Chapitre 10 : 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.

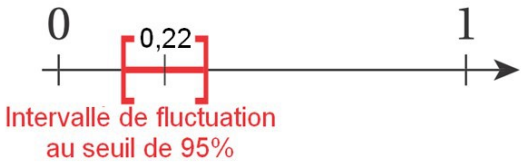
Définition :
Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience sur l'ensemble des personnes ou objets sur lesquels porte l'étude statistique (la population).
Un échantillon issu d'une population est donc l'ensemble de quelques éléments de cette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

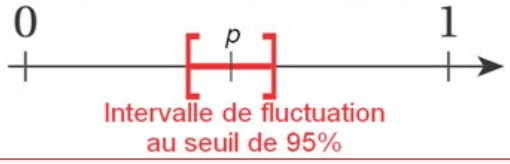
Dans ce cas, la **fréquence observée** f est $\frac{41}{200} = \frac{20,5}{100} = 20,5\% = 0,205$

Pour un échantillon de taille 200, l'**intervalle de fluctuation** de la fréquence f 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 la fréquence observée f est l'intervalle $I_{f_{95}} = \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$

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 $I_{f_{95}} = \left[0,22 - \frac{1}{\sqrt{200}} ; 0,22 + \frac{1}{\sqrt{200}} \right]$ soit de façon approchée $[14,9 ; 29,1]$

arrondi par défaut — arrondi par excès

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 de femmes dans une entreprise qui respecte la parité est $p = 0,5$

Entreprise A : $f_{obs} = 0,43$ sur un échantillon de taille $n = 100$

L'intervalle de fluctuation au seuil de 95% de la fréquence observée de femmes dans une entreprise de taille 100 qui respecte la parité est :

$$I_{f_{95}} = \left[0,5 - \frac{1}{\sqrt{100}} ; 0,5 + \frac{1}{\sqrt{100}}\right]$$

$$= [0,4 ; 0,6] \qquad \text{on a } 0,43 \in I_{f_{95}} \text{ donc}$$

l'entreprise A est considérée comme respectant la parité.

Entreprise B : $f_{obs} = 0,46$ pour un échantillon de taille $n = 2500$.

L'intervalle de fluctuation au seuil de 95% de la fréquence observée de femmes dans une entreprise de taille 2500 qui respecte la parité est :

$$I_{f_{95}} = \left[0,50 - \frac{1}{\sqrt{2500}} ; 0,50 + \frac{1}{\sqrt{2500}}\right]$$

$$= [0,48 ; 0,52] \qquad 0,46 \notin I_{f_{95}}$$

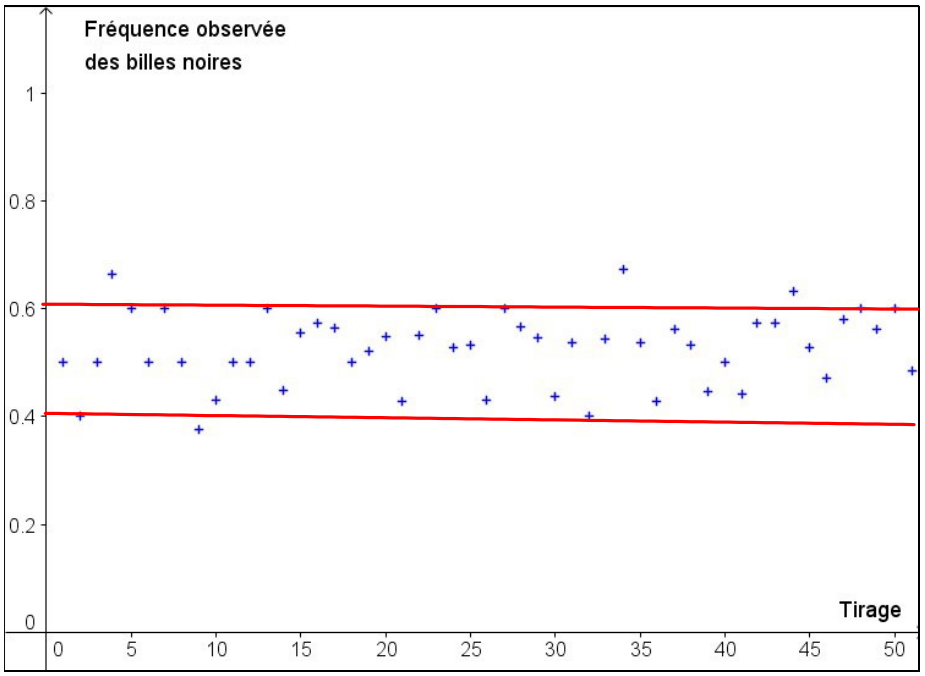
donc l'entreprise B est considérée comme ne respectant pas bien la parité.

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 $I_c = \left[f_{obs}-\frac{1}{\sqrt{n}};f_{obs}+\frac{1}{\sqrt{n}}\right]$ 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) On constitue un échantillon de taille $n = 250$ pour lequel on mesure une fréquence observée d'intentions favorables au candidat A $f_{obs} = \frac{138}{250} = 0{,}552$

Déterminons l'intervalle de confiance de la proportion p de votes favorables au candidat A.

$$I_{c_{95}} = \left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$$

$$I_{c_{95}} = \left[0{,}552 - \frac{1}{\sqrt{250}} ; 0{,}552 + \frac{1}{\sqrt{250}} \right] = [0{,}488 ; 0{,}616]$$

par défaut par excès

$50\% \in I_c$ donc le candidat A n'est pas sûr de gagner l'élection.

Exercice 7 + Exercice 10

2) Plus la taille de l'échantillon est importante moins l'amplitude de l'intervalle de confiance est importante : le sondage est donc plus fiable.

$I_c = \left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$ avec $n = 1000$ et $f_{obs} = \frac{538}{1000} = 0{,}538$

$I_c = [0{,}506 ; 0{,}567]$ L'intervalle de confiance au seuil de 95% met en évidence que le pourcentage d'opinions favorables au candidat dépasse 50%.
Cela lui donne de fortes chances d'être élu (avec un risque inférieur à 5% de se tromper en croyant victoire suite à ce sondage)