Intervalle de fluctuation des fréquences. Estimation

Ce que dit le programme :

CONTENUS	CAPACITÉS ATTENDUES	COMMENTAIRES
Intervalle de fluctuation	• Démontrer que si la variable aléatoire X_n suit la loi $\mathcal{B}(n, p)$, alors pour tout α dans]0,1[on a : $\lim\limits_{n\to\infty} P\left(\frac{X_n}{n}\in I_n\right)=1-\alpha$ où I_n désigne l'intervalle $\left[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ • Connaître l'intervalle de fluctuation asymptotique (*) au seuil de 95% : $\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ où p désigne la proportion dans la population.	La démonstration ci-contre donne l'expression d'un intervalle de fluctuation asymptotique (*) au seuil $1-\alpha$ de la variable aléatoire fréquence $F_n=\frac{X_n}{n}$ qui, à tout échantillon de taille n, associe la fréquence obtenue f. Avec les exigences usuelles de précision, on pratique cette approximation dès que $n\geq 30$, $np\geq 5$ et $n(1-p)\geq 5$. En majorant $1{,}96\sqrt{p(1-p)}$, on retrouve l'intervalle de fluctuation présenté en classe de seconde. La problématique de prise de décision, déjà rencontrée, est travaillée à nouveau avec l'intervalle de fluctuation asymptotique.
Estimation Intervalle de confiance (*). Niveau de confiance.	• Estimer par intervalle une proportion inconnue à partir d'un échantillon. • Déterminer une taille d'échantillon suffisante pour obtenir, avec une précision donnée, une estimation d'une proportion au niveau de confiance 0,95.	• Les attendus de ce paragraphe sont modestes et sont à exploiter en lien avec les autres disciplines. • Il est intéressant de démontrer que, pour une valeur de p fixée, l'intervalle $\left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}}\right]$ contient, pour n assez grand, la proportion p avec une probabilité au moins égale à 0,95. On énonce alors que p est élément de l'intervalle $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}}\right]$ avec un niveau de confiance de plus de 95 %, où f désigne la fréquence observée sur un échantillon de taille n. Avec les exigences usuelles de précision, on utilise cet intervalle dès que $n\geq 30$, $np\geq 5$ et $n(1-p)\geq 5$. La simulation de sondages sur tableur permet de sensibiliser aux fourchettes de sondage. Il est important de noter que, dans d'autres champs, on utilise l'intervalle $\left[f-1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}} ; f+1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$ qu'il n'est pas possible de justifier dans ce programme. [⇆] [SVT] Analyse de graphiques où les données sont fournies par des intervalles de confiance. (AP) *Prise de décision lors de la comparaison de deux proportions (par exemple lors d'un essai thérapeutique).*

(*) Avec les notations précédentes :
Un intervalle de fluctuation asymptotique de la variable aléatoire F_n au seuil $1-\alpha$ 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.

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 $1-\alpha$, intervalle aléatoire déterminé à partir de la variable aléatoire fréquence F_n qui, à tout échantillon de taille n, associe la fréquence.

Les intervalles de confiance considérés ici sont centrés en la fréquence observée f.

I. Introduction

Dans une population, on étudie un ***caractère donné*** présent dans cette population.
On distingue les deux situations suivantes :

1ère situation : Intervalle de fluctuation et Conformité d'une hypothèse.

On connaît la proportion effective ***p*** du caractère dans la population totale. Soit $n \geqslant 1$.

On prélève un échantillon de taille ***n*** et on calcule la fréquence observée f_{obs}. On recommence l'épreuve un grand nombre de fois. On constate que ces fréquences observées f_{obs} varient en étant "suffisamment proches" de ***p*** dans 95% des cas, par exemple.

Le "*suffisamment proches*" dépend naturellement de la taille de l'échantillon.

On dit que les fréquences observées f_{obs} fluctuent dans un certain intervalle I.

Ou encore que I est l'intervalle de fluctuation des fréquences « au seuil de 95% » ou bien « au risque de 5% ».

Par conséquent : Si on prélève un échantillon de taille ***n*** et on calcule la fréquence observée f_{obs}. On peut émettre l'hypothèse H_1 que « pour ce caractère, *l'échantillon est représentatif* ou *conforme* à la population générale ». On prend la ***décision*** suivante :

- Si $f_{obs} \in I$, on peut affirmer que « ***au seuil de*** 95%, on peut ***accepter l'hypothèse*** H_1 de conformité » (réponse positive)
- Si $f_{obs} \notin I$, on peut affirmer que « ***au risque (d'erreur) de*** 5%, on peut ***rejeter l'hypothèse*** H_1 de conformité » (réponse négative).

Nous avons vu en Seconde :

Si la proportion effective du caractère ***p*** $\in$ **[0,2 ; 0,8]** et si on prélève un échantillon de taille ***n*** **> 25**, alors l'intervalle de fluctuation des fréquences f_{obs} ***au seuil de*** **95%** est donné par :

$$I = \left[p - \frac{1}{\sqrt{n}} \, ; \, p + \frac{1}{\sqrt{n}} \right] \text{, c'est-à-dire : } \mathbf{P}(f_{obs} \in \mathrm{I}) = 0{,}95 = 95\%.$$

Nous allons améliorer ce résultat.

Exemple 2nde : On lance une pièce de monnaie 2500 fois. On note la fréquence d'apparition du côté « Face ». On recommence cette opération plusieurs fois, en notant à chaque la fréquence d'apparition du côté « Face ». On constate que ces fréquences ne sont pas identiques.
1°) Déterminer l'intervalle de fluctuation de ces fréquences au seuil de 95%, en supposant que la pièce de monnaie soit parfaitement équilibrée.

2°) Si on note 1100 apparitions du côté « Face » sur les 2500 lancers, peut-on en déduire que la pièce soit « parfaitement équilibrée » ? Justifier votre réponse.

2ème situation : Estimation de la proportion effective par Intervalle de confiance.

On ne connaît pas la proportion effective ***p*** du caractère dans la population totale.

Nous allons faire ***un sondage sur un échantillon de taille n*** dans cette population. Sous certaines hypothèses sur ***n*** et ***p***, nous pouvons faire une estimation de la proportion p par intervalle de confiance.

Nous avons vu en Seconde :

Si la proportion effective ***p*** du caractère est comprise entre 0,2 et 0,8 et si on prélève des échantillons de taille ***n*** **> 25**, alors 95% des intervalles associés $\left[f - \frac{1}{\sqrt{n}} \, ; \, f + \frac{1}{\sqrt{n}} \right]$ contiennent ***p***.

Par conséquent :

On prélève un échantillon de taille ***n***, et on calcule la fréquence observée f_{obs} du caractère dans cet échantillon. Si $n > 25$ et $0,2 < p < 0,8$ alors la proportion p appartient à ***l'intervalle de confiance*** **IC** ***au seuil de*** **95%,** avec **IC** = $\left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$, c'est-à-dire : $P(p \in IC) = 0,95 = 95\%$.

Exemple historique : source (Institut de sondage IPSOS : http://www.ipsos.fr)

ÉLECTIONS PRÉSIDENTIELLES 2002 EN FRANCE :
DATE DU TERRAIN: Le 21 avril 2002.
ECHANTILLON : 1089 personnes constituant un échantillon national représentatif de la population française âgée de 18 ans et plus et inscrite sur les listes électorales.
METHODE : Echantillon interrogé par téléphone. Méthode des quotas (sexe, âge, profession du chef de famille, catégorie d'agglomération et région).
INTENTIONS DE VOTE AU 1er TOUR (Institut IPSOS) : **20%** pour J.CHIRAC, **18%** pour L.JOSPIN et **14%** pour J.M.LEPEN. On se prépare à un duel entre J.CHIRAC et L.JOSPIN au 2ème tour.

1. Déterminer les intervalles de confiance au niveau 95% donnant des estimations de la proportion de vote pour chacun des trois candidats.
2. Les résultats effectifs du 1er tour sont les suivants : **19,88%** pour J.CHIRAC, **16,18%** pour L.JOSPIN et **16,68%** pour J.M.LEPEN. Ces résultats sont-ils conformes avec les intervalles de confiance trouvés dans la question 1° ?
3. Analyser la situation avec les outils de la classe de 2nde. Qui a voté quoi ?. Les motivations de vote.
4. Reprendre l'exercice à la fin du chapitre pour analyser la situation avec les outils de la classe de Terminale.

II. Intervalle de fluctuation

2.1) Le théorème de MOIVRE-LAPLACE

Dans une population, on étudie un ***caractère donné,*** présent dans cette population. Dans ce paragraphe, on suppose qu'***on connaît la proportion effective p du caractère.***

Pour tout entier naturel non nul ***n***, on prélève un ***échantillon aléatoire de taille n,*** cest-à-dire en effectuant ***n*** tirages successifs indépendants (avec remise) dans cette population. *En général, on suppose que l'effectif total de la population est suffisamment grand pour supposer que les tirages sont effectués avec remise.*

on appelle X_n la variable aléatoire qui compte le nombre de succès, c'est-à-dire le nombre d'individus ayant le caractère étudié.

X_n suit la loi binomiale $\mathcal{B}(n, p)$, de paramètres n et p. On sait que :

$$E(X_n) = np, \quad V(X_n) = \sigma^2 = np(1-p) \quad \text{et} \quad \sigma = \sqrt{V(X_n)} = \sqrt{np(1-p)}$$

Enfin, on définit une nouvelle variable aléatoire Z_n de la manière suivante :

$$Z_n = \frac{X_n - m}{\sigma} = \frac{X_n - np}{\sqrt{np(1-p)}}$$

Z_n est une variable aléatoire ***centrée réduite*** associée à X_n : $E(Z_n) = 0$ et $\sigma(Z_n) = 1$.

Enfin, nous avons besoin aussi de la variable aléatoire ***centrée réduite Z*** qui suit la loi normale $\mathcal{N}(0,1)$ et définie dans la chapitre précédent par :

$$P(a \leqslant Z \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Théorème 1 de Moivre-Laplace :

Soit p un nombre réel (fixé) compris entre 0 et 1. On suppose que, pour tout entier $n \geqslant 1$, X_n une variable aléatoire qui suit une loi binomiale de paramètres n et p.

On appelle $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ la variable aléatoire centrée réduite associée à X_n .

Alors, pour tous nombres réels a et b tels que $a < b$, on a :

$$\boxed{\lim_{n \to +\infty} P(a \leqslant Z_n \leqslant b) = P(a \leqslant Z \leqslant b)}$$

c'est-àdire : $$\boxed{\lim_{n \to +\infty} P\left(a \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b\right) = P(a \leqslant Z \leqslant b)}$$

où Z est la variable aléatoire qui suit la loi normale centrée réuite $\mathcal{N}(0,1)$.

Théorème admis.

Par conséquent :

Corollaire 1 du théorème de Moivre-Laplace

Soit p un nombre réel (fixé) compris entre 0 et 1. On suppose que, pour tout entier $n \geqslant 1$, X_n une variable aléatoire qui suit une loi binomiale de paramètres n et p. On appelle $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ la variable aléatoire centrée réduite associée à X_n . Alors, si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, pour tous nombres réels a et b tels que $a < b$, on a l'approximation suivante :

$$\boxed{P(a \leqslant Z_n \leqslant b) \simeq P(a \leqslant Z \leqslant b)}$$

où Z est la variable aléatoire qui suit la loi normale centrée réuite $\mathcal{N}(0,1)$,

ou encore : $$\boxed{P\left(a \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \simeq P(a \leqslant Z \leqslant b)}$$

Conséquence immédiate du théorème ci-dessus.

2.2) Intervalle de fluctuation des fréquences

Soit $\boldsymbol{p}$ un nombre réel (fixé) compris entre 0 et 1. Pour tout entier $n \geqslant 1$, on prélève d'une manière aléatoire un échantillon de taille n et on calcule la ***fréquence observée*** du caractère dans l'échantillon.

Soit $\boldsymbol{X_n}$ une variable aléatoire égale au ***nombre de succès dans l'échantillon***. X_n suit la loi binomiale $\mathcal{B}(n, p)$ de paramètres n et p.

De même, on appelle $\boldsymbol{F_n}$ la variable aléatoire égale à ***la fréquence du succès dans l'échantillon***. On peut calculer sa moyenne et son écart-type à partir de ceux de X_n.

Comme pour tout $n \geqslant 1$: $F_n = \dfrac{X_n}{n} = \dfrac{1}{n} X_n$ et $E(aX) = a\,E(X)$, $V(a\,X) = a^2 V(X)$,

on obtient :

$$E(F_n) = p,\ V(F_n) = \frac{p(1-p)}{n} \quad \text{et} \quad \sigma(F_n)=\sqrt{V(F_n)}=\sqrt{\frac{p(1-p)}{n}}=\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Soit α un nombre réel (fixé) compris entre 0 et 1. On désigne par I_n l'intervalle :

$$I_n = \left[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

où $P(u_\alpha \leqslant Z \leqslant u_\alpha)=1-\alpha$ et Z suit la loi normale $\mathcal{N}(0,1)$.

Théorème 2

Soit α et p deux nombres réels (fixés) compris entre 0 et 1. On suppose que, pour tout entier $n \geqslant 1$, X_n une variable aléatoire qui suit une loi binomiale de paramètres n et p. Soit F_n est la variable aléatoire égale à la ***la fréquence du succès dans l'échantillon de taille n***. Alors, avec les notations ci-dessus, on a :

$$\boxed{\lim_{n\to+\infty} P(F_n \in I_n)=1-\alpha}$$

c'est-à-dire :

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

ou encore :

$$\boxed{\lim_{n\to+\infty} P\left(p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)=1-\alpha}$$

Définition :

Soit α un nombre réel (fixé) compris entre 0 et 1. L'intervalle I_n ainsi défini :

$$I_n = \left[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

s'appelle ***l'intervalle de fluctuation asymptotique des fréquences au seuil de* 1– α** (pour une réponse positive) ou ***au risque de* α** (pour une réponse négative).

Démonstration du théorème 2 (ROC)

En trois "petites" étapes :

- On transforme la double inégalité du théorème de Moivre-Laplace, de X_n en $F_n=\frac{X_n}{n}$, entre a et b, puis on écrit l'égalité des probabilités des deux événements ;
- On écrit la formule du théorème de Moivre-Laplace avec la limite de probabilités ;
- On introduit α et u_α avec la loi normale centrée réduite.

1ère étape : Pour tous nombres réels a et b tels que $a < b$, on a :

$$a \leqslant Z_n \leqslant b \quad \text{(ssi)} \quad a \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b$$

$$\text{(ssi)} \quad a\sqrt{np(1-p)} \leqslant X_n - np \leqslant b\sqrt{np(1-p)}$$

(ssi) $\quad \dfrac{a\sqrt{np(1-p)}}{n} \leqslant \dfrac{X_n - np}{n} \leqslant \dfrac{b\sqrt{np(1-p)}}{n}$ en divisant partout par n

(ssi) $\quad a\dfrac{\sqrt{np(1-p)}}{\sqrt{n^2}} \leqslant \dfrac{X_n}{n} - p \leqslant b\dfrac{\sqrt{np(1-p)}}{\sqrt{n^2}}$ en posant $n=\sqrt{n^2}$ au dénom.

(ssi) $\quad a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{X_n}{n} - p \leqslant b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ après simplification

(ssi) $\quad a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n - p \leqslant b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ avec $F_n = \dfrac{X_n}{n}$

(ssi) $\quad p + a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ en rajoutant p aux trois termes.

Par conséquent : $P\left(a \leqslant \dfrac{X_n - np}{\sqrt{np(1-p)}} \leqslant b\right) = P\left(p - a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$

2ème étape : D'après le théorème de Moivre-Laplace, on sait que :

$$\lim_{n\to+\infty} P\left(a \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b\right) = P(a \leqslant Z \leqslant b)$$

par suite : $\boxed{\lim_{n\to+\infty} P\left(p - a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = P(a \leqslant Z \leqslant b)}$

3ème étape : Soit α un nombre réel (fixé) compris entre 0 et 1. On sait qu'il existe un nombre réel u_α tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$. On pose alors : $a = -u_\alpha$ et $b = u_\alpha$

On obtient alors :

$$\boxed{\lim_{n\to+\infty} P\left(p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha.}$$

CQFD

2.3) Intervalle de fluctuation au seuil de 95%

Corollaire 1 (♥)

Si α = 0,05, $u_{0,05} = 1,96$. Donc si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, alors l'*intervalle de fluctuation asymptotique des fréquences au seuil de* 95% est donné par :

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

Démonstration :

C'est une conséquence directe du théorème précédent pour α = 0,05.

Or, d'après le cours [chapitre précédent], on sait que $u_{0,05} = 1,96$. d'où le résultat.

2.4) Prise de décision

Si on prélève un échantillon de taille ***n*** et on calcule la fréquence observée f_{obs}. On peut émettre l'hypothèse H_1 que « pour ce caractère, *l'échantillon est* ***représentatif*** ou ***conforme*** à la population générale ». On prend la ***décision*** suivante :

- Si $f_{obs} \in I$, on peut affirmer que « ***au seuil de*** 95%, on peut ***accepter l'hypothèse*** H_1 de conformité » (réponse positive)
- Si $f_{obs} \notin I$, on peut affirmer que « ***au risque (d'erreur) de*** 5%, on peut ***rejeter l'hypothèse*** H_1 de conformité » (réponse négative).

Exemple : [*Baccalauréat S Asie 18 juin 2013*] (*source APMEP*)

À des fins publicitaires, un grossiste affiche sur ses plaquettes : « 88 % de notre thé est garanti sans trace de pesticides ».

Un inspecteur de la brigade de répression des fraudes souhaite étudier la validité de l'affirmation. À cette fin, il prélève 50 boîtes au hasard dans le stock du grossiste et en trouve 12 avec des traces de pesticides. On suppose que, dans le stock du grossiste, la proportion de boîtes sans trace de pesticides est bien égale à 0,88.

On note F la variable aléatoire qui, à tout échantillon de 50 boîtes, associe la fréquence des boîtes ne contenant aucune trace de pesticides.

1°) Donner l'intervalle de fluctuation asymptotique de la variable aléatoire F au seuil de 95 %.

2°) L'inspecteur de la brigade de répression peut-il décider, au seuil de 95 %, que la publicité est mensongère ?

Corrigé :

1°) On vérifie tout d'abord que les hypothèses sont satisfaites :

- $n = 50$ et $50 > 30$ et $p = 0{,}88$;
- $np = 50 \times 0{,}88 = 44$ et $44 > 5$;
- $n(1-p) = 50 \times 0{,}12 = 6$ et $6 > 5$.

On sait alors que l'intervalle de fluctuation asymptotique au seuil de 95 % est égal à :

$$I_{50} = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

$$I_{50} = \left[0{,}88 - 1{,}96\frac{\sqrt{0{,}88 \times 0{,}12}}{\sqrt{50}} ; 0{,}88 + 1{,}96\frac{\sqrt{0{,}88 \times 0{,}12}}{\sqrt{50}}\right]$$

Conclusion : L'intervalle de fluctuation asymptotique au seuil de 95 % est égal, au centième près, à : $I_{50} = [0{,}79 ; 0{,}98]$.

2°) Calcul de la fréquence observée : $f_{obs} = \dfrac{50-12}{50} \simeq 0{,}76$.

L'inspecteur de la brigade de répression constate une proportion de lots sans pesticides d'environ $f_{obs} \approx 0{,}76$. Or $0{,}76 \notin I_{50}$.

Par conséquent, l'inspecteur de la brigade de répression constate donc qu'au risque de 5 %, l'hypothèse de conformité est rejetée.

Conclusion : l'inspecteur de la brigade de répression décide qu'au risque de 5 %, la publicité du grossiete est mensongère.

2.5) Conséquence

Corollaire 2 (ROC) : Version simplifiée vue en seconde

Si $\alpha = 0{,}05$, $u_{0,05} = 1{,}96$. Donc, pour n assez grand, c'est-àdire si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, alors l'intervalle $I = \left[p - 1/\sqrt{n}\,;\, p + 1/\sqrt{n}\right]$ contient les fréquences observées avec une probabilité au moins égale à 0,95. Une approximation de l'*intervalle de fluctuation asymptotique des fréquences au seuil de* 95% est donné par :

$$I = \left[p - \frac{1}{\sqrt{n}}\,;\, p + \frac{1}{\sqrt{n}}\right]$$

Démonstration : On suppose que les conditions du corollaire 1 sont vérifiées.

On définit une fonction f par : $f(p) = 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ ou $f(p) = \dfrac{1{,}96}{\sqrt{n}}\sqrt{p-p^2}$

La fonction f est définie et continue sur l'intervalle [0 ; 1] et dérivable sur]0 ; 1[et pour tout $p \in$]0 ; 1[on a : $f'(x) = \dfrac{1-2p}{2\sqrt{p(1-p)}}$. Le signe de $f'(x)$ est le même que celui de $1-2p$. On obtient le tableau de variations suivant :

p	0		1/2		1
$f'(p)$	‖	+	0	–	‖
$f(p)$	0	↗	$f(1/2)$	↘	0

Or, $f\left(\dfrac{1}{2}\right) = \dfrac{1{,}96}{\sqrt{n}} \times \sqrt{\dfrac{1}{2} \times \dfrac{1}{2}}$ donc $f\left(\dfrac{1}{2}\right) = \dfrac{1{,}96}{2\sqrt{n}} < \dfrac{1}{\sqrt{n}}$.

La fonction f admet donc un maximum pour $p = 1/2$. Donc, pour tout entier naturel n non nul et tout nombre réel $p \in$ [0 ; 1] on a : $f(p) \leqslant f\left(\dfrac{1}{2}\right) < \dfrac{1}{\sqrt{n}}$ puisque $1{,}96 < 2$.

Par conséquent : $\left[p - 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\, p + 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[p - \dfrac{1}{\sqrt{n}}\,;\, p + \dfrac{1}{\sqrt{n}}\right]$

Or, si A et B sont deux événements tels que $A \subset B$, alors $P(A) \leqslant P(B)$.

Or, on sait que $P\left(\left[p - 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\, p + 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) = 0{,}95$

donc :

$$P\left(\left[p - \frac{1}{\sqrt{n}}\,;\, p + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0{,}95$$

Conclusion : Si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, alors l'*intervalle de fluctuation asymptotique des fréquences F_n au seuil de* 95% est contenu dans l'intervalle :

$$I_n = \left[p - \frac{1}{\sqrt{n}}\,;\, p + \frac{1}{\sqrt{n}}\right] .$$

III. Estimation d'une proportion

3.1) Intervalle de confiance

Soit ***n*** un entier naturel non nul. On cherche à ***estimer*** ***la proportion effective p*** (***inconnue*** *dans ce pragraphe*) d'un caractère dans une population donnée.

Pour tout entier naturel non nul ***n***, on prélève un ***échantillon aléatoire de taille n,*** cest-à-dire en effectuant ***n*** tirages successifs indépendants (avec remise) dans cette population. *En général, on suppose que l'effectif total de la population est suffisamment grand pour supposer que les tirages sont effectués avec remise.*

Sous certaines hypothèses sur ***n*** et ***p***, nous pouvons faire une estimation de la proportion effective ***p*** par intervalle de confiance.

Définition :

Soit α un nombre réel (fixé) compris entre 0 et 1. Alors, ***un intervalle de confiance*** pour une proportion *p* à un ***niveau de confiance 1– α***, est la ***réalisation*** d'un intervalle, noté **IC** (*avec deux lettres*) qui, pour *n* assez grand, contient la proportion *p* avec une probabilité au moins égale à 1– α, c'est-à-dire :

$$P(p \in IC) \geqslant 1-\alpha$$

Seul le cas **α = 0,05** est au programme de Terminale, c'est-à-dire un ***niveau de confiance de 95%***. Cela signifie que, si on réalisait un grand nombre de sondages de même taille *n*, et on calculait les intervalles de confiance correspondants, alors ***plus de 95% de ces intervalles contiennent la proportion p***.

3.2) Estimation de p par intervalle de confiance au niveau 95%

Soit ***n*** un entier naturel non nul. On prélève un ***échantillon aléatoire de taille n***.

On appelle X_n une variable aléatoire « nombre de succès », alors X_n suit une loi binomiale de paramètres *n* et *p*. Soit F_n est la variable aléatoire égale à la ***la fréquence du succès dans l'échantillon de taille n***. Alors, on a les résultats suivants :

Propriété ***admise*** :

Soit *p* un nombre réel (fixé) compris entre 0 et 1 et ***n*** un entier naturel non nul.
Alors, pour n suffisamment grand, l'intervalle aléatoire $\left[F_n-\frac{1}{\sqrt{n}};F_n+\frac{1}{\sqrt{n}}\right]$ contient la proportion ***p*** avec une probabilité au moins égale à 95%.

Conséquence pratique :

Si f désigne la fréquence observée sur un échantillon de taille *n* alors, avec les exigences usuelles de précision, c'est-à-dire : si $n \geqslant 30$, $nf \geqslant 5$ et $n(1-f) \geqslant 5$, l'intervalle **IC** = $\left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right]$ est ***un intervalle de confiance*** pour la proportion ***p*** au ***niveau de confiance de 95%***. *Cet intervalle est aussi appelé « **fourchette de sondage** ».*

Exemple : Dans un échantillon de 100 voitures, prélevé au hasard dans un parc de voitures de location, on a constaté que 88 n'ont pas eu de sinistre au cours des 6 derniers mois.
Donner un intervalle de confiance de la proportion *p* au niveau de confiance de 95%, de voitures n'ayant pas eu de sinistre au cours des 6 derniers mois. On suppose que le nombre de véhicule est suffisamment grand pour effectuer des tirages sans remise.

Corrigé. On calcule la fréquence observée : $f=\frac{88}{100}$.

On vérifie les hypothèses : on a n = 100, donc si $n\geqslant 30$; nf = 88, donc $nf\geqslant 5$ et $n(1–f)$ = 12, donc $n(1-f)\geqslant 5$.

Par conséquent une réalisation de l'intervalle de confiance pour la proportion p au niveau de confiance de 95% est donnée par :

IC = $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]=\left[0{,}88-\frac{1}{\sqrt{100}};0{,}88+\frac{1}{\sqrt{100}}\right]$ donc IC = $[0{,}78\,;\,0{,}98]$.

3.3) Déterminer la taille d'un échantillon

Comment déterminer une taille d'échantillon suffisante pour obtenir, avec une précision donnée, une estimation d'une proportion au niveau de confiance 0,95.

La précision de l'estimation est donnée par l'amplitude de l'intervalle de confiance et dépend de la taille ***n*** de l'échantillon.

Définition :

On appelle ***amplitude d'un intervalle*** [a ; b] *la longueur* ***l = b – a*** de cet intervalle.

Propriété 1. :

On considère un échantillon de taille ***n***. L'aplitude d'un intervalle de confiance au niveau de confiance 95% est égale à $l=\frac{2}{\sqrt{n}}$.

En effet : $l = b - a = \left(f+\frac{1}{\sqrt{n}}\right)-\left(f-\frac{1}{\sqrt{n}}\right)=\frac{2}{\sqrt{n}}$.

Application : Déterminer une taille d'échantillon suffisante pour obtenir, avec une précision d'amplitude inférieure à 0,06, une estimation d'une proportion au niveau de confiance 0,95.

En effet, $l\leqslant 0{,}06\Leftrightarrow\frac{2}{\sqrt{n}}\leqslant 0{,}06\Leftrightarrow\frac{4}{n}\leqslant 0{,}06^2\Leftrightarrow n\geqslant\frac{4}{0{,}0036}\Leftrightarrow n\geqslant 1111{,}11$

Conclusion : Pour obtenir une estimation d'une proportion au niveau de confiance 0,95 avec une précision d'amplitude inférieure ou égale à 6%, il faut prélever un échantillon d'une taille d'au moins 1112 individus.