LIMIT THEOREMS FOR QUANTILE EMPIRICAL PROCESSES

Igor Borisov
Sobolev Institute of Mathematics, Novosibirsk, RUSSIA

Saint Petersburg, June 10–14, 2013

The problem

Let $X_1, X_2, \ldots$ be a sequence of identically distributed random variables (r.v-s). Denote by $F(x)$ the distribution function of X_1. Let $X_{1:n} \leq \ldots \leq X_{n:n}$ be the corresponding order statistics based on the sample $X_1, \ldots, X_n$. Consider the following two empirical distribution functions (d.f.-s):

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i < t),$$

$$E_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\mathbf{E}X_{i:n} < t) \ \text{ if } \ \mathbf{E}|X_1| < \infty,$$

where $\mathbb{I}(\cdot)$ is the indicator of an event (including the nonrandom case). The **nonrandom empirical d.f.** $E_n(t)$ was introduced by W. Hoeffding (1953). He proved that, **in the iid case**,

$$\lim_{n\to\infty} \int g(x) dE_n(x) = \mathbf{E}g(X_1) \qquad (1)$$

for any continuous function g such that the modulus $|g|$ has a **convex majorant** integrable w.r.t. F.

The main goal is to extend (1) to the case of dependent r.v-s $\{X_i\}$.

Preliminaries

For an arbitrary distribution function $G(x)$ we introduce the quantile transform (generalized inverse function)

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

The values $G^{-1}(0)$ and $G^{-1}(1)$ may be defined as the corresponding limits.
Recall the main property of the quantile transforms. Hereinafter we denote by ω a random variable having the $(0, 1)$-uniform distribution. Then
1) The random variable $G^{-1}(\omega)$ has the distribution function $G(x)$.
2) If $G_n \xrightarrow{w} G$ then $G_n^{-1}(\omega) \to G^{-1}(\omega)$ a.s.

We study limit behavior of the quantile empirical process $F_n^{-1}(t)$, $t \in (0, 1)$, and its mean function $\mathbf{E}F_n^{-1}(t) \equiv E_n^{-1}(t)$ as well as of some functionals of the processes under consideration.

Introduce the following notation:

$$\|g\|_p := \left(\int_0^1 |g(t)|^p dt \right)^{1/p} = \left(\mathbf{E}|g(\omega)|^p \right)^{1/p}, \quad p \geq 1.$$

Lemma 1 (I.S.Borisov and A.V.Shadrin, 1996).
Given arbitrary d.f.-s G_1 and G_2,

$$\|G_1^{-1} - G_2^{-1}\|_p^p \leq C(p) \int_R |G_1(t) - G_2(t)||t|^{p-1} dt, \tag{2}$$

where
1) $C(p) = p2^{p-1}$ *for any* $p \geq 1$ *(in the case of odd numbers p, inequality* (2) *with the same constant was obtained by Sh. Ebralidze,* 1970*);*
2) *The value* $C(p) = p2^{p-1}$ *is unimprovable in the class of centered distributions for every* $p \geq 1$;
3) *If* $p = 1$ *then the sign "*$\leq$*" in* (2) *can be replaced by "*$=$*"* *(Yu. V. Prohorov,* 1956*);*
4) $C(p) = p$ *if there exists* $t \in [0, 1]$ *such that* $G_1^{-1}(t) = G_2^{-1}(t) = 0$ *(the case of zero medians, i.e.,* $t = 1/2$*, has been studied by Sh. Ebralidze,* 1970*).*

Corollary 1. *Let* $\{X_i\}$ *be a stationary ergodic sequence and* $\mathbf{E}|X_1|^p < \infty$, $p \geq 1$. *Then, as* $n \to \infty$,

$$\|F_n^{-1} - F^{-1}\|_p \to 0 \ \ a.s.$$

Corollary 2. *Let* $\{X_i\}$ *be a sequence of identically distributed random variables such that* $\mathbf{E}|F_n(t) - F(t)| \to 0$ *as* $n \to \infty$ *for all* $t \in R$. *Then, under the restriction* $\mathbf{E}|X_1|^p < \infty$, $p \geq 1$, *as* $n \to \infty$,

$$\|E_n^{-1} - F^{-1}\|_p^p \leq \int_0^1 \mathbf{E}|F_n^{-1}(t) - F^{-1}(t)|dt$$

$$\leq C(p) \int_R \mathbf{E}|F_n(t) - F(t)||t|^{p-1} dt \to 0. \qquad (3)$$

Corollary 3. *Under the restrictions of Corollary* 2 *for* $p = 2$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{Var} X_{i:n} = \|F^{-1}\|_2^2 - \|E_n^{-1}\|_2^2 \to 0 \qquad (4)$$

since $\|F^{-1}\|_2^2 = \mathbf{E}X_1^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}X_{i:n}^2$ *and* $\|E_n^{-1}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{E}X_{i:n})^2$; *moreover, in view of* (3) *and the triangle inequality,*

$$\left|\|E_n^{-1}\|_2 - \|F^{-1}\|_2\right| \leq \|E_n^{-1} - F^{-1}\|_2 \to 0.$$

Corollary 4. *Under the restrictions of Corollary* 2 *for* $p = 1$, *the convergence* $E_n \xrightarrow{w} F$ *is valid, i.e.,*

$$\lim_{n\to\infty} \int g(x) dE_n(x) = \mathbf{E} g(X_1)$$

for all ***continuous bounded*** *functions* g.

It remains only to verify the uniform integrability condition.

Lemma 2. *For any sequence of identically distributed r.v-s* $\{X_i\}$ *with finite mean, and for any nonnegative convex function h such that* $\mathbf{E}h(X_1) < \infty$*, the following upper bound is valid:*

$$\Delta_n := \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{E}X_{i:n})\mathbb{I}(h(\mathbf{E}X_{i:n}) > K) \leq 3\mathbf{E}h(X_1)\mathbb{I}(h(X_1) > K/2)$$

for all $K > 0$.

Proof. Denote

$$h_K(x) := (h(x) - K)\mathbb{I}(h(x) > K) = \max\{h(x) - K, 0\}.$$

The function $h_K(x)$ is convex. Applying Jensen's inequality twice, we obtain

$$\Delta_n = \frac{1}{n}\sum_{i=1}^{n} h_K(\mathbf{E}X_{i:n}) + \frac{K}{n}\sum_{i=1}^{n}\mathbb{I}(h(\mathbf{E}X_{i:n}) > K)$$

$$\leq \mathbf{E}h_K(X_1) + \frac{K}{n}\sum_{i=1}^{n}\mathbb{I}(\mathbf{E}h(X_{i:n}) > K).$$

Finally,

$$\frac{K}{n}\sum_{i=1}^{n}\mathbb{I}(\mathbf{E}h(X_{i:n})>K)\leq\frac{K}{n}\sum_{i=1}^{n}\mathbb{I}(\mathbf{E}h(X_{i:n})\mathbb{I}(h(X_{i:n})>K/2)>K/2)$$

$$\leq\frac{2}{n}\sum_{i=1}^{n}\mathbf{E}h(X_{i:n})\mathbb{I}(h(X_{i:n})>K/2)=2\mathbf{E}h(X_1)\mathbb{I}(h(X_1)>K/2)$$

since $\mathbb{I}(\zeta>N)\leq\zeta/N$ for all $N>0$ and any nonnegative r.v. ζ (including the case $\zeta\equiv const$).
Thus,

$$\Delta_n\leq\mathbf{E}h_K(X_1)+2\mathbf{E}h(X_1)\mathbb{I}(h(X_1)>K/2)\leq 3\mathbf{E}h(X_1)\mathbb{I}(h(X_1)>K/2).$$

The lemma is proved.

Main result

So, we have proved the Hoeffding's result under very mild restrictions.

Theorem 1. *Let* $\{X_i\}$ *be a sequence of identically distributed r.v-s (arbitrarily correlated) with finite mean. Let* h *be a nonnegative convex function and* g *be a continuous function such that* $|g(x)| \leq h(x)$ *and* $\mathbf{E}h(X_1) < \infty$. *If* $\mathbf{E}|F_n(t) - F(t)| \to 0$ *as* $n \to \infty$ *for all* $t \in R$ *then*

$$\lim_{n\to\infty} \int g(x) dE_n(x) = \mathbf{E}g(X_1). \tag{5}$$

Remark. It is clear that

$$\mathbf{E}|F_n(t) - F(t)| \leq \frac{1}{n} \left(\sum_{i,j=1}^{n} \left[\mathbf{P}(\max\{X_i, X_j\} < t) - F^2(t) \right] \right)^{1/2}. \tag{6}$$

In particular, if $\mathbf{P}(\max\{X_i, X_j\} < t) \to F^2(t)$ as $|i - j| \to \infty$ for all $t \in R$ then the main restriction of the theorem is fulfilled.

Rates of convergence

We now study the rate of convergence in the relation (5).
Denote

$$\Delta_n(g) := |Eg(X_1) - \frac{1}{n}\sum_{i=1}^{n} g(EX_{i:n})|.$$

We study the case of locally Lipschitz functions $g \in Lip(K_\alpha)$, i.e.,

$$|g(x) - g(y)| \leq K(x, y)|x - y|,$$

with the following restriction on the constant:
$K(x, y) \leq B(1 + (|x| \vee |y|)^\alpha), \quad \alpha \geq 0.$
In the sequel, we assume that $B = 1$ (without loss of generality).

Theorem 2. *Let* $\{X_i\}$ *be a sequence of identically distributed r.v-s (arbitrarily correlated) and* $|X_1| \le c$ *with probability* 1*. Let* $g \in Lip(K_\alpha)$ *and* $\mathbf{E}|g(X_1)| < \infty$*. Then*

$$\Delta_n(g) \le (1 + c^\alpha) \int_{-c}^{c} \mathbf{E}|F_n(t) - F(t)| dt. \tag{7}$$

Remark. For pairwise independent $\{X_i\}$, we deduce from (6) that

$$\mathbf{E}|F_n(t) - F(t)| \le \frac{1}{\sqrt{n}} \sqrt{F(t)(1 - F(t))}.$$

So, in this case, we can specify the upper bound in (7):

$$\Delta_n(g) \le \frac{1 + c^\alpha}{\sqrt{n}} \int_{-c}^{c} \sqrt{F(t)(1 - F(t))} dt. \tag{8}$$

Put $\delta(c) := \mathbf{E}|X_1|^{1+\alpha}\mathbb{I}(|X_1| > c)$.

Theorem 3. *Let $g \in Lip(K_\alpha)$ and $\mathbf{E}|X_1|^{1+\alpha} < \infty$. Then, for pairwise independent $\{X_i\}$,*

$$\Delta_n(g) = O(\delta(c_n))$$

as $n \to \infty$, where c_n is a solution to one of the following equations:

$$c^{\frac{\alpha+1}{2}} = \delta(c)n^{1/2} \quad \textit{if} \ \ 0 < \alpha < 1,$$

$$c \log c = \delta(c)n^{1/2} \quad \textit{if} \ \ \alpha = 1,$$

$$c^{\alpha} = \delta(c)n^{1/2} \quad \textit{if} \ \ \alpha > 1.$$

Corollary. *Under the conditions of Theorem 3, let* $E|X_1|^{1+\alpha+r} < \infty, r > 0$ *and* $\alpha + r > 1$. *Then*

$$\Delta_n(g) \leq Cn^{-\frac{r}{2(\alpha+r)}}.$$

In particular, if $g_o(x) = x^2$ *(the case* $\alpha = 1$*) and* $E|X_1|^{2+r} < \infty, r > 0$, *then*

$$n\Delta_n(g_o) = \sum_{i=1}^{n} \mathbf{Var} X_{i:n} \leq Cn^{1-\frac{r}{2(1+r)}}.$$

Remark. In the iid case, there is a sufficiently large class of marginal distributions such that the value $n\Delta_n(g)$ is bounded uniformly on n.

THANK YOU SO MUCH!