Variable aléatoire, estimation ponctuelle et par intervalle

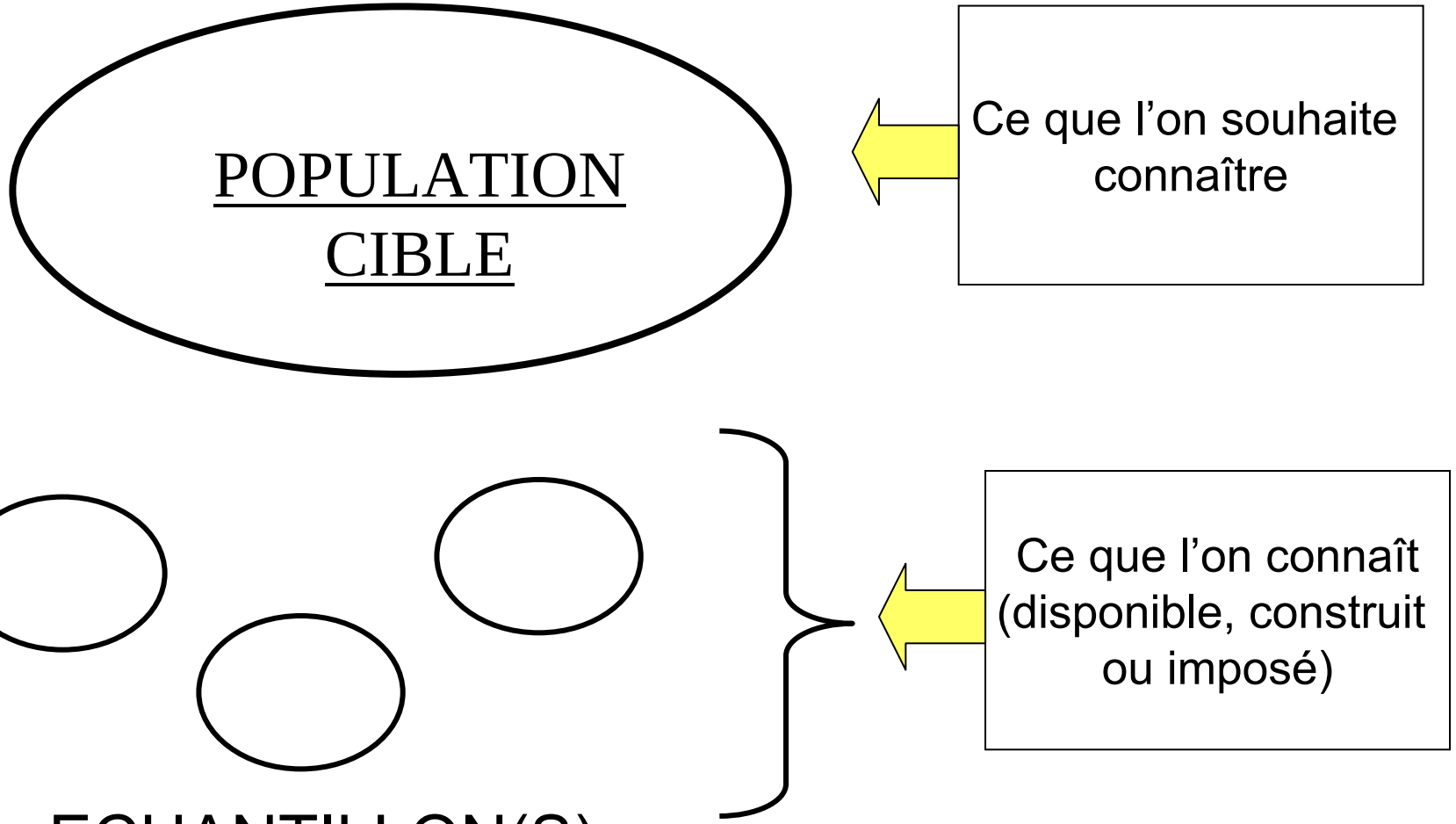
Professeur E. Albuisson

> CHU et Faculté de Médecine

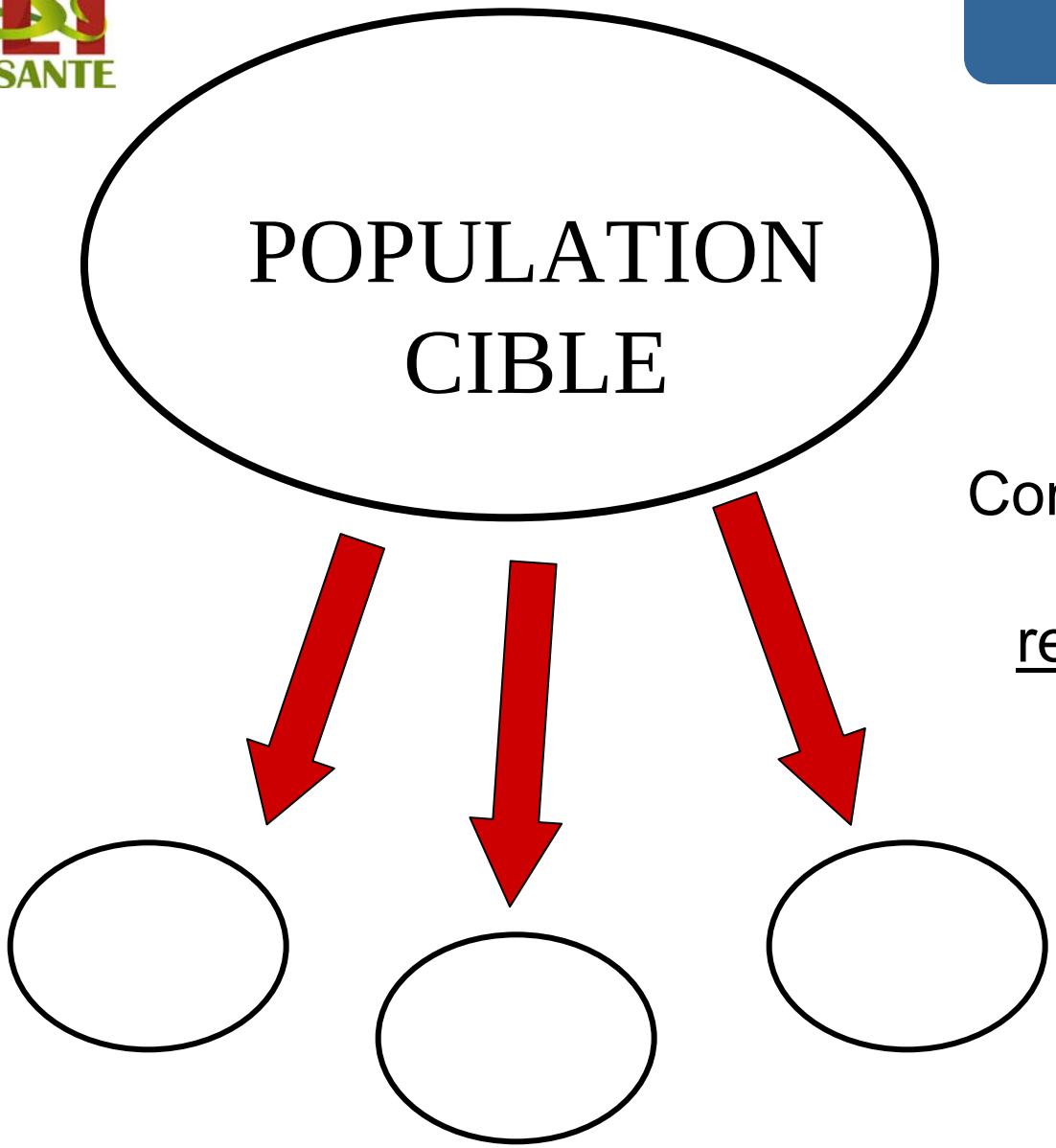
Nancy-Université
Université Henri Poincaré

ministère éducation nationale enseignement supérieur recherche

Position du problème

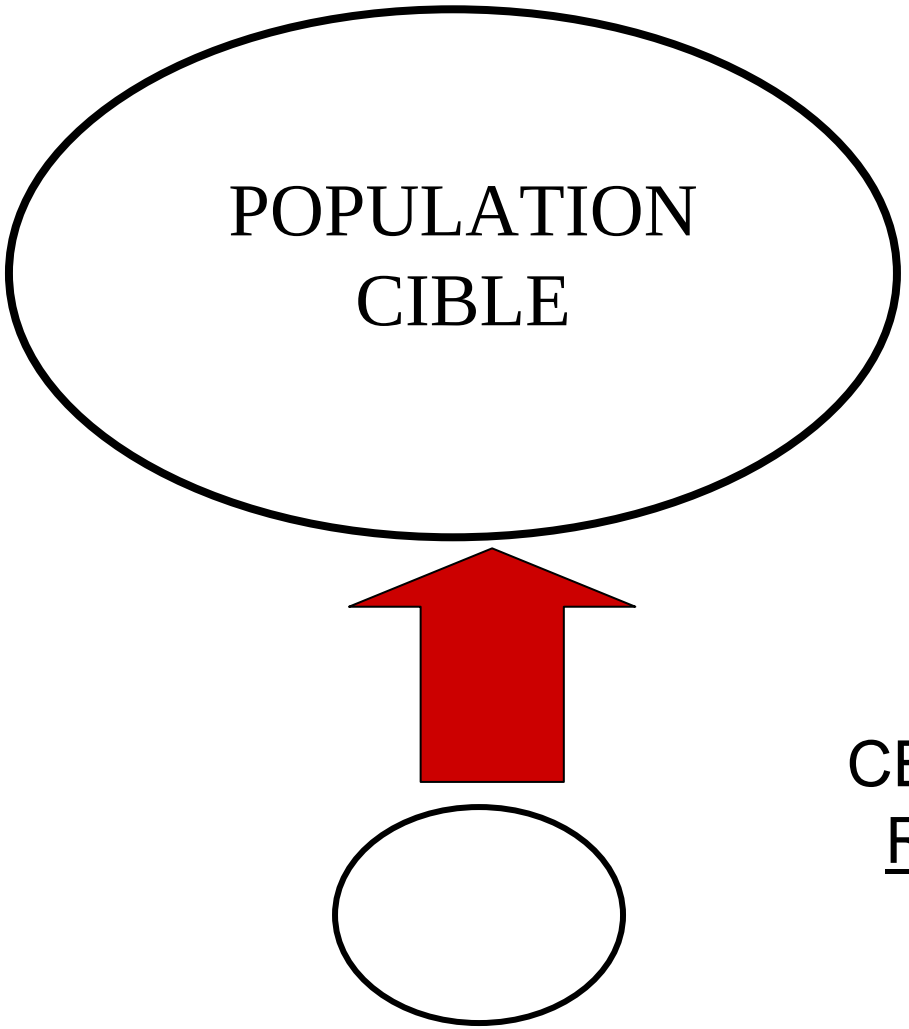


Représentativité ?



Comment obtenir un (ou des) échantillon(s) <u>représentatif(s) de la population cible?</u>

Représentativité ?



CET ECHANTILLON EST-IL REPRESENTATIF DE LA POPULATION CIBLE?

OBTENIR UN (DES)ECHANTILLON(S) REPRESENTATIF(S) ?

« Photographie en réduction » de la population cible

Equiprobabilité pour chaque individu de la population d'en faire partie.

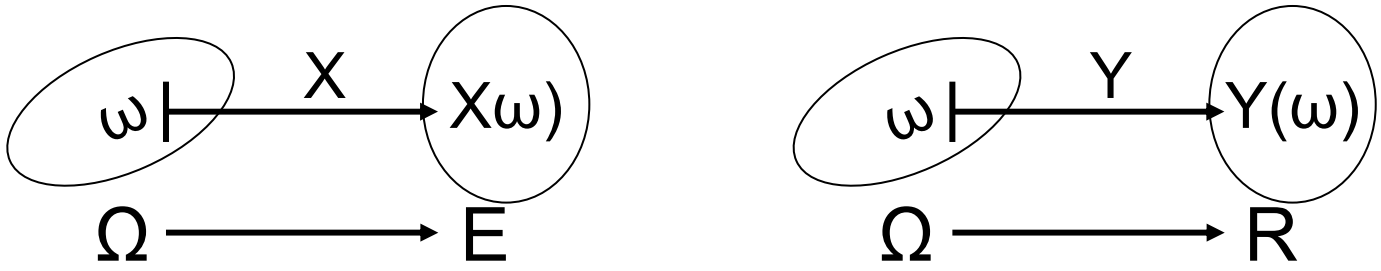
Le tirage au sort est la meilleure méthode pour garantir cette équiprobabilité

Equilibrer les caractères connus et les caractères inconnus

Sinon BIAIS avec des résultats (ex: différences constatées) non extrapolables à la population cible.

Variable aléatoire

Application qui, munie de son argument, permet de connaitre le résultat d'une expérience (épreuve) aléatoire.



Ex: Variable aléatoire entière: Application X de Ω dans E qui à ω fait correspondre X(ω). (ex: X correspond à la lecture de la face supérieure d'un dé après le lancé du dé avec X (face supérieure obtenue) = 1)

Ex: Variable aléatoire réelle: Application Y de Ω dans R
qui à ω fait correspondre Y(ω). (ex: Y correspond au dosage de la glycémie avec Y (patient nommé Martin)= 0.90 g/l)

Variable aléatoire

La variable aléatoire permet d'obtenir une valeur mais *elle n'est pas* cette valeur (X=3).

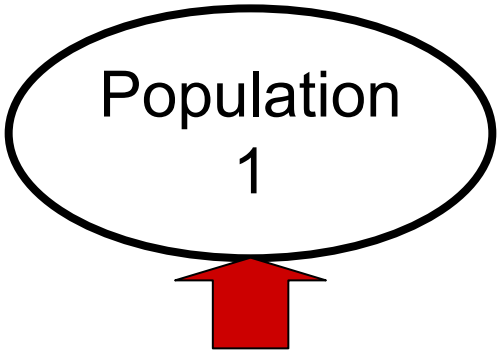
En conséquence, les variables aléatoires (les applications) seront notée en lettres majuscules :

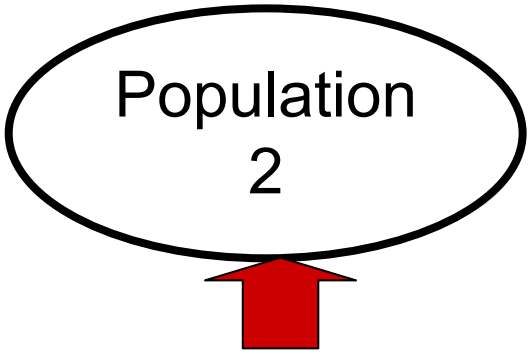
X, M, S,…et

leurs réalisations (les valeurs prises par l'application une fois celle-ci munie de son argument) en lettres minuscules: x, m, s, …

(ex: La réalisation de X est X(ω)=x=1)

LA DIFFERENCE AU SENS STATISTIQUE. A QUEL NIVEAU SE POSER LA QUESTION ?





ECHANTILLON 1	ECHANTILLON 2
Effectif n1	Effectif n2
Valeur observée sur l'échantillon: 47	Valeur observée sur l'échantillon: 61

LA STATISTIQUE INFERENTIELLE
LES HYPOTHESES…

…SE POSENT AU NIVEAU DES POPULATIONS+++

LES COMPARAISONS SE SITUENT AU NIVEAU DES POPULATIONS…

…ET PAS DES ECHANTILLONS

(Même si les statistiques des échantillons sont utilisées en pratique)

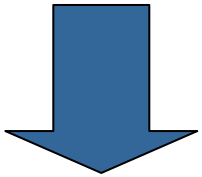
DEUX INTERROGATIONS MAJEURES:

1/LA DISTRIBUTION DU CRITERE (VARIABLE X ETUDIEE)
En particulier distribution normale?

2/LA VARIABILITE DU CRITERE (VARIABLE X ETUDIEE)

Démarche

1/ LE BUT DE L'ETUDE ?



2/ DEFINIR
LA POPULATION CIBLE

3/ CHOIX DU (DES)
CRITERE(S)
PERTINENCE/BUT, OBJECTIVITÉ

4/DISTRIBUTION DU (DES) CRITERE(S) ET
VARIABILITE DU (DES) CRITERE(S)

Démarche (exemple)

1/EQUILIBRE DE L'HTA SOUS TRAITEMENT A ?

2/ HYPERTENDUS

3/ PRESSION ART. SYSTOLIQUE
PRESSION ART. DIASTOLIQUE
TRIGLYCERIDES

DISTRIBUTION DE LA P.A.S. ? VARIABILITE DE LA P.A.S. ?
DISTRIBUTION DE LA P.A.D. ? VARIABILITE DE LA P.A.D. ?
DISTRIBUTION DES TRIGLYCERIDES ? VARIABILITE DES TRIGLYCERIDES ?

Distribution du (des) critère(s)

1/ QUALITATIF

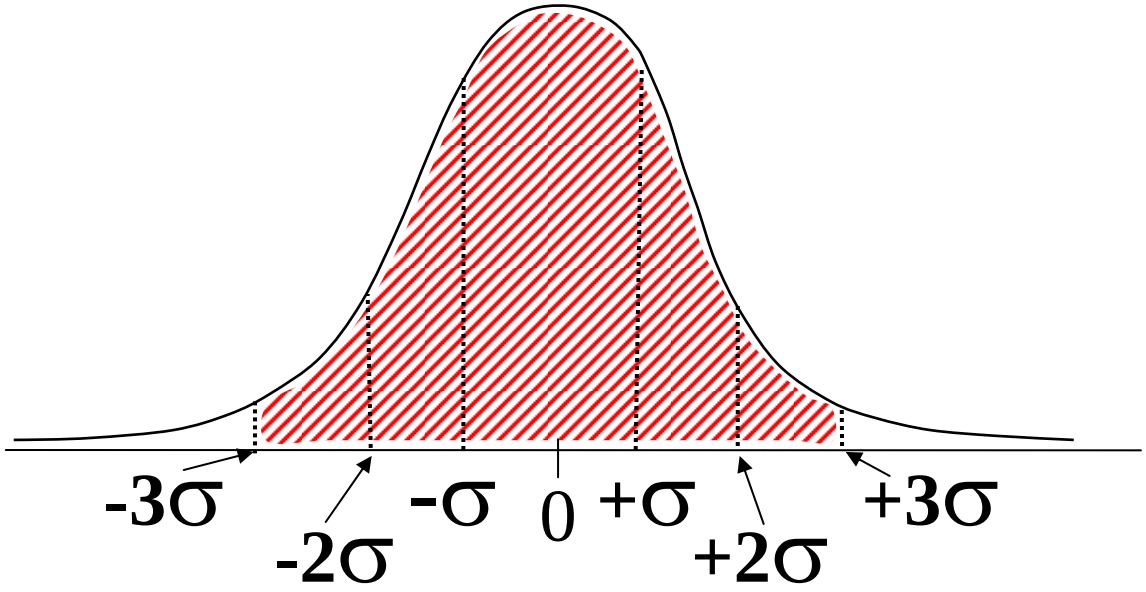
DISTRIBUTION
EFFECTIFS, POURCENTAGES

2/ QUANTITATIF

*DISTRIBUTION **NORMALE** (SYMETRIQUE) :*
MOYENNE ARITHMETIQUE,VARIANCE (ECART-TYPE)

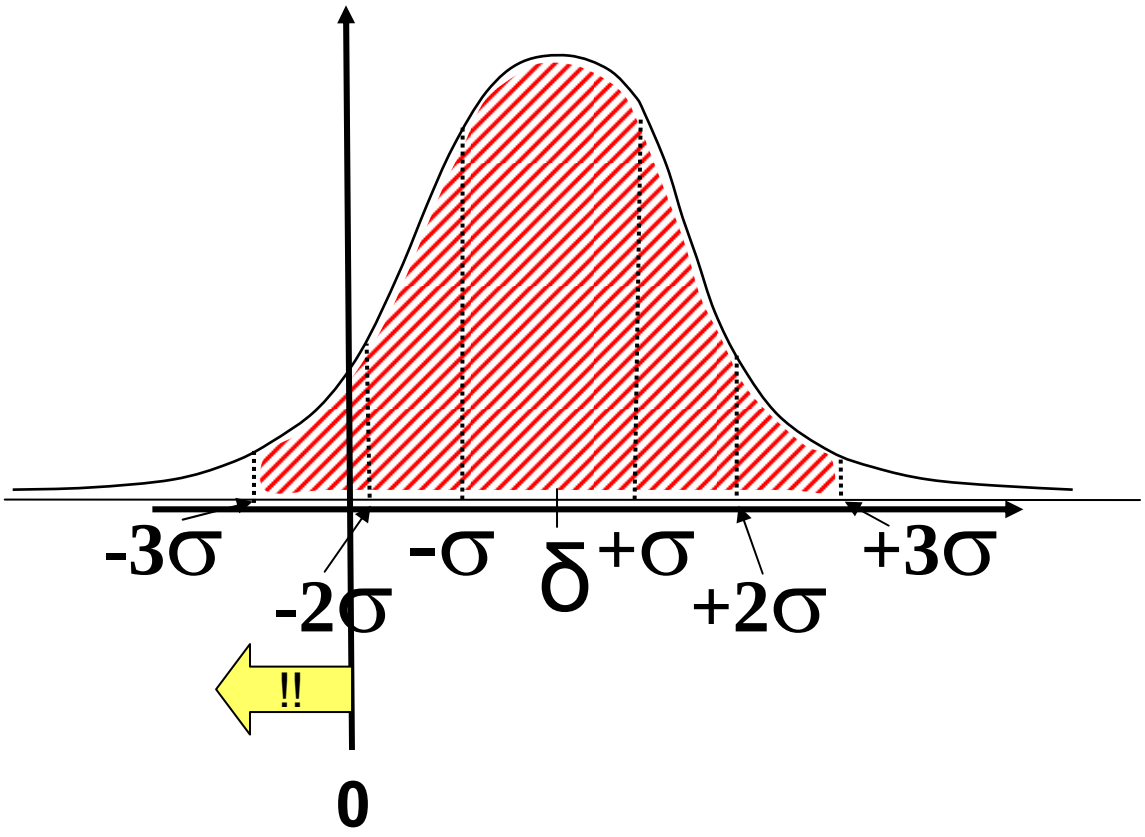
*DISTRIBUTION **NON NORMALE** (NON SYMETRIQUE)*
MOYENNE ARITHMETIQUE, MEDIANE, MIN, MAX.

ATTENTION AUX ORDRES DE GRANDEUR DES ECARTS A LA MOYENNE



Moyenne +/- 1σ	(table : 0,994)	68,0%
Moyenne +/- 2σ	(table : 1,96)	95,0%
Moyenne +/- 3σ	(table : 3)	99,7%

ATTENTION AUX VALEURS IMPOSSIBLES



Théorème Central Limite (TCL)

Quelle que soit la distribution

d'une variable aléatoire X

sa moyenne M sur un échantillon de taille n

suit asymptotiquement (pour n infini)

une loi normale $\mathrm{N}(\mu , \frac{\sigma}{\sqrt{n}})$

Standard Error = SE = $\frac{\sigma}{\sqrt{n}}$ = écart type de la moyenne

Asymptotique (pour n infini!)… en fait dés **n = 30**

('grand' échantillon)

Théorème central limite
Généralisation

Soit la somme :

Sn = X1 + X2 + X 3 + …+ Xn

Avec X1, X2, X 3, …, Xn étant n variables aléatoires, indépendantes, de variance finie

Si n → infini

Alors **Sn** → loi normale

Quelle que soit la loi des Xi

VARIABILITE

DU SUJET LUI-MEME

Interaction(s) …Sujet, Pathologie , Traitement, Environnement, Temps,…

DE LA MESURE

Précision de la mesure, Expérience

REPRODUCTIBILITE DE LA MESURE

Stabilité de l'instrument de mesure, expérience

Référence (ex : lot témoin)

MOYENNE ARITHMETIQUE (échantillon)

$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

VALEURS x_i réalisations de la variable aléatoire X dans l'échantillon

n VALEURS Effectif de l'échantillon

VARIANCE D'ECHANTILLON

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n}$$

VARIANCE
=
MOYENNE
DES
CARRES
DES ECARTS A LA
MOYENNE

s : ECART TYPE ou
Standard deviation (SD)

VARIANCE D'ECHANTILLON

AUTRE FORMULATION
DE LA VARIANCE
D'ECHANTILLON
=
MOYENNE DES CARRES
- LE CARRE DE LA
MOYENNE

(FORMULATION PLUS
PRATIQUE POUR LES
CALCULS)

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - m^2$$

NOTATIONS POUR MOYENNE ET VARIANCE

	MOYENNE (position)	VARIANCE (dispersion)
POPULATION (PARAMETRES)	μ	σ^2
⇑	↑	↑
ESTIMATION PONCTUELLE	?	?
⇑	↑	↑
ECHANTILLON (STATISTIQUES)	$m = \overline{x} = \frac{\Sigma x_i}{n}$	$s^2 = \frac{\Sigma x_i^2}{n} - m^2$

QUALITES D'UN ESTIMATEUR

- Un estimateur D_n du paramètre δ est **SANS BIAIS** si $E(D_n)= \delta$

- Un estimateur D_n du paramètre δ est **CONVERGENT** si $E(D_n – \delta)^2$ tend vers 0 quand n tend vers l'infini

La moyenne arithmétique M est un bon estimateur de μ car SANS BIAIS et CONVERGENT

Ex: SANS BIAIS

$$E(M_n) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)$$

$$= \frac{1}{n} E\left(X_1+X_2+X_3+\ldots X_n\right)$$

$$= \frac{1}{n}\left(E(X_1)+E(X_2)+E(X_3)+\ldots E(X_n)\right)$$

$$= \frac{1}{n} n \mu$$

$$= \mu$$

QUALITES D'UN ESTIMATEUR

La variance $S_n^2 = \frac{\sum_{i=1}^{n}(X_i - M)^2}{n}$

est-elle un bon estimateur de σ² ?

SANS BIAIS ?

$E(S^2_n) = \ldots = \frac{n-1}{n}\sigma^2$

Estimateur biaisé

La variance S^2_n n'est pas un bon estimateur de σ²

Nécessité d'une correction

$$\frac{n}{n-1}S_n^2 = \frac{\sum_{i=1}^{n}(X_i - M)^2}{n-1}$$

Cette expression corrigée par n/(n-1) est un bon estimateur de σ²

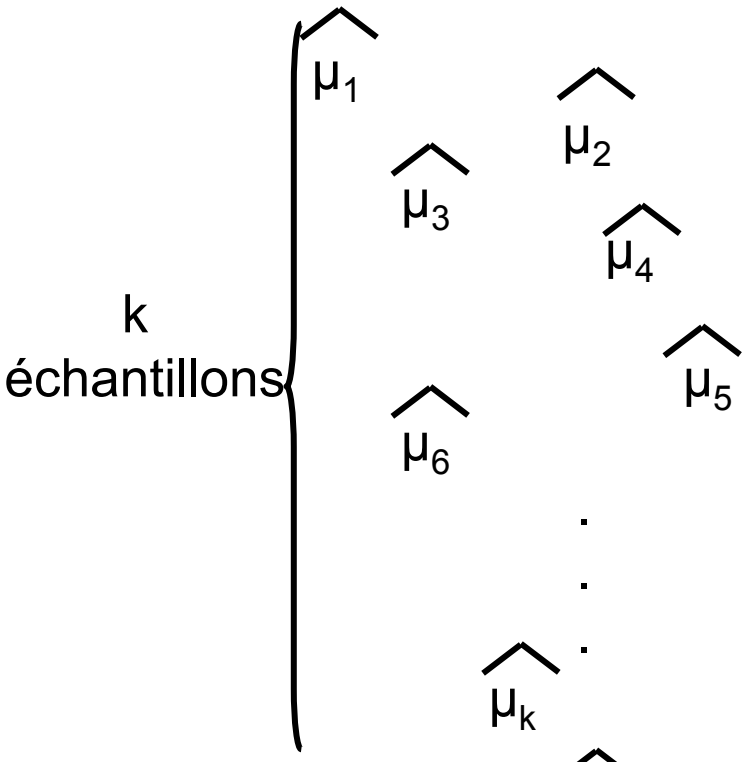
NOTATIONS POUR MOYENNE ET VARIANCE

	MOYENNE (position)	VARIANCE (dispersion)
POPULATION (PARAMETRES)	μ	σ^2
$\Uparrow$	$\uparrow$	$\uparrow$
ESTIMATION PONCTUELLE	$\hat{\mu} = m$	$\hat{\sigma^2} = \frac{n}{n-1} s^2$
$\Uparrow$	$\uparrow$	$\uparrow$
ECHANTILLON (STATISTIQUES)	$m = \bar{x} = \frac{\Sigma x_i}{n}$	$s^2 = \frac{\Sigma x_i^2}{n} - m^2$

Interprétation graphique
moyenne variance

ESTIMATION PONCTUELLE

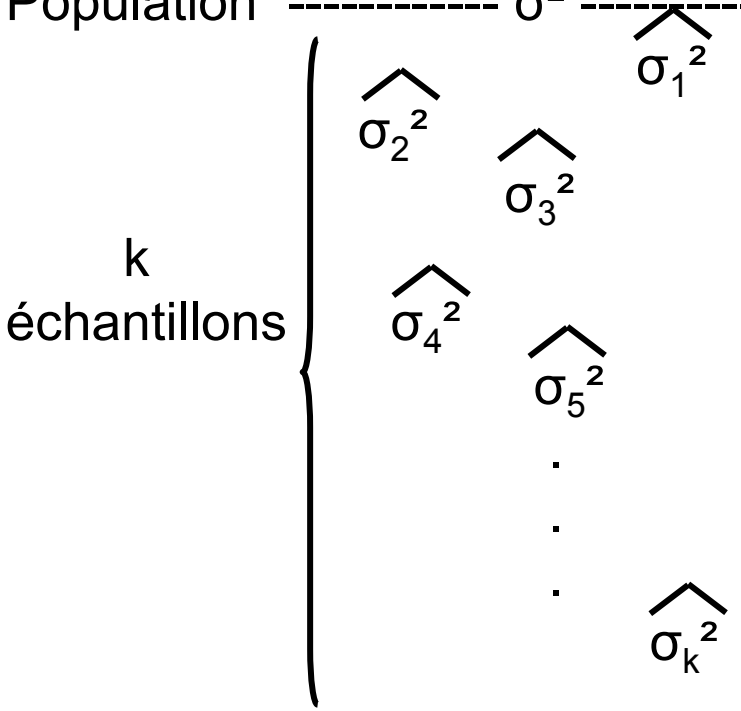
Population ---------- μ ----------



PLUS n_i ⇗ PLUS $\hat{\mu}_i = m_i$
sera proche de μ

ESTIMATION PONCTUELLE

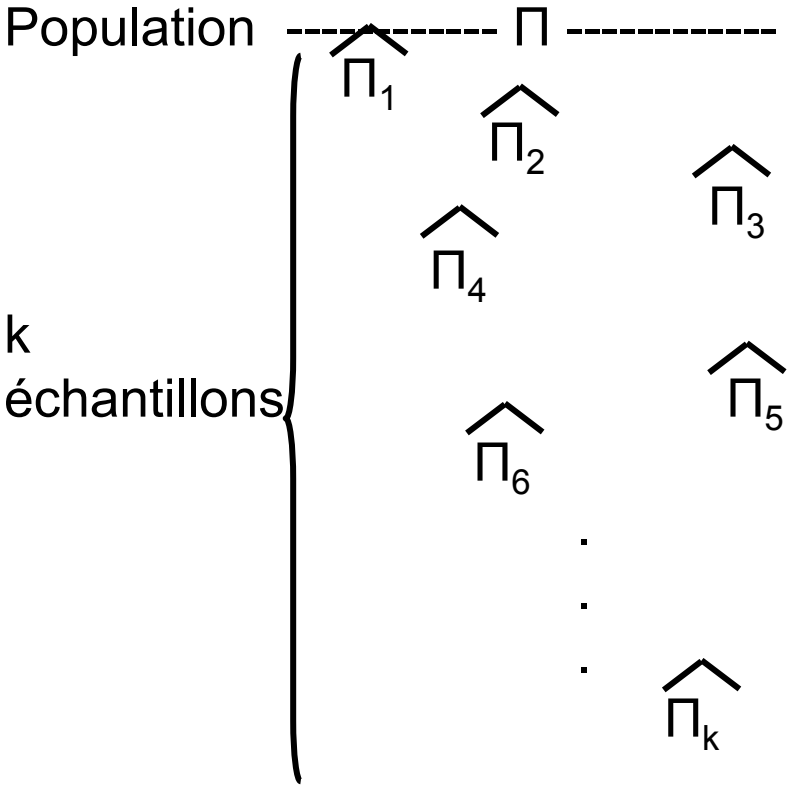
Population ---------- σ^2 ----------



PLUS n_i ⇗ PLUS $\hat{\sigma}_i^2 = \dfrac{n_i}{n_i - 1} s_i^2$
sera proche de σ^2

Interprétation graphique
proportion (variable qualitative)

ESTIMATION PONCTUELLE



PLUS n_i ⇗ PLUS $\hat{\Pi}_i = p_i$ sera proche de Π

Π est le paramètre de la population

$\hat{\Pi}$ est l'estimation ponctuelle de Π

$\hat{\Pi} = p$

avec p étant le pourcentage observé sur l'échantillon

Notion de risque dans l'estimation

Dire qu'un estimateur représente un paramètre

Quel est le **risque** de se tromper en disant cela ?

Quel est le **risque acceptable** de se tromper en disant cela ?

L'effectif ne suffit pas.
Notion de **risque de première espèce :**
Le risque α

Risque de première espèce ou risque α

Risque **choisi à 5 % sauf Indication contraire**

Risque ***acceptable :***
Convention, Contexte

	Ce que l'on dit	
Réalité ⇨ :	**Vrai**	**Faux**
	1- α	**α**
	La confiance	**Le risque**

Comment tenir compte de ce risque ?

L'estimation du paramètre **par un intervalle**

L'intervalle de confiance (1- α)
Symétrique (en général)
de part et d'autre de l'estimateur ponctuel

Exemple pour la moyenne arithmétique:

[----$\hat{\mu}$----]

Estimation par intervalle (de confiance) d'un paramètre

Il faut connaître :

La **nature** et la **distribution** de la variable

Le **risque α** consenti : (1- α) de confiance

Il faut obtenir

L'estimateur ponctuel du paramètre
(calcul sur l'échantillon)

La variance de (l'estimateur - le paramètre)
en utilisant un (autre) paramètre ou en utilisant l'estimateur ponctuel de ce (ou cet autre) paramètre calculé sur l'échantillon

Estimation par intervalle (de confiance) de la moyenne μ

Variable quantitative X (ex : dosage) de distribution normale N (μ , σ)

Le risque α consenti (5%) donne 95% de confiance

Il faut obtenir

L'estimateur ponctuel du paramètre

($\hat{\mu}$ = m avec m calculée sur l'échantillon)

La variance de (M – μ)

soit en utilisant le paramètre σ^2 si il est connu

soit en utilisant son estimation ponctuelle avec s^2 calculée sur l'échantillon.

Principe général pour le calcul de l'intervalle de confiance

S'intéresser à la différence :

D = L'estimateur - le paramètre

LA CENTRER (- sa moyenne)

ET LA REDUIRE (div. par son écart-type)

$$\text{on obtient} \quad \frac{\text{D - (sa moyenne)}}{\text{son écart type}}$$

CALCUL DE L'INTERVALLE DE CONFIANCE DE LA MOYENNE μ

Différence : (M – μ)

Sa moyenne : E(M- μ)

= E(M) – E(μ)

= E(M) – μ

= μ – μ

= 0

Sa variance : Var(M - μ)

= Var(M) + Var(μ)

= Var(M) + 0

$$= \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

La différence M-μ
suit une loi normale N (0, $\frac{\sigma}{\sqrt{n}}$)

Si σ² connue

$$\frac{(M-\mu)-0}{\frac{\sigma}{\sqrt{n}}}$$

Loi Normale

$$\frac{|M-\mu|}{\frac{\sigma}{\sqrt{n}}} = |U|_{\alpha}$$

On a la probabilité (1- α) pour que $|M-\mu| \leq |U|_{\alpha} \frac{\sigma}{\sqrt{n}}$

Réalisation sur un échantillon :

$$|m-\mu| \leq |u|_{\alpha} \frac{\sigma}{\sqrt{n}}$$

L'estimation par intervalle de μ est :

$$m-|u|_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu \leq m+|u|_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Conditions: M suit une loi normale

σ^2 connue

Exemple

Sur un échantillon de 35 sujets, on a dosé une substance sérique. On observe une moyenne m = 24 UI/l

On sait par la littérature : article de référence

$\sigma^2 = 14\ (UI/l)^2$

Quel est l'intervalle de confiance à 95% de μ ?

(Table: $u_\alpha = 1.96$ pour α=5%)

En appliquant les formules précédentes, on obtient:

$$24 - 1.96\sqrt{\frac{14}{35}} = 22.76\ UI/l$$

$$24 + 1.96\sqrt{\frac{14}{35}} = 25.24\ UI/l$$

Il y a 95 chances sur 100 pour que :

$22.76\ UI/l \leq \mu \leq 25.24\ UI/l$

La différence M-μ
suit une loi normale N (0, $\frac{\sigma}{\sqrt{n}}$)

Si σ² inconnue

$$\frac{(M-\mu)-0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

$$\frac{\hat{\sigma}^2}{n} = \frac{\frac{n}{(n-1)}S^2}{n} = \frac{S^2}{(n-1)}$$

Loi de Student

$$\frac{|M-\mu|}{\frac{S}{\sqrt{n-1}}} = |T|_{\alpha,(n-1)ddl}$$

On a la probabilité (1- α) pour que $|M-\mu| \leq |T|_{\alpha} \frac{S}{\sqrt{n-1}}$

avec (n-1) ddl

Réalisation sur un échantillon :

$$|m-\mu| \leq |t|_{\alpha,(n-1)ddl} \frac{s}{\sqrt{n-1}}$$

L'estimation par intervalle de μ est :

$$m-|t|_{\alpha} \frac{s}{\sqrt{n-1}} \leq \mu \leq m+|t|_{\alpha} \frac{s}{\sqrt{n-1}}$$

avec (n-1) ddl

Conditions: M suit une loi normale

Exemple

Sur un échantillon de 17 sujets, on a dosé une substance sérique. (Loi normale).

On observe

une moyenne m = 24 UI/l et

une variance $s^2 = 19\ (UI/l)^2$

Quel est l'intervalle de confiance à 95% de μ ?

(Table $t_{\alpha,16\ ddl} = 2.12$)

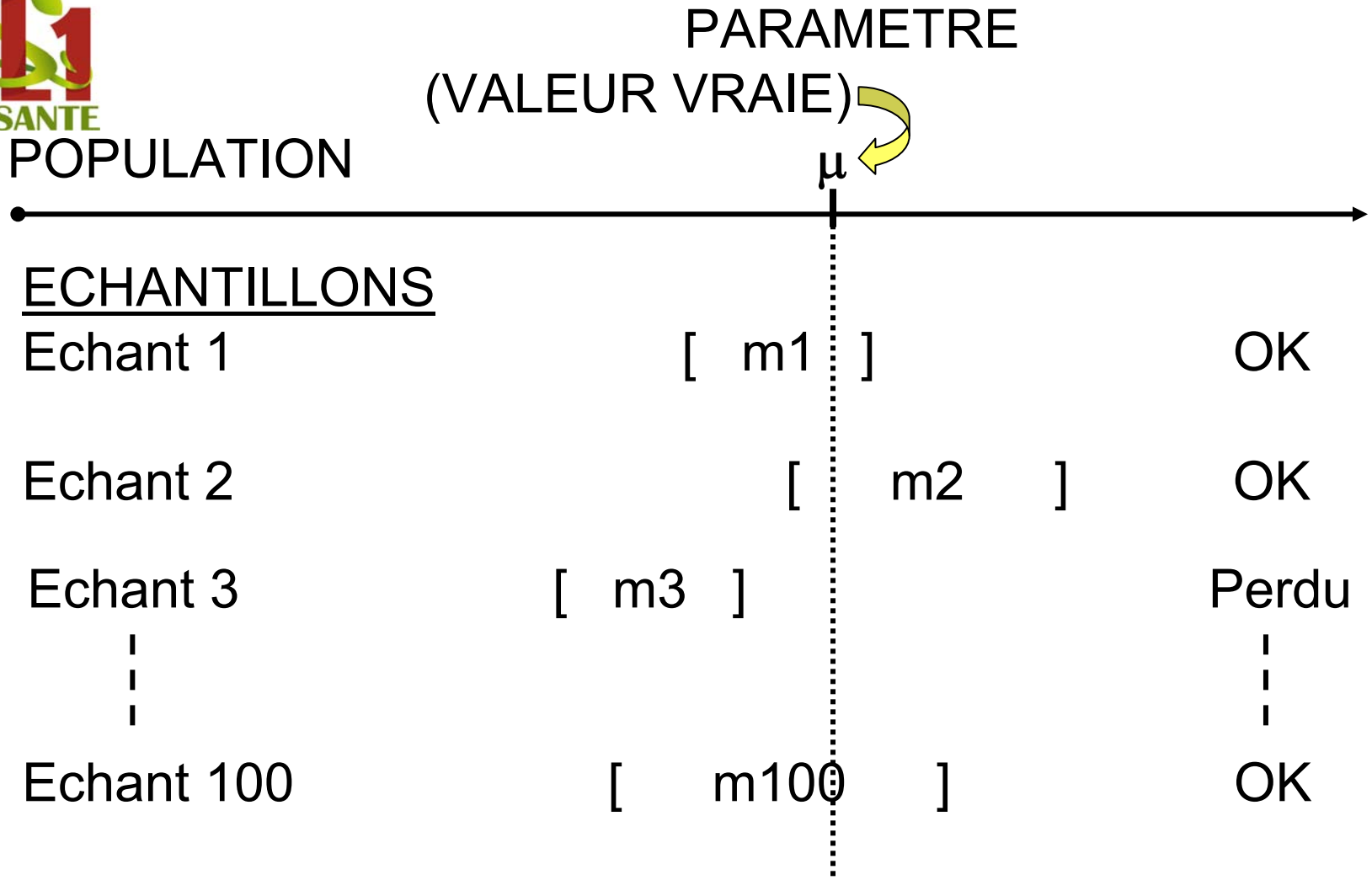
En appliquant les formules précédentes, on obtient:

$$24 - 2.12 \sqrt{\frac{19}{16}} = 21.69\ UI/l$$

$$24 + 2.12 \sqrt{\frac{19}{16}} = 26.31\ UI/l$$

Il y a 95 chances sur 100 pour que :

$$21.69\ UI/l \leq \mu \leq 26.31\ UI/l$$



INTERVALLE DE CONFIANCE IC: Si α=5% il y a 5 intervalles sur 100 qui ne contiendront pas μ

Estimation par intervalle (de confiance) de la proportion Π

Variable qualitative (ex : pile ou face)
Le risque α consenti (5%) donne 95% de confiance

Il faut obtenir
L'estimateur ponctuel du paramètre

($\hat{\Pi}$ = p avec p calculé sur l'échantillon)

La variance de (P - Π) en utilisant son estimateur ponctuel calculé sur l'échantillon

$$Var(p) = \frac{p(1-p)}{n}$$

avec p calculé sur l'échantillon

CALCUL DE L'INTERVALLE DE CONFIANCE DE LA PROPORTION Π

Différence : (P – Π)

Sa moyenne : E(P - Π)

= E(P) – E(Π)

= E(P) – Π

= Π – Π

= 0

Sa variance : Var(P- Π)

= Var(P) + Var(Π)

= Var(P) + 0

$$= \frac{\Pi(1-\Pi)}{n} + 0 = \frac{\Pi(1-\Pi)}{n}$$

La différence (P – Π)
suit une loi normale N (0, $\sqrt{\frac{\Pi(1-\Pi)}{n}}$

$$\frac{(P-\Pi)-0}{\sqrt{\frac{\Pi(1-\Pi)}{n}}}$$

Loi Normale

$$\frac{|P-\Pi|}{\sqrt{\frac{\Pi(1-\Pi)}{n}}} = |U|_{\alpha}$$

On a la probabilité (1- α) pour que

$$|P-\Pi| \leq |U|_{\alpha} \sqrt{\frac{\Pi(1-\Pi)}{n}}$$

Réalisation sur un échantillon :

$$\left|p-\Pi\right| \leq \left|u\right|_{\alpha}\sqrt{\frac{p(1-p)}{n}}$$

L'estimation par intervalle de Π est :

avec

Borne inférieure (Binf.) = $p-\left|u\right|_{\alpha}\sqrt{\frac{p(1-p)}{n}}$

$p+\left|u\right|_{\alpha}\sqrt{\frac{p(1-p)}{n}}$ = Borne supérieure (Bsup.)

Conditions: n=effectif, Binf. et Bsup. bornes de l'intervalle

n Binf. ≥ 5 ; n(1-Binf.) ≥ 5; n Bsup. ≥ 5; n(1-Bsup.) ≥ 5

Exemple

Sur un échantillon de 60 sujets on a observé 25 sujets porteurs du génotype G29.

p = 41.7 %

Quel est l'intervalle de confiance à 95% de Π ?

En appliquant les formules précédentes, on obtient:

$$0.417 - 1.96\sqrt{\frac{0.417(1-0.417)}{60}} = 0.29$$

$$0.417 + 1.96\sqrt{\frac{0.417(1-0.417)}{60}} = 0.54$$

Il y a 95 chances sur 100 pour que : 29 % ≤ Π ≤ 54 %

Avec 60×0.29 et 60×0.71et 60×0.54 et 60×0.46 tous ≥ 5

ATTENTION DANS LA LITTÉRATURE
$m \pm ?$

S'agit-il de

$m \pm SD$?

$m \pm SE$?

$IC = m \pm |u_\alpha| \times SE$? ($|u_\alpha|$ ou $|t_\alpha|$)

La différence n'est pas négligeable

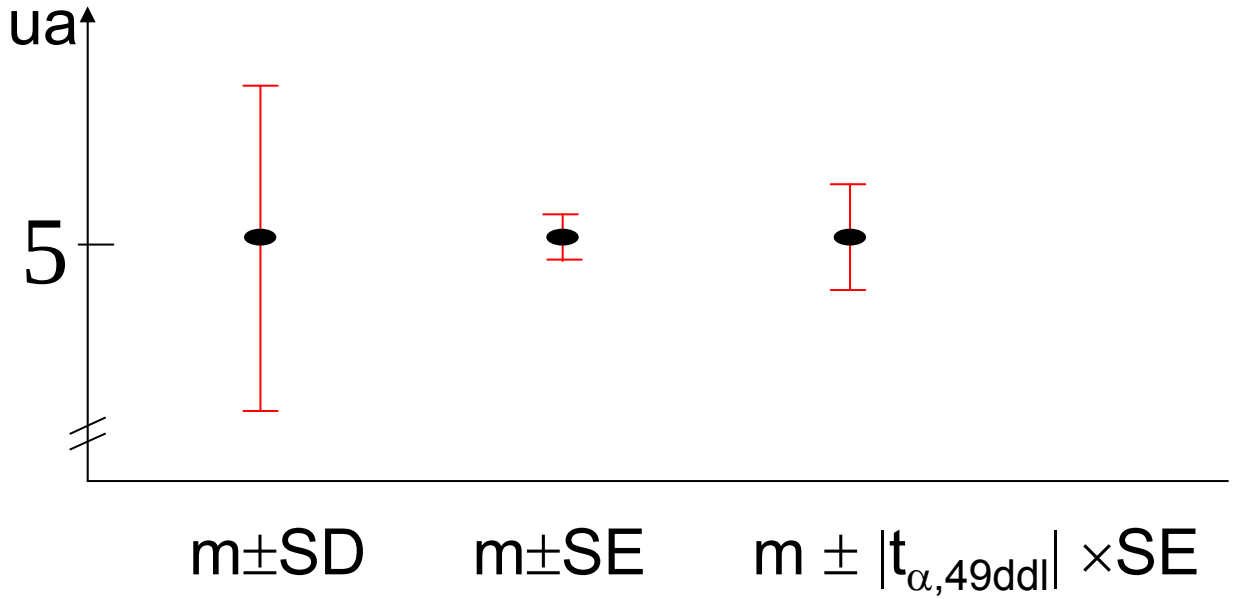
ATTENTION DANS LA LITTÉRATURE
m ± ?

Exemple : m = 5 ua n = 50 s^2 = 30 ua^2

$m \pm SD \rightarrow 5 \pm 5{,}48$ (en ua)

$m \pm SE \rightarrow 5 \pm 0{,}78$ (en ua)

$m \pm |t_{\alpha,49ddl}| \times SE \rightarrow 5 \pm 1{,}53$ (en ua)



ATTENTION DANS LA LITTÉRATURE
m ± ?

A noter que m ± SD donne 5 ± 5,48

soit

borne inf.	-0.48
borne sup.	10.48

Une valeur négative est-elle possible /contexte ?

Loi normale ?

Conséquences dans les études cliniques

Lecture critique d'articles médicaux

Quelle taille pour l'échantillon ?

Calcul du nombre de sujets nécessaires.

POPULATION

Problèmes:
Hétérogénéité
Variabilité

ECHANTILLON

EFFECTIF SUFFISANT ?

LE NOMBRE DE SUJETS NECESSAIRES (n)
= f (VARIABILITE, …)

SI VARIABILITE ↗ ALORS **n** ↗

Exemple: **Nombre de sujets nécessaires pour avoir une précision voulue**

But : estimer la moyenne μ du dosage de la protéine A chez des sujets diabétiques avec une précision de ± 0.10 et avec une confiance de 95% ?
On connaît $\sigma^2 = 2$ UI²
(loi normale)
Taille de l'échantillon ?

$$|m - \mu| \leq |u|_\alpha \frac{\sigma}{\sqrt{n}} \leq i$$

Précision

$$u_\alpha^2 \frac{\sigma^2}{n} \leq i^2$$

$$n \geq u_\alpha^2 \frac{\sigma^2}{i^2}$$

ex:

$$n \geq 1.96^2 \frac{2}{0.01}$$

$$n \approx 769 sujets$$

Nancy-Université
Université Henri Poincaré

ministère éducation nationale enseignement supérieur recherche