Utilisation de la loi normale en statistiques

Notion d'échantillon de taille *n*

Dans une population, la proportion d'individus ayant un caractère particulier est p.
On prélève au hasard n individus de cette population.
Chaque individu a ou n'a pas ce caractère.
Cette expérience aléatoire peut s'assimiler à un schéma de Bernoulli bien qu'il n'y ait pas exactement indépendance entre les tirages puisqu' une fois la personne est choisie la proportion p change dans la population restante (tirages successifs sans remise au lieu de tirages successifs avec remise).
Si n est petit par rapport au nombre total d'individus dans la population, cette simplification est légitime car p ne change quasiment pas.
Si X_n est la variable aléatoire qui donne le nombre de personnes ayant ce caractère sur les n , alors X_n suit la loi binomiale de paramètres (n,p).
X_n prend ses valeurs dans {0 , 1 , .. , n-1 , n} soit n+1 valeurs possibles.
$F_n = \frac{Xn}{n}$ qui correspond à la fréquence est une variable aléatoire qui prend ses valeurs dans $\{0 , \frac{1}{n} , \frac{2}{n} , .. , \frac{n-1}{n} , 1\}$.
Un échantillon de taille *n* est la réalisation de cette expérience aléatoire et la fréquence observée du caractère dans cet échantillon est donc une valeur prise par F_n.

Intervalle de fluctuation (de fréquence ou d'échantillonnage)

La fréquence observée change d'un échantillon à un autre, c'est ce que l'on appelle la fluctuation de la fréquence ou la fluctuation d'échantillonnage.

Définition 1

L'intervalle de fluctuation au seuil 1 – α est tout intervalle I pour lequel
$P(F_n \in I) \geq 1 - \alpha$ où $\alpha \in]0,1[$.

Par exemple en prenant α = 0,05, I est un intervalle de fluctuation au seuil 0,95 ou au seuil de 95%.
Cela qui signifie que dans au moins 95% des échantillons de taille n la fréquence observée est dans I.

Définition 2

I_n est un intervalle de fluctuation asymptotique au seuil 1 – α quand
$\lim_{n\to+\infty} P(F_n \in I_n) = 1 - \alpha$ (asymptotique car il s'agit d'une limite).

Le programme donne comme intervalle de fluctuation asymptotique au seuil 1-α :
$I_n = [\, p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} , p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \,]$
u_α est le réel strictement positif défini par : $P(- u_\alpha \leq X \leq u_\alpha) = 1-\alpha$ où X suit la loi normale centrée réduite N(0,1).
(Les démonstrations sont faites en TleS à partir du théorème de Moivre Laplace: Voir annexe).

Ainsi en prenant α = 0,05 on a $u_\alpha \approx 1{,}96$ et donc :
$\lim_{n\to+\infty} P(F_n \in [\, p - 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} , p + 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \,]) = 0{,}95$.
En prenant n et p tels que $n \geq 30$ $np \geq 5$ et $n(1\text{-}p) \geq 5$ on a comme approximation
$P(F_n \in [\, p - 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} , p + 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \,]) \approx 0{,}95$ (mais pas forcément $\geq$)
C'est-à-dire, sous ces conditions sur n et p et dans à peu près 95% des cas, la fréquence observée sur un échantillon de taille n est dans $[\, p - 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} , p + 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \,]$.

Remarques

1) Cet intervalle est centré en p et sa taille est d'autant plus petite que n est grand.
Ce dernier point se comprend aisément, en effet, si la taille n de l'échantillon est proche de la taille de la population entière, la fréquence observée sera aussi proche de p !

2) En étudiant sur [0 ; 1] la fonction f définie par $f(x) = \sqrt{x(1-x)}$, on trouve aisément que $\sqrt{p(1-p)}$ est maximale pour p=0,5 et vaut alors 0,5.
Comme 1,96x0,5<1 on obtient que $\frac{\sqrt{p(1-p)}}{\sqrt{n}} < \frac{1}{\sqrt{n}}$
et donc, pour tout p dans [0 ; 1] ,
$[\,p - 1{,}96\,\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,,\ p + 1{,}96\,\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,] \subset [\,p - \frac{1}{\sqrt{n}}\,,\ p + \frac{1}{\sqrt{n}}\,]$
et donc

$$P(F_n \in [\,p - 1{,}96\,\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,, p + 1{,}96\,\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,]) < P(F_n \in \left[\,p - \frac{1}{\sqrt{n}}\,, \quad p + \frac{1}{\sqrt{n}}\right])$$

On retrouve l'intervalle de fluctuation de seconde car le premier membre de cette inégalité vaut à peu près 0,95 (sa limite vaut 0,95 quand n tend vers +∞)
On peut même démontrer qu'il existe un entier n à partir duquel

$$P(F_n \in \left[\,p - \frac{1}{\sqrt{n}}\,, \quad p + \frac{1}{\sqrt{n}}\right]) \geq 0{,}95$$

Pour voir la démonstration

Applications

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.**

Prise de décision

On admet que dans la population d'enfants de 11 à 14 ans d'un département français le pourcentage d'enfants ayant déjà eu une crise d'asthme dans leur vie est de 13%.
Un médecin d'une ville de ce département est surpris du nombre important d'enfants le consultant ayant des crises d'asthme et en informe les services sanitaires. Ceux–ci décident d'entreprendre une étude et d'évaluer la proportion d'enfants de 11 à 14 ans ayant déjà eu des crises d'asthme.
Ils sélectionnent de manière aléatoire 100 jeunes de 11 à 14 ans de la ville.
La règle de décision prise est la suivante : si la proportion observée est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 95% alors une investigation plus complète sera mise en place afin de rechercher les facteurs de risque pouvant expliquer cette proportion élevée.
1) Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeunes de 11 à 14 ans ayant eu une crise d'asthme dans un échantillon de taille 100.
solution : [0,06 ; 0,20]
2) L'étude réalisée auprès des 100 personnes a dénombré 19 jeunes ayant déjà eu des crises d'asthme. Que pouvez-vous conclure ?
Solution : la valeur 0,19 est à l'intérieur de l'intervalle de fluctuation asymptotique au seuil de 95%, On en conclut que la règle de décision choisie ne prévoit pas de réaliser une enquête supplémentaire.
3) Le médecin n'est pas convaincu par cette conclusion et déclare que le nombre de personnes interrogées était insuffisant pour mettre en évidence qu'il y avait plus de jeunes ayant eu des crises d'asthme que dans le reste du département.
Combien faudrait-il prendre de sujets pour qu'une proportion observée de 19% soit en dehors de l'intervalle de fluctuation asymptotique ?
Solution : il faut et il suffit que la borne supérieure de l'intervalle asymptotique de fluctuation soit inférieure à 0,19 ce qui équivaut à $0{,}13 + 1{,}96\,\frac{\sqrt{0{,}13\times 0{,}87}}{\sqrt{n}} < 0{,}19$ soit n > 120
La taille doit donc être de 121 sujets au minimum si on souhaite mettre en évidence une proportion anormalement élevée dans la ville étudiée.

Annexe :

$F_n \in [\, p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, ,\, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,] \Leftrightarrow X_n \in [n\,p - u_\alpha \sqrt{np(1-p)}\, ,n\,p + u_\alpha \sqrt{np(1-p)}\,]$ car $F_n = \frac{X_n}{n}$

donc $P(F_n \in [\, p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, ,\, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,])$

$= P(X_n \in [n\,p - u_\alpha \sqrt{np(1-p)}\, ,n\,p + u_\alpha \sqrt{np(1-p)}\,])$

$= P(Z_n \in [\, -u_\alpha\, ,\, u_\alpha\,]$ où $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$

or $\lim_{n\to+\infty}(\, P(\, Z_n \in [\, -u_\alpha\, ,u_\alpha\,]\,)\,) = P(\, Z \in [\, -u_\alpha\, ,\, u_\alpha\,])$ où Z est la loi normale centrée réduite et d'après le théorème de Moivre Laplace

Mais $P(\, Z \in [\, -u_\alpha\, ,\, u_\alpha\,]) = 1-\alpha$ d'après le cours.

donc $\lim_{n\to+\infty} P(F_n \in [\, p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, ,\, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,]) = 1-\alpha$ (c.q.f.d.)

Intervalle de confiance

On part de ce simple changement d'écriture

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

L'intervalle $\left[\, F_n - \frac{1}{\sqrt{n}}\, ,\, F_n + \frac{1}{\sqrt{n}}\,\right]$ est un intervalle aléatoire : il va varier selon l'échantillon.

A chaque échantillon on obtient un intervalle $\left[\, f - \frac{1}{\sqrt{n}}\, ,\, f + \frac{1}{\sqrt{n}}\,\right]$ où f est la fréquence observée pour cet échantillon.

Or on a vu qu'il existe un rang n à partir duquel $P(F_n \in \left[\, p - \frac{1}{\sqrt{n}}\, ,\, p + \frac{1}{\sqrt{n}}\,\right]) \geq 0{,}95$

donc , pour n suffisamment grand , pour au moins 95% des échantillons de taille n, p se trouve dans $\left[\, f - \frac{1}{\sqrt{n}}\, ,\, f + \frac{1}{\sqrt{n}}\,\right]$

ou encore pour moins de 5% d'entre eux, p ne se trouve pas dans $\left[\, f - \frac{1}{\sqrt{n}}\, ,\, f + \frac{1}{\sqrt{n}}\,\right]$

Définition 3

Soit f la fréquence observée sur un échantillon de taille n alors

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

Remarques

1) Cet intervalle sera utilisé avec comme conditions $n \geq 30$ $nf \geq 5$ et $n(1-f) \geq 5$

2) Il existe d'autres intervalles de confiance au même niveau

3) L'amplitude de l'intervalle de fluctuation asymptotique est évidemment la même que celle de l'intervalle de confiance, donc on peut trouver n tel que cette amplitude soit au plus a

Par exemple pour a = 0,04 on trouve $\frac{2}{\sqrt{n}} \leq 0{,}04$ et donc $n \geq 2500$

ce qui signifie que l'échantillon doit être au moins de taille 2500 pour avoir un intervalle de confiance d'amplitude 0,04

(On pourrait prendre l'intervalle de fluctuation avec $u_\alpha \approx 1{,}96$ pour un seuil de 95%)

4) L'échantillon de taille n qui sera utilisé doit être représentatif.

C'est-à-dire doit correspondre à un tirage aléatoire de n personnes.

Dans la pratique on peut utiliser un autre caractère dont la proportion p' dans la population entière est connue, par exemple femmes ou hommes.

On regarde alors si la fréquence de femmes, par exemple, dans l'échantillon de taille n choisi se trouve bien dans l'intervalle de fluctuation obtenu à partir de n et de p'.

Si oui on utilise cet échantillon pour obtenir l'intervalle de confiance pour l'autre caractère étudié.

Applications

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

Exemple de détermination d'un intervalle de confiance

Prenons un cas très classique : un sondage politique précédant le premier tour d'une élection présidentielle.

Le 18 avril 2002, l'institut IPSOS effectue un sondage dans la population en âge de voter.

On constitue un échantillon de 1000 personnes (inscrites sur les listes électorales) que l'on suppose choisies ici de manière aléatoire. Ce n'est pas le cas en pratique mais le principe reste le même que dans cet exemple.(voir remarque 4)

Les résultats partiels en sont les suivants :

Sur les 1000 personnes

135 ont déclaré vouloir voter pour Jean-Marie Le Pen

195 ont déclaré vouloir voter pour Jacques Chirac

170 ont déclaré vouloir voter pour Lionel Jospin.

On peut déterminer trois intervalles de confiance au niveau de confiance de 95% :

Jean-Marie Le Pen $[0{,}135-0{,}032 ; 0{,}135+0{,}032] = [0{,}103 ; 0{,}167]$

Jacques Chirac $[0{,}195-0{,}032 ; 0{,}195+0{,}032]=[0{,}163 ; 0{,}227]$

Lionel Jospin $[0{,}170-0{,}032 ; 0{,}170+0{,}032]=[0{,}138 ; 0{,}202]$.

Donc la valeur unique en pourcentage donnée par l'institut est entachée d'une imprécision de +/–3 points.

En examinant les trois intervalles trouvés, on peut a posteriori dire que le vrai résultat (16,9%,19,9%,16,2%) est compatible avec ceux-ci pour Jacques Chirac et Lionel Jospin car leurs résultats sont dans les intervalles correspondants. En revanche, le résultat de Jean-Marie Le Pen est légèrement supérieur à la borne supérieure de son intervalle de confiance (mais l'institut CSA lui donnait 14%, ce qui donne un intervalle [0,108 ; 0,172] qui contient son score réel).