Estimation

How to get information from samples

{Arnaud.Legrand,Jean-Marc.Vincen}t@imag.fr

Université de Grenoble-Alpes, UFR IM^{2}AG
M2R-MOSIG Scientific Methodology

November 2016

PROTOCOL VALIDATION

Typical random access protocol to a common channel (CSMA family)

```
Protocol-Send (M)
  while Message is not sent
    Send(M)
    if Collision
      W=Random (I_n)
      n = n + 1
      I_{n+1} = g(n, I_n)
      Wait (W)
```

What should be the *amount of time* ?

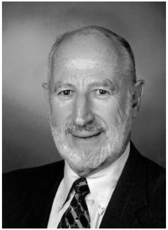
Protocol dimensioning

Waiting time :

- Random
- Uniformly distributed on an interval $[0, I_n]$
- Length of the interval depends on the number of collisions
- Adaptive scheme $I_{n+1} = 2 \times I_n$,
- I_0 fixed, characteristic of the protocol

PROTOCOL HISTORY

University of Hawaiì 1970 http://www.hicss.hawaii.edu/



Norman Abramson et al.
Use of a radio network to provide computer communications without centralization or vacations

Ancestor of CSMA/CD (ethernet), CSMA/CA (WiFi)...

QUANTITATIVE SPECIFICATION VALIDATION

Experiment

Propose an experiment to check the specification of the protocol

Estimation

How could I_0 be estimated ?

Decision

How could you conclude on the validity of the implementation of the protocol ?

SAMPLES

The observations : a sequence of n waiting times.
We suppose that the experiment has been well driven (no outliers, ...) and the observed values denoted by

$$\{x_1, \cdots, x_n\}.$$

The stochastic model : the observations are considered to be realizations of independent random variables with the same probability law F

$$\{X_1, \cdots, X_n\}$$

Question : What could be said on F from the observations $\{x_1, \cdots, x_n\}$?

A priori knowledge :

- The shape of the law is known and depends on some parameter(s) unknown : **parametric estimation**

  ex: X follows a uniform distribution on $[0, \theta[$ with θ unknown

$$F_\theta(x) = \begin{cases} 0 & x \leqslant 0; \\ \frac{1}{\theta} x & x \in [0, \theta[; \\ 1 & \theta \leqslant x. \end{cases}$$

- The shape of the law is unknown and some parameters are under study (expectation, variance, moments,...) : **non-parametric estimation**

BASIC CONCEPTS

- A **statistic** is a function of the observations : $t_n(x_1, \cdots, x_n)$, it usually summarize some parameter of the distribution.
- An **estimator** is a random variable $T_n = t(X_1, \cdots, X_n)$ (model of the statistic)

Example

$t(x_1, \cdots, x_n) = \max_{1 \leqslant i \leqslant n} x_i$ is a statistic on the samples;

and $T_n = \max_{1 \leqslant i \leqslant n} X_i$ the corresponding estimator.

Law of T_n under the hypothesis X_i uniformly distributed on $[0, \theta[$:

$F_n^\theta(x) = \mathbb{P}(\max_{1 \leqslant i \leqslant n} X_i \leqslant x) = \frac{1}{\theta^n} x^n$ using independence and uniformity law of X_i;

and density

$$f_n^\theta(x) = \frac{1}{\theta^n} n x^{n-1}.$$

BIAS

An estimator T_n of some parameter θ is **unbiased** iff

$$\mathbb{E}T_n = \theta$$

Example

Bias of estimator T_n

$$\mathbb{E}T_n = \int_0^\theta x f_n^\theta(x)dx = \int_0^\theta x \frac{1}{\theta^n} n x^{n-1} dx = \frac{1}{\theta^n} n \left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1}\theta \neq \theta.$$

T_n is a biased estimator, on average it underestimate θ on average. But, for large samples the bias decreases to 0.

$\lim_{n\to+\infty} \mathbb{E}T_n = \theta$ T_n is aymptotically unbiased (or consistent)

To compensate the bias,

$$T'_n = \frac{n+1}{n} T_n,$$

which is unbiased.

RISK

The quality T_n of an estimator is evaluated by the **risk** function

$$R(T_n) = \mathbb{E}(T_n - \theta)^2 = \mathbb{E}(T_n - \mathbb{E}T_n)^2 + (\mathbb{E}T_n - \theta)^2 = \mathbb{V}ar T_n + (\mathbb{E}T_n - \theta)^2$$

- $\mathbb{V}ar T_n$: the concentration of the distribution
- $(\mathbb{E}T_n - \theta)^2$: impact of the bias.

For an unbiased estimator

$$R(T_n) = \mathbb{V}ar T_n$$

Risk of T'_n

For the example

$$\begin{aligned}\mathbb{V}ar T_n &= \mathbb{E}(T_n - \mathbb{E}T_n)^2 = \mathbb{E}T_n^2 - (\mathbb{E}T_n)^2 = \int_0^\theta x^2 f_n^\theta(x)dx - \left(\frac{n}{n+1}\theta\right)^2 \\ &= \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2\end{aligned}$$

$$\begin{aligned}\mathbb{V}ar T'_n &= \mathbb{V}ar \frac{n+1}{n} T_n = \frac{(n+1)^2}{n^2} \mathbb{V}ar T_n = \frac{(n+1)^2}{n^2}\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2 \\ &= \frac{1}{n(n+2)}\theta^2\end{aligned}$$

ANOTHER ESTIMATOR

Consider

$$U_n = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

$$\mathbb{E}U_n = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{2}{n}\sum_{i=1}^{n} \frac{\theta}{2} = \theta.$$

U_n is an unbiased estimator of θ.

Risk of U_n

For the example

$$\begin{aligned}
\mathbb{V}ar U_n &= \mathbb{V}ar\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right); \\
&= \left(\frac{2}{n}\right)^2 \sum_{i=1}^{n} \mathbb{V}ar X_i \text{ because of the independence of } X_i; \\
&= \frac{4}{n^2} n \frac{\theta^2}{12} = \frac{\theta^2}{3n}.
\end{aligned}$$

The risk associated to U_n is much larger than the risk of T_n' so we will prefer T_n'.