Estimation d'un paramètre

Pr BOUZBID S

Prérequis

Connaître

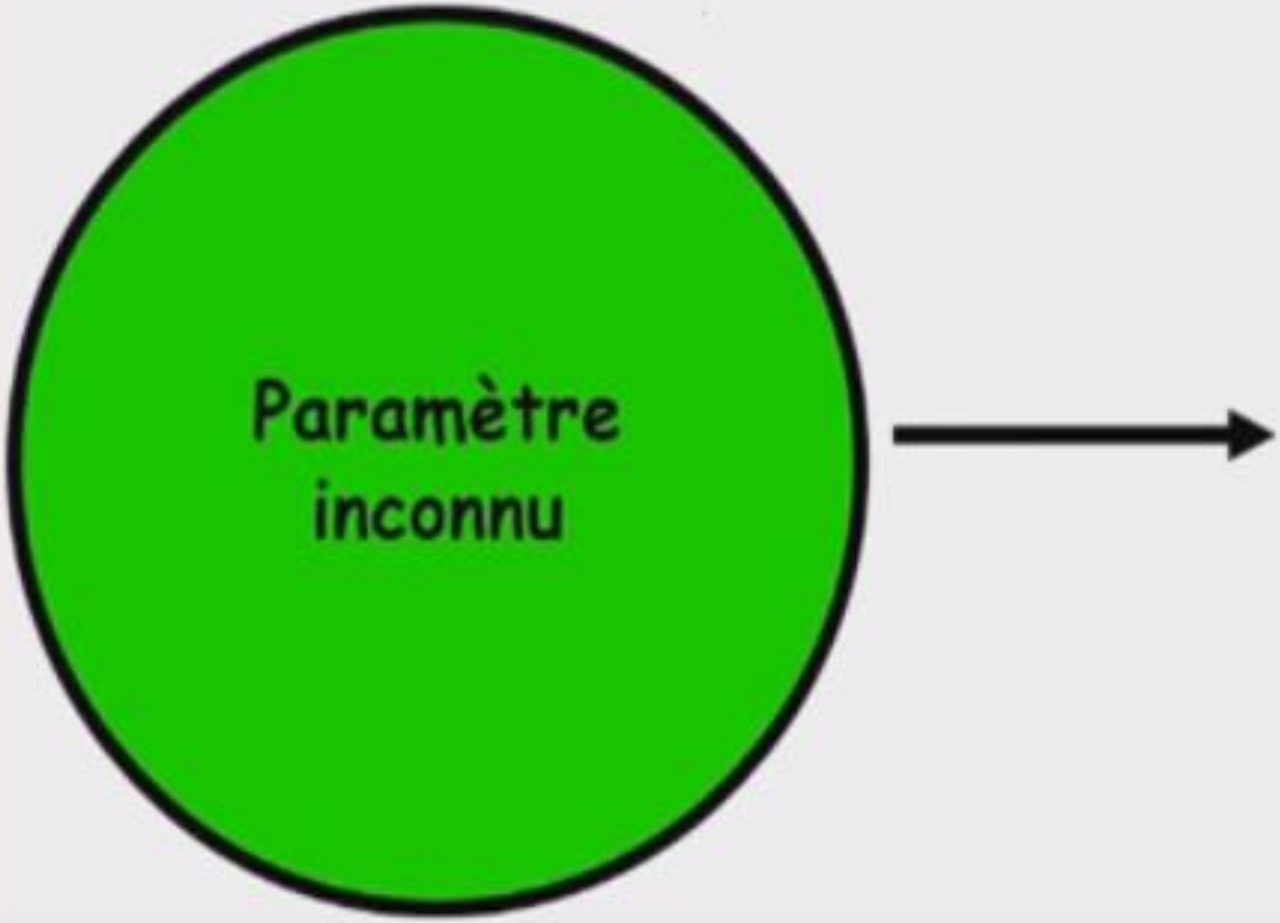
- les paramètres descriptifs d'une population :

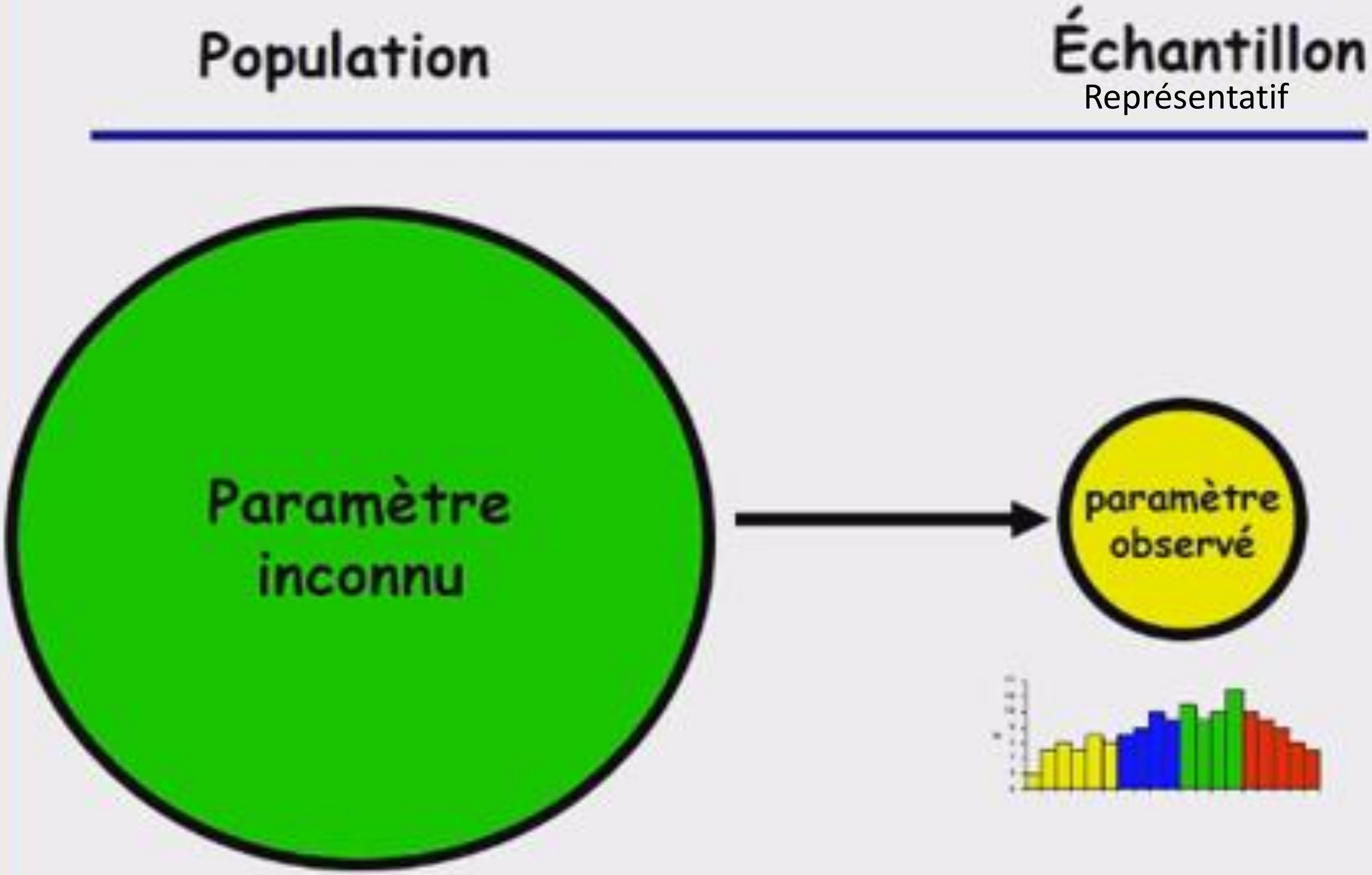
Paramètres de position
- pourcentage
- moyenne

Paramètres de dispersion
- variance
- écart-type

- la loi normale et ses propriétés

Population





Estimer sans trop se tromper les paramètres de la population à partir d'un échantillon

Objectifs du cours

A l'issue du cours, l'auditeur devrait être capable de :

- d'estimer un pourcentage
- d'estimer une moyenne
- de se fixer un risque d'erreur pour réaliser une estimation par intervalle de confiance

Exemple 1

Étude des troubles du sommeil chez les enfants de 2 à 3 ans en Isère

Population : 14 000 enfants de 2 à 3 ans.

Échantillon : 540 enfants.

dont 86 présentaient des troubles du sommeil

Mesure des paramètres sur l'échantillon :

1) Pourcentage d'enfants avec troubles du sommeil : $p_o = 16\%$

Quel est le vrai pourcentage de troubles du sommeil dans la population des enfants de 2 à 3 ans en Isère ?

Exemple 2

Étude des troubles du sommeil chez les enfants de 2 à 3 ans en Isère

Population : 14 000 enfants de 2 à 3 ans.

Échantillon : 540 enfants.

Mesure des paramètres sur l'échantillon :

2) durée *moyenne* du temps de sommeil : m = 11,7 heures

Problème : comment estimer la moyenne
du temps de sommeil des enfants de 2 à 3 ans dans la population
à partir des résultats observés sur l'échantillon ?

Problématique

Exemple 1

- Trouble du sommeil
- Unité : *Oui/Non*
- Variable dichotomique
- Paramètre : pourcentage

Problématique

Exemple 1	Exemple 2
• Trouble du sommeil	• Durée de sommeil
• Unité : *Oui/Non*	• Unité : heures
• Variable dichotomique	• Variable quantitative
• Paramètre : pourcentage	• Paramètre : moyenne

Comment estimer un paramètre inconnu

à partir d'un paramètre observé sur un échantillon ?

Échantillon

Sous-groupe de la population d'étude
dont tous les individus ont été tirés au sort

Échantillon aléatoire simple

Tous les individus de la population d'étude ont la même probabilité d'être tirés.

Toutes les formules, s'appliquent pour le type d'échantillonage aléatoire simple.

Paramètres à estimer

Variable qualitative binaire :   pourcentage

Variable binaire

Paramètre inconnu dans la population :

pourcentage P

Paramètre observé dans l'échantillon :

pourcentage p_o

Question : comment estimer P en connaissant p_o ?

Fluctuation d'échantillonnage d'un pourcentage

p_1 p_3 p_2

P ?

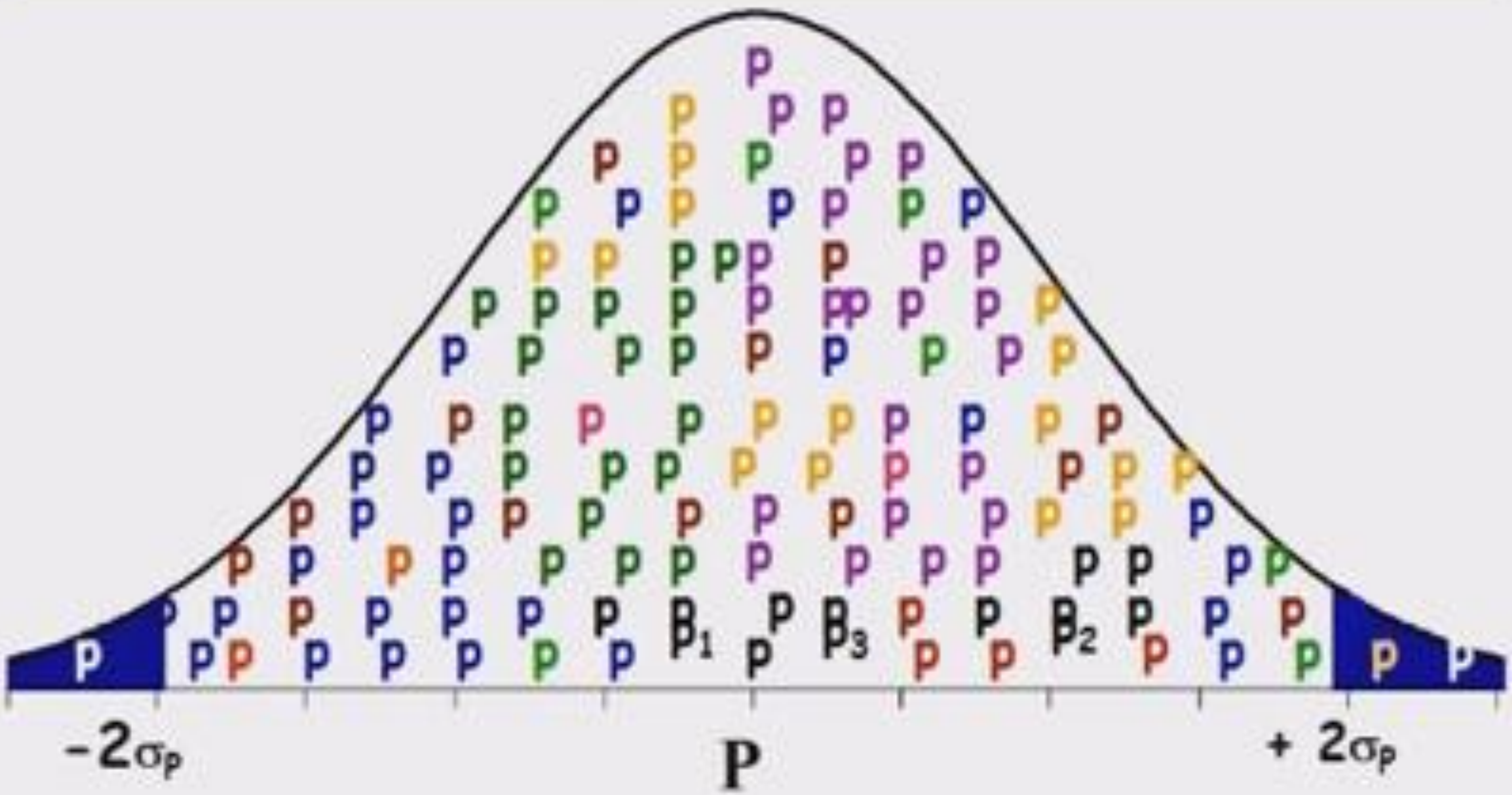
Fluctuation d'échantillonnage d'un pourcentage

Théorème central limite

Un pourcentage p observé sur un échantillon de taille n

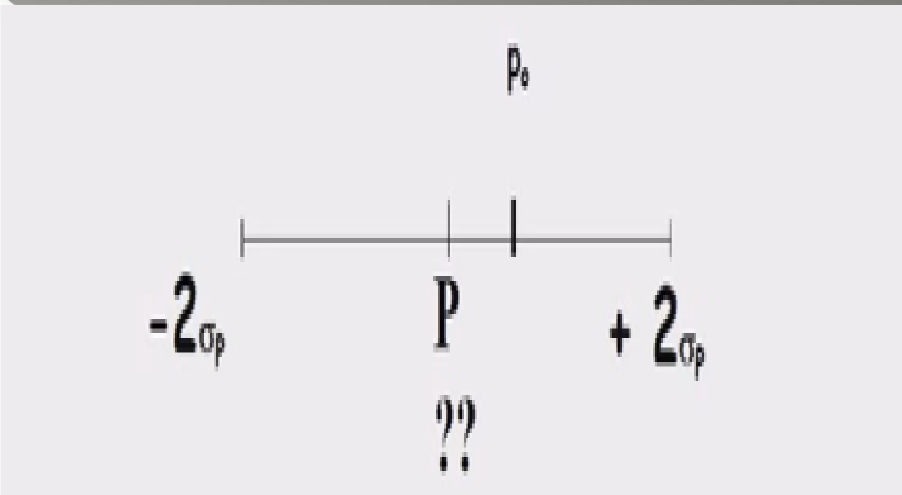
- est une variable aléatoire
- suivant une loi normale
- centrée sur le pourcentage P

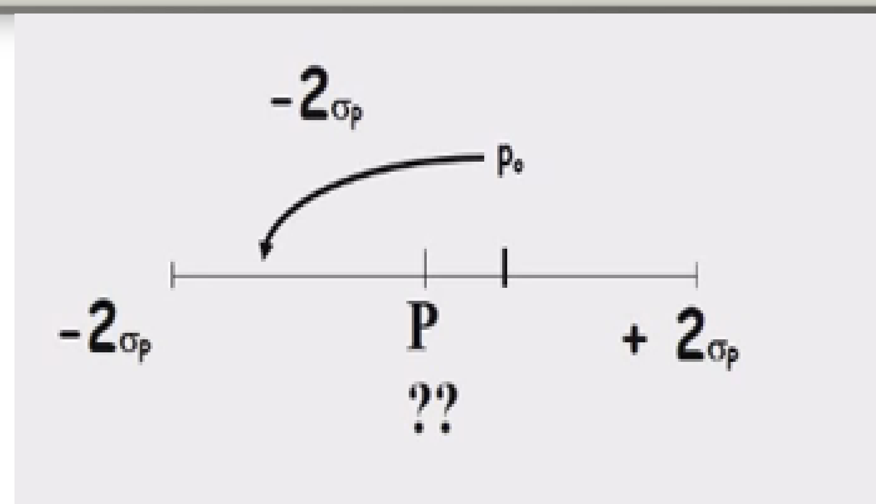
Fluctuation d'échantillonnage d'un pourcentage

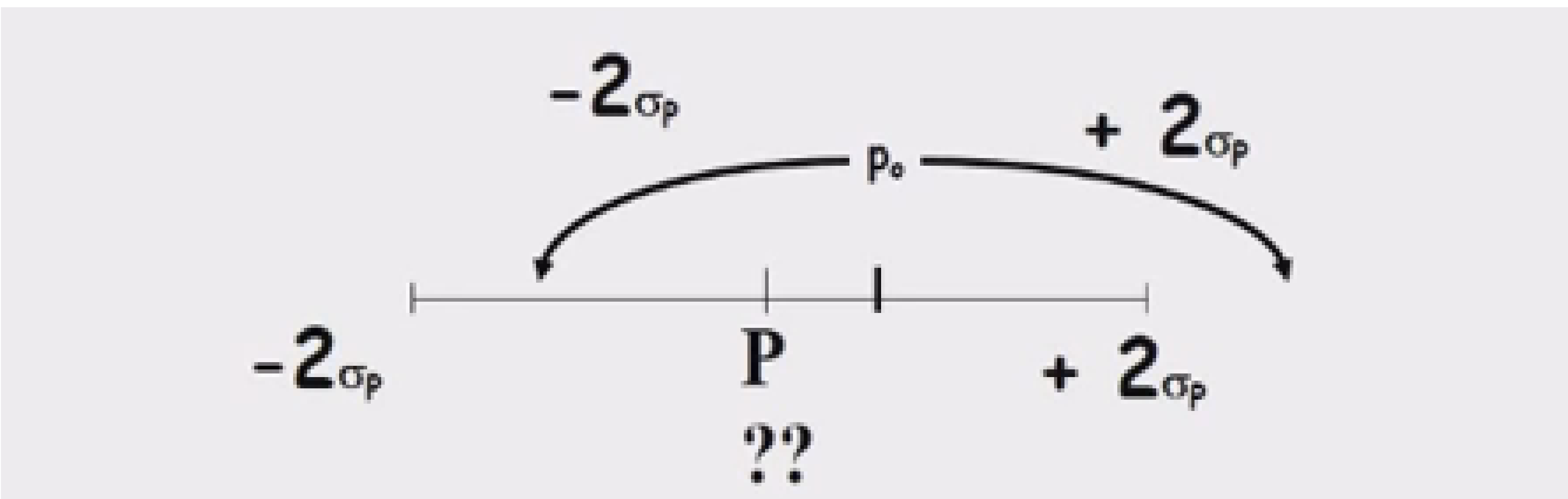


95 % des valeurs de p_o sont comprises entre -2 et +2 écarts type de P

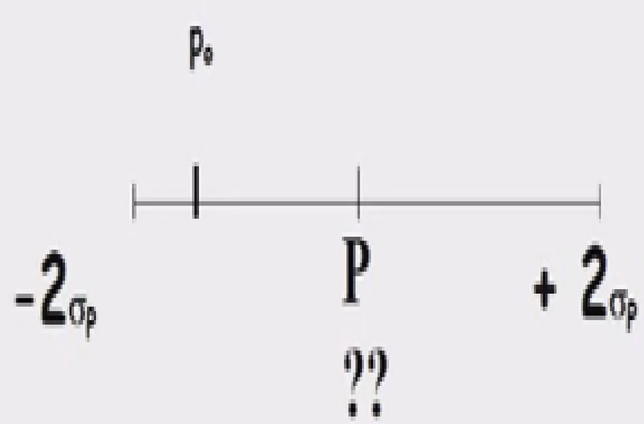
Que faire avec une seule observation p_o ?

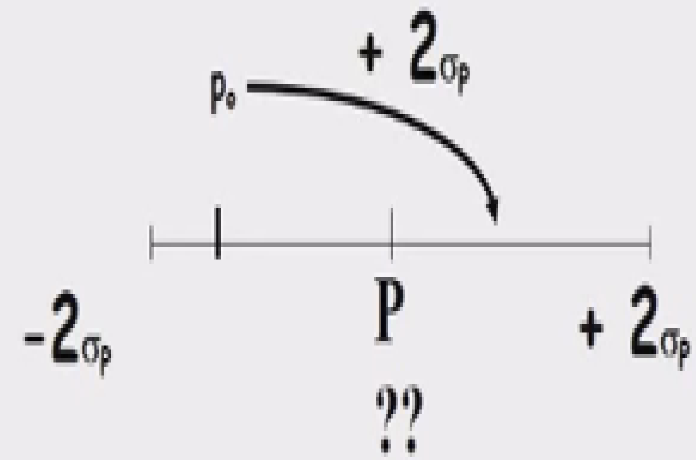


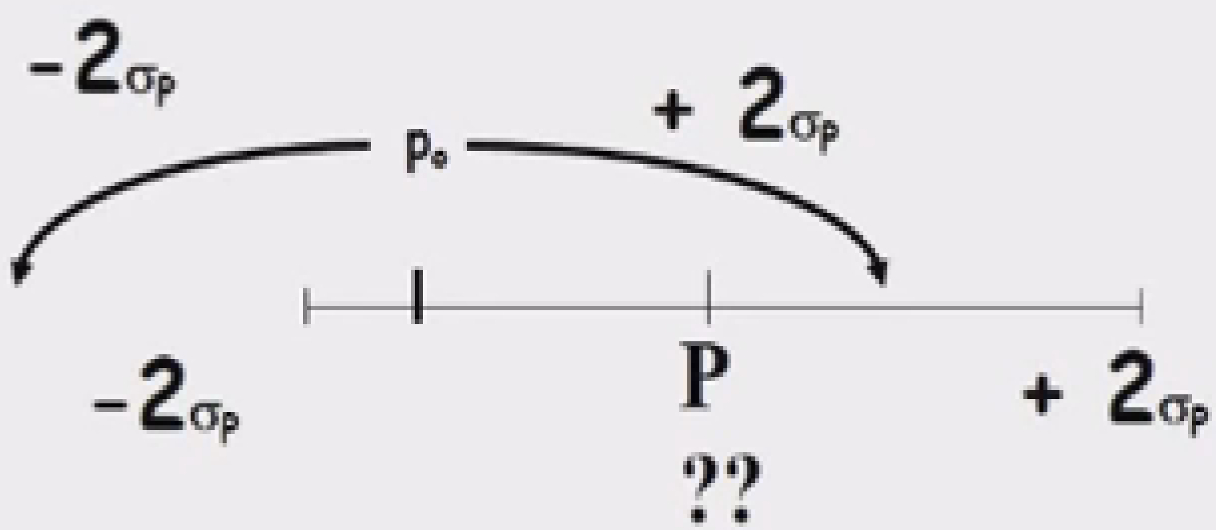




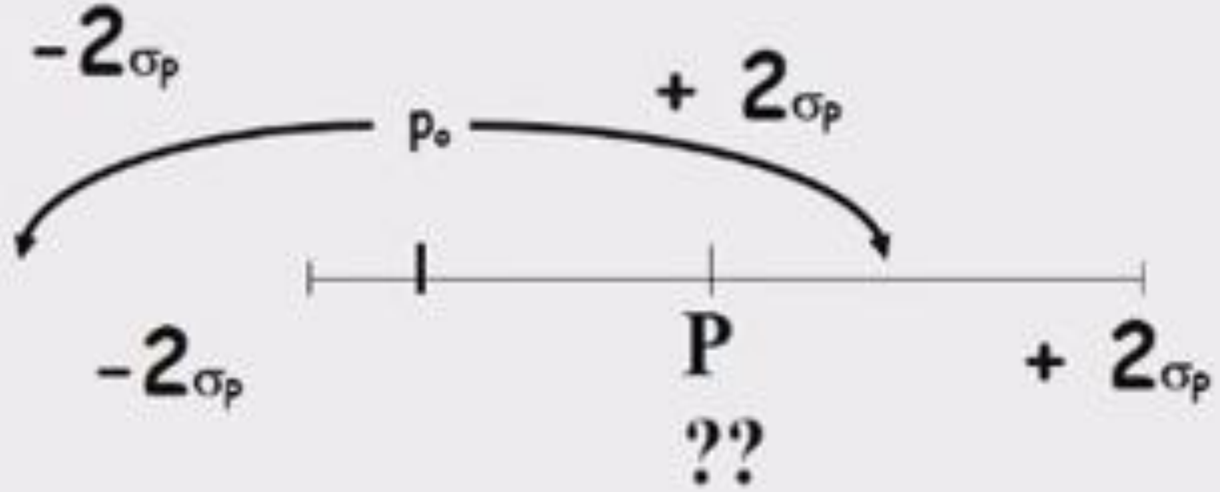
Que faire avec une seule observation p_o ?







Que faire avec une seule observation p_o ?



$$p_o - 2\sigma_p < P < p_o + 2\sigma_p$$

dans 95 % des cas

$$p_o - 2\sigma_p < P < p_o + 2\sigma_p$$

écart type de P ????

$$\sigma_p \approx \sqrt{\frac{p_o\,(1 - p_o)}{n}}$$

Bon estimateur

n : taille de l'échantillon

p_o : pourcentage observé

Intervalle de confiance à 95% d'un pourcentage

$$p_o - 2\sqrt{\frac{p_o(1-p_o)}{n}} < \mathbf{P} < p + 2\sqrt{\frac{p_o(1-p_o)}{n}}$$

Intervalle de confiance à 95% d'un pourcentage

$$p_o - 2\sqrt{\frac{p_o(1-p_o)}{n}} < \mathbf{P} < p + 2\sqrt{\frac{p_o(1-p_o)}{n}}$$

$$\mathbf{P} = p_o \pm 2\sqrt{\frac{p_o(1-p_o)}{n}}$$

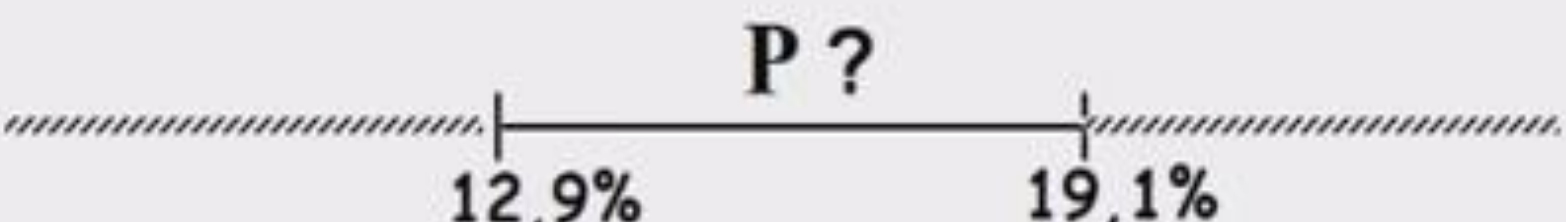
Ex 1

Etude des troubles du sommeil chez les enfants de 2 à 3 ans en Isère

Pourcentage d'enfants avec troubles du sommeil : $p_o = 16\%$
Taille de l'échantillon n = 540

$$IC95\% : \left[0{,}16 - 2\sqrt{\frac{0{,}16 \times 0{,}84}{540}} \;;\; 0{,}16 + 2\sqrt{\frac{0{,}16 \times 0{,}84}{540}} \right]$$

IC95% : [0,129 ; 0,191]



Conditions d'application du calcul de l'intervalle de confiance d'un pourcentage

1) $np_{inf} >= 5$ ET $n(1-p_{sup}) >=5$

Condition d'application (1)

Pourcentage d'enfants avec troubles du sommeil : p = 16%
Taille de l'échantillon n = 540

IC95% : [0,129 ; 0,191]

$\downarrow$ pi $\quad$ $\downarrow$ ps

npi = 540 x 0,129 = 69,7 (> 5)

n(1-ps) = 540 x (1 - 0,191) = 103,1 (> 5)

Conditions d'application du calcul de l'intervalle de confiance d'un pourcentage

1) $np_{inf} >= 5$ ET $n(1-p_{sup}) >=5$

sinon utiliser méthode binomiale exacte

- tables (J. Bouyer)
- logiciels

Conditions d'application du calcul de l'intervalle de confiance d'un pourcentage

1) $np_{inf} >= 5$ ET $n(1-p_{sup}) >=5$

2) Taille de l'échantillon n négligeable par rapport à la taille de la population N

$n/N < 10\%$

Condition d'application (2)

Taille de la population $N = 14\ 000$

Taille de l'échantillon $n = 540$

$n/N = 540 / 14\ 000 = 0{,}039 = 3{,}9\% < 10\%$

Lorsque la taille de l'échantillon n est grande par rapport à la taille N population

$$n/N > 10\%$$

$$\sigma_p \approx \sqrt{\frac{p_0\,(1-p_0)}{n}}\;\boxed{\sqrt{\frac{N-n}{N-1}}}$$

facteur d'exhaustivité

Formule approchée

N, négligeable /n, d'où N/D=1

Lorsque la taille de l'échantillon n est grande par rapport à la taille N population

$$n/N > 10\%$$

$$\sigma_p \approx \sqrt{\frac{p_o \left(1 - p_o\right)}{n}} \boxed{\sqrt{\frac{N - n}{N - 1}}}$$

facteur d'exhaustivité

Quand n tend vers N

$$p \approx P$$

I.C. tend vers 0


P

Exemple 2

Étude des troubles du sommeil chez les enfants de 2 à 3 ans en Isère

Moyenne de la durée du sommeil	m = 11,7 heures
Écart-type	s = 1,3 heures
Taille de l'échantillon	n = 540

Paramètres à estimer

Variable qualitative binaire : pourcentage

Variable quantitative :	moyenne

Variable quantitative

Paramètre inconnu dans la population :

moyenne μ
écart-type σ

Paramètre observé dans l'échantillon :

moyenne m
écart-type s

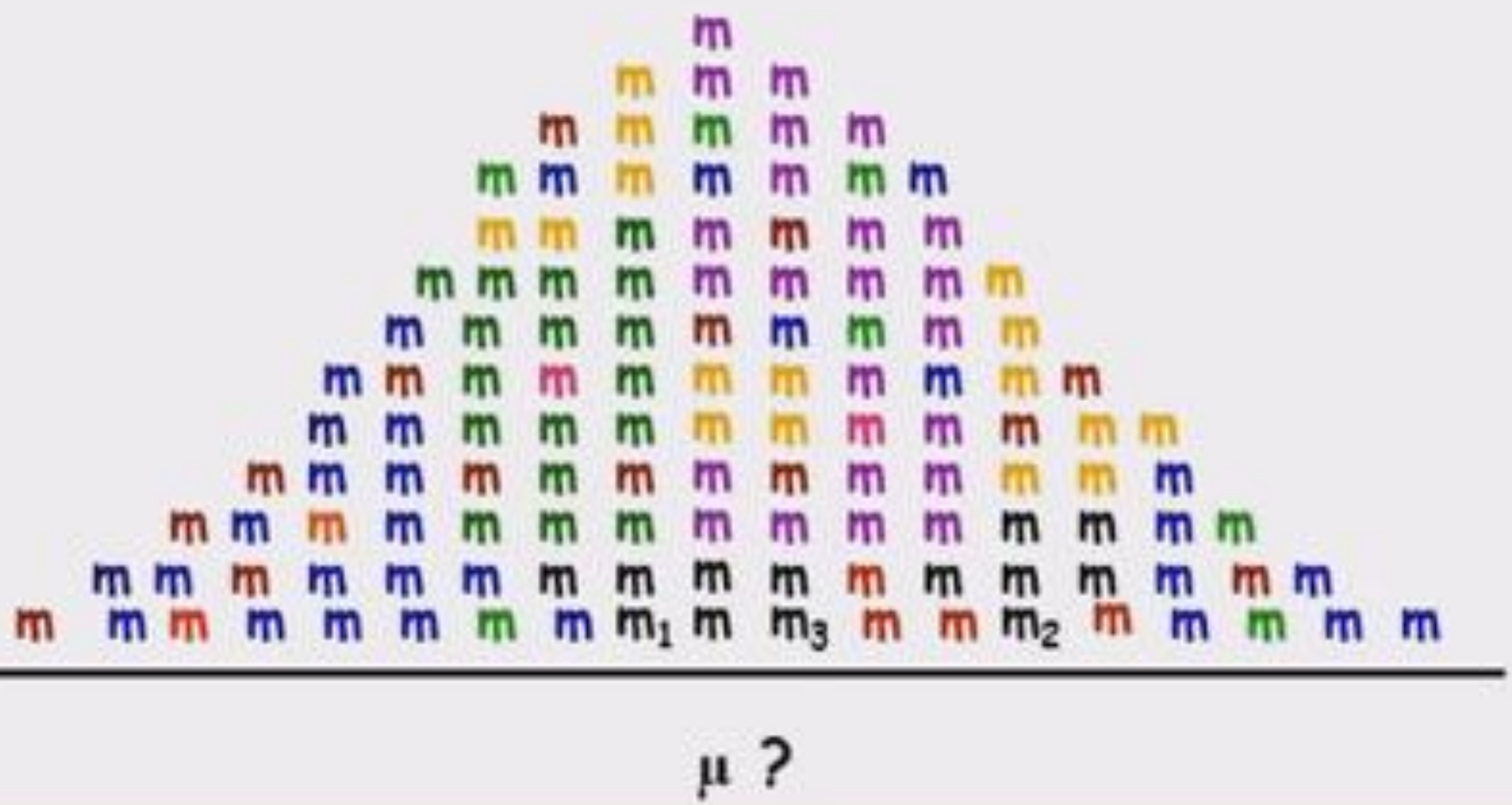
Question : comment estimer μ en connaissant m ?

Fluctuation d'échantillonnage d'une moyenne

m_1 m_3 m_2

μ ?

Fluctuation d'échantillonnage d'une moyenne

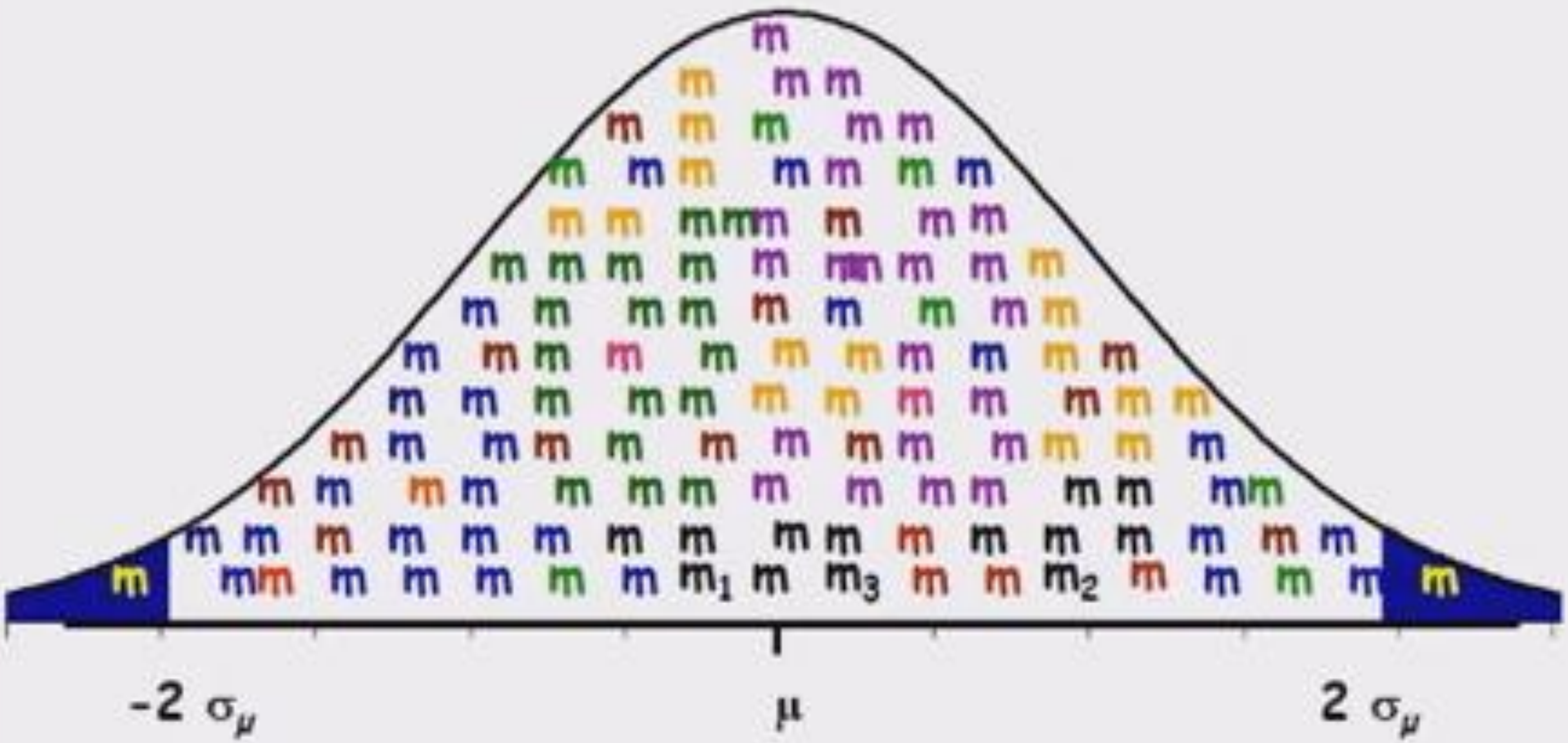


Théorème central limite

La moyenne d'un échantillon

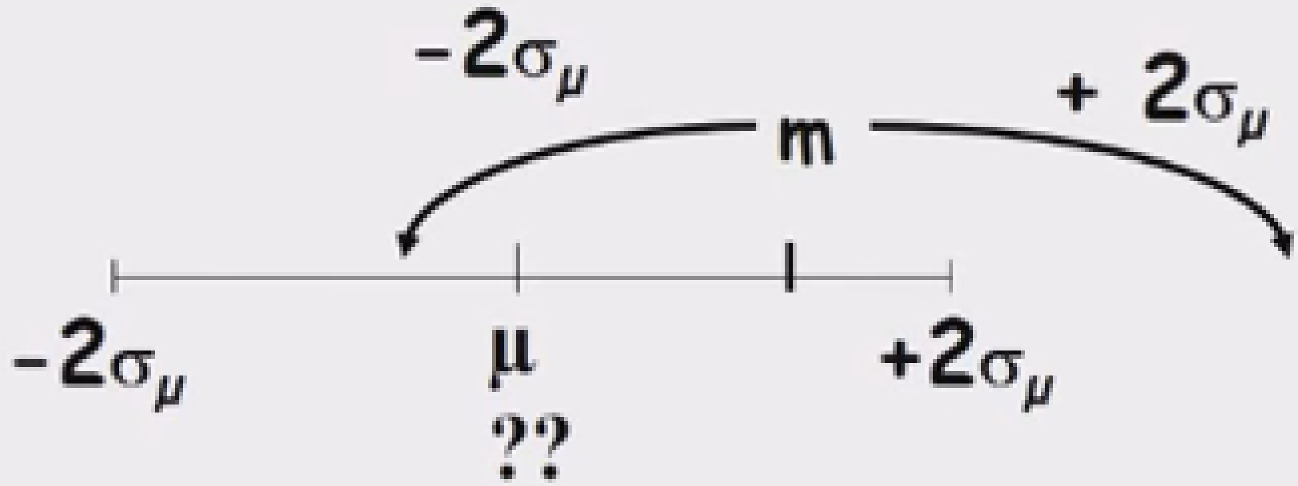
- est une variable aléatoire
- suivant une loi normale
- centrée sur la moyenne μ

Fluctuation d'échantillonnage d'une moyenne

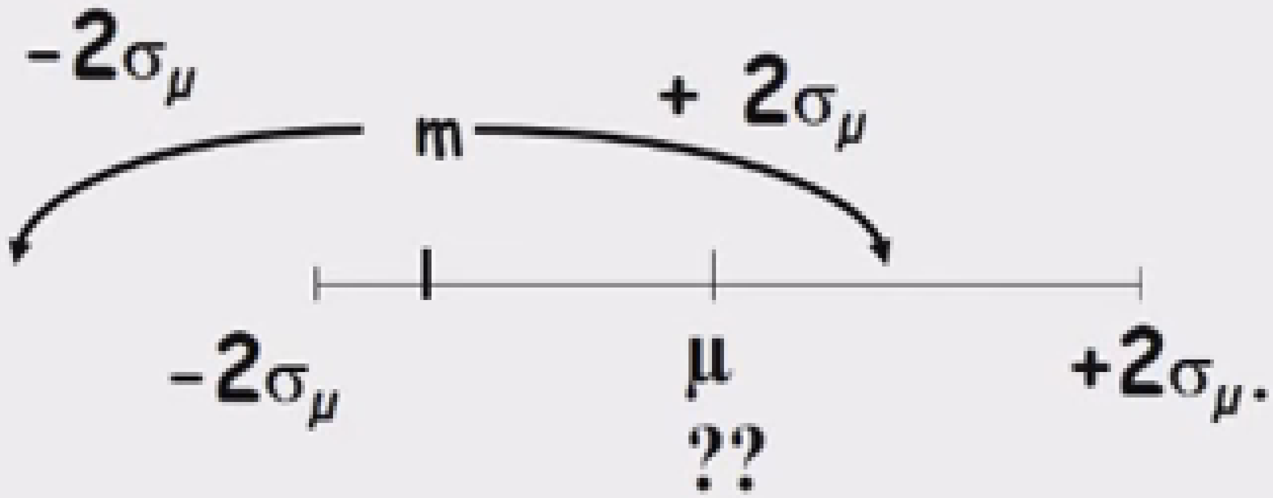


95 % des valeurs de m sont comprises entre -2 et +2 écarts type de μ

Que faire avec une seule observation m ?



Que faire avec une seule observation m ?



$$m - 2\sigma_\mu \quad < \quad \mu \quad < \quad m + 2\,\sigma_\mu$$

$$m - 2\,\sigma_\mu \quad < \quad \mu \quad < \quad m + 2\,\sigma_\mu$$

$\sigma_\mu \equiv$ écart type de μ ????

$$m - 2\,\sigma_\mu \quad < \quad \mu \quad < \quad m + 2\,\sigma_\mu$$

σ_μ = écart type de μ ????

$$\sigma_\mu \approx \text{ standard error } = \frac{s}{\sqrt{n}}$$

s : écart type de l'échantillon
n : taille de l'échantillon

$$m - 2\,\sigma_\mu \quad < \quad \mu \quad < \quad m + 2\,\sigma_\mu$$

σ_μ = écart type de µ ????

$$\sigma_\mu \approx \text{standard error} = \frac{s}{\sqrt{n}}$$

s : écart type de l'échantillon
n : taille de l'échantillon

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

Echantillon

$$s = \sqrt{\frac{\sum (x - m)^2}{N}}$$

population

Calcul d'un écart-type s avec Excel

fonction Excel : =ECARTYPE(A1:A9)

	A			
1	2			
2	7			
3	8		2,76	divisé par n-1
4	7			
5	9	ECARTYPEP :	2,60	divisé par N
6	1			
7	6			
8	8			
9	7			

Intervalle de confiance à 95% d'une moyenne

$$m - 2\frac{s}{\sqrt{n}} \quad < \quad \mu \quad < \quad m + 2\frac{s}{\sqrt{n}}$$

$$\mu = m \pm 2 \frac{s}{\sqrt{n}}$$

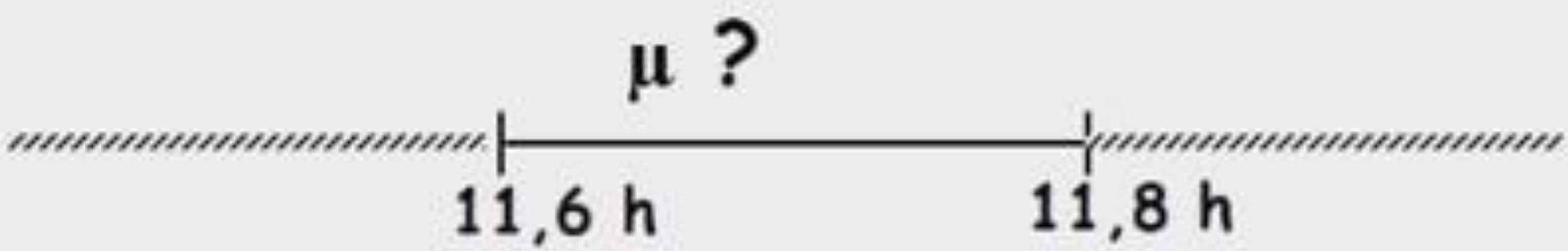
Exemple 2

Étude des troubles du sommeil chez les enfants de 2 à 3 ans en Isère

Moyenne de la durée du sommeil	m = 11,7 heures
Écart-type	s = 1,3 heures
Taille de l'échantillon	n = 540

$$\text{IC } 95\% : \left[11{,}7 - 2\,\frac{1{,}3}{\sqrt{540}}\ ;\quad 11{,}7 + 2\,\frac{1{,}3}{\sqrt{540}}\right]$$

$$\text{IC } 95\% : \quad [\ 11{,}6\ ;\ 11{,}8\ ]$$



Conditions d'application du calcul de l'intervalle de confiance d'une moyenne

- Taille de l'échantillon $n > 30$
- Taille de l'échantillon n négligeable par rapport à la taille de la population N ($n/N < 10\%$)

Lorsque la taille de l'échantillon n est grande par rapport à la taille N population

$$n/N > 10\%$$

$$\sigma_\mu \approx \frac{s}{\sqrt{n}} \boxed{\sqrt{\frac{N - n}{N - 1}}}$$

facteur d'exhaustivité

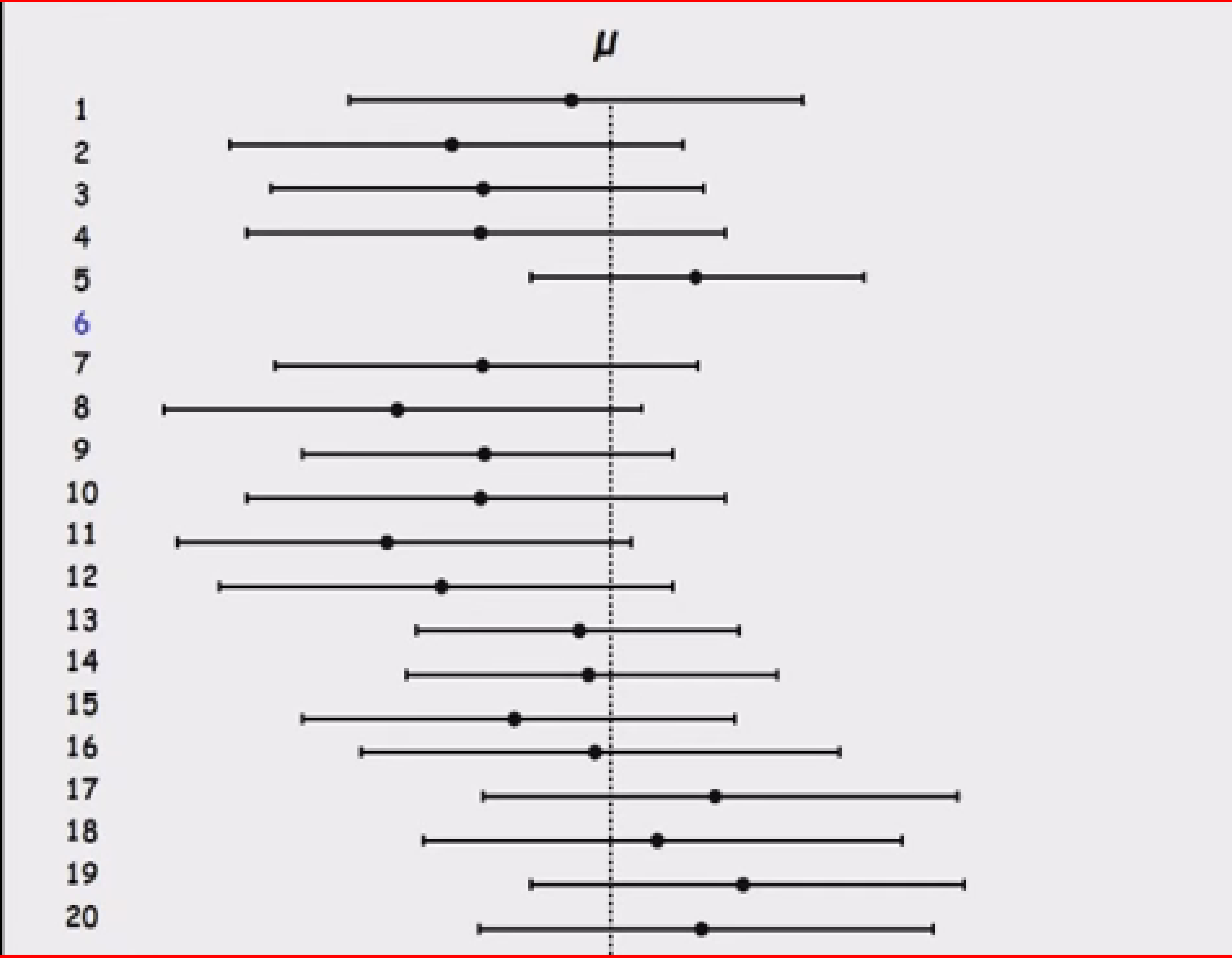

μ

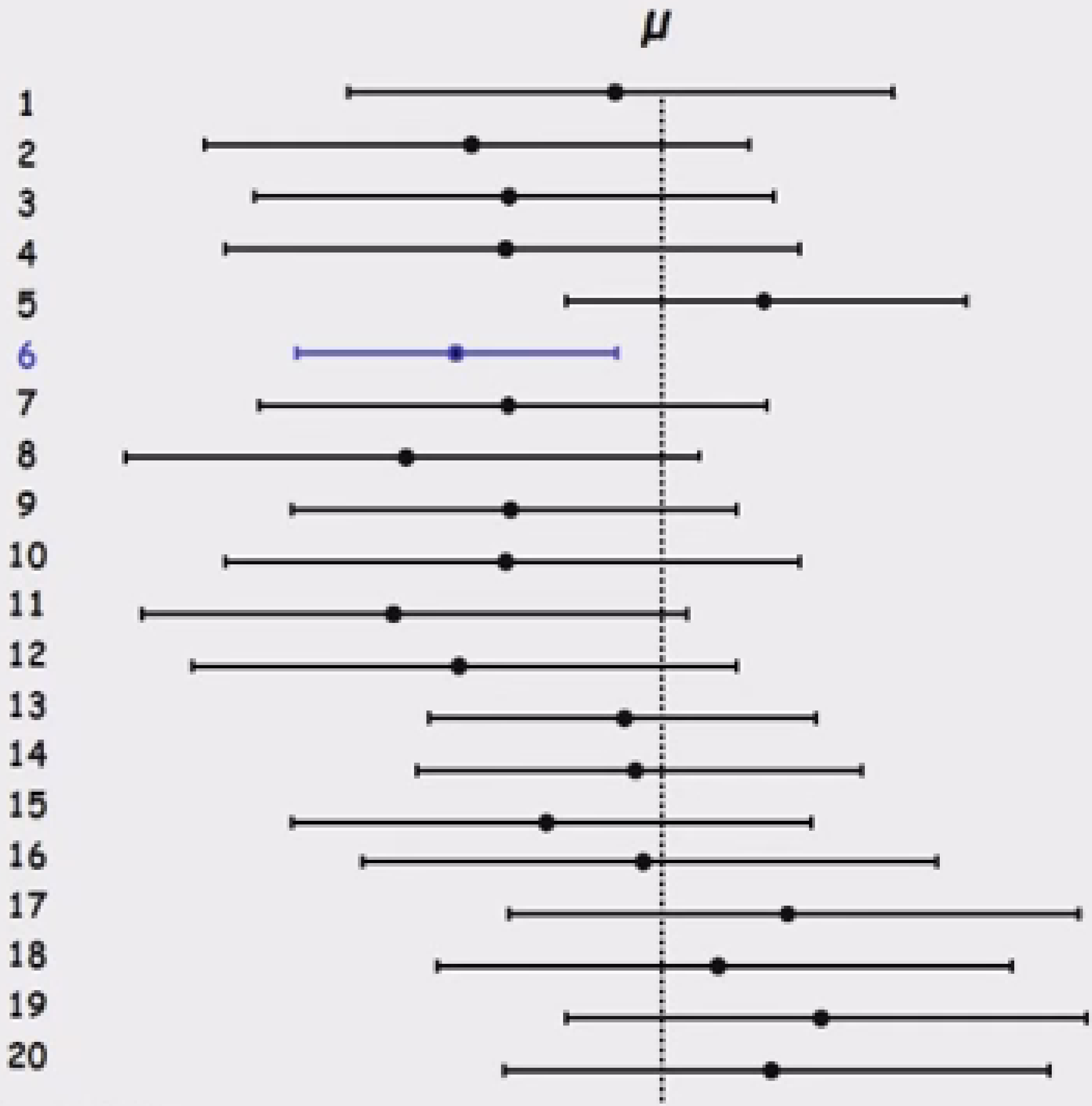
Signification d'un intervalle de confiance à 95%

- Il y a 95 chances sur 100 pour que le paramètre inconnu μ ou P soit contenu dans cet intervalle.

- Il y a 5 chances sur 100 pour que le paramètre inconnu μ ou P soit extérieur à cet intervalle.

- Il y a donc 5 chances sur 100 de se tromper.


μ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

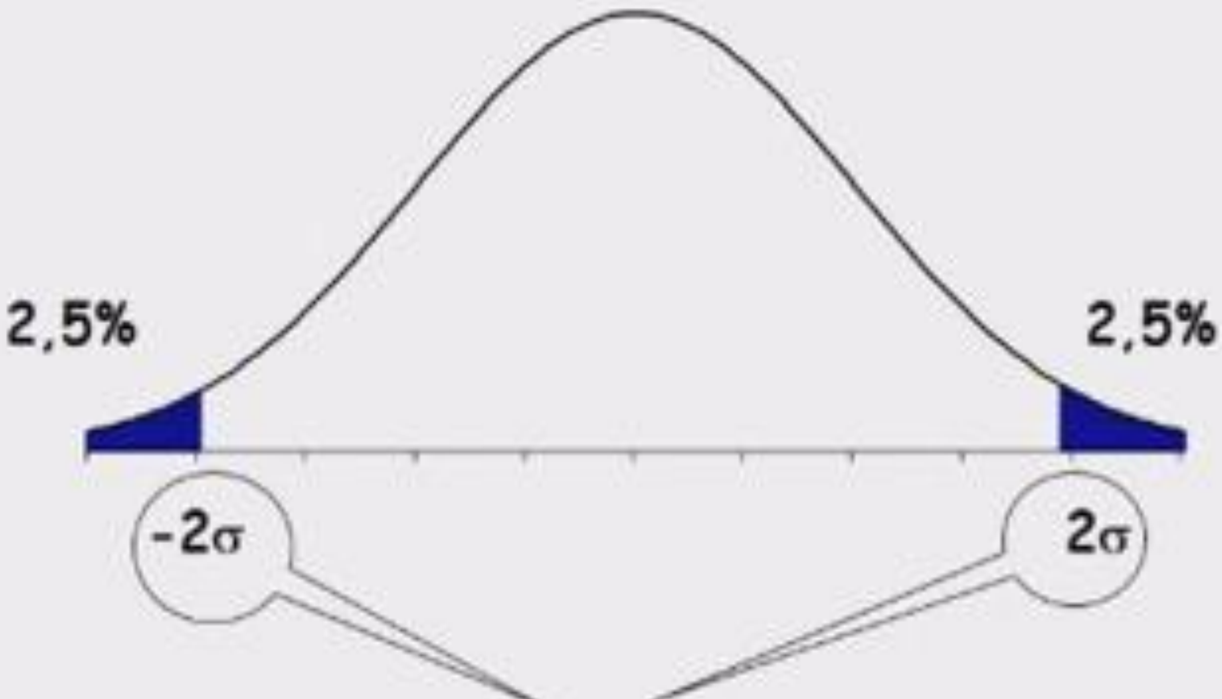

μ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Le risque d'erreur

lorsqu'on estime un paramètre

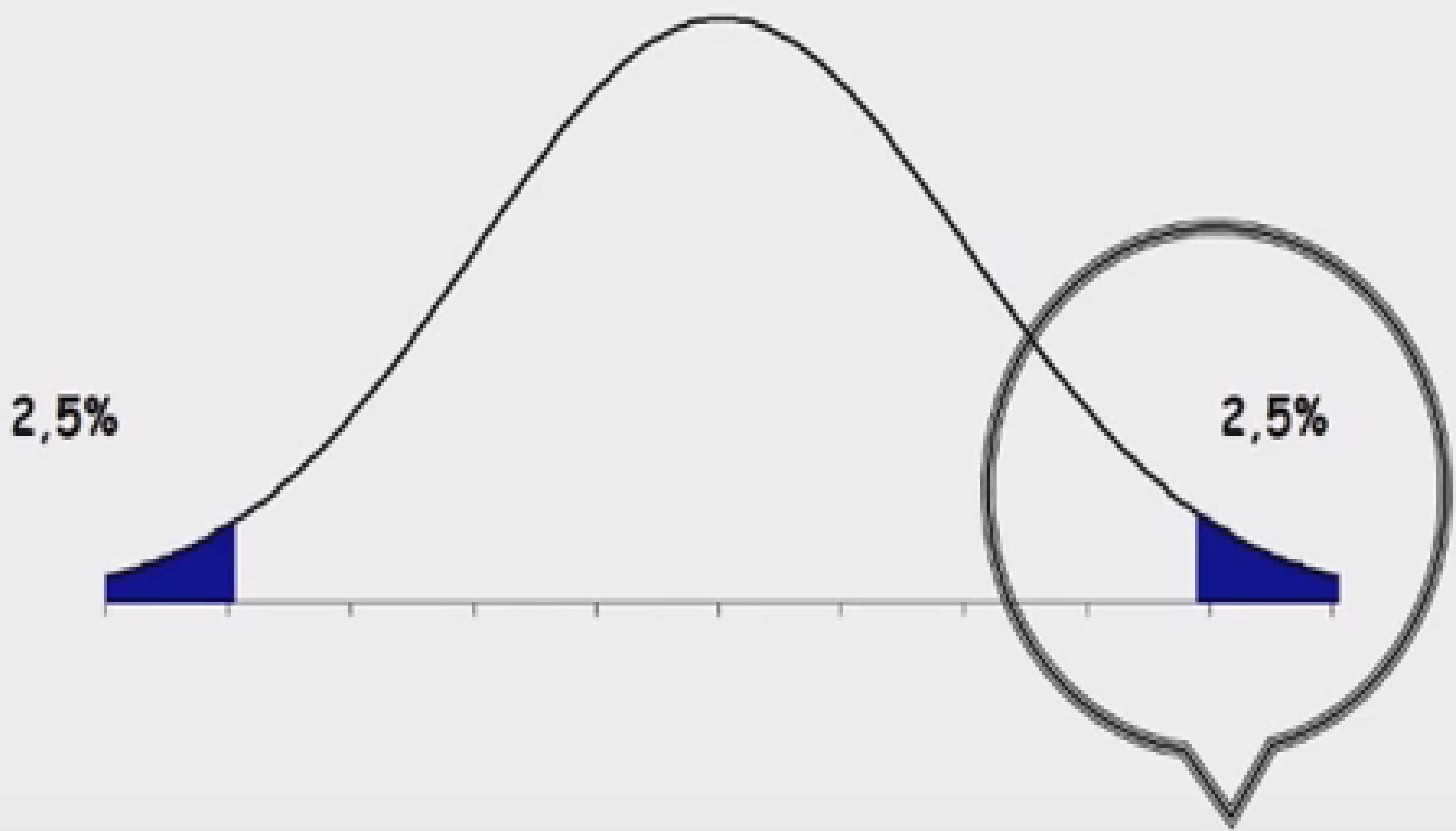
Le risque consenti α

- On appelle risque α le risque d'erreur consentie
- Nous avons pris jusqu'à maintenant α=5%:   I.C. à 95%
- Car nous avions choisi la valeur seuil ± 2 écarts type

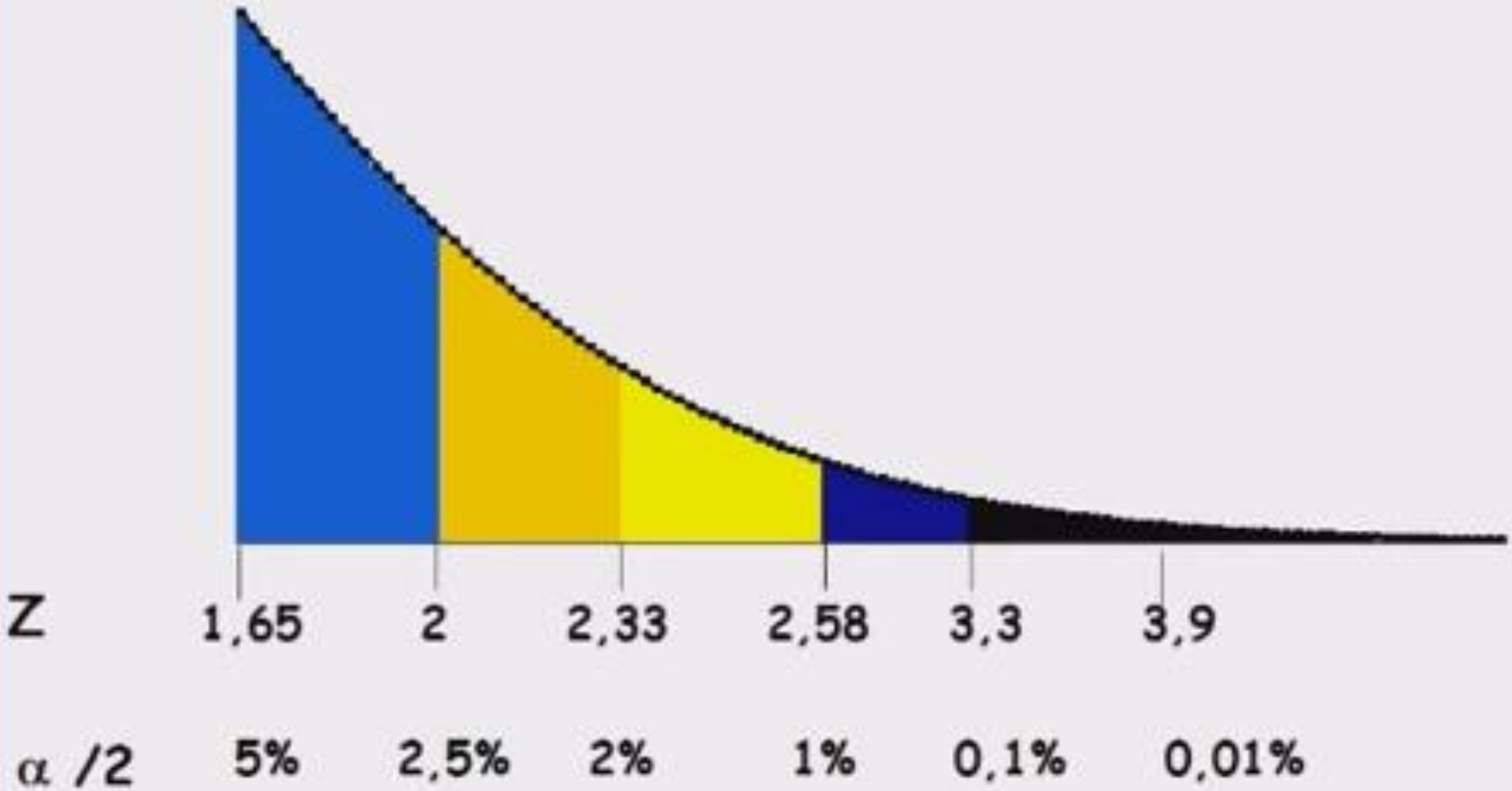


Valeur seuil choisie $Z_{5\%} = 2$

Valeur z de la loi normale réduite


2,5%
2,5%

Valeurs de Z pour quelques risques usuels



Ça dépendra du risque d'erreur que je choisis
Plus Z est élevé plus le risque est faible

Table de Z (risque alpha bilatéral)

lpha	20%	10%	5%	4%	3%	2%	1%	0,05%	0,01%	0,001%
Z	1,28	1,64	1,96	2,05	2,17	2,33	2,58	3,48	3,89	4,06

Lire les tables de l'écart réduit

Intervalles de confiance pour un seuil α de 5%

moyenne $\quad \mu = m \pm 2 \dfrac{s}{\sqrt{n}}$

$$Z_{5\%}$$

pourcentage $\quad P = p_o \pm 2 \sqrt{\dfrac{p_o \; (1-p_o)}{n}}$

Intervalle de confiance formules générales

moyenne $$\mu = m \pm Z_{\alpha} \frac{s}{\sqrt{n}}$$

pourcentage $$P = p_o \pm Z_{\alpha} \sqrt{\frac{p_o \left(1 - p_o\right)}{n}}$$

Comment obtenir une meilleure précision donc un intervalle de confiance plus étroit ?

$$\mu = m \pm \boxed{Z_\alpha \frac{s}{\sqrt{n}}}$$

$$P = p_o \pm \boxed{Z_\alpha \sqrt{\frac{p_o(1-p_o)}{n}}}$$

Précision

Plus le terme qui vient après le **±**
(Zα x Ecart Type) est petit, meilleure est la précision

Comment obtenir une meilleure précision donc un intervalle de confiance plus étroit ?

$$\mu = m \pm \boxed{Z_\alpha \frac{s}{\sqrt{n}}}$$

Précision

$$P = p_o \pm \boxed{Z_\alpha \sqrt{\frac{p_o(1-p_o)}{n}}}$$

1. Diminuer Z, mais c'est augmenter le risque α

2. Augmenter la taille de l'échantillon n qui se trouve au dénominateur de la précision.

© Thierry Ancelle 2015

Calcul de la taille d'un échantillon pour estimer un paramètre

Calcul de la taille d'un échantillon

Exemple : variable binaire

$$i = Z_{\alpha}\sqrt{\frac{p(1-p)}{n}}$$

$$n = p\,(1 - p)\,\frac{Z_{\alpha}^{2}}{i^{2}}$$

Exemple 3

Fréquence du surpoids chez les élèves de 6ème en Haute Savoie

On désire estimer cette fréquence avec une précision de 3%

Quelle taille d'échantillon est nécessaire ?

$$n = p\,(1 - p)\,\frac{Z_{\alpha}^{2}}{i^{2}}$$

Exemple 3

Fréquence du surpoids chez les élèves de 6ème en Haute Savoie

On désire estimer cette fréquence avec une précision de 3%

Quelle taille d'échantillon est nécessaire ?

- $i = 3\% = 0{,}03$

$$n = p\,(1 - p)\,\frac{Z_{\alpha}^{2}}{i^{2}}$$

- IC 95% -----> $Z_{\alpha} = 2$

Reste **p** qu'on recherche ???

Exemple 3

Fréquence du surpoids chez les élèves de $6^{ème}$ en Haute Savoie

On désire estimer cette fréquence avec une précision de 3%

Quelle taille d'échantillon est nécessaire ?

$$n = p\,(1 - p)\,\frac{Z_\alpha^2}{i^2}$$

- $i = 3\% = 0{,}03$
- IC 95% -----> $Z_\alpha = 2$
- p ??? littérature = 17 % = 0,17

Ordre de grandeur

$$n = 0{,}17\,(1 - 0{,}17)\,\frac{2^2}{0{,}03^2} = 627$$

Si aucune idée sur **p , première enquête ,** on fait rentrer dans la formule **, p=0,5** ce qui donne le produit p *(1-p), le produit **le plus élevé** possible, et donc le chiffre **n** le plus élevé possible.

Exemple 3

Fréquence du surpoids chez les élèves de 6ème en Haute Savoie

On désire estimer cette fréquence avec une précision de 1 %

Quelle taille d'échantillon est nécessaire ?

$$n = p\,(1 - p)\,\frac{Z_\alpha^2}{i^2}$$

$i = 1\% = 0{,}01$

IC 95% -----> $Z_\alpha = 2$

p ??? littérature = 17 % = 0,17

$$n = 0{,}17\,(1 - 0{,}17)\,\frac{2^2}{0{,}01^2} = \mathbf{5644}\ !!$$

Calcul de la taille d'un échantillon

Variable quantitative

Précision : $i = Z_\alpha \dfrac{s}{\sqrt{n}}$

$$n = s^2 \frac{Z_\alpha^2}{i^2}$$

Exemple 3

Fréquence du surpoids chez les élèves de $6^{ème}$ en Haute Savoie

On désire estimer la moyenne du poids des élèves avec une précision de 1 kg

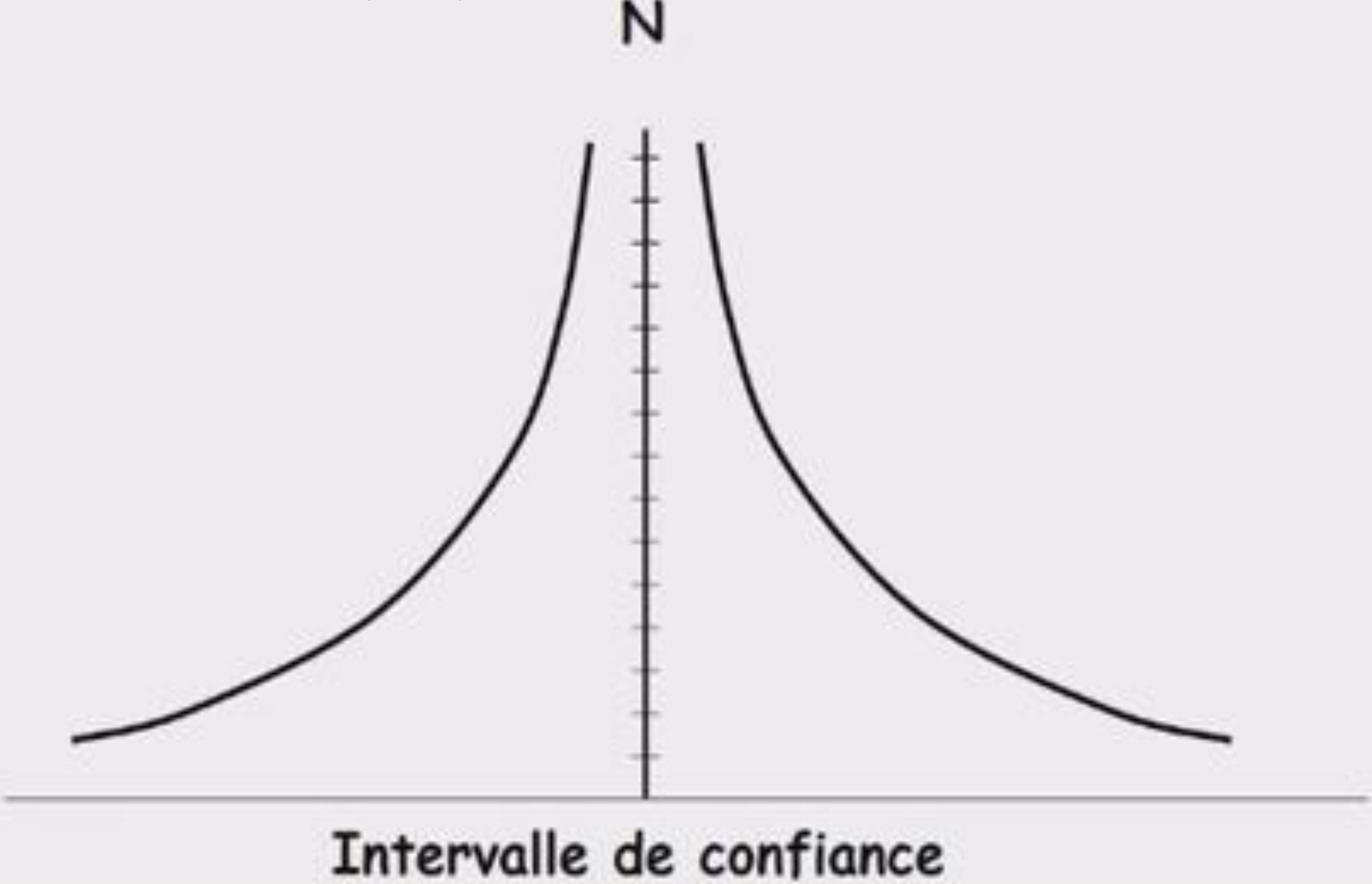
Quelle taille d'échantillon est nécessaire ?

$$n = s^2 \frac{Z_\alpha^2}{i^2}$$

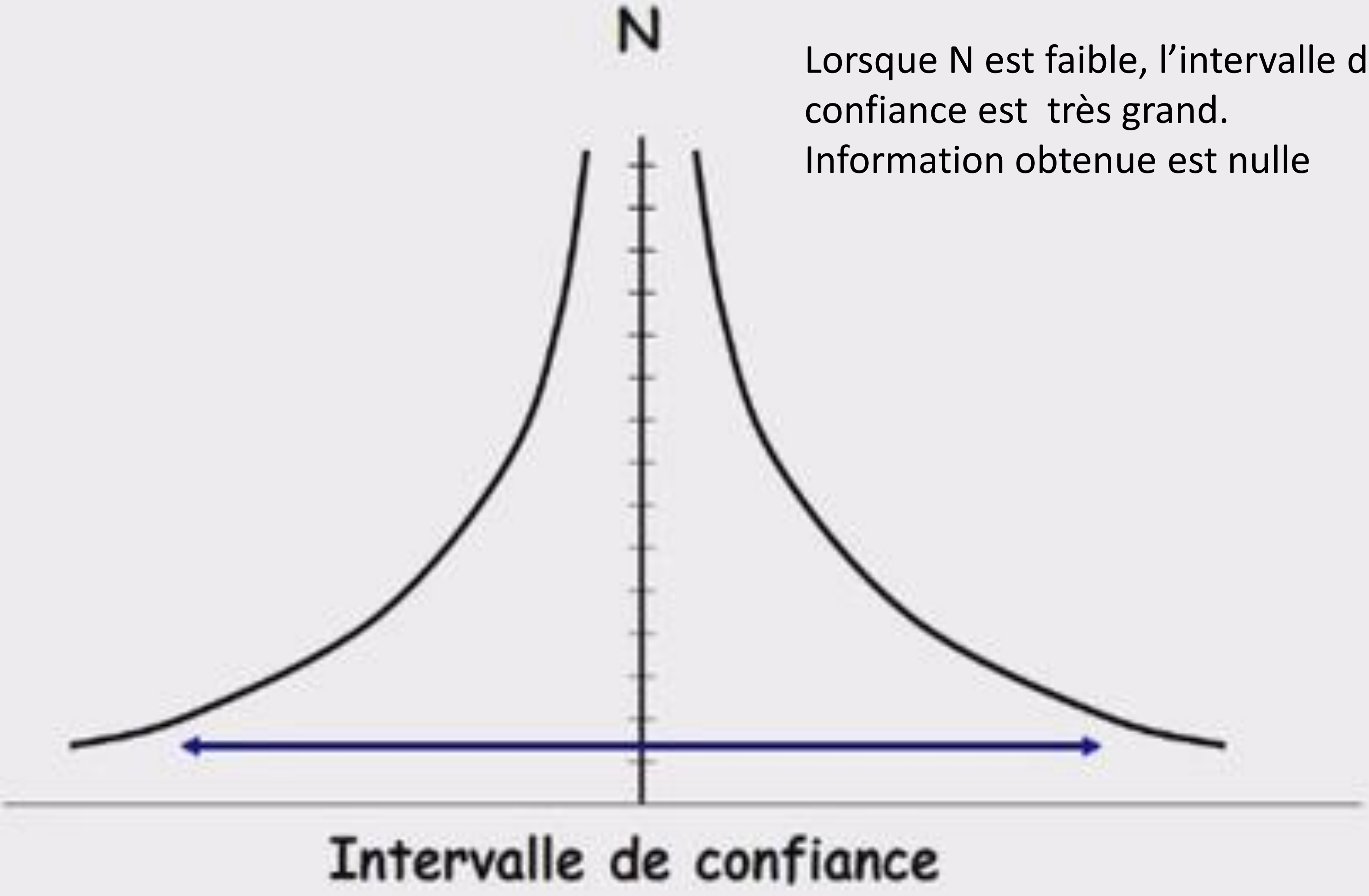
- i = 1 kg
- IC 95% -----> $Z_\alpha = 2$
- s ??? littérature, étude pilote

Précision obtenue en fonction de la taille de l'échantillon

Jusqu'où peut-on élever la taille de l'échantillon ??

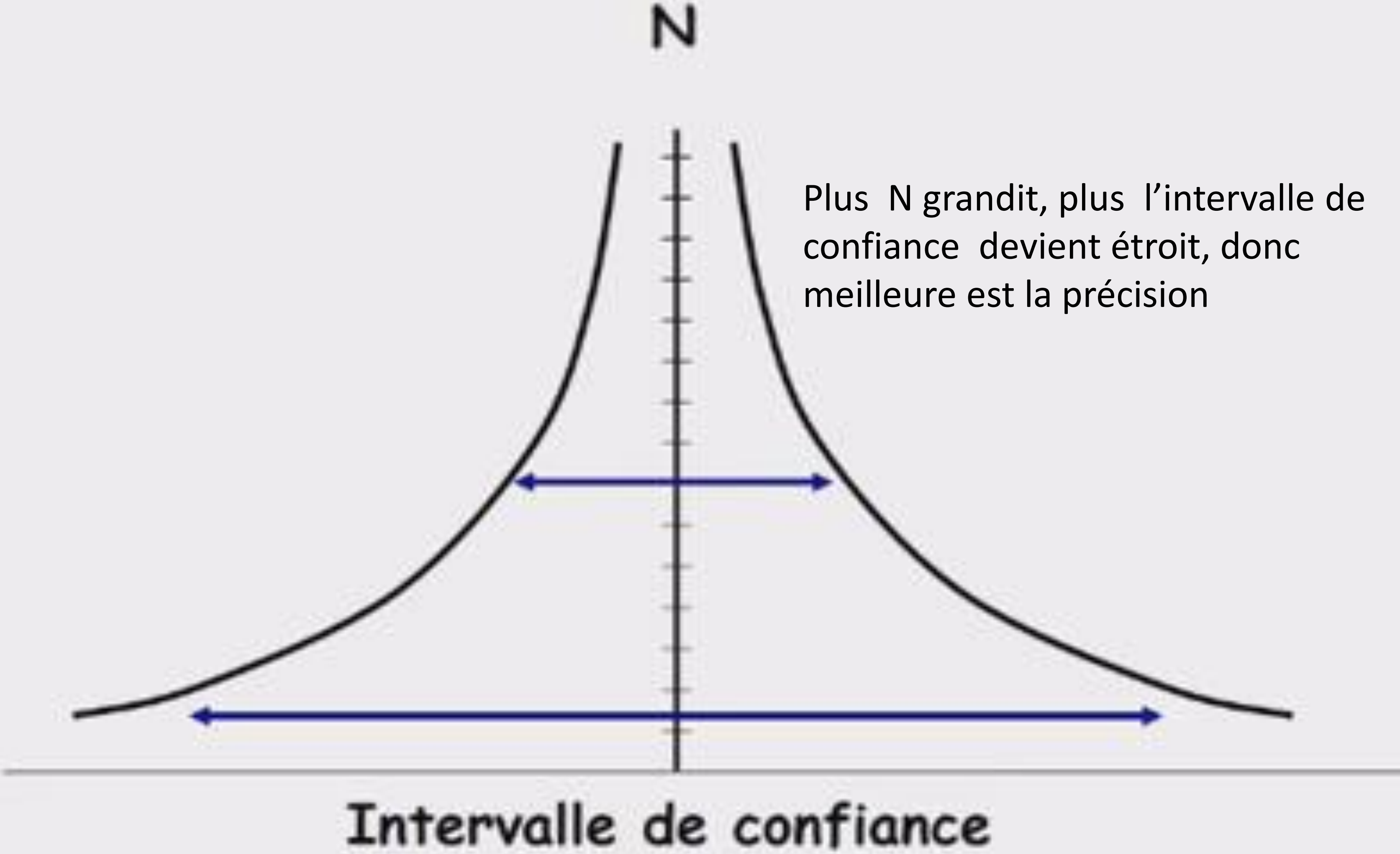


Précision obtenue en fonction de la taille de l'échantillon



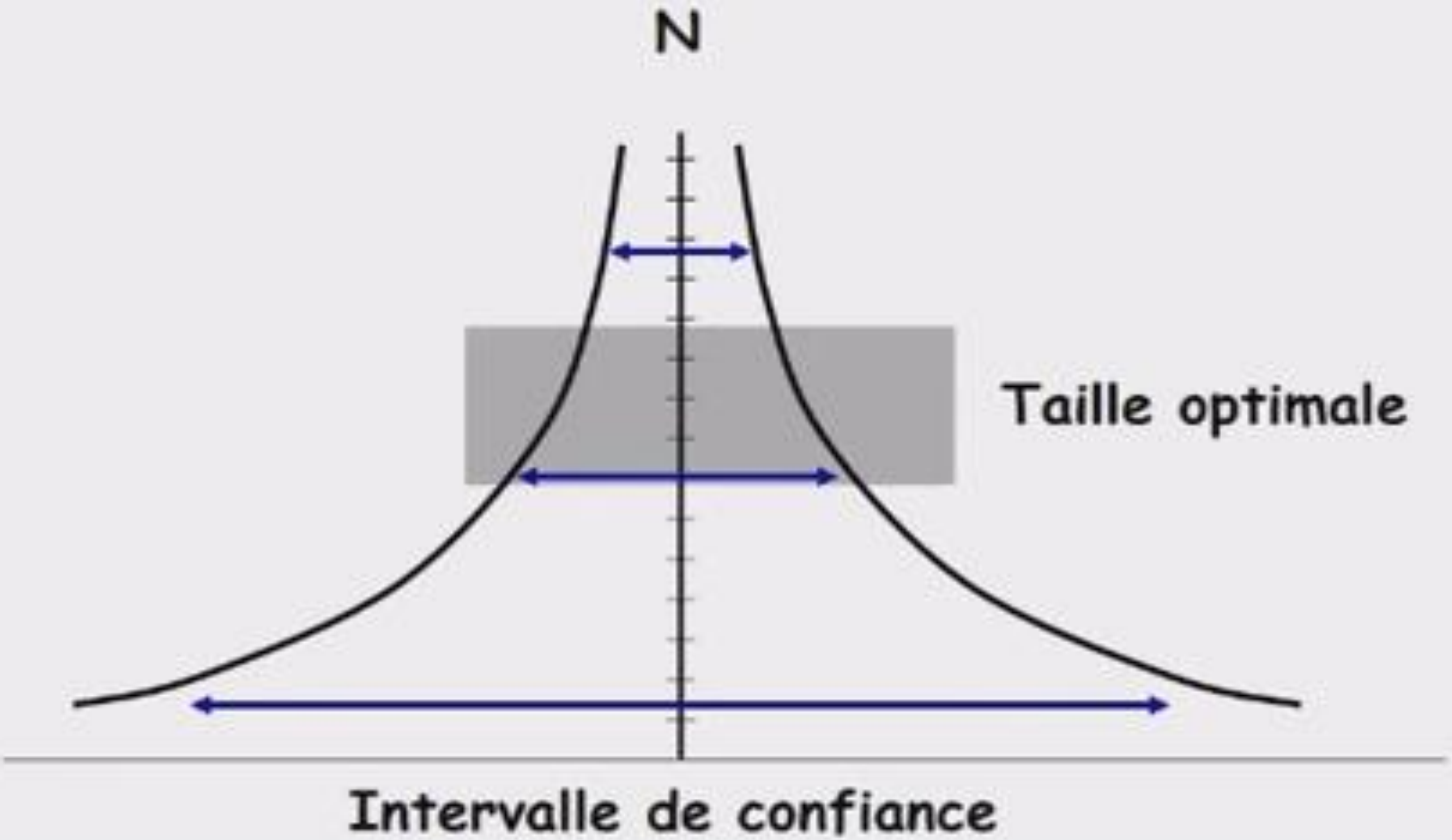
Lorsque N est faible, l'intervalle de confiance est très grand. Information obtenue est nulle

Précision obtenue en fonction de la taille de l'échantillon



Plus N grandit, plus l'intervalle de confiance devient étroit, donc meilleure est la précision

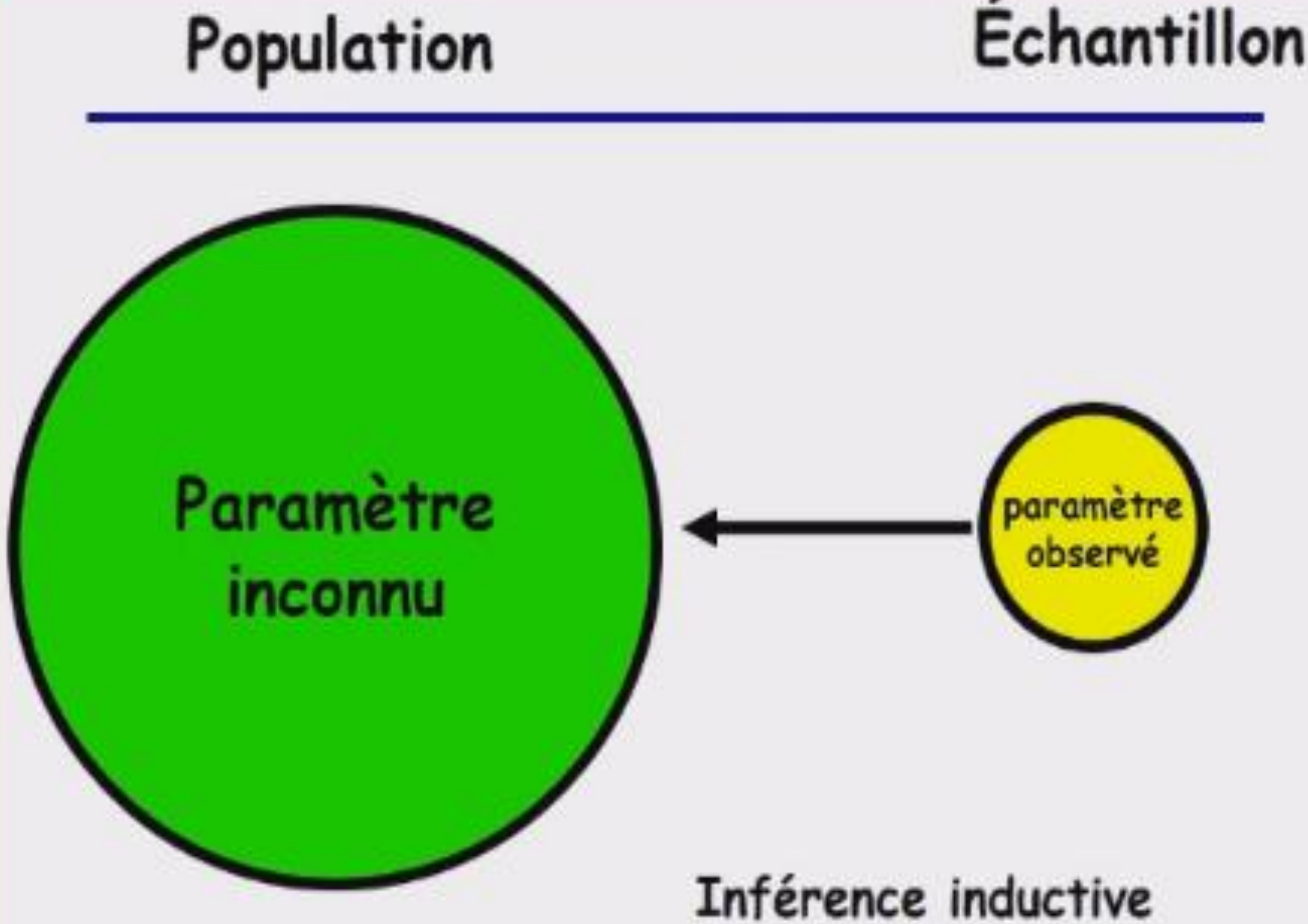
Précision obtenue en fonction de la taille de l'échantillon



Il y a une limite optimale à ne pas dépasser.
A partir d'un certain seuil , on remarque que la charge de travail augmente énormément pour une amélioration dérisoire de la précision

Utiliser Paramètres mesurées à partir d'un échantillon

Échantillon



Tirer par inférence inductive, une estimation sur les paramètres réels de population