Chapitre 14 : Echantillonage et estimation

0-RAPPELS

Voir livre Transmath TS page 435

.

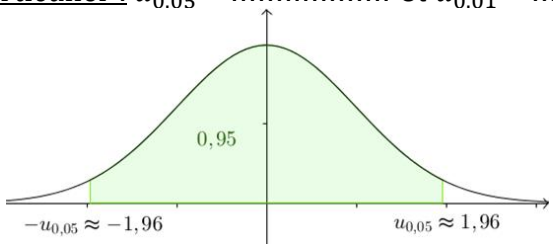
1-COMPLÉMENTS SUR LA LOI NORMALE

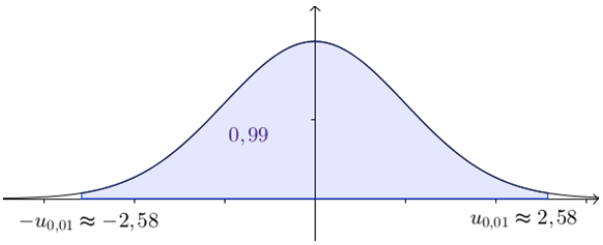
11. Intervalles centrés sur l'espérance

Théorème: Soit X de loi $\mathcal{N}(\mu, \sigma^2)$,
pour tout nombre réel α de]0;1[, $P(X \in [u - u_\alpha\sigma; u + u_\alpha\sigma]) = 1 - \alpha$
où u_α est l'unique nombre >0 tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$

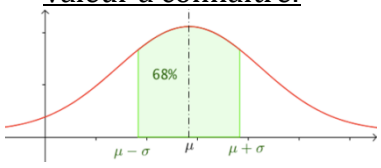
Démonstration:
Soit X de loi $\mathcal{N}(\mu, \sigma^2)$ alors $Y = \frac{X-\mu}{\sigma}$ de loi $\mathcal{N}(0,1)$.
Soit α de]0;1[.
il existe un unique réel positif u_α tel que que $P(-u_\alpha \leq Y \leq u_\alpha) = 1 - \alpha$ (Propriété de la loi normale)
⇔
⇔
⇔
-- CQDF –

Cas particulier : $u_{0.05}$ ≅..... et $u_{0.01}$ ≅.....

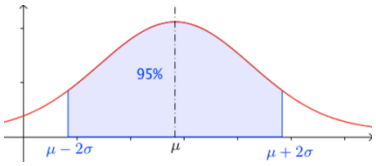




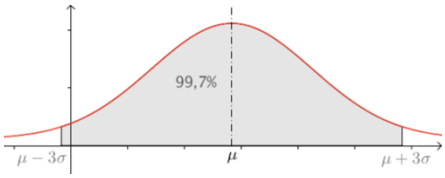
Valeur à connaitre:



$P(X \in [u - \sigma; u + \sigma])$ =.....



$P(X \in [u - 2\sigma; u + 2\sigma])$ =.....



$P(X \in [u - 3\sigma; u + 3\sigma])$ =.....

12. Loi de la fréquence de succès F_n

Définition : X_n est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n, p)$
La variable aléatoire $F_n = \frac{X_n}{n}$ s'appelle la variable aléatoire fréquence de succès pour un schéma de Bernoulli de paramètres n et

Loi de F_n :
pour tout n et k tels que $0 \leqslant k \leqslant n$, $P\left(F_n = \frac{k}{n}\right) = P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Théorème: Soient Xn est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n, p) et\ F_n$ la fréquence de succès, alors :

- L'**espérance** de F_n est $\mathrm{E}(F_n) = p$
- La **variance** de F_n est $\mathrm{V}(F_n) = \frac{p(1-p)}{n}$
- **L'écart type** F_n est $\sigma = \sqrt{V(F_n)} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Démonstration :

$\mathrm{E}(F_n) =$.....

$\mathrm{V}(F_n) =$.....

$\sigma = \sqrt{V(F_n)} = \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$

--CQFD--

13. Approximation de la loi de F_n par une loi normale

Règle: Soient Xn est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n,p) et\ F_n$ la fréquence de succès, n≥30, np≥5 n(1-p)≥5, la loi de la fréquence de succès F_n peut être approchée par un loi normale $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$,

Démonstration :

-- CQFD--

Exemple :

On effectue 100 lancers d'une pièce équilibrée, on appelle succèsl'apparition de Pile. F_{100} indique la fréquence de succès dans un schéma de Bernoulli d'ordre n=100 et de paramètre p =0,5. Approximer E(F_{100}) et V(F_{100}).

Verifions les trois conditions:

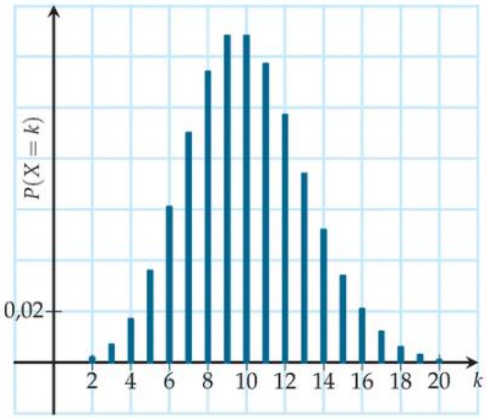
D'apres la règle ci-dessus, F_{100} peut être approximé par .....;

Activité

Une entreprise fabrique des vis en acier. Elle affirme que 5 % des vis qu'elle produit ont un défaut à la fin de la chaîne de production.
La responsable du **contrôle qualité** prélève un échantillon de n vis (au vu du grand nombre de vis produites, on assimile ce prélèvement à un tirage avec remise) pour effectuer une analyse.

Partie A : Avec $n = 200$

La contrôleuse prélève un échantillon de 200 vis au hasard.
On note X la variable aléatoire donnant le nombre de vis avec défaut qui suit donc une loi binomiale de paramètres $n = 200$ et $p = 0{,}05$ sous l'hypothèse de l'énoncé, et $F = \dfrac{X}{200}$ la variable aléatoire donnant la fréquence des vis défectueuses dans l'échantillon.
On donne les éléments ci-dessous obtenus avec un tableur : un graphique représentant en partie cette loi et un tableau de valeurs (les probabilités sont arrondies à 10^{-3}).



k	$P(X \leqslant k)$
2	0,002
3	0,009
4	0,026
5	0,062
6	0,124
7	0,213
8	0,327
9	0,455
10	0,583

k	$P(X \leqslant k)$
11	0,700
12	0,796
13	0,870
14	0,922
15	0,956
16	0,976
17	0,988
18	0,994
19	0,997

1) Peut-on dire que la probabilité qu'elle relève exactement 3 vis avec défaut est élevée ?

2) La probabilité qu'elle relève entre 6 et 10 vis avec défaut est-elle supérieure à 0,6 ?

3) a) Trouver la plus petite valeur de l'entier b tel que $P(4 \leqslant X \leqslant b) \geqslant 0{,}95$.

b) En déduire un intervalle $[f_1 \,;\, f_2]$ tel que $P(f_1 \leqslant F \leqslant f_2) \geqslant 0{,}95$.

On dit que l'intervalle $[0{,}02 \,;\, 0{,}08]$ est un intervalle de fluctuation au seuil de 95 % de la variable aléatoire F donnant la fréquence : cela signifie qu'il y a au moins 95 % de chance que la fréquence des vis défectueuses (dans cet échantillon de 200 vis) soit dans cet intervalle.

4) a) Proposer un autre intervalle $[c \,;\, d]$ (avec c et d entiers) tel que $P(c \leqslant X \leqslant d) \geqslant 0{,}95$.

b) En déduire un autre intervalle de fluctuation au seuil de 95 % de la variable aléatoire F.

5)

En Première, pour déterminer un tel intervalle, on appliquait la méthode suivante :

- on cherche le plus petit entier a tel que $P(X \leqslant a) > 0{,}025$;
- on cherche le plus petit entier b tel que $P(X \leqslant b) \geqslant 0{,}975$;
- on calcule l'intervalle de fluctuation au seuil de 95 % donné par $\left[\dfrac{a}{n} \,;\, \dfrac{b}{n}\right]$ où n correspond au paramètre n de la loi binomiale utilisée.

Lequel des deux intervalles trouvés aux questions **3b** et **4b** obtient-on avec cette méthode ?

Partie B : Avec n quelconque et des résultats de Terminale

La contrôleuse prélève un échantillon de n vis au hasard et on note :

- X_n la variable aléatoire donnant le nombre de vis défectueuses, qui suit donc une loi binomiale de paramètres n et $p = 0{,}05$;
- $F_n = \dfrac{X_n}{n}$ la variable aléatoire donnant la fréquence des vis défectueuses dans l'échantillon ;
- $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ la variable aléatoire centrée réduite associée à X_n.

1) En utilisant le théorème de Moivre-Laplace, que peut-on dire de $P(u \leqslant Z_n \leqslant v)$ quand n tend vers $+\infty$?

2) Que peut-on en déduire pour $P(-1{,}96 \leqslant Z_n \leqslant 1{,}96)$ quand n tend vers $+\infty$?

3) En déduire un intervalle $[f_1 \,;\, f_2]$ dépendant de p et n tel que $P(f_1 \leqslant F_n \leqslant f_2) \approx 0{,}95$ quand n tend vers $+\infty$ (on admettra que $P(f_1 \leqslant F_n \leqslant f_2) \geqslant 0{,}95$ quand n tend vers $+\infty$).

Cet intervalle obtenu grâce à une limite est dit **intervalle de fluctuation asymptotique** de F_n au seuil de 95 %.

Lorsque l'on réalise n tirages avec remise (ou n tirages assimilables à des tirages avec remise, comme c'est le cas ici), cet intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence des succès est donné par $\left[p - 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \,;\, p + 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ où p est la probabilité d'un succès.

.....

.....

.....

.....

.....

.....

.....

Partie C : Comparaison des intervalles obtenus

1) a) Calculer l'intervalle de fluctuation asymptotique au seuil de 95 % dans le cas $n = 200$ (arrondir à 10^{-3}).

b) Comparer cet intervalle avec celui obtenu par la méthode de Première de la partie **A** .

2) On donne le tableau ci-dessous regroupant des intervalles de fluctuation au seuil de 95 % obtenus avec la méthode de Première pour $p = 0{,}05$ et différentes valeurs de n :

Pour $n =$	30	500	1000
Intervalle de fluctuation	[0 ; 0,133]	[0,032 ; 0,07]	[0,037 ; 0,064]
Intervalle de fluctuation asymptotique			

a) Recopier et compléter le tableau en calculant les intervalles de fluctuation asymptotiques au seuil de 95 % correspondant à chaque valeur de n.

b) Que peut-on dire des intervalles présents dans chaque colonne quand n augmente ?

> Cet intervalle de fluctuation asymptotique est plus facile à déterminer que l'intervalle de fluctuation obtenu avec la méthode de Première. On estime qu'il en donne une approximation satisfaisante lorsque $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Partie D : Prise de décision

La contrôleuse a finalement choisi de prélever 400 vis et 26 d'entre elles ont un défaut.
Elle demandera un nouveau réglage des machines si la fréquence observée n'est pas dans l'intervalle de fluctuation asymptotique au seuil de 95 %.
Que va t-elle décider ?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

II/ ECHANTILLONAGE ET PRISE DE DÉCISION

21. Intervalle de fluctuation asymptotique au seuil 1-α

Dans ce paragraphe, on suppose que la proportion *p* du caractère étudié est connue.

Définition : Soit X_n une variable aléatoire qui suit la loi binomiale B (n;p) et α un nombre réel de l'intervalle]0;1[et et b des nombres réels.
[a;b] est **un intervalle de fluctuation de X_n au seuil 1 - α ⇔ P(a ≤ Xn ≤ b) ≥ 1 - α**

Propriété : Si la variable aléatoire X_n suit la loi binomiale B(n;p) avec p dans l'intervalle]0;1[alors pour tout nombre réel α de]0;1[, $\lim\limits_{n\to+\infty} P(\frac{X_n}{n} \in I_n) = 1-\alpha$
où I_n désigne l'intervalle $[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$
u_α est le réel tel que P(- u_α ≤ Z ≤ u_α) = 1- α lorsque de loi normale centrée réduite N (0;1)

Démonstration (Exigible au Bac) :

X_n suit la loi binomiale $B(n;p)$ donc la suite de variables aléatoires $Z_n = \dfrac{X_n - E(X_n)}{\sigma(X_n)}$ suit une loi normale centrée réduite $N(0;1)$ et d'après le théorème de Moivre-Laplace, on a :

$\lim\limits_{n\to+\infty} P(a \le Z_n \le b) = \int_a^b \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, pout tous réels a et b avec $a < b$.

Or $Z_n = \dfrac{X_n - E(X_n)}{\sigma(X_n)} = \dfrac{X_n - np}{\sqrt{np(1-p)}} = \dfrac{n\left(\frac{X_n}{n} - p\right)}{n\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \dfrac{F_n - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}$.

Donc $\lim\limits_{n\to+\infty} P\left(p + a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \le F_n \le p + b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \int_a^b \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

Comme, pour tout réel $\alpha \in]0;1[$, il existe un unique réel positif u_α tel que $P(-u_\alpha \le X \le u_\alpha) = 1-\alpha$ où X suit une loi normale centrée réduite $N(0;1)$, on a :

$\int_{-u_\alpha}^{u_\alpha} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1-\alpha$.

En prenant $a = -u_\alpha$ et $b = u_\alpha$, on a : $\lim\limits_{n\to+\infty} P\left(p - u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \le F_n \le p + u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha$

.—CQFD--

En conséquence:

Définition : Soit X_n une variable aléatoire qui suit la loi binomiale B (n;p) avec p dans lintervalle]0;1[et α un nombre réel de l'intervalle]0;1[.
L'intervalle $I_{n=}$ $[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ est appelé **un intervalle de fluctuation asymptotique au seuil 1-α** de la varible aléatoire fréquence $F_n = \frac{X_n}{n}$ qui, à tout échantillon de taille n associe la fréquence obtenue f.

Valeurs particulières :
On obtient les intervalle de fluctuation asymptotique suivant :

- $I_n = [p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ au seuil de 95% (car $u_{0,05} \approx 1{,}96$)

- $I_n = [p - 2{,}58\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 2{,}58\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ au seuil de 99% (car $u_{0,01} \approx 2{,}58$)

Remarque : En pratique ces deux intervalles permettent des prises de décisions au seuil de 95% ou de 99% sous les conditions suivantes : n ≥ 30 ; np ≥ 5 et n(1-p) ≥ 5

Application :
On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne est ***p* = 0,3**.
On tire successivement avec remise *n* = 50 boules.
Soit X_{50} la variable aléatoire dénombrant le nombre de boules blanches tirées.
X_{50} suit **la loi binomiale** $\mathcal{B}(50,\ 0{,}3)$.
Déterminer l'intervalle de fluctuation asymptotique au seuil de 95%. Interpréter ce résultat.
Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% pour n = 500 tirages.
.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2.2 Regle de prise de décision

Dans ce paragraphe, la proportion du caractère étudié n'est pas connue mais est supposée être égale à *p*. La prise de décision consiste à valider ou invalider l'hypothèse faite sur la proportion *p*.

Propriété : Règle de décision :

On considère une population dans laquelle on souhaite savoir si la proportion d'individus vérifiant une certaine propriété est p : c'est l'hypothèse à tester.
Pour cela, on détermine d'abord sous cette hypothèse un intervalle de fluctuation asymptotique I (à un certain seuil) de la fréquence du caractère dans un échantillon de taille n prélevé dans la population (en admettant que ce prélèvement est assimilable à des tirages avec remise). Puis on observe effectivement cette fréquence f dans un échantillon donné et :

- si $f \notin I$ alors on rejette l'hypothèse que la proportion est p au seuil considéré ;
- si $f \in I$ alors on ne rejette pas l'hypothèse que la proportion est p au seuil considéré.

Autrement dit: rejeter un hypothèse à un seuil défini⇔le risque d'érreyur (rejetter à tort) est d'environ 5%

Application :

Le pourcentage de personnes du groupe sanguin O dans la population française est de 43 %.
On souhaite déterminer si l'on peut faire la même hypothèse pour d'autres populations, en étudiant des échantillons de 250 personnes dans ces populations (dont on suppose qu'elles sont suffisamment grandes pour assimiler ces prélèvements d'échantillons à des tirages avec remise).
On note X la variable aléatoire donnant le nombre de personnes du groupe O dans un échantillon de 250 personnes issu d'une population dont 43 % des individus sont du groupe O.

1) a) Quelle loi suit X ?
   b) Déterminer l'intervalle de fluctuation asymptotique de la fréquence des individus du groupe O au seuil de 95 % dans un tel échantillon. Arrondir à 10^{-3} près.
2) a) On observe pour un échantillon de la population canadienne une proportion de 47 % d'individus du groupe O.
   Peut-on rejeter l'hypothèse que 43 % des Canadiens sont du groupe O ?
   b) On observe pour un échantillon de Basques : 138 individus du groupe O parmi les 250 personnes de l'échantillon.
   Peut-on rejeter l'hypothèse que 43 % des Basques sont du groupe O ?

III/ ESTIMATION PAR INTERVALLE DE CONFIANCE

1. Estimation

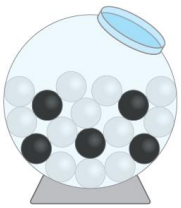
Précedemment, on a vu comment, à partir de données sur la population, on pouvait déduire des informations sur les échantillons extraits de cette population.

Mais, dans la plupart des cas, on est confronté au problème inverse : on ne peut pas (par manque de temps, de moyens, par impossibilité matérielle) mesurer une variable sur la population entière. On prélève alors des échantillons sur lesquels on fait des mesures et des calculs, et on veut en déduire des informations sur la population entière.

On va donc chercher à estimer des paramètres inconnus de la population.

Estimation : On estime la proportion p par un intervalle de confiance déterminé à partie de f et de n (taille des échantillons) selon un niveau de confiance 1 – α

Echantillon		**Population**
On calcule la fréquence f des individus ayant ce caractère	*Estimation* → ← *Echantillonnage*	*Proportion inconnue p d'individus ayant un certain caractère*

Echantillonnage – Prise de décision		**Estimation**
- Une urne contient un très grand nombre de boules blanches et de boules noires dont **on connaît la proportion** *p* de boules blanches. On tire avec remise *n* boules (échantillon) et on observe la fréquence d'apparition des boules blanches. Cette fréquence observée appartient à un intervalle, appelé **intervalle de fluctuation** de centre *p*. - Dans le cas où on ne connaît pas la proportion *p* mais on est capable de faire une hypothèse sur sa valeur, on parle de **prise de décision**. On veut par exemple savoir si un dé est bien équilibré. On peut faire l'hypothèse que l'apparition de chaque face est égale à 1/6 et on va tester cette hypothèse à l'aide d'une expérience (d'acception/rejet de l'hypothèse de départ).		Une urne contient un très grand nombre de boules blanches et de boules noires dont **on ignore la proportion** *p* de boules blanches. On tire avec remise *n* boules dans le but d'estimer la proportion *p* de boules blanches. On obtient ainsi une fréquence d'apparition qui va nous permettre d'estimer la proportion p à l'aide d'un **intervalle de confiance**. 

2. Intervalle de confiance

Propriété :

X_n est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n,p)$.

$F_n = \frac{X_n}{n}$ est la fréquence associée à X_n.

Pour *n* suffisamment grand, *p* appartient à l'intervalle $J_n = [F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}]$ avec une probabilité supérieure ou égale à 0,95.

Autrement dit: la proportion inconnue est telle que $P(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}) \cong 0.95$, lorsque F_n est la fréquence d'une loi binomiale $\mathcal{B}(n,p)$.

Démonstration :

Définition :
Soit f_{obs} une fréquence observée du caractère étudié sur un échantillon de taille *n*.
On **appelle intervalle de confiance de la proportion *p* au niveau de confiance 0,95**, l'intervalle $[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}}]$.

Remarques :

- Cet intervalle de confiance a pour centre..... et pour amplitude.....
- Un niveau de confiance 0,95 signifie que dans ..... cas sur ....., on affirme à juste titre que *p* appartient à l'intervalle de confiance.
- Il n'est pas vrai d'affirmer que *p* est égal au centre de l'intervalle de confiance. Il n'est pas possible d'évaluer la position de *p* dans l'intervalle de confiance.
- *p* étant inconnu, il n'est pas possible de vérifier si les conditions énoncées sur *n* et *p* en introduction de chapitre sont vérifiées. Cependant, il faudra les vérifier sur la fréquence observée *f* : $n \geq 30$, n×f≥5, et n×(1-f)≥5
- $[f_{obs} - \frac{1\sqrt{f_{obs}(1-f_{obs})}}{\sqrt{n}} ; f_{obs} + \frac{1\sqrt{f_{obs}(1-f_{obs})}}{\sqrt{n}}]$ est aussi un intervalle de confiance de la proportion *p* au niveau de confiance 0,95

Regle: Pour estimer une proportion p inconnue à partie d'un échantillon, on utilise en général l'encadrement fourni par un intervalle de confiance au niveau de confiance 0,95

Application 1 : *Estimer une proportion inconnue par un intervalle de confiance*
Un institut de sondage interroge 1052 personnes entre les deux tours de l'élection présidentielle sur leur intention de vote. 614 déclarent avoir l'intention de voter pour Martine Phinon.
En supposant que les votes seront conformes aux intentions, la candidate a-t-elle raison de croire qu'elle sera élue ?

Application 2 : *Déterminer une taille d'échantillon suffisante pour obtenir une estimation d'une proportion*
Un constructeur automobile fait appel à un institut de sondage afin de mesurer le degré de satisfaction du service après-vente. L'institut souhaite estimer la proportion de clients satisfaits au niveau de confiance 0,95 avec une amplitude d'au plus 5 centièmes.
Combien de personnes au minimum faut-il interroger ?