Hypothesis Testing I

Maxim Panov

Skoltech

November, 2020

Skoltech

Outline

Hypothesis Testing

- In hypothesis testing, we start with some default theory called a null hypothesis.
- We ask if the data provide sufficient evidence to reject the theory.
- If not we say that we are failing to reject the null.

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$.

$$H_0 \colon p = \frac{1}{2} \text{ (null hypothesis) vs. } H_1 \colon p \neq \frac{1}{2} \text{ (alternative hypothesis).}$$

Consider the statistic $T = |\hat{p}_n - \frac{1}{2}|$.

If T is reasonably large (threshold value will be defined later), then we reject H_0.

Hypothesis Testing

Definition

Statistical Hypothesis or simply **Hypothesis** is an assumption about the probability distribution from which the observed data sample was generated.

Suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1:

$$\Theta = \Theta_0 \bigcup \Theta_1, \ \Theta_0 \bigcap \Theta_1 = \emptyset.$$

We wish to test

$$H_0 \colon \theta \in \Theta_0 \ \text{ vs. } \ H_1 \colon \theta \in \Theta_1.$$

Simple and Composite Hypotheses

We wish to test

$$H_0 \colon \theta \in \Theta_0 \quad \text{vs.} \quad H_1 \colon \theta \in \Theta_1.$$

We consider 2 types of statistical hypotheses:

- **Simple hypothesis** uniquely determines the distribution function on the set under consideration.
  - For example, $\theta = \theta_0$ is a simple hypothesis.
  - Thus, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta \neq \theta_0\}$.

- **Composite hypothesis** asserts that a distribution belongs to a certain set of distributions.
  - For example, $\Theta_0 = \{\theta > \theta_0\}$ vs. $\Theta_1 = \{\theta < \theta_0\}$ is composite hypotheses.

Types of Tests

Definition

Hypothesis testing is the process of deciding whether the statistical hypothesis under consideration contradicts the observed data sample.

In general, statistical tests can be parametric or nonparametric:

- **Parametric test** — a test that assumes that the sample is generated by a distribution from a given parametric family.
  - In particular, there are many criteria for analyzing samples from a normal distribution.
- **Nonparametric test** — a criterion that does not rely on additional distributional assumptions.

Statistical Criterion and Its Properties

Definition

Statistical criterion is a strict mathematical rule by which a statistical hypothesis is accepted or rejected.

- **Data:** $X^n = \{X_1, \ldots, X_n\}$ are i.i.d. random variables from some distribution F with density $f(x;\theta), \theta \in \Theta$.
- **Problem:** $H_0 \colon \theta \in \Theta_0$ vs. $H_1 \colon \theta \in \Theta_1$.
- **Statistical criterion:**

$$\varphi \colon \mathbb{R}^n \to \{0, 1\}.$$

  - If $\varphi(X^n) = 0$ then we accept null hypothesis H_0.
  - If $\varphi(X^n) = 1$ then we reject H_0 (accept its alternative H_1).

Statistical Criterion

- In this regard, each criterion φ corresponds to a certain partitioning of the sample space χ into two mutually complementary sets χ_0 and χ_1.
- χ_0 consists of those observations of x for which H_0 is accepted:

$$\chi_0 = \{X^n \colon \varphi(X^n) = 0\}.$$

- χ_1 of those for which H_0 is rejected (critical region):

$$\chi_1 = \{X^n \colon \varphi(X^n) = 1\}.$$

- Thus, any criterion for testing the hypothesis H_0 is uniquely specified by the corresponding critical region χ_1.

General Principle of Decision Making

Question: How to choose the criterion φ or alternatively the critical area χ_1?

It is usually done based on the *general principle of decision making*:

- if the data observed in an experiment is unlikely under the hypothesis H_0, then
  - it is considered that the hypothesis H_0 does not agree with the data;
  - H_0 is rejected in this case.
- Otherwise, the data is considered to agree with H_0, and H_0 is accepted.

Statistical Criterion and Its Properties

In accordance with the general principle of decision-making, the critical region χ_1 is chosen so that the probability

$$\mathbb{P}(X^n \in \chi_1 \mid H_0)$$

is small.

Definition

A criterion is said to have **a significance level** α if

$$\mathbb{P}(X^n \in \chi_1 \mid H_0) \leq \alpha.$$

Significance level corresponds to the probability of rejecting the null when it is true.

Such an error is usually called Type I error.

Significance Level

Independent of the criterion, we can make

- the right decision;
  or
- make one of two errors: Type I or Type II.

This is illustrated in the table below:

		True hypothesis	
		H_0	H_1
Criterion application result	H_0	OK	Type II error
	H_1	Type I error	OK

- Type I error is usually called “false positive” (in Russian usually “false alarm”);
- Type II error is usually called “false negative” (in Russian usually “missed target”);

Power Function

- Let the criterion have a critical region χ_1.
- Let $\mathcal{F} = F_0 \cup F_1$ be the set of all admissible distributions of the sample X^n.
- Let F_0 (respectively, F_1) be the set of distributions satisfying the hypothesis H_0 (respectively, H_1).

Definition

Functional

$$W(F) = W(F; \chi_1) = \mathbb{P}(X^n \in \chi_1 \mid F), \quad F \in \mathcal{F}$$

is called **the power function** of a test.

In other words, the power function of the test shows the probability that the observed value of X^n falls into the critical region χ_1, when F is its true distribution.

Size of Test

It is easy to express the probabilities of both error types through the power function:

- $W(F)$ is **Type I error probability** given $F \in F_0$.
- $1 - W(F)$ is **Type II error probability** given $F \in F_1$.

Definition

Size of a test:

$$\alpha = \sup_{F \in F_0} W(F).$$

We conclude, that if the size of the criterion does not exceed α, then its level is α.

How to Construct a Good Test?

- It is logical to strive to construct the criterion so as to minimize the probability of errors of both types.
- However, for a fixed sample size, the sum of the probabilities of both types of errors cannot be made arbitrarily small.
- Therefore, we are guided by a rational principle of choosing a critical area.

Definition

Of all the critical regions satisfying a given level of significance, the one for which the probability of a Type II error is minimal is selected.

Example: Testing Mean of Normal Distribution

Example

- Consider $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where σ is fixed.
- Consider 2 hypotheses $H_0\colon \mu \leq 0$ vs. $H_1\colon \mu > 0$.
- Consider the criterion – H_0 rejected if $T = \bar{X} > c$.
- Then the critical region $\chi_1 = \{(x_1, \ldots, x_n)\colon T(x_1, \ldots, x_n) > c\}$.
- Hence,

$$W(\mu) = \mathbb{P}_\mu(\bar{X} > c) = \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right)$$

$$= \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right),$$

where Z is a random variable with standard normal distribution.

Example: Testing Mean of Normal Distribution

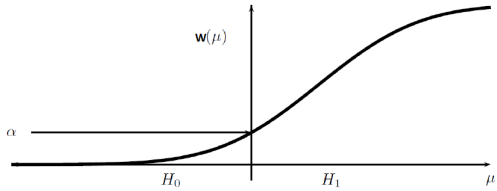


Figure: $W(\mu)$

Example

Hence it is easy to see that the size of the criterion is $W(0) = 1 - \Phi(\frac{\sqrt{n}c}{\sigma})$.

For the criterion size to be α, we must set $c = c_\alpha = \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}$.

Most Powerful Criterion

Definition

The most powerful criterion – is the criterion of maximum power relative to the hypothesis H_1 among all criteria of size α.
In other words, it is the criterion having the minimum Type II error $1 - W(F)$ for any $F \in F_1$.

- In specific tasks, the most powerful criterion is not always achievable.
- Therefore, in real tasks, it is often necessary to restrict ourselves to more moderate requirements.

Properties of Hypothesis Tests

- The minimum such requirement is the requirement of unbiasedness.

Definition

The statistical test is called **unbiased**, if for any alternative data distribution we should fall into the critical region with a greater probability than under the null hypothesis.

- In the case of large samples, the consistency condition is also important.

Definition

The statistical test is called **consistent**, if, in the case of

1. the alternative hypothesis H_1 is true,
2. and the large number of observations is provided,

we will fall into the critical region with a probability close to 1
(i.e., rejecting the null hypothesis, we will make the correct decision).

Usual Form of the Statistical Criterion

- In the example above the critical region was defined as

$$\chi_1 = \{(x_1, \ldots, x_n) \colon T(x_1, \ldots, x_n) > c\},$$

  i.e. it is completely determined by the values of the single statistic

$$T(X^n) = T(x_1, \ldots, x_n).$$

- The majority of the statistical criteria are of this type.
- It is convenient to redefine the critical region in terms of the statistics and corresponding significance level α:

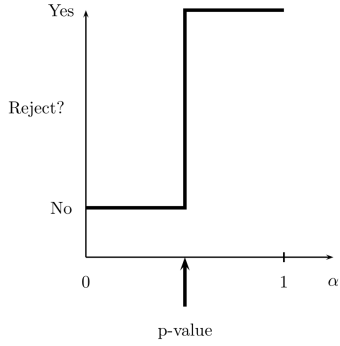
$$R_\alpha = \{t \colon \quad t = T(X^n) > c_\alpha\}.$$

P-value

Definition

Let for each $\alpha \in (0, 1)$ there is a criterion of size α corresponding to some statistic $T(X^n)$ with critical region R_α.
Then

$$\text{p-value} = \inf\{\alpha \colon T(X^n) \in R_\alpha\}.$$



- Thus, p-value is the lowest level of significance at which H_0 can still be rejected.
- The smaller the p-value, the more likely H_0 should be rejected.

P-value

Typical values for $\mathrm{p\text{-}value}$:

- $< 0.01 \rightarrow H_0$ is obviously not true;
- $0.01 - 0.05 \rightarrow H_0$ is not true;
- $0.05 - 0.10 \rightarrow H_0$ is rather not true;
- $> 0.1 \rightarrow$ nothing definite can be said about the hypothesis H_0.

The large $\mathrm{p\text{-}value}$ does not confirm the hypothesis H_0.

A large $\mathrm{p\text{-}value}$ appears if:

- H_0 is true;
- H_0 is not true, but the power of the criterion is insufficient.

P-values in Action

Theorem

Let the criterion of size α, constructed for the statistic $T(X^n)$, has the following form: H_0 is rejected if $T(X^n) > c_\alpha$.

Then

$$\text{p-value} = \sup_{F \in F_0} \mathbb{P}(T(X^n) \geq T(x^n) \mid F),$$

where x^n is an observation of X^n.

If $F_0 = F$, then

$$\text{p-value} = \mathbb{P}(T(X^n) \geq T(x^n) \mid F_0).$$

That is, p-value is the probability (when the hypothesis H_0 is fulfilled) that the statistic $T(X^n)$ will take a value greater than or equal to the one that was observed in practice (observation x^n).

Examples of Statistical Tests

After a short introduction to the theory of hypothesis testing, let us turn directly to various statistical criteria.

In this lecture, the following statistical criteria will be considered:

- Hypothesis Testing for Normal Distribution.
- Student's t-test.
- Wald's criterion.

Hypothesis Testing for Normal Distribution

- Consider $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where σ is fixed.
- We have already considered the composite hypothesis:

$$H_0 \colon \mu \leq 0 \quad \text{vs.} \quad H_1 \colon \mu > 0.$$

- What if we focus on the simple hypothesis:

$$H_0 \colon \mu = \mu_0 \quad \text{vs.} \quad H_1 \colon \mu \neq \mu_0.$$

- Two important cases:
  - σ^2 is known;
  - σ^2 is unknown;

Case of known σ^2

Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where

- parameter μ is unknown;
- parameter σ^2 is known.

Problem:

$$H_0 \colon \mu = \mu_0 \quad \text{vs.} \quad H_1 \colon \mu \neq \mu_0.$$

Then the criterion statistics:

$$T(X^n) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}.$$

Under the null hypothesis:

$$T(X^n) \sim \mathcal{N}(0, 1).$$

The null hypothesis is rejected if $|T(X^n)| > z_{\alpha/2}$.

Student's t-distribution

Definition (Student's t-distribution)

Let $Z_0, Z_1, \ldots, Z_k$ be independent standard normal random variables.
Consider

$$V = \frac{Z_0}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} Z_i^2}},$$

then V is distributed according Student's distribution (t-distribution) with k degrees of freedom.

The density of Student's distribution with k degrees:

$$f(v) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{v^2}{k})^{\frac{k+1}{2}}}.$$

For $k = 1$, the t-distribution coincides with the Cauchy distribution.

Student's t-test

Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where parameters (μ, σ^2) are unknown.

$$H_0 \colon \mu = \mu_0 \quad \text{vs.} \quad H_1 \colon \mu \neq \mu_0.$$

Let S_n^2 denote the sample variance. Then the criterion statistics

$$T(X^n) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$$

has the Student's t-distribution with $n - 1$ degrees of freedom.

The main hypothesis is rejected if

$$|T(X^n)| > t_{n-1,\alpha/2},$$

where $t_{n-1,\alpha/2}$ is the quantile of the Student's t-distribution.

Wald's Criterion

Let

- θ — scalar parameter;
- $\widehat{\theta}_n$ — its estimate;
- $\widehat{se}$ — standard error estimate $\widehat{\theta}_n$.

Hypothesis:

$$H_0 \colon \theta = \theta_0 \quad \text{vs.} \quad H_1 \colon \theta \neq \theta_0.$$

Definition (Wald's criterion of size α)

If $\widehat{\theta}_n$ is an asymptotically normal parameter estimate θ, i.e.

$$W = \frac{\widehat{\theta}_n - \theta_0}{\widehat{se}} \to \mathcal{N}(0, 1), \quad n \to \infty,$$

then the hypothesis H_0 is rejected if $|W| > z_{\alpha/2}$.

Wald's Criterion

Theorem

Asymptotically, the size of the Wald's test is α, i.e.

$$W(\theta) = \mathbb{P}(|W| > z_{\alpha/2} \mid \theta_0) \to \alpha, \quad n \to \infty.$$

Proof.

Provided that $\theta = \theta_0$, due to the asymptotic normality of the estimate, $\frac{\widehat{\theta}_n - \theta_0}{\widehat{se}} \rightsquigarrow \mathcal{N}(0,1)$. Therefore, the probability of rejecting the main hypothesis when it is actually correct is:

$$\mathbb{P}(|W| > z_{\alpha/2} \mid \theta_0) = \mathbb{P}\left(\frac{|\widehat{\theta}_n - \theta_0|}{\widehat{se}} > z_{\alpha/2} \mid \theta_0\right) \to$$
$$\to \mathbb{P}(|Z| > z_{\alpha/2}) = \alpha.$$

Example: Comparison of Mean Values

- Consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ – two independent samples from general populations.
- Their means equal μ_1 and μ_2 correspondingly.
- s_1^2 and s_2^2 — sample variances.
- Let $\delta = \mu_1 - \mu_2$.

$$H_0 \colon \delta = 0 \quad \text{vs.} \quad H_1 \colon \delta \neq 0; \qquad \widehat{\delta} = \bar{X} - \bar{Y}; \qquad \widehat{se} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

Hypothesis H_0 is rejected if $|W| > z_{\alpha/2}$, where

$$W = \frac{\widehat{\delta} - 0}{\widehat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}.$$

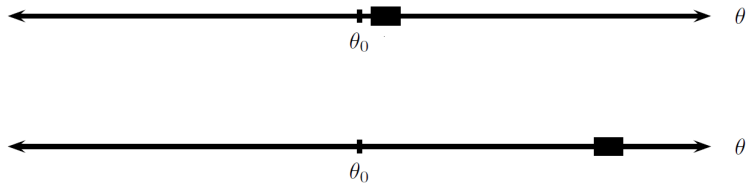
Wald's Criterion

Theorem

Wald's criterion of size α rejects the hypothesis $H_0 \colon \theta = \theta_0$ in favor of $H_1 \colon \theta \neq \theta_0$, if and only if $\theta_0 \notin C_n$, where

$$C_n = (\widehat{\theta}_n - \widehat{se}\ z_{\alpha/2}, \widehat{\theta}_n + \widehat{se}\ z_{\alpha/2}).$$

Thus, testing a hypothesis is equivalent to checking whether the value of θ_0 falls within the confidence interval.



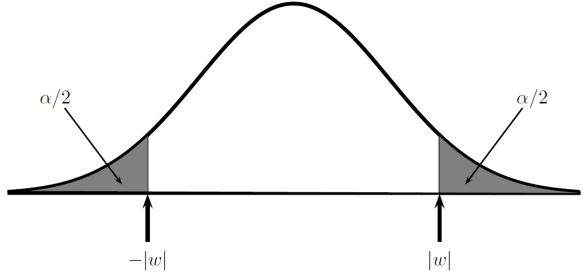
P-values for Wald's criterion

Theorem

Let $w = \frac{\widehat{\theta}_n - \theta_0}{\widehat{se}}$ *be an observed value of Wald's statistic* W.
Then

$$\text{p-value} = \mathbb{P}(|W| > |w| \mid \theta_0) \simeq \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|),$$

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



Thank you for your attention!