Echantillonnage

I) Position du problème

Dans une population on sait que le caractère C est présent avec la proportion p (on pourrait dire que la probabilité de choisir un individu ayant le caractère C est p).

On prélève un échantillon de taille n dans cette population.

Deux questions peuvent se poser :

• Que peut-on dire de la fréquence f du caractère C dans cet échantillon ?

• Si on peut calculer cette fréquence f peut-on affirmer que cet échantillon est représentatif de la population ?

II) Rappels de seconde

En classe de seconde, on a observé que sur un grand nombre d'échantillons de taille n, 95% au moins fournissent une fréquence f appartenant à l'intervalle $[\ p - \frac{1}{\sqrt{n}}\ ; p + \frac{1}{\sqrt{n}}]$ sous certaines conditions pour n et p.

On dispose donc en termes de probabilité du résultat suivant :

Pour $n \geq 25$ et $0{,}2 \leq p \leq 0{,}8$ lorsqu'on prélève un échantillon de taille n dans une population ou la probabilité du caractère est p, la fréquence f du caractère sur cet échantillon appartient à l'intervalle $[p - \frac{1}{\sqrt{n}}\ ; p + \frac{1}{\sqrt{n}}\ ]$ avec une probabilité supérieure ou égale à 0,95

Remarque :

Ce résultat permet de répondre aux deux questions évoquées au I) sous les conditions données sur n et p

• Dans un échantillon de taille n on a plus de 95% de chances d'avoir une fréquence dans l'intervalle donné.

• Si la fréquence du caractère dans l'échantillon de taille n est dans l'intervalle donné, on peut prévoir que l'échantillon est représentatif de la population au seuil de confiance de 95 %

Exemples

1) On sait que dans la population française les naissances de garçons ont pour probabilité p = 0,512. On prélève un échantillon de 100 enfants.

a) Donner un intervalle donnant la fréquence probable au seuil de 95%

b) Quel est le nombre de garçons probables dans cet échantillon ?

Solution :

a) D'après le résultat précédent un intervalle de fluctuation est

$[\ 0{,}512 - \frac{1}{\sqrt{100}}\ ;\ 0{,}512 + \frac{1}{\sqrt{100}}\ ] = [\ 0{,}412\ ;\ 0{,}612\ ]$

b) On peut donc estimer que dans cet échantillon on doit avoir au seuil de 95% entre 42 et 61 garçons.

2) On lance un dé 200 fois et dans les résultats on constate que l'on a obtenu 60 fois la face 6. Que peut-on dire de cette série de lancers ?

Solution

En théorie la proportion de « face 6 » est P = $\frac{1}{6}$ donc un intervalle de fluctuation au seuil de 95 % correspondant à cet échantillon est [$\frac{1}{6} - \frac{1}{\sqrt{200}}$; $\frac{1}{6} + \frac{1}{\sqrt{200}}$] cet intervalle est inclus dans l'intervalle [0,096 ; 0,238]

Donc en théorie on devrait avoir entre 0,096 x 200 ≈ 20 et 0,238 x 200 ≈ 47 fois la « face 6 » dans un tel échantillon.

On peut donc affirmer que cet échantillon n'est pas représentatif au seuil de 95 %, autrement dit on pourrait affirmer que le dé ne semble pas bien équilibré.

III) Lien avec la loi binomiale. Intervalle de fluctuation

En classe de première, le tirage au hasard d'un individu dans une population qui peut présenter un caractère C avec une probabilité p est assimilable à une épreuve de Bernoulli de paramètre p où le succès S est « avoir le caractère C ».

Le prélèvement d'un échantillon de taille n, dans cette population s'assimile alors à un schéma de Bernoulli de paramètres n et p et la variable aléatoire X qui compte le nombre de succès suit la loi binomiale $\mathcal{B}$ (n , p).

La variable aléatoire $F = \frac{X}{n}$ représente alors la fréquence aléatoire du succès S sur un échantillon de taille n.

D'après le résultat de seconde, on a P($p - \frac{1}{\sqrt{n}} \leq F \leq p + \frac{1}{\sqrt{n}}$) ≥ 0,95, et on dit que [$p - \frac{1}{\sqrt{n}}$; $p + \frac{1}{\sqrt{n}}$] est un intervalle de fluctuation au seuil de 95 %

1) Définition

Soit X une variable aléatoire qui suit une loi binomiale $\mathcal{B}$ (n , p) et $F = \frac{X}{n}$ la variable aléatoire qui représente la fréquence aléatoire du succès.
Un intervalle de fluctuation de F au seuil de 95 % est un intervalle :

- **De la forme [$\frac{a}{n}$; $\frac{b}{n}$] où a et b sont des entiers entre 0 et n ;**
- **Tel que P($\frac{a}{n} \leq F \leq \frac{b}{n}$) ≥ 0,95 ce qui équivaut à P($a \leq X \leq b$) ≥ 0,95.**

Remarque:

En pratique : on s'efforce d'obtenir l'intervalle [$\frac{a}{n}$; $\frac{b}{n}$] de plus faible amplitude, pour cela, il suffit de chercher les plus petits entiers a et b tels que :

P($X \leq a$) ≥ 0,025 et P($X \leq b$) ≥ 0,975 ce qui implique bien P($a \leq X \leq b$) ≥ 0,95

Exemple :

On considère un paquet de cartes contenant 3 cœurs et 7 piques, on effectue 100 tirages d'une carte en remettant à chaque fois la carte dans le paquet. A l'aide du tableur, on veut déterminer un intervalle de fluctuation au seuil de 95 % de la fréquence d'une carte de cœur dans l'échantillon prélevé.

Solution :

Le nombre X de cartes de cœur suit la loi binomiale $\mathcal{B}$ (100 ; 0,3)

La fréquence de « cartes de cœur » est donné par la variable aléatoire $F = \frac{X}{100}$

On cherche deux entiers a et b tels que P($\frac{a}{100} \leq F \leq \frac{b}{100}$) ≥ 0,95 autrement dit tels que P($a \leq X \leq b$) ≥ 0,95

k	P(X=k)	P(X<=k)
1	1,3862E-14	1,3862E-14
2	2,94073E-13	3,07935E-13
3	4,11703E-12	4,42496E-12
4	4,27877E-11	4,72126E-11
5	3,52081E-10	3,99294E-10
6	2,38912E-09	2,78842E-09
7	1,37497E-08	1,65381E-08
8	6,85027E-08	8,50408E-08
9	3,00107E-07	3,85148E-07
10	1,17042E-06	1,55557E-06
11	4,10406E-06	5,65963E-06
12	1,30451E-05	1,87047E-05
13	3,7845E-05	5,65497E-05
14	0,000100791	0,000157341
15	0,000247659	0,000405
16	0,000563866	0,000968865
17	0,001194068	0,002162933
18	0,002359706	0,004522639
19	0,004364569	0,008887208
20	0,007575645	0,016462853
21	0,0123684	0,028831253
22	0,019034486	0,047865739
23	0,027665029	0,075530767
24	0,038039414	0,113570182
25	0,049559923	0,163130104
26	0,061269135	0,22439924
27	0,071966921	0,296366161
28	0,080412019	0,376778179
29	0,085561557	0,462339736
30	0,086783865	0,549123601
31	0,083984385	0,633107986
32	0,07761057	0,710718556
33	0,068539205	0,779257761
34	0,05788395	0,837141712
35	0,046779682	0,883921394
36	0,036198564	0,920119958
37	0,026834457	0,946954414
38	0,019066588	0,966021002
39	0,012990422	0,979011424
40	0,008490169	0,987501593
….	….	….
….	….	…..

Ci-contre une copie d'écran du tableur avec les valeurs prise par la variable X et les valeurs des probabilités cumulées P($X \leq k$)

La partie ombrée montre que :

- P($X \leq 21$) ≥ 0,025
- P($X \leq 39$) ≥0,975

Donc P ($21 \leq X \leq 39$) ≥ 0,975

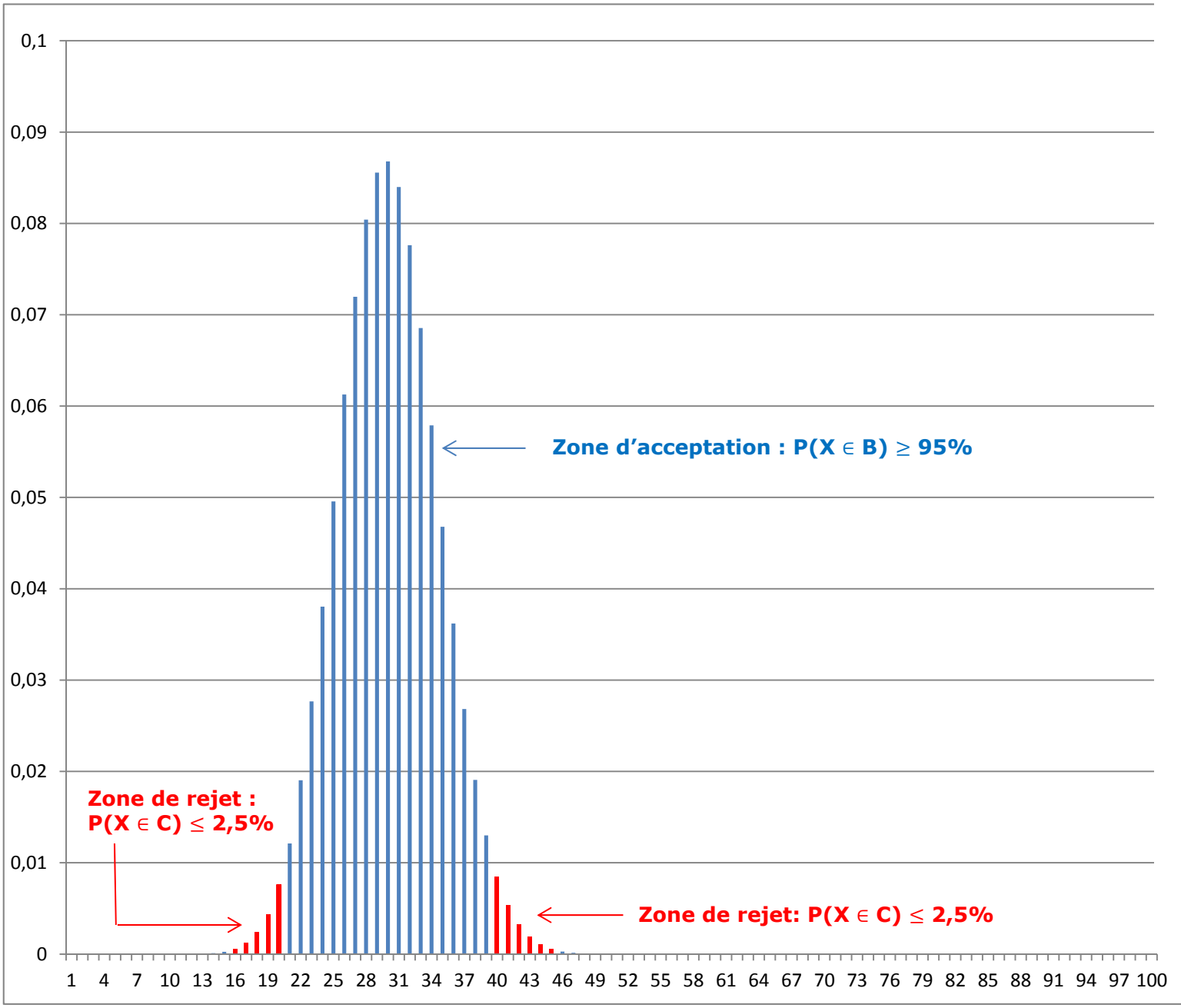
Ou encore P($0,21 \leq F \leq 0,39$) ≥ 0,95

Donc [0,21 ; 0,39] est un intervalle de fluctuation au seuil de 95 % de « carte de cœur »

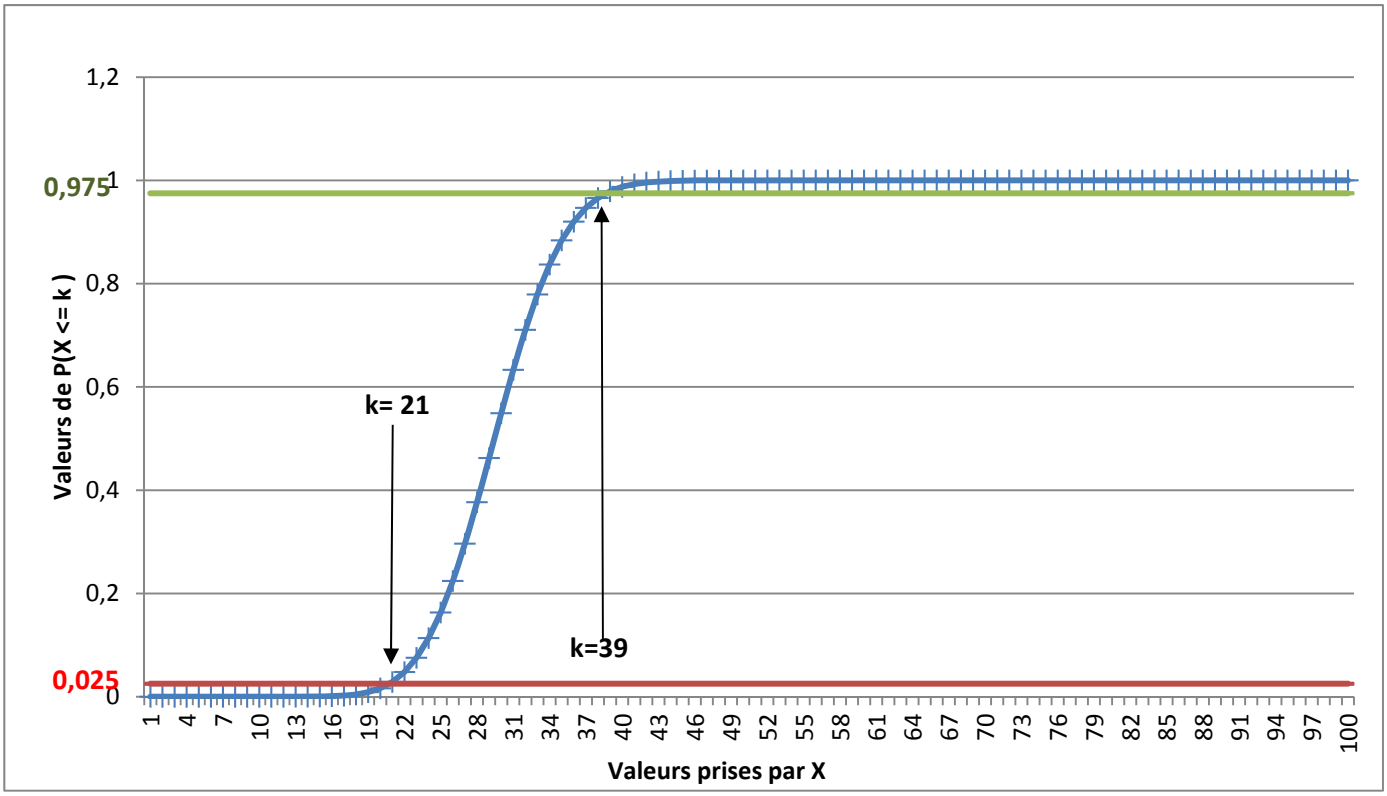
Le risque de voir la fréquence de « carte de cœur » sortir de cet intervalle est inférieur à 5 %

Un échantillon de 100 cartes contenant moins de 21 cartes de cœur ou plus de 39 cartes de cœur ne serait pas représentatif au seuil de 95%.

L'histogramme ci-dessous illustre en bleu la zone d'acceptation au seuil de 95% et en rouge la zone de rejet :



Le graphique ci-dessous représente les probabilités cumulées P($X \leq k$) en fonction de k.
Il illustre le choix de l'intervalle de fluctuation.



2) Détermination d'un intervalle de fluctuation à l'aide de la loi binomiale. Comparaison avec celui donné en seconde

Exemple :

Sur l'exemple précédent on a n = 100 et p = 0,3

- L'intervalle donné en seconde est I = [$p - \frac{1}{\sqrt{n}}$; $p + \frac{1}{\sqrt{n}}$] soit I = [0,2 ; 0,4]
- L'intervalle trouvé ci-dessus est J = [0,21 ; 0,39]

■ On peut constater que les deux intervalles obtenus sont très voisins

■ On peut aussi constater que J est inclus dans I

Ceci conforte le résultat vu en seconde plus simple à utiliser mais plus approximatif.

3) Prise de décision à partir d'un échantillon

Exemple

On cherche à savoir si une pièce est bien équilibrée.

■ On fait l'hypothèse que la pièce est bien équilibrée donc la probabilité d'obtenir PILE est p = 0,5

■ On lance n fois cette pièce et on détermine la fréquence f de PILE obtenue

■ On se fixe le seuil 95 % et on détermine l'intervalle de fluctuation à l'aide de la loi

$\mathcal{B}$ (n ; 0,5)

■ On prend une décision :

• Si f n'est pas dans l'intervalle de fluctuation, on rejette l'hypothèse que la pièce est bien équilibrée avec un risque de se tromper de 5 %

• Si f est dans l'intervalle de fluctuation, on ne rejette pas l'hypothèse que la pièce est bien équilibrée (on ne dit pas que on accepte cette hypothèse car le risque de se tromper en l'acceptant est inconnu).

Remarque :

C'est avec ce type de prise de décision (mais avec des méthodes beaucoup plus compliquées et beaucoup plus précises) qu'on détermine l'efficacité de certains médicaments ou les effets secondaires de ces médicaments.