_m - \frac{\mathbf{1}_m \mathbf{1}_m^\top}{m} \right) \sqrt{n_i}(Y - \mu \mathbf{1}_m) \right)^\top \right] \\
&= \sqrt{n_i} \left(I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i} \right) \mathbb{E}\left[(X_i - \mu \mathbf{1}_{n_i})(Y - \mu \mathbf{1}_m)^\top \right] \left(I_m - \frac{\mathbf{1}_m \mathbf{1}_m^\top}{m} \right)^\top.
\end{aligned}$$

Note that only the ith entry of Y is correlated with X_i, and $\mathbb{E}\left[(X_i - \mu \mathbf{1}_{n_i})(Y_i - \mu) \right] = \frac{\sigma^2}{n_i} \mathbf{1}_{n_i}$. Therefore, the above quantity is equal to

$$= \frac{\sigma^2}{\sqrt{n_i}} \left(I_{n_i} - \frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top}{n_i} \right) \left[\mathbf{0}_{n_i} \cdots \mathbf{0}_{n_i} \ \mathbf{1}_{n_i} \ \mathbf{0}_{n_i} \cdots \mathbf{0}_{n_i} \right] \left(I_m - \frac{\mathbf{1}_m \mathbf{1}_m^\top}{m} \right)^\top = \mathbf{0}_{n_i \times m}$$

where $\mathbf{0}_d$ denotes the all-zeros vector in dimension d. The claimed independence then follows easily.

Finally, by a standard characterization of the F-distribution with $m-1$ and $n-m$ degrees of freedom, denoted by $F_{m-1,n-m}$, we conclude that

$$\frac{\frac{1}{m-1}\sum_{i=1}^m n_i (Y_i - \bar{Y})^2}{\frac{1}{n-m}\sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - Y_i)^2} \sim F_{m-1,n-m}.$$

To define a test at significance level $\alpha \in (0,1)$, we can simply accept H_0 if the above quantity is no larger than f_α, where f_α denotes the $(1-\alpha)$-quantile of $F_{m-1,n-m}$.

2.3 Wald test

Let θ be a real-valued parameter. Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}_\theta$, let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of θ. Define $\hat{\sigma}_n := \sqrt{\mathsf{Var}(\hat{\theta}_n)}$. Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Assume that $\hat{\theta}_n$ is asymptotically normal in the sense

$$\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

For $\alpha \in (0,1)$, the Wald test at asymptotic significance level α is defined to be

$$\phi(X_1, \ldots, X_n) := \begin{cases} 1 & \text{if } \left|\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n}\right| > z_{\alpha/2}, \\ 0 & \text{if } \left|\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n}\right| \leq z_{\alpha/2}. \end{cases}$$

To be more precise, a test ϕ is said to have an asymptotic significance level α if

$$\mathbb{P}_{\theta_0}\{\phi(X_1, \ldots, X_n) = 1\} \to \alpha \quad \text{as } n \to \infty.$$

This is the case for the Wald test because

$$\mathbb{P}_{\theta_0}\left\{\left|\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n}\right| > z_{\alpha/2}\right\} \to \mathbb{P}\{|Z| > z_{\alpha/2}\} = \alpha$$

where $Z \sim \mathcal{N}(0,1)$.

Moreover, the power $\beta(\theta)$ at $\theta \neq \theta_0$ is approximately

$$\begin{aligned} \mathbb{P}_\theta\left\{\left|\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n}\right| > z_{\alpha/2}\right\} &= \mathbb{P}_\theta\left\{\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} > z_{\alpha/2}\right\} + \mathbb{P}_\theta\left\{\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} < -z_{\alpha/2}\right\} \\ &= \mathbb{P}_\theta\left\{\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} > \frac{\theta_0 - \theta}{\hat{\sigma}_n} + z_{\alpha/2}\right\} + \mathbb{P}_\theta\left\{\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} < \frac{\theta_0 - \theta}{\hat{\sigma}_n} - z_{\alpha/2}\right\} \\ &= 1 - \Phi\Big(\frac{\theta_0 - \theta}{\hat{\sigma}_n} + z_{\alpha/2}\Big) + \Phi\Big(\frac{\theta_0 - \theta}{\hat{\sigma}_n} - z_{\alpha/2}\Big). \end{aligned}$$

Since $\hat{\sigma}_n \to 0$ in probability, the power converges to 1.

The associated confidence interval at asymptotic confidence level $1 - \alpha$ is

$$(\hat{\theta}_n - \hat{\sigma}_n z_{\alpha/2}, \hat{\theta}_n + \hat{\sigma}_n z_{\alpha/2}).$$

Comparing paired binary variables Consider i.i.d. pairs of binary random variables (X_i, Y_i) for $i = 1, \ldots, n$. Let $\delta := \mathbb{E}[X_i] - \mathbb{E}[Y_i]$ and consider testing $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$. The plug-in estimators of δ and the variance are, respectively,

$$\hat{\delta}_n = \frac{1}{n}\sum_{i=1}^n (X_i - Y_i), \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - Y_i - \hat{\delta}_n)^2.$$

The test statistic for the Wald test is $\hat{\delta}_n / \hat{\sigma}_n$.

2.4 Goodness-of-fit tests

Suppose that we observe i.i.d. $Z_1, \ldots, Z_n \sim \mathcal{P}$ for a distribution $\mathcal{P}$ with density $f(x)$. Consider the nonparametric setting: We test $H_0 : \mathcal{P} = \mathcal{P}_0$ against $H_1 : \mathcal{P} \neq \mathcal{P}_0$, where there is no parametric assumption on $\mathcal{P}_0$.

2.4.1 Pearson's chi-squared test

If the distribution of interest is discrete, then the test introduced in this section can be applied directly. If the distribution is continuous, we can pre-process the data as follows. For example, in the real-valued case, let $B_1, \dots, B_k$ form a partition of $\mathbb{R}$. For each $i \in [k]$, let $p_i := \int_{B_i} f(x)\,dx$. Moreover, let $X_i := |\{j \in [n] : Z_j \in B_i\}$. Then X_i/n should be close to p_i. Therefore, the essence of goodness-of-fit testing is the following discrete testing problem for multinomial data.

Consider i.i.d. categorical random variables random variables $Y_1, \dots, Y_n$, each taking value i with probability p_i for $i \in [k]$. Here we have $p_i > 0$ for each $i \in [k]$ and $\sum_{i=1}^k p_i = 1$. We are interested in testing $H_0 : p_i = p_i^0$ for all $i \in [k]$ against $H_1 : p_i \neq p_i^0$ for some $i \in [k]$. Let X_i be the number of Y_j that are equal to i. Then $X_i \sim \mathsf{Bin}(n, p_i)$ marginally. Consider the test statistic

$$T := \sum_{i=1}^k \frac{(X_i - np_i^0)^2}{np_i^0}.$$

To obtain the asymptotic distribution of T at $p = p^0$, we first compute the covariance

$$\mathbb{E}[(X_i - np_i)(X_j - np_j)] = \mathbb{E}[X_i X_j] - n^2 p_i p_j.$$

It holds that

$$\mathbb{E}[X_i X_j] = \mathbb{E}\left[\Big(\sum_{\ell=1}^n \mathbb{1}\{Y_\ell = i\}\Big)\Big(\sum_{m=1}^n \mathbb{1}\{Y_m = j\}\Big)\right] = \sum_{\ell=1}^n \sum_{m=1}^n \mathbb{P}\{Y_\ell = i,\, Y_m = j\}.$$

For $i \neq j$, we have

$$\mathbb{E}[X_i X_j] = \sum_{\ell \neq m} p_i p_j = n(n-1)p_i p_j.$$

For $i = j$, we have

$$\mathbb{E}[X_i^2] = \sum_{\ell = m} p_i + \sum_{\ell \neq m} p_i^2 = np_i + n(n-1)p_i^2.$$

Therefore, if we define a random vector $V \in \mathbb{R}^n$ with $V_i := \frac{X_i - np_i}{\sqrt{n}}$, then

$$\mathbb{E}[V_i V_j] = \frac{np_i \mathbb{1}\{i = j\} + n(n-1)p_i p_j - n^2 p_i p_j}{n} = p_i \mathbb{1}\{i = j\} - p_i p_j.$$

We conclude that the covariance matrix of V is

$$\mathbb{E}[VV^\top] = \mathsf{Diag}(p) - pp^\top.$$

Note that $X_k = n - \sum_{i=1}^{k-1} X_i$, so it suffices to consider the first $k-1$ coordinates of V. Let Σ denote the top-left $(k-1) \times (k-1)$ principal minor of $\mathbb{E}[VV^\top]$. Then it is not hard to check that

$$\Sigma^{-1} = \mathsf{Diag}(p_{-k})^{-1} + \frac{1}{p_k}\mathbf{1}\mathbf{1}^\top,$$

where $p_{-k} = (p_1, \dots, p_{k-1})^\top$ and $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^{k-1}$. By the central limit theorem,

$$\Sigma^{-1/2} V_{-k} \xrightarrow{d} \mathcal{N}(0, I_{k-1}),$$

and then by the continuous mapping theorem,

$$V_{-k}^\top \Sigma^{-1} V_{-k} = \|\Sigma^{-1/2} V_{-k}\|_2^2 \xrightarrow{d} \chi_{k-1}^2.$$

The left-hand side is equal to

$$\sum_{i=1}^{k-1} \frac{1}{p_i} \frac{(X_i - np_i)^2}{n} + \frac{1}{p_k} \frac{(\sum_{i=1}^{k-1}(X_i - np_i))^2}{n} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} =: T.$$

Therefore, Pearson's chi-squared test at significance level $\alpha \in (0,1)$ is

$$\phi(X_1, \dots, X_k) := \begin{cases} 1 & \text{if } T > q_\alpha, \\ 0 & \text{if } T \le q_\alpha, \end{cases}$$

where q_α denotes the $(1-\alpha)$-quantile of χ_{k-1}^2.

2.4.2 Kolmogorov–Smirnov test

Let F denote the CDF of $\mathcal{P}$, and let F_n denote the empirical CDF defined by

$$F_n(x) = \frac{1}{n} \big| \{ j \in [n] : Z_j \le x \} \big|.$$

Then we can define the Kolmogorov–Smirnov statistic

$$D_n := \max_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

It is standard that $U_j := F(Z_j) \sim \mathsf{Unif}(0,1)$. As a result, we have

$$\begin{aligned} \mathbb{P}\{D_n \le t\} &= \mathbb{P}\left\{ \max_{x \in \mathbb{R}} |F_n(x) - F(x)| \le t \right\} \\ &= \mathbb{P}\left\{ \max_{x \in \mathbb{R}} \left| \frac{|\{j \in [n] : Z_j \le x\}|}{n} - F(x) \right| \le t \right\} \\ &= \mathbb{P}\left\{ \max_{x \in \mathbb{R}} \left| \frac{|\{j \in [n] : U_j \le F(x)\}|}{n} - F(x) \right| \le t \right\} \\ &= \mathbb{P}\left\{ \max_{y \in (0,1)} \left| \frac{|\{j \in [n] : U_j \le y\}|}{n} - y \right| \le t \right\}. \end{aligned}$$

This shows that the distribution of D_n does not depend on the distribution $\mathcal{P}$.

Moreover, it is known that

$$\sqrt{n} D_n \xrightarrow{d} \mathcal{K} \quad \text{as } n \to \infty,$$

where $\mathcal{K}$ denotes the Kolmogorov distribution with CDF

$$F_{\mathcal{K}}(x) = \frac{\sqrt{2\pi}}{x} \sum_{\ell=1}^{\infty} \exp\left(- \frac{(2\ell - 1)^2 \pi^2}{8x^2} \right).$$

Then the Kolmogorov–Smirnov test at significance level $\alpha \in (0,1)$ is defined by

$$\phi(Z_1, \dots, Z_n) := \begin{cases} 1 & \text{if } \sqrt{n} D_n > k_\alpha, \\ 0 & \text{if } \sqrt{n} D_n \le k_\alpha, \end{cases}$$

where k_α denotes the $(1-\alpha)$-quantile of $\mathcal{K}$.

2.5 Nonparametric tests for two samples

2.5.1 Wilcoxon signed-rank test

Given i.i.d. real-valued data $X_1, \dots, X_n \sim \mathcal{P}$, suppose that we are interested in testing between

- $H_0 : \mathcal{P}$ is symmetric around zero;
- H_1 : otherwise.

First, if we would like to test symmetry around any $m_0 \in \mathbb{R}$, it suffices to subtract m_0 from each X_i to reduce the problem to the above setting. Moreover, the above task appears frequently when we deal with paired data: Suppose that we have i.i.d. pairs $(Y_1, Z_1), \dots, (Y_n, Z_n)$. If each pair is exchangeable in the sense that (Y_i, Z_i) and (Z_i, Y_i) have the same distribution, then $X_i := Y_i - Z_i$ has a distribution which is symmetric around zero. Note that we do not have to assume that Y_i and Z_i are independent.

The Wilcoxon signed-rank test uses a statistic T defined as follows. Let the sign function be defined by

$$\mathsf{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

Moreover, consider the order statistics $X_{(1)}, \dots, X_{(n)}$ by absolute values, which are obtained by reordering $X_1, \dots, X_n$ so that

$$|X_{(1)}| \le \cdots \le |X_{(n)}|.$$

Then the signed-rank statistic T is defined to be

$$T := \sum_{i=1}^{n} i \cdot \mathsf{sign}(X_{(i)}).$$

It is easily seen that this statistic does not depend on the particular distribution $\mathcal{P}$ as long as it is symmetric around zero.

To derive a test using the statistic T, assume $\mathbb{P}\{X_i = 0\} = 0$ for simplicity. Under H_0, the signs $I_i := \mathsf{sign}(X_{(i)})$ are independent Rademacher random variables. We can compute

$$\mathbb{E}[T] = \sum_{i=1}^{n} i \cdot \mathbb{E}[I_i] = 0, \qquad \mathsf{Var}(T) = \sum_{i=1}^{n} i^2 \cdot \mathsf{Var}(I_i) = \frac{1}{6} n(n+1)(2n+1).$$

By the central limit theorem, a test at asymptotic significance level $\alpha \in (0, 1)$ is

$$\phi(X_1, \dots, X_n) := \begin{cases} 1 & \text{if } \frac{|T|}{\sqrt{\mathsf{Var}(T)}} > z_{\alpha/2}, \\ 0 & \text{if } \frac{|T|}{\sqrt{\mathsf{Var}(T)}} \le z_{\alpha/2}, \end{cases}$$

where $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of $\mathcal{N}(0, 1)$.

2.5.2 Mann–Whitney U test, aka Wilcoxon rank-sum test

The purpose of the Mann–Whitney U test is similar to that of the Wilcoxon signed-rank test, but the setup is different. Suppose that we have two i.i.d. samples $Y_1, \dots, Y_m \sim \mathcal{P}$ and $Z_1, \dots, Z_n \sim \mathcal{Q}$, where the sample sizes may not be the same. Moreover, assume that the two samples are independent from each other. We are interested in testing $H_0 : \mathcal{P} = \mathcal{Q}$ against $H_1 : \mathcal{P} \neq \mathcal{Q}$. The Mann–Whitney U statistic is defined by

$$U := \sum_{i=1}^{m} \sum_{j=1}^{n} \mathsf{sign}(Y_i - Z_j).$$

Another equivalent test statistic is the Wilcoxon rank-sum statistic defined as follows. Consider the two samples $Y_1, \dots, Y_m, Z_1, \dots, Z_n$ combined. For $i \in [m]$, suppose that Y_i is the R_ith smallest number in the combined sample. Thus R_i represents the rank of Y_i and is an integer between 1 and $m + n$. Then we can define the Wilcoxon rank-sum statistic to be

$$S := \sum_{i=1}^{m} R_i.$$

To see that S is equivalent to U, we assume that all the data points are distinct with probability 1 for simplicity. Then we have

$$\begin{aligned}
S = \sum_{i=1}^{m} R_i &= \sum_{i=1}^{m} \Bigg(1 + \sum_{j=1}^{m} \mathbb{1}\{Y_j < Y_i\} + \sum_{j=1}^{n} \mathbb{1}\{Z_j < Y_i\}\Bigg) \\
&= \sum_{i=1}^{m} \Bigg(1 + \sum_{j \in [m] \setminus \{i\}} \frac{1}{2}\Big(\mathsf{sign}(Y_i - Y_j) + 1\Big) + \sum_{j=1}^{n} \frac{1}{2}\Big(\mathsf{sign}(Y_i - Z_j) + 1\Big)\Bigg) \\
&= m + \frac{m(m-1)}{2} + \frac{mn}{2} + \frac{1}{2} \sum_{i=1}^{m} \sum_{j \in [m] \setminus \{i\}} \mathsf{sign}(Y_i - Y_j) + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathsf{sign}(Y_i - Z_j) \\
&= \frac{m(m+n+1) + U}{2}.
\end{aligned}$$

Hence, S and U are related to each other linearly.

It remains to focus on the statistic U. Let $I_{ij} := \mathsf{sign}(Y_i - Z_j)$. Under H_0, it is clear that

$$\mathbb{E}[U] = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}[I_{ij}] = 0.$$

For the variance, we have

$$\mathsf{Var}(U) = \mathbb{E}\left[\Bigg(\sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij}\Bigg)\Bigg(\sum_{k=1}^{m} \sum_{\ell=1}^{n} I_{k\ell}\Bigg)\right] = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\ell=1}^{n} \mathbb{E}[I_{ij} I_{k\ell}].$$

If $i \neq k$ and $j \neq \ell$, then $\mathbb{E}[I_{ij} I_{k\ell}] = 0$. If $i = k$ and $j = \ell$, then $\mathbb{E}[I_{ij} I_{k\ell}] = 1$. If $i = k$ and $j \neq \ell$, then

$$\begin{aligned}
\mathbb{E}[I_{ij} I_{k\ell}] &= \mathbb{E}[\mathsf{sign}(Y_i - Z_j) \cdot \mathsf{sign}(Y_i - Z_\ell)] \\
&= \mathbb{P}\{Y_i > \max(Z_j, Z_k) \text{ or } Y_i < \min(Z_j, Z_k)\} - \mathbb{P}\{Z_j < Y_i < Z_\ell \text{ or } Z_\ell < Y_i < Z_j\}.
\end{aligned}$$

By applying the CDF F of the underlying distribution, we can again assume without loss of generality that Y_i and Z_j are uniform random variables on $[0,1]$. Thus the above expectation is equal to

$$\int_0^1 \int_0^1 \big(1 - |z_j - z_k| - |z_j - z_k|\big)\, dz_j\, dz_k = \frac{1}{3}.$$

In summary, we obtain

$$\mathsf{Var}(U) = mn + \frac{mn(n-1)}{3} + \frac{nm(m-1)}{3} = \frac{mn(m+n+1)}{3}.$$

Using Gaussian approximation, we can again apply the Z-test with the statistic U.

2.6 Testing with permutations

2.6.1 Wald–Wolfowitz runs test

A common task is to determine whether a sequence of observations is "sufficiently random". Let us consider a generic example. Suppose that we observe binary outcomes $X_1, \dots, X_n \in \{0,1\}$ and are interested in testing between

- $H_0 : X_1, \dots, X_n$ are i.i.d. $\mathsf{Ber}(p)$ random variables for some parameter $p \in (0,1)$;
- H_1 : otherwise.

Note that H_0 is a composite hypothesis that includes any possible $p \in (0,1)$, and here we are not interested in estimating p. Instead, we simply would like to determine whether the binary observations are truly random.

The idea of the Wald–Wolfowitz runs test is as follows. First, we condition on the number of ones, denoted by m, and so the number of zeros is $\ell = n - m$. Under this conditioning, if H_0 were true, the sequence $X_1, \dots, X_n$ would be a uniform random permutation of m ones and ℓ zeros. Next, to test whether the observed sequence is indeed uniformly random, we consider the number of runs R in the sequence, where a run is a maximal subsequence consisting of either all ones or all zeros. For example, the sequence 1110000110001 has five runs.

With m ones and ℓ zeros, the number of sequences having $R = r$ is

$$N(r) := \begin{cases} 2\binom{m-1}{k-1}\binom{\ell-1}{k-1} & \text{if } r = 2k, \\ \binom{m-1}{k-1}\binom{\ell-1}{k} + \binom{m-1}{k}\binom{\ell-1}{k-1} & \text{if } r = 2k+1. \end{cases}$$

Therefore, under H_0, the probability that the observed sequence has a number of runs between r_1 and r_2 is

$$\left(\sum_{r=r_1}^{r_2} N(r)\right) \Big/ \binom{m+\ell}{m}.$$

At significance level $\alpha \in (0,1)$, we can choose r_1 and r_2 such that the above quantity is roughly $1-\alpha$. Then we accept H_0 if $r_1 \le R \le r_2$.

A computationally easier alternative is (again) to use Gaussian approximation; one can compute

$$\mathbb{E}[R] = \frac{2m\ell}{m+\ell} + 1, \qquad \mathsf{Var}(R) = \frac{2m\ell(2m\ell - m - \ell)}{(m+\ell)^2(m+\ell-1)}.$$

2.6.2 Permutation test

The permutation test refers to a class of nonparametric tests which involve computing the values of a test statistic under all possible permutations of the observed data points. To demonstrate a typical setting, let us revisit the problem of testing whether two distributions are the same. Suppose that we observe i.i.d. $X_1, \dots, X_m \sim \mathcal{P}$ and i.i.d. $Y_1, \dots, Y_\ell \sim \mathcal{Q}$ where the two samples are independent. We test $H_0 : \mathcal{P} = \mathcal{Q}$ against $H_1 : \mathcal{P} \neq \mathcal{Q}$.

Let $T = T(X_1, \dots, X_m, Y_1, \dots, Y_\ell)$ be some test statistic, e.g., $T = |\bar{X} - \bar{Y}|$ if the means of $\mathcal{P}$ and $\mathcal{Q}$ are different. Let $n = m + \ell$ and consider all $n!$ permutations of the data points. Namely, we shuffle $X_1, \dots, X_m, Y_1, \dots, Y_\ell$, compute the mean of the first m observations and the mean of the remaining ℓ observations, and compute the difference between the two means. Let the test statistics be denoted by $T_1, \dots, T_{n!}$.

Under H_0, all these statistics are identically distributed, so all the values of these statistics are equally likely. In particular, the rank R of T among $T_1, \dots, T_{n!}$ is uniformly random. At significance level $\alpha \in (0, 1)$, we can choose r_1 and r_2 such that $\frac{r_2 - r_1}{n!} \approx 1 - \alpha$. Then we accept H_0 if $r_1 \leq R \leq r_2$.

Although the above example is specific, the same reasoning is valid in broader settings. Let $X_1, \dots, X_n$ denote the observations. Suppose that under H_0, the distribution of a test statistic $T(X_1, \dots, X_n)$ is invariant when we permute the observations (and it is not invariant under H_1). Then the same procedure applies. Finally, note that computing all $n!$ test statistics is hard if n is large. Therefore, in contrast to the previous tests that work well for a large sample, the permutation test is usually employed when the sample size is small.

Chapter 3

Extensions of the basic setup

3.1 Bayesian hypothesis testing

3.1.1 Simple hypotheses

Recall the simply hypothesis testing problem between $H_0 : X \sim p_0$ and $H_1 : X \sim p_1$, where p_0 and p_1 are densities with respect to a reference measure μ on $\mathcal{X}$. We now introduce the Bayesian point of view: Suppose that the binary parameter θ has prior distribution $\Pi = \mathsf{Ber}(\pi_1)$ for $\pi_1 \in [0, 1]$. That is, prior to observing any data, we have $\theta = 0$ with probability π_0 and $\theta = 1$ with probability $\pi_1 = 1 - \pi_0$. As before, for a test ϕ, we let

$$\alpha_\phi := \mathbb{E}_0[\phi(X)], \qquad \beta_\phi := \mathbb{E}_1[\phi(X)].$$

Moreover, let $W_0 \geq 0$ be the cost of type I error and $W_1 \geq 0$ be the cost of type II error. Then the Bayes risk of the test ϕ with respect to the prior Π is

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

We say that a test ϕ_Π is Bayes optimal with respect to the prior Π if it minimizes the Bayes risk $R_\Pi(\phi)$ over all tests.

Theorem 3.1. *In the above setting, a Bayes optimal test with respect to the prior* Π *is given by*

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

Moreover, any other Bayes optimal test must coincide with ϕ_Π *on the set*

$$\mathcal{D} := \{x \in \mathcal{X} : W_1\pi_1 p_1(x) \neq W_0\pi_0 p_0(x)\}$$

μ*-almost everywhere, and it may take arbitrary values in* $[0, 1]$ *on the set* $\mathcal{X} \setminus \mathcal{D}$.

Proof. Note that the Bayes risk can be written as

$$R_\Pi(\phi) = \int_{\mathcal{X}} W_0\pi_0\phi\, p_0\, d\mu + \int_{\mathcal{X}} W_1\pi_1(1 - \phi)p_1\, d\mu = \int_{\mathcal{X}} (W_0\pi_0 p_0 - W_1\pi_1 p_1)\phi\, d\mu + W_1\pi_1.$$

Therefore, Bayes optimal tests ϕ minimize

$$\int_{\mathcal{X}} (W_0\pi_0 p_0 - W_1\pi_1 p_1)\phi \, d\mu.$$

It is not hard to see that the test defined above is Bayes optimal. □

In other words, the Bayes optimal test ϕ_Π has region of rejection

$$\Big\{x \in \mathcal{X} : L(x) \geq \frac{W_0\pi_0}{W_1\pi_1}\Big\}, \quad \text{where } L(x) := \frac{p_1(x)}{p_0(x)}.$$

3.1.2 Composite hypotheses

Next, we turn to testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$ given $X \sim \mathcal{P}_\theta$ for $\theta \in \Theta$. For this composite hypothesis testing problem, we take the Bayesian approach and assume a prior distribution Π on Θ. Suppose that Π has density π with respect to a reference measure ν on Θ, and $\mathcal{P}_\theta$ has density p_θ with respect to a reference measure μ on $\mathcal{X}$. As $\theta \sim \Pi$ is a random variable now, it is better to write the density of X as a conditional density:

$$p(x \mid \theta) \equiv p_\theta(x), \quad x \in \mathcal{X}, \, \theta \in \Theta.$$

The posterior distribution of θ given $X = x$ has density

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

by the Bayes formula, where

$$p(x) := \int_\Theta p(x \mid \theta) \, d\Pi(\theta).$$

In general, we have

$$d\Pi(\theta \mid x) = \frac{p(x \mid \theta)}{p(x)} d\Pi(\theta).$$

We continue to let W_0 and W_1 be the costs of the two types of errors. Define the power function of a test ϕ as

$$\beta_\phi(\theta) := \mathbb{E}[\phi(X) \mid \theta], \quad \theta \in \Theta.$$

The risk function of ϕ is defined as

$$R(\theta; \phi) := W_0\beta_\phi(\theta) \cdot \mathbb{1}\{\theta \in \Theta_0\} + W_1(1 - \beta_\phi(\theta)) \cdot \mathbb{1}\{\theta \in \Theta_1\}.$$

The Bayes risk of ϕ with respect to the prior Π is

$$R_\Pi(\phi) := \int_\Theta R(\theta; \phi) \, d\Pi(\theta) = \int_{\Theta_0} W_0\beta_\phi(\theta) \, d\Pi(\theta) + \int_{\Theta_1} W_1(1 - \beta_\phi(\theta)) \, d\Pi(\theta).$$

As before, Bayes optimal tests are the minimizers of the Bayes risk.

Theorem 3.2. *In the above setting, a Bayes optimal test with respect to the prior* Π *is given by*

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

Moreover, any other Bayes optimal test must coincide with ϕ_Π *on the set*

$$\mathcal{D} := \{x \in \mathcal{X} : W_1\Pi(\Theta_1 \mid x) \neq W_0\Pi(\Theta_0 \mid x)\}$$

μ*-almost everywhere, and it may take arbitrary values in* $[0,1]$ *on the set* $\mathcal{X} \setminus \mathcal{D}$.

Proof. Rewrite the Bayes risk as

$$\begin{aligned} R_\Pi(\phi) &= \int_{\Theta_0} W_0\beta_\phi(\theta)\, d\Pi(\theta) + \int_{\Theta_1} W_1(1-\beta_\phi(\theta))\, d\Pi(\theta) \\ &= \int_\Theta \beta_\phi(\theta)\Big(W_0 \cdot \mathbb{1}\{\theta \in \Theta_0\} - W_1 \cdot \mathbb{1}\{\theta \in \Theta_1\}\Big)\, d\Pi(\theta) + W_1\Pi(\Theta_1). \end{aligned}$$

It suffices to consider minimizing the first term

$$\begin{aligned} &\int_\Theta \beta_\phi(\theta)\Big(W_0 \cdot \mathbb{1}\{\theta \in \Theta_0\} - W_1 \cdot \mathbb{1}\{\theta \in \Theta_1\}\Big)\, d\Pi(\theta) \\ &= \int_\Theta \int_{\mathcal{X}} \phi(x)p(x \mid \theta)\, d\mu(x) \cdot \Big(W_0 \cdot \mathbb{1}\{\theta \in \Theta_0\} - W_1 \cdot \mathbb{1}\{\theta \in \Theta_1\}\Big)\, d\Pi(\theta) \\ &= \int_{\mathcal{X}} \int_\Theta \phi(x)\Big(W_0 \cdot \mathbb{1}\{\theta \in \Theta_0\} - W_1 \cdot \mathbb{1}\{\theta \in \Theta_1\}\Big)p(x)\, d\Pi(\theta \mid x)\, d\mu(x) \\ &= \int_{\mathcal{X}} \phi(x)\Big(W_0 \cdot \Pi(\Theta_0 \mid x) - W_1 \cdot \Pi(\Theta_1 \mid x)\Big)p(x)\, d\mu(x). \end{aligned}$$

The conclusion follows easily as before. □

In other words, the Bayes optimal test ϕ_Π has region of rejection

$$\Big\{x \in \mathcal{X} : T(x) \geq \frac{W_0}{W_1}\Big\}, \quad \text{where } T(x) := \frac{\Pi(\Theta_1 \mid x)}{\Pi(\Theta_0 \mid x)}.$$

Here $T(x)$ is the ratio between posterior probabilities of Θ_1 and Θ_0 conditional on $X = x$.

Finally, if we define $\pi_\ell = \Pi(\Theta_\ell)$ and

$$p_\ell(x) = \int_{\Theta_\ell} \frac{p_\theta(x)}{\pi_\ell}\, d\Pi(\theta)$$

for $\ell \in \{0,1\}$, then the Bayes optimal tests for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ with respect to the prior Π coincide with those for the simple hypothesis testing problem between $H_0 : X \sim p_0$ and $H_1 : X \sim p_1$ with respect to the prior $\mathsf{Ber}(\pi_1)$.

3.2 Sequential testing

Consider the hypothesis testing problem between

- $H_0 : X_1, \ldots, X_n, \ldots$ i.i.d. from $\mathcal{P}_0$;
- $H_1 : X_1, \ldots, X_n, \ldots$ i.i.d. from $\mathcal{P}_1$.

Here the observations $X_1, \ldots, X_n, \ldots$ come in a sequential fashion, and at time n, we test H_0 against H_1 based on the observations that we have seen so far.

3.2.1 Stopping time and sequential test

A stopping time $\tau = \tau(X_1, X_2, \dots)$ is a random variable taking values in $\{0, 1, 2, \dots\}$ such that the event $\{\tau = n\}$ depends only on the observations $X_1, \dots, X_n$. This is not saying that τ itself is a function of $X_1, \dots, X_n$ for any fixed n. For example, $\tau = \inf\{n : X_n \geq 1\}$ is a stopping time: If n is the first time that X_n is at least 1, then we set τ to be n. However, if all $X_1, \dots, X_n$ are less than 1, then we have no idea about the value of τ yet. Moreover, $\tau = \sup\{n : X_n \geq 1\}$ is obviously not a stopping time.

A sequential test is a pair $(\tau, \{\phi_n\}_{n=0}^{\infty})$, where τ is a stopping time and $\phi_n = \phi_n(X_1, \dots, X_n)$ is a test based on the first n observations. For brevity, we denote the sequential test by (τ, ϕ) or simply ϕ. It does the following: If $\tau = n$, then we stop at time n and apply ϕ_n to make a decision between H_0 and H_1.

An equivalent formulation of the sequential test (τ, ϕ) is a sequence of random decisions $\nu = \{\nu_j\}_{j=0}^{\infty}$, where $\nu_j \in \{\mathsf{c}, 0, 1, *\}$. If $\tau = n$, then

- $\nu_j = \mathsf{c}$ for $j < n$, meaning that ν decides to continue and request an additional observation;
- $\nu_n \in \{0, 1\}$, meaning that ν makes a decision between H_0 and H_1;
- $\nu_j = *$ for $j > n$, meaning that the test has already stopped.

Using the terminology of statistical decision theory, we refer to each instance of ν as an action, and the sequential test is equivalent to a decision rule that generates a distribution on actions.

3.2.2 Bayes optimal sequential test

Consider the Bayesian setup with prior $\Pi = \mathsf{Ber}(\pi)$ on $\{0, 1\}$ where $\pi \in [0, 1]$. When the sequential test ϕ stops at τ and makes a decision ϕ_τ, suppose that it incurs a loss

$$L(\theta, \phi) := \lambda\tau + W_0\phi_\tau \mathbb{1}\{\theta = 0\} + W_1(1 - \phi_\tau)\mathbb{1}\{\theta = 1\} = \lambda\tau + W_\theta \phi_\tau^{1-\theta}(1 - \phi_\tau)^\theta,$$

where $\lambda > 0$ is a constant representing the cost of one observation and we set $0^0 = 1$ by convention. Note that the above formula is valid even when we consider randomized tests. This risk of the sequential test is

$$R(\theta, \phi) := \mathbb{E}_\theta[L(\theta, \phi)] = \mathbb{E}_\theta\left[\lambda\tau + W_\theta\phi_\tau(X_1, \dots, X_\tau)^{1-\theta}\big(1 - \phi_\tau(X_1, \dots, X_\tau)\big)^\theta\right].$$

The Bayes risk with respect to the prior Π is

$$R_\Pi(\phi) := \mathbb{E}[R(\theta, \phi)] = (1 - \pi)\,\mathbb{E}_0[L(0, \phi)] + \pi\,\mathbb{E}_1[L(1, \phi)].$$

We aim for a Bayes optimal sequential test which, by definition, minimizes the Bayes risk over all sequential tests.

3.2.3 Analysis of the minimum Bayes risk

First, consider the trivial case $\tau = 0$, which means that the test does not use any data. For $\alpha \in (0, 1)$, we can simply consider a constant test $\phi_\alpha = \alpha$. Its Bayes risk is

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

Therefore, the minimum Bayes risk is

$$\rho_0(\pi) = \min\{(1-\pi)W_0, \pi W_1\},$$

which is achieved by either ϕ_0 or ϕ_1. To be more precise,

$$\rho_0(\pi) = \begin{cases} \pi W_1 & \text{if } \pi \le \frac{W_0}{W_0+W_1}, \\ (1-\pi)W_0 & \text{if } \pi > \frac{W_0}{W_0+W_1}. \end{cases}$$

Next, consider the case $\tau \ge 1$, which means that the test uses at least one observation. Let be the minimum Bayes risk be

$$\rho_+(\pi) := \inf_\phi R_\Pi(\phi) = \inf_\phi \big((1-\pi)\,\mathbb{E}_0[L(0,\phi)] + \pi\,\mathbb{E}_1[L(1,\phi)]\big).$$

This is clear that $\rho_+(\pi) \ge \lambda > 0$. Moreover, $\rho_+(\pi)$ is a concave function in π because it is an infimum of linear functions. Therefore, the graphs of $\rho_+(\pi)$ and $\rho_0(\pi)$ intersect

- at two points if $\frac{W_0W_1}{W_0+W_1} > \rho_+\big(\frac{W_0}{W_0+W_1}\big)$;
- at one point if $\frac{W_0W_1}{W_0+W_1} = \rho_+\big(\frac{W_0}{W_0+W_1}\big)$;
- at zero point if $\frac{W_0W_1}{W_0+W_1} < \rho_+\big(\frac{W_0}{W_0+W_1}\big)$.

Therefore, we can find $0 < \gamma_0 \le \gamma_1 < 1$ such that $\rho_+(\pi) < \rho_0(\pi)$ for $\pi \in (\gamma_0, \gamma_1)$ and $\rho_+(\pi) \ge \rho_0(\pi)$ otherwise. This leads to the following result.

Proposition 3.3. *A Bayes optimal sequential test (τ, ϕ) must, at time 0,*

- *stop and accept H_0 if $\pi \le \gamma_0$;*
- *stop and reject H_0 if $\pi \ge \gamma_1$;*
- *continue to request one observation if $\pi \in (\gamma_0, \gamma_1)$.*

While the above result does not seem useful, the intuition extends to the following more general and useful setting. Let $\pi_0 := \pi$. By the Bayes formula, at time 1, the posterior distribution of θ on $\{0,1\}$ is $\mathsf{Ber}(\pi_1)$ where

$$\pi_1 := \mathbb{P}\{\theta = 1 \mid X_1\} = \frac{\pi \cdot p_1(X_1)}{(1-\pi)\cdot p_0(X_1) + \pi \cdot p_1(X_1)}.$$

Inductively, we see that at time n, the posterior distribution of θ is $\mathsf{Ber}(\pi_n)$ where

$$\pi_n := \mathbb{P}\{\theta = 1 \mid X_1, \dots, X_n\} = \frac{\pi \cdot p_1(X_1)\cdots p_1(X_n)}{(1-\pi)\cdot p_0(X_1)\cdots p_0(X_n) + \pi \cdot p_1(X_1)\cdots p_1(X_n)}.$$

Also, recall that the cost of an observation λ and the costs of the two types of errors W_0 and W_1 are all constant. Therefore, conditional on the event $\{\tau \ge n\}$, a Bayes optimal sequential test is expected to behave at time n in a way similar to its behavior at time 0, but with $\mathsf{Ber}(\pi_n)$ being the new prior at time n. As a result, it is plausible that the following theorem holds.

Theorem 3.4. *Define a stopping time*

$$\tau := \inf\big\{n \geq 0 : \pi_n \notin (\gamma_0, \gamma_1)\big\}.$$

For all $n \geq 0$, conditional on $\tau = n$, let ϕ_n be the test that

- *accepts H_0 if $\pi_n \leq \gamma_0$;*
- *rejects H_0 if $\pi_n \geq \gamma_1$.*

Then (τ, ϕ) is the unique Bayes optimal test in the sense that any other Bayes optimal test agrees with (τ, ϕ) with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$.

Here we say that two sequential tests (τ, ϕ) and (τ', ϕ') agree with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$ if $\mathbb{P}_i\{\tau = \tau'\} = 1$ and $\mathbb{P}_i\{\phi_n(X_1, \ldots, X_n) = \phi'_n(X_1, \ldots, X_n) \mid \tau = \tau' = n\} = 1$ for $i = 1, 2$. The proof of Theorem 3.4 is based on an inductive argument, but we omit it here.

3.2.4 Likelihood ratios for sequential testing

Observe that the posterior π_n can be expressed in terms of the likelihood ratio

$$L_n = L_n(X_1, \ldots, X_n) := \frac{p_1(X_1) \cdots p_1(X_n)}{p_0(X_1) \cdots p_0(X_n)}.$$

Namely, we have

$$\frac{1}{\pi_n} = \frac{1-\pi}{\pi}\frac{1}{L_n} + 1, \qquad L_n = \frac{1-\pi}{\pi}\frac{\pi_n}{1-\pi_n}.$$

As a result,

- $\pi_n \leq \gamma_0$ if and only if $L_n \leq \Gamma_0 := \frac{1-\pi}{\pi}\frac{\gamma_0}{1-\gamma_0}$;
- $\pi_n \geq \gamma_1$ if and only if $L_n \geq \Gamma_1 := \frac{1-\pi}{\pi}\frac{\gamma_1}{1-\gamma_1}$.

Therefore, Theorem 3.4 can be rewritten as follows.

Theorem 3.5. *Define a stopping time*

$$\tau(\Gamma_0, \Gamma_1) := \inf\big\{n \geq 0 : L_n \notin (\Gamma_0, \Gamma_1)\big\}.$$

For all $n \geq 0$, conditional on $\tau(\Gamma_0, \Gamma_1) = n$, let ϕ_n be the test that

- *accepts H_0 if $L_n \leq \Gamma_0$;*
- *rejects H_0 if $L_n \geq \Gamma_1$.*

Then (τ, ϕ) is the unique Bayes optimal test.

To further study a likelihood-ratio sequential test ϕ of the above form, let $\alpha_0(\phi)$ denote the probability of a type I error and $\alpha_1(\phi)$ denote the probability of a type II error. Then the Bayes risk of ϕ is

$$R_\Pi(\phi) = (1-\pi)\big(\lambda\,\mathbb{E}_0[\tau] + W_0\,\alpha_0(\phi)\big) + \pi\big(\lambda\,\mathbb{E}_1[\tau] + W_1\,\alpha_1(\phi)\big). \tag{3.1}$$

The following result says that, among all sequential tests achieving certain average type I and type II errors, the likelihood-ratio sequential test does it the fastest.

Theorem 3.6. *Fix $0 < \Gamma_0 \le \Gamma_1 < \infty$. Let $\phi(\Gamma_0, \Gamma_1)$ denote the likelihood-ratio sequential test defined above, and let $\tau(\Gamma_0, \Gamma_1)$ denote its stopping time. For any sequential test (τ, ϕ) such that*

$$\alpha_0(\phi) \le \alpha_0\big(\phi(\Gamma_0, \Gamma_1)\big), \qquad \alpha_1(\phi) \le \alpha_1\big(\phi(\Gamma_0, \Gamma_1)\big),$$

it holds that

$$\mathbb{E}_0[\tau] \ge \mathbb{E}_0\left[\tau(\Gamma_0, \Gamma_1)\right], \qquad \mathbb{E}_0[\tau] \ge \mathbb{E}_0\left[\tau(\Gamma_0, \Gamma_1)\right].$$

Proof. We only provide a sketch of the proof. Recall that in the last section, γ_0 and γ_1 are defined as the points where $\rho_0(\pi)$ and $\rho_+(\pi)$ intersect. Then $\Gamma_i := \frac{1-\pi}{\pi}\frac{\gamma_i}{1-\gamma_i}$ for $i = 0, 1$. Based on properties of $\Gamma_i(W_0, W_1)$ as a function of W_0 and W_1, one can show that W_0 and W_1 can be chosen so that $\Gamma_i(W_0, W_1)$ is equal to any value of Γ_i. By Theorem 3.5, we have that $\phi(\Gamma_0, \Gamma_1)$ is the Bayes optimal test for the risk with this choice of weights. The conclusion follows by virtue of (3.1). □

Ideally, we would like to set the thresholds Γ_0 and Γ_1 to obtain the Bayes optimal likelihood-ratio sequential test or to achieve certain average type I and type II errors. Neither of these tasks is easy in general because the thresholds are implicitly defined. The following result provides an approximate solution to the latter problem.

Proposition 3.7. *We have*

$$\alpha_0\big(\phi(\Gamma_0, \Gamma_1)\big) \le \frac{1}{\Gamma_1}\Big(1 - \alpha_1\big(\phi(\Gamma_0, \Gamma_1)\big)\Big), \qquad \alpha_1\big(\phi(\Gamma_0, \Gamma_1)\big) \le \Gamma_0\Big(1 - \alpha_0\big(\phi(\Gamma_0, \Gamma_1)\big)\Big).$$

Proof. For $0 < \Gamma_0 \le \Gamma_1 < \infty$, define

$$S_n := \big\{(x_1, \dots, x_n) : L_k(x_1, \dots, x_k) \in (\Gamma_0, \Gamma_1) \text{ for } k = 1, \dots, n-1, \text{ and } L_n(x_1, \dots, x_n) \ge \Gamma_1\big\}.$$

Then we obtain

$$\begin{aligned}
\alpha_0\big(\phi(\Gamma_0, \Gamma_1)\big) &= \sum_{n=1}^{\infty} \int_{S_n} p_0(x_1) \cdots p_0(x_n)\, d\mu(x_1) \cdots d\mu(x_n) \\
&\le \frac{1}{\Gamma_1} \sum_{n=1}^{\infty} \int_{S_n} p_1(x_1) \cdots p_1(x_n)\, d\mu(x_1) \cdots d\mu(x_n) \\
&= \frac{1}{\Gamma_1}\Big(1 - \alpha_1\big(\phi(\Gamma_0, \Gamma_1)\big)\Big).
\end{aligned}$$

The proof of the second inequality is similar. □

Consequently, if $\alpha_i\big(\phi(\Gamma_0, \Gamma_1)\big) = \alpha_i$ for $i = 0, 1$, then

$$\alpha_0 \le \frac{1}{\Gamma_1}(1 - \alpha_1), \qquad \alpha_1 \le \Gamma_0(1 - \alpha_0),$$

which implies that

$$\Gamma_0 \ge \Gamma_0' := \frac{\alpha_1}{1 - \alpha_0}, \qquad \Gamma_1 \le \Gamma_1' := \frac{1 - \alpha_1}{\alpha_0}.$$

Applying the above proposition again, we obtain

$$\begin{aligned}
\alpha_0\big(\phi(\Gamma_0', \Gamma_1')\big) &\le \frac{1}{\Gamma_1'}\Big(1 - \alpha_1\big(\phi(\Gamma_0', \Gamma_1')\big)\Big) \le \frac{1}{\Gamma_1'} = \frac{\alpha_0}{1 - \alpha_1}, \\
\alpha_1\big(\phi(\Gamma_0', \Gamma_1')\big) &\le \Gamma_0'\Big(1 - \alpha_0\big(\phi(\Gamma_0', \Gamma_1')\big)\Big) \le \Gamma_0' = \frac{\alpha_1}{1 - \alpha_0}.
\end{aligned}$$

Therefore, if α_i is small for $i = 0, 1$, then choosing the thresholds Γ_0' and Γ_1' for the likelihood-ratio sequential test yields the desired average errors approximately.

3.3 Generalized Neyman–Pearson lemma

Consider integrable functions $f_1, \dots, f_{m+1} : \mathcal{X} \to \mathbb{R}$ and real numbers $\alpha_1, \dots, \alpha_m$. The generalized Neyman–Pearson lemma describes the solutions of the optimization problem

$$\max_{\phi:\mathcal{X}\to[0,1]} \int \phi f_{m+1}\, d\mu \quad \text{s.t.} \quad \int \phi f_j\, d\mu = \alpha_j,\ j = 1, \dots, m. \tag{3.2}$$

The integrals in this section are all over $\mathcal{X}$ unless otherwise specified.

Theorem 3.8. *Let Φ_α denote the set of tests ϕ which satisfy the constraints in* (3.2)*. If Φ_α is nonempty, then the following statements hold:*

1. *There exists a function $\phi^* : \mathcal{X} \to [0,1]$ that solves* (3.2)*.*

2. *If*
$$\phi^*(x) := \begin{cases} 1 & \text{if } f_{m+1}(x) > \sum_{j=1}^m u_j f_j(x), \\ 0 & \text{if } f_{m+1}(x) < \sum_{j=1}^m u_j f_j(x), \end{cases} \tag{3.3}$$
for constants $u_1, \dots, u_m \in \mathbb{R}$ and $\phi^* \in \Phi_\alpha$, then ϕ^* solves (3.2)*.*

3. *If the above ϕ^* is defined with $u_j \geq 0$ for $j = 1, \dots, m$ and $\phi^* \in \Phi_\alpha$, then ϕ^* solves the problem*
$$\max_{\phi:\mathcal{X}\to[0,1]} \int \phi f_{m+1}\, d\mu \quad \text{s.t.} \quad \int \phi f_j\, d\mu \leq \alpha_j,\ j = 1, \dots, m.$$

4. *Let Φ denote the set of all tests $\phi : \mathcal{X} \to [0,1]$. The set*
$$\mathcal{C} := \Big\{ \Big(\int \phi f_1\, d\mu, \dots, \int \phi f_m\, d\mu \Big) : \phi \in \Phi \Big\}$$
is a closed and convex subset of $\mathbb{R}^m$. If $(\alpha_1, \dots, \alpha_m)$ belongs to $\mathsf{relint}(\mathcal{C})$, the relative interior of the set $\mathcal{C}$, then there exist constants $u_1, \dots, u_m \in \mathbb{R}$ and a test ϕ^* satisfying (3.3) *that solves* (3.2)*. Moreover, the condition* (3.3) *is necessary for any solution of* (3.2)*.*

3.3.1 Proof of the theorem

1. Note that
$$\sup_{\phi\in\Phi_\alpha} \int \phi f_{m+1}\, d\mu \leq \int |f_{m+1}|\, d\mu < \infty,$$
and there exists a sequence of tests ϕ_n such that
$$\int \phi_n f_{m+1}\, d\mu \to \sup_{\phi\in\Phi_\alpha} \int \phi f_{m+1}\, d\mu$$
as $n \to \infty$. Then with some standard arguments in analysis and topology, we can show that there exists a subsequence of tests ϕ_{n_k} and a test $\phi^* \in \Phi_\alpha$ such that
$$\int \phi_{n_k} f\, d\mu \to \int \phi^* f\, d\mu$$
as $k \to \infty$ for all integrable functions f. Applying the above convergence to $f = f_{m+1}$ proves the first claim.

2. Let ϕ^* be a test satisfying (3.3). Then for any test $\phi \in \Phi_\alpha$, we have

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

since the integrand is nonnegative. This implies that

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

since $\phi^*, \phi \in \Phi_\alpha$. This proves the second claim.

3. This follows similarly.

4. This part is more advanced. We first prove that $\mathcal{C}$ is convex and closed; the rest will be proved later in this section. The set Φ of all tests $\phi : \mathcal{X} \to [0,1]$ is convex, and the mapping

$$\phi \mapsto \Big(\int \phi f_1\, d\mu, \dots, \int \phi f_m\, d\mu\Big)$$

is linear. As the image of a convex set under a linear mapping, the set $\mathcal{C}$ is convex. To see that $\mathcal{C}$ is closed, suppose that there is a sequence of tests ϕ_n such that

$$\Big(\int \phi_n f_1\, d\mu, \dots, \int \phi_n f_m\, d\mu\Big) \to v$$

as $n \to \infty$ for a point $v \in \mathcal{C}$. It suffices to show that $v \in \mathcal{C}$. For this, we again extract a subsequence ϕ_{n_k} such that $\int \phi_{n_k} f\, d\mu \to \int \phi f\, d\mu$ as $k \to \infty$ for some $\phi \in \Phi$ and all integrable functions f. It then follows that $v = (\int \phi f_1\, d\mu, \dots, \int \phi f_m\, d\mu) \in \mathcal{C}$.

To prove the final part of the theorem, we need some preparations. Let

$$\mathcal{D} := \Big\{ \Big(\int \phi f_1\, d\mu, \dots, \int \phi f_m\, d\mu, \int \phi f_{m+1}\, d\mu\Big) : \phi \in \Phi \Big\}.$$

Similarly, $\mathcal{D}$ is a closed and convex subset of $\mathbb{R}^{m+1}$. Fix $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathsf{relint}(\mathcal{C})$. Define

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

and

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

By Part 1 of the theorem, $c^+(\alpha) \in \mathcal{D}$; similarly, $c^-(\alpha) \in \mathcal{D}$. Then, by the convexity of $\mathcal{D}$, the segment connecting $c^+(\alpha)$ and $c^-(\alpha)$ is contained in $\mathcal{D}$. We therefore have the following lemma.

Lemma 3.9. *The function $c^+_{m+1}(\alpha)$ is concave and the function $c^-_{m+1}(\alpha)$ is convex; in particular, both are continuous. Moreover, the set $\mathcal{D}$ can be represented as*

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

Informally, the set $\mathcal{D}$ has the set $\mathcal{C}$ as its "base". The "upper" boundary of $\mathcal{D}$ consists of graphs of the concave functions $c^+_{m+1}(\alpha)$, $\alpha \in \mathcal{C}$, and the "lower" boundary of $\mathcal{D}$ consists of graphs of the convex functions $c^-_{m+1}(\alpha)$, $\alpha \in \mathcal{C}$.

We split the rest of the proof into two cases.

Case 1: $c^+_{m+1}(\alpha) = c^-_{m+1}(\alpha)$ for some $\alpha \in \mathsf{relint}(\mathcal{C})$. Let us start with the following lemma.

Lemma 3.10. *If $c^+_{m+1}(\alpha) = c^-_{m+1}(\alpha)$ for some $\alpha \in \mathsf{relint}(\mathcal{C})$, then $c^+_{m+1}(\beta) = c^-_{m+1}(\beta)$ for all $\beta \in \mathcal{C}$. In this case, $c_{m+1}(\beta) := c^+_{m+1}(\beta) = c^-_{m+1}(\beta)$ is a linear function and thus can be written $c_{m+1}(\beta) = \langle u, \beta \rangle$ for some $u \in \mathbb{R}^m$. Moreover, the set $\mathcal{D}$ belongs to a hyperplane in $\mathbb{R}^{m+1}$ passing through the origin $\mathbf{0} \in \mathbb{R}^{m+1}$.*

Proof. We only sketch the proof. By the concavity of $c^+_{m+1}(\beta)$ and the convexity of $c^-_{m+1}(\beta)$, one can prove the first claim, e.g., via a contrapositive argument. Since $c_{m+1}(\beta)$ is both convex and concave, it is not hard to see that we must have

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

Moreover, the origin is contained in $\mathcal{D}$ since we can take $\phi \equiv 0$. Setting $\alpha = \mathbf{0}$ in the above equation yields

$$c_{m+1}(\beta) = c_{m+1}(0) + \langle u, \beta \rangle = \langle u, \beta \rangle,$$

finishing the proof. □

Any test $\phi \in \Phi$ corresponds to $x \in \mathcal{D}$ via the relation $x_j = \int \phi f_j \, d\mu$. By the above lemma, we have $x_{m+1} = \sum_{j=1}^m u_j x_j$ for some $u \in \mathbb{R}^m$. It then follows that

$$\int \phi f_{m+1} \, d\mu = \int \phi \cdot \Big(\sum_{j=1}^m u_j f_j \Big) \, d\mu$$

for any $\phi \in \Phi$. Therefore, we must have $f_{m+1} = \sum_{j=1}^m u_j f_j$ almost everywhere, and the condition (3.3) holds trivially.

Case 2: $c^-_{m+1}(\alpha) < c^+_{m+1}(\alpha)$ for all $\alpha \in \mathsf{relint}(\mathcal{C})$. In this case, there exists a hyperplane H in $\mathbb{R}^{m+1}$ passing through the point $c^+(\alpha)$ such that the set $\mathcal{D}$ lies on one side of H and that H does not intersect $\mathsf{relint}(\mathcal{D})$. In other words, there is a vector $w = (u, \lambda) \in \mathbb{R}^{m+1}$ for $u \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$ such that

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

and

$$\langle x, w \rangle \geq \langle c^+(\alpha), w \rangle \quad \text{for } x \in \mathcal{D}, \tag{3.4a}$$
$$\langle x, w \rangle > \langle c^+(\alpha), w \rangle \quad \text{for } x \in \mathsf{relint}(\mathcal{D}). \tag{3.4b}$$

We claim that $\lambda \neq 0$. To see this, suppose that $\lambda = 0$. Then H is orthogonal to the hyperplane containing $\mathcal{C}$, so H contains the segment between $c^+(\alpha)$ and $c^-(\alpha)$. However, it is not hard to show that a point in the middle of the segment is in $\mathsf{relint}(\mathcal{D})$, which yields a contradiction.

As a result, we can rescale the vector w so that $\lambda = -1$, let $x = (\tilde{x}, x_{m+1})$, and rewrite (3.4) as

$$\langle \tilde{x}, u \rangle - x_{m+1} \geq \langle \alpha, u \rangle - c^+_{m+1}(\alpha) \quad \text{for } x \in \mathcal{D}, \tag{3.5a}$$
$$\langle \tilde{x}, u \rangle - x_{m+1} > \langle \alpha, u \rangle - c^+_{m+1}(\alpha) \quad \text{for } x \in \mathsf{relint}(\mathcal{D}). \tag{3.5b}$$

Let $\phi^* \in \Phi_\alpha$ be a test maximizing $\int \phi f_{m+1} \, d\mu$. Then we have $c^+_{m+1}(\alpha) = \int \phi^* f_{m+1} \, d\mu$. For

$$x = (\tilde{x}, x_{m+1}) = \Big(\int \phi f_1 \, d\mu, \ldots, \int \phi f_m \, d\mu, \int \phi f_{m+1} \, d\mu \Big) \in \mathcal{D}$$

where $\phi \in \Phi$, we can further derive from (3.5) that

$$\int \phi^* \Big(f_{m+1} - \sum_{j=1}^m u_j f_j \Big)\, d\mu = c_{m+1}^+(\alpha) - \langle \alpha, u \rangle \geq x_{m+1} - \langle \tilde{x}, u \rangle = \int \phi \Big(f_{m+1} - \sum_{j=1}^m u_j f_j \Big)\, d\mu$$

for $\phi \in \Phi$. For this, it is necessary and sufficient that ϕ^* satisfies (3.3).

3.3.2 Application to two-sided testing

For $\theta \in \Theta \subset \mathbb{R}$, let X be from an exponential family with density

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

for $x \in \mathcal{X}$, where $T(x) \in \mathbb{R}$, $h(x) > 0$, and

$$Z(\theta) := \int_{\mathcal{X}} \exp\big(\theta\, T(x)\big)\, h(x)\, d\mu(x) < \infty.$$

Note that $Z(\theta)$ is a convex function, so we may assume that Θ is an open interval in $\mathbb{R}$. Here $T(X)$ is known as the sufficient statistic of this exponential family. We assume that the function $T(x)$ is not constant to avoid the trivial case.

Recall that there is no UMP test for the two-sided testing problem between $H_0 : \theta \in \Theta_0 = [\theta_1, \theta_2]$ and $H_1 : \theta \in \Theta_1 = [\theta_1, \theta_2]^c$. In this case, we may instead seek an unbiased test ϕ whose power function $\beta_\phi(\theta)$ satisfies $\beta_\phi(\theta) \leq \alpha$ for $\theta \in \Theta_0$ and $\beta_\phi(\theta) \geq \alpha$ for $\theta \in \Theta_1$. Let us now consider a more ambitious goal.

A test $\phi^* : \mathcal{X} \to [0, 1]$ is called a uniformly most powerful unbiased (UMPU) test at significance level $\alpha \in (0, 1)$ for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, if it is unbiased and satisfies $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$ at all $\theta \in \Theta_1$ for any unbiased test ϕ at significance level α.

Any UMP test is UMPU, but the converse is not true. We have the following result which is a consequence of the generalized Neyman–Pearson lemma.

Theorem 3.11. *Let $\mathcal{P}_\theta$ denote the exponential family defined above. Let $X \sim \mathcal{P}_\theta$. For $\theta_1, \theta_2 \in \Theta$, $\theta_1 < \theta_2$, and $\alpha \in (0, 1)$, there exists a UMPU test ϕ^* at significance level α for testing $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1 : \theta \notin [\theta_1, \theta_2]$. Moreover, ϕ^* satisfies*

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [\tau_1, \tau_2], \\ 0 & \text{if } T(x) \in (\tau_1, \tau_2), \end{cases} \tag{3.6}$$

for constants $\tau_1, \tau_2 \in \mathbb{R}$.

Proof. In the case of exponential family, the power function $\beta_\phi(\theta)$ is continuous for any test ϕ. Therefore, if ϕ is unbiased, then $\beta_\phi(\theta_1) = \beta_\phi(\theta_2) = \alpha$. Moreover, we have $\beta_\phi(\theta) = \int_{\mathcal{X}} \phi\, p_\theta\, d\mu$. Fix $\theta' > \theta_2$, and consider the following optimization problem:

$$\max_\phi \int \phi\, p_{\theta'}\, d\mu \quad \text{s.t.} \int \phi\, p_{\theta_1}\, d\mu = \int \phi\, p_{\theta_2}\, d\mu = \alpha,$$

where all the integrals are over the sample space $\mathcal{X}$.

We will use the generalized Neyman–Pearson lemma to show that the above problem has a solution ϕ^* of the form:

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

for certain constants u_1 and u_2. To this end, it suffices to check that (α, α) is in the relative interior of the convex set

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

where Φ denotes the set of all tests. We may assume that $p_{\theta_1} \neq p_{\theta_2}$. By considering the power β_+^1 of the most powerful test at significance level α for testing $H_0 : \theta = \theta_1$ against $H_1 : \theta = \theta_2$, we see that $(\alpha, \beta_+^1) \in \mathcal{C}$ where $\beta_+^1 > \alpha$. Similarly, considering the least powerful test gives $(\alpha, \beta_-^1) \in \mathcal{C}$ for $\beta_-^1 < \alpha$. Swapping θ_1 and θ_2 yields $(\beta_+^0, \alpha) \in \mathcal{C}$ where $\beta_+^0 > \alpha$ and $(\beta_-^0, \alpha) \in \mathcal{C}$ for $\beta_-^0 < \alpha$. Then the convexity of $\mathcal{C}$ guarantees that (α, α) is in the relative interior of $\mathcal{C}$.

Next, note that the condition $p_{\theta'}(x) > u_1 p_{\theta_1}(x) + u_2 p_{\theta_2}(x)$ is equivalent to

$$\frac{1}{Z(\theta')} e^{\theta' T(x)} > \frac{u_1}{Z(\theta_1)} e^{\theta_1 T(x)} + \frac{u_2}{Z(\theta_2)} e^{\theta_2 T(x)}. \tag{3.7}$$

We now split the proof into several cases:

- $u_1 \leq 0$ and $u_2 \leq 0$: The condition (3.7) always holds, so $\phi^*(x) \equiv 1$ and the test is not unbiased.

- $u_1 > 0$ and $u_2 \leq 0$: The condition (3.7) can be rewritten as

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

  The left-hand side is an increasing function in $T(x)$, so the region of rejection of ϕ^* is determined by $T(x) > c$ for a constant c. However, as we have seen in Theorem 1.2, such a test is known to have a strictly increasing power function and cannot satisfy $\beta_{\phi^*}(\theta_1) = \beta_{\phi^*}(\theta_2) = \alpha$.

- $u_1 \geq 0$ and $u_2 > 0$: Rewriting the condition (3.7) as

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

  we can use a similar argument to show that this is impossible.

- $u_1 < 0$ and $u_2 > 0$: In view of the above reformulation of the condition (3.7), since the left-hand side is convex, the test ϕ^* must be of the form (3.6).

A similar analysis can be applied to maximizing the power at $\theta' < \theta_1$, and also to minimizing the power at $\theta' \in (\theta_1, \theta_2)$. The conclusion is that ϕ^* of the form (3.6) maximizes the power outside $[\theta_1, \theta_2]$ and minimizes the power inside (θ_1, θ_2) subject to the constraints that the power is equal to α at θ_1 and θ_2. This shows that ϕ^* is the UMPU test at significance level α.

There was a caveat in the above analysis: For different θ', the constants τ_1 and τ_2 might not be the same, so ϕ^* is not well-defined. However, this is not a problem. It can be shown that we may take

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [\tau_1, \tau_2], \\ \gamma_1 & \text{if } T(x) = \tau_1, \\ \gamma_2 & \text{if } T(x) = \tau_2, \\ 0 & \text{if } T(x) \in (\tau_1, \tau_2), \end{cases}$$

for constants $\gamma_1, \gamma_2 \in [0, 1]$; furthermore, the constants $\tau_1, \tau_2, \gamma_1, \gamma_2$ are determined by the constraint $\beta_{\phi^*}(\theta_1) = \beta_{\phi^*}(\theta_2) = \alpha$ (and thus do not depend on θ'). □

3.3.3 Testing equality

We continue to consider the exponential family defined in the last subsection. The testing problem between $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$ is the limiting case of the two-sided testing problem above, with $\theta_1 = \theta_2 = \theta_0$. It is possible to develop a UMPU test for this problem, where the power function satisfies $\beta_\phi(\theta_0) = \alpha$ and $\beta'_\phi(\theta_0) = 0$. The power function is differentiable for any test in the case of one-parameter exponential family, so the condition $\beta'_\phi(\theta_0) = 0$ is a consequence of the fact that $\beta_\phi(\theta)$ is minimized at θ_0. The following identity holds.

Proposition 3.12. *For any test ϕ and any $\theta \in \Theta$, we have*

$$\beta'_\phi(\theta) = \mathsf{Cov}_\theta\big(\phi(X), T(X)\big).$$

Proof. We have

$$\begin{aligned} \beta'_\phi(\theta) &= \frac{d}{d\theta} \int \phi \, p_\theta \, d\mu \\ &= \frac{d}{d\theta} \int \phi(x) \, \frac{1}{Z(\theta)} \exp\big(\theta \, T(x)\big) \, h(x) \, d\mu(x) \\ &= \int \phi(x) \, \frac{T(x)}{Z(\theta)} \exp\big(\theta \, T(x)\big) \, h(x) \, d\mu(x) - \int \phi(x) \, \frac{Z'(\theta)}{Z(\theta)^2} \exp\big(\theta \, T(x)\big) \, h(x) \, d\mu(x) \\ &= \int \phi \, T \, p_\theta \, d\mu - \frac{Z'(\theta)}{Z(\theta)} \int \phi \, p_\theta \, d\mu \\ &= \mathbb{E}_\theta[\phi(X) \, T(X)] - \frac{Z'(\theta)}{Z(\theta)} \, \mathbb{E}_\theta[\phi(X)]. \end{aligned}$$

Moreover,

$$\frac{Z'(\theta)}{Z(\theta)} = \frac{1}{Z(\theta)} \cdot \frac{d}{d\theta} \int \exp\big(\theta \, T(x)\big) \, h(x) \, d\mu(x) = \frac{1}{Z(\theta)} \int T(x) \exp\big(\theta \, T(x)\big) \, h(x) \, d\mu(x) = \mathbb{E}_\theta[T(X)].$$

It follows that

$$\beta'_\phi(\theta) = \mathbb{E}_\theta[\phi(X) \, T(X)] - \mathbb{E}_\theta[T(X)] \; \mathbb{E}_\theta[\phi(X)] = \mathsf{Cov}_\theta\big(\phi(X), T(X)\big).$$

□

Therefore, the condition $\beta'_\phi(\theta_0) = 0$ is equivalent to saying that $\phi(X)$ and $T(X)$ are uncorrelated for $X \sim p_{\theta_0}$. In addition, if $\beta_\phi(\theta_0) = \mathbb{E}_{\theta_0}[\phi(X)] = \alpha$, then

$$\mathbb{E}_{\theta_0}[\phi(X)\,T(X)] = \alpha\,\mathbb{E}_{\theta_0}[T(X)]$$

or

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

To find a UMPU test at significance level α, we can solve the following optimization problem for $\theta \neq \theta_0$:

$$\max_\phi \int \phi\,p_\theta\,d\mu \quad \text{s.t.} \int \phi\,p_{\theta_0}\,d\mu = \alpha, \quad \int \phi\,T\,p_{\theta_0}\,d\mu = \alpha \int T\,p_{\theta_0}\,d\mu.$$

A similar analysis based on the generalized Neyman–Pearson lemma yields a UMPU test of the form

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [\tau_1, \tau_2], \\ \gamma_1 & \text{if } T(x) = \tau_1, \\ \gamma_2 & \text{if } T(x) = \tau_2, \\ 0 & \text{if } T(x) \in (\tau_1, \tau_2), \end{cases}$$

where the constants $\gamma_1, \gamma_2 \in [0, 1]$ and τ_1, τ_2 are determined by the constraints in the above optimization problem. The conclusion is stated as the following theorem.

Theorem 3.13. *Let $\mathcal{P}_\theta$ be a one-parameter exponential family, where θ belongs to an open interval $\Theta \subset \mathbb{R}$. Let $X \sim \mathcal{P}_\theta$. For $\theta_0 \in \Theta$ and $\alpha \in (0, 1)$, there exists a UMPU test ϕ^* at significance level α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Moreover, ϕ^* can be taken to have the above form.*

We conclude this section by revisiting the one-sided problem of testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Note that the one-sided test ϕ based on a statistic $T(X)$ can be viewed as an increasing function of T:

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

where $\gamma \in [0, 1]$. Then we can show that such a test ϕ does not satisfy $\mathsf{Cov}(\phi(T), T) = 0$ using the following result.

Proposition 3.14. *Let T be a real-valued random variable, and let $\phi : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then we have $\mathsf{Cov}(\phi(T), T) \geq 0$. Moreover, the inequality is strict unless $\phi(T)$ is constant almost surely.*

Proof. Since $T - \mathbb{E}[T]$ is a random variable and $\phi(T) - \mathbb{E}[\phi(T)]$ is increasing in $T - \mathbb{E}[T]$, we may assume without loss of generality that $\mathbb{E}[T] = 0$ and $\mathbb{E}[\phi(T)] = 0$. Then we have

$$\mathsf{Cov}(\phi(T), T) = \int_{\mathbb{R}} \phi(t) \cdot t\,d\mathcal{P}(t) = \int_{\mathbb{R}} \big(\phi(t) - \phi(0)\big) t\,d\mathcal{P}(t),$$

since $\int_{\mathbb{R}} t\,d\mathcal{P}(t) = 0$. As $\phi(t)$ is increasing, the integrand is nonnegative everywhere on $\mathbb{R}$. The conclusion follows. □

Consequently, we have confirmed that the one-sided test is not UMPU for testing equality.

3.4 Testing in higher dimensions

So far, our discussion has mainly focused on the case where the parameter θ is real-valued. Let us now consider testing in higher dimensions.

3.4.1 Multivariate exponential family

We start with a technical lemma.

Lemma 3.15. *Let U be a random vector in $\mathbb{R}^\ell$ and T be a random vector in $\mathbb{R}^m$. Under H_i for $i = 0, 1$, let $\mathcal{P}_i$ be the joint distribution of (U, T). Let $\mathcal{Q}_i$ be the marginal distribution of T. Let $\mathcal{R}_{i,t}$ be the conditional distribution of U given $T = t$. Let the corresponding lower-case letters denote the densities of the above distributions. Suppose that*

$$\frac{p_1(u,t)}{p_0(u,t)} = a(u)b(t)$$

for functions $a(u) > 0$ and $b(t)$. Then we have

$$\frac{q_1(t)}{q_0(t)} = b(t)\,\mathbb{E}_0[a(U) \mid T = t], \qquad \frac{r_{1,t}(u)}{r_{0,t}(u)} = \frac{a(u)}{\mathbb{E}_0[a(U) \mid T = t]}. \tag{3.8}$$

Proof. For any $B \subset \mathbb{R}^m$, we have

$$\begin{aligned}
\mathbb{P}_1\{T \in B\} &= \mathbb{E}_0[\mathbb{1}\{T \in B\}\, a(U)\, b(T)] \\
&= \mathbb{E}_0\Big[\,\mathbb{E}_0[a(U) \mid T] \cdot \mathbb{1}\{T \in B\}\, b(T)\Big] \\
&= \int_B b(t)\ \mathbb{E}_0[a(U) \mid T = t]\, q_0(t)\, dt,
\end{aligned}$$

so the first equation in (3.8) holds.

Moreover, for any $C \subset \mathbb{R}^\ell$, we have

$$\begin{aligned}
\mathbb{P}_1\{T \in B,\, U \in C\} &= \mathbb{E}_0[\mathbb{1}\{T \in B,\, U \in C\}\, a(U)\, b(T)] \\
&= \mathbb{E}_0\Big[\,\mathbb{E}_0[\mathbb{1}\{U \in C\}\, a(U) \mid T] \cdot \mathbb{1}\{T \in B\}\, b(T)\Big] \\
&= \int_B \Big(\int_C a(u)\, r_{0,t}(u)\, du\Big)\, b(t)\, q_0(t)\, dt \\
&= \int_B \Big(\int_C a(u)\, r_{0,t}(u)\, du\Big)\, \frac{q_1(t)}{\mathbb{E}_0[a(U) \mid T = t]}\, dt,
\end{aligned}$$

where the last equality follows from the first equation in (3.8). On the other hand, we have

$$\mathbb{P}_1\{T \in B,\, U \in C\} = \int_B \Big(\int_C r_{1,t}(u)\, du\Big) q_1(t)\, dt.$$

The above two displays together imply that

$$r_{1,t}(u) = \frac{a(u)\, r_{0,t}(u)}{\mathbb{E}_0[a(U) \mid T = t]},$$

proving the second equation in (3.8). □

Next, consider an $(\ell + m)$-parameter exponential family on $\mathbb{R}^n$ with densities

$$p_{\theta,\eta}(x) = h(x) \exp\Big(\theta^\top U(x) + \eta^\top T(x) - A(\theta,\eta)\Big), \quad x \in \mathbb{R}^n, \tag{3.9}$$

where $\theta, U(x) \in \mathbb{R}^\ell$ and $\eta, T(x) \in \mathbb{R}^m$. The following result gives the marginal and conditional distributions of the sufficient statistics U and T.

Proposition 3.16. *If $X \sim p_{\theta,\eta}$, then there exist measures λ_θ on $\mathbb{R}^m$ and ν_t on $\mathbb{R}^\ell$ such that:*

- *With $\theta \in \mathbb{R}^\ell$ fixed, the marginal distributions of T form an m-parameter exponential family with densities*
$$q_{\theta,\eta}(t) = \exp\Big(\eta^\top t - A(\theta,\eta)\Big), \quad t \in \mathbb{R}^m,$$
with respect to the measure λ_θ.

- *The conditional distributions of U given $T = t$ form an ℓ-parameter exponential family with densities*
$$r_{\theta,t}(u) = \exp\Big(\theta^\top u - A_t(\theta)\Big), \quad u \in \mathbb{R}^\ell,$$
with respect to the measure ν_t for some function $A_t(\theta)$. In particular, these densities do not depend on $\eta \in \mathbb{R}^m$.

Proof. Let $\omega_{\theta,\eta}$ denote the joint distribution of (U,T) under $p_{\theta,\eta}$. Fix parameters θ_0 and η_0. Then

$$\frac{\omega_{\theta,\eta}}{\omega_{\theta_0,\eta_0}} = \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}} = \exp\Big((\theta-\theta_0)^\top u + (\eta-\eta_0)^\top t + A(\theta_0,\eta_0) - A(\theta,\eta)\Big).$$

By the first equation in (3.8),

$$\begin{aligned}\frac{q_{\theta,\eta}(t)}{q_{\theta_0,\eta_0}(t)} &= \exp\Big((\eta-\eta_0)^\top t + A(\theta_0,\eta_0) - A(\theta,\eta)\Big) \int \exp\Big((\theta-\theta_0)^\top u\Big)\, r_{\theta_0,t}(u)\, du \\ &= \exp\Big(\eta^\top t - A(\theta,\eta)\Big) \int \exp\Big((\theta-\theta_0)^\top u - \eta_0^\top t + A(\theta_0,\eta_0)\Big)\, r_{\theta_0,t}(u)\, du.\end{aligned}$$

Therefore, $q_{\theta,\eta}(t)$ is of the desired form once we take λ_θ to be the measure such that

$$\frac{1}{q_{\theta_0,\eta_0}(t)} = \int \exp\Big((\theta-\theta_0)^\top u - \eta_0^\top t + A(\theta_0,\eta_0)\Big)\, r_{\theta_0,t}(u)\, du.$$

Moreover, by the second equation in (3.8),

$$\frac{r_{\theta,t}(u)}{r_{\theta_0,t}(u)} = \frac{\exp\big((\theta-\theta_0)^\top u\big)}{\int \exp\big((\theta-\theta_0)^\top w\big)\, r_{\theta_0,t}(w)\, dw} = \exp\Big(\theta^\top u - A_t(\theta)\Big) \exp(-\theta_0^\top u),$$

where

$$A_t(\theta) := \log \int \exp\big((\theta-\theta_0)^\top w\big)\, r_{\theta_0,t}(w)\, dw.$$

Then $r_{\theta,t}(u)$ is of the desired form once we take ν_t to be the measure such that

$$\frac{1}{r_{\theta_0,t}(u)} = \exp(-\theta_0^\top u).$$

This completes the proof. □

3.4.2 UMPU tests in higher dimensions

Suppose that we observe data X from a full-rank exponential family with density (3.9), where $\theta \in \mathbb{R}$ (i.e., $\ell = 1$) and $\eta \in \mathbb{R}^m$. We now derive tests by conditioning on T. By the above proposition, the condition distribution of U given $T = t$ is from a one-parameter exponential family. According to the theory that has been developed, a UMP conditional test for $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ is given by

$$\phi_1(X) := \begin{cases} 1 & \text{if } U > c(T), \\ \gamma(T) & \text{if } U = c(T), \\ 0 & \text{if } U < c(T), \end{cases}$$

where $c(\cdot)$ and $\gamma(\cdot)$ are determined by the constraint

$$\mathbb{E}_{\theta_0,\eta}[\phi_1 \mid T = t] = \alpha.$$

Moreover, for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, a UMPU conditional test is given by

$$\phi_2(X) := \begin{cases} 1 & \text{if } U \notin [c_1(T), c_2(T)]^c, \\ \gamma_1(T) & \text{if } U = c_1(T), \\ \gamma_2(T) & \text{if } U = c_2(T), \\ 0 & \text{if } U \in (c_1(T), c_2(T)), \end{cases}$$

where $c_i(\cdot)$ and $\gamma_i(\cdot)$, $i = 1, 2$, are determined by the constraints

$$\mathbb{E}_{\theta_0,\eta}[\phi_2 \mid T = t] = \alpha, \qquad \mathbb{E}_{\theta_0,\eta}[\phi_2\, U \mid T = t] = \alpha\, \mathbb{E}_{\theta_0,\eta}[U \mid T = t].$$

Theorem 3.17. *The tests ϕ_1 and ϕ_2 are UMPU at significance level α for their respective testing problems.*

Proof. We provide a sketch of the proof. Consider the test ϕ_1. Since $\mathbb{E}_{\theta_0,\eta}[\phi_1 \mid T] = \alpha$, we have $\mathbb{E}_{\theta_0,\eta}[\phi_1] = \alpha$. In addition, the conditional power function $\mathbb{E}_{\theta,\eta}[\phi_1 \mid T]$ is increasing in θ, so the power function $\mathbb{E}_{\theta,\eta}[\phi_1]$ is also increasing. It follows that ϕ_1 is subject to significance level α. Moreover, let ϕ be an unbiased test so that $\mathbb{E}_{\theta_0,\eta}[\phi] = \alpha$. Using the completeness of the exponential family with respect to T, one can show that in fact $\mathbb{E}_{\theta_0,\eta}[\phi \mid T] = \alpha$. Since ϕ_1 is UMPU conditional on T, it holds that $\mathbb{E}_{\theta,\eta}[\phi_1 \mid T] \ge \mathbb{E}_{\theta,\eta}[\phi \mid T]$ for any $\theta > \theta_0$. We then obtain $\mathbb{E}_{\theta,\eta}[\phi_1] \ge \mathbb{E}_{\theta,\eta}[\phi]$ for any $\theta > \theta_0$. Therefore, ϕ_1 is UMP among all unbiased tests, i.e., it is UMPU.

Next, consider the test ϕ_2. Recall that

$$1 = \int p_{\theta,\eta}(x)\, d\mu(x) = \int h(x)\, \exp\Big(\theta^\top U(x) + \eta^\top T(x) - A(\theta,\eta)\Big)\, d\mu(x),$$

so

$$e^{A(\theta,\eta)} = \int h(x)\, \exp\Big(\theta^\top U(x) + \eta^\top T(x)\Big)\, d\mu(x).$$

Differentiating both sides with respect to θ yields

$$e^{A(\theta,\eta)} \frac{\partial A(\theta,\eta)}{\partial \theta} = \int U(x) \cdot h(x)\, \exp\Big(\theta^\top U(x) + \eta^\top T(x)\Big)\, d\mu(x).$$

As a result, we obtain

$$m(\theta,\eta) := \frac{\partial A(\theta,\eta)}{\partial \theta} = \int U(x)\cdot p_{\theta,\eta}(x)\,d\mu(x) = \mathbb{E}_{\theta,\eta}[U].$$

Moreover,

$$\begin{aligned}\frac{\partial\,\mathbb{E}_{\theta,\eta}[\phi]}{\partial\theta} &= \int \phi(x)\frac{\partial}{\partial\theta}p_{\theta,\eta}(x)\,d\mu(x)\\ &= \int \phi(x)\big(U(x)-m(\theta,\eta)\big)p_{\theta,\eta}(x)\,d\mu(x)\\ &= \mathbb{E}_{\theta,\eta}\left[\phi\cdot\big(U-m(\theta,\eta)\big)\right].\end{aligned}$$

If ϕ is unbiased, then $\mathbb{E}_{\theta_0,\eta}[\phi]=\alpha$ and

$$0 = \mathbb{E}_{\theta_0,\eta}\left[\phi\cdot\big(U-m(\theta_0,\eta)\big)\right] = \mathbb{E}_{\theta_0,\eta}[\phi\,U] - \alpha\,m(\theta_0,\eta) = \mathbb{E}_{\theta_0,\eta}[(\phi-\alpha)U].$$

By completeness again, the above equation holds even if the expectations are taken conditionally on T, i.e.,

$$\mathbb{E}_{\theta_0,\eta}[\phi\mid T]=\alpha, \qquad \mathbb{E}_{\theta_0,\eta}[\phi\,U\mid T] = \alpha\,\mathbb{E}_{\theta_0,\eta}[U\mid T].$$

Under these constraints, we know that ϕ_2 is UMP conditional on T:

$$\mathbb{E}_{\theta,\eta}[\phi_2\mid T] \geq \mathbb{E}_{\theta,\eta}[\phi\mid T].$$

Therefore, we can conclude that $\mathbb{E}_{\theta,\eta}[\phi_2]\geq\mathbb{E}_{\theta,\eta}[\phi]$, i.e., ϕ_2 is UMPU. □

3.4.3 Application to the *t*-test

Let us apply the general theory to the t-test for Gaussian means. Suppose that $X_1,\ldots,X_n$ are i.i.d. $\mathcal{N}(\mu,\sigma^2)$ random variables where μ and σ^2 are unknown. Consider testing $H_0:\mu\leq 0$ against $H_1:\mu>0$. The joint density of $X=(X_1,\ldots,X_n)$ is

$$\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(\frac{-\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}\Big) = \frac{1}{(2\pi)^{n/2}}\exp\Big(\frac{\mu}{\sigma^2}U(x)-\frac{1}{2\sigma^2}T(x)-\frac{n\mu^2}{2\sigma^2}-n\log\sigma\Big)$$

where $U(x)=\sum_{i=1}^n x_i$ and $T(x)=\sum_{i=1}^n x_i^2$. This is of the form (3.9) with $\theta=\mu/\sigma^2$ and $\eta=-1/(2\sigma^2)$. By the above general theory, a UMPU test is

$$\phi(X) := \begin{cases}1 & \text{if } U\geq c(T),\\ 0 & \text{if } U<c(T),\end{cases}$$

where $c(\cdot)$ is chosen so that

$$\mathbb{P}_0\{U\geq c(T)\mid T\}=\alpha.$$

It is intuitive and can be shown rigorously that X conditional on $T=t$ is uniform over the sphere $\{x\in\mathbb{R}^n:\|x\|_2=\sqrt{t}\}$. Hence $X/\sqrt{T}$ is uniform over the unit sphere in $\mathbb{R}^n$. Let $\mathcal{Q}$ denote the distribution of $U/\sqrt{T}=\mathbf{1}^\top X/\sqrt{T}$, and let q_α denote the $(1-\alpha)$-quantile of $\mathcal{Q}$. Define $c(T):=q_\alpha\sqrt{T}$. Then the above constraint is satisfied.

This UMPU test is in fact equivalent to the t-test for Gaussian means:

$$\phi(X) := \begin{cases} 1 & \text{if } \sqrt{n}\bar{X}/S \geq t_{\alpha,n-1}, \\ 0 & \text{if } \sqrt{n}\bar{X}/S < t_{\alpha,n-1}, \end{cases}$$

where $t_{\alpha,n-1}$ is the $(1-\alpha)$-quantile of the t-distribution with $n-1$ degrees of freedom. To see the equivalence, note that the test statistic in the t-test is

$$\frac{\bar{X}}{S/\sqrt{n}} = \frac{U/n}{S/\sqrt{n}},$$

where

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

Thus we can rewrite the test statistic as

$$\frac{\bar{X}}{S/\sqrt{n}} = \frac{U/n}{\sqrt{(T-U^2/n)/(n-1)}} = \frac{\sqrt{n-1}\,\mathsf{sign}(U)}{\sqrt{nT/U^2 - 1}},$$

which is a strictly increasing function in $U/\sqrt{T}$. Thresholding $\frac{\bar{X}}{S/\sqrt{n}}$ is therefore equivalent to thresholding $U/\sqrt{T}$ provided that the two tests both have power α under $\mathbb{P}_0$. We conclude that the t-test is UMPU.

Chapter 4

Large-sample theory

4.1 Hellinger distance and testing errors

4.1.1 Definitions and properties

Let p and q be two densities with respect to a measure μ on the sample space $\mathcal{X}$, identified with their respective distributions. Define

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

Then $H(p,q) = \sqrt{H^2(p,q)}$ is called the Hellinger distance between p and q. Since $\int p = 1$, we can think of $\sqrt{p}$ as a function with unit norm in the Hilbert space $L_2(\mu)$. Then $H(p,q)$ is the natural L_2-distance between $\sqrt{p}$ and $\sqrt{q}$. Moreover, we define

$$A(p,q) := \int_{\mathcal{X}} \sqrt{pq} \, d\mu,$$

which is called the Hellinger affinity between p and q. In other words, $A(p,q)$ is the cosine of the angle between $\sqrt{p}$ and $\sqrt{q}$ in the space $L_2(\mu)$. The following properties hold:

- $H^2(p,q) = 2(1 - A(p,q))$;
- $0 \leq A(p,q) \leq 1$;
- $0 \leq H^2(p,q) \leq 2$;
- $H^2(p,q) = 0$ if and only if $A(p,q) = 1$ if and only if $p = q$ μ-almost everywhere;
- $H^2(p,q) = 2$ if and only if $A(p,q) = 0$ if and only if $pq = 0$ μ-almost everywhere.

Furthermore, define

$$\mathsf{TV}(p,q) := \frac{1}{2} \int_{\mathcal{X}} |p - q| \, d\mu,$$

which is called the total variation distance between p and q. Here we view a density p as a function with unit norm in $L_1(\mu)$. Up to a factor 2, the total variation distance is the natural L_1-distance between p and q.

Proposition 4.1. *We have*

$$\mathsf{TV}(p,q) \le H(p,q)\sqrt{1-\frac{H^2(p,q)}{4}}.$$

Proof. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
A(p,q)^2 = \Big(\int \sqrt{pq}\,d\mu\Big)^2 &= \Big(\int \sqrt{\min(p,q)\cdot\max(p,q)}\,d\mu\Big)^2 \\
&\le \Big(\int \min(p,q)\,d\mu\Big)\Big(\int \max(p,q)\,d\mu\Big) \\
&= \Big(\int \frac{p+q-|p-q|}{2}\,d\mu\Big)\Big(\int \frac{p+q+|p-q|}{2}\,d\mu\Big) \\
&= \Big(1-\frac{1}{2}\int |p-q|\,d\mu\Big)\Big(1+\frac{1}{2}\int |p-q|\,d\mu\Big) \\
&= 1-\Big(\frac{1}{2}\int |p-q|\,d\mu\Big)^2 = 1-\mathsf{TV}(p,q)^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathsf{TV}(p,q) \le \sqrt{1-A(p,q)^2} &= \sqrt{1-A(p,q)}\,\sqrt{1+A(p,q)} \\
&= \sqrt{\frac{H^2(p,q)}{2}}\sqrt{2-\frac{H^2(p,q)}{2}} = H(p,q)\sqrt{1-\frac{H^2(p,q)}{4}}.
\end{aligned}$$

□

4.1.2 Bounding errors in hypothesis testing

Consider a simple hypothesis testing problem between $H_0 : X \sim p_0$ and $H_1 : X \sim p_1$. The Hellinger distance between p_0 and p_1 is closely related to the errors in this testing problem. For a test ϕ, let $\alpha_\phi := \mathbb{E}_0[\phi]$ and $\beta_\phi := \mathbb{E}_1[\phi]$. Then $\max(\alpha_\phi, 1-\beta_\phi)$ is the maximum of the expected type I and type II errors. We now bound this quantity using the Hellinger distance.

Proposition 4.2. *Let $L(X) := p_1(X)/p_0(X)$ be the likelihood ratio. The likelihood-ratio test*

$$\phi_c(X) := \begin{cases} 1 & \text{if } L(X) \ge c, \\ 0 & \text{if } L(X) < c, \end{cases} \tag{4.1}$$

satisfies

$$\max(\alpha_{\phi_c}, 1-\beta_{\phi_c}) \le \max(\sqrt{c}, 1/\sqrt{c})\,A(p_0,p_1) = \max(\sqrt{c}, 1/\sqrt{c})\Big(1-\frac{H^2(p_0,p_1)}{2}\Big).$$

Proof. We have

$$\alpha_{\phi_c} = \mathbb{P}_0\{L(X) \ge c\} \le \mathbb{E}_0\sqrt{\frac{L(X)}{c}} \le \frac{1}{\sqrt{c}}\int \sqrt{\frac{p_1}{p_0}}\,p_0\,d\mu = \frac{1}{\sqrt{c}}\,A(p_0,p_1).$$

Similarly, one can show that

$$1-\beta_{\phi_c} \le \sqrt{c}\,A(p_0,p_1),$$

finishing the proof. □

On the other hand, we can prove a lower bound complementing the above upper bound.

Proposition 4.3. *Let Φ denote the set of all tests based on the observation X. We have*

$$\inf_{\phi\in\Phi} \max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2}\Big(1 - \mathsf{TV}(p_0,p_1)\Big) \geq \frac{1}{2}\left(1 - H(p_0,p_1)\sqrt{1-\frac{H^2(p_0,p_1)}{4}}\right).$$

Proof. For any test ϕ, we have

$$\max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2}(\alpha_\phi + 1 - \beta_\phi) = \frac{1}{2} + \frac{1}{2}(\alpha_\phi - \beta_\phi) = \frac{1}{2} + \frac{1}{2}\int \phi\cdot(p_0 - p_1)\,d\mu.$$

To minimize the right-hand side, we take the likelihood-ratio test ϕ_1 defined in (4.1), which gives

$$\max(\alpha_\phi, 1-\beta_\phi) \geq \frac{1}{2} - \frac{1}{2}\int_{p_1\geq p_0} (p_1 - p_0)\,d\mu.$$

Since $\int p_0\,d\mu = \int p_1\,d\mu = 1$, we have

$$\int_{p_1\geq p_0} (p_1 - p_0)\,d\mu = \int_{p_0>p_1} (p_0 - p_1)\,d\mu = \frac{1}{2}\int |p_0 - p_1|\,d\mu.$$

It follows that

$$2\max(\alpha_\phi, 1-\beta_\phi) \geq 1 - \frac{1}{2}\int |p_0 - p_1|\,d\mu = 1 - \mathsf{TV}(p_0,p_1).$$

Applying Proposition 4.1 then completes the proof. □

4.1.3 Tensorization and large-sample analysis

Let $\mu^{(n)}$ denote the product measure $\mu\times\cdots\times\mu$ on the space $\mathcal{X}^n = \mathcal{X}\times\cdots\times\mathcal{X}$. Let $p^{(n)}$ denote the product density on $\mathcal{X}^n$ defined by

$$p^{(n)}(x_1,\ldots,x_n) = p(x_1)\cdots p(x_n).$$

Proposition 4.4. *For any densities p and q and any integer $n\geq 1$, we have*

$$A(p^{(n)}, q^{(n)}) = A(p,q)^n, \qquad H^2(p^{(n)}, q^{(n)}) = 2\left(1 - \left(1 - \frac{H^2(p,q)}{2}\right)^n\right).$$

In particular, if $A(p,q) < 1$ or equivalently $H^2(p,q) > 0$, then we have $A(p^{(n)}, q^{(n)}) \to 0$ and $H^2(p^{(n)}, q^{(n)}) \to 2$ as $n\to\infty$.

Proof. We have

$$\begin{aligned} A(p^{(n)}, q^{(n)}) &= \int_{\mathcal{X}^n} \sqrt{p^{(n)}(x)\,q^{(n)}(x)}\,d\mu^{(n)}(x) \\ &= \int_{\mathcal{X}}\cdots\int_{\mathcal{X}} \sqrt{p(x_1)\,q(x_1)\cdots p(x_n)\,q(x_n)}\,d\mu(x_1)\cdots d\mu(x_n) = A(p,q)^n. \end{aligned}$$

Moreover,

$$H^2(p^{(n)}, q^{(n)}) = 2(1 - A(p^{(n)}, q^{(n)})) = 2(1 - A(p,q)^n) = 2\left(1 - \left(1 - \frac{H^2(p,q)}{2}\right)^n\right).$$

□

The above property is known as tensorization. We now discuss the consequence of tensorization in hypothesis testing.

Given two densities p_n and q_n with respect to μ on $\mathcal{X}$ (that are allowed to depend on n), consider testing $H_0^{(n)}$: i.i.d. $X_1, \dots, X_n \sim p_n$ against $H_1^{(n)}$: i.i.d. $X_1, \dots, X_n \sim q_n$. A sequence of tests $\phi^{(n)}$ is called consistent if

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

Proposition 4.5. *Let*

$$L_n(X_1, \dots, X_n) := \frac{q_n(X_1) \cdots q_n(X_n)}{p_n(X_1) \cdots p_n(X_n)}$$

be the likelihood ratio between the product densities. For any $c \in (0, \infty)$, define

$$\phi_c^{(n)}(X_1, \dots, X_n) := \begin{cases} 1 & \text{if } L_n(X_1, \dots, X_n) \geq c, \\ 0 & \text{if } L_n(X_1, \dots, X_n) < c. \end{cases} \tag{4.2}$$

Then we have

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

As a result, if $H(p_n, q_n) > b$ for a constant $b > 0$, then the sequence of tests $\phi_c^{(n)}$ is consistent.

Proof. This follows immediately from above results. □

Next, we turn to the case where

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

Theorem 4.6. *1. If $\sqrt{n}\,\delta_n \to \infty$ as $n \to \infty$, then there exists a sequence of test $\phi^{(n)}$ such that*

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

2. If $\sqrt{n}\,\delta_n \leq B$ for a constant $B > 0$, then there exists a constant $b > 0$ such that

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

where the infimum is taken over all test $\phi^{(n)}$ based on the observations $X_1, \dots, X_n$.

3. If $\sqrt{n}\,\delta_n \to 0$ as $n \to \infty$, then

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

Proof. For the first claim, consider again the test (4.2). Then we have

$$\max(\alpha_{\phi_c^{(n)}}, 1 - \beta_{\phi_c^{(n)}}) \leq \max(\sqrt{c}, 1/\sqrt{c})\, A(p_n, q_n)^n = \max(\sqrt{c}, 1/\sqrt{c}) \left(1 - \frac{\delta_n^2}{2}\right)^n,$$

since $A(p_n, q_n) = 1 - \frac{H^2(p_n, q_n)}{2}$. The above bound goes to 0 as $n \to \infty$ if $\sqrt{n}\,\delta_n \to \infty$.

For the other two claims, we apply Proposition 4.3 to obtain

$$\inf_{\phi^{(n)}} \max(\alpha_{\phi^{(n)}}, 1 - \beta_{\phi^{(n)}}) \geq \frac{1}{2}\left(1 - H(p_n^{(n)}, q_n^{(n)})\sqrt{1 - \frac{H^2(p_n^{(n)}, q_n^{(n)})}{4}}\right).$$

By tensorization, we have

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

If $\sqrt{n}\,\delta_n \leq B$, then $\left(1 - \frac{\delta_n^2}{2}\right)^n \geq \left(1 - \frac{B}{2n}\right)^n$, which is lower bounded by a positive constant for all large n. Therefore, $H^2(p_n^{(n)}, q_n^{(n)})$ is bounded away from 2 for all large n. This in turn yields the second claim.

Finally, if $\sqrt{n}\,\delta_n \to 0$ as $n \to \infty$, then $\left(1 - \frac{\delta_n^2}{2}\right)^n \to 1$, $H^2(p_n^{(n)}, q_n^{(n)}) \to 0$, and the third claim follows. □

4.2 Revisiting likelihood-ratio tests

4.2.1 Setup

Let $X_1, \ldots, X_n$ be i.i.d. observations from the distribution with density p_{θ^*} where $\theta^* \in \Theta$. The likelihood function is

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

The maximum likelihood estimator (MLE) is

$$\hat{\theta} := \operatorname*{argmax}_{\theta \in \Theta} L_n(\theta).$$

Consider the composite hypothesis testing problem between $H_0 : \theta^* \in \Theta_0$ and $H_1 : \theta^* \in \Theta_1$, where $\Theta_0 \subset \Theta$ and $\Theta_1 = \Theta \setminus \Theta_0$. In this case, we can define the likelihood ratio to be

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

Let the MLEs under H_0 and H_1 be

$$\hat{\theta}_0 := \operatorname*{argmax}_{\theta \in \Theta_0} L_n(\theta), \qquad \hat{\theta}_1 := \operatorname*{argmax}_{\theta \in \Theta_1} L_n(\theta),$$

respectively. Then the likelihood ratio is

$$\frac{L_n(\hat{\theta}_1)}{L_n(\hat{\theta}_0)} = \frac{p_{\hat{\theta}_1}(X_1) \cdots p_{\hat{\theta}_1}(X_n)}{p_{\hat{\theta}_0}(X_1) \cdots p_{\hat{\theta}_0}(X_n)}.$$

The log-likelihood ratio is

$$\tilde{\Lambda}_n := \log \frac{\sup_{\theta \in \Theta_1} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} = \log \frac{L_n(\hat{\theta}_1)}{L_n(\hat{\theta}_0)} = \sum_{i=1}^n \log \frac{p_{\hat{\theta}_1}(X_i)}{p_{\hat{\theta}_0}(X_i)}.$$

Moreover, it is often more convenient to study a modified version of the log-likelihood ratio

$$\Lambda_n := 2\log\frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)} = 2\log\frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)} = 2\sum_{i=1}^{n}\log\frac{p_{\hat{\theta}}(X_i)}{p_{\hat{\theta}_0}(X_i)}.$$

It is not hard to see that

$$\Lambda_n = 2\max\{\tilde{\Lambda}_n, 0\}.$$

Similar to the cases discussed before, we will reject the null if Λ_n exceeds a threshold c. More precisely, at significance level $\alpha \in (0,1)$, we would like to have

$$\mathbb{P}_{\theta^*}\{\Lambda_n \geq c\} \leq \alpha \quad \text{for all } \theta^* \in \Theta_0.$$

This is more difficult to achieve than before because we now have a composite hypothesis. However, we will show that under certain models, the asymptotic distribution of the log-likelihood can be characterized for $\theta^* \in \Theta_0$, allowing us to overcome the aforementioned difficulty.

4.2.2 Examples

Before introducing the general theory, let us discuss two simple examples.

Gaussian Let $X \sim \mathcal{N}(\theta^*, I_d)$ where $\theta^* \in \Theta \subset \mathbb{R}^d$. The likelihood is

$$L(\theta) = \frac{1}{(2\pi)^{d/2}}\exp\Big(\frac{-\|X-\theta\|_2^2}{2}\Big),$$

and the log-likelihood is

$$\log L(\theta) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\|X-\theta\|_2^2.$$

Therefore, the MLE is

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta}\|X-\theta\|_2^2 = \Pi_\Theta X,$$

where Π_Θ denotes the projection of X onto the set Θ. If Θ is closed and convex, then the projection is well-defined. Then we have

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

where $\mathsf{dist}(X,\Theta)$ denotes the ℓ_2-distance from X to the set Θ. It follows that

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

Consider the special case where $\Theta = \mathbb{R}^d$ and Θ_0 is a subspace of $\mathbb{R}^d$ of dimension $k < d$. In this case,

$$\Lambda = \mathsf{dist}^2(X,\Theta_0) = \|\Pi_{\Theta_0^\perp} X\|_2^2,$$

where $\Theta_0^\perp$ denotes the orthogonal complement of Θ_0. Under $H_0 : \theta \in \Theta_0$, we have

$$\Lambda = \|\Pi_{\Theta_0^\perp}(\theta^* + Z)\|_2^2 = \|\Pi_{\Theta_0^\perp} Z\|_2^2 \sim \chi^2_{d-k},$$

where $Z \sim \mathcal{N}(0, I_d)$. Therefore, the distribution of Λ does not depend on $\theta^* \in \Theta_0$. To obtain a test at significance level α, we can choose c to be the $(1-\alpha)$-quantile of χ^2_{d-k} and reject H_0 if $\Lambda \geq c$.

Multinomial Let $X = (X_1, \ldots, X_k)$ be from a multinomial distribution with parameters $n \geq 1$ and $\theta^* \in \Theta$, where the parameter space is the probability simplex

$$\Theta := \Big\{\theta = (\theta_1, \ldots, \theta_k) : \theta_i \geq 0 \text{ for } i \in [k], \sum_{i=1}^k \theta_i = 1\Big\}.$$

The likelihood is

$$L_n(\theta) = \frac{n!}{X_1! \cdots X_k!} \theta_1^{X_1} \cdots \theta_k^{X_k}.$$

The MLE is

$$\begin{aligned}
\hat{\theta} &= \operatorname*{argmax}_{\theta \in \Theta} \Big(\log \frac{n!}{X_1! \cdots X_k!} + X_1 \log \theta_1 + \cdots + X_k \log \theta_k \Big) \\
&= \operatorname*{argmin}_{\theta \in \Theta} \Big(- X_1 \log \theta_1 - \cdots - X_k \log \theta_k \Big) \\
&= \operatorname*{argmin}_{\theta \in \Theta} \Big(\frac{X_1}{n} \log \frac{X_1/n}{\theta_1} + \cdots + \frac{X_k}{n} \log \frac{X_k/n}{\theta_k} \Big).
\end{aligned}$$

For $\theta, \theta' \in \Theta$, the Kullback–Leibler (KL) divergence between θ' and θ is defined as

$$\mathsf{KL}(\theta', \theta) := \sum_{i=1}^k \theta_i' \log \frac{\theta_i'}{\theta_i}.$$

Then we have

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathsf{KL}(X/n, \theta),$$

where $X/n = (X_1/n, \ldots, X_k/n)$ is the vector of frequencies. Moreover, it is not hard to use Jensen's inequality to show that

$$\mathsf{KL}(\theta', \theta) \geq 0$$

with equality achieved if and only if $\theta = \theta'$. As a result, we see that $\hat{\theta} = X/n$.

For a subset $\Theta_0 \subset \Theta$, we have

$$\hat{\theta}_0 = \operatorname*{argmin}_{\theta \in \Theta_0} \mathsf{KL}(X/n, \theta) = \operatorname*{argmin}_{\theta \in \Theta_0} \mathsf{KL}(\hat{\theta}, \theta).$$

The modified log-likelihood ratio is

$$\Lambda_n = 2 \log \frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)} = 2 \sum_{i=1}^k X_i \log \frac{\hat{\theta}_i}{(\hat{\theta}_0)_i} = 2n \sum_{i=1}^k \hat{\theta}_i \log \frac{\hat{\theta}_i}{(\hat{\theta}_0)_i} = 2n \, \mathsf{KL}(\hat{\theta}, \hat{\theta}_0) = 2n \inf_{\theta \in \Theta_0} \mathsf{KL}(\hat{\theta}, \theta).$$

Consider the special case $\Theta_0 = \{\theta^{(0)}\}$ where $\theta^{(0)}$ has positive entries. Then

$$\Lambda_n = 2n \sum_{i=1}^k \hat{\theta}_i \log \frac{\hat{\theta}_i}{\theta_i^{(0)}} = 2n \sum_{i=1}^k \hat{\theta}_i \log \left(1 + \frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}} \right).$$

The Taylor expansion gives

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

where $r(x) \to 0$ as $x \to 0$. Moreover, we have

$$\hat{\theta}_i - \theta_i^{(0)} = O_p(1/\sqrt{n})$$

where O_p means having an order in probability as $n \to \infty$. Combining these facts gives

$$\log\left(1 + \frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}}\right) = \frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}} - \frac{1}{2}\left(\frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}}\right)^2 + o_p(1/n)$$

as $n \to \infty$. Therefore,

$$\begin{aligned}
\Lambda_n &= 2n \sum_{i=1}^k \hat{\theta}_i \frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}} - n \sum_{i=1}^k \hat{\theta}_i \left(\frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}}\right)^2 + o_p(1) \\
&= 2n \sum_{i=1}^k \theta_i^{(0)} \frac{\hat{\theta}_i - \theta_i^{(0)}}{\theta_i^{(0)}} + 2n \sum_{i=1}^k \frac{(\hat{\theta}_i - \theta_i^{(0)})^2}{\theta_i^{(0)}} - n \sum_{i=1}^k \frac{(\hat{\theta}_i - \theta_i^{(0)})^2}{\theta_i^{(0)}} - n \sum_{i=1}^k \frac{(\hat{\theta}_i - \theta_i^{(0)})^3}{(\theta_i^{(0)})^2} + o_p(1) \\
&= n \sum_{i=1}^k \frac{(\hat{\theta}_i - \theta_i^{(0)})^2}{\theta_i^{(0)}} + o_p(1).
\end{aligned}$$

The main term is in fact the same statistic used for the goodness-of-fit in Section 2.4.1. We can show that it converges to χ^2_{k-1}. Therefore, the likelihood-ratio test is asymptotically equivalent to Pearson's chi-squared test in this case.

4.3 Asymptotic theory for likelihood-ratio tests

Let us first introduce the Kullback–Leibler (KL) divergence more generally. The KL divergence between probability distributions with densities p and q is defined as

$$\mathsf{KL}(p, q) := \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

Proposition 4.7. *For any densities p and q, we have*

$$\mathsf{KL}(p, q) \geq H^2(p, q).$$

Proof. Since $\log x \leq 2(\sqrt{x} - 1)$ for $x \geq 0$, we obtain

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

Therefore,

$$-\mathsf{KL}(p, q) = \mathbb{E}_p \log \frac{q(X)}{p(X)} \leq 2\Big(\mathbb{E}_p \sqrt{\frac{q(X)}{p(X)}} - 1\Big) = 2\Big(\int \sqrt{pq}\, d\mu - 1\Big) = 2(A(p, q) - 1) = -H^2(p, q).$$

□

4.3.1 Consistency of the MLE

We continue to use the same notation as in the previous section. Recall that the MLE is

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta\in\Theta} L_n(\theta) = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i).$$

Suppose that $\mathbb{E}_{\theta^*}|\log p_\theta(X)| < \infty$ for all $\theta, \theta^* \in \Theta$. Then by the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \to \mathbb{E}_{\theta^*} \log p_\theta(X)$$

almost surely as $n \to \infty$. Moreover, we would like

$$\theta^* = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{\theta^*} \log p_\theta(X). \tag{4.3}$$

To guarantee this, note that

$$\mathbb{E}_{\theta^*} \log p_{\theta^*}(X) - \mathbb{E}_{\theta^*} \log p_\theta(X) = \mathsf{KL}(p_{\theta^*}, p_\theta) \geq H^2(p_{\theta^*}, p_\theta).$$

Therefore, if the model $\{p_\theta : \theta \in \Theta\}$ is identifiable, i.e., $H^2(p_\theta, p_{\theta'}) > 0$ for any distinct $\theta, \theta' \in \Theta$, then θ^* is the unique maximizer in (4.3). Furthermore, under some regularity assumptions of the model, we can show that

$$\hat{\theta}_n \to \theta^*$$

in probability or almost surely as $n \to \infty$, in which case we say that the MLE $\hat{\theta}_n$ is consistent.

4.3.2 Asymptotic normality of the MLE

We now briefly discuss a more refined property of the MLE $\hat{\theta}_n$ called asymptotic normality. Suppose that Θ is an open subset of $\mathbb{R}^d$. Suppose that the model $\{p_\theta : \theta \in \Theta\}$ is quadratic mean differentiable (QMD) at $\theta \in \Theta$, i.e., there is a function $\psi_\theta : \mathcal{X} \to \mathbb{R}^d$ such that

$$\left\| \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta}\right) - \langle \psi_\theta, h\rangle \right\|_{L_2} = o(\|h\|_2)$$

as $\|h\|_2 \to 0$. Informally, for $d = 1$, this is saying that

$$\lim_{h\to 0} \frac{\sqrt{p_{\theta+h}} - \sqrt{p_\theta}}{h} = \psi_\theta, \qquad \frac{\partial}{\partial\theta}\sqrt{p_\theta} = \psi_\theta,$$

and furthermore,

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

The Fisher information matrix is defined as

$$I(\theta) := \mathbb{E}_\theta\left[\Big(\frac{\partial}{\partial\theta}\log p_\theta\Big)\Big(\frac{\partial}{\partial\theta}\log p_\theta\Big)^\top\right] = 4\int_{\mathcal{X}} \psi_\theta\psi_\theta^\top\, d\mu.$$

Define the log-likelihood ratio process as

$$Z_n(\theta; u) := \log \frac{\prod_{i=1}^n p_{\theta+\frac{u}{\sqrt{n}}}(X_i)}{\prod_{i=1}^n p_\theta(X_i)} = \sum_{i=1}^n \Big(\log p_{\theta+\frac{u}{\sqrt{n}}}(X_i) - \log p_\theta(X_i) \Big),$$

where $u \in \mathbb{R}^d$ and $\theta, \theta + \frac{u}{\sqrt{n}} \in \Theta$. Then we have

$$\hat{u}_n := \operatorname*{argmax}_{u \in \mathbb{R}^d} Z_n(\theta^*; u) = \sqrt{n}(\hat{\theta}_n - \theta^*).$$

The following result is due to Le Cam (see Chapter 7 of [vdV00]).

Theorem 4.8 (Local asymptotic normality)**.** *In the above setting, for any $u \in \mathbb{R}^d$, we have*

$$Z_n(\theta^*; u) = u^\top Y_n(\theta^*) - \frac{1}{2} u^\top I(\theta^*)\, u + o_p(1)$$

as $n \to \infty$, where

$$Y_n(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log p_\theta(X_i).$$

Moreover, as a result the central limit theorem,

$$Y_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta^*))$$

as $n \to \infty$, and for any $u \in \mathbb{R}^d$,

$$Z_n(\theta^*; u) \xrightarrow{d} u^\top I(\theta^*)^{1/2} Z - \frac{1}{2} u^\top I(\theta^*)\, u$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0, I_d)$.

The following heuristic derivation then yields the asymptotic normality of the MLE $\hat{\theta}_n$:

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta^*) = \hat{u}_n &= \operatorname*{argmax}_{u \in \mathbb{R}^d} Z_n(\theta^*; u) \\
&\overset{d}{\approx} \operatorname*{argmax}_{u \in \mathbb{R}^d} \Big(u^\top I(\theta^*)^{1/2} Z - \frac{1}{2} u^\top I(\theta^*)\, u \Big) \\
&= I(\theta^*)^{-1/2} \operatorname*{argmax}_{v \in \mathbb{R}^d} \Big(v^\top Z - \frac{1}{2} \|v\|_2^2 \Big) \\
&= I(\theta^*)^{-1/2} Z \sim \mathcal{N}(0, I(\theta^*)^{-1}),
\end{aligned}$$

provided that $I(\theta^*)$ is invertible, so we obtain

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

4.3.3 Wilks' theorem

Let Θ be an open subset of $\mathbb{R}^d$ and L be a subspace of $\mathbb{R}^d$ of dimension $k < d$. Consider testing $H_0 : \theta^* \in \Theta \cap L$ against $H_1 : \theta^* \in \Theta \setminus L$. Let $\hat{\theta}_n$ denote the MLE for the whole model and let $\hat{\theta}_{n,0}$ denote the MLE under H_0.

Theorem 4.9. *In the above setting, under some additional regularity assumptions, we have that for any $\theta^* \in \Theta \cap L$,*

$$\Lambda_n := 2 \log \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_{n,0})} \xrightarrow{d} \chi^2_{d-k} \qquad \text{as } n \to \infty.$$

Proof. We provide a sketch of the proof. Let

$$\hat{u}_n := \operatorname*{argmax}_{u \in \mathbb{R}^d} Z_n(\theta^*; u) = \sqrt{n}(\hat{\theta}_n - \theta^*), \qquad \hat{u}_{n,0} := \operatorname*{argmax}_{u \in L} Z_n(\theta^*; u) = \sqrt{n}(\hat{\theta}_{n,0} - \theta^*).$$

Then

$$\Lambda_n = 2 \log \frac{L_n(\hat{\theta}_n)}{L_n(\theta^*)} - 2 \log \frac{L_n(\hat{\theta}_{n,0})}{L_n(\theta^*)} = 2\big(Z_n(\theta^*; \hat{u}_n) - Z_n(\theta^*; \hat{u}_{n,0})\big).$$

By local asymptotic normality, we have

$$Z_n(\theta^*; u) = Q_n(\theta^*; u) + o_p(1), \qquad \text{where } Q_n(\theta; u) := u^\top Y_n(\theta) - \frac{1}{2} u^\top I(\theta)\, u.$$

It follows that

$$\Lambda_n = 2 \sup_{u \in \mathbb{R}^d} Z_n(\theta^*; u) - 2 \sup_{u \in L} Z_n(\theta^*; u) = 2 \sup_{u \in \mathbb{R}^d} Q_n(\theta^*; u) - 2 \sup_{u \in L} Q_n(\theta^*; u) + o_p(1).$$

Furthermore, note that

$$2\, Q_n(\theta; u) = \|I(\theta)^{-1/2} Y_n(\theta)\|_2^2 - \|I(\theta)^{-1/2} Y_n(\theta) - I(\theta)^{1/2} u\|_2^2.$$

As a consequence,

$$2 \sup_{u \in \mathbb{R}^d} Q_n(\theta^*; u) = \|I(\theta^*)^{-1/2} Y_n(\theta^*)\|_2^2$$

and

$$2 \sup_{u \in L} Q_n(\theta^*; u) = \|I(\theta^*)^{-1/2} Y_n(\theta^*)\|_2^2 - \inf_{u \in L} \|I(\theta^*)^{-1/2} Y_n(\theta^*) - I(\theta^*)^{1/2} u\|_2^2.$$

We therefore obtain

$$\Lambda_n = \inf_{u \in L} \|I(\theta^*)^{-1/2} Y_n(\theta^*) - I(\theta^*)^{1/2} u\|_2^2 + o_p(1) = \mathsf{dist}^2\left(I(\theta^*)^{-1/2} Y_n(\theta^*), \tilde{L}\right) + o_p(1),$$

where $\tilde{L}$ denotes the k-dimensional subspace $I(\theta^*)^{1/2} L = \{I(\theta^*)^{1/2} x : x \in L\}$.

Local asymptotic normality also gives that $I(\theta^*)^{-1/2} Y_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, I_d)$. We conclude that

$$\Lambda_n \xrightarrow{d} \mathsf{dist}^2(Z, \tilde{L}) = \|\Pi_{\tilde{L}^\perp} Z\|_2^2 \sim \chi^2_{d-k},$$

where $Z \sim \mathcal{N}(0, I_d)$. □

4.4 Bahadur's efficiency and Stein's regime

4.4.1 Efficiency of likelihood-ratio tests

Given i.i.d. $X_1, \dots, X_n \sim p_\theta$, consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ where Θ_0 and Θ_1 are disjoint finite sets. Suppose that we use a test statistic $T_n = T_n(X_1, \dots, X_n)$ and reject H_0 if T_n exceeds some threshold $t \in \mathbb{R}$. Define a function

$$\alpha_n(t) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\{T_n \geq t\}$$

which is assumed to be continuous and decreasing. Define $t_n(\alpha)$ to be threshold such that

$$\alpha_n(t_n(\alpha)) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\{T_n \geq t_n(\alpha)\} = \alpha.$$

Lemma 4.10. *The random variable $\alpha_n(T_n)$ is the p-value.*

Proof. Recall that in general, the p-value is defined as $\inf\{\alpha : X \in S_1(\alpha)\}$, where $S_1(\alpha)$ is the region of rejection of the test at significance level α. Then the p-value for T_n is $\inf\{\alpha : T_n \geq t_n(\alpha)\}$. Note that $T_n \geq t_n(\alpha)$ if and only if $\sup_{\theta \in \Theta_0} \mathbb{P}'_\theta\{T'_n \geq T_n\} \leq \alpha$, where T'_n is an i.i.d. copy of T_n and the probability $\mathbb{P}'_\theta$ is with respect to T'_n. Hence we can write the p-value as

$$\inf\Big\{\alpha : \sup_{\theta \in \Theta_0} \mathbb{P}'_\theta\{T'_n \geq T_n\} \leq \alpha\Big\} = \sup_{\theta \in \Theta_0} \mathbb{P}'_\theta\{T'_n \geq T_n\} = \alpha_n(T_n).$$

□

Since $\alpha_n(\cdot)$ is decreasing, for $\theta \in \Theta_0$, we have

$$\mathbb{P}_\theta\{\alpha_n(T_n) \leq \alpha\} = \mathbb{P}_\theta\{T_n \geq t_n(\alpha)\} \leq \alpha.$$

In particular, the test that rejects H_0 when $\alpha_n(T_n) \leq \alpha$ has significance level α. Next, we would like this test to be as powerful as possible. In other words, for $\theta \in \Theta_1$, we would like $\mathbb{P}_\theta\{\alpha_n(T_n) \leq \alpha\}$ to be as small as possible.

The quality of a test based on T_n can be characterized by the rate of decay of $\alpha_n(T_n)$ as $n \to \infty$ at $\theta \in \Theta_1$. More precisely, given i.i.d. observations $X_1, \dots, X_n \sim \mathcal{P}_\theta$ where $\theta \in \Theta_1$, we define

$$B_\theta(T) := -\lim_{n \to \infty} \frac{1}{n} \log \alpha_n(T_n)$$

if the limit exists in probability. This quantity is called the Bahadur slope of the statistic T_n. The larger the slope, the better the test. The test that has the largest slope is called Bahadur efficient.

Recall that the log-likelihood ratio is

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

and the KL divergence between densities p and q is

$$\mathsf{KL}(p, q) = \mathbb{E}_p\Big[\log \frac{p(X)}{q(X)}\Big].$$

The following theorem characterizes the optimal Bahadur slope using the KL divergence and shows that the likelihood-ratio test is Bahadur efficient.

Theorem 4.11. *In the above setting, let $T_n = T_n(X_1, \dots, X_n)$ be a statistic. Then for any $\theta \in \Theta_1$,*

$$B_\theta(T) \le \min_{\theta' \in \Theta_0} \mathsf{KL}(p_\theta, p_{\theta'})$$

almost surely, with the equality achieved for $T_n = \tilde{\Lambda}_n$.

Proof. We sketch the proof for the case $\Theta_0 = \{\theta'\}$. For $\theta \in \Theta_1$, the log-likelihood ratio is

$$\Lambda_n'(\theta) = \log \frac{p_\theta(X_1) \cdots p_\theta(X_n)}{p_{\theta'}(X_1) \cdots p_{\theta'}(X_n)}.$$

By the law of large numbers,

$$\frac{1}{n}\Lambda_n'(\theta) = \frac{1}{n}\sum_{i=1}^n \log \frac{p_\theta(X_i)}{p_{\theta'}(X_i)} \to \mathsf{KL}(p_\theta, p_{\theta'}).$$

Take constants $B > A > \mathsf{KL}(p_\theta, p_{\theta'})$. We have

$$\begin{aligned}\mathbb{P}_\theta\{\alpha_n(T_n) \le e^{-nB},\, \Lambda_n'(\theta) \le nA\} &= \mathbb{E}_{\theta'}\left[\mathbb{1}\{\alpha_n(T_n) \le e^{-nB},\, \Lambda_n'(\theta) \le nA\} \cdot e^{\Lambda_n'(\theta)}\right] \\ &\le e^{nA}\, \mathbb{P}_{\theta'}\{\alpha_n(T_n) \le e^{-nB}\} = e^{n(A-B)} \to 0.\end{aligned}$$

Consequently, as $n \to \infty$, we have $\alpha_n(T_n) \ge e^{-nB}$ so that

$$-\frac{1}{n}\log \alpha_n(T_n) \le B.$$

This holds for any $B > \mathsf{KL}(p_\theta, p_{\theta'})$, so the desired inequality follows.

Moreover, for $T_n = \tilde{\Lambda}_n$, we have

$$\alpha_n(t) = \mathbb{P}_{\theta'}\{\tilde{\Lambda}_n \ge t\}.$$

By a union bound and Markov's inequality,

$$\alpha_n(t) \le \sum_{\theta \in \Theta_1} \mathbb{P}_{\theta'}\{\Lambda_n'(\theta) \ge t\} \le \sum_{\theta \in \Theta_1} \mathbb{P}_{\theta'}\{e^{\Lambda_n'(\theta)} \ge e^t\} \le \sum_{\theta \in \Theta_1} e^{-t}\, \mathbb{E}_{\theta'}[e^{\Lambda_n'(\theta)}] = |\Theta_1| \cdot e^{-t}.$$

As a result, $\alpha_n(\tilde{\Lambda}_n) \le |\Theta_1| \cdot e^{-\tilde{\Lambda}_n}$, and it follows that

$$-\frac{1}{n}\log \alpha_n(\tilde{\Lambda}_n) \ge -\frac{1}{n}\log|\Theta_1| + \frac{1}{n}\tilde{\Lambda}_n \ge -\frac{1}{n}\log|\Theta_1| + \frac{1}{n}\Lambda_n'(\theta) \to \mathsf{KL}(p_\theta, p_{\theta'}).$$

□

4.4.2 Chernoff–Stein lemma

The above result is sometimes formulated in a different way. Suppose that $X_1, \dots, X_n$ are i.i.d. observations from p under H_0 and from q under H_1. For a test $\phi^{(n)}$, let $\alpha_{\phi^{(n)}} = \mathbb{E}_p[\phi^{(n)}]$ and $\beta_{\phi^{(n)}} = \mathbb{E}_q[\phi^{(n)}]$ as before. Then we define

$$V_\alpha := -\lim_{n\to\infty} \inf_{\phi^{(n)}:\alpha_{\phi^{(n)}} \le \alpha} \frac{1}{n}\log\left(1 - \beta_{\phi^{(n)}}\right)$$

provided that the limit exists. The limit

$$V := \lim_{\alpha \to 0} V_\alpha$$

is called Stein's exponent. Note that the quantity V_α is very similar to the Bahadur slope, but they are different. In particular, V_α is defined as the limit of a sequence of numbers (exponents of expected type II errors), while the Bahadur slope is defined as the limit of a sequence of random variables (exponents of p-values). Nevertheless, they are used to describe essentially the same phenomenon. The following theorem, called the Chernoff–Stein lemma, is similar to the above theorem about the Bahadur efficiency.

Theorem 4.12. *In the above setting, we have*

$$V_\alpha = \mathsf{KL}(p, q)$$

for all $\alpha \in (0, 1)$. *Consequently,*

$$V = \mathsf{KL}(p, q).$$

Proof. The proof is similar to that of the above result about the Bahadur efficiency. □

If n is large, the Chernoff–Stein lemma implies that

$$\mathsf{KL}(p, q) = V_\alpha \approx -\frac{1}{n} \log \big(1 - \beta_{\phi^{(n)}}\big).$$

As a result, if we aim for $1 - \beta \leq e^{-k}$ for example, then we should take the sample size n to be approximately $\frac{k}{\mathsf{KL}(p,q)}$.

4.5 Chernoff's regime and large deviation

In the last section, we fix the significance level α and study the rate of decay of the p-value or the type II error. Recall that the decay is of the form e^{-nE} where E is an exponent (the Bahadur slope or the Stein exponent, which is equal to a KL divergence). We now turn to studying the decay of expected type I and type II errors simultaneously: Suppose that $\alpha \to 0$ at a rate e^{-nE_0} and $1 - \beta \to 0$ at a rate e^{-nE_1}; we aim to find E_0 and E_1. There is obviously a trade-off between E_0 and E_1 because we cannot make them large simultaneously. The previous section corresponds to the case $E_0 = 0$.

Before studying rates of decay of testing errors, we first analyze the tail probability of a sum of independent random variables in this section.

4.5.1 Chernoff bound

For i.i.d. $X_1, \ldots, X_n$, large deviation theory focuses on obtaining inequalities of the form

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \gamma\right\} = e^{-nE(\gamma)+o(n)},$$

where $E(\gamma)$ is the rate function defined by

$$E(\gamma) = -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \gamma\right\}.$$

The usual Chernoff's bound does the following: for any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \gamma\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^{n} X_i\right) \geq \exp(n\lambda\gamma)\right\} \\ &\leq \exp(-n\lambda\gamma)\ \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] \\ &= \exp\Big(-n\lambda\gamma + n\log \mathbb{E}[\exp(\lambda X_1)]\Big).\end{aligned}$$

Minimizing this over $\lambda \geq 0$ yields an upper bound. (We also need a matching lower bound, which will be discussed in the next section.)

4.5.2 Cumulant generating function

The key quantity in the above upper bound is

$$\psi_X(\lambda) := \log \mathbb{E}[\exp(\lambda X)]$$

which is called the log moment generating function (log-MGF) or the cumulant generating function (CGF) of a random variable X. We now state some facts about the CGF without proofs.

Theorem 4.13. *Let X be a non-constant random variable. Suppose that the CGF ψ_X of X exists. Then it has the following properties:*

1. *The CGF ψ_X is convex and continuous.*
2. *The CGF ψ_X is strictly convex and thus ψ'_X is strictly increasing.*
3. *The CGF ψ_X is infinitely differentiable and*

$$\psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{E[e^{\lambda X}]} = e^{-\psi_X(\lambda)}\,\mathbb{E}[Xe^{\lambda X}].$$

   In particular, $\psi_X(0) = 0$ and $\psi'_X(0) = \mathbb{E}[X]$.

4. *If $a \leq X \leq b$ almost surely, then $a \leq \psi'_X \leq b$.*
5. *Conversely, if $a \leq \psi'_X \leq b$, then $a \leq X \leq b$ almost surely. Therefore, the essential support of the distribution of X equals the closure of the range of ψ'_X.*
6. *Given n i.i.d. copies of X, let $\bar{X}$ be the sample mean. The Chernoff bound holds:*

$$\mathbb{P}\{\bar{X} \geq \gamma\} \leq \exp(-n(\lambda\gamma - \psi_X(\lambda))) \quad \text{for any } \lambda \geq 0.$$

Next, we define the rate function $\psi^*_X : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ as the Legendre–Fenchel transform (i.e., convex conjugate) of the CGF ψ_X:

$$\psi^*_X(\gamma) = \sup_{\lambda \in \mathbb{R}} \Big(\lambda\gamma - \psi_X(\lambda)\Big).$$

Theorem 4.14. *The rate function ψ^*_X has the following properties:*

1. *Let*
$$a = \inf_{\lambda\in\mathbb{R}} \psi'_X(\lambda) = \operatorname{essinf}(X), \qquad b = \sup_{\lambda\in\mathbb{R}} \psi'_X(\lambda) = \operatorname{esssup}(X).$$
Then
$$\psi^*_X(\gamma) = \begin{cases} \lambda\gamma - \psi_X(\lambda) \text{ for } \lambda \text{ s.t. } \gamma = \psi'_X(\lambda), & \text{if } \gamma \in (a,b), \\ -\log \mathbb{P}\{X = \gamma\}, & \text{if } \gamma = a \text{ or } b, \\ \infty, & \text{if } \gamma \notin [a,b]. \end{cases}$$

2. *The rate function $\psi^*_X(\gamma)$ is strictly convex and strictly positive except that $\psi^*_X(\mathbb{E}[X]) = 0$.*

3. *The rate function $\psi^*_X(\gamma)$ is decreasing for $\gamma \in (a, \mathbb{E}[X])$ and increasing for $\gamma \in (\mathbb{E}[X], b)$.*

4. *Given n i.i.d. copies of X, let $\bar{X}$ be the sample mean. The Chernoff bound implies that, for $\gamma \geq \mathbb{E}[X]$, we have*
$$\mathbb{P}\{\bar{X} \geq \gamma\} \leq \exp(-n\,\psi^*_X(\gamma)).$$

We skip the proofs for all but the last statement. For the last statement, by the Chernoff bound, it suffices to show that

$$\psi^*_X(\gamma) = \sup_{\lambda\in\mathbb{R}} \Big(\lambda\gamma - \psi_X(\lambda)\Big) = \sup_{\lambda\geq 0} \Big(\lambda\gamma - \psi_X(\lambda)\Big).$$

To this end, note that the derivative of the objective function with respect to λ is $\gamma - \psi'_X(\lambda)$. Recall that $\psi'_X(\lambda)$ is increasing and $\psi'_X(0) = \mathbb{E}[X] \leq \gamma$. Thus the objective function in the above optimization problem is increasing for $\lambda \leq 0$. The result follows.

4.5.3 Tilted distribution

As a preparation for the following sections, let us formally introduce tilted distributions which have already appeared above. For a random variable $X \sim \mathcal{P}$ and a constant $\lambda \in \mathbb{R}$, we define the tilted distribution $\mathcal{P}_\lambda$ by

$$d\mathcal{P}_\lambda(x) = \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]}\, d\mathcal{P}(x) = e^{\lambda x - \psi_X(\lambda)} d\mathcal{P}(x).$$

In other words, if $\mathcal{P}$ has density p, then the density of $\mathcal{P}_\lambda$ is given by $p_\lambda(x) = e^{\lambda x - \psi_X(\lambda)} p(x)$. Tilting is also called exponential tilting, Esscher tilting, or the Esscher transform. In addition, note that $\{p_\lambda : \lambda \in \mathbb{R}\}$ is an exponential family.

Theorem 4.15. *For $X \sim \mathcal{P}$, the tilted distribution $\mathcal{P}_\lambda$ has the following properties:*

1. *The CGF of $\mathcal{P}_\lambda$ is*
$$\psi_\lambda(u) = \psi_X(\lambda + u) - \psi_X(\lambda).$$

2. *Tilting trades mean for divergence in the following sense:*
$$\begin{aligned} &\mathbb{E}_\lambda[X] = \psi'_X(\lambda) < \mathbb{E}[X] \quad \text{for } \lambda < 0, \\ &\mathbb{E}_\lambda[X] = \psi'_X(\lambda) > \mathbb{E}[X] \quad \text{for } \lambda > 0, \\ &\mathsf{KL}(\mathcal{P}_\lambda, \mathcal{P}) = \psi^*_X(\psi'_X(\lambda)) = \psi^*_X(\mathbb{E}_\lambda[X]). \end{aligned}$$

3. If $\mathbb{P}\{X < s\} > 0$ *and* $\mathbb{P}\{X > t\} > 0$*, then for any* $\varepsilon > 0$*,*

$$\mathbb{P}_\lambda\{X \geq s + \varepsilon\} \to 0 \quad \text{as } \lambda \to -\infty,$$
$$\mathbb{P}_\lambda\{X \leq t - \varepsilon\} \to 0 \quad \text{as } \lambda \to \infty.$$

4. If $X_\lambda \sim \mathcal{P}_\lambda$*, then*

$$X_\lambda \xrightarrow{d} \operatorname{essinf}(X) = a \quad \text{as } \lambda \to -\infty,$$
$$X_\lambda \xrightarrow{d} \operatorname{essinf}(X) = b \quad \text{as } \lambda \to \infty.$$

4.6 Information projection and large deviation exponent

In the previous section, we used the Chernoff bound to obtain an upper bound on the large deviation probability. In particular, the exponent was expressed in terms of the Legendre–Fenchel transform of the CGF. We now discuss a different method which gives a formula for the exponent in terms of the information projection.

Theorem 4.16. *For i.i.d.* $X_1, \dots, X_n \sim \mathcal{P}$*, let* $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$*. Then for any* $\gamma \in \mathbb{R}$*, we have*

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\bar{X} \geq \gamma\} = \min_{\mathcal{Q}:\mathbb{E}_\mathcal{Q}[X]\geq\gamma} \mathsf{KL}(\mathcal{Q}, \mathcal{P}),$$

and the same conclusion holds with both $\geq$ *replaced by* $>$*.*

This theorem is in the same spirit as the Bahadur efficiency and the Chernoff–Stein lemma, and so is its proof. The optimization problem $\min_{\mathcal{Q}\in\mathcal{E}} \mathsf{KL}(\mathcal{Q}, \mathcal{P})$ is called the information projection, where $\mathcal{E}$ denotes a convex set of distributions.

Theorem 4.17. *Suppose that there exists* $\mathcal{Q}^* \in \mathcal{E}$ *such that* $\mathsf{KL}(\mathcal{Q}^*, \mathcal{P}) = \min_{\mathcal{Q}\in\mathcal{E}} \mathsf{KL}(\mathcal{Q}, \mathcal{P})$*. Then for any* $\mathcal{Q} \in \mathcal{E}$*, we have*

$$\mathsf{KL}(\mathcal{Q}, \mathcal{P}) \geq \mathsf{KL}(\mathcal{Q}, \mathcal{Q}^*) + \mathsf{KL}(\mathcal{Q}^*, \mathcal{P}).$$

Proof. We may assume without loss of generality that $\mathsf{KL}(\mathcal{Q}, \mathcal{P}) < \infty$, which also implies that $\mathsf{KL}(\mathcal{Q}^*, \mathcal{P}) < \infty$. For $\lambda \in [0, 1]$, let $\mathcal{Q}_\lambda := (1-\lambda)\mathcal{Q}^* + \lambda\mathcal{Q}$. Assuming densities exist for simplicity, we have

$$\mathsf{KL}(\mathcal{Q}_\lambda, \mathcal{P}) = \mathbb{E}_p\Big[\frac{q_\lambda}{p} \log \frac{q_\lambda}{p}\Big] = \mathbb{E}_p\Big[\frac{(1-\lambda)q^* + \lambda q}{p} \log \frac{(1-\lambda)q^* + \lambda q}{p}\Big],$$

so

$$\frac{\partial}{\partial\lambda} \mathsf{KL}(\mathcal{Q}_\lambda, \mathcal{P}) = \mathbb{E}_p\Big[\Big(\frac{1}{p} \log \frac{(1-\lambda)q^* + \lambda q}{p} + \frac{1}{p}\Big)(q - q^*)\Big].$$

Since Q^* is the minimizer, the derivative at $\lambda = 0$ is nonnegative. Therefore,

$$\begin{aligned}
0 &\leq \mathbb{E}_p\Big[\Big(\frac{1}{p} \log \frac{q^*}{p} + \frac{1}{p}\Big)(q - q^*)\Big] \\
&= \mathbb{E}_q\Big[\log \frac{q^*}{p}\Big] - \mathbb{E}_{q^*}\Big[\log \frac{q^*}{p}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{q}{p}\Big] - \mathbb{E}_q\Big[\log \frac{q}{q^*}\Big] - \mathbb{E}_{q^*}\Big[\log \frac{q^*}{p}\Big] \\
&= \mathsf{KL}(\mathcal{Q}, \mathcal{P}) - \mathsf{KL}(\mathcal{Q}, \mathcal{Q}^*) - \mathsf{KL}(\mathcal{Q}^*, \mathcal{P}).
\end{aligned}$$

The conclusion follows. □

To determine the large deviation exponent more explicitly, we need to solve the information projection $\min_{\mathcal{Q}\in\mathcal{E}} \mathsf{KL}(\mathcal{Q},\mathcal{P})$ where $\mathcal{E} = \{\mathcal{Q} : \mathbb{E}_{\mathcal{Q}}[X] \geq \gamma\}$.

Theorem 4.18. *Given $X \sim \mathcal{P}$, let $b = \sup_{\lambda\in\mathbb{R}} \psi'_X(\lambda) = \operatorname{esssup}(X)$. The information projection over $\{\mathcal{Q} : \mathbb{E}_{\mathcal{Q}}[X] \geq \gamma\}$ satisfies*

$$\min_{\mathcal{Q}:\mathbb{E}_{\mathcal{Q}}[X]\geq\gamma} \mathsf{KL}(\mathcal{Q},\mathcal{P}) = \begin{cases} 0 & \text{if } \gamma < \mathbb{E}[X], \\ \psi_X^*(\gamma) & \text{if } \mathbb{E}[X] \leq \gamma < b, \\ -\log \mathbb{P}\{X = b\} & \text{if } \gamma = b, \\ \infty & \text{if } \gamma > b. \end{cases}$$

Moreover, if $\mathbb{E}[X] \leq \gamma < b$, then the minimizer $\mathcal{Q}$ is equal to the tilted distribution $\mathcal{P}_\lambda$ defined by $d\mathcal{P}_\lambda(x) = \exp(\lambda x - \psi_X(\lambda))\, d\mathcal{P}(x)$.

Proof. We assume for simplicity that $\mathcal{P}$ and $\mathcal{Q}$ have densities p and q respectively.

First case: Taking $\mathcal{Q} = \mathcal{P}$ gives $\mathsf{KL}(\mathcal{Q},\mathcal{P}) = 0$.

Fourth case: We have $q(x) > 0$ and $p(x) = 0$ on a nontrivial subset of (b,∞), which gives that $\mathbb{E}_q[\log \frac{q}{p}] = \infty$.

Third case: If $\mathbb{P}_p\{X = b\} = 0$, then the situation is similar to the previous case, namely, $q(x) > 0$ and $p(x) = 0$ on a nontrivial subset of (b,∞).

If $\mathbb{P}_p\{X = b\} > 0$, then $\mathbb{P}_q\{X \leq b\} = 1$ because otherwise $\mathsf{KL}(\mathcal{Q},\mathcal{P}) = \infty$ by a similar argument. To have $\mathbb{E}_q[X] = b$, it must hold that $\mathbb{P}_q\{X = b\} = 1$. Therefore, $\mathbb{E}_q[\log \frac{q}{p}] = \log \frac{1}{\mathbb{P}_p\{X=b\}}$.

Second case: Let λ be such that $\gamma = \psi'_X(\lambda) = \mathbb{E}_\lambda[X]$, where $\mathbb{E}_\lambda$ is with respect to the tilted distribution $p_\lambda(x) = \exp(\lambda x - \psi_X(\lambda))\, p(x)$. Moreover, the first-order optimality condition implies that $\psi_X^*(\gamma) = \lambda\gamma - \psi_X(\lambda)$. For any $\mathcal{Q}$ such that $\mathbb{E}_q[X] \geq \gamma$, we have

$$\begin{aligned} \mathsf{KL}(\mathcal{Q},\mathcal{P}) = \mathbb{E}_q\left[\log \frac{q\cdot p_\lambda}{p\cdot p_\lambda}\right] &= \mathsf{KL}(\mathcal{Q},\mathcal{P}_\lambda) + \mathbb{E}_q\left[\log \frac{p_\lambda}{p}\right] \\ &= \mathsf{KL}(\mathcal{Q},\mathcal{P}_\lambda) + \mathbb{E}_q[\lambda X - \psi_X(\lambda)] \\ &\geq \mathsf{KL}(\mathcal{Q},\mathcal{P}_\lambda) + \lambda\gamma - \psi_X(\lambda) \\ &= \mathsf{KL}(\mathcal{Q},\mathcal{P}_\lambda) + \psi_X^*(\gamma) \geq \psi_X^*(\gamma), \end{aligned}$$

where both inequalities become equalities if $\mathcal{Q} = \mathcal{P}_\lambda$, and the last inequality is an equality if and only if $\mathcal{Q} = \mathcal{P}_\lambda$. This not only proves the second case of claim 1 but also claims 2 and 3. □

The above theorems combined yield the following result.

Corollary 4.19. *For i.i.d. $X, X_1, \ldots, X_n \sim \mathcal{P}$, let $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$. If $\mathbb{E}[X] \leq \gamma < \operatorname{esssup}(X)$, then we have*

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\bar{X} \geq \gamma\} = \psi_X^*(\gamma).$$

In particular, the Chernoff bound is tight.

This result is precisely the reason why the Legendre–Fenchel transform of the CGF is called the rate function. Moreover, in retrospect, it is not surprising that the Chernoff bound is tight: It employs the same type of change of measure as in the definition of the tilted distribution, and the information projection argument shows that the tilted distribution is the best change of measure.

The tool of information projection also has the following more general consequence, called Sanov's theorem.

Theorem 4.20. *Consider i.i.d. random variables $X_1, \dots, X_n \sim \mathcal{P}$. Let $\hat{\mathcal{P}} := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical distribution, where δ_{X_i} denotes the delta measure (i.e., the point mass) at X_i. Let $\mathcal{E}$ be a convex set of distributions. Then under some regularity assumptions on $\mathcal{P}$ and $\mathcal{E}$, we have*

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\hat{\mathcal{P}} \in \mathcal{E}\} = \min_{\mathcal{Q}\in\mathcal{E}} \mathsf{KL}(\mathcal{Q}, \mathcal{P}).$$

For example, a sufficient set of regularity assumptions is the following: The sample space $\mathcal{X}$ is a Polish space, and the set $\mathcal{E}$ is weakly closed and has a nonempty interior. We skip the proof.

4.7 Implication of large deviation on testing errors

Recall the setup where we observe i.i.d. $X_1, \dots, X_n$ from p under H_0 and from q under H_1. For a test $\phi^{(n)}$, we would like the expected type I error $\alpha_{\phi^{(n)}} = \mathbb{E}_p[\phi^{(n)}]$ to decay as e^{-nE_0} and the expected type II error $1 - \beta_{\phi^{(n)}}$ where $\beta_{\phi^{(n)}} = \mathbb{E}_q[\phi^{(n)}]$ to decay as e^{-nE_1}. There is a trade-off between E_0 and E_1 as they cannot be arbitrarily large at the same time.

Let us define

$$T = \log \frac{q(X)}{p(X)}, \qquad T_i = \log \frac{q(X_i)}{p(X_i)}$$

so that the log-likelihood is

$$\log \frac{q(X_1)\cdots q(X_n)}{p(X_1)\cdots p(X_n)} = \sum_{i=1}^n T_i.$$

The CGF of T under p is

$$\psi_p(\lambda) = \log \mathbb{E}_p[e^{\lambda T}] = \log \int_{\mathcal{X}} p(x)^{1-\lambda} q(x)^\lambda \, d\mu(x),$$

and the rate function is

$$\psi_p^*(\gamma) = \sup_{\lambda\in\mathbb{R}} \big(\lambda\gamma - \psi_p(\lambda)\big).$$

Note that $\psi_p(0) = \psi_p(1) = 0$. Since $\psi_p(\lambda)$ is convex, it is finite for $\lambda \in [0, 1]$.

The CGF $\psi_p(\lambda)$ is related to the Rényi divergence, defined by

$$D_\lambda(p, q) := \frac{1}{\lambda - 1} \log \mathbb{E}_q\left[\left(\frac{p(X)}{q(X)}\right)^\lambda\right] = \frac{1}{\lambda - 1} \log \int_{\mathcal{X}} p(x)^\lambda q(x)^{1-\lambda} \, d\mu(x)$$

where $\lambda \neq 1$. It can be shown that $D_\lambda(p, q) \geq 0$, and by L'Hôpital's rule,

$$\lim_{\lambda\to 1} D_\lambda(p, q) = \lim_{\lambda\to 1} \frac{\mathbb{E}_q[(p/q)^\lambda \log(p/q)]}{\mathbb{E}_q[(p/q)^\lambda]} = \mathbb{E}_p[(p/q)\log(p/q)] = \mathsf{KL}(p, q).$$

Moreover, we have

$$\psi_p(\lambda) = (\lambda - 1)\, D_\lambda(q, p) = -\lambda\, D_{1-\lambda}(p, q).$$

This gives another explanation why $\psi_p(0) = \psi_p(1) = 0$ and $\psi_p(\lambda) < 0$ for $\lambda \in (0, 1)$. It follows that

$$\psi_p'(0) = \lim_{\lambda\to 0} \frac{\psi_p(\lambda)}{\lambda} = -\lim_{\lambda\to 0} D_{1-\lambda}(p, q) = -\mathsf{KL}(p, q)$$

and similarly

$$\psi_p'(1) = \mathsf{KL}(q, p).$$

Theorem 4.21. *In the above setting, the best achievable exponents (E_0, E_1) are characterized by*

$$E_0(\gamma) = \psi_p^*(\gamma), \qquad E_1(\gamma) = \psi_p^*(\gamma) - \gamma,$$

where $\gamma \in [-\mathsf{KL}(p,q), \mathsf{KL}(q,p)]$. Moreover, $E_0(\gamma)$ is increasing and $E_1(\gamma)$ is decreasing.

Proof. The idea is to apply large deviation theory to the sum $\bar{T} = \frac{1}{n}\sum_{i=1}^n T_i$. First, consider the likelihood-ratio test $\phi^{(n)}$ which rejects H_0 if $\bar{T} \geq \gamma$ for a constant $\gamma \in \mathbb{R}$. Then the Chernoff bound implies that

$$\alpha_{\phi^{(n)}} = \mathbb{P}_p\{\bar{T} \geq \gamma\} \leq \exp(-n\,\psi_p^*(\gamma))$$

if

$$\gamma \geq \mathbb{E}_p[T] = \mathbb{E}_p\Big[\log \frac{q(X)}{p(X)}\Big] = -\,\mathsf{KL}(p,q).$$

Similarly, we have

$$1 - \beta_{\phi^{(n)}} = \mathbb{P}_q\{\bar{T} < \gamma\} \leq \exp(-n\,\psi_q^*(\gamma))$$

if

$$\gamma \leq \mathbb{E}_q[T] = \mathbb{E}_q\Big[\log \frac{q(X)}{p(X)}\Big] = \mathsf{KL}(q,p).$$

It remains to note that

$$\psi_q(\lambda) = \log \mathbb{E}_q[e^{\lambda \log(q/p)}] = \log \mathbb{E}_p[e^{\lambda \log(q/p)} \cdot (q/p)] = \log \mathbb{E}_p[e^{(\lambda+1)\log(q/p)}] = \psi_p(\lambda+1)$$

and so

$$\psi_q^*(\gamma) = \sup_{\lambda \in \mathbb{R}} \Big(\lambda\gamma - \psi_p(\lambda+1)\Big) = \sup_{\lambda \in \mathbb{R}} \Big(\lambda\gamma - \psi_p(\lambda)\Big) - \gamma = \psi_p^*(\gamma) - \gamma.$$

Therefore, the exponents $E_0(\gamma) = \psi_p^*(\gamma)$ and $E_1(\gamma) = \psi_p^*(\gamma) - \gamma$ are achievable.

Conversely, it can be shown that the exponents achieved by the likelihood-ratio tests are in fact optimal. □

The above result has the following consequence in the Bayesian setting.

Corollary 4.22. *Consider the prior $\Pi = \mathsf{Ber}(\pi_1)$ over the two hypotheses for $\pi_1 \in (0,1)$. Let $\pi_0 = 1 - \pi_1$. Define the optimal Bayes risk as*

$$R_n^*(\Pi) := \min_{\phi^{(n)}} \Big(\pi_0\,\alpha_{\phi^{(n)}} + \pi_1\,(1 - \beta_{\phi^{(n)}})\Big).$$

Then we have

$$-\lim_{n\to\infty} \frac{1}{n} \log R_n^*(\Pi) = \max_{\gamma} \min\{E_0(\gamma), E_1(\gamma)\} = \psi_p^*(0) = -\min_{\lambda \in [0,1]} \psi_p(\lambda).$$

The quantity $\psi_p^*(0)$ is called the Chernoff exponent and does not depend on the prior.

We also have the following equivalent formulations of the optimal exponents for the testing errors.

Theorem 4.23. *In the above setting, the best exponents (E_0, E_1) can be stated in the following equivalent forms:*

1. *$E_0 = \mathsf{KL}(p_\lambda, p)$ and $E_1 = \mathsf{KL}(p_\lambda, q)$, where p_λ denotes the tilted distribution obtained from tilting p along T towards q for $\lambda \in [0,1]$, defined by*

$$p_\lambda(x) := e^{\lambda T - \psi_p(\lambda)} p(x) = p(x)^{1-\lambda} q(x)^\lambda e^{-\psi_p(\lambda)}.$$

2. *$E_0 \in [0, \mathsf{KL}(q,p)]$ and*

$$E_1 = E_1^*(E_0) = \min_{q' : \mathsf{KL}(q',p) \le E_0} \mathsf{KL}(q', q).$$

Proof. We give a sketch of the proof.

1. Fix λ and define $\gamma = \gamma(\lambda) = \mathbb{E}_{p_\lambda}[T]$. Then it can be shown that

$$\mathsf{KL}(p_\lambda, p) = \psi_p^*(\gamma)$$

   whereas

$$\mathsf{KL}(p_\lambda, q) = \mathbb{E}_{p_\lambda}[\log(p_\lambda/q)] = \mathbb{E}_{\mathbb{P}_\lambda}[\log(p_\lambda/p) - \log(p/q)] = \mathsf{KL}(p_\lambda, p) - \mathbb{E}_{p_\lambda}[T] = \psi_p^*(\gamma) - \gamma.$$

   Moreover, as $\lambda \in [0,1]$, we have $\gamma = \mathbb{E}_{p_\lambda}[T] \in [-\mathsf{KL}(p,q), \mathsf{KL}(q,p)]$.

2. For simplicity, suppose that q^* achieves the minimum in the definition of E_1, and that $q^* \neq p$ and $q^* \neq q$. Then we have

$$\mathsf{KL}(q^*, q) \le \mathsf{KL}(p, q)$$

   in view of the objective of the minimization problem and

$$\mathsf{KL}(q^*, p) \le E_0 \le \mathsf{KL}(q, p)$$

   in view of the constraint of the minimization problem. It follows that

$$\mathbb{E}_{q^*}[T] = \mathbb{E}_{q^*}\left[\log \frac{q^*}{p} \frac{q}{q^*}\right] = \mathsf{KL}(q^*, p) - \mathsf{KL}(q^*, q) \in [-\mathsf{KL}(p,q), \mathsf{KL}(q,p)].$$

   It can be shown that there is a unique tilted distribution p_λ satisfying

$$\mathbb{E}_{p_\lambda}[T] = \mathbb{E}_{q^*}[T], \qquad \mathsf{KL}(p_\lambda, p) \le \mathsf{KL}(q^*, p), \qquad \mathsf{KL}(p_\lambda, q) \le \mathsf{KL}(q^*, q).$$

   We then conclude that $q^* = p_\lambda$ and the result follows from the first part.

□

Chapter 5

Modern topics in testing and inference

5.1 Multiple testing and FDR control

Suppose that we have m tests for testing between H_{0i} and H_{1i} for $i = 1, \dots, m$. Let p_i denote the p-value of the ith test. Recall that for $\alpha \in (0,1)$, each individual test is at significance level α if it rejects the null when $p_i \le \alpha$. Instead, to have the overall probability of falsely rejecting any H_{0i} for $i = 1, \dots, m$ less than or equal to α, by a union bound, we need to reject H_{0i} if $p_i \le \alpha/m$. This is referred to as the Bonferroni method. However, the requirement that we do not falsely reject any null is too strict when m is large.

5.1.1 False discovery rate

Rather than disallowing any false rejection of the null, a more practical idea is to control the rate of false rejection. For the m tests indexed by $i = 1, \dots, m$, let R denote the number of times we reject the null H_{0i}, and let V denote the number of times the rejection is wrong, i.e., the truth is H_{0i}. Then the false discovery proportion (FDP) is defined to be V/R, and the false discovery rate (FDR) is defined to be $\mathbb{E}[V/R]$.

Before introducing a method to control the FDR, we first establish a basic fact about p-values. For a family of tests ϕ_α each at significance level $\alpha \in (0,1)$, consider the region of rejection $S_1(\alpha) := \{x \in \mathcal{X} : \phi_\alpha(x) = 1\}$. Recall that the p-value is defined as $\hat{p}(X) := \inf\{\alpha : X \in S_1(\alpha)\}$.

Lemma 5.1. *Suppose that under H_0, we have $\mathbb{P}\{X \in S_1(\alpha)\} = \alpha$ for all $\alpha \in (0,1)$. Then, under H_0, we have $\mathbb{P}\{\hat{p}(X) \le t\} = t$ for $t \in (0,1)$. In other words, the p-value $\hat{p}(X)$ is uniformly distributed over $(0,1)$.*

Proof. For any $t \in (0,1)$, we have $\hat{p}(X) \le t$ if and only if $X \in S_1(t)$, so the result follows. □

Let us consider the following procedure, which is called the Benjamini–Hochberg method:

1. Let $p_{(1)} < p_{(2)} < \cdots < p_{(m)}$ be the order statistics of the p-values.
2. Define $\ell_i := \frac{\alpha\, i}{m\, C_m}$ where $C_m = 1$ if the p-values are independent and $C_m = \sum_{j=1}^{m} (1/j)$ otherwise. Then define $R^* := \max\{i \in [m] : p_{(i)} \le \ell_i\}$.

3. Reject all null hypotheses H_{0i} for which $p_i \le p_{(R^*)}$.

Theorem 5.2. *Consider the Benjamini–Hochberg (BH) method described above. Let V^* denote the number of times the rejection is wrong, i.e., the truth is H_{0i}. The FDR is defined to be $\mathbb{E}\left[\frac{V^*}{\max\{R^*,1\}}\right]$. Then we have $\mathbb{E}\left[\frac{V^*}{\max\{R^*,1\}}\right] \le \alpha$. In other words, the BH method achieves an FDR at most α.*

This result is quite general—it does not rely on the underlying statistical model nor the number of correct null hypotheses. We will prove the above theorem for the case where the p-values are independent (and thus $C_m = 1$). Let us first discuss the intuition, assuming $R^* \ge 1$. Up to a relabeling of the hypotheses, we may assume that the hypotheses $H_{01}, \dots, H_{0m_0}$ are true and $H_{1(m_0+1)}, \dots, H_{1m}$ are true. The key to proving the above theorem is establishing the inequality

$$\mathbb{E}[V^*/R^* \mid p_{m_0+1}, \dots, p_m] \le \frac{m_0}{m}\alpha,$$

which holds regardless of the values of $p_{m_0+1}, \dots, p_m$. It then follows that $\mathbb{E}[V^*/R^*] \le \alpha$. The intuition of the above inequality simply lies in the critical condition $R^*\frac{\alpha}{m} \approx V^*\frac{1}{m_0}$, which holds because we reject at level $R^*\frac{\alpha}{m}$ and for the m_0 null hypotheses, the p-values are uniform over $(0,1)$.

5.1.2 Analysis of the Benjamini–Hochberg method

One way to analyze the BH method is through continuous-time stochastic processes. To this end, we first introduce martingales (informally, without using any measure theory). Let $\{W_t : t \ge 0\}$ be a continuous-time stochastic process, i.e., an infinite collection of random variables W_t indexed by $t \ge 0$. The stochastic process $\{W_t : t \ge 0\}$ is called a martingale if $\mathbb{E}[|W_t|] < \infty$ and

$$\mathbb{E}\left[W_t \mid \{W_r : 0 \le r \le s\}\right] = W_s$$

for any $0 \le s \le t$. A random variable τ taking values in $[0, \infty)$ is called a stopping time if the event $\{\tau = t\}$ only depends on $\{W_s : 0 \le s \le t\}$. The optional stopping theorem states that, if $\tau \le T$ for a constant T, then $\mathbb{E}[W_\tau] = \mathbb{E}[W_0]$.

An equivalent formulation of the BH method For $t \in [0,1]$, define

$$R(t) := |\{i \in [m] : p_i \le t\}|,$$

where $|\cdot|$ denotes the cardinality of a set. Moreover, define

$$Q(t) := \frac{mt}{\max\{R(t), 1\}}, \qquad t_\alpha := \sup\{t \in [0,1] : Q(t) \le \alpha\}.$$

We claim that the BH method is equivalent to rejecting H_{0i} for which $p_i \le t_\alpha$.

First, suppose that the supremum t_α satisfies $R(t_\alpha) = 0$. Then we have

$$t_\alpha = \sup\left\{t \in [0,1] : mt \le \alpha\right\} = \alpha/m.$$

In this case, $p_i > t_\alpha = \alpha/m$ for all $i \in [m]$ by the definition of $R(t_\alpha)$, so we accept all null hypotheses H_{0i} according to the above rule. Then we need to show that the BH method does the same.

To this end, note that we have already showed that $p_{(1)} > \alpha/m$. For any $i = 2, \ldots, m$, since $\frac{\alpha\, i}{m} > t_\alpha = \alpha/m$, by the definition of t_α, we obtain

$$\alpha < Q(\alpha\, i/m) = \frac{\alpha\, i}{\max\{R(\alpha\, i/m), 1\}}.$$

It follows that $i > R(\alpha\, i/m) = |\{j \in [m] : p_j \le \alpha\, i/m\}|$, and so $p_{(i)} > \alpha\, i/m$. Consequently, $R^* = 0$ in the BH method and all null hypotheses are accepted.

Next, assume $R(t_\alpha) \ge 1$. By the definition of $R(t_\alpha)$, we have $t_\alpha \ge p_{(1)}$, so at least one null hypothesis is rejected. Moreover, by the definition of t_α, we have $\alpha \ge Q(p_{(1)}) = m p_{(1)}$ and thus $p_{(1)} \le \alpha/m$. It follows that $R^* \ge 1$ and the BH method also rejects at least one null hypothesis.

To prove equivalence of the two rejection rules, it suffices to show that $p_{(R^*)} \le t_\alpha < p_{(R^*+1)}$. Towards this end, we note that

$$t_\alpha = \sup\left\{t \in [p_{(1)}, 1] : \frac{mt}{R(t)} \le \alpha\right\} = \sup\left\{t \in [p_{(1)}, 1] : |\{i \in [m] : p_i \le t\}| \ge mt/\alpha\right\}$$

by the definition of $R(t)$. Since

$$|\{i \in [m] : p_i \le p_{(R^*)}\}| = R^* \ge m\, p_{(R^*)}/\alpha$$

by the definition of R^*, we obtain that $p_{(R^*)} \le t_\alpha$. On the other hand, the definition of R^* also implies that

$$|\{i \in [m] : p_i \le p_{(R^*+1)}\}| = R^* + 1 < m\, p_{(R^*+1)}/\alpha,$$

so $p_{(R^*)} > t_\alpha$. This completes the proof of the claim.

Auxiliary stochastic processes Define $\mathcal{I}_0 := \{i \in [m] : H_{0i} \text{ is true}\}$ and

$$V(t) := |\{i \in [m] : p_i \le t,\, H_{0i} \text{ is true}\}| = \sum_{i \in \mathcal{I}_0} \mathbb{1}\{p_i \le t\}, \qquad W(t) := \frac{V(t)}{t}.$$

We claim that

$$\mathbb{E}\left[W(t) \mid \{W(r)\}_{r=s}^1\right] = W(s)$$

for any $0 \le t \le s \le 1$. In other words, $\{W_t\}_{t=0}^1$ is a backward martingale as t decreases from 1 to 0.

To see this, recall that p_i for $i \in \mathcal{I}_0$ are i.i.d. uniform distributions over $[0, 1]$. For $0 \le t \le s \le 1$, let us condition on $\{V(r)\}_{r=s}^1$, i.e., condition on the set of p-values $\{p_i : i \in \mathcal{I}_0,\, p_i > s\}$. For the remaining $V(s)$ p-values, by symmetry, each of them is uniformly distributed over $[0, s]$. As a result, the conditional expectation of $\mathbb{1}\{p_i \le t\}$ is t/s. We obtain that

$$\mathbb{E}\left[V(t) \mid \{V(r)\}_{r=s}^1\right] = V(s) \cdot t/s.$$

The claim then follows easily.

Since we are going backward in time from $t = 1$ to $t = 0$, the quantity t_α is the first moment t such that $Q(t)$ falls below α. Therefore, t_α is a stopping time. By the optional stopping theorem,

$$\mathbb{E}[W(t_\alpha)] = \mathbb{E}[W(1)] = |\mathcal{I}_0|.$$

Finishing the proof By the definitions of $Q(t)$ and t_α, we have

$$\alpha = Q(t_\alpha) = \frac{m\, t_\alpha}{\max\{R(t_\alpha), 1\}}$$

and so

$$\frac{V(t_\alpha)}{\max\{R(t_\alpha), 1\}} = \frac{\alpha\, V(t_\alpha)}{m\, t_\alpha}.$$

In addition, note that $R^* = R(t_\alpha)$ and $V^* = V(t_\alpha)$ in the BH method. Taking the expectation of the above equation yields

$$\mathbb{E}\left[\frac{V^*}{\max\{R^*, 1\}}\right] = \mathbb{E}\left[\frac{V(t_\alpha)}{\max\{R(t_\alpha), 1\}}\right] = \mathbb{E}\left[\frac{\alpha\, V(t_\alpha)}{m\, t_\alpha}\right] = \frac{\alpha}{m}\,\mathbb{E}[W(t_\alpha)] = \frac{\alpha}{m}|\mathcal{I}_0| \le \alpha.$$

Therefore, we have proved Theorem 5.2.

5.1.3 False coverage rate

Recall that confidence regions are obtained by reformulating hypothesis testing in terms of a region covering the true parameter. In the same vein, when dealing with multiple testing, we can reformulate the FDR to obtain a related notion called the false coverage rate (FCR).

To be more specific, suppose that we have m inference problems and would like to construct a confidence region $\mathcal{C}_i$ such that $\theta_i \in \mathcal{C}_i$ with high probability for $i = 1, \ldots, m$. Let $S_i \in \{0, 1\}$ be the unknown indicator of whether i is selected for coverage. Then we can define $R := \sum_{i=1}^m S_i$ and $V := \sum_{i=1}^m S_i \cdot \mathbb{1}\{\theta_i \notin \mathcal{C}_i\}$. The false coverage proportion (FCP) is defined to be V/R, and the FCR is defined to be $\mathbb{E}[V/R]$.

5.2 Variable selection

5.2.1 Conditional randomization testing

Consider the setting that a response Y may depend on a large number of covariates $X_1, \ldots, X_d$. Our goal is to select a subset of variables that the response truly depends on. To formulate the problem, we define a null variable X_i to be one such that

$$Y \perp X_i \mid X_{-i},$$

that is, X_i is independent of Y conditional on $X_{-i} = (X_j : j \in [d],\, j \neq i)$. For example, in a linear model

$$Y = \sum_{i=1}^d \beta_i X_i + \varepsilon$$

where ε is random noise, the set of null variables is $\{X_i : i \in [d],\, \beta_i = 0\}$.

There are many methods for such a variable selection or feature selection problem. We introduce conditional randomization testing (CRT) in this section. Suppose that we know the conditional

distribution of $X_i \mid X_{-i}$. Then we can sample $\tilde{X}_i$ from this distribution (conditionally independently from X_i). If X_i is a null variable, then

$$\begin{aligned} p(X_i, X_{-i}, Y) &= p(X_i \mid X_{-i}, Y) \cdot p(X_{-i}, Y) \\ &= p(X_i \mid X_{-i}) \cdot p(X_{-i}, Y) \\ &= p(\tilde{X}_i \mid X_{-i}) \cdot p(X_{-i}, Y) = p(\tilde{X}_i, X_{-i}, Y). \end{aligned}$$

Therefore, given $\tilde{X}_i$, we can test whether

$$(X_1, \ldots, X_i, \ldots, X_d, Y) \stackrel{d}{=} (X_1, \ldots, \tilde{X}_i, \ldots, X_d, Y)$$

to decide if X_i is a null variable.

Given a feature importance score $T(\cdot)$, the following procedure obtains the p-values for testing whether X_i is null:

1. Compute the score $t_i^* := T(X_i, X_{-i}, Y)$.
2. For $k = 1, \ldots, K$, sample $\tilde{X}_i^{(k)} \sim X_i \mid X_{-i}$ and compute the score $t_i^{(k)} := T(\tilde{X}_i^{(k)}, X_{-i}, Y)$.
3. Compute the p-value
$$p_i := \frac{|\{k \in [K] : t_i^{(k)} \geq t_i^*\}| + 1}{K + 1}.$$

Under the null, all the scores t_i^* and $t_i^{(k)}$ for $k \in [K]$ are identically distributed. As a result, p_i is uniform over $\{\frac{1}{K+1}, \frac{2}{K+1}, \ldots, 1\}$. We reject the null if p_i is too extreme.

For example, in a linear model, the score function may be defined as the magnitude of the estimated coefficient for each coordinate.

Note that the p-values are not independent because the copies of $\tilde{X}_i$ are sampled conditional on X_{-i}. This can potentially be an issue if we would like to obtain finer properties of the test. A more serious issue with CRT is its computational complexity. For example, in linear regression, if we use the magnitude of a coefficient as the feature importance score (or any other usual choice), then we would solve linear regression $K \times d$ times in total for testing all the variables $X_1, \ldots, X_d$. Moreover, if d is large, we typically need K to be large as well, making the computational complexity even worse.

5.2.2 Knockoffs

We continue the setup from above. Consider a simple example: $d = 2$ and

$$Y = X_2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \qquad X_1, X_2 \sim \mathcal{N}(0, 1), \qquad \mathbb{E}[X_1 X_2] = 0.5.$$

For an estimator $(\hat{\beta}_1, \hat{\beta}_2)$ of the coefficients, let us use $|\hat{\beta}_i|$ as the score function. In this example, X_1 is a null variable, but $|\beta_1|$ may not be small. This is because X_1 is correlated with X_2 and may influence Y through X_2.

The knockoff approach proposed by [BC15] aims to resolve this in a way that is computationally more efficient than CRT. More precisely, a set of variables $\tilde{X}_1, \ldots, \tilde{X}_d$ are called model-X knockoffs if the following two statements hold:

1. Pairwise exchangeability: If X_i is a null variable, then

$$(X_i, X_{-i}, \tilde{X}_i, \tilde{X}_{-i}) \stackrel{d}{=} (\tilde{X}_i, X_{-i}, X_i, \tilde{X}_{-i});$$

2. Response independence: $Y \perp \tilde{X} \mid X$, where $X = (X_1, \ldots, X_d)$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$.

Having constructed $\tilde{X}$, we compute the importance scores

$$Z_i = T_i(X, \tilde{X}, Y), \qquad \tilde{Z}_i = T_{d+i}(X, \tilde{X}, Y), \qquad i \in [d],$$

for the actual observed variables X_i and the knockoffs $\tilde{X}_i$. For example, in a linear model, we concatenate X and $\tilde{X}$ to obtain $2d$ covariates and regress Y on all of them; then we can define $Z_i = |\hat{\beta}_i|$ and $\tilde{Z}_i = |\hat{\beta}_{d+i}|$ for some estimator $\hat{\beta}$ of the coefficients. As a result of the pairwise exchangeability, we have $(Z_i, \tilde{Z}_i) \stackrel{d}{=} (\tilde{Z}_i, Z_i)$ if X_i is a null variable.

Next, we construct knockoff-adjusted scores $W_i = w_i(Z_i, \tilde{Z}_i)$ via some anti-symmetric function $w_i : \mathbb{R}^d \to \mathbb{R}$, i.e., $w_i(Z_i, \tilde{Z}_i) = -w_i(\tilde{Z}_i, Z_i)$. For example, a simple choice is

$$W_i = w_i(Z_i, \tilde{Z}_i) = Z_i - \tilde{Z}_i.$$

Lemma 5.3. *For any null variable X_i, the distribution W_i is symmetric so that $\mathsf{sign}(W_i)$ is a Rademacher random variable. Moreover, conditional on $|W| = (|W_1|, \ldots, |W_d|)$, we have that $\mathsf{sign}(W_1), \ldots, \mathsf{sign}(W_d)$ are independent.*

We now introduce a procedure for testing $H_1, \ldots, H_d$ while controlling the FDR, where H_i denotes the null hypothesis that X_i is a null variable. For $t > 0$, define

$$S^+(t) := \{i \in [d] : W_i \geq t\}, \qquad S^-(t) := \{i \in [d] : W_i \leq -t\},$$

and

$$\hat{F}(t) := \frac{|S^-(t)| + 1}{|S^+(t)| \vee 1}.$$

This is an estimator of the FDP $F(t)$ defined by

$$F(t) := \frac{|\{i \in [d] : X_i \text{ is null }, W_i \geq t\}|}{|S^+(t)| \vee 1} \approx \frac{|\{i \in [d] : X_i \text{ is null }, W_i \leq -t\}|}{|S^+(t)| \vee 1} \leq \hat{F}(t).$$

Finally, for $\alpha \in (0, 1)$, we define

$$\hat{S} := \{i \in [d] : W_i \geq \tau_\alpha\}, \qquad \tau_\alpha := \min\left\{t > 0 : \hat{F}(t) \leq \alpha\right\}.$$

Theorem 5.4. *If we reject all $i \in \hat{S}$, then the FDR is controlled below α.*

The knockoff approach has some advantages over other methods. First, the approach and the theoretical guarantee are model-free, as we put essentially no assumptions on (X, Y). Second, it requires only a regression of Y on $(X, \tilde{X})$ and is computationally more efficient than CRT.

Proof. We provide a sketch of the proof of the theorem. Let $\mathcal{H}_0 := \{i \in [d] : X_i \text{ is null}\}$. Then

$$F(\tau_\alpha) = \frac{|\mathcal{H}_0 \cap S^+(\tau_\alpha)|}{|S^+(\tau_\alpha)| \vee 1} \le \frac{|\mathcal{H}_0 \cap S^+(\tau_\alpha)|}{|\mathcal{H}_0 \cap S^-(\tau_\alpha)| + 1} \cdot \frac{|S^-(\tau_\alpha)| + 1}{|S^+(\tau_\alpha)| \vee 1} \le \alpha \cdot \frac{|\mathcal{H}_0 \cap S^+(\tau_\alpha)|}{|\mathcal{H}_0 \cap S^-(\tau_\alpha)| + 1}.$$

To prove $\mathbb{E}[F(\tau_\alpha)] \le \alpha$, it suffices to show that

$$\mathbb{E}\Big[\frac{V^+(\tau_\alpha)}{V^-(\tau_\alpha) + 1}\Big] \le 1,$$

where $V^+(\tau_\alpha) := |\mathcal{H}_0 \cap S^+(\tau_\alpha)|$ and $V^-(\tau_\alpha) := |\mathcal{H}_0 \cap S^-(\tau_\alpha)|$.

Similar to the proof for the Benjamini–Hochberg method in FDR control, we argue that $\frac{V^+(t)}{V^-(t)+1}$ is a supermartingale as t increases from 0 to 1. Therefore, the optional stopping theorem implies

$$\mathbb{E}\Big[\frac{V^+(\tau_\alpha)}{V^-(\tau_\alpha) + 1}\Big] \le \mathbb{E}\Big[\frac{V^+(0)}{V^-(0) + 1}\Big] = \mathbb{E}\Big[\frac{V^+(0)}{|\mathcal{H}_0| - V^+(0) + 1}\Big].$$

Furthermore, by exchangeability, we have $V^+(0) \sim \mathsf{Bin}(|\mathcal{H}_0|, 1/2)$. Then it is not hard to do an explicit computation to show that the above expectation is bounded by 1. □

5.3 Selective inference

5.3.1 False coverage rate and confidence intervals

We consider the situation where we use data to select some parameters and then form confidence intervals for the selected parameters. As a motivating example, consider the Gaussian sequence model

$$Y_i \sim \mathcal{N}(\theta_i, 1), \qquad i = 1, \dots, n.$$

For $\alpha \in (0, 1)$ and each $i \in [n]$, we can construct a confidence interval for θ_i at confidence level $1 - \alpha$:

$$\mathsf{CI}_i(\alpha) := (Y_i - z_{\alpha/2}, Y_i + z_{\alpha/2}),$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of $\mathcal{N}(0, 1)$. Moreover, suppose that we would like to select a subset of parameters θ_i that are nonzero. A natural selection is

$$S := \{i \in [n] : 0 \notin \mathsf{CI}_i(\alpha)\}.$$

For a selected variable, a measure of the quality of the confidence interval is the conditional coverage

$$\mathbb{P}\{\theta_i \in \mathsf{CI}_i(\alpha) \mid i \in S\}.$$

Note that the conditioning distorts the coverage: If θ_i is close to 0, then the coverage is low. In particular, regardless of how small α is, the conditional coverage goes to 0 as $\theta_i \to 0$. This shows that it is impossible to achieve good conditional coverage for each individual parameter.

Instead, we consider the FCR introduced before, redefined here as

$$\mathbb{E}\Big[\frac{V}{R \vee 1}\Big],$$

where $R = |S|$ is the number of selected parameters and V is the number of the selected parameters that are not covered by the corresponding confidence intervals, i.e.,

$$V = |\{i \in S : \theta_i \notin \mathsf{CI}_i(\alpha)\}|.$$

Two remarks about the FCR:

- If $S = [n]$ and $R = n$ (i.e., without selection), the FCR is automatically controlled:

$$\mathbb{E}\left[\frac{V}{n}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}\{\theta_i \notin \mathsf{CI}_i(\alpha)\}}{n}\right] \leq \alpha,$$

  because by the definition of a confidence interval, we have $\mathbb{P}\{\theta_i \notin \mathsf{CI}_i(\alpha)\} \leq \alpha$.

- Bonferroni's method replaces α by α/n and therefore controls the FCR:

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq \mathbb{E}[V] \leq \sum_{i=1}^n \mathbb{P}\{\theta_i \notin \mathsf{CI}_i(\alpha/n)\} \leq \alpha.$$

Just as for FDR control, Bonferroni's method is not desirable here, because it results in very wide confidence intervals. Instead, we introduce a less conservative method. Suppose that each parameter θ_i is associated with a test statistic T_i for $i \in [n]$, and we let $T = (T_1, \ldots, T_n)$. Consider the following procedure [BY05]:

1. Apply any selection rule to obtain the selection set $S = S(T)$.

2. For $i \in S$, compute

$$R_i := \min_t \{|S(T_{-i}, t)| : i \in S(T_{-i}, t)\},$$

   where T_{-i} denotes the set of test statistics T without T_i.

3. For $i \in S$, define the FCR-adjusted confidence interval for θ_i to be $\mathsf{CI}_i(R_i\,\alpha/n)$.

The second step may appear to be complex, but it is often the case that $R_i = R = |S|$ for reasonable selection rules. Before establishing the general theorem, we consider two extreme cases:

- If $R = n$, then we make no adjustment and still use the confidence interval $\mathsf{CI}_i(\alpha)$.

- If $R = 1$, then we get Bonferroni's method for the one selected confidence interval $\mathsf{CI}_i(\alpha/n)$.

Theorem 5.5. *Suppose that the statistics T_i are independent for $i \in [n]$. The adjusted confidence intervals defined in the above procedure achieve an FCR at most α.*

Proof. We can write the FCR as

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] = \sum_{i=1}^n \mathbb{E}[X_i], \qquad X_i := \frac{\mathbb{1}\{i \in S,\, \theta_i \notin \mathsf{CI}_i(R_i\,\alpha/n)\}}{|S| \vee 1}.$$

It suffices to prove that $\mathbb{E}[X_i] \leq \alpha/n$. Since $R_i \leq |S| = |S(T)|$ by definition, we have

$$X_i = \sum_{k=1}^n \frac{\mathbb{1}\{i \in S,\, \theta_i \notin \mathsf{CI}_i(k\,\alpha/n),\, R_i = k\}}{|S|} \leq \sum_{k=1}^n \frac{\mathbb{1}\{\theta_i \notin \mathsf{CI}_i(k\,\alpha/n),\, R_i = k\}}{k}.$$

Conditional on T_{-i}, it holds that

$$\mathbb{E}[X_i \mid T_{-i}] \le \sum_{k=1}^{n} \frac{\mathbb{P}\{\theta_i \notin \mathsf{CI}_i(k\,\alpha/n)\} \cdot \mathbb{1}\{R_i = k\}}{k} \le \sum_{k=1}^{n} \frac{k\,\alpha/n \cdot \mathbb{1}\{R_i = k\}}{k} \le \alpha/n.$$

It follows that $\mathbb{E}[X_i] = \alpha/n$ and the proof is complete. □

Back to the example of $Y_i \sim \mathcal{N}(\theta_i, 1)$, we simply have $T_i = Y_i$ and $R_i = R = |S|$. Suppose that $\theta_i = \theta$ for all $i \in [n]$. If $\theta \to \infty$, then $R = n$ and the confidence interval are not adjusted. If $\theta = 0$, then $R = 0$ with probability $1 - \alpha$. In both cases, the FCR is α. In between, it can be shown that the FCR is lower bounded by $\alpha/2$, so the procedure is not overly conservative.

5.3.2 Post-selection inference

Setup and classical confidence interval Consider the linear regression model

$$y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n \times d}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. We assume that σ is known for simplicity. A subset $M \subset [d]$ is called a model. In the classical setting, we fix a model M and fit the model with data:

$$\hat{\beta}_M := (X_M^\top X_M)^{-1} X_M^\top Y,$$

where X_M denotes the matrix consisting of the columns of X with indices in M. With M fixed, it is not hard to see that

$$\hat{\beta}_M \sim \mathcal{N}\big(\beta_M, \sigma^2 (X_M^\top X_M)^{-1}\big), \quad \text{where } \beta_M := (X_M^\top X_M)^{-1} X_M^\top X\beta.$$

It follows that

$$\frac{(\hat{\beta}_M)_i - (\beta_M)_i}{\sigma\sqrt{(X_M^\top X_M)^{-1}_{ii}}} = \frac{((X_M^\top X_M)^{-1} X_M^\top \varepsilon)_i}{\sigma\sqrt{(X_M^\top X_M)^{-1}_{ii}}} = \frac{w_i^\top \varepsilon}{\sigma \|w_i\|_2} \sim \mathcal{N}(0, 1),$$

where

$$w_i := X_M (X_M^\top X_M)^{-1} e_i.$$

As a result, we can construct a confidence interval

$$\Big((\hat{\beta}_M)_i - z_{\alpha/2} \cdot \sigma\|w_i\|_2, (\hat{\beta}_M)_i + z_{\alpha/2} \cdot \sigma\|w_i\|_2\Big)$$

that contains $(\beta_M)_i$ with probability $1 - \alpha$. If we would like the above confidence interval to be valid for all $i \in M$, we can replace α by $\alpha/|M|$.

POSI confidence interval In practice, we need to select a model $\hat{M} = \hat{M}(Y)$ based on the data. Then the above confidence interval may not be valid. Post-selection inference (POSI) is a procedure that constructs confidence intervals CI_i such that

$$\mathbb{P}\left\{(\beta_{\hat{M}})_i \in \mathsf{CI}_i \text{ for all } i \in \hat{M}\right\} \ge 1 - \alpha$$

for any data-dependent selected model $\hat{M}$. The POSI confidence interval is of the form

$$\Big((\hat{\beta}_{\hat{M}})_i - K_{\alpha/2} \cdot \sigma \|\hat{w}_i\|_2, (\hat{\beta}_{\hat{M}})_i + K_{\alpha/2} \cdot \sigma \|\hat{w}_i\|_2\Big),$$

where $K_{\alpha/2} = K_{\alpha/2}(X)$ is a certain constant defined in POSI and

$$\hat{w}_i := X_{\hat{M}}(X_{\hat{M}}^\top X_{\hat{M}})^{-1} e_i.$$

See [BBB+13] for more details. It turns out that if α is a small constant, then

$$\sqrt{2 \log d} \lesssim K_\alpha(X) \lesssim \sqrt{d}.$$

If the design matrix X is orthogonal or consists of i.i.d. Gaussian entries, then $K_\alpha(X)$ is close to the lower bound. An advantage of POSI is that it is valid for any model selection process, while disadvantages include that it is very conservative and that it is difficult to compute $K_{\alpha/2}$ in practice.

POSI for LASSO (This part is non-rigorous and contains some errors.) To obtain confidence intervals that are narrower than those from POSI, we can restrict our attention to specific procedures for model selection $\hat{M}$. Let us consider the LASSO estimator

$$\hat{\beta} := \operatorname*{argmin}_{\beta'} \Big(\frac{1}{2}\|Y - X\beta'\|_2^2 + \lambda \|\beta'\|_1\Big)$$

and set

$$\hat{M} := \{i \in [d] : \hat{\beta}_i \neq 0\}.$$

We would like to construct confidence intervals for entries of

$$\beta_{\hat{M}} := (X_{\hat{M}}^\top X_{\hat{M}})^{-1} X_{\hat{M}}^\top X\beta$$

Towards this end, the paper [LSST16] studies the estimate

$$(\hat{\beta}_{\hat{M}})_i = \hat{w}_i^\top Y \sim \mathcal{N}(\hat{w}_i^\top X\beta, \sigma^2 \|\hat{w}_i\|_2^2), \quad \text{where } \hat{w}_i = X_{\hat{M}}(X_{\hat{M}}^\top X_{\hat{M}})^{-1} e_i.$$

For post-selection inference, we need to figure out the conditional distribution of $\hat{w}_i^\top Y$ given that $i \in \hat{M}$. This is difficult, but we can further condition on other events to make the conditional distribution tractable. Namely, we first condition on $\hat{M}$ and also the signs of entries of $\hat{\beta}$. This conditioning can be expressed as a set of linear constraints $AY \leq b$ for a matrix A and a vector b using the KKT conditions. As a result, $\hat{\beta}_{\hat{M}} \mid \{\hat{M}, \mathsf{sign}(\hat{\beta})\}$ is a truncated multivariate Gaussian. Furthermore, to focus on $\hat{w}_i^\top Y$, we can condition on the projection of Y onto $\hat{w}_i^\perp$. It turns out that the distribution of the univariate truncated Gaussian

$$\hat{w}_i^\top Y \mid \big\{\hat{M}, \mathsf{sign}(\hat{\beta}), \Pi_{\hat{w}_i^\perp}(Y)\big\}$$

can be described explicitly. Consequently, confidence intervals can be constructed around $(\hat{\beta}_{\hat{M}})_i = \hat{w}_i^\top Y$ to cover $(\beta_{\hat{M}})_i$ for $i \in \hat{M}$.

5.4 *e*-value

5.4.1 Definition and the associated test

We introduce a new concept called the e-value, which is closely related to the p-value. The motivation for e-values is to address the optional continuation problem, i.e., deciding whether to collect new data and do further testing based on previous test outcomes. In such a sequential setting, p-values can often be misleading because tests at different stages are not independent.

Suppose that we observe data $X \sim \mathcal{P}$ and would like to test a null hypothesis $H_0 : \mathcal{P} \in \mathcal{H}_0$. A nonnegative random variable $E = E(X)$ is called an e-variable for testing H_0 if

$$\sup_{\mathcal{P} \in \mathcal{H}_0} \mathbb{E}_{\mathcal{P}}[E(X)] \leq 1.$$

The value that an e-variable takes is called an e-value. To compare this to the p-value, we can define the p-value in the following way. A random variable P is called a p-random variable for testing H_0 if

$$\sup_{\mathcal{P} \in \mathcal{H}_0} \mathbb{P}_{\mathcal{P}}\{P(X) \leq \alpha\} \leq \alpha \quad \text{for all } \alpha \in (0, 1).$$

The value that a p-variable takes is called a p-value. In other words, a p-variable is a variable that stochastically dominates a uniform variable over $(0, 1)$.

Lemma 5.6. *If $E(X)$ is an e-variable, then $1/E(X)$ is a p-variable.*

Proof. We have

$$\mathbb{P}\{1/E(X) \leq \alpha\} = \mathbb{P}\{E(X) \geq 1/\alpha\} \leq \alpha\, \mathbb{E}[E(X)] \leq \alpha$$

by Markov's inequality. □

Recall that we reject H_0 if $P(X) \leq \alpha$ at significance level $\alpha \in (0, 1)$. Therefore, we can reject H_0 if $E(X) \geq 1/\alpha$. This is called the safe test, in the following sense. Since Markov's inequality is often not tight, $\mathbb{P}\{E(X) \geq 1/\alpha\}$ may be much smaller than α. That is, $E(X)$ rarely exceeds $1/\alpha$, so we rarely reject H_0. Therefore, the safe test given by the e-value is usually conservative.

5.4.2 Bayes factor

Given $X \sim \mathcal{P}_\theta$, consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Let Π_0 and Π_1 denote prior distributions over Θ_0 and Θ_1 respectively. Define the marginal densities as

$$p_{\Pi_r}(x) := \int_{\theta \in \Theta_r} p_\theta(x)\, d\Pi_r(\theta), \qquad r = 0, 1.$$

The ratio between marginal likelihoods is called the Bayes factor:

$$\frac{p_{\Pi_1}(x)}{p_{\Pi_0}(x)}.$$

We can reject H_0 if the Bayes factor is large.

The Bayes factor is not an e-value in general. However, in the case where $\Theta_0 = \{\theta_0\}$, if

$$E(X) := \frac{p_{\Pi_1}(x)}{p_{\theta_0}(x)},$$

then

$$\mathbb{E}_{\theta_0}[E(X)] = \int_{\mathcal{X}} p_{\Pi_1}(x)\, d\mu(x) = 1.$$

Hence the Bayes factor $E(X)$ is an e-value. The safe test rejects H_0 if $E(X) \geq 1/\alpha$.

Note that the safe test is not the Neyman–Pearson likelihood-ratio test in general. Recall that the likelihood-ratio test rejects H_0 if $\mathbb{E}(X) \geq \tau$, where τ is defined so that

$$\mathbb{P}_{\theta_0}\{E(X) \geq \tau\} = \alpha.$$

The safe test is, again, more conservative typically. Consider the following examples.

Simple Gaussian mean testing Suppose that we observe i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$. Consider testing $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1 > 0$. Let $X = (X_1, \ldots, X_n)$. The e-variable is the likelihood ratio

$$E(X) = \prod_{i=1}^{n} \exp\Big(\theta_1 X_i - \frac{\theta_1^2}{2}\Big).$$

Fix $\alpha = 0.05$. Then the safe test rejects H_0 if

$$E(X) \geq 20 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} \Big(\theta_1 X_i - \frac{\theta_1^2}{2}\Big) \geq \log 20 \approx 3 \quad \Longleftrightarrow \quad \bar{X} \geq \frac{\theta_1}{2} + \frac{\log 20}{\theta_1}.$$

This is much more conservative than the likelihood-ratio test, i.e., the Z-test in this case, which rejects H_0 if

$$\bar{X} \geq \frac{z_{0.05}}{\sqrt{n}}, \qquad z_{0.05} \approx 1.64.$$

Composite Gaussian mean testing Continuing with the above setting, suppose that we now have $H_1 : \theta \in \mathbb{R} \setminus \{0\}$ with a prior $\Pi_1 = \mathcal{N}(0, 1)$. The Bayes factor is given by

$$E(x) = \frac{\int_{\mathbb{R}} p_\theta(x) p_0(\theta)\, d\theta}{p_0(x)},$$

where $p_\theta(x)$ denotes the density of $\mathcal{N}(\theta, 1)$. It is not hard to compute

$$\log E(X) = -\frac{\log(n+1)}{2} + \frac{n^2}{2(n+1)} \bar{X}^2.$$

The safe test rejects H_0 if

$$|\bar{X}| \geq \sqrt{\frac{2(n+1)}{n^2}\Big(\frac{\log(n+1)}{2} + \log 20\Big)} \approx \sqrt{\frac{6 + \log n}{n}}$$

for n large. This is slightly more conservative than the Z-test, which rejects H_0 if

$$|\bar{X}| \geq \frac{z_{0.025}}{\sqrt{n}}, \qquad z_{0.025} \approx 1.96.$$

While e-values are more conservative, they come with several statistical advantages:

- It is easier to compute e-values than p-values for high-dimensional problems with more complicated models;
- e-values allow us to perform sequential inference as we will see in the next section;
- e-values are more robust to model misspecification than p-values;
- e-values rely on expectations, which are robust to data dependence, whereas p-values rely on tail probabilities, which are not.

5.4.3 Composite null

In the above Bayesian setup, suppose that we have a composite null $H_0 : \theta \in \Theta_0$ where $|\Theta_0| > 1$. Recall that $\frac{p_{\Pi_1}(x)}{p_{\theta_0}(x)}$ is an e-variable for any parameter θ_0, but the Bayes factor $E(X) = \frac{p_{\Pi_1}(x)}{p_{\Pi_0}(x)}$ is not necessarily an e-variable. To obtain an e-variable in this case, we consider the reverse information projection

$$\theta_0^* := \operatorname*{argmin}_{\theta \in \Theta_0} \mathsf{KL}(p_{\Pi_1}, p_\theta).$$

(Recall that the KL divergence is not symmetric and the information project is defined with respect to the first argument.) An e-variable achieving

$$\sup_E \mathbb{E}_{p_{\Pi_1}}[\log E] \quad \text{s.t.} \quad \sup_{\theta \in \Theta_0} \mathbb{E}_{p_\theta}[E] \leq 1$$

is said to be optimal relative to p_{Π_1}.

Theorem 5.7. *Suppose that the above reverse information projection θ_0^* exists. Then*

$$E(X) := \frac{p_{\Pi_1}(X)}{p_{\theta_0^*}(X)}$$

is an e-variable. Furthermore, it is optimal relative to p_{Π_1}.

See [GdHK20]. Maximizing $\log E$ has an advantage over maximizing E: It avoids E taking values close to 0 since $\log(\cdot)$ tends to $-\infty$ near 0. This is important because, as we will see in the next section, we often take the product of multiple e-values.

5.5 Applications of e-values

5.5.1 Optional continuation with e-values

Suppose that data (X_1, Z_1), (X_2, Z_2), ... are collected sequentially. For example, X_i may denote the outcome of an experiment and Z_i may denote the cost of the experiment. We will compute an e-value once in a while, after obtaining a batch of data. Let the batch sizes be $n_1, n_2, \ldots$, and let $N_t := \sum_{i=1}^t n_i$ for any integer $t \geq 0$. Let E_i denote an e-value computed after obtaining the i-th batch of data, i.e., up to observing (X_{N_i}, Z_{N_i}), such that

$$\mathbb{E}\left[E_i \mid (X_1, Z_1), \ldots, (X_{N_{i-1}}, Z_{N_{i-1}})\right] \leq 1.$$

We define

$$V_t := \prod_{i=1}^{t} E_i, \qquad t \geq 0.$$

For any stopping rule (that may depend on the data up to the present), let τ be the number of batches collected when we stop. We report the final result V_τ.

Proposition 5.8. *The discrete-time stochastic process $\{V_t\}_{t=0}^{\infty}$ is a nonnegative supermartingale. Moreover, for any stopping time τ, we have $\mathbb{E}[V_\tau] \leq 1$ so that V_τ is an e-value.*

Proof. For any integer $t \geq 0$, let $\mathcal{F}_t$ denote the filtration at time t. Then we have

$$\mathbb{E}[V_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[V_{t-1}E_t \mid \mathcal{F}_{t-1}] = V_{t-1}\,\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \leq V_{t-1},$$

since E_t is an e-value computed for the t-th batch of data. This says that V_t is a supermartingale by definition. By the optional stopping theorem, we obtain $\mathbb{E}[V_\tau] \leq \mathbb{E}[V_0] = 1$. □

Corollary 5.9. *(Ville's inequality) For any $\alpha \in (0, 1)$, we have*

$$\mathbb{P}\left\{\sup_{t\geq 0} V_t \geq 1/\alpha\right\} \leq \alpha.$$

Proof. Define a stopping time $\tau := \inf\{t \geq 0 : V_t \geq 1/\alpha\}$. Then we have $\sup_{t\geq 0} V_t \geq 1/\alpha$ if and only if $V_\tau \geq 1/\alpha$. By Markov's inequality and the optional stopping theorem,

$$\mathbb{P}\{V_\tau \geq 1/\alpha\} \leq \alpha\,\mathbb{E}[V_\tau] \leq \alpha\,\mathbb{E}[V_0] = \alpha.$$

□

In summary, regardless of how E_i depends on the past and what stopping rule we use, the e-value V_τ gives a test at significance level α, i.e., the safe test that rejects H_0 if $V_\tau \geq 1/\alpha$. We now consider an example.

Multi-armed bandit Suppose that there are K arms. The reward returned by arm k at time i, if pulled, is denoted by $X_{k,i}$. We employ strategy $(k_i)_{i\geq 1}$ which pulls arm k_i at time $i \geq 1$. This strategy gives independent rewards $X_{k_i,i}$ for $i \geq 0$. The goal is to quickly detect arms with means greater than 1 to maximize the profit. For $k \in [K]$, the null hypothesis H_0 is that arm k has mean reward at most 1. The running reward for arm k at time t is

$$M_{k,t} = \prod_{i\in[t]:k_i=k} X_{k,i}.$$

Since the strategy may depend on past outcomes, the process $M_{k,t}$ can be very complicated. However, we still have a valid e-value $M_{k,\tau}$ for any stopping time τ and any $k \in [K]$. We can reject the null if $M_{k,t} \geq 1/\alpha$.

5.5.2 FDR control with e-values

Suppose that we are interested in testing multiple null hypotheses $H_1, \ldots, H_n$ and have obtained respective e-values $e_1, \ldots, e_n$. Denote the order statistics as $e_{(1)} \geq \cdots \geq e_{(n)}$. To control the FDR at level $\alpha \in (0, 1)$, the e-Benjamini–Hochberg method rejects hypotheses with the largest $\hat{k}$ e-values, where

$$\hat{k} := \max\left\{i \in [n] : \frac{i\, e_{(i)}}{n} \geq \frac{1}{\alpha}\right\}.$$

This is essentially the same as the Benjamini–Hochberg method, except that p_i is replaced by $1/e_i$.

Theorem 5.10. *The e-Benjamini–Hochberg method achieves an FDR at most $\alpha\, n_0/n \leq \alpha$, where n_0 is the number of true null hypotheses.*

Proof. As before, we let R denote the number of hypotheses that are rejected, and let V denote the number of null hypotheses that are rejected. The FDP is equal to

$$\frac{V}{R \vee 1} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R \vee 1},$$

where $\mathcal{H}_0$ is the set of indices of null hypotheses and V_i is the indicator of the event that the i-th hypothesis is rejected. Suppose that $R \geq 1$ without loss of generality. For any $H_{(i)}$ rejected (which corresponds to $e_{(i)}$), we have

$$\frac{1}{R} \leq \frac{1}{i} \leq \frac{\alpha\, e_{(i)}}{n}.$$

It follows that

$$\frac{V}{R} \leq \sum_{i \in \mathcal{H}_0} \frac{V_{(i)}\, \alpha\, e_{(i)}}{n} \leq \sum_{i \in \mathcal{H}_0} \frac{\alpha\, e_{(i)}}{n} = \frac{\alpha}{n} \sum_{i \in \mathcal{H}_0} e_{(i)}.$$

Since each $e_{(i)}$ is an e-value so that $\mathbb{E}[e_{(i)}] \leq 1$, we have

$$\mathbb{E}\left[\frac{V}{R}\right] \leq \frac{\alpha}{n} \sum_{i \in \mathcal{H}_0} \mathbb{E}[e_{(i)}] \leq \frac{\alpha}{n}\, n_0.$$

□

Note that the proof does not require independence of the e-values. Similar to the case of simple hypothesis testing, using e-values is safer than using p-values but may not be as powerful.

5.6 Conformal inference

5.6.1 Prediction interval

Suppose that data (X, Y) follows a joint distribution $\mathcal{P}$, where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. For $\alpha \in (0, 1)$, a $(1 - \alpha)$ prediction interval for Y is a set $\mathcal{C}(X)$ such that $\mathbb{P}\{Y \in \mathcal{C}(X)\} \geq 1 - \alpha$. Typically, we observe a training set of data $\{(X_i, Y_i)\}_{i=1}^n$ and use it to construct a prediction interval $\mathcal{C}(X_{n+1})$ for Y_{n+1} for a test point (X_{n+1}, Y_{n+1}) not in the training set.

For a regression function $\mu(x;\theta)$ and a regularizer $\mathcal{R}(\theta)$, we can consider the regularized least squares fit

$$\hat{\mu}(x) := \mu(x;\hat{\theta}), \qquad \hat{\theta} := \operatorname*{argmin}_{\theta} \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \mu(X_i;\theta)\big)^2 + \mathcal{R}(\theta).$$

This gives a prediction $\hat{\mu}(X_{n+1})$ of Y_{n+1}.

To obtain a prediction interval, let us consider a potential procedure:

1. Fit any regression model to obtain $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ using the training set $\{(X_i, Y_i)\}_{i=1}^n$.
2. For $i \in [n]$, set $R_i := |Y_i - \hat{\mu}(X_i)|$.
3. Let Δ be the $\lceil (n+1)(1-\alpha) \rceil$-th smallest value of $\{R_i\}_{i=1}^n$.
4. Define the prediction interval to be $\mathcal{C}(X_{n+1}) := [\hat{\mu}(X_{n+1}) - \Delta, \hat{\mu}(X_{n+1}) + \Delta]$.

Unfortunately, this simple method does not give a valid prediction interval due to dependence issue.

5.6.2 Split conformal

To fix the above method, we can split the data into a training set (in-sample data) and a hold-out set (out-of-sample data). Then the following method, which we refer to as the split conformal method, can be applied:

1. Partition $[n]$ into two disjoint sets $\mathcal{I}_1$ and $\mathcal{I}_2$ of sizes n_1 and n_2 respectively.
2. Fit any regression model to obtain $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ using the set $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$.
3. For $i \in \mathcal{I}_2$, set $R_i := |Y_i - \hat{\mu}(X_i)|$.
4. Let Δ be the $\lceil (n_2+1)(1-\alpha) \rceil$-th smallest value of $\{R_i\}_{i \in \mathcal{I}_2}$.
5. Define the prediction interval to be $\mathcal{C}(X_{n+1}) := [\hat{\mu}(X_{n+1}) - \Delta, \hat{\mu}(X_{n+1}) + \Delta]$.

A set of random variables $\{Z_i\}_{i=1}^n$ is called exchangeable if for any permutation $\pi : [n] \to [n]$, we have

$$(Z_1, \ldots, Z_n) \overset{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(n)}).$$

If the random variables are i.i.d., then they are obviously exchangeable. Moreover, if Z is a random variable and ε_i are i.i.d. random variables, then $Z_i := Z + \varepsilon_i$ are exchangeable, where $i \in [n]$.

Theorem 5.11. *If the data points $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, then the interval constructed by the split conformal method is a valid prediction interval, i.e., $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$.*

A conformity score $s(z) \in \mathbb{R}$ is a quantity that measures how much z corresponds to previous observations. For example, given $Z_1, \ldots, Z_{n+1}$, a high score $s(Z_{n+1})$ means that Z_{n+1} conforms to $Z_1, \ldots, Z_n$, while a low score means the opposite.

Proposition 5.12. *Let $Z_1, \ldots, Z_{n+1}$ be exchangeable observations. Suppose that we have conformity scores $s(Z_1), \ldots, s(Z_{n+1})$ which are distinct almost surely. Then the conformal p-value*

$$p(Z_{n+1}) := \frac{|\{i \in [n] : s(Z_i) \leq s(Z_{n+1})\}| + 1}{n+1}$$

is uniform over $\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, 1\}$. As a result, $p(Z_{n+1})$ is a valid p-value.

Proof. Since $Z_1, \ldots, Z_{n+1}$ are exchangeable, so are the conformity scores $s(Z_1), \ldots, s(Z_{n+1})$. This implies that the rank of $s(Z_{n+1})$ is uniform over $[n+1]$, so the result follows. □

Proof of Theorem 5.11. Let $Z_i = (X_i, Y_i)$ and define the conformity score

$$s(Z_i) = -|Y_i - \hat{\mu}(X_i)|.$$

Note that implicitly $\hat{\mu}$ and s are functions of $\{Z_i\}_{i \in \mathcal{I}_1}$. It is not hard to see that $\{s(Z_i)\}_{i \in \mathcal{I}_2 \cup \{n+1\}}$ are exchangeable. Then setting

$$p(Z_{n+1}) := \frac{|\{i \in \mathcal{I}_2 : s(Z_i) \le s(Z_{n+1})\}| + 1}{n_2 + 1}$$

gives the conformal p-value. We have

$$Y_{n+1} \notin \mathcal{C}(X_{n+1}) \quad \iff \quad R_{n+1} := |Y_{n+1} - \hat{\mu}(X_{n+1})| > \Delta,$$

i.e., R_{n+1} is larger than the $\lceil (n_2+1)(1-\alpha) \rceil$-th smallest value of $\{R_i\}_{i \in \mathcal{I}_2}$. In other words, $s(Z_{n+1})$ is smaller than the $\lceil (n_2+1)(1-\alpha) \rceil$-th largest value of $\{s(Z_i)\}_{i \in \mathcal{I}_2}$. It follows that

$$p(Z_{n+1}) \le \frac{n_2 - \lceil (n_2+1)(1-\alpha) \rceil + 1}{n_2 + 1} \le 1 - (1-\alpha) = \alpha.$$

We conclude that

$$\mathbb{P}\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\} \le \mathbb{P}\{p(Z_{n+1}) \le \alpha\} \le \alpha$$

by Proposition 5.12. □

5.6.3 Quantile regression

Before presenting an improved method, let us first introduce a way to produce a quantile estimate. Consider a continuous random variable Y with CDF F and a quantile $q_\alpha := F^{-1}(\alpha)$ for $\alpha \in (0,1)$. Define the pinball loss

$$\rho_\alpha(z) := z\big(\alpha - \mathbb{1}\{z < 0\}\big) = \begin{cases} \alpha z & \text{if } z \ge 0, \\ (1-\alpha)(-z) & \text{if } z < 0. \end{cases}$$

We claim that

$$q_\alpha = \operatorname*{argmin}_u \mathbb{E}[\rho_\alpha(Y-u)] = \operatorname*{argmin}_u \left((1-\alpha) \int_{-\infty}^{u} (u-y)\, dF(y) + \alpha \int_u^{\infty} (y-u)\, dF(y) \right).$$

To see this, it suffices to check the first-order condition

$$(1-\alpha) \int_{-\infty}^{u} dF(y) - \alpha \int_u^{\infty} dF(y) = (1-\alpha)F(u) - \alpha(1-F(u)) = F(u) - \alpha,$$

which yields $u = F^{-1}(\alpha) = q_\alpha$. Thus, minimizing the expected pinball loss recovers the quantile.

Motivated by this fact, we consider the following estimator of a quantile in a regression model. For a quantile regression function $f(x;\theta)$ and a regularizer $\mathcal{R}(\theta)$, define

$$\hat{q}_\alpha(x) := f(x;\hat{\theta}), \qquad \hat{\theta} := \operatorname*{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha\big(Y_i - f(X_i,\theta)\big) + \mathcal{R}(\theta).$$

5.6.4 Conformal quantile regression

A more recently proposed method, conformal quantile regression [RPC19], improves on the split conformal method. The algorithm is as follows:

1. Partition $[n]$ into two disjoint sets $\mathcal{I}_1$ and $\mathcal{I}_2$ of sizes n_1 and n_2 respectively.
2. Apply any quantile regression method to obtain lower and upper quantiles $\hat{q}_{\alpha_l}$ and $\hat{q}_{\alpha_u}$ using the set $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1}$.
3. For $i \in \mathcal{I}_2$, set $E_i := \max\{\hat{q}_{\alpha_l}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_u}(X_i)\}$.
4. Let Δ be the $\lceil (n_2+1)(1-\alpha)\rceil$-th smallest value of $\{E_i\}_{i\in\mathcal{I}_2}$.
5. Define the prediction interval to be $\mathcal{C}(X_{n+1}) := [\hat{q}_{\alpha_l}(X_{n+1}) - \Delta, \hat{q}_{\alpha_u}(X_{n+1}) + \Delta]$.

The second sample $\{(X_i, Y_i)\}_{i\in\mathcal{I}_2}$ is called the calibration set, which is used to conformalize the prediction interval obtained from the training set $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1}$.

Theorem 5.13. *If the data points $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, then the interval constructed by conformal quantile regression is a valid prediction interval, i.e., $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1-\alpha$.*

Proof. Let

$$E_{n+1} := \max\{\hat{q}_{\alpha_l}(X_{n+1}) - Y_{n+1}, Y_{n+1} - \hat{q}_{\alpha_u}(X_{n+1})\}.$$

It is not hard to see that $\{E_i\}_{i\in\mathcal{I}_2\cup\{n+1\}}$ are exchangeable. By the construction of the prediction interval, we have $Y_{n+1} \notin \mathcal{C}(X_{n+1})$ if and only if $E_{n+1} > \Delta$. The latter condition means that E_{n+1} is larger than the $\lceil (n_2+1)(1-\alpha)\rceil$-th smallest value of $\{E_i\}_{i\in\mathcal{I}_2}$. This happens with probability at most α by virtue of exchangeability (in a way similar to the proof of Theorem 5.11). □

Moreover, if the non-conformity scores E_i are distinct almost surely, then the prediction interval is nearly perfectly calibrated in the sense that

$$1-\alpha \leq \mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \leq 1-\alpha+\frac{1}{n_2+1}.$$

Conformal quantile regression produces adaptive intervals, while the split conformal method produces intervals of constant width. As a result, the intervals produced by conformal quantile regression are typically narrower and have better conditional coverage.

Chapter 6

Testing in networks

6.1 Detection of a planted clique in a graph

6.1.1 The planted clique model

Consider an undirected graph on n vertices. Let the vertex set be denoted by $[n] = \{1, \dots, n\}$ and the adjacency matrix be denoted by $A \in \{0,1\}^{n \times n}$. We identify the adjacency matrix A with the graph itself. We say that A is an Erdős–Rényi graph with edge density $1/2$ and write $A \sim G(n, 1/2)$, if the edges $(A_{ij})_{i<j}$ are independent $\mathsf{Ber}(1/2)$ random variables.

A clique is a complete subgraph in a graph. In other words, the induced subgraph of A with vertex set $K \subset [n]$ is a clique if $A_{ij} = 1$ for any distinct $i, j \in K$. The planted clique model $G(n, 1/2, k)$ can be described as follows: In an Erdős–Rényi graph, take a uniformly random subset $K \subset [n]$ of size k and replace the subgraph with vertex set K with a clique. As a result, we obtain a graph A with

$$A_{ij} = \begin{cases} 1 & \text{if } i, j \in K, \\ \mathsf{Ber}(1/2) & \text{otherwise,} \end{cases}$$

where the random edges are independent $\mathsf{Ber}(1/2)$ variables. In particular, the Erdős–Rényi model $G(n, 1/2)$ is equivalent to the planted clique model $G(n, 1/2, 0)$ with no clique.

Detection of a planted clique refers to the problem of determining whether the observed graph contains a planted clique of size k. In the language of hypothesis testing, we test the null hypothesis $H_0 : A \sim G(n, 1/2, 0)$ against the alternative hypothesis $H_1 : A \sim G(n, 1/2, k)$. Let us consider the asymptotic regime where $n \to \infty$ and use $o(1)$ to denote a vanishing quantity. The difficulty of this detection problem is clearly related to the size k of the planted clique:

- If $k = 0$ or k is too small, then it is impossible to distinguish H_1 from H_0.
- If $k = n$ or k is sufficiently large, then it is easy to distinguish H_1 from H_0.
- What is the threshold k above which we can distinguish H_1 from H_0 with probability $1 - o(1)$ given infinite computational power? We call this threshold the statistical (or information-theoretic) threshold.
- What is the threshold k above which we can distinguish H_1 from H_0 with probability $1 - o(1)$ using a polynomial-time algorithm? We call this threshold the computational threshold.
- Are the above two thresholds the same?

6.1.2 Statistical threshold

Under H_0, each edge is present with probability $1/2$ independently in the graph A. There are 45 possible edges of A between vertices $1, \dots, 10$, so the induced subgraph of A on the vertex set $[10]$ is a clique with probability 2^{-45}. As $n \to \infty$, there are infinitely many groups of 10 vertices in A, so A contains a clique of size 10 with probability $1 - o(1)$. Therefore, if under H_1 a clique of size $k = 10$ is planted in addition, there is no significant difference between H_0 and H_1.

This motivates us to study the clique number $\omega(A)$ which is defined to be the size of the largest clique in A. If $\omega(A)$ is bounded by k_0 with probability $1 - o(1)$ under H_0 and k is larger than k_0, then detection of the planted clique of size k is possible under H_1.

Theorem 6.1. *Let $A \sim G(n, 1/2)$. For any constant $\varepsilon > 0$, we have*

$$\mathbb{P}\{\omega(A) \le (2+\varepsilon)\log_2 n\} \to 1 \quad \text{as } n \to \infty.$$

Proof. For any fixed subset $S \subset [n]$ of size k, it holds that

$$\mathbb{P}\{A_{ij} = 1 \text{ for all distinct } i, j \in S\} = 2^{-\binom{k}{2}}.$$

Since there are $\binom{n}{k}$ subsets of $[n]$ of size k, we obtain

$$\mathbb{P}\{\omega(A) \ge k\} \le \mathbb{P}\{A \text{ contains a clique of size } k\} \le \binom{n}{k} \cdot 2^{-\binom{k}{2}} \le n^k \, 2^{-\frac{k(k-1)}{2}}.$$

For $k := \lfloor (2+\varepsilon)\log_2 n \rfloor$, we have

$$\log_2(n^k 2^{-\frac{k(k-1)}{2}}) = k \log_2 n - k(k-1)/2 \to -\infty$$

as $n \to \infty$, so the conclusion holds. □

This result immediately implies that there is a consistent test for distinguishing H_1 from H_0.

Corollary 6.2. *Suppose that $k > (2+\varepsilon)\log_2 n$ for a constant $\varepsilon > 0$. Then*

$$\mathbb{P}_0\{\omega(A) > (2+\varepsilon)\log_2 n\} + \mathbb{P}_1\{\omega(A) \le (2+\varepsilon)\log_2 n\} \to 0 \quad \text{as } n \to \infty.$$

In other words, the test that rejects H_0 if and only if $\omega(A) > (2+\varepsilon)\log_2 n$ achieves vanishing type I and type II errors.

Later we will show that if $k \le (2-\varepsilon)\log_2 n$ for a constant $\varepsilon > 0$, then there is no test that achieves vanishing type I and type II errors. Thus the statistical threshold for planted clique detection is tightly characterized.

6.2 Spectral methods

Let us continue considering the planted clique model. The issue with the above test based on the clique number $\omega(A)$ is that it cannot be efficiently computed: In general, finding the largest clique in A entails an exhaustive search which takes exponential time. We now consider an efficient spectral method that succeeds in detecting the planted clique for much larger k.

Define an affine transform W of the adjacency matrix A by

$$W_{ij} := \begin{cases} 2A_{ij} - 1 & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

Then under H_0, the entries $(W_{ij})_{i<j}$ are i.i.d. Rademacher variables, while under H_1, for $i < j$,

$$W_{ij} = \begin{cases} 1 & \text{if } i, j \in K, \\ \mathsf{Rademacher} & \text{otherwise,} \end{cases}$$

where $K \subset [n]$ denotes the vertex set of the planted clique. We will use the spectral norm $\|W\|$ (i.e., the largest singular value of W) as the test statistic.

6.2.1 Spectral norm of the noise

We first assume H_0 and study $\|W\|$.

Theorem 6.3. *Under H_0, there is an absolute constant $C > 0$ such that*

$$\mathbb{P}\{\|W\| \leq C\sqrt{n}\} \geq 1 - \exp(-n) \to 1 \qquad \text{as } n \to \infty.$$

We say that $\mathcal{N} \subset \mathcal{B}_n$ is an ε-net of the unit ball $\mathcal{B}_n := \{u \in \mathbb{R}^n : \|u\|_2 \leq 1\}$, if for any $u \in \mathcal{B}_n$, there exists $v \in \mathcal{N}$ such that $\|u - v\|_2 \leq \varepsilon$.

Lemma 6.4. *If $\mathcal{N}$ is a ε-net of $\mathcal{B}_n$, then*

$$\|W\| \leq \frac{1}{1 - 2\varepsilon} \max_{v \in \mathcal{N}} v^\top W v.$$

Proof. There exists $u \in \mathcal{B}_n$ such that $\|W\| = u^\top W u$. Choose $v \in \mathcal{N}$ such that $\|u - v\|_2 \leq \varepsilon$. Then we have

$$\|W\| = u^\top W u = v^\top W v + (u - v)^\top W v + u^\top W (u - v) \leq v^\top W v + \varepsilon \|W\| + \varepsilon \|W\|.$$

Rearranging this inequality finishes the proof. □

Lemma 6.5. *For $\varepsilon \in (0, 1)$, there exists an ε-net $\mathcal{N}$ of $\mathcal{B}_n$ that has cardinality $|\mathcal{B}| \leq (1 + 2/\varepsilon)^n$.*

Proof. We successively pick points in $\mathcal{B}_n$ that are at least distance ε away from each other until we cannot do so; call this set of points $\mathcal{N}$. (The set $\mathcal{N}$ is called a maximal ε-packing of $\mathcal{B}_n$.) Note that for any other point $u \in \mathcal{B}_n$, there must exist $v \in \mathcal{N}$ such that $\|u - v\|_2 \leq \varepsilon$ because otherwise we can still add u in $\mathcal{N}$. Hence, by definition, $\mathcal{N}$ is an ε-net.

Next, consider the balls D_v of radius $\varepsilon/2$ centered at $v \in \mathcal{N}$. These balls are disjoint because different points in $\mathcal{N}$ are at least distance ε away from each other. Moreover, the union of all D_v for $v \in \mathcal{N}$ is contained in the ball A of radius $1 + \varepsilon/2$ centered at the origin. Consequently,

$$|\mathcal{N}| \leq \frac{\mathsf{Vol}(A)}{\mathsf{Vol}(D_v)} = \Big(\frac{1 + \varepsilon/2}{\varepsilon/2}\Big)^n = \big(1 + 2/\varepsilon\big)^n,$$

finishing the proof. □

Lemma 6.6. *Suppose that $X_1, \ldots, X_N$ are i.i.d. Rademacher random variables. For any fixed vector $z \in \mathbb{R}^N$, we have*

$$\mathbb{P}\{z^\top X > t\} \le \exp\Big(\frac{-t^2}{2\|z\|_2^2}\Big).$$

Proof. It is not hard to check

$$\mathbb{E}[e^{\gamma X_i}] = \cosh(\gamma) \le \exp(\gamma^2/2).$$

Then, by Chernoff's bound, we have

$$\begin{aligned}
\mathbb{P}\Big\{\sum_{i=1}^n z_i X_i \ge t\Big\} &= \mathbb{P}\Big\{\exp\Big(\lambda \sum_{i=1}^N z_i X_i\Big) \ge \exp(\lambda t)\Big\} \\
&\le \exp(-\lambda t)\ \mathbb{E}\Big[\exp\Big(\lambda \sum_{i=1}^N z_i X_i\Big)\Big] \\
&= \exp(-\lambda t) \prod_{i=1}^N \exp(\lambda^2 z_i^2/2).
\end{aligned}$$

Choosing $\lambda = t/\|z\|_2^2$ yields the conclusion. □

Proof of Theorem 6.3. For a fixed vector $v \in \mathcal{B}_n$, we have

$$v^\top W v = \sum_{i,j=1}^n W_{ij} v_i v_j = 2 \sum_{i<j} W_{ij} v_i v_j.$$

By Lemma 6.6 and $\sum_{i<j} v_i^2 v_j^2 \le \frac{1}{2}\sum_{i,j=1}^n v_i^2 v_j^2 \le \frac{1}{2}$, it holds

$$\mathbb{P}\{v^\top W v > 2t\} \le \exp(-t^2).$$

Then Lemmas 6.4 and 6.5 with $\varepsilon = 1/4$ imply that

$$\mathbb{P}\{\|W\| > 4t\} \le \mathbb{P}\Big\{\max_{v \in \mathcal{N}} v^\top W v > 2t\Big\} \le 9^n \exp(-t^2).$$

Choosing $t = C\sqrt{n}$ for a large constant $C > 0$ completes the proof. □

6.2.2 The spectral test

As a result of Theorem 6.3, if $\|W\|$ is larger than $C\sqrt{n}$ under H_1, then we can distinguish H_1 from H_0. Too see how large $\|W\|$ is under H_1, consider its expectation $\mathbb{E}[W]$ specified by

$$\mathbb{E}[W_{ij}] = \begin{cases} 1 & \text{if } i,j \in K,\ i \ne j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\xi := \frac{1}{\sqrt{k}} \mathbf{1}_K \in \{0,1\}^n$ be the unit vector defined by $\xi_i = \frac{1}{\sqrt{k}}$ if $i \in K$ and $\xi_i = 0$ if $i \notin K$. Then $\mathbb{E}[W]$ can by obtained from $k\,\xi\xi^\top$ by replacing its diagonal entries with zeros. Consequently,

$$\|\,\mathbb{E}[W]\| \ge \|k\,\xi\xi^\top\| - \|\,\mathbb{E}[W] - k\,\xi\xi^\top\| = k - 1.$$

Moreover, the matrix $W - \mathbb{E}[W]$ has Rademacher entries except that $W_{ij} - \mathbb{E}[W_{ij}] = 0$ if $i, j \in K$ or $i = j$. Similar to Theorem 6.3, we have

$$\mathbb{P}\{\|W - \mathbb{E}[W]\| \le C\sqrt{n}\} \to 1 \quad \text{as } n \to \infty$$

for an absolute constant $C > 0$ under H_1. By the triangle inequality

$$\|W\| \ge \|\,\mathbb{E}[W]\| - \|W - \mathbb{E}[W]\|,$$

we obtain the following theorem.

Theorem 6.7. *Under H_1, there is an absolute constant $C > 0$ such that if $k > 2\,C\sqrt{n} + 1$, then*

$$\mathbb{P}\{\|W\| > C\sqrt{n}\} \to 1 \quad \text{as } n \to \infty.$$

Combining Theorems 6.3 and 6.7 immediately yields a consistent test.

Corollary 6.8. *There is an absolute constant $C > 0$ such that the following holds. Suppose that $k > 2\,C\sqrt{n} + 1$. Then*

$$\mathbb{P}_0\{\|W\| > C\sqrt{n}\} + \mathbb{P}_1\{\|W\| \le C\sqrt{n}\} \to 0 \quad \text{as } n \to \infty.$$

In other words, the test that rejects H_0 if $\|W\| > C\sqrt{n}$ is consistent.

It is widely conjectured that the computational threshold $C\sqrt{n}$ is in fact tight up to a constant factor even though it is significantly larger than the statistical threshold $2\log_2 n$. That is to say, if $k \le c\sqrt{n}$ for a certain absolute constant $c > 0$, then there may be no polynomial-time algorithm that can distinguish H_1 from H_0 with vanishing type I and type II errors. If this is indeed the case, we say that there is a statistical-to-computational gap for planted clique detection.

6.3 Lower bounds for testing in random graphs

In this section, we introduce tools for proving statistical and computational lower bounds for testing problems involving random graphs.

6.3.1 Fourier basis of functions on a random graph

Define a standardized version of the adjacency matrix $\bar{A} \in \mathbb{R}^{n \times n}$ by

$$\bar{A}_{ij} := 2A_{ij} - 1$$

for $i, j \in [n]$. For $A \sim G(n, 1/2)$, we have

$$\mathbb{E}[\bar{A}_{ij}] = 0, \qquad \mathsf{Var}(\bar{A}_{ij}) = 1.$$

We now define a Fourier basis of functions of $(A_{ij})_{i<j}$ (or, equivalently, functions of $(\bar{A}_{ij})_{i<j}$). Let $S \subset \binom{[n]}{2}$ denote an edge set; S can also be viewed as the subgraph of the complete graph K_n induced by this edge set. Next, define

$$\phi_S(A) := \prod_{(i,j)\in S} \bar{A}_{ij}.$$

Proposition 6.9. *The set $\{\phi_S : S \subset \binom{[n]}{2}\}$ forms an orthonormal basis of the set of real-valued functions of $(A_{ij})_{i<j}$ with respect to the inner product $\langle f, g\rangle := \mathbb{E}[f(A)\, g(A)]$, where $A \sim G(n, 1/2)$.*

Furthermore, let $V_{\le D}$ denote the set of polynomials in $(A_{ij})_{i<j}$ that have degrees at most D. Then $\{\phi_S : S \subset \binom{[n]}{2}, |S| \le D\}$ forms an orthonormal basis of $V_{\le D}$.

Proof. It is not hard to check that

$$\mathbb{E}[\phi_S(A)\,\phi_T(A)] = \mathbb{E}\left[\prod_{(i,j)\in S} \bar{A}_{ij} \cdot \prod_{(i',j')\in T} \bar{A}_{i'j'}\right] = \begin{cases} 1 & \text{if } S = T, \\ 0 & \text{if } S \neq T. \end{cases}$$

Moreover, there are $2^{\binom{n}{2}}$ possible values that the graph $(A_{ij})_{i<j}$ can take, so the set of real-valued functions of A has dimension $\binom{n}{2}$. There are precisely $\binom{n}{2}$ functions in the defined basis, so this proves the first claim. The second claim holds by the orthogonality and the fact that the set $\{\phi_S : S \subset \binom{[n]}{2}, |S| \le D\}$ spans $V_{\le D}$. □

In the sequel, we also use the norm $\|\cdot\|$ associated with the above inner product $\langle\cdot,\cdot\rangle$.

6.3.2 Statistical lower bounds

Consider testing between two distributions $\mathbb{P}_0 = G(n, 1/2)$ and $\mathbb{P}_1$ of random graphs (we will set $\mathbb{P}_1 = G(n, 1/2, k)$ later). Recall that whether there exists a consistent test depends on the total variation distance between $\mathbb{P}_0$ and $\mathbb{P}_1$: For any fixed adjacency matrix $B \in \{0,1\}^{n\times n}$, let $p_r(B)$ be the probability that $\mathbb{P}_r$ generates B where $r = 0, 1$. Then we have

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) = \frac{1}{2}\sum_B |p_0(B) - p_1(B)| = 1 - \sum_B \min\big\{p_0(B), p_1(B)\big\}$$

where the sum is over all possible adjacency matrices B, and

$$\inf_\phi \big(\mathbb{P}_0\{\phi = 1\} + \mathbb{P}_1\{\phi = 0\}\big) = 1 - \mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1)$$

where the infimum is taken over all possible tests ϕ.

Proposition 6.10. *Let $L(A) := \frac{p_1(A)}{p_0(A)}$ be the likelihood ratio, and consider its norm $\|L\|$ defined by*

$$\|L\|^2 := \mathbb{E}_0\left[\Big(\frac{p_1(A)}{p_0(A)}\Big)^2\right].$$

If $\|L\| \le C$ for a constant $C > 0$, then there exists $c > 0$ such that $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \le 1 - c$. As a result, there exists no consistent test that distinguishes $\mathbb{P}_1$ from $\mathbb{P}_0$ as $n \to \infty$.

Proof. For brevity, we write sums as integrals. Then $1 - \mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) = \int \min(p_0, p_1)$. We have

$$\begin{aligned}\Big(\int \sqrt{p_0 p_1}\Big)^2 &= \Big(\int \sqrt{\min(p_0, p_1)\cdot\max(p_0, p_1)}\Big)^2 \\ &\le \int \min(p_0, p_1) \cdot \int \max(p_0, p_1) \\ &\le \int \min(p_0, p_1) \cdot \int (p_0 + p_1) = 2\int \min(p_0, p_1),\end{aligned}$$

where we used the Cauchy–Schwarz inequality. Moreover,

$$\Big(\int \sqrt{p_0 p_1}\Big)^2 = \exp\Big(2\log\int\sqrt{p_0p_1}\Big) = \exp\Big(2\log\int_{p_0p_1>0} p_1\sqrt{\frac{p_0}{p_1}}\Big)$$
$$\geq \exp\Big(2\int_{p_0p_1>0} p_1\log\sqrt{\frac{p_0}{p_1}}\Big) = \exp\Big(-\int_{p_0p_1>0} p_1\log\frac{p_1}{p_0}\Big),$$

where we used Jensen's inequality. Applying Jensen's inequality again, we obtain

$$\int_{p_0p_1>0} p_1\log\frac{p_1}{p_0} \leq \log\int_{p_0p_1>0} p_1\frac{p_1}{p_0} = \log\int p_0\Big(\frac{p_1}{p_0}\Big)^2 = \log\|L\|^2.$$

Combining everything, if $\|L\|^2 \leq C$, then $\int\sqrt{p_0p_1} \geq c'$ and so $\int\min(p_0,p_1) \geq c$ for constants $c, c' > 0$. We conclude that $\mathsf{TV}(\mathbb{P}_0,\mathbb{P}_1) \leq 1-c$. □

6.3.3 Computational lower bounds

By the above result, to establish a statistical lower bound against all tests, it suffices to control the norm $\|L\|$. Next, we show that, if the goal is to establish a lower bound against polynomial tests of degrees at most D, it suffices to consider the norm of the projected likelihood ratio $\|L_{\leq D}\|$. To be more precise, recall that $V_{\leq D}$ denotes the set of polynomials in $(A_{ij})_{i<j}$ that have degrees at most D. Define the function $L_{\leq D}(A)$ to be the projection of the likelihood ratio $L(A)$ onto $V_{\leq D}$. Then we have

$$\|L_{\leq D}\| = \max_{f\in V_{\leq D},\,\|f\|\leq 1} \mathbb{E}_0[L(A)f(A)] = \max_{f\in V_{\leq D},\,\|f\|\leq 1} \mathbb{E}_1[f(A)].$$

We say that a polynomial $f(A)$ in the entries $(A_{ij})_{i<j}$ strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$ if

$$\sqrt{\max\big\{\mathsf{Var}_0(f(A)),\mathsf{Var}_1(f(A))\big\}} = o\Big(\Big|\,\mathbb{E}_1[f(A)] - \mathbb{E}_0[f(A)]\Big|\Big)$$

as $n\to\infty$; see [BAH+22]. Note that if the above condition holds, say, with $\mathbb{E}_1[f(A)] > \mathbb{E}_0[f(A)]$, then we can take $\tau := \frac{1}{2}\big(\mathbb{E}_1[f(A)] + \mathbb{E}_0[f(A)]\big)$, and by Chebyshev's inequality,

$$\mathbb{P}_0\{f(A) > \tau\} \leq \mathbb{P}_0\Big\{\big|f(A) - \mathbb{E}_0[f(A)]\big| > \frac{1}{2}\big(\mathbb{E}_1[f(A)] - \mathbb{E}_0[f(A)]\big)\Big\}$$
$$\leq \frac{4\,\mathsf{Var}_0(f(A))}{(\mathbb{E}_1[f(A)] - \mathbb{E}_0[f(A)])^2} = o(1).$$

Similarly, $\mathbb{P}_1\{f(A)\leq\tau\} = o(1)$, so the test that rejects $\mathbb{P}_0$ if $f(A) > \tau$ is consistent.

Proposition 6.11. *If $\|L_{\leq D}\| \leq C$ for a constant $C > 0$, then there exists no polynomial $f \in V_{\leq D}$ that strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$.*

Proof. Suppose there is a polynomial $f(A)$ that strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$. Without loss of generality, we can standardize $f(A)$ under $\mathbb{P}_0$. Then $\mathbb{E}_0[f(A)] = 0$, $\mathsf{Var}_0(f(A)) = \|f\|^2 = 1$, and $|E_1[f(A)]| \to \infty$ as $n\to\infty$ by the strong separation. This contradicts $\|L_{\leq D}\| \leq C$. □

6.4 Statistical-to-computational gap for detecting a planted clique

Using tools from the last section, we provide evidence supporting the conjectured statistical-to-computational gap for detection of a planted clique.

6.4.1 Low-degree polynomials

Let us first answer the question: To predict the computational threshold of a problem, what degree D should we consider? It is conjectured in the literature that taking $D = \mathrm{polylog}(n)$ will yield a good prediction. One rationale behind this conjecture is that polynomials of logarithmic degrees approximate spectral methods sufficiently well.

Proposition 6.12. *Let $\mathbb{P}_0 = G(n, 1/2)$ and $\mathbb{P}_1 = G(n, 1/2, k)$, where $k \geq 4\,C\sqrt{n}$ for a large constant $C > 0$. For $D = \lceil 10 \log_2 n \rceil$, there is a degree-$D$ polynomial $f(A)$ such that*

$$\mathbb{P}_0\{f(A) > \tau\} + \mathbb{P}_1\{f(A) \leq \tau\} \to 0 \quad \text{as } n \to \infty$$

for a threshold τ. Furthermore, $f(A)$ strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$.

Proof. Suppose that $A \sim \mathbb{P}_1 = G(n, 1/2, k)$. Recall that $\|\,\mathbb{E}[\bar{A}]\| \geq k - 1$ and $\|\bar{A} - \mathbb{E}[\bar{A}]\| \leq C\sqrt{n}$ for a constant $C > 0$ with high probability. We used the statistic $\|\bar{A}\|$ to distinguish between $\mathbb{P}_0$ and $\mathbb{P}_1$, but $\|\bar{A}\|$ is not a polynomial in the entries of A. Nevertheless, we introduce a degree-D polynomial that can be used as a test statistic in replace of $\|\bar{A}\|$.

Let $\lambda_i(M)$ denotes the ith largest eigenvalue of M. Consider the polynomial

$$f(A) := \mathsf{tr}(\bar{A}^D) = \sum_{i=1}^{n} \lambda_i(\bar{A})^D = \lambda_1(\bar{A})^D \Big[1 + \sum_{i=2}^{n} \Big(\frac{\lambda_i(\bar{A})}{\lambda_1(\bar{A})}\Big)^D\Big].$$

By Weyl's inequality,

$$\lambda_1(\bar{A}) \geq \lambda_1(\mathbb{E}[\bar{A}]) - \|\bar{A} - \mathbb{E}[\bar{A}]\| \geq k - 1 - C\sqrt{n} \geq 2.9\,C\sqrt{n}$$

and

$$|\lambda_i(\bar{A})| \leq |\lambda_i(\mathbb{E}[\bar{A}])| + \|\bar{A} - \mathbb{E}[\bar{A}]\| \leq 1 + C\sqrt{n} \leq 1.1\,C\sqrt{n}.$$

Therefore, $\lambda_1(\bar{A}) \geq 2|\lambda_i(\bar{A})|$ for any $i \geq 2$. If $D \geq 10 \log_2 n$, then $\big(\frac{|\lambda_i(\bar{A})|}{\lambda_1(\bar{A})}\big)^D \leq n^{-10}$ for $i \geq 2$. It follows that

$$f(A) \geq \lambda_1(\bar{A})^D (1 - n^{-9}) \geq (2\,C\sqrt{n}\,)^D.$$

On the other hand, suppose $A \sim G(n, 1/2)$. Then

$$f(A) = \sum_{i=1}^{n} \lambda_i(\bar{A})^D \leq n\|\bar{A}\|^D \leq n(C\sqrt{n}\,)^D,$$

since we showed that $\|\bar{A}\| \leq C\sqrt{n}$ with high probability. Note that $2^D \geq n^{10} > n$ for $D \geq 10 \log_2 n$, so the degree-D polynomial $f(A)$ gives a consistent test between $\mathbb{P}_0$ and $\mathbb{P}_1$.

We omit the proof of strong separation between $\mathbb{P}_0$ and $\mathbb{P}_1$ by $f(A)$. In fact, strong separation is only a second-moment condition, while we have already shown exponential tail bounds for $\lambda_1(\bar{A})$ and thus can control $f(A)$ under either $\mathbb{P}_0$ or $\mathbb{P}_1$. □

6.4.2 Establishing the lower bounds

In view of Propositions 6.10 and 6.11, to prove statistical and computational lower bounds for the planted clique problem, it suffices to bound $\|L\|$ and $\|L_{\le D}\|$ respectively. Recall that $V_{\le D}$ denotes the set of polynomials in $(A_{ij})_{i<j}$ that have degrees at most D, and $\{\phi_S : S \subset \binom{[n]}{2}, |S| \le D\}$ is an orthonormal basis of $V_{\le D}$. As a result,

$$\|L_{\le D}\|^2 = \sum_{S \subset \binom{[n]}{2}, |S| \le D} \langle L, \phi_S \rangle^2 = \sum_{S \subset \binom{[n]}{2}, |S| \le D} \mathbb{E}_0[L(A)\,\phi_S(A)]^2 = \sum_{S \subset \binom{[n]}{2}, |S| \le D} \mathbb{E}_1[\phi_S(A)]^2.$$

Moreover, the degree D is at most $\binom{n}{2}$, and $\{\phi_S : S \subset \binom{[n]}{2}\}$ is an orthonormal basis of the set of all functions of A. Hence we have $\|L\| = \|L_{\le \binom{n}{2}}\|$.

Theorem 6.13. *Let $\mathbb{P}_0 = G(n, 1/2)$ and $\mathbb{P}_1 = G(n, 1/2, k)$. Consider the likelihood ratio $L(A) = \frac{p_1(A)}{p_0(A)}$. Then we have the following results:*

- *If $k \le (2-\varepsilon)\log_2 n$ for $\varepsilon > 0$, then $\|L\|^2 \le 2$.*
- *If $k \le n^{1/2-\varepsilon}$ for $\varepsilon > 0$ and $D = o\big((\frac{\log n}{\log\log n})^2\big)$, then $\|L_{\le D}\|^2 \le 4$.*

Proof. Since $\|L_{\le D}\|^2 = \sum_{S \subset \binom{[n]}{2}, |S| \le D} \mathbb{E}_1[\phi_S(A)]^2$, we study $\mathbb{E}_1[\phi_S(A)]$. Let $K \subset [n]$ denote the vertex set of the clique under $\mathbb{P}_1$. If either i or j is not in K, then $A_{ij} \sim \mathsf{Ber}(1/2)$ and $\mathbb{E}[\bar{A}_{ij} \mid z] = 0$; otherwise, $A_{ij} = 1$ and $\bar{A}_{ij} = 1$. Therefore, by the independence of A_{ij} conditional on K, we have

$$\mathbb{E}_1[\phi_S(A)] = \mathbb{E}\left[\prod_{(i,j)\in S} \mathbb{E}_1[\bar{A}_{ij} \mid K]\right] = \mathbb{P}\{i, j \in K \text{ for all } (i,j) \in S\}$$

which is precisely the probability that the clique K contains all vertices of S viewed as a graph. Let $v(S)$ denote the number of vertices of S. Note that we must have $v(S) \le k$ for otherwise the above probability is zero. Since K is by definition a uniformly random subset of $[n]$ of size k under $\mathbb{P}_1$, we obtain that for $v(S) \le k$,

$$\mathbb{E}_1[\phi_S(A)] = \frac{\binom{n-v(S)}{k-v(S)}}{\binom{n}{k}} = \frac{k(k-1)\cdots(k-v(S)+1)}{n(n-1)\cdots(n-v(S)+1)} \le (k/n)^{v(S)}.$$

- First, consider the case $D = \binom{n}{2}$ so that $\|L\| = \|L_{\le D}\|$. Then we have

$$\|L\|^2 = \sum_{S \subset \binom{[n]}{2}} \mathbb{E}_1[\phi_S(A)]^2 \le \sum_{S : v(S) \le k} (k/n)^{2v(S)} = \sum_{m=0}^{k} \sum_{S : v(S) = m} (k/n)^{2m} \le \sum_{m=0}^{k} n^m 2^{km/2} (k/n)^{2m},$$

  where the last step holds because there are at most $\binom{n}{m} 2^{\binom{m}{2}} \le n^m 2^{m^2/2} \le n^m 2^{km/2}$ graphs S with $v(S) = m$. Furthermore, if $k \le (2-\varepsilon)\log_2 n$, then

$$n 2^{k/2} (k/n)^2 \le n\, n^{1-\varepsilon/2} \Big(\frac{2\log_2 n}{n}\Big)^2 = \frac{(2\log_2 n)^2}{n^{\varepsilon/2}} \le 1/2.$$

  We conclude that

$$\|L\|^2 \le \sum_{m=0}^{k} (1/2)^m \le 2.$$

- Next, consider the low-degree case where $D = o\big((\frac{\log n}{\log\log n})^2\big)$. For brevity, we assume $\sqrt{D}$ is an integer. For $m \le 2\sqrt{D}$, there are at most $\binom{n}{m}2^{\binom{m}{2}} \le n^m 2^{m^2} \le n^m 2^{m\sqrt{D}}$ graphs S such that $v(S) = m$. For $2\sqrt{D} < m \le 2D$, there are at most $\binom{n}{m}\binom{m}{2}^D \le n^m m^{2D}$ graphs S such that $v(S) = m$ and $|S| \le D$. It follows that

$$\|L_{\le D}\|^2 = \sum_{S \subset \binom{[n]}{2}, |S| \le D} (k/n)^{2v(S)} = \sum_{m=0}^{2D} \sum_{v(S)=m, |S| \le D} (k/n)^{2m}$$
$$\le \sum_{m=0}^{2\sqrt{D}} n^m 2^{m^2} (k/n)^{2m} + \sum_{2\sqrt{D}}^{2D} n^m m^{2D} (k/n)^{2m}.$$

For the first term, note that for $D = o\big((\frac{\log n}{\log\log n})^2\big)$ and $k \le n^{1/2-\varepsilon}$, we have

$$n 2^{2\sqrt{D}} (k/n)^2 \le n e^{o(\log n)} (k/n)^2 \le n^{1+o(1)} n^{-1-2\varepsilon} \le 1/2.$$

Therefore,

$$\sum_{m=0}^{2\sqrt{D}} n^m 2^{m^2} (k/n)^{2m} \le \sum_{m=0}^{2\sqrt{D}} (n 2^{2\sqrt{D}} (k/n)^2)^m \le 2.$$

For the second term, note that for $m = 2\sqrt{D}$, we have

$$n^{2\sqrt{D}} (2\sqrt{D})^{2D} (k/n)^{4\sqrt{D}} = \big(n(k/n)^2 (2\sqrt{D})^{\sqrt{D}}\big)^{2\sqrt{D}}$$
$$\le \big((k^2/n)(\log n)^{o(\frac{\log n}{\log\log n})}\big)^{2\sqrt{D}} \le \big(n^{-2\varepsilon} n^{o(1)}\big)^{2\sqrt{D}} \le 1.$$

Moreover, for $2\sqrt{D} \le m < 2D$, we have

$$\frac{n^{m+1}(m+1)^{2D}(k/n)^{2(m+1)}}{n^m m^{2D} (k/n)^{2m}} \le n(k/n)^2 \Big(1 + \frac{1}{2\sqrt{D}}\Big)^{2D}$$
$$\le (k^2/n)\, e^{\sqrt{D}} \le n^{-2\varepsilon} e^{o(\log n)} \le n^{-2\varepsilon} n^{o(1)} \le 1/2.$$

We conclude that

$$\sum_{2\sqrt{D}}^{2D} n^m m^{2D} (k/n)^{2m} \le 2.$$

The two terms combined yield that $\|L_{\le D}\|^2 \le 4$.

□

We summarize the statistical-to-computational gap for planted clique detection as follows.

Corollary 6.14. *Consider testing between $\mathbb{P}_0 = G(n, 1/2)$ and $\mathbb{P}_1 = G(n, 1/2, k)$.*

- *If $k \ge (2+\varepsilon)\log_2 n$ for $\varepsilon > 0$, then there is a consistent test.*
- *If $k \le (2-\varepsilon)\log_2 n$ for $\varepsilon > 0$, then there is no consistent test.*
- *If $k \ge Cn^{1/2}$ for a large constant $C > 0$, then there is a polynomial of degree $O(\log n)$ that strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$.*
- *If $k \le n^{1/2-\varepsilon}$ for $\varepsilon > 0$, then there is no polynomial of degree $o\big((\frac{\log n}{\log\log n})^2\big)$ that strongly separates $\mathbb{P}_0$ and $\mathbb{P}_1$.*

Bibliography

[BAH+22] Afonso S Bandeira, Ahmed El Alaoui, Samuel B Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *arXiv preprint arXiv:2205.09727*, 2022.

[BBB+13] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

[BC15] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[BY05] Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[Efr12] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[GdHK20] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

[JN20] Anatoli Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization*. Princeton University Press, 2020.

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[LSST16] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[PW19] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Preprint*, 2019.

[Ros20] Sheldon M Ross. *Introduction to probability and statistics for engineers and scientists*. Academic Press, 2020.

[RPC19] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

[vdV00] Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Was04] Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.