Statistique Inférentielle
Chapitre 21 : Estimateur

Notations

Population : Notations

- Les lettres majuscules sont réservées pour la population.
- Population U avec un effectif N

$$U = \{ U_i \; tels \; que \; i = 1 \ldots N \}$$

- Les U_i sont des données certaines mais inconnues.
- Si on étudie un caractère de cette population - la variable d'intérêt, on la représente par $\mathfrak{I}$.

$$\mathfrak{I}: U \to \Re: U_i \to Y_i (i = 1 \ldots N)$$

- Les valeurs individuelles de cette variable sont notées

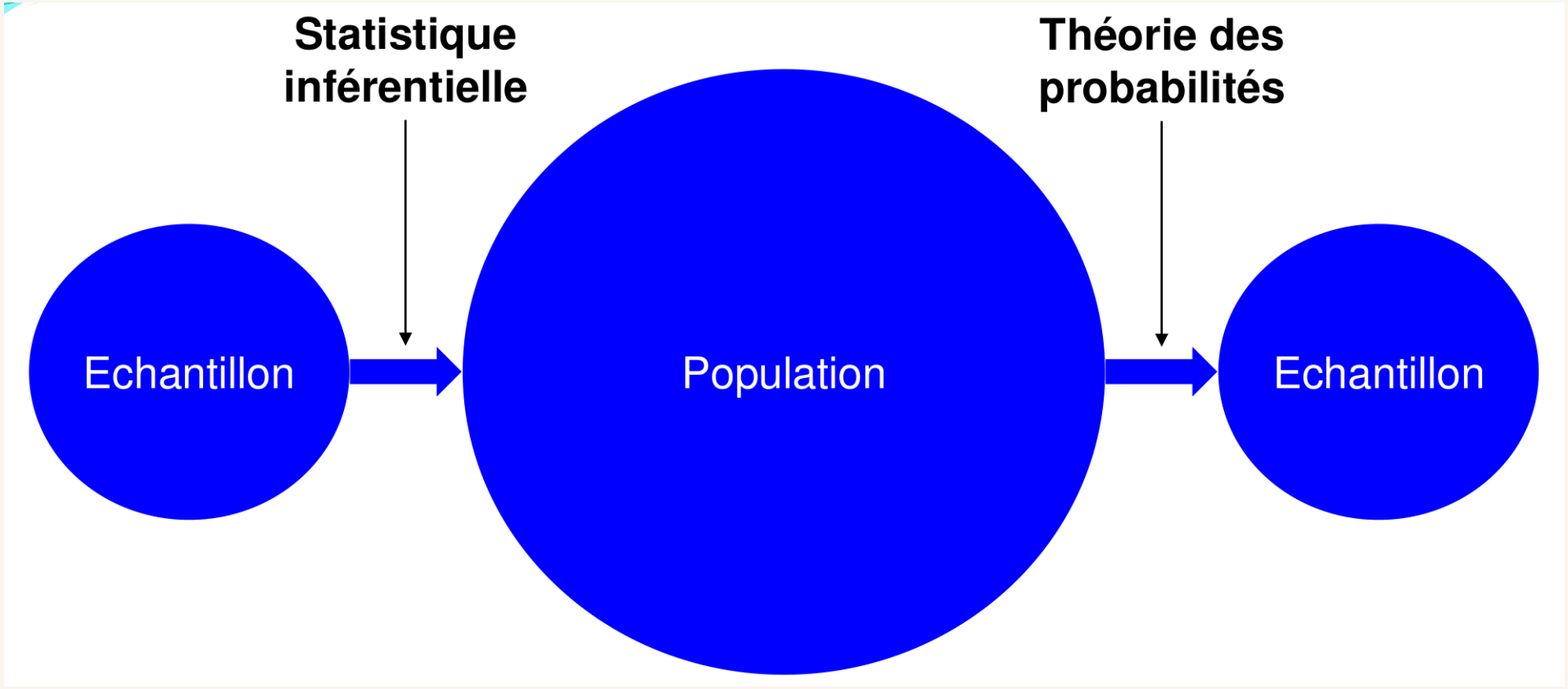
$$Y_i (i = 1 \ldots N)$$

Échantillon : Notations

- Les lettres minuscules sont réservées pour l'échantillon.
- On prélève un échantillon de taille n.
- Les unités tirées seront notées $u_1, u_2, \ldots, u_n$.
- Les valeurs observées sur ces unités particulières sont $y_1, y_2, \ldots, y_n$.
- Les u_i sont connues mais aléatoires.
- Par conséquent, y_i est une variable aléatoire, dont les valeurs possibles sont y_i $(i = 1 \ldots n)$.

- $f = \frac{n}{N}$ est le taux de sondage
  - n est la taille de l'échantillon
  - N est la taille de la population

Population versus Echantillon



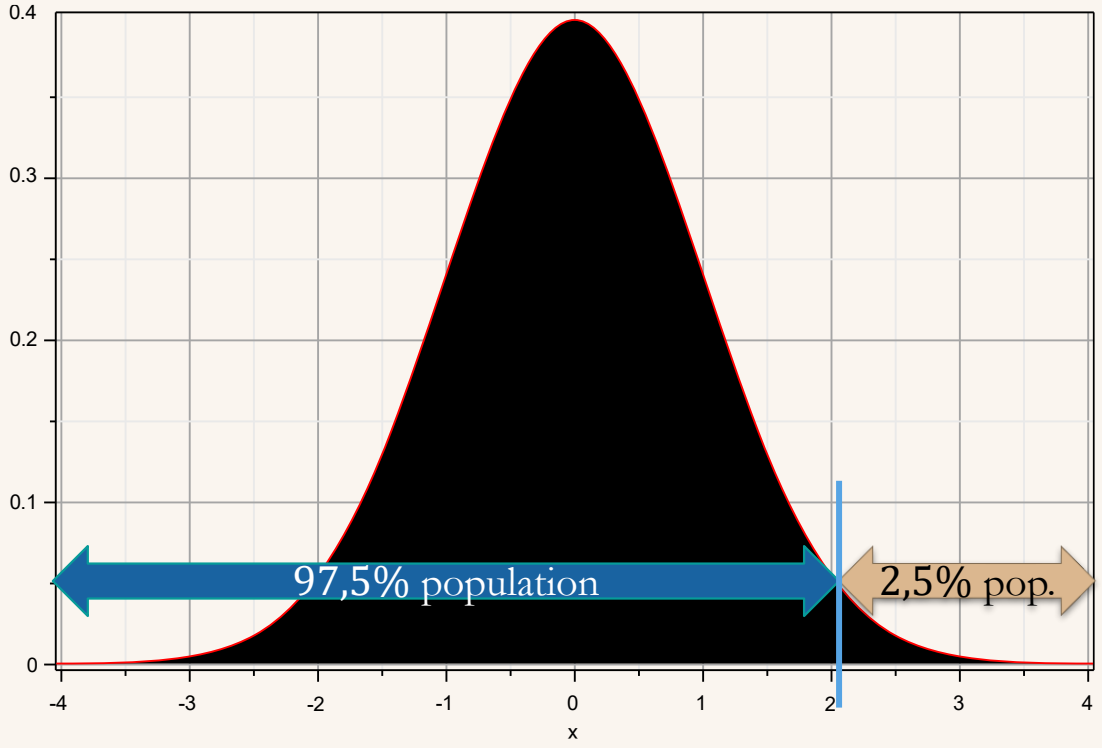
Quantiles

Quantile

- Les quantiles divisent un jeu de données en intervalles contenant le même nombre de données.
- Par exemple :
  - quartiles = les trois quantiles qui divisent le jeu de données en quatre groupes de taille égale.
  - médiane = le quantile qui divise le jeu de données en deux groupes de taille égale.

Quantile

- Le quantile d'ordre *p* (0<p<1) est une valeur qui partage le jeu de données d'effectif *n* en deux sous-ensembles contenant un nombre d'observations égal à *n.p* et *n.(1-p)*.
  - Càd contenant une proportion d'observations égale à *p* et *1-p*.
- Par exemple
  - La médiane est le quantile d'ordre 0,5.
  - Le quantile d'ordre 0,975 partage les données en deux sous-ensembles contenant 97,5% et 2,5% des données.
  - Le quantile d'ordre 0,975 vaut environ 2 dans la loi normale centrée réduite (cf schéma)

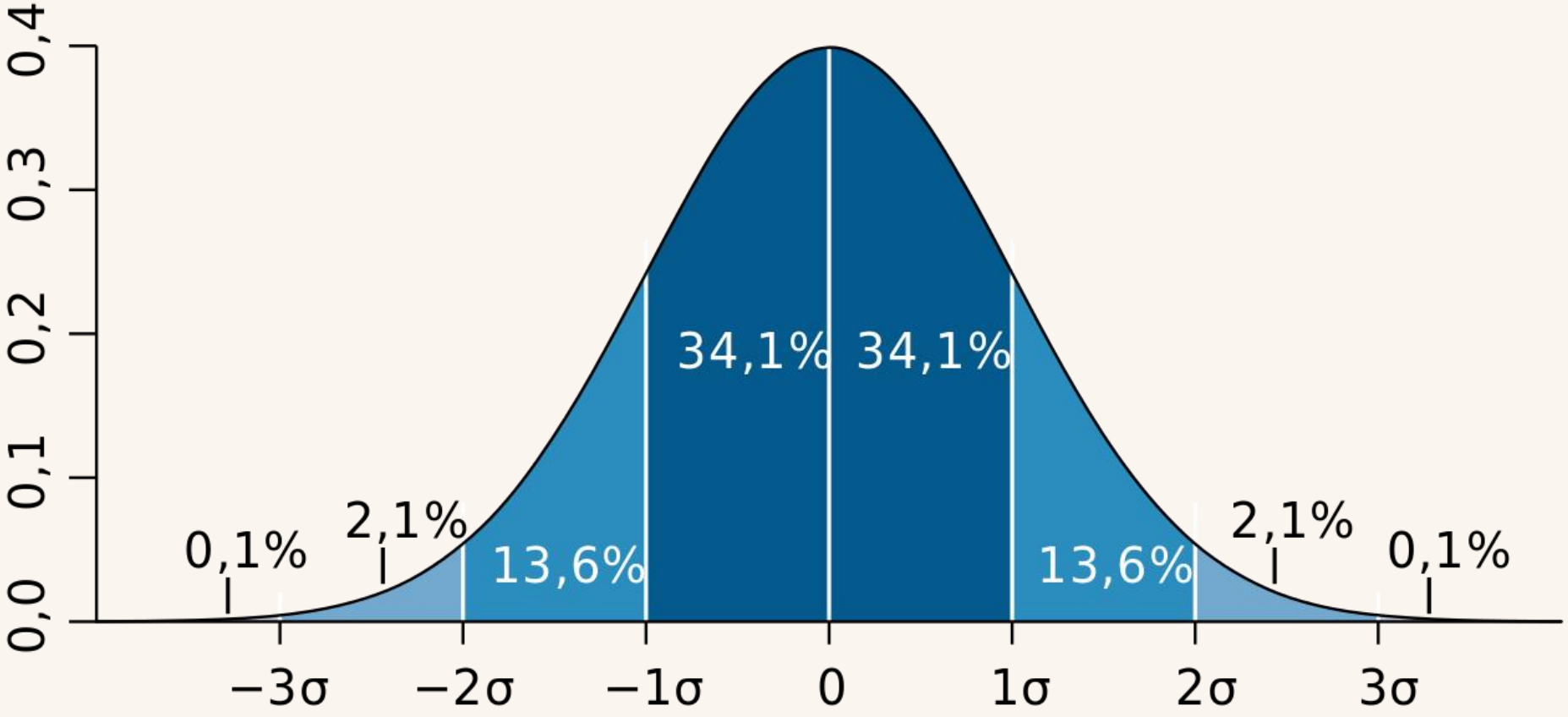


Théorème Central Limite

source : Wikipedia

https://fr.wikipedia.org/wiki/Théorème_central_limite

Rappel : la loi de Laplace-Gauss



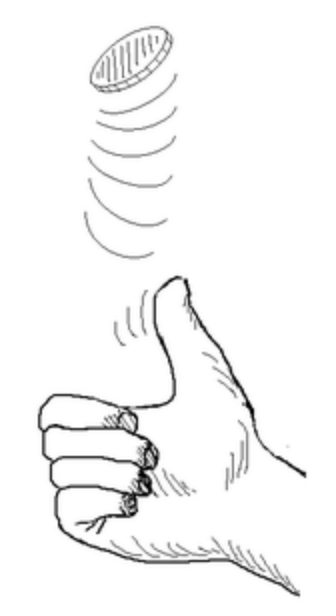
Principe

- Une somme d'un grand nombre de variables aléatoires indépendantes et identiquement distribuées tend vers la loi de Laplace-Gauss.
- Enfin, bon
  - assez souvent …
  - sous des certaines hypothèses
- Explication de l'omniprésence de la loi normale dans la nature
  - de nombreux phénomènes naturels sont le résultat de l'addition d'un grand nombre de petites perturbations aléatoires.

Principe

- Universalité
  - Si n variables aléatoires suivent déjà une loi normale d'espérance μ, alors leur somme suit évidemment aussi une loi normale de paramètre $n\mu$.
  - Mais, si des variables ne suivent pas la loi normale, alors … le théorème s'applique souvent aussi !

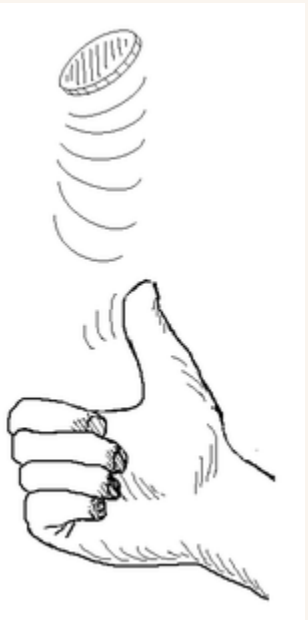
Exemple : Pile ou face



- Modélisation
  - Valeurs considérées : pile =0 ; face = 1
  - Chaque face à 50% de chance d'être tirée.
  - On procède à n tirages et on fait la somme des valeurs.
- Exemple typique d'une loi binomiale
  - cf chapitre 03
- Mais que se passe-t-il si le nombre d'expérience devient très élevé ?

Exemple : Pile ou face

- si un seul tirage
  - n = 1



Résultats d'un tirage

Résultat tirage n°1	Somme
0	0
1	1

Fréquences pour un tirage

Valeurs de la somme	Nombre d'apparitions	Fréquence
0	1	1/2 = 0,5 (50 %)
1	1	1/2 = 0,5 (50 %)

Exemple : Pile ou face

- si trois tirages
  - $n = 3$

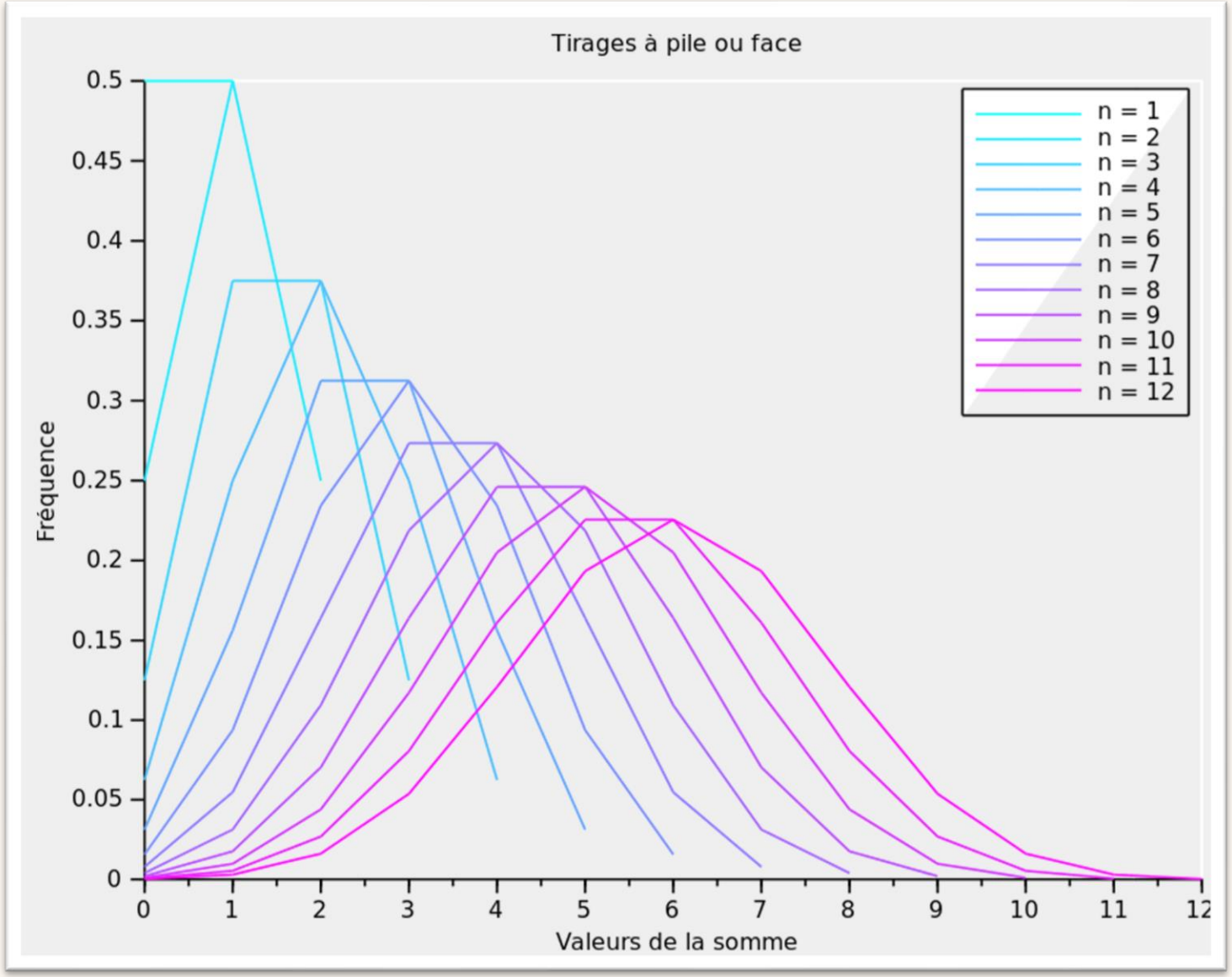


Résultats et fréquences de trois tirages

Résultat tirage n°1	Résultat tirage n°2	Résultat tirage n°3	Somme		Valeurs de la somme	Nombre d'apparitions	Fréquence
0	0	0	**0**		0	1	0,125 (12,5 %)
0	0	1	**1**		1	3	0,375 (37,5 %)
0	1	0	**1**		2	3	0,375 (37,5 %)
0	1	1	**2**		3	1	0,125 (12,5 %)
1	0	0	**1**				
1	0	1	**2**				
1	1	0	**2**				
1	1	1	**3**				

Exemple : Pile ou face

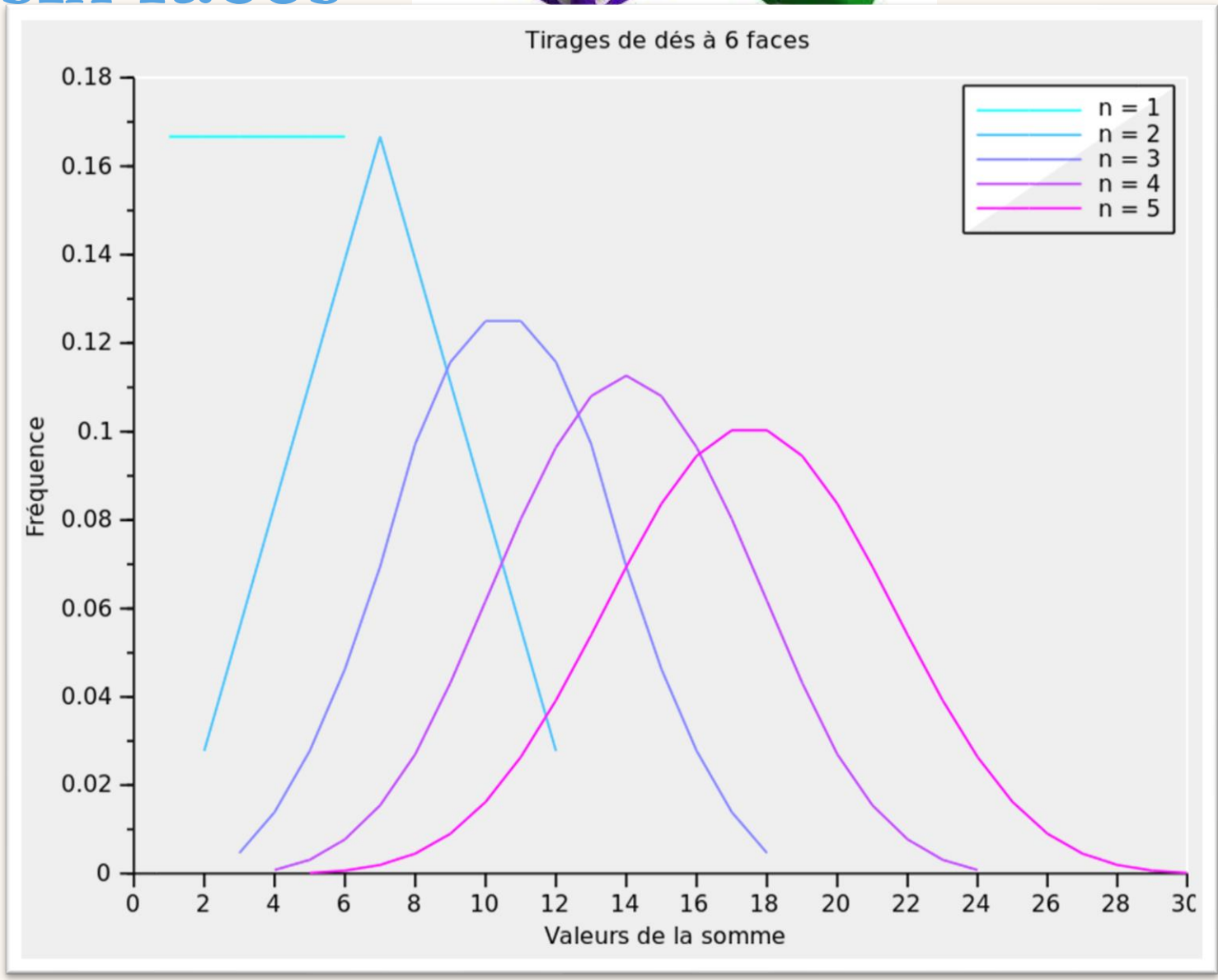
- on va jusque 12 tirages
  - n = 1 à 12
- on tend vers la courbe en cloche



Exemple : le dé à six faces

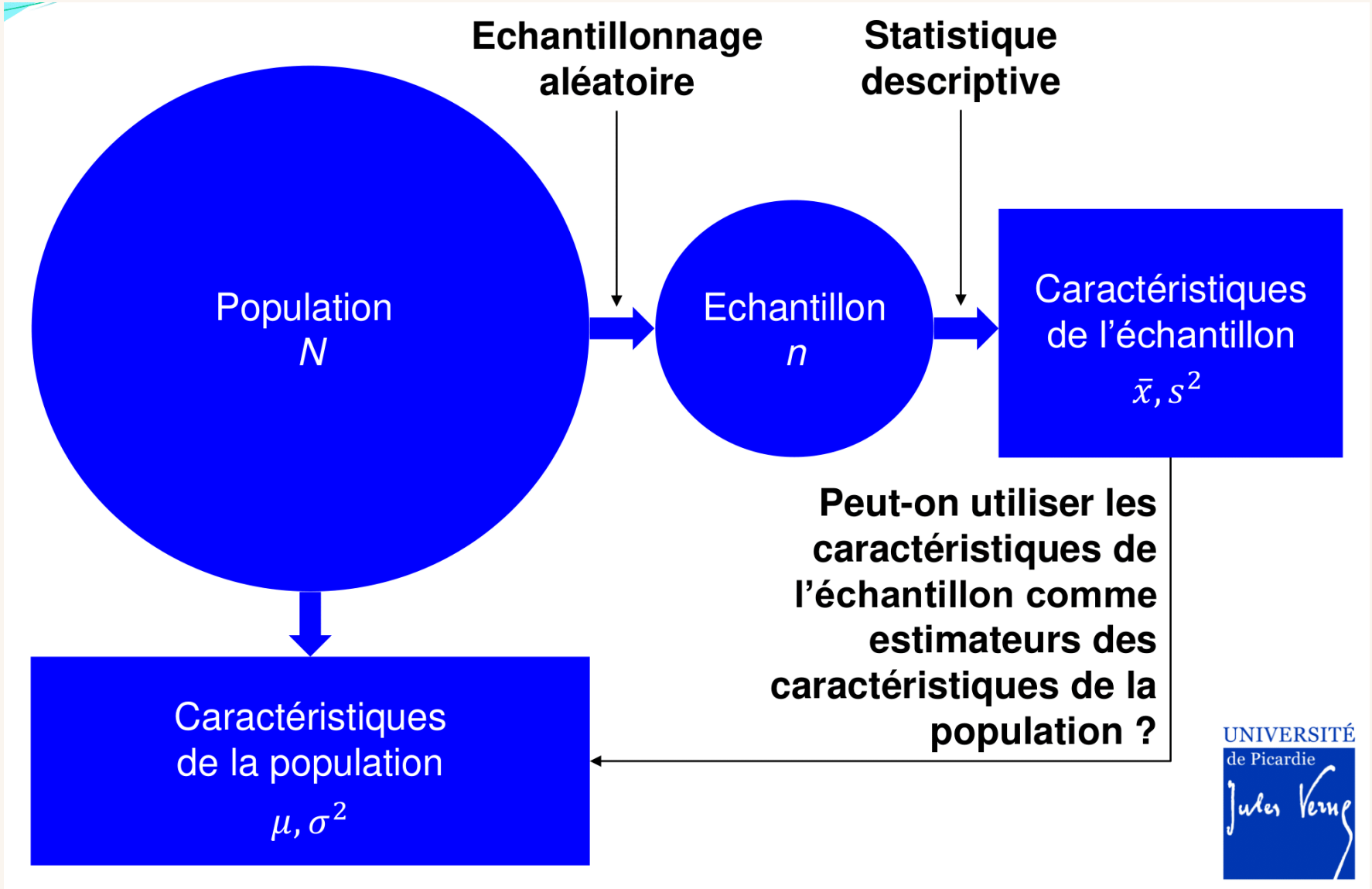


- on fait la somme des n tirages
- on va jusque 12 tirages
  - n = 1 à 12
- on tend vers la courbe en cloche



Indicateurs numériques

Concept d'estimation



UNIVERSITÉ de Picardie Jules Verne

Population

- Moyenne de la population de la variable d'intérêt :

$$\bar{Y} = \frac{1}{N}\sum_{i \in U} Y_i$$

- Variance $$\sigma^2 = \frac{1}{N}\sum_{i \in U}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{N}\sum_{i \in U} Y_i{}^2 - \bar{Y}^2$$

- Ces formules sont en fait celles utilisées au chapitre 01.
- Mais ici … au niveau de la population totale, elle sont inconnues. Leur intérêt est donc purement théorique.

Echantillon

- Moyenne estimée de la population , observée sur l'échantillon :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Variance (empirique) estimée de la population , observée sur l'échantillon :

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

  - Pourquoi n -1 ? Parce que c'est un échantillon.
  - La variance d'une population est plus grande que la variance d'un échantillon de cette population.
  - Une démonstration mathématique existe, qui sort du cadre de ce cours.

- Ces valeurs, au niveau de l'échantillon, sont connues et calculables.

Bilan de l'estimation

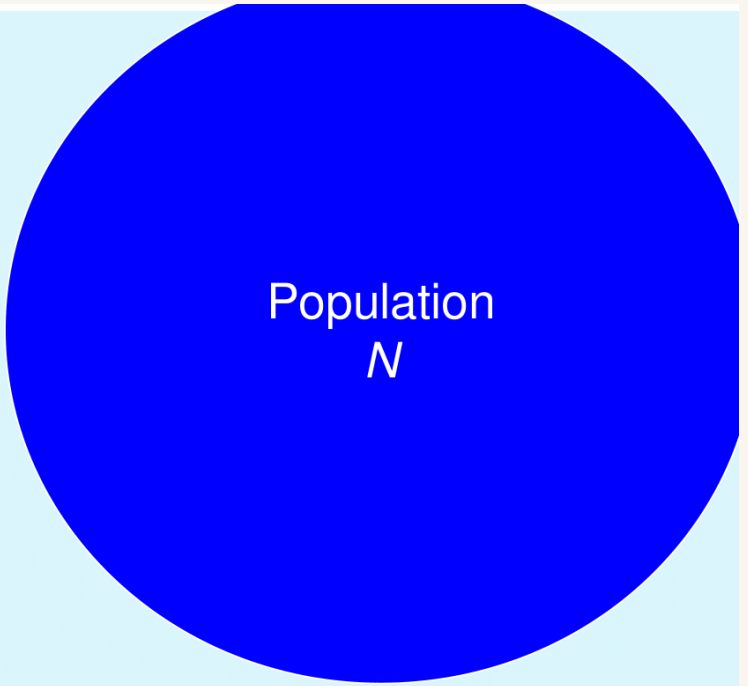
Statistique inférentielle :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$s^2 < \hat{\sigma}^2 \approx \sigma^2$ — Vrai si $n > 30$

$\bar{x} \approx \mu$

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



Statistique descriptive :

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



UNIVERSIT
de Picardie
Jules Vern

Probabilité égale avec ou sans remise

Probabilité égale avec remise = PEAR

Probabilité égale sans remise = PESR

Seul PESR est étudié au cours.

PEAR / PESR

- Sondage aléatoire simple
- Probabilité égale de tirer chacun des échantillons
- Tous les échantillons de taille n, prélevés au sein d'une population d'effectif N, sont réalisables avec la même probabilité.

Probabilité Egale Avec Remise (PEAR)

- Sondage aléatoire avec remise
- Principe
  - chaque unité de la population reçoit un numéro, de 1 à N, un "identifiant"
  - on choisit un nombre entre 1 et N, l'individu dont le numéro correspond à ce nombre est introduit dans l'échantillon
  - on recommence n fois l'opération, on obtient alors un échantillon de taille n , constitué de n individus, certains étant peut-être doublons.

nombre échantillons possibles = N * n

Probabilité Egale Sans Remise (PESR)

- Sondage aléatoire sans remise
- Principe
  - chaque unité de la population reçoit un numéro, de 1 à N, un "identifiant"
  - on choisit un nombre entre 1 et N
  - si ce numéro a déjà été tiré,
    alors l'individu correspondant n'est pas introduit à nouveau dans l'échantillon et on recommence le tirage
  - sinon, si ce numéro n'a pas encore été tiré,
    alors l'individu correspondant est introduit dans l'échantillon
  - on recommence n fois l'opération, on obtient alors un échantillon de n individus tous différents.

$$\text{nombre échantillons possibles} = C_n^p = \frac{n!}{(n-p)!p!}$$

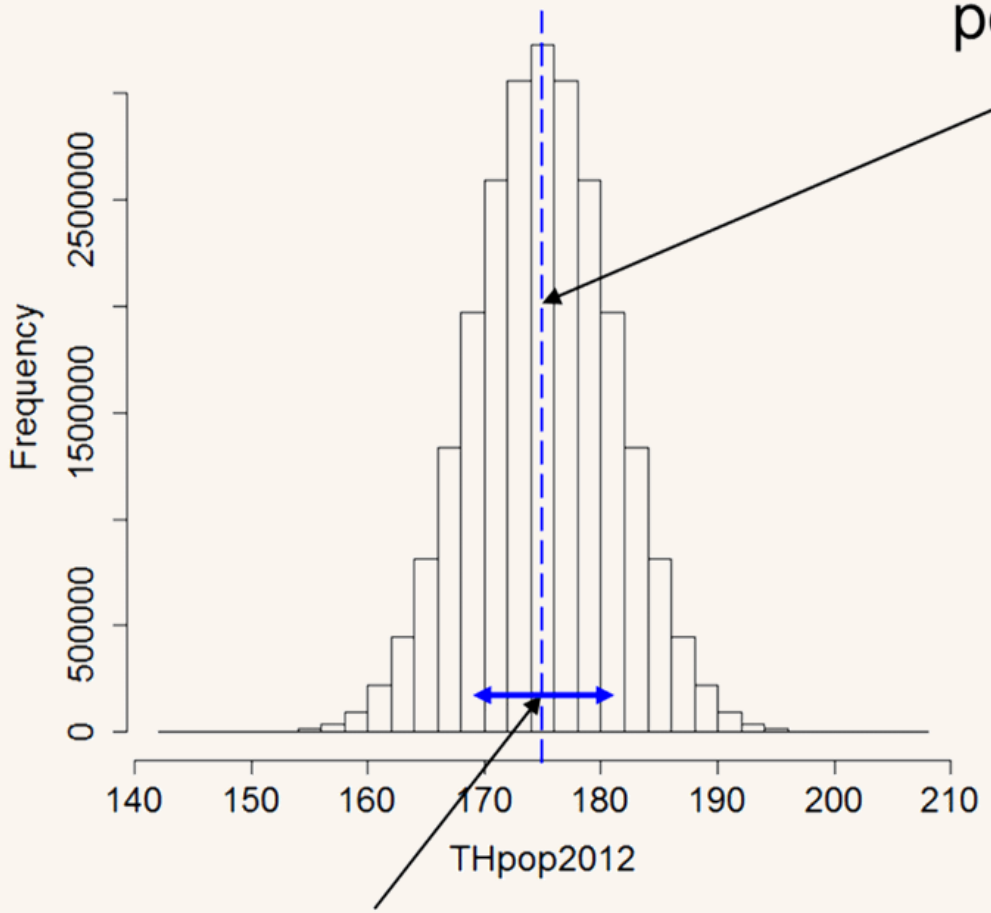
Précision de la moyenne estimée

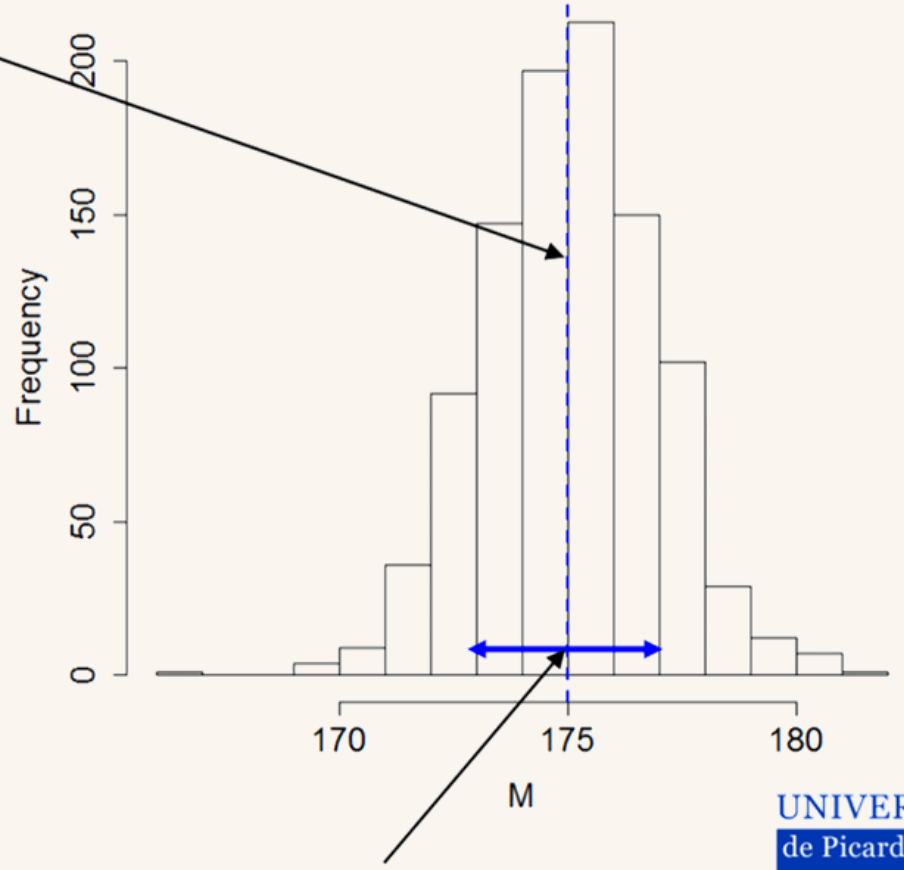
… à partir de la moyenne de l'échantillon.

Taille (cm) de 24,4 millions d'hommes adulte en France (population)

Moyennes des tailles (cm) de 1000 échantillons de 10 individus chacun

Moyenne de la population



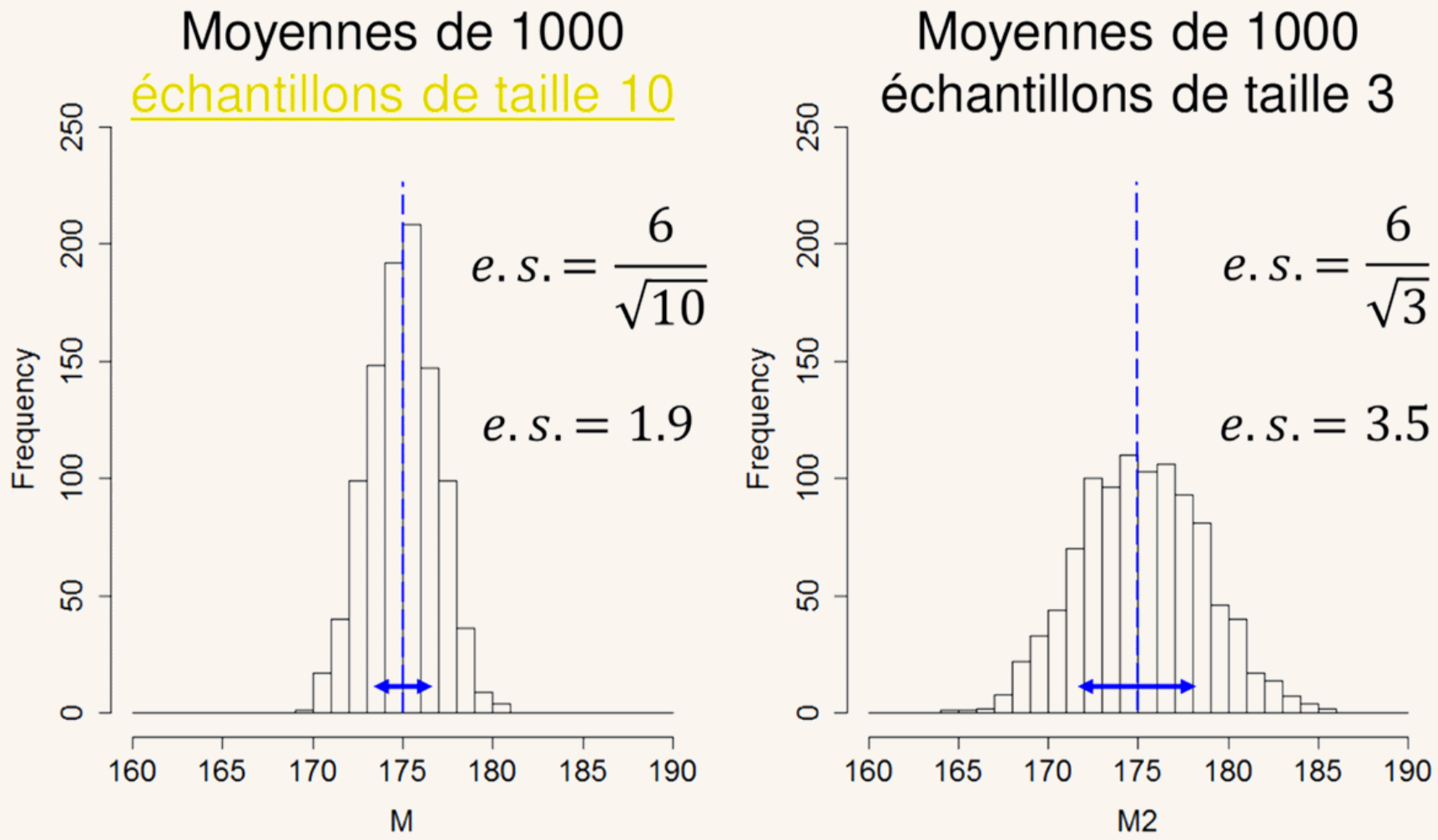


Ecart type des données autour de la moyenne

Erreur standard de la moyenne = ecart type

UNIVERSITÉ de Picardie Jules Verne

Comparez les deux distributions de moyenne ? Qu'en pensez-vous ?



L'erreur standard de la moyenne permet d'apprécier la précision de la moyenne et cette précision augmente avec la taille de l'échantillon

UNIVERSITÉ de Picardie Jules Verne

Problème de l'estimation

- Dans quelle mesure l'estimation (de la moyenne ou autre) d'une population, par l'usage d'un échantillon, est-elle précise ?
- Soit une population dont on voudrait estimer un paramètre Y à partir d'un échantillon x_1, x_2, …, x_n prélevé dans cette population.
  - Exemple : Y = salaire moyen de la population
- Cette estimation de Y sera donc une valeur qui dépendra de x_1, x_2, …, x_n et variera suivant l'échantillon choisi.
  - Dit autrement : Si vous recommencez p fois l'expérience, alors vous obtiendrez p estimations différentes de Y.
- L'estimation de Y pour chaque échantillon est elle aussi une variable aléatoire que nous appellerons "estimateur de Y".

Problème de l'estimation de la moyenne $\bar{Y}$ de la population à partir d'un échantillon aléatoire

- Il y a autant d'estimateur (différents) $\hat{\bar{Y}}$ que d'échantillons.
- Tous ces estimateurs sont dispersés autour de $\bar{Y}$ avec une variance $\hat{V}\left(\hat{\bar{Y}}\right)$ estimée elle aussi.

Notation	Statistique	PEAR	PESR
$\hat{\bar{Y}}$	Estimateur de la moyenne $\bar{Y}$	$\bar{y}$	$\bar{y}$
$\hat{V}\left(\hat{\bar{Y}}\right)$	Estimateur de la variance de cette moyenne (et non de la population !)	$\frac{s^2}{n}$	$(1-f).\frac{s^2}{n}$
$\hat{\sigma}(\hat{\bar{Y}})$	Estimateur de l'écart-type autour cette moyenne	$\sqrt{\frac{s^2}{n}}$	$\sqrt{(1-f).\frac{s^2}{n}}$

Exemple d'estimation (1/3)

Soit une population $U = \{U_1,\ U_2,\ U_3\}$ avec $Y_1 = 2$, $Y_2 = 4$, $Y_3 = 12$ → N=3

Quelle est la moyenne de la population ?

$$\overline{Y} = \frac{Y}{N} = \frac{2+4+12}{3} = \frac{18}{3} = 6$$

Quelle est la variance de la population ?

$$\sigma^2 = \frac{1}{N}\sum_{i\in U}\left(Y_i - \overline{Y}\right)^2 = \frac{1}{3}.(2-6)^2 + \frac{1}{3}.(4-6)^2 + \frac{1}{3}.(12-6)^2 = \frac{16}{3} + \frac{4}{3} + \frac{36}{3} = \frac{56}{3} = 18{,}7$$

Pour rappel, dans la réalité, ces valeurs ne sont pas calculables, la population étant trop nombreuse ou inaccessible.

Exemple d'estimation (2/3)

Soit le prélèvement d'un échantillon aléatoire de taille 2. Supposons que U_2 et U_3 aient été tirés.

Taux de sondage ?

$$f = \frac{n}{N} = \frac{2}{3}$$

Moyenne de l'échantillon ?

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{2}.(4 + 12) = \frac{16}{2} = 8$$

Dispersion empirique observée sur l'échantillon ?

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{1}.(4-8)^2 + \frac{1}{1}.(12-8)^2 = 16 + 16 = 32$$

Exemple d'estimation (3/3)

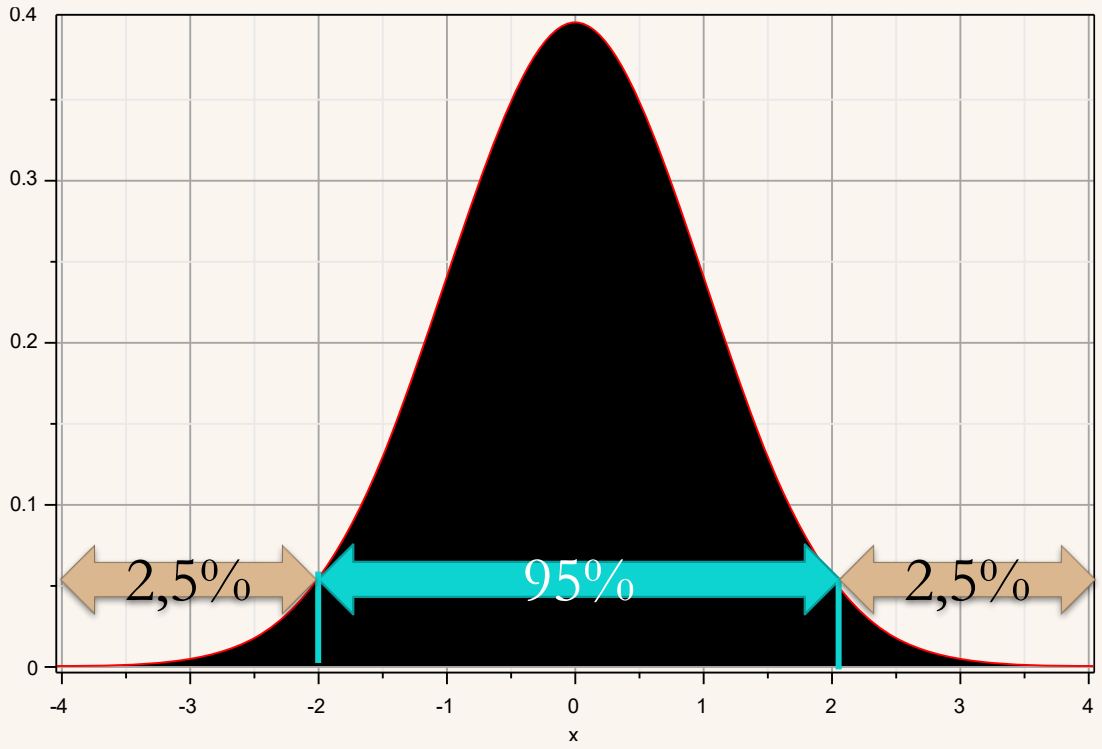
Statistique	PEAR	PESR
$\hat{\bar{Y}}$	8	8
Estimateur de la variance de la moyenne $\hat{V}\left(\hat{\bar{Y}}\right)$	$\frac{s^2}{n}$ $\frac{32}{2} = 16$	$(1-f).\frac{s^2}{n}$ $\left(1-\frac{2}{3}\right).\frac{32}{2} = \frac{1}{3}.\frac{32}{2} = \frac{16}{3} = 5{,}3$

Intervalle de confiance

Intervalle de confiance

- Il est d'usage de produire, pour une statistique estimée, un intervalle de confiance à un niveau de confiance donné.
  - cas habituel : niveau de confiance de 95%.
- Théorème central limite : rappel
  - La somme d'un grand nombre de variables aléatoires indépendantes se rapproche de la loi de Laplace-Gauss.

Intervalle de confiance



- Intuitivement, en se rappelant qu'il y a une probabilité de 95% que la moyenne $\bar{Y}$ se trouve dans un intervalle d'environ 2 écarts-types autour de l'estimation de la moyenne $\hat{\bar{Y}}$, on peut écrire :

$$I_c \cong \hat{\bar{Y}} \pm 2 \,.\, \hat{\sigma}(\hat{\bar{Y}})$$

ou

$$\bar{Y} \in [\hat{\bar{Y}} - 2 \,.\, \hat{\sigma}(\hat{\bar{Y}}); \hat{\bar{Y}} + 2 \,.\, \hat{\sigma}(\hat{\bar{Y}})]$$

à 95% de chances

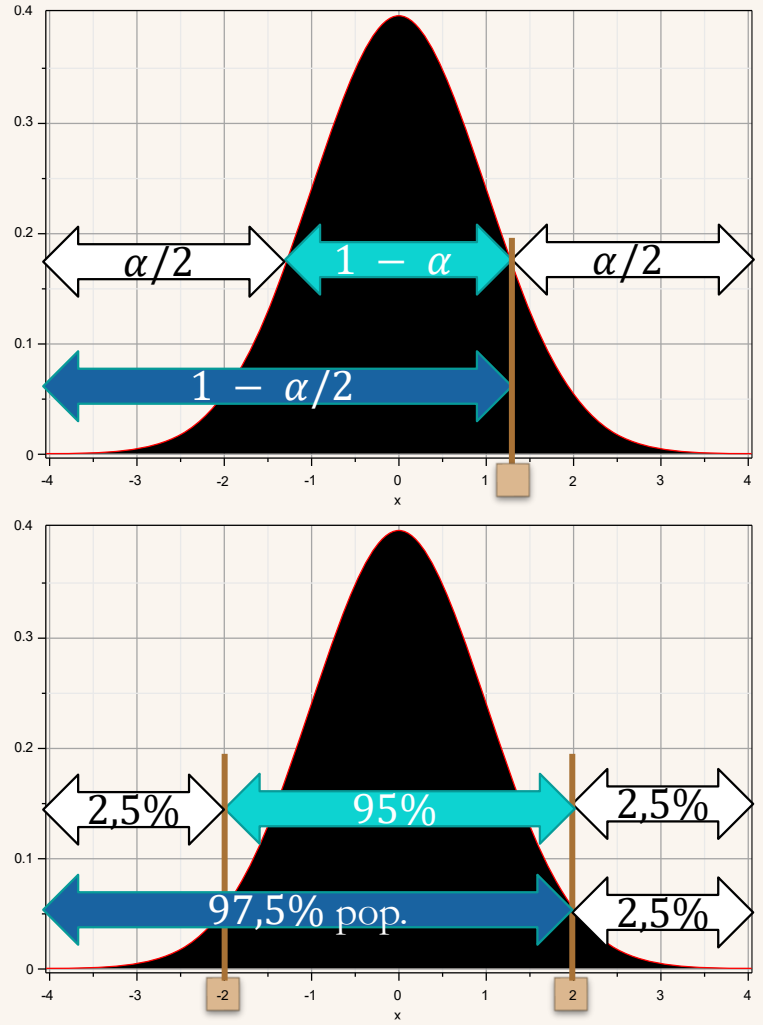
Intervalle de confiance

- L'usage parle du niveau de confiance "1 - α"
  - $\alpha \in]0,1[$
  - Par exemple, niveau de confiance à $1 - \alpha = 95\%$ signifie

$$1 - \alpha = 0{,}95 \quad \alpha = 0{,}05 \rightarrow \frac{\alpha}{2} = 0{,}025$$

$$\rightarrow 1 - \frac{\alpha}{2} = 0{,}975$$

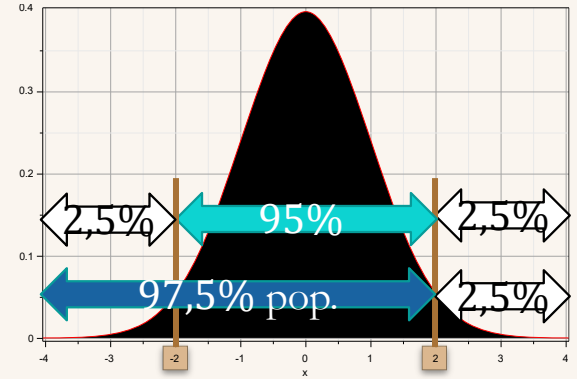
- $Z_{1-\frac{\alpha}{2}}$ = quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite
  - Par exemple, $Z_{0,975} = 1{,}96$
- L'intervalle de confiance au niveau de confiance 1 - α est donné par : $I_c = \bar{Y} \pm Z_{1-\frac{\alpha}{2}}.\hat{\sigma}(\bar{Y})$



Intervalle de confiance : exemple avec tableur (1/2)

- Recherchons la valeur d'un quantile si le niveau de confiance est de 95%

$$1 - \alpha = 0{,}95 \ \ldots$$
$$1 - \alpha/2 = 0{,}975$$



- Fonction "quantile" dans le tableur
  - La fonction "distribution normale" du tableur :
    - gSheet : **`NORMSINV( x )`**
    - Excel FR : **`LOI.NORMALE.STANDARD.INVERSE.N( x )`**
  - On donne 0,975 en paramètre **`x`**
  - La valeur du quantile correspondant s'affiche : $Z_{1-\alpha/2} = Z_{0{,}975} = \mathbf{1{,}9600}$
- Si nous cherchons la moyenne à partir d'un échantillon PESR, l'intervalle de confiance à 95% est donné par :

$$I_c = \hat{\bar{Y}} \pm 1{,}96 \sqrt{(1-f).\frac{s^2}{n}}$$

Intervalle de confiance : exemple avec tableur (2/2)

- On veut estimer le nombre moyen de T-shirt possédés par les étudiants de l'ISFCE (= variable d'intérêt).
- Intervalle de confiance à 95%

- Caractéristiques :
  - nombre d'étudiants : 1000
  - nombre d'étudiants sondés : 50
  - moyenne échantillon : 47,2
  - variance échantillon : 196

$$\hat{V}(\bar{y}) = (1-f).\frac{s^2}{n} = \left(1-\frac{50}{1000}\right).\frac{196}{50} = \frac{950}{1000}.\frac{196}{50} = 3{,}724$$

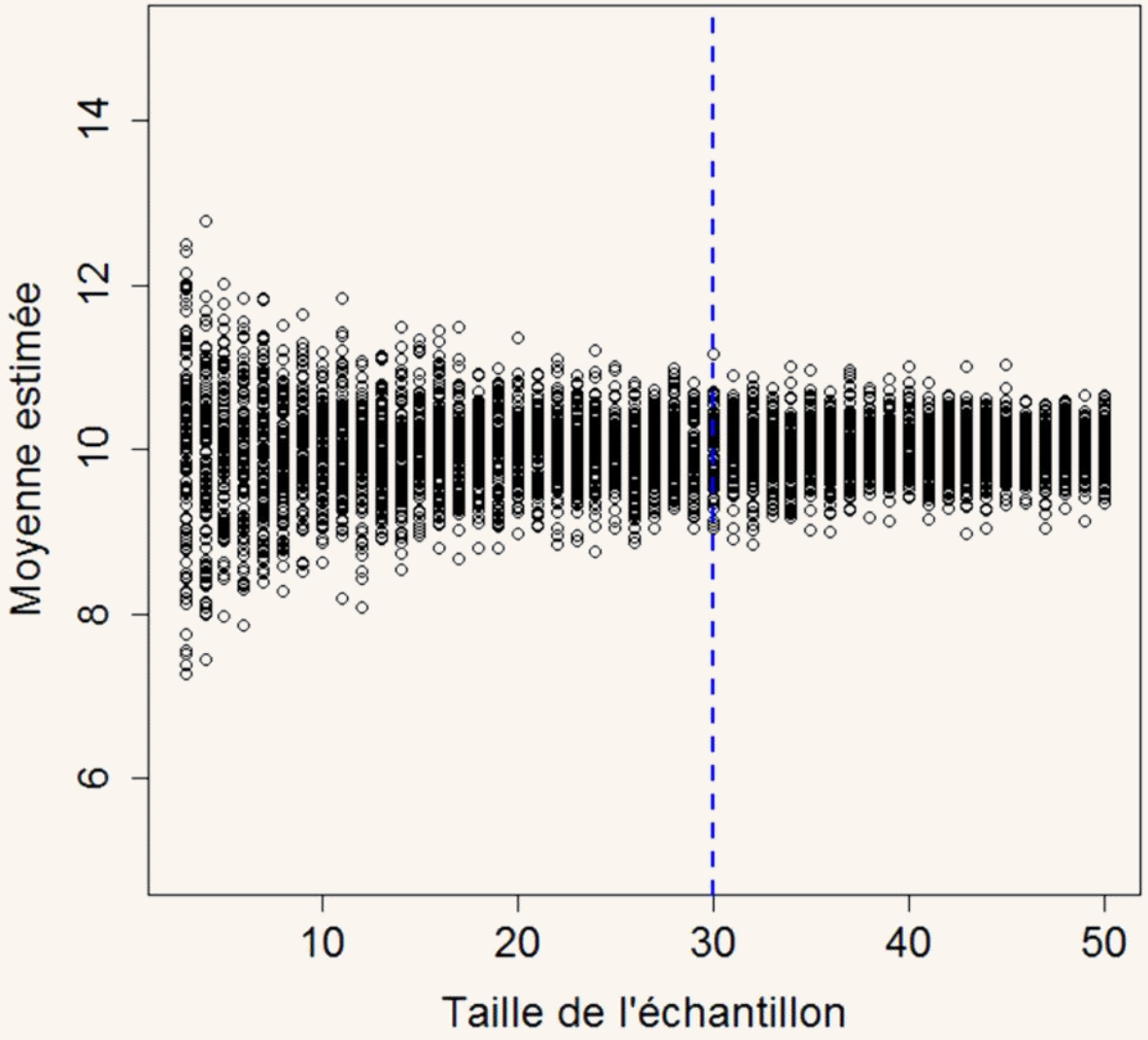
$$I_c = 47{,}2 \pm 1{,}96\sqrt{3{,}724} = 47{,}2 \pm 1{,}96.1{,}93 = 47{,}2 \pm 3{,}8 = [43{,}4; 51]$$

Petits échantillons

Imprécision sur la variance

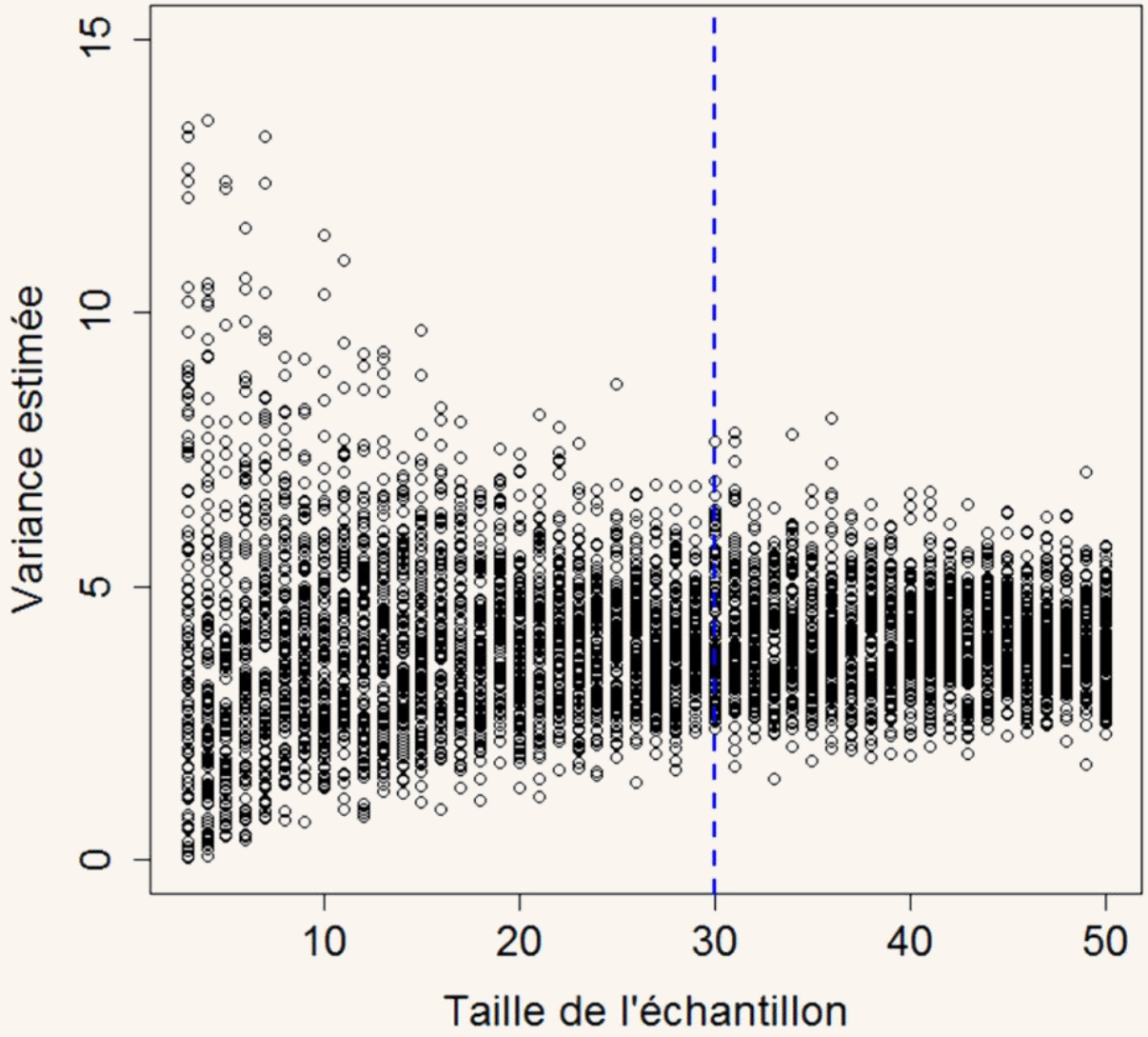
Loi de Student

Petits échantillons : estimation moyenne



UNIVERSITÉ de Picardie Jules Verne

Petits échantillons : estimation variance



UNIVERSITÉ de Picardie Jules Verne

Quid des petits échantillons ?

- Si n descend en dessous de 30
  - la sous-estimation de σ^2 devient trop important
  - l'erreur sur l'estimateur de la moyenne augmente fortement
  - l'intervalle de confiance devient imprécis
  - on va sous-estimer l'étendue réelle de l'intervalle de confiance
- La loi de Student corrige ce phénomène.

Précision et taille d’un échantillon

Précision

- Précision absolue (en unité)

$$P_a = Z_{1-\frac{\alpha}{2}} . \sqrt{V(\hat{\bar{Y}})} = P_r . \hat{\bar{Y}}$$

- Précision relative (en %)

$$P_r = \left[Z_{1-\frac{\alpha}{2}} . \sqrt{V(\hat{\bar{Y}})} \right] . \frac{1}{\hat{\bar{Y}}} = \frac{P_a}{\hat{\bar{Y}}}$$

- On peut écrire alors :

$$I_c = \hat{\bar{Y}} \pm Z_{1-\frac{\alpha}{2}} . \hat{\sigma}\left(\hat{\bar{Y}}\right) = \hat{\bar{Y}} \pm P_a = \hat{\bar{Y}} . (1 \pm P_r)$$

Taille d'un échantillon PEAR

- Comment calculer le nombre d'individus n à interroger pour obtenir une précision absolue convenue P_a ?
- En mode PEAR, voici l'équation :

$$Z_{1-\frac{\alpha}{2}}.\sqrt{\widehat{V(\hat{\bar{Y}})}} = Z_{1-\frac{\alpha}{2}}.\sqrt{\frac{s^2}{n}} = Z_{1-\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} = P_a$$

$$n = \left(Z_{1-\frac{\alpha}{2}}.\frac{s}{P_a}\right)^2$$

Taille d'un échantillon PESR

- En mode PESR, voici l'équation :

$$Z_{1-\frac{\alpha}{2}}.\sqrt{\widehat{V(\hat{\bar{Y}})}} = Z_{1-\frac{\alpha}{2}}.\sqrt{(1-f)\frac{s^2}{n}} = P_a \text{ avec } f = \frac{n}{N}$$

- Après quelques calculs indignes d'un étudiant du supérieur, on trouve :

$$\left(1-\frac{n}{N}\right).\frac{1}{n} = \left(\frac{P_a}{s\,.\,Z_{1-\frac{\alpha}{2}}}\right)^2 \rightarrow \boxed{n = \frac{(s.Z_{1-\frac{\alpha}{2}})^2.N}{N.P_a{}^2+(s.Z_{1-\frac{\alpha}{2}})^2}}$$

- Si N est beaucoup plus grand que n (au moins 100 fois), alors on peut simplifier la formule :

$$\boxed{n = \left(Z_{1-\frac{\alpha}{2}}.\frac{s}{P_a}\right)^2}$$

Taille d'un échantillon : exemple

- Si on désire une précision absolue de 5 pour un intervalle de confiance de 90%, dont on connaît la valeur S = 34 d'un échantillon PEAR, de combien d'individus a-t-on besoin ?

$1 - \alpha = 0{,}90$
$\alpha = 0{,}10$
$\alpha/2 = 0{,}05$
$1 - \alpha/2 = 0{,}95$

$$Z_{1-\alpha/2} = Z_{0{,}95} = 1{,}645$$

$$n = \left(1{,}645.\frac{34}{5}\right)^2 = 125{,}1$$

Soit 126 personnes à interroger

Estimation d’une proportion

Proportion & Population

- Soit une population U partagée en deux classes K1 et K2.
- Posons P , la proportion de valeurs dans K1. (0<P<1)
- Considérons la variable Y définie ainsi

$$Y = \begin{cases} 1 \; si \; U_i \; appartient \; \text{à} \; K1 \\ 0 \; dans \; le \; cas \; contraire \end{cases}$$

- $\bar{Y}$ (moyenne de la variable Y) = P
  - on note aussi Q = 1 - P
- Vu que $Y_i^2 = Y_i$, on a aussi :

$$\sigma^2 = \frac{1}{N}\sum Y_i^2 \; - \; \overline{Y^2} = P \; - P^2$$

$$\sigma^2 = P \, . \, (\, 1 \; - P \,) = P \, . \, Q$$

Proportion & Échantillon

- La moyenne et la variance de la population totale peuvent être estimées par les formules suivantes :

$$\hat{P} = p = \bar{y} = \frac{1}{n}.\sum_{k=1}^{n} y_k \qquad \text{avec} \qquad y_k = \begin{cases} 1 \; si \; U_k \; appartient \text{ à } K1 \\ 0 \; si \; U_k \; appartient \text{ à } K2 \end{cases}$$

$$\hat{Q} = 1 - p$$

Statistique	PEAR	PESR
$V\left(\hat{\bar{Y}}\right)$	$\frac{\sigma^2}{n} \longrightarrow \frac{PQ}{n}$	$(1-f).\frac{S^2}{n} \longrightarrow \frac{N-n}{N-1}\frac{P.Q}{n}$
$\hat{V}\left(\hat{\bar{Y}}\right)$	$\frac{s^2}{n} \longrightarrow \frac{pq}{n-1}$	$(1-f).\frac{s^2}{n} \longrightarrow (1-f)\frac{p.q}{n-1}$

Exemple (cf complément)

- On s'intéresse à une variable représentant le soutien à un programme politique X, en Belgique. On interroge 1000 personnes (PESR) parmi les 10.000.000 de Belges. Après dépouillement des résultats, on constate que ce programme est soutenu par 20% des personnes interrogées. Le but est de trouver un intervalle de confiance autour de cette valeur, si on travaille à du 95%.

Exemple

p = 0,20 et donc q = 1 - p = 0,80

→

$$\hat{V}(\hat{P}) = (1-f)\frac{p.q}{n-1} = \left(1-\frac{1000}{1000000}\right)\frac{0{,}20.0{,}80}{1000-1} = \frac{9999000}{10000000}\frac{0{,}16}{999}$$

$$= 0{,}00016$$

→

$$I_c = \hat{P} \pm Z_{1-\alpha/2}.\sqrt{\hat{V}(\hat{P})} = 0{,}20 \pm 1{,}96.\sqrt{0{,}00016}$$

$$= 0{,}20 \pm 1{,}96\ .\ 0{,}01265. = 0{,}20 \pm 0{,}025$$

$$= [0{,}175; 0{,}225]$$