

# ANALYSE STATISTIQUE DES INCERTITUDES DE MESURE

## LMAFY1101

Anouar El Ghouch

LSBA, UCLouvain

# PLAN

VOCABULAIRE ET NOTATIONS

MODÉLISATION STATISTIQUE DES MESURES

LES INTERVALLES DE CONFIANCE

LA MÉTHODE BOOTSTRAP

PROPAGATION DES ERREURS

# VOCABULAIRE ET NOTATIONS

- Idéalement, lors d'un mesurage (processus de mesure), on cherche à trouver **la vraie valeur** d'une grandeur, càd. la valeur que l'on obtiendrait si le mesurage était parfait.
- L'expérience montre qu'aucune mesure, aussi soigneuse soit-elle, n'est totalement exemptée d'incertitudes et/ou d'erreurs même minimales.
- **Un mesurage n'étant jamais parfait**, la vraie valeur d'une grandeur est toujours inconnue.

Ainsi, lorsqu'on mesure la même grandeur plusieurs fois, il n'est pas rare d'observer **des variations dans les résultats**. Ces variations peuvent être attribuées à de multiples causes. Par exemple,

- Les conditions expérimentales : température, pression, humidité, luminosité, etc.,
- Le ou les opérateur(s): qualification, expérience, fatigue, temps de réaction, etc.,
- Le ou les instrument(s) de mesure: précision, sensibilité, réglage, etc..

Ces variations incontrôlables sont observées même dans des conditions dites de répétabilité (même procédure opératoire, mêmes opérateurs, même système de mesure, mêmes conditions de fonctionnement, même laboratoire, ...).

Quoique c'est discutable, on utilisera par la suite le terme "**erreurs de mesures**" ou "**incertitudes de mesure**" pour qualifier de telles variations.

Les erreurs de mesure appartiennent à deux catégories principales:

- Les **erreurs systématiques (ou biais)** qui affectent toutes les mesures toujours de la même manière (sur ou sous évaluation):
  - appareil de mesure non conforme, mauvais réglage, graduation erronée...

Dans un système de mesure correctement conçu, les erreurs systématiques sont négligeables.

- Les **erreurs (ou variations) aléatoires** incontrôlables et parfois inexplicables, se produisent en sens divers et affectent chaque mesure de manière différente:
  - positionnement toujours différent de l'œil de l'observateur par rapport aux graduations d'un appareil, variations des réflexes de l'observateur qui manipule un chronomètre...

Les erreurs aléatoires sont pratiquement toujours présentes.

Pour les scientifiques, mesurer une grandeur n'est pas simplement proposer une valeur (approximative) mais aussi lui associer une incertitude qui qualifie la qualité de la mesure.

En général, tout résultat d'un mesurage (sérieux) doit s'écrire sous la forme suivante:

$$x_m \pm \delta x, \text{ où}$$

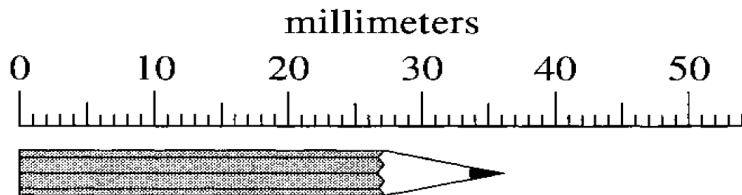
- $x_m$  est un chiffre qui représente notre meilleure estimation
- $\delta x$  est un chiffre qui indique l'incertitude associée à cette mesure

En générale, cela signifie que la vraie valeur de la grandeur se situe vraisemblablement dans l'intervalle

$$[x_m - \delta x, x_m + \delta x]$$

Dans certaines situations simples, des évaluations raisonnables de l'incertitude ne nécessitent qu'un peu de bon sens.

## EXEMPLE 1



Mesure d'une longueur à l'aide d'une règle.

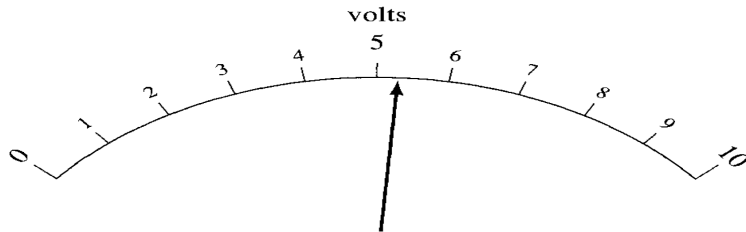
Nous pourrions raisonnablement opter pour une longueur ( $l$ ) plus proche de 36 millimètres que de 35 ou 37 millimètres, aucune lecture plus précise n'étant possible, donnant ainsi en conclusion,

(mesure de  $l$ ) :  $36 \pm 0.5$  mm

- Meilleure estimation de longueur = 36 mm.
- Incertitude = 0.5 mm.
- Intervalle vraisemblable =  $[35.5, 36.5]$  mm.



## EXEMPLE 2



Lecture sur un voltmètre.

De même, pour le voltmètre, une conclusion raisonnable pour la tension ( $t$ ) lue peut être :

(mesure de  $t$ ) :  $5.3 \pm 0.1$  V

- Meilleure estimation de tension = 5.3 V.
- Incertitude = 0.1 V.
- Intervalle vraisemblable = [5.2, 5.4] V.

# INCERTITUDES ET MESURES RÉITÉRÉES

Dans ces deux derniers exemples, le problème est facile, car il se limite à la simple lecture d'un point situé sur une échelle graduée.

Mais, en pratique, beaucoup de mesures font intervenir des incertitudes bien plus difficiles à évaluer.

Par exemple, lors de la mesure d'un intervalle de temps à l'aide d'un chronomètre, la principale source d'incertitude ne réside pas dans la difficulté de lecture du cadran mais provient de notre propre temps de réaction, inconnu, lors du lancement et de l'arrêt du chronomètre.

**Une des meilleures manières d'évaluer l'incertitude associée à une mesure est de la répéter afin d'en comparer les valeurs.**

**EXEMPLE :** À l'aide d'un chronomètre, on effectue une série de 10 mesures de la période d'un pendule. Voici les résultats (en secondes) :

2.3, 2.4, 2.5, 2.4, 2.2, 2.8, 2.3, 2.7, 2.6, 2.5

Il semble raisonnable de choisir la moyenne  $x_m = \bar{x} = 2.47\text{s}$  comme meilleure estimation de la vraie valeur de période.

**Mais notre objectif n'est pas seulement de donner une estimation ponctuelle (ici 2.47s) mais d'accompagner cette estimation par une mesure d'incertitude pour pouvoir, à la fin, écrire le résultat sous la forme  $x_m \pm \delta x$ .**

**MÉTHODE :** Approche statistique basée principalement sur la modélisation de ce phénomène aléatoire et le calcul des variances.

# MODÉLISATION STATISTIQUE DES MESURES

Le point de vue moderne consiste à voir le résultat d'**une mesure comme une variable aléatoire**. Plus précisément, toute mesure peut être considérée comme la somme de la vraie valeur plus une erreur:

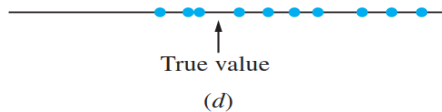
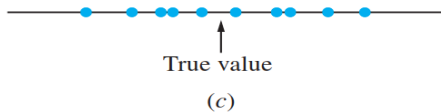
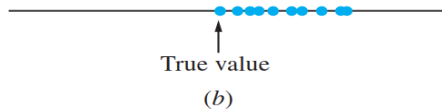
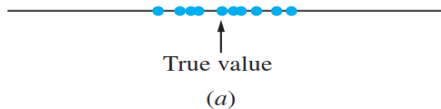
$$\text{valeur mesurée} = \text{vraie valeur} + \text{erreur}$$

Sous forme d'équation, cela donne

$$X = \mu + \epsilon$$

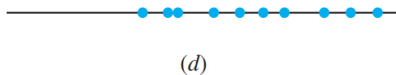
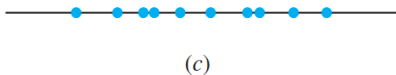
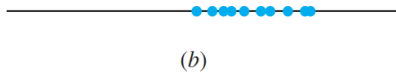
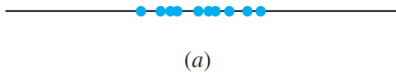
où, l'erreur,  $\epsilon$  est une variable aléatoire:

- $E(\epsilon) = E(X) - \mu = \text{biais}$  d'estimation. Le biais est une mesure de la **non-justesse** du mesurage: plus le biais est faible, plus la méthode est juste.  $\text{biais} = 0 \rightarrow$  mesures répétées centrées autour de la vraie valeur  $\mu$ .  $\text{biais} > (<) 0 \rightarrow$  sur(sous)-estimation systématique.
- $\text{Var}(\epsilon) = \text{Var}(X) = \sigma^2 = \text{variance}$  (théorique) des mesures.  $\sigma$  quantifie le **manque de précision** du processus de mesure, on l'appelle parfois l'**incertitude** statistique ou aléatoire.  $\sigma = 0 \rightarrow$  absence de perturbations aléatoires  $\leftrightarrow$  mesures répétées qui ne varient pas. Si  $\sigma$  est grand alors les mesures répétées seront fortement dispersées  $\leftrightarrow$  peu de précision.



(a) Le biais et l'incertitude sont tous deux faibles. (b) le biais est important; l'incertitude est petite. (c) le biais est petit; l'incertitude est grande. (d) Le biais et l'incertitude sont tous deux importants.

Dans la vie réelle, nous ne connaissons pas la vraie valeur. Ainsi, le graphique des mesures serait ...



(a) et (b) sont plus précis. Mais sans informations supplémentaires sur la valeur réelle, nous ne pouvons rien dire sur le biais.

Le biais est difficile (impossible) à évaluer, car il nécessite la connaissance de la vraie valeur  $\mu$  qui est inconnue ! L'estimation du biais est essentiellement un processus d'étalonnage (des appareils de mesure), pour lequel **des informations externes sont nécessaires**. Par exemple, mesurer à plusieurs reprises une quantité standard dont la vraie valeur est connue puis estimer le biais comme la différence entre la moyenne des mesures et la vraie valeur de référence.

À partir d'ici, nous supposons, sauf indication contraire, que le biais a été réduit à un niveau négligeable  $\Leftrightarrow$   $X = \mu + \epsilon$ , avec  $E(\epsilon) = 0$ .

# MODÉLISATION DES MESURES ITÉRÉES

Supposons qu'après une série de  $n$  mesures indépendantes effectuées **dans des conditions identiques** (même opérateur, même matériel, ...) on obtienne les valeurs

$$x_1, x_2, \dots, x_n$$

**HYPOTHÈSE FONDAMENTALE :** Chaque  $x_i$  est une réalisation particulière de la **variable aléatoire**

$$X_i = \text{résultat de la mesure numéro } i.$$

Typiquement, les v.a.  $X_i$ ,  $i = 1, \dots, n$  sont supposées être **indépendantes et identiquement distribuées (i.i.d.)** de moyenne  $\mu = E(X_i)$  et de variance  $\sigma^2 = \text{Var}(X_i)$ .

Ce qui est équivalent à dire que

$$X_i = \mu + \epsilon_i$$

et que les  $\epsilon_i$  sont i.i.d. avec  $E(\epsilon_i) = 0$  et  $\text{Var}(\epsilon_i) = \sigma^2$ .



Soit

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$$

la moyenne de note échantillon aléatoire  $\{X_1, \dots, X_n\}$ .

Si  $\epsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, n$ , alors  $X_i \sim N(\mu, \sigma^2)$ . Dans ce cas, on sait que,

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

Ce qui est équivalent à écrire que

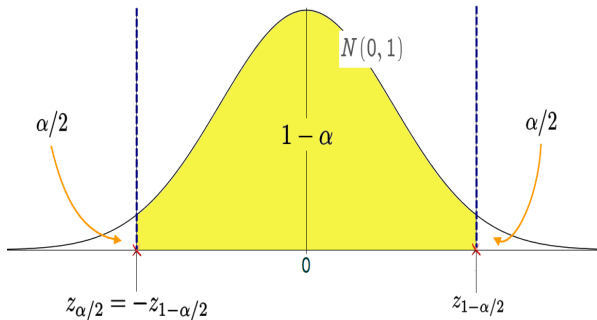
$$\boxed{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)}$$

Ce résultat s'avère très utile par la suite.

# LES INTERVALLES DE CONFIANCE

On sait que si  $Z$  est une v.a.  
 $N(0, 1)$ , alors

$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha,$$
$$\forall \alpha \in (0, 1)$$



Puisque  $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ,

$$\Rightarrow P(-z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

$$\Leftrightarrow P(\bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

## L'intervalle aléatoire

$$\bar{X}_n \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = \left[ \bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

est ce qu'on appellera un intervalle de confiance (IC) pour  $\mu$  de niveau de confiance  $1 - \alpha$ .

- Si  $\alpha = 0.32$ , alors  $z_{1-\alpha/2} = 0.99 \Rightarrow$  IC à 68% de confiance.
- Si  $\alpha = 0.05$ , alors  $z_{1-\alpha/2} = 1.96 \Rightarrow$  IC à 95% de confiance.
- Si  $\alpha = 0.01$ , alors  $z_{1-\alpha/2} = 2.58 \Rightarrow$  IC à 99% de confiance.

Le niveau de confiance utilisé le plus couramment (en statistique) est de 95%.

## EXEMPLE

Supposons que les mesures de la période du pendule proviennent d'une distribution Normale avec  $\sigma = 0.2$  :

2.3, 2.4, 2.5, 2.4, 2.2, 2.8, 2.3, 2.7, 2.6, 2.5

Calculons un IC à 95% pour  $\mu$  :

```
> x <- c(2.3, 2.4, 2.5, 2.4, 2.2, 2.8, 2.3, 2.7, 2.6, 2.5)
> mx <- mean(x)
> alpha <- 0.05
> dx <- qnorm(1 - alpha/2) * (0.2/sqrt(10))
```

```
> c(mx, dx)
[1] 2.470 0.124
```

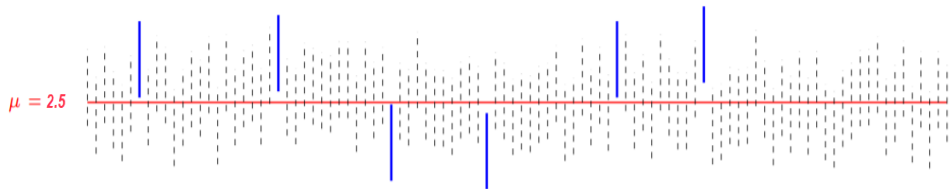
On conclut que la période du pendule ( $\mu$ ) est estimée à 2.470 avec une incertitude (statistique) de 0.124 et on écrit

$$\mu = 2.470 \pm 0.124, \text{ avec } 95\% \text{ de confiance}$$

Autrement dit, un intervalle de confiance à 95% pour  $\mu$ , la vraie valeur de la période, est donné par [2.346, 2.594].

# INTERPRÉTATION

- On simule 100 échantillons de taille  $n = 10$  selon une loi  $N(2.5, 0.04)$
- On prend  $\alpha = 0.05$  et on calcule pour chaque échantillon l'intervalle  $\bar{x} \pm 1.96 \frac{0.2}{\sqrt{10}}$



- Sur les 100 intervalles calculés, environ 95 contiennent  $\mu$ , alors que 5 ne la contiennent pas.

- Lorsqu'on calcule à partir de notre échantillon un intervalle de confiance au niveau 95%, on espère que notre intervalle ne soit pas parmi les 5% des intervalles qui ne contiennent pas  $\mu$ .
- *Dans l'exemple précédent, sur la base de l'échantillon étudié, et les hypothèses formulées, on est quasi-sûr que  $\mu$ , la vraie période du pendule, est comprise entre 2.346 et 2.594. Le risque (statistique) de se tromper est de 5%.*
- Si ce risque de 5% nous paraît trop important, on pourra élever le niveau de confiance à, par exemple, 99% :

```
> mean(x) + c(-1, 1) * qnorm(1 - 0.01/2) * (0.2/sqrt(10))  
[1] 2.31 2.63
```

On note que  $2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ , la longueur d'un IC, est gouvernée par trois facteurs :

- Lorsque la confiance souhaitée augmente ( $\alpha$  diminue), la longueur de l'IC augmente. Or, un intervalle de confiance trop large ne sera pas très utile (car peu informatif).
- À mesure que la taille de l'échantillon (nombres de mesures)  $n$  augmente, la longueur de l'IC diminue.
- À mesure que l'écart-type ( $\sigma$ ), qui reflète les variations dans les mesures, augmente, la longueur de l'IC augmente.



## REMARQUE

Pour une variable Normale de moyenne  $\mu$  et de variance  $\sigma^2$ , on a déjà vu que l'intervalle suivant

$$[\mu - 1.96\sigma, \mu + 1.96\sigma]$$

contient 95% de la population (des valeurs possibles de  $X$ ).

Il ne faudra pas confondre cet intervalle, qu'on peut appeler **intervalle de tolérance**, avec l'intervalle de confiance à 95%.

Quelle est la différence ?

## CAS OÙ $\sigma^2$ EST INCONNUE

La validité et le calcul de l'intervalle  $\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  nécessitent trois conditions :

1.  $X_i$  sont i.i.d.
2.  $X_i \sim N(\mu, \sigma^2)$
3.  $\sigma^2$  est connue

**PROBLÈME :**  $\sigma^2$ , la variance théorique (de la population d'intérêt), est pratiquement toujours inconnue.

**SOLUTION :** Utiliser  $S_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$  la variance empirique des  $n$  mesures

Malheureusement, de façon générale,

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \not\sim N(0, 1).$$

En effet, l'estimation de  $\sigma$  entraîne une incertitude supplémentaire qu'on ne peut pas négliger **sauf dans le cas où notre taille d'échantillon est "suffisamment grande"**.

On peut montrer que pour corriger cela, il suffit de remplacer, dans la formule précédente, la loi normale par la loi dite de **Student**.

Plus précisément, on sait montrer que

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

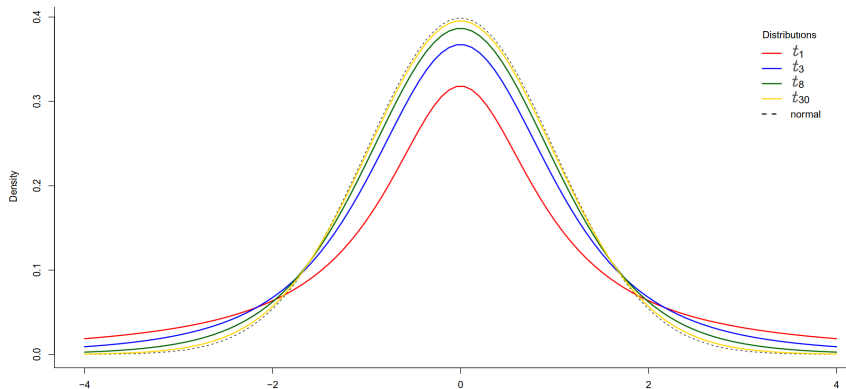
où  $t_{n-1}$  est la distribution **t de Student** à  $n - 1$  degrés de libertés, où  $n$  est le nombre de mesures dont on dispose.

La loi de Student est une loi dérivée de la loi normale standard, mais qui se distingue de ce dernier par des queues (ailes) plus épaisses.

Aussi, comme la normale standard, la densité Student est symétrique autour de 0 (sa moyenne).

Cette loi admet un seul paramètre qu'on appelle **degré de liberté**. Pour indiquer loi de Student, on utilise la lettre **t** indexée par le degré de liberté. Par exemple, on écrit  $t_8$  pour dire une loi de Student à 8 degrés de libertés.

Voici la densité de Student pour différents degrés de liberté :



$t_n \rightarrow N(0, 1)$  quand  $n \rightarrow \infty$ . Il est quasiment impossible de distinguer un  $t_{30}$  d'un  $N(0, 1)$ . En pratique, on n'utilise la Student que lorsque  $n < 30$ .

Intuitivement, l'utilisation de  $t$  à la place de  $N(0, 1)$  traduit une plus grande incertitude due au fait que  $\sigma$  est inconnu et  $n$  est petit.

En conclusion, lorsqu'on dispose d'un **petit nombre de mesures** qui provient d'une **loi Normale** et qu'**on ne connaît pas la variance** théorique (de la population) alors l'IC de  $\mu$  est donné par

$$\bar{X} \pm t_{n-1}(1 - \alpha/2) \frac{S}{\sqrt{n}},$$

où  $t_{n-1}(1 - \alpha/2)$  est le quantile d'ordre  $(1 - \alpha/2)$  d'un  $t_{n-1}$ .

$\alpha$	0.32 (68%)	0.05 (95%)	0.01 (99%)
n	$t_{n-1}(1 - \alpha/2)$	$t_{n-1}(1 - \alpha/2)$	$t_{n-1}(1 - \alpha/2)$
2	1.82	12.71	63.66
4	1.19	3.18	5.84
10	1.05	2.26	3.25
31	1.01	2.04	2.75
1000	0.99	1.96	2.58

Dans notre exemple du pendule (**sans l'hypothèse  $\sigma = 0.2$** ) cela donne:

```
> dx <- qt(1 - alpha/2, 9) * (sd(x)/sqrt(10))  
> dx
```

```
[1] 0.135
```

On conclut que

$\mu = 2.470 \pm 0.135$ , avec 95% de confiance.

$\Leftrightarrow$  un intervalle de confiance à 95% pour  $\mu$ , la vraie valeur de la période, est donné par [2.335, 2.605].

Avec R, vous pouvez obtenir ce dernier intervalle directement en utilisant la fonction `t.test` :

```
> t.test(x, conf.level = 0.95)$conf
```

```
[1] 2.33 2.61
```

```
attr(,"conf.level")
```

```
[1] 0.95
```

## RÉCAPITULATIF ET REMARQUES

Lorsqu'un mesurage, **avec erreur supposé normal**, est effectué  $n$  ( $n \geq 1$ ) fois de façon identique et indépendante, le résultat se donne sous la forme

$$x_m \pm \delta x \equiv \bar{x}_n \pm k\sigma_{\bar{x}_n}$$

- $k = 1, 2$  ou  $2.6$ , en fonction de la confiance souhaitée (68%, 95% ou 99%)
- $\sigma_{\bar{x}_n} = \sqrt{\text{Var}(\bar{X}_n)} = \sigma/\sqrt{n}$ , l'écart-type de la moyenne  $\bar{X}_n$ .
- $\sigma = \sqrt{\text{Var}(X)}$  est l'écart-type de  $X$ . S'il est inconnu, on le remplacera par  $s$ , et **on ajustera  $k$  en fonction de  $n$  et de la confiance souhaitée.**
- Par la suite, sans autre précision ou contre indication,  $k$  sera considérée, par défaut, égale à 1  $\rightarrow$  **par défaut,  $x_m \pm \delta x$  représente le résultat du calcul  $\bar{x}_n \pm \sigma_{\bar{x}_n}$ .**
- Dans le cas d'une **mesure unique** ( $n = 1$ ) cette expression se simplifie à  $x \pm \sigma$ .
- $\sigma$  est *l'incertitude associée à une mesure* alors que  $\sigma_{\bar{x}_n}$  est *l'incertitude associée à la moyenne* de  $n$  mesures.

## INCERTITUDE RELATIVE

La quantité

$$\delta^r x = \frac{\delta x}{x_m},$$

souvent exprimée en % après multiplication par 100, est appelée incertitude relative. L'avantage de cette quantité est qu'elle est **sans unité** ce qui facilite la lecture et l'interprétation des résultats. Aussi, elle permet de comparer des incertitudes sur des mesures avec des unités et/ou des échelles différentes.

**EXEMPLE:** voici deux mesures

(mesure x):  $21 \pm 1$  g    et    (mesure y):  $0.021 \pm 0.001$  Kg

Puisque

$$\delta x / x_m = \delta y / y_m = 0.05$$

On peut écrire ces deux mesures en utilisant l'incertitude relative sous la forme suivante

(mesure x):  $21 \text{ g} \pm 5\%$     et    (mesure y):  $0.021 \text{ Kg} \pm 5\%$



# LA MÉTHODE BOOTSTRAP

La construction d'un intervalle de confiance par la formule classique (basé sur la Normal ou la Student) peut conduire à un intervalle biaisé (i.e., un intervalle avec une probabilité inférieure à celle attendue de contenir la vraie valeur).

Une meilleure alternative est d'utiliser la méthode dite de **Bootstrap**.

Le "bootstrap" est une technique d'inférence statistique très générale et très robuste permettant, entre autres, de construire des intervalles de confiance.

C'est une technique relativement récente qui date de la fin des années 1970, époque où la possibilité de **calculs intensifs** devient abordable.

Le bootstrap est basé sur le **rééchantillonnage** des données observées.

Si notre échantillon contient  $n$  valeurs, alors, dans sa version la plus simple, le bootstrap consiste à

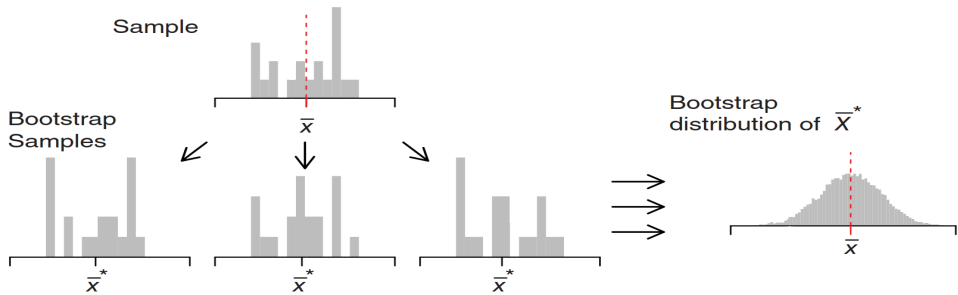
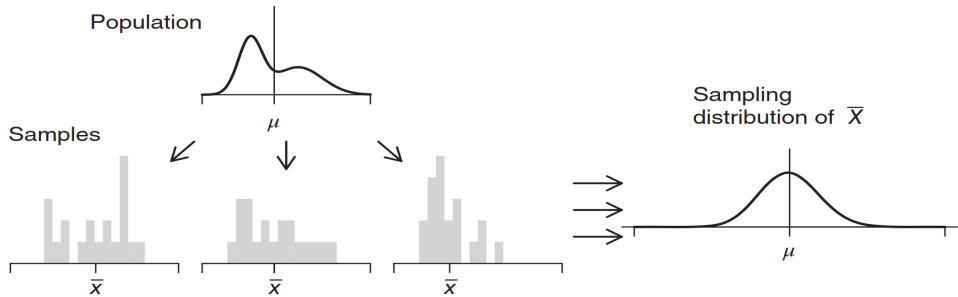
(1) Tirer au sort **avec remise**, *parmi les valeurs observées*, une série **de même taille  $n$** . On obtient ainsi ce qu'on appelle un **échantillon bootstrap**. Dans ce dernier, certaines valeurs (de la série originale) apparaîtront plusieurs fois alors que d'autres valeurs ne seront pas présentes.

(2) On calcule la moyenne empirique de l'échantillon bootstrap obtenue en (1).

On répète ces deux opérations  $B$  fois, où  **$B$  est un grand nombre** (par exemple 1000). On obtient ainsi  $B$  moyennes ( $\bar{x}_b^*$ ,  $b = 1, \dots, B$ ) qui correspondent aux  $B$  échantillons bootstrap. L'étoile indique que la moyenne est calculée non pas à partir de l'échantillon original, mais à partir d'un échantillon bootstrap.

**POURQUOI RÉÉCHANTILLONNER:** Le principe derrière cette idée est le suivant. Comme l'échantillon d'origine donne une approximation de la distribution de population à partir de laquelle il a été prélevé, les (re)échantillons de cet échantillon donnent une approximation de ce que nous obtiendrions si nous prélevions de nombreux échantillons dans la population.

Ainsi, à partir de l'ensemble des  $\bar{x}_b^*$  on peut espérer construire une approximation (bootstrap) de la **distribution d'échantillonnage de la moyenne**. Si cette approximation est valide alors on pourra l'utiliser pour, par exemple, calculer un intervalle de confiance.



## INTERVALLE DE CONFIANCE PAR BOOTSTRAP

Il existe plusieurs méthodes pour construire un IC à partir de  $\{\bar{x}_b^*, b = 1, \dots, B\}$ . Parmi lesquelles il y a :

- LA MÉTHODE "NORMALE" :

$$\bar{x}_n \pm z_{1-\alpha/2} \sqrt{v_{\bar{x}}^*},$$

où  $v_{\bar{x}}^* = \frac{1}{B} \sum_{b=1}^B (\bar{x}_b^* - \bar{x})^2$  n'est rien d'autre qu'une estimation de  $\text{Var}(\bar{X}_n)$ .

- LA MÉTHODE "PERCENTILE" :

Cette méthode consiste simplement à calculer les quantiles empiriques d'ordres  $\alpha/2$  et  $1 - \alpha/2$  de la série  $\{\bar{x}_b^*, b = 1, \dots, B\}$ . Ces valeurs sont utilisées, respectivement, comme borne inférieure et borne supérieure de l'intervalle de confiance à  $100(1 - \alpha)\%$ .

On comprendra mieux cela en examinant un exemple.

# BOOTSTRAPER "MANUELLEMENT"

Reprenons les données de la période du pendule :

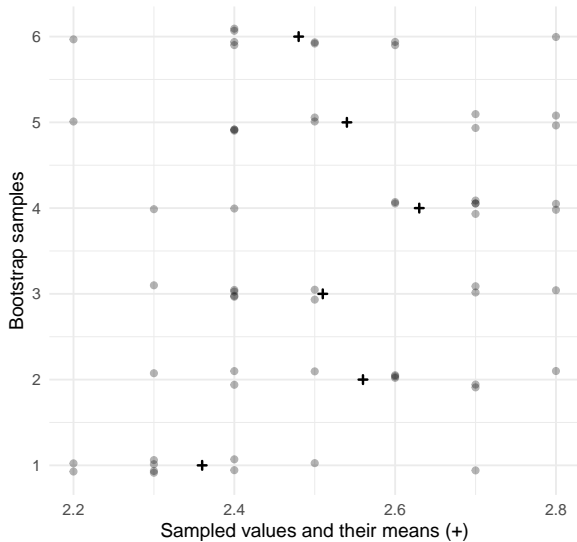
```
> x <- c(2.3, 2.4, 2.5, 2.4, 2.2, 2.8, 2.3, 2.7, 2.6, 2.5)
> boot.samp <- replicate(2000, sample(x, replace = TRUE)) # 2000 échantillons bootstrap
> boot.samp[, 1:6] # les 6 premiers échantillons bootstrap
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	2.4	2.5	2.7	2.7	2.7	2.2
[2,]	2.3	2.4	2.4	2.8	2.5	2.4
[3,]	2.3	2.6	2.5	2.6	2.5	2.4
[4,]	2.2	2.7	2.3	2.8	2.4	2.6
[5,]	2.3	2.4	2.4	2.7	2.8	2.6
[6,]	2.7	2.6	2.8	2.4	2.4	2.4
[7,]	2.5	2.7	2.7	2.6	2.4	2.5
[8,]	2.2	2.8	2.4	2.7	2.2	2.4
[9,]	2.3	2.6	2.5	2.7	2.7	2.5
[10,]	2.4	2.3	2.4	2.3	2.8	2.8

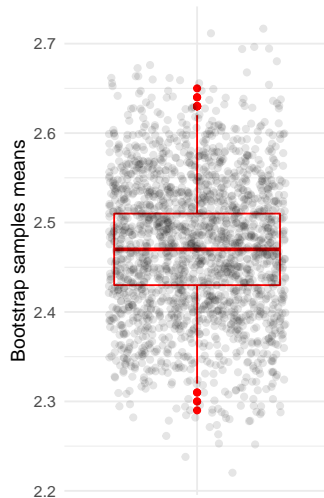
```
> boot.mean <- colMeans(boot.samp) # Moyennes des 2000 échantillons.
> boot.mean[1:6] # les 6 premiers moyennes
```

```
[1] 2.36 2.56 2.51 2.63 2.54 2.48
```

Les 6 premiers échantillons bootstrap



Les 2000 moyennes





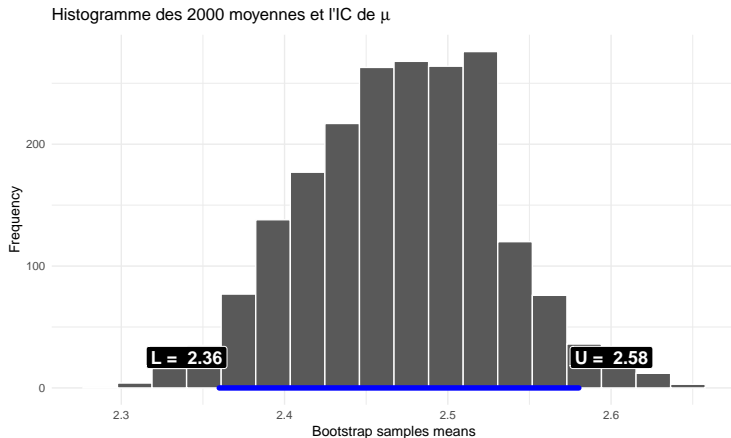
```
> # Normale CI  
> mean(x) + c(-1, 1) * qnorm(1 - alpha/2) * sqrt(mean((boot.mean - mean(x))^2))
```

```
[1] 2.36 2.58
```

```
> # Percentile CI  
> quantile(boot.mean, probs = c(alpha/2, 1 - alpha/2))
```

```
2.5% 97.5%
```

```
2.36 2.58
```



# BOOTSTRAP FACILE AVEC R

Il existe plusieurs packages et fonctions qui permettent de bootstraper facilement dans R. Voici un exemple.

```
> require(bootstrap)
> boot.mean <- bootstrap(x, 2000, mean)$thetastar # Moyennes des 2000 échantillons.
> quantile(boot.mean, probs = c(alpha/2, 1 - alpha/2))

2.5% 97.5%
2.36  2.58
```

Notez que l'IC obtenu à l'aide du bootstrap est quasi identique à celui obtenu auparavant à l'aide de la formule classique (basé sur la Student), mais cela n'est pas nécessairement le cas, en général.

## REMARQUES

Il n'est pas nécessaire de supposer la normalité des données pour utiliser la méthode du bootstrap.

Très souvent, le bootstrap est plus facile à utiliser/comprendre et donne de meilleurs résultats.

La même idée s'applique à plein d'autres statistiques. Par exemple, pour construire un IC pour la médiane, il suffit de remplacer, dans le code ci-dessus, `mean` par `median`.

Le bootstrap est une méthode asymptotique, elle ne peut pas fonctionner si la taille d'échantillon est "trop faible". La taille minimale dépend de la nature des données et de la nature du problème. Aussi, le nombre  $B$  des répliques doit être grand.

Le bootstrap n'est pas une recette miraculeuse qui résoudrait tous les problèmes! Plusieurs exemples (simples) où le bootstrap ne fonctionne pas ont été signalés dans la littérature.

# PROPAGATION DES ERREURS

- La plupart des grandeurs physiques ne sont habituellement pas accessibles directement mais s'obtiennent en combinant (à l'aide d'une formule qui traduit une certaine loi) plusieurs quantités.
- Par exemple, l'évaluation de l'**aire**  $A$  d'un rectangle requiert les mesures de sa **longueur**  $L$  et de sa **largeur**  $I$  avant de calculer  $A = LI$ . De même, la **vitesse**  $V$  d'un mobile revient à mesurer sa **distance**  $D$  parcourue au bout d'un **temps**  $T$  ce qui permet d'en déduire  $V = D/T$ .
- Ainsi, les mesures intéressantes s'effectuent en deux étapes: (1) mesure directe (répétée ou non) d'une ou plusieurs grandeurs et (2) un calcul final conduisant au résultat (grandeur) cherché.
- Chacune des grandeurs mesurées a une certaine incertitude. **Ces incertitudes se propagent au résultat final**. L'erreur qui touche ce dernier dépend des incertitudes individuelles.

- On veut déterminer de quelle manière chacune de ces incertitudes se répercute sur la grandeur finale.
- Dans l'exemple de l'aire du rectangle  $A$ . Les distances mesurées comportent une incertitude  $\delta L$  sur la longueur et  $\delta I$  sur la largeur. Mais comment déduire l'incertitude  $\delta A$  sur l'aire calculée en multipliant ces deux distances ?
- De façon plus générale, supposons qu'on cherche

$$\mu = g(\mu_1, \mu_2),$$

où  $\mu_1$  et  $\mu_2$  sont les vraies valeurs des deux grandeurs nécessaires pour le calcul de la grandeur convoitée  $\mu$ .  $g$  étant une fonction quelconque et connue qui traduit une certaine loi physique.

- Si  $\mu_1$  et  $\mu_2$  étaient connues, alors on pourrait obtenir  $\mu$  sans aucune difficulté. Malheureusement ceci est impossible puisque, en pratique, les vraies valeurs sont inconnues.

- Soient  $X_i = \mu_1 + \epsilon_{1i}$ ,  $i = 1, \dots, n_1$  et  $Y_i = \mu_2 + \epsilon_{2i}$ ,  $i = 1, \dots, n_2$ , deux séries de mesures des deux grandeurs en question ( $n_1 \geq 1$  et  $n_2 \geq 1$ ).
- Si nous disposons des mesures individuelles  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , alors nous pouvons calculer  $g(X_i, Y_i)$  et appliquer directement ce qui a été vu précédemment.
- Le plus souvent (c'est ce qu'on va considérer par la suite) les résultats sont fournis sous la forme

$$\bar{X} \pm \sigma_{\bar{X}},$$

$$\bar{Y} \pm \sigma_{\bar{Y}},$$

où  $\bar{X} = n_1^{-1} \sum_i X_i$  et  $\bar{Y} = n_2^{-1} \sum_i Y_i$ .

- La meilleure estimation de la grandeur cherchée qu'on pourrait obtenir en combinant ces deux quantités est

$$Z = g(\bar{X}, \bar{Y}).$$

Notez que  $Z = \mu + \epsilon$  avec  $\epsilon = g(\bar{X}, \bar{Y}) - g(\mu_1, \mu_2)$ .

Suivant la démarche exposée dans la section précédente, le résultat final de ce mesurage, qui combine X et Y, peut être exprimé sous la forme

$$Z \pm \sigma_Z$$

Pour calculer effectivement cet intervalle, censé contenir la vraie valeur  $\mu$ , il faut trouver  $\sigma_Z = \sqrt{\text{Var}(Z)}$ . Ce qui n'est pas toujours une tâche facile.



## CAS D'UNE COMBINAISON LINÉAIRE

Supposons que  $Z$  peut être écrit comme suit

$$Z = c_0 + c_1 \bar{X} + c_2 \bar{Y},$$

où  $c_0, c_1$ , et  $c_2$  sont des constantes, et  $\bar{X}$  et  $\bar{Y}$  sont **indépendantes**. Dans ce cas, il est possible de calculer (exactement)  $\sigma_Z$  en utilisant la formule suivante.

$$\sigma_Z = \sqrt{c_1^2 \text{Var}(\bar{X}) + c_2^2 \text{Var}(\bar{Y})}.$$

**EXEMPLE :** La longueur  $L$  et la largeur  $I$  d'un rectangle sont données respectivement par  $4.2 \pm 0.1$  cm et  $7.6 \pm 0.3$  cm. Calculez le **périmètre** et précisez son incertitude.

Le périmètre est donné par  $P = 2 \times (L + I)$ . Nous pouvons l'estimer par  $2 \times (4.2 + 7.6) = 23.6$  cm. Son incertitude est donnée par

$$\sigma_P = 2\sqrt{\sigma_L^2 + \sigma_I^2} \approx 0.63 \text{ cm}$$

Nous concluons que le périmètre est de  $23.6 \pm 0.63$  cm.

## CAS D'UNE COMBINAISON NON-LINÉAIRE

Dans le **cas général**, où, pour une fonction  $g$  quelconque,

$$Z = g(\bar{X}, \bar{Y}), \quad g : \mathbb{R}^2 \rightarrow \mathbb{R},$$

nous ne pouvons pas calculer exactement  $\text{Var}(Z)$  mais nous pouvons l'approximer.

En effet, puisque  $\text{Var}(Z) = \text{Var}(\epsilon)$ , avec  $\epsilon = g(\bar{X}, \bar{Y}) - g(\mu_1, \mu_2)$ , nous pouvons montrer, en utilisant un développement de Taylor d'ordre 1, que

$$\sigma_Z \approx \sqrt{\left(\frac{\partial Z}{\partial \bar{X}}\right)^2 \sigma_{\bar{X}}^2 + \left(\frac{\partial Z}{\partial \bar{Y}}\right)^2 \sigma_{\bar{Y}}^2},$$

où  $\frac{\partial Z}{\partial \bar{X}}$  est la dérivée partielle de  $Z = g(\bar{X}, \bar{Y})$  par rapport à la première variable, i.e., la dérivée qui s'obtient en dérivant  $g$  par rapport à  $\bar{X}$ , tout en maintenant  $\bar{Y}$  constant.

C'est ce qu'on appelle la **formule générale de propagation d'incertitudes**. Il est important de noter que **cette formule n'est valide que dans le cas où les mesures (i.e. les variables  $\bar{X}$  et  $\bar{Y}$ ) sont indépendantes**.

**EXEMPLE 1:** Le rayon  $R$  d'un cercle est mesuré à  $5 \pm 0.01$  cm. Estimez l'aire du cercle et trouvez l'incertitude dans cette estimation.

L'aire est donnée par  $A = \pi R^2$ . Basée sur la valeur  $R$  mesurée, l'aire du cercle est

$$\begin{aligned} A \pm \left| \frac{dA}{dR} \right| 0.01 \\ &= \pi 5^2 \pm (2\pi 5)(0.01) \\ &= 78.54 \pm 0.31 \text{ cm}^2 \equiv [78.23, 78.85] \text{ cm}^2 \end{aligned}$$

**REMARQUE:** Puisque la fonction  $g : x \rightarrow \pi x^2$  est croissante sur  $[0, \infty)$ , nous pouvons obtenir un intervalle de confiance pour  $A$  en appliquant simplement la transformation  $g$  sur les bornes de l'intervalle de confiance de  $R$ . Ce qui donne

$$[\pi 4.99^2, \pi 5.01^2] = [78.23, 78.85] \text{ cm}^2$$

**EXEMPLE 2:** La **masse** d'une roche est mesurée à  $m = 674 \pm 1$  g et son **volume** est mesuré à  $V = 261 \pm 0.1$  mL. Estimer la **densité** de la roche et trouver l'incertitude dans l'estimation.

Puisque la densité est donnée par  $D = m/V$  et que

$$\frac{\partial D}{\partial m} = \frac{1}{V} = 0.0038 \text{ mL}^{-1} \quad \text{et} \quad \frac{\partial D}{\partial V} = -\frac{m}{V^2} = -0.0099 \text{ g/mL}^2$$

On conclut que la densité de la roche est

$$\begin{aligned} & 674/261 \pm \sigma_D \\ & = 2.582 \pm \sqrt{(0.0038)^2(1)^2 + (-0.0099)^2(0.1)^2} \\ & = 2.582 \pm 0.004 \text{ g/mL} \end{aligned}$$