

Mathematical Statistics 2, Part III: Hypothesis testing

Statistics Team TSE

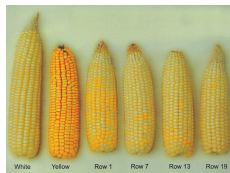
2022–2023

Syllabus

- 1 Introduction to hypothesis testing
 - An introductory example
 - Definition
 - Terminology
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory

Historical Example : Zea mays plant

- 1878 : Darwin recorded some data on the heights of Zea mays plants.
- Effect of **cross**/**self**-fertilization on the height of Zea mays.



- Experiment : select **cross**/**self**-fertilized plants, grow them in the same pot, and then later measure their heights.
- Statistical question : are cross-fertilized plants generally taller than self-fertilized plants ?

Historical Example : Zea mays plant

- Y : height of cross-fertilized plants, Z : height of self-fertilized plants
-

$$\mathbb{E}[Y] \geq \mathbb{E}[Z]?$$

- Denote $X = Y - Z$, that mimics the difference of height of plants grown in the same pot and define

$$\mu := \mathbb{E}[X]$$

- Record $X_1 = Y_1 - Z_1, \dots, X_n = Y_n - Z_n$
- $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu > 0$
- This is not exactly an estimation problem. We have to take a decision instead.

Syllabus

- 1 Introduction to hypothesis testing
 - An introductory example
 - Definition
 - Terminology
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory

What is testing ?

Consider a parametric statistical model, indexed by θ .

Testing is deciding between two (mutually exclusive) hypotheses :

- the **null hypothesis** \mathcal{H}_0
- the **alternative hypothesis** \mathcal{H}_1 .

This defines a **testing problem**.

These hypotheses are claims about the distribution, i.e. about θ .

		Reality	
		H_0 False	H_0 True
Test	Reject H_0	Correct rejection H_0 ✓ = Power = $1 - \beta$	Type I error = α ✗
	Accept H_0	Type II error ✗	Correct acceptance of H_0 ✓

A toy Gaussian example

Let (X_1, X_2, \dots, X_n) be a random sample from the $\mathcal{N}(\mu, 1)$ distribution. We can test $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$.

($\bar{X} = 0$ is not a hypothesis!)

- A hypothesis is **simple** when it fully fixes the distribution of the observations.
- A hypothesis that is not simple is said to be **composite**.

In this practical example, \mathcal{H}_0 is a simple hypothesis while \mathcal{H}_1 is composite.

A toy Gaussian example

To choose between \mathcal{H}_0 and \mathcal{H}_1 , we consider $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

A **test** is a rule for deciding whether or not to reject \mathcal{H}_0 in favor of \mathcal{H}_1 .

Here, it seems natural that

- (i) a small value of $|\bar{X}_n|$ is compatible with \mathcal{H}_0 ,
- (ii) a large value of $|\bar{X}_n|$ provides evidence in favor of \mathcal{H}_1 .

We say that \bar{X}_n is the **test statistic**.

A resulting decision rule would be of the following form :

$$\text{reject } \mathcal{H}_0 \text{ iff } |\bar{X}_n| \geq s_n,$$

where $s_n > 0$ is a given **threshold** (or **cutoff**).

Key observation : under \mathcal{H}_0 , all is known about $\mathcal{L}(\bar{X}_n)$!

A test on the mean of a Gaussian sample

Let \mathcal{X} be the sample space (i.e., the set of possible observations).

Formally, a test will be a map $\varphi : \mathcal{X} \rightarrow \{0, 1\}$, where

- $r = 0$ stands for non-rejection of \mathcal{H}_0 , i.e. accept \mathcal{H}_0
- $r = 1$ stands for rejection of \mathcal{H}_0 , i.e. accept \mathcal{H}_1

In the example above, the values in $\mathcal{X} = \mathbb{R}^n$ corresponding to the condition $|\bar{X}_n| > s_n$ define the **critical region** (or **rejection region**) \mathcal{R} :

$$\mathcal{R} = \{(X_1, \dots, X_n) \in \mathcal{X} : |\bar{X}_n| \geq s_n\} = \varphi^{-1}(\{1\}).$$

The complement of this region, $\mathcal{A} = \varphi^{-1}(\{0\})$, is the **“acceptance” region**.

Syllabus

- 1 Introduction to hypothesis testing
 - An introductory example
 - Definition
 - Terminology
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory

Qualities of a hypothesis

A **simple** hypothesis fully characterizes the distribution of the observations :
For $X \sim N(\mu, 1)$, $\mathcal{H}_0 : \mu = 0$ and $\mathcal{H}_1 : \mu = 1$ are simple hypotheses.

A **composite** hypothesis is one that is not simple : for $X \sim N(\mu, \sigma^2)$,

$$\mathcal{H}_0 : \mu \leq 0, \quad \mathcal{H}_1 : \mu > 0,$$

$$\mathcal{H}_0 : \mu = 0, \quad \mathcal{H}_1 : \mu = 1,$$

all are composite hypotheses.

Unilateral or **one-sided** testing problem : $\mathcal{H}_0 : \mu \leq 0$ vs $\mathcal{H}_1 : \mu > 0$
 (“ \mathcal{H}_1 lies on one side of \mathcal{H}_0 only”)

Bilateral or **two-sided** testing problem : $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$
 (“ \mathcal{H}_1 lies on both sides of \mathcal{H}_0 ”)

Classical testing problems

- **Testing about means** : is a mean equal to a given prespecified value ?
Are two means equal to each other ?
- **Testing about proportions** : is a proportion equal to a given value ?
Are two proportions equal to each other ?
- **Testing about variances** : is a variance equal to a given value ?
Are two variances equal to each other ?
- **Goodness-of-fit testing** : is a sample arising from a normal distribution ? From some other prespecified class of distributions ?
- **Independence testing** : are two variables mutually independent ?

Methods for constructing tests

- Neyman–Pearson tests
- Likelihood ratio tests
- Wald tests
- Lagrange multiplier tests

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
 - Definitions
 - Examples
 - Test comparison
 - Dissymmetry of \mathcal{H}_0 and \mathcal{H}_1
 - Neyman principle
 - Computing the significance level
 - Sample size dependency
 - From CI to tests and reciprocally
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory

General framework

In a statistical model $(\mathcal{X}, \mathcal{F}, \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\})$, we partition Θ into

$$\Theta = \Theta_0 \cup \Theta_1,$$

which defines the testing problem

$\mathcal{H}_0 : \theta \in \Theta_0$ (null hypothesis)

$\mathcal{H}_1 : \theta \in \Theta_1$ (alternative hypothesis)

(The statistician must decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$).

The set of decisions is modelled by $\{0, 1\}$.

A decision rule (a test φ) is therefore the indicator function of a subset of \mathcal{X} called the rejection region (described by a condition on a test statistic).

Loss function

There are two possible types of errors.

Type 1 error : \mathcal{H}_0 holds true and the statistician decides to reject \mathcal{H}_0 .
The cost of this loss is assumed to be equal to L_1 .

Type 2 error : \mathcal{H}_1 holds true and the statistician decides to accept \mathcal{H}_0 .
The cost of this loss is assumed to be equal to L_2 .

Of course, if the statistician makes the right decision, then the loss is zero.

Risk function

For a test φ , associated to the critical region \mathcal{R} , the resulting **risk function** $\theta \mapsto R(\theta, \varphi)$ is then as follows :

- for $\theta \in \Theta_0$, $R(\theta, \varphi) = L_1 \times \mathbb{P}_\theta[\varphi = 1]$
- for $\theta \in \Theta_1$, $R(\theta, r_C) = L_2 \times \mathbb{P}_\theta[\varphi = 0] = L_2 \times (1 - \mathbb{P}_\theta[\varphi = 1])$

The risk is therefore completely determined by the **power function**

$$\beta_\varphi : \theta \in \Theta \longmapsto \mathbb{P}_\theta[\varphi = 1].$$

- **Type 1 risk at $\theta \in \Theta_0$** : $\alpha_\varphi(\theta) = \mathbb{P}_\theta[\varphi = 1]$
Size of the test :

$$\alpha(\varphi) = \sup_{\theta \in \Theta_0} \alpha_\varphi(\theta)$$

- **Type 2 risk at $\theta \in \Theta_1$** : $1 - \beta_\varphi(\theta)$

Computing a power function : Example 1

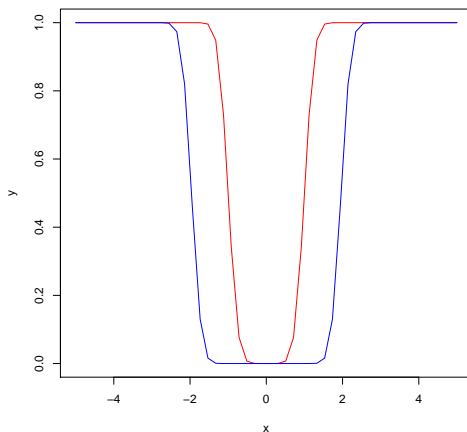
In the Gaussian example above, the power function is β_φ :

$$\begin{aligned}\mu \mapsto \mathbb{P}_\mu[\varphi = 1] &= \mathbb{P}_\mu[|\bar{X}_n| > s] = \mathbb{P}[|\mathcal{N}(\mu, \tfrac{1}{n})| > s] \\ &= \mathbb{P}[\mathcal{N}(\mu, \tfrac{1}{n}) < -s] + \mathbb{P}[\mathcal{N}(\mu, \tfrac{1}{n}) > s] \\ &= \mathbb{P}[\mathcal{N}(0, 1) < -\sqrt{n}(s + \mu)] + \mathbb{P}[\mathcal{N}(0, 1) > \sqrt{n}(s - \mu)] \\ &= \Phi(-\sqrt{n}(s + \mu)) + 1 - \Phi(\sqrt{n}(s - \mu)),\end{aligned}$$

where $\Phi(x) = \mathbb{P}[\mathcal{N}(0, 1) < x]$ is the c.d.f. of the $\mathcal{N}(0, 1)$ distribution.

Plot of a power function

Plot of this power function for $(n, s) = (25, 1)$ and $(n, s) = (25, 1.96)$



Comparing tests

The test φ (with rejection region \mathcal{R}) is preferred to the test φ' (with critical region \mathcal{R}') if and only if

$$\alpha_{\varphi}(\theta) \leq \alpha_{\varphi'}(\theta) \quad \forall \theta \in \Theta_0 \quad \text{and} \quad \beta_{\varphi}(\theta) \geq \beta_{\varphi'}(\theta) \quad \forall \theta \in \Theta_1,$$

or equivalently, if and only if

$$\mathbb{P}_{\theta}[\varphi = 1] \leq \mathbb{P}_{\theta}[\varphi' = 1] \quad \forall \theta \in \Theta_0 \quad \text{and} \quad \mathbb{P}_{\theta}[\varphi = 1] \geq \mathbb{P}_{\theta}[\varphi' = 1] \quad \forall \theta \in \Theta_1.$$

The comparison is therefore independent of L_1 and L_2 and only involves the power function of each test.

The ideal test

The ideal test would be one for which

- (i) the power function takes value 0 on Θ_0 , and
- (ii) the power function takes value 1 on Θ_1

(such a test would indeed be preferred to any other test).

But **there is no such test!** (except in degenerate cases where one can discriminate with certainty between \mathcal{H}_0 and \mathcal{H}_1).

However, one may use the above concept to restrict to **admissible tests** (a test is said to be admissible if no other test will be preferred to it).

Computing a power function : Example 2

We consider the Bernoulli model $\mathcal{B}(\theta)$, $\theta \in [0, 1]$ and

$$\mathcal{H}_0 = \{\theta \leq \theta_0\} \quad \text{and} \quad \mathcal{H}_1 = \{\theta > \theta_0\}.$$

We use the statistics

$$T = \sum_{i=1}^n X_i \sim \mathcal{B}(n, \theta),$$

and define the test φ_c :

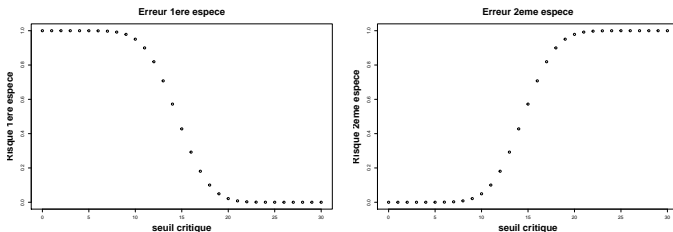
$$\varphi_c = \mathbf{1}_{T \geq c}.$$

We observe that $\theta \mapsto \beta_{\varphi_c}(\theta) = \mathbb{P}_\theta[\varphi_c = 1]$ is an **increasing** function of θ :

$$\alpha(\varphi_c) = \sup_{\theta \leq \theta_0} \beta_{\varphi_c}(\theta) = \beta_{\varphi_c}(\theta_0).$$

Computing a power function : Example 2

$$\mathcal{H}_0 = \{\theta \leq \theta_0\} \quad \text{and} \quad \mathcal{H}_1 = \{\theta > \theta_0\}.$$



Neyman approach : we fix an amount α of admissible Type 1 error and then calibrate the test φ_c accordingly :

$$c_\alpha = \inf\{c : \beta_{\varphi_c}(\theta_0) \leq \alpha\}.$$

Numerical example

- $\alpha = 0.05$, $n = 1000$, $\theta_0 = 0.5$, $c_\alpha = 526$
- $\alpha = 0.01$, $n = 1000$, $\theta_0 = 0.5$, $c_\alpha = 537$

Computing a power function : Example 2

Switch the position of \mathcal{H}_0 and \mathcal{H}_1 .

$$\tilde{\mathcal{H}}_0 = \{\theta > \theta_0\} \quad \text{and} \quad \tilde{\mathcal{H}}_1 = \{\theta \leq \theta_0\}.$$

$\tilde{\varphi}_c$ becomes

$$\tilde{\varphi}_c = \mathbf{1}_{T \leq c}.$$

The power function $\beta_{\tilde{\varphi}_c}(\theta) = \mathbb{P}_{\theta}[\tilde{\varphi}_c = 1]$, decreases with θ .

$$\alpha(\tilde{\varphi}_c) = \sup_{\theta > \theta_0} \beta_{\tilde{\varphi}_c}(\theta) = \mathbb{P}_{\theta_0}[\tilde{\varphi}_c = 1] = \mathbb{P}_{\theta_0}[T \leq c]$$

Neyman approach : Numerical example

- $\alpha = 0.05$, $n = 1000$, $\theta_0 = 0.5$, $c_{\alpha} = 474$
- $\alpha = 0.01$, $n = 1000$, $\theta_0 = 0.5$, $c_{\alpha} = 463$

Computing a power function : Example 2

Imagine that we observe with $n = 1000$

$$T = \sum_{i=1}^n X_i = 510.$$

Define $\theta_0 = 0.5$, and $\alpha = 0.05$.



$$\mathcal{H}_0 = \{\theta \leq \theta_0\} \quad \text{and} \quad \mathcal{H}_1 = \{\theta > \theta_0\}.$$

We are led to accept \mathcal{H}_0 since $510 < 526$.



$$\tilde{\mathcal{H}}_0 = \{\theta > \theta_0\} \quad \text{and} \quad \tilde{\mathcal{H}}_1 = \{\theta \leq \theta_0\}.$$

We are led to accept $\tilde{\mathcal{H}}_0$ too since $510 > 474$.

The choice of \mathcal{H}_0 is therefore fundamental !

Neyman principle with a level of a test

To calibrate a statistical test, we need to :

- Specify a statistical model $\mathbb{P}_\theta, \theta \in \Theta_0$
- Identify \mathcal{H}_0 and \mathcal{H}_1 (or equivalently Θ_0 and Θ_1)
- Propose a family of tests φ_c , c being a threshold.
- **Significance level** or **size** of a test : α (maximal Type 1 error).
- **Neyman principle**. Find a test φ_{c_α} such that :

$$\beta_{c_\alpha} := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\varphi_c = 1] \leq \alpha.$$

Important point : the calibration of the test φ_{c_α} essentially depends on \mathcal{H}_0 (and not on \mathcal{H}_1) !

P-Value

In practice, the Neyman principle leads to an indexation of tests with α :

$$\alpha \longmapsto \varphi_\alpha$$

The region of rejection \mathcal{R}_α may also be parametrized with the help of α :

$$\mathcal{R}_\alpha := \{\varphi_\alpha = 1\}$$

It is an easy exercise to observe in practical situation that :

$$\alpha' < \alpha \implies \mathcal{R}_{\alpha'} \subset \mathcal{R}_\alpha.$$

- If $\alpha = 1$, we always reject \mathcal{H}_0 (regardless the sample) :

$$\mathcal{R}_1 = \mathcal{X}^n$$

- If $\alpha = 0$, we never reject \mathcal{H}_0 (regardless the sample) :

$$\mathcal{R}_0 = \emptyset$$

P-Value

Definition : A *p-value* of a sample X is the smallest value of α that leads to reject \mathcal{H}_0 :

$$p - value(X) = \inf \{ \alpha \in [0, 1] : X \in \mathcal{R}_\alpha \}.$$

It is a level of significance of a test.

- For $\alpha < p - value(X)$, we accept \mathcal{H}_0 .
- For $\alpha > p - value(X)$, we reject \mathcal{H}_0 .

Hence, given a $p - value(X)$ small is in favor of rejecting \mathcal{H}_0

Computing the significance level of a test

In our Gaussian example, with $n = 25$ (and $s \geq 0$) :

- The significance level of the test is $\alpha = \mathbb{P}_0[\varphi_s] = 2\Phi(-5s)$
(which gives $\mathbb{P}_0[\varphi_{0.8}] = 6.3 \times 10^{-5}$ or $\mathbb{P}_0[\varphi_{0.5}] = 0.012$)
- If we observe $\bar{X}_n = 1.5$, then the empirical significance level is
 $\mathbb{P}_0[|\bar{X}_n| \geq 1.5] = 2\Phi(-5 \times 1.5) = 6.3 \times 10^{-14}$

In this example :

- If $s \nearrow$, then the significance level \searrow
- If $|\bar{X}_n| \nearrow$, then the empirical level \searrow
- If the true value of $|\mu| \nearrow$, then the Type 2 risk \searrow

Testing using a level or a p-value

In practice, there are two ways to conduct a test :

- 1 given a target level α , find the corresponding critical region and see whether or not the observation belongs to this critical region.
- 2 compute the empirical significance level (p-value) and let the final user choose its own level. The rule is then to reject \mathcal{H}_0 if the p-value is strictly smaller than the chosen level.

If the test rejects \mathcal{H}_0 , then the test is said to be **significant**.

Testing with large samples

“Unless the data follow a parametric model extremely closely, almost any model will be rejected when using a sufficiently large set of data”.

In our running example (X_1, X_2, \dots, X_n are i.i.d. $\mathcal{N}(\mu, 1)$) :

If $0 < s < 1$, then $\mathbb{P}_\mu[\varphi_s = 1] \rightarrow 1$ as $n \rightarrow \infty$: with a large enough sample of size n , we will asymptotically reject \mathcal{H}_0 under the alternative with probability 1 !

Conclusion : the threshold should be chosen as a function of n .

CI \implies Tests

$$\mathcal{H}_0 := \{\theta = \theta_0\} \quad \text{and} \quad \mathcal{H}_1 := \{\theta \neq \theta_0\}$$

Imagine that we are able to build a CI of level $1 - \alpha$

$$\mathbb{P}_\theta[\theta \in [LB(X), UB(X)]] = 1 - \alpha.$$

We can define $\varphi_\alpha = \mathbf{1}_{\theta \notin [LB(X), UB(X)]}$.

As a construction, the test satisfies : $\mathbb{P}_{\theta_0}[\varphi_\alpha = 1] = \alpha$.

Example Let $X = (X_1, \dots, X_n)$, where the X_i 's are i.i.d. $\mathcal{N}(\mu, 1)$.

$$[LB(X), UB(X)] = \left[\bar{X}_n - z_{1-\alpha/2} \frac{1}{\sqrt{n}}, \bar{X}_n + z_{1-\alpha/2} \frac{1}{\sqrt{n}} \right]$$

is a CI at confidence level $1 - \alpha$.

Hence, for testing $\mathcal{H}_0 : \mu = \mu_0$ vs $\mathcal{H}_1 : \mu \neq \mu_0$,

$$\varphi_\alpha = \mathbf{1}_{\mu_0 \notin [LB(X), UB(X)]}$$

is a test of size α .

Tests \implies CIs

$$\mathcal{H}_0 := \{\theta = \nu\} \quad \text{and} \quad \mathcal{H}_1 := \{\theta \neq \nu\}$$

Imagine that we are able to test at level α with a test φ_ν .

$$\mathbb{P}_\nu[\varphi_\nu = 1] = \alpha$$

We observe X and introduce :

$$CI(X) = \{\nu : \varphi_\nu(X) = 0\}$$

We then observe that

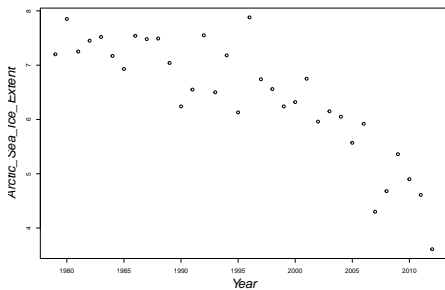
$$\forall \nu \quad \mathbb{P}_\nu[\nu \in CI(X)] = \mathbb{P}_\nu[\varphi_\nu(X) = 0] = 1 - \alpha$$

$CI(X)$ is then a CI of guarantee $1 - \alpha$.

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model**
- 4 Neyman–Pearson theory

A simple regression example



$Y_i = \beta_0 + \beta_1 t_i + \sigma \epsilon_i(\theta)$ where

$\theta = (\beta_0, \beta_1, \sigma)$ and

$t_i = \text{Year}_i - 1979$.

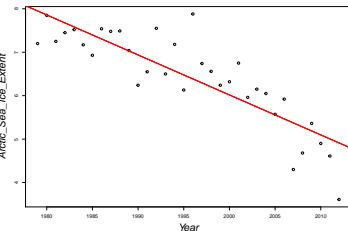
We assume that $\{\epsilon_i(\theta)\}_{i=1}^n$ i.i.d. standard Gaussian.

MSE :

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 t_i)^2$$

$$\hat{\sigma}^2 = (n - 2)^{-1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 t_i)^2$$

A simple regression example



$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \frac{\sigma^2}{ns} \begin{pmatrix} \frac{\|\mathbf{t}\|^2}{n} & -\bar{\mathbf{t}} \\ -\bar{\mathbf{t}} & 1 \end{pmatrix} \right)$$

where $\mathbf{1} = (1, \dots, 1)^T$ and $\mathbf{t} = (t_1, \dots, t_n)^T$,
 $s = n^{-1}\|\mathbf{t}\|^2 - (\bar{\mathbf{t}})^2$ and $\bar{\mathbf{t}} = n^{-1} \sum_{i=1}^n t_i$.

We shall verify (not so obvious) :

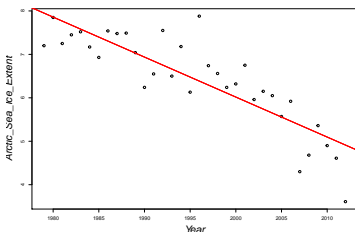
$(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$ and

$(n-2)\hat{\sigma}^2/\sigma^2$ and $(\hat{\beta}_1, \hat{\beta}_2)$ are independent.

	2.5 %	97.5 %	
(Intercept)	7.55	8.34	$\frac{n\sqrt{s}(\hat{\beta}_1 - \beta_1)}{\ \mathbf{t}\ \hat{\sigma}} \quad \text{and} \quad \frac{\sqrt{ns}(\hat{\beta}_2 - \beta_2)}{\hat{\sigma}}$
Year-1979	-0.11	-0.07	

are pivotal statistics.

A simple regression example



$\frac{\sqrt{ns}(\hat{\beta}_2 - \beta_2)}{\hat{\sigma}}$ is a pivotal function for any $\theta = (\beta_1, \beta_2, \sigma^2)$. We shall verify that :

$$\mathbb{P}_\theta \left(\beta_2 \in \left[\hat{\beta}_2 \pm \frac{\hat{\sigma} t_{1-\alpha/2}^{n-2}}{\sqrt{ns}} \right] \right) = 1 - \alpha.$$

We consider the test :

$$\mathcal{H}_0 := \{\beta_2 = 0\} \text{ vs } \mathcal{H}_1\{\beta_2 \neq 0\}.$$

For this given sample, we obtain : p-value :

$$1.76 \cdot 10^{-10}!$$

Should we decide to accept or reject \mathcal{H}_0 ?

From the construction CI \implies test, we deduce that :

$$\phi(Z) = \begin{cases} 1 \\ 0 \end{cases} \left[\hat{\beta}_2 \pm \frac{\hat{\sigma} t_{1-\alpha/2}^{n-2}}{\sqrt{ns}} \right]$$

has level α .

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory**
 - Likelihood Ratio Test
 - Composite hypotheses : UMP tests for MLR families
 - Composite hypotheses : likelihood ratio tests
 - Three classical tests

Introduction to the optimality problem

Towards an optimal testing strategy ?

Consider testing $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta = \theta_1$. Fix $\alpha \in (0, 1)$.

The **Neyman principle** consists in searching, among all tests with level $\leq \alpha$, the one with (uniformly) most power under \mathcal{H}_1 .

Said differently, we are looking for φ_α^\star such that :

- φ_α^\star has size α :

$$\mathbb{P}_{\theta_0}[\varphi_\alpha^\star = 1] \leq \alpha$$

- φ_α^\star is optimal in this class :

$$\forall \psi : \mathbb{P}_{\theta_0}[\psi = 1] \leq \alpha \quad \mathbb{P}_{\theta_1}[\psi = 1] \leq \mathbb{P}_{\theta_1}[\varphi_\alpha^\star = 1]$$

φ_α^\star is referred to as a **Uniformly Most Powerful** test of size α .

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 **Neyman–Pearson theory**
 - **Likelihood Ratio Test**
 - Composite hypotheses : UMP tests for MLR families
 - Composite hypotheses : likelihood ratio tests
 - Three classical tests

Likelihood Ratio Test

Warning : we assume that \mathbb{P}_θ is absolutely continuous w.r.t. Lebesgue measure and > 0 . This may be relaxed (beyond the scope of the lecture)...

Definition : For testing $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta = \theta_1$, the LRT r_K defined as

$$r_K = \mathbf{1} \left\{ \frac{L(\theta_0; x)}{L(\theta_1; x)} \leq K \right\} = \mathbf{1} \left\{ \frac{L(\theta_1; x)}{L(\theta_0; x)} \geq 1/K \right\}$$

(where $L(\theta; x)$ is the likelihood function of the model) is a **Neyman test associated to the threshold $K(> 0)$** .

Remarks :

- The quantity $L(\theta_0; x)/L(\theta_1; x)$ is called a **likelihood ratio**
- This is a likelihood-based rule : it rejects \mathcal{H}_0 if x is sufficiently more likely under θ_1 than under θ_0
- Choosing K fully defines the Neyman test and its level

The Neyman–Pearson Theorem

Neyman–Pearson Theorem : For any $\alpha \in]0, 1[$, a K_α exists such that

- The Neyman test r_{K_α} satisfies :

$$\mathbb{P}_{\theta_0}[r_{K_\alpha} = 1] = \alpha$$

- r_{K_α} is UMP.
- The Neyman test is unbiased, i.e. its power under the alternative is larger than or equal to its level α .

One further result :

- If T is a sufficient statistic, then any Neyman test is a function of T .

The Neyman–Pearson Theorem on Gaussian distributions

Assume that :

$$\mathcal{H}_0 := \{\mathcal{N}(\mu_0, \sigma_0^2)\} \quad \text{vs} \quad \mathcal{H}_1 := \{\mathcal{N}(\mu_1, \sigma_1^2)\}.$$

The LR is given by :

$$\forall x \in \mathbb{R} \quad r(x) = \frac{\sigma_0}{\sigma_1} \exp \left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2 + \frac{1}{2\sigma_0^2}(x - \mu_0)^2 \right).$$

The rejection region is given by :

$$-\frac{1}{2\sigma_1^2}(x - \mu_1)^2 + \frac{1}{2\sigma_0^2}(x - \mu_0)^2 > \log K_\alpha + \frac{1}{2} \log \left(\frac{\sigma_1^2}{\sigma_0^2} \right)$$

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 Neyman–Pearson theory
 - Likelihood Ratio Test
 - Composite hypotheses : UMP tests for MLR families
 - Composite hypotheses : likelihood ratio tests
 - Three classical tests

MLR families of distributions

Definition : A family of distributions with densities $\{f_{\theta}(x) : \theta \in \Theta\}$, where Θ is an interval, has **the monotone likelihood ratio (MLR) property** if there exists a statistic $T = T(X)$ such that any likelihood ratio

$$\frac{f_{\theta_2}(X)}{f_{\theta_1}(X)}, \quad \text{with } \theta_1 < \theta_2,$$

is a monotone (\nearrow or \searrow) function of T .

Examples : the Poisson family, the Gamma family, etc.

The Lehmann theorem

(1) For testing $\mathcal{H}_0 : \theta \leq \theta_0$ vs $\mathcal{H}_1 : \theta > \theta_0$, in a family having the MLR property with LRs \nearrow (resp., \searrow) in T , any test with critical region

$$\varphi_s = \mathbf{1}_{x \in \mathcal{X} : T(x) \geq \text{(resp., } \leq) s}$$

is UMP at its own level (this level is then equal to $\mathbb{P}_{\theta_0}[\varphi_s = 1]$).

(2) For testing $\mathcal{H}_0 : \theta \geq \theta_0$ vs $\mathcal{H}_1 : \theta < \theta_0$, in a family having the MLR property with LRs \nearrow (resp., \searrow) in T , any φ_s :

$$\varphi_s = \mathbf{1}_{T(x) \leq \text{(resp., } \geq) s}$$

is UMP at its own level (still equal to $\mathbb{P}_{\theta_0}[\varphi_s = 1]$).

The Lehmann theorem

Consider testing $\mathcal{H}_0 : \theta \leq \theta_0$ vs $\mathcal{H}_1 : \theta > \theta_0$.

When LR's are  (resp., ) in T ,

$$\frac{L(\theta_0; x)}{L(\theta_1; x)} \leq K \Leftrightarrow \frac{L(\theta_1; x)}{L(\theta_0; x)} \geq 1/K \Leftrightarrow T(x) \geq (\text{resp., } \leq) s,$$

so that, in both cases, the UMP test obtained from the Lehmann theorem coincides with the Neyman test for testing $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta = \theta_1$ (irrespective of the chosen $\theta_1 > \theta_0$).

The Lehmann theorem in exponential families

In the exponential family with densities

$$f_{\theta}(x) = C(\theta)h(x) \exp(Q(\theta)T(x)),$$

with $Q(\theta) \nearrow$ in θ , the test depends only on the sufficient statistic $T(x)$.

Distribution	T for one observation	T for n n i.i.d. observations
Poisson $P(\theta)$	X	\bar{X}
Binomial $B(n, \theta)$	X	\bar{X}
$N(\theta, \sigma^2)$, with σ^2 known	X	\bar{X}
Gamma $\gamma(\theta)$	$\log X$	$n^{-1} \sum_{i=1}^n \log X_i$

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 **Neyman–Pearson theory**
 - Likelihood Ratio Test
 - Composite hypotheses : UMP tests for MLR families
 - **Composite hypotheses : likelihood ratio tests**
 - Three classical tests

Likelihood ratio tests

Consider testing

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad \text{vs} \quad \mathcal{H}_1 : \theta \in \bar{\Theta}_0,$$

where Θ_0 is an affine subspace of Θ with dimension q .

Examples :

- For $\Theta = \{\theta \in \mathbb{R}\}$, testing $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta \neq \theta_0$
- For $\Theta = \{\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \mathbb{R}^2\}$, testing $\mathcal{H}_0 : \theta_1 - \theta_2 = 2$ vs $\mathcal{H}_1 : \theta_1 - \theta_2 \neq 2$

Likelihood ratio tests

In the same way as the MLE of θ is defined as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; x),$$

we define the restricted MLE of θ under \mathcal{H}_0 as

$$\tilde{\theta} = \arg \max_{\theta \in \Theta_0} L(\theta; x).$$

The likelihood ratio statistic is then

$$\lambda = \frac{L(\tilde{\theta}; x)}{L(\hat{\theta}; x)}.$$

The likelihood ratio test rejects \mathcal{H}_0 when $\lambda \leq s$
(where the threshold s is to be determined by the target level of the test).

Likelihood ratio tests : an example

For a random sample X_1, \dots, X_n from the $\mathcal{N}(\mu, \sigma^2)$ distribution with both μ and σ^2 unknown, the LR test for testing $\mathcal{H}_0 : \mu = \mu_0$ vs $\mathcal{H}_1 : \mu \neq \mu_0$ is associated with the critical region

$$C_s = \{(X_1, \dots, X_n) : |T| \geq s\},$$

where we let

$$T = \frac{\bar{X} - \mu_0}{\sqrt{s^2/n}}.$$

This is equivalent to the classical Student test !

(Beware of the two different quantities s above)

Likelihood ratio tests : asymptotic properties

Consider testing $\mathcal{H}_0 : \theta \in \Theta_0$ vs $\mathcal{H}_1 : \theta \in \bar{\Theta}_0$.

We assume that we are in a sampling model with densities satisfying $f_\theta(x) > 0 \forall x \forall \theta$. Then, under some mild regularity assumptions,

$$-2 \log(\lambda) \xrightarrow{\mathcal{L}} \chi_{p-q}^2 \quad \text{under } \mathcal{H}_0,$$

with $p = \dim(\Theta)$ and $q = \dim(\Theta_0)$ (so that $p - q$ is the number of parameters “fixed by the constraint \mathcal{H}_0 ”).

This allows us to construct tests with an asymptotic level equal to α :

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[C] = \alpha$$

Likelihood ratio tests and the Fisher test

Fix $z \in \mathbb{R}^r$ and R an $r \times p$ full-rank matrix ($r < p$). Consider then testing $\mathcal{H}_0 : R\beta = z$ vs $\mathcal{H}_1 : R\beta \neq z$ in the classical linear model

$$Y = X\beta + \varepsilon.$$

Then the **Fisher test** rejects \mathcal{H}_0 for large values of

$$F = \frac{\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2/r}{\|Y - X\hat{\beta}\|^2/(n-p)},$$

where $\hat{\beta} = (X'X)^{-1}X'Y$ is the Gaussian MLE of β (the usual OLS) and $\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(z - R\hat{\beta})$ is the restricted MLE of β .

Likelihood ratio tests and the Fisher test

Here, the likelihood ratio statistic is

$$\lambda = \left(\frac{\|Y - X\hat{\beta}\|^2}{\|Y - X\tilde{\beta}\|^2} \right)^{n/2}$$

and it can be checked that

$$F = \frac{n-p}{r} \left(\frac{1}{\lambda^{2/n}} - 1 \right).$$

Clearly, $\lambda \leq K$ is equivalent to $F \geq C$, so that the Fisher test is equivalent to the likelihood ratio test.

Syllabus

- 1 Introduction to hypothesis testing
- 2 Tests and decision theory
- 3 Application to statistical testing in linear model
- 4 **Neyman–Pearson theory**
 - Likelihood Ratio Test
 - Composite hypotheses : UMP tests for MLR families
 - Composite hypotheses : likelihood ratio tests
 - **Three classical tests**

Three classical tests

Further developments possible...

Single-parameter case : In a model with parameter $\theta \in \mathbb{R}$, we consider testing $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta \neq \theta_0$.

Multi-parameter case : In a model with parameter $\theta \in \mathbb{R}^p$, we consider testing $\mathcal{H}_0 : R\theta = z$ vs $\mathcal{H}_1 : R\theta \neq z$, with R a full-rank $r \times k$ matrix and $z \in \mathbb{R}^r$.

For these situations, there exist three classical tests :

- the Likelihood Ratio test (LRT)
- the Wald test (WT)
- the Lagrange Multiplier test (LMT)

(the last test is also called the *score test*).