

# Cours d'Éléments de Statistique

Jean Christophe Meunier

## MODULE 7 Théorie de l'échantillonnage

1<sup>ère</sup> Bac, Commerce Extérieur  
Année Académique 2015-2016



### A. Introduction

## De l'échantillonnage en statistique inférentielle

- Hypothèse de la statistique inférentielle
  - Population considérée comme infinie
    - échantillon = approche 'palpable'
  - les variables statistiques qui décrivent la population peuvent être considérées comme des v.a.
  - La répartition des valeurs de ces v.a. sont caractérisées par des lois de probabilités
    - La répartition d'une variable statistique  $X$  sur la population est décrite par une loi de probabilité possédant des caractéristiques (...  $X$ ,  $\sigma^2_X$  ou autres paramètres résumant la distribution)
- La qualité de l'inférence repose en premier lieu sur la qualité du sondage effectué  $\Rightarrow$  théorie de l'échantillonnage

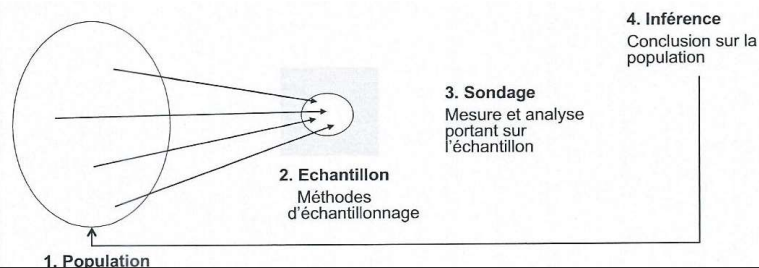
## De l'échantillonnage en statistique inférentielle

- L'objectif de la statistique inférentielle :
  - identifier/approcher/utiliser les lois de probabilité, au vu d'un échantillon de valeurs des variables obtenues
  - Méthodes d'estimation : Théorie d'échantillonnage
    - permettent d'approcher les lois ou certaines de leurs caractéristiques (ex : approcher , à partir de l'échantillon, l'espérance  $E(Y)$  de la variable  $Y$ =salaire,...)
  - Méthodes de Test d'hypothèse :
    - permettent de confirmer ou d'infirmer des hypothèses faites sur ces lois (ex : décider si, au vu de l'échantillon, l'affirmation «  $E(Y)=1500$  euros » est plausible.)

## B. Théorie d'échantillonnage

### Echantillon

- Si travail sur population :
  - On décrit avec certitude les caractéristiques de la population
    - Statistiques descriptive
    - Souvent illusoire : travaille long et coûteux (ex. recensement obligatoire)
- Travail sur échantillon
  - Plus réaliste, mais recours statistique inférentielle
- Enjeux majeurs
  - Récolter de bons échantillons et déterminer leur qualité



## Qualité de l'échantillon

- Objectif de l'inférence :
  - Estimer les paramètres de la population (e.g.  $\mu$ ,  $\sigma$ ), à partir d'échantillons pris dans population et en utilisant lois de probabilité
- Qualité souhaitée de l'échantillon :
  - Représentatif de la population
  - Non biaisé
  - Dont moyenne  $\bar{X}$  et variance  $\sigma_x^2$  bons estimateurs de  $\mu$  et  $\sigma^2$
  - Qualité de l'estimation dépend de la qualité de l'échantillon
- !! Privilégier qualité (sondage répondu sérieusement) à la quantité (grand échantillon mais biaisé)

## Échantillonnages aléatoires (1)

- Reposent sur le hasard
  - unités statistiques sont désignées par le hasard et ont toutes la même chance d'être choisies
  - permettre de calculer la marge d'erreur échantillonnale : généralisation possible
- Une base de sondage est
  - liste des individus à partir de laquelle on prélève un échantillon.
  - Cette liste détermine la population observée.

## Échantillonnage aléatoires (2)

1. L'échantillonnage aléatoire simple
  - choisir aléatoirement des individus dans la base de sondage
2. L'échantillonnage aléatoire systématique
  - unités statistiques sont choisies à intervalle régulier dans la base de sondage
3. L'échantillonnage aléatoire stratifié
  - population divisible en groupes distincts (strates) : âge, ethnie, etc...
  - prélever un échantillon ayant même composition que population
  - Très coûteux
4. L'échantillonnage aléatoire par grappes
  - Grappe : sous-ensembles non homogènes de population définis selon proximité (ex. critère de proximité : choix aléatoire d'hôpitaux)
  - Choisir aléatoirement groupes plutôt que unités isolées
  - Économie de temps et d'argent

## Échantillonnage aléatoires (3)

### • Exemple d'échantillonnage aléatoire simple

#### Exemple

Soit la population de 100 étudiants inscrits à l'EPFC pour laquelle nous observons le résultat final (cote sur 100) à l'issue de leur formation.

Nous souhaitons estimer la cote moyenne et la variance de cette population à partir d'un échantillon de 5 étudiants.

#### Résolution

L'échantillonnage aléatoire de 5 étudiants choisis parmi la population définie peut concrètement être réalisé de 2 manières:

1. Tirer 5 jetons dans une urne contenant 100 jetons distincts (jetons nominatifs)
  2. Attribuer un numéro d'ordre à chacun des 100 étudiants, de 00 à 99, et choisir au hasard 5 numéros grâce à une table de nombres aléatoires (voir **Table 1** en annexe). Par exemple, si on commence à la ligne 3, on obtiendrait 72, 20, 47, 33 et 84.
- Chacun des 100 étudiants possède la même probabilité d'être tiré, soit  $1/100$ .

## Échantillonnages non-aléatoires (1)

- Peu coûteux, rapides et simples
- Généralisation peu opportune
- Impossible de calculer la marge d'erreur échantillonnale

## Échantillonnages non-aléatoires (2)

1. L'échantillonnage accidentel
  - Sujet se trouve au mauvais endroit au mauvais moment
  - Ex : Enquêteur à la sortie du magasin entre 12h et 14h
2. L'échantillonnage par volontaire
  - une des méthodes les plus utilisées
  - sujets décident de faire partie de l'étude de leur propre gré.
3. L'échantillonnage par quotas
  - respecte les proportions de population, mais choix 'accidentel' des sujets
  - Ex : Enquêteur à la sortie du magasin mais s'arrête quand 50-50 fille-garçon comme dans population générale
4. L'échantillonnage au jugé
  - technique basée sur les connaissances du chercheur sur la population
  - chercheur choisit les sujets qu'il juge possédant les caractéristiques typiques de la population

## Théorie de l'échantillonnage

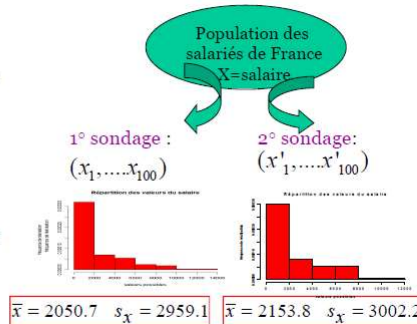
- Etude des liaisons existant entre une population et les échantillons de cette population, prélevés par sondage
- Notion d'échantillon aléatoire

Quelle que soit la technique d'échantillonnage utilisée, le contenu du jeu de données prélevé varie d'un sondage à l'autre

On pourrait répéter le sondage un grand nombre de fois, on obtiendrait la plupart du temps une répartition différente des valeurs prélevées.

Le résultat d'un sondage est aléatoire

Sondage de 100 salariés



## Théorie de l'échantillonnage

- Raisonnement
  - Point de départ : On tire tous les échantillons aléatoires possibles (de taille  $n$ ) d'une population connue
  - Chaque échantillon est une réalisation aléatoire dont les paramètres (ex:  $\bar{x}$ ,  $\sigma^2$ ) varie autour des paramètres réels de la population connue (ex:  $\bar{x}$ ,  $\sigma^2$ )
  - En connaissant, les paramètres de chaque échantillon et ceux de la population
    - On peut avoir une idée de l'erreur de l'extrapolation entre échantillon considéré et la population

## Théorie de l'échantillonnage

- Exemple :

- Population de 4 sujets sur une variable x

Nom	Variable X
Albert	2
Carlos	6
Hakim	10
Sylvie	14

- On prélève avec remise\* tous les échantillons possibles de taille  $n=2$

\* Arrangement avec remise =  $n^p = 4^2 = 16$  échantillons aléatoires possibles

## Théorie de l'échantillonnage

- Valeurs des 16 échantillons ( $n=2$ )

2, 2	2, 6	2, 10	2, 14	6, 2	6, 6	6, 10	6, 14
10, 2	10, 6	10, 10	10, 14	14, 2	14, 6	14, 10	14, 14

- Moyenne des 16 échantillons

2	4	6	8	4	6	8	10
6	8	10	12	8	10	12	14

- Distribution d'échantillonnage

$\bar{X}_i$	2	4	6	8	10	12	14
$n_i$	1	2	3	4	3	2	1

$\bar{X}_i$  (et non  $X_i$ ) : chaque réalisation aléatoire  $\bar{X}_i$  (moyenne d'échantillon  $n=2$ ) de  $\mu$  (moyenne population  $N = 4$ )



## Théorie de l'échantillonnage

- Calcul moyenne

- Moyenne population

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

$$= (2+6+10+14)/4 = 8$$

Nom	Variable X
Albert	2
Carlos	6
Hakim	10
Sylvie	14

- Moyenne des moyennes d'échantillonnage

$$\text{moyenne } \bar{x} = \frac{\sum_{i=1}^n \bar{x}_i}{n} = \frac{\sum_{i=1}^n n_i \bar{x}_i}{n}$$

- = (2+4+6+8+4+6+8+10+6+8+10+12+8+10+12+14)/16 = 8

2	4	6	8	4	6	8	10
6	8	10	12	8	10	12	14

## Théorie de l'échantillonnage

- Calcul variance

- Variance population

$$\sigma_{\text{population}}^2 = \frac{(2-8)^2 + (6-8)^2 + (10-8)^2 + (14-8)^2}{16} = 20$$

Nom	Variable X
Albert	2
Carlos	6
Hakim	10
Sylvie	14

- Variance des moyennes d'échantillonnage

$$\sigma_{\bar{x} \text{ d'échantillonnage}}^2 = \frac{(2-8)^2 + \dots + (14-8)^2}{16} = 10$$

2	4	6	8	4	6	8	10
6	8	10	12	8	10	12	14

## Théorie de l'échantillonnage

- Calcul écart-type → Erreur Standard (SE)

–  $\sigma_{\text{population}} = \sqrt{20}$  et  $\sigma_{\bar{x} \text{ d'échantillonnage}} = \sqrt{10}$

$$\sigma_{\text{population}} = \sqrt{20} = \sqrt{2 * 10} = \sqrt{2} * \sqrt{10} = \sqrt{2} * \sigma_{\bar{x} \text{ d'échantillonnage}}$$

$$\sigma_{\text{population}} = \sqrt{n} * \sigma_{\bar{x} \text{ d'échantillonnage}}$$

$$SE = \sigma_{\bar{x} \text{ d'échantillonnage}} = \frac{\sigma_{\text{population}}}{\sqrt{n}}$$

Plus la taille des échantillons tirés est grande plus SE diminue

**SE = erreur standard** = exprime dans quel 'ampleur' la moyenne de l'échantillon (  $\bar{X}$  ) varie autour de la moyenne réelle de la population (  $\mu$  )

## Théorie de l'échantillonnage

- Calcul écart-type → Erreur Standard (SE)

- Tirage des échantillons avec ou sans remise

1. Si tirage échantillons avec remise : formule telle quelle

$$SE = \sigma_{\bar{x} \text{ d'échantillonnage}} = \frac{\sigma_{\text{population}}}{\sqrt{n}}$$

2. Si tirage échantillons sans remise : ajout d'un coefficient correcteur

$$SE = \sigma_{\bar{x} \text{ d'échantillonnage}} = \frac{\sigma_{\text{population}}}{\sqrt{n}} \sqrt{\frac{N-n}{n-1}}$$

Coefficient correcteur

Par facilité on peut noter  $\sigma_{\text{échantillon}} \rightarrow s$  et  $\sigma_{\text{population}} \rightarrow \sigma$

# Théorème central limite

## Théorème central limite

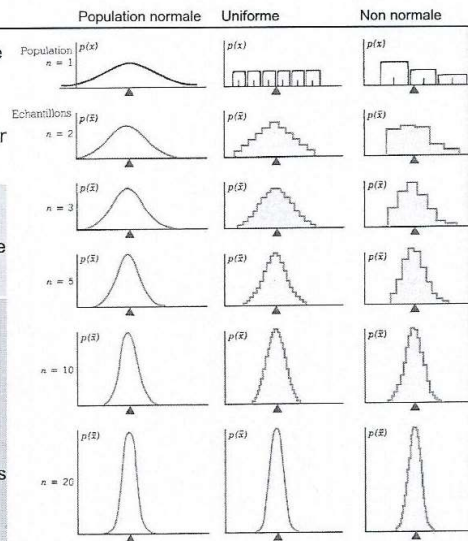
La distribution d'échantillonnage  $p(\bar{X})$  diffère de la distribution  $p(x)$  de la population.

En effet, à mesure que la taille  $n$  des échantillons augmente, la forme de la distribution d'échantillonnage se modifie.

Lorsque la population mère est normale, ou lorsque la taille de l'échantillon est grande ( $n > 10$  ou  $20$ ), la distribution d'échantillonnage a une forme normale ou approximativement normale.

Théorème central limite (règle d'approximation normale)

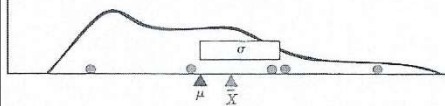
Dans les échantillons aléatoires de taille  $n$ , la moyenne de l'échantillon  $\bar{X}$  varie autour de la moyenne de la population  $\mu$  avec un écart type égal à  $\sigma/\sqrt{n}$  (où  $\sigma$  est l'écart type de la population). Donc, quand  $n$  augmente, la distribution d'échantillonnage est de plus en plus concentrée sur  $\mu$  et proche de la distribution normale



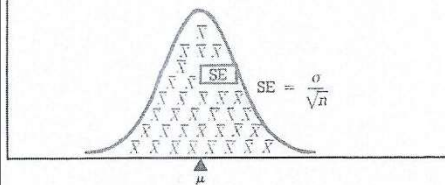
# Théorème central limite

## Utilité du théorème central limite

Population en gris avec la moyenne  $\bar{X}$  d'un échantillon déterminé



Distribution d'échantillonnage en bleu (avec les  $\bar{X}$  de tous les échantillons possibles).



Comme on le voit sur la figure ci-contre, la distribution de  $\bar{X}$  (distribution d'échantillonnage) est approximativement normale, et plus concentrée autour de  $\mu$  que celle de la population.

### Utilité du théorème central limite

Il nous permet d'utiliser les tables normales pour déterminer dans quelle mesure une moyenne  $\bar{X}$  de l'échantillon permet une estimation précise de la moyenne  $\mu$  de la population.

Connaissant  $n$ ,  $\mu$  et  $\sigma$ , on peut tracer une distribution normale qui approche suffisamment la distribution des moyennes d'échantillons  $\bar{X}$

## Exercice 1

Une population d'une grande université a une taille moyenne  $\mu$  de 69 pouces et un écart type  $\sigma$  de 3,22 pouces.

Si **un** échantillon aléatoire de  $n = 10$  individus est prélevé, quelle est la probabilité pour que la moyenne de **cet** échantillon  $\bar{X}$  s'écarte de 2 pouces de la moyenne de la population  $\mu$  ?

### Résolution

La taille de l'échantillon étant suffisamment grande, on peut appliquer le théorème central limite et approcher la distribution de l'ensemble des moyennes de tous les échantillons par une distribution normale de moyenne

$$\mu = 69$$

et d'écart type

$$s = \frac{\sigma}{\sqrt{n}} = \frac{3,22}{\sqrt{10}} = 1,02$$

On cherche la probabilité pour que  $\bar{X}$  soit compris entre 67 et 71.

Calculons d'abord la probabilité pour que  $\bar{X}$  soit supérieure à 71, en centrant et réduisant la distribution:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{71 - 69}{1,02} = 1,96$$

## Exercice 1 : solution

Cela signifie que la valeur critique 71 pour  $\bar{X}$  est à environ 2 fois SE.

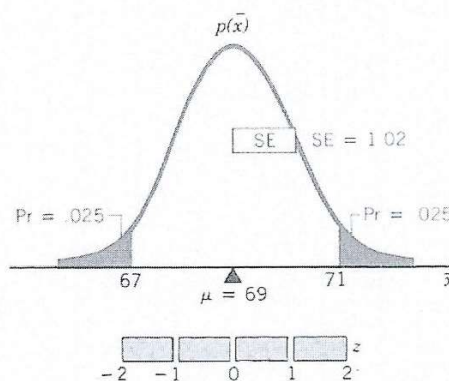
D'après la table de la loi normale centrée réduite (annexe, table IV), on trouve que la probabilité pour que  $Z$  excède 1,96 est de 0,25 (car  $Z_{0,25} = 1,96$ ).

En raison de la symétrie de la distribution normale, la probabilité pour que  $\bar{X} < 67$  est aussi égale à 0,25

La probabilité pour que  $67 < \bar{X} < 71$  correspondant à la partie centrale dans la figure ci-contre est donc:

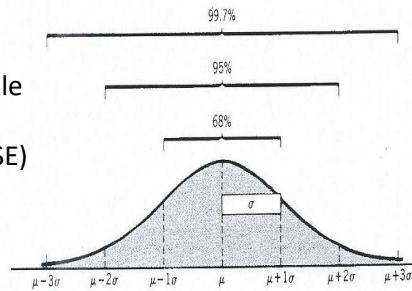
$$\Pr(67 < \bar{X} < 71) = 1 - 0,025 - 0,025 = 0,950$$

Cela signifie qu'il y a 95% de chance pour que la moyenne de l'échantillon s'écarte de moins de 2 pouces de la moyenne de la population.



## Intervalle de confiance (IC)

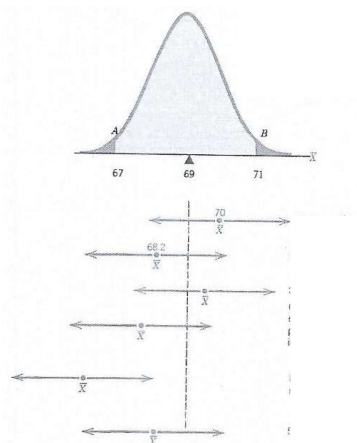
- SE permet de construire un IC autour de la moyenne de la population ( $\mu$ )
  - Sachant que la distribution d'échantillonnage approche la normale
  - Et connaissant la moyenne ( $\bar{X}$ ) de l'échantillon et l'erreur standard (SE)



- On sait que parmi les échantillons tirés aléatoirement de cette population :
  - Environ 68 % auront une moyenne comprise entre  $\left[ \bar{x} - \frac{\sigma(X)}{\sqrt{n}}; \bar{x} + \frac{\sigma(X)}{\sqrt{n}} \right]$
  - Environ 95 % auront une moyenne comprise entre  $\left[ \bar{x} - 2\frac{\sigma(X)}{\sqrt{n}}; \bar{x} + 2\frac{\sigma(X)}{\sqrt{n}} \right]$
  - Environ 99,7 % auront une moyenne comprise entre  $\left[ \bar{x} - 3\frac{\sigma(X)}{\sqrt{n}}; \bar{x} + 3\frac{\sigma(X)}{\sqrt{n}} \right]$

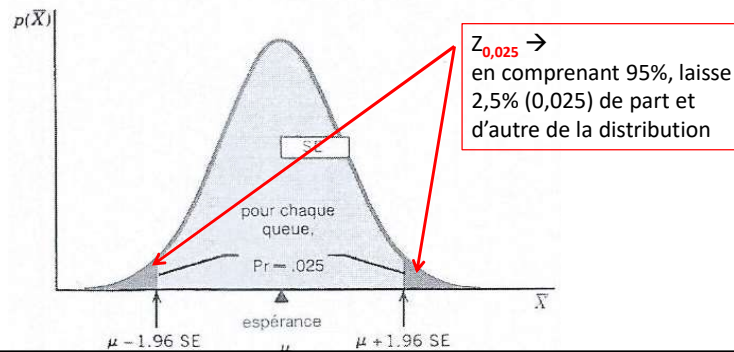
## Intervalle de confiance (IC)

- Par convention l'IC le plus courant (de référence) est 95%
  - La logique** : trouver un intervalle autour de  $\mu$  dans lequel  $\bar{X}$  serait compris dans 95 % des cas



## Comment trouver IC de 95% ? (1)

- On recherche une marge d'erreur pour laquelle
  - $P(\mu - \text{marge d'erreur} < \bar{X} < \mu + \text{marge d'erreur}) = 95\%$
- On sait que  $\bar{X}$  suit loi normale et doté d'une SE
  - Selon loi normale centrée réduite la valeur de z qui comprend 95 % de la distribution est 1,96  $\rightarrow Z_{0,025} = 1,96$  (voir tableau IV)



## Comment trouver IC de 95% ? (2)

- Dès lors  $P(\mu - \text{marge d'erreur} < \bar{X} < \mu + \text{marge d'erreur}) = 95\%$  équivaut à
  - $P(\mu - z_{0,025} SE < \bar{X} < \mu + z_{0,025} SE) = 95\%$
  - $P(\mu - 1,96 SE < \bar{X} < \mu + 1,96 SE) = 95\%$
- On peut transformer l'IC en prenant  $\mu$  comme référence
 

$\mu - 1,96 SE < \bar{X}$  donnant  $\mu < \bar{X} + 1,96 SE$

et

$\bar{X} < \mu + 1,96 SE$  donnant  $\mu > \bar{X} - 1,96 SE$

Dès lors

$P(\bar{X} - 1,96 SE < \mu < \bar{X} + 1,96 SE) = 95\%$
- Intervalle de confiance à 95%

$$\mu = \bar{X} \pm 1,96 SE = \bar{X} \pm z_{0,025} SE$$

– Et, sachant que  $SE = \frac{\sigma}{\sqrt{n}}$

$$IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$



### Résumé de la démarche

1. Le paramètre  $\mu$  de la population est constant et reste constant. La variable aléatoire, c'est l'intervalle d'estimation car son centre  $\bar{X}$  est une variable aléatoire.
2. Pour comprendre comment il est construit, on peut écrire l'intervalle de confiance comme

$$\mu = \bar{X} \pm z_{0,025} SE$$

où  $z_{0,025}$  est la valeur qui laisse de côté 2,5% de la queue de distribution à gauche et à droite de la distribution.

Cette équation donne le prototype de tous les intervalles de confiance étudiés.

Une autre forme très utile est:

Intervalle de confiance à 95% : 
$$\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

3. A mesure que la taille de l'échantillon s'accroît,  $\bar{X}$  est tel que l'écart type de l'échantillon  $\sigma/\sqrt{n}$  devient plus petit et que l'intervalle de confiance se réduit. Cet accroissement de la précision est le gain résultant de l'augmentation de la taille de l'échantillon
4. Si l'on souhaite accroître le seuil de confiance, on élargit l'intervalle de confiance.

## Exercice 2

On considère un échantillon suffisamment petit qui fera l'objet d'un suivi répété, afin d'estimer le nombre de repas pré-cuisinés dans une grande ville. Le taux de réponse élevé signifie que l'échantillon est essentiellement aléatoire. On a enregistré, lors de l'enquête, le nombre de repas « prêts à consommer » pris par chaque individu au cours de la semaine précédente. En résumé, on obtient  $\bar{X} = 0,82$ ,  $\sigma = 0,5$  et  $n = 180$ .

Calculer l'intervalle de confiance à 95% pour la moyenne de la population totale de cette ville.

Solution

La formule  $\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$  peut être utilisée car on connaît l'écart type de la population

Où  $z_{0,025}$  est la valeur de  $z$  correspondant à une probabilité de 2,5%, soit 1,96.

On peut alors calculer les bornes de l'intervalle recherché:

$$\begin{array}{c} \bar{X} \\ \hline \bar{X} - z_{0,025} \frac{\sigma}{\sqrt{n}} \qquad \qquad \bar{X} + z_{0,025} \frac{\sigma}{\sqrt{n}} \end{array}$$

## Exercice 2 : Solution

- Dans ce cas, on connaît

$\bar{X} = 0,82$  et  $n = 180$  : moyenne et taille de l'échantillon

$\sigma = 0,5$  : écart-type de la population

Dès lors,

$$\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}} = 0,82 \pm 1,96 \frac{0,50}{\sqrt{180}} = 0,82 \pm 1,96 * 0,037 = 0,82 \pm 0,07$$

CI à 95 % de  $\mu = 0,82 \pm 0,07$

- Problème !

- Dans la réalité, le  $\sigma$  réel de la population est rarement connu
- On recourt à  $\sigma_{\text{échantillon}}$  (qu'on peut aussi noter  $s$ ) comme approximation de  $\sigma_{\text{population}}$  (qu'on peut aussi noter  $\sigma$ )
- 2 cas de figure (voir diapos suivantes):
  - Grands échantillons :  $n \geq 100$
  - Petits échantillons :  $n < 100$

## $\sigma_{\text{population}}$ inconnu et grands échantillons ( $n \geq 100$ )

- Quand  $n \geq 100$

- $s$  est une bonne approximation de  $\sigma^*$

$$IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm z_{0,025} \frac{s}{\sqrt{n}}$$

\*Rappel : par facilité on peut noter  $\sigma_{\text{échantillon}} \rightarrow s$  et  $\sigma_{\text{population}} \rightarrow \sigma$



## $\sigma_{\text{population}}$ inconnu et petits échantillons ( $n < 100$ )

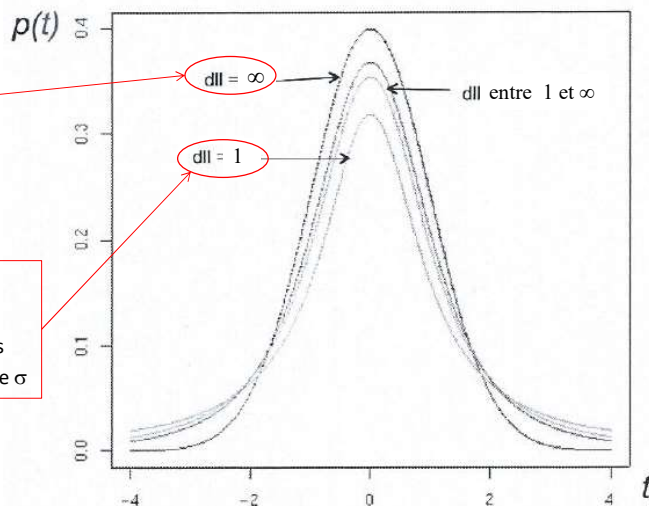
- Quand  $n < 100$ 
  - 's' est toujours utilisé comme approximation de ' $\sigma$ '
  - Mais comme  $n$  petit, incertitude supplémentaire car  $s$  peut plus facilement fluctuer autour de  $\sigma$
- Pour obtenir IC à 95%, l'intervalle doit être élargi
  - Remplacer  $z_{0.025}$  de la distribution normale par  $t_{0.025}$  de la distribution t de Student qui est plus élargie
  - On utilise la statistique  $t = \frac{\bar{X} - \mu}{s} \times \sqrt{n}$  et non la statistique  $z = \frac{\bar{X} - \mu}{s}$
- IC à 95%
  - $IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm t_{0.025} \frac{s}{\sqrt{n}}$
  - Valeurs de  $t$  (table V) tabulée selon degré de liberté (ddl)
  - $ddl = n - 1$

## Distribution de la loi t de Student

- ddl : quantité d'information pour estimer  $s$

Si  $ddl \rightarrow \infty$   
Loi t suit loi z  
(normale centrée réduite)

Si ddl petit  
distribution plus élargie  
compense pour les fluctuations  
plus importantes de  $s$  autour de  $\sigma$



## Synthèse des cas de figure pour calcul d'IC

1.  $\sigma$  est connu

$$IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

2.  $\sigma$  est inconnu (grands échantillons ( $n \geq 100$ ))

$$IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm z_{0,025} \frac{s}{\sqrt{n}}$$

3.  $\sigma$  est inconnu (petits échantillons ( $n < 100$ ))

$$IC \text{ à } 95\% \text{ de } \mu = \bar{X} \pm t_{0,025} \frac{s}{\sqrt{n}}$$

Quel que soit le cas,  
plus  $n$  augmente :  
**IC se réduit**  
Et, dès lors  
**Estimation plus précise de**  
 $\mu$  par  $\bar{X}$

## Exercice 3 (idem que exercice 2 mais $\sigma$ inconnu)

On considère un échantillon suffisamment petit qui fera l'objet d'un suivi répété, afin d'estimer le nombre de repas pré-cuisinés dans une grande ville. Le taux de réponse élevé signifie que l'échantillon est essentiellement aléatoire. On a enregistré, lors de l'enquête, le nombre de repas « prêts à consommer » pris par chaque individu au cours de la semaine précédente. En résumé, on obtient  $\bar{X} = 0,82$ ,  $s = 0,48$  et  $n = 180$ . Calculer l'intervalle de confiance à 95% pour la moyenne de la population totale de cette ville.

Solution

La formule  $\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$  ne peut être utilisée car on ne connaît pas l'écart type de la population.

Par contre, on connaît l'écart type de l'échantillon  $s$ .

Lorsque  $n \geq 100$ , on peut utiliser, en bonne approximation, remplacer  $\sigma$  par  $s$  :

$$\text{Si } n \geq 100, \text{ alors } \mu \approx \bar{X} \pm z_{0,025} \frac{s}{\sqrt{n}}$$

On a donc:

$$\mu \approx \bar{X} \pm z_{0,025} \frac{s}{\sqrt{n}} = 0,82 \pm 1,96 \frac{0,48}{\sqrt{180}} = 0,82 \pm 0,07$$

## Exercice 4

Dans une classe très nombreuse, on extrait un échantillon de 4 notes: 64, 66, 89 et 77.

Calculer un intervalle de confiance à 95% pour la moyenne de toute la classe.

Que peut-on conclure ? Comment améliorer l'estimation ?

Solution

Comme  $n = 4$ ,  $ddl = 3$  et donc, d'après la table V, on trouve que  $t_{0,025} = 3,18$ .

Pour les échantillons, la variance  $s^2$  est donnée par:

Note observée $X$	$(X - \bar{X})$	$(X - \bar{X})^2$
64	-10	100
66	-8	64
89	15	225
77	3	9
$\bar{X} = \frac{296}{4} = 74$	$\sqrt{\quad}$	$s^2 = \frac{398}{3} = 132,7$

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

Correction pour calcul  $s^2$   
échantillon :  
**Important qd n est petit**  
**Qd n grand (>100), correction**  
**négligeable**

## Exercice 4 : solution

On a obtenu  $\bar{X} = 74$  et  $s^2 = 132,7$ .

Après substitution dans:

Intervalle de confiance à 95%

$$\mu = \bar{X} \pm t_{0,025} \frac{s}{\sqrt{n}}$$

on obtient:

$$\mu = 74 \pm 3,18 (\sqrt{132,7})/\sqrt{4} = 74 \pm 18$$

Cela signifie qu'avec un intervalle à 95%, la note moyenne est comprise entre 56 et 92.

On peut améliorer l'estimation en augmentant la taille de l'échantillon.

## Exercice 5

Quelle est la taille d'échantillon nécessaire pour estimer la moyenne d'une population dans un intervalle de 10 unités, avec un degré de confiance de 95%, sachant que l'écart type de la population est égal à 200 ?

Solution

L'intervalle de confiance à 95 % lorsque  $\sigma$  est connu est donné par:

$$\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}. \text{ Comme l'intervalle est de 10 unités, on peut écrire:}$$

$$2 \times z_{0,025} \frac{\sigma}{\sqrt{n}} = 10$$

$$n = \left( \frac{2 \times z_{0,025} \times \sigma}{10} \right)^2 = \left( \frac{1,96 \times 400}{10} \right)^2 = 78,4^2 = 6147$$

## Formules relatives à population et à échantillon

Moyenne de population

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

Variance de population

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

Ecart type de population

$$\sigma = \sqrt{\sigma^2}$$

Moyenne d'échantillon

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Variance d'échantillon

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

Ecart type d'échantillon

$$s = \sqrt{S^2}$$