Niveau :

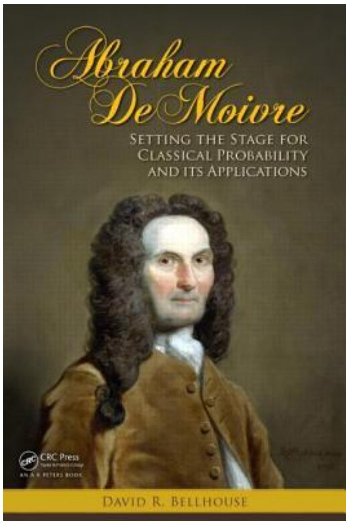
Terminale S

Titre Cours :

Chapitre 11
Echantillonnage et estimation

Année :

2015-2016

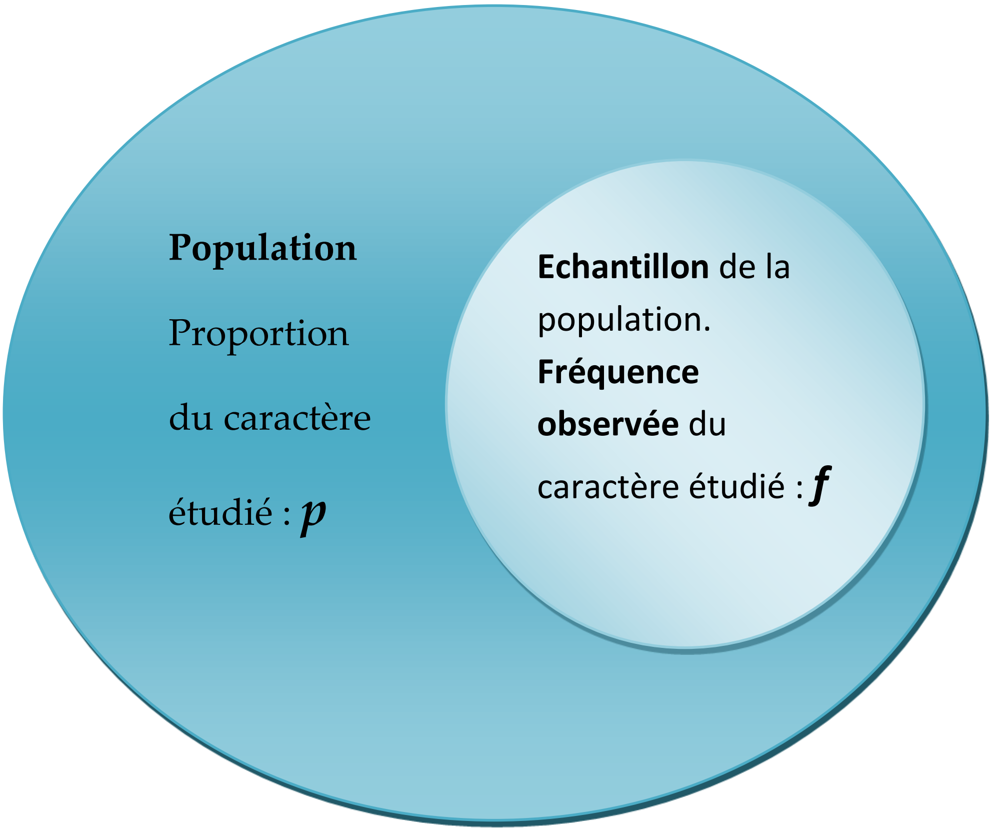


Abraham de Moivre (1667-1754)
Mathématicien Français

Citation du moment :

«Un probabiliste est une personne qui peut avoir la tête dans un four et les pieds pris dans la glace tout en disant qu'en moyenne il se sent bien» (Anonyme : Livre de Patrick Bogaert)

I. Situation



On souhaite étudier un caractère dans une population (Ex : Ceux qui ont voté pour un candidat, …….).
Deux possibilités s'offrent à nous :

- **Echantillonnage** : Soit on connaît la **proportion p** d'individus ayant ce caractère dans l'ensemble de la population et on souhaite savoir dans quel intervalle va fluctuer (**intervalle de fluctuation**) la fréquence de ce caractère dans un échantillon obtenu. Si la fréquence *f* de l'échantillon est dans l'intervalle de fluctuation, alors on dira que l'échantillon est représentatif de la population.

- **Estimation** : Soit on ne connaît pas la proportion d'individus ayant ce caractère dans la population. On prélève un échantillon de la population et on détermine la **fréquence** de ce caractère apparaissant dans l'échantillon. Cette fréquence nous permet d'estimer **un intervalle de confiance** dans lequel fluctue la **proportion p** du caractère dans la population totale. Cela permet, par exemple, de faire des sondages et d'estimer le pourcentage de vote pour un candidat, sur une population complète.

Définition 01 : Intervalle de fluctuation (p connue ou conjecturé)
Soit α un réel de l'intervalle $]0,1[$
L'intervalle de fluctuation au seuil de $100(1-\alpha)$ %, relatif aux échantillons de taille n, est l'intervalle centré autour de p qui contient la fréquence observée f dans un échantillon de taille n avec une probabilité égale à $1-\alpha$.

Définition 01 : Intervalle de confiance (p est inconnu)
Soit α un réel de l'intervalle $]0,1[$
L'intervalle de confiance au seuil de $100(1-\alpha)$ %, relatif aux échantillons de taille n, est l'intervalle centré autour de f où se situe la proportion p du caractère dans la population avec une probabilité égale à $1-\alpha$.

II. Rappels de seconde et de première S

1. Seconde

Soit p la proportion effective d'un caractère d'une population comprises entre 0,2 et 0,8 et f la fréquence du caractère dans un échantillon de taille n supérieure ou égale à 25. **L'intervalle de fluctuation** au seuil de 95 % $(\alpha = 0.05)$ est

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

Soit f la fréquence observée d'un caractère d'un échantillon (de taille ≥ 25) comprises entre 0,2 et 0,8 et p la proportion du caractère dans la population.

L'intervalle de confiance au seuil de 95 % $(\alpha = 0.05)$ est

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

Exemple 01 :

Dans un village de campagne à proximité d'industries chimiques, il est né entre l'année 2001 et 2005, 164 enfants dont 56 garçons.

Cela est-il normal ?

Exemple 02 :

Le 4 mai 2007, sit deux jours avant le second tour des élections présidentielles, on publie le sondage suivant réalisé auprès de 992 personnes :

Ségolène Royal : 45 %   Nicolas Sarkozy : 55 %

Comment pouvez-vous interpréter ce sondage ?

2. Première S

Le tirage au hasard d'un individu dans une population est une épreuve de Bernoulli de paramètre p, où le succès est l'issue « avoir le caractère étudié »

Le prélèvement au hasard d'un échantillon de taille n dans cette population s'assimile dans à un schéma de Bernoulli de paramètre n et p.

La variable aléatoire X, qui compte le nombre d'individus ayant le caractère étudié, suit une loi Binomiale $B(n,p)$

La fréquence du caractère étudié dans l'échantillon est donc donnée par la variable aléatoire $F=\frac{X}{n}$

Définition : Soit X une variable aléatoire qui suit la loi binomiale $B(n,p)$ et $F=\frac{X}{n}$ la variable aléatoire qui correspond à la fréquence f du caractère étudié.

L'intervalle de fluctuation de f au seuil 95 % fourni par la loi binomial est le plus petit intervalle de la forme $\left[\frac{a}{n};\frac{b}{n}\right]$ où a et b sont des entiers entre 0 et n tels que :

$P\left(F<\frac{a}{n}\right)\leq 0,025$ et $P\left(F>\frac{b}{n}\right)\leq 0,025$ et donc $P\left(\frac{a}{n}\leq F\leq\frac{b}{n}\right)\geq 0,95$

Remarque : Le prélèvement au hasard d'un échantillon de taille n consiste en n répétitions d'un tirage avec remise. Mais si les tirages s'effectuent sans remise, l'assimilation à un schéma de Bernoulli est encore possible lorsque la population est suffisamment grande en regard de n.

Définition : Soit f la fréquence observée d'un caractère d'un échantillon (de taille ≥ 25) comprises entre 0,2 et 0,8 et p la proportion du caractère dans la population. **L'intervalle de confiance** au seuil de 95 % $(\alpha=0.05)$ est

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

III. Théorème de Moivre-Laplace

Théorème

Soit n un entier naturel non nul et p une probabilité. On note X une variable aléatoire réelle qui suit une loi Binomiale $B(n,p)$.

Si on note $Z=\dfrac{X-np}{\sqrt{np(1-p)}}$ alors Z est une variable aléatoire centrée réduite et pour tous réels a et b tels que $a\leq b$, on a : $\lim\limits_{n\mapsto+\infty} P(a\leq Z\leq b)=\int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$

En d'autres termes, lorsqu'on travaille sur un échantillon très grand, on peut faciliter ses calculs en remplaçant la loi binomiale par la loi normale (on parlera de convergence en loi). Z suit donc une loi normale centrée réduite.

En pratique, on considère que pour $n\geq 30$ l'approximation est valable. En classe de terminale S on prendra comme conditions que $\begin{cases} n\geq 30 \\ np\geq 5 \\ np(1-p)\geq 5 \end{cases}$

IV. Intervalle de fluctuation asymptotique et intervalle de confiance en terminale S

Soit X_n une variable aléatoire suivant une loi binomiale $B(n,p)$ où $p\in]0;1[$

On suppose que $n\geq 30$, $np\geq 5$ et $np(1-p)\geq 5$

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

D'après le théorème de Moivre-Laplace :

Z_n suit une loi ……………………………………

et

$P\left(Z_n\in[-u_\alpha;u_\alpha]\right)=$

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

Démontrez que $Z_n \in [-u_\alpha ; u_\alpha] \Leftrightarrow \dfrac{X_n}{n} \in I_n$:

Définition :

L'intervalle $I_n = \left[p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ contient la fréquence $F_n = \dfrac{X_n}{n}$ avec une probabilité qui se rapproche de $1-\alpha$ lorsque n augmente : On dit que c'est un intervalle de fluctuation asymptotique de F_n au seuil $1-\alpha$

Prenons $\alpha = 0,05$ et donc $1-\alpha = 0,95$

Quelle est la valeur de u_α ? $u_\alpha =$

Définition :

Soit X_n une variable aléatoire suivant une loi binomiale $B(n,p)$ où $p \in]0;1[$

On suppose que $n \geq 30$, $np \geq 5$ et $np(1-p) \geq 5$

L'intervalle de fluctuation asymptotique au seuil de 95 % est donc :

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

1. Justifiez que $P\left(p-2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leq\frac{X_n}{n}\leq p+2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\geq 0{,}95$

2. Démontrer que pour tout $p\in]0;1[$, $p(1-p)\geq\frac{1}{4}$

3. En déduire que $P\left(p-\frac{1}{\sqrt{n}}\leq\frac{X_n}{n}\leq p+\frac{1}{\sqrt{n}}\right)\geq 0{,}95$

4. En déduire que $\left(\frac{X_n}{n}-\frac{1}{\sqrt{n}}\leq p\leq\frac{X_n}{n}+\frac{1}{\sqrt{n}}\right)\geq 0{,}95$

Définition : Soit f la fréquence observée d'un caractère d'un échantillon (de taille ≥ 30) telle que $nf \geq 5$, $nf(1-f) \geq 5$ et p la proportion du caractère dans la population.

L'intervalle de confiance au niveau de confiance 0,95 est

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

Précision d'une estimation :

Quelle est l'amplitude de l'intervalle de confiance au niveau 95 % ? ……………

Conclusion :

Plus la taille de l'échantillon est grand et plus l'intervalle de confiance au niveau 95 % est …………………