Chapitre 18 : Loi normale et estimation.

I-Rappels de seconde

1. Echantillons

Lorsqu'on travaille sur une population de grande taille, il est rarement possible d'avoir accès aux données relatives à l'ensemble de la population.

On utilise alors un échantillon de cette population.

Définition Un échantillon de taille n est une sélection de n individus choisis "au hasard" dans une population.

Exemple

On étudie la répartition mâle/femelle d'une population de truites peuplant une rivière.

Il est pratiquement impossible de recenser toutes les truites de la rivière. On décidera donc de travailler sur un échantillon en prélevant, par exemple, 100 truites.

La taille de l'échantillon doit être suffisamment élevée pour fournir des résultats fiables (mais pas trop pour ne pas entrainer un surcroit de travail important !)

2. Intervalle de fluctuation et intervalle de confiance

Si l'on effectue plusieurs échantillonnage de même taille sur une même population, on obtiendra en général des fréquences légèrement différentes pour un caractère donné.

Voici, par exemple, les résultats que l'on pourrait obtenir en prélevant 5 échantillons de 100 truites :

Echantillons	échantillon n°1	échantillon n°2	échantillon n°3	échantillon n°4	échantillon n°5
Pourcentage de truites femelles	52%	55%	42%	50%	48%

Ce phénomène s'appelle la fluctuation d'échantillonnage.

Le résultat suivant précise cette notion :

Théorème On note p la proportion d'un caractère dans une population donnée.

définition On prélève un échantillon de taille n de cette population et on note f la fréquence du caractère dans l'échantillon.

Si $0,2 \leq p \leq 0,8$ et si $n \geq 25$ alors, dans au moins 95% des cas, f appartient à l'intervalle : $\boxed{I=\left[p-\frac{1}{\sqrt{(n)}};p+\frac{1}{\sqrt{(n)}}\right]}$.

I est appelé l'intervalle de fluctuation au seuil 95%

Remarques

- On applique le théorème ci-dessus si on connaît la proportion p du caractère dans la population.
- Bien retenir la signification de chacune des variables :
  - p = proportion du caractère dans l'ensemble de la population
  - f = fréquence du caractère dans l'échantillon
  - n = taille de l'échantillon
- Au niveau Seconde, les intervalles de fluctuation seront toujours demandés au seuil de 95%.
  Ce seuil a été choisi car :
  - il conduit à une formule assez simple
  - on peut considérer comme "raisonnablement fiable" un résultat validé dans 95% des cas
- En terminale, on verra un intervalle de fluctuation plus précis !

Autrement dit :

Si l'on fait des simulations et que la probabilité d'un résultat est p, 95 % des résultats trouvés sont dans l'intervalle $\boxed{I=\left[p-\frac{1}{\sqrt{(n)}};p+\frac{1}{\sqrt{(n)}}\right]}$. Cet intervalle s'appelle l'intervalle de fluctuation au seuil 95%.

Réciproquement :

Si on fait une simulation pour estimer une probabilité p, on aura 95% de chances que p se trouve dans l'intervalle $\boxed{I=\left[f-\frac{1}{\sqrt{(n)}};f+\frac{1}{\sqrt{(n)}}\right]}$ où f est le résultat de la simulation sur un échantillon de dimension n. Cet intervalle s'appelle l'intervalle de confiance au seuil de 95%.

Cet intervalle de confiance sera réutilisé en terminale.

Exemple 1 : Intervalle de fluctuation

Supposons que notre rivière contienne 50% de truites femelles (et donc 50% de mâles...). Pour nos échantillons de taille 100, n=100≥25 ; par ailleurs p=0,5 ∈ [0,2 ; 0,8] Donc l'intervalle de fluctuation au seuil de 95% sera :

I=

Exemple 2 : Intervalle de _ _ _ _ _ _ _.

Supposons qu'on lance un dé bien équilibré, et qu'on compte le nombre de résultats supérieurs à 4. Cet événement aura 2 chances sur 6, donc 1 sur 3 de se produire, soit une probabilité p = 1/3 ≈ 33%. Si l'on fait un millier d'échantillons de 100 lancers aléatoires, l'intervalle de fluctuation sera I=_ _ _ _ _ _ _ _=_ _ _ _ _ _ _
donc 95% des résultats seront entre _ _ _% et _ _ _%.

Exemple 3 : Intervalle de _ _ _ _ _ _ _.

Supposons qu'un sondage sur 100 personnes donne 35 voix pour le candidat A, quel sera l'intervalle de _ _ _ _ _ _ _ _ _ correspondant ?

I=

Même question si on interroge 1 000 personnes.

Exercice 1

Deux entreprises recrutent leur personnel dans un bassin d'emploi où il y a autant d'hommes que de femmes. Dans l'entreprise A il y a 100 employés dont 43 femmes. Dans l'entreprise B, il y a 2500 employés dont 1150 femmes (soit 46%). Il semble que l'entreprise B respecte mieux la parité que l'entreprise A.

Il s'agit en réalité de savoir si les entreprises sont composées de personnes choisies au hasard dans la population sans favoriser un sexe au détriment de l'autre. Pour le savoir, on considère que chaque entreprise est un échantillon de la population.

1. Quelle est la proportion *p* de femmes dans la population ?
2. Quelles sont les tailles de ces échantillons ?
3. Quels sont leurs intervalles de fluctuation ?
4. Expliquer pourquoi ces intervalles permettent de dire que l'entreprise A respecte mieux la parité que l'entreprise B.

Exercice 2

Une société fabrique des écrans plasma. En moyenne, 21% des écrans sont défectueux. Lors d'un contrôle d'un lot de 40 écrans, 14 sont défectueux.

1. Calculer la proportion d'écrans défectueux dans ce lot.
2. Faut-il s'en inquiéter ?
3. Afin de tester une nouvelle machine, on contrôle 400 écrans au hasard. 28% d'entre eux sont défectueux.

   Ce résultat semble satisfaisant. Qu'en pensez-vous ?

Exercice 3

Une rhino-pharyngite guérit naturellement en moins de 5 jours dans 60% des cas. On veut tester un médicament censé abréger la durée de la maladie. Pour cela, on administre le médicament à 1000 personnes. Pour 63% d'entre elles, la guérison a eu lieu en moins de 5 jours. Que penser de l'efficacité de ce médicament ?

Exercice 4

On veut estimer la proportion p de foyers disposant en France d'un abonnement internet. 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.

Exercice 5

Une urne contient des boules blanches et des boules noires. On aimerait connaître la proportion p des boules blanches. Pour cela, on effectue 100 tirages avec remise dans cette urne.On obtient 32 boules blanches. Estimez p à l'aide de l'intervalle de confiance au niveau 0,95.

Exercice 2 : 1°) La proportion d'écrans défectueux dans ce lot de 40 écrans est :

$$f = \frac{14}{40} = 0,35 = 35\%.$$

2°) Lors de la fabrication, 21% des écrans sont défectueux.
Donc $p = 0,21$ (probabilité qu'un écran soit défect.)
Si on prend un échantillon de 40 écrans, l'intervalle de fluctuation au seuil de 95% est

$$I = \left[0,21 - \frac{1}{\sqrt{40}} ; 0,21 + \frac{1}{\sqrt{40}}\right]$$

(on peut utiliser cette formule car $p \in [0,2 ; 0,8]$ et $n = 40 \geq 25$)

$$I = [0,05 ; 0,37]$$

La proportion trouvée dans le lot de 40 écrans et de 0,35 qui est bien dans l'intervalle de fluctuation, cela est donc normal.

3°) Ici on ne connait pas la probabilité qu'un écran soit défectueux (Ils sont fabriqués par une autre machine). Par contre on connait la fréquence d'écrans défectueux dans un lot de 400 : elle est de : $f = 0,28$.

L'intervalle de confiance est donc : $\left[0,28 - \frac{1}{\sqrt{400}} ; 0,28 + \frac{1}{\sqrt{400}}\right]$

C'est à dire : $[0,23 ; 0,33]$.

On sait que la probabilité p qu'un écran soit défectueux est dans cet intervalle, on peut donc affirmer que cette nouvelle machine produit plus d'écrans défectueux que la première (affirmation au risque de 5% de se tromper)

Exercice 3 : La probabilité qu'une rhino-pharyngite guérisse naturellement en moins de 5 jours est : $p = 96$.
Dans l'échantillon des 1000 personnes, la fréquence de guérison en moins de 5 jours est : $f = 0,63$.
Or l'intervalle de fluctuation au seuil de 5% est :

$$\left[p - \frac{1}{\sqrt{1000}} ; p + \frac{1}{\sqrt{1000}}\right] = [0,56 ; 0,63]$$

La fréquence $f = 0,63$ est dans cette intervalle, cela permet d'affirmer (au risque de 5%) que le médicament n'a pas d'efficacité prouvée.

Exercice 4 On veut estimer p (on sait que $0,5 \leq p < 0,7$)
→ on peut donc appliquer l'intervalle de confiance si $n \geq 15$
On veut un résultat à 0,01 près pour p.
Or l'intervalle de confiance est : $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$.
Il est de longueur $\frac{2}{\sqrt{n}}$. Il faut donc que :

$$\frac{2}{\sqrt{n}} \leq 0,01 \Leftrightarrow \frac{2}{0,01} \leq \sqrt{n} \Leftrightarrow 40000 \leq n.$$

Donc l'échantillon doit comporter au moins 40000 foyers.

Exercice 5 : D'après le cours p appartient à l'intervalle de confiance : $\left[f - \frac{1}{\sqrt{100}} ; f + \frac{1}{\sqrt{100}}\right]$ dans 95% des cas.

Donc $0,32 - 0,1 \leq p \leq 0,32 + 0,1$

Donc $p \in [0,22 ; 0,42]$ (avec un niveau de confiance de 95%)

II - Intervalle de fluctuation

① Notion d'intervalle de fluctuation.

Définition : X est une v.a. qui suit la loi binomiale $\mathcal{B}(n;p)$. α est un nombre de $]0;1[$ et a et b sont deux réels.

Dire que l'intervalle $[a;b]$ est un intervalle de fluctuation de X au seuil de $1-\alpha$ signifie que :

$$P(a \leq X \leq b) \geq 1-\alpha.$$

Remarque : L'intervalle de fluctuation introduit en classe de première entre dans le cadre général de cette définition, celui ou en seconde est un intervalle approché

② Intervalle de fluctuation asymptotique.

n est un entier naturel et p un réel de $]0;1[$.

Théorème : X_n est un variable aléatoire qui suit la loi binomiale $\mathcal{B}(n;p)$

Pour tout réel $\alpha \in]0;1[$

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$$

où $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

et u_α est le nombre tel que $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$

où Z suit la loi normale $\mathcal{N}(0;1)$

Démonstration :

X_n suit la loi binomiale $\mathcal{B}(n;p)$

donc $E(X_n) = np$ et $\sigma(X_n) = \sqrt{np(1-p)}$

Posons $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$

Z_n est la v.a. centrée réduite de X_n.

D'après le théorème de Moivre-Laplace ;

$$\lim_{n \to +\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq Z \leq u_\alpha)$$

où Z suit la loi normale $\mathcal{N}(0;1)$

or $P(-u_\alpha \leq Z_n \leq u_\alpha)$

$$= P\left(-u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha\right)$$

$$= P\left(-u_\alpha \sqrt{np(1-p)} \leq X_n - np \leq u_\alpha \sqrt{np(1-p)}\right)$$

$$= P\left(np - u_\alpha \sqrt{np(1-p)} \leq X_n \leq np + u_\alpha \sqrt{np(1-p)}\right)$$

$$= P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

$$= P\left(\frac{X_n}{n} \in I_n\right)$$

Donc $\lim_{n \to +\infty} P\left(\frac{X_n}{n} \in I_n\right) = P(-u_\alpha \leq Z \leq u_\alpha)$

Définition : L'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ est appelé intervalle de fluctuation asymptotique au seuil de confiance $1-\alpha$, de la variable aléatoire fréquence $F_n = \frac{X_n}{n}$ qui à tout échantillon de taille n associe la fréquence observée.

Remarque : Cet intervalle contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.

Cette approximation est valable en pratique dès que $n \geq 30$ $np \geq 5$ et $n(1-p) \geq 5$.

Dans le cas où $\alpha = 0{,}05$ alors $1-\alpha = 0{,}95$ et $u_{0{,}05} = 1{,}96$.

On en déduit un intervalle de fluctuation asymptotique au seuil de 95%.

Propriété : Un intervalle de fluctuation asymptotique au seuil de confiance de 95% de la fréquence F_n d'un caractère dans un échantillon de taille n est :

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

où p désigne la proportion de ce caractère dans la population

Remarque : L'intervalle de fluctuation au seuil de confiance de 95% ou en seconde est :

$$I = \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$$

on peut montrer que $I_n \subset I$

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

Exercice : Montrons que $I_n \subset I$.

Pour cela, définissons la fonction $f(p) = \sqrt{p(1-p)}$ définie pour $p \in [0;1]$.

a) Calculer $f'(p)$.

$$f'(p) = \frac{u'}{2\sqrt{u}} = \frac{1-2p}{2\sqrt{p(1-p)}}$$

b) En déduire les variations de f sur $]0;1[$.

p	0		$\frac{1}{2}$		1
$1-2p$		$+$	0	$-$	
$f(p)$	0	↗	$f(\frac{1}{2})$	↘	0

$$f\left(\frac{1}{2}\right) = \sqrt{\frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}$$

c) Le maximum est $\frac{1}{2}$

d) Donc $f(p) \leq \frac{1}{2}$

$$\frac{f(p)}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}}$$

$$\frac{1{,}96 f(p)}{\sqrt{n}} \leq \frac{1{,}96 \times 1}{2\sqrt{n}} \leq \frac{1}{\sqrt{n}}$$

3-Prise de décision

On considère une population dans laquelle on fait l'hypothèse H suivante :

H="la proportion du caractère étudié dans la population est p"

On observe la fréquence f de ce caractère dans un échantillon de taille n.

On note I l'intervalle de fluctuation de la fréquence au seuil de 95%.

La règle de décision est la suivante :

- Si $f \in I$, on considère que l'hypothèse n'est pas remise en cause, on l'accepte.
- Si $f \notin I$, on rejette l'hypothèse selon laquelle cette proportion vaut p.

Un exemple : Une marque de bonbons chocolatés vend des paquets constutués de bonbons de cinq couleur différentes. Les bonbons de couleur marron sont annoncés comme représentant 20% de l'ensemble des bonbons.

Pour vérifier cette information, les élèves d'une classe de terminale ont observé un échantillon, que l'on peut considérer comme aléatoire, de 690 bonbons. Ils ont dénombré 140 bonbons marron.

Que peut-on penser de la proportion annoncée ?

Solution : On est typiquement dans un cas de prise de décision.

1°) Quelle est l'hypothèse H que l'on veut vérifier ?

H="

2°) Déterminer l'intervallede fluctuation au seuil de 95% :

$$p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} =$$

$$p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} =$$

Donc l'intervalle de fluctuation est : I=

3°) Calculer la fréquence observée f :

f=

4°) conclure.

III- Estimation.

1-Echantillon et estimation

Dans une population on s'intéresse à la proportion p d'individus ayant un certain caractère.

Cette proportion p est inconnue, et aucun élément ne permet d'émettre une hypothèse sur sa valeur.

On prélève alors un échantillon de taille n dans cette population, avec remise (ou, si la population est assez grande, on admet que l'on peut considérer qu'il s'agit de tirages avec remise).

Du fait de la fluctuation d'échantillonnage, la fréquence f observée varie d'un échantillon à l'autre et est probablement différente de la proportion inconnue p.
On estime alors la proportion p par un i**ntervalle de confiance** obtenu à partir de la fréquence observée f, avec un certain niveau de confiance.

2-Intervalle de confiance

Définition : L'intervalle $I=\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de la proportion inconnue p au niveau de confiance 0,95.

Exemple1 : Un laboratoire pharmaceutique met en place un test pour estimer l'efficacité d'un nouveau médicament contre les migraines.
Deux groupes de 125 patient souffrant de migraines, considérés comme des échantillons aléatoires, participent à ce test. On administre aux patients du groupe A le nouveau médicament, alors que les patients du groupe B reçoivent un placebo.

Au bout de 4 jours de traitement, 73 patients du groupe A et 64 patients du groupe B déclarent ressentir une diminution de l'intensité des migraines.

Au niveau de confiance 0,95, le test permet-il d'estimer que le médicament est efficace contre la migraine ?

Solution :

•Intervalle de confiance pour le groupe A

la taille de l'échantillon est n=

la fréquence des patients qui ont ressenti une amélioration est : f=

L'intervalle de confiance est donc :

$f-\frac{1}{\sqrt{n}}=$

$f+\frac{1}{\sqrt{n}}=$ I=

•<u>Intervalle de confiance pour le groupe B</u>

la taille de l'échantillon est n=

la fréquence des patients qui ont ressenti une amélioration est : f=

L'intervalle de confiance est donc :

$f-\frac{1}{\sqrt{n}}=$

$f+\frac{1}{\sqrt{n}}=$ J=

•<u>Conclusion</u> :

<u>Exemple2</u> : Déterminer la taille d'un échantillon

Le gérant d'une salle de spectacle veut estimer la proportion de places vendus non occupées afin d'organiser la surréservation. Quelle taille d'échantillon de ses clients doit-il étudier pour obtenir un intervalle de confiance au niveau de confiance 0,95 de longueur au plus 0,04 ?

<u>Solution :</u>

•Mise en équation :

•Résolution :

•Conclusion :