ECHANTILLONNAGE

Le principe :

On considère par exemple l'expérience suivante consistant à lancer plusieurs fois un dé et à noter si la face supérieure affichée est un 4 ou un autre nombre.

La valeur supposée et théorique de la probabilité d'obtenir un 4 est $\frac{1}{6}$.

La mise en défaut ou non de cette expérience, nous permettra d'affirmer s'il est raisonnable de penser que le dé est pipé ou ne l'est pas.

En réalisant l'expérience un certain nombre de fois (échantillon), on mesure la fréquence d'apparition du 4. Si la fréquence et la valeur théorique sont trop "éloignées" (dépassent un seuil fixé) alors on peut rejeter la valeur théorique et considérer que le dé est pipé. Dans le cas inverse, on considère qu'il ne l'est pas.

I. Notion d'échantillon

Exemple :
Si, sur l'ensemble des cartes à puce produites par une entreprise en une semaine, on en prélève 200, on dit que cet ensemble de 200 cartes à puce constitue un **échantillon de taille 200** de la population de toutes les cartes à puce produites en une semaine.

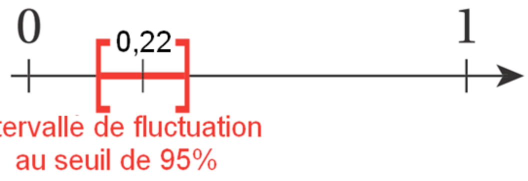
Définition :
Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience sur l'ensemble des personnes ou objets sur lesquels porte l'étude statistique (la population).
Un échantillon issu d'une population est donc l'ensemble de quelques éléments de cette population.

II. Intervalle de fluctuation

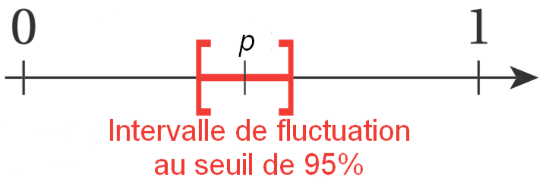
On suppose que 22% des cartes à puce produites par l'entreprise sont défectueuses.
La **proportion théorique** p est donc égale à 22%.
On prélève un échantillon de taille 200 parmi cette production et on compte le nombre de cartes à puce défectueuses parmi cet échantillon. Ce nombre est égal à 41.

Dans ce cas, la **fréquence observée** f est égale à $\frac{41}{200} = 0,205$.

Pour un échantillon de taille 200, l'**intervalle de fluctuation** de la fréquence p des cartes à puce défectueuses au seuil de 95 %, est un intervalle de centre 0,22 tel que les fréquences observées se trouvent dans cet intervalle pour 95 % des échantillons de taille 200.



Définition :
L'intervalle de fluctuation au seuil de 95% d'une fréquence d'un échantillon de taille n est l'intervalle centré autour de la proportion théorique p tel que la fréquence observée f se trouve dans l'intervalle avec une probabilité égale à 0,95.



Propriété :
Pour 0,2 < p < 0,8 et n > 25, l'intervalle de fluctuation au seuil de 95% de f est l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$.

Cela signifie qu'on a une probabilité de 0,95 pour que la fréquence observée se trouve dans l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$.

Remarque :
L'amplitude de cet intervalle est égale à $\frac{2}{\sqrt{n}}$.

Dans l'exemple précédent, l'intervalle de fluctuation au seuil de 95% de p = 0,22 est $\left[0,22-\frac{1}{\sqrt{200}};0,22+\frac{1}{\sqrt{200}}\right]$ soit de façon approchée [0,15 ; 0,29].

Méthode : Prendre une décision à partir d'un échantillon

Vidéo https://youtu.be/BIlBtFIVUAY

Deux entreprises A et B recrutent dans un bassin d'emploi où il y a autant de femmes que d'hommes, avec la contrainte du respect de la parité.
Dans l'entreprise A, il y a 100 employés dont 43 femmes (soit 43 %).
Dans l'entreprise B, il y a 2500 employés dont 1150 femmes (soit 46 %).

Or, 46 % est plus proche de 50 % que 43 % : les chiffres parlent d'eux-mêmes !
Si on admet que la parité, c'est exactement 50 % de femmes, il est vrai que B est plus proche que A.
Peut-on alors affirmer que l'entreprise B respecte mieux la parité que l'entreprise A ?
(D'après document ressource « Prob-stat » - Juin 2009)

La proportion théorique p est égale à 0,5 (50% de femmes).

Pour l'entreprise A :
La taille de l'échantillon n est égale à 100.
La fréquence observée f est égale à 0,43.

Pour l'entreprise B :
La taille de l'échantillon n est égale à 2500.
La fréquence observée f est égale à 0,46.

Pour chaque entreprise, peut-on affirmer que la fréquence de femmes respecte la parité ?
Pour y répondre, on va vérifier dans chaque cas si la fréquence observée f se situe dans l'intervalle de fluctuation au seuil de 95%.

Pour l'entreprise A :
L'intervalle de fluctuation au seuil de 95% de p = 0,5 est :

$$I_f = \left[0,5 - \frac{1}{\sqrt{100}} ; 0,5 + \frac{1}{\sqrt{100}}\right] = [0,4 ; 0,6] \quad \text{donc} \quad f = 0,43 \in I_f$$

Pour l'entreprise B :
L'intervalle de fluctuation au seuil de 95% de p = 0,5 est :

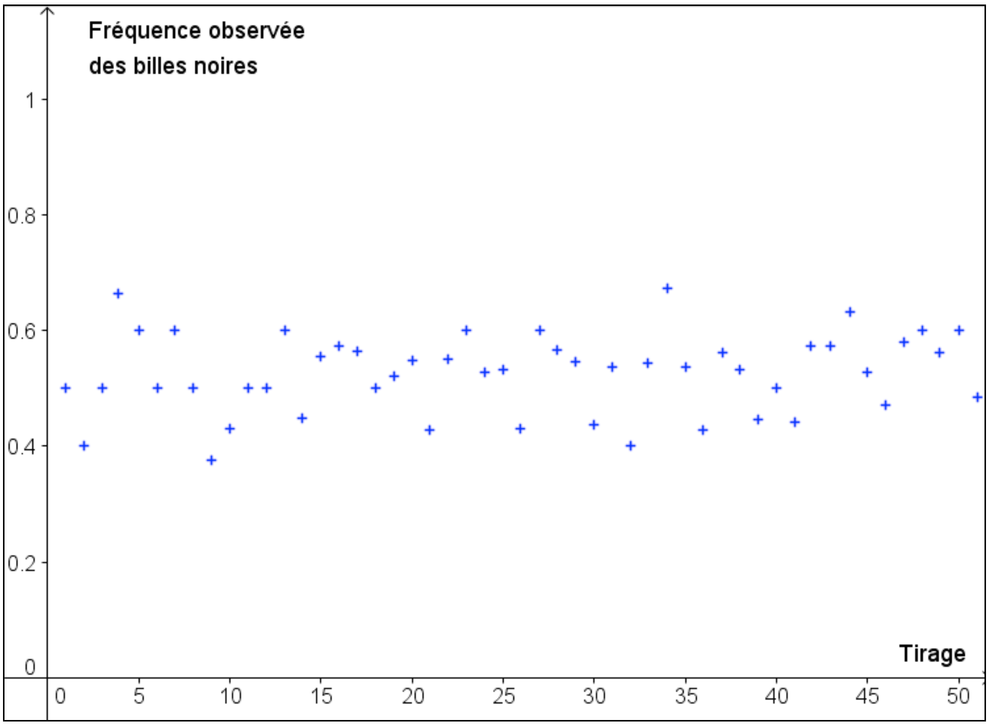
$$I_f = \left[0,5 - \frac{1}{\sqrt{2500}} ; 0,5 + \frac{1}{\sqrt{2500}}\right] = [0,48 ; 0,52] \quad \text{donc} \quad f = 0,46 \notin I_f .$$

La valeur 43% est donc dans l'intervalle de fluctuation de l'entreprise A alors que la valeur 46% n'est pas dans l'intervalle de fluctuation de l'entreprise B.
La proportion de 46% s'observe donc dans moins de 5% des échantillons de taille 2500.
On peut alors rejeter l'hypothèse que l'entreprise B respecte la parité.
Par contre, pour l'entreprise A, on peut accepter cette hypothèse.

III. Intervalle de confiance

Exemple :
Un jeu consiste à tirer 100 billes d'un sac contenant 300 billes noires et 300 billes blanches.
L'expérience peut être simulée avec un tableur afin d'effectuer rapidement un grand nombre de tirage.
Pour cet échantillon de taille 100, on compte le nombre de billes noires et on calcule la fréquence observée f.
On pourrait ainsi vérifier que, dans 95 % des cas, la fréquence des billes noires dans l'échantillon appartient à l'intervalle :
$\left[0{,}5-\frac{1}{\sqrt{100}};0{,}5+\frac{1}{\sqrt{100}}\right]$ soit : [0,4 ; 0,6] où p = 0,5 et n = 100.

NUAGE DE POINTS DES FREQUENCES OBSERVEES DES BILLES NOIRES POUR 50 TIRAGES EFFECTUES



Définition :
Soit p la proportion théorique tel que 0,2 < p < 0,8.
On considère la fréquence observée f pour un échantillon donné de taille n > 25.
L'intervalle $I_C=\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ est appelé un intervalle de confiance (ou fourchette de sondage) de p au niveau 0,95.

Propriété :
95 % des intervalles de confiance associés aux échantillons de taille n possibles ayant comme fréquence observée f contiennent la proportion théorique p.

Méthode : Estimer une proportion inconnue

Vidéo https://youtu.be/mo1Vb60lho8

1) Avant les élections, le candidat A commande un sondage effectué sur 250 personnes.
138 personnes interrogées déclarent avoir l'intention de voter pour le candidat A.
Le candidat A peut-il espérer être élu ?
2) Le candidat A commande un second sondage effectué sur 1000 personnes pour lequel 538 personnes déclarent avoir l'intention de voter pour lui.
Le candidat A peut-il espérer être élu ?

1) Soit p la proportion théorique d'électeurs pour le candidat A.

La fréquence observée est égale à $f = \frac{138}{250} = 0,552$

L'intervalle de confiance de p au seuil de 0,95 est :

$I_C = \left[0,552 - \frac{1}{\sqrt{250}} ; 0,552 + \frac{1}{\sqrt{250}}\right]$ soit de façon approchée [0,49 ; 0,62].

On a donc : 0,49 < p < 0,62. Il est donc possible que le candidat A ne soit pas élu.

2) La fréquence observée est égale à $f = \frac{538}{1000} = 0,538$

L'intervalle de confiance de p au seuil de 0,95 est :

$I_C = \left[0,538 - \frac{1}{\sqrt{1000}} ; 0,538 + \frac{1}{\sqrt{1000}}\right]$ soit de façon approchée [0,51 ; 0,57].

On a donc : 0,51 < p < 0,57.
La proportion théorique évaluée est supérieure à 50%. Le candidat A peut donc espérer être élu puisque 95% des échantillons possibles de taille 1000 seraient compris dans cet intervalle.

Exercice 1
Une urne contient des boules de différentes couleurs dont 75% de boules rouges.
Cyril tire une boule au hasard, note la couleur et la remet dans l'urne.
Il prétend avoir effectué cette expérience 60 fois et avoir obtenu 35 boules rouges.
Son frère Paulo affirme qu'il n'a pas fait l'expérience sérieusement.
On se propose de vérifier s'il a de bonnes raisons de l'affirmer.
1) Déterminer la proportion théorique p et la taille n de l'échantillon.
2) Calculer la fréquence observée f.
3) Calculer l'intervalle de fluctuation au seuil de 95% I_f.
4) Vérifier si la fréquence observée f appartient à l'intervalle de fluctuation I_f et conclure.

Exercice 2
La proportion de personnes aux cheveux châtains en France est d'environ 50%.
On a observé un échantillon de 150 personnes dont 89 ont les cheveux châtains.
Cet échantillon est-il représentatif de la population ?

Exercice 3
Dans une classe de 37 élèves, un délégué de classe a été élu avec 60% des voix.
Parmi les 17 filles, 11 d'entre elles ont voté pour ce délégué.
Les filles sont-elles représentatives des résultats des élections de délégués ?

Exercice 4
En 1976, dans un comté du Texas, Rodrigo Partida était condamné à huit ans de prison.
Il attaqua ce jugement au motif que la désignation des jurés de ce comté était discriminante à l'égard des Hispano-américains.
En effet, 79,1% de la population de ce comté était d'origine hispanique.
Alors que, sur les 870 personnes convoquées pour être juré, il n'y eut que 339 personnes d'origine hispanique. Qu'en pensez-vous ?
(D'après IREM de Paris Nord)

Exercice 5
Une maladie guérit naturellement dans 70% des cas. Un laboratoire souhaite tester l'efficacité d'un nouveau médicament.
Pour cela, on administre ce médicament à 500 personnes. Pour 77% d'entre elles, la guérison a eu lieu.
Que penser de l'efficacité de ce médicament ?

Exercice 6
Un centre commercial n'attire que 22% de clients hors de la communauté urbaine.
Souhaitant élargir sa clientèle, le centre commercial s'agrandit (nouveaux magasins, cinéma, restaurants, …)
Après les travaux, voulant connaître l'impact de ses investissements, 300 clients sont interrogés : 72 d'entre eux habitent hors de la communauté urbaine.
Peut-on affirmer que l'agrandissement a eu un impact sur la fréquentation des clients hors communauté urbaine.

Exercice 7
Un fournisseur d'accès à Internet disposait de 25% de part de marché avant l'arrivée d'un nouveau concurrent.
Après l'arrivée de ce concurrent, il effectue une enquête sur un échantillon de 200 foyers et obtient 19% de part de marché.
Peut-il considérer que l'arrivée de ce nouveau concurrent lui a fait perdre des parts de marché.

Exercice 8
Un musée national connaît une proportion de visiteurs français égale à 71%.
Durant l'été, le musée propose une exposition temporaire sur le thème de l'Egypte ancienne.
On souhaite connaître son impact sur la fréquentation des visiteurs français.
Décider si la nouvelle exposition a eu un impact dans les cas suivants :
1) Sur un échantillon de 50 visiteurs, 82% sont français.
2) Sur un échantillon de 500 visiteurs, 77% sont français.

Exercice 9
Léanne a lancé une pièce de monnaie 50 fois. Elle a obtenu 32 fois « pile ».
A-t-elle de bonnes raisons de s'étonner du résultat.

Exercice 10
Une urne contient un grand nombre des boules rouges et des boules blanches.
Sans les compter, on souhaiterait connaître la proportion p de boules rouges.
Pour cela, on effectue 100 tirages avec remise dans cette urne. On obtient 41 boules rouges.
Estimez p à l'aide de l'intervalle de confiance au niveau 0,95.

Exercice 11
On effectue un sondage auprès de 800 personnes pour leur demander leur intention de vote aux prochaines élections. 54% d'entre elles déclarent de façon ferme vouloir voter pour le candidat A.
Le candidat A peut-il espérer être élu ?

Exercice 12
Le conseil municipal d'une commune organise un sondage effectué au hasard sur 400 personnes pour leur demander s'ils sont favorables à de nouveaux investissements de voirie. 235 personnes donnent un avis favorables.
Donner un intervalle de confiance au seuil de 95% de la proportion p des personnes favorables aux investissements de voirie. Que peut conclure le conseil municipal ?

Exercice 13
Pour mesurer l'audience d'une émission télévisée (audimat), on installe des appareils électroniques chez certains téléspectateurs.
Une chaîne de télévision a mesuré à partir de 1000 appareils et a relevé une audience de 31% sur le créneau 19/20h.
Donner une fourchette de l'audimat p sur ce créneau au seuil de 95%.

Exercice 14
On veut estimer la proportion p de jeunes de 10 à 15 ans disposant d'un forfait mobile. On sait que p est compris entre 50% et 70%.
Quelle doit être la taille minimale de l'échantillon pour obtenir un résultat avec une précision de 1% au seuil de 0,95. Une telle enquête est-elle envisageable ?