23.10.

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 (In)
      n = n+1
      In+1 = g(n, In)
      Wait (W)
```

What should be the amount of time?

- Protocol dimensioning

  Waiting time:
  - random
  - uniformly distributed on an interval $[0, I_n]$
  - ...

Samples

The observation: a sequence of n waiting times

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

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

$\{X_1, \dots, X_n\}$

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

▷ A priori knowledge:

- the shape of the law is known and depends on some parameter values: parametric estimation

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

- the shape of the law is unknown and some parts are under study: non-parametric estimation

BASIC CONCEPTS

- a statistic is a function of observations : $f_n(x_1, \dots, x_n)$; usually summarizes some param. of the distribution
- an estimator is a random variable $T_n = t(X_1, \dots, X_n)$ (model of statistic)

Example:

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

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

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

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

and density $f_n^\theta(x) = \frac{1}{\theta^n} n \cdot x^{n-1}$

BIAS

- An estimator T_n of some parameter θ is UNBIASED if $\mathbb{E} T_n = \theta$

Example:

$$\mathbb{E} T_n = \int_0^\theta x f_n^\theta(x) dx = \dots = \frac{n}{n+1}\theta \neq \theta \ !$$

T_n is biased estimator, on average it underestimate θ

But, for large samples, the bias decreases to 0

$\lim_{n \to +\infty} \mathbb{E} T_n = \theta$, T_n is **asymptotically unbiased**

to compensate the bias,

$T_n' = \frac{n+1}{n} T_n$ which is unbiased

RISK

The quality T_n of an estimator is evalued by Risk function:

$$R(T_n) = \mathbb{E}(T_n - \theta)^2 = \mathbb{E}(T_n - \mathbb{E}T_n)^2 + \underbrace{(\mathbb{E}T_n - \theta)^2}_{\text{bias}} = \mathrm{Var} T_n + (\mathbb{E}T_n - \theta)^2$$

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

For an unbiased estimator $R(T_n) = \mathrm{Var}(T_n)$

$$\text{Var } T_n = \mathbb{E}(T_n - \mathbb{E}T_n)^2 = \mathbb{E}T_n^2 - (\mathbb{E}T_n)^2 = \int_0^\theta x^2 f_{T_n}(x)dx - \left(\frac{n}{n+1}\theta\right)^2 = \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2$$

$$\text{Var } T_n' = \text{Var } \frac{n+1}{n} T_n = \frac{(n+1)^2}{n^2} \text{Var } T_n = \frac{(n+1)^2}{n^2}\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2 = \frac{1}{n(n+2)}\theta^2$$

Another estimator

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

$$\mathbb{E}U_n = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{2}{n}\sum_{i=1}^{n}\frac{\theta}{2} = \theta \quad \Rightarrow U_n \text{ is unbiased estimator of } \theta$$

Risk of U_n:

$$\text{Var } U_n = \text{Var}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{2}{n}\right)^2 \sum \text{Var } X_i \quad \text{because of the independence of } X_i$$

$$= \frac{4}{n^2} n \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

← the risk is much larger than the risk of T_n'
↳ so we prefer T_n'

Synthesis

Parametric Estimation

- Assumptions: the results of experiments are modeled by a sequence of independent identically distributed with a KNOWN probability law
- An estimator
- The bias of an estimator is expected difference between the estimator & the real value $E_\theta(T_n - \theta)$
- The risk is the expectation of the error around the real value $E_\theta(T_n - \theta)^2$

Non-parametric Estimation - estimation of the mean

- Assumptions: the results of experiments are modeled by a sequence of independent identically distributed with a unknown probability law, but with some properties mean m & variance σ^2
- The average $\frac{1}{n}\sum X_i$ is an unbiased estimator of m
- The error $\frac{1}{n}\sum X_i - m$ converges almost surely to 0 (strong law of large numbers)
- The law of errors satisfies the CLT (CI and confidence intervals):

$$\frac{1}{\sqrt{n}}\sum \frac{(X_i - m)^2}{\sigma^2} \xrightarrow{d} N(0,1)$$

Data Statistics Introduction

* 3 parts:
  1. The Problem : The Data Set - Building a DataSet (DATA I
  2. The Problem : The Data Set - How data was produced
  3. Explorations

1.

DATA PRODUCTION

- Why this data has been produced? (purpose)
- Which approach has been used? (method)
- How this dataset has been practically produced ? (observations)

ANALYSIS OF THE SET OF VARIABLES

- Identification of variable types
- Identification of the variable role
- Identification of the variables semantics

ANALYSIS OF THE TYPE OF VARIABLES

- Nominal Variables : classification, membership
- Ordinal Variables : comparison ! level
- Quantitative Variables : quantities

USAGE OF VARIABLES

- Response Variables → response time, duration ...
- Explanatory Variables → size, load (affect response)
- Univariate → 1 variable is used
  Multivariate → several -11-

TO BE CONTINUED