
ÉCHANTILLONNAGE

Complément au cours du livre p. 205 et suivantes

 Les parties décorées sont à rendre pour le **samedi 9 mai**.

1. Introduction

Notre vie quotidienne est soumise à des décisions portant sur des valeurs chiffrées (actuellement, le nombre de personnes contaminées influe sur les décisions prises par le gouvernement).

MAIS il est bon de comprendre ce que signifient les chiffres obtenus à l'aide d'un sondage d'une partie de la population : c'est l'objet de ce chapitre.

2. Échantillon - fluctuation

Généralement on ne peut pas observer chaque individu d'une population, on étudie donc un **échantillon** de taille n .

Chaque individu de l'échantillon est choisi **au hasard** dans la population.

exemple : On veut tester si un couple de dés sont équilibrés. Pour cela, on s'intéresse à la somme des points des faces supérieures et on observe la réalisation de l'événement « la somme est supérieure ou égale à 7 ».

On simule 100 lancers de dés à l'aide d'un tableur et on observe la fréquence de l'événement : « la somme est supérieure ou égale à 7 ».



1. Créer une feuille tableur comme le modèle ci-contre (il faudra la joindre à votre mail).
2. Le premier lancer des dés est en ligne 7, on veut simuler 100 lancers : jusqu'à quelle ligne faudra-t-il copier les formules ? ligne **106**
3. En cellule **A8**, écrire une formule qui permet d'obtenir le numéro du lancer (les nombres dans la colonne **A** vont de 1 en 1) et qui pourra être recopier vers le bas. **=A7 + 1**
4. Dans les cellules **B7** et **C7** écrire la formule
=ENT(ALEA() * 6) + 1
qui permet de simuler un lancer de dé.
5. Dans la cellule **D7** écrire la formule qui permet de calculer la somme des points obtenus en **B7** et **C7**. **=B7 + C7**
6. recopier vers le bas, autant de fois que nécessaire les formules précédentes.
7. la cellule **D4** contient une formule qui donne la fréquence d'apparition des sommes supérieures ou égales à 7.
Déterminer parmi les formules suivantes, celle qui permet d'obtenir cette fréquence et expliquer pourquoi les autres ne conviennent pas :
 - (a) **=NB.SI(A7:A106; ">=7") / 100** non : la colonne **A** compte le nombre de lancers.
 - (b) **=NB.SI(D7:D106; ">=7")** non : on veut une fréquence, il faut diviser par le nombre de lancers.

	A	B	C	D
1	somme de dés			
2				
3				
4	fréquence somme ≥ 7			
5				
6	n° du lancer	Dé 1	Dé 2	somme
7		1	3	4
8				7
9				
10				

- (c) =NB.SI(D7:D106; ">=7") / 100 OUI
 (d) =NB.SI(D7:D106; ">7") / 100 non : on peut être égal à 7
 (e) =NB.SI(D7:D106; ">=7") / 100 non : les ; signifient « et » et non « jusqu'à ».

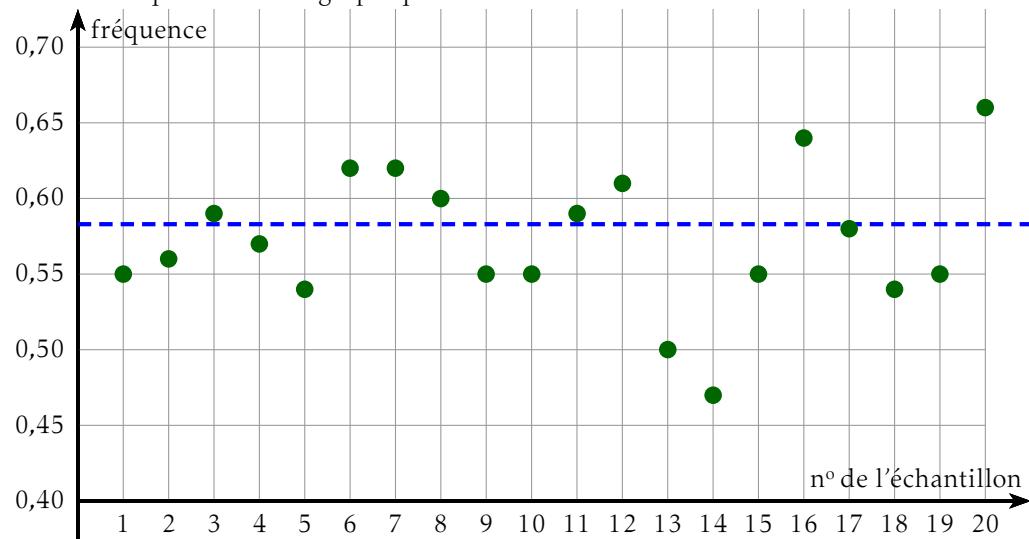
8. l'appui sur la touche F9 du clavier permet de simuler une autre centaine de lancers de dés.

20 appuis permettent donc de créer 20 échantillons de 100 lancers : écrire les fréquences obtenues.

...

9. de mon côté, j'ai obtenu les fréquences suivantes : 0,55 ; 0,56 ; 0,59 ; 0,57 ; 0,54 ; 0,62 ; 0,62 ; 0,6 ; 0,55 ; 0,55 ; 0,59 ; 0,61 ; 0,5 ; 0,47 ; 0,55 ; 0,64 ; 0,58 ; 0,54 ; 0,55 ; 0,66 ; que j'ai représentés dans un graphique.

Placer vos fréquences sur le graphique.



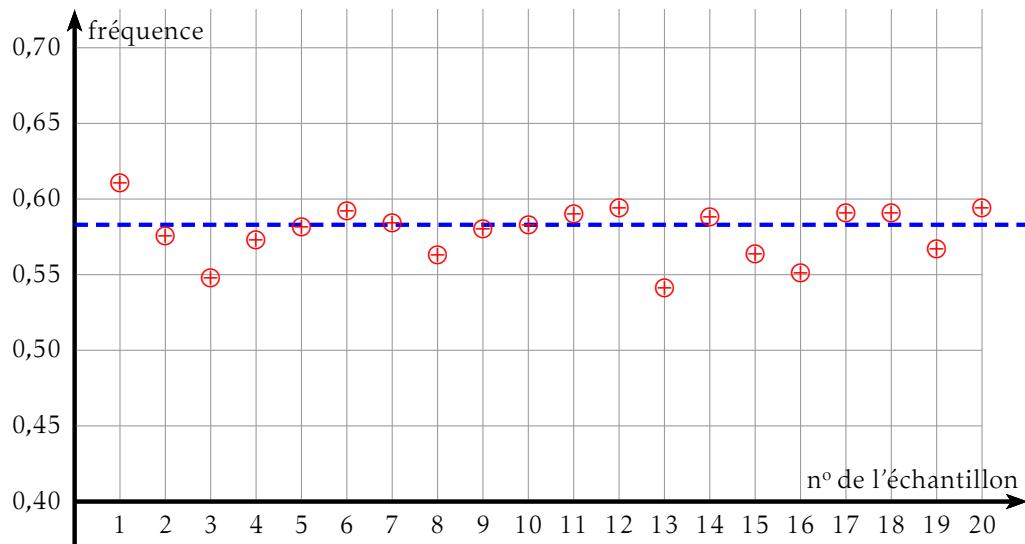
10. Sur l'ensemble de ces 40 valeur, déterminer la plus petite fréquence obtenue et la plus grande.

À retenir

Deux échantillons de même taille, issus de la même expérience, ne sont généralement pas identiques.

On appelle **fluctuation d'échantillonnage** les variations de fréquences observées.

Le graphique suivant représente les fréquences obtenues pour la répétition de vingt expériences de 1 000 lancers (c'est à dire des échantillons de taille $n = 1\ 000$).



Cette expérience est particulière, car on peut calculer la *probabilité* (ou *fréquence théorique*) d'obtenir une somme supérieure ou égale à 7.

11. Compléter le tableau ci-contre qui donne la somme des points obtenus par le dé bleu et le dé rouge.
12. Calculer la probabilité p d'obtenir une somme supérieure ou égale à 7 (arrondir à 10^{-3}). $p = \frac{21}{36} \simeq 0,583$
13. Sur chacun des graphiques précédents, tracer la droite (parallèle à l'axe des abscisses) d'équation $y = p$.

À retenir

Plus la taille de l'échantillon augmente, plus les fréquences observées se rapprochent de la fréquence théorique (ou de la probabilité)



•	2	3	4	5	6	7
•	3	4	5	6	7	8
••	4	5	6	7	8	9
••	5	6	7	8	9	10
•••	6	7	8	9	10	11
••••	7	8	9	10	11	12

3. Prise de décision - Intervalle de fluctuation

3.1 Théorie

On émet une hypothèse sur la valeur de la proportion p du caractère étudié. On considère donc p comme connu.

Un échantillon de taille n est prélevé dans la population et on observe une fréquence f du caractère étudié.

En reprenant l'exemple précédent, on n'effectue **qu'une série** de 100 lancers (si on le fait à la main, avec des vrais dés, cela prend du temps!).

Pour un sondage réel, on n'a ni le temps, ni l'argent de répéter les expériences.

Question : Peut-on, à partir de la fréquence observée f , valider la conjecture (l'hypothèse) faite sur p ?

À retenir

Propriété : Soit p la proportion effective d'un caractère dans une population telle que $p \in [0,2; 0,8]$.

f est la fréquence observée du caractère dans un échantillon de taille $n \geq 25$ de cette population.

Alors peut montrer que f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité de 95 %.

Cet intervalle est l'**intervalle de fluctuation de f au seuil de 95 %.**



On veut tester si deux dés sont bien équilibrés. Pour cela on étudie la probabilité d'obtenir une somme supérieure ou égale à 7.

14. En théorie, il faut trouver une fréquence proche de $p = \dots$
0,583 (la valeur trouvée à la question 12).

f est la première valeur que vous avez trouvée à l'aide du tableau, c'est donc

n est le nombre de lancers effectués pour obtenir f , c'est donc
.... 100

L'intervalle de fluctuation est $I_F = \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right] = \dots$

$$\left[0,583 - \frac{1}{\sqrt{100}} ; 0,583 + \frac{1}{\sqrt{100}} \right] = [0,483 ; 0,683]$$

Si $f \in I_F$, alors on accepte l'hypothèse que les dés sont bien équilibrés, sinon on rejette cette hypothèse.

Remarquez que dans les deux cas il y a un risque de se tromper : on n'est pas certain à 100 % du résultat !

3.2 Une histoire vraie

Entre 1999 et 2003, 132 enfants sont nés dans la réserve indienne d'Aamjiwnaag, située au Canada. Sur ces 132 naissances, il y a eu 46 garçons.



On pose la question : « parmi 132 naissances, le fait que 46 d'entre elles correspondent à des garçons est dans la norme. »

15. À l'aide de l'information donnée dans la quatrième colonne de l'article déterminer la probabilité qu'un nouveau né soit un garçon, c'est à dire donner une valeur approchée à 10^{-3} du rapport $p = \frac{\text{nombre de garçons}}{\text{nombre de naissances}}$.

on lit 105 garçons pour 100 filles, donc $p = \frac{105}{105 + 100} \approx 0,512$

16. Justifier qu'une valeur approchée de la fréquence observée est $f \approx 0,348$.

La fréquence observée est $f = \frac{46}{132} \approx 0,348$

17. Déterminer les bornes de l'intervalle de fluctuation au seuil de 95 %.

$$\text{IF} = \left[0,5 - \frac{1}{\sqrt{132}} ; 0,5 + \frac{1}{\sqrt{132}} \right] \approx [0,412 ; 0,587].$$

18. À la question « Le nombre de naissances de garçons est-il dans la norme ? » que permet de répondre l'étude mathématique ? Quelle est la réponse que donnent les études de Santé (colonne 2 de l'article) ?

$f \notin \text{IF}$. On rejette l'hypothèse parmi 132 naissances, le fait que 46 d'entre elles correspondent à des garçons est dans la norme. »

<https://www.aamjiwnaang.ca/>
<https://aamjiwnaangsolidarity.org/>
<https://www.theglobeandmail.com/life/the-mystery-of-the-missing-boys/article20395694/>

4. Estimation

4.1 Théorie

À retenir

On cherche une information sur une population de taille $n \geq 25$ à partir de l'étude d'un échantillon.

Si f est la fréquence observée ($f \in [0,2 ; 0,8]$) alors on montre que l'intervalle de confiance au seuil de 95 % pour la proportion réelle p du caractère dans la population est donnée par :

$$I_C = \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right].$$

4.2 Applications

4.2.1 Une vidéo d'explication

<https://www.youtube.com/watch?v=2Dcv7UbOVNs>

4.2.2 Élections du début du siècle

Avant les élections présidentielles de mai 2002, les différents instituts de sondages publient les résultats suivants : (résultats des 13-16 mars 2002 réalisés par la SOFRES pour *le Nouvel Observateur*).

Candidat	intentions de votes en %
Lionel Jospin (Gauche)	21
Jacques Chirac (Droite)	23,5
Jean-Marie Le Pen (FN)	10

Sources : <http://www.france-politique.fr/sondages-electoraux-presidentielle-2002.htm>

À l'issue du premier tour, Jean-Marie Le Pen obtient 16,8% des voix, alors que Lionel Jospin n'en obtient que 16,2%...

Les sondages mentent-ils ? (Pour certains cette élection est particulière car depuis un des candidats d'extrême droite est présent au second tour des élections présidentielles...)



On cherche à savoir si « les sondages mentent », c'est à dire : « la proportion réelle $p = 16,2\%$ correspondant aux votes pour Lionel Jospin appartient-elle à l'intervalle de confiance ? ».

N'ayant pas d'information sur la taille de l'échantillon, on considère que $n = 100$.

19. Déterminer f la fréquence observée des intentions de votes pour Lionel Jospin.

La fréquence observée est $f = 0,21$

20. Déterminer les bornes de l'intervalle de confiance dans ce cas.

$$\text{IC} = \left[0,21 - \frac{1}{\sqrt{100}} ; 0,21 + \frac{1}{\sqrt{100}} \right] \approx [0,11 ; 0,31]$$

21. Conclure.

Comme $16,2 \in \text{IC}$ on ne peut pas affirmer que les sondages se sont trompés !

Correction des copies



BO.KY : 16/30 : Bon début. Dommage qu'il manque les deux dernières parties.

Cours 2. Tableur : 2. réponse incohérente avec les formules choisies ensuite...

7. il ne peut pas y avoir deux formules qui donnent un résultat différent !

12. Tbien

Cours 3. Fluctuation :

Cours 4. Estimations :



KO.DA : 16/30 : Bon début. Dommage qu'il manque les deux dernières parties.

Cours 2. Tableur :

Cours 3. Fluctuation :

Cours 4. Estimations :



KR.DI : 22/30 : Très bon travail. Revoir partie 3 du cours.

Cours 2. Tableur :

Cours 3. Fluctuation : 14. $n = 100$ ou $n = 20$? Attention notation
[...]

17. revoir l'intervalle

18. le statisticien constate que la fréquence observée n'est pas dans l'intervalle de fluctuation ; les interprétations sont faites par d'autres...

Cours 4. Estimations : 20. Attention notation [...]

21. j'ai un doute sur ta conclusion.



OM.AN : 29/30 : Excellent travail! Félicitations! Bel effort de compression des fichiers! (il ne reste plus qu'à les nommer...)

Cours 2. Tableur :

Cours 3. Fluctuation : 18. le statisticien constate que la fréquence observée n'est pas dans l'intervalle de fluctuation ; les interprétations sont faites par d'autres...

Cours 4. Estimations : 21. donc les sondages mentent-ils ?