Echantillonnage et estimation : Séance 1

1 Intervalle de fluctuation et prise de décision

a. Définition de l'intervalle de fluctuation asymptotique

Problématique : Dans une certaine population, la proportion d'individus présentant le caractère de $\mathcal{C}$ est p. Que peut-on dire de la fréquence f du caractère $\mathcal{C}$ sur un échantillon aléatoire de taille n ?

La variable aléatoire X_n qui à un échantillon de taille n associe le nombre d'individus présentant le caractère $\mathcal{C}$ suit la loi binomiale de paramètres n et p. En effet le choix au hasard d'un échantillon est assimilé à un tirage avec remise. De plus en posant $F_n = \dfrac{X_n}{n}$, on définit la variable aléatoire « fréquence du caractère C dans l'échantillon ».

Remarque

Dans ce contexte la **proportion p est connue ou supposée connue**.

Théorème-Définition 1.

Soient X_n une variable aléatoire suivant la loi $\mathcal{B}(n,p)$, $F_n = \dfrac{X_n}{n}$ et Z la loi normale centrée réduite. Soit $\alpha \in]0,1[$ et u_α le réel tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$. Alors :

$$\lim_{n\to+\infty} P(F_n \in I_n) = 1-\alpha \text{ avec } I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

L'intervalle I_n est appelé intervalle de fluctuation asymptotique de la fréquence F_n au seuil de $1-\alpha$.

Démonstration *A traiter sur feuille annexe*

Remarques

- Ce théorème signifie que pour n grand la variable aléatoire F_n prend ses valeurs dans I_n avec une probabilité proche de $1-\alpha$ ou de manière équivalente qu'elle prend ses valeurs en dehors de I_n avec une probabilité proche de α.
- On dit indifféremment « au seuil de $1-\alpha$ » ou « au seuil de confiance $1-\alpha$ » ou « au niveau de confiance de $1-\alpha$ ». On peut aussi dire « au risque de α ».
- On admet que pour $n \geqslant 30$,$np \geqslant 5$ et $n(1-p) \geqslant 5$ on peut approcher $P(F_n \in I_n)$ par $1-\alpha$. On vérifiera donc que ces trois conditions sont bien remplies pour appliquer cette approximation.
- Si les trois conditions ne sont pas remplies on appliquera l'intervalle de fluctuation vu en première avec la loi binomiale (voir p 387 de votre livre).

Exemple

L'intervalle de fluctuation au seuil de 95% (A connaître)

On veut l'intervalle de fluctuation au seuil de 95%, on a donc $1-\alpha = 0,95$, donc $\alpha = 0,05$. Or on a vu en I 3d) qu'une valeur approchée de $u_{0,05}$ est 1,96. On prendra par conséquent :

$$I_n = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

Remarque

On peut montrer que l'intervalle ci-dessus est contenu dans l'intervalle de fluctuation au seuil de 95% vu en seconde :

$$\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$$

.

Travail en autonomie

Savoir-faire 1 et 2 p 359 + 17 p 365

b. Prise de décision

La prise de décision concerne une hypothèse que l'on fait sur la fréquence p d'un caractère $\mathcal{C}$ d'une population et se fait à partir d'un échantillon de taille n pour lequel la fréquence de $\mathcal{C}$ est f, en utilisant l'intervalle de fluctuation asymptotique I_n.

La procédure est la suivante : (prise de décision au seuil de 95%)

- On formule l'hypothèse : « la fréquence du caractère $\mathcal{C}$ dans la population est p ».
- On vérifie : $n \geqslant 30$,$np \geqslant 5$ et $n(1-p) \geqslant 5$.
- On calcule $I_n = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$] en arrondissant les bornes.
- On calcule la fréquence f observée sur l'échantillon de taille n.
- On applique la règle de décision au seuil de 95% : Si $f \in I$, l'hypothèse est acceptée, si $f \notin I$, l'hypothèse est rejetée.

Remarques

- On peut rencontrer parfois l'expression « décision au seuil de 5% » mais dans ce cas le seuil en question est le seuil de risque. Le risque devant être petit on comprendra selon le cas s'il s'agit du seuil de risque ou du seuil de confiance.
- Le théorème du a) prouve que l'on peut rejeter l'hypothèse à tort avec une probabilité de 0,05 (risque d'erreur 5%).
- D'autres seuils sont possibles et pratiqués en particulier le seuil de décision de 99%.

Exercice 1

Pour créer ses propres bijoux, on peut acheter un kit contenant des perles de cinq couleurs différentes, dans des proportions affichées sur le paquet. Ainsi, les perles de couleur marron sont annoncées comme représentant 20% de l'ensemble des perles. Les élèves d'une classe de Terminale ont voulu vérifier cette information. Pour cela, ils ont choisi d'observer un échantillon aléatoire de perles et de construire un intervalle de fluctuation asymptotique au seuil de 95% pour la proportion de perles marron. Ils ont donc constitué un échantillon, que l'on peut considérer aléatoire, de 690 perles. Ils ont dénombré 140 perles marron. La règle de décision est la suivante : si la proportion de perles marron dans l'échantillon n'appartient pas à l'intervalle de fluctuation, on rejette l'hypothèse selon laquelle les perles marron représentaient 20% des perles pendant la période où les kits utilisés pour l'expérience ont été produits.

1. Déterminer l'intervalle de fluctuation asymptotique J au seuil de 95% pour la proportion de perles marron.
2. Calculer la proportion de perles marron dans l'échantillon. Que peut-on en conclure ?
3. Dans le même échantillon, il y avait 152 perles jaunes pour une proportion annoncée de 20% et 125 perles rouges, pour une proportion annoncée de 10%. Que peut-on conclure de ces résultats ?

Exercice 2

Dans un casino, il a été décidé que les « machines à sous » doivent être réglées sur une fréquence de gain du joueur de $g = 0,06$. Une fréquence inférieure est supposée faire « fuir le client », et une fréquence supérieure est susceptible de ruiner le casino. Deux contrôleurs différents vérifient une même machine. Le premier a joué 120 fois et gagné 14 fois, le second a joué 400 fois et gagné 30 fois. En utilisant des intervalles de fluctuation asymptotiques au seuil 95%, examiner dans chaque cas la décision à prendre par le contrôleur, à savoir accepter ou rejeter l'hypothèse $g = 0,06$.

Travail en autonomie

Savoir-faire 3 et 4 p 361 + 29 p 366

2 Intervalle de confiance et estimation

a. Définition de l'intervalle de confiance

Problématique : Dans une certaine population, la fréquence d'individus présentant le caractère $\mathcal{C}$ sur un échantillon donné de taille n est f. Que peut-on dire de la proportion p de C dans la population ?

Remarque

Dans ce contexte la **proportion p est inconnue**.

Lemme 1.

Soit X_n une variable suivant la loi $\mathcal{B}(n,p)$ et $F_n = \dfrac{X_n}{n}$. Pour tout $p \in]0,1[$, il existe n_0 entier naturel tel que :

$$\text{Si } n \geqslant n_0, \text{ alors : } P\left(\left[p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0,95$$

Théorème 1.

Soit X_n une variable suivant la loi $\mathcal{B}(n,p)$ et $F_n = \dfrac{X_n}{n}$. Pour tout $p \in]0,1[$, il existe n_0 entier naturel tel que :

$$\text{Si } n \geqslant n_0, \text{ alors : } P\left(\left[F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0,95$$

Démonstration .....

.....

.....

Interprétation : F_n étant la fréquence observée sur un échantillon de taille n, si $n \geqslant n_0$, p est dans l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}, F_n + \dfrac{1}{\sqrt{n}}\right]$ avec un niveau de confiance de plus de 95%.

Définition 2.

Soit f la fréquence du caractère $\mathcal{C}$ sur un échantillon de taille n. L'intervalle $\left[f - \dfrac{1}{\sqrt{n}}, f + \dfrac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance à 95% de la proportion inconnue p dans la population.

Remarques

- On utilise cet intervalle dès que $n \geqslant 30$,$nf \geqslant 5$ et $n(1-f) \geqslant 5$.
- La précision de cet intervalle est égale à sa longueur à savoir $\dfrac{2}{\sqrt{n}}$.

Travail en autonomie

Savoir-faire 6 p 363

b. Estimation d'une proportion, utilisation de l'intervalle de confiance

Remarque

On estime la proportion p par un intervalle de confiance déterminé à partir de f et de n selon un niveau de confiance de 95%.

Exercice 3
Dans une urne contenant des boules rouges et bleues en proportions inconnues, on effectue des tirages au hasard avec remise.

1. Après avoir effectué 100 tirages, on compte 52 boules rouges et 48 boules bleues. Donner un intervalle de confiance à 95% de la proportion p de boules rouges dans l'urne.

2. Combien faudrait-il, au minimum, effectuer de tirages pour obtenir un intervalle de confiance à 95% de longueur inférieure ou égale à 2.10^{-2} (c'est-à-dire une précision d'au moins 0,02) ?

Exercice 4
Une usine fabrique des pièces métalliques, qui sont censées résister à certaines contraintes mécaniques. Le responsable de fabrication souhaite estimer le taux de pièces défectueuses concernant la résistance mécanique dans la production.
Pour cela, il utilise la méthode par intervalle de confiance au niveau 95%, en extrayant au hasard n pièces en fin de production, qui sont soumises à contrainte mécanique jusqu'à la rupture. En fonction du niveau de contrainte à la rupture, on décide de la nature défectueuse ou pas de la pièce.

1. Chaque pièce testée étant détruite, le responsable souhaite minorer la taille de l'échantillon testé, tout en ayant un intervalle de confiance de longueur inférieure à 0,1. Quelle taille d'échantillon peut-on lui conseiller ?

2. II est finalement décidé de mener l'étude sur 500 pièces ; on en trouve 40 défectueuses. Quel intervalle de confiance, au niveau de confiance 95%, obtient-on ? (Arrondir les bornes à 10^{-4})

3. L'année précédente, à l'issue d'un problème grave de rupture d'une pièce, une large étude avait débouché sur 130 pièces défectueuses dans un échantillon de 1 000. Peut-on supposer que la mise en place de nouvelles procédures de fabrication a vraiment diminué la proportion de pièces défectueuses ? (Arrondir les bornes à 10^{-4})

Exercice 5
Un usine produit des billes de trois couleurs différentes, rouges, vertes et bleues, qui sont mélangées. On prélève un échantillon de 1000 billes dans le stock et on compte 250 rouges. On recommence et on trouve 300 vertes, puis encore pour obtenir 350 bleues.

1. Peut- on affirmer à 95% qu'il y a plus de vertes que de rouges ?

2. Peut- on affirmer à 95% qu'il y a plus de bleues que de rouges ?

Exercice 6
Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour Le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée de 18 ans et plus et inscrite sur les listes électorales. On suppose cet échantillon constitué de manière aléatoire (même si en pratique cela n'est pas le cas). Les intentions de vote au premier tour pour les principaux candidats sont les suivantes : 20% pour J. Chirac, 18% pour L. Jospin et 14% pour J.-M. Le Pen. Les médias se préparent pour un second tour entre J. Chirac et L. Jospin.

1. Déterminer pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 de la proportion inconnue d'électeurs ayant l'intention de voter pour lui.

2. Le 21 avril, les résultats du premier tour des élections sont les suivantes : $19,88\%$ pour J.Chirac, $16,18\%$ pour L. Jospin et $16,86\%$ pour J.-M. Le Pen. Les pourcentages de voix recueillies par chaque candidat sont-ils bien dans les intervalles de confiance précédents ?

3. Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 0,95, l'un de ces trois candidats pour le second tour ?

Travail en autonomie
Savoir-faire 5 et 7 p 363