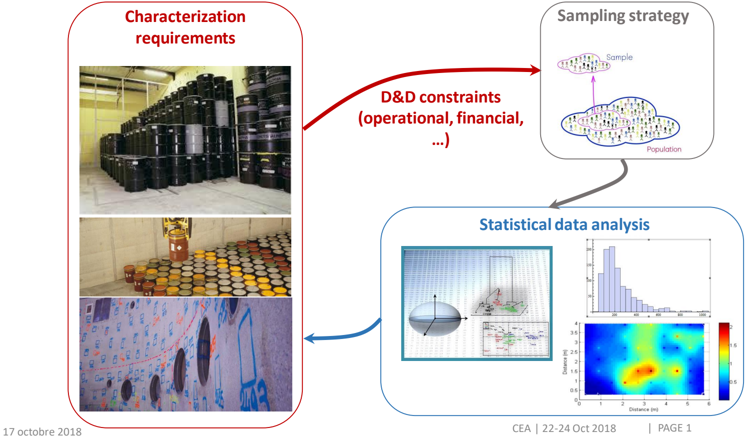
Statistical characterization for nuclear dismantling applications with small data sets

G. Blatman[1], T. Delage[1], B. Iooss[1], N. Pérot[2]

[1] EDF R&D, France; [2] CEA/DEN, France

SFdS/MASCOT-NUM meeting - Big ideas for small data

Context


Characterization requirements
D&D constraints (operational, financial, ...)
Sampling strategy
Sample
Population
Statistical data analysis
Distance (m)
17 octobre 2018
CEA | 22-24 Oct 2018 | PAGE 1

Problem statement

Context : Radiological characterization of contaminated elements from nuclear facilities

Problem : Small number of available data

$\longrightarrow$ Inappropriate statistical tools (e.g. Gaussian approximation) to determinate risk confidence bounds

Ex : The 2σ rule (95% of values inside $\pm 2\sigma$) works in the Gaussian case

Risks of a wrong estimation of the contamination : Under-estimation (impact on safety) or over-estimation (impact on economic cost)

Strategy : Resort to robust inequalities which only depend on weak assumptions about the statistical distribution of the measured quantity

Outline

Outline

1. Prediction, tolerance and confidence intervals
2. Application to real measurements

Probabilistic framework

- Consider a set of measures $\mathcal{X} = \{X_1, \ldots, X_n\}$ of a given quantity
- They are assumed to be independent copies of a continuous random variable X with unknown distribution (but finite mean and variance)
- In the context of risk analysis, it is relevant to estimate from the data the three following kinds of probabilistic intervals :

Unilateral prediction interval	Unilateral tolerance interval	Bilateral confidence interval on $\mu = \mathbb{E}[X]$
$\mathbb{P}[X \leqslant s] \geqslant \gamma$	$\mathbb{P}[\mathbb{P}[X \leqslant s] \geqslant \gamma] \geqslant \beta$	$\mathbb{P}[s_1 \leqslant \mu \leqslant s_2] \geqslant \gamma$

s, s_1, s_2 : threshold values γ, β : prescribed probabilities (e.g. 95%)

$\alpha = 1 - \gamma$: probabilistic risk bound ; then $\mathbb{P}[X \geqslant s] \leqslant \alpha$

Probabilistic framework

- Consider a set of measures $\mathcal{X} = \{X_1, \ldots, X_n\}$ of a given quantity
- They are assumed to be independent copies of a continuous random variable X with unknown distribution (but finite mean and variance)
- In the context of risk analysis, it is relevant to estimate from the data the three following kinds of probabilistic intervals :

Unilateral prediction interval	Unilateral tolerance interval	Bilateral confidence interval on $\mu = \mathbb{E}[X]$
$\mathbb{P}[X \leqslant s] \geqslant \gamma$	$\mathbb{P}[\mathbb{P}[X \leqslant s] \geqslant \gamma] \geqslant \beta$	$\mathbb{P}[s_1 \leqslant \mu \leqslant s_2] \geqslant \gamma$

s, s_1, s_2 : threshold values γ, β : prescribed probabilities (e.g. 95%)

$\alpha = 1 - \gamma$: probabilistic risk bound ; then $\mathbb{P}[X \geqslant s] \leqslant \alpha$

Intervals based on Gaussian approximation

Notation : $X \sim \mathsf{N}(\mu, \sigma)$, $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$, $z_u = u$-$\mathcal{N}(0,1)$-quantile

Gaussian case with known (μ, σ) : The exact α-prediction interval is

$$s = \mu + \sigma z_{1-\alpha}$$

Gaussian case with unknown (μ, σ) : The exact α/β-tolerance interval is

$$s = \bar{X}_n + t_{n-1,\beta,\sqrt{n} z_{1-\alpha}} \frac{S_n}{\sqrt{n}}$$

However these are only approximations if X is not Gaussian, which may reveal poor if n is small and/or X is highly skewed

Intervals based on Gaussian approximation

Notation : $X \sim \mathsf{N}(\mu, \sigma)$, $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$, $z_u = u$-$\mathcal{N}(0,1)$-quantile

Gaussian case with known (μ, σ) : The exact α-prediction interval is

$$s = \mu + \sigma z_{1-\alpha}$$

Gaussian case with unknown (μ, σ) : The exact α/β-tolerance interval is

$$s = \bar{X}_n + t_{n-1,\beta,\sqrt{n}z_{1-\alpha}} \frac{S_n}{\sqrt{n}}$$

However these are only approximations if X is not Gaussian, which may reveal poor if n is small and/or X is highly skewed

Intervals based on Gaussian approximation

Notation : $X \sim \mathsf{N}(\mu, \sigma)$, $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}\mathsf{ar}[X]$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$, $z_u = u$-$\mathcal{N}(0,1)$-quantile

Gaussian case with known (μ, σ) : The exact α-prediction interval is

$$s = \mu + \sigma z_{1-\alpha}$$

Gaussian case with unknown (μ, σ) : The exact α/β-tolerance interval is

$$s = \bar{X}_n + t_{n-1,\beta,\sqrt{n}z_{1-\alpha}} \frac{S_n}{\sqrt{n}}$$

However these are only approximations if X is not Gaussian, which may reveal poor if n is small and/or X is highly skewed

Intervals based on Gaussian approximation

Notation : $X \sim \mathsf{N}(\mu, \sigma) \quad , \quad \mu = \mathbb{E}[X] \quad , \quad \sigma^2 = \mathbb{V}\mathrm{ar}[X]$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$, $z_u = u$-$\mathcal{N}(0,1)$-quantile

Gaussian case with known (μ, σ) : The exact α-prediction interval is

$$s \;=\; \mu + \sigma z_{1-\alpha}$$

Gaussian case with unknown (μ, σ) : The exact α/β-tolerance interval is

$$s \;=\; \bar{X}_n + t_{n-1,\beta,\sqrt{n}z_{1-\alpha}} \frac{S_n}{\sqrt{n}}$$

However these are only approximations if X is not Gaussian, which may reveal poor if n is small and/or X is highly skewed

Intervals based on concentration inequalities

We recall that we look for $\mathbb{P}[X \geqslant s] \leqslant \alpha$

Concentration inequalities give

$$s = \bar{X}_n + t \quad \text{and} \quad \alpha = \left(1 + \frac{t^2}{kS_n^2}\right)^{-1}$$

with $t \geq 0$ and k a positive constant

In practice, either s is fixed, either α is fixed (then t is directly recovered)

Inequality name	Value of k	Assumptions
Bienaymé-Chebyshev (BC)	1	None
Camp-Meidell (CM)	$4/9$	Unimodal pdf
Van Dantzig (VD)	$3/8$	Convex pdf tails

Note : Camp-Meidell inequality gives the so-called "3σ rule"

Intervals based on concentration inequalities

We recall that we look for $\mathbb{P}[X \geqslant s] \leqslant \alpha$

Concentration inequalities give

$$s = \bar{X}_n + t \quad \text{and} \quad \alpha = \left(1 + \frac{t^2}{kS_n^2}\right)^{-1}$$

with $t \geq 0$ and k a positive constant

In practice, either s is fixed, either α is fixed (then t is directly recovered)

Inequality name	Value of k	Assumptions
Bienaymé-Chebyshev (BC)	1	None
Camp-Meidell (CM)	$4/9$	Unimodal pdf
Van Dantzig (VD)	$3/8$	Convex pdf tails

Note : Camp-Meidell inequality gives the so-called "3σ rule"

Intervals based on concentration inequalities

We recall that we look for $\mathbb{P}[X \geqslant s] \leqslant \alpha$

Concentration inequalities give

$$s = \bar{X}_n + t \quad \text{and} \quad \alpha = \left(1 + \frac{t^2}{kS_n^2}\right)^{-1}$$

with $t \geq 0$ and k a positive constant

In practice, either s is fixed, either α is fixed (then t is directly recovered)

Inequality name	Value of k	Assumptions
Bienaymé-Chebyshev (BC)	1	None
Camp-Meidell (CM)	$4/9$	Unimodal pdf
Van Dantzig (VD)	$3/8$	Convex pdf tails

Note : Camp-Meidell inequality gives the so-called "3σ rule"

Intervals based on concentration inequalities

We recall that we look for $\mathbb{P}\left[X \geqslant s\right] \leqslant \alpha$

Concentration inequalities give

$$s = \bar{X}_n + t \quad \text{and} \quad \alpha = \left(1 + \frac{t^2}{kS_n^2}\right)^{-1}$$

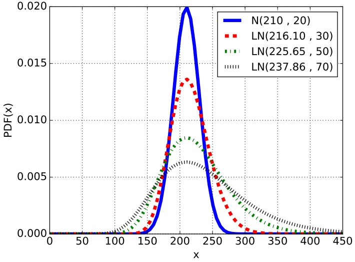
with $t \geq 0$ and k a positive constant

In practice, either s is fixed, either α is fixed (then t is directly recovered)

Inequality name	Value of k	Assumptions
Bienaymé-Chebyshev (BC)	1	None
Camp-Meidell (CM)	$4/9$	Unimodal pdf
Van Dantzig (VD)	$3/8$	Convex pdf tails

Note : Camp-Meidell inequality gives the so-called "3σ rule"

Examples of risk estimates with known distributions of X



$\alpha = 0.05$	$\mathcal{N}(210, 20)$	$\mathcal{LN}(216, 30)$	$\mathcal{LN}(226, 50)$	$\mathcal{LN}(238, 70)$
Gauss	0.05	0.04	0.04	0.03
BC	0.27	0.25	0.23	0.23
CM	0.14	0.13	0.12	0.12
VD	0.12	0.11	0.10	0.10

Extension to tolerance intervals by bootstrapping

A β-confidence level is required due to the empirical estimation of the mean and standard deviation

$$\mathbb{P}\left[\mathbb{P}\left[X \geqslant s\right] \leqslant \alpha\right] \geqslant \beta$$

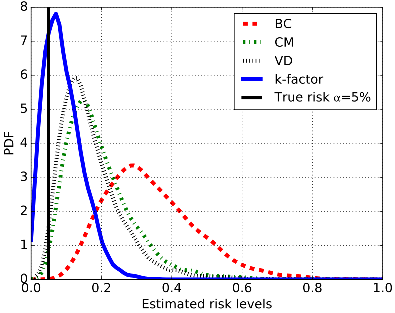
From sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, we repeat B times (e.g. $B = 500$) :

- Create a new n-size sample $\mathcal{X}'$ by sampling with replacement in $\mathcal{X}$,
- Compute $\bar{X}_n$ and S_n,
- If s (resp. α) is fixed, compute t and α (resp. s)

From the B-size sample of α values (resp. s values), take the β-quantile of α (resp. s)

Numerical experiments : $X \sim \mathcal{LN}(238, 70)$ and $n = 30$

Keep $\beta = 95\%$-quantile of bootstrap sample ($B = 500$) of α estimates



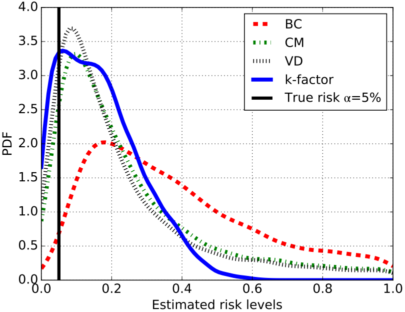
Statistical distributions of the quantiles ($N = 5000$ repetitions)

Proportion of non-conservative estimates of the exact $\alpha = 5\%$:

k-factor	BC	CM	VD
0.25	0.00	0.00	0.01

Numerical experiments : $X \sim \mathcal{LN}(238, 70)$ and $n = 10$

Keep $\beta = 95\%$-quantile of bootstrap sample ($B = 500$) of α estimates



Statistical distributions of the quantiles ($N = 5000$ repetitions)

Proportion of non-conservative estimates of the exact $\alpha = 5\%$:

k-factor	BC	CM	VD
0.17	0.01	0.07	0.10

Outline

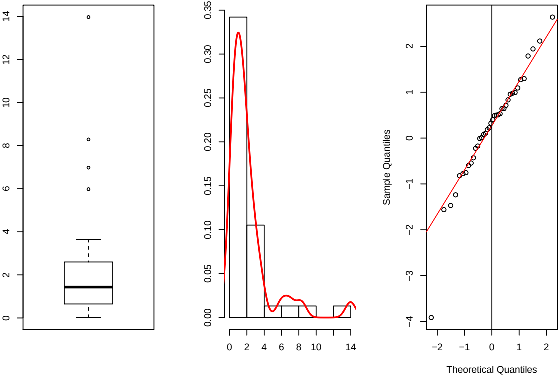
Data : H_2 flow rates of radioactive waste drums

Evaluation of H_2 flow rates (in l / drum / year) required for disposal in final waste repositories

Population of several thousands drums

Data : H_2 flow rates of radioactive waste drums

Measures on a random sample of size $n = 38$, $(\bar{X}_n, S_n) = (2.18, 2.67)$



Adequacy to a parametric distribution (as the log-normal one) is rejected by statistical tests

Some results obtained with the Camp-Meidell inequality

Estimation of the risk α of threshold exceedance ($\beta = 0.95$) :

$$s = 5 \text{ gives } \alpha = 58\%$$
$$s = 10 \text{ gives } \alpha = 11\%$$
$$s = 15 \text{ gives } \alpha = 4\%$$

Estimation of the relative error on the mean flow rate (the empirical mean is equal to 2.18 l/drum/year) :

- 31% = relative error on the estimation of the mean H_2 flow rate with $(\alpha, \beta) = (0.75, 0.95)$
- 93 = sample size required to reach a 20%-relative error on the estimation of the mean with $(\alpha, \beta) = (0.75, 0.95)$

Some results obtained with the Camp-Meidell inequality

Estimation of the risk α of threshold exceedance ($\beta = 0.95$) :

$$s = 5 \text{ gives } \alpha = 58\%$$
$$s = 10 \text{ gives } \alpha = 11\%$$
$$s = 15 \text{ gives } \alpha = 4\%$$

Estimation of the relative error on the mean flow rate (the empirical mean is equal to 2.18 l/drum/year) :

- $31\% =$ relative error on the estimation of the mean H_2 flow rate with $(\alpha, \beta) = (0.75, 0.95)$
- $93 =$ sample size required to reach a 20%-relative error on the estimation of the mean with $(\alpha, \beta) = (0.75, 0.95)$

Conclusions and prospects

- Be careful with the Gaussian approximation especially for small data samples
- Concentration inequalities provide robust risk bound and confidence interval for the mean
- Their degrees of conservatism are linked to explicit assumptions on the distribution of the studied variable

- Apply more sophisticated concentration inequalities in order to give tighter bounds

Bibliography

- G. Blatman, T. Delage, B. Iooss and N. Pérot. Probabilistic risk bounds for the characterization of radiological contamination. *The European Journal of Physics - Nuclear Sciences & Technology (EPJ-N)*, 3, 23, 2017

- P-C. Pupion and G. Pupion, Méthodes statistiques applicables aux petits échantillons, Hermann, 2010
- G.J. Hahn and W.Q. Meeker, Statistical intervals. A guide for practitionners, Wiley-Interscience, 1991