Copyright © Tous droits réservés
Ce document, ou toute partie de celui-ci, ne peut être reproduit ou utilisé de quelque manière que ce soit sans l'autorisation écrite expresse du propriétaire.

Échantillonnage

Capacités attendues en fin de chapitre :

- Observer la loi des grands nombres à l'aide d'une simulation sur Python ou tableur
- Simuler N échantillons de taille n d'une expérience aléatoire à deux issues.
- Donner un intervalle de fluctuation ou intervalle de confiance.

Le mathématicien du chapitre :

Georges Gallup (1901-1984) est un statisticien et un sociologue américain. Il fonde l'American Institute of public opinion qui organise des sondages basés sur des échantillons représentatifs de la population pour prédire les opinions.
L'échantillonnage se résume assez simplement par : « Le hasard est capricieux, mais seulement au coup par coup »

Il est parfois impossible ou trop coûteux de recueillir des données sur l'ensemble d'une population. On étudie alors un ***échantillon*** de cette population à l'aide d'un sondage.

1) Modélisation de la situation.

- Lorsqu'on étudie une partie de la population, on dit qu'on étudie un échantillon de cette population
- Le nombre d'individus formant l'échantillon est appelé l'effectif relatif, noté n.

Notation : On note p la proportion de la population vérifiant le critère étudié et f_{obs} la fréquence observée de l'échantillon vérifiant ce critère.

Théorème de la stabilisation des fréquences :

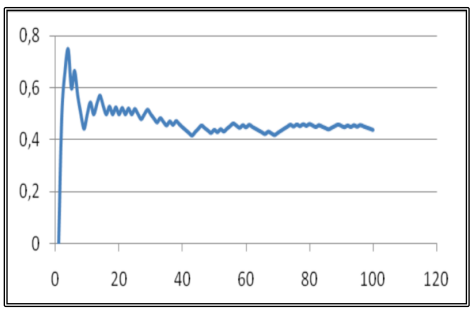
Plus la taille de l'échantillon est grande, plus la fréquence observée f_{obs} se rapproche de la proportion p.

Exemples :

- Un enfant jette un dé à six faces deux fois de suite. Il obtient deux fois de suite le chiffre 1. Peut-on affirmer pour autant qu'à chaque fois qu'il va jeter ce dé, il tombera sur le 1 ?
- On interroge à la sortie du lycée les 10 premiers élèves qui sortent à 16 h 30. Aucun ne se déclare fumeur… Peut-on pour autant affirmer qu'aucun élève ne fume au Michel Platini ????

Le problème de ces deux exemples précédents provient de la taille de l'échantillon étudié. Impossible d'avoir des certitudes avec un effectif relatif réduit.

- Pour un sondage, on sait que la proportion de personnes ayant répondu « oui » était de 0,4. On connait donc $p = 0{,}4$. Voici une représentation des valeurs de f_{obs} en fonction de la taille de l'échantillon. On constate que plus la taille de l'échantillon est grande, plus f_{obs} se rapproche de p.



Copyright © Tous droits réservés
Ce document, ou toute partie de celui-ci, ne peut être reproduit ou utilisé de quelque manière que ce soit sans l'autorisation écrite expresse du propriétaire.

2) Estimation d'une fréquence inconnue à partir de la population

Dans cette partie, la proportion d'un certain caractère dans une population est ***connue***.

Théorème de l'intervalle de fluctuation : Certaines conditions sont indispensables :

La taille n de l'échantillon doit être supérieure à 25, p doit appartenir à l'intervalle $[0{,}2;\ 0{,}8]$
Alors, dans ces conditions, dans plus de 95 % des cas,
f_{obs} appartient à l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$

Remarque :

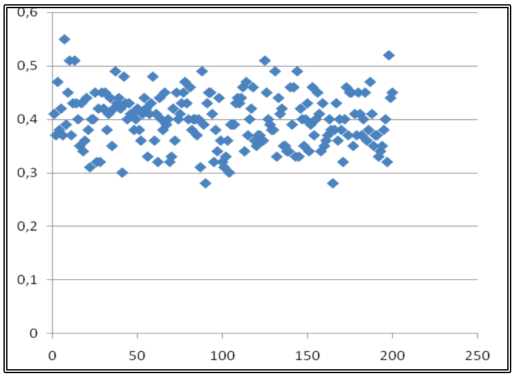
Avec les conditions à respecter, on ne peut jamais donner un intervalle de fluctuation lorsqu'on jette un dé. En effet, la proportion d'obtenir le chiffre 3 vaut $p=\frac{1}{6}$ et n'entre pas dans l'intervalle $[0{,}2;\ 0{,}8]$

Exemple :

Pour le sondage précédent $p=0{,}4$, on a effectué 200 fois un sondage sur 100 personnes (donc $n=100$). On a représenté ici les 200 valeurs de f_{obs} obtenues.

$$p-\frac{1}{\sqrt{n}}=0{,}4-\frac{1}{\sqrt{100}}=0{,}3$$
$$p+\frac{1}{\sqrt{n}}=0{,}4+\frac{1}{\sqrt{100}}=0{,}5$$

On obtient donc : $I_F=[0{,}3;\ 0{,}5]$



Il doit donc y avoir au moins $\frac{95}{100}\times 200=190$ valeurs comprises entre 0,3 et 0,5.

Attention :

On arrondit toujours la borne inférieure ***par défaut*** et la borne supérieure ***par excès***, quel que soit le nombre derrière. Sauf demande contraire, les bornes sont données au millième.

Prise de décision :

Dans une population d'effectif N, on suppose que la proportion d'un certain caractère est p. On souhaite juger cette hypothèse (valider ou rejeter). Pour cela, on prélève dans la population, au hasard et avec remise, un échantillon de taille n sur lequel on observe la fréquence f_{obs} de ce caractère.

Règle de décision :

- Si $f_{obs}\in I_F$, on ***accepte*** l'hypothèse que la proportion est p au seuil de 95%.
- Si $f_{obs}\notin I_F$, on ***rejette*** cette hypothèse au seuil de risque de 5%.

Exercice :

On souhaite savoir si une entreprise exerce une discrimination à l'embauche vis-à-vis du personnel féminin. S'il n'y a pas de discrimination, la proportion de femmes dans cette entreprise devrait être représentative de la proportion de femmes dans la population active. On admet que la proportion de femmes dans la population active est 0,5.

1. En utilisant l'intervalle de fluctuation au seuil 0,95, déterminer si une entreprise contenant 1235 femmes sur 2540 salariés exerce une discrimination à l'égard des femmes.
2. Quel doit être le nombre minimal d'employés dans cette entreprise pour que la proportion f_{obs} de femmes appartienne à l'intervalle de fluctuation $[0{,}49;\ 0{,}51]$?

Copyright © Tous droits réservés
Ce document, ou toute partie de celui-ci, ne peut être reproduit ou utilisé de quelque manière que ce soit sans l'autorisation écrite expresse du propriétaire.

Solution :

1) On évalue d'abord si les conditions sont bien respectées : $n = 2540$ et $p = 0{,}5$
On calcule la fréquence observée : $f_{obs} = \frac{1235}{2540} \approx 0{,}486$.L'intervalle de fluctuation est : $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right] = \left[0{,}5 - \frac{1}{\sqrt{2540}}; 0{,}5 + \frac{1}{\sqrt{2540}}\right] \approx [0{,}48; 0{,}52]$ On accepte donc l'hypothèse initiale qu'il y a 50 % de femme dans l'entreprise.
Donc il n'y a pas de discrimination à l'embauche.
2) On cherche donc la valeur de *n* tel que $\frac{1}{\sqrt{n}} = 0{,}01$ soit donc $n = 10000$

Exemple :

L'institut national de prévention et d'éducation pour la santé (INPES) a étudié la proportion de fumeurs quotidiens de tabac en France en 2010. Chez les 15-19 ans, 26% des garçons et 20% des filles sont des fumeurs quotidiens.
Sur un échantillon de 1000 lillois de cette tranche d'âge, dont 450 sont des filles, on a dénombré 178 fumeurs quotidiens chez les garçons et 98 chez les filles.
On fait l'hypothèse que dans la ville de Lille, la proportion de fumeurs quotidiens chez les 15-19 ans est de 26% pour les garçons et de 20% pour les filles.
Répondre aux questions suivantes pour les garçons puis pour les filles.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la fréquence de fumeurs quotidiens chez les 15-19 ans (on n'oubliera pas de vérifier les conditions d'application de la règle de prise de décision).

2. Pouvez-vous considérer, au seuil de risque de 5%, que la fréquence observée de jeunes fumeurs quotidiens lillois dans cet échantillon est en accord avec la proportion de jeunes fumeurs quotidiens de la population française ?

Solution :

Pour les garçons :

On vérifie d'abord que les conditions d'application sont vérifiées :$n = 550, p = 0{,}26$
On peut ainsi définir l'intervalle de fluctuation asymptotique au seuil 0,95 dans un échantillon de taille $n = 550$: $\left[0{,}26 - \frac{1}{\sqrt{550}}; 0{,}26 + \frac{1}{\sqrt{550}}\right] \approx [0{,}217; 0{,}303]$
La fréquence observée, fréquence des garçons fumeurs dans l'échantillon considéré, est égale à : $f_{obs} = \frac{178}{550} \approx 0{,}3236$
Comme cette fréquence n'appartient pas à l'intervalle de fluctuation asymptotique au seuil 0,95, l'hypothèse selon laquelle 26 % des garçons sont des fumeurs est rejetée.

Pour les filles :

On vérifie d'abord que les conditions d'application sont vérifiées : $n = 450, p = 0{,}2$
On peut ainsi définir l'intervalle de fluctuation asymptotique au seuil 0,95 dans un échantillon de taille $n = 450$: $\left[0{,}2 - \frac{1}{\sqrt{450}}; 0{,}2 + \frac{1}{\sqrt{450}}\right] \approx [0{,}152; 0{,}248]$
La fréquence observée, fréquence des filles fumeurs dans l'échantillon considéré, est égale à $f_{obs} = \frac{98}{450} \approx 0{,}218$
Comme cette fréquence appartient à l'intervalle de fluctuation asymptotique au seuil 0,95, l'hypothèse selon laquelle 20 % des filles sont des fumeurs est acceptée.

3) *Estimation d'une proportion inconnue à partir d'un échantillon*

Dans cette partie, la proportion d'un certain caractère dans une population est ***inconnue***.

Propriété :

Les deux propriétés suivantes sont équivalentes :

$$(P_1) f_{obs} \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right] \Leftrightarrow (P_2) p \in \left[f_{obs} - \frac{1}{\sqrt{n}}; f_{obs} + \frac{1}{\sqrt{n}}\right]$$

Copyright © Tous droits réservés
Ce document, ou toute partie de celui-ci, ne peut être reproduit ou utilisé de quelque manière que ce soit sans l'autorisation écrite expresse du propriétaire.

Démonstration :

Il suffit de remarquer que $p \in \left[f_{obs} - \frac{1}{\sqrt{n}}; f_{obs} + \frac{1}{\sqrt{n}}\right]$si et seulement si la distance entre p et f_{obs} est inférieure ou égale à $\frac{1}{\sqrt{n}}$, c'est-à-dire si et seulement si $f_{obs} \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$

Théorème de l'intervalle de confiance : Certaines conditions sont indispensables :

- La taille n de l'échantillon doit être supérieure à 25,
- f_{obs} doit appartenir à l'intervalle $[0{,}2; 0{,}8]$

Alors, dans ces conditions, $I_C = \left[f_{obs} - \frac{1}{\sqrt{n}}; f_{obs} + \frac{1}{\sqrt{n}}\right]$ contient p avec une probabilité d'au moins 0,95.

Remarques :

- Le niveau de confiance 0,95 nous indique simplement que dans plus de 95 cas sur 100, on énonce à juste titre que la proportion inconnue p appartient à l'intervalle de confiance considéré. Cependant, on ne peut faire aucun pronostic sur une localisation possible de cette proportion p dans l'intervalle. En particulier, ***la proportion inconnue n'est pas nécessairement le centre de l'intervalle***.
- A chaque tirage d'un échantillon, on obtient un I_C différent.

Exemple :

Après avoir examiné 100 poissons pêchés dans un lac, on a constaté que 20% d'entre eux étaient malades. Par quel intervalle de confiance au niveau de confiance de 95% peut-on estimer la proportion de poissons malades dans le lac ?
On estime la population de poissons dans le lac suffisamment importante pour assimiler le prélèvement des 100 poissons à des tirages successifs avec remise.

Solution :

On a donc : $n = 100$ et $f_{obs} = 0{,}2$; On vérifie que les conditions sont vérifiées :
On a donc : $I_C = \left[0{,}2 - \frac{1}{\sqrt{100}}; 0{,}2 + \frac{1}{\sqrt{100}}\right] = [0{,}1; 0{,}3]$
$I_C$est un intervalle de confiance de la proportion des poissons malades du lac au niveau de confiance de 95%.

Exercice :

Une urne contient des boules rouges et des boules noires. On cherche à estimer la proportion p (inconnue) de boules rouges de l'urne à l'aide de tirages successifs avec remise. Combien de tirages suffit-il d'effectuer pour déterminer un encadrement de p d'amplitude 0,1 au niveau de confiance de 95% ?

Solution :

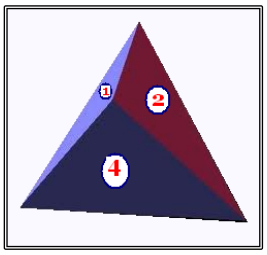
On rappelle l'expression de l'intervalle de confiance : $I_C = \left[f_{obs} - \frac{1}{\sqrt{n}}; f_{obs} + \frac{1}{\sqrt{n}}\right]$
L'amplitude de l'intervalle étant $\frac{2}{\sqrt{n}}$, je cherche n afin que : $\frac{2}{\sqrt{n}} \leq 0{,}1$soit alors $n \geq 400$.
Ainsi, à partir de 400 tirages, on aura un encadrement de la proportion des boules rouges de l'urne avec une précision de 0,1 avec un niveau de confiance de 0,95.

Copyright © Tous droits réservés
Ce document, ou toute partie de celui-ci, ne peut être reproduit ou utilisé de quelque manière que ce soit sans l'autorisation écrite expresse du propriétaire.

4) Simulation

1) Dé tétraédrique

On souhaite réaliser la simulation d'un jet de dé tétraédrique à l'aide de la TI. A l'aide des résultats des élèves, on déterminera si les f_{obs} sont dans l'intervalle de fluctuation à 95%



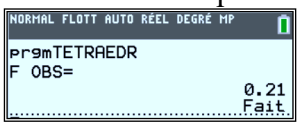
On rappelle qu'un dé tétraédrique est composé de 4 faces identiques ayant la forme d'un triangle équilatéral. Chaque face a donc la même probabilité d'apparaître lors d'un lancer.
Si on s'intéresse à la face numérotée 4, on a donc : $p = 0{,}25$
On va procéder à la simulation de 100 jetés de dés par chaque élève de la classe. On a donc $n = 100$
On obtient alors $I_F = \left[0{,}25 - \frac{1}{\sqrt{100}}; 0{,}25 + \frac{1}{\sqrt{100}}\right] = [0{,}15; 0{,}35]$

Remarquons que les conditions pour obtenir un intervalle de fluctuation sont bien vérifiées.

On construit alors sur la TI le programme ci-contre :
On initialise d'abord les listes au cas où elles aient été utilisées auparavant.

On simule le tirage d'une face à l'aide `5:nbrAléatEnt(`

On obtient alors pour chaque élève de la classe une f_{obs}



Sur les …. élèves de la classe, il y en a ….. qui ont une f_{obs} en dehors de l'intervalle de fluctuation soit un pourcentage de $\frac{\cdots}{\cdots} \times = \cdots$ ce qui est cohérent avec les 95 %

```
NORMAL FLOTT AUTO RÉEL DEGRÉ MP
PROGRAM:TETRAEDR
:EffListe L1,L2
:100→dim(L1)
:For(I,1,100)
:nbrAléatEnt(1,4)→L1(I)
:End
:0→X
:For(J,1,100)
:If L1(J)=4
:Then
:X+1→X
:End
:End
:X/100→F
:Disp "F OBS=",F
```

2) Pile ou face

A l'aide du programme ci-dessous, chaque élève réalise la simulation de 50 échantillons de taille 200 d'un lancer de pièce. Ici, $n = 200$ et $p = 0{,}5$. Les conditions étant vérifiées, on obtient l'intervalle de fluctuation à 95 % : $I_F = \left[0{,}5 - \frac{1}{\sqrt{200}}; 0{,}5 + \frac{1}{\sqrt{200}}\right] \approx [0{,}429; 0{,}571]$

La touche nbrAleat permet d'obtenir au hasard un nombre de l'intervalle [0;1]

On peut faire apparaître les 50 valeurs des f_{obs} en traçant le graphique. On trace les droites $y = 0{,}429$ et $y = 0{,}571$

On observe sur le graphique que seules 2 f_{obs} sont en dehors de l'intervalle de fluctuation soit donc 4 %

```
NORMAL FLOTT AUTO RÉEL RAD MP
PROGRAM:PILEFACE
:EffListe L1,L2
:For(J,1,50)
:0→X
:For(I,1,200)
:If NbrAléat<0.5
:Then
:X+1→X
:End
:End
:J→L1(J)
:X/200→L2(J)
:End
```