Incertitudes des statistiques du lycée au GUM ou que vient f... ici ce $\frac{a}{\sqrt{3}}$?

11 juin 2013

Intention

Éclairer la notion d'incertitude au programme de physique des classes de première année à partir des connaissances des élèves en statistiques et probabilités (et un peu plus).

- le vocabulaire de la métrologie
- rappels de statistiques et probabilités
- retour sur les notions liées à l'incertitude
- un exemple type

Le symbole ↻ signifie :
« demander à l'auditoire plus compétent que moi... »

Intention

Les documents dont je me suis servi

- *Évaluation des données de mesures - Guide pour l'expression de l'incertitude de mesure*
  Document mis en ligne par le BIPM (référence JCGM 100 : 2008) ;
  mots clés pour une recherche sur la toile : **GUM, BIPM**.
- *Programmes de TS en maths*
- *Programmes de CPGE première année, en physique*

Vocabulaire : mesures

mesurande	grandeur particulière soumise à *mesurage* ie : dont on veut *estimer* la valeur
estimer	proposer un couple : (*valeur*, *incertitude*)
valeur vraie	valeur du *mesurande*
mesurage	ensemble d'opérations ayant pour but de d'*estimer* la valeur d'une grandeur
incertitude de...	caractéristique de *dispersion* (notion statistique/probabiliste)

Vocabulaire : erreur

La notion **d'erreur de mesurage** est centrale, même si le but poursuivi dans cet exposé conduira à peu en parler : c'est l'écart entre la valeur du mesurande et la valeur obtenue par mesurage. Comme la valeur du mesurande, dont l'existence est un postulat, n'est jamais connue, l'erreur n'est jamais observable. Cela reste néanmoins une valeur aléatoire soumise à étude statistique.
On distingue deux composantes de l'erreur : une *composante aléatoire* et *une composante systématique* (on dit aussi *erreur aléatoire, erreur systématique*).

Vocabulaire : erreurs

• L'erreur aléatoire provient de variations *multiples* non *prévisibles*. Son *espérance mathématique* (ou sa valeur moyenne) est égale à 0.
• La composante systématique provient d'un *effet reconnu* qui fait que l'on estime que l'erreur moyenne ne sera pas nulle (ce qui constitue un biais) : position de l'observateur par rapport à un instrument analogique, effet connu de l'instrument, variation due à la température, ↻...
• La somme de ces deux composantes de l'erreur est donc une somme de variables aléatoires dont l'espérance est

$$E(e_a + e_s) = E(e_a) + E(e_s) = e_s.$$

D'une façon générale, en statistique, on appelle *biais* l'écart entre la moyenne d'un *estimateur* et la moyenne de la variable aléatoire qu'il est censé estimer.

Vocabulaire : erreurs

Définitions du GUM

Une erreur aléatoire est le résultat d'un mesurage moins la moyenne d'un nombre infini de mesurages du même mesurande, effectué dans des conditions de répétabilité.

Une erreur systématique est la moyenne qui résulterait d'un nombre infini de mesurages d'un même mesurande, effectué dans des conditions de répétabilité, moins une valeur (vraie) du mesurande.

Vocabulaire : incertitudes

L'*incertitude* est une grandeur qui permet d'estimer la *dispersion* des résultats de différents mesurages réalisés dans les mêmes conditions. C'est tout naturellement que la première définition d'une incertitude, *l'incertitude-type,* se confond avec la notion *d'écart-type.*
Le tableau qui va suivre donne les définitions à adopter, sur lesquelles nous reviendrons après quelques rappels de statistiques et probabilités...

incertitude-type	c'est l'écart-type de la valeur du mesurage
... de type A	c'est la composante obtenue par analyse statistique de séries de mesurages *(variance empirique)*
... de type B	c'est la composante obtenue par d'autres moyens (calculs de probabilité, la loi de la *valeur* étant modélisée *a priori* par exemple...)
incertitude composée	incertitude-type de l'estimation d'une grandeur $f(x_1, ..., x_n)$, les $(x_i)_i$ étant soumis à mesurage
incertitude élargie	voir intervalle de confiance.

Voyons maintenant ce que tout cela signifie d'un point de vue statistique...

Statistiques : Écart-type d'une population

• Considérons un ensemble fini de valeurs $x = (x_k)_{1 \leq k \leq N}$. Sa moyenne, sa variance et son écart-type sont définies par

$$\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_k, \quad Var(x) = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})^2 \text{ et } \sigma(x) = [Var(x)]^{1/2}.$$

• Lorsqu'on veut estimer ces paramètres sans pour autant faire les calculs de façon exhaustive (sondage par exemple) on procède en prélevant un *échantillon* de taille $n \leq N$.

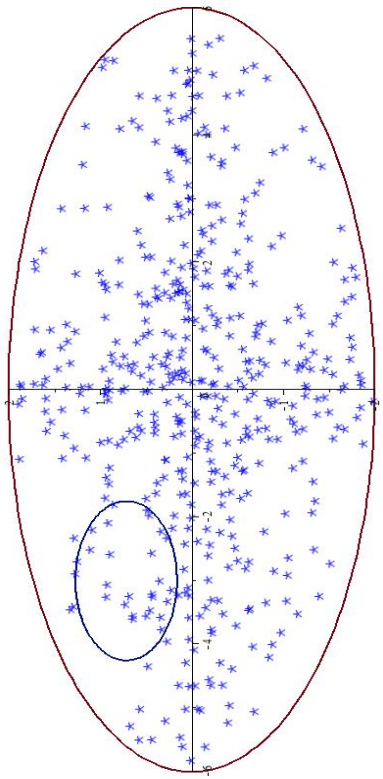
Estimations à partir d'un échantillon

• Un *échantillon* de taille n d'éléments de $(x_k)_k$ est suite n tirages *indépendants* $X_1, ..., X_n$. On distinguera **trois** notions :

• la moyenne, la variance de la population totale ;

• la moyenne empirique, la variance empirique définies pour l'échantillon : ce sont elles-mêmes des variables aléatoires dont les valeurs dépendent du tirage ;

• les estimateurs de la moyenne, de la variance de la population complète calculés à partir d'un échantillon...



Échantillon d'une loi (ou d'une variable aléatoire)

Définition

Soit X une variable aléatoire quelconque. Un échantillon de taille n de X est une suite $(X_1, ..., X_n)$ de variables aléatoires

- *indépendantes*
- *de même loi que X.*

⋆ Estimations à partir d'un échantillon : moyenne et variance empiriques, *estimateur sans biais de la variance*...

$E(X_i) = \bar{x}$	$Var(X_i) = Var(x)$
$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$
$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\hat{\sigma}_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$

Lois de probabilité, variables aléatoires

Lois discrètes	Lois continues
$P(X = \alpha_i) = p_i$	$P(X \in]\alpha, \beta]) = \int_{\alpha}^{\beta} f(t)\,dt$
$\sum_i p_i = 1$	$\int_{-\infty}^{\infty} f(t)\,dt = 1$
$\bar{X} = E(X) = \sum_i \alpha_i P(X = \alpha_i)$	$\bar{X} = E(X) = \int_{-\infty}^{\infty} t\, f(t)\,dt$
$Var(X) = E\left((X - \bar{X})^2\right)$	$Var(X) = E\left((X - \bar{X})^2\right)$
$Var(X) = \sum_i p_i(\alpha_i - \bar{X})^2$	$Var(X) = \int_{-\infty}^{\infty} f(t)(t - \bar{X})^2\,dt$

Propriétés

- On a toujours :

$$E(X+Y) = E(X) + E(Y) \text{ et } Var(X) = E(X^2) - E^2(X)$$

- Formule de transfert

$$E(\phi \circ X) = \int_{-\infty}^{\infty} \phi(t)\, f(t)\, dt$$

- X et Y sont *non-corrélées* lorsque $E(X\,Y) = E(X)\,E(Y)$ et

$$E(X\,Y) = E(X)\,E(Y) \Rightarrow Var(X+Y) = Var(X) + Var(Y).$$

Propriétés

- Deux événements A et B sont *indépendants* lorsque

$$P(A \cap B) = P(A) \times P(B).$$

- Deux variables aléatoires X et Y sont *indépendantes* lorsque, pour tout couple d'intervalles (I, J) de $\mathbb{R}$, les événements $P(X \in I)$ et $P(Y \in J)$ sont indépendants.

Théorème
Si X et Y sont deux variables aléatoires indépendantes,

$$E(X\,Y) = E(X)\,E(Y) \text{ et } Var(X + Y) = Var(X) + Var(Y)$$

Loi faible des grands nombres

Théorème

Soit $(X_i)_i$ une suite de variables aléatoires indépendantes suivant la loi de X. On suppose que X admet une variance finie, alors la suite des moyennes empiriques converge en probabilité vers $E(X) = m$ et

$$P(|\hat{\mu}_n - m| \geq \varepsilon) \leq \frac{Var(X)}{n \cdot \varepsilon^2} \tag{1}$$

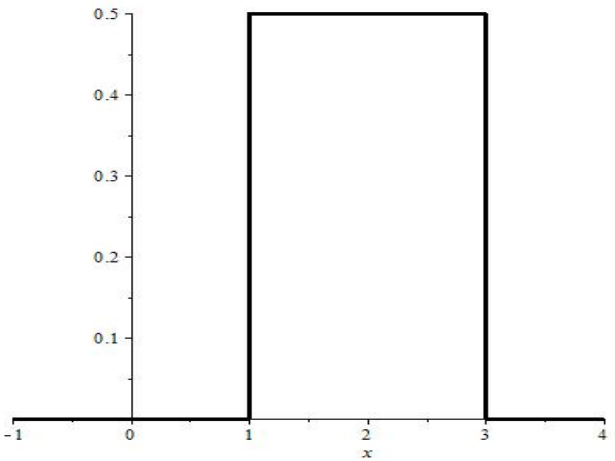
A rapprocher de la définition de l'erreur aléatoire ou de l'erreur systématique dans le GUM.

loi uniforme d'amplitude $2a$

• La *loi uniforme* sur un intervalle $[a_1, a_2]$ a pour densité la fonction définie par

$$f(t) = \begin{cases} \frac{1}{a_2 - a_1} & \text{si } a_1 \leq t \leq a_2 \\ 0 & \text{sinon.} \end{cases}$$

$$\boxed{u = \frac{a}{\sqrt{3}}}$$



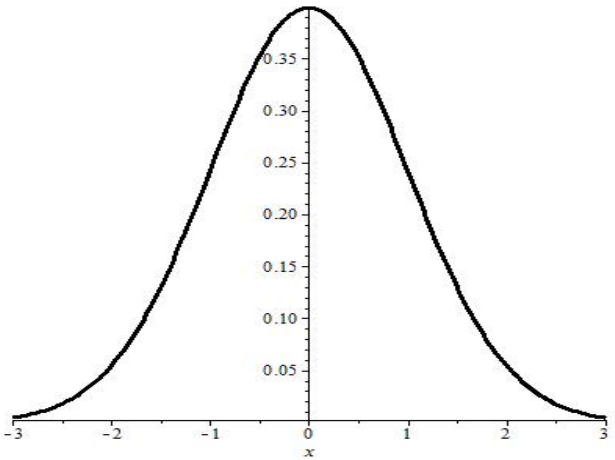
• C'est cette loi que la norme impose quand on peut seulement estimer les limites d'une grandeur (a est la *précision*)

$$\mu = \int_{a_1}^{a_2} t f(t)\, dt = \frac{a_1 + a_2}{2}; \sigma^2 = \int_{a_1}^{a_2} (t - \mu)^2 f(t)\, dt = \frac{(a_2 - a_1)^2}{12} = \frac{a^2}{3}$$

loi normale $\mathcal{N}(\mu, \sigma)$

• La loi de *Laplace Gauss*, de moyenne μ et d'écart-type σ a pour densité

$$f(t) = \frac{1}{\sqrt{2\,\pi\,\sigma}}\, e^{-1/2\,\frac{(x-\mu)^2}{\sigma^2}}$$



• Son importance vient du *théorème central limite* dont les élèves ont une version particulière avec le théorème de *Moivre-Laplace* en terminale... C'est le modèle de l'accumulation d'erreurs de mesurage indépendantes.

⋆ Théorème central limite

Théorème

On considère un échantillon $(X_1, ..., X_n)$ d'une loi de probabilité d'espérance 0 et d'écart-type $\sigma > 0$ (ie : les X_i suivent la même loi et sont indépendantes). Alors, la variable aléatoire

$$\zeta_n^* = \frac{S_n}{\sigma\sqrt{n}} = \frac{X_1 + X_2 + ... + X_n}{\sigma\sqrt{n}}$$

converge en loi vers la loi normale centrée réduite :

$$\lim_{n\to+\infty} P(\zeta_n^* \in [a, b]) = \frac{1}{\sigma\sqrt{2\,\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}\, dt$$

Théorème de Moivre Laplace (énoncé en Terminale)

- On dit que x suit une loi de *Bernoulli* de paramètre p si $P(X = 0) = p$ et $P(X = 1) = 1 - p$.
- La loi *Binomiale* $\mathcal{B}(n, p)$ est la loi d'une *somme* $X_n = \sum_{k=1}^{n} x_i$ de n variables aléatoires *indépendantes* qui suivent chacune une loi de Bernoulli de paramètre p : $P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $E(X_n) = p$, $Var(X_n) = n = (1 - p)$.
- Dans le cas de cette somme le théorème central limite s'exprime :

$$\lim_{n \to +\infty} P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \in [a, b] \right) = \frac{1}{\sigma\sqrt{2\,\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dt$$

Intervalles de fluctuation

- Comme les élèves apprennent en terminale que si $\alpha \in]0,1[$, il existe $u_{\alpha>0}$ tel que $P(X \in [-u_\alpha, , u_\alpha]) = 1-\alpha$.
- ils savent déterminer un intervalle I_n tel que

$$\lim_{n\to+\infty} \left(\frac{X_n}{n} \in I_n \right) = 1-\alpha ...$$

- Exemple : comme $u_{0,05} \approx 1,96$,

$$\frac{X_n}{n} \in \left[p - 1.96\sqrt{\frac{p(1-p}{n}}, p + 1.96\sqrt{\frac{p(1-p}{n}} \right]$$

avec une probabilité de 0.95. C'est ce que l'on appelle un intervalle de confiance à 95%

Intervalles de confiance et incertitudes élargies

• Une *incertitude élargie* est obtenue en multipliant l'écart-type ou l'incertitude type d'un facteur k de telle sorte que $P(x \in [X - k\,\sigma, x + k\,\sigma]) = 1 - \alpha$ ($1 - \alpha$ est le *niveau de confiance*).

• Au programme de terminale : connaître les valeurs de $P(X \in [\mu - k\sigma, \mu + k\sigma])$ pour $k = 1, 2, 3...$

Détente

Exercice : se saisir de la planche d'exercices

Retour sur les incertitudes de type A, de type B

- Méthode statistiques d'estimation : ce sont elles, avec la loi des grands nombres, qui nous donnent les outils pour une estimation de type A. ↻
- Le GUM conseille de prendre lorsque seule la *précision* a d'un appareil est connue (ce qui est noté $\pm a$), pour incertitude $u = \frac{a}{\sqrt{3}}$.
- Une analyse des sources d'erreurs peut conduire à accepter de modéliser par une loi normale dont les paramètres seront estimés par échantillonnage... ↻

Ce sont là des incertitudes de type B...

Incertitude composée

• On appelle *incertitude composée*, une incertitude sur la valeur d'un mesurande de la forme $y = f(x_1, ..., x_n)$ lorsque l'estimation de y est donnée par $Y = f(X_1, ..., X_n)$ où les $(X_i)_i$ sont des valeurs observées par mesurage des $(x_1, ..., x_n)$.

Incertitude composée par approximation linéaire ou par simulation

• Lorsque f est une fonction de classe $\mathcal{C}^2$ et les $(X_i)_i$ **ne sont pas corrélées,** le GUM conseille de calculer l'incertitude sur le mesurage de y avec

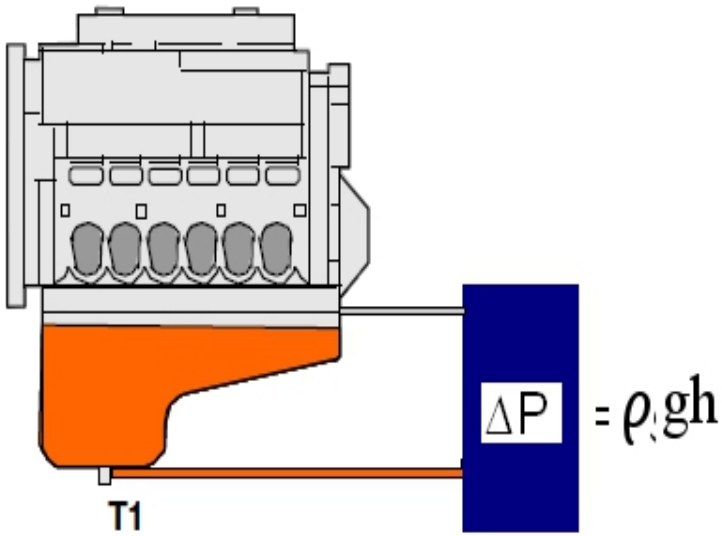
$$u_c^2(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) \tag{2}$$

• Pour estimer $Y = f(X_1, ..., X_n)$ lorsque des mesures des X_i sont données avec des incertitude de type B, on peut aussi procéder par *simulation* pour estimer y.

Un exemple de calcul d'incertitude composée

• Pour déterminer les pertes d'huile d'un moteur de bateau, on pratique des mesures régulières dont on souhaite évaluer l'incertitude.

• Un capteur de pression relève la hauteur d'huile au dessus d'un certain point. Le carter est cartographié et les volumes en fonction de cette hauteur sont donc connus.
L'expression de la variation est une fonction $f\left(\rho_0, \rho, P, P1, P2, M1, M2\right)$



$$M = \frac{\left(-\rho_0 P + \rho_0 \alpha \rho g + P1\, \rho\right)\left(M2 - M1\right)}{\rho_0 \left(-P2 + P1\right)} + M1 \qquad (3)$$

Deux solutions proposées

- Calcul d'incertitude composée par la formule

$$u_c^2(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) \tag{4}$$

avec un logiciel de calcul formel pour obtenir une expression des dérivées partielles

- Simulation stochastique de la variable aléatoire $f(X_1, ..., X_n)$, les X_i étant simulées à partir des hypothèses sur leurs lois de probabilités.

Simulation : programme MAPLE

```
S1 := 0 ; S2 := 0 ;
for k from 1 to N do
        construction de la liste des arguments simulés
        Xnum := NULL ;
        for e in [X] do
                Xnum := Xnum,Sim[e]() ;
        od ;
        calcul de F(rho0,rho,P,P1,P2,M1,M2)
        M :=evalf(F(Xnum)) ;
        moments
        S1 :=S1+M ;
        S2 :=S2+M**2 ;
od :
Moy := evalf(S1/N) ;
Var := evalf((S2-S1**2/N)/(N-1)) ;
```