Institute of Geophysics
Polish Academy of Sciences

On accuracy of upper quantiles estimation

Iwona Markiewicz
Witold G. Strupczewski
Krzysztof Kochanek

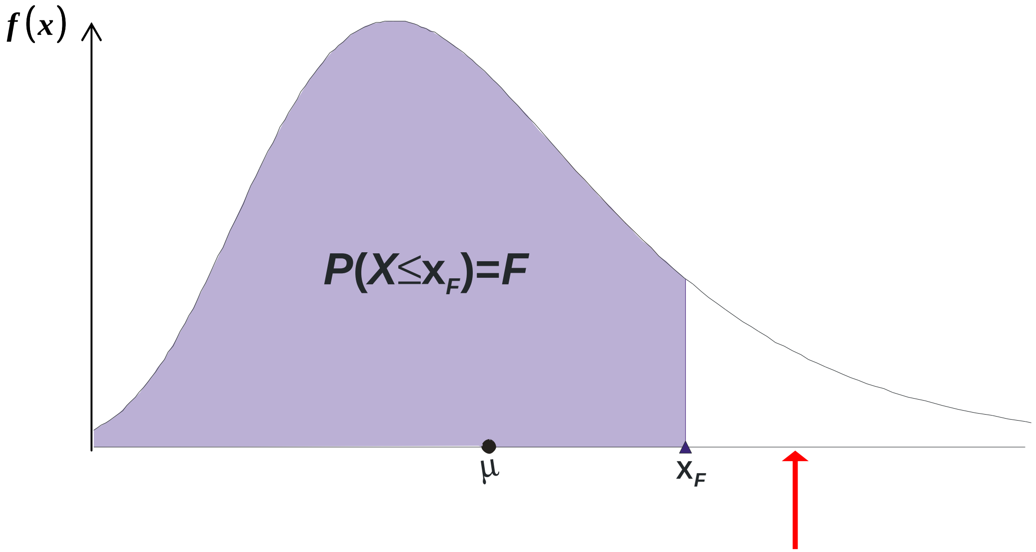
International Workshop ADVANCES IN STATISTICAL HYDROLOGY, May 23-25, 2010 Taormina, Italy

Contents

- Introduction
  - Objective
  - Estimation methods
  - Probability distributions
- Accuracy of upper quantile estimates - TRUE hypothetical pdf
  - Two-parameter distribution
  - Three-parameter distribution
- Accuracy of upper quantile estimates - FALSE hypothetical pdf
  - Two-parameter distribution
  - Three-parameter distribution
- Conclusions

Introduction

Flood Frequency Analysis (FFA) = estimation of upper quantiles of peak flows probability distribution, obtained from annual or partial duration series; assumed PDF is the statistical hypothesis



x_F - F quantile

$$P(X \le x_F) = \int_{-\infty}^{x_F} f(x)\,dx = F$$

Return period *T*

$$T = \frac{1}{1-F} = \frac{1}{P(X > x_F)}$$

$$T = 10 \leftrightarrow F = 0.9$$

$$T = 100 \leftrightarrow F = 0.99$$

$$T = 1000 \leftrightarrow F = 0.999$$

Problem

- Properties of estimation methods are identified with the case of known population PDF
- Hypothetical model differs from the true one !
  - upper part of PDF is outside the scope of actual observation range
  - measured peak flows are error-corrupted data and their quality information is low
  - no simple statistical model can reproduce the data set in its entire range of variability
  - probability of correct identification of PDF on the basis of short hydrological samples is very low

↓

Traditional approach based on the knowledge of theoretical distribution is not acceptable

Objective

THEORETICAL PROPERTIES OF ESTIMATION METHODS VARY SIGNIFICANTLY IN THE CASE OF MODEL MISSPECIFICATION

Estimation methods

- Method of (conventional) moments - MOM
- Method of linear moments - LMM
- Maximum likelihood method - MLM
- Method built on mean deviation - MDM

MDM	Location	Dispersion	Skewness
Measure	μ	$\delta_\mu = \int_{-\infty}^{+\infty} \lvert x-\mu \rvert \, dF(x)$	$\delta_S = \mu - x_{0.5}$
Dimensionless measure	-	$\delta C_V = \frac{\delta_\mu}{\mu}$	$\delta C_S = \frac{\delta_S}{\delta_\mu}$

Markiewicz, I. and Strupczewski ,W.G. (2009). Dispersion measures for flood frequency analysis. *Physics and Chemistry of the Earth*, 34: 670-678. DOI 10.1016/j.pce.2009.04.003.
Markiewicz, I., Strupczewski, W.G., Kochanek, K. and Singh, V.P. (2006). Relations between three dispersion measures used in flood frequency analysis. *Stochastic Environamental Research and Risk Assessment*, 20: 391-405. DOI 10.1007/s00477-006-0033-x.

Probability distributions

Distribution	Probability density function (PDF)
Log-normal 3 (LN3) $\varepsilon = 0$: log-normal 2 (LN2)	$f(x) = \frac{1}{(x-\varepsilon)b\sqrt{2\pi}} \exp\left[-\frac{(\ln(x-\varepsilon)-m)^2}{2b^2}\right]$ m - scale, $b > 0$ - shape; $\varepsilon < x < \infty$
Generalized extreme values (GEV) $\varepsilon = 0$: log-Gumbel (LG)	$f(x) = \frac{1}{\alpha}\left[-\frac{\kappa}{\alpha}(x-\varepsilon)\right]^{1/\kappa-1} \exp\left\{-\left[-\frac{\kappa}{\alpha}(x-\varepsilon)\right]^{1/\kappa}\right\}$ $\alpha > 0$ - scale, $\kappa < 0$ - shape; $\varepsilon < x < \infty$

True hypothetical distribution

- **Two-parameter distributions**
  ***T* = LN2, *H* = LN2 and *T* = LG, *H* = LG**

log-normal2, log-Gumbel
$\mu > 0$
$C_V = 0.2, 0.6, 1.0$
$N = 20\,(10)\,100$

MC = 20,000

→ $\delta RMSE(\hat{x}_{0.99}),\ \delta B(\hat{x}_{0.99})$

- **Three-parameter distributions**
  ***T* = LN3, *H* = LN3 and *T* = GEV, *H* = GEV**

log-normal3, GEV
$\mu = 0,\ \sigma = 1$
$C_S = 2.0, 4.0$
$N = 20\,(10)\,100$

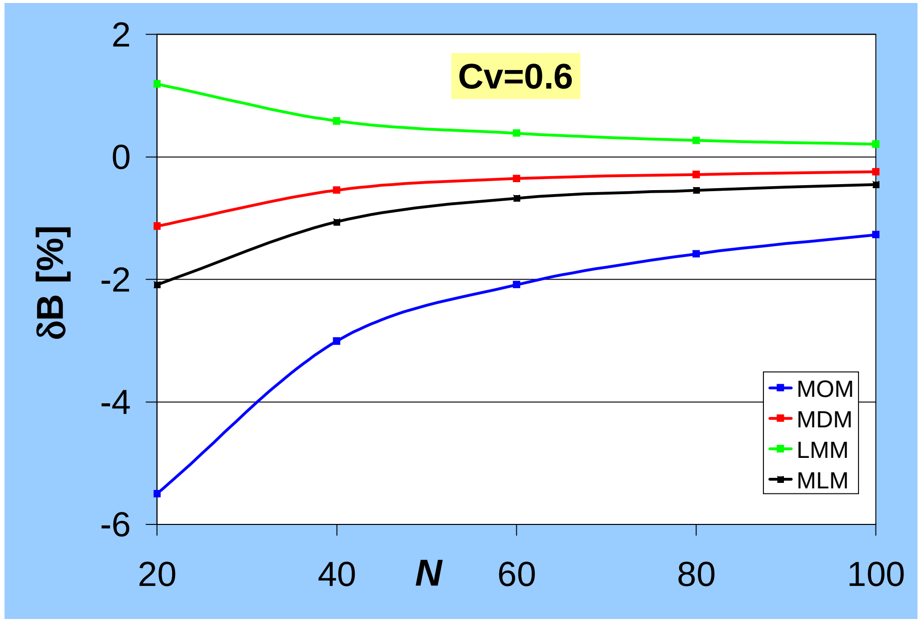
MC = 20,000

→ $\delta RMSE(\hat{x}_{0.99}),\ \delta B(\hat{x}_{0.99})$

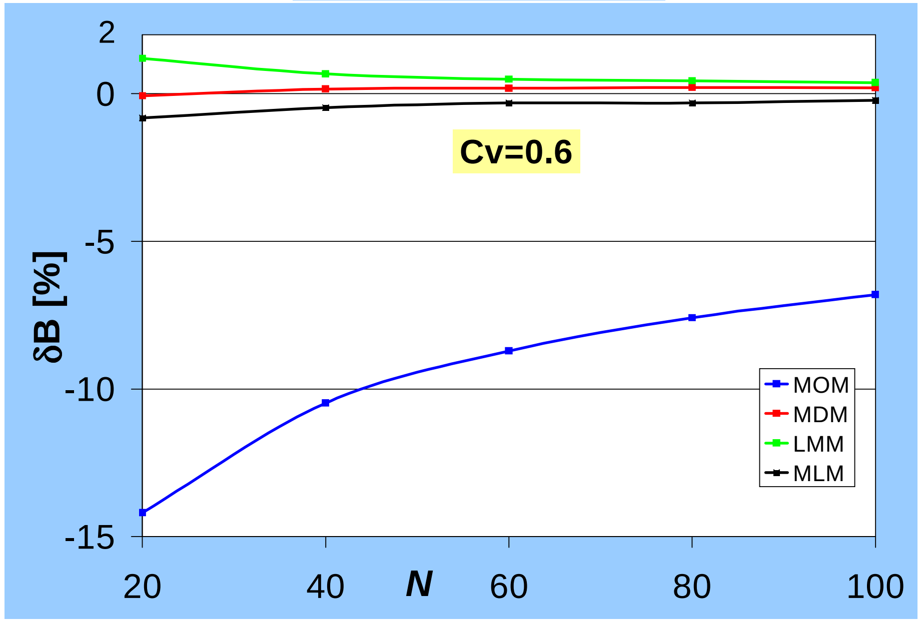
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta B\left(\hat{x}_{0.99}\right)=\frac{E\left(\hat{x}_{0.99}-x_{0.99}\right)}{x_{0.99}}$$

***T* = LN2, *H* = LN2**



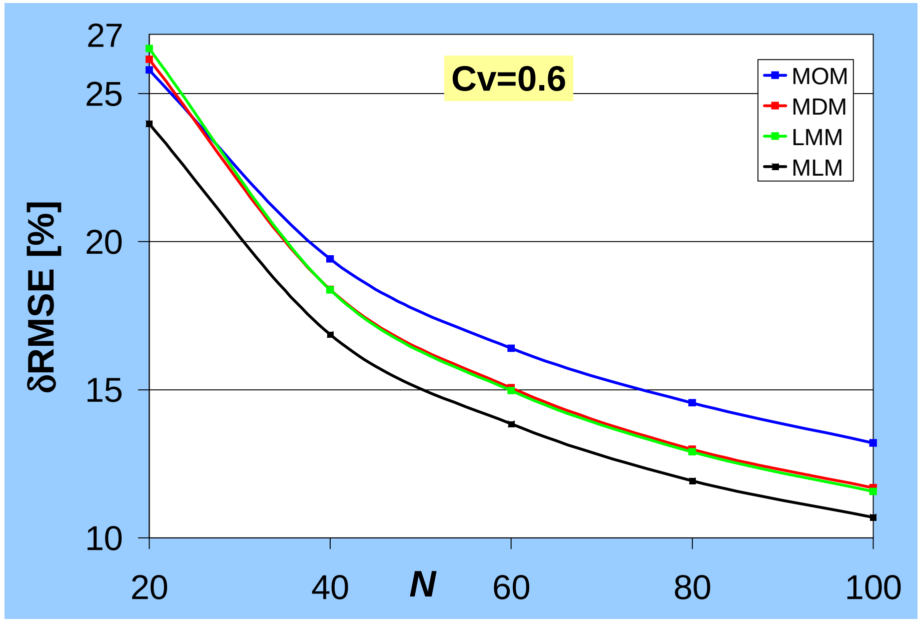
***T* = LG, *H* = LG**



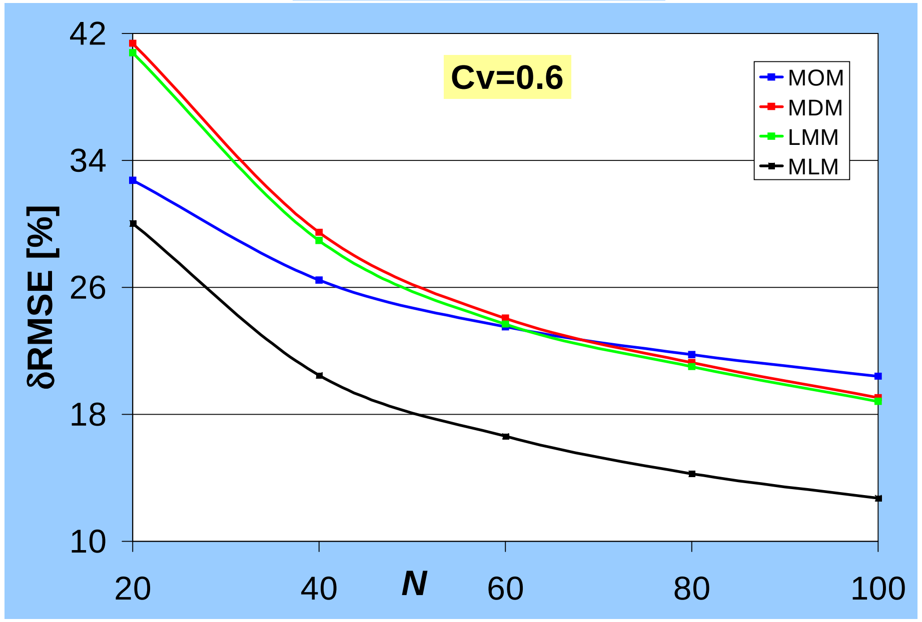
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta RMSE\left(\hat{x}_{0.99}\right) = \frac{\sqrt{E\left(\hat{x}_{0.99} - x_{0.99}\right)^2}}{x_{0.99}}$$

***T* = LN2, *H* = LN2**



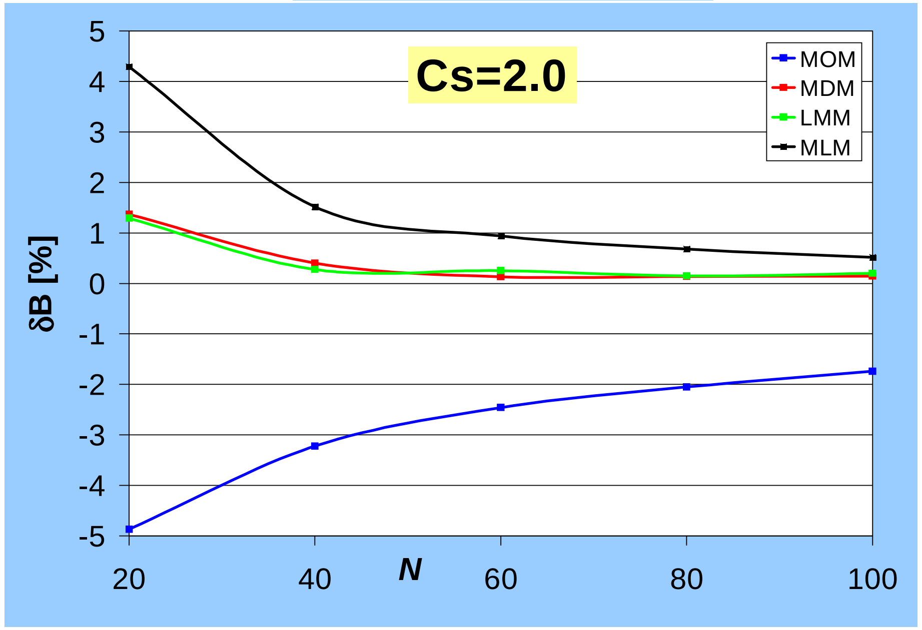
***T* = LG, *H* = LG**



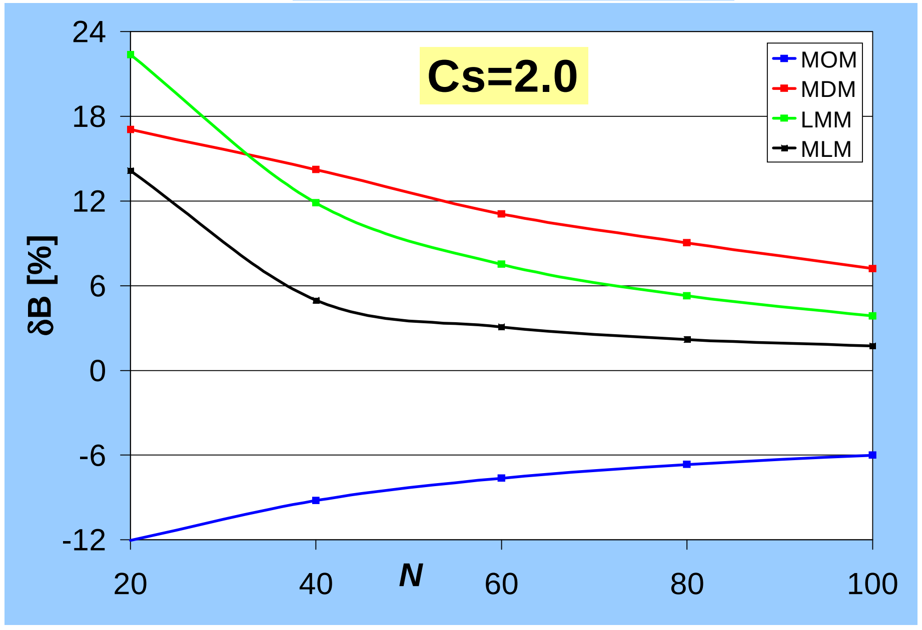
Accuracy of upper quantile estimates three-parameter PDFs

$$\delta B(\hat{x}_{0.99}) = \frac{E(\hat{x}_{0.99} - x_{0.99})}{x_{0.99}}$$

***T* = LN3, *H* = LN3**



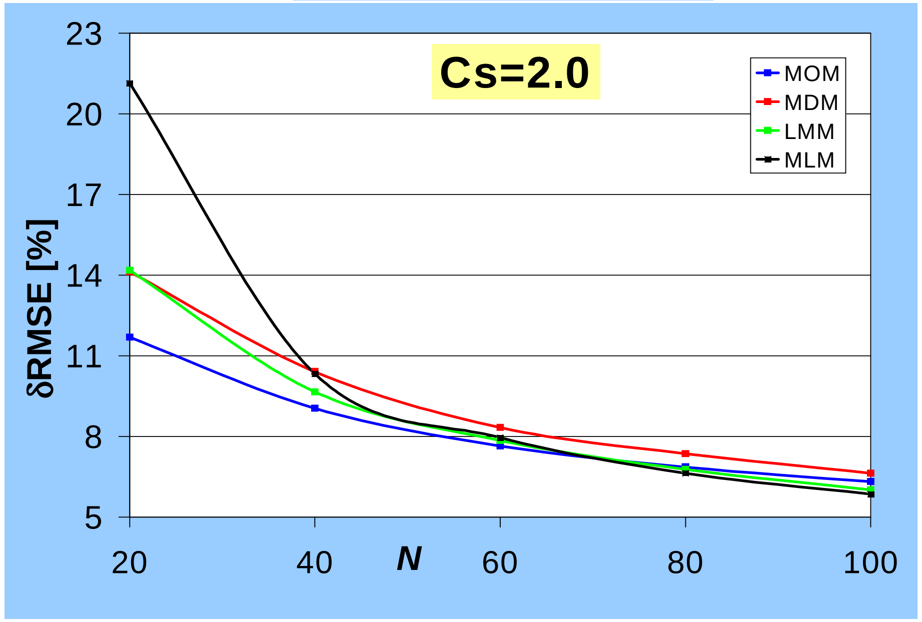
***T* = GEV, *H* = GEV**



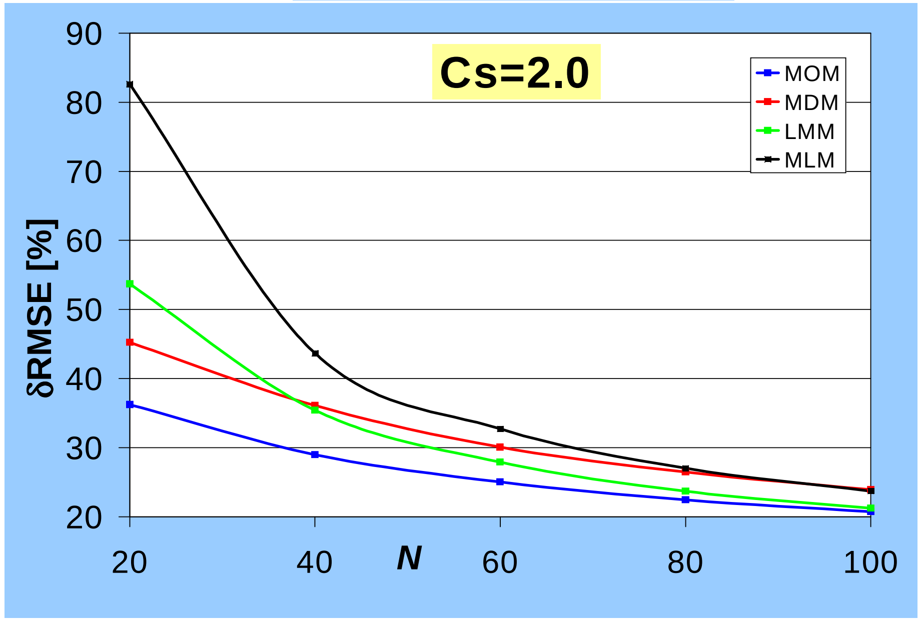
Accuracy of upper quantile estimates three-parameter PDFs

$$\delta RMSE\left(\hat{x}_{0.99}\right) = \frac{\sqrt{E\left(\hat{x}_{0.99} - x_{0.99}\right)^2}}{x_{0.99}}$$

T = LN3, H = LN3



T = GEV, H = GEV



False hypothetical distribution

- **Two-parameter distributions**
  ***T* = LN2, *H* = LG and *T* = LG, *H* = LN2**

log-normal2, log-Gumbel
$\mu > 0$
$C_V = 0.2, 0.6, 1.0$
$N = 20\,(10)\,100$

MC = 20,000

→ $\delta RMSE(\hat{x}_{0.99}),\ \delta B(\hat{x}_{0.99})$

- **Three-parameter distributions**
  ***T* = LN3, *H* = GEV and *T* = GEV, *H* = LN3**

log-normal3, GEV
$\mu = 0,\ \sigma = 1$
$C_S = 2.0, 4.0$
$N = 20\,(10)\,100$

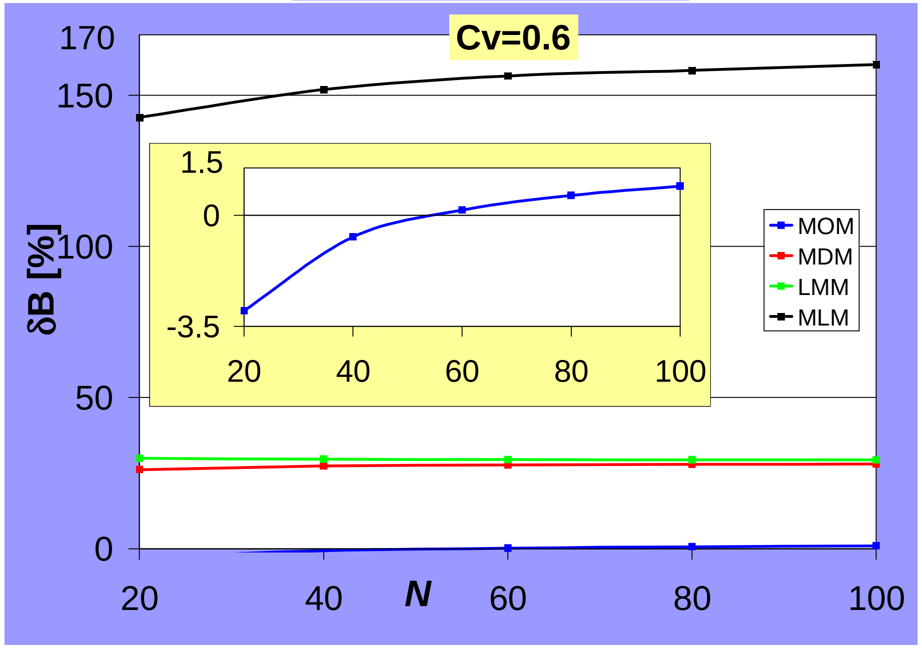
MC = 20,000

→ $\delta RMSE(\hat{x}_{0.99}),\ \delta B(\hat{x}_{0.99})$

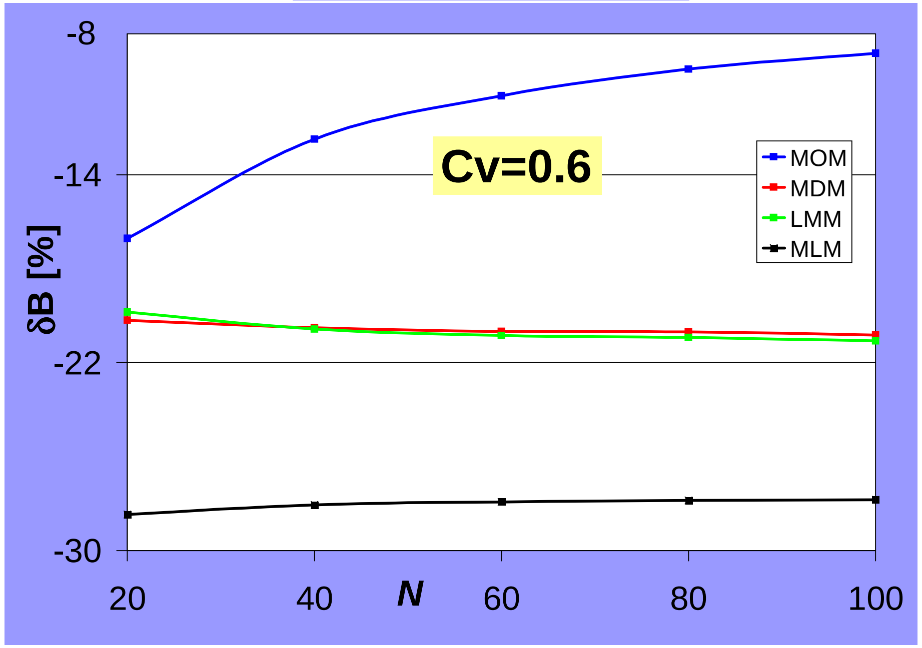
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta B(\hat{x}_{0.99}) = \frac{E(\hat{x}_{0.99} - x_{0.99})}{x_{0.99}}$$

***T* = LN2, *H* = LG**



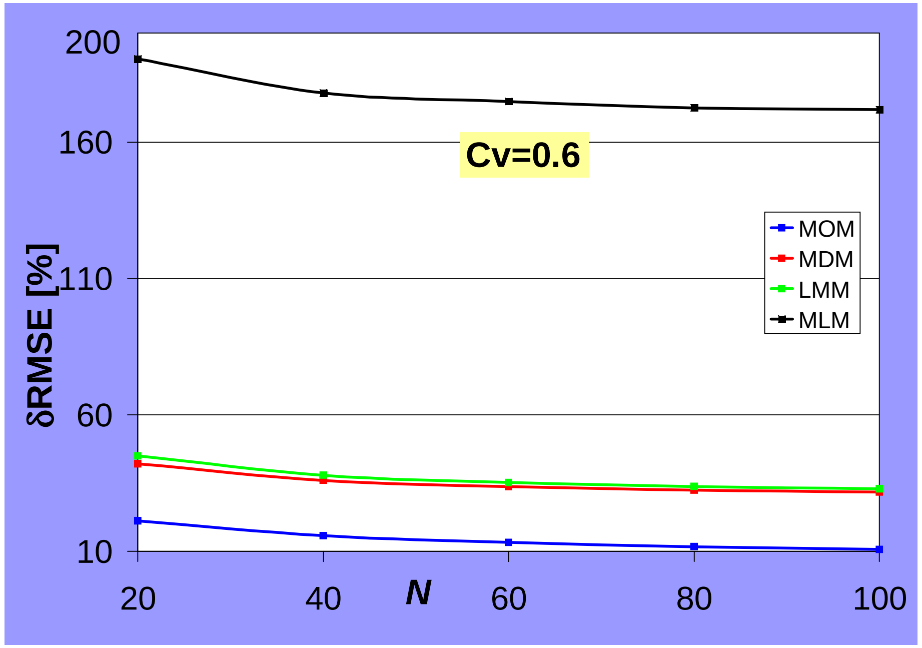
***T* = LG, *H* = LN2**



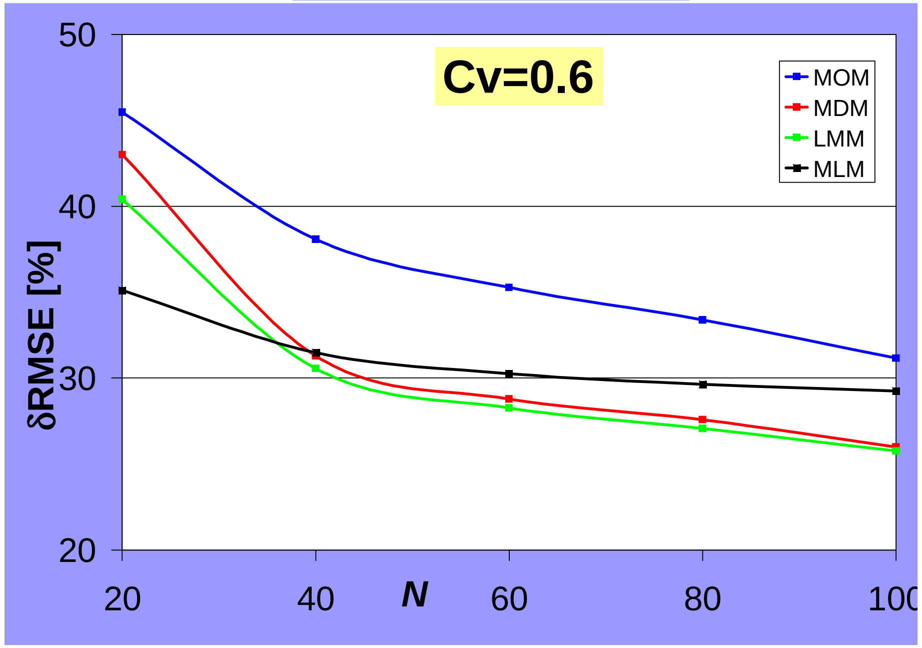
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta RMSE\left(\hat{x}_{0.99}\right)=\frac{\sqrt{E\left(\hat{x}_{0.99}-x_{0.99}\right)^2}}{x_{0.99}}$$

T = LN2, H = LG



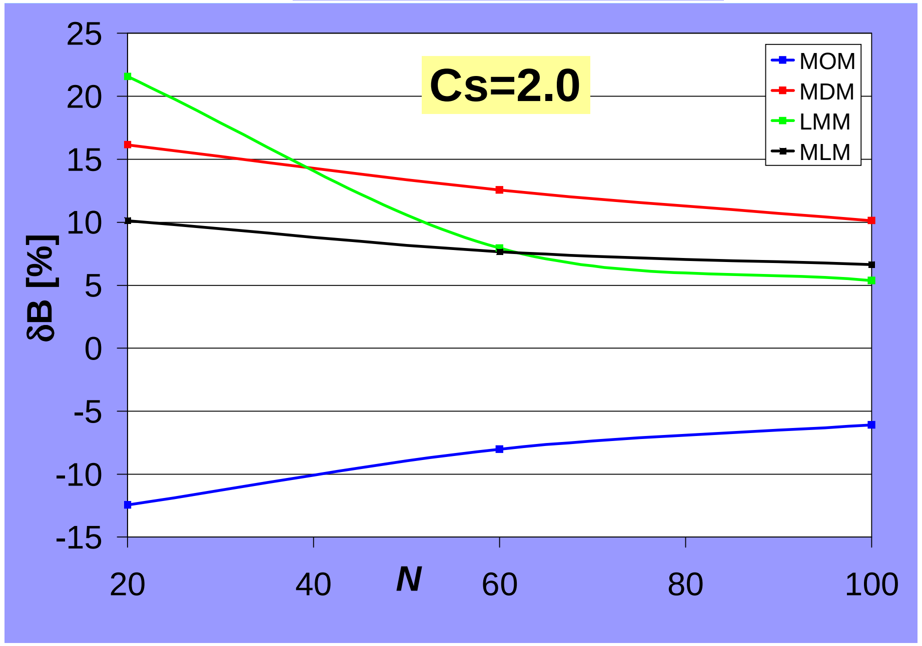
T = LG, H = LN2



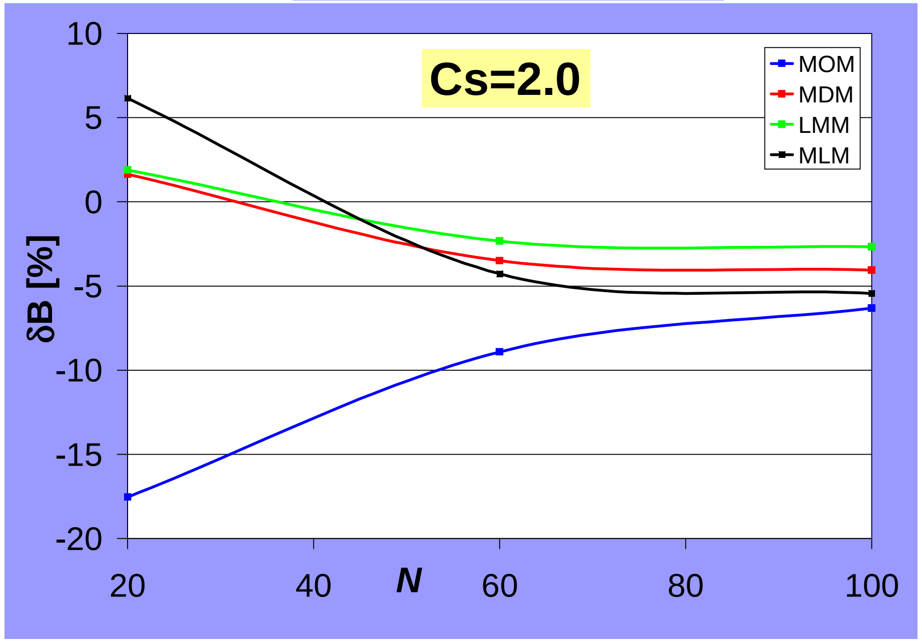
Accuracy of upper quantile estimates three-parameter PDFs

$$\delta B(\hat{x}_{0.99}) = \frac{E(\hat{x}_{0.99} - x_{0.99})}{x_{0.99}}$$

T = LN3, H = GEV



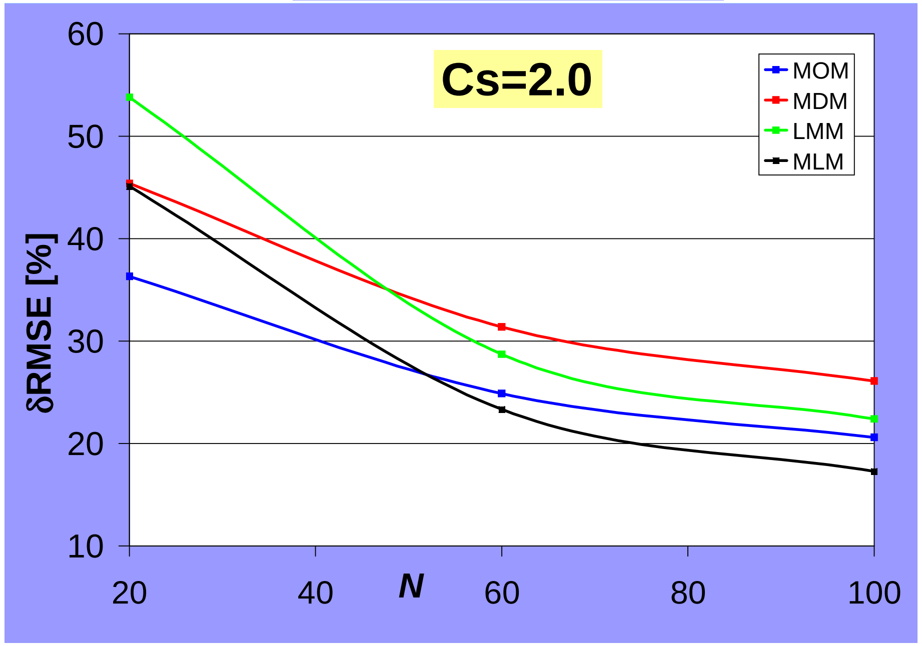
T = GEV, H = LN3



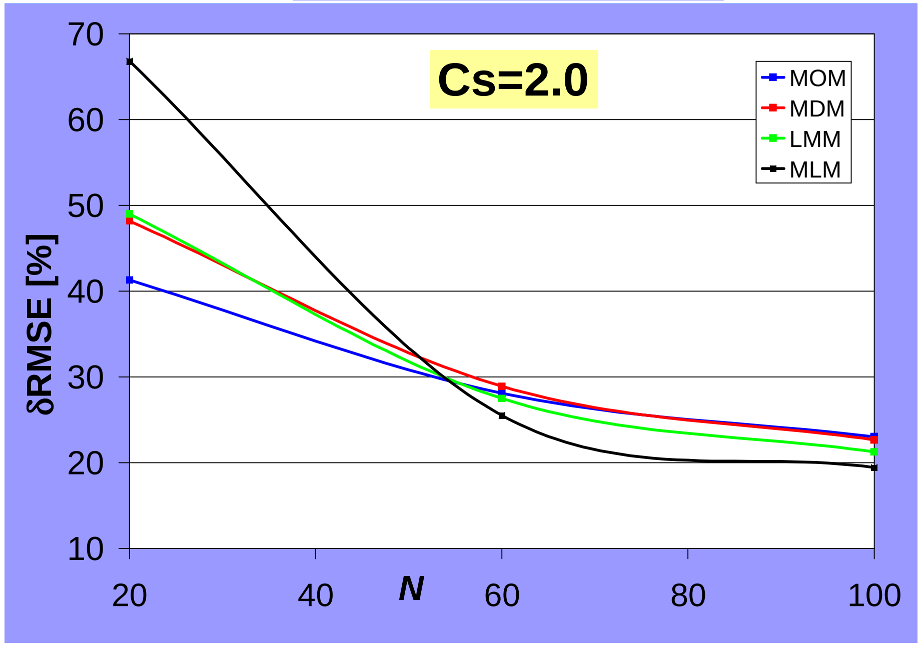
Accuracy of upper quantile estimates three-parameter PDFs

$$\delta RMSE\left(\hat{x}_{0.99}\right)=\frac{\sqrt{E\left(\hat{x}_{0.99}-x_{0.99}\right)^2}}{x_{0.99}}$$

T = LN3, H = GEV



T = GEV, H = LN3



Conclusions

- ✓ **Ranking of estimation methods in respect to upper quantile accuracy depends on:**
  - type of distributions, both real and hypothetical
  - number of distribution parameters
  - sample size

- ✓ **For two-parameter distributions, in the case of model misspecification, the MLM yields the highest bias of quantile estimates regardless on the sample size, while the MOM the smallest one**

- ✓ **Presented analysis can be a source of information about the properties of selected distribution and estimation (D/E) procedures**

- ✓ **Studies should be extended for other distributions**

THANK YOU