Échantillonnage - Estimation | 2nde

Introduction : Ce chapitre répond à deux types de questions :

♠ Exemple d'introduction 1. **Vu le sondage, le candidat sera-t-il élu ou non ?** (Oral)

Avant une élection, un sondage indique qu'un candidat a 56 % d'intentions de votes. Mais si un sondage sur 4 personnes ne nous dit pas la même chose que s'il s'agit d'un sondage sur 1000 personnes. Remarquons aussi que la proportion de votants favorables au candidats dépend de l'échantillon choisi. Combien de personnes faut-il interroger pour être sûr(e) qu'il sera élu ? Peut-on être sûr(e) à 100% ? non. Alors « sûr » dans quel sens ?

♠ Exemple d'introduction 2. **Discrimination ou non ?** (Oral)

Au vu de la proportion de femmes ou de noirs dans une entreprise, comment établir s'il y a discrimination ou non ? A partir de quelle valeur la proportion est-elle anormalement faible et révèle une discrimination ? Peut-on être sûr(e) à 100% ? non. Alors « sûr » dans quel sens ?

Table des matières

I. Modélisation de la situation

Définition [1].
- Un ***échantillon*** est une partie de la population étudiée. Les individus de l'échantillon sont choisis au hasard pour obtenir un échantillon représentatif de l'ensemble de la population.
- Le nombre d'individus formant l'échantillon s'appelle la ***taille de l'échantillon***.

♠ Exemple d'introduction 3. Oral, différencier P et f.

- intentions de votes pour un candidat donné
- Pourcentage d'obèses? Critère = obèse ou non.
- Pourcentage de malades guéris par un nouveau médicament? Critère = guéris ou non.

Notation [2].
- On note p la proportion de la population (totale) vérifiant le critère étudié.
- On note f la proportion de l'échantillon vérifiant le critère étudié (f comme fréquence).

Remarque : Pour une population donnée, p est fixe mais si on fait plusieurs sondages, on trouve des valeurs de f légèrement différentes à chaque fois.

A quoi cela sert ? f permet souvent d'estimer p . Pour avoir une approximation raisonnablement précise, il faut que l'échantillon soit suffisamment grand, voir [5].

Théorème de stabilisation des fréquences [3].
Plus la taille de l'échantillon est grande, plus f se rapproche de p .

II. Intervalle de fluctuation, intervalle de confiance

Ce sont en fait deux versions du même théorème (et voilà pourquoi les conditions d'applications sont semblables),

▪ l'une [4] servant (1) à dire si la proportion observée f semble compatible avec une populations dans laquelle la proportion est p et (2) à prévoir l'étendue des fluctuations entre divers échantillons d'une même population;

▪ l'autre [5] servant à évaluer p connaissant f, avec une précision d'autant meilleure que l'échantillon est grand.

A. Intervalle de fluctuation: Quelles fluctuations peut-on attendre entre divers échantillons tirés d'une même population ?

On l'utilise quand p , **la proportion dans l'ensemble de la population, est connue** ; l'hypothèse à vérifier pour utiliser le théorème porte sur p et on obtient un encadrement de f, la proportion observée dans l'échantillon.

théorème de l'intervalle de fluctuation au seuil de 95% [4].

si
- la taille n de l'échantillon est supérieure ou égale à 25 ($n \geqslant 25$) et si
- la proportion dans la population vérifie $0{,}2 \leqslant p \leqslant 0{,}8$

alors, dans ces conditions, dans plus de 95% des cas, la fréquence observée f vérifie

$$f \in \left[p - \frac{1}{\sqrt{n}} ;\ p + \frac{1}{\sqrt{n}} \right] \text{ càd } p - \frac{1}{\sqrt{n}} \leqslant f \leqslant p + \frac{1}{\sqrt{n}} .$$

Intervalle de fluctuations au seuil de 95% :Il existe des fluctuations entre échantillons, mais, pour 95% au moins des échantillons, la proportion observée se trouve dans cet intervalle.

♠ Application 4. Si les conditions d'application du théorème sont vérifiées mais que pourtant la **proportion observée f n'est PAS dans l'intervalle de fluctuation**, cela indique que soit dans la populations initiale la proportion n'etait pas égale à p soit que l'échantillon n'a pas été choisi au hasard soit que l'on est tombé sur un des rares[1] échantillons pour lesquels cela se produit.

Exemple: Discrimination ou non? Pratique: On calcule (ou on connait) une valeur théorique pour p et et on regarde si la proportion observée f est dans l'intervalle de fluctuation. Si les échantillons sont reéllement choisis au hasard, donc en l'absence de discrimination, pour au moins 95% des échantillons la fréquence abservées est dans l'intervalle de fluctuation. Si elle ne l'est pas, cela indique qu'il y a probablement[2] discrimination.

Remarque : $p - \frac{1}{\sqrt{n}} \leqslant f \leqslant p + \frac{1}{\sqrt{n}} \Leftrightarrow f - \frac{1}{\sqrt{n}} \leqslant p \leqslant f + \frac{1}{\sqrt{n}}$.

B. Intervalle de confiance: Pour évaluer la proportion dans l'ensemble de la population connaissant la proportion observée dans l'échantillon

On l'utilise quand p , **la proportion dans l'ensemble de la population, n'est PAS connue**; l'hypothèse à vérifier pour utiliser le théorème porte sur f, la proportion observée dans l'échantillon (*et pas sur p. Heureusement! Vu que p est inconnu, ce serait impossible à vérifier!*) et on obtient un encadrement de p.

théorème de l'intervalle de confiance au seuil de 95% [5].

si
- la taille n de l'échantillon est supérieure ou égale à 25 ($n \geqslant 25$) et
- la proportion dans l'échantillon vérifie $0{,}2 \leqslant f \leqslant 0{,}8$

alors, dans ces conditions, dans plus de 95% des cas, la proportion dans la population (totale) vérifie $p \in \left[f - \frac{1}{\sqrt{n}} ;\ f + \frac{1}{\sqrt{n}} \right]$ càd $f - \frac{1}{\sqrt{n}} \leqslant p \leqslant f + \frac{1}{\sqrt{n}}$.

Intervalle de confiance au niveau 95%: On a confiance que p se trouve dans cet intervalle pour au moins 95% des échantillons observés.

♠ Application 5. Estimation de p d'après les resultats d'un sondage.

Rappel sur les méthodes de travail : Travail quotidien = Reprendre ce qui a été fait en classe (en faisant des **restitutions** jusqu'à savoir retrouver **sans aucune aide** les définitions et les propriétés vues ou révisées au cours de la séance et refaire **sans aucune aide** les exercices faits en classe au cours de la séance.)

1 *« rares » car avec le choix du seuil de 95%, il y a au plus 5% des échantillons pour lesquels f n'est pas dans l'intervalle de fluctuations.*

2 *« Probablement » car on travaille au seuil de 95% et pas 100%.*

Illustration de la notion d'intervalle de fluctuation via une simulation

Cette simulation vise à illustrer le théorème suivant :

théorème de l'intervalle de fluctuation au seuil de 95%

si

- la taille n de l'échantillon est supérieure ou égale à 25 ($n \geqslant 25$) et si
- la proportion dans la population vérifie $0{,}2 \leqslant p \leqslant 0{,}8$

alors, dans ces conditions,

dans plus de 95% des cas, $f \in \left[p - \frac{1}{\sqrt{n}} ;\ p + \frac{1}{\sqrt{n}} \right]$ càd $p - \frac{1}{\sqrt{n}} \leqslant f \leqslant p + \frac{1}{\sqrt{n}}$.

Intervalle de fluctuations au seuil de 95% :Il existe des fluctuations entre échantillons, mais dans au moins 95% des échantillons, la proportion observée se trouve dans cet intervalle.

Dans cette formule,

- f est la *fréquence observée* : Par exemple, si on lance 100 fois une pièce, et que pile apparaît 59 fois, on dit que la *fréquence d'apparition de « pile »* est $f = \frac{59}{100} = 0{,}59$.
- p est la *fréquence théorique*. Lorsqu'on étudie un caractère dans une population c'est la proportion de la population présentant ce caractère et dans le cas du lancer d'une pièce, c'est la *probabilité*. Lors du lancer d'une pièce, nous avons deux possibilités : pile ou face. La probabilité d'avoir pile est de une chance sur deux d'où $p = \frac{1}{2} = 0{,}5$
- n est le *nombre de répétitions de l'expérience*. Par exemple, si on lance 100 fois la pièce, $n = 100$. Cette formule nous indique que si on lance 100 fois une pièce non truquée (c'est-à-dire pour laquelle on exactement autant de chance d'obtenir « pile » que « face »), l'intervalle de fluctuation à 95% a pour bornes : $\begin{cases} p - \frac{1}{\sqrt{n}} = 0{,}5 - \frac{1}{\sqrt{100}} = 0{,}5 - 0{,}1 = 0{,}4 \\ p + \frac{1}{\sqrt{n}} = 0{,}5 + \frac{1}{\sqrt{100}} = 0{,}5 + 0{,}1 = 0{,}6 \end{cases}$

■ Partie 1. Programmation de la simulation par une calculatrice du lancer d'un dé

1) Avant de programmer, familiarisons-nous avec les nombres aléatoires et la modélisation du problème.

Pour comprendre la situation, commençons par un cas simple sans programmation :

(1) Votre machine peut générer des nombres aléatoires : Rand (TI) ou Ran# (Casio) est un nombre aléatoire compris entre 0 et 1 (Plus précisément, $0 \leqslant Rand < 1$.) [voir Manuel Repères 2nde, page 167]

(2) On choisit une règle qui indique si le nombre obtenu correspond à « pile » ou à « face » : Par exemple, on peut dire que si Rand (ou Ran#) < 0.5, la pièce est tombée sur « pile ».

(3) Lancez 4 fois votre pièce virtuelle (= demandez 4 nombres aléatoires à votre machine), calculez à la main la fréquence d'apparition de pile.

2) Écrire un programme qui simule le lancer d'une pièce.

Vous allez réaliser un programme sur votre calculatrice qui affiche 4 nombres aléatoires et vous calculerez à la main la fréquence d'apparition de pile.

3) Écrire un programme qui simule le lancer d'une pièce 25 fois de suite et qui calcule la fréquence d'apparition de « pile » : Améliorez le programme précédent pour qu'il calcule (et affiche à la fin !) la fréquence d'apparition de « pile ».

4) A quoi s'attend-on pour la fréquence observée lorsqu'on augmente le nombre de lancers ?. . . . .

. . . .

■ **Partie 2. Fluctuations de la fréquence d'apparition de « PILE » quand on réalise plusieurs simulations**

1) Avec 25 lancers : Grâce à votre programme, vous allez alors simuler le lancer de 25 pièces à plusieurs reprises et noter la fréquence d'apparition de pile dans le tableau :

Série de 25 lancers	1	2	3	4	5	6	7	8	9	10
f										

Quel est le pourcentage de ces valeurs comprises dans l'intervalle $\left[0{,}5-\frac{1}{\sqrt{25}};0{,}5-+\frac{1}{\sqrt{25}}\right]$?

On commencera par calculer $0{,}5-\frac{1}{\sqrt{25}}$ et $0{,}5+\frac{1}{\sqrt{25}}$, puis on complétera la phrase suivante :

Pour ...... échantillons sur les 10 c'est à dire ...... % des échantillons, la fréquence observée est dans l'intervalle ....

2) Avec 100 lancers : Vous allez alors simuler le lancer de 100 pièces à plusieurs reprises et noter la fréquence d'apparition de pile.

Série de 100 lancers	1	2	3	4	5	6	7	8	9	10
f										

Quel est le pourcentage de ces valeurs comprises dans l'intervalle $\left[0{,}5-\frac{1}{\sqrt{100}};0{,}5-+\frac{1}{\sqrt{100}}\right]$?

Réponse : Pour ...... échantillons sur les 10 c'est à dire ...... % des échantillons, la fréquence observée est dans l'intervalle ....

■ **Partie 3. Fluctuations de la fréquence d'un événement quand on lance un dé équilibré**

Écrire un programme qui simule le lancer d'un dé.
Vous étudierez la fréquence d'apparition de la face « 5 » ou de la face « 6 ». Donnez vos conclusions.

Coupe de pouce : La fonction « partie entière », notée Int sur vos machines retourne le plus petit entier inférieur ou égal au nombre choisi. Ainsi, Int(3)=3 ; Int(21,8)=21 ; Int(3,5)=3 ; Int(87,25)=87...etc.

TI 89 : int(1+6*Rand()) ; Casio fx-9860G : Int(1+6*Ran#)

Résumé de cours

Échantillonnage ***p connue*** Utilisation = Prise de décision à partir d'une échantillon : On regarde si la valeur observée pour f est « raisonnable » *(Pour voir s'il y a discrimination contre les femmes dans une entreprise, pour voir si un jury a vraiment été choisi au hasard...etc)*	**Estimation** ***p inconnue*,** Utilisation = Estimation : On estime p au moyen de f. *(pour estimer la proportion p de gens qui vont voter pour un candidat à partir des résultats d'un sondage...etc)*
Si $n\geqslant 25$ et $0{,}2\leqslant p\leqslant 0{,}8$ alors pour au moins 95% des échantillons, la fréquence **observée** vérifie $f\in\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$ ↿ *Intervalle de **fluctuation** au seuil de 95%*	Si $n\geqslant 25$ et $0{,}2\leqslant f\leqslant 0{,}8$ alors pour au moins 95% des échantillons, la fréquence **théorique** vérifie $p\in\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ ↿ *Intervalle de **confiance*** au seuil de 95%

Découverte de la notion d'intervalle de fluctuation via une simulation

■ Partie 1. Programmation de la simulation par une calculatrice du lancer d'un dé

1) Avant de programmer, familiarisons-nous avec les nombres aléatoires et la modélisation du problème.

Pour comprendre la situation, commençons par un cas simple sans programmation :

(4) Votre machine peut générer des nombres aléatoires : Rand (TI) ou Ran# (Casio) est un nombre aléatoire compris entre 0 et 1 (Plus précisément, $0 \leqslant Rand < 1$.)

(5) On choisit une règle qui indique si le nombre obtenu correspond à « pile » ou à « face » : Par exemple, on peut dire que si Rand (ou Ran#) < 0.5, la pièce est tombée sur « pile ».

(6) Lancez 4 fois votre pièce virtuelle (= demandez 4 nombres aléatoires à votre machine), calculez à la main la fréquence d'apparition de pile.

2) Écrire un programme qui simule le lancer d'une pièce.

Vous allez réaliser un programme sur votre calculatrice qui affiche 4 nombres aléatoires et vous calculerez à la main la fréquence d'apparition de pile.

3) Écrire un programme qui simule le lancer d'une pièce 25 fois de suite et qui calcule la fréquence d'apparition de « pile » : Améliorez le programme précédent pour qu'il calcule (et affiche à la fin !) la fréquence d'apparition de « pile ».

4) A quoi s'attend-on pour la fréquence observée lorsqu'on augmente le nombre de lancers ?. . . . .
. . . .

■ Partie 2. Fluctuations de la fréquence d'apparition de « PILE » quand on réalise plusieurs simulations

1) Avec 25 lancers : Grâce à votre programme, vous allez alors simuler le lancer de 25 pièces à plusieurs reprises et noter la fréquence d'apparition de pile dans le tableau :

Série de 25 lancers	1	2	3	4	5	6	7	8	9	10
f										

Quel est le pourcentage de ces valeurs comprises dans l'intervalle $\left[0{,}5-\frac{1}{\sqrt{25}} ; 0{,}5+\frac{1}{\sqrt{25}}\right]$?

On commencera par calculer $0{,}5-\frac{1}{\sqrt{25}}$ et $0{,}5+\frac{1}{\sqrt{25}}$, puis on complétera la phrase suivante :

Pour . . . . . . échantillons sur les 10 c'est à dire . . . . . . % des échantillons, la fréquence observée est dans l'intervalle . . . .

2) Avec 100 lancers : Vous allez alors simuler le lancer de 100 pièces à plusieurs reprises et noter la fréquence d'apparition de pile.

Série de 100 lancers	1	2	3	4	5	6	7	8	9	10
f										

Quel est le pourcentage de ces valeurs comprises dans l'intervalle $\left[0{,}5-\frac{1}{\sqrt{100}} ; 0{,}5+\frac{1}{\sqrt{100}}\right]$?

Réponse : Pour . . . . . . échantillons sur les 10 c'est à dire . . . . . . % des échantillons, la fréquence observée est dans l'intervalle . . . .

■ Partie 3. Fluctuations de la fréquence d'un événement quand on lance un dé équilibré

Écrire un programme qui simule le lancer d'un dé.

Vous étudierez la fréquence d'apparition de la face « 5 » ou de la face « 6 ». Donnez vos conclusions.

Coupe de pouce : La fonction « partie entière », notée Int sur vos machines retourne le plus petit entier inférieur ou égal au nombre choisi. Ainsi, Int(3)=3 ; Int(21,8)=21 ; Int(3,5)=3 ; Int(87,25)=87...etc.

TI : int(1+6*Rand()) ou RanInt(1,6) ou EntAlea(1,6) chercher dans le catalogue (touche « catalog »);
Casio fx-9860G : Int(1+6*Ran#), chercher dans le catalogue (touche « catalog »).