Corrigés du problème de distribution d'échantillonnage de la moyenne $\overline{X}$
(usine textile)

Lecture de l'énoncé :

X = Longueur des morceaux de tissu en cm

Population mère :
$E(X) = m = 90$ la gen mor ceau de tissu en moyenne
$\sigma = 0,60$ l'ecat type
$N = 10\,000$ (1er cas de figue)
$N = 2\,000$ (2ème cas de figure)

Echantillon :
$n = 200$ echantillon aleatoire
$\bar{x} = 90,30$ moyenne observé

a) 1ère Question : P ($\overline{X} \leq 89,90$) ?

Procédure pour N = 10 000 :

1) C'est un problème de distribution d'échantillonnage de la moyenne car :
- la question porte sur une moyenne ;
- on travaille sur un échantillon ;
- on connaît les paramètres de la population mère ($m = 90$, $\sigma = 0,60$)

2) $X \sim N(90 , 0,60)$ car :
- $n = 200 > 60 \Rightarrow$ Théorème central limite
- il s'agit d'une production en série et standardisée. On peut appliquer le critère d'atomicité qui est un des critères liés à la loi Normale.

3) Taux de sondage = $\frac{n}{N} = \frac{200}{10\,000} = 0,02 < 0,05 \Rightarrow$ on ne doit donc pas appliquer le facteur d'exhaustivité.

4) $X \sim N(90 , 0,60) \Rightarrow \overline{X} \sim N(E(\overline{X}) = 90 ; \sigma_{\overline{X}} = \frac{0,60}{\sqrt{200}} = 0,04243)$

$$\Rightarrow \mathbf{P(\overline{X} \leq 89,90)} = P(T \leq \frac{89,90 - 90}{0,04243}) = P(T \leq -2,36)$$
$$= 1 - P(T \leq +2,36) = 1 - 0,9909$$
$$= \mathbf{0,91\ \% = 0,0091.}$$

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

Toujours pb 1

TD 2+3

Procédure pour N = 2 000 :

1) Même réponse que pour N = 10 000.

2) Même réponse que pour N = 10 000.

3) Taux de sondage = $\frac{200}{2\,000} = 0{,}10 > 0{,}05 \Rightarrow$ on va appliquer le facteur d'exhaustivité.

4) $X \sim N(90\ ,\ 0{,}60) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = 90\ ;\ \sigma_{\overline{X}} = \frac{0{,}60}{\sqrt{200}}\sqrt{\frac{2\,000-200}{2\,000-1}} = 0{,}04026\right)$

$\Rightarrow$ **P** $(\overline{X} \leq 89{,}90) = P\left(T \leq \frac{89{,}90-90}{0{,}04026}\right) = P(T \leq -2{,}48) = 1 - P(T \leq +2{,}48) = 1 - 0{,}9934$

= 0,66 % = 0,0066.

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

b) 2ème Question : Montrer que l'échantillon est représentatif de la population mère pour N = 10 000.

- Prendre un risque de 5% de se tromper $\Rightarrow$ l'intervalle symétrique que l'on veut respecter doit couvrir 95% des situations possibles.

$\Rightarrow P(m - t\ \sigma_{\overline{X}} \leq \overline{X} \leq m + t\ \sigma_{\overline{X}}) = 0{,}95.$

- m = 90
- t de la loi Normale pour un intervalle symétrique de 95% = ± 1,96

(Pour trouver + t :

- rechercher dans la table de la loi N la probabilité la plus proche de 97,5% = 0,975.
- on voit que 0,975 se situe à l'intersection de la ligne 1,9 et de la colonne 0,06

$\Rightarrow$ + t = + 1,96).

- $\sigma_{\overline{X}}$ = 0,04243 (pour N = 10 000)

$\Rightarrow$ **Intervalle de confiance** = [90 - (1,96)(0,04243) ; 90 + (1,96)(0,04243)]

= [89,917 ; 90,083]

95% des moyennes de morceaux de tissu sont dans cet intervalle de confiance

- Comme la moyenne de l'échantillon $\bar{x}$ = 90,30 se trouve à l'extérieur de l'intervalle de confiance, on peut dire, au risque de 5% de se tromper, que l'échantillon n'est pas représentatif de la population mère.

Dans ce cas, on remet les 200 morceaux de tissu dans la population mère (car ici, les morceaux testés sont encore utilisables) et il faut de nouveau tirer un échantillon aléatoire de 200 morceaux de tissu afin de refaire la procédure.

Corrigés du problème de distribution d'échantillonnage d'une différence de moyennes $\overline{X_1} - \overline{X_2}$ (piles électriques)

Lecture de l'énoncé :

X_1 = durée d'utilisation des piles de la société 1 en heures
X_2 = durée d'utilisation des piles de la société 2 en heures

Populations mères :
$m_1 = 230$ $\sigma_1 = 30$
$m_2 = 210$ $\sigma_2 = 20$

Echantillons :
$n_1 = 100$
$n_2 = 125$

Question : $P(\overline{X_1} - \overline{X_2} \geq 30)$?

Procédure :

1) Problème de distribution d'échantillonnage d'une différence de moyennes car :

— la question porte sur une comparaison de moyennes ;

- on travaille sur échantillons ;
- on connaît tous les paramètres des populations mères (m_1 et m_2 , σ_1 et σ_2).

2) $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$ car :

- $n_1 = 100 > 60$ et $n_2 = 125 > 60 \Rightarrow$ Théorème central limite ;
- les piles sont fabriquées en très grandes séries $\Rightarrow$ critère d'atomicité $\Rightarrow$ loi Normale.

3) Taux de sondage < 5% car :

- même si on ne connaît pas exactement N_1 et N_2 on sait que les tailles des populations mères sont très importantes vue la fabrication en très grande séries ;

- de plus, ici, les piles testées sont perdues à la vente (car elles sont vidées de toute énergie, donc, dans cette situation de produits testés perdus à la vente, on aura toujours un taux de sondage très faible bien inférieur à 5%.

4) comme $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = 230 - 210 = 20 \text{ , } \sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{30^2}{100} + \frac{20^2}{125}} = 3{,}493\right)$$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N(20 \text{ , } 3{,}493).$$

$$\Rightarrow \mathbf{P}(\overline{X_1} - \overline{X_2} \geq 30) = P\left(T \geq \frac{30 - 20}{3{,}493}\right) = P(T \geq 2{,}86) = 1 - P(T \leq 2{,}86) = 1 - 0{,}9979$$

= 0,21% = 0,0021.

La probabilité est donc très faible et on peut considérer que les sociétés 1 et 2 respectent leur cahier des charges et que l'écart de durée de vie moyenne est tendanciellement inférieur à 30h.

Lecture de l'énoncé :

X = Taux de factures non réglées dans les 10 j ouvrables suivant l'échéance.
Population mère :
Pour le 1) : p= 0,12 ; q=1-0,12 = 0, 88
N = plusieurs dizaines de milliers.
Pour le 2) : p= 0,09 ; q=1-0,09=0,91.
Echantillon :
Pour le 1) : n = 500 et f=0,14.
Pour le 2) : n=220 et f= $\frac{25}{220}$ =0,1136.

1)a) Déterminer l'intervalle de confiance à 95% et commenter.

Intervalle de confiance à 95% = risque de 5% $\Rightarrow$ $P(E(F) - t\,\sigma_F \leq F \leq E(F) + t\,\sigma_F) = 0,95$.

- **Problème de distribution d'échantillonnage d'une proportion** car : la question porte sur un taux ; on travaille sur échantillon ; on connaît les paramètres de la population mère (p = 0,12). Connaissant p, on connaît q et on connaît σ.
- **n = 500 > 60 $\Rightarrow$ Théorème central limite $\Rightarrow$ X ~ N.**
- **Taux de sondage < 5%** car la société gère plusieurs dizaines de milliers de factures, donc N est important et $\frac{n}{N}=\frac{500}{N} < 0,05$, donc pas de facteur d'exhaustivité.
- $X \sim N \Rightarrow F \sim N\left(0,12\ ,\ \sqrt{\frac{0,12\,x\,0,88}{500}} = 0,0145\right)$

Intervalle de confiance à 95% = [0,12 – (1,96)(0,0145) , 0,12 + (1,96)(0,0145)]
= **[0,0916 ; 0,1484]**

Le taux de factures non réglées dans les 10 j ouvrables suivant l'échéance va de 9,16% (hypothèse optimiste) à 14,84% (hypothèse pessimiste), en prenant un risque de 5% de se tromper.
Comme f = 0,14 est à l'intérieur de cet intervalle de confiance, l'échantillon est jugé représentatif de a population mère au risque de 5%.
Dans ce cas, on considère que les bases de raisonnement à ce sujet n'ont pas à être actualisées (on remarque toutefois que 0,14 est proche de la borne supérieure : peut être est-ce le signe d'un changement sur les habitudes de règlement des clients).

1)b) Si risque de 3% $\Rightarrow$ intervalle de confiance de 0,97

$\Rightarrow P(E(F)-t\sigma_F \leq F \leq E(F)+t\sigma_F)=0,97$.
Intervalle de confiance à 97% = [0,12-(2,17)(0,0145) ;0,12+ (2,17)(0,0145)]
= **[0,0885 ; 0,1515]**
$\Rightarrow$ **comme f = 0,14, dans l'intervalle donc l'échantillon est toujours jugé représentatif, au risque de 3%.**

2)a) En prenant un risque de 3%, l'échantillon du benchmark est-il représentatif ?

$f = \frac{25}{220} = 0,1136$
$\Rightarrow P(E(F)-t\sigma_F \leq F \leq E(F)+t\sigma_F)=0,97$.

$E(F) = 0,09$; $\sigma_F = \sqrt{\frac{0,09x0,91}{220}} = 0,0193.$

Intervalle de confiance à 97% = [0,09-(2,17)(0,0193);0,09+(2,17)(0,0193)]
= [0,0481;0,1319].
⇒ comme f = 0,1136 est à l'intervalle de confiance, l'échantillon est jugé représentatif au risque de 3%.

2)b) Quelle est la probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark ?
$p_1 = 0,12$ $p_2 = 0,09$
$n_1 = 500$ $n_2 = 220$
Question : P (F_1-$F_2 \leq 0,01$) ?

Procédure :

X_1 = taux de règlement hors délai pour société A.
X_2 = taux de règlement hors délai pour benchmark.
- Problème de distribution d'échantillonnage d'une différence de taux car :
• Nous gérons des taux à partir d'échantillons
• Nous devons comparer les taux de la société A et du benchmark
• Nous connaissons les paramètres des populations mères car nous connaissons p_1 et p_2.
- X_1 et X_2 suivent une loi Normale par application du théorème central limite ($n_1 = 500 > 60$ et $n_2 = 220 > 60$).
- Les taux de sondage pour la société A comme pour le benchmark sont < 5% car ils ont des dizaines de milliers de factures.
- X_1 et $X_2 \sim N \Rightarrow F_1$-$F_2 \sim N (E_{(F1-F2)} = 0,12-0,09 = 0,03$;

$\sigma_{F_1-F_1} = \sqrt{\frac{0,12x0,88}{500} + \frac{0,09x0,91}{220}} = 0,0242).$

P (F_1-$F_2 \leq 0,01$) $= P\left(T \leq \frac{0,01-0,03}{0,0242}\right) = P (T \leq -0,83) = P (T \geq +0,83)$

$= 1-P (T \leq +0,83) = 1-0,7967 =$ **0,2033.**

La probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark est de 20,33%.

Corrigés des problèmes d'estimation de la moyenne de la population mère

1)

Lecture de l'énoncé :

X = Durée de vie du tube cathodique d'une marque de TV, en heures.

Population mère :
m pas connue
$\sigma = 450$

Echantillon :
n = 55
$\bar{x} = 9\,500$

Question :
$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}90$

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N mais σ connu et n = 55 > 30).

⇒ t = 1,645.

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{55}} = 60{,}678.$

⇒ **Intervalle de confiance** = [9 500 – (1,645)(60,678) ; 9 500 + (1,645)(60,678)]
= **[9 400,18 ; 9 599,82].**

La moyenne des durées de vie de la population mère se situe entre 9 400,18 heures et 9 599,82 heures dans 90% des situations possibles (ou en prenant un risque de 10% de se tromper).

On peut dire aussi que :

- **5% des tubes auront une durée de vie < 9 400,18 heures ;**
- **5% des tubes auront une durée de vie > 9 599,82 heures.**

2)

Lecture de l'énoncé :

X = durée de vie des tubes cathodiques d'une marque de TV, en heures.

Population mère :

m pas connue

$\sigma = 450$.

Echantillon :

n = 25

$\bar{x} = 9\,500$.

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$.

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N, n = 25 < 30 mais σ connu = 450).

⇒ t = 2,575.

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{25}} = 90$.

⇒ **Intervalle de confiance** = [9 500 – (2,575)(90) ; 9 500 + (2,575)(90)]

= **[9 268,25 ; 9 731,75].**

Au risque de 1%, m va fluctuer entre ces 2 durées.

3)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

$n = 60$

$\bar{x} = 9\,450$

$\sigma' = 446{,}234$

Question :

$P(\bar{x} - t\,\sigma_{\bar{X}} \leq m \leq \bar{x} + t\,\sigma_{\bar{X}}) = 0{,}95.$

Procédure :

1) Problème d'estimation de m et de l'écart type car on ne connaît pas les paramètres de la population mère.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x} = 9\,450$

t de la loi Normale car $n = 60 > 30$ ⇒ $t = 1{,}96$.

$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s.

$$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{60}{59}(446{,}234)^2 = 202\,499{,}779 \Rightarrow s = \sqrt{202\,499{,}779} = 449{,}999 \approx 450$$

$$\Rightarrow \sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{450}{\sqrt{60}} = 58{,}094$$

⇒ **Intervalle de confiance** = [9 450 – (1,96)(58,094) ; 9 450 + (1,96)(58,094)]

= **[9 336,13 ; 9 563,87].**

Au risque de 5%, m se situera entre ces 2 durées.

4)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

n = 25

$\bar{x} = 9\,500$

$\sigma' = 440{,}908.$

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$

Procédure :

1) Problème d'estimation de m et σ car ces paramètres ne sont pas connus.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x} = 9\,500$

On utilise le t de Student Fisher car les 3 conditions sont réunies (**X ~ N,** σ pas connu, n = 25 < 30)

⇒ t = 2,797 (lecture de la table de St avec un nombre de ° de liberté = 25-1 = 24 et un risque de 1%).

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s ⇒ $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$.

$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{25}{24}(440{,}908)^2 = 202\,499{,}8588$

⇒ $s = \sqrt{202\,499{,}8588} = 449{,}9998 \approx 450.$

⇒ $\sigma_{\bar{x}} = \frac{450}{\sqrt{25}} = 90.$

⇒ **Intervalle de confiance** = [9 500 – (2,797)(90) ; 9 500 + (2,797)(90)]

= **[9 248,27 ; 9 751,73]**.

Au risque de 1%, m se situera entre ces 2 durées.

Corrigés des problèmes d'estimation de la proportion de la population mère

X = Taux de femmes utilisant la marque de lessive A dans une grande ville.

Lecture de l'énoncé :

Population mère :

p pas connue donc on ne connaît pas q, donc on ne connaît pas σ.

Echantillon :

n = 500

f = 0,35

Questions :

a) $P(f - t\,\sigma_F \leq p \leq f + t\,\sigma_F) = 0,95$.

b) Taille minimale de l'échantillon = calcul de n.

a)

Procédure :

1) Problème d'estimation de p car on ne connaît pas le taux de la population mère (c'est donc aussi et automatiquement un problème d'estimation de σ).

2) X ~ N car :

- la lessive est un produit de grande consommation, fabriqué en très grandes séries et donc standardisé $\Rightarrow$ critère d'atomicité.
- $n = 500 > 60 \Rightarrow$ Théorème central limite.

3) Comme il s'agit d'une grande ville, N est supposé important

$\Rightarrow$ **Taux de sondage < 5%** $\Rightarrow$ pas de facteur d'exhaustivité.

4)

- f = 0,35
- t de la loi N car σ pas connu et X ~ N, mais n = 500 > 30 donc nous n'avons pas les 3 conditions d'utilisation d'une loi de St $\Rightarrow$ t = 1,96.
- $\sigma_F = \sqrt{\frac{f(1-f)}{n-1}} = \sqrt{\frac{0,35(1-0,35)}{500-1}} = 0,02135.$

$\Rightarrow$ **Intervalle de confiance** = [0,35 – (1,96)(0,02135) ; 0,35 + (1,96)(0,02135)]

= **[0,3082 ; 0,3918].**

Au risque 5%, il y a entre 30,82 % et 39,18% des femmes de cette ville qui préfèrent la marque de lessive A.

b) Calcul de n

Pour calculer n il faut connaître l'erreur globale qui a été déterminée par le client, en accord avec le prestataire de l'étude.

Bien souvent le client désire une erreur la plus faible possible mais le prestataire doit rester réaliste et ne pas s'engager sur une erreur qu'il ne pourra pas respecter.

Cette erreur globale fait donc l'objet d'une négociation dès le départ sachant que plus l'erreur globale est faible, plus n est élevée et donc plus il faut interviewer de personnes donc plus le coût de l'étude s'élève.

Ici l'erreur globale e = 0,02 = t σ_F.

$$\Rightarrow 0{,}02 = 1{,}96\sqrt{\frac{(0{,}35)(0{,}65)}{n-1}}$$

Pour supprimer la racine carrée, nous allons élever les deux termes de cette égalité au carré et nous allons isoler n afin de trouver sa valeur :

$$\Rightarrow 0{,}02^2 = 1{,}96^2 \frac{(0{,}35)(0{,}65)}{n-1}$$

$$\Rightarrow n - 1 = 1{,}96^2 \frac{(0{,}35)(0{,}65)}{0{,}02^2}$$

$$\Rightarrow n = 1{,}96^2 \frac{(0{,}35)(0{,}65)}{0{,}02^2} + 1$$

$$\Rightarrow \mathbf{n = 2\,185{,}91 \approx 2\,186.}$$

Pour avoir une erreur globale de 2%, il faut interviewer 2 186 dames de cette ville.

Remarque :

Dans la première question, on interviewait 500 dames $\Rightarrow e = 1{,}96\sqrt{\frac{(0{,}35)(0{,}65)}{499}} = 0{,}042.$

$\Rightarrow$ pour n = 500 , e = 0,042
pour n = 2 186 , e = 0,02.

On vérifie ainsi une loi énoncée souvent en marketing : « Lorsque l'on veut diviser par 2 l'erreur globale, il faut multiplier par 4 le nombre de personnes à interviewer ».
Bien sûr, cette loi ne nous donne qu'une approximation de n compte tenu du choix de l'erreur globale.
Le seul moyen de connaître exactement n est de le calculer comme indiqué.

INDUCTION ou INFERENCE STATISTIQUE

I - INTRODUCTION

1) Définitions

a) L'induction (ou inférence) statistique

Ensemble des méthodes statistiques qui ont pour but de tirer des conclusions ou d'aider à prendre des décisions au sujet d'une population mère à partir d'un échantillon aléatoire prélevé dans cette population.

b) La population mère

Ensemble de tous les individus concernés par le phénomène économique étudié.

Exemples :

Tous les électeurs d'un territoire pour une élection politique ;
Tous les consommateurs potentiels pour la vente d'un produit ;
Tous les produits fabriqués pour une série de production....

- **La connaissance exacte des paramètres de la population mère** (Moyenne, Ecart-type, Proportion, pour les processus gaussiens) **demande l'analyse exhaustive de celle-ci = un recensement, et une actualisation constante.**
- Dans la pratique, les recensements sont rares à cause du coût induit et du temps nécessaire à l'analyse.

De plus, dans certains cas, il y a même une impossibilité matérielle de réalisation : taille trop importante, non connaissance des limites du champ d'étude ou alors destruction des éléments testés.

c) Les échantillons

Il existe **2 grandes catégories de méthodes d'échantillonnages** :

- **L'échantillonnage non aléatoire** : l'analyste utilise son expérience et son jugement pour constituer l'échantillon avec, le plus souvent, **l'application de statistiques officielles visant à recréer l'existant (la méthode des quotas)**, avec tous les risques de non représentativité de celui-ci.
- **L'échantillonnage aléatoire ou probabiliste : il permet de calculer précisément l'erreur due à l'échantillonnage et, par conséquent, de juger de la valeur l'information partielle obtenue et donc de la représentativité de l'échantillon**, lorsque l'on observe toutes les conditions de l'étude.

On parle d'échantillon aléatoire simple (SAS), lorsque :

- Chaque unité de la population mère a la même probabilité d'être sélectionnée dans l'échantillon ;
- Chaque échantillon de même taille, tiré de la population mère, a la même probabilité d'être choisi.

2) L'échantillon aléatoire simple peut être tiré avec ou sans remise

a) L'échantillon aléatoire simple avec remise = Echantillon non exhaustif

Chaque unité est remise dans la population mère après avoir été observée et avant qu'une autre unité soit choisie.

Avantage : Probabilité de base p est constante.

Inconvénient : Mauvaise gestion du temps et risque de tirer au sort plusieurs fois la même unité..

Dans ce cas, il y a indépendance entre les résultats d'un tirage à l'autre et chaque unité conserve la même probabilité d'être sélectionnée $\Rightarrow$ le processus est stationnaire.

b) L'échantillon aléatoire simple sans remise = Echantillon exhaustif

L'unité tirée au sort n'est pas remise dans la population mère.
Il n'y a plus d'indépendance d'un tirage à l'autre et, pour chaque unité particulière, la probabilité de base d'être choisie d'un tirage à l'autre est modifiée.
Pour les probabilités de gestion :

- Si le taux de sondage $(\frac{n}{N}) < 5\%$: on suppose l'indépendance entre les résultats d'un tirage à l'autre et on ne modifie rien → **on ne tient pas compte du processus de changement des probabilités.**

- Si le taux de sondage est $\geq 5\%$ **: on tient compte de la détérioration de la probabilité et on applique un facteur de correction, le facteur d'exhaustivité** $\frac{N-n}{N-1}$.

3) Il existe 2 grandes catégories de problèmes

a) Les problèmes de distribution d'échantillonnage

- On connaît la valeur de certains paramètres de la population mère.
- On cherche à induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans l'échantillon.
- On cherche à répondre à la question de la représentativité de l'échantillon.

b) Les problèmes d'estimation

- On ne connaît pas la valeur des paramètres de la population mère.
- On étudie statistiquement les paramètres de l'échantillon tiré au sort, et on essaye d'induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans la population mère ⇒ on cherche donc, à partir des valeurs de l'échantillon, la valeur des paramètres de la population mère, compte tenu d'un risque choisi et géré.
Il existe 2 **grands types d'approches complémentaires :**
- L'estimation ponctuelle ;
- L'estimation par intervalle de confiance.

4) Présentation des symboles utilisés

Pour la population mère :
- Taille = N ;
- Moyenne = m ;
- Variance = σ^2 **;**
- Ecart-type = σ **;**
- Taux = p.

Pour l'échantillon :
- Taille = n ;
- Moyenne = $\bar{x}$ **;**
- Variance = σ'^2 **;**
- Ecart-type = σ' **;**
- Taux = f.

II – LES PROBLEMES DE DISTRIBUTION D'ECHANTILLONNAGE

1) <u>On connaît les paramètres de la population mère</u>

a) Il y a C_N^n échantillons de taille n et ils ont tous leurs caractéristiques et leurs valeurs ($\bar{x}, \sigma'$, f) qui peuvent être différentes.

- Le problème de distribution d'échantillonnage gère le risque de prendre un seul échantillon parmi les C_N^n possibles, en essayant de tenir compte de la diversité des résultats possibles sur tous les échantillons de taille n.

- Dans ce cas, on utilise un outil $\overline{X}$, pour la gestion des moyennes, qui est sensé représenter toutes les moyennes ($\bar{x}$) d'échantillons de taille n.

- De même, on utilise un outil F, pour la gestion des taux, sensé représenter tous les taux (f) d'échantillons de taille n.

b) Procédure à suivre

- Qualifier le problème ;

- Prouver que le phénomène X suit une loi Normale (X ~ N) ;

- Chercher le taux de sondage $\frac{n}{N}$;

- Calculer la probabilité demandée ou montrer que l'échantillon est représentatif de la population mère en prenant un risque de se tromper.

2) <u>Problème de distribution d'échantillonnage de la moyenne</u>

Ce type de problème se pose lorsque nous gérons une question portant sur une moyenne à partir d'un échantillon.

On utilise la variable aléatoire $\overline{X}$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif (tirage avec remise) ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

$$\text{Si } X \sim N(m, \sigma) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = m ; \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\right)$$

Avec $E(\overline{X})$ = moyenne de toutes les moyennes $\bar{x}$ d'échantillons de taille n.

$\sigma_{\overline{X}}$ = dispersion moyenne de l'ensemble des $\bar{x}$ autour de $E(\overline{X})$.

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X \sim N(m, \sigma) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = m ; \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right)$$

3) Problème de distribution d'échantillonnage d'une différence de moyennes

Ce type de problème se pose lorsque nous avons à gérer une question portant sur une comparaison de moyennes à partir d'échantillons.
La différence permet de comparer 2 éléments.
On utilise la variable aléatoire $\overline{X_1} - \overline{X_2}$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

Si $X_1 \sim N(m_1, \sigma_1)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = m_1 - m_2 ;\ \sigma_{(\overline{X_1} - \overline{X_2})} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$X_2 \sim N(m_2, \sigma_2)$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

Si $X_1 \sim N(m_1, \sigma_1)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = m_1 - m_2 ;\ \sigma_{(\overline{X_1} - \overline{X_2})} = \sqrt{\frac{\sigma_1^2}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2}\frac{N_2 - n_2}{N_2 - 1}}\right)$$

$X_2 \sim N(m_2, \sigma_2)$

4) Problème de distribution d'échantillonnage d'une proportion

Ce type de problème se pose lorsque nous gérons une question portant sur des proportions (%) à partir d'échantillons.
Une proportion est un nombre relatif c'est-à-dire un rapport de 2 chiffres : o peut dire aussi un taux, un ratio, une fréquence…
On appelle :

- **p la proportion de la population mère ;**
- **f la fréquence de l'échantillon de taille n.**

F représente la variable aléatoire qui gère toutes les valeurs possibles de toutes les fréquences f de tous les échantillons de taille n.
F est l'équivalent pour les taux de $\overline{X}$ pour les moyennes.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

Si $X \sim N \Rightarrow F \sim N\left(E_{(F)} = p ;\ \sigma_F = \sqrt{\frac{pq}{n}}\right)$ avec $q = 1 - p$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X \sim N \Rightarrow F \sim N\left(E_{(F)} = p ;\ \sigma_F = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

5) Problème de distribution d'échantillonnage d'une différence de proportions

Ce type de problème se pose lorsque nous avons à gérer une question portant sur une comparaison de taux à partir d'échantillons.
On utilise la variable aléatoire $F_1 - F_2$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

$$\text{Si } X_1 \sim N \quad X_2 \sim N \Rightarrow F_1 - F_2 \sim N\left(E_{(F_1-F_2)} = p_1 - p_2 ;\ \sigma_{F_1-F_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X_1 \sim N \quad X_2 \sim N \Rightarrow F_1 - F_2 \sim N\left(E_{(F_1-F_2)} = p_1 - p_2 ;\ \sigma_{F_1-F_2} = \sqrt{\frac{p_1q_1}{n_1}\frac{N_1-n_1}{N_1-1} + \frac{p_2q_2}{n_2}\frac{N_2-n_2}{N_2-1}}\right)$$

III – LES PROBLEMES D'ESTIMATION

- On ne connaît pas tous les paramètres de la population mère et on cherche à les estimer.
Le problème d'estimation gère, en plus, les risques d'élargir les connaissances que l'on a de l'échantillon à l'ensemble de la population mère.

- Il permet d'estimer les paramètres de la population mère :

- **m et σ, pour les moyennes ;**
- **p pour les proportions ;**

à partir des observations de l'échantillon :

- **$\bar{x}$ et σ', pour les moyennes ;**
- **f pour les proportions.**

1) Présentation de toutes les situations possibles

a) Pour les moyennes, on peut rencontrer 4 situations :

- On connaît m et σ : Problème de distribution d'échantillonnage de la moyenne et de l'écart-type.

- On connaît m et on ne connaît pas σ : Problème de distribution d'échantillonnage de la moyenne et un problème d'estimation de l'écart-type.

- On ne connaît pas m et on connaît σ : Problème d'estimation de la moyenne et un problème de distribution d'échantillonnage de l'écart-type.

- On ne connaît pas m et σ : Problème d'estimation de la moyenne et de l'écart-type.

Il faudra donc bien lire l'énoncé et étudier la situation pour voir les informations dont on dispose et en déduire la bonne qualification du problème à résoudre.

b) Pour les taux, on peut rencontrer 2 situations :

- On connaît p donc, on connaît tout (q = 1 - p) : Problème de distribution d'échantillonnage d'une proportion.

- On ne connaît pas p donc, on ne connaît rien : Problème d'estimation d'une proportion.

2) L'estimation ponctuelle

- Lorsque l'on ne connaît pas les paramètres de la population mère, on va d'abord les estimer à partir des informations statistiques obtenues par l'observation de l'échantillon.

Bien sûr, ces statistiques sont loin d'être parfaites et surtout ne peuvent refléter tous les résultats possibles (en effet, chaque échantillon de taille n peut avoir ses propres valeurs).

a) L'estimateur ponctuel de m = $\bar{x}$.

b) L'estimateur ponctuel de p = f.

c) L'estimateur ponctuel de la Variance de la population mère = s^2

- Si le tirage est non exhaustif ou si le tirage est exhaustif avec un taux de sondage < 5% :

- **Calcul de la variance statistique de l'échantillon =** $\sigma'^2 = \frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2$
- **Calcul de s^2 :** $\boxed{s^2 = \frac{n}{n-1}\sigma'^2}$
- **Calcul de s :** $\boxed{s = \sqrt{s^2}}$.

- Si le tirage est exhaustif avec un taux de sondage ≥ 5% :

- **Calcul de la variance statistique de l'échantillon =** $\sigma'^2 = \frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2$
- **Calcul de s^2 :** $\boxed{s^2 = \frac{N-1}{n-1}\frac{n}{N}\sigma'^2}$

▪ **Calcul de s :** $\boxed{s = \sqrt{s^2}}$.

3) L'estimation par intervalle de confiance

a) Estimation de la moyenne de la population mère m

$$\boxed{P(\bar{x} - t\,\sigma_{\bar{x}} < m < \bar{x} + t\,\sigma_{\bar{x}}) = \alpha\,\%}$$

Procédure :

- Qualifier le problème ;

- Prouver que le phénomène X suit une loi Normale ($X \sim N$) ;

- Chercher le taux de sondage $\frac{n}{N}$;

- Calcul d'estimation des paramètres de la population mère, en utilisant les estimateurs ponctuels et l'intervalle de confiance : $P(\bar{x} - t\,\sigma_{\bar{x}} < m < \bar{x} + t\,\sigma_{\bar{x}}) = \alpha\,\%$

▪ **Calcul de** $\bar{x} = \frac{\sum n_i x_i}{N}$ avec $N = \sum n_i$

▪ **Trouver t : dans les problèmes d'estimation, 2 situations sont possibles :**

Lorsque les 3 conditions suivantes sont réunies :

- **$X \sim N$**
- **σ pas connu,**
- **$n < 30$: on lit le t dans la table de Student Fisher : en fonction du risque accepté et du nombre de degrés de liberté (n – 1).**

Exemple :

Dans la table de Student de la page 12, si n = 19 et que le risque est de 10% :

On prend la ligne 18 = 19 - 1

la colonne 0,95

et à l'intersection de la ligne 18 et de la colonne 0,95, on lit le t de Student = 1,734.

Dans tous les autres cas de figure : on utilise la table de Laplace-Gauss = la loi Normale.

▪ **Calcul de** $\sigma_{\bar{x}}$ **: 2 cas de figure :**

- **Si σ est connu** : pas de problème : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- **Si σ n'est pas connu :**
  - on calcule s, estimateur ponctuel de l'écart type de la population mère.
  - on fait $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$.

b) Estimation de la proportion de la population mère p

$$\boxed{P\,(f - t\,\sigma_F < p < f + t\,\sigma_F) = \alpha\,\%}$$

Procédure :

- **Qualifier le problème ;**
- **Prouver que le phénomène X suit une loi Normale** $(X \sim N)$;
- **Chercher le taux de sondage** $\frac{n}{N}$;
- **Calcul d'estimation des paramètres de la population mère** : en utilisant les estimateurs ponctuels et l'intervalle de confiance :

$$P\,(f - t\,\sigma_F < p < f + t\,\sigma_F) = \alpha\,\%$$

- **calcul de** $f = \frac{n_i}{\sum n_i}$
- **trouver t** : idem que pour les moyennes : 2 situations.
- **calcul de** $\sigma_F = \sqrt{\frac{f(1-f)}{n-1}}$.

Corrigés du problème de distribution d'échantillonnage de la moyenne $\overline{X}$
(usine textile)

Lecture de l'énoncé :

X = Longueur des morceaux de tissu en cm

Population mère :
$E(X) = m = 90$ la long mor cea de Tissu en moyenne
$\sigma = 0,60$ l'ecat type
$N = 10\,000$ (1er cas de figue)
$N = 2\,000$ (2ème cas de figure)

Echantillon :
$n = 200$ echantillon aleatoire
$\bar{x} = 90,30$ moyenne observé

a) 1ère Question : P ($\overline{X} \leq 89,90$) ?

Procédure pour N = 10 000 :

1) C'est un problème de distribution d'échantillonnage de la moyenne car :
- la question porte sur une moyenne ;
- on travaille sur un échantillon ;
- on connaît les paramètres de la population mère ($m = 90$, $\sigma = 0,60$)

2) X ~ N (90 , 0,60) car :
- $n = 200 > 60 \Rightarrow$ Théorème central limite
- il s'agit d'une production en série et standardisée. On peut appliquer le critère d'atomicité qui est un des critères liés à la loi Normale.

3) Taux de sondage $= \frac{n}{N} = \frac{200}{10\,000} = 0,02 < 0,05 \Rightarrow$ on ne doit donc pas appliquer le facteur d'exhaustivité.

4) $X \sim N(90 , 0,60) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = 90 ; \sigma_{\overline{X}} = \frac{0,60}{\sqrt{200}} = 0,04243\right)$

$$\Rightarrow \mathbf{P}(\overline{X} \leq 89,90) = P\left(T \leq \frac{89,90 - 90}{0,04243}\right) = P(T \leq -2,36)$$
$$= 1 - P(T \leq +2,36) = 1 - 0,9909$$
$$= \mathbf{0,91\ \% = 0,0091.}$$

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

Aoujous pb 1

TD 2+3

Procédure pour N = 2 000 :

1) Même réponse que pour N = 10 000.

2) Même réponse que pour N = 10 000.

3) Taux de sondage = $\frac{200}{2\,000} = 0{,}10 > 0{,}05 \Rightarrow$ on va appliquer le facteur d'exhaustivité.

4) $X \sim N(90 , 0{,}60) \Rightarrow \overline{X} \sim N(E(\overline{X}) = 90 ; \sigma_{\overline{X}} = \frac{0{,}60}{\sqrt{200}} \sqrt{\frac{2\,000-200}{2\,000-1}} =$ 0,04026)

$\Rightarrow$ **P** $(\overline{X} \leq 89{,}90) = P(T \leq \frac{89{,}90-90}{0{,}04026}) = P(T \leq -2{,}48) = 1 - P(T \leq +2{,}48) = 1 - 0{,}9934$ **= 0,66 % = 0,0066.**

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

b) 2ème Question : Montrer que l'échantillon est représentatif de la population mère pour N = 10 000.

- Prendre un risque de 5% de se tromper $\Rightarrow$ l'intervalle symétrique que l'on veut respecter doit couvrir 95% des situations possibles.

$\Rightarrow P(m - t\,\sigma_{\overline{X}} \leq \overline{X} \leq m + t\,\sigma_{\overline{X}}) = 0{,}95.$

- m = 90
- t de la loi Normale pour un intervalle symétrique de 95% = ± 1,96

(Pour trouver + t :

- rechercher dans la table de la loi N la probabilité la plus proche de 97,5% = 0,975.
- on voit que 0,975 se situe à l'intersection de la ligne 1,9 et de la colonne 0,06

$\Rightarrow$ + t = + 1,96).

- $\sigma_{\overline{X}}$ = 0,04243 (pour N = 10 000)

$\Rightarrow$ **Intervalle de confiance** = [90 - (1,96)(0,04243) ; 90 + (1,96)(0,04243)]
= **[89,917 ; 90,083]**

95% des moyennes de morceaux de tissu sont dans cet intervalle de confiance

- Comme la moyenne de l'échantillon $\bar{x}$ = 90,30 se trouve à l'extérieur de l'intervalle de confiance, on peut dire, au risque de 5% de se tromper, que l'échantillon n'est pas représentatif de la population mère.

Dans ce cas, on remet les 200 morceaux de tissu dans la population mère (car ici, les morceaux testés sont encore utilisables) et il faut de nouveau tirer un échantillon aléatoire de 200 morceaux de tissu afin de refaire la procédure.

Corrigés du problème de distribution d'échantillonnage d'une différence de moyennes $\overline{X_1} - \overline{X_2}$ (piles électriques)

Lecture de l'énoncé :

X_1 = durée d'utilisation des piles de la société 1 en heures
X_2 = durée d'utilisation des piles de la société 2 en heures

Populations mères :
$m_1 = 230$ $\sigma_1 = 30$
$m_2 = 210$ $\sigma_2 = 20$

Echantillons :
$n_1 = 100$
$n_2 = 125$

Question : $P(\overline{X_1} - \overline{X_2} \geq 30)$?

Procédure :

1) Problème de distribution d'échantillonnage d'une différence de moyennes car :

— la question porte sur une comparaison de moyennes ;

- on travaille sur échantillons ;
- on connaît tous les paramètres des populations mères (m_1 et m_2 , σ_1 et σ_2).

2) $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$ car :

- $n_1 = 100 > 60$ et $n_2 = 125 > 60 \Rightarrow$ Théorème central limite ;
- les piles sont fabriquées en très grandes séries $\Rightarrow$ critère d'atomicité $\Rightarrow$ loi Normale.

3) Taux de sondage < 5% car :

- même si on ne connaît pas exactement N_1 et N_2 on sait que les tailles des populations mères sont très importantes vue la fabrication en très grande séries ;

- de plus, ici, les piles testées sont perdues à la vente (car elles sont vidées de toute énergie, donc, dans cette situation de produits testés perdus à la vente, on aura toujours un taux de sondage très faible bien inférieur à 5%.

4) comme $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = 230 - 210 = 20 \text{ , } \sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{30^2}{100} + \frac{20^2}{125}} = 3{,}493\right)$$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N(20 \text{ , } 3{,}493).$$

$$\Rightarrow P(\overline{X_1} - \overline{X_2} \geq 30) = P\left(T \geq \frac{30 - 20}{3{,}493}\right) = P(T \geq 2{,}86) = 1 - P(T \leq 2{,}86) = 1 - 0{,}9979$$

= 0,21% = 0,0021.

La probabilité est donc très faible et on peut considérer que les sociétés 1 et 2 respectent leur cahier des charges et que l'écart de durée de vie moyenne est tendanciellement inférieur à 30h.

Lecture de l'énoncé :

X = Taux de factures non réglées dans les 10 j ouvrables suivant l'échéance.
Population mère :
Pour le 1) : p= 0,12 ; q=1-0,12 = 0, 88
N = plusieurs dizaines de milliers.
Pour le 2) : p= 0,09 ; q=1-0,09=0,91.
Echantillon :
Pour le 1) : n = 500 et f=0,14.
Pour le 2) : n=220 et f=$\frac{25}{220}$=0,1136.

1)a) Déterminer l'intervalle de confiance à 95% et commenter.

Intervalle de confiance à 95% = risque de 5% $\Rightarrow$ $P(E(F) - t\,\sigma_F \leq F \leq E(F) + t\,\sigma_F) = 0,95$.

- **Problème de distribution d'échantillonnage d'une proportion** car : la question porte sur un taux ; on travaille sur échantillon ; on connaît les paramètres de la population mère (p = 0,12). Connaissant p, on connaît q et on connaît σ.
- **n = 500 > 60 $\Rightarrow$ Théorème central limite $\Rightarrow$ X ~ N.**
- **Taux de sondage < 5%** car la société gère plusieurs dizaines de milliers de factures, donc N est important et $\frac{n}{N}=\frac{500}{N} < 0,05$, donc pas de facteur d'exhaustivité.

- $X \sim N \Rightarrow F \sim N\left(0,12\ ,\ \sqrt{\frac{0,12\,x\,0,88}{500}} = 0,0145\right)$

Intervalle de confiance à 95% = [0,12 – (1,96)(0,0145) , 0,12 + (1,96)(0,0145)]
= **[0,0916 ; 0,1484]**

Le taux de factures non réglées dans les 10 j ouvrables suivant l'échéance va de 9,16% (hypothèse optimiste) à 14,84% (hypothèse pessimiste), en prenant un risque de 5% de se tromper.

Comme f = 0,14 est à l'intérieur de cet intervalle de confiance, l'échantillon est jugé représentatif de a population mère au risque de 5%.

Dans ce cas, on considère que les bases de raisonnement à ce sujet n'ont pas à être actualisées (on remarque toutefois que 0,14 est proche de la borne supérieure : peut être est-ce le signe d'un changement sur les habitudes de règlement des clients).

1)b) Si risque de 3% $\Rightarrow$ intervalle de confiance de 0,97

$\Rightarrow P(E(F)-t\sigma_F \leq F \leq E(F)+t\sigma_F)=0,97$.

Intervalle de confiance à 97% = [0,12-(2,17)(0,0145) ;0,12+ (2,17)(0,0145)]
= **[0,0885 ; 0,1515]**

$\Rightarrow$ **comme f = 0,14, dans l'intervalle donc l'échantillon est toujours jugé représentatif, au risque de 3%.**

2)a) En prenant un risque de 3%, l'échantillon du benchmark est-il représentatif ?

$f = \frac{25}{220} = 0,1136$

$\Rightarrow P(E(F)-t\sigma_F \leq F \leq E(F)+t\sigma_F)=0,97$.

$E(F) = 0,09$; $\sigma_F = \sqrt{\frac{0,09x0,91}{220}} = 0,0193.$

Intervalle de confiance à 97% = [0,09-(2,17)(0,0193);0,09+(2,17)(0,0193)]
= [0,0481;0,1319].
⇒ comme f = 0,1136 est à l'intervalle de confiance, l'échantillon est jugé représentatif au risque de 3%.

2)b) Quelle est la probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark ?
$p_1 = 0,12$ $p_2 = 0,09$
$n_1 = 500$ $n_2 = 220$
Question : P (F_1-$F_2 \leq 0,01$) ?

Procédure :

X_1 = taux de règlement hors délai pour société A.
X_2 = taux de règlement hors délai pour benchmark.

- Problème de distribution d'échantillonnage d'une différence de taux car :

• Nous gérons des taux à partir d'échantillons
• Nous devons comparer les taux de la société A et du benchmark
• Nous connaissons les paramètres des populations mères car nous connaissons p_1 et p_2.

- X_1 et X_2 suivent une loi Normale par application du théorème central limite ($n_1 = 500 > 60$ et $n_2 = 220 > 60$).

- Les taux de sondage pour la société A comme pour le benchmark sont < 5% car ils ont des dizaines de milliers de factures.

- X_1 et $X_2 \sim N \Rightarrow F_1$-$F_2 \sim N\ (E_{(F1-F2)} = 0,12-0,09 = 0,03$;

$\sigma_{F_1-F_2} = \sqrt{\frac{0,12x0,88}{500} + \frac{0,09x0,91}{220}} = 0,0242).$

P (F_1-$F_2 \leq 0,01$) $= P\ (T \leq \frac{0,01-0,03}{0,0242}) = P\ (T \leq -0,83) = P\ (T \geq +0,83)$

$= 1-P\ (T \leq +0,83) = 1-0,7967 =$ **0,2033.**

La probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark est de 20,33%.

Corrigés des problèmes d'estimation de la moyenne de la population mère

1)

Lecture de l'énoncé :

X = Durée de vie du tube cathodique d'une marque de TV, en heures.

Population mère :
m pas connue
$\sigma = 450$

Echantillon :
n = 55
$\bar{x} = 9\,500$

Question :
$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}90$

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ .

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N mais σ connu et n = 55 > 30).

⇒ t = 1,645.

- $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{450}{\sqrt{55}} = 60{,}678.$

⇒ **Intervalle de confiance** = [9 500 – (1,645)(60,678) ; 9 500 + (1,645)(60,678)]
= **[9 400,18 ; 9 599,82].**

La moyenne des durées de vie de la population mère se situe entre 9 400,18 heures et 9 599,82 heures dans 90% des situations possibles (ou en prenant un risque de 10% de se tromper).

On peut dire aussi que :

- **5% des tubes auront une durée de vie < 9 400,18 heures ;**
- **5% des tubes auront une durée de vie > 9 599,82 heures.**

2)

Lecture de l'énoncé :

X = durée de vie des tubes cathodiques d'une marque de TV, en heures.

Population mère :

m pas connue

$\sigma = 450$.

Echantillon :

n = 25

$\bar{x} = 9\,500$.

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$.

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N, n = 25 < 30 mais σ connu = 450).

⇒ t = 2,575.

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{25}} = 90$.

⇒ **Intervalle de confiance** = [9 500 – (2,575)(90) ; 9 500 + (2,575)(90)]

= **[9 268,25 ; 9 731,75].**

Au risque de 1%, m va fluctuer entre ces 2 durées.

3)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

n = 60

$\bar{x}$ = 9 450

σ' = 446,234

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0,95.$

Procédure :

1) Problème d'estimation de m et de l'écart type car on ne connaît pas les paramètres de la population mère.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x}$ = 9 450

t de la loi Normale car n = 60 > 30 ⇒ t = 1,96.

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s.

$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{60}{59}(446,234)^2 = 202\,499,779 \Rightarrow s = \sqrt{202\,499,779} = 449,999 \approx 450$

$\Rightarrow \sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{450}{\sqrt{60}} = 58,094$

⇒ **Intervalle de confiance** = [9 450 – (1,96)(58,094) ; 9 450 + (1,96)(58,094)]

= **[9 336,13 ; 9 563,87].**

Au risque de 5%, m se situera entre ces 2 durées.

4)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

n = 25

$\bar{x} = 9\,500$

$\sigma' = 440{,}908.$

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$

Procédure :

1) Problème d'estimation de m et σ car ces paramètres ne sont pas connus.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x} = 9\,500$

On utilise le t de Student Fisher car les 3 conditions sont réunies (**X ~ N,** σ pas connu, n = 25 < 30)

⇒ t = 2,797 (lecture de la table de St avec un nombre de ° de liberté = 25-1 = 24 et un risque de 1%).

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s ⇒ $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$.

$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{25}{24}(440{,}908)^2 = 202\,499{,}8588$

⇒ $s = \sqrt{202\,499{,}8588} = 449{,}9998 \approx 450.$

⇒ $\sigma_{\bar{x}} = \frac{450}{\sqrt{25}} = 90.$

⇒ **Intervalle de confiance** = [9 500 – (2,797)(90) ; 9 500 + (2,797)(90)]

= **[9 248,27 ; 9 751,73].**

Au risque de 1%, m se situera entre ces 2 durées.

INDUCTION ou INFERENCE STATISTIQUE

I - INTRODUCTION

1) Définitions

a) L'induction (ou inférence) statistique

Ensemble des méthodes statistiques qui ont pour but de tirer des conclusions ou d'aider à prendre des décisions au sujet d'une population mère à partir d'un échantillon aléatoire prélevé dans cette population.

b) La population mère

Ensemble de tous les individus concernés par le phénomène économique étudié.

Exemples :

Tous les électeurs d'un territoire pour une élection politique ;
Tous les consommateurs potentiels pour la vente d'un produit ;
Tous les produits fabriqués pour une série de production....

- **La connaissance exacte des paramètres de la population mère** (Moyenne, Ecart-type, Proportion, pour les processus gaussiens) **demande l'analyse exhaustive de celle-ci = un recensement, et une actualisation constante.**
- Dans la pratique, les recensements sont rares à cause du coût induit et du temps nécessaire à l'analyse.

De plus, dans certains cas, il y a même une impossibilité matérielle de réalisation : taille trop importante, non connaissance des limites du champ d'étude ou alors destruction des éléments testés.

c) Les échantillons

Il existe **2 grandes catégories de méthodes d'échantillonnages** :

- **L'échantillonnage non aléatoire** : l'analyste utilise son expérience et son jugement pour constituer l'échantillon avec, le plus souvent, **l'application de statistiques officielles visant à recréer l'existant (la méthode des quotas)**, avec tous les risques de non représentativité de celui-ci.
- **L'échantillonnage aléatoire ou probabiliste : il permet de calculer précisément l'erreur due à l'échantillonnage et, par conséquent, de juger de la valeur l'information partielle obtenue et donc de la représentativité de l'échantillon**, lorsque l'on observe toutes les conditions de l'étude.

On parle d'échantillon aléatoire simple (SAS), lorsque :

- Chaque unité de la population mère a la même probabilité d'être sélectionnée dans l'échantillon ;
- Chaque échantillon de même taille, tiré de la population mère, a la même probabilité d'être choisi.

2) L'échantillon aléatoire simple peut être tiré avec ou sans remise

a) L'échantillon aléatoire simple avec remise = Echantillon non exhaustif

Chaque unité est remise dans la population mère après avoir été observée et avant qu'une autre unité soit choisie.

Avantage : Probabilité de base p est constante.

Inconvénient : Mauvaise gestion du temps et risque de tirer au sort plusieurs fois la même unité..

Dans ce cas, il y a indépendance entre les résultats d'un tirage à l'autre et chaque unité conserve la même probabilité d'être sélectionnée $\Rightarrow$ le processus est stationnaire.

b) L'échantillon aléatoire simple sans remise = Echantillon exhaustif

L'unité tirée au sort n'est pas remise dans la population mère.

Il n'y a plus d'indépendance d'un tirage à l'autre et, pour chaque unité particulière, la probabilité de base d'être choisie d'un tirage à l'autre est modifiée.

Pour les probabilités de gestion :

- Si le taux de sondage $(\frac{n}{N}) < 5\%$: on suppose l'indépendance entre les résultats d'un tirage à l'autre et on ne modifie rien → **on ne tient pas compte du processus de changement des probabilités.**

- Si le taux de sondage est $\geq 5\%$ **: on tient compte de la détérioration de la probabilité** et **on applique un facteur de correction, le facteur d'exhaustivité** $\frac{N-n}{N-1}$.

3) Il existe 2 grandes catégories de problèmes

a) Les problèmes de distribution d'échantillonnage

- On connaît la valeur de certains paramètres de la population mère.

- On cherche à induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans l'échantillon.

- On cherche à répondre à la question de la représentativité de l'échantillon.

b) Les problèmes d'estimation

- On ne connaît pas la valeur des paramètres de la population mère.

- On étudie statistiquement les paramètres de l'échantillon tiré au sort, et **on essaye d'induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans la population mère ⇒ on cherche donc, à partir des valeurs de l'échantillon, la valeur des paramètres de la population mère, compte tenu d'un risque choisi et géré.**

Il existe **2 grands types d'approches complémentaires** :

- L'estimation ponctuelle ;

- L'estimation par intervalle de confiance.

4) Présentation des symboles utilisés

Pour la population mère :

- Taille = N ;

- Moyenne = m ;

- Variance = σ^2 **;**

- Ecart-type = σ **;**

- Taux = p.

Pour l'échantillon :

- Taille = n ;

- Moyenne = $\bar{x}$ **;**

- Variance = σ'^2 **;**

- Ecart-type = σ' **;**

- Taux = f.

II – LES PROBLEMES DE DISTRIBUTION D'ECHANTILLONNAGE

1) <u>On connaît les paramètres de la population mère</u>

a) Il y a C_N^n échantillons de taille n et ils ont tous leurs caractéristiques et leurs valeurs ($\bar{x}, \sigma'$, f) qui peuvent être différentes.

- Le problème de distribution d'échantillonnage gère le risque de prendre un seul échantillon parmi les C_N^n possibles, en essayant de tenir compte de la diversité des résultats possibles sur tous les échantillons de taille n.

- Dans ce cas, on utilise un outil $\overline{X}$, pour la gestion des moyennes, qui est sensé représenter toutes les moyennes ($\bar{x}$) d'échantillons de taille n.

- De même, on utilise un outil F, pour la gestion des taux, sensé représenter tous les taux (f) d'échantillons de taille n.

b) Procédure à suivre

- Qualifier le problème ;

- Prouver que le phénomène X suit une loi Normale ($X \sim N$) ;

- Chercher le taux de sondage $\frac{n}{N}$;

- Calculer la probabilité demandée ou montrer que l'échantillon est représentatif de la population mère en prenant un risque de se tromper.

• 2) <u>Problème de distribution d'échantillonnage de la moyenne</u>

Ce type de problème se pose lorsque nous gérons une question portant sur une moyenne à partir d'un échantillon.

On utilise la variable aléatoire $\overline{X}$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif (tirage avec remise) ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

$$\text{Si } X \sim N(m, \sigma) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = m ; \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

Avec $E(\overline{X})$ = moyenne de toutes les moyennes $\bar{x}$ d'échantillons de taille n.

$\sigma_{\overline{X}}$ = dispersion moyenne de l'ensemble des $\bar{x}$ autour de $E(\overline{X})$.

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X \sim N(m, \sigma) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = m ; \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right)$$

3) Problème de distribution d'échantillonnage d'une différence de moyennes

Ce type de problème se pose lorsque nous avons à gérer une question portant sur une comparaison de moyennes à partir d'échantillons.

La différence permet de comparer 2 éléments.

On utilise la variable aléatoire $\overline{X_1} - \overline{X_2}$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

Si $X_1 \sim N(m_1, \sigma_1)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = m_1 - m_2 ;\ \sigma_{(\overline{X_1} - \overline{X_2})} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$X_2 \sim N(m_2, \sigma_2)$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

Si $X_1 \sim N(m_1, \sigma_1)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = m_1 - m_2 ;\ \sigma_{(\overline{X_1} - \overline{X_2})} = \sqrt{\frac{\sigma_1^2}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2}\frac{N_2 - n_2}{N_2 - 1}}\right)$$

$X_2 \sim N(m_2, \sigma_2)$

4) Problème de distribution d'échantillonnage d'une proportion

Ce type de problème se pose lorsque nous gérons une question portant sur des proportions (%) à partir d'échantillons.

Une proportion est un nombre relatif c'est-à-dire un rapport de 2 chiffres : o peut dire aussi un taux, un ratio, une fréquence…

On appelle :

- **p la proportion de la population mère ;**
- **f la fréquence de l'échantillon de taille n.**

F représente la variable aléatoire qui gère toutes les valeurs possibles de toutes les fréquences f de tous les échantillons de taille n.

F est l'équivalent pour les taux de $\overline{X}$ pour les moyennes.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

Si $X \sim N \Rightarrow F \sim N\left(E_{(F)} = p ;\ \sigma_F = \sqrt{\frac{pq}{n}}\right)$ avec $q = 1 - p$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X \sim N \Rightarrow F \sim N\left(E_{(F)} = p\,;\ \sigma_F = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

5) Problème de distribution d'échantillonnage d'une différence de proportions

Ce type de problème se pose lorsque nous avons à gérer une question portant sur une comparaison de taux à partir d'échantillons.
On utilise la variable aléatoire $F_1 - F_2$ pour gérer ce type de problème.

a) Si l'échantillonnage est non exhaustif ou si l'échantillonnage est exhaustif avec un taux de sondage < 5%

$$\text{Si } X_1 \sim N \quad X_2 \sim N \Rightarrow F_1 - F_2 \sim N\left(E_{(F_1-F_2)} = p_1 - p_2\,;\ \sigma_{F_1-F_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$$

b) Si l'échantillonnage est exhaustif (tirage sans remise) avec un taux de sondage ≥ 5%

$$\text{Si } X_1 \sim N \quad X_2 \sim N \Rightarrow F_1 - F_2 \sim N\left(E_{(F_1-F_2)} = p_1 - p_2\,;\ \sigma_{F_1-F_2} = \sqrt{\frac{p_1q_1}{n_1}\frac{N_1-n_1}{N_1-1} + \frac{p_2q_2}{n_2}\frac{N_2-n_2}{N_2-1}}\right)$$

III – LES PROBLEMES D'ESTIMATION

- On ne connaît pas tous les paramètres de la population mère et on cherche à les estimer.
Le problème d'estimation gère, en plus, les risques d'élargir les connaissances que l'on a de l'échantillon à l'ensemble de la population mère.

- Il permet d'estimer les paramètres de la population mère :

- **m et σ, pour les moyennes ;**
- **p pour les proportions ;**

à partir des observations de l'échantillon :

- **$\bar{x}$ et σ', pour les moyennes ;**
- **f pour les proportions.**

1) Présentation de toutes les situations possibles

a) Pour les moyennes, on peut rencontrer 4 situations :

- On connaît m et σ : Problème de distribution d'échantillonnage de la moyenne et de l'écart-type.

- On connaît m et on ne connaît pas σ : Problème de distribution d'échantillonnage de la moyenne et un problème d'estimation de l'écart-type.

- On ne connaît pas m et on connaît σ : Problème d'estimation de la moyenne et un problème de distribution d'échantillonnage de l'écart-type.

- On ne connaît pas m et σ : Problème d'estimation de la moyenne et de l'écart-type.

Il faudra donc bien lire l'énoncé et étudier la situation pour voir les informations dont on dispose et en déduire la bonne qualification du problème à résoudre.

b) Pour les taux, on peut rencontrer 2 situations :

- On connaît p donc, on connaît tout (q = 1 - p) : Problème de distribution d'échantillonnage d'une proportion.

- On ne connaît pas p donc, on ne connaît rien : Problème d'estimation d'une proportion.

2) L'estimation ponctuelle

- Lorsque l'on ne connaît pas les paramètres de la population mère, on va d'abord les estimer à partir des informations statistiques obtenues par l'observation de l'échantillon.

Bien sûr, ces statistiques sont loin d'être parfaites et surtout ne peuvent refléter tous les résultats possibles (en effet, chaque échantillon de taille n peut avoir ses propres valeurs).

a) L'estimateur ponctuel de m = $\bar{x}$.

b) L'estimateur ponctuel de p = f.

c) L'estimateur ponctuel de la Variance de la population mère = s^2

- Si le tirage est non exhaustif ou si le tirage est exhaustif avec un taux de sondage < 5% :

- **Calcul de la variance statistique de l'échantillon =** $\sigma'^2 = \frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2$
- **Calcul de s^2 :** $\boxed{s^2 = \frac{n}{n-1}\sigma'^2}$
- **Calcul de s :** $\boxed{s = \sqrt{s^2}}$.

- Si le tirage est exhaustif avec un taux de sondage ≥ 5% :

- **Calcul de la variance statistique de l'échantillon =** $\sigma'^2 = \frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2$
- **Calcul de s^2 :** $\boxed{s^2 = \frac{N-1}{n-1}\frac{n}{N}\sigma'^2}$

▪ **Calcul de s :** $\boxed{s = \sqrt{s^2}}$.

3) L'estimation par intervalle de confiance

a) Estimation de la moyenne de la population mère m

$$\boxed{P(\bar{x} - t\,\sigma_{\bar{x}} < m < \bar{x} + t\,\sigma_{\bar{x}}) = \alpha\,\%}$$

Procédure :

- **Qualifier le problème ;**
- **Prouver que le phénomène X suit une loi Normale** ($X \sim N$) ;
- **Chercher le taux de sondage** $\frac{n}{N}$;
- **Calcul d'estimation des paramètres de la population mère,** en utilisant les estimateurs ponctuels et l'intervalle de confiance : $P(\bar{x} - t\,\sigma_{\bar{x}} < m < \bar{x} + t\,\sigma_{\bar{x}}) = \alpha\,\%$

▪ **Calcul de** $\bar{x} = \frac{\sum n_i x_i}{N}$ avec $N = \sum n_i$

▪ **Trouver t : dans les problèmes d'estimation, 2 situations sont possibles :**

Lorsque les 3 conditions suivantes sont réunies :

- $X \sim N$
- σ **pas connu,**
- **n < 30 : on lit le t dans la table de Student Fisher : en fonction du risque accepté et du nombre de degrés de liberté (n – 1).**

Exemple :
Dans la table de Student de la page 12, si n = 19 et que le risque est de 10% :
On prend la ligne 18 = 19 - 1
la colonne 0,95
et à l'intersection de la ligne 18 et de la colonne 0,95, on lit le t de Student = 1,734.

Dans tous les autres cas de figure : on utilise la table de Laplace-Gauss = la loi Normale.

▪ **Calcul de** $\sigma_{\bar{x}}$ **: 2 cas de figure :**

- **Si** σ **est connu** : pas de problème : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- **Si** σ **n'est pas connu :**
  - on calcule s, estimateur ponctuel de l'écart type de la population mère.
  - on fait $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$.

b) Estimation de la proportion de la population mère p

$$\boxed{P\,(f - t\,\sigma_F < p < f + t\,\sigma_F) = \alpha\,\%}$$

Procédure :

- **Qualifier le problème ;**
- **Prouver que le phénomène X suit une loi Normale** $(X \sim N)$;
- **Chercher le taux de sondage** $\frac{n}{N}$;
- **Calcul d'estimation des paramètres de la population mère** : en utilisant les estimateurs ponctuels et l'intervalle de confiance :

$$P\,(f - t\,\sigma_F < p < f + t\,\sigma_F) = \alpha\,\%$$

- **calcul de** $f = \frac{n_i}{\sum n_i}$
- **trouver t** : idem que pour les moyennes : 2 situations.
- **calcul de** $\sigma_F = \sqrt{\frac{f(1-f)}{n-1}}$.

Corrigés du problème de distribution d'échantillonnage de la moyenne $\overline{X}$
(usine textile)

Lecture de l'énoncé :

X = Longueur des morceaux de tissu en cm

Population mère :
E(X) = m = 90 la long mor ceau de tissu en moyenne
σ = 0,60 l'ecat type
N = 10 000 (1er cas de figue)
N = 2 000 (2ème cas de figure)

Echantillon :
n = 200 echantillon aleatoire
$\overline{x}$ = 90,30 moyenne observé

a) **1ère Question : P ($\overline{X} \leq 89{,}90$) ?**

Procédure pour N = 10 000 :

1) C'est un problème de distribution d'échantillonnage de la moyenne car :
- la question porte sur une moyenne ;
- on travaille sur un échantillon ;
- on connaît les paramètres de la population mère (m = 90, σ = 0,60)

2) X ~ N (90 , 0,60) car :
- n = 200 > 60 $\Rightarrow$ Théorème central limite
- il s'agit d'une production en série et standardisée. On peut appliquer le critère d'atomicité qui est un des critères liés à la loi Normale.

3) Taux de sondage = $\frac{n}{N} = \frac{200}{10\,000} = 0{,}02 < 0{,}05 \Rightarrow$ on ne doit donc pas appliquer le facteur d'exhaustivité.

4) $X \sim N(90 , 0{,}60) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = 90 ; \sigma_{\overline{X}} = \frac{0{,}60}{\sqrt{200}} = 0{,}04243\right)$

$$\Rightarrow \mathbf{P}(\overline{X} \leq 89{,}90) = P\left(T \leq \frac{89{,}90 - 90}{0{,}04243}\right) = P(T \leq -2{,}36)$$
$$= 1 - P(T \leq +2{,}36) = 1 - 0{,}9909$$
$$= \mathbf{0{,}91\ \% = 0{,}0091.}$$

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

Aoyons pb 1

TD 2+3

Procédure pour N = 2 000 :

1) Même réponse que pour N = 10 000.

2) Même réponse que pour N = 10 000.

3) Taux de sondage = $\frac{200}{2\,000} = 0{,}10 > 0{,}05 \Rightarrow$ on va appliquer le facteur d'exhaustivité.

4) $X \sim N(90\ ,\ 0{,}60) \Rightarrow \overline{X} \sim N\left(E(\overline{X}) = 90\ ;\ \sigma_{\overline{X}} = \frac{0{,}60}{\sqrt{200}}\sqrt{\frac{2\,000-200}{2\,000-1}} = 0{,}04026\right)$

$\Rightarrow$ **P** $(\overline{X} \leq 89{,}90) = P\left(T \leq \frac{89{,}90-90}{0{,}04026}\right) = P\ (T \leq -2{,}48) = 1 - P(T \leq +2{,}48) = 1 - 0{,}9934$

= 0,66 % = 0,0066.

Comme cette probabilité est faible, on peut considérer que la machine réalise correctement le travail programmé et n'a pas besoin d'être de nouveau réglée.

b) 2ème Question : Montrer que l'échantillon est représentatif de la population mère pour N = 10 000.

- Prendre un risque de 5% de se tromper $\Rightarrow$ l'intervalle symétrique que l'on veut respecter doit couvrir 95% des situations possibles.

$\Rightarrow P(m - t\ \sigma_{\overline{X}} \leq \overline{X} \leq m + t\ \sigma_{\overline{X}}) = 0{,}95.$

- m = 90
- t de la loi Normale pour un intervalle symétrique de 95% = ± 1,96

(Pour trouver + t :

- rechercher dans la table de la loi N la probabilité la plus proche de 97,5% = 0,975.
- on voit que 0,975 se situe à l'intersection de la ligne 1,9 et de la colonne 0,06

$\Rightarrow$ + t = + 1,96).

- $\sigma_{\overline{X}} = 0{,}04243$ (pour N = 10 000)

$\Rightarrow$ **Intervalle de confiance** = [90 - (1,96)(0,04243) ; 90 + (1,96)(0,04243)]

= [89,917 ; 90,083]

95% des moyennes de morceaux de tissu sont dans cet intervalle de confiance

- Comme la moyenne de l'échantillon $\bar{x}$ = 90,30 se trouve à l'extérieur de l'intervalle de confiance, on peut dire, au risque de 5% de se tromper, que l'échantillon n'est pas représentatif de la population mère.

Dans ce cas, on remet les 200 morceaux de tissu dans la population mère (car ici, les morceaux testés sont encore utilisables) et il faut de nouveau tirer un échantillon aléatoire de 200 morceaux de tissu afin de refaire la procédure.

Corrigés du problème de distribution d'échantillonnage d'une différence de moyennes $\overline{X_1} - \overline{X_2}$ (piles électriques)

Lecture de l'énoncé :

X_1 = durée d'utilisation des piles de la société 1 en heures
X_2 = durée d'utilisation des piles de la société 2 en heures

Populations mères :
$m_1 = 230$ $\sigma_1 = 30$
$m_2 = 210$ $\sigma_2 = 20$

Echantillons :
$n_1 = 100$
$n_2 = 125$

Question : $P(\overline{X_1} - \overline{X_2} \geq 30)$?

Procédure :

1) Problème de distribution d'échantillonnage d'une différence de moyennes car :

— la question porte sur une comparaison de moyennes ;

- on travaille sur échantillons ;
- on connaît tous les paramètres des populations mères (m_1 et m_2 , σ_1 et σ_2).

2) $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$ car :

- $n_1 = 100 > 60$ et $n_2 = 125 > 60 \Rightarrow$ Théorème central limite ;
- les piles sont fabriquées en très grandes séries $\Rightarrow$ critère d'atomicité $\Rightarrow$ loi Normale.

3) Taux de sondage < 5% car :

- même si on ne connaît pas exactement N_1 et N_2 on sait que les tailles des populations mères sont très importantes vue la fabrication en très grande séries ;

- de plus, ici, les piles testées sont perdues à la vente (car elles sont vidées de toute énergie, donc, dans cette situation de produits testés perdus à la vente, on aura toujours un taux de sondage très faible bien inférieur à 5%.

4) comme $X_1 \sim N(230, 30)$ et $X_2 \sim N(210, 20)$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N\left(E_{(\overline{X_1} - \overline{X_2})} = 230 - 210 = 20 \text{ , } \sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{30^2}{100} + \frac{20^2}{125}} = 3{,}493\right)$$

$$\Rightarrow \overline{X_1} - \overline{X_2} \sim N(20 \text{ , } 3{,}493).$$

$$\Rightarrow P(\overline{X_1} - \overline{X_2} \geq 30) = P\left(T \geq \frac{30 - 20}{3{,}493}\right) = P(T \geq 2{,}86) = 1 - P(T \leq 2{,}86) = 1 - 0{,}9979$$

= 0,21% = 0,0021.

La probabilité est donc très faible et on peut considérer que les sociétés 1 et 2 respectent leur cahier des charges et que l'écart de durée de vie moyenne est tendanciellement inférieur à 30h.

Lecture de l'énoncé :

X = Taux de factures non réglées dans les 10 j ouvrables suivant l'échéance.
Population mère :
Pour le 1) : p= 0,12 ; q=1-0,12 = 0, 88
N = plusieurs dizaines de milliers.
Pour le 2) : p= 0,09 ; q=1-0,09=0,91.
Echantillon :
Pour le 1) : n = 500 et f=0,14.
Pour le 2) : n=220 et f=$\frac{25}{220}$=0,1136.

1)a) Déterminer l'intervalle de confiance à 95% et commenter.

Intervalle de confiance à 95% = risque de 5% $\Rightarrow$ P (E(F) – t $\sigma_F \leq F \leq$ E(F) + t σ_F) = 0,95.

- **Problème de distribution d'échantillonnage d'une proportion** car : la question porte sur un taux ; on travaille sur échantillon ; on connaît les paramètres de la population mère (p = 0,12). Connaissant p, on connaît q et on connaît σ.
- **n = 500 > 60 $\Rightarrow$ Théorème central limite $\Rightarrow$ X ~ N.**
- **Taux de sondage < 5%** car la société gère plusieurs dizaines de milliers de factures, donc N est important et $\frac{n}{N}=\frac{500}{N}$ < 0,05, donc pas de facteur d'exhaustivité.

- X ~ N $\Rightarrow$ F ~ N (0,12 , $\sqrt{\frac{0,12\,x\,0,88}{500}}$ = 0,0145)

Intervalle de confiance à 95% = [0,12 – (1,96)(0,0145) , 0,12 + (1,96)(0,0145)]
= **[0,0916 ; 0,1484]**

Le taux de factures non réglées dans les 10 j ouvrables suivant l'échéance va de 9,16% (hypothèse optimiste) à 14,84% (hypothèse pessimiste), en prenant un risque de 5% de se tromper.
Comme f = 0,14 est à l'intérieur de cet intervalle de confiance, l'échantillon est jugé représentatif de a population mère au risque de 5%.
Dans ce cas, on considère que les bases de raisonnement à ce sujet n'ont pas à être actualisées (on remarque toutefois que 0,14 est proche de la borne supérieure : peut être est-ce le signe d'un changement sur les habitudes de règlement des clients).

1)b) Si risque de 3% $\Rightarrow$ intervalle de confiance de 0,97

$\Rightarrow$ P (E(F)-t$\sigma_F \leq$ F $\leq$E(F)+tσ_F)=0,97.
Intervalle de confiance à 97% = [0,12-(2,17)(0,0145) ;0,12+ (2,17)(0,0145)]
= **[0,0885 ; 0,1515]**
$\Rightarrow$ **comme f = 0,14, dans l'intervalle donc l'échantillon est toujours jugé représentatif, au risque de 3%.**

2)a) En prenant un risque de 3%, l'échantillon du benchmark est-il représentatif ?

f = $\frac{25}{220}$ = 0,1136
$\Rightarrow$ P (E(F)-t$\sigma_F \leq$ F $\leq$E(F)+tσ_F)=0,97.

$E(F) = 0,09$; $\sigma_F = \sqrt{\frac{0,09x0,91}{220}} = 0,0193.$

Intervalle de confiance à 97% = [0,09-(2,17)(0,0193);0,09+(2,17)(0,0193)]
= [0,0481;0,1319].
⇒ comme f = 0,1136 est à l'intervalle de confiance, l'échantillon est jugé représentatif au risque de 3%.

2)b) Quelle est la probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark ?
$p_1 = 0,12$ $p_2 = 0,09$
$n_1 = 500$ $n_2 = 220$
Question : P (F_1-$F_2 \leq 0,01$) ?

Procédure :

X_1 = taux de règlement hors délai pour société A.
X_2 = taux de règlement hors délai pour benchmark.
- Problème de distribution d'échantillonnage d'une différence de taux car :
• Nous gérons des taux à partir d'échantillons
• Nous devons comparer les taux de la société A et du benchmark
• Nous connaissons les paramètres des populations mères car nous connaissons p_1 et p_2.
- X_1 et X_2 suivent une loi Normale par application du théorème central limite ($n_1 = 500 > 60$ et $n_2 = 220 > 60$).
- Les taux de sondage pour la société A comme pour le benchmark sont < 5% car ils ont des dizaines de milliers de factures.
- X_1 et $X_2 \sim N \Rightarrow F_1$-$F_2 \sim N (E_{(F1-F2)} = 0,12-0,09 = 0,03$;

$\sigma_{F_1-F_2} = \sqrt{\frac{0,12x0,88}{500} + \frac{0,09x0,91}{220}} = 0,0242).$

P (F_1-$F_2 \leq 0,01$) $= P\left(T \leq \frac{0,01-0,03}{0,0242}\right) = P (T \leq -0,83) = P (T \geq +0,83)$

$= 1-P (T \leq +0,83) = 1-0,7967 =$ **0,2033.**

La probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark est de 20,33%.

Corrigés des problèmes d'estimation de la moyenne de la population mère

1)

Lecture de l'énoncé :

X = Durée de vie du tube cathodique d'une marque de TV, en heures.

Population mère :
m pas connue
$\sigma = 450$

Echantillon :
n = 55
$\bar{x} = 9\,500$

Question :
$P(\bar{x} - t\,\sigma_{\bar{X}} \leq m \leq \bar{x} + t\,\sigma_{\bar{X}}) = 0{,}90$

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N mais σ connu et n = 55 > 30).

⇒ t = 1,645.

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{55}} = 60{,}678.$

⇒ **Intervalle de confiance** = [9 500 – (1,645)(60,678) ; 9 500 + (1,645)(60,678)]
= **[9 400,18 ; 9 599,82].**

La moyenne des durées de vie de la population mère se situe entre 9 400,18 heures et 9 599,82 heures dans 90% des situations possibles (ou en prenant un risque de 10% de se tromper).

On peut dire aussi que :

- **5% des tubes auront une durée de vie < 9 400,18 heures ;**
- **5% des tubes auront une durée de vie > 9 599,82 heures.**

2)

Lecture de l'énoncé :

X = durée de vie des tubes cathodiques d'une marque de TV, en heures.

Population mère :

m pas connue

$\sigma = 450$.

Echantillon :

n = 25

$\bar{x} = 9\,500$.

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$.

Procédure :

1) Problème d'estimation de m et problème de distribution d'échantillonnage de l'écart type car :

- on travaille sur échantillon ;
- la question porte sur une moyenne ;
- on ne connaît pas m (donc il faut l'estimer) et on connaît σ.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

- $\bar{x} = 9\,500$
- On utilise le t de la loi N car les 3 conditions de Student-Fisher ne sont pas respectées (X~ N, n = 25 < 30 mais σ connu = 450).

⇒ t = 2,575.

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{25}} = 90$.

⇒ **Intervalle de confiance** = [9 500 – (2,575)(90) ; 9 500 + (2,575)(90)]

= **[9 268,25 ; 9 731,75].**

Au risque de 1%, m va fluctuer entre ces 2 durées.

3)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

n = 60

$\bar{x}$ = 9 450

σ' = 446,234

Question :

$P(\bar{x} - t\,\sigma_{\bar{X}} \leq m \leq \bar{x} + t\,\sigma_{\bar{X}}) = 0,95.$

Procédure :

1) Problème d'estimation de m et de l'écart type car on ne connaît pas les paramètres de la population mère.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x}$ = 9 450

t de la loi Normale car n = 60 > 30 ⇒ t = 1,96.

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s.

$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{60}{59}(446,234)^2 = 202\,499,779 \Rightarrow s = \sqrt{202\,499,779} = 449,999 \approx 450$

$\Rightarrow \sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{450}{\sqrt{60}} = 58,094$

⇒ **Intervalle de confiance** = [9 450 – (1,96)(58,094) ; 9 450 + (1,96)(58,094)]

= **[9 336,13 ; 9 563,87].**

Au risque de 5%, m se situera entre ces 2 durées.

4)

Lecture de l'énoncé :

Population mère :

m pas connue et σ pas connu.

Echantillon :

n = 25

$\bar{x} = 9\,500$

$\sigma' = 440{,}908$.

Question :

$P(\bar{x} - t\,\sigma_{\bar{x}} \leq m \leq \bar{x} + t\,\sigma_{\bar{x}}) = 0{,}99$

Procédure :

1) Problème d'estimation de m et σ car ces paramètres ne sont pas connus.

2) X ~ N car il s'agit d'une production de masse et standardisée ⇒ critère d'atomicité ⇒ loi Normale.

3) Taux de sondage < 5% car :

- production de masse ⇒ N est importante
- les articles testés sont perdus à la vente.

4)

$\bar{x} = 9\,500$

On utilise le t de Student Fisher car les 3 conditions sont réunies (**X ~ N,** σ pas connu, n = 25 < 30)

⇒ t = 2,797 (lecture de la table de St avec un nombre de ° de liberté = 25-1 = 24 et un risque de 1%).

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ mais ici σ n'est pas connu, donc il faut l'estimer par s ⇒ $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$.

$s^2 = \frac{n}{n-1}\sigma'^2 = \frac{25}{24}(440{,}908)^2 = 202\,499{,}8588$

⇒ $s = \sqrt{202\,499{,}8588} = 449{,}9998 \approx 450$.

⇒ $\sigma_{\bar{x}} = \frac{450}{\sqrt{25}} = 90$.

⇒ **Intervalle de confiance** = [9 500 – (2,797)(90) ; 9 500 + (2,797)(90)]

= **[9 248,27 ; 9 751,73]**.

Au risque de 1%, m se situera entre ces 2 durées.

Corrigés des problèmes d'estimation de la proportion de la population mère

X = Taux de femmes utilisant la marque de lessive A dans une grande ville.

Lecture de l'énoncé :
Population mère :
p pas connue donc on ne connaît pas q, donc on ne connaît pas σ.

Echantillon :
n = 500
f = 0,35

Questions :
a) $P(f - t\,\sigma_F \leq p \leq f + t\,\sigma_F) = 0,95$.
b) Taille minimale de l'échantillon = calcul de n.

a)
Procédure :
1) Problème d'estimation de p car on ne connaît pas le taux de la population mère (c'est donc aussi et automatiquement un problème d'estimation de σ).
2) X ~ N car :
- la lessive est un produit de grande consommation, fabriqué en très grandes séries et donc standardisé $\Rightarrow$ critère d'atomicité.
- n = 500 > 60 $\Rightarrow$ Théorème central limite.

3) Comme il s'agit d'une grande ville, N est supposé important
$\Rightarrow$ **Taux de sondage < 5%** $\Rightarrow$ pas de facteur d'exhaustivité.

4)
- f = 0,35
- t de la loi N car σ pas connu et X ~ N, mais n = 500 > 30 donc nous n'avons pas les 3 conditions d'utilisation d'une loi de St $\Rightarrow$ t = 1,96.
- $\sigma_F = \sqrt{\frac{f(1-f)}{n-1}} = \sqrt{\frac{0,35(1-0,35)}{500-1}} = 0,02135.$

$\Rightarrow$ **Intervalle de confiance** = [0,35 – (1,96)(0,02135) ; 0,35 + (1,96)(0,02135)]
= [0,3082 ; 0,3918].
Au risque 5%, il y a entre 30,82 % et 39,18% des femmes de cette ville qui préfèrent la marque de lessive A.

b) Calcul de n
Pour calculer n il faut connaître l'erreur globale qui a été déterminée par le client, en accord avec le prestataire de l'étude.
Bien souvent le client désire une erreur la plus faible possible mais le prestataire doit rester réaliste et ne pas s'engager sur une erreur qu'il ne pourra pas respecter.
Cette erreur globale fait donc l'objet d'une négociation dès le départ sachant que plus l'erreur globale est faible, plus n est élevée et donc plus il faut interviewer de personnes donc plus le coût de l'étude s'élève.

Ici l'erreur globale $e = 0{,}02 = t\ \sigma_F$.

$$\Rightarrow 0{,}02 = 1{,}96\sqrt{\frac{(0{,}35)(0{,}65)}{n-1}}$$

Pour supprimer la racine carrée, nous allons élever les deux termes de cette égalité au carré et nous allons isoler n afin de trouver sa valeur :

$$\Rightarrow 0{,}02^2 = 1{,}96^2\ \frac{(0{,}35)(0{,}65)}{n-1}$$

$$\Rightarrow n - 1 = 1{,}96^2\ \frac{(0{,}35)(0{,}65)}{0{,}02^2}$$

$$\Rightarrow n = 1{,}96^2\ \frac{(0{,}35)(0{,}65)}{0{,}02^2} + 1$$

$$\Rightarrow \mathbf{n = 2\,185{,}91 \approx 2\,186.}$$

Pour avoir une erreur globale de 2%, il faut interviewer 2 186 dames de cette ville.

Remarque :

Dans la première question, on interviewait 500 dames $\Rightarrow e = 1{,}96\sqrt{\frac{(0{,}35)(0{,}65)}{499}} = 0{,}042.$

$\Rightarrow$ pour $n = 500$, $e = 0{,}042$
pour $n = 2\,186$, $e = 0{,}02$.

On vérifie ainsi une loi énoncée souvent en marketing : « Lorsque l'on veut diviser par 2 l'erreur globale, il faut multiplier par 4 le nombre de personnes à interviewer ».
Bien sûr, cette loi ne nous donne qu'une approximation de n compte tenu du choix de l'erreur globale.
Le seul moyen de connaître exactement n est de le calculer comme indiqué.

APPLICATIONS DU COURS

I - PROBLEMES DE DISTRIBUTION D'ECHANTILLONNAGE

1) Moyenne $\overline{X}$

Dans une usine textile, on utilise une machine automatique pour couper des morceaux de tissu.
Pour une série de fabrication, lorsque la machine est correctement ajustée, la longueur des morceaux de tissu doit être en moyenne de 90 cm avec un écart-type de 0,60 cm.
Pour contrôler la longueur des morceaux de tissu, on tire dans la production d'une journée, un échantillon aléatoire de 200 morceaux de tissu.

a) Calculer la probabilité que la moyenne de l'échantillon soit au plus égale à 89,90 cm, ceci dans 2 situations :

- production de la journée = 10 000 morceaux ;
- production de la journée = 2 000 morceaux.

b) Si la moyenne observée sur cet échantillon est de 90,30 cm, celui-ci est-il représentatif de la population mère, en prenant un risque de 5% de se tromper (avec N = 10 000).

2) Différence de moyennes, $\overline{X_1} - \overline{X_2}$

2 sociétés fabriquent des piles électriques d'un certain format.
Les piles de la société 1 ont une durée d'utilisation moyenne de 230 heures avec un écart-type de 30 heures.
Les piles de la société 2 ont une durée d'utilisation moyenne de 210 heures avec un écart-type de 20 heures.
Quelle est la probabilité que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 100 piles de la société 1 soit d'au moins 30 heures de plus que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 125 piles de la société 2 ?

3) Proportion p

Le directeur financier d'une société sait par expérience que 12% des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance.
Le chiffre d'affaires s'étant accru sensiblement, le directeur financier veut vérifier si la situation va évoluer.
Il fait prélever un échantillon aléatoire de 500 factures, à partir duquel il constate que 14% des factures ne sont pas réglées dans les délais.
Déterminer l'intervalle de confiance à 95% et commenter ce résultat, sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

TD4

1) Le directeur financier d'une société A sait par expérience que 12% des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance.
Le chiffre d'affaires s'étant accru sensiblement, le directeur financier veut vérifier si la situation va évoluer.
Il fait prélever un échantillon aléatoire de 500 factures, à partir duquel il constate que 14% des factures ne sont pas réglées dans les délais.

a) Déterminer l'intervalle de confiance à 95% et commenter ce résultat, sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

b) Même question mais avec un risque de 3%.

2) Le benchmark (la référence) dans ce secteur d'activité, a un taux de non règlement dans les 10 jours ouvrables suivant l'échéance de 9%.
Lors d'un travail de réactualisation de ses données, le responsable financier a fait tirer et examiner 220 factures au hasard parmi des dizaines de milliers.
Après analyse, on a constaté que 25 de celles-ci n'étaient pas réglées dans le délai de 10 jours ouvrables suivant l'échéance.

a) En prenant un risque de 3%, l'échantillon du benchmark est-il représentatif ?

b) Quelle est la probabilité que le taux de non règlement de la société A soit au plus de 1% supérieur à celui du benchmark ?

II - PROBLEMES D'ESTIMATION

1) Moyenne

a) Soit X la variable aléatoire « durée de vie d'un tube cathodique d'une marque de TV.

On ne connaît pas la moyenne des durée de vie des tubes, par contre, on connaît l'écart-type de la distribution des durées de vie de la population mère : $\sigma = 450$ heures.

Pour un échantillon de 55 tubes prélevés au hasard, on a calculé que la durée de vie moyenne était de 9 500 heures.

Estimer la durée de vie moyenne de la population des tubes fabriqués par un intervalle de confiance à 90%.

b) Reprenons le même exemple, mais cette fois, l'échantillon est de taille $n = 25$ (les autres éléments ne changeant pas).

Estimer la durée de vie moyenne de la population des tubes fabriqués par un intervalle de confiance à 99%.

c) Cette fois, l'écart-type de la population mère n'est plus connu.

A partir d'un échantillon de taille $n = 60$, nous avons $\bar{x} = 9\,450$ heures et $\sigma' = 446{,}234$ heures.

Estimer à l'aide d'un intervalle de confiance à 95% la moyenne des durées de vie de la population mère.

d) L'écart-type de la population mère n'étant pas connu, on prélève un échantillon aléatoire de taille $n = 25$.

La moyenne des durées de vie de cet échantillon est égale à 9 500 heures et l'écart-type est de 440, 908 heures.

Estimer à l'aide d'un intervalle de confiance à 99% la durée de vie moyenne des tubes fabriqués.

2) Proportion

a) Les responsables d'une étude de marché portant sur le produit lessive essayent de connaître les conditions de pénétration de l'agglomération toulonnaise pour ce qui est de la cible des femmes de 35 à 60 ans.

Ils interviewent au hasard 500 dames de ce cœur de cible et constatent que 175 d'entre elles préfèrent utiliser la marque de lessive A.

Déterminer l'intervalle de confiance à 95% de la proportion des femmes de 35 à 60 ans de cette agglomération préférant la marque de lessive A.

b) Supposons qu'avant de tirer l'échantillon, les responsables de l'étude aient décidé d'estimer la proportion à $\pm$ 2% près.

Quelle devrait être dans ce cas la taille minimale de l'échantillon, en désirant toujours avoir un intervalle de confiance à 95% et en considérant toujours que $f = 0{,}35$.

Inférence statistique

Politique de rémunération d'une grande société

Une grande société fait le recensement de tous les salaires mensuels bruts, versés aux employés de ses différentes succursales, implantées en métropole.

Le nombre total des salariés recensés est de 3 521.

La distribution des salaires mensuels bruts, dans cette société, est la suivante :

Montant des salaires mensuels bruts (€)	1 762	2 357	2 863	3 695	4 345	5 684
Probabilité	0,15	0,23	0,35	0,22	0,04	0,01

Remarque :
Les probabilités sont établies à partir des fréquences observées sur l'ensemble de la population mère.

La direction des ressources humaines vous demande si elle peut choisir la succursale implantée à Tours, comme échantillon représentatif, pour ce qui concerne la politique salariale de la société au niveau national, en prenant un risque de 13%, avec un risque de 4% pour les montants les plus élevés et un risque de 9% pour les montants les plus faibles.

A ce titre, elle vous fournit les statistiques des salaires mensuels bruts pour l'ensemble des salariés de la succursale de Tours :

Montant des salaires mensuels bruts (€)	1 643	2 156	2 747	3 841	4 369
Effectif concerné	19	36	42	17	2

Rappel de formules :

Statistique descriptive :

Moyenne = $\bar{x} = \frac{\sum n_i x_i}{\sum n_i}$

Variance = $\sigma^2 = \frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2$

Ecart-type = $\sigma = \sqrt{\sigma^2}$

Probabilités :

Moyenne = m = $E(X) = \sum p_i X_i$

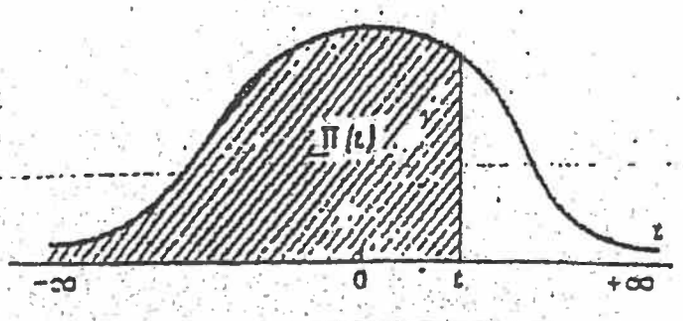
Variance = $\sigma^2 = \sum p_i X_i^2 - E^2(X)$

Ecart-type = $\sigma = \sqrt{\sigma^2}$

TABLE DE LA FONCTION INTÉGRALE DE LA LOI DE LAPLACE-GAUSS $\mathcal{N}(0,1)$

Probabilité d'une valeur inférieure à t :

$$\Pr\{T < t\} = \Pi(t)$$



t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,608 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 0	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,8485	0,850 8	0,853 1	0,855 4	0,8577	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 8	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 9	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,985 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6

TABLE POUR LES GRANDES VALEURS DE t

t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$\Pi(t)$	0,998 65	0,999 04	0,999 31	0,999 52	0,999 66	0,999 76	0,999 841	0,999 928	0,999 968	0,999 997

NOTA. — La table donne les valeurs de $\Pi(t)$ pour t positif. Lorsque t est négatif il faut prendre le complément à l'unité de la valeur lue dans la table.

Exemple : pour $t = 1{,}37$ $\quad \Pi(t) = 0{,}914\,7$;
pour $t = -1{,}37$ $\quad \Pi(t) = 0{,}085\,0$.

11

LOI NORMALE – LOI de LAPLACE-GAUSS

1) Conditions d'application pure
   - il doit s'agir de **variables continues** ;
   - la variable doit **dépendre d'un grand nombre de causes indépendantes, dont les effets s'additionnent et dont aucune n'est prépondérante** ;
   - il faut un **grand nombre d'observations**.

Remarque :
Dans la vie des affaires, ces conditions sont rarement réunies.
La loi Normale est le plus souvent utilisée par approximation (voir les conditions sur les fiches) **ou par application du théorème central limite** (voir le point 4 de cette fiche).

2) **Si** $X \sim N(E(X), \sigma X)$ (lire si X suit une loi Normale de paramètres E(X) et σX), **on passe par une variable centrée réduite** $\boxed{t = \frac{X - E(X)}{\sigma X}}$ **avec E(T) = 0 et** $\sigma T = 1$ (c'est pour cela que Gauss et Laplace ont pu faire une table).
La table de la loi Normale donne les valeurs de la fonction de répartition $\pi(t) = P(T \leq t)$.
Propriété : **il y a une parfaite symétrie de X par rapport à E(X) = mode = médiane et de T par rapport à E(T) = 0.**

3)
Si $X_1 \sim N(E(X_1), \sigma X_1)$
$X_2 \sim N(E(X_2), \sigma X_2)$ } $\Rightarrow X_1 + X_2 \sim N(E(X_1)+E(X_2), \sqrt{\sigma^2_{X_1} + \sigma^2_{X2}})$
$\Rightarrow X_1 - X_2 \sim N(E(X_1)-E(X_2), \sqrt{\sigma^2_{X_1} + \sigma^2_{X_2}})$

avec X_1 **et** X_2 **indépendantes.**

Si $X \sim N(E(X), \sigma X) \Rightarrow aX + b \sim N(aE(X) + b, \sqrt{a^2\sigma^2_X})$

4) Théorème central limite

a) **Si** $X_i \sim N(E(X_i), \sigma X_i)$

$Y = \sum X_i$ } $\Rightarrow Y \sim N(E(Y), \sigma Y)$ **avec** $E(Y) = \sum E(X_i)$, $\sigma Y = \sqrt{\sum \sigma^2_{Xi}}$

Ceci représente le fait qu'un groupe (Y) de variables (X) suivant toutes une loi Normale, suit lui-même une loi Normale.

b) **Si les variables** X_i **ne suivent pas toutes une loi Normale, Y suivra quand même une loi Normale, si le nombre d'observations n est suffisant (50 à 60 au minimum).**

loi de Student

Table 6 : Quantiles $t_{\nu;p}$ de la variable t_ν

$$F(t_{\nu;p}) = P(t_\nu \leqslant t_{\nu;p}) = p$$

ν	p												
	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	0.289	0.445	0.617	0.817	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.778	3.747	4.604	7.173	8.610
5	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.500	4.785	5.408
8	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.897	3.355	4.501	5.041
9	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.813	2.228	2.764	3.169	4.144	4.587
11	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.394	0.538	0.694	0.870	1.080	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.625	2.977	3.787	4.141
15	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.603	2.947	3.733	4.073
16	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.584	2.921	3.686	4.015
17	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.611	3.922
19	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.540	2.861	3.579	3.883
20	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.390	0.532	0.685	0.858	1.060	1.320	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.389	0.530	0.682	0.853	1.054	1.310	1.696	2.040	2.453	2.744	3.375	3.634
32	0.255	0.389	0.530	0.682	0.853	1.054	1.309	1.694	2.037	2.449	2.739	3.365	3.622
33	0.255	0.389	0.530	0.682	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.389	0.529	0.682	0.852	1.053	1.307	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.389	0.529	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.340	3.591
36	0.255	0.388	0.529	0.681	0.852	1.052	1.306	1.688	2.028	2.435	2.720	3.333	3.582
37	0.255	0.388	0.529	0.681	0.851	1.051	1.305	1.687	2.026	2.431	2.715	3.326	3.574
38	0.255	0.388	0.529	0.681	0.851	1.051	1.304	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.388	0.529	0.681	0.851	1.050	1.304	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.705	3.307	3.551
41	0.255	0.388	0.529	0.681	0.850	1.050	1.303	1.683	2.020	2.421	2.701	3.301	3.544
42	0.255	0.388	0.528	0.680	0.850	1.049	1.302	1.682	2.018	2.419	2.698	3.296	3.538
43	0.255	0.386	0.528	0.680	0.850	1.049	1.302	1.681	2.017	2.416	2.695	3.291	3.532
44	0.255	0.388	0.528	0.680	0.850	1.049	1.301	1.680	2.015	2.414	2.692	3.286	3.526
45	0.255	0.388	0.528	0.680	0.850	1.049	1.301	1.679	2.014	2.412	2.690	3.282	3.520
46	0.255	0.388	0.528	0.680	0.850	1.048	1.300	1.679	2.013	2.410	2.687	3.277	3.515
47	0.255	0.388	0.528	0.680	0.849	1.048	1.300	1.678	2.012	2.408	2.685	3.273	3.510
48	0.255	0.388	0.528	0.680	0.849	1.048	1.299	1.677	2.011	2.407	2.682	3.269	3.505
49	0.255	0.388	0.528	0.680	0.849	1.048	1.299	1.677	2.010	2.405	2.680	3.265	3.500

Synthèse de cours concernant la gestion d'un intervalle symétrique

---

1) La notion de risque est associée à la notion d'intervalle de confiance.
Un risque de se tromper de 5% veut dire que l'on envisage, dans les calculs d'ignorer 5% des situations possibles ⇒ cela veut dire aussi que l'on désire prendre en compte, dans l'intervalle de confiance, 95% des situations possibles.
Si le risque est de 1%, l'intervalle de confiance couvre 99% des situations possibles.
Si le risque est de 30%, l'intervalle de confiance couvre 70% des situations possibles et ainsi de suite....
Ce sont donc des notions tout à fait complémentaires.

Dans la pratique, c'est le responsable de l'étude qui choisit le risque, en fonction de :
- l'expérience que l'on a sur le phénomène étudié ;
- des caractéristiques propres au contexte étudié (sur le marché de la téléphonie mobile, le risque est par nature beaucoup plus élevé que sur le marché des lessives) ;
- de l'aversion au risque du décideur.

Fondamentalement, il existe deux types d'intervalle de confiance :
- **les intervalles symétriques** : les bornes de l'intervalle sont équidistantes de l'axe de symétrie (E(X) et E(T) pour la loi Normale) et le risque est équitablement réparti de part et d'autre de l'intervalle (50-50) ;
- **les intervalles non symétriques** : les bornes de l'intervalle ne sont pas équidistantes de l'axe de symétrie et le risque est réparti différemment selon que les valeurs minimales ou maximales soient jugées plus ou moins risquées.

Dans la pratique, c'est le responsable de l'étude qui choisit le type d'intervalle de confiance et qui donne la répartition du risque à appliquer.

Schéma représentant un risque de 5% avec intervalle de confiance symétrique :

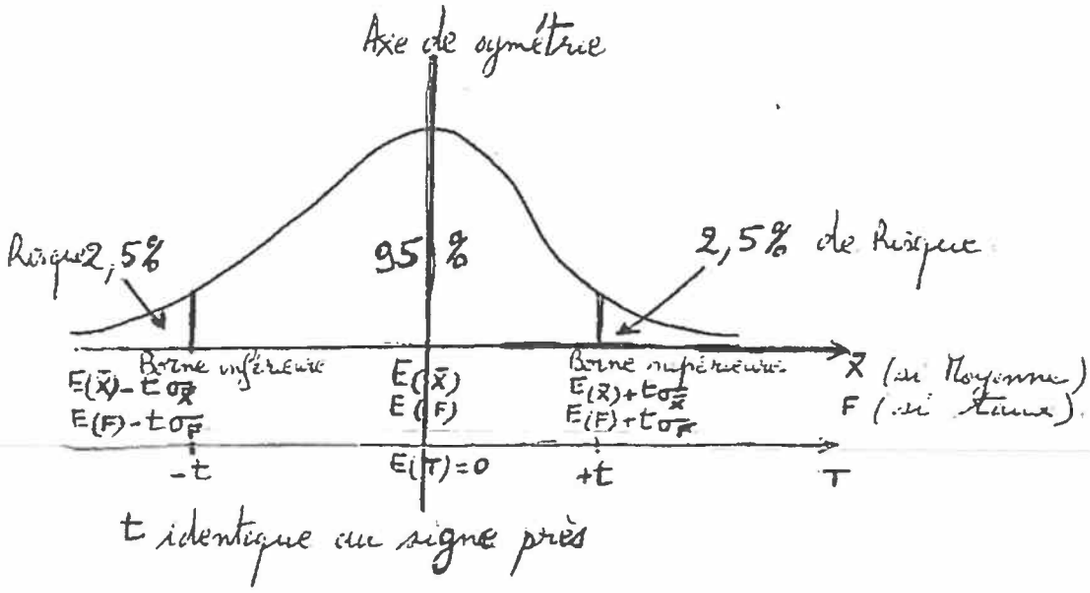
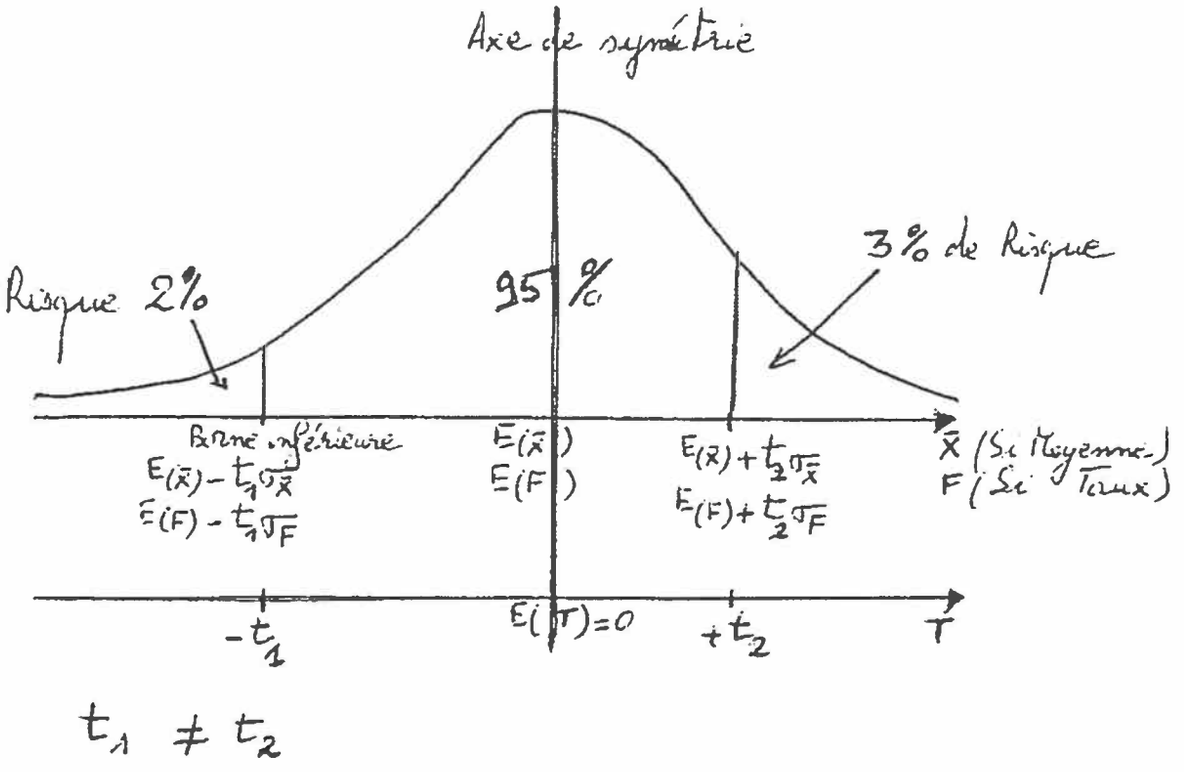


Schéma représentant un risque de 5% avec intervalle non symétrique, en considérant la répartition du risque suivante : 3% du risque sur les valeurs maximales et 2% du risque sur les valeurs minimales :



2) Procédure à appliquer lorsque l'on désire vérifier si un échantillon est représentatif de la population mère.

a) Calcul des bornes de l'intervalle de confiance compte tenu du risque choisi et du type d'intervalle de confiance ;

b) Calcul de la moyenne de l'échantillon $\bar{x}$ (ou du taux de l'échantillon f, si le problème porte sur un taux) ;

c) 2 cas de figure :

▪ **si $\bar{x}$ se trouve dans l'intervalle de confiance** (entre la borne inférieure et la borne supérieure), **on conclut que l'échantillon est représentatif de la population mère au risque qui a été choisi.**
Dans ce cas, on continue le travail entrepris ;

▪ **si $\bar{x}$ se trouve à l'extérieur de l'intervalle de confiance,** on conclut que **l'échantillon n'est pas représentatif de la population mère au risque choisi.**
Dans ce cas, on ne continue pas le travail et on retire un échantillon aléatoire.