# Large deviations for $L$-statistics

F.Gamboa, H.Boistard

## Scheme

1. $L$-statistics : definition, expression, limit theorems.

2. Large deviations : some tools from the large deviations theory, a large deviations principle for some $L$-statistics.

3. Example : the uniform law.

## Definition of an L-statistics

$(X_i), i = 1 \ldots n$ an i.i.d. sample with distribution function $F$ , and $X_{(1)} \leq \cdots \leq X_{(n)}$ the associated order statistics.

**Definition 1** *An* L-statistics *is of the form :*

$$A_n = \frac{1}{n} \sum_{i=1}^{n} c_{n,i} b(X_{(i)}),$$

*where $c_{n,i}$ are some possibly $k$-dimensional coefficients and $b$ is some function from $\mathbb{R}$ to $\mathbb{R}$.*

In many examples $b$ is the identity and $c_{n,i} \simeq a(\frac{i}{n})$ for some bounded function $a$ on $[0, 1]$.

## Examples of $L$-statistics

- $\alpha$-trimmed mean : $\frac{1}{n-2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor+1}^{n-\lfloor \alpha n \rfloor} X_{(i)}$.

The corresponding function is $a(t) = \begin{cases} 0 & \text{for } t < \alpha \text{ or } t > 1-\alpha, \\ \frac{1}{1-2\alpha} & \text{for } t \in [\alpha, 1-\alpha]. \end{cases}$

- a part of D'Agostino's goodness-of-fit test statistics :

$$D = \frac{\sum_{i=1}^{n} (i - (n+1)2^{-1}) X_{(i)}}{n^2 S_n},$$

where $S_n^2$ is the sample variance.
The corresponding function is $a(t) = t - \frac{1}{2}$.

• Gini's difference mean : $(X_i), i = 1 \dots n$ an i.i.d. sample with distribution function $F$.

A dispersion parameter : $\theta = E(|X_1 - X_2|)$ and its estimator

$$T_n = \frac{1}{C_n^2} \sum_{i<j} |X_i - X_j| = \frac{1}{C_n^2} \sum_{i=1}^{n} (-n + 2i - 1) X_{(i)}.$$

The corresponding function $a$ is

$$a(t) = 4(t - \frac{1}{2}).$$

## Expression with the quantile function

For $G$ a distribution function, the associated quantile function is defined by its left continuous left inverse

$$G^{-1}(t) = \inf\{x : G(x) \geq t\} \text{ for } t \in [0,1].$$

The empirical distribution function defined for $x \in \mathbb{R}$ by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

admits as an inverse $F_n^{-1} : t \mapsto X_{(i)}$ for $t \in ]\frac{i-1}{n}, \frac{i}{n}]$.

$$A_n \simeq \frac{1}{n} \sum_{i=1}^{n} a(\frac{i}{n}) b(X_{(i)}) = \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} a(\frac{i}{n}) b(F_n^{-1}(t)) dt \simeq \int_0^1 a(t) b(F_n^{-1}(t)) dt.$$

## Limit properties for L-statistics

• Helmers (1978-1981), Vandemaele and Veraverbeke (1982) : approximation of $L$-statistics with $U$-statistics, and obtention of Berry-Esseen bounds. Conditions on the coefficients $c_{n,i}$, $b$ the identity.

• Shorack and Wellner (1986) : treatment via empirical processes and obtention of weak and strong laws of large numbers, CLT and law of iterated logarithm. Conditions of boundedness of $a$ and $b$.

• More recently, for instance weaker sufficient conditions on $b$ (under rather strong conditions on $a$) obtained by Li, Rao and Tomkins (2001) for the obtention of the CLT and the LIL.

## Large deviations

$(X, \mathcal{B})$ is a topological space with its borelian sigma-algebra.

**Definition 2** *A sequence* $(P_n)$ *of probability measures on* $(X, \mathcal{B})$ *satisfies a Large Deviations Principle (LDP) with speed* $n$ *if there exists a function* $I : X \to [0, +\infty]$ *lower-semicontinuous, called rate function, such that for all* $A \in \mathcal{B}$

$$-\inf_{x \in \mathring{A}} I(x) \leq \liminf_{n \to +\infty} \frac{1}{n} \log P_n(A) \leq \limsup_{n \to +\infty} \frac{1}{n} \log P_n(A) \leq -\inf_{x \in \bar{A}} I(x)$$

The rate function $I$ is said to be *good* if its level sets are compact.

A sequence of random variables $(X_n)$ on a probability space $(\Omega, \mathcal{A}, P)$, is said to satisfy a LDP if the sequence of probability measures defined by $P_n = \mathcal{L}(X_n)$ satisfies a LDP.

## Tools from the theory of large deviations

**Theorem 1** ***Contraction principle*** *Let* $(P_n)$ *be a sequence of probability measures on* $(X, \mathcal{B})$ *which satisfies a LDP with good rate function* $I$*,* $Y$ *a metric space and* $f : X \to Y$ *a continuous function.*

*Then the sequence of probability measures on* $Y$ $P_n \circ f^{-1}$ *satisfies a LDP with good rate function*

$$J(y) = \inf\{I(x) : f(x) = y\} \text{ for } y \in Y.$$

**Definition 3** ***Exponential equivalence*** *Let* $(X_n)$ *and* $(Y_n)$ *be random variables on a probability space* $(\Omega, \mathcal{A}, P)$, *with value in some metric space* $(Y, d)$. $(X_n)$ *and* $(Y_n)$ *are called* exponentially equivalent *if for every* $\epsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P(d(X_n, Y_n) > \epsilon) = -\infty.$$

**Theorem 2** *If a LDP with a good rate function holds for the random variables* $(X_n)$ *which are asymptotically equivalent to the r.v.* $(Y_n)$, *then the same holds for* $(Y_n)$.

## Sanov's theorem

The space $\mathbb{P}(\mathbb{R})$of all probability measures on $\mathbb{R}$ is equipped with the topology of weak convergence of probability measures.
Let $(X_i), i = 1 \ldots n$ be an i.i.d. sample with law $P \in \mathbb{P}(\mathbb{R})$.

**Definition 4** *The* empirical measure *associated to this sample is*

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

**Theorem 3** *$\nu_n$ satisfies a LDP with good rate function*

$$I(Q) = \begin{cases} \int \log(\frac{dQ}{dP}) dQ & \text{for } Q \ll P \text{ and } \log(\frac{dQ}{dP}) \in L_1(Q), \\ +\infty & \text{else.} \end{cases}$$

## Exponentially equivalent statistics

**Proposition 1** *Let* $T_{n,a} = \int_0^1 a(t)F_n^{-1}(t)dt$ *and* $A_n = \frac{1}{n}\sum_{i=1}^n a(\frac{i}{n})X_{(i)}$.
*Suppose the following hypotheses are satisfied :*
*(I) regularity of* $a$ *: for all* $n$ *there exists* $b_n = o(1/n)$ *such that for all* $i = 1 \dots n$

$$\|\frac{1}{n}a(\frac{i}{n}) - \int_{\frac{i-1}{n}}^{\frac{i}{n}} a(t)dt\| \leq b_n.$$

*(II) the domain of definition of the Laplace transform of* $|X_i|$ *is not reduced to* $\{0\}$,
*then for every* $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \log P(\|A_n - T_{n,a}\| > \epsilon) = -\infty.$$

## LDP for $T_{n,a}$ obtained by contraction

**Theorem 4** *Suppose that*
*(III)* $F$ *has compact support included in* $[-M, M]$ *and that*
*(IV)* $a$ *is bounded on* $[0, 1]$,
*then* $T_{n,a} = \int_0^1 a(t)F_n^{-1}(t)dt$ *satisfies a LDP with good rate function*

$$J(C) = \inf\{I(G) : G \text{ is a d.f. on } \mathbb{R} \text{ s.t. } \int_0^1 a(t)G^{-1}(t)dt = C\}.$$

## Statement of the LDP for $A_n$

Suppose that hypotheses

(I) regularity of $a$

(III) boudedness of the support of $F$

(IV) boundedness of $a$

are satisfied. Then $A_n$ satisfies a LDP with good rate function

$$J(C) = \inf\{I(G) : G \text{ is a d.f.on } \mathbb{R} \text{ s.t. } \int_0^1 a(t)G^{-1}(t)dt = C\}.$$

## Example : the uniform law on $[0,1]$

We denote the uniform d.f. by $F : F(t) = t$ for $t \in [0,1]$. Let $\mathcal{X}$ be the set of all $x = G^{-1}$ for $G \ll F$. The elements of this set are derivable almost everywhere. For $x = G^{-1} \in \mathcal{X}$, let

$$K(x) := I(G) = -\int_0^1 \log x'(t)dt.$$

**Minimization problem** $(P)$ **:** minimize

$$K(x) = \begin{cases} -\int_0^1 \log x'(t)dt & \text{for } x \in \mathcal{X} \\ +\infty & \text{elsewhere,} \end{cases}$$

under the (possibly $k$-dimensional) constraint : $\int_0^1 a(t)x(t)dt = C$.
$J(C)$ is the value of the minimum.

## Result of the minimization

Suppose that $\int_0^1 a(t)dt = 0$, put $A(t) = \int_t^1 a(s)ds$.

**Proposition 2** *The minimum value of the minimization problem* $(P)$ *is*

$$J(C) = 1 + \sup_{\lambda \in \mathbb{R}, \mu \in \mathbb{R}^k} \lambda + < \mu, C > + \int_0^1 \log(-\lambda - < \mu, A(s) >) ds$$

*where* $<,>$ *denotes the usual scalar product in* $\mathbb{R}^k$.

## Sketch of the proof

• formulation of $(P)$ in terms of $x' = y$.

$(\tilde{P})$ Minimize $\tilde{K}(y) = -\int_0^1 \log y(t)dt$ under the constraints

$$\int_0^1 A(s)y(s)ds = C, 0 \leq \int_0^1 y(t)dt \leq 1.$$

• fix $\int_0^1 y(t)dt = \alpha$, and obtain minimization problem $(P_\alpha)$, minimum found by a duality argument.

• minimax argument to find the minimum (over $\alpha$) of the minimum values for each $(P_\alpha)$

• final discussion.

## References


– R.Helmers, Ann. Prob., Vol. 9, No2 (1981).
– M.Vandemaele, N.Veraverbeke, Ann. Prob., Vol. 10, No2 (1982).
– G.R.Shorack, J.A.Wellner, Empirical Processes With Applications to Statistics, Wiley (1986).
– D.Li, M.B.Rao, R.J.Tomkins, Journ. Mult. Analysis, 78 (2001).
– J.M.Borwein, A.S.Lewis, SIAM J. Optimization, Vol. 3, No2 (1993).