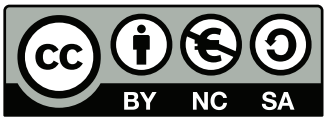
Métodos Avanzados en Estadística

Estudiante

Máster en Ciencia de Datos

Universidad Autónoma de Madrid

tu-web.es



Métodos Avanzados en Estadística

Estudiante

Máster en Ciencia de Datos

Universidad Autónoma de Madrid

tu-web.es

Índice

Parte I.
Teoría

$$P(|\hat{\theta}_n - \theta| > \epsilon) = P(|\hat{\theta}_n - \theta|^2 > \epsilon^2) \leq \frac{E|\hat{\theta} - \theta|^2}{\epsilon^2} = \frac{sesgo^2(\hat{\theta}) + Var(\hat{\theta})}{\epsilon^2}$$

Si tenemos

$$\sqrt{n}(\bar{X_n} - \mu) \to^d N(0, \sigma^2),$$

como $\sqrt{n} \to \infty$ y la diferencia tiende a cero, nos está indicando que la velocidad con la que $\bar{X_n}$ se acerca a μ es la misma con la que $\frac{1}{\sqrt{n}}$ va a cero

Sean $X_1, \ldots, X_n \sim U(0, \theta)$. Sabemos que $\mu = \theta/2$ y que $\sigma^2 = \theta^2/12$. Antes de nada, sabemos que por LDGN $\bar{X} \to^p \mu$ por lo que $2\bar{X} \to^d \theta$ (es consistente). Entonces, por TCL, tenemos que $\sqrt{n}(\bar{x} - \theta/2) \to^d N(0, \theta^2/12)$ así que $\sqrt{n}(2\bar{x} - \theta) \to^d N(0, \theta^2/3)$. Este el resultado teórico.

Si consideramos $n = 20, \theta = 10$, tenemos que $\theta^2/3n = 100/60$ así que todos los valores de nuestra distribución estrán en el intervalo $[10 \pm 2.6] \sim [7.4, 12.6]$.

1. Bootstrap

Our goal in bootstrapping will be to approximate the distribution of the estimator $T = T(x_1, \ldots, x_n; F)$. The idea would be to take different samples from the distribution and then compute the

Hence, we would like to approximate

$$H_n(x) = P_F(T(X_1, \ldots, X_n; F) \leq x)$$

If we knew the distribution function F, we could generate samples and the generate the histogram of the distribution of T. However, in this case, F is unknown. In bootstrap, we will make use of the empiric distribution F_n that the initial sample that we have provides. Hence, we can approximate:

$$\hat{H_n}(x) = P_{F_n}(T(X_1^*, \ldots, X_n^*, F_n) \leq x)$$

This is called *ideal bootstrap*.

We can generate new samples by extracting elements from the original sample **with replacement**. We get new samples where all the elements belong to the initial sample but some elements can be repeated. This way, we can approximate the value of $\hat{H_n}(x)$. We compute for each generated sample $T^{*(b)} = T(X_1^{*b}, \ldots, X_n^{*b}; F_n)$ and, lastly:

$$\hat{H_n}(x) \approx \frac{1}{B} \sum_{b=1}^{B} I_{T^{*b} \leq x}$$

Recall that we are approximating $\hat{H}_n$. We have then two approximations

$$H_n(x) \approx \hat{H}_n(x) \approx \tilde{H}_B(x)$$

the first one is approximating F by F_n. Then, we approximate H_n by $\hat{H}_n$. The first one is problematic since it requires a large amount of samples and **regularidad (translate)**.

1.1. Variance bootstrapping

We can also estimate the variance of an estimator θ: $Var_F(\hat{\theta})$. The process is approximately the same, we firstly compute the ideal bootstrap and then the approximation based in B re-samples is:

$$Var_{F_n}(\hat{\theta^*}) \approx \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\theta}_j^* - \bar{\theta^*})^2,$$

where $\hat{\theta}_j^*$ is the estimator for the re-sample j.

2. PDF Estimation

Let $X_1, \ldots, X_n$ be an iid sample from the distribution F with p.d.f. f. We woud like to estimate f making no previous assumptions, just making use of the data. Our goal will be to compute an estimator $\hat{f}$ such that $\hat{f} \approx f$. Let $h \approx 0$, then

$$P(x-h \leq X \leq x+h) = \int_{x-h}^{x+h} f(t)dt \approx 2hf(x)$$

that means that we can approximate $f(x)$ as follows:

$$f(x) \approx \frac{1}{2h}P(x-h \leq X \leq x+h).$$

If we replace the probability by the proportion we obtain an p.d.f. estimator:

$$\hat{f}(x) = \frac{1}{2h}\frac{\#\{i : |x-X_i| < h\}}{n} = \frac{1}{2hn}\#\{i : \frac{|x-X_i|}{h} \leq 1\}.$$

If $K(x) = \frac{1}{2}\mathbb{I}_{\{|x|\leq 1\}}$, where $\mathbb{I}_A$ is the indicator function over A, then

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

2.1. Estimating the kernel as a convolution

Let $U = Y + Z$ where:

1. $Y \sim F_n$ where F_n is the empiric distribution function
2. Z's density function is $K_h(x) = h^{-1}K(x/h)$.
3. Y and Z are independent.

that is, making a small perturbation on the empiric distribution. We can understand the kernel estimator as a slightly modified(smoothed) empiric distribution so that it is continuous.

Proposición 2.1. In the previous conditions, the p.d.f. of U is the kernel estimator $\hat{f}$.

Demostración. Firstly, we recall that

$$F_u(x) = P(u \leq x) = P(y + z \leq x)$$

now, using the total probability formula, we obtain

$$P(y + z \leq x) = \frac{1}{n}\sum_{i=1}^{n} P(y + z \leq x | y = x_i) = \frac{1}{n}\sum_{i=1}^{n} P(z \leq x - x_i)$$

in the last step, we can *forget* the conditional since Z, Y are independent. Then, we can use the definition of the probability to obtain:

$$\frac{1}{n}\sum_{i=1}^{n} P(z \leq x - x_i) = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x-x_i} k(\frac{u}{h})du$$

We can now derivate the distribution function to obtain

$$f_u(x) = F_u'(x) = \frac{1}{nh} = \sum_{i=1}^{n} k(\frac{x - x_i}{h}) = \hat{f}(x)$$

□

Using this, we can have an *algorithm* to obtain an iid random variable with p.d.f. $\hat{f}$ following the next steps:

1. Pick a random element with the same probability in the set $X_1, \ldots, X_n$. Let the result be X^*.
2. Simulate items to obtain Z with distribution K_h
3. Compute $U = X^* + Z$

2.2. Integrating the MSE

In general, we use the bias and the variance of an estimator to determine its *goodness*. Recall that

1. The bias of $\hat{f}(x)$ is $E[\hat{f}(x)]-f(x)$
2. The variance of $\hat{f}(x)$ is $E[(\hat{f}(x)-E[\hat{f}(x)])^2]$
3. The MSE $E[\hat{f}(x)-f(x)]^2$ verifies:

$$MSE = Bias^2[\hat{f}(x)] + Var[\hat{f}(x)].$$

We would like to determine the MSE in terms of the hyperparameters h, n and the smoothness of f. If we use higer values for h, then we would have a high bias since we would be forgetting the data. If we use very small values for h, our estimator would be very colse to the data, we would have a high variance and very small bias.

2.2.1. Other criteria

There are also other criteria to integrate the MSE. For instance, considering the $||\dot{|}|_p$ in L_p, we can see that the ECMI is the expectation of the L_2 distance between f and $\hat{f}$ squared, that is:

$$ECMI(\hat{f}) = E[||f-\hat{f}||_2^2]$$

It is easily shown seeing that:

$$E(||f-\hat{f}||_2^2) = E[\int |f(x)-\hat{f}(x)|^2 dx]$$

and, using Fubini's theorem, we obtain:

$$E[\int |f(x)-\hat{f}(x)|^2 dx] = \int E[(f(x)-\hat{f}(x))^2]dx = \int MSE(x)dx$$

3. Bias and Variance approximations

We will make the following assumptions for this section:

- K is a symmetric function with $\int K(u)du = 1$, $\int uK(u)du = 0$.
- $\sigma_K^2 = \int u^2K(u)du < \infty$ and $d_k = ||K||_2^2 = \int K(u)^2du < \infty$.
- f is twice *derivable* with continuous derivative.

Proposición 3.1. The bias... (diapositiva 21/37)

Demostración. Then, the approximation of the bias can be obtained as follows:

$$E[\hat{f}(x)] = E\left[\frac{1}{nh} = \sum K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{n}E[K\left(\frac{x-x_i}{h}\right] = \frac{1}{n}\int K\left(\frac{x-t}{h}f(t)dt\right)$$

we apply a *variable change* $w = \frac{x-t}{h}$, $dw = \frac{-dt}{h}$, $t = x - wh$, we obtain

$$\frac{1}{n}\int K\left(\frac{x-t}{h}f(t)dt\right) = \int K(w)f(x-wh)dw$$

and, applying Taylor's theorem (forgetting the terms that have h^3 and so on since h will be very small),

$$\int K(w)f(x-wh)dw = \int K(w)\left[f(x) - whf'(x) + \frac{w^2h^2}{2}f''(x)'\right]dw = f(x) + \frac{h^2}{2}f''(x)\int w^2K(w)dw$$

where the last integral equals σ_k^2, so

$$Bias(\hat{f}(x)) = \frac{h^2}{2}f''(x)\sigma_k^2.$$

As we can see, the bias of $\hat{f}$ depends highly on h.

□

Proposición 3.2. The variance..

Demostración. We can see that

$$Var[\hat{f}(x)] = Var\left[\frac{1}{nh} = \sum K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{h^2n}Var\left[K\left(K(\frac{x-x_i}{h})\right)\right] = \frac{1}{h^2n}\left[E\left[K^2(\frac{x-x_i}{h})\right] - E^2\left[K(\frac{x-}{h}\right.\right.$$

which is approximately

$$\frac{1}{h^2n}E\left[K^2(\frac{x-x_i}{h})\right] = \frac{1}{nh^2}\int K^2(\frac{x-t}{h})f(t)dt$$

and, applying the same *variable change*, and Taylor's theorem:

$$= \frac{1}{nh^2}\int K^2(w)f(x-hw)dw \approx \frac{1}{nh}\int K^2(w)\left[f(x) - hwf'(x) + \frac{h^2w^2}{2}f''(x)\right]dw$$

Since we are dividing by nh, we can forget the term $f'(x)$ term, so we finally obtain:

$$Var[\hat{f}] = \frac{1}{nh}f(x)d_k$$

□

Using both previous propositions, we obtain:

Proposición 3.3. The approximation of the MSEI is:

$$MSEI(\hat{f}) \approx \frac{h^4}{4}\sigma_k^4\int f''(x)^2dx + \frac{d_K}{nh} = \frac{h^4}{4}\sigma_K||f''||_2^2 + \frac{||K||_2^2}{nh}$$

The idea now is to optimize this function with respect h. We have to derivate this function respect to h and make it equal to 0, that is:

$$\frac{4h^3}{4}\sigma_K^4||f''||_2^2 = \frac{||K||_2^2}{nh^2}$$

and the optimal value is

$$h^* = cn^{-1/5} \to 0$$

which implies

$$nh^* = cn^{4/5} \to \infty$$

The conclusions are:

- h^* has to go to zero but slowly
- The speed at which h goes to zero in this case is $n^{-4/5}$, which is pretty slower than just n.

3.0.1. Optimal kernel

What is more, when computing the IMSE, the factor that depends on K is $\sigma(K) = \sigma_K^{4/5}||K||_2^{8/5}$. It can be proved that if $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$.

To optimize the kernel, we have the following variational problem:

$$\min ||K||_2^2 \quad s.t.$$

The solution to this problem is the Epanechnikov

$$K^*(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right) \quad \text{if } -\sqrt{5} \leq u \leq \sqrt{5}$$

The choice of the kernel does not hardly affect the value of the IMSE.

3.0.2. Choosing the smoothing parametern

To select the optimal smoothing parameter there are many methods:

1. Plug in:
    - Suppose that $f \approx N(\mu, \sigma)$
    - Non parametric methods
2. Cross validation

Plug-in supposing Normal.-

We assume that

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp ...$$

So, we substitute $\|f''\|_2^2$ in the expression of the optimal parameter h^* and , since we already know that the value of the norm, we obtain

$$h^* = \sigma(4/(3n))^{1/5} \approx \sigma(1.0456)n^{-1/5}$$

So, since the number is approximately 1, it is enough to estimate σ to estimate h^*. Some literature (*Silverman*) proposes to use $\hat{\sigma} = \min\{s, \hat{\sigma}_{ri}\}$ where $\hat{\sigma}_{ri}$ is the interquartilic range (estandardized so that it converges to σ).

Non-parametric plug-in.-

We fix a preliminar smoothing parameter g, using for instance the previous approximation. This way, we obtain an auxiliar estimator $\hat{f}_g(x)$. Using this estimator, we now estimate

$$\|\hat{f''}\|_2^2 = \int \hat{f}_g''(x)^2 dx$$

Cross Validation

The intuitive idea under cross validation is to divide the population in two splits and we use one of them to obtain information of how good is the estimator computed with the first one. We can then compute

$$MSEI(\hat{f}) = E\left[\int \hat{f}(x;h)^2 dx\right] - 2E\left[\int \hat{f}(x;h) f(x) dx\right] + \int f(x)^2 dx$$

Again, we would like to find h^* such that this IMSE is small. As we can see, the last term only depends on the *true* density function. For each h, our sample is fixed, so the expectation is reduced to a single term and we would like to minimize

$$C(h) = \int \hat{f}(x;h)^2 dx - \frac{2}{n}\sum \hat{f}_{(-i)}(X_i;h)$$

where $f_{(-i)}$ is the estimator computed with all the observations except X_i. The last term comes from the following reasoning. Let $x_1, \ldots, x_n \to \hat{f}$, and let us have another point x_{n+1}. Then, we can see that:

$$\int \hat{f}(x;h) f(x) dx = E\left[\hat{f}(x_{n+1};h) | x_1, \ldots x_n\right]$$

If we take expectation in both sides and use the known formula, we obtain:

$$E\left[\int \hat{f}(x;h) f(x) dx\right] = E\left[\hat{f}(x_{n+1};h)\right]$$