Cours de Statistique Inférentielle

Jean Christophe meunier

Module 2
Théorie d'échantillonnage & Tests d'hypothèse

2ème Bac, Commerce Extérieur
Année académique 2015-2016

EPFC

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, σ^2_X ou autres paramètres résumant la distribution)
- La qualité de l'inférence repose en premier lieu sur la qualité du sondage effectué ⇒ théorie de l'échantillonnage

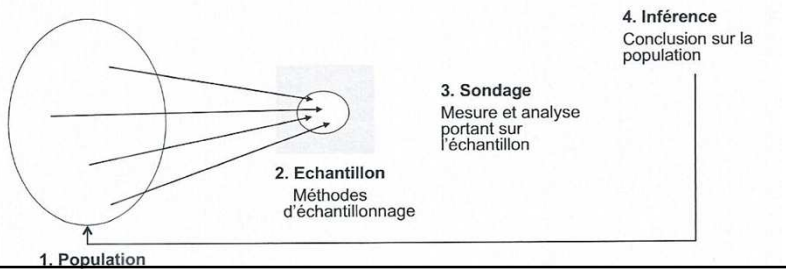
De l'échantillonnage en statistique inférentielle

- 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 couteux (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. μ, σ), 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 $\overline{X}$ et variance σ_x^2 bons estimateurs de μ et σ^2
  - Qualité de l'estimation dépend de la qualité de l'échantillon
- !! Privilégier qualité à la quantité
  - sondage répondu sérieusement par un échantillon modeste plutôt que 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.

Échantillonnages 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 qu'unités isolées
   - Économie de temps et d'argent

Échantillonnages 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

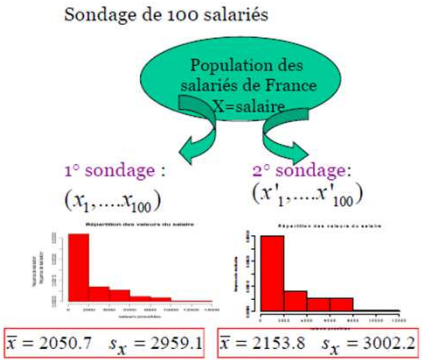
Théorie de l'échantillonnage

- 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



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: $\bar{x}$, σ^2) varie autour des paramètres réels de la population connue (ex: $\bar{x}$, σ^2)
  - En connaissant les paramètres de chaque échantillon et ceux de la population
    - On peut avoir une idée de l'erreur d'extrapolation entre l'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 μ (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

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

- = (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

$$\sigma^2_{population} = \frac{(2-8)^2 + (6-8)^2 + (10-8)^2 + (14-8)^2}{4} = 20$$

Nom	Variable X
Albert	2
Carlos	6
Hakim	10
Sylvie	14

  - Variance des moyennes d'échantillonnage

$$\sigma^2_{\bar{x}\ d\ échantillonnage} = \frac{(2-8)^2 + \cdots + (14-8)^2}{16} = 10$$

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

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

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

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

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 (μ)

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

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

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

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

Coefficient correcteur

Par facilité on peut noter $\sigma_{échantillon}$ → s et $\sigma_{population}$ → σ

Théorème central limite

Théorème central limite

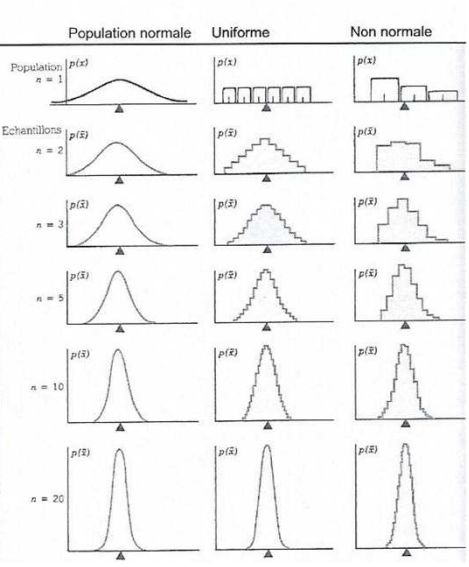
La distribution d'échantillonnage p($\bar{X}$) diffère de la distribution p(x) de la population.

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.

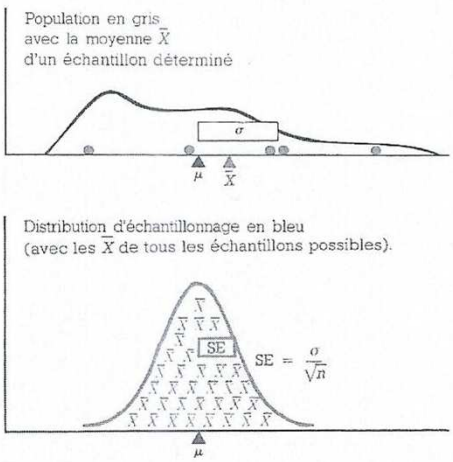
Théorème central limite (règle d'approximation normale)

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



Théorème central limite

Utilité du théorème central limite



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 μ 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 μ de la population.

Connaissant n, μ et σ, 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 μ de 69 pouces et un écart type σ 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 μ ?

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.

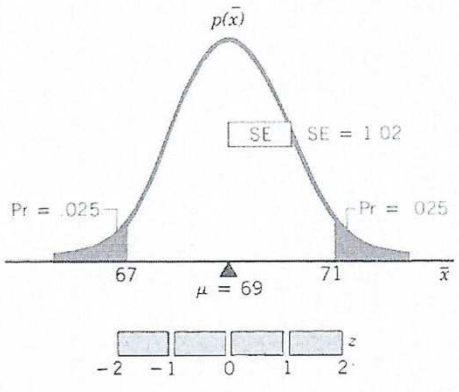
D'après la table de la loi normale centrée réduite (annexe, table IV), on trouve que la probabilité pour que Z excède 1,96 est de 0,25 (car $Z_{0,25} = 1,96$).

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:

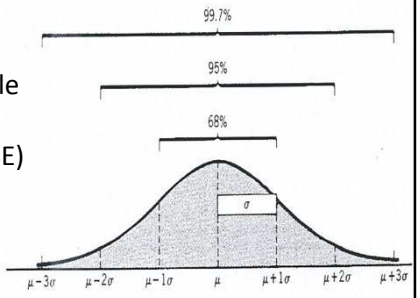
$$\Pr(67 < \bar{X} < 71) = 1 - 0,025 - 0,025 = 0,950$$

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)

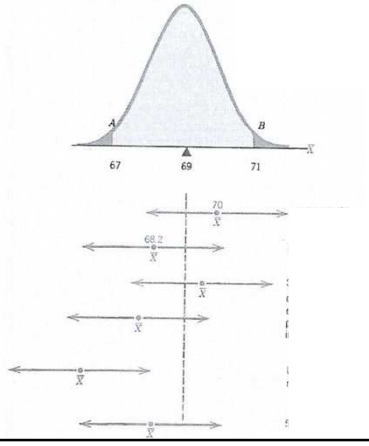
- SE permet de construire un IC autour de la moyenne de la population (μ)
  - Sachant que la distribution d'échantillonnage approche la normale
  - Et connaissant la moyenne ($\overline{X}$) de l'échantillon et l'erreur standard (SE)



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

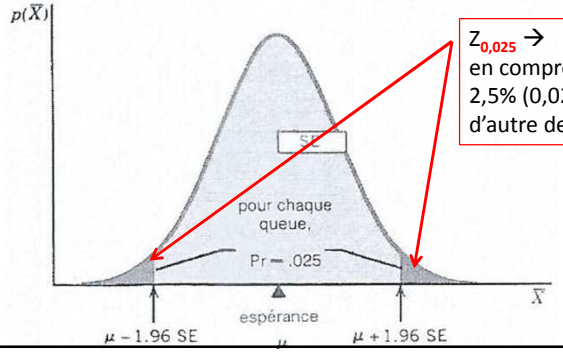
Intervalle de confiance (IC)

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



Comment trouver IC de 95% ? (1)

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



$Z_{0,025}$ →
en comprenant 95%, laisse 2,5% (0,025) de part et d'autre de la distribution

Comment trouver IC de 95% ? (2)

- Dès lors *P*(μ-marge d'erreur < $\bar{X}$ < μ + marge d'erreur) = 95% équivaut à
  - P(μ- $z_{0.025}$ SE < $\bar{X}$ < μ + $z_{0.025}$ SE) = 95%
  - P(μ- 1,96 SE < $\bar{X}$ < μ + 1,96 SE) = 95%
- On peut transformer l'IC en prenant μ comme référence

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

  et

  $\bar{X}$ < μ + 1,96 SE donnant μ > X – 1,96 SE

  Dès lors

  P($\bar{X}$- 1,96 SE < μ < $\bar{X}$ + 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 μ 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 a 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 $n = 180$.

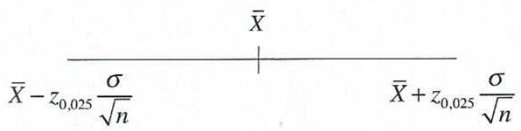
Calculer l'intervalle de confiance à 95% pour la moyenne de la population totale de cette ville.

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:



Exercice 2 : Solution

- Dans ce cas, on connait

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

  σ = 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 μ = 0,82 ± 0,07

- Problème !
  - Dans la réalité, le σ 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 σ)
  - 2 cas de figure (voir dias suivantes):
    - Grands échantillons : n ≥ 100
    - Petits échantillons : n < 100

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

- Quand n ≥ 100
  - s est une bonne approximation de σ*

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

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

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

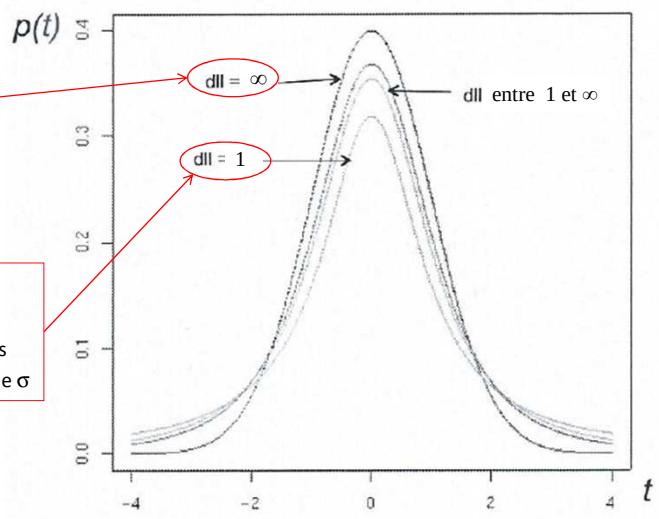
- Quand n < 100
  - 's' est toujours utilisé comme approximation de 'σ'
  - Mais comme n petit, incertitude supplémentaire car s peut plus facilement fluctuer autour de σ
- Pour obtenir IC à 95%, l'intervalle doit être élargi
  - Remplacer $z_{0.025}$ de la distribution normale par $t_{0.025}$ de la 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%
  - $IC\ \text{à}\ 95\%\ de\ \mu = \bar{X} \pm t_{0,025} \frac{s}{\sqrt{n}}$
  - Valeurs de t (table V) tablées selon degré de liberté (ddl)
  - ddl = n – 1

Distribution de la loi t de Student

- ddl : quantité d'information pour estimer s

Si ddl → ∞
Loi t suit loi z
(normale centrée réduite)

Si ddl petit
distribution plus élargie
compense pour les fluctuations
plus importantes de s autour de σ



Synthèse des cas de figure pour calcul d'IC

1. σ est connu

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

2. σ est inconnu (grands échantillons (n ≥ 100))

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

3. σ est inconnu (petits échantillons (n < 100))

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

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

Exercice 3 (idem que exercice 2 mais σ 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 ≥ 100, on peut utiliser, en bonne approximation, remplacer σ par s :

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

On a donc:

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

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
$\bar{X} = \frac{296}{4} = 74$	0√	$s^2 = \frac{398}{3} = 132.7$

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

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

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 σ est connu est donné par:

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

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

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

Formules relatives à la population et à l'échantillon

Moyenne de population

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

Moyenne d'échantillon

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

Variance de population

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

Variance d'échantillon

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

Ecart type de population

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

Ecart type d'échantillon

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

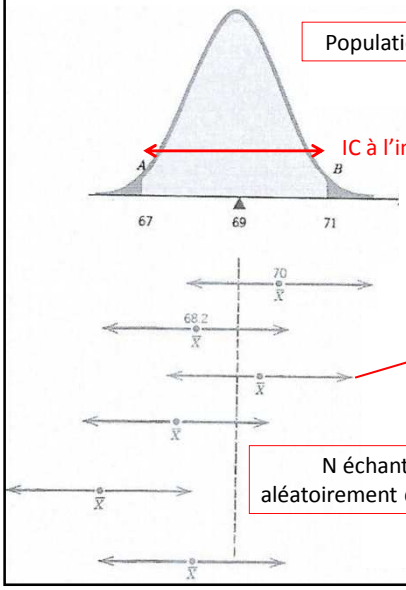
C. Tests d'hypothèse

Principe du test d'hypothèse (1)

1. On suppose d'abord qu'il existe 2 hypothèses: l'hypothèse nulle H_0 et l'hypothèse alternative H_A (ou alternative de recherche)
   - H_0 $E(\bar{X}) = \mu$, c'est-à-dire que l'échantillon estime bien la population
   - H_A $E(\bar{X}) \neq \mu$, c'est-à-dire que l'échantillon ne représente pas bien la population
2. La procédure de test commence par supposer que H_0 est vraie
3. Le but de la procédure est de déterminer s'il y a suffisamment d'évidence statistique pour rejeter l'affirmation que H_A est vraie
4. Deux type de décisions sont alors possibles:
   1. Il y a suffisamment d'évidence pour affirmer que H_A est vraie, et on rejette H_0
   2. Il n'y a pas suffisamment d'évidence pour affirmer que H_A est vraie, et on garde H_0

L'intervalle de confiance (IC) sert de base pour estimer l'ensemble des hypothèses acceptables

Principe du test d'hypothèse (2)

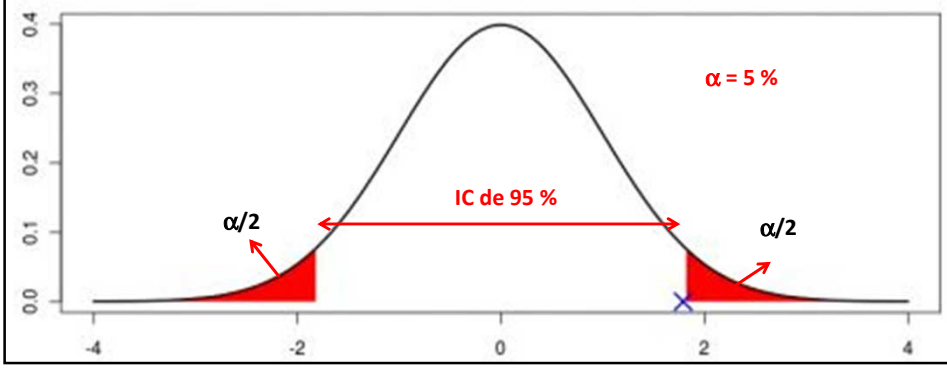


H_0 : cet échantillon fait bien partie de la population

H_A : cet échantillon ne fait pas partie de la population

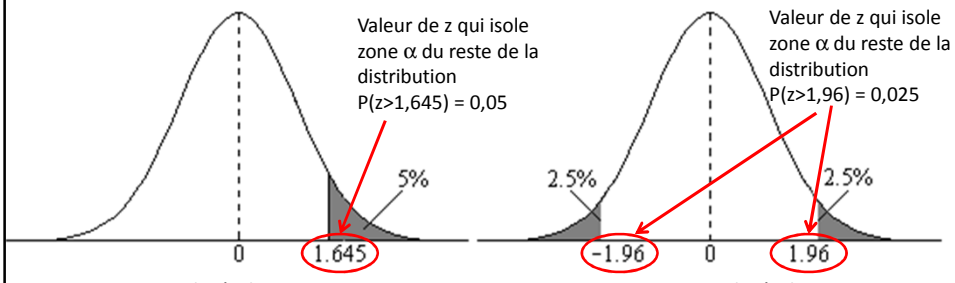
Probabilité critique

- Pour un IC donné, la zone de rejet (α) de l'H_0 représente la probabilité inverse à l'IC
  - Pour IC 95%
    - α = 5 % (2,5 % à l'extrême inférieure et 2,5 % à l'extrême supérieure)
  - α représente la probabilité critique au-delà de laquelle H_0 rejetée



Probabilité critique

- Test unilatéral vs. bilatéral

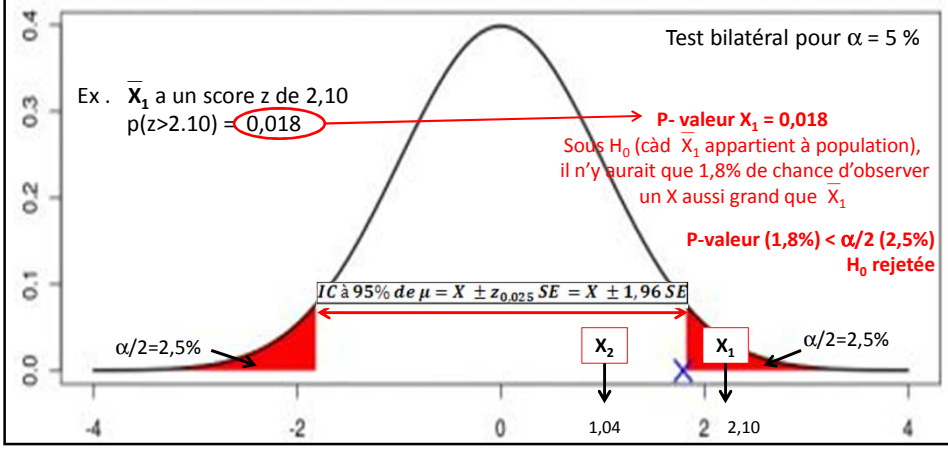


Unilatéral :
On a une idée sur le sens de la différence
H0 : $\bar{X} = \mu$
Ha : $\bar{X} > \mu$ ou $\bar{X} < \mu$

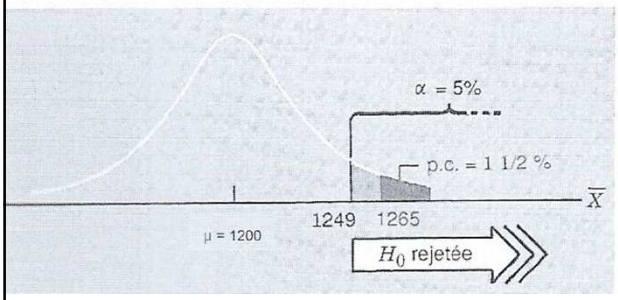
Bilatéral :
Pas d'idée sur le sens de la différence
H0 : $\bar{X} = \mu$
Ha : $\bar{X} \neq \mu$

Probabilité critique et p-valeur

- La **p-valeur** d'un test représente la probabilité que la valeur observée $\bar{X}$ appartient bien à la population (μ), sous H_0
  - Si p-valeur < probabilité critique (α) H_0 rejeté au profit de H_A



Probabilité critique et p-valeur



On rejette H_0 si $p_c \leq \alpha$

Probabilité critique et p-valeur

- Interprétation de la p-valeur
  - A partir de quelle p-valeur, peut-on conclure que l'hypothèse alternative (H_A) est vraie (donc rejeter H_0) ?

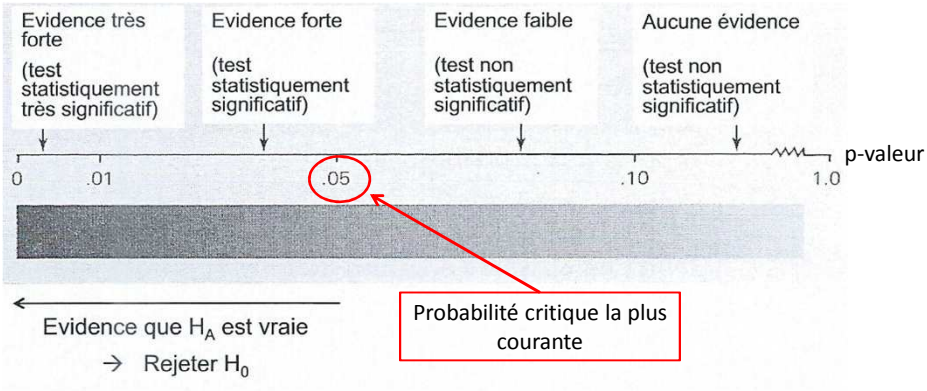


Illustration (1)

Un procédé de fabrication courant a produit des millions de tubes TV, dont la durée de vie moyenne est μ = 1200 heures et l'écart type σ = 300 heures.

Un nouveau procédé fournit un échantillon de 100 tubes avec une moyenne $\bar{X}$ = 1 265 heures.

Cet échantillon fait-il apparaître le nouveau procédé comme meilleur ?

Solution

Commençons par voir si la valeur de $\bar{X}$ observée est une bonne estimation de μ. Pour cela, nous devons calculer l'intervalle de confiance à 95% centré sur $\bar{X}$

$$\mu = \bar{X} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

$$1265 \pm 59$$

Comme on peut le voir, cet intervalle n'encadre pas μ = 1200.

On peut donc conclure que l'hypothèse nulle H_0 : le processus est le même (pas meilleur) peut être rejetée.

Illustration (2)

Dans un test classique, la moyenne μ et l'écart type σ de la population sont connus.

On procède en 3 étapes:

1. On définit d'abord: (1) l'hypothèse nulle H_O, (2) la taille de l'échantillon n et (3) le seuil d'erreur du test noté α.
2. On suppose provisoirement que l'hypothèse nulle H_O est vraie. On trace la distribution (connaissant μ et σ de la population) et on définit arbitrairement la zone de rejet à partir du seuil d'erreur choisi, α. Tout cela se fait avant toute observation de données.
3. On considère maintenant l'échantillon. Si la valeur observée $\bar{X}$ tombe dans la région de rejet (partie grisée), elle est alors jugée suffisamment en conflit avec l'hypothèse nulle pour que celle-ci soit rejetée. Sinon, H_O n'est pas rejetée.

$$Z_c = Z_{0,05} = 1,64 = \frac{\bar{X}_c - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

d'où on tire: $\bar{X}_c = 1249$

H_0 est rejetée si $\bar{X} > \bar{X}_c$

Test unilatéral

Tout α est situé dans la queue de droite

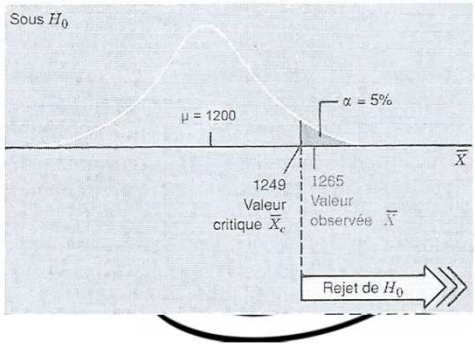


Illustration (3)

SOLUTION

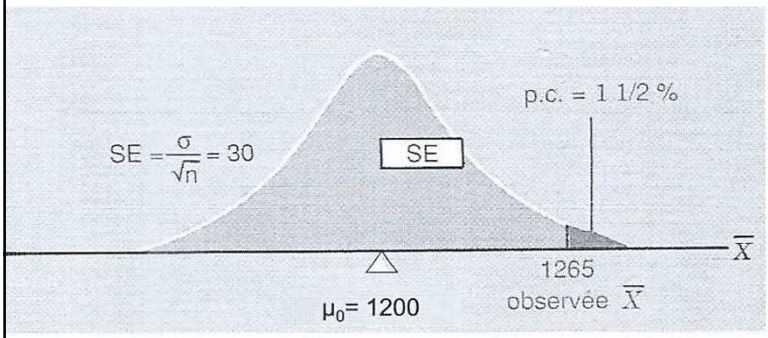
Sur la Figure 9-1 on a représenté la distribution hypothétique de $\bar{X}$, sous H_0. (Pour la première fois, on présente ici la convention selon laquelle les distributions hypothétiques sont représentées en blanc pâle.) Selon le théorème central limite, cette distribution est normale, avec une moyenne $\mu_0 = 1\ 200$, et un écart type d'échantillon $SE = \sigma/\sqrt{n} = 300/\sqrt{100} = 30$. On emploie ces valeurs pour centrer et réduire la valeur observée $\bar{X} = 1\ 265$:

$$Z = \frac{\bar{X} - \mu_0}{SE} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$= \frac{1\ 265 - 1\ 200}{30} = 2.17 \qquad (9\text{-}11)$$

Donc $\Pr(\bar{X} \geq 1\ 265) = \Pr(Z \geq 2.17) = .015$ (9-12)

Illustration (4)



Cela signifie que si le nouveau procédé n'était pas meilleur (sous H_0), il n'y aurait que 1,5% de chance d'observer $\bar{X}$ aussi grand que 1265.

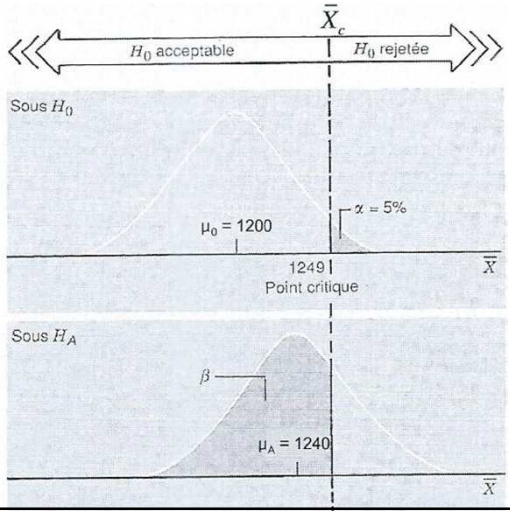
Ce 1,5% est appelé probabilité critique pc unilatérale correspondant à l'hypothèse alternative.

La pc résume le degré de concordance entre les données et H_0.

pc = Pr (la valeur d'échantillon est égale à la valeur réellement observée sous H_0)

Erreurs de première et deuxième espèces

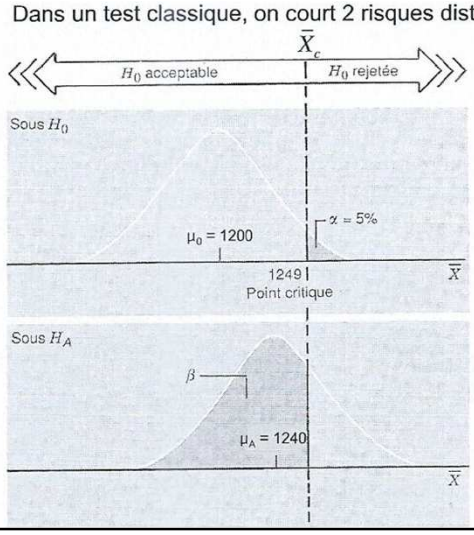
- Quid si H_0 acceptée/rejetée à tort ?



Erreurs de première et deuxième espèces

- Puissance d'un test statistique

Dans un test classique, on court 2 risques distincts lors de la prise de décision:



Si en réalité, H_0 est vraie (sous H_0)
Erreur de *type 1*: rejeter H_0 à tort lorsque $\bar{X} > \bar{X}_c$ (zone ombrée)
Probabilité de l'erreur de type 1 = α

Si en réalité, H_A est vraie (sous H_A)
Erreur de *type 2*: admettre H_0 à tort lorsque $\bar{X} < \bar{X}_c$ (zone ombrée)
Probabilité de l'erreur de type 2 = β

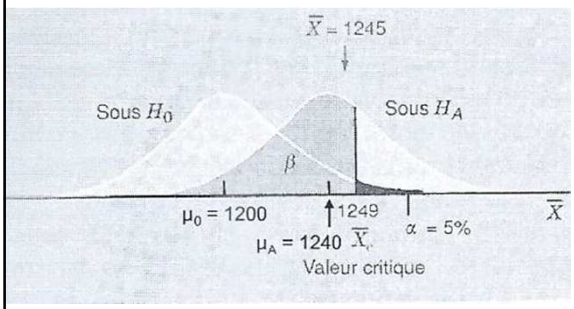
Puissance d'un test = $1 - \beta$
Pr de rejeter H0 lorsque H0 est fausse

Erreurs de première et deuxième espèces

Etat du monde	Décision	
	H_0 acceptable	**H_0 rejetée**
Si H_0 est vraie	Décision correcte Probabilité = (1 – α) = seuil de confiance	Erreur de type 1 Probabilité = α = **seuil du test**
Si H_0 est fausse	Erreur de type 2 Probabilité = β	Décision correcte Probabilité = (1 – β) = **puissance du test**

Erreurs de première et deuxième espèces

- Interdépendance des 2 types d'erreurs

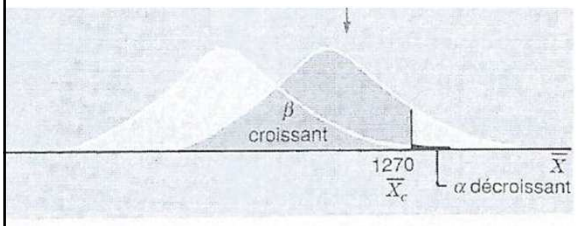


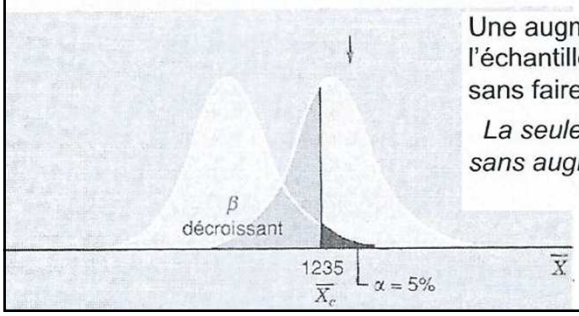
En déplaçant la valeur critique X_c vers la droite, la réduction de α entraîne l'accroissement de β

α et β sont interdépendants: réduire l'une augmente l'autre

Erreurs de première et deuxième espèces

- Interdépendance des 2 types d'erreurs





Une augmentation de la taille de l'échantillon n entraîne la réduction de β sans faire augmenter α

La seule manière de réduire une erreur sans augmenter l'autre est de rassembler plus d'information

Exercice 5

Calculer la zone de rejet et la p_c pour les situations suivantes et interpréter les résultats:

a) H_0: $\mu = 1{,}000$
H_1: $\mu \neq 1{,}000$
$\sigma = 200$, $n = 100$, $\bar{x} = 980$, $\alpha = 0{,}10$

b) H_0: $\mu = 50$
H_1: $\mu > 50$
$\sigma = 5$, $n = 9$, $\bar{x} = 51$, $\alpha = 0{,}05$

c) H_0: $\mu = 15$
H_1: $\mu < 15$
$\sigma = 2$, $n = 25$, $\bar{x} = 14.3$, $\alpha = .10$

d) H_0: $\mu = 100$
H_1: $\mu \neq 100$
$\sigma = 10$, $n = 100$, $\bar{x} = 100$, $\alpha = .05$

e) H_0: $\mu = 70$
H_1: $\mu > 70$
$\sigma = 20$, $n = 100$, $\bar{x} = 80$, $\alpha = .01$

f) H_0: $\mu = 50$
H_1: $\mu < 50$
$\sigma = 15$, $n = 100$, $\bar{x} = 48$, $\alpha = .05$

Exercice 5 : solution

a) H_0: $\mu = 1{,}000$

H_1: $\mu \neq 1{,}000$

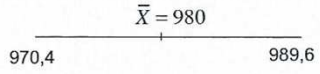
$\sigma = 200$, $n = 100$, $\bar{x} = 980$, $\alpha = 0{,}10$

Comme σ est connu, on peut utiliser la formule suivante pour calculer l'intervalle de confiance à 90% (1 – α) :

$$\mu = \bar{X} \pm z_{0,05} \frac{\sigma}{\sqrt{n}}$$

Dans les tables IV, $z_{0,05}$ = 0,480. L'intervalle de confiance centré sur $\bar{X} = 980$ s'étend donc sur :

$$\pm z_{0,05} \frac{\sigma}{\sqrt{n}} = \pm 0{,}480 \frac{200}{\sqrt{100}} = \pm 9{,}6$$



La valeur μ = 1000 (sous H_0) n'est pas comprise dans cet intervalle et doit, en principe être rejetée.

Cependant, calculons la pc pour savoir si on peut vraiment rejeter H_0

Exercice 5 : solution

Si H_0 est vraie, alors, on devrait trouver une probabilité non négligeable de trouver des valeurs de $\bar{X} < 980$

Quelle est la probabilité pour que l'on trouve des valeurs encore plus petites que 980 ?

$p_c = \Pr(\bar{X} < 980)$ sous H_0

Pour cela, on réduit et centre la variable X: $z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} = \frac{-20}{200}10 = -1$

Donc:

$$p_c = \Pr(\bar{X} < 980) = \Pr(z < -1) = \Pr(z > 1) = 0{,}159$$

La pc étant largement supérieure à 5%, il est très peu évident (probable) que H_1 soit vraie et on ne peut donc rejeter H_0