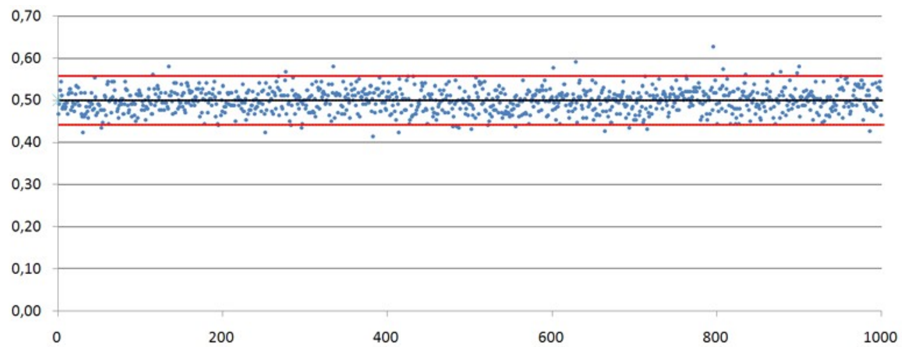
INTERVALLE(S) DE FLUCTUATION



I. Définition

DÉFINITION.

Soit $X \sim \mathcal{B}(n;p)$ et $\alpha \in\ ]0;1[$.

On appelle ***intervalle de fluctuation au seuil 1-α*** un intervalle I qui vérifie :

$$p(X \in I) \geqslant 1-\alpha$$

En Première, vous avez défini ainsi l'intervalle de fluctuation au seuil 0,95 :

DÉFINITION.

Soit $X \sim \mathcal{B}(n;p)$ et $\alpha \in\ ]0;1[$.

On appelle ***intervalle de fluctuation au seuil 0,95*** l'intervalle $I=\left[\frac{a}{n},\frac{b}{n}\right]$ où :

- a est le plus grand entier vérifiant $p(X<a) \leqslant 0{,}025$
- b est le plus petit entier vérifiant $p(X>b) \leqslant 0{,}025$.

Une autre définition équivalente, et bien plus pratique, est la suivante :

On appelle ***intervalle de fluctuation au seuil 0,95*** l'intervalle $I=\left[\frac{a}{n},\frac{b}{n}\right]$ où :

- a est le plus petit entier vérifiant $p(X \leqslant a) > 0{,}025$
- b est le plus petit entier vérifiant $p(X \leqslant b) \geqslant 0{,}975$.

Exemple : en première partie de soirée, une série a attiré près de 6,2 millions de téléspectateurs soit 34 % de part d'audience. Paul pense faire un sondage auprès de 100 habitants de son village, pour savoir la proportion de ceux qui ont regardé cette série.

Soit X la variable aléatoire qui correspond au nombre de téléspectateurs qui ont regardé cette série dans un échantillon de 100 personnes ayant regardé la télévision en première partie de soirée.
Le nombre de téléspectateurs en première partie de soirée est suffisamment important pour considérer que la variable X suit la loi binomiale de paramètres $n =$ ……… et $p =$ ……… .
Le plus petit entier a tel que $\mathrm{p}(X \leqslant a) > 0{,}025$ est ………
et, le plus petit entier b tel que $\mathrm{p}(X \leqslant a) \geqslant 0{,}975$ est ……… .
Un intervalle de fluctuation à 95 % de la fréquence des téléspectateurs qui ont regardé cette série dans un échantillon de taille 100 est donc :

………………………………

De la même manière, déterminer un intervalle de fluctuation à 90 % :

L'intérêt de l'intervalle de Première est qu'il fournit un intervalle pour toutes les valeurs de n et de p, alors que l'intervalle $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right]$ de Seconde n'est pas adapté pour les « petites binomiales », comme nous le verrons un peu plus loin.

On peut également adapter cette méthode à d'autres pourcentages et déterminer des intervalles de fluctuation à 90 %, à 99 %…, suivant les besoins. Ce n'était pas possible en Seconde.

Cet intervalle de fluctuation est alors parfois difficile à déterminer pour de grandes valeurs de n .
Par exemple :

- sur une calculatrice *Casio Graph 35+*, $\binom{500}{350}$ affiche *Erreur math*
mais le menu DIST-BINM permet d'avoir des valeurs approchées des $\mathrm{p}(X=k)$ …
- sur *Xcas*, la probabilité $\mathrm{p}(X \leqslant 101)$ où $X \sim \mathscr{B}(n ; 0{,}052)$ n'est bien calculée que jusqu'à $n=3975$.

Il nous faudrait donc un autre outil permettant de conclure dans ces cas-là.

II. Intervalle de fluctuation asymptotique

II.1 Définition

DÉFINITION.

Soit $X \sim \mathcal{B}(n;p)$ et $\alpha \in \,]0;1[$. On note $F_n = \dfrac{X_n}{n}$.

On appelle ***intervalle de fluctuation asymptotique*** de la v.a.r. F_n ***au seuil de 1-α*** un intervalle déterminé à partir de n et p et qui contient F_n avec une probabilité d'autant plus proche de 1-α que n est grand.

Remarque : il n'existe pas un unique intervalle de fluctuation asymptotique à un seuil donné !
On emploiera l'expression « l'intervalle de fluctuation asympt. » pour désigner celui utilisé en Terminale.

II.2 Approximation à l'aide de la loi normale

ROC

PROPRIÉTÉ. Si $X_n \sim \mathcal{B}(n;p)$, alors pour tout $\alpha \in \,]0;1[$:

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

où : $F_n = \dfrac{X_n}{n}$

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

u_α désigne le réel tel que $p(-u_\alpha \leqslant T \leqslant u_\alpha) = 1-\alpha$ où $T \sim \mathcal{N}(0\,;\,1)$.

Démonstration :

Soit $X_n \sim \mathcal{B}(n;p)$ et $\alpha \in \,]0;1[$. On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.

On admet qu'il existe un unique réel $u_\alpha > 0$ tel que $p(-u_\alpha \leqslant T \leqslant u_\alpha) = 1-\alpha$.

1. En utilisant le théorème de Moivre-Laplace, démontrer que : $\lim\limits_{n\to+\infty} p(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1-\alpha$.

2. Démontrer que : $p(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = p\left(p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{X_n}{n} \leqslant p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$

3. Compléter le théorème démontré : $\lim\limits_{n\to+\infty} p\left(\dfrac{X_n}{n} \in I_n\right) = 1-\alpha$ où $I_n = \dots$

Remarque : on admet que l'on peut utiliser l'approximation $p\left(\dfrac{X_n}{n} \in I_n\right) \approx 1-\alpha$ lorsque :

$$n \geqslant 30 \quad ; \quad np \geqslant 5 \quad ; \quad n(1-p) \geqslant 5$$

(l'approximation étant d'autant meilleure que n est grand)

D'après la propriété de la loi normale centrée réduite, on sait que $u_{0,05} \approx 1{,}96$, d'où :

DÉFINITION.
L'intervalle de fluctuation asymptotique au seuil 0,95 de la variable aléatoire fréquence est :

On a également vu que $u_{0,01} \approx 2{,}58$, donc on pourrait aussi dire qu'un intervalle de fluctuation asymptotique au seuil 0,99 est :

II.3 <u>Intervalle de fluctuation de la classe de Seconde</u>

En majorant $1{,}96\sqrt{p(1-p)}$ par 1, on retrouve l'intervalle de fluctuation vu en Seconde :

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

<u>Démonstration</u> :

Soit $f(p)=p(1-p)=-p^2+2p$.

1. Étudier les variations de la fonction *f*.

2. En déduire que $0 \leqslant p(1-p) \leqslant \frac{1}{4}$ puis que $1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{1}{\sqrt{n}}$ et $-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \geqslant -\frac{1}{\sqrt{n}}$.

3. Conclure.

L'intervalle vu en Seconde est donc une approximation de l'intervalle de fluctuation de T°S.

Remarque : ces deux intervalles sont centrés sur la proportion p du caractère étudié dans la population.
Ce n'est pas nécessairement le cas de l'intervalle de fluctuation déterminé en Première.

II.4 Critiques : tout ça est-il vraiment rigoureux ?

Critique n°1 : tirages avec remise ?

Dans les exemples, les tirages sont souvent effectués sans remise (par exemple, lors d'un sondage).
La taille des échantillons considérés étant faible par rapport à la taille de la population totale, on assimile souvent les tirages réalisés à des tirages avec remise et on peut alors appliquer les résultats précédents.

Critique n°2 : « environ 95 % » ?

On pose $\mathrm{I}_n=\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ et $\mathrm{J}_n=\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$.
Prenons $\alpha=0{,}05$, le cas le plus utilisé.
En T^le, on a vu qu'on utilise l'approximation $\mathrm{p}(\mathrm{F}_n\in\mathrm{I}_n)\approx0{,}95$ lorsque : $n\geqslant30$; $n\,p\geqslant5$; $n(1-p)\geqslant5$.
En 2^nde, on a vu qu'on utilise l'approximation $\mathrm{p}(\mathrm{F}_n\in\mathrm{J}_n)\approx0{,}95$ lorsque $n\geqslant25$ et $0{,}2\leqslant p\leqslant0{,}8$.
Mais cette valeur approchée est-elle toujours (souvent ?) au-dessus de 1-α ? (ce serait le mieux pour prendre des décisions). Puisque $\lim_{n\to+\infty}\mathrm{p}(F_n\in\mathrm{I}_n)=0{,}95$, on peut démontrer qu'il existe un entier n_0 tel que si $n\geqslant n_0$ alors $\mathrm{p}(F_n\in\mathrm{I}_n)\geqslant0{,}95$. Mais cet entier n_0 dépend de p, et peut varier considérablement.
A l'aide d'algorithmes, on peut obtenir le tableau suivant :

p	0,35	0,36	0,37	0,38	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5
n_0	31	30	36	64	56	81	90	120	143	209	271	288	304	399	399	529

Voici quelques exemples où le résultat est inférieur à 95 %, et même où les conditions ne sont pas vérifiées mais la probabilité cherchée est de quasiment 100 % !

n	p	$n\geqslant30$, $n\,p\geqslant5$ et $n(1-p)\geqslant5$?	$\mathrm{p}(\mathrm{F}_n\in\mathrm{I}_n)$	$n\geqslant25$ et $0{,}2\leqslant p\leqslant0{,}8$?	$\mathrm{p}(\mathrm{F}_n\in\mathrm{J}_n)$
141	0,422	oui	**≈ 0,9397**	oui	≈ 0,9597
222	0,241	oui	**≈ 0,9408**	oui	≈ 0,9817
1530	0,1	oui	**≈ 0,9450**	**non**	≈ 0,9992
1700	0,006	oui	≈ 0,9600	**non**	≈ 1
30	0,484	oui	**≈ 0,9341**	oui	**≈ 0,9341**
30	0,48	oui	**≈ 0,9340**	oui	≈ 0,9550
34	0,8087	oui	**≈ 0,9234**	**non**	≈ 0,9918
528	0,5	oui	**≈ 0,9499**	oui	**≈ 0,9499**
28	0,59	**non**	**≈ 0,9474**	oui	**≈ 0,9474**

En programmant des algorithmes (par exemple sur *Xcas*), on peut déterminer qu'en faisant varier p de 0 à 1, avec un pas de 0,0001, et en faisant varier *n* de 1 à 2000 :

- sur les 19 288 321 couples $(n;p)$ qui vérifient les conditions $n\geqslant30$, $n\,p\geqslant5$ et $n(1-p)\geqslant5$, 9 251 445 donnent $\mathrm{p}(\mathrm{F}_n\in\mathrm{I}_n)<0{,}95$. Cela fait tout de même environ 47,96 % des couples !
  Il y a même 102 411 couples qui donnent $\mathrm{p}(\mathrm{F}_n\in\mathrm{I}_n)<0{,}94$.

- sur les 11 856 000 couples $(n;p)$ qui vérifient les conditions $n\geqslant25$ et $0{,}2\leqslant p\leqslant0{,}8$, 43 336 donnent $\mathrm{p}(\mathrm{F}_n\in\mathrm{J}_n)<0{,}95$. Cela fait seulement environ 0,37 % des couples !

• la plus grande valeur de n telle que $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right) < 0{,}95$ est $n = 528$ (et $p = 0{,}5$).
On trouve une probabilité d'environ 0,9499.

• la probabilité minimale des $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right)$ (en vérifiant les conditions $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$) est atteinte lorsque $n = 30$ et $p = 0{,}484$. On trouve une probabilité d'environ 0,9341.

• la plus grande valeur de n telle que $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{I}_n\right) < 0{,}95$ est $n = 2000$ (lorsque $p = 0{,}9973$).
On trouve une probabilité d'environ 0,9470.

• la probabilité minimale des $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{I}_n\right)$ (en vérifiant les conditions $n \geqslant 30$, $n\,p \geqslant 5$ et $n(1-p) \geqslant 5$) est atteinte lorsque $n = 34$ et $p = 0{,}8087$
On trouve une probabilité d'environ 0,9234.

• en faisant varier n de 25 à 2000 et p de 0 à 1 tous les 0,001, on trouve 3000 couples $(n\,;p)$ qui vérifient les conditions de Seconde mais pas les conditions de Terminale ; et on trouve 746 287 couples $(n\,;p)$ qui vérifient les conditions de Terminale mais pas les conditions de Seconde, soit environ 38 % des couples…
Par contre, on montre facilement que si $n \geqslant 30$: si $0{,}2 \leqslant p \leqslant 0{,}8$ alors $n\,p \geqslant 5$ et $n(1-p) \geqslant 5$.
Autrement dit, si $n \geqslant 30$, si $(n\,;p)$ vérifie les conditions de Seconde, alors il vérifie celles de T$^{\text{le}}$.

À RETENIR

• pour l'intervalle J_n vu en Seconde, dès que $n > 528$ on a $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right) \geqslant 0{,}95$.

• même si l'intervalle I_n vu en Terminale donne des probabilités plus souvent en dessous de 95 % qu'avec l'intervalle J_n de Seconde, cet intervalle I_n a une amplitude bien plus petite…
Le résultat est donc plus précis.

III. Prise de décision

III.1 Conjecturer une proportion et valider/invalider cette hypothèse

On considère un caractère dont la proportion dans la population est supposée être égale à p.
La prise de décision consiste, à partir d'un échantillon de taille n, à valider ou non cette hypothèse faire sur la proportion p :

1) On calcule la fréquence observée f du caractère dans cet échantillon.

2) Si les conditions d'approximation $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ sont vérifiées, on détermine l'intervalle de fluctuation asymptotique au seuil 0,95.
Sinon, on peut déterminer l'intervalle de fluctuation étudié en Seconde ou en Première…

3) On applique la règle suivante :

- Si $f \notin \mathrm{I}$ alors on rejette l'hypothèse faite sur p.
  Dans ce cas, **il y a un risque[1] de se tromper de 5 % :**
  la probabilité qu'on rejette à tort l'hypothèse faite sur p alors qu'elle est vraie (proba. cond.) est environ égale à 5 %.

- Si $f \in \mathrm{I}$ alors on accepte l'hypothèse faite sur p.
  Dans ce cas, **le risque d'erreur n'est pas quantifié[2] !**

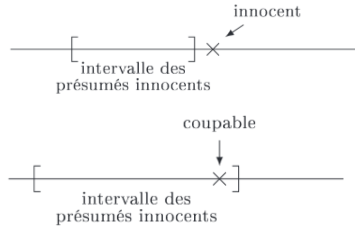
Le tableau ci-dessous représente bien les deux types d'erreurs :

On fait l'hypothèse $p=p_0$		
Réalité ↓ Décision →	$p=p_0$	$p \neq p_0$
$p=p_0$	OK	erreur : rejeter à tort
$p \neq p_0$	erreur : valider à tort	OK

Le risque de valider à tort n'est pas quantifiable car cela signifie que nous avons validé l'hypothèse $p=p_0$ alors que $p \neq p_0$: mais alors on ne sait pas combien vaut p ! On ne peut donc pas calculer la probabilité conditionnelle de valider cette hypothèse sachant qu'elle est fausse…

Pourquoi ne pas abaisser le seuil de rejet d'une hypothèse ?

On pourrait penser qu'il suffit de réduire le risque d'erreur (de première espèce) de rejeter une hypothèse à tort, de façon à n'avancer que des hypothèses très fiables.
Mais en faisant cela, on augmente le risque de commettre une autre erreur (de seconde espèce) : accepter l'hypothèse alors qu'elle est fausse !
Une analogie simple suffit à faire comprendre la situation : une prise de décision est comme un jugement au tribunal (hypothèse = le prévenu est présumé innocent). Il y a deux risques au jugement : celui de condamner un innocent (rejet à tort de l'hypothèse – première espèce), ou d'innocenter un coupable (acceptation à tort de l'hypothèse – seconde espèce).



Plus on réduit un risque, plus on augmente l'autre : ainsi, la décision que l'on doit prendre est un compromis adapté à la situation.
Voilà pourquoi le seuil de 5 % est souvent utilisé.

1 On parle de « risque de première espèce ». Ce risque est défini à l'avance (le plus souvent 1 % ou 5 %).
2 On parle de « risque de seconde espèce ».
À taille d'échantillon égale, si l'on diminue le risque de première espèce, on augmente le risque de seconde espèce…

III.2 Tester la conformité d'un échantillon par rapport à la population

On considère un caractère dont la proportion dans la population est connue, égale à p .
La prise de décision consiste, à partir d'un échantillon de taille n , à valider ou non sa représentativité.

1) On calcule la fréquence observée f du caractère dans cet échantillon.

2) Si les conditions d'approximation $n \geqslant 30$, $n\,p \geqslant 5$ et $n(1-p) \geqslant 5$ sont vérifiées, on détermine l'intervalle de fluctuation asymptotique au seuil 0,95.
Si les conditions d'approximation ne sont pas vérifiées, on peut déterminer l'intervalle de fluctuation étudié en Seconde ou en Première…

3) On applique la règle suivante :

- Si $f \in \mathrm{I}$ alors on considère que l'échantillon est représentatif de la population.
- Si $f \notin \mathrm{I}$ alors on considère que l'échantillon n'est pas représentatif de la population.

Remarque : ici, on ne fait pas d'hypothèse sur la probabilité théorique p, puisqu'on la connaît.
On ne commet dont aucune erreur…