Igor Kortchemski
Probality theory,
Autumn 2023
ETH

Chapter 8: A glimpse on statistical theory

(not part of the exam)

Outline: 1) Estimation
2) Confidence intervals

Until now, we have studied i.i.d. random variables with known distributions. In statistical theory the situation is different: we observe a sequence of values which we assume to be the realization of an i.i.d. sequence of r.v (called a sample) but with unknown law. Using the sample, we would like to estimate the unknown law, or decide to accept or reject a hypothesis that concerns it.

1) Estimation

In practice, it often happens that the unknown law belongs to a certain family of probability measures depending on a parameter θ.

For example, a company would like to commercialize a new product, and we would like to estimate the proportion $\theta \in [0,1]$ of the population susceptible of buying the product.

Definition A statistical model is a space Ω equipped with a σ-field $\mathcal{F}$ and a family of probability measures $(P_\theta)_{\theta \in \Theta}$. We say that Θ is the space of parameters.

Examples:
- $\Theta = [0,1]$ and P_θ is the law of $\mathrm{Ber}(\theta)$
- $\Theta = (0,\infty)$ and P_θ is the law of $\mathrm{Exp}(\theta)$
- $\Theta = \mathbb{R} \times \mathbb{R}_+$ and $P_{(m,\sigma^2)}$ is the law of $\mathcal{N}(m,\sigma^2)$

Definition A sample of size n is a sequence of realizations $X_1(\omega), \dots, X_n(\omega)$ of random variables $X_1, \dots, X_n$

In practice, we often have data, which is assumed to be a sample of iid r.v $X_1, \dots, X_n$ under P_θ, with θ unknown.

Definition • An estimator is a function d with values in the space of parameters Θ which depends on the sample, i.e. of the form $d(X_1, \dots, X_n)$

It is without bias when $\forall \theta \in \Theta$, $E_\theta[d(X_1, \dots, X_n)] = \theta$ (here E_θ denotes expectation with respect to P_θ)

It is strongly consistent if $\forall \theta \in \Theta$ under P_θ, $d(X_1, \dots, X_n) \xrightarrow[n \to \infty]{a.s} \theta$

END OF LECTURE 27

Example In the statistical model $\Theta = [0,1]$ and P_θ is the Bernoulli law of parameter θ,

$d(X_1, \dots, X_n) = \dfrac{X_1 + \dots + X_n}{n}$ is an unbiased estimator, strongly consistent (by the law of large numbers), called the empirical mean, and often denoted by $\overline{X_n}$.

2) Confidence intervals

In practice, we do not just give a numerical estimation of a parameter, but a "small" interval in which the parameter should be

In statistics, we use the term "confidence term". More precisely, let us consider a statistical model $(P_\theta, \theta \in \Theta)$ and a sample $(X_1, \dots, X_n)$ of size n.

We fix a confidence level $1-\alpha$, where $\alpha \in (0,1)$ represents the probability of error that we tolerate.

A confidence interval of level $1-\alpha$ is an interval $I(X_1, \dots, X_n) = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$ such that

$$P_\theta(\theta \in I(X_1, \dots, X_n)) \geq 1-\alpha \text{ for every } \theta \in \Theta.$$

Of course, for a given size of a sample, we hope for a high level of confidence and a small interval (these two conditions being antagonistic)

Concretely, one often starts with an estimator $d(X_1, \dots, X_n)$ et one tries to measure the "error" of this estimator to find an interval around $d(X_1, \dots, X_n)$ which has the desired level of confidence.

Example In the statistical model where $\Theta = [0,1]$ and P_θ follows the Bernoulli law of parameter θ, using the estimator $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$, the Bienaymé-Tchebychev inequality gives:

$$P_\theta(|\bar{X}_n - \theta| \geq \varepsilon) \leq \frac{Var_\theta(\bar{X}_n)}{\varepsilon^2} = \frac{\theta - \theta^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

Thus, if the confidence level $1-\alpha$ is fixed, we need $\frac{1}{4n\varepsilon^2} \leq \alpha$ and then

$$(*) \quad \left[\bar{X}_n - \frac{1}{2\sqrt{n\alpha}}, \bar{X}_n + \frac{1}{2\sqrt{n\alpha}}\right]$$ is a confidence interval of level $1-\alpha$.

We can also use the Central Limit Theorem to build asymptotic confidence levels i.e whose level is asymptotically $1-\alpha$ when $n \to \infty$: $\forall \theta \in \Theta$, $\liminf_{n\to\infty} P_\theta(\theta \in I(X_1, \dots, X_n)) \geq 1-\alpha$.

Let us just give the main idea:

If $Z_n \xrightarrow[n\to\infty]{(d)} N(0,1)$, then $\forall q > 0$ we have $\mathbb{P}(|Z_n| > q) \xrightarrow[n\to\infty]{} \mathbb{P}(|N(0,1)| > q)$.

We then choose q_α such that $\mathbb{P}(|N(0,1)| > q_\alpha) = \alpha$ (for example for $\alpha = 0{,}05$ $q_\alpha \simeq 1{,}96$).

Then $\mathbb{P}(-q_\alpha \leq Z_n \leq q_\alpha) \xrightarrow[n\to\infty]{} 1-\alpha$, which allows to build asymptotic confidence intervals of level $1-\alpha$.

Example In the statistical model where $\Theta = [0,1]$ and P_θ follows the Bernoulli law of parameter θ, we have $\frac{\sqrt{n}}{\sqrt{\theta(1-\theta)}}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{(d)} N(0,1)$ by the central limit, so $\mathbb{P}\left(\theta \in \left[\bar{X}_n - \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} q_\alpha, \bar{X}_n + \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} q_\alpha\right]\right) \xrightarrow[n\to\infty]{} 1-\alpha$

Problem: this interval depends on θ, unknown.

Solution 1: choose a larger interval, here using $\theta(1-\theta) \leq \frac{1}{4}$, we get $I(X_1, \dots, X_n) = \left[\bar{X}_n - \frac{q_\alpha}{\sqrt{2n}}, \bar{X}_n + \frac{q_\alpha}{\sqrt{2n}}\right]$

This interval can be smaller than $(*)$ but it is asymptotic (not exact)

Solution 2: replace occurences of θ by a strongly consistent estimator. Indeed, by Slutsky's theorem,

$$\frac{\sqrt{n}}{\sqrt{\bar{X}_n - \bar{X}_n^2}}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{(d)} N(0,1), \text{ so } \mathbb{P}\left(\theta \in \left[\bar{X}_n - \frac{\sqrt{\bar{X}_n - \bar{X}_n^2}}{\sqrt{n}} q_\alpha, \bar{X}_n + \frac{\sqrt{\bar{X}_n - \bar{X}_n^2}}{\sqrt{n}} q_\alpha\right]\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

The price to pay is then often the convergence slows down.