

Chapitre: Echantillonnage et estimation

1 Echantillonnage et intervalle de fluctuation

Dans tout le paragraphe p est un réel de l'intervalle]0; 1[et n un entier strictement positif.

1.1 Contexte

Propriété 1

Dans une certaine population, la proportion d'individus présentant le caractère C est p.

Un **échantillon aléatoire de taille** n est construit à partir de n tirages aléatoires avec remise dans la population (parfois les tirages sont sans remise mais la taille de la population est assez grande pour les assimiler à des tirages avec remise).

La variable aléatoire X_n donnant le nombre d'individus présentant le caractère C dans un échantillon aléatoire de taille n suit la loi binomiale $\mathscr{B}(n;p)$. On a : $\mathbb{E}(X_n) = np$ et $\mathbb{V}(X_n) = np(1-p)$.

La variable aléatoire $F_n = \frac{X_n}{n}$ donne alors la fréquence du caractère C dans l'échantillon aléatoire de taille n tiré de la population.

On a :
$$\mathbb{E}(F_n) = p$$
 et $\mathbb{V}(F_n) = \frac{p(1-p)}{n}$ donc $\sigma_{F_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

1.2 Intervalle de fluctuation exact avec la loi binomiale, méthode de Première

Propriété 2

Soit X_n une loi binomiale $\mathcal{B}(n, p)$ qui correspond à la réalisation d'un échantillon de taille n (avec remise) dans une population comportant un caractère avec une proportion p.

Un intervalle de fluctuation au seuil de 0,95 de la fréquence (du caractère) observée dans l'échantillon est l'intervalle $\left[\frac{a}{n};\frac{b}{n}\right]$,

• *a* est le plus petit entier tel que

$$\mathrm{P}(X_n \leq a) > 0,025$$

• b est le plus petit entier tel que

$$P(X_n \leqslant b) \geqslant 0,975$$

En effet, on a alors:

$$P\left(\frac{a}{n} \leqslant \frac{X_n}{n} \leqslant \frac{b}{n}\right) \geqslant 0,95$$

Exemple 1

1. Utiliser le programme suivant pour déterminer la borne inférieure de l'intervalle de fluctuation exact pour la loi binomiale \mathcal{B} (100; 0.3).



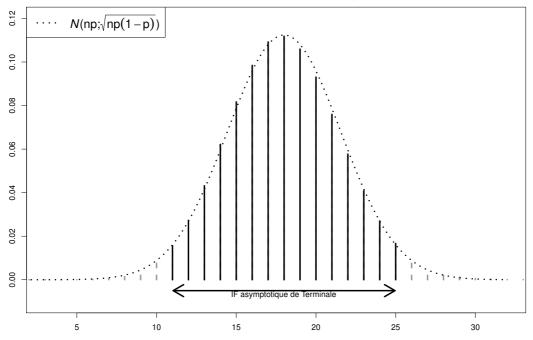
```
Programme TEXAS
                                                  Programme Casio
                                                  ?→ N
Prompt N
                                                  ?→ P
Prompt P
                                                  0→ K
0→ K
While binomFRep(N,P,K) \leq 0.025
                                                  While binominalCD(K,N,P)\leq 0.025
K+1 \rightarrow K
                                                 K+1 \rightarrow K
End
                                                 WhileEnd
Disp K
                                                 K 🗸
```

- **2.** Adapter le programme précédent pour déterminer la borne supérieure de l'intervalle de fluctuation exact pour la loi binomiale \mathcal{B} (100; 0.3).
- **3.** Compléter la fonction Python ci-dessous pour qu'elle retourne l'intervalle de l'intervalle de fluctuation exact pour la loi binomiale $\mathcal{B}(n; p)$.



1.3 Intervalle de fluctuation asymptotique, méthode de Terminale

Pour la loi B(60 ; 0.3) , F au seuil de 0,95 de première : [0.183333333333333333 ; 0.41666666666667] soit une probabilité de : 0.966547864257626 IF de Terminale : [0.184044836251248 ; 0.415955163748752] soit une probabilité de : 0.949884630350361



Théorème 1

Soit $\alpha \in]0$; 1[et u_{α} le réel tel que $\mathbb{P}(-u_{\alpha} \leq Z \leq u_{\alpha}) = 1 - \alpha$ où Z suit la loi normale $\mathcal{N}(0; 1)$. Si la variable aléatoire X_n suit la loi binomiale $\mathcal{B}(n; p)$ avec p dans l'intervalle]0; 1[alors :

$$\lim_{n \to +\infty} \mathbb{P}\left(\frac{X_n}{n} \in I_n\right) = 1 - \alpha \qquad \text{où} \qquad I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \tag{1}$$

Preuve: ROC

$$\frac{X_n}{n} \in I_n \iff p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\frac{X_n}{n} \in I_n \iff np - u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \leqslant X_n \leqslant np + u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\frac{X_n}{n} \in I_n \iff np - u_\alpha \sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha \sqrt{np(1-p)}$$



En posant $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ qui est la variable aléatoire centrée réduite associée à X_n il vient

$$\frac{X_n}{n} \in I_n \Longleftrightarrow -u_\alpha \leqslant Z_n \leqslant u_\alpha$$

Ainsi on $a \mathbb{P}(\frac{X_n}{n} \in I_n) = \mathbb{P}(-u_\alpha \leq Z_n \leq u_\alpha).$

Or d'après le théorème de Moivre-Laplace on $a\lim_{n\to+\infty}\mathbb{P}(-u_{\alpha}\leqslant Z_n\leqslant u_{\alpha})=\int_{-u}^{u_{\alpha}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx.$

$$D'ou \lim_{n \to +\infty} \mathbb{P}\left(\frac{X_n}{n} \in I_n\right) = \lim_{n \to +\infty} \mathbb{P}\left(-u_\alpha \leqslant Z_n \leqslant u_\alpha\right) = \int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Cette dernière intégrale est égale à $\mathbb{P}(-u_{\alpha} \leq Z \leq u_{\alpha})$ où Z suit la loi normale $\mathcal{N}(0; 1)$. Or $\mathbb{P}(-u_{\alpha} \leq Z \leq u_{\alpha}) = 1 - \alpha$ par définition de u_{α} .

$$Donc \lim_{n \to +\infty} \mathbb{P} \left(\frac{X_n}{n} \in I_n \right) = 1 - \alpha.$$

Définition 1

Un intervalle de fluctuation asymptotique de la variable aléatoire $F_n = \frac{X_n}{n}$ au seuil de $1 - \alpha$ est un intervalle déterminé à partir de p et de n et qui contient F_n avec une probabilité d'autant plus proche de $1 - \alpha$ que n est grand.

L'intervalle I_n du théorème précédent est donc un intervalle de fluctuation asymptotique de F_n au seuil de $1-\alpha$.

Propriété 3

Soit une variable aléatoire X_n suivant la loi binomiale $\mathcal{B}(n\,;\,p)$ avec p dans l'intervalle]0;1[et soit

$$I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$
 un intervalle de fluctuation « asymptotique au seuil de

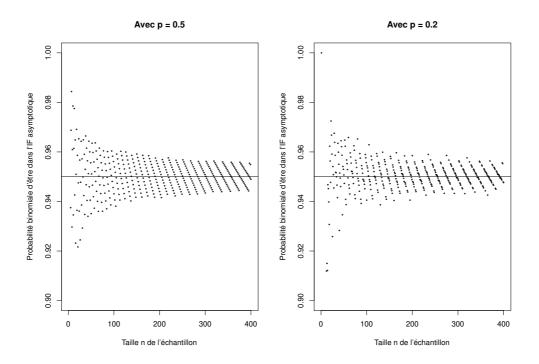
 $1-\alpha$ » de la variable aléatoire fréquence $F_n=\frac{X_n}{n}$.

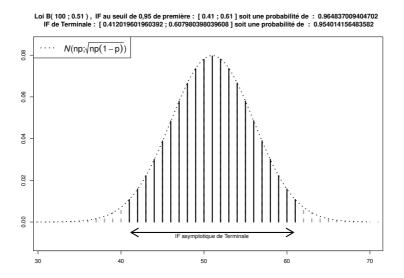
Sous les conditions, $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$, on peut utiliser l'approximation $\mathbb{P}(F_n \in I_n) \approx 1-\alpha$. Ainsi, l'intervalle de fluctuation asymptotique I_n peut être considéré comme un intervalle de fluctuation de F_n au seuil de $1-\alpha$.

Un cas particulier important est celui de l'intervalle de fluctuation asymptotique au seuil de **0,95** qui $\sqrt{n(1-n)}$

est
$$\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$
.







Exemple 2

1. Compléter la fonction Python if_asymptotique pour qu'elle retourne un intervalle de fluctuation asymptotique de terminale au seuil de 0,95, calculé avec la propriété précédente

```
from scipy.stats import norm
from math import sqrt

def invNorm(s):
    """Retourne le réel u tel que que P(X <= u)= s si X suit la loi</pre>
```



```
normale N(0,1)"""
  return norm.ppf(s)

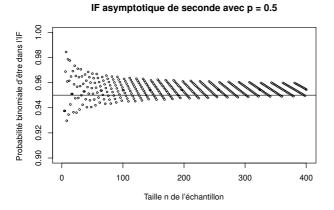
def if_asymptotique(n, p, s=0.95):
    """Retourne un intervalle de fluctuation asymptotique au seuil s
  pour la variable aléatoire fréquence X/n d'une v.a. suivant une loi
  binomiale de paramètres n et p"""
    u = invNorm(......)
    binf = ......
    bsup = ......
    return [binf, bsup]
```

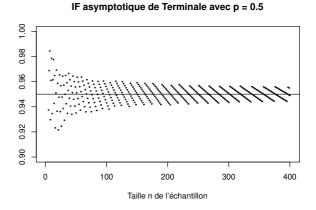
- **2.** Dans une maternité, on admet qu'il naît en moyenne 51 % des garçons. On fait le point sur la proportion de garçons toutes les 100 naissances.
 - **a.** Quelle loi suit la variable aléatoire *X* donnant le nombre de garçons dans un échantillon de 100 naissances?
 - **b.** Déterminer l'intervalle de fluctuation exact de la variable aléatoire fréquence $F = \frac{X}{100}$ associée à X.
 - **c.** Avec la loi normale, déterminer l'intervalle de fluctuation asymptotique au seuil de 95 % de la variable aléatoire fréquence $F = \frac{X}{100}$ pour un échantillon de taille 100.
- **3.** La proportion d'un caractère dans une population est p = 0,6. Déterminer un intervalle de fluctuation asymptotique de la fréquence de ce caractère dans les échantillons de taille 100, prélevés au hasard et avec remise :
 - **a.** au seuil de 0,8;

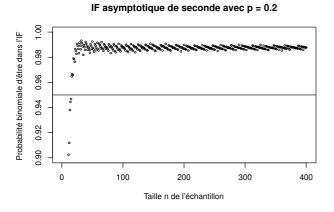
b. au seuil de 0, 9.

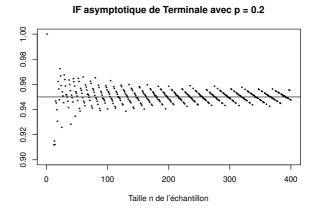


1.4 Lien avec l'intervalle de fluctuation de Seconde









Théorème 2

Si la variable aléatoire X_n suit la loi $\mathcal{B}(n; p)$, alors pour tout réel $p \in]0; 1[$, il existe un entier n_0 tel que, si $n \ge n_0$ alors $\mathbb{P}(p - \frac{1}{\sqrt{n}} \le \frac{X_n}{n} \le p + \frac{1}{\sqrt{n}}) > 0,95$ c'est-à-dire $\mathbb{P}(p - \frac{1}{\sqrt{n}} \le F_n \le p + \frac{1}{\sqrt{n}}) > 0,95$.

On justifie ainsi l'intervalle de fluctuation donnée en seconde : $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ pour la fréquence f d'un caractère dans un échantillon aléatoire réalisé avec remise de taille n extrait d'une population où la proportion du caractère est p.

Preuve:

Soit X_n suit la loi $\mathscr{B}(n;p)$, on définit $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ la variable centrée réduite associée à X_n . D'après le théorème de Moivre-Laplace, si on pose $a_n = \mathbb{P}(-2 \leqslant Z_n \leqslant 2)$ alors $\lim_{n \to +\infty} a_n = \mathbb{P}(-2 \leqslant Z \leqslant 2)$ où Z suit la loi $\mathscr{N}(0;1)$.



De plus d'après les propriétés de symétrie de la loi normale, si on pose $\lim_{n\to+\infty}a_n=\ell$ alors on a $\ell=\mathbb{P}\big(-2\leqslant Z\leqslant 2\big)=2\mathbb{P}\big(Z\leqslant 2\big)-1\approx 0,9544$ à 10^{-3} près avec les tables de la loi \mathcal{N} $(0\,;\,1)$. On a donc $\lim_{n\to+\infty}a_n=\ell>0,954$.

Soit ε un réel tel que ε < 0,004, puisque (a_n) converge vers ℓ , il existe un entier n_0 (dépendant de ε) tel que pour tout $n \ge n_0$ on ait :

$$a_n \in [\ell - \varepsilon; \ell - \varepsilon] \quad \text{c'est-} \hat{a} \text{-dire} \quad a_n > \ell - \varepsilon$$
 (2)

Or $\ell > 0,954$ et $\varepsilon < 0,004$, donc d'après (2) on a pour tout $n \ge n_0$:

$$a_n > 0.954 - 0.004$$
 c'est-à-dire $a_n > 0.95$ (3)

Or
$$a_n = \mathbb{P}\left(p - \frac{2}{\sqrt{n}}\sqrt{p(1-p)} \leqslant \frac{X_n}{n} \leqslant p + \frac{2}{\sqrt{n}}\sqrt{p(1-p)}\right)$$
.

Soit ψ la fonction trinôme définie sur [0;1] par $\psi(p)=p(1-p)=p-p^2$, ψ est positive et admet un maximum, atteint en $\frac{-1}{2\times (-1)}=\frac{1}{2}$.

Donc pour tout $p \in [0; 1]$ on $a \in \psi(p) = p(1-p) \leqslant \psi\left(\frac{1}{2}\right) = \frac{1}{4}$.

Par croissance de la fonction racine carrée sur \mathbb{R}^+ on a pour tout $p \in [0; 1]$:

$$\sqrt{p(1-p)} \leqslant \sqrt{\frac{1}{4}} = \frac{1}{2} \quad donc \, pour \, n > 0 \quad \frac{2}{\sqrt{n}} \sqrt{p(1-p)} \leqslant \frac{1}{\sqrt{n}}$$
 (4)

D'après (4), Les intervalles $J_n = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]$ et $I_n = \left[p - \frac{2}{\sqrt{n}} \sqrt{p(1-p)}; p + \frac{2}{\sqrt{n}} \sqrt{p(1-p)} \right]$ ont même centre et J_n a une plus grande amplitude donc $I_n \subset J_n$ donc $\mathbb{P}\left(\frac{X_n}{n} \in J_n\right) \geqslant \mathbb{P}\left(\frac{X_n}{n} \in I_n\right)$.

Or d'après (3), pour tout entier $n \ge n_0$, $a_n = \mathbb{P}(\frac{X_n}{n} \in I_n) > 0,95$ donc

$$\mathbb{P}\left(\frac{X_n}{n} \in J_n\right) = \mathbb{P}\left(p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}\right) > 0,95.$$



2 Prise de décision à partir d'un échantillon

Méthode Contexte et règle de décision

On fait l'hypothèse que la proportion d'un caractère *C* dans une population est *p*.

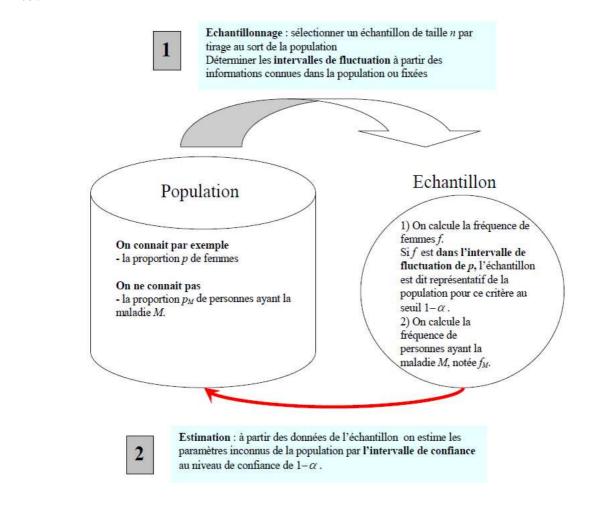
Sous cette hypothèse, le nombre d'individus X_n présentant le caractère C dans un échantillon aléatoire de taille n prélevé avec remise dans la population suit la loi $\mathcal{B}(n; p)$ et on peut déterminer un intervalle de

fluctuation I au seuil de 95 % de la variable aléatoire fréquence $F_n = \frac{X_n}{n}$ (un intervalle de fluctuation exact ou une approximation par un intervalle de fluctuation asymptotique).

On **observe** une fréquence *f* du caractère dans l'échantillon prélevé.

Pou tester l'hypothèse « La proportion du caractère dans la population est p » on adopte la règle de décision suivante :

- si *f* ∈ I (ce qui arrive dans 95 % des cas), on considère que l'hypothèse selon laquelle la proportion est *p* dans la population n'est pas remise en question et on l'accepte;
- si $f \notin I$, on **rejette** l'hypothèse selon laquelle cette proportion vaut p avec **un risque d'erreur de 5** %.



Exemple 3 Polynésie juin 2016

Pour des raisons pratiques, l'astronome responsable du club souhaiterait installer un site d'observation sur les hauteurs d'une petite ville de 2 500 habitants. Mais la pollution lumineuse due à l'éclairage public nuit à la



qualité des observations. Pour tenter de convaincre la mairie de couper l'éclairage nocturne pendant les nuits d'observation, l'astronome réalise un sondage aléatoire auprès de 100 habitants et obtient 54 avis favorables à la coupure de l'éclairage nocturne.

L' astronome fait l'hypothèse que 50 % de la population du village est favorable à la coupure de l'éclairage nocturne. Le résultat de ce sondage l'amène-t-il à changer d'avis?

Méthode Détermination pratique de l'intervalle de fluctuation au seuil de 95 %

- Si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, on utilise l'intervalle de fluctuation asymptotique au seuil de 95 % déterminé à la propriété 2: $\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
- Sinon on utilise l'intervalle de fluctuation exact déterminé en classe de Première avec la loi binomiale $\mathcal{B}(n; p)$ c'est-à-dire l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ où :
 - * a est le plus petit entier tel que P (X \leq a) > 0,025
 - * *b* est le plus petit entier tel que $P(X \le b) \ge 0,975$.

Exemple 4 Métropole juin 2014

La chaîne de production d'un laboratoire fabrique, en très grande quantité, le comprimé d'un médicament. La chaîne de production a été réglée dans le but d'obtenir au moins 97 % de comprimés conformes. Afin d'évaluer l'efficacité des réglages, on effectue un contrôle en prélevant un échantillon de 1 000 comprimés dans la production. La taille de la production est supposée suffisamment grande pour que ce prélèvement puisse être assimilé à 1 000 tirages successifs avec remise.

Le contrôle effectué a permis de dénombrer 53 comprimés non conformes sur l'échantillon prélevé. Ce contrôle remet-il en question les réglages faits par le laboratoire? On pourra utiliser un intervalle de fluctuation asymptotique au seuil de 95 %.



3 Intervalle de confiance

3.1 Principe général

On utilise un intervalle de fluctuation lorsque la proportion p du caractère étudié dans la population est connue ou si l'on fait une hypothèse sur sa valeur.

On utilise un intervalle de confiance lorsque l'on veut estimer une proportion inconnue dans une population à partir de la fréquence f observée sur un échantillon : c'est le problème des sondages.

3.2 Propriétés

Propriété 4

Soit F_n la variable aléatoire fréquence qui à tout échantillon aléatoire de taille n extrait avec remise d'une population dans laquelle la proportion d'un caractère est p, associe la fréquence obtenue.

Pour n assez grand, l'intervalle $\left[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}\right]$, la proportion p avec une probabilité au moins égale à 0,95.

Preuve: ROC

n est un entier strictement positif, $p \in [0; 1]$, X_n suit la loi $\mathcal{B}(n; p)$ et d'après le théorème 2, $F_n = \frac{X_n}{n}$ vérifie pour n assez grand : $\mathbb{P}(p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}) \geqslant 0,95$.

Or on
$$a: p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}.$$
Donc pour n assez grand on $a: \mathbb{P}(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}) \geqslant 0,95.$

Définition 2

Un intervalle de confiance pour une proportion p à un niveau de confiance $1-\alpha$ est la réalisation, à partir d'un échantillon, d'un intervalle aléatoire contenant la proportion p avec une probabilité supérieure ou égale à $1-\alpha$. Cet intervalle aléatoire est déterminé à partir de la variable aléatoire fréquence $F_n = \frac{X_n}{n}$.

En particulier, pour $n \ge 30$, $np \ge 5$ et $n(1-p) \ge 5$, si f est la fréquence observée d'un caractère dans un échantillon de aléatoire de taille n extrait avec remise d'une population dans laquelle la proportion d'un caractère est p, alors l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion p au niveau de confiance de 0,95 = 95%.

En pratique on ne connaît pas \boldsymbol{p} et on mesure la fréquence \boldsymbol{f} du caractère dans l'échantillon.

Sous les conditions $n \ge 30$, $nf \ge 5$ et $n(1-f) \ge 5$ on considère que $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance pour p au niveau de confiance 0,95.



3.3 Exemples

Exemple 5 Centres Étrangers juin 2016

Un institut effectue un sondage pour connaître, dans une population donnée, la proportion de personnes qui sont favorables à un projet d'aménagement du territoire. Pour cela, on interroge un échantillon aléatoire de personnes de cette population, et l'on pose une question à chaque personne.

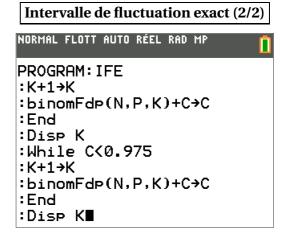
On suppose que n personnes ont répondu à la question ,et on admet que ces personnes constituent un échantillon aléatoire de taille n (où n est un entier naturel supérieur à 50).

Parmi ces personnes, 29 % sont favorables au projet d'aménagement.

- 1. Donner un intervalle de confiance, au niveau de confiance de 95 %, de la proportion de personnes qui sont favorables au projet dans la population totale.
- **2.** Déterminer la valeur minimale de l'entier *n* pour que l'intervalle de confiance, au niveau de confiance de 95 %, ait une amplitude inférieure ou égale à 0,04.

ANNEXE: programmes pour calculatrices TI

Intervalle de fluctuation exact (1/2) NORMAL FLOTT AUTO RÉEL RAD MP PROGRAM: IFE :Prompt N,P :0→K :binomFdp(N,P,K)→C :While C≤0.025 :K+1→K :binomFdp(N,P,K)+C→C :End :Disp K :While C<0.975



Intervalle de fluctuation asymptotique

```
PROGRAM: IFA
:Prompt N,P,A
:invNormale((1+A)/2,0,1,GA
UCHE) > U
:Disp P-U*\(\tau(P*(1-P)/N)\)
:Disp P+U*\(\tau(P*(1-P)/N)\)
:
```



Table des matières

1	Echantillonnage et intervalle de fluctuation	1
	1.1 Contexte	1
	1.2 Intervalle de fluctuation exact avec la loi binomiale, méthode de Première	1
	1.3 Intervalle de fluctuation asymptotique, méthode de Terminale	3
	1.4 Lien avec l'intervalle de fluctuation de Seconde	
2	Prise de décision à partir d'un échantillon	9
3	Intervalle de confiance	11
	3.1 Principe général	11
	3.2 Propriétés	
	3.3 Evemples	