Mathématiques en Terminale S
Échantillonnage et estimation

Table des matières

Section 1

Principe et rappel :

La notion de fluctuation d'échantillonnage a été introduite en seconde puis en première.

L'intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, est l'intervalle centré autour de p, proportion du caractère dans la population, où se situe, avec une probabilité égale à 0,95, la fréquence observée dans un échantillon de taille n.

1.1 En seconde :

En seconde l'intervalle de fluctuation au seuil de 95% est l'intervalle $[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$.

Exemple : Dans une urne comportant 40% de boules blanches, on tire au hasard une boule,on regarde sa couleur et on remet la boule. En répétant n fois cette expérience, on obtient un échantillon de taille n. On souhaite savoir si la proportion initiale de boules blanches est conservée dans l'échantillon.

.....
.....
.....

Commentaire : Pour obtenir un échantillon de taille n, on répète donc de façon identique et indépendante, la même expérience qui livre à chaque réalisation, un calcul de fréquence. Il s'agit alors d'un tirage avec remise.
Néanmoins, lorsque le tirage est effectué dans une population à fort effectif, la remise ou non de l'individu affecte très peu les probabilités des événements suivants. On considère donc que dans ce cas qu'un tirage sans remise s'assimile à un tirage avec remise.

1.2 En première :

Exemple :**Cas de la loi binomiale** Le responsable de la maintenance des machines à sous d'un casino doit vérifier qu'un certain type de machine est bien réglé sur une fréquence de succès de 0,06. Pour cela il veut établir un programme qui lui fournira, en fonction de n (nombre de coups joués) et de p (probabilité de succès), un intervalle de fluctuation, au seuil de 95%, de la fréquence de succès. Cela lui permettra de prendre la décision de régler chaque machine pour laquelle il aura observé, dans l'historique des jeux, une fréquence de succès se situant en dehors de cet intervalle de fluctuation.
On cherche le plus petit entier a pour lequel $P(X \leq a)$ est strictement supérieur à 0,025 et le plus petit entier b pour lequel $P(X \leq b)$ est supérieur ou égal à 0,975. Ces recherches peuvent s'effectuer à la calculatrice ou avec un programme. Lors du contrôle d'une machine, le technicien constate qu'elle a fourni 8 succès sur 65 jeux, soit une fréquence observée de succès d'environ 0,12. L'intervalle de fluctuation de la variable fréquence fourni par le programme précédent est [0,015 ; 0,123]. Bien que la fréquence observée de succès soit de 0,12, la règle de décision n'amène pas à remettre en question le réglage de la machine.
Le théorème de Moivre-Laplace va permettre de donner un intervalle de fluctuation calculable directement, sous réserve que n soit assez grand. Comme il est obtenu grâce à une convergence, on le qualifie d'intervalle de fluctuation asymptotique.

Section 2

Intervalle de fluctuation asymptotique :

Propriétés

On considère une variable aléatoire X_n qui suit la loi binomiale $\mathcal{B}(n,p)$ avec p dans l'intervalle $]0;1[$.Alors pour tout α dans $]0;1[$, on a :

$$\lim_{n\to+\infty} P(\frac{X_n}{n} \in I_n) = 1-\alpha \text{ où } I_n =]p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}[$$

Remarque :Quand $n > 30$, $np > 5$ et $n(1-p) \geq 5$, il est courant de faire les calculs impliquant une variable binomiale en la remplaçant par une variable suivant une loi normale de mêmes espérance et variance. Ce sont les conditions requises pour l'application du théorème de Moivre Laplace.

Démonstration :

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$. Alors d'après le théorème de Moivre Laplace,

$$\lim_{n\to+\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$$

Or : $P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq \dfrac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha) = P(-u_\alpha\sqrt{np(1-p)} \leq X_n - np \leq u_\alpha\sqrt{np(1-p)})$

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

Exemple concret d'utilisation :

Une compagnie aérienne possède des A340 (longs courriers) d'une capacité de 300 places. Cette compagnie a vendu n billets pour le vol 2012.

La probabilité pour qu'un acheteur se présente à l'embarquement est p et les comportements des acheteurs sont indépendants les uns des autres.

On note X_n la variable aléatoire désignant le nombre d'acheteurs d'un billet se présentant à l'embarquement.

La compagnie cherche à optimiser le remplissage de l'avion en vendant éventuellement plus de places que la capacité totale de l'avion (surréservation ou surbooking) soit ici $n > 300$. Comme il y a évidemment un risque que le nombre de passagers munis d'un billet se présentant à l'embarquement excède 300, la compagnie veut maîtriser ce risque.

1. Quelle loi suit la variable X_n ?

.....
.....
.....

2. On suppose que $0,5 < p < 0,95$. Écrire l'intervalle de fluctuation asymptotique I_n de $\frac{X_n}{n}$ au seuil de 0,95.

.....
.....
.....

3. Montrer que si $I_n \subset [0; \frac{300}{n}]$ alors la probabilité que le nombre de passagers se présentant à l'embarquement excède 300 est proche de 0,05.

.....
.....
.....

4. On cherche à déterminer la valeur de n maximale permettant de satisfaire la condition de l'inclusion $I_n \subset [0; \frac{300}{n}]$.

   (a) Montrer que $I_n \subset [0; \frac{300}{n}] \Rightarrow pn + 1.96\sqrt{n}\sqrt{p(1-p)} - 300 \leq 0$

.....
.....
.....

(b) On pose $f(x) = px + 1.96\sqrt{x}\sqrt{p(1-p)} - 300$. Montrer qu'il existe un unique entier n_0 tel que si $n \leq n_0$ alors $f(n) > 0$ et si $n \geq n_0$ alors $f(n) < 0$.

.....
.....
.....

(c) Déterminer à la calculatrice les valeurs de n_0 pour $p = 0,85; p = 0,9; p = 0,95$.

.....
.....
.....

Section 3

Estimation :

3.1 Principe :

Il est souvent difficile pour des raisons à la fois financières et logistiques de pouvoir recueillir des données sur la population toute entière. Le plus souvent, on se contente de travailler sur un échantillon, c'est à dire une fraction ou sous-ensemble de cette population. Ceci présente bien sûr des avantages en termes de faisabilité et de coût, mais impose des contraintes pour que l'information recueillie au niveau de l'échantillon (estimation) soit la plus proche possible de celle de la population entière (paramètre). La démarche pratique est donc la suivante :

- on sélectionne un échantillon de la population que l'on étudie, on appelle cela l'échantillonnage.
- on vérifie, selon les cas, à partir d'intervalles de fluctuation que l'échantillon ainsi obtenu est représentatif de la population pour des critères qui sont connus dans la population.

Exemple :

On souhaite estimer la prévalence du surpoids dans une ville V, c'est-à-dire la proportion de personnes ayant une masse trop importante par rapport à leur taille. Pour cela 460 personnes ont été sélectionnées de manière aléatoire à partir de la liste des logements connue par la municipalité, c'est-à-dire que le fait d'avoir été sélectionné pour participer à l'étude est uniquement dû au hasard. On admet que cette procédure permet d'assimiler la sélection des personnes interrogées à un schéma de Bernoulli.
Un enquêteur s'est déplacé au sein de chaque logement après avoir convenu d'un rendez-vous afin de recueillir les informations nécessaires à l'enquête.

1. Dans un premier temps, l'enquêteur va s'assurer que l'échantillon est représentatif de la population qu'on étudie sur des informations qu'on peut vérifier et qui sont en lien avec le critère étudié. Dans le cas présent on peut connaître par exemple la proportion d'hommes et de femmes dans la population de la ville, ainsi que la répartition selon l'âge en demandant à la municipalité qui se référera aux

informations du recensement. Parallèlement on peut comptabiliser le nombre d'hommes et de femmes dans l'échantillon ainsi que la répartition selon l'âge.

	Hommes	Femmes	Total
Echantillon	200	260	460
	Moins de 60 ans	Plus de 60 ans	Total
Echantillon	352	108	460

On sait que, dans la population, il y a 46% d'hommes et 20% de personnes de plus de 60 ans.

(a) Déterminer l'intervalle de fluctuation asymptotique au seuil 0,95 de la variable aléatoire « proportion de femmes » dans un échantillon aléatoire de taille 460 sélectionné au sein de la population de cette ville.

.....
.....
.....

(b) Calculer la proportion de femmes dans l'échantillon et vérifier si cette valeur appartient à l'intervalle de fluctuation.

.....
.....
.....

(c) Déterminer l'intervalle de fluctuation asymptotique au seuil 0,95 de la variable aléatoire « proportion de personnes âgées de plus de 60 ans » dans un échantillon aléatoire de taille 460 sélectionné au sein de la population de cette ville.

.....
.....
.....

(d) Calculer la proportion de personnes de plus de 60 ans dans l'échantillon et vérifier si cette valeur appartient à l'intervalle de fluctuation.

.....
.....
.....

(e) Si pour chacune des variables, genre et âge, l'intervalle de fluctuation asymptotique au seuil de 95% contient la valeur de l'échantillon on considère que l'échantillon est représentatif de la population pour cette information. Quelle est donc la conclusion pour le cas étudié ici ?

.....
.....
.....

La représentativité sur deux critères ne signifie évidemment pas la représentativité sur tous les critères et dans tous les cas, il est peu vraisemblable qu'un échantillon de 460 sujets soit représentatif pour tous les critères. Les résultats obtenus sur un échantillon ne peuvent pas remplacer les résultats exacts d'un recensement. Cependant la vérification précédente sur des critères importants permet de considérer que l'échantillon retenu est structuré comme la population étudiée, au regard de certains critères.

2. La première étape de ce travail a donc été de sélectionner un échantillon qui soit accepté comme représentatif de la population. Ainsi les informations qui seront obtenues à partir de cet échantillon seront généralisables, avec un certain nombre de précautions, à l'ensemble de la population dont il est extrait. Dans le cas de l'étude présentée ici, on souhaite estimer la proportion de personnes en surpoids ; pour cela il est tout d'abord important de définir le surpoids. La définition du surpoids donnée par l'OMS (Organisation Mondiale de la Santé) est la suivante : une personne est considérée en surpoids si son IMC (Indice de masse corporelle) est supérieur à 25. L'IMC se calcule de la manière suivante : masse en kg/(taille en m)^2. La proportion de personnes en surpoids dans l'échantillon étudié est de 29,5%. Comme il s'agit d'un calcul réalisé à partir des données d'un échantillon on sait que cette valeur ne correspond pas exactement à la valeur de la prévalence dans la population, car si nous avions pris un autre échantillon nous aurions obtenu une autre valeur. Pour cette raison il est nécessaire de communiquer un intervalle qui sera obtenu à partir des informations observées et pour lequel on puisse dire avec un « niveau de confiance » supérieur à 0,95 qu'il contient la vraie valeur de la prévalence du surpoids dans la ville. Si f est la fréquence observée dans l'échantillon une expression de cet intervalle, qui sera appelé intervalle de confiance, est :

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

où n est la taille de l'échantillon.
Déterminer un intervalle de confiance au niveau de confiance de 95

.....
.....
.....

3.2 Définition :

On considère une variable aléatoire X_n qui suit la loi binomiale $\mathcal{B}(n;p)$. Alors pour tout p dans $]0;1[$ et pour n suffisamment grand,

$$P(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}) \geq 0.95$$

ce qui équivaut à :

$$P(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}) \geq 0.95$$

en posant $F_n = \frac{X_n}{n}$.

Définition

Pour une valeur de p fixée dans l'intervalle $]0;1[$, l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}\right]$ contient pour n assez grand, la proportion p avec une probabilité supérieure ou égale à 0.95.

À partir de cet intervalle, on obtient en calculant la fréquence f dans un échantillon de taille n, une réalisation $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ de cet intervalle.

Définition

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

Il existe d'autres manières de déterminer un intervalle de confiance d'une proportion ;dans d'autres champs disciplinaires on utilise l'intervalle :

$$\left]f - 1.96\frac{\sqrt{f(1-f)}}{\sqrt{f}}; f + 1.96\frac{\sqrt{f(1-f)}}{\sqrt{f}}\right[$$

3.3 Longueur de l'intervalle de fluctuation :

L'intervalle de fluctuation asymptotique $I_n = \left]p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right[$ a pour longueur :

$$2u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Donc pour α et n fixés, la longueur de I_n varie comme $\sqrt{p(1-p)}$. Elle est donc maximale quand $p = \frac{1}{2}$ et d'autant plus faible que p est proche de 0 ou de 1. Quelques valeurs de la longueur $2u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ pour $n = 1000$:

	$p = 0,1$	$p = 0,3$	$p = 0,4$	$p = 0,5$
$\alpha = 0,05$	0,037	0,057	0,061	0,062
$\alpha = 0,01$	0,049	0,075	0,08	0,082

Si on cherche à estimer par intervalle, au niveau de confiance 0,95, une valeur de p dont on sait qu'elle est plutôt proche de 0,5 (cas du second tour de l'élection présidentielle), on a un intervalle de confiance, appelé dans ce cas fourchette de sondage, d'amplitude proche de 0,06. Si on cherche à estimer une valeur de p sans doute inférieure à 0,1 (cas des petits candidats du premier tour), on a une fourchette d'amplitude proche de 0,04.
On constate sur le tableau précédent que, n étant fixé, l'augmentation du niveau de confiance augmente simultanément la longueur de l'intervalle de confiance, ce qui est un résultat général facile à justifier (et à concevoir).

3.4 Taille minimale de l'échantillon pour une précision donnée :

On étudie d'abord la taille minimale de l'échantillon pour avoir une longueur donnée a de l'intervalle de fluctuation pour un seuil ou un niveau de confiance fixé.

Intervalle de la classe de seconde : $I_n = \left]p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right[$.

On cherche n tel que $\dfrac{2}{\sqrt{n}} \leq a$ ce qui équivaut à $n \geq \dfrac{4}{a^2}$. Quelques valeurs de n pour a fixé et quelque soit p :

Valeurs de a	0.06	0.04	0.02	0.01
Valeurs de n	1112	2500	10000	40000

On montre que dans ce cas l'amplitude de l'intervalle de fluctuation est la même que celle de l'intervalle de confiance.Donc, avec un niveau de confiance de 0,95, pour obtenir un intervalle de confiance d'amplitude 0,06, il faut un échantillon de taille 1112 au moins.

Intervalle de la classe de Terminale : $I_n = \left]p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right[$.

On cherche n tel que $2u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leq a$ ce qui équivaut à $n \geq \dfrac{4u_\alpha^2 p(1-p)}{a^2}$. Quelques valeurs :

Pour $p = 0.5$

Valeurs de a	0.06	0.04	0.02	0.01
Valeurs de n si $\alpha = 0,05$	1067	2401	9604	38416
Valeurs de n si $\alpha = 0,01$	1849	4161	16641	66664

Pour $p = 0.1$

Valeurs de a	0.06	0.04	0.02	0.01
Valeurs de n si $\alpha = 0,05$	385	865	3458	13830
Valeurs de n si $\alpha = 0,01$	666	1498	5991	23964

Section 4

Conclusion :

On utilise un intervalle de fluctuation lorsque la proportion p dans la population est connue ou si l'on fait une hypothèse sur sa valeur.

On utilise un intervalle de confiance lorsque l'on veut estimer une proportion inconnue dans une population.