CHAPTER

18

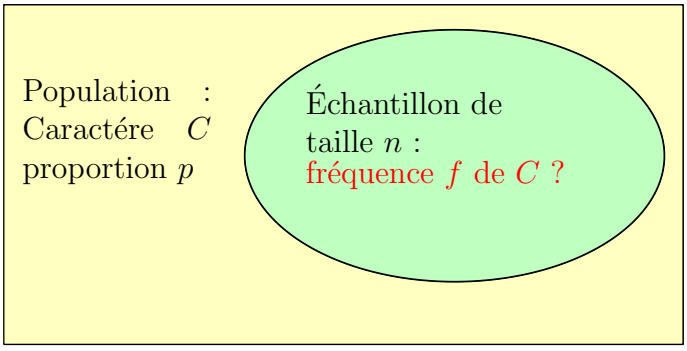
Fluctuation et estimation

Lorsqu'on s'intéresse à une caractèristique d'une population, il est parfois impossible de tester chaque individu, on est alors amené à travailler sur des échantillons.L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

1 Fluctuation d'échantillonnage

1 1 Rappels

Dans une population donnée où la proportion d'individus présentant un caractère C est p, on prélève un échantillon de taille n.
En classe de seconde, on a observé que sur un grand nombre d'échantillon de taille n et sous certaines conditions, 95% au moins fournissent une fréquence f appartenant à l'intervalle $[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}]$.



En classe de première : Le tirage aléatoire d'un individu dans un population est assmillé à une épreuve de Bernoulli, le prélevement au hasard d'un échantillon de taille n dans cette population correspond à un schéma de Bernoulli de paramètres n et p. La variable aléatoire X qui compte le nombre de succès avoir le caractère C , suit la loi binomiale $\mathscr{B}(n \ ; \ p)$.

La variable aléatoire fréquence $F = \dfrac{X}{n}$ représente la fréquence aléatoire du succès sur un échantillon de taille n.
Or (cf cours seconde) on a $P(p - \frac{1}{\sqrt{n}} \leqslant F \leqslant \ p + \frac{1}{\sqrt{n}}) \geqslant 95\%$,
l'intervalle $[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}]$ est un intervalle de fluctuation de F.

Définition 1

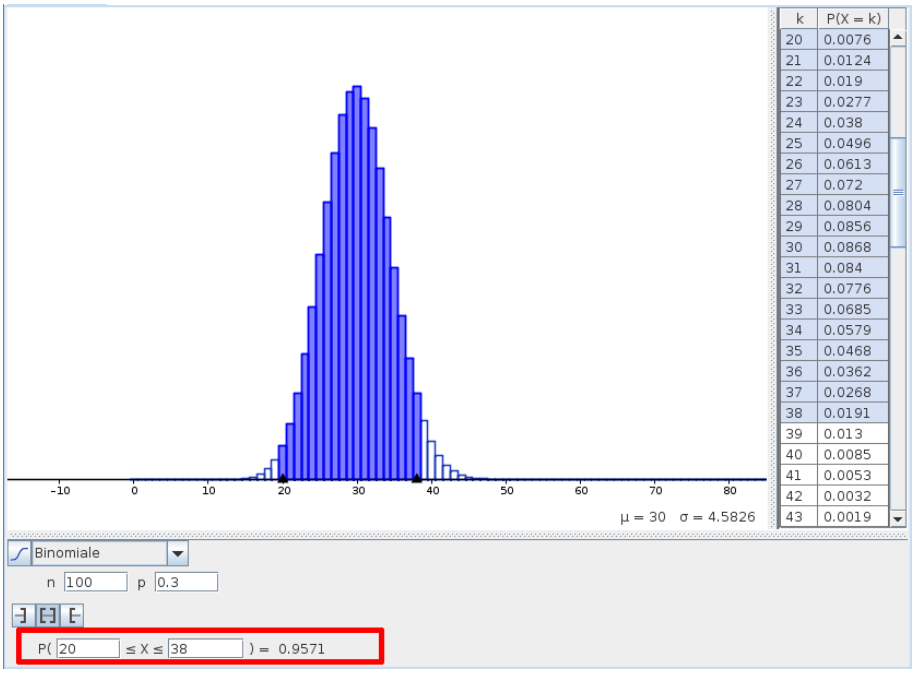
Soit X une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n \ ; \ p)$. et $F = \dfrac{X}{n}$ la variable aléatoire fréquence du succès.
Un intervalle de fluctuation de F au seuil de 95% est un intervalle

- $\left[\dfrac{a}{b} \ ; \ \dfrac{b}{n}\right]$, avec a et b deux entiers compris entre 0 et n ;
- tel que $P\left(\dfrac{a}{n} \leqslant F \leqslant \dfrac{b}{n}\right) \geqslant 95\%$ soit $P(a \leqslant X \leqslant b) \geqslant 95\%$

Pour déterminer un intervalle de fluctuation il suffit de déterminer deux entier a et b tels que $P(a \leqslant X \leqslant b) \geqslant 95\%$.

Exemple. Une urne contient 3 boules rouges et 7 boules blanches, on effectue 100 tirages au hasard avec remise. Déterminer un intervalle de fluctuation au seuil de 95% de la fréquence d'apparition d'une boule rouge dans l'échantillon prélevé.

À l'aide du logiciel GeoGebra, on représente une loi binomiale $\mathscr{B}(100\ ;\ 0,3)$.



On détermine deux entiers $a = 20$ et $b = 38$ qui conviennent,
en effet $P(20 \leqslant X \leqslant 38) \geqslant 95\%$

Dans la suite de ce chapitre, on suppose que la taille de l'échantillon n est la proportion p du caractère C vérifient :

$$n \geqslant 30,\quad n \times p \geqslant 5\ \ etn \times (1-p) \geqslant 5$$

2 Intervalle de fluctuation asymptotique et Test

2 1 Intervalle de fluctuation asymptotique

Définition 2

Pour tout réel α tel que $0 < \alpha < 1$, un intervalle de fluctuation asymptotique de la variable aléatoire F au seuil de $1-\alpha$ est un intervalle dépendant uniquement de n et de p qui contient F avec une probabilité proche de $1-\alpha$ quand n est grand.

Propriété 1

Pour tout réel $\alpha \in]0\ ;\ 1[$, il existe un unique réel u_α tel que la probabilité que la variable aléatoire fréquence F prenne des valeurs dans l'intervalle $I_n = \left[p - u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\ ;\ p + u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\right]$ se rapproche de $1-\alpha$ quand la taille n de léchantillon devient grand
$\lim\limits_{n\to+\infty} P(F \in I_n) = 1-\alpha$

Propriété 2

IFA au seuil de 95%
l'intervalle de fluctuation asymptotique au seuil de 95% de la variable aléatoire fréquence F est :

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

2 2 Test

Dans cette partie on souhaite vérifier, à l'aide d'échantillon de taille n , si on peut raisonnablement penser que la proportion p de la population est bien celle annoncée. On construit un **test** qui va nous permettre d'énoncer une règle de décision concernant cette proportion.

- H_0 : la proportion de la population présentant le caractère C est p . (Hypothèse H_0)
- On établit un IFA au seuil α,

  (le plus souvent $\alpha = 95\%$ soit $\left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \ ; \ p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
- On énnonce le test :

  Si la fréquence observée f de l'échantillon appartient à l'IFA au seuil α **on accepte** l'hypothèse H_0.

  Si la fréquence observée f de l'échantillon n'appartient pas à l'IFA au seuil α **on rejette** l'hypothèse H_0 c'est-à-dire que la proportion de la population n'est pas p au risque de 5% de se tromper.

Remarques.

- Lorsqu'on rejette l'hypothèse H_0 au risque de 5%, on peut rejetter à tort l'hypothèse (rejet sachant qu'elle est vraie) avec une probabilité proche de 0,05.
- Par contre lorqu'on accepte l'hypothèse H_0 on ne connait pas la probabilité d'erreur.

  En effet si la proportion $p = p_0$ cela implique que la fréquence f d'échantillon de taille n appartient à l'intervale de fluctuation asymptote au seuil de 95 % mais on ne sait rien de la réciproque :

  si $f \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \ ; \ p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ n'implique pas forcément que $p = p_0$ avec une probabilité de 95%.

3 Estimation

Lorsque la proportion p d'un caractère C d'une population est inconnue et qu'on est dans l'impossibilité de tester l'ensemble de cette population, on fait ce que l'on appelle une estimation par intervalle de confiance.

Définition 3

Pour tout réel α tel que $0 < \alpha < 1$, un **intervalle de confiance** de la proportion p au niveau de confiance $1 - \alpha$ est un intervalle issue d'un échantillon de taille n contenant la proportion p avec une probabilité supérieur ou égale à $1 - \alpha$.

Remarques. L'intervalle de confiance n'est pas unique et il dépent de l'échantillon aléatoire choisi.
Seul l'intervale de confiance au niveau de confiance de 95% est au programme de la classe de terminale.

Propriété 3

L'intervalle de confiance au niveau de confiance de 95% est :

$$\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right]$$

avec f la fréquence observée d'un échantillon de taille n.

Démonstration. soit F la variable aléatoire qui a un échantillon de taille n associe la fréquence f de cet échantillon de taille n.
Pour n est assez grand, la probabilité $P(p-\frac{1}{\sqrt{n}} \leqslant F \leqslant p+\frac{1}{\sqrt{n}})$ est supérieur ou égale à 95%. Ce qui peut se réécrire :

$$p - \frac{1}{\sqrt{n}} \leqslant F \leqslant p + \frac{1}{\sqrt{n}}$$
$$-\frac{1}{\sqrt{n}} \leqslant F - p \leqslant +\frac{1}{\sqrt{n}}$$
$$-F - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -F + \frac{1}{\sqrt{n}}$$
$$F + \frac{1}{\sqrt{n}} \geqslant -p \geqslant F - \frac{1}{\sqrt{n}}$$
$$F - \frac{1}{\sqrt{n}} \leqslant p \leqslant F + \frac{1}{\sqrt{n}}$$

d'où $P(p - \frac{1}{\sqrt{n}} \leqslant F \leqslant p + \frac{1}{\sqrt{n}}) \geqslant 95\% \iff P(F - \frac{1}{\sqrt{n}} \leqslant p \leqslant F + \frac{1}{\sqrt{n}}) \geqslant 95\%$
L'intervalle $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right]$ étant une réalisation de l'intervalle aléatoire on en déduit donc qu'il s'agit d'un intervalle de confiance au niveau de confiance de 95%. □

Exemple. Comparaison de taux de germination

Un maraîcher achète un lot de semences de tomates pour produire des plants de tomates. Il lui reste des semences de l'année précédente, dont il doit contrôler le taux de germination pour pouvoir les utiliser avec les autres.
Il faut donc comparer les taux de germination des semences des deux années.
Une stratégie consiste à calculer et à comparer les intervalles de confiances des taux de germination (qui sont des proportions) des plants de l'année précédente

Si les deux intervalles ne se recoupent pas, on peut conclure à une différence de taux de germination entre les semences des deux origines. Il faudra alors les planter séparèment.
Pour faire cette comparaison le maraîcher prélève, aléatoirement dans les semences de l'année, un échantillon de 200 graines qu'il met à germer. Il constate que 185 graines germent. Il prélève ensuite, aléatoirement dans les semences de l'année précédente, un échantillon de 200 graines qu'il met à germer. Il constate que 150 graines germent.

1. Déterminer un intervalle de confiance, au niveau de confiance de 95%, du taux de germination p_a du lot de semences de l'année.
2. Même question pour le lot de semences de l'année précédente p_b. Conclure.