# Concentration inequalities for order statistics

## Using the entropy method and Rényi's representation

Maud Thomas[1]

in collaboration with Stéphane Boucheron[1]

[1]LPMA Université Paris-Diderot

Workshop Nachwuchsforscherinnen in Stochastik
10-11 October 2013

# Background: order statistics

- Sample: $X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$

## Order statistics

$X_{(1)} \geq \ldots \geq X_{(n)}$ non-increasing rearrangement of $X_1, \ldots, X_n$.

- $X_{(1)}$: sample maximum
- $X_{(n/2)}$: sample median
- For all $k$, $\mathbb{P}\{X_{(k)} \leq t\} = \sum_{i=k}^{n} \binom{n}{i} F^i(t)(1 - F(t))^{n-i}$
- Classical statistic theory and Extreme Value Theory provide:
  - Asymptotic distributions
  - Convergence of moments

## Goal

derive simple, non-asymptotic variance/tail bounds for order statistics.

# Gaussian case and the Poincaré's inequality

- Suppose $X_i \sim \mathcal{N}(0,1), 1 \leq i \leq n$.

## Poincaré's inequality

If $Z := f(X_1, \ldots, X_n)$ with $f : \mathbb{R}^n \to \mathbb{R}$ L- Lipschitz, then $\mathrm{Var}[Z] \leq L^2$.

- $f\colon (x_1, \ldots, x_n) \to x_{(k)}$ is 1- Lipschitz $\Rightarrow \mathrm{Var}[X_{(k)}] \leq 1$.
- Extreme Value Theory: $\mathrm{Var}[X_{(1)}] = O\left(\frac{1}{\log n}\right)$.
- Classical statistic theory: $\mathrm{Var}[X_{(n/2)}] = O\left(\frac{1}{n}\right)$.

## We do not understand (clearly)

in which way the maximum is a smooth function of the sample.

# Order statistics and spacings

## Proposition (Boucheron, T. (2012))

*For all* $0 < k \leq n/2$

$$\mathrm{Var}[X_{(k)}] \leq k\mathbb{E}\left[\left(X_{(k)} - X_{(k+1)}\right)^2\right] = k\mathbb{E}[{\Delta_k}^2]\,.$$

## Remark

Without any assumption such as:

- $F$ belongs to the max-domain of attraction of an extreme value distribution $G$, *i.e*

$$\lim_{n\to+\infty} F^n(a_n x + b_n) = G(x)$$

  for every continuity point $x$ of $G$.

- $(X_{(k)})$ is a sequence of
  extreme order statistics, if $k$ fixed, $n \to \infty$;
  central order statistics, if $k/n \to p \in (0,1)$ while, $n \to \infty$;
  intermediate order statistics, if $k/n \to 0$, $k \to \infty$.

# Rényi's representation

## Rényi's representation (Rényi (1953))

Let $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ be the order statistics of an independent sample of the standard exponential distribution, then

$$\left(Y_{(n)}, \ldots, Y_{(i)}, \ldots, Y_{(1)}\right) \sim \left(\frac{E_n}{n}, \ldots, \sum_{k=i}^{n} \frac{E_k}{k}, \ldots, \sum_{k=1}^{n} \frac{E_k}{k}\right)$$

where $E_1, \ldots, E_n$ are i.i.d standard exponential random variables.

## Representation for order statistics

If $Y_{(1)} \geq \ldots \geq Y_{(n)}$ are the order statistics of an exponential sample, then

$$F^{\leftarrow}\left(1 - e^{-Y_{(1)}}\right) \geq \ldots \geq F^{\leftarrow}\left(1 - e^{-Y_{(n)}}\right)$$

are distributed as the order statistics of a sample drawn according to $F$.

# Variance bound for order statistics when the hazard rate is non-decreasing

$V_k = k\Delta_k^2$: the Efron-Stein estimate of the variance of $X_{(k)}$.

## Theorem (Boucheron, T. (2012))

*If $F$ is log-concave, then for $1 \leq k \leq n/2$,*

$$\mathrm{Var}\left[X_{(k)}\right] \leq \mathbb{E}V_k \leq \frac{2}{k}\mathbb{E}\left[\left(\frac{1}{h(X_{(k+1)})}\right)^2\right].$$

*where $h = \frac{F'}{1-F}$.*

## Theorem (Absolute values of Gaussian - Boucheron, T. (2012))

*Let $n \geq 3$, let $X_{(1)}$ be the maximum of absolute values of $n$ standard independent Gaussian random variables,*

$$\mathrm{Var}[X_{(1)}] \leq \frac{1}{\log 2}\frac{8}{\log(2n) - \log(1 + 4\log\log(2n))}.$$

# Tail bounds

- Rényi's representation implies that: $(Y_{(1)}, \ldots, Y_{(n)})$ are sub-gamma as a sum of gamma variables.

## Sub-gamma on the right tail with variance factor $v$ and scale parameter $c$

$$\log \mathbb{E} e^{\lambda(X - \mathbb{E}X)} \leq \frac{\lambda^2 v}{2(1 - c\lambda)} \text{ for every } \lambda \quad \text{such that} \quad 0 < \lambda < 1/c\,.$$

## Bernstein's inequality

for $t > 0$, $\mathbb{P}\left\{X \geq \mathbb{E}X + \sqrt{2vt} + ct\right\} \leq \exp(-t)\,.$

- Expect $(X_{(1)}, \ldots, X_{(n)})$ to have a sub-gamma behavior, and so to satisfy a Bernstein-type inequality :
- Entropy method with the modified logarithmic Sobolev inequality of Massart (2000) and a "decoupling" inequality .

Thank you for your attention !