Statistics
MDI220
4. Hypothesis Testing

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

We present methods to decide whether data support a given hypothesis.

1 Statistical test

The objective of a statistical test is to decide whether some hypothesis, called the null hypothesis, can be rejected in favor of an alternative hypothesis, in view of the observations. The null and alternative hypotheses are respectively denoted by H_0 and H_1.

The null hypothesis H_0 is that considered as true in the absence of any observation (default choice).

Example. *There is the same proportion of male and female births (null hypothesis, H_0).*

The null and alternative hypotheses do not play the same role. If the null hypothesis is rejected in favor of the alternative hypothesis, the statistician must provide some evidence for this. Let $\delta(x) \in \{0, 1\}$ be the decision, based on the observation x:

$$\delta(x) = \begin{cases} 0 & \rightarrow \text{ accept } H_0 \\ 1 & \rightarrow \text{ reject } H_0 \text{ in favor of } H_1. \end{cases}$$

The type-I error rate is the rate of false positive $\rightarrow$ $\alpha = \mathrm{P}(\delta(X) = 1|H_0)$.
The type-II error rate is the rate of false negative $\rightarrow$ $\beta = \mathrm{P}(\delta(X) = 0|H_1)$.

Example. *Let H_0 = "There is the same proportion of male and female births".*
A type-I error is to falsely decide from data that there is not the same proportion of male and female births.

There is a trade-off between type-I and type-II errors. A trivial test deciding 0 for any observation has type-I error rate $\alpha = 0$ but type-II error rate $\beta = 1$; similarly, a trivial test deciding 1 for any observation has type-II error rate $\beta = 0$ but type-I error rate $\alpha = 1$.

The Neyman-Pearson principle consists in controlling the type-I error rate, that is to find a statistical test minimizing the type-II error rate β with a constraint on the type-I error rate α. The constraint on α is called the *level* of the test (e.g., 1%); the quantity $1 - \beta$ is called the *power* of the test.

In practice, the significance of the test on observed data is assessed through the so-called p-value.

The p-value is the probability of sampling a value at least as extreme as the true observation under the null hypothesis H_0.

Example. *Consider a test of level $\alpha = 1\%$. The null hypothesis is rejected whenever the p-value is less than 1%. For a p-value of $1‰$, the null hypothesis is rejected, with high confidence. In words, the type-I error rate (false positive) is $1‰$, much less than the target of 1%.*

2 Parametric model

In parametric models, the hypotheses form subsets of the parameters:

$$\begin{array}{l} H_0 \to \Theta_0 \subset \Theta \\ H_1 \to \Theta_1 \subset \Theta \end{array} \quad \Theta_0 \cap \Theta_1 = \emptyset$$

Following Neyman-Pearson's principle, the objective is to control the worst type-I error rate:

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\delta(X) = 1).$$

The quality of the test is then assessed through the type-II error rates:

$$\forall \theta \in \Theta_1, \quad \beta(\theta) = P_\theta(\delta(X) = 0).$$

A statistical test δ is *uniformly most powerful* (UMP) at level α if for any other test δ' such that:

$$\sup_{\theta \in \Theta_0} P_\theta(\delta'(X) = 1) \leq \alpha,$$

we have:

$$\forall \theta \in \Theta_1, \quad P_\theta(\delta'(X) = 0) \geq \beta(\theta).$$

The existence of a UMP test is not guaranteed, except in some specific cases. In the rest of the document, we consider a statistical model with some real parameter $\theta \in \mathbb{R}$. We assume that the model is dominated and denote by p_θ the probability density function.

3 Simple hypotheses

The hypotheses are simple if $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.

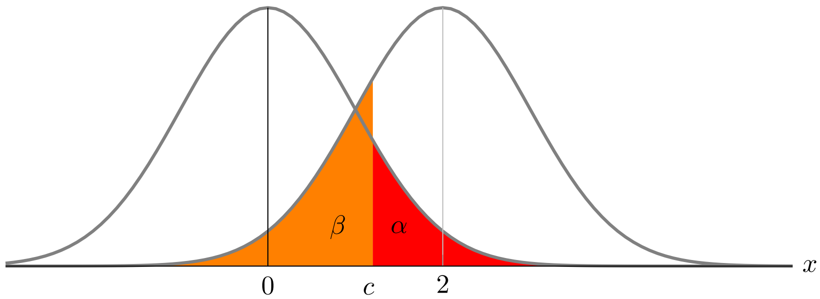
For simple hypotheses, it is always possible to find a UMP test. Denote by $p_0 = p_{\theta_0}$ and $p_1 = p_{\theta_1}$ the probability measures under hypotheses H_0 and H_1.

The test $\delta(x) = 1_{\left\{\frac{p_1(x)}{p_0(x)} > c\right\}}$ is UMP, for any $c > 0$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$. The hypotheses are $H_0 = \{\theta = 0\}$ and $H_1 = \{\theta = 2\}$. There is a single observation. The likelihood ratio is:*

$$\frac{p_1(x)}{p_0(x)} = \frac{e^{-\frac{(x-2)^2}{2}}}{e^{-\frac{x^2}{2}}} \propto e^{4x}$$

This is an increasing function of x so any test of the form $\delta(x) = 1_{\{x>c\}}$ is UMP.



Denoting by P_0 the probability measure under H_0, we get the type-I error rate:

$$\alpha = P_0(\delta(X) = 1) = P_0(X > c) = P(Z > c),$$

with $Z \sim \mathcal{N}(0, 1)$. Similarly, with P_1 the probability measure under H_1, the type-II error rate is given by:

$$\beta = P_1(\delta(X) = 0) = P_1(X \le c) = P(Z \le c - 2).$$

Taking $c = 2$ for instance, we get $\alpha \approx 2\%$ and $\beta = 50\%$. For $x = 3$, the null hypothesis is rejected. The p-value of this decision is $\approx 1‰$.

4 One-tailed test

A one-tailed test has the form $\Theta_0 = \{\theta \le \theta_0\}$ and $\Theta_1 = \{\theta > \theta_0\}$.

For one-tailed hypothesis testing, it is possible to find a UMP test provided the likelihood ratio is monotone, in the sense that:

$$\forall \theta' > \theta, \quad \frac{p_{\theta'}(x)}{p_\theta(x)} = f(T(x)),$$

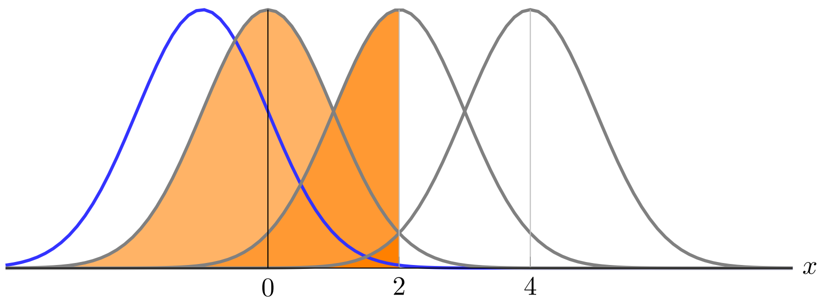
where T is a statistic (independent of θ) and f is an increasing function.

If the likelihood ratio is monotone, the test $\delta(x) = 1_{\{T(x)>c\}}$ is UMP, for any $c > 0$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$. The hypotheses are $H_0 = \{\theta \le 0\}$ and $H_1 = \{\theta > 0\}$. The likelihood ratio is:*

$$\forall \theta' > \theta, \quad \frac{p_{\theta'}(x)}{p_\theta(x)} = \frac{e^{-\frac{(x-\theta')^2}{2}}}{e^{-\frac{(x-\theta)^2}{2}}} \propto e^{2(\theta'-\theta)x}.$$

This is an increasing function of x so any test of the form $\delta(x) = 1_{\{x>c\}}$ is UMP.



The type-I error rate is given by:

$$\forall \theta \leq 0, \quad P_\theta(\delta(X) = 1) = P_\theta(X > c) = P(Z > c - \theta),$$

with $Z \sim \mathcal{N}(0, 1)$, so that:

$$\alpha = \sup_{\theta \leq 0} P_\theta(\delta(X) = 1) = P(Z > c).$$

Similarly, the type-II error rate is given by:

$$\forall \theta \geq 0, \quad \beta(\theta) = P_\theta(\delta(X) = 0) = P_\theta(X \leq c) = P(Z \leq c - \theta).$$

Taking $c = 2$ for instance, we get $\alpha \approx 2\%$ and $\beta(0^+) \approx 98\%, \beta(2) = 50\%, \beta(4) \approx 2\%$.

5 Two-tailed test

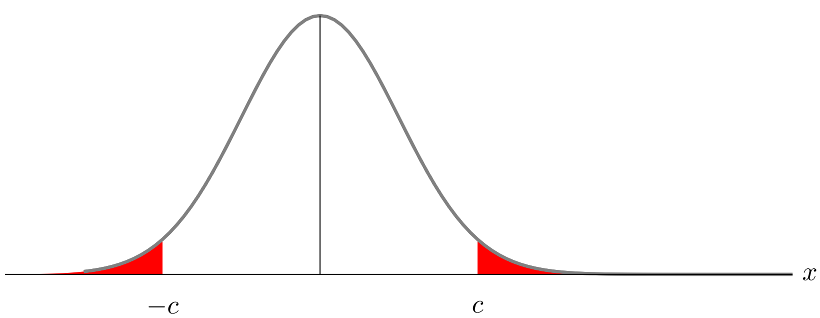
A two-tailed test has the form $\Theta_0 = [\theta_1, \theta_2]$ for some $\theta_1 \leq \theta_2$ and $\Theta_1 = \mathbb{R} \setminus \Theta_0$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$. The hypotheses are $H_0 = \{\theta = 0\}$ and $H_1 = \{\theta \neq 0\}$. The type-I error rate of the test $\delta(x) = 1_{\{|x| > c\}}$ is:*

$$\alpha = P_0(|X| > c) = 2P(Z > c)$$

with $Z \sim \mathcal{N}(0, 1)$. Now the type-II error rates are:

$$\forall \theta \neq 0, \quad \beta(\theta) = P_\theta(|X| \leq c) = P(Z \in [\theta - c, \theta + c]).$$



This example shows that there is no UMP test in general. Consider the other test $\delta'(x) = 1_{\{x > c'\}}$. Assume that c' is chosen in such a way that δ' has level α. Then δ (or any other test of level α) cannot be more powerful than δ' for all $\theta \neq 0$. The reason is that the power of δ' is:

$$\forall \theta \neq 0, \quad \beta'(\theta) = P_\theta(X \leq c') = P(Z \leq c' - \theta) \to 0 \quad \text{when } \theta \to +\infty.$$

We need to impose another assumption on the test.

A test of level α is *unbiased* if $P_\theta(\delta(X) = 1) \geq \alpha$ for all $\theta \in \Theta_1$.

If the test is unbiased, the probability of rejecting H_0 against H_1 is always higher under H_1 (that is, for all parameters $\theta \in \Theta_1$). Observe that the previous test δ' is biased since

$$P_\theta(X > c') = P(Z > c' - \theta) \to 0 \quad \text{when } \theta \to -\infty.$$

We are looking for a uniformly most powerful among all unbiased tests (UMPU test).

Consider a parametric model in the exponential family $p_\theta(x) = h(x)e^{\eta(\theta)T(x) - A(\theta)}$ with $T(x) \in \mathbb{R}$ and η an increasing function. There exists a UMPU test of the form $\delta(x) = 1_{\{T(x) \notin [c_1, c_2]\}}$.

Example. *Consider the Gaussian model with hypothesis* $H_0 = \{\theta = 0\}$ *and* $H_1 = \{\theta \neq 0\}$. *We have:*

$$p_\theta(x) \propto e^{-\frac{1}{2}(x-\theta)^2} \propto e^{-\frac{x^2}{2}} e^{x\theta}.$$

The corresponding function $\eta(\theta) = \theta$ *is increasing, so there exists a UMPU test of the form* $\delta(x) = 1_{\{x \notin [c_1, c_2]\}}$. *By symmetry, this test has the form* $\delta(x) = 1_{\{|x| > c\}}$, *as considered previously.*