Chap 14 – Intervalle de fluctuation / Estimation

L'échantillonnage sert généralement à évaluer la proportion d'une population présentant une certaine caractéristique (sondage pour prévoir l'issue d'une élection) ou à vérifier qu'un groupe (échantillon) est ou n'est pas représentatif de la population à certains égards (est-ce qu'un médicament est efficace ? est-ce qu'une partie de la population est plus prône à développer une certaine pathologie ? etc.).
Généralement la fréquence f observée dans un échantillon n'est pas exactement la même que p celle dans la population dont on elle est extraite. En général elle fluctue autour de p. La question que l'on est amené à se poser c'est « à quel point peut-elle s'éloigner de p avant que l'on déclare l'échantillon incompatible avec notre hypothèse de travail (paramètre de la loi) ? »

I. Rappels :Théorème de Moivre-Laplace

Théorème de Moivre-Laplace
Pour tout nombre entier naturel n, X_n est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n;p)$. On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, variable aléatoire centrée réduite associée à X_n.

Alors pour tous nombre réels a et b tels que $a < b$, $\lim\limits_{n\to+\infty} P(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

Ce théorème nous permettra de faire le lien entre la loi binomiale qui nous servira à créer des échantillons et la loi normale qui nous permettra de les interpréter.

Dans la pratique (domaine de validité)
On ne s'intéressera pas à une limite, mais à une variable aléatoire X donnée.
Si elle est binomiale de paramètres n et p avec $\boldsymbol{n \geq 30}$, $\boldsymbol{np \geq 5}$ et $\boldsymbol{n(1-p) \geq 5}$ alors :

$$P\left(a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b\right) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

II. Intervalles de fluctuation

Rappel :
en seconde et en première on avait définit l'intervalle de fluctuation de la fréquence au seuil de 95% comme étant l'intervalle $I_{95\%}$ centré sur la proportion tel que l'on soit sur à au moins 95% qu'un échantillon non biaisé donne sa fréquence de réussite dans $I_{95\%}$. i.e. $P(f \in I_{95\%}) \geq 0{,}95$
En seconde on avait définit l'intervalle de fluctuation à l'aide de la formule suivante :
$I_{95\%} = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$
En première S : pour trouver l'intervalle de fluctuation lié à une variable aléatoire X_n(suivant une binomiale) on avait cherché a le dernier entier tel que $P(X_n \leq a) \leq 2{,}5\%$ et le b le premier entier tel que $P(X_n \leq b) \geq 97{,}5\%$. On avait alors : $I_{95\%} = \left[\frac{a}{n}; \frac{b}{n}\right]$ car $F_n = \frac{X_n}{n}$ la variable aléatoire associée à la fréquence d'occurrence de la caractéristique vérifie bien $P\left(\frac{a}{n} \leq F_n \leq \frac{b}{n}\right) \geq 95\%$. C'est assez pénible du coup, surtout si n est grand on aimerait avoir une approche rapide comme celle de seconde .

Histoire de rentabiliser nos efforts : on va généraliser l'intervalle de fluctuation au seuil de 0,95 à n'importe quel seuil de la forme $1 - \alpha$. La valeur α correspond à la probabilité tolérable de faire une erreur de type I

autrement dit un faux négatif (c'est-à-dire que l'on considère que l'échantillon et sa fréquence sont incompatibles avec les paramètre de la variable aléatoire , alors que ça n'est pas le cas)

ROC : On sait que si Z suit une loi $\mathcal{N}(0;1)$ pour tout nombre α compris entre 0 et 1 il existe un unique nombre u_α tel que : $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$

Et donc d'après le théorème de Moivre Laplace on sait que $P\left(-u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha\right)$ converge vers $P(-u_\alpha \leq Z \leq u_\alpha)$ cherchons maintenant l'intervalle de fluctuation de la fréquence encadrant $F_n = \frac{X_n}{n}$

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

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

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

Ainsi $P\left(p - u_\alpha\sqrt{\frac{p(1-p)}{n}} \leq \mathrm{F_n} \leq p + u_\alpha\sqrt{\frac{p(1-p)}{n}}\right)$ converge vers $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$

D'où :

Corollaire 1 du théorème de Moivre Laplace :

Soit Z une variable aléatoire suivant la loi $\mathcal{N}(0;1)$, et u_α est le réel tel $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$

Si X_n suit une loi binomiale $\mathcal{B}(n;p)$, alors la fréquence de succès $F_n = \frac{X_n}{n}$ vérifie :

$$\lim_{n\to+\infty} P\left(p - u_\alpha\sqrt{\frac{p(1-p)}{n}} \leq \mathrm{F_n} \leq p + u_\alpha\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Dans l'application pratique un peu plus haut nous avons vu que si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ alors la loi $\frac{X_n - np}{\sqrt{np(1-p)}}$ était équivalente une loi normale et donc nous aurons $\left[p - u_\alpha\sqrt{\frac{p(1-p)}{n}}; p + u_\alpha\sqrt{\frac{p(1-p)}{n}}\right]$pour intervalle de fluctuation et si on se place au seuil de 95% il sera : $\left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}; p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$

Propriété

Il existe un entier n_0 à partir duquel $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ est un intervalle de fluctuation au seuil de 95%¨.

Démonstration

On sait que si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ autrement dit à partir d'un entier n_0 alors $F_n = \frac{X_n}{n}$ admettra comme intervalle de fluctuation au seuil de 95% : $\left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}; p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$

Soit f la fonction définie sur $]0;1[$ par $f(p) = 1{,}96\sqrt{p(1-p)}$ on aura donc $f'(p) = \frac{1{,}96(1-2p)}{\sqrt{p(1-p)}}$

On en déduit que la fonction est croissante jusqu'à ½ et décroissante après donc elle culmine en ½ pour lequel elle prend la valeur $\frac{1{,}96}{2}$ qui est plus petite que 1 donc $\forall p \in]0;1[$ $1{,}96\sqrt{p(1-p)} < 1$ etdonc $1{,}96\sqrt{\frac{p(1-p)}{n}} < \frac{1}{\sqrt{n}}$ et $\frac{-1}{\sqrt{n}} < -1{,}96\sqrt{\frac{p(1-p)}{n}}$ et donc $\left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}; p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$

donc $0{,}95 \leq P\left(F_n \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}; p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]\right) < P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]\right)$ (inégalité stricte car l'inclusion est stricte)

Conclusion pour $\forall n \geq n_0, P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]\right) > 0{,}95$

III. Intervalle de confiance

Plaçons nous dans la situation où l'on ne connaisse pas la probabilité de réussite d'une épreuve (ou encore la proportion p de présence d'une caractéristique dans une population donnée. A l'aide d'un sondage on va chercher à estimer cette valeur.

Un sondage sur n personnes d'une population présentant une proportion p d'une caractéristique donnée est assimilé une variable aléatoire suivant une loi $\mathcal{B}(n; p)$. On sait que pour n suffisamment grand on aura :

$P\left(p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}\right) > 0{,}95$

$p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \Leftrightarrow \begin{cases} p - \frac{1}{\sqrt{n}} \leq F_n \\ F_n \leq p + \frac{1}{\sqrt{n}} \end{cases} \Leftrightarrow \begin{cases} p \leq F_n + \frac{1}{\sqrt{n}} \\ F_n - \frac{1}{\sqrt{n}} \leq p \end{cases} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$

Donc $P\left(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}\right) > 0{,}95$

Définition

Si le sondage sur n personnes nous donne une fréquence de réussite de f alors $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de la proportion p avec un niveau de confiance 0,95.

Remarques :

Cet intervalle n'est pas optimal, par contre $\left[f - 1{,}96\sqrt{\frac{f(1-f)}{n}}; p + 1{,}96\sqrt{\frac{f(1-f)}{n}}\right]$ nous permet de rester avec le même niveau de confiance, tout en gagnant en précision.

Les calculs sont tout de même plus longs, et il est plus compliqué de prévoir la taille minimale de l'échantillon pour obtenir un intervalle d'une précision donnée

Exercice :

On veut un encadrement à un centième près de la valeur de la proportion p, combien de personne doit on interroger au minimum ?

$\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ a pour amplitude $\frac{2}{\sqrt{n}}$ .or on veut un encadrement au centième près donc on veut :

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

On devra donc interroger au moins 40 000 personnes pour avoir la précision attendue.

donc $0{,}95 \leq P\left(F_n \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}; p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]\right) < P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]\right)$ (inégalité stricte car l'inclusion est stricte)

Conclusion pour $\forall n \geq n_0, P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]\right) > 0{,}95$

III Intervalle de confiance

Plaçons nous dans la situation où l'on ne connaisse pas la probabilité de réussite d'une épreuve (ou encore la proportion p de présence d'une caractéristique dans une population donnée. A l'aide d'un sondage on va chercher à estimer cette valeur.

Un sondage sur n personnes d'une population présentant une proportion p d'une caractéristique donnée est assimilé une variable aléatoire suivant une loi $\mathcal{B}(n; p)$. On sait que pour n suffisamment grand on aura :

$P\left(p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}\right) > 0{,}95$

$p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \Leftrightarrow \begin{cases} p - \frac{1}{\sqrt{n}} \leq F_n \\ F_n \leq p + \frac{1}{\sqrt{n}} \end{cases} \Leftrightarrow \begin{cases} p \leq F_n + \frac{1}{\sqrt{n}} \\ F_n - \frac{1}{\sqrt{n}} \leq p \end{cases} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$

Donc $P\left(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}\right) > 0{,}95$

Définition

Si le sondage sur n personnes nous donne une fréquence de réussite de f alors $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de la proportion p avec un niveau de confiance 0,95.

Remarques :

Cet intervalle n'est pas optimal, par contre $\left[f - 1{,}96\sqrt{\frac{f(1-f)}{n}}; p + 1{,}96\sqrt{\frac{f(1-f)}{n}}\right]$ nous permet de rester avec le même niveau de confiance, tout en gagnant en précision.

Les calculs sont tout de même plus longs, et il est plus compliqué de prévoir la taille minimale de l'échantillon pour obtenir un intervalle d'une précision donnée

Exercice :

On veut un encadrement à un centième près de la valeur de la proportion p, combien de personne doit on interroger au minimum ?

$\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ a pour amplitude $\frac{2}{\sqrt{n}}$ .or on veut un encadrement au centième près donc on veut :

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

On devra donc interroger au moins 40 000 personnes pour avoir la précision attendue.