Statistics of Measurement

- ***Measurement*** ***Mesure***
- ***uncertainty*** ***incertitude***
- ***Calibration*** ***Etalonnage***

Philippe Breuil, 2009

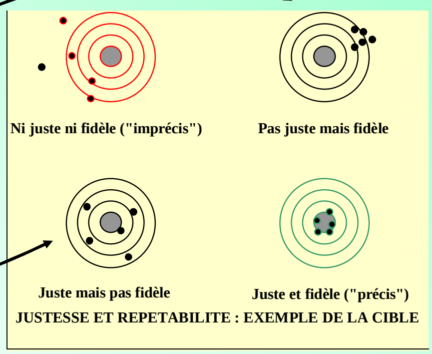
Error & Uncertainty…

Measurement x of a variable which real value is x_R:

Error: variable $\varepsilon = x_R - x$

$$x = x_R + \varepsilon_B + \varepsilon_A$$

- Erreur systématic error (bias): ε_B
  Erreur sytématique
- Random or accidental error: ε_A
  Erreur aléatoire



Random or systematic error?

Erreur systématique ou aléatoire?

- Random error is unpredictable, and its average value is null
- The systematic or unpredictable character of the error can depend on the context …

3

Random or systematic error?

- Ex: Mass measurement with a balance:

Error: $\varepsilon = \varepsilon_a + \varepsilon_T + \varepsilon_o + \varepsilon_u + \varepsilon_s$

ε_a: not explained random error
ε_T: Drift with température
ε_o: « operator » error
ε_u: Error appropriate for the device
ε_s: Systematic error of the serial

Conditions of experiments	ε_a	ε_T	ε_o	ε_u	ε_s
1 opérator, 1 day, 1 device					
Idem + several days					
Idem + several operators					
Idem + tests on different devices					

Systematic
random

4

Caracterisation of the measurement *(and of random error):*

- Estimated mean of one real value x_R

Moyenne estimée

$$\bar{x} = \frac{1}{n}\sum_1^n x_i$$

$$\lim_{n\to\infty}(\bar{x}) = x_R$$

x_R = *real value*

(***Law of large numbers***)
(Loi des grands nombres)

(if random error)

5

Standard deviation of random error
Ecart-type de l'erreur aléatoire:

Standard deviation of random error:

$$\sigma = \sqrt{\left\langle (x_i - x_R)^2 \right\rangle} = \sqrt{\frac{1}{n}\sum_1^n (x_i - x_R)^2}$$

Real value
In priori unknown…

Estimated **standard deviation of the random error**

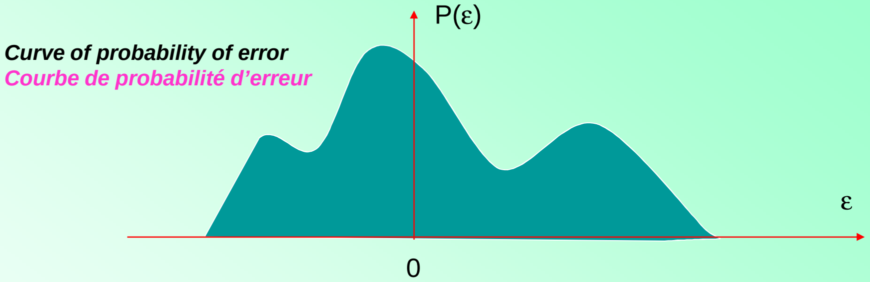
$$s = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$$

Variance *(of random error)* **:** $V = \sigma^2$

Relative standard deviation : $\sigma_r = \frac{\sigma}{|x|}$
(associated with a measurement x)

6

Distributions of errors

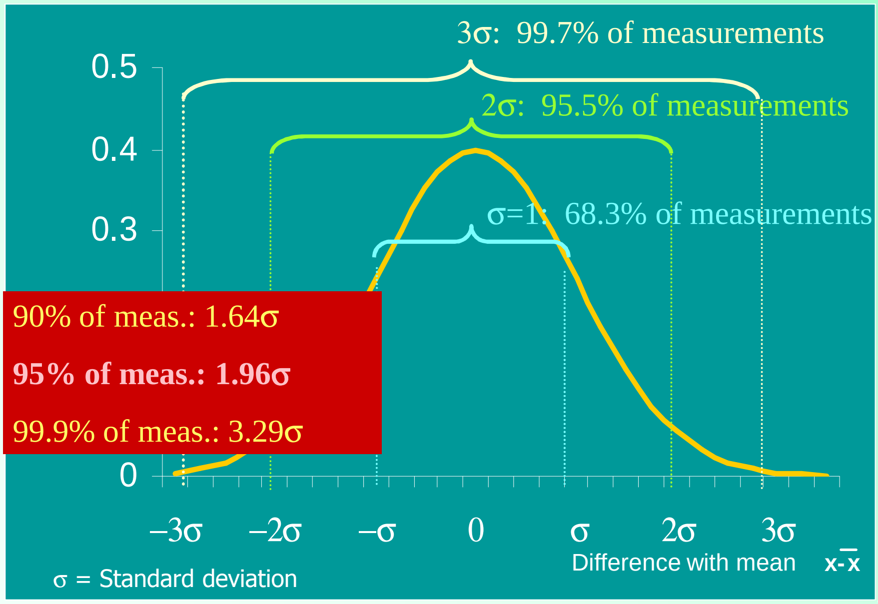


- Gaussian or normal*,
- Uniform,
- Poissonnian,
- Etc…

* The most wide-spread, thanks to Central limit theorem …

Gaussian distribution (or normal)

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}$$



Central Limit Theorem
Théorème central limite

- If a variable is the resultant of a big number of causes, small, in additive effect, this variable tends to a gaussian law.

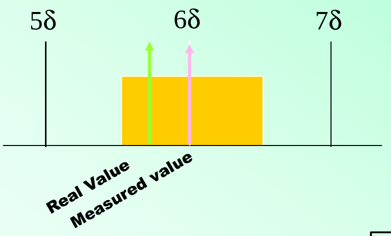
- It is because of this interpretation that the gaussian law is very often used as model (regrettably not always for good reasons).

Demo 9

Uniform distribution of error
Distribution d'erreur uniforme

Equiprobable error on an intervall

Relatively rare, except discretisation error:

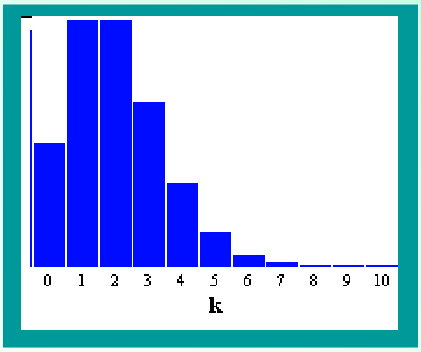


$$\sigma_d = \sqrt{\frac{1}{\delta}\int_{-\frac{\delta}{2}}^{\frac{\delta}{2}} x^2 dx} = \frac{\delta}{2\sqrt{3}} = 0.29\delta$$

10

Poisson distribution

- Ex: counting of nonsimultaneous events (radioactive decay , queue…)



$$P(r) = \frac{\mu^r e - \mu}{r!}$$

Mean: μ
Variance: μ

demo

11

Significance tests

Tests d'hypohèse

Tests of an hypothesis (« null hypothesis », *ex: 2 samples have the same mean*) from a finite number of samples, with random error.
The result of the test is not absolute but is a probability which is an help in order to validate or not the initial hypothesis**, so it is never a proof.**

Parametric Tests *(hypothesis on distribution + or – necessary)*	
1-sample T-test	**Comp. Sample to reference value, confidence interval**
2-samples T-test	**Comparaison of 2 samples**
F-Test	Comparaison of the variance of 2 samples
ANOVA	Analysis of variance: analysis of the variances of K samples, comparison of the means
Chi-Square test	Utilisation in particular for the verification of distribution hypothesis
Grubbs' test	Detection of outliers *(« valeurs aberrantes »)*
non paramétric tests *(no d'hypothesis on the distribution)*	
Test Wilcoxon.M.W	Comparaison of 2 samples, *rank method*

12

one-sample T-test (a)

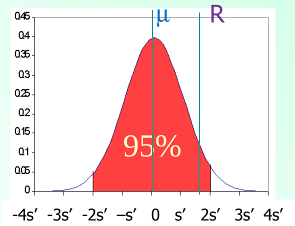
Comparaison of the mean of samples to a reference value R, we suppose that the distribution is normal.*
Hypothesis: « the difference between the mean and a reference value is due only to random errors »

This hypothesis is good if its probability is > 95% (as an example)

n measurements: mean μ, estimated standard deviation of each measurement: s

The estimated standard deviation of the mean is: $$s' = \frac{s}{\sqrt{n}}$$

normal law:



Hypothesis good if:

$$|\mu - R| \le 1.96 \frac{s}{\sqrt{n}}$$

(proba of 95%)

Only if s can be calculated with enough precision (n>25)

13

** Unusefull if n big*

one-sample T-test (b)

Comparaison of the mean of samples to a reference value R, we suppose that the distribution is normal.*
Hypothesis: « the difference between the mean and a reference value is due only to random errors »

$$|\mu - R| \le t.\frac{s}{\sqrt{n}}$$

(t=1.96 for prob. of 95% et n big)

Where t is the Student coefficient

Deg. freedom= n-1

Excel

=LOI.STUDENT.INVERSE(0.05;10)
(french version)
1-proba deg.free.

	Coefs de Student t				
intervalle de conf.	90.0%	95.0%	98.0%	99.0%	99.9%
p	0.1	0.05	0.02	0.01	0.001
deg lib					
1	6.31	12.71	31.82	63.66	636.62
2	2.92	4.30	6.96	9.92	31.60
3	2.35	3.18	4.54	5.84	12.92
4	2.13	2.78	3.75	4.60	8.61
5	2.02	2.57	3.36	4.03	6.87
6	1.94	2.45	3.14	3.71	5.96
7	1.89	2.36	3.00	3.50	5.41
8	1.86	2.31	2.90	3.36	5.04
9	1.83	2.26	2.82	3.25	4.78
10	1.81	2.23	2.76	3.17	4.59
12	1.78	2.18	2.68	3.05	4.32
14	1.76	2.14	2.62	2.98	4.14
17	1.74	2.11	2.57	2.90	3.97
20	1.72	2.09	2.53	2.85	3.85
30	1.70	2.04	2.46	2.75	3.65
40	1.68	2.02	2.42	2.70	3.55
50	1.68	2.01	2.40	2.68	3.50
100	1.66	1.98	2.36	2.63	3.39
100000	1.64	1.96	2.33	2.58	3.29

two-sample T-test

Comparison of 2 means, distribution is supposed normal.*
Hypothesis: « the difference between the 2 means is due only to random errors »

	samp 1	samp 2
Nb meas.	n_1	n_2
Est. Standard dev.	s_1	s_2
Est. mean	μ_1	μ_2

Comparason of μ_1 and μ_2:
1-sample-test
Comparaison of $|\mu_1$-$\mu_2|$ to 0

The standard deviation of $|\mu_1$-$\mu_2|$ is then:

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(«combined standard uncertainty »)

The number of degrees of freedom is approximatively: n_1+n_2-2

(or, for perfectionnists:

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}$$

** Inutile si n_1 et n_2 grands*

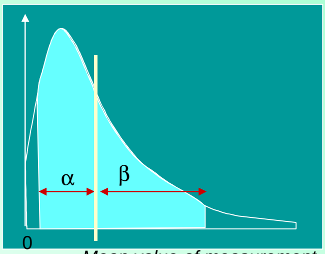
15

Confidence interval
Intervalle de confiance

interval in which measurement has a probability P of being in

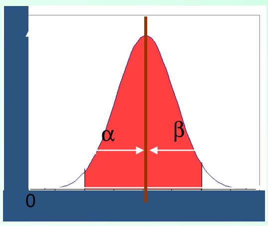
$$[x-\alpha, x+\beta]$$

The calculation of A and B according to the probability P (generally 95%) depends on the law of distribution on the error



Mean value of measurement

gaussian law: very large majority of the cases



•Symetric Distribution :

α=β= uncertainty

•Mean value of measurement = the most probable

16

Writing of a measurement (gaussian distribution)

The result of a measurement must contain **4 éléments :**

Ex : C_{NO} = 125.3 ppb ± 1.7 ppb (95% *or* k=2)

1 2 3 4

1 : **Numerical value with a correct number of decimals**

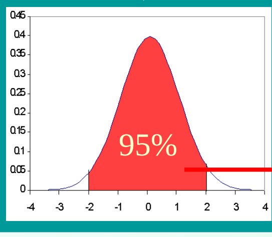
2 : **Unit**

3 : **expanded uncertainty*** = t.σ

4 : **Coverage factor** t**

(from one sample test)

Probability in % so that measurement is out of the interval $[x - ts, x + ts]$



Gaussian law

**Incertitude élargie*

***Coefficient d'élargissement*

	Coefs de Student t				
intervalle de conf.	90.0%	95.0%	98.0%	99.0%	99.9%
p	0.1	0.05	0.02	0.01	0.001
deg lib					
1	6.31	12.71	31.82	63.66	636.62
2	2.92	4.30	6.96	9.92	31.60
3	2.35	3.18	4.54	5.84	12.92
4	2.13	2.78	3.75	4.60	8.61
5	2.02	2.57	3.36	4.03	6.87
6	1.94	2.45	3.14	3.71	5.96
7	1.89	2.36	3.00	3.50	5.41
8	1.86	2.31	2.90	3.36	5.04
9	1.83	2.26	2.82	3.25	4.78
10	1.81	2.23	2.76	3.17	4.59
12	1.78	2.18	2.68	3.05	4.32
14	1.76	2.14	2.62	2.98	4.14
17	1.74	2.11	2.57	2.90	3.97
20	1.72	2.09	2.53	2.85	3.85
30	1.70	2.04	2.46	2.75	3.65
40	1.68	2.02	2.42	2.70	3.55
50	1.68	2.01	2.40	2.68	3.50
100	1.66	1.98	2.36	2.63	3.39
100000	1.64	1.96	2.33	2.58	3.29

Number of points - 1 used for the calculation of σ

17

Evaluation of uncertainty :

- **Evaluation by statistical analysis of series of measurements (« type A »)**
  (generaly measurements, but sometimes simulations)

- **Evaluation by calculation of the effect on final uncertainty of the various sources of uncertainty** (« type B ») **which can be estimated :**

  With a type A method,

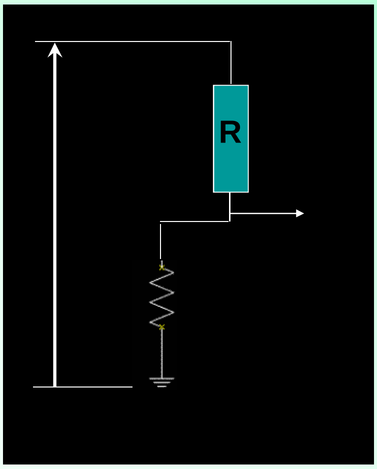
  From manufacturer's specifications etc…

It is then necessary to know the « combined standard uncertainty » (lois de propagation de l'erreur)…

18

Evaluation by statistical analysis of series of measurements

Example: measurement of a resistor:



$$V = E\frac{R_0}{R+R_0} \longrightarrow R = R_0\frac{E-V}{V}$$

Measurement and calculation with a big number N of standard resistors (R known)

Calculation of standard deviation σ

95% uncertainty = t.σ, # 2σ if N>15

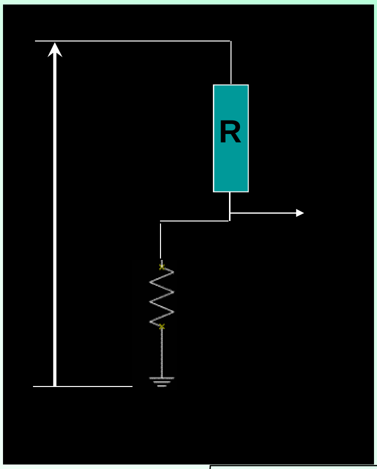
Does not include:

-Uncertainty of Vmeter (V)

-Uncertainties on E, R_0, standard resistors…

Evaluation by calculation of the effect on final uncertainty of the various sources of uncertainty

Example: measurement of a resistor:



$$R = R_0\frac{E - V}{V}$$

σ(R_0), σ(E), σ(V) known

combined standard uncertainty

lois de propagation de l'erreur

$$\sigma(R) = \sqrt{\left(\frac{E-U}{U}\sigma(R_0)\right)^2 + \left(\frac{R_0}{V}\sigma(E)\right)^2 + \left(\frac{R_0E}{V^2}\sigma(V)\right)^2}$$

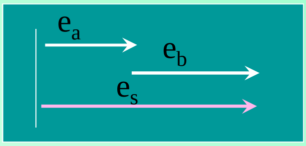
Explanation

combined standard uncertainty for *random & independant errors* (1)

Lois de propagation des erreurs *aléatoires indépendantes*

Example of sum: s=a+b

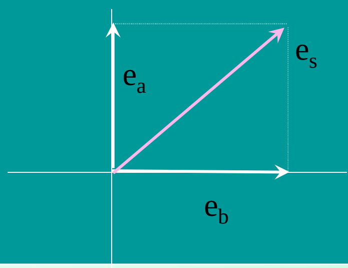
errors are added algebraically : $e_s = e_a + e_b$



And Standard deviations?

***Random error*:**

$$\sigma(s) \leq \sigma(a) + \sigma(b)$$



Standard deviation of random error of the sum of variables

Example of sum: s=a+b (suite)

Variance: $Var(a) = \frac{1}{n-1}\sum(a_i - \bar{a})^2$ $\quad Var(b) = \frac{1}{n-1}\sum(b_i - \bar{b})^2$

$$Var(s) = \frac{1}{n-1}\sum(a_i - \bar{a} + b_i - \bar{b})^2 = \frac{1}{n-1}\left(\sum(a_i - \bar{a})^2 + \sum(b_i - \bar{b})^2 + 2\sum((a_i - \bar{a})(b_i - \bar{b}))\right)$$

$$Var(a+b) = Var(a) + Var(b) + 2Cov(a,b)$$

=0 if independent errors

Variances *add*,

→ ***And standard deviations?***

$$\sigma(a+b) = \sqrt{\sigma(a)^2 + \sigma(b)^2}$$

***quadratic sum** of standard deviations*

22

combined standard uncertainty for *random & independant errors* (2)

Linear Combination: $y = \sum_i a_i x_i$ ➡ $\sigma\ (y)^2 = \sum_i (a_i \sigma(x_i))^2$

Ex: ***sum or difference***:
$y = a + b - c$ ➡ $\sigma(y) = \sqrt{\sigma(a)^2 + \sigma(b)^2 + \sigma(c)^2}$

Fundamental application in instrumentation : The mean ➡
la moyenne

23

combined standard uncertainty for *random & independant errors* (3)

Fundamental application in instrumentation : The mean

$\bar{x} = \frac{1}{n}\sum_1^n x_i$ ➡ $\sigma(\bar{x}) = \frac{1}{n}\sqrt{\sum_1^n \sigma(x)^2} = \frac{\sigma(x)}{\sqrt{n}}$

Law of large numbers

Theorem of the mean:

If a series of measurements has a random and independent error with standard deviation S, then the mean of N of these measurements has a random error which standard deviation is divided by the square root of N.

Théorème de la moyenne:

Si une série de mesures a une erreur aléatoire et indépendante d'écart-type σ, alors la moyenne de n de ces mesures a une erreur aléatoire dont l'écart-type est divisé par racine carrée de n.

Demo 24 Exc

combined standard uncertainty for *random & independant errors* (4)

Products & powers :

Interesting property with relative standard deviations: $\frac{\sigma(x)}{x}$

$$y = A\prod_i x_i^{\alpha_i} \Rightarrow \left(\frac{\sigma(y)}{y}\right)^2 = \sum_i \left(\alpha_i \frac{\sigma(x_i)}{x_i}\right)^2$$

Quadratic sum of ***relative standard deviations***

25

combined standard uncertainty for *random & independant errors (5)*

Unspecified Function:

$$y = f(x_1,..x_i,..)$$

$$\sigma^2(y) = \sum_1^n \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma^2(x_i)$$

General case (non independent variables):

$$\sigma^2(y) = \sum_1^n \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma^2(x_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j} Cov(x_i, x_j)$$

!

(rarely used...)

26

Calculation of uncertainty with *simulation of random error*

Ex: $y=f(x_1,x_2)$,

x_1 et x_2 have independent random errors characterized by their standard deviation s_1 et s_2

- Simulate some experiments with the same mean value x_1 et x_2 + random error

$yi=f(x_1+\varepsilon_1,x_2+\varepsilon_2)$ → ***N times***

- Simulated errors ε_i are calculated using a random number generator, with adequate distribution (generally gaussian*)

- **Calculation of the standard deviation of the y_i**

** Box-Müller Algorithm :* $\varepsilon = \sqrt{-2\text{Ln}(\text{Rnd})} * \cos(2\pi \text{Rnd}')$

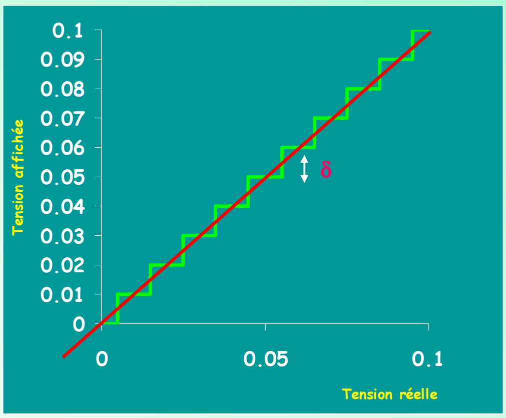
Rnd et Rnd': uniform distributions 0-1 $\bar{\varepsilon} = 0 \quad \sigma\varepsilon = 1$

Ex: débitmètres

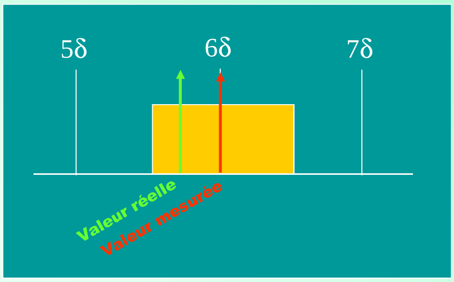
Discretization uncertainty
l'incertitude due à la discrétisation :

- **Discretization: possible values are a discrete (= finite) set.**
- ***Resolution* δ of the system,**
  (≠ *uncertainty*, ≠ *sensitivity*.

- ***Résolution* δ du système,**
  (≠ *incertitude*, ≠ *sensibilité*.



Discretization uncertainty : (2)



maxi error: δ/2

Standard deviation of discretization error :

$$\sigma_d = \frac{\delta}{2\sqrt{3}}$$

Total standard deviation:

$$\sigma(X) = \sqrt{\sigma^2(x) + \sigma_d^2}$$

29

Ecart type de l'erreur due à la discrétisation: (3) significant decimals

Writing of a value with a finite number of digits ➡ **Discretization error**

For a measurement, standard deviation of the « digit » error must be small beside the initial standard deviation

But the numerical systems can give much significant digits...

Ex: measurement = 438.2659872

Standard deviation mesure = 0.55 (0.125 %)

Must we write:

438.2659872 ?

438.265 ?

438 ?

30

Ecart type de l'erreur due à la discrétisation: (4) significant decimals : example

Ex: mesure = 438.2659872 **Standard deviation = σ(x)= 0.55 (0.125 %)**

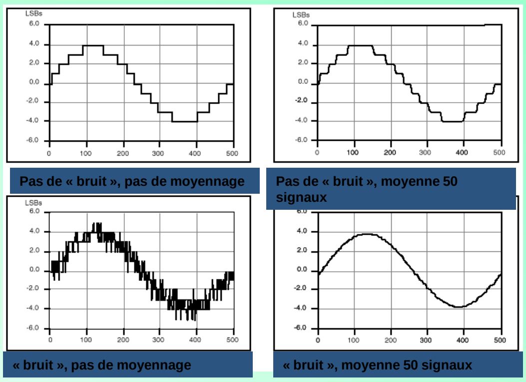
Displayed measurement ***resolution***	S.D. discrétisation error $\sigma_d = \frac{\delta}{2\sqrt{3}}$	S.D. final error $\sigma(X) = \sqrt{\sigma^2(x) + \sigma_d^2}$	%of error due to discretization
438.265987 ***0.000001***	2.9 10-7	0.55	$< 10^{-3}$ %
438.27 ***0.01***	2.9 10-3	0.550008	10^{-3} %
438.3 ***0.1***	0.029	0.5506	0.14 %
438 ***1***	0.29	0.62	*13 %*
440 ***10***	2.89	2.94	*434 %*

•Displayed uncertainty must not have more than 2 significant digits

« Dithering »

demo

The theorem of the mean allows too to reduce discretization (quantification) error:

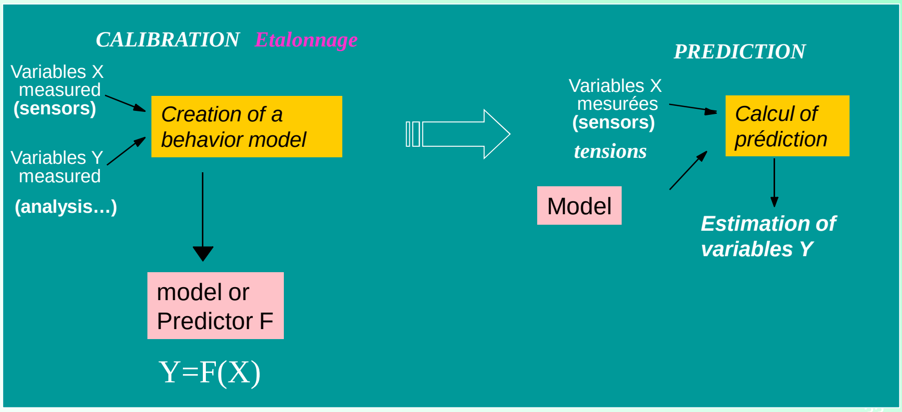


… if the signal before digitalization contains a noise (random and independent) of standard deviation higher than the initial resolution…

Random error allows here to improve « precision » of the measurement ! ! !

32

Calibration: Creation of a *behavior model*
Etalonnage: Création d'un modèle de *comportement*



The least Squares
Les moindres carrés

Least squares method:

Goal: modelization of: Y = F(X),
In fact, from experiments, we have Y = F(X) + E ***where E =***
« residual » or « modelization error »

minimal

$$E = Y - F(X) = \left(y_j^i - f_j(x_1^i, \ldots, x_j^i)\right)$$

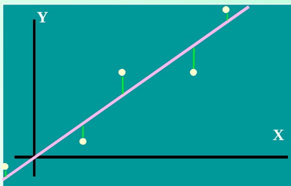
Euclidean distance is generally used:

$$\sum_{i,j} \left(y_j^i - f_j(x_1^i, \ldots, x_k^i)\right)^2$$ **« minimal »**

Linear regression
La régression linéaire

- Monovariable (y=f(x)) + linear hypothesis (y=ax+b)
- 2 cas:

Demo Excel



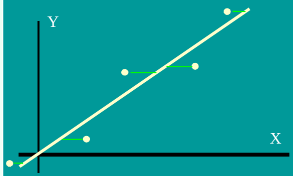
Minimisations des erreurs sur Y

y/x Regression :

$$y = ax + b + \varepsilon_y$$

Slope
pente

Intercept point
ordonnée à l'origine



Minimisation des erreurs sur X

x/y Regression :

$$\hat{x} = a'y + b' + \varepsilon_x \Rightarrow \hat{y} = \frac{1}{a'}x - \frac{b'}{a'}$$

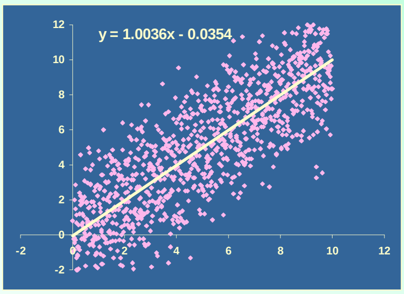
Attention, a y/x régression with errors on x, is source of a systématic error on the slope…

Linear regression: y/x ou x/y?

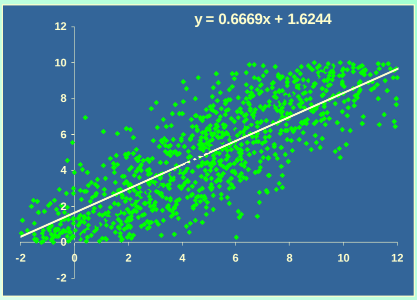
Demo

Attention, a y/x régression with errors on x, is source of a systématic error on the slope…

Y/X Regression



X/Y Regression



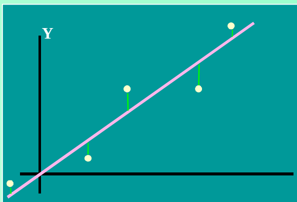
Negligible difference in a majority of cases…

Caractéristics of the y/x régression :

- Estimations of a et b:

$$a = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$$b = \overline{y} - a\overline{x}$$



Minimisations des erreurs sur Y

$$\hat{y} = ax + b$$

(it is possible too to know standard deviation of a and b)

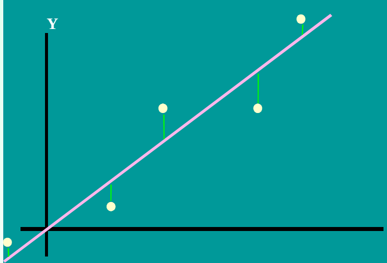
37

Evaluation of the adjustement quality of a linear regression:

Variances:

Var(Y) = Var(aX+b) + Var(ε)

Total information = modelized information + residual information



determination coefficient:

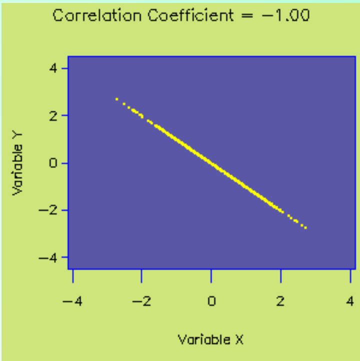
$$R_d = \frac{\text{Var(modelized info.)}}{\text{Var(total info.)}}$$

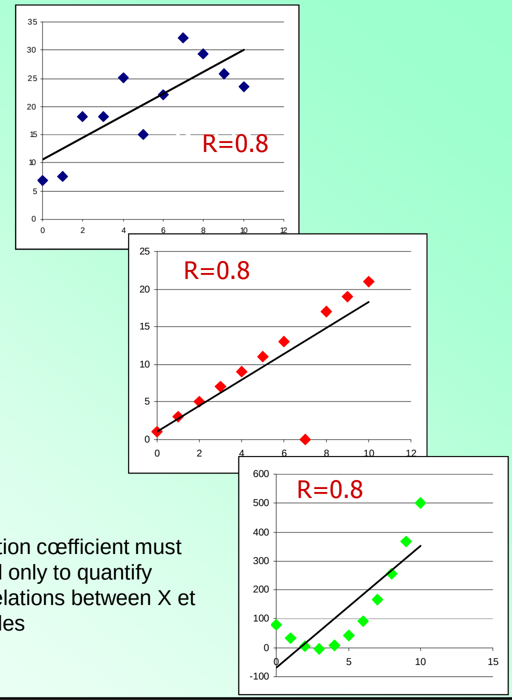
correlation coefficient:

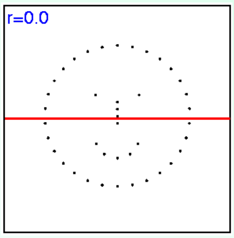
$$R_c = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

$R_d = R_c^2$

Correlation coef., utilisation:

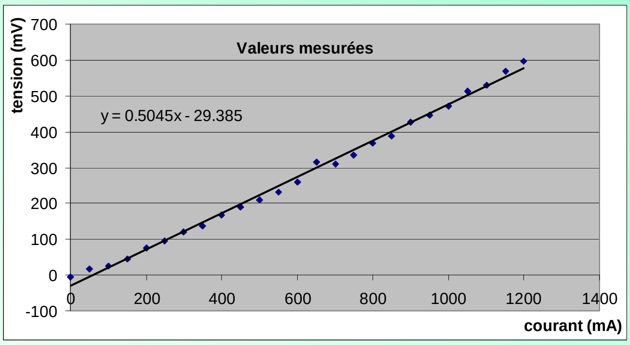


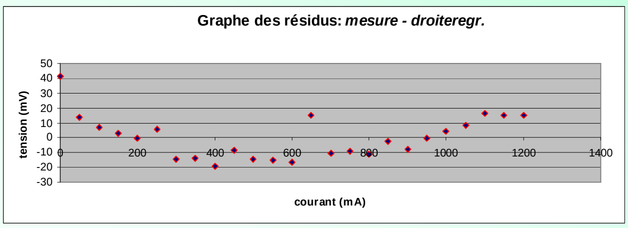




Correlation cœfficient must be used only to quantify linear relations between X et Yvariables

Linear Regression : residuals





Residuals = not modelized information, ideally random error

$$R_i=y_i-(a.x_i+b)$$

Detection of outliers (points aberrants)
To eliminate after checking and with precautions...

Verification of linearity
To correct if required with nonlinear regression

other methods for modelization:

- Multivariables:
  - Linear Multivariable
  - Multivariable Regression, principal components analysis
- Monovariable not linear
- Polynomial, power law, unspecified function

41

multivariate linear Régression :

- n Xvariables, p experiments

Matrix writing, for the p experiments: **$Y = A.X + \varepsilon_y$** To be minimized

$$\begin{cases} y_1 = a_1 x_{1,1} + \ldots + a_n x_{n,1} \\ y_p = a_1 x_{1,p} + \ldots + a_n x_{n,p} \end{cases}$$

➢**p<n: no solution…**

➢**p=n: maybe a solution:** **$A = Y.X^{-1}$** Residual ε_y nul…

(Ex: 2 points to calculate a straight line…)

➢**p>n: X not a square matrix…**

But, there may be a least squares solution!

$$A = Y.X^T.(X.X^T)^{-1}$$

$X^T.(X.X^T)^{-1}$ is called « pseudo-inverse » of X

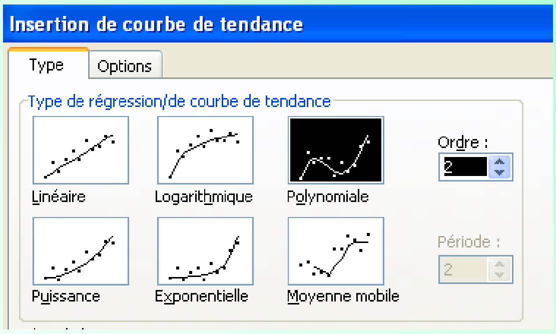
•**Excel function "DROITEREG"**

42

Not linear Regression : 1 "simple« case

- f function may be decomposed in a linear combination of monovariate, ex polynomial:
  - $y=(a_0)+a_1.x+a_2.x^2+\ldots+a_n.x^n$
- Then $x, x^2, \ldots x^n$ may considered as linearly independant and the problem is linear multivariate one.

Ex: courbe de tendance d'Excel:



Demo Excel

43

Non linear Regression : general case

One must then use algorithms of optimization.

Ex: avec le solveur d'Excel

44

Biblio & useful links

Books:

"Statistics for analytical Chemistry" 3 rd ed. J.C. Miller and J.N. Miller John Wiley & Sons, 1998.

"Multivariate Statistical Methods, A Primer" B.F.J. Monley, Chapman & Hall 1986.

"Modélisation et estimation des erreurs de mesure", M Neuilly, CETAMA, Lavoisier, Paris 1993.

WWW:

http://www.emse.fr/%7Epbreuil/capmes/index.htmlCe document + exercices sous Excel

http://physics.nist.gov/cuu/Uncertainty/index.html: excellent site simple et succinct mais de référence sur la calcul des incertitudes, ce document s'en est beaucoup inspiré. Un cours plus complet de statistiques est disponible à : http://www.itl.nist.gov/div898/handbook/

http://rfv.insa-lyon.fr/~jolion/STAT/poly.html : cours complet et touffu de probas & stats.

http://math.uc.edu/~brycw/classes/147/blue/tools.htm : collection de liens (dont beaucoup morts…) , books online et applets.

http://www.helsinki.fi/~jpuranen/links.html#psy: très grande collection de liens « stats » classés.

http://my.execpc.com/~helberg/statistics.html: même chose…

http://www.math-info.univ-paris5.fr/smel/simulations/cadre_simulations.html: cours, articles exemples et surtout "applets" sur les stats

45