Chapitre 5

Estimation

Estimation ponctuelle

Exemple: on souhaite connaître le poids moyen des chats forestiers en France

X = « poids d'un chat forestier ». On cherche $\mu=E(X)$ (moyenne de tous les chats de France)

$\Rightarrow$ Echantillonage de n individus: $x_1,\ldots, x_n$

A partir de ces données, quelle est la valeur la plus raisonnable pour μ ? (si on devait essayer de deviner)

$\Rightarrow$ C'est la moyenne empirique $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$

On dit que $\bar{x}$ est une estimation de μ. On écrit $\hat{\mu} = \bar{x}$

Eviter les confusions

On a définit jusqu'à maintenant 3 types de moyenne: il ne faut pas les confondre.

1) La moyenne théorique (ou réelle) μ: c'est celle que l'on souhaite connaître mais qu'on ne connaîtra jamais car il faudrait mesurer TOUS les chats

2) La moyenne observée ou empirique $\bar{x}$: c'est une mesure uniquement descriptive de notre échantillon (permet de résumer nos données)

3) La moyenne estimée $\hat{\mu}$: traduit une tentative de notre part d'essayer de deviner la moyenne théorique (en s'appuyant sur des critères mathématiques)

Dans le cas de la moyenne, on a moyenne estimée = moyenne observée. Ce n'est pas vrai tout le temps (p.e. faux pour la variance)

La notion d'estimateur

L'estimation de notre moyenne dépend de notre échantillon.

L'estimation serait différente si on avait attrapé des chats différents.

$\Rightarrow$ On parle de variable aléatoire $\overline{X} = \frac{X_1 + ... + X_n}{n}$

(X_i étant la variable aléatoire: « taille du i-ème chat capturé »)

On dit que $\overline{X}$ est un **estimateur** de la moyenne

Dans le cas de la moyenne on dit que l'estimateur est sans biais car

$$E(\overline{X}) = \frac{E(X_1) + ... + E(X_n)}{n} = \frac{\mu + ... + \mu}{n} = \frac{n\mu}{n} = \mu$$

Une expérience simple

Question: quelle est la taille moyenne des étudiants (hommes) de Mathsv de l'année dernière ?

On prélève au hasard N étudiants et on mesure leur taille x_1,…, x_n

On calcule la moyenne empirique $(x_1+\ldots+ x_n)/n$

On essaie avec différentes valeurs de n (n=4, n=20 et peut être au besoin n=50)

***Rappel*: ce que l'on cherche c'est la moyenne de TOUS les étudiants (un peu moins de 200).**

La moyenne empirique (ou observée) sert d'ESTIMATION de la moyenne réelle recherchée

Que nous apprennent les résultats ?

1) Il n'y a aucune chance que la moyenne observée nous donne exactement la moyenne recherchée: elle est différente d'un échantillon à l'autre (c'est une variable aléatoire).

En effet: si on prend 4 étudiants, on peut tomber par malchance sur 4 étudiants particulièrement grands (ou petits) et donc sur-estimer (ou sous-estimer) notre moyenne recherchée

Que nous apprennent les résultats ?

1) Il n'y a aucune chance que la moyenne observée nous donne exactement la moyenne recherchée: elle est différente d'un échantillon à l'autre (c'est une variable aléatoire)

2) Plus on prend d'étudiants, moins l'écart entre deux estimations va être grand

En effet: avec 4 étudiants, on pouvait tomber par malchance sur des étudiants tous très grands. Avec 50 étudiants c'est très improbable: on aura certainement des grands, des petits et des moyens et tout cela va se compenser.

Que nous apprennent les résultats ?

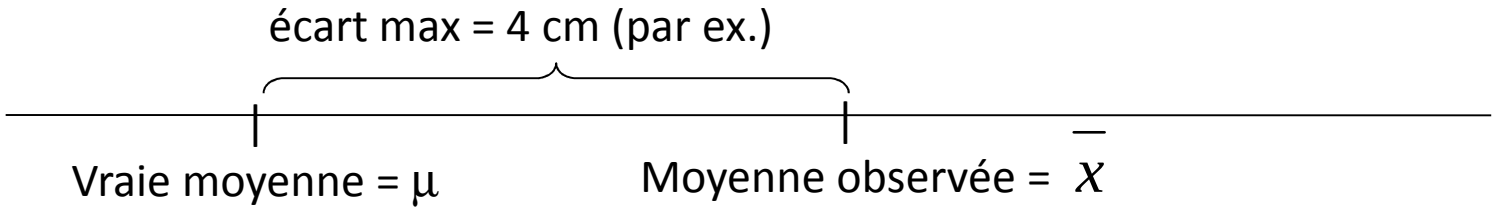
1) Il n'y a aucune chance que la moyenne observée nous donne exactement la moyenne recherchée: elle est différente d'un échantillon à l'autre (c'est une variable aléatoire)

2) Plus on prend d'étudiants, moins l'écart entre deux estimations va être grand

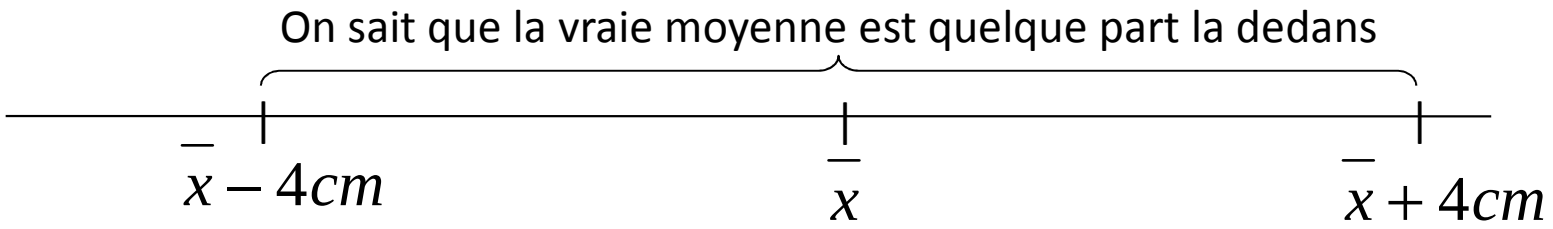
3) **Moralité**: on ne peut pas connaître la moyenne de manière exacte: il y a une imprécision dans notre estimation (**effet d'échantillonnage**). Par contre, plus on prend d'individus, moins on va s'écarter de la véritable moyenne (meilleure précision)

Et les maths dans tout ça ?

Les maths permettent d'évaluer de combien la moyenne estimée à partir d'un échantillon peut au maximum s'éloigner de la valeur réelle. Ceci permet, à partir de la moyenne observée, de deviner la vraie moyenne recherchée.



La plupart du temps on ne sait pas où est la vraie moyenne, mais:



Bien sûr ceci est basé sur un calcul de probabilité…

On met tout ça en équation…

On appelle X la variable aléatoire « taille des étudiants homme de mathsv ». On suppose que X~N(μ,σ)

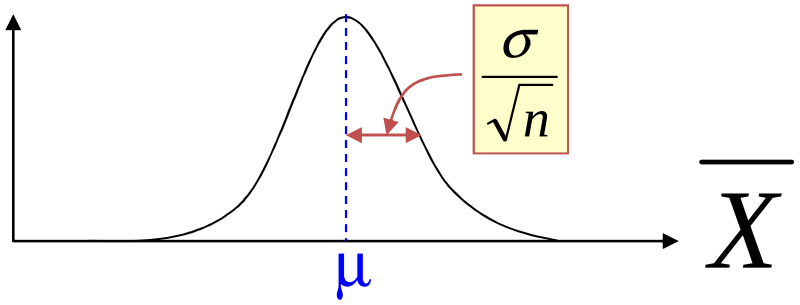
On choisit n individu.

On appelle X_i = « taille du i-ème étudiant choisi ».

Tous les X_i suivent aussi une N(μ,σ)

On montre qu'alors: $\overline{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

μ est alors la valeur moyenne attendu pour X et $\sigma / \sqrt{n}$ décrit (à peu près) l'écart moyen que l'on attend entre la moyenne estimée et la moyenne réelle



On met tout ça en équation…

Pour se ramener à une loi normale centrée réduite on centre et on réduit:

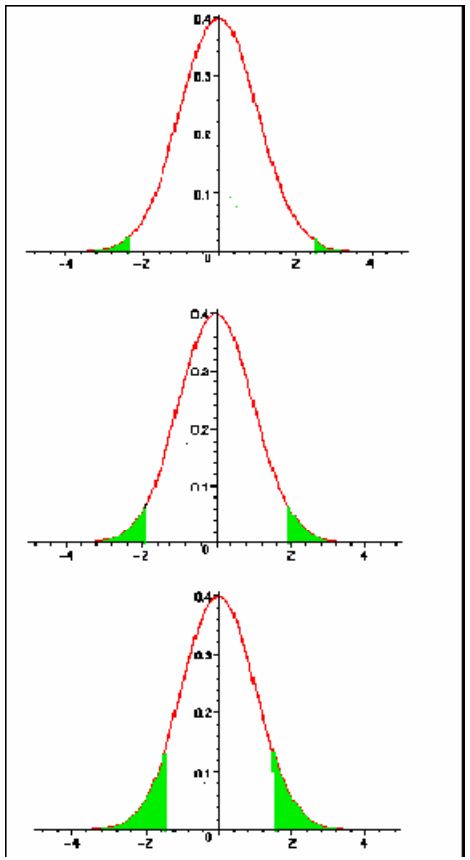
$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Trouver un Z = 0.5 signifierait que, dans notre échantillon, la valeur moyenne estimée s'écarte de la moyenne réelle (ou théorique) de 0.5 fois ce à quoi on pourrait s'attendre normalement.

Z décrit **l'écart réduit** entre la moyenne réelle et la moyenne observée

L’écart réduit étant distribué comme une loi N(0,1), il peut prendre toutes les valeurs possibles (écart entre moyenne observée et théorique aussi grand qu’on veut)

Cependant, en y regardant de plus près (cf Tables):



Il y a 99 chances sur 100 que la valeur observée (z) de Z soit entre -2.58 et +2.58

Il y a 95 chances sur 100 que la valeur observée (z) de Z soit entre -1.96 et +1.96

Il y a 90 chances sur 100 que la valeur observée (z) de Z soit entre -1.64 et +1.64

Ca veut dire que:

1) En théorie, ma moyenne estimée peut être très loin de ma moyenne réelle, MAIS:

2) En pratique, la moyenne estimée a une probabilité très forte de n'être « pas trop loin » de la moyenne réelle

3) En acceptant de prendre un risque de me tromper, je peux donner un intervalle dans lequel doit se situer ma valeur moyenne

Exemple: en ayant 5% de risque de se tromper, on peut dire que -1.96<z<+1.96 (z = l'écart réduit observé) et donc:

$$-1.96 < z < +1.96 \quad \Leftrightarrow -1.96 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < +1.96$$

$$\Leftrightarrow -1.96 \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < +1.96 \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow -\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Généralisation:

On appelle α le risque que l'on est prêt à prendre (on parle de **risque de première espèce**)

On appelle ε_α la valeur telle que si Z~N(0,1), alors $P(-\varepsilon_\alpha<Z<+\varepsilon_\alpha)=1-\alpha$

Alors on montre qu'avec une probabilité α de se tromper on a:

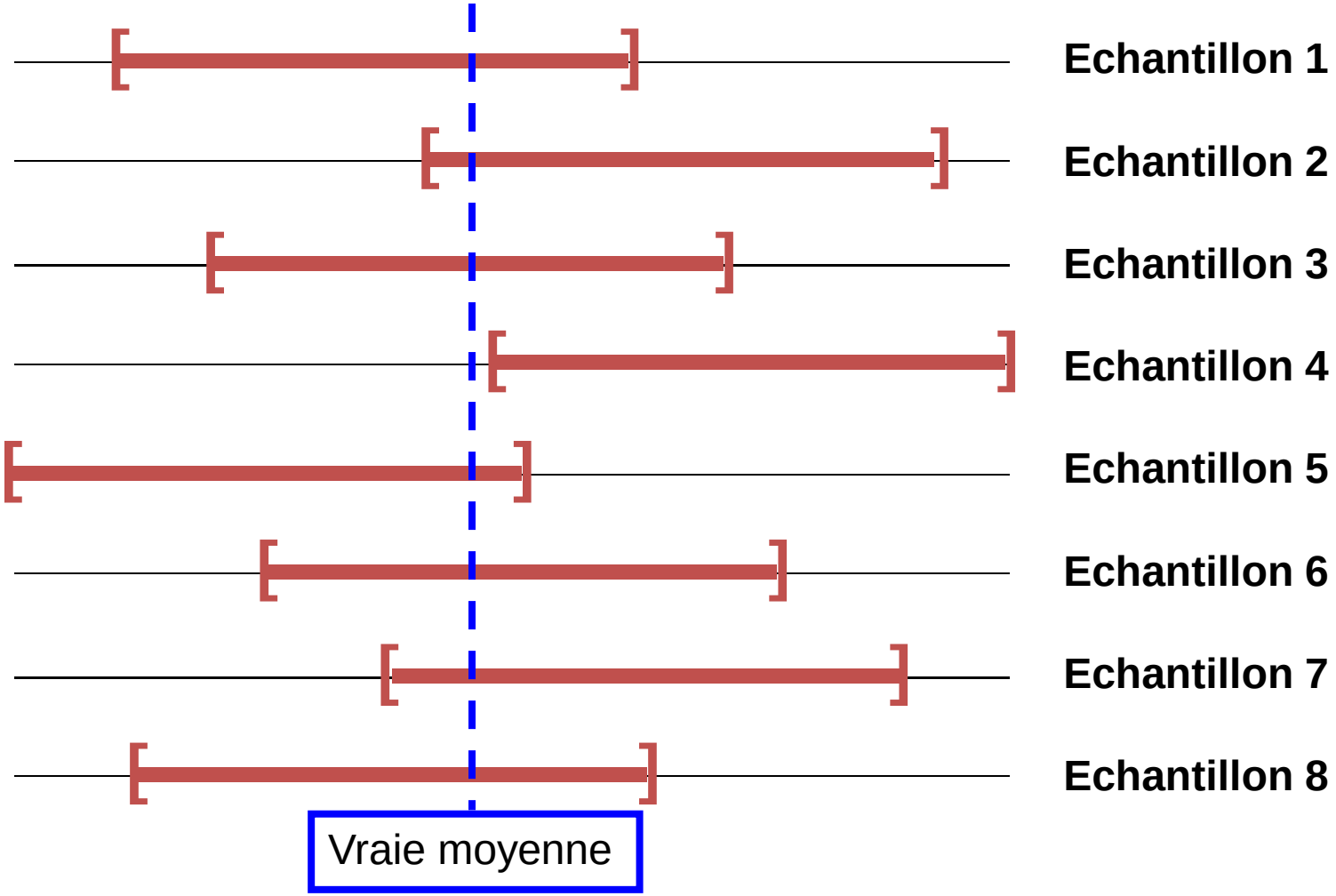
$$\bar{x}-\varepsilon_\alpha\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+\varepsilon_\alpha\frac{\sigma}{\sqrt{n}}$$

L'intervalle $\left[\bar{x}-\varepsilon_\alpha\frac{\sigma}{\sqrt{n}};\bar{x}+\varepsilon_\alpha\frac{\sigma}{\sqrt{n}}\right]$ est appelé intervalle de confiance au

risque α de la moyenne **réelle.** Il permet de quantifier **l'effet d'échantillonnage** sur l'estimation de la moyenne.

Remarque:

La valeur de l'IC dépend de l'échantillon. Si on répétait un échantillonnage de même taille 100 fois, on aurait des intervalles de confiance qui contiennent bien la vraie moyenne dans 95 cas (à peu près) si α=0.05



Un exemple…

Si on connaît pas la variance on fait quoi …?

Bah on l'estime…

Mesures $x_1, \ldots, x_n$

Intuitivement, on voudrait prendre

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

Pour des raisons mathématiques, il vaut mieux prendre

$$\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Si on connaît as la variance on fait quoi …?

L'estimation de la variance dépend aussi de l'échantillon. Pour faire simple on appelle aussi $\hat{\sigma}^2$ l'estimateur de la variance

Comme tout à l'heure on va considérer l'écart réduit (on remplace l'écart type par son estimation $\hat{\sigma}$)

$$T = \frac{\overline{X} - \mu}{\hat{\sigma} / \sqrt{n}}$$

Contrairement à tout à l'heure, l'écart réduit n'est plus distribué comme une N(0,1) car on a divisé par $\hat{\sigma}$ qui est une variable aléatoire

Si on connaît as la variance on fait quoi …?

On montre mathématiquement que $\hat{\sigma}^2$ est distribué suivant un χ^2 à n-1 ddl

$$\Rightarrow T = \frac{\overline{X} - \mu}{\hat{\sigma} / \sqrt{n}} \sim T(n-1 \ ddl)$$

$\Rightarrow$ C'est la raison pour laquelle on utilise la lettre T (au lieu de Z quand la variance est connue)

On en déduit que: $$IC_{\alpha}(\mu) = \left[\overline{x} - t_{\alpha}^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} ; \overline{x} + t_{\alpha}^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

avec t_{α}^{n-1} la valeur telle que si T~T(n-1 ddl), alors $P(-t_{\alpha}^{n-1} < T < t_{\alpha}^{n-1}) = 1 - \alpha$

Remarque: si n≥30

1) On a plus besoin de supposer que la variable mesurée (et donc les X_i) suit une loi normale
2) On peut approcher T par une loi normale (cf dia plus loin)

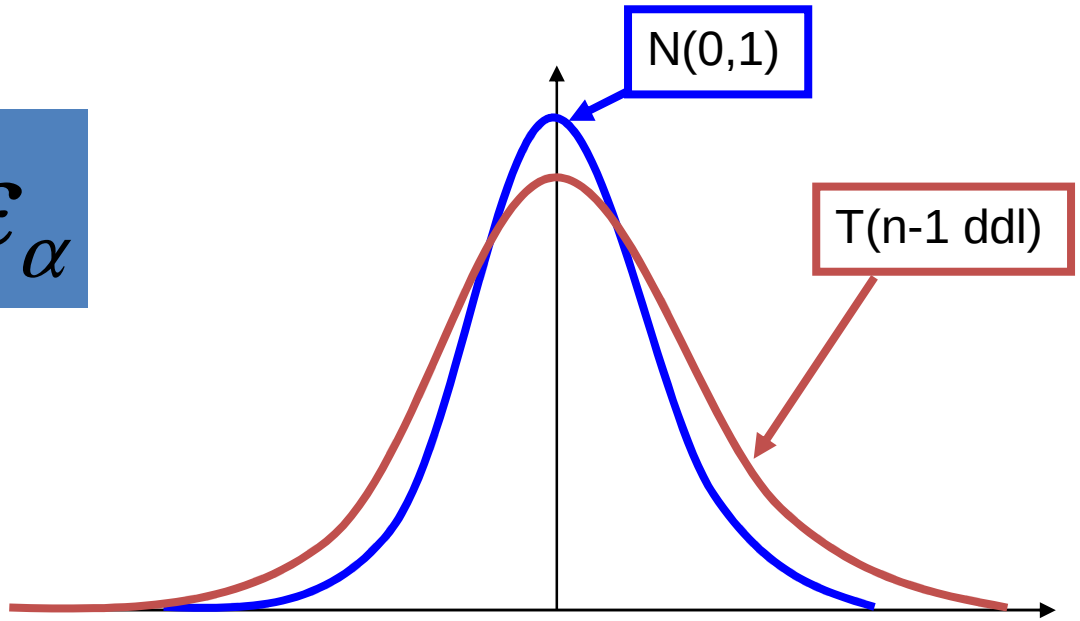
Si on connaît as la variance on fait quoi …?

Remarque 2:

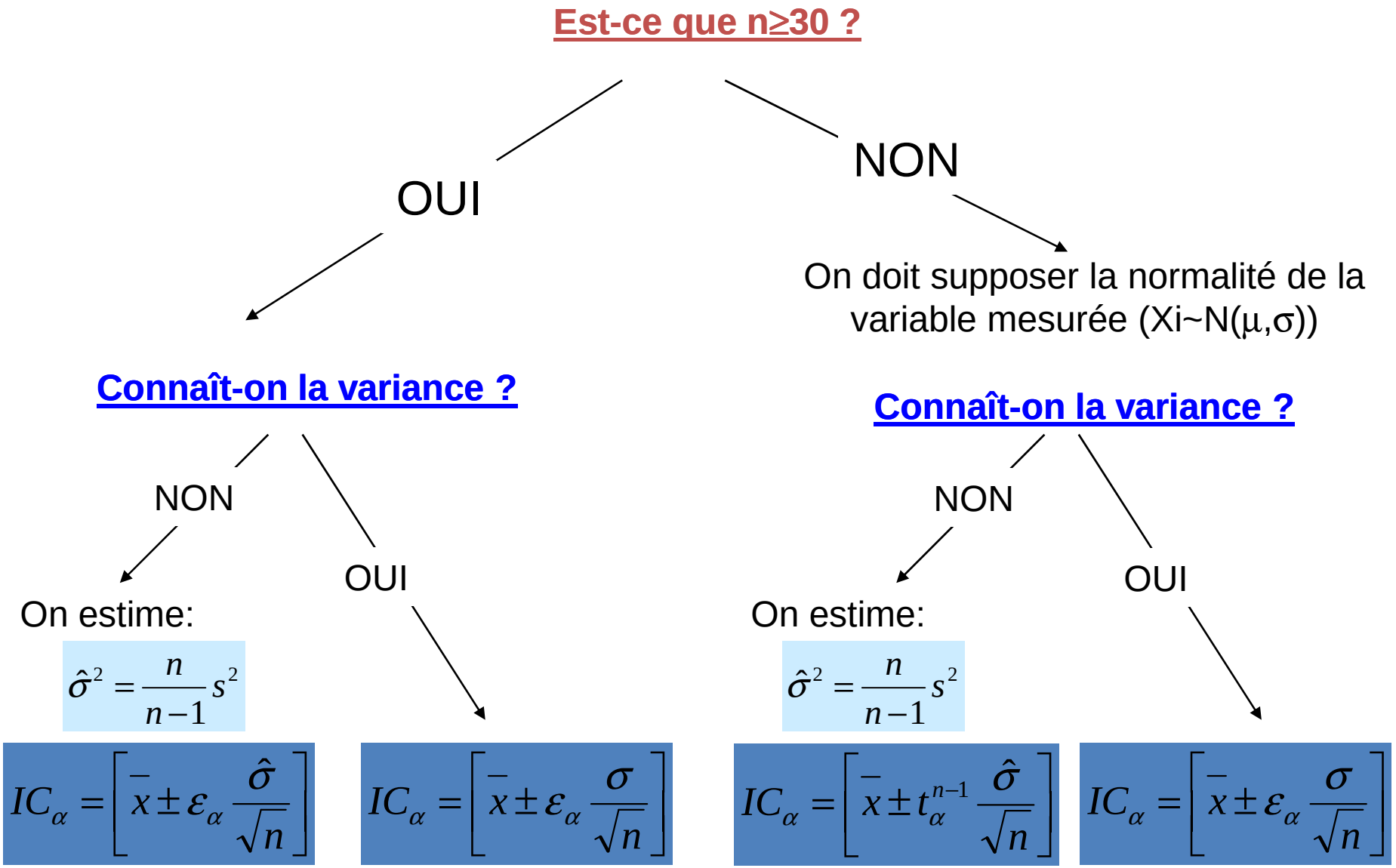
Le fait de ne pas connaître la variance induit une imprécision supplémentaire qui se traduit par un intervalle de confiance plus large.

En effet, on a toujours

$$t_{\alpha}^{n-1} > \varepsilon_{\alpha}$$



Récapitulatif



Application

Essayons de deviner la taille moyenne des étudiants de mathsv de l’année dernière

On calcule $\hat{\sigma} =$

VRAIE VALEUR = ...

Intervalle de confiance d'une proportion

Exemple: on souhaite connaître la fréquence d'un allèle dans la population

On échantillonne n individus

⇒ Estimation de la proportion de porteurs de l'allèle ?

⇒*Fréquence observée*: $f = \frac{\text{Nombre de porteurs de l'allèle dans l'échantillon}}{\text{Nombre d'individus échantillonnés}} = \frac{x}{n}$

Estimation de la fréquence réelle: $\hat{p} = f$

Comme pour l'estimation d'une moyenne il existe une incertitude (ou erreur) associée à cette estimation. Il faut quantifier cette erreur (**effet d'échantillonnage**)

⇒ MATHS (calcul de probabilité)

Remarque: Comme pour l'estimation de la moyenne et de la variance, on fait la distinction entre l'estimateur (variable aléatoire) et l'estimation (quantité observée réalisation de l'estimateur). Pour **simplifier** les notations on note les deux $\hat{p}$

Intervalle de confiance d'une proportion

D'un point de vue mathématique c'est une probabilité que l'on estime

$\Rightarrow$ En prenant un individu au hasard dans la population, quelle est la probabilité (notée **p**) qu'il porte l'allèle?

On échantillonne n individus

On définit la variable aléatoire X = « nombre d'individus porteurs de l'allèle »

On a X~B(n,p)

Si n ≥ 30, np ≥ 5 et nq ≥ 5, on peut faire l'approximation:

$$X \sim N(np, \sqrt{npq}) \qquad \text{où } q = 1 - p$$

Donc $\hat{p} = \frac{X}{n} \sim N(p, \sqrt{\frac{pq}{n}})$

Imprécision de l'estimation liée à l'échantillonnage

Intervalle de confiance d'une proportion

Comme précédemment (cf estimation moyenne), on centre et on réduit pour se ramener à une N(0,1)

$$\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Ici, on ne connaît pas l'écart type de $\hat{p}$, donc on l'estime:

$$Ecart\ type\ réel\ = \sqrt{\frac{pq}{n}} \quad \rightarrow \quad estimation\ = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$avec\ \hat{q} = 1 - \hat{p}$$

Comme n≥30, on a à peu près:

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \sim N(0,1)$$

Intervalle de confiance d'une proportion

Comme précédemment, comme Z~N(0,1), on peut a $P(\varepsilon_\alpha < Z < \varepsilon_\alpha) = 1 - \alpha$

Par un calcul similaire à celui de l'intervalle de confiance de la moyenne, on obtient l'intervalle de confiance de la proportion p:

$$IC_\alpha(p) = \left[\hat{p} - \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} ; \hat{p} + \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} \right]$$

Remarques:

1) ici, on a qu'une seule formule car on doit forcément avoir $n \geq 30$

1) Si $n < 30$, il faut faire appel à d'autres méthodes (hors programme)

Intervalle de confiance d'une proportion

Exemple: sondage élection présidentielle

2 candidats: A et B
p = proportion de la population (française votante) qui vote pour A

Estimation sur 10,000 sondés:

$\hat{p} = 0.52$

$IC_{0.05}(p) = ?$