Inférence statistique

Cours 6

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

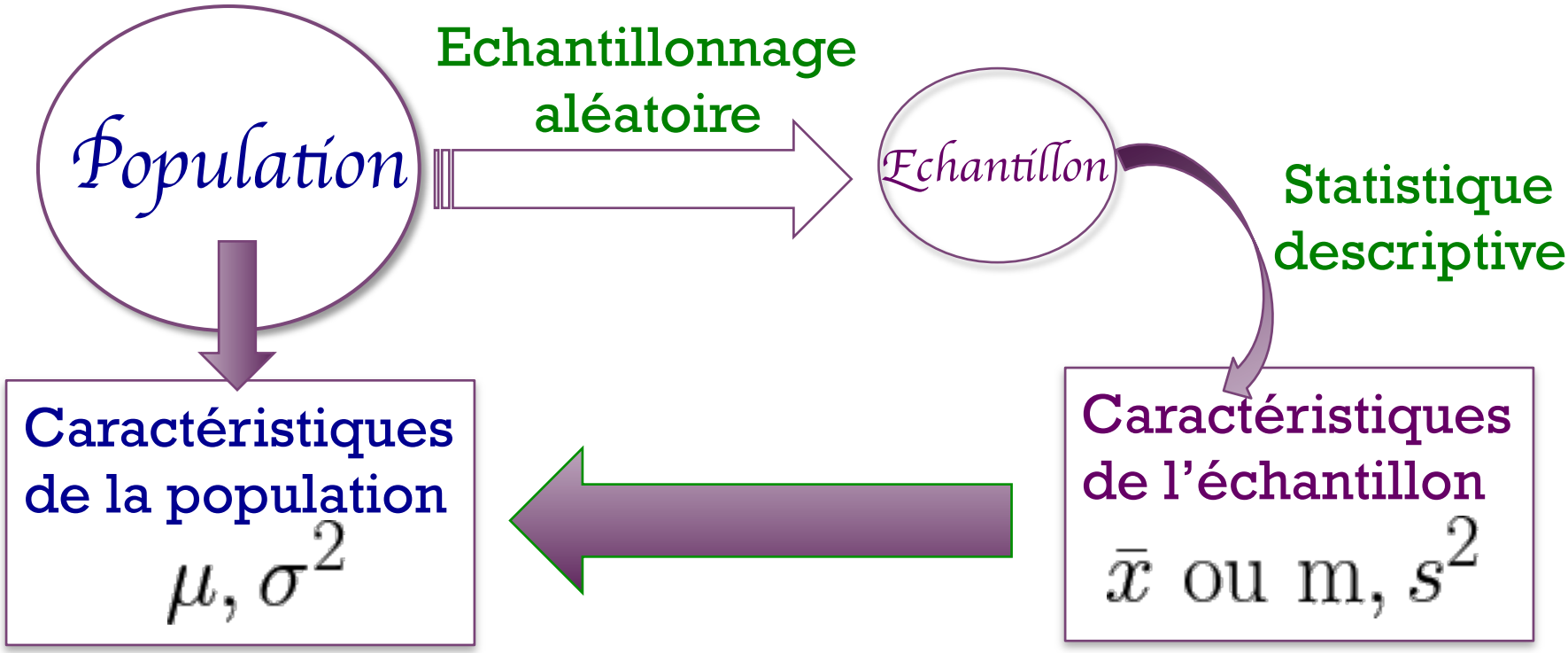
- Définition
  Inférer: généraliser les résultats obtenus sur un échantillon à l'ensemble de la population parente dont cet échantillon est extrait

- 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

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

- L'écart-type σ de la population

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

ou

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

Le 1/N n'est appliqué que sur le 1er terme

Intermède
(écart-type population & échantillon)

- Soit m_l la moyenne des x'_i de l'échantillon l

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

Echantillon 1

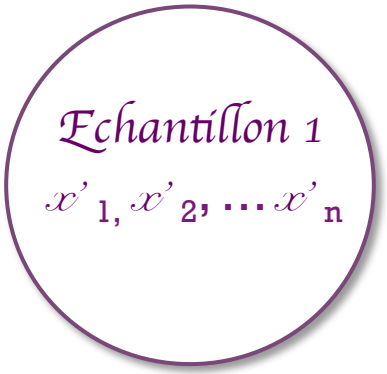
$x'_1, x'_2, \ldots x'_n$

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

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

Intermède
(écart-type population & échantillon)

- Soit m_l la moyenne des x'_i de l'échantillon l



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

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

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{i=n} x^2 - n \times m^2\right)$$

- 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)

- Attention
  la formule de la covariance dans le cas d'un échantillon change aussi:

$$cov(x,y) = \frac{1}{n-1} \times (\sum_{i=1}^{i=n} x_i \times y_i - n \times \bar{x} \times \bar{y})$$

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.

Remarque

- Toujours dans ce même zoo on aurait pu avoir cet échantillon : 9,13, 12, 8, 20 ou bien celui là :10,11,12,13,14…
- Si on calcule la variance de chaque échantillons, celle-ci peut prendre diverses valeurs s, qui tantôt sous-estiment, tantôt surestiment σ
- On pourrait penser que ces valeurs sont centrées sur σ ce qui n'est pas le cas. La moyenne possibles s^2 de la variance de l'échantillon est:

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

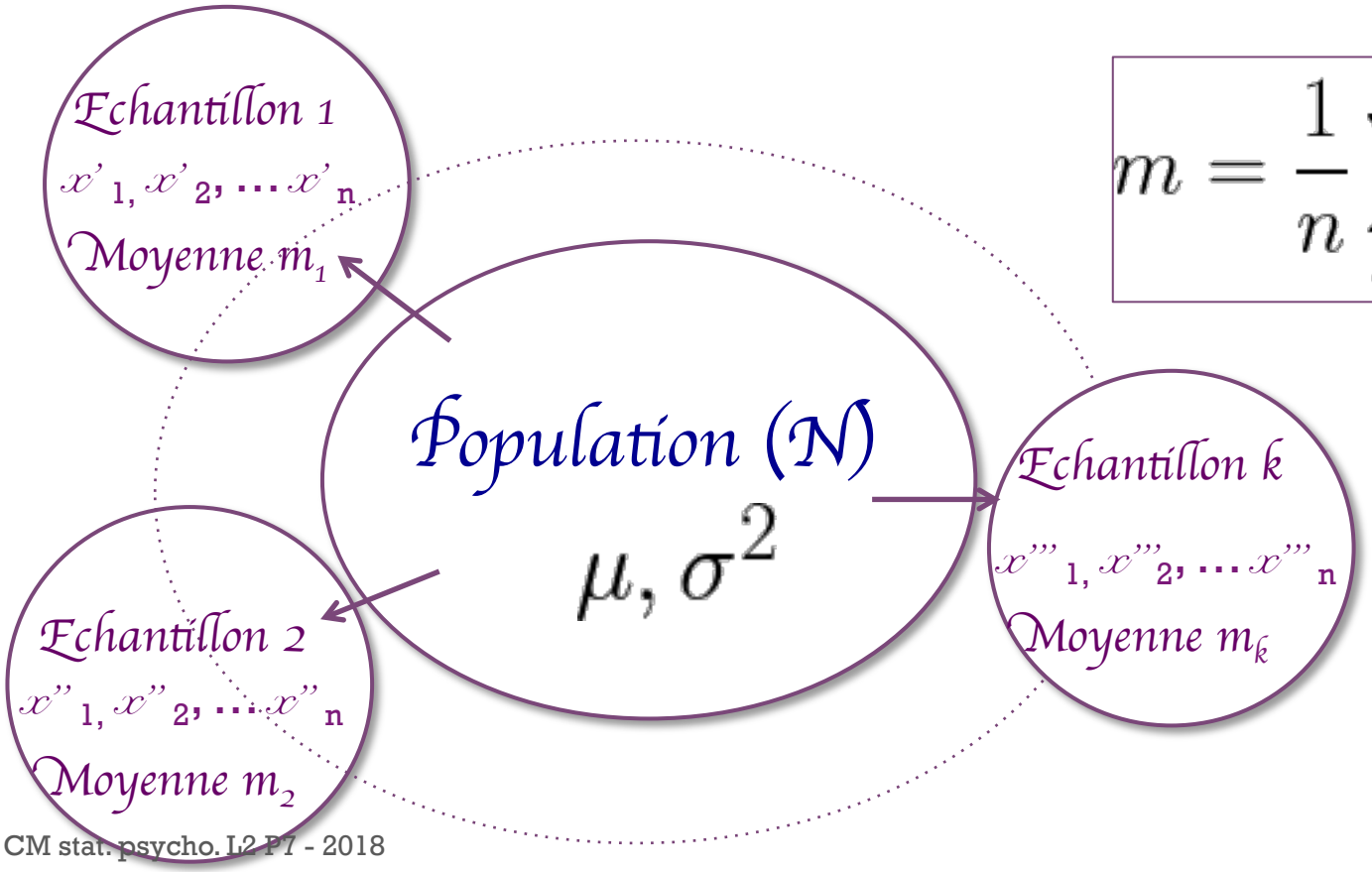
- La moyenne des variances d'échantillon n'est pas la variance de la population, mais la variance de la population multipliée par (n−1)/n

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, …, m_k) vont se distribuer normalement

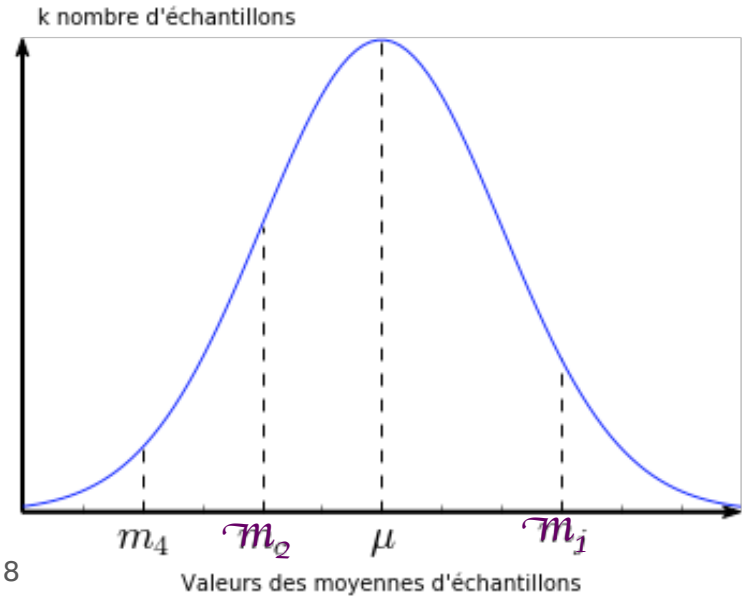


$$m = \frac{1}{n}\sum_{i=1}^{n} x'_i$$

Distribution d'échantillonnage de la moyenne

- Les statistiques-moyennes (m_1 ,m_2 , m_3 … m_k) vont s'organiser dans une distribution dite distribution d'échantillonnage des moyennes ou distribution des moyennes échantillonnales

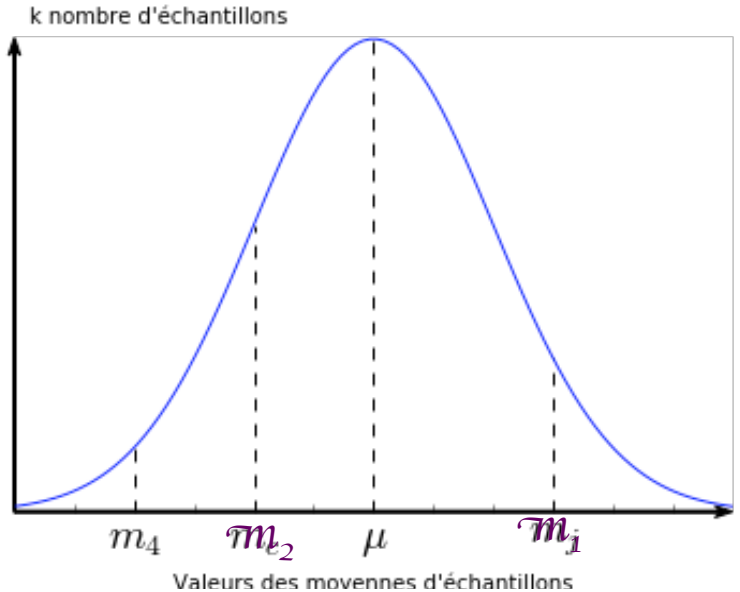
- Remarque: il s'agit ici d'une distribution normale constituée, non plus par les valeurs de x, mais par des moyennes m d'échantillons qui sont devenues notre nouvelle variable



Distribution d'échantillonnage de la moyenne

- La moyenne de cette distribution des moyennes est la meilleure estimation que nous puissions donner de la moyenne de l'ensemble parent μ

- 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➔∞ 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 souvent 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 indépendant 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

- Au niveau des écarts type on a donc $\sigma_m = \frac{\sigma}{\sqrt{n}}$
  - ✓ Il s'agit de l'écart-type de la moyenne

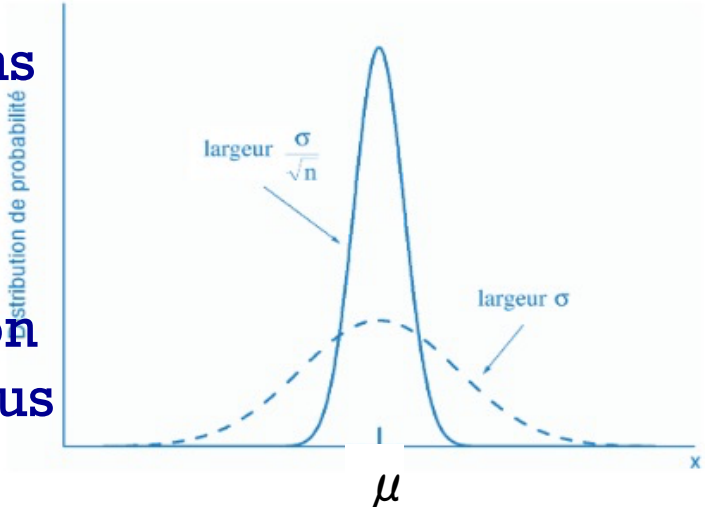
- Plus n est grand plus il devient probable que la moyenne d'un échantillon particulier se rapproche de la moyenne de la population: c'est la loi des grands nombres

- 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 $\sigma/\sqrt{n}$

- 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.



- A mesure que la taille de l'échantillon augmente, nous avons accès à une plus grande quantité d'informations pour estimer la moyenne de la population. Par conséquent, la différence probable entre la vraie valeur de la moyenne de la population et la moyenne échantillonnale diminue.

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
- Contrairement à la moyenne, la variance de l'échantillon s^2 est un estimateur biaisé, donc peu fiable, de la variance de la population σ^2
- On admettra que la meilleur estimation de σ calculé à partir de s est donné par:

$$\sigma = s \times \sqrt{\frac{n}{n-1}}$$

- Remarque si n est grand $\sigma \sim s$

Exemple

- On demande à un échantillon de 8 étudiants de L2 de psycho. de mesurer avec un chrono le temps entre deux images successives projetées sur un écran.

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

On donne :

$$\sum_{i=1}^{i=8} x_i^2 = 329391$$

$$\sum_{i=1}^{i=8} x_i = 1623$$

Exemple

- On demande à un échantillon de 8 étudiants de L2 de psycho. de mesurer avec un chrono le temps entre deux images successives projetées sur un écran.

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

Le temps moyen est donc de m=1623/8=202,90 ms.
Pour calculer l'écart-type, on utilise la formule en 1 sur (n-1):

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{i=n} x^2 - n \times m^2\right)$$

Exemple

$$s^2 = \frac{1}{7} \times \left(\sum_{i=1}^{i=8} x_i^2 - 8 \times \sum_{i=1}^{i=8} x_i\right)$$

$s^2=(1/7)\times(329391-8\times202.90^2)$ =17,83 donc s=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

- Si un 9eme étudiants relève t_9=203 ms nous pouvons prendre s=4,22 ms comme incertitude

- On pourrait indiquer avec :
  - ✓ 68 % de confiance que t_9 se trouve dans l'intervalle: $t_9 = 203 \pm 4{,}22$ ms
  - ✓ 95 % de confiance que t_9 se trouve dans l'intervalle: $t_9 = 203 \pm 2\times4{,}22$ ms $= 203 \pm 8{,}44$ ms

- ATTENTION ceci n'est pas tout à fait exact: vu le faible échantillon, j'aurais utiliser la table de Student pour calculer le niveau de confiance (voir en fin de cours)

Exemple

- L'incertitude associée à la moyenne est donnée par l'écart-type divisé par √n

- C'est l'écart-type de la moyenne:

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

L'écart-type de la moyenne vaut donc, dans notre exemple, 4,22/√8 =1,49 ms
La valeur moyenne du temps mesurer entre deux images successives son écart-type valent donc:

202,90 ± 1,49 ms

La meilleur estimation est donc comprise entre 201,41 et 204,39 ms

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

Intervalles de confiance

- 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

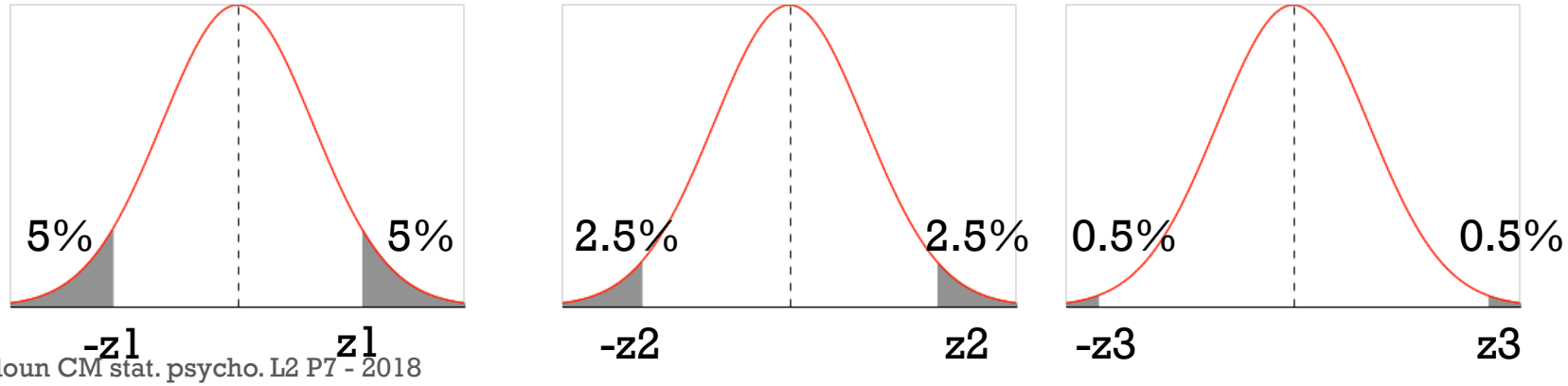
- Limite est appelée seuil de probabilité et est désignée par le symbole α

Intervalles de confiance

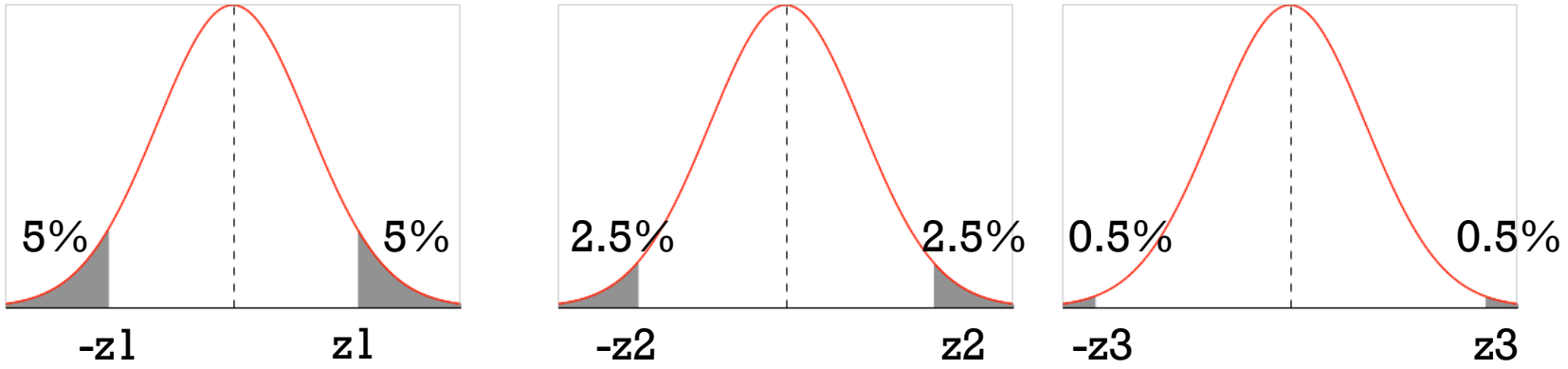
- Définition Wiki:
  Permet de définir une marge d'erreur entre les résultats d'un sondage et un relevé exhaustif de la population totale.
  Plus généralement, l'intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon.

Intervalles de confiance de la moyenne

- En psychologie, les seuils les plus couramment employés sont:
  - ✓ $\alpha_1 = 0.10$ c.à.d. 10% des valeurs extrêmes sont écartées
  - ✓ $\alpha_2 = 0.05$ c.à.d. 5% des valeurs extrêmes sont écartées
  - ✓ $\alpha_3 = 0.01$ c.à.d. 1% des valeurs extrêmes sont écartées



Intervalles de confiance de la moyenne



Si on ne dispose pas de la table bilatérale, pour trouver z1 tel que P(-z1< z < z1)=0.9 avec la table des z unilatérale on utilise P(-z1< z < z1) = 2×P (z<z1)-1 (cf. cours 5)

- P(-z1< z < z1) = 0.9 => P(z<z1)=(1+0,9)/2 =0.95 soit z1=1,65
- P(-z2< z < z2) = 0.95 => P(z<z1)=(1+0,95)/2 =0.975 z1=1,96
- P(-z3< z < z3) = 0.99 => P(z<z1)=(1+0,99)/2 =0.995 z1=2,58

Intervalles de confiance de la moyenne

- Intervalle de confiance de μ lorsque σ est connu

  A 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-α (0.90, 0.95 ou 0.99 par exemple) la véritable valeur de μ

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

  - ✓ Avec α_1=0.1 $\quad m-1{,}65\,\sigma/\sqrt{n} \leq \mu \leq m+1{,}65\,\sigma/\sqrt{n}$
  - ✓ Avec α_2=0.05 $\quad m-1{,}96\,\sigma/\sqrt{n} \leq \mu \leq m+1{,}96\,\sigma/\sqrt{n}$
  - ✓ Avec α_3=0.01 $\quad m-2{,}58\,\sigma/\sqrt{n} \leq \mu \leq m+2.58\,\sigma/\sqrt{n}$

- Exemple

  Dans un test en classe de 6eme sur une population scolaire nationale, la variance des notes obtenues est de 225 (σ=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 α_1=0.05 et α_2=0.01 ?

1) α_1=0.05

Limite de confiance inférieure
m-1,96 $\sigma/\sqrt{n}$=102,5-1,96x15/10=99,56

Limite de confiance supérieure
m+1,96 $\sigma/\sqrt{n}$=102,5+1,96x15/10=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 (σ=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 α_1=0.05 et α_2=0.01 ?

1) α_2=0.01

Limite de confiance inférieure
m-1,96 σ /$\sqrt{n}$=102,5-2,58x15/10=98,63

Limite de confiance supérieure
m+1,96 σ /$\sqrt{n}$=102,5+2,58x15/10=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.

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 ($n \geq 30$), la distribution de l'échantillonnage est gaussienne.

2. Par contre dans le cas de petits effectifs ($n < 30$), et lorsque la distribution parente est normale, les caractéristiques des échantillons impliquent qu'on ait recours à une autre famille de distributions : celles de Student.

1er 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

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

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

- L'intervalle de confiant pour μ sera donc

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

- ✓ Pour α_1=0.05 $\quad m-1{,}96\ s_x/\sqrt{n} \leq \mu \leq m+1{,}96\ s_x/\sqrt{n}$
- ✓ Pour α_2=0.01 $\quad m-2{,}58\ s_x/\sqrt{n} \leq \mu \leq m+2.58\ s_x/\sqrt{n}$

2e 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 dit du t de Student
  - ✓ $m - t\, s/\sqrt{n} \leq \mu \leq m + t\, s/\sqrt{n}$
  - ✓ s étant l'écart-type de l'échantillon, calculé avec n-1 au dénominateur
  - ✓ La valeur de t est fournie par la table de Student avec α et $\upsilon = n-1$ connus
- Table de coefficient de Student (quelques valeurs)

n	2	3	4	5	7	8	10	20	30
t_{68}	1,31	1,18	1,13	1,10	1,07	1,06	1,04	1,02	1,01
t_{95}	4,30	3,18	2,77	2,57	2,36	2,30	2,22	2,08	2,04
t_{99}	9,92	5,84	4,60	4,03	3,50	3,35	3,16	2,84	2,74

2e cas : petits échantillons (n<30)

- Remarque

Dans l'exemple des temps mesurés (8 mesures) on a trouvé:

$$202{,}90 \pm 1{,}49 \text{ ms}$$

Le tableau donne pour $\upsilon = 8\text{-}1 = 7$ mesures pour 68% de confiance t=1,07. On doit donc écrire en toute rigueur: $202{,}90 \pm 1{,}07 \times 1{,}49$ ms. Soit

$$\underline{202{,}90 \pm 1{,}60 \text{ ms avec 68\% de confiance}}$$

2e cas : petits échantillons (n<30)

- Exemple

La moyenne des notes obtenues à un test par 25 élèves d'une classe est 21,2 avec un écart-type égal à 6,5. Quelles sont les limites de confiance de la moyenne de l'ensemble-parent ? :

On calcule $\sigma_m = 6,5 / \sqrt{25} = 1,3$

Pour N = 25, on a $\upsilon = 25-1 = 24$. Au seuil $\alpha = 0.05$ pour $\upsilon = 24$, on donne la valeur t de Student de 2,06. Les limites de confiance de la moyenne μ sont donc

- ✓ 21,2-2,06×1,3=18,53
- ✓ 21,2+2,06×1,3= 23,88

Soit $18,53 \leq \mu \leq 23,88$

Intervalles de confiance de l'écart-type

- En tirant au hasard k échantillons de même effectif n, on a une distribution des estimations de σ
- Si n, effectif de chacun des échantillons est grand, typiquement supérieur à 80, on démontre que les distributions d'échantillonnage de l'écart-type est centrée sur σ de l'ensemble parent:

Pas à l'exam

$$\sigma_s = \frac{\sigma}{\sqrt{2n}}$$

Intervalles de confiance de l'écart-type

- Si on ne connaît pas σ et pour calculer les limites de confiance on utilisera (z dans la table bilatérale):

$$s - z_\alpha \frac{s}{\sqrt{2n}} \leq \sigma \leq +z_\alpha \frac{s}{\sqrt{2n}}$$

Pas à l'exam

- Pour des échantillons plus petits, le problème est plus complexe, car il est lié à une distribution dit du chi 2 χ^2