Uncertainty

Florent Bourgeois

UNCERTAINTY

Master Class 2020 – LGC

Florent Bourgeois, Prof.

https://lgc.cnrs.fr/annuaire/florent-bourgeois/

http://lgc.cnrs.fr

UNCERTAINTY

"*The only certainty is uncertainty*"

~ Pliny the Elder (AD 23/24 – 79)

"Gaius Plinius Secundus (AD 23/24–79), called Pliny the Elder (/ˈplɪni/), was a Roman author, **a naturalist** and natural philosopher, a naval and army commander of the early Roman Empire, and a friend of emperor Vespasian. He wrote the encyclopedic Naturalis Historia (Natural History), which became an editorial model for encyclopedias. He spent most of his spare time **studying, writing, and investigating natural** and geographic **phenomena** in the field." (Quoted source: Wikipedia)

http://lgc.cnrs.fr

UNCERTAINTY

A fact of nature

Field observations, i.e. "real data" are uncertain, always !

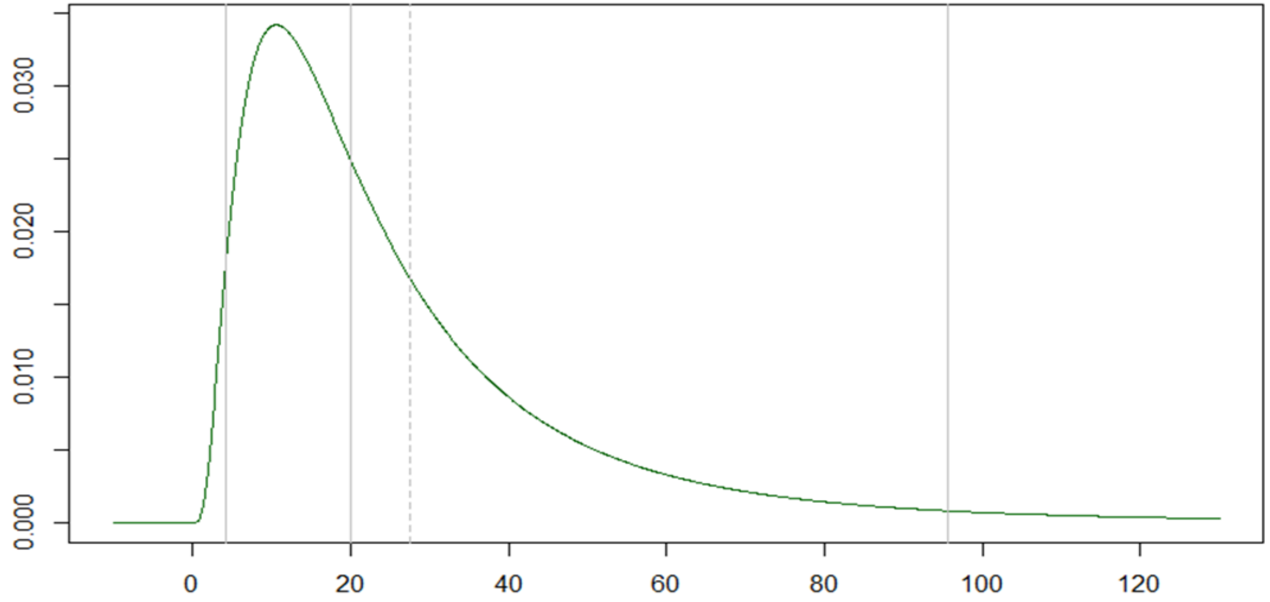
Scientists and engineers use real data in order to derive working solutions to real-world problems. Uncertainties are therefore inevitable in science and engineering.

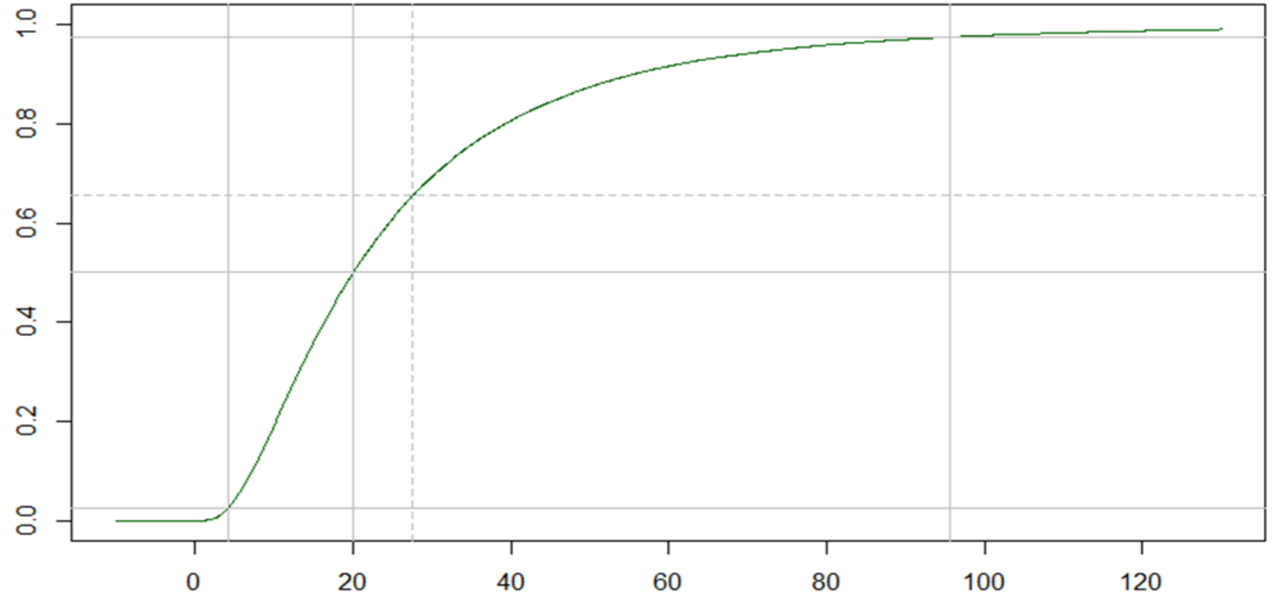
No real data should be trusted nor reported without error bars !

Real data can be abstracted as random variables (r.v.), which are characterized fully by probability distribution functions.

Once a probability distribution function is assigned to a variable of interest, all is known about the said variable.

Error bars are estimated from the r.v.'s probability distribution function



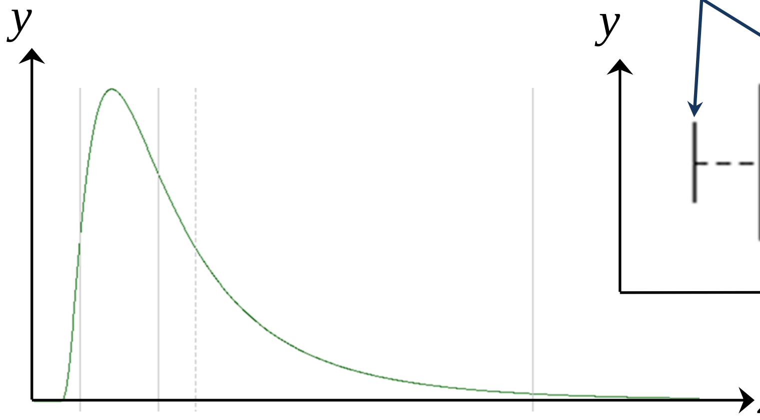


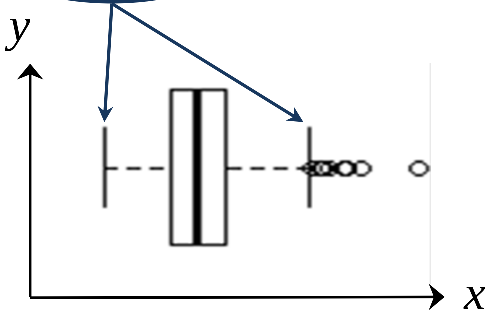
LABORATOIRE
DE GÉNIE
CHIMIQUE
TOULOUSE • UMR 5503

UNCERTAINTY

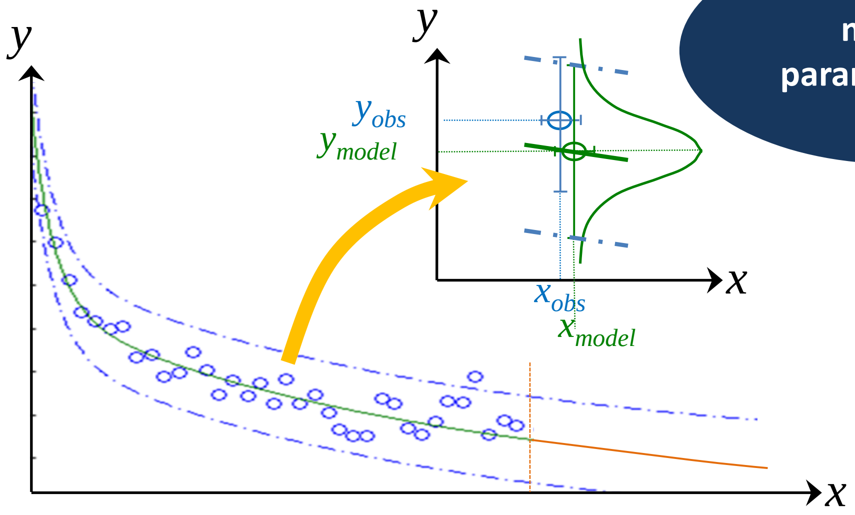
The error bar

Estimated models parameters are r.v.

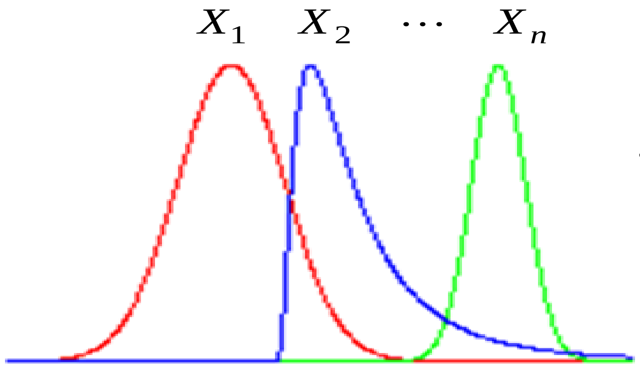


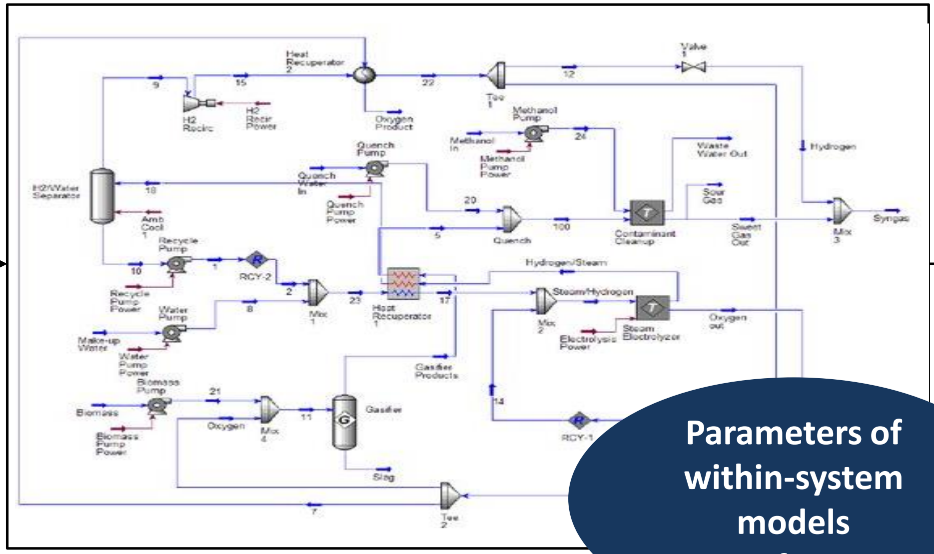


An observation



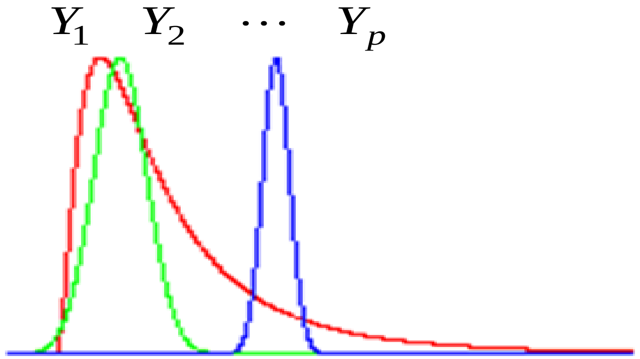
A model





A system

Parameters of within-system models are often r.v.



http://lgc.cnrs.fr

UNCERTAINTY

Computational statistics

- **Computational statistics** merges statistics with numerical methods.

- Computational statistics' methods and tools make analysis and manipulation of statistical data possible without mastering the (elitist) art of formal statistical (paper-and-pencil) derivations, i.e. without a formal mathematician training in statistics (up to a point…).

- Statistical analyses and calculations can now be performed using dedicated and easily accessible computing tools.

- Such tools offer powerful capabilities for visualizing statistical data and results. "Graphical statistics" makes statistical concepts (distribution functions, confidence intervals, inference testing,…) easier to undertand.

Draw your statistics !

- That said, it remains essential to understand statistical concepts in order to make a sound use of computational statistics and carry out quality analysis of real data.

UNCERTAINTY

Computational statistics tools

- Primary function : statistical data manipulation, analysis, modelling and visualisation
- Now a full-fledged scientific computing tool
- Free, all platforms (https://www.r-project.org/)
- Large and growing community of users and developers (https://www.r-bloggers.com/) and a myriad of dedicated self-learning websites and textbooks.

- Recommended GUI : RStudio (http://www.rstudio.com/)
- Interactive web applications with Shiny (https://shiny.rstudio.com/)
- Notebooks
- MSExcel® interoperability with BERT (https://oatao.univ-toulouse.fr/24353/)

UNCERTAINTY

An unworthy illustrative example

model: $Y = 48\dfrac{X_1 X_2 X_3}{X_4 X_5}$

i	X_i (obs.)	ΔX_i
1	0.5	0.001
2	0.05	0.0002
3	5.5	0.28
4	0.0544	0.0005
5	18.7	0.28

$\Rightarrow Y = 6.49$ $\pm \Delta Y$

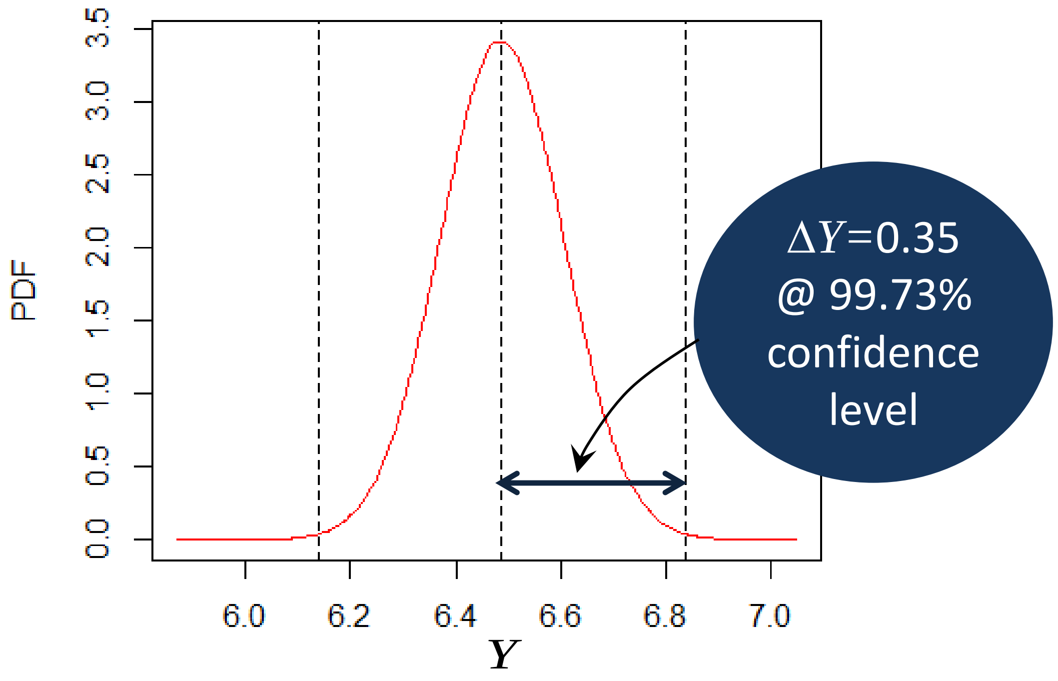
Paper-and-pencil solution

- By propagation of variance (simple derivation, but derivation nonetheless, just making a point here…)
- $\Delta Y = Y\sqrt{\sum_{i=1}^{5}\left(\dfrac{\Delta X_i}{X_i}\right)^2} = 0.35$
- Is this derivation correct? Are there underlying assumptions? Is there a confidence associated?

Monte-Carlo solution

- Hyp. : X_i's are normally distributed
- Hyp. : ΔX_i's are given at the 99.73% confidence level
- Monte-Carlo simulation:
  - •Select a sample size, say $n = 10^6$
  - •Generate n sets $[X_1, \ldots, X_5]$ by random sampling
  - •Calculate n values of Y using the model equation
  - •Analyse the statistics of the $n \times Y$ values

$\Rightarrow$



http://lgc.cnrs.fr

LABORATOIRE DE GÉNIE CHIMIQUE
TOULOUSE • UMR 5503

UNCERTAINTY

An unworthy illustrative example

model: $Y = 48\dfrac{X_1 X_2 X_3}{X_4 X_5}$

i	X_i (obs.)	ΔX_i
1	0.5	0.001
2	0.05	0.0002
3	5.5	0.28
4	0.0544	0.0005
5	18.7	0.28

$\Rightarrow Y = 6.49$ $\pm \Delta Y$

Paper-and-pencil solution

- By propagation of variance (simple derivation, but derivation nonetheless, just making a point here…)
- $\Delta Y = Y\sqrt{\sum_{i=1}^{5}\left(\dfrac{\Delta X_i}{X_i}\right)^2} = 0.35$
- Is this derivation correct? What are the underlying assumptions? Is there a confidence associated?

Monte-Carlo solution

- Hyp. : X_i's are normally distributed
- Hyp. : ΔX_i's are given at the 99.73% confidence level
- Monte-Carlo simulation:
  - Select a sample size, say $n = 10^6$
  - Generate n sets $[X_1, \ldots, X_5]$ by random sampling
  - Calculate n values of Y using the model equation
  - Analyse the statistics of the $n\times Y$ values

$\Rightarrow$

```
library(mvtnorm)
meanX<-c(0.5,0.05,5.5,0.0544,18.7) # Xi
deltaX<-c(1e-3,2e-4,2.8e-1,5e-4,2.8e-1) # ΔXi
n<-1e6 # Monte-Carlo sample size
prob<-0.9973 # Confidence associated with ΔX
varX<-diag((deltaX/qnorm(1-(1-prob)/2))^2) # X ~ Normal
Xsim<-rmvnorm(n,mu=X,sigma=varX) # generate n×[X1,…, X5]
Ysim<-48*Xsim[,1]*Xsim[,2]*Xsim[,3]/(Xsim[,4]*Xsim[,5]) # generate n×Y
q <- quantile(ecdf(Ysim), probs=c((1- prob)/2,1-(1- prob)/2)) # quantiles
Ysimpdf <- density(Ysim) # PDF
print(paste("deltaY=",q[2]-mean(Ysim)))
plot(Ysimpdf$x, Ysimpdf$y, type="l", col="red", ylab="PDF", xlab="Y")
abline(v=c(mean(Ysim), q), lty="dashed")
```

Sensitivity analysis of hypotheses

http://lgc.cnrs.fr

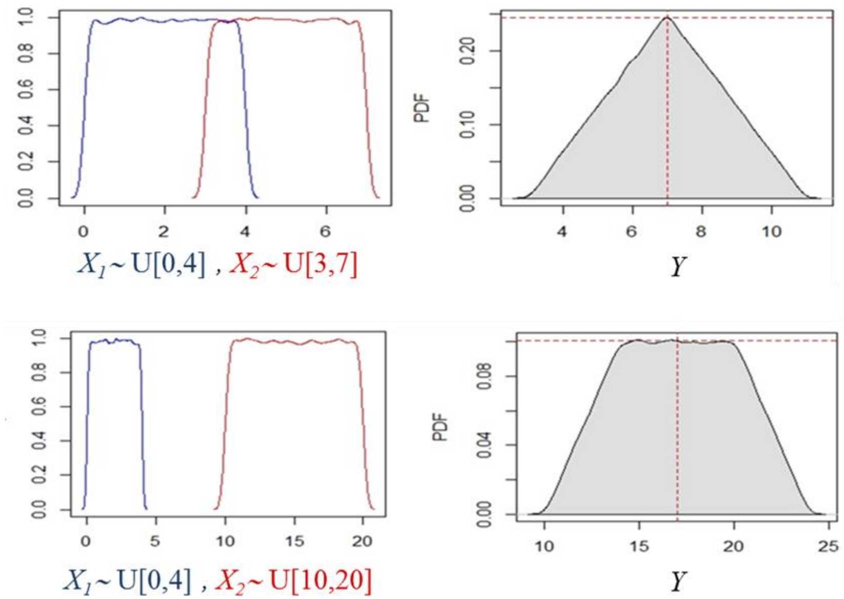
LABORATOIRE DE GÉNIE CHIMIQUE
TOULOUSE • UMR 5503

UNCERTAINTY

A worthier illustrative example

X_1 and X_2 are uniform random variables.
What is the uncertainty ΔY of $Y = X_1 + X_2$?
What is the distribution of Y ?

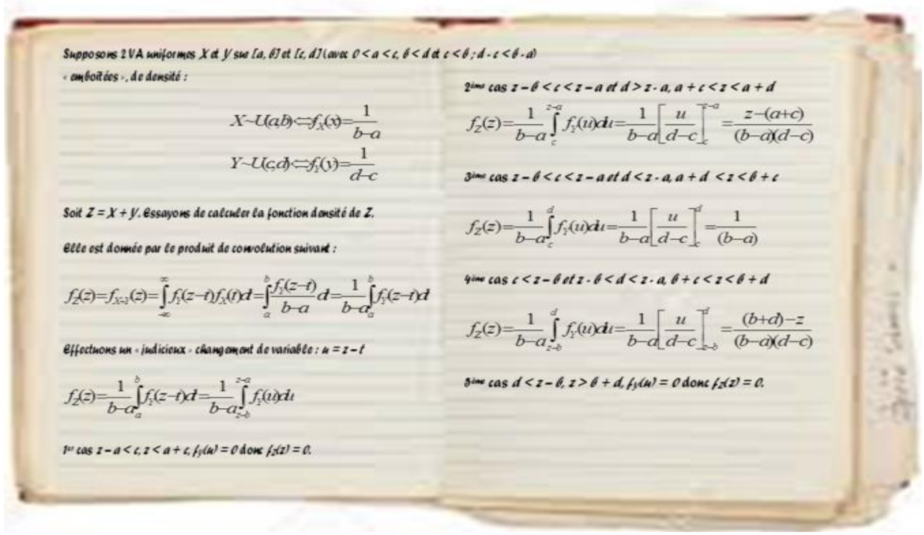
Monte-Carlo solution



Quantiles, uncertainties and distribution function parameters can be estimated numerically from a **Monte-Carlo sample**. Here, trapezoid[10,13,22,24].

http://lgc.cnrs.fr

Paper-and-pencil solution



- Let us now consider the situation where the r.v. have different distributions? What would the uncertainty ΔY be in this case ? The distribution of Y ?

$$\text{model: } Y = 48\frac{X_1 X_2 X_3}{X_4 X_5}$$

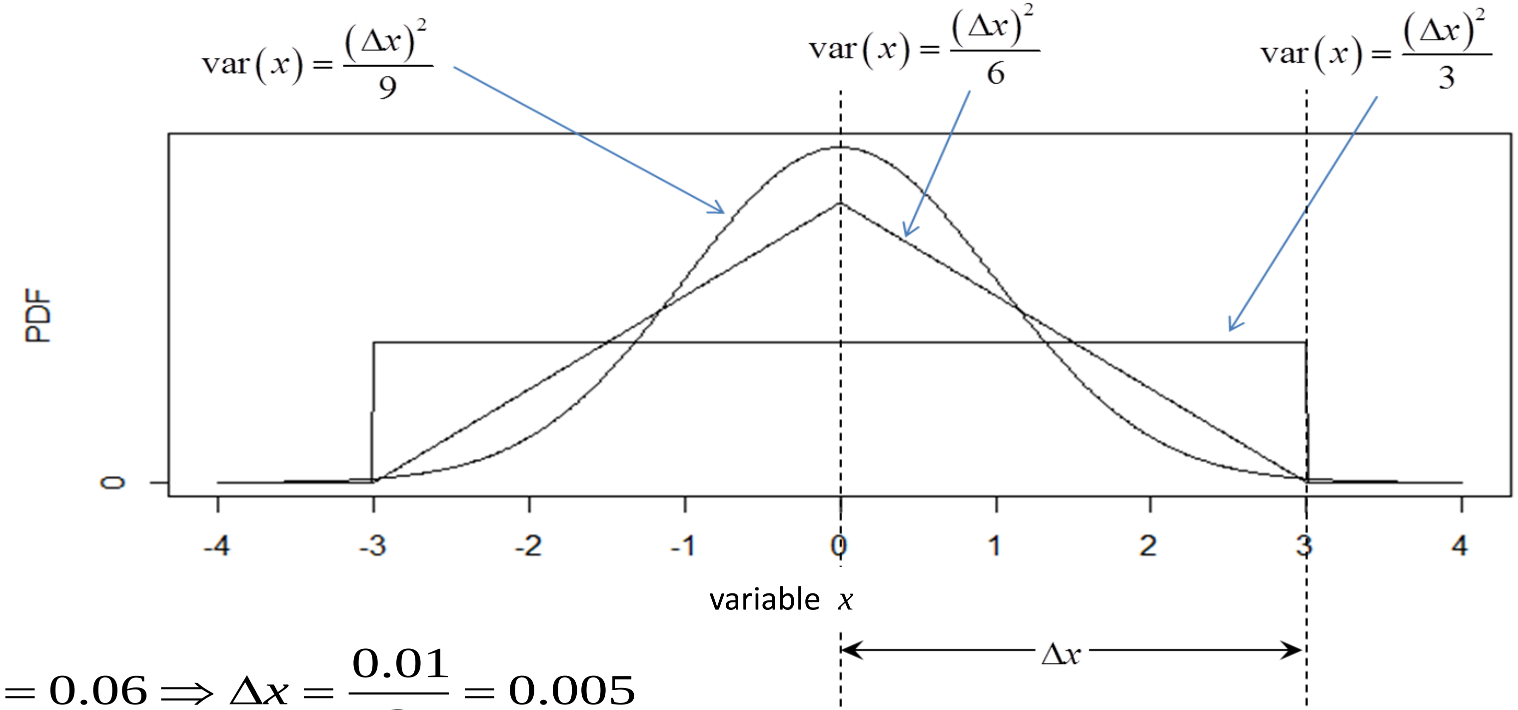
Which pathway would you rather walk ?

LABORATOIRE DE GÉNIE CHIMIQUE TOULOUSE • UMR 5503

UNCERTAINTY

Standards

***Ref.* : Guide for Uncertainty Measurement (norme ISO GUM)** / Guide pour l'expression de l'incertitude de mesure - Annex F: Practical guidance on evaluating uncertainty components. (2008 versions: EN / FR)



- Significant figures : $x = 0.06 \Rightarrow \Delta x = \frac{0.01}{2} = 0.005$

- A prudent solution when one does not know the statistical distribution from which an observation is drawn is to use the triangular isosceles distribution.

UNCERTAINTY

Generation of correlated random variables

Model parameters are always correlated !

- Their correlation comes from the model equation itself.

  Example: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$

- Model parameters are correlated through a **variance-covariance matrix**

- Taking the correlation of model parameters into account, with their variance-covariance matrix, is critical when using a model in a **sensitivity analysis**.

  - Example shows the output of a **Monte-Carlo simulation** (sample size = 10,000) for 3 normally distributed r.v., with and without correlation.
    $X_1 \sim N\left(0,\sqrt{3}\right); X_2 \sim N\left(5,\sqrt{2}\right); X_3 \sim N\left(10,1\right)$

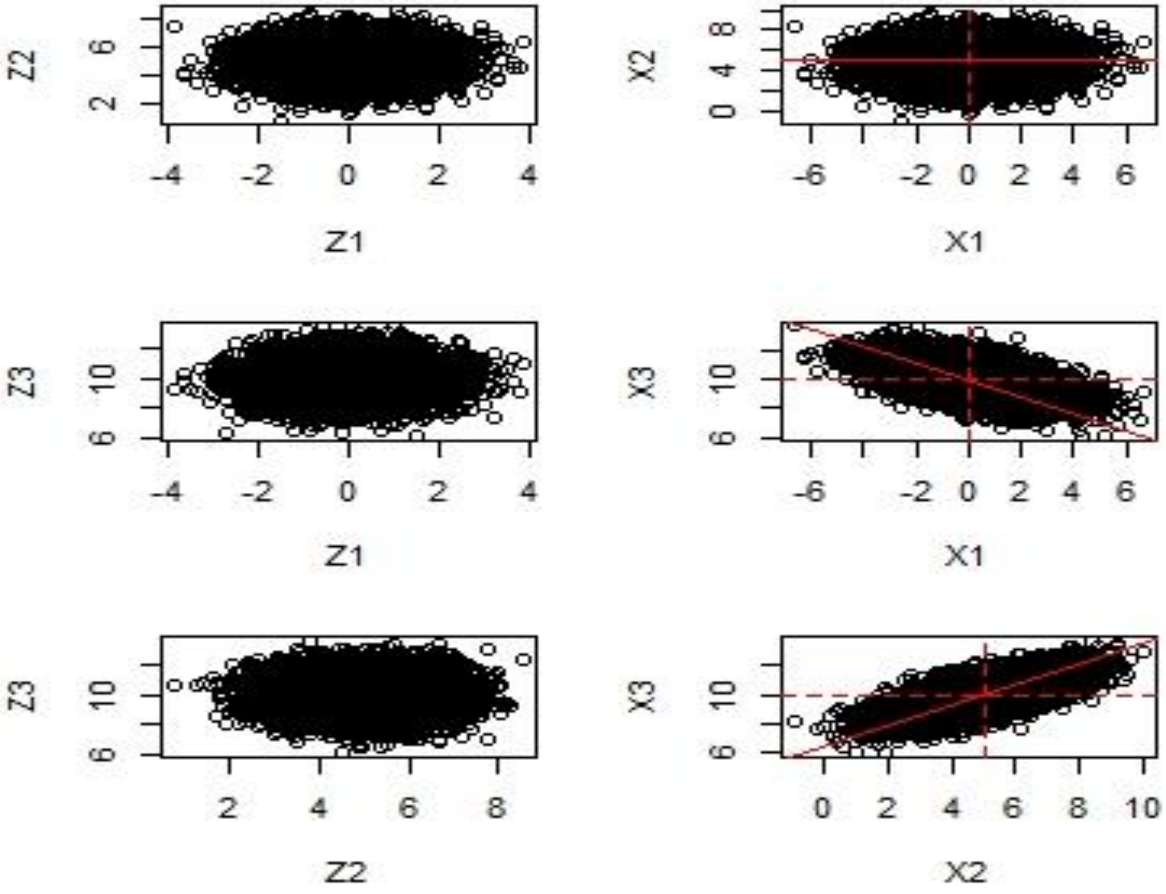
  - This can be done with any type of r.v.

Uncorrelated r.v.

$$\Sigma = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

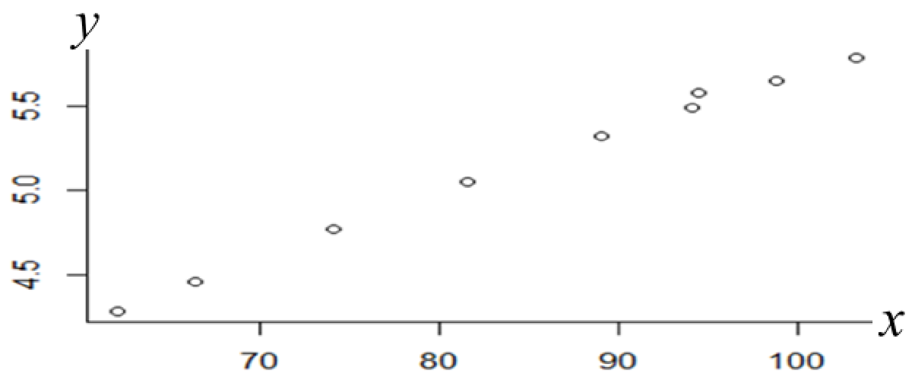
Correlated r.v.

$$\Sigma = \begin{pmatrix} 3 & 0 & -1 \\ 0 & 2 & 1 \\ -1 & 1 & 1 \end{pmatrix}$$



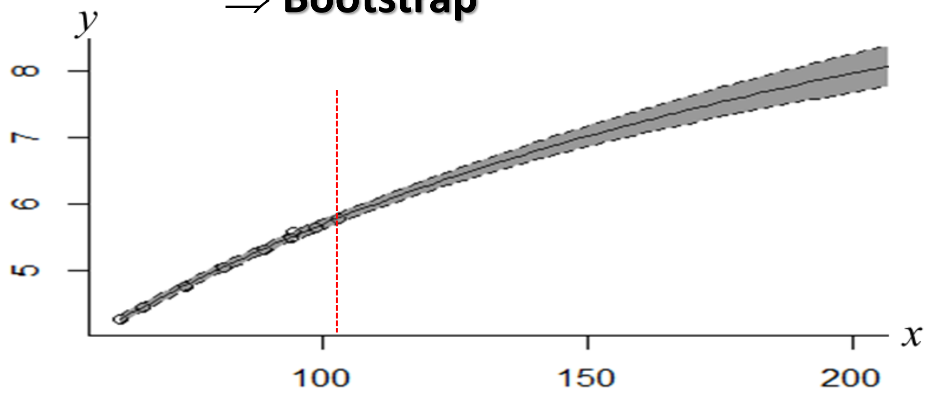
http://lgc.cnrs.fr

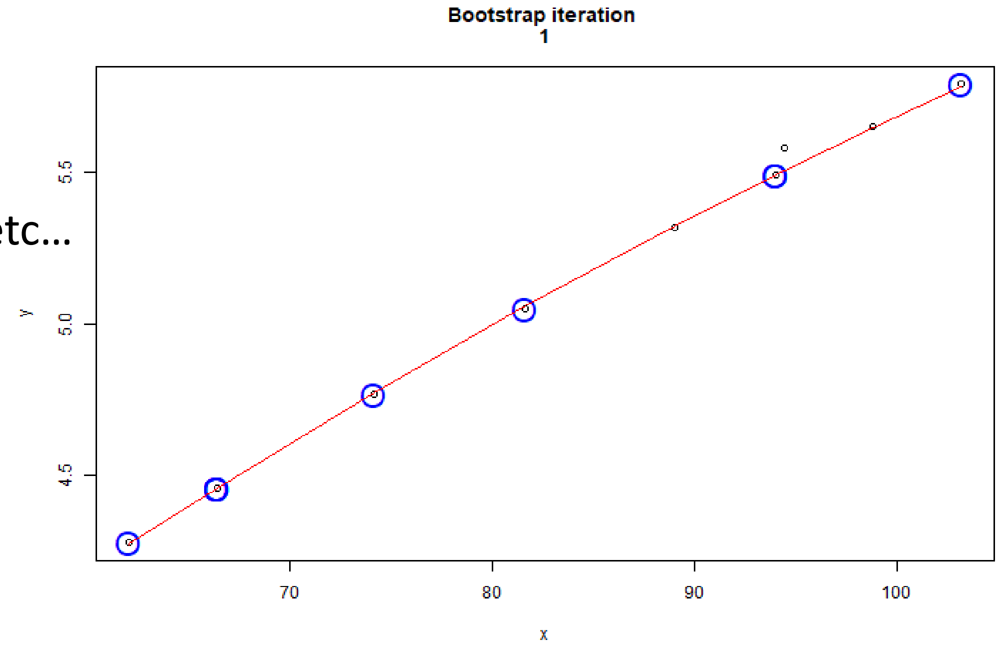
UNCERTAINTY



Modelling

- **One set** of observations (x,y)
- Consider the simplest of cases, i.e. independent identically distributed (i.i.d.) Normal r.v. $\Rightarrow$ estimation of model parameters (expected values and uncertainties) and model confidence/prediction intervals by least-squares estimation.
  - Linear models with errors in Y only
    $\Rightarrow$ analytical solutions exist
  - Linear models with errors in X and Y, non-linear models
    $\Rightarrow$ numerical solutions only
    $\Rightarrow$ Levenberg-Marquardt, Simulated annealing, etc…
    $\Rightarrow$ **Bootstrap**





- Broad applicability of the **bootstrap method** to estimation problems in engineering systems' analysis

http://lgc.cnrs.fr

UNCERTAINTY

Wrap-up

- Uncertainty is a fact of life, and uncertainties are inevitable in science and engineering.
- Computational statistics brings analysis of real data at your fingertips, **with uncertainties !**
- **Today's keywords :** Uncertainty; Random variable; Probability distribution function; Monte-Carlo method; Sensitivity analysis; Bootstrap; Variance-covariance matrix; Computational statistics; **R**

Last words

- **Quality data are priceless !**
- With data analysis, and particularly with computational statistics, take nothing for granted ! Always question hypotheses and test them.
- Outliers are your worst foes !
    - Detection of outliers / Sampling, data collection and manipulation
- Small number of observations (sample size)

UNCERTAINTY

Master Class 2020 – LGC

Florent Bourgeois, Prof.

https://lgc.cnrs.fr/annuaire/florent-bourgeois/