STATISTICAL INFERENCE

Adair Antonio da Silva Neto

January 30, 2023

Contents

Chapter 1

Basic Concepts

1.1 Samples, Statistics, and Estimators

What is statistical inference? Consider X a random variable (r.v.) with probability density function (p.d.f.) or probability function (p.f.) $f(x \mid \theta)$ in which θ is an unknown parameter. **Statistical inference** consists of specifying one or more values for θ given a set of observed values of X.

In this text we deal with two kinds of problems:

1. In a **estimation** problem, the goal is to find, using some specified criteria, values that adequately represent the unknown parameters.

2. In a **hypothesis testing** problem, the goal is to verify the validity of claims about one or more values of the unknown parameter.

To deal with these problems, first, it is necessary to introduce the concepts of population and random sample.

Definition 1.1.1 (Population and Sample). **Population** is the set of values of an observable characteristic associated with a collection of individuals or objects of interest. Any subset of a population is called a **sample**.

Definition 1.1.2 (Random Sample). A sequence $X_1, \ldots, X_n$ of independent and identically distributed (i.i.d.) random variables with p.d.f. or p.f. $f(x \mid \theta)$ is called a **random sample** with size n of the distribution of X.

The idea is to use the sample $X_1, \ldots, X_n$ to gain information about the parameter θ.

Definition 1.1.3 (Likelihood Function). The joint density (or probability) function

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = f(x_1 \mid \theta) \cdots f(x_n \mid \theta)$$

is called the **likelihood function** of θ corresponding to the observed sample $x = x_1, \ldots, x_n$

and is denoted by

$$\mathrm{L}(\theta, x) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Definition 1.1.4 (Statistic)**.** Any function of the sample that does not depend on unknown parameters is called a **statistic**.

Example 1.1.1. Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be a random sample of the r.v. X with p.d.f. (or p.f.) $f(x \mid \theta)$. The following are examples of statistics:

1. The sample minimum $\mathrm{X}_1 = \min(\mathrm{X}_1, \ldots, \mathrm{X}_n)$;
2. The sample maximum $\mathrm{X}_n = \max(\mathrm{X}_1, \ldots, \mathrm{X}_n)$;
3. The sample median $\tilde{\mathrm{X}} = \mathrm{med}(\mathrm{X}_1, \ldots, \mathrm{X}_n)$;
4. The sample mean $\overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i$;
5. The sample variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{X}_i - \overline{\mathrm{X}})^2$.

Definition 1.1.5 (Parameter Space)**.** The set Θ in which θ assumes its values is called a **parameter space**.

Definition 1.1.6 (Point estimator)**.** A **point estimator** (or simply an **estimator**) for θ is any statistic that assumes values on Θ.

It is usual in many situations to estimate a function $g(\theta)$. Naturally, a point estimator for $g(\theta)$ is any statistic that assumes values on the possible values for $g(\theta)$.

One of the main goals of statistics is to find a ‘reasonable’ point estimator for an unknown parameter θ or a function of it $g(\theta)$. A common procedure to assess the quality of a point estimator is its mean squared error (MSE).

Definition 1.1.7 (Mean Squared Error)**.** The **mean squared error** of a point estimator $\hat{\theta}$ for the parameter θ is given by

$$\mathrm{MSE}[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Definition 1.1.8 (Bias)**.** The **bias** of an estimator $\hat{\theta}$ is defined as

$$\mathrm{B}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

We say that an estimator $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}[\hat{\theta}] = \theta$ for all $\theta \in \Theta$, i.e., $\mathrm{B}(\hat{\theta}) = 0$.

If

$$\lim_{n \to \infty} \mathrm{B}(\hat{\theta}) = 0, \quad \forall\ \theta \in \Theta$$

we say that the point estimator $\hat{\theta}$ is **assintotically unbiased** for θ.

The next result shows an easier way to compute the mean squared error.

Theorem 1.1.1. Let $\hat{\theta}$ be a point estimator for θ. Then

$$\mathrm{MSE}[\hat{\theta}] = \mathrm{Var}[\hat{\theta}] + \mathrm{B}^2(\hat{\theta})$$

Remark that if $\hat{\theta}$ is an unbiased estimator for θ, then $\mathrm{MSE}[\hat{\theta}] = \mathrm{Var}[\hat{\theta}]$.

Using the MSE to compare estimators, we say that $\hat{\theta}_1$ is **better** than $\hat{\theta}_2$ if

$$\mathrm{MSE}[\hat{\theta}_1] \leq \mathrm{MSE}[\hat{\theta}_2]$$

for all θ, replacing $\leq$ with $<$ for at least one value of θ. In this case, the point estimator $\hat{\theta}_2$ is said to be **inadimissible**.

If there exists an estimator $\hat{\theta}^*$ such that for all $\hat{\theta} \neq \hat{\theta}^*$,

$$\mathrm{MSE}[\hat{\theta}^*] \leq \mathrm{MSE}[\hat{\theta}]$$

for all θ, replacing $\leq$ with $<$ for at least one value of θ, we say that $\hat{\theta}^*$ is **optimal** for θ. Notice that if the estimators are unbiased, then $\hat{\theta}^*$ is the **uniformly minimum-variance unbiased estimator** if

$$\mathrm{Var}[\hat{\theta}^*] \leq \mathrm{Var}[\hat{\theta}]$$

for all θ, replacing $\leq$ with $<$ for at least one value of θ.

Chapter 2

Efficient Estimators and Sufficient Statistics

In this chapter, we'll find the lower bound of the variance of unbiased estimators, which will be called an efficient estimator. However, these are found only for the exponential family of distributions.

Moreover, what criteria can we impose on the estimator in order to find an optimal one? Using the smallest MSE, we will see that the point estimator should be a function of sufficient statistics, which can be understood as statistics that condense the data without losing information, i.e., they are as informative of the parameter (or distribution) as the whole sample.

2.1 Efficient Estimators

Definition 2.1.1 (Efficient Estimators)**.** The **efficiency** of an unbiased point estimator $\hat{\theta}$ for the parameter θ is

$$e(\hat{\theta}) = \frac{\mathrm{LI}(\theta)}{\mathrm{Var}[\hat{\theta}]}$$

where $\mathrm{LI}(\theta)$ is the lower bound of the variance of the unbiased estimators of θ.

When $\mathrm{LI}(\theta) = \mathrm{Var}[\hat{\theta}]$ we say that $\hat{\theta}$ is **efficient**. In this case, $e(\hat{\theta}) = 1$.

To compute $\mathrm{LI}(\theta)$, the following result is useful.

Theorem 2.1.1. If the support $\mathrm{A}(x) = \{x : f(x \mid \theta) > 0\}$ is independent of θ and it is possible to exchange the order of operations of derivation and integration under the distribution of the r.v. X, then

$$\mathrm{LI}(\theta) = \frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(\mathrm{X}|\theta)}{\partial \theta}\right)^2\right]}$$

The conditions above will be called **regularity conditions**.

Definition 2.1.2 (Score function and Fisher information)**.** The **score function** is defined as

$$s(\mathrm{X},\theta) = \frac{\partial \ln f(\mathrm{X} \mid \theta)}{\partial \theta}$$

The **Fisher information** is defined as

$$\mathrm{I}_{\mathrm{F}}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln f(\mathrm{X} \mid \theta)}{\partial \theta}\right)^2\right]$$

Theorem 2.1.2. Under the regularity conditions, the expected value of the score function equals zero, i.e.,

$$\mathbb{E}\left[\frac{\partial \ln f(\mathrm{X} \mid \theta)}{\partial \theta}\right] = 0$$

Consequently,

$$\mathrm{I}_{\mathrm{F}}(\theta) = \mathrm{Var}\left[\frac{\partial \ln f(\mathrm{X} \mid \theta)}{\partial \theta}\right]$$

Another important result states that

Theorem 2.1.3.

$$\mathbb{E}\left[\left(\frac{\partial \ln f(\mathrm{X} \mid \theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln f(\mathrm{X} \mid \theta)}{\partial \theta^2}\right]$$

Using this result, it can be easily verified that, given a random sample, the total Fisher information of θ with respect to (w.r.t.) the observed sample is the sum of the Fisher information of the n observations, i.e., $n\mathrm{I}_{\mathrm{F}}(\theta)$.

Theorem 2.1.4 (Information Inequality (Cramér-Rao))**.** Under the regularity conditions,

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{n\mathrm{I}_{\mathrm{F}}(\theta)}$$

This inequality shows a way to verify whether a given estimator is efficient.

2.2 Sufficient Statistics

Definition 2.2.1 (Sufficient Statistic)**.** The statistic $\mathrm{T} = \mathrm{T}(\mathrm{X}_1, \dots, \mathrm{X}_n)$ is **sufficient** for θ if the conditional distribution of $\mathrm{X}_1, \dots, \mathrm{X}_n$ given θ is independent of θ.

Intuitively, T contains all information that the random sample $\mathrm{X}_1, \dots, \mathrm{X}_n$ has about θ. To obtain sufficient statistics, we may use the following theorem.

Theorem 2.2.1 (Neyman Factorization Criterion)**.** Let $\mathrm{X}_1, \dots, \mathrm{X}_n$ be a random sample of the distribution of the r.v. X with p.d.f. (or p.f.) $f(x \mid \theta)$ and likelihood function $\mathrm{L}(\theta, x)$. Then the statistics $\mathrm{T} = \mathrm{T}(\mathrm{X}_1, \dots, \mathrm{X}_n)$ is sufficient for θ if, and only if, (iff.)

$$\mathrm{L}(\theta, x) = h(x_1, \dots, x_n) g_\theta(\mathrm{T}(x_1, \dots, x_n))$$

with h depending only on $x_1,\dots,x_n$ and g_θ depending on θ and $x_1,\dots,x_n$ only through T.

What happens when our parameter θ is a vector?

Theorem 2.2.2 (Multiparametric Factorization Criterion)**.** Let $X_1,\dots,X_n$ be a random sample of the distribution of the r.v. X with p.d.f. (or p.f.) $f(x \mid \theta)$ and likelihood function $L(\theta,x)$. Then the r-dimensional statistics $T = (T_1,\dots,T_r)$, with $T_i = (X)$, is jointly sufficient for θ iff.

$$L(\theta,x) = h(x_1,\dots,x_n)g_\theta(T_1(x),\dots,T_r(x))$$

with h depending only on $x_1,\dots,x_n$ and g_θ depending on θ and $x = (x_1,\dots,x_n)$ only through T.

In many cases, the dimension r equals the dimension of the parameter space Θ. However, there are some situations in which the dimension of Θ is smaller than r.

Definition 2.2.2 (Equivalent Statistics)**.** Two statistics T_1 and T_2 are **equivalent** if there is a bijection between them.

Note that if T_1 and T_2 are equivalent and T_1 is sufficient for θ, then T_2 is also sufficient for θ. The same holds in the multidimensional case.

2.3 Exponential Families

The Binomial, Normal, Exponential and Poisson (and others!) can be considered special cases of a more general family of distributions.

Definition 2.3.1 (Exponential Family)**.** A distribution of a r.v. X belongs to the **unidimensional exponential family** of distributions if we can write its p.d.f. (or p.f.) as

$$f(x \mid \theta) = e^{c(\theta)T(x)+d(\theta)+S(x)}, \quad x \in A \tag{2.1}$$

in which $c, d,$ T and S are real-valued functions.

Note that $d(\theta)$ is associated with the normalization constant of the density.

Theorem 2.3.1. Let $X_1,\dots,X_n$ be a random sample of the distribution of the r.v. X belonging to the exponential family, i.e., with p.d.f. (or p.f.) given by (2.1). Then the joint distribution of $X_1,\dots,X_n$ is given by

$$f(x_1,\dots,x_n \mid \theta) = e^{c^*(\theta)\sum_{i=1}^n T(x_i)+d^*(\theta)+S^*(x)} \tag{2.2}$$

which also belong to the exponential family with

$$T(X) = \sum_{i=1}^n T(x_i), \quad c^*(\theta) = c(\theta),$$

$$d^*(\theta) = nd(\theta), \quad \text{and } S^*(x) = \sum_{i=1}^n S(x_i)$$

Notice that by (2.2) and considering

$$h(x_1,\dots,x_n) = e^{\sum_{i=1}^{n} \mathrm{S}(x_i)} \prod_{i=1}^{n} \chi_{\mathrm{A}}(x_i) \quad \text{and} \quad g_\theta(\mathrm{T}) = e^{c(\theta)\sum_{i=1}^{n} \mathrm{T}(x_i)+nd(\theta)}$$

it follows, by the factorization criterion, that $\mathrm{T}(\mathrm{X}) = \sum_{i=1}^{n} \mathrm{T}(\mathrm{X}_i)$ is the sufficient statistic for θ.

Generalizing the concept of exponential family for the multidimensional case, we have the following definition.

Definition 2.3.2 (Exponential Family (Multidimensional case)). A distribution of an r.v. X belongs to the *k*-**dimensional exponential family** of distributions if we can write its p.d.f. (or p.f.) as

$$f(x \mid \theta) = e^{\sum_{j=1}^{k} c_j(\theta)\mathrm{T}_j(x)+d(\theta)+\mathrm{S}(x)}, \quad x \in \mathrm{A} \tag{2.3}$$

in which c_j, T_j, d and S are real-valued functions and $j = 1,\dots,k$.

As in the unidimensional case, samples of k-dimensional exponential families have distributions that belong to the k-dimensional exponential family.

Theorem 2.3.2. Let $\mathrm{X}_1,\dots,\mathrm{X}_n$ be a random sample of the distribution of the r.v. X belonging to the k-dimensional exponential family, i.e., with p.d.f. (or p.f.) given by (2.3). Then the joint distribution of $\mathrm{X}_1,\dots,\mathrm{X}_n$ is given by

$$f(x_1,\dots,x_n \mid \theta) = e^{\sum_{j=1}^{k} c_j^*(\theta)\sum_{i=1}^{n} \mathrm{T}_j(x_i)+d^*(\theta)+\mathrm{S}^*(x)} \tag{2.4}$$

which also belongs to the exponential family with

$$\mathrm{T}_j(\mathrm{X}) = \sum_{i=1}^{n} \mathrm{T}_j(x_i), \quad c_j^*(\theta) = c_j(\theta),$$

$$d^*(\theta) = nd(\theta), \quad \text{and } \mathrm{S}^*(x) = \sum_{i=1}^{n} \mathrm{S}(x_i)$$

In this case, $(\mathrm{T}_1,\dots,\mathrm{T}_k)$ is jointly sufficient for θ.

2.4 Estimators Based on Sufficient Statistics

Let $\mathrm{T} = \mathrm{T}(\mathrm{X}_1,\dots,\mathrm{T}_n)$ be a sufficient statistic for θ and $\mathrm{S} = \mathrm{S}(\mathrm{X}_1,\dots,\mathrm{X}_n)$ be an estimator for θ that is not a function of T. Then

$$\hat{\theta} = \mathbb{E}[\mathrm{S} \mid \mathrm{T}] \tag{2.5}$$

is an estimator for θ, i.e., a function of T that does not depend on θ. In fact, since T is sufficient, the conditional distribution of $\mathrm{X}_1,\dots,\mathrm{X}_n$ given T is independent of θ. Also notice that S is a function only of $\mathrm{X}_1,\dots,\mathrm{X}_n$.

Moreover, if S is an unbiased estimator for θ, then $\hat{\theta}$ is also unbiased for θ.

Theorem 2.4.1 (Rao-Blackwell)**.** If S is an unbiased estimator for θ, then

$$\mathrm{Var}[\hat{\theta}] \leq \mathrm{Var}[\mathrm{S}], \quad \forall\ \theta \tag{2.6}$$

Using the concept of complete statistics with the definition of sufficiency, we can obtain an optimal estimator, i.e., the unbiased estimator with uniform minimum variance.

Definition 2.4.1 (Complete Statistics)**.** A statistic $\mathrm{T} = \mathrm{T}(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ is said to be **complete** w.r.t. a family $f(x \mid \theta) : \theta \in \Theta$ if the only real-valued function g defined on the domain of T such that $\mathbb{E}[g(\mathrm{T})] = 0$, for all θ, is the zero function, i.e., $g(\mathrm{T}) \equiv 0$ with probability one.

Theorem 2.4.2. Suppose that X has a distribution belonging to the k-dimensional exponential family. Then the statistic

$$\mathrm{T}(\mathrm{X}) = \left(\sum_{i=1}^{n} \mathrm{T}_1(\mathrm{X}_i), \ldots, \sum_{i=1}^{n} \mathrm{T}_k(\mathrm{X}_i) \right)$$

is sufficient for θ. If the variation domain of $(c_1(\theta), \ldots, c_k(\theta))$ contains a k-dimensional rectangle, then the statistic T(X) is also complete.

Notice that in the unidimensional case, the variation domain of $c(\theta)$ must contain an interval of the real line. In the bidimensional case, a square, and so on.

Theorem 2.4.3 (Lehmann-Scheffé)**.** Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be a random sample of the distribution of the r.v. X with p.d.f. (or p.f.) $f(x \mid \theta)$, T a sufficient and complete statistic, and S an unbiased estimator for θ. Then $\hat{\theta} = \mathbb{E}[\mathrm{S} \mid \mathrm{T}]$ is the unique unbiased estimator of θ based on T and is the **uniformly minimum-variance unbiased estimator** (UMVUE) for θ.

Chapter 3

Estimation Methods

In this chapter, we tackle the question of how to fabricate estimators. Two methods will be considered: the maximum likelihood and the method of moments.

3.1 Maximum Likelihood

The definition of the maximum likelihood estimator is straightforward.

Definition 3.1.1 (Maximum Likelihood Estimator)**.** The **maximum likelihood estimator** (MLE) of θ is the value $\hat{\theta} \in \Theta$ that maximizes the likelihood function $\mathrm{L}(\theta, x)$.

We denote the **log-likelihood function** of θ by

$$l(\theta, x) = \ln \mathrm{L}(\theta, x)$$

It can be easily verified that the value of θ that maximizes $\mathrm{L}(\theta, x)$ also maximizes $l(\theta, x)$. In the uniparametric case, in which Θ is an interval of the real line, and $l(\theta, x)$ can be derived, the maximum likelihood estimator can be obtained as the root of the **likelihood equation**:

$$l'(\theta, x) = \frac{\partial l(\theta, x)}{\partial \theta} = 0 \tag{3.1}$$

To verify that the solution of this equation is a maximum point, it is necessary to verify that

$$l''(\theta, x) = \left. \frac{\partial^2 \ln \mathrm{L}(\theta, x)}{\partial \theta^2} \right|_{\theta = \hat{\theta}} < 0 \tag{3.2}$$

In more complicated cases, the solution to (3.1) is obtained using numerical methods. If Θ is discrete or the maximum of $l(\theta, x)$ occurs at the borders of Θ, then we have to inspect the graph of the likelihood function.

One of the numerical methods is the **score method**. Let $\mathrm{U}(\theta)$ denote the score function. By the (3.1), we know that for the maximum likelihood estimator $\hat{\theta}$, $\mathrm{U}(\hat{\theta}) = 0$. Using the Newton-Raphson method, we obtain the following iterative procedure

$$\theta_{j+1} = \theta_j - \frac{\mathrm{U}(\theta_j)}{\mathrm{U}'(\theta_j)} \tag{3.3}$$

In some cases, replacing $\mathrm{U}'(\theta_j)$ on (3.3) by $\mathbb{E}[\mathrm{U}'(\theta_j)]$, i.e., the Fisher information on θ_j multiplied by -1, may render a simplified procedure.

3.2 Properties of Maximum Likelihood Estimators

The next theorems state some properties of the MLE.

Theorem 3.2.1 (MLE is a function of a sufficient statistic)**.** Let $T = T(X_1,\dots,X_n)$ be a sufficient statistic for θ. Then the maximum likelihood estimator $\hat{\theta}$, if it exists, is a function of T.

The proof follows immediately from the factorization criterion.

Theorem 3.2.2 (Invariance Principle)**.** Let g be an invertible function defined on Θ. If $\hat{\theta}$ is a maximum likelihood estimator of θ, then $g(\hat{\theta})$ is a maximum likelihood estimator for $g(\theta)$.

In large samples, the MLE of θ and $g(\theta)$ are approximately unbiased, whose variance coincides with the corresponding lower bounds of the variance of the unbiased estimators of θ and $g(\theta)$. Put another way, in large samples, the MLE is efficient.

Theorem 3.2.3 (Distribution in Large Samples)**.** If the sample is large and the regularity conditions are met, then

$$\sqrt{n}(\hat{\theta}-\theta) \overset{a}{\sim} N\left(0, \frac{1}{I_F(\theta)}\right)$$

and

$$\sqrt{n}(g(\hat{\theta})-g(\theta)) \overset{a}{\sim} N\left(0, \frac{(g'(\theta))^2}{I_F(\theta)}\right)$$

where $\overset{a}{\sim}$ means assymptotic distribution.

Theorem 3.2.4 (Likelihood for Independent Samples)**.** If $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ are independent random samples related to a parameter of interest, then

$$L(\theta,x,y) = L(\theta,x)L(\theta,y)$$

and

$$l(\theta,x,y) = l(\theta,x) + l(\theta,y)$$

3.3 Multiparametric Case

In the multiparametric case, under the regularity conditions, we can obtain the MLE of $\theta_1,\dots,\theta_r$ by solving the equations

$$\frac{\partial \ln L(\theta,x)}{\partial \theta_i} = 0, \quad i = 1,\dots,r$$

If the support of the distribution of X depends on θ or the maximum occurs at the borders of Θ, then it is necessary to inspect the graph of the likelihood function.

In the case of only two parameters, using the equation

$$\frac{\partial \ln L(\theta_1,\theta_2,x)}{\partial \theta_1} = 0$$

we obtain a solution for θ_1 as a function of θ_2, which we denote by $\hat{\theta}_1(\theta_2)$. Substituting the solution for θ_1 in the joint likelihood, we have a function that depends only on θ_2, i.e.,

$$g(\theta_2, x) = l(\hat{\theta}_1(\theta_2), \theta_2, x)$$

This function can be used to obtain an MLE for θ_2, thus reducing the problem to the uniparametric case.

3.4 Maximum Likelihood in the Exponential Family

If the distribution of X belongs to the unidimensional exponential family, then the MLE of θ based on the random sample $X = (X_1, \ldots, X_n)$ is the solution of

$$\mathbb{E}[T(X)] = T(X)$$

For the k-parametric case, the MLE for $\theta_1, \ldots, \theta_k$ follow from the solution of the equations

$$\mathbb{E}[T_j(X)] = T_j(X), \quad j = 1, \ldots, k$$

3.5 The Method of Moments

Let $r \geq 1$ and consider

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r \quad \text{and} \quad \mu_r = \mathbb{E}[X^r]$$

the r-th sample moment of a random sample $X_1, \ldots, X_r$ and the r-th population moment, respectively.

The method of moments consists in obtaining estimators for $\theta = (\theta_1, \ldots, \theta_k)$ solving the equations

$$m_r = \mu_r, \quad r = 1, \ldots, k$$

3.6 Consistent Estimators

A consistent estimator is one that, as the sample size gets bigger, the estimators get as close to the parameters as we want. To define it, we use the concept of convergence in probability.

Definition 3.6.1 (Consistent Estimator)**.** We say that the estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is **consistent** with the parameter θ if

$$\lim_{n\to\infty} P[\,|\,\hat{\theta} - \theta\,| > \varepsilon\,] = 0$$

To verify this property, the Chebyshev inequality is often used.

Chapter 4

Decision Theory: the Minimax and Bayes Principles

In this chapter, we study the problems of estimation and hypothesis testing with Game Theory, in which the adversaries are the statistician and nature.

Our goal is to minimize the risk function, which is dual to the problem of maximizing the utility function, in the language of economics.

It is not always possible to minimize the risk function uniformly on θ. Thus, we'll study two alternative procedures, the minimax (to avoid the maximum risk) and the Bayes principle, which assumes that nature uses a random procedure.

4.1 Basic Elements

A decision problem consists of:

1. A nonempty set Θ as the state space, which is the parameter space. Nature chooses a value θ in this set.

2. A nonempty set $\mathfrak{A}$ of all possible actions that the statistician may take. In an estimation problem, $\mathfrak{A} = \Theta$. If the problem is hypothesis testing, $\mathfrak{A}$ generally consists of either accepting or rejecting the formulated hypothesis.

3. A function $d : \mathfrak{X} \longrightarrow \mathfrak{A}$, called **decision function** (or **procedure**), in which $\mathfrak{X}$ is the sample space associated with a r.v. X, corresponding to an experiment idealized by the statistician to obtain information about the choice of θ done by nature. We define $\mathfrak{D}$ as the class of all possible decision functions. In this class, the statistician looks for a procedure that is 'better' according to some criteria.

4. A real valued function $l(\theta, a)$, defined on $\Theta \times \mathfrak{A}$, called the **loss function**, satisfying $l(\theta, a) \geq 0$, for all $\theta \in \Theta$ and $a \in \mathfrak{A}$, with equality when $a = \theta$ (i.e., the correct action is taken).

Intuitively, the function $l(\theta, a)$ represents the loss incurred by the statistician by taking action a when θ is the choice made by nature. The loss function is also called **error function** or **cost function**.

Example 4.1.1 (Loss Functions)**.** The following are examples of often-used loss functions.

- Squared loss: $l(\theta,a)=(\theta-a)^2$;
- Absolute loss: $l(\theta,a)=|\theta-a|$;
- $l(\theta,a)=c(\theta)|\theta-a|^r$, where $c(\theta)>0$, $r>0$.

Since the loss function depends on the unknown parameter θ, it is not possible to minimize it directly. Thus, we'll try to reduce the risk function.

Definition 4.1.1 (Risk Function)**.** The **risk function** corresponding to the procedure d and the loss function $l(\theta,a)$ is given by

$$\mathrm{R}(\theta,d)=\mathbb{E}[l(\theta,d(\mathrm{X}))]=\sum_{\{x\in\mathfrak{X}\}} l(\theta,d(x))f(x\mid\theta) \tag{4.1}$$

in the discrete case and

$$\mathrm{R}(\theta,d)=\mathbb{E}[l(\theta,d(\mathrm{X}))]=\int_{\mathfrak{X}} l(\theta,d(x))f(x\mid\theta)\,\mathrm{d}x \tag{4.2}$$

in the continuous case.

Remark that the risk function is the mean loss over the sample space $\mathfrak{X}$, and is a function of the parameter θ.

Definition 4.1.2 (Better Procedure)**.** We say that a procedure d_1 is **better** than a procedure d_2 when

$$\mathrm{R}(\theta,d_1)\le\mathrm{R}(\theta,d_2),\quad \forall\ \theta \tag{4.3}$$

and the strict inequality holds for at least one θ.

If these conditions are met, we say that the procedure d_2 is **inadimissible**. Moreover, if a procedure d_1 is better than all other procedures $d_2\in\mathfrak{D}$, we say that d_1 is the **best procedure** on $\mathfrak{D}$.

4.2 The Minimax Principle

To secure us from the maximum risk, we may use the minimax principle.

Definition 4.2.1 (Minimax Principle)**.** A procedure d_0 is said to be a **minimax procedure** on a class $\mathfrak{D}$ of procedures if

$$\sup_{\theta\in\Theta}\mathrm{R}(\theta,d_0)=\inf_{d\in\mathfrak{D}}\sup_{\theta\in\Theta}\mathrm{R}(\theta,d)$$

The idea is simply to compare the maximum risk of each procedure.

4.3 The Bayes Principle

Suppose that nature uses a random mechanism to choose a value for θ. This random procedure is represented by an **a priori distribution** with p.d.f. (or p.f.) denoted by $\pi(\theta)$.

Definition 4.3.1 (Bayes Risk)**.** The **Bayes risk** of the procedure d, w.r.t. the loss $l(\theta, d)$ is given by

$$r(\pi, d) = \mathbb{E}_{\pi}[\mathrm{R}(\theta, d)] = \sum_{\{\theta \in \Theta\}} \mathrm{R}(\theta, d)\pi(\theta)$$

in the discrete case, and

$$r(\pi, d) = \int_{\Theta} \mathrm{R}(\theta, d)\pi(\theta)\ \mathrm{d}\theta$$

in the continuous case.

Notice that if $\mathrm{R}(\theta, d)$ is constant (i.e. independent of θ), then $r(\pi, d) = \mathrm{R}(\theta, d)$.

Definition 4.3.2 (Bayes Decision Function)**.** A decision function d_{B} is said to be a **Bayes decision function** w.r.t. the a priori π and the class $\mathfrak{D}$ of decision functions if

$$r(\pi, d_{\mathrm{B}}) = \min_{d \in \mathfrak{D}} r(\pi, d)$$

Using the squared loss, it is possible to characterize the estimators in the class $\mathfrak{D}$ of all decision functions.

Theorem 4.3.1. Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be a random sample of the r.v. X, with p.d.f. $f(x \mid \theta)$. Let us consider for θ the a priori distribution with p.d.f. $\pi(\theta)$. Then, w.r.t. the squared loss, the Bayes procedure (estimator) in the class $\mathfrak{D}$ of all decision functions is given by

$$d_{\mathrm{B}}(\mathrm{X}) = \mathbb{E}[\theta \mid \mathrm{X}]$$

i.e., it is the conditional expectation of θ given $\mathrm{X}_1, \ldots, \mathrm{X}_n$, which is called the **a posteriori distribution** of θ.

Notice that since

$$f(x \mid \theta)\pi(\theta) = f(x, \theta) = \pi(\theta \mid x)g(x)$$

it follows that

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)}{g(x)} = \frac{f(x \mid \theta)\pi(\theta)}{g(x)} \tag{4.4}$$

and that

$$g(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta)\ d\theta$$

The density $\pi(\theta \mid x)$ is called the **a posteriori probability density function**.

The Theorem 4.3.1 can be generalized for a function of θ as follows

$$d_{\mathrm{B}}(x) = \mathbb{E}[\tau(\theta) \mid \mathrm{X}]$$

However, Bayes estimators are not invariant, i.e., if $\hat{\theta}$ is a Bayes estimator for θ, then $\tau(\hat{\theta})$ is not necessarily a Bayes estimator for $\tau(\theta)$.

The next result relates the Bayes estimators with sufficient statistics.

Theorem 4.3.2. Let $X_1, \ldots, X_n$ be a random sample of the r.v. X, with p.d.f. (or p.f.) $f(x \mid \theta)$. And let $T = T(X_1, \ldots, X_n)$ be a sufficient statistic for θ. We consider for θ the a priori distribution with p.d.f. (or p.f.) $\pi(\theta)$. Then the Bayes estimator of θ w.r.t. squared loss is a function of T.

In fact, for any loss function, the Bayes estimator only depends on X through T.

Chapter 5

Interval Estimates

In this chapter, we deal with the problem of parameter estimation using the concept of confidence intervals. There are two main approaches. For the classical one, we'll use special random variables called pivotal quantities. For the Bayesian approach, a posterior distribution will be used.

We start this chapter by studying some properties of the sample mean and the variance from normal populations and then we study methods to construct intervals.

5.1 Samples of Normal Populations

Theorem 5.1.1. Let $X_1, \dots, X_n$ be a random sample of the distribution $N(\mu, \sigma^2)$. Then

1. $\overline{X}$ and S^2 are independent;
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$;
3. $\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}$.

where χ^2_ν denotes a r.v. with chi-square distribution with ν degrees of freedom, i.e., with p.d.f. given by

$$f(y \mid \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}, \quad y > 0$$

and t_ν denotes a r.v. with Student's t-distribution with ν degrees of freedom, i.e., with p.d.f. given by

$$f(y \mid \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} (1 + t^2/\nu)^{-(\nu+1)/2}, \quad -\infty < t < \infty$$

and $\overline{X} = \sum_{i=1}^n X_i/n$ and $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$.

5.2 The Method of Pivotal Quantity

Definition 5.2.1 (Pivotal Quantity). A random variable $Q(X, \theta)$ is a **pivotal quantity** for a parameter θ if its distribution is independent of θ.

To construct intervals, notice that, for each fixed $\gamma = 1 - \alpha$, we can find λ_1 and λ_2 in the distribution of $Q(X, \theta)$ such that

$$P[\lambda_1 \leq Q(X, \theta) \leq \lambda_2] = \gamma \tag{5.1}$$

Since $Q(X, \theta)$ is independent of θ, it follows that λ_1 and λ_2 also do not depend on θ. Moreover, if for each X there exist $t_1(X)$ and $t_2(X)$ such that

$$\lambda_1 \leq Q(X, \theta) \leq \lambda_2 \iff t_1(X) \leq \theta \leq t_2(X)$$

then, from (5.1),

$$P[t_1(X) \leq \theta \leq t_2(X)] = \gamma \tag{5.2}$$

Thus, $[t_1(X), t_2(X)]$ is a random interval that contains θ with probability (or **confidence coefficient**) $\gamma = 1 - \alpha$.

Remark that in the discrete case, we can choose λ_1 and λ_2 such that (5.1) is satisfied for a confidence coefficient greater or equal to γ, but as close as possible. An alternative, when n is sufficiently big, is to consider the confidence intervals based on the distribution of the maximum likelihood estimator.

5.3 Intervals for Normal Populations

Let $X_1, \ldots, X_n$ be a random sample with distribution $N(\mu, \sigma^2)$. Assuming that σ^2 is known, a pivotal quantity for μ based on the sufficient statistic $\overline{X}$ is given by

$$Q(X, \mu) = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

which has distribution $N(0, 1)$.

Therefore, given a confidence coefficient γ, we can determine λ_1 and λ_2 such that

$$P\left[\lambda_1 \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq \lambda_2\right] = \gamma$$

Since the distribution $N(0, 1)$ is symmetric, the smallest interval is symmetric. Let $\lambda_1 = -z_{\alpha/2}$ and $\lambda_2 = z_{\alpha/2}$, with $P[Z \leq z_{\alpha/2}] = 1 - \alpha/2$. The smallest interval is given by

$$\left[\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

On the other hand, if σ^2 is unknown, then by the Theorem 5.1.1

$$Q(X, \mu) = \frac{\overline{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

which is a pivotal quantity.

Since the t-student distribution is symmetric, we want to find a symmetric interval. To do that, let $\lambda_1 = -t_{\alpha/2}$ and $\lambda_2 = t_{\alpha/2}$ with $P[T \le t_{\alpha/2}] = 1-\alpha/2$, $T \sim t_{n-1}$. The smallest interval is given by

$$\left[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Regarding σ^2 and considering μ unknown, by the Theorem 5.1.1,

$$Q(X, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^n_{n-1}$$

is a pivotal quantity for σ^2.

To find a symmetric interval, consider $\lambda_1 = q_1$ and $\lambda_2 = q_2$, with $P[\chi^2_{n-1} \ge q_2] = [\chi^2_{n-1} \le q_q] = \alpha/2$. The smallest interval is given by

$$\left[\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1}\right]$$

5.4 Approximate Confidence Intervals

It is possible to construct approximate confidence intervals for a parameter θ based on the asymptotic distribution of the MLE $\hat{\theta}$ of θ. Using the Theorem 3.2.3, we have that

$$\frac{\hat{\theta} - \theta}{\sqrt{(nI_F(\theta))^{-1}}} \overset{a}{\sim} N(0,1)$$

Since $I_F(\theta)$ may depend on θ, we replace it with $\hat{\theta}$, obtaining

$$Q(X, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{(nI_F(\hat{\theta}))^{-1}}} \overset{a}{\sim} N(0,1)$$

In this way, $Q(X, \theta)$ is a pivotal quantity with a distribution approximately equal to the distribution $N(0,1)$ in big samples.

W.r.t. a function $g(\theta)$, we may consider as a pivotal quantity

$$Q(X, g(\theta)) = \frac{g(\hat{\theta}) - g(\theta)}{\sqrt{\frac{(g'(\hat{\theta}))^2}{nI_F(\hat{\theta})}}} \overset{a}{\sim} N(0,1)$$

5.5 Bayesian Confidence Intervals

Now let us consider for θ an a priori density function $\pi(\theta)$. By (4.4), the a posteriori density function for θ is

$$\pi(\theta \mid X) = \frac{\prod_{i=1}^n f(x_i \mid \theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f(x_i \mid \theta)\pi(\theta)\, d\theta}$$

Definition 5.5.1 (Bayesian Confidence Interval)**.** An interval $[t_1, t_2]$ is a **Bayesian confidence interval** for θ with confidence coefficient $\gamma = 1 - \alpha$ if

$$\int_{t_1}^{t_2} \pi(\theta \mid \mathrm{X})\, d\theta = \gamma$$

The Bayesian interval of the smallest length is known as the **highest posterior density (HPD) interval**. In general, computational methods are needed to obtain an HPD interval.

Chapter 6

Hypothesis Testing

The main question of this chapter is whether to accept or reject a given claim based on a set of evidence.

6.1 Statistical Formulation

Definition 6.1.1 (Statistical Hypothesis)**.** A **statistical hypothesis** is any claim about the probability distribution of one or more random variables.

We denote by H_0, called the **null hypothesis**, the hypothesis of interest. If H_0 is rejected, we accept as true the alternative hypothesis H_1.

We'll say that the distribution of an r.v. X is **totally specified** if we know the p.d.f. (or p.f.) $f(x \mid \theta)$ and $\theta \in \Theta$. And we say that the distribution is **partially specified** if we know $f(x \mid \theta)$ but not θ.

Associated with the hypothesis H_0 and H_1, we define the sets Θ_0 and Θ_1. In this language, H_0 claims that $\theta \in \Theta_0$ and H_1 claims that $\theta \in \Theta_1$.

Definition 6.1.2 (Statistical Hypothesis Test)**.** A decision function $d : \mathfrak{X} \longrightarrow \{a_0, a_1\}$, in which a_0 corresponds to the action of considering the hypothesis H_0 as true and a_1 corresponds to the action of considering the hypothesis H_1 as true, is called a **statistical hypothesis test**.

In this definition, $\mathfrak{X}$ denotes the sample space associated with the sample $\mathrm{X}_1, \ldots, \mathrm{X}_n$ and the function d divides the space into two sets

$$\mathrm{A}_0 = \{(x_1, \ldots, x_n) \in \mathfrak{X} : d(x_1, \ldots, x_n) = a_0\}$$

and

$$\mathrm{A}_1 = \{(x_1, \ldots, x_n) \in \mathfrak{X} : d(x_1, \ldots, x_n) = a_1\}$$

Notice that A_0 and A_1 form a partition of $\mathfrak{X}$. Since A_0 contains the sample points that induce the acceptance of H_0, we call A_0 the **acceptance region** and A_1, the **rejection region** or **critical region**.

In the case that $\mathrm{H}_0 : \theta = \theta_0$ (simple) and $\mathrm{H}_1 : \theta = \theta_1$ (simple) and considering the loss function $l(\theta,d)=0$ or 1, if the correct decision is taken, then the risk function is given by

$$\begin{aligned}R(\theta_0,d)&=\mathbb{E}[l(\theta_0,d)]=P[X\in A_1\mid\theta_0]\\&=P_{H_0}[\text{Reject }H_0]=\alpha\end{aligned}$$

If the incorrect decision is taken,

$$\begin{aligned}R(\theta_1,d)&=\mathbb{E}[l(\theta_1,d)]=P[X\in A_0\mid\theta_1]\\&=P_{H_1}[\text{Accept }H_0]=\beta\end{aligned}$$

The risks α and β are known as the **type I error** and **type II error**, respectively. These errors are summarized in the following table.

Table 6.1: Types of error in Hypothesis Testing

Decision	H_0 is true	H_0 is false
Accept H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

Definition 6.1.3 (Power of a Test)**.** The **power of a test** with critical region A_1 to test $H_0:\theta=\theta_0$ against $H_1:\theta=\theta_1$ is given by

$$\pi(\theta_1)=P_{H_1}[X\in A_1]=P[X\in A_1\mid\theta_1]$$

Remark that $\pi(\theta_1)=1-\beta$, where β is the probability of committing a type II error.

6.2 More powerful tests

Fixed a probability of an error of type I, α, known as the **test level**, we may look for the critical region A_1^*, that has the smallest probability of an error of type II, i.e., the greatest power among all tests with level lesser or equal to α.

Remark that in the discrete case

$$\alpha(A_1)=P_{H_0}[X\in A_1]=\sum_{x\in A_1}f(x\mid\theta_0)\quad\text{and}\quad\beta(A_1)=\sum_{x\in A_0}f(x\mid\theta_1)$$

The next result shows a test that minimizes the linear combination of the errors of the type $a\alpha+b\beta$, with a and b known.

Lemma 6.2.1. Consider the test with the critical region

$$A_1^*=\left\{x:\frac{L_1(x)}{L_0(x)}\geq\frac{a}{b}\right\}$$

where a and b are specified and $b>0$. Then, for any other test with critical region A_1, we have that

$$a\alpha(A_1^*)+b\beta(A_1^*)\leq a\alpha(A_1)+b\beta(A_1)$$

with

$$L_1(x) = \prod_{i=1}^{n} f(x_i \mid \theta_1) \quad \text{and} \quad L_0(x) = \prod_{i=1}^{n} f(x_i \mid \theta_0) \tag{6.1}$$

We now present the **most powerful** (MP) test with level α to test H_0 against H_1.

Lemma 6.2.2 (Neyman-Pearson). Consider the test with the critical region

$$A_1^* = \left\{ x : \frac{L_1(x)}{L_0(x)} \geq k \right\}$$

with L_0 and L_1 given as in (6.1). Then A_1^* is the best critical region with level $\alpha = \alpha(A_1^*)$ to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, i.e., $\beta(A_1^*) \leq \beta(A_1)$ for any other test A_1 with $\alpha(A_1) \leq \alpha$.

This test is known as the **likelihood-ratio test (LRT)**.

Definition 6.2.1 (Descriptive Level). The **descriptive level**, denoted by $\hat{\alpha}$, is the smallest significance level α for which the null hypothesis H_0 is rejected.

6.3 Uniformly most powerful tests

First, let us consider the case of a simple null hypothesis and a composed alternative hypothesis, i.e., $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \Theta_1$.

Definition 6.3.1 (Uniformly most powerful test). A test A_1^* is said to be a **uniformly most powerful test** (UMP) to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$, if it is MP of level α to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \theta_1$, for all $\theta_1 \in \Theta_1$.

Put another way, the critical region A_1^* should not depend on θ_1 in any $\theta_1 \in \Theta_1$.

Definition 6.3.2 (Power Function). The **power function** $\pi(\theta)$ with critical region A_1^* to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$ is given by

$$\pi(\theta) = P_\theta[X \in A_1^*]$$

i.e. the probability of rejecting H_0 for $\theta \in \Theta$. Notice that $\pi(\theta_0) = \alpha$.

Now, we consider the more general case in which both hypotheses are composed, i.e., in which we test $H_0 : \theta = \Theta_0$ against $H_1 : \theta \in \Theta_1$.

Theorem 6.3.1. If $X_1, \ldots, X_n$ has a distribution that belongs to the exponential family, then the UMP to test $H_0 : \theta = \Theta_0$ against $H_1 : \theta > \theta_0$ is also a UMP to test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

Also, the UMP to test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ is UMP to test $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$.

6.4 Generalized Likelihood-ratio Tests

In this section, we consider a general case where the hypotheses of interest are $H_0 : \theta = \Theta_0$ against $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_1 \cup \Theta_2$, $\Theta_1 \cap \Theta_2 = \emptyset$, $\Theta_1 \neq \emptyset$, and $\Theta_2 \neq \emptyset$.

The **generalized likelihood-ratio test (GLRT)** can be defined as the test with critical region given by

$$A_1^* = \left\{ x : \frac{\sup_{\theta\in\Theta_1} L(\theta,x)}{\sup_{\theta\in\Theta_0} L(\theta,x)} \geq c \right\}$$

Notice that when the hypothesis are simple, the GLRT coincides with the LRT.

Since,

$$\frac{\sup_{\theta\in\Theta} L(\theta,x)}{\sup_{\theta\in\Theta_0} L(\theta,x)} = \max\left\{ 1, \frac{\sup_{\theta\in\Theta_1} L(\theta,x)}{\sup_{\theta\in\Theta_0} L(\theta,x)} \right\}$$

we can also define the GLRT as

$$A_1^* = \left\{ x : \lambda(x) = \frac{\sup_{\theta\in\Theta_0} L(\theta,x)}{\sup_{\theta\in\Theta} L(\theta,x)} \leq c \right\} \tag{6.2}$$

Remark that $0 \leq \lambda(x) \leq 1$. If H_0 is true, then $\lambda(x)$ should be close to 1. And if H_0 is false, the denominator must be big when compared to the numerator, and thus $\lambda(x)$ must be close to zero.

To determine c on (6.2), we need to solve

$$\alpha = \sup_{\theta\in\Theta_0} P[\lambda(X) \leq c]$$

and, to do that, we need the distribution of the statistic $\lambda(X)$, which, in general, is not easily obtained. Thus, we can find a function h strictly increasing on the domain of $\lambda(x)$ such that $h(\lambda(X))$ has a simple form and a known distribution.

To implement the GLRT, we need the following procedure

1. Obtain the MLE $\hat{\theta}$ of θ;
2. Obtain the MLE $\hat{\theta}_0$ of θ, when $\theta \in \Theta_0$;
3. Compute
$$\lambda(X) = \frac{L(\hat{\theta}_0, X)}{L(\hat{\theta}, X)};$$
4. Find the function h;
5. Obtain c solving the equation $\alpha = P_{H_0}[h(\lambda(X)) \leq c]$.

The next theorem gives an asymptotic distribution of the statistic of the GLRT, solving the problem of finding the distribution of $\lambda(X)$ for large distributions.

Theorem 6.4.1. Let $X_1, \ldots, X_n$ be a random sample of the r.v. X with p.d.f. $f(x \mid \theta)$. Under the regularity conditions, if $\theta \in \Theta_0$, the distribution of the statistic $-2\log\lambda(X)$ converges to the distribution chi-square when the size of the sample n goes to infinity. The number of degrees of freedom of the limit distribution is the difference between the number of not specified

> parameters on Θ and the number of not specified parameters on Θ_0.

Notice that we can obtain a confidence interval using a hypothesis test and vice-versa.

Using this idea, we can formulate the problem from the Bayesian viewpoint using a Bayesian confidence interval. To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we construct the Bayesian interval for θ. If θ_0 is in the interval, we accept H_0 and reject it otherwise.

Bibliography

[BS01] Heleno Bolfarine and Mônica Carneiro Sandoval. *Introdução à Inferência Estatística*. SBM, 2 edition, 2001.

[Was04] Larry Wasserman. *All of Statistics: a Concise Course in Statistical Inference*. Springer, 2004.