

Inférence statistique

Cours 7

Conventions & symboles

- ◆ Lorsqu'on se réfère à la population-parente, on utilisera les paramètres qui seront désignés par des lettres grecques
 - ✓ μ (mu) qui désigne la moyenne,
 - ✓ σ^2 (sigma minuscule carré) qui désigne la variance,
 - ✓ σ (sigma minuscule) qui désigne l'écart-type de la population.

- ◆ Lorsqu'on se réfère à des échantillons, on utilisera les paramètres qui seront désignés par des lettres latines
 - ✓ m qui désigne une moyenne calculée sur un échantillon,
 - ✓ s^2 qui désigne la variance de l'échantillon,
 - ✓ s qui désigne l'écart-type de cet échantillon.

Introduction: le raisonnement inférentiel

◆ Définition

Inférer: généraliser les résultats obtenus sur un échantillon à l'ensemble de la population parente dont cet échantillon est extrait

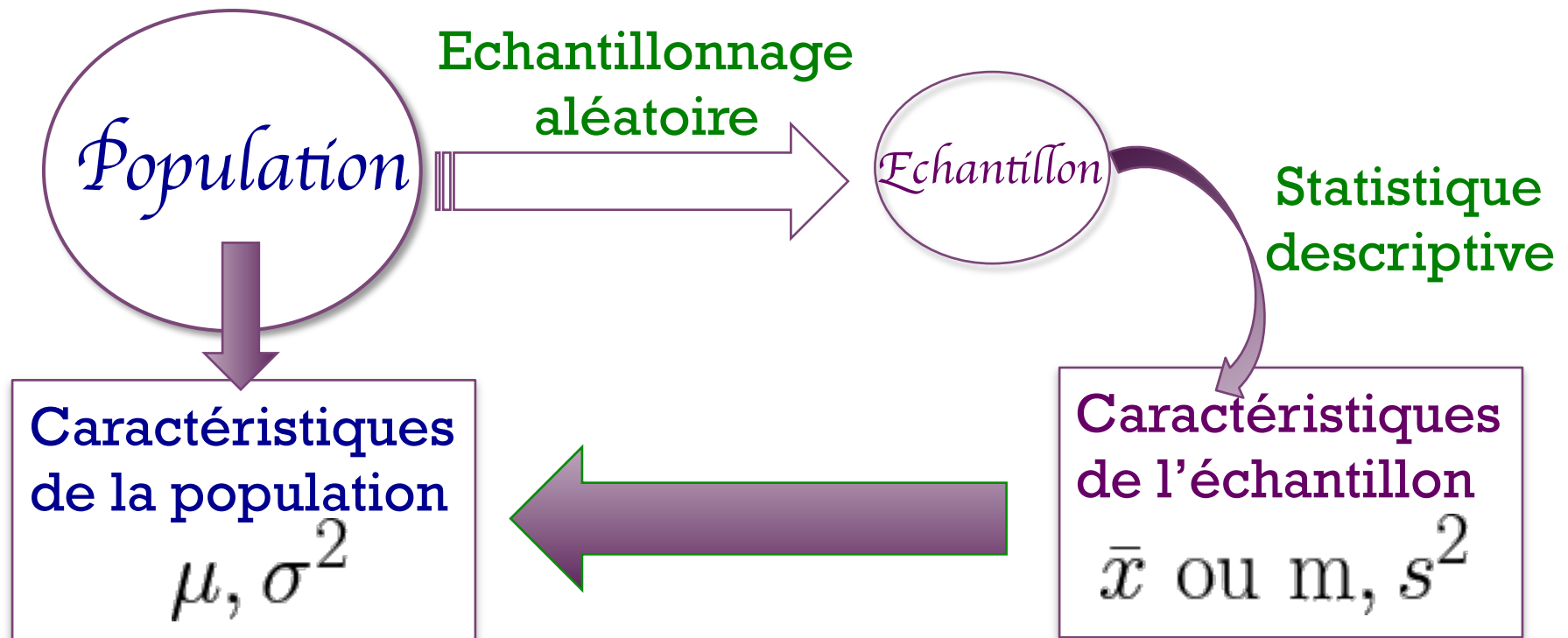
◆ Généralisation : possible seulement sous certaines conditions qui doivent être vérifiées par des tests statistiques

◆ L'objectif est double

- ✓ Confirmer (ou infirmer) que l'échantillon appartient bien à la population générale
- ✓ Dans le cas contraire, déterminer la population parente réelle

◆ L'inférence sur une moyenne constitue le type d'inférences le plus fréquent dans les sciences humaines

Inférences sur une moyenne



- ◆ Il s'agit de comparer une moyenne m obtenue sur un échantillon avec la moyenne correspondante μ dans la population parente

Intermède

(écart-type population & échantillon)

- ◆ Soit la moyenne des x_i de la population

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

Le $1/N$ n'est appliqué que sur le 1^{er} terme

- ◆ L'écart-type σ de la population

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

ou

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

Intermède

(écart-type population & échantillon)

- ◆ Soit m_1 la moyenne des x'_i de l'échantillon 1

$$m = \frac{1}{n} \sum_{i=1}^n x'_i$$

Echantillon 1
 x'_1, x'_2, \dots, x'_n

- ◆ Dès que nous opérons sur de échantillons avec des effectifs plus restreints on utilisera l'écart type (corrigé)

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

- ◆ A la limite où l'on effectue une mesure ($n=1$), l'écart type s est alors indéfini, ce qui reflète effectivement notre ignorance de l'incertitude commise après une seule mesure

Intermède

(écart-type population & échantillon)

$$s^2 = \frac{n-1}{n} \sigma^2$$

- ◆ La moyenne des variances d'échantillon n'est pas la variance de la population, mais la variance de la population multipliée par $(n-1)/n$

Exemple

- ◆ On a relevé l'âge (ans) des 6 lézards d'un zoo : 1,2,2,1,3,3.
 - ✓ Quel est l'âge moyen des lézards du zoo ?
 $m=2,0$ ans
 - ✓ Quel est l'écart-type de l'âge des lézards ?
 $s=0,82$ ans
On utilise ici la formule avec le dénominateur en $1/N$ car c'est toute la population dont il est question.

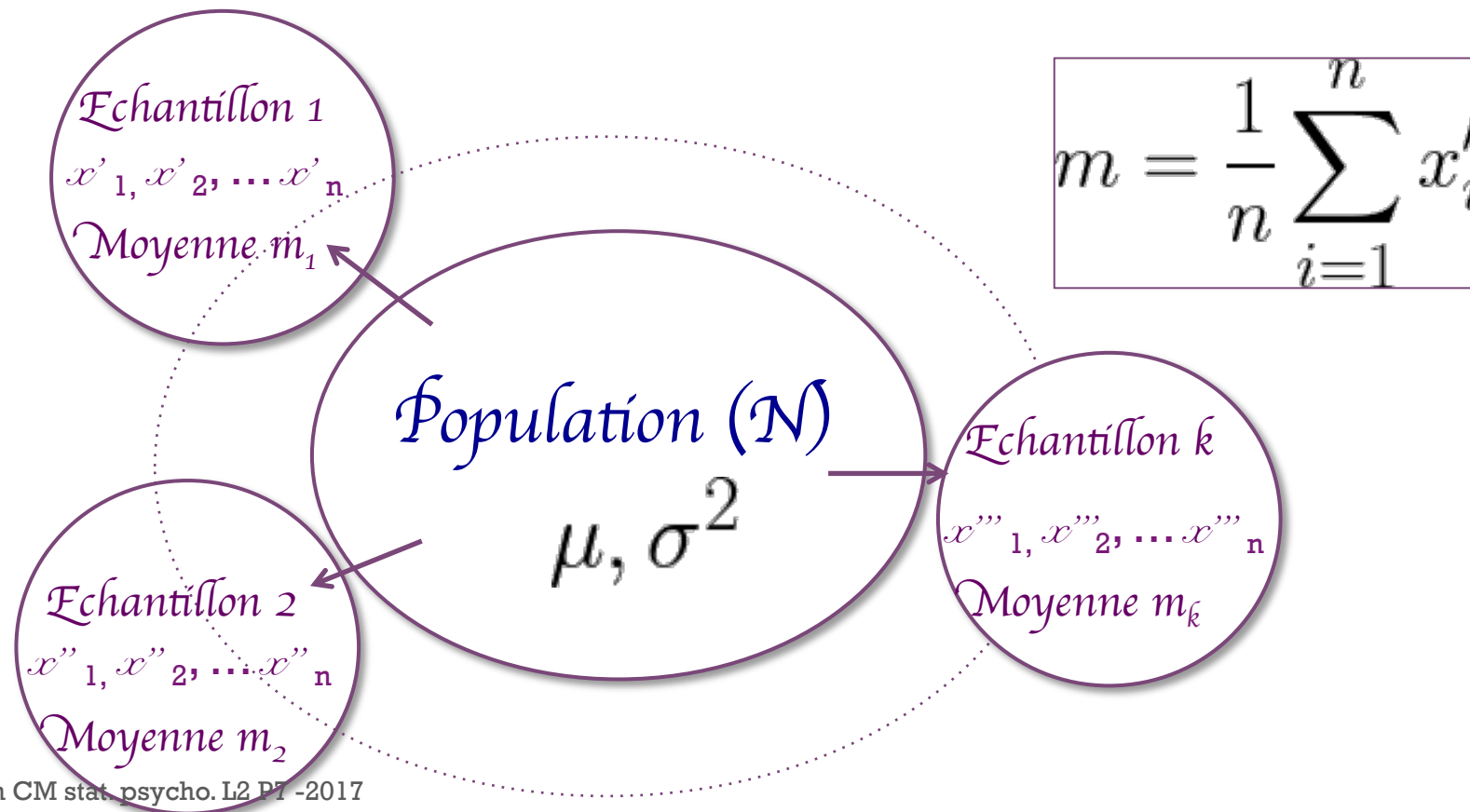
- ◆ Toujours dans le même zoo on a choisi au hasard 5 zèbres parmi les 42 que compte le zoo et on a relevé leur âge (ans): 8,11, 17, 7, 19.
 - ✓ D'après cet échantillon quel est l'âge moyen des zèbres du zoo?
 $m=12,4$ ans
 - ✓ Quel est l'écart-type de l'âge des zèbres ?
 $s=5,36$ ans on prend ici $1/(n-1)$ car c'est un échantillon de la population.

Distribution d'échantillonnage de la moyenne

- ◆ L'objectif de l'inférence statistique est d'estimer avec le moins d'erreur possible les paramètres la moyenne et l'écart type d'une population.
- ◆ Ils utilisent pour cela des échantillons tirés de la population
- ◆ La moyenne de l'échantillon constituera une estimation de la moyenne de la population.
- ◆ Il est important de garder à l'esprit que la moyenne d'un échantillon est une variable aléatoire, elle varie d'un échantillon à l'autre.

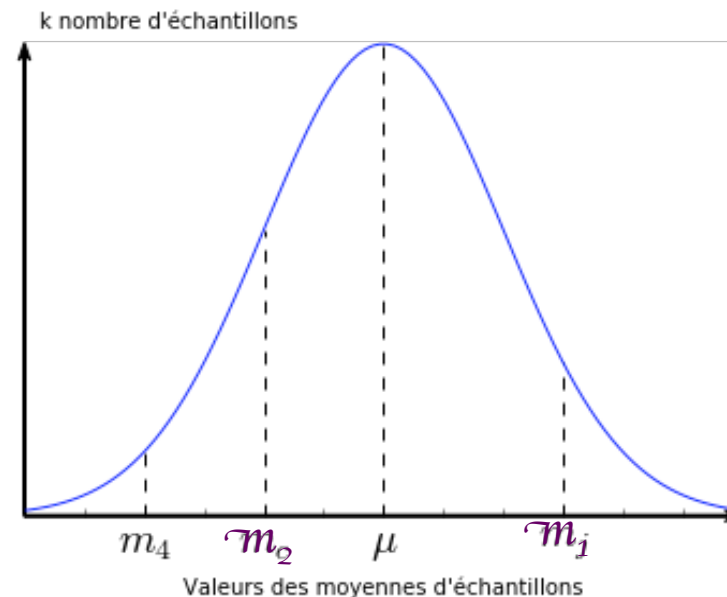
Distribution d'échantillonnage de la moyenne

- ◆ Si, d'une population parente normale, nous tirions au hasard un nombre élevé d'échantillons k avec tous le même effectif n , leurs moyennes (m_1, \dots, m_k) vont se distribuer normalement



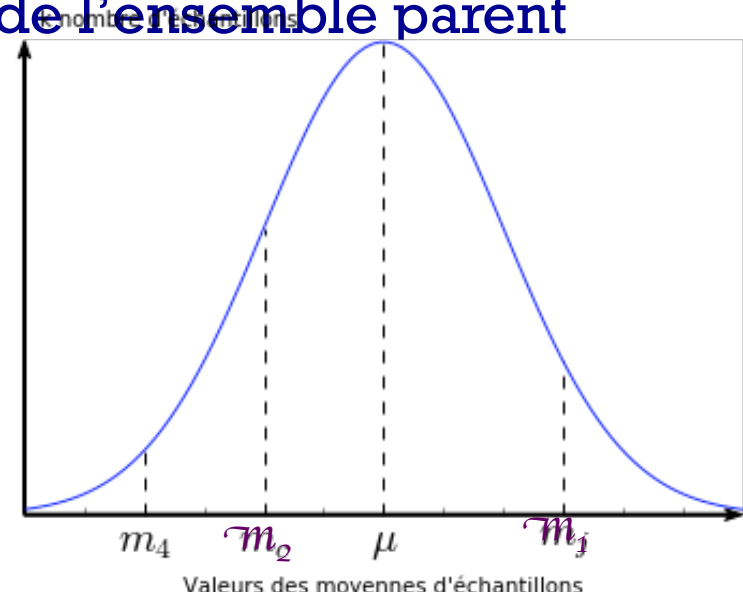
Distribution d'échantillonnage de la moyenne

- ◆ Les statistiques-moyennes ($m_1, m_2, m_3 \dots m_k$) vont s'organiser dans une distribution dite distribution d'échantillonnage des moyennes ou distribution des moyennes échantillonnales
- ◆ Remarquer qu'il s'agit ici d'une distribution normale constituée par des moyennes m d'échantillons



Distribution d'échantillonnage de la moyenne

- ◆ Les statistiques-moyennes ($m_1, m_2, m_3 \dots m_k$) vont s'organiser dans une distribution dite distribution d'échantillonnage des moyennes ou distribution des moyennes échantillonnales
- ◆ Remarquer qu'il s'agit ici d'une distribution normale constituée par des moyennes m d'échantillons
- ◆ La moyenne de cette distribution des moyennes est la meilleure estimation de la moyenne μ de l'ensemble parent
- ◆ Estimateur non-biaisé et convergent
 - ✓ Plus k augmente et plus μ se rapproche de la "vraie" moyenne de l'ensemble parent.
 - ✓ A la limite $k \rightarrow \infty$ la moyenne de la distribution des moyennes est la même que la moyenne de la population.



L'erreur-type de la distribution d'échantillonnage

- ◆ L'écart-type de la distribution d'échantillonnage est appelée erreur-type ou erreur standard
- ◆ Cette grandeur représente l'amplitude des fluctuations aléatoires constatées sur les échantillons
 - ✓ En désignant par σ_m^2 la variance de la distribution d'échantillonnage on a

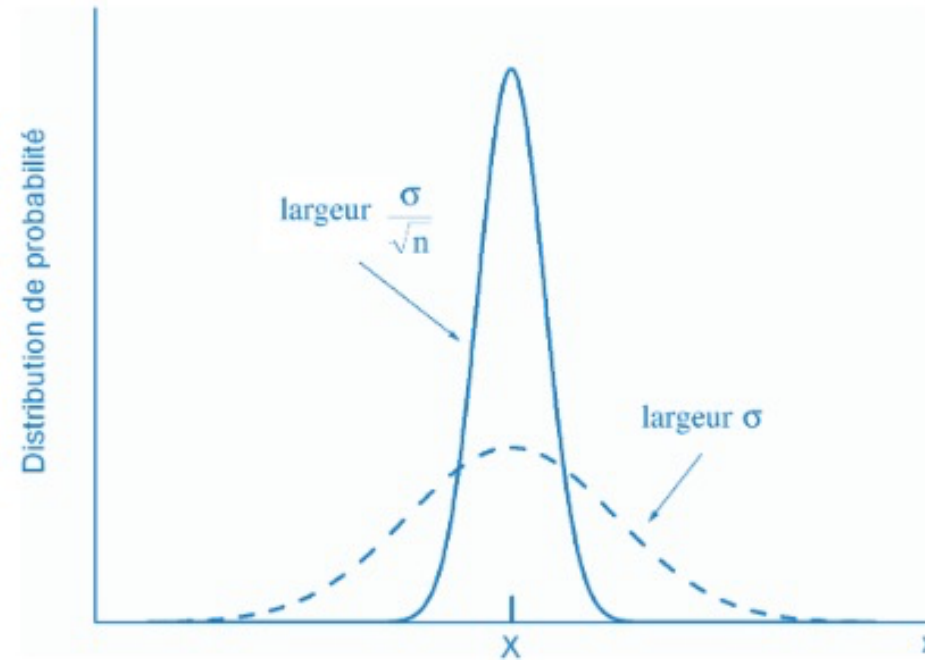
$$\sigma_m^2 = \frac{\sigma^2}{n}$$

- ✓ La variance de la distribution d'échantillonnage pour des échantillons indépendants de même effectif N est égale à la variance de la population divisée par l'effectif de l'échantillon

L'erreur-type de la distribution d'échantillonnage

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

- ◆ Au niveau des écarts type on a donc
- ◆ Plus n est grand plus il devient probable que la moyenne d'un échantillon particulier se rapproche de la moyenne de la population: c'est la loi des grands nombres
- ◆ Il s'agit de l'écart-type de la moyenne
- ◆ σ_m représente l'incertitude sur la détermination de la valeur vraie m à partir de n mesures, cette détermination est donc \sqrt{n} fois plus précise que celle obtenue à partir d'une mesure unique
- ◆ Si l'on extrait d'une population de moyenne μ et d'écart-type σ un échantillon de taille n , la moyenne de cet échantillon est une variable aléatoire de moyenne μ et d'écart-type σ/\sqrt{n}



- ◆ Lorsqu'on effectue une mesure unique, la valeur trouvée suit la distribution de probabilité représentée en pointillés.
- ◆ Si on réalise plusieurs déterminations de la moyenne de n mesures, elles suivent la distribution en trait plein.

L'erreur-type de la distribution d'échantillonnage

- ◆ Cependant dans une majorité de cas, σ est inconnu et doit être estimé à partir de s_x

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

$$\sigma_m = \frac{s}{\sqrt{n}}$$

Exemple

- ◆ On demande à 8 étudiants de psycho. de mesurer avec un chrono le temps entre deux images successives projetées sur un écran. Calculer l'écart type de la moyenne.

i (n° étudiant)	1	2	3	4	5	6	7	8
Temps trouvée (ms)	202	201	210	199	204	208	201	198

Le temps moyen est $m=202,90$ ms.

Pour calculer l'écart-type

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

$s_x = 4,22$ ms, soit $202,90 \pm 4,22$ ms.

L'incertitude sur toute nouvelle mesure du temps vaut dorénavant 4,22 ms.

C'est un intervalle de confiance sur les valeurs.

Exemple

- ◆ Pour calculer l'écart type de la moyenne on doit d'abord calculer σ

$$\sigma_m = \frac{s}{\sqrt{n}}$$

L'écart-type de la moyenne vaut donc $4,22/\sqrt{8} = 1,49$ ms

La valeur moyenne et son écart-type vaut donc:

$$202,90 \pm 1,50 \text{ ms}$$

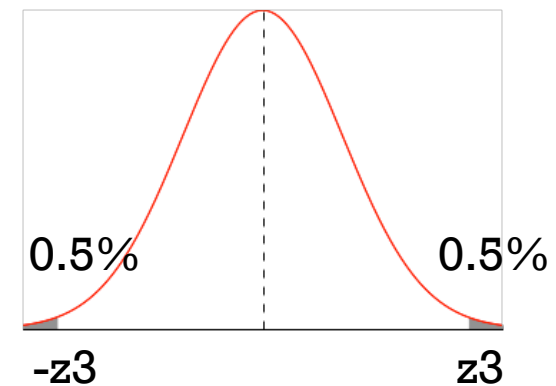
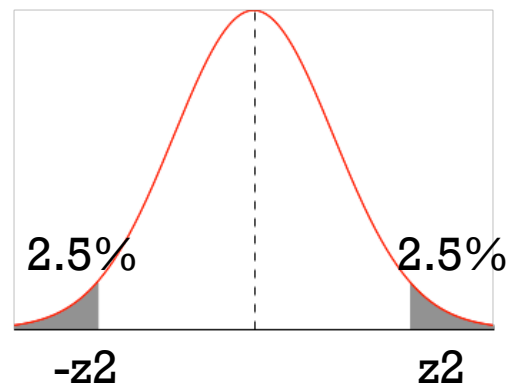
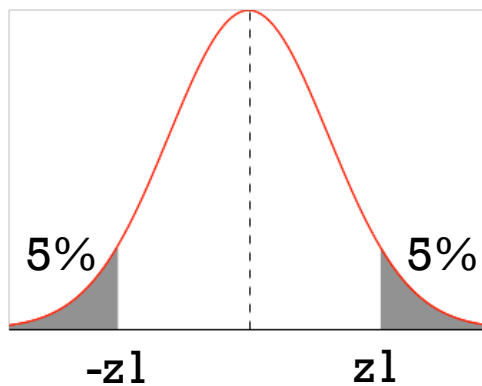
C'est un intervalle de confiance sur la moyenne des valeurs

Seuil de signification

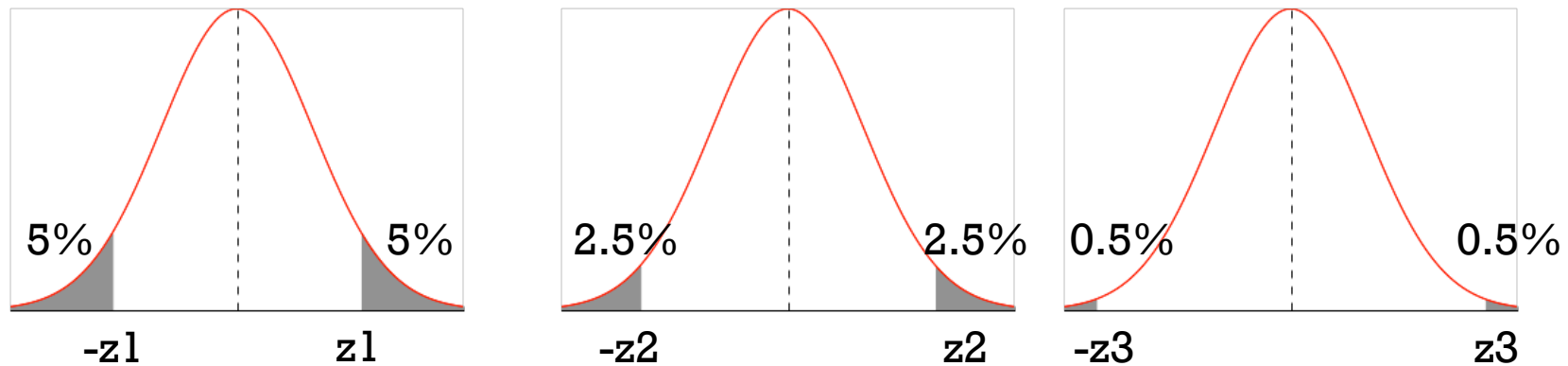
- ◆ Lorsque l'on connaît une distribution d'échantillonnage, les différentes estimations tirées des échantillons se répartissent suivant un modèle de distribution qui nous permet de connaître les probabilité d'apparition de chacune d'elles
- ◆ Pour éviter de retenir des estimations très peu probables, on va fixer une limite au-delà de laquelle il convient d'éliminer des valeurs qui du fait de leur rareté, risquent de conduire à des estimations aberrantes des paramètres
- ◆ Limite est appelée seuil de probabilité et est désignée par le symbole α

Seuil de signification

- ◆ On psychologie, les seuils les plus couramment employés sont:
 - ✓ $\alpha_1 = 0.10$ soit $z_1=1.65$: 10% des valeurs extrêmes sont écartées
 - ✓ $\alpha_2 = 0.05$ soit $z_1=1.96$: 5% des valeurs extrêmes sont écartées
 - ✓ $\alpha_3 = 0.01$ soit $z_3=2.58$: 1% des valeurs extrêmes sont écartées



Intervalle de confiance de la moyenne



Pour trouver z_1 tel que $P(-z_1 < z < z_1) = 0.9$ avec la table des z unilatérale on utilise $P(-z_1 < z < z_1) = 2 \times P(z < z_1) - 1$ (c.f. cours 6)

- ◆ $P(-z_1 < z < z_1) = 0.9 \Rightarrow P(z < z_1) = (1 + 0.9) / 2 = 0.95$ soit $z_1 = 1,65$
- ◆ $P(-z_1 < z < z_1) = 0.95 \Rightarrow P(z < z_1) = (1 + 0.95) / 2 = 0.975$ $z_1 = 1,96$
- ◆ $P(-z_1 < z < z_1) = 0.99 \Rightarrow P(z < z_1) = (1 + 0.99) / 2 = 0.995$ $z_1 = 2,58$

Intervalle de confiance de la moyenne

◆ Intervalle de confiance de μ lorsque σ est connu

À partir d'une moyenne d'échantillon observée m , on va pouvoir déterminer un intervalle dont cette valeur sera le centre et qui contiendra avec une confiance définie $1 - \alpha$ (0.90, 0.95 ou 0.99 par exemple) la véritable valeur de μ

À partir d'un échantillon de taille n et de moyenne m on aura donc:

✓ Avec $\alpha_1 = 0.1$ $m - 1,65 \sigma / \sqrt{n} \leq \mu \leq m + 1,65 \sigma / \sqrt{n}$

✓ Avec $\alpha_2 = 0.05$ $m - 1,96 \sigma / \sqrt{n} \leq \mu \leq m + 1,96 \sigma / \sqrt{n}$

✓ Avec $\alpha_3 = 0.01$ $m - 2,58 \sigma / \sqrt{n} \leq \mu \leq m + 2,58 \sigma / \sqrt{n}$

Intervalles de confiance de la moyenne

◆ Intervalle de confiance de μ lorsque σ est inconnu

On va devoir estimer σ à partir de s .

◆ Deux cas sont à distinguer selon la taille de l'échantillon:

1. Lorsque les effectifs sont assez grands ($n > 30$), la distribution de l'échantillonnage est gaussienne.
2. Par contre dans le cas de petits effectifs, mais lorsque la distribution parente est normale, les caractéristiques des échantillons impliquent qu'on ait recours à une autre famille de distributions : celles de Student.

◆ Exemple

Dans un test en classe de 6eme sur une population scolaire nationale, la variance des notes obtenues est de 225 ($\sigma = 15$). Sur un groupe de 100 élèves de 6eme, la moyenne des notes obtenues est de 102,5.

Dans quel intervalle de confiance peut-on situer μ aux seuils de $\alpha_1 = 0.05$ et $\alpha_2 = 0.01$?

1) $\alpha_1 = 0.05$

Limite de confiance inférieure

$$m - 1,96 \sigma / \sqrt{n} = 102,5 - 1,96 \times 15 / 10 = 99,56$$

Limite de confiance supérieure

$$m + 1,96 \sigma / \sqrt{n} = 102,5 + 1,96 \times 15 / 10 = 105,44$$

Ce qui donne donc $99,56 \leq \mu \leq 105,44$ avec une probabilité de 95%

◆ Exemple

Dans un test en classe de 6eme sur une population scolaire nationale, la variance des notes obtenues est de 225 ($\sigma = 15$). Sur un groupe de 100 élèves de 6eme, la moyenne des notes obtenues est de 102,5.

Dans quel intervalle de confiance peut-on situer μ aux seuils de $\alpha_1 = 0.05$ et $\alpha_2 = 0.01$?

1) $\alpha_2 = 0.01$

Limite de confiance inférieure

$$m - 1,96 \sigma / \sqrt{n} = 102,5 - 2,58 \times 15 / 10 = 98,63$$

Limite de confiance supérieure

$$m + 1,96 \sigma / \sqrt{n} = 102,5 + 2,58 \times 15 / 10 = 106,37$$

Ce qui donne donc $98,63 \leq \mu \leq 106,37$ avec une probabilité de 99%

- ◆ On note que l'intervalle de confiance s'élargit quand on veut obtenir une confiance plus grande. Mais cette situation n'est pas toujours la plus avantageuse car, dans ce cas, la précision sur une valeur particulière est plus faible

1^{er} cas : grands échantillons ($n > 30$)

- ◆ L'erreur type sur le paramètre μ est l'écart-type de la distribution des statistiques m calculées sur des échantillons au hasard de même effectif n .

On l'estime donc en calculant

$$\sigma_m^2 = \frac{s_x^2}{n}$$

ce qui donne, au niveau de l'erreur-type estimée

$$\sigma_m = \frac{s_x}{\sqrt{n}}$$

s_x étant l'écart-type de l'échantillon, calculé avec $n-1$ au dénominateur.

1^{er} cas : grands échantillons ($N > 30$)

- ◆ L'intervalle de confiance pour μ sera donc

$$m - z_{\alpha} s_x / \sqrt{n} \leq \mu \leq m + z_{\alpha} s_x / \sqrt{n}$$

- ✓ Pour $\alpha_1 = 0.05$ $m - 1,96 s_x / \sqrt{n} \leq \mu \leq m + 1,96 s_x / \sqrt{n}$

- ✓ Pour $\alpha_2 = 0.01$ $m - 2,58 s_x / \sqrt{n} \leq \mu \leq m + 2,58 s_x / \sqrt{n}$

2^e cas : petits échantillons ($n < 30$)

- ◆ L'estimation de σ par s est médiocre lorsque N est petit
- ◆ Il convient de se référer à une nouvelle variable appelée t
 - ✓ Récapitulées en fonction de n dans un table die du t de Student
 - ✓ $m - t s_x / \sqrt{n} \leq \mu \leq m + t s_x / \sqrt{n}$
 - ✓ La valeur de t est fournie par la table de Student avec α et $\nu = n - 1$ connus