TS — Échantillonnage

Le 3-4-2019

Remarque : En donnant un intervalle de confiance, vous avez 5 % de risque que paramètre ne soit pas dans l'intervalle ou 95 % de chance qu'il y soit. Cela permet de donner une fourchette de valeurs d'un paramètre inconnu.

Seuil de confiance	Risque	t_α
99 %	1 %	2,58
95 %	5 %	1,96
90 %	10 %	1,65

→ Table des lois normales

Le mardi 21-5-2019

Lois normale centrée réduite ; intervalles centrés en 0

Dans tout le paragraphe, on note P la loi normale centrée réduite sur $\mathbb{R}$.

① On note $u_{0,05}$ le réel positif tel que $P\left(\left[-u_{0,05}\,;u_{0,05}\right]\right)=0,95$.

Illustration graphique

On ne peut pas calculer $u_{0,05}$ de manière exacte. Il faut utiliser la calculatrice.

Pour les calculatrices mises à jour, on tape :

invNormale(0.95, 0, 1, CENTER)

Affichage : {– 1.959963986 1.959963986}

On peut aussi utiliser la commande résol .

On retiendra que la valeur arrondie au centième de $u_{0,05}$ est 1,96.

② $P\left(\left[-u_{0,01}\,;u_{0,01}\right]\right)=0,99$

invNormale(0.99, 0, 1, CENTER)

Affichage : {– 2.575829303 2.575829303}

③ Cas général avec α

Lorsque α tend vers 1, u_α tend vers $+\infty$.

④ Lien avec les plages de normalité

Présentation rapide :

Le jeudi 25 avril 2019

Dans la population française, on estime que la proportion de gauchers est d'environ 19 %.
Déterminer un intervalle de fluctuation au seuil de 95 % de la fréquence de gauchers dans un échantillon aléatoire de taille 5000.

X : variable aléatoire qui compte le nombre de gauchers dans un échantillon aléatoire de taille 5000.

X suit la loi binomiale de paramètres $n=5000$ et $p=0,19$.

$\mathrm{E}(\mathrm{X})=np=950$

$\mathrm{V}(\mathrm{X})=n\times p\times q=769,5$ (on laisse en décimal)

Avec la loi binomiale

invBinom

$a=896$ $b=1005$

$\dfrac{a}{n}=0,1792$ $\dfrac{b}{n}=0,201$

$[0,1792\,;0,201]$

Avec la loi normale

On remplace X par une variable aléatoire Y qui suit la loi normale d'espérance $\mu=950$ et de variance $\sigma^2=769,5$ (donc d'écart-type $\sigma=\sqrt{769,5}$).

On cherche l'intervalle centré sur μ contenant 95 % des valeurs.

On trouve 895,6308675 (affichage calculatrice) et 1004,369132.

Les valeurs trouvées correspondent à peu près à celles obtenues par la méthode utilisant la loi binomiale.

On pourrait éventuellement utiliser une correction de continuité.

On divise ces deux valeurs par 5000.

0,1791261735 0,20087382264

$[0,1791261735\,;0,20087382264]$

Rappels sur les notions antérieures :

En 2^e, on a mis en évidence la notion de fluctuation d'échantillonnage (fluctuation d'une fréquence, d'une moyenne, d'un écart-type…) : variabilité naturelle (c'est-à-dire due au hasard) de la distribution de fréquences d'une expérience aléatoire. On a aussi mis en évidence lors de simulations un intervalle de fluctuation pour une fréquence (notamment grâce à des graphiques en nuages de points).

On a vu en 1ère la notion d'intervalle de fluctuation d'une fréquence en utilisant la loi binomiale (méthode des échantillons aléatoires) et on a vu l'application à la prise de décision sur la fréquence d'un échantillon.

Cette année, nous avons étudié les lois normales.

Objectifs de ce chapitre :

- Utiliser la loi normale pour définir un intervalle de fluctuation d'une manière plus simple (notion d'intervalle de fluctuation asymptotique d'une fréquence par calcul) ;

- Appliquer la notion à la prise de décision sur échantillon (on aborde la notion de « test d'hypothèse ») ;

- Aborder la notion d'intervalle de confiance d'une proportion.

Le cours comporte une partie assez théorique.

Le cours s'organise en plusieurs parties :

- Intervalle de fluctuation asymptotique d'une fréquence

- Prise de décision sur échantillon

- Estimation d'une proportion par intervalle de confiance

- Comparaison de deux proportions.

Plan du chapitre :

I. Propriété fondamentale

1°) Énoncé

X_n est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;\,p)$.

On pose $F_n = \dfrac{X_n}{n}$; F_n représente la fréquence de succès pour un schéma de Bernoulli de paramètres *n* et *p*.

Pour tout réel $\alpha \in]0\,;1[$, on a $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1-\alpha$ où I_n désigne l'intervalle

$\left[p - u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\,;\,p + u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\right]$ (intervalle de centre *p* et de rayon $u_\alpha\sqrt{\dfrac{p(1-p)}{n}}$).

Rappel : u_α est le réel positif tel que $P(-u_\alpha \leqslant T \leqslant u_\alpha) = 1-\alpha$ avec T, variable aléatoire qui suit la loi normale centrée réduite (l'indice indique juste que *u* dépend de α).

2°) Démonstration

$$\begin{aligned} F_n \in I_n &\Leftrightarrow p - u_\alpha\sqrt{\frac{p(1-p)}{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha\sqrt{\frac{p(1-p)}{n}} \\ &\Leftrightarrow np - u_\alpha\sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha\sqrt{np(1-p)} \\ &\Leftrightarrow -u_\alpha\sqrt{np(1-p)} \leqslant X_n - np \leqslant u_\alpha\sqrt{np(1-p)} \\ &\Leftrightarrow -u_\alpha \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant u_\alpha \end{aligned}$$

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$, variable aléatoire centrée réduite associée à X_n.

On a donc $P(F_n \in I_n) = P(-u_\alpha \leqslant Z_n \leqslant u_\alpha)$.

D'après le théorème de Moivre-Laplace, $\lim\limits_{n \to +\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = \displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}}\,e^{-\frac{x^2}{2}}\,dx$.

En effet, le théorème de Moivre-Laplace dit que pour tout couple (a,b) de réels tels que $a \leqslant b$ on a :

$\lim\limits_{n \to +\infty} P(a \leqslant Z_n \leqslant b) = \displaystyle\int_a^b \frac{1}{\sqrt{2\pi}}\,e^{-\frac{x^2}{2}}\,dx$).

Or par définition de u_α, $\displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}}\,e^{-\frac{x^2}{2}}\,dx = 1-\alpha$.

On en conclut que $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1-\alpha$.

On notera que la variable aléatoire $F_n = \dfrac{X_n}{n}$ suit une loi qui ne porte pas de nom (mais dont la représentation en diagramme en bâtons ressemble à celle la loi binomiale).

3°) Rappel sur la détermination de u_α (faire un graphique pour visualiser la situation)

$\alpha \in]0\,;1[$

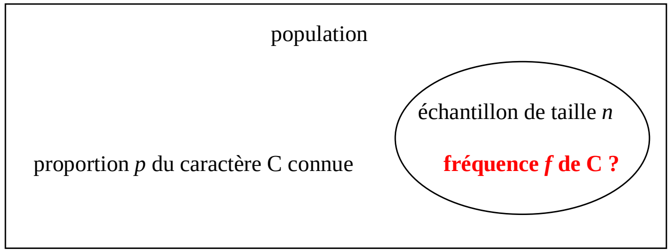
$$\begin{aligned} P(-u_\alpha \leqslant T \leqslant u_\alpha) = 1-\alpha &\Leftrightarrow 2P(0 \leqslant T \leqslant u_\alpha) = 1-\alpha \\ &\Leftrightarrow 2\left[P(T \leqslant u_\alpha) - P(T < 0)\right] = 1-\alpha \quad \text{(utilisation des aires éventuelle)} \\ &\Leftrightarrow 2\left[P(T \leqslant u_\alpha) - 0,5\right] = 1-\alpha \\ &\Leftrightarrow 2P(T \leqslant u_\alpha) - 1 = 1-\alpha \\ &\Leftrightarrow 2P(T \leqslant u_\alpha) - 1 = 1-\alpha \\ &\Leftrightarrow P(T \leqslant u_\alpha) = 1-\frac{\alpha}{2} \end{aligned}$$

$\alpha \in]0\,;1[$ donc $\dfrac{\alpha}{2} \in \left]0\,;\dfrac{1}{2}\right[$.

Par suite, $1-\dfrac{\alpha}{2} \in \left]\dfrac{1}{2}\,;1\right[$.

Par la propriété d'inversion de F, on sait qu'il existe un unique réel $u_\alpha > 0$ tel que $P(T \leqslant u) = 1-\dfrac{\alpha}{2}$.

II. Intervalle de fluctuation asymptotique d'une fréquence



On reprend les notations du paragraphe précédent.

En particulier, on reprend l'intervalle I_n qui a été donné dans la propriété.

1°) Définition

L'intervalle $\left[p - u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\,;\,p + u_\alpha\sqrt{\dfrac{p(1-p)}{n}}\right]$ est appelé l'intervalle de fluctuation asymptotique de la variable aléatoire F_n au seuil $1-\alpha$.

Il s'agit d'un intervalle de centre *p*.

On peut calculer l'amplitude de cet intervalle :

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

2°) Approximation

Cet intervalle contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que *n* est grand.

Nous admettrons sans démonstration (car la démonstration dépasse le cadre du programme) que cette approximation est valable dès que les 3 conditions suivantes sont remplies :

$n \geqslant 30$;
$np \geqslant 5$;
$n(1-p) \geqslant 5$.

On observe sans peine que, plus le seuil est élevé (c'est-à-dire plus le risque α est petit), plus l'amplitude de l'intervalle de fluctuation est grande.

3°) Cas particulier très important : seuil de 95 %

Dans le cas où $\alpha = 0{,}05$, on a : $1-\alpha = 0{,}95$.

L'intervalle de fluctuation asymptotique au seuil de 95 % est $\left[p-u_{0,05}\sqrt{\frac{p(1-p)}{n}}\ ;\ p+u_{0,05}\sqrt{\frac{p(1-p)}{n}}\right]$.

Dans le chapitre sur la loi normale, on a donné une valeur approchée de $u_{0,05}$ à connaître par cœur :

$u_{0,05} \approx 1{,}96$ (valeur arrondie au millième).

On remplace fréquemment $u_{0,05}$ par cette valeur approchée, bien qu'avec les calculatrices actuelles il soit possible d'obtenir une valeur approchée plus précise de $u_{0,05}$.

On peut donc énoncer la propriété suivante à retenir par cœur :

On assimile fréquemment l'intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence d'un caractère dans un échantillon de taille *n* à l'intervalle $\left[p-1{,}96\sqrt{\frac{p(1-p)}{n}}\ ;\ p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$ où *p* désigne la proportion de ce caractère dans la population.

Comme on travaillera la plupart du temps avec un seuil de 95 %, on utilisera très souvent cet intervalle, qu'il faut donc connaître par cœur.

Il s'agit d'un intervalle de centre *p*.

4°) Exemple

La proportion de droitiers dans la population française est de 85 %.
Déterminons l'intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence des droitiers dans les échantillons de taille 400.

$n = 400$

$p = 0{,}85 \qquad 1-p = 0{,}15$

$np = 400 \times 0{,}85 = 340$

$n(1-p) = 400 \times 0{,}15 = 60$

On a donc :

$n \geqslant 30$
$np \geqslant 5$
$n(1-p) \geqslant 5$

Par conséquent, les trois conditions sont vérifiées ; on peut appliquer la formule du cours.
Avec la calculatrice, on trouve :

$p-1{,}96\sqrt{\frac{p(1-p)}{n}} \approx 0{,}815$ (valeur décimale approchée au millième par défaut)

$p+1{,}96\sqrt{\frac{p(1-p)}{n}} \approx 0{,}885$ (valeur décimale approchée au millième par excès)

On en déduit l'intervalle de fluctuation asymptotique cherché : $[0{,}815\ ;\ 0{,}885]$.

On pourrait comparer avec l'intervalle de fluctuation « exact » fourni par la loi binomiale.

Avec la loi binomiale, on trouve, en utilisant les notations habituelles : $a = 326$ et $b = 354$.
On a : $\frac{a}{400} = 0{,}815$ et $\frac{b}{400} = 0{,}885$.
Donc dans ce cas, l'intervalle de fluctuation « exact » fourni par la loi binomiale correspond à l'intervalle de fluctuation asymptotique.

5°) Remarque

Quand les conditions d'utilisation de l'intervalle de fluctuation asymptotique ne sont pas réunies, on détermine un intervalle de fluctuation (intervalle de fluctuation « exact ») avec la loi binomiale.

6°) Autre cas particulier : seuil de 99 %

Dans le cas où $\alpha = 0,01$, on a : $1-\alpha = 0,99$.

Dans le chapitre sur la loi normale, on a donné une valeur approchée de $u_{0,01}$ à connaître par cœur :

$u_{0,01} \approx 2,58$ (valeur arrondie au millième).

On remplace fréquemment $u_{0,01}$ par cette valeur approchée.

On peut donc énoncer la propriété suivante qu'il n'est pas nécessaire de retenir par cœur (quoiqu'elle soit très facile à retenir) :

On assimile fréquemment l'intervalle de fluctuation asymptotique au seuil de 99 % de la fréquence d'un caractère dans un échantillon de taille n à l'intervalle $\left[p-2,58\sqrt{\dfrac{p(1-p)}{n}}\ ;\ p+2,58\sqrt{\dfrac{p(1-p)}{n}}\right]$ où p désigne la proportion de ce caractère dans la population.

III. Intervalle de fluctuation d'une fréquence « simplifié »

1°) Intervalle simplifié au seuil de 95 %

Le maximum de la fonction $x \mapsto x(1-x) = x - x^2$ est égal à $\dfrac{1}{4}$; il est obtenu pour $x = \dfrac{1}{2}$.

On a donc : $\forall x \in \mathbb{R} \quad x(1-x) \leqslant \dfrac{1}{4}$.

Par conséquent, $\sqrt{p(1-p)} \leqslant \dfrac{1}{2}$ d'où $1,96\sqrt{\dfrac{p(1-p)}{n}} \leqslant 2 \times \dfrac{\frac{1}{2}}{\sqrt{n}}$; par suite $1,96\sqrt{\dfrac{p(1-p)}{n}} \leqslant \dfrac{1}{\sqrt{n}}$ donc l'intervalle de fluctuation $\left[p-1,96\sqrt{\dfrac{p(1-p)}{n}}\ ;\ p+1,96\sqrt{\dfrac{p(1-p)}{n}}\right]$ est inclus dans l'intervalle $\left[p-\dfrac{1}{\sqrt{n}}\ ;\ p+\dfrac{1}{\sqrt{n}}\right]$.

L'intervalle de fluctuation « simplifié » d'une fréquence au seuil de 95 % est $\left[p-\dfrac{1}{\sqrt{n}}\ ;\ p+\dfrac{1}{\sqrt{n}}\right]$ avec les notations précédentes.

On retrouve la formule de l'intervalle de fluctuation simplifié donnée en 2^{e} (qui permet d'effectuer les calculs assez facilement sous les conditions qui ont été précisées au préalable).

On peut remarquer que cet intervalle a pour amplitude $\dfrac{2}{\sqrt{n}}$.

On pourrait calculer cet intervalle pour l'exemple des droitiers.

On a vu que l'on peut agrandir l'intervalle de fluctuation asymptotique au seuil 0,95 en majorant $1,96\sqrt{p(1-p)}$ par 1.

On retient la propriété :

X_n est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\ ;\ p)$.

Il existe un entier naturel n_0 tel que si $n \geqslant n_0$, alors $P\left(p-\dfrac{1}{\sqrt{n}} \leqslant \dfrac{X_n}{n} \leqslant p+\dfrac{1}{\sqrt{n}}\right) \geqslant 0,95$.

2°) Intervalle simplifié au seuil de 99 %

De même, on peut agrandir l'intervalle de fluctuation asymptotique au seuil de 99 % en majorant $2,58\sqrt{p(1-p)}$ par $\dfrac{3}{2}$ (en effet, $2,58 < 3$ et $\sqrt{p(1-p)} < \dfrac{1}{2}$).

On obtient l'intervalle de fluctuation « simplifié » au seuil de 99 % : $\left[p-\dfrac{3}{2\sqrt{n}}\ ;\ p+\dfrac{3}{2\sqrt{n}}\right]$.

Cette formule n'est pas à savoir par cœur.

IV. Prise de décision sur échantillon

On aborde dans ce paragraphe la notion de *preuve statistique* (avec la notion de « test d'hypothèse »).

1°) Exemple

On considère qu'une machine qui sert à fabriquer des objets fonctionne de manière satisfaisante si la proportion d'objets défectueux est de 1 pour 1000.
Sur un échantillon de 10 000 objets, on a trouvé 15 objets défectueux.
On veut savoir si la machine est bien réglée.

On va répondre en utilisant les intervalles de fluctuation.
On va répondre avec un seuil de 0,95.

$n = 10000$

$p = 0,001$

$1-p = 0,999$

On a :

$n \geqslant 30$
$np \geqslant 5$
$n(1-p) \geqslant 5$

L'intervalle de fluctuation asymptotique au seuil de 0,95 de la fréquence d'objets défectueux dans un échantillon de taille 10 000 est $\mathrm{I} = [0,0003\ ;\ 0,0017]$.

On énonce la règle de décision (ou « test d'hypothèse ») :

On note f la fréquence d'objets défectueux dans un échantillon aléatoire de taille 10 000.

Si $f \notin \mathrm{I}$ on rejette l'« hypothèse » $p = 0{,}0001$ au seuil de risque de 5 %.

Si $f \in \mathrm{I}$, on ne rejette pas l'hypothèse (on ne dit pas qu'on l'accepte).

On applique la règle.
Ici, $f = 0{,}0015$.
On a $f \in \mathrm{I}$.
On peut considérer que la machine est bien réglée.

2°) Généralisation

On cherche à savoir au seuil de décision de 5 % (ou à un autre seuil précisé à l'avance), si la proportion p du caractère C dans la population vaut $p = p_0$ ou non, à partir d'un échantillon de taille n.
On suppose que $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$.

On procède ainsi :

• On détermine l'intervalle $\left[p_0 - 1{,}96\sqrt{\dfrac{p_0(1-p_0)}{n}} \, ; \, p + 1{,}96\sqrt{\dfrac{p_0(1-p_0)}{n}} \right]$.

• On calcule la fréquence de C sur l'échantillon de taille n prélevé.

• On applique la règle de décision (ou test d'hypothèse) avec un risque de 5 % :

Si $f \notin \mathrm{I}$, on rejette l'hypothèse $p = p_0$.

Si $f \in \mathrm{I}$, on ne la rejette pas.

La probabilité de rejeter à tort l'hypothèse $p = p_0$ est environ égal à 5 %.
Le seuil de décision correspond à ce risque.

3°) Remarques

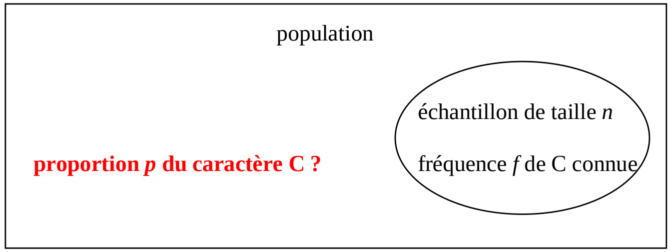
• Lorsque les trois conditions $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$ ne sont pas vérifiées, on utilise l'intervalle de fluctuation à l'aide de la loi binomiale $\mathscr{B}(n\,;p)$ vu en 1ère.

• On peut aussi travailler avec d'autres seuils.

V. Estimation d'une proportion par intervalle de confiance

Ce paragraphe constitue une approche de la *statistique inférentielle*.

1°) Exposé du problème



Attention, ce type de type de schéma ne montre pas le caractère aléatoire de l'échantillon.

2°) Mathématisation

X_n est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;p)$.

On pose $\mathrm{F}_n = \dfrac{\mathrm{X}}{n}$ (comme on l'a dit au début du chapitre, F_n suit une loi qui ne porte pas de nom mais dont la représentation en diagramme en bâtons ressemble à celle la loi binomiale).

On va travailler avec l'intervalle de fluctuation asymptotique simplifié.

Une équivalence importante :

$$p - \frac{1}{\sqrt{n}} \leqslant \mathrm{F}_n \leqslant p + \frac{1}{\sqrt{n}} \Leftrightarrow \mathrm{F}_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant \mathrm{F}_n + \frac{1}{\sqrt{n}}$$

On peut aussi travailler avec la valeur absolue : $\mathrm{F}_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant \mathrm{F}_n + \dfrac{1}{\sqrt{n}} \Leftrightarrow \left| \mathrm{F}_n - p \right| \leqslant \dfrac{1}{\sqrt{n}}$.

D'après la propriété du **III. 1°)**, on a :

Il existe un entier naturel non nul n_0 tel que si $n \geqslant n_0$, alors $P\left(\mathrm{F}_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant \mathrm{F}_n + \dfrac{1}{\sqrt{n}} \right) \geqslant 0{,}95$.

3°) Définition

On note f la fréquence du caractère C dans un échantillon de taille n.

L'intervalle $\left[f-\frac{1}{\sqrt{n}}\,;\, f+\frac{1}{\sqrt{n}}\right]$ est l'intervalle de confiance à 95 % (niveau de confiance) de la proportion inconnue p dans la population.

4°) Conditions de validité

➤ **Condition de validité à vérifier au début (avant calculs) :**

$$n \geqslant 30$$

➤ **Conditions de validité à vérifier *a posteriori* (après calculs) :**

$$np \geqslant 5$$

à vérifier aux bornes

$$n(1-p) \geqslant 5$$

2 conditions suffisent pour vérifier les conditions de validité *a posteriori* :

$$np_{\min} \geqslant 5$$

$$n(1-p_{\max}) \geqslant 5$$

où $p_{\min}$ désigne la borne inférieure (borne de gauche) de l'intervalle de confiance (c'est-à-dire $f-\frac{1}{\sqrt{n}}$) ;

$p_{\max}$ désigne la borne supérieure (borne de droite) de l'intervalle de confiance (c'est-à-dire $f+\frac{1}{\sqrt{n}}$).

Ce sont ces deux conditions que l'on vérifie a posteriori.

5°) Exemple

Dans une population donnée, on s'intéresse à la proportion de blonds.
Dans un échantillon de taille 100 extrait de cette population, on a trouvé 22 blonds.

Déterminer l'intervalle de confiance de la proportion de blonds dans la population au niveau 95 %.

Solution :

$n = 100$

On calculer une estimation ponctuelle de la fréquence dans l'échantillon.

$f = \frac{22}{100} = 0,22$ (fréquence observée de blonds dans l'échantillon)

On vérifie la condition $n \geqslant 30$ avant calcul.

On détermine l'intervalle de confiance I de la proportion de blonds dans cet échantillon.

$$\mathrm{I} = \left[0,22-\frac{1}{\sqrt{100}}\,;\,0,22+\frac{1}{\sqrt{100}}\right]$$

$$\mathrm{I} = [0,12\,;\,0,32]$$

On vérifie les conditions de validité *a posteriori*.

$$np_{\min} \geqslant 5$$
$$n(1-p_{\max}) \geqslant 5$$

On peut donner une interprétation concrète du résultat en disant que la proportion de blonds dans la population est comprise entre 0,12 et 0,32 avec un niveau de confiance de 0,95.

On parle d'**estimation** par intervalle de confiance.

6°) Remarques

• L'intervalle de confiance dépend de la taille de l'échantillon mais pas de la taille de la population.

• La précision de l'intervalle de confiance, donné par sa longueur (c'est-à-dire son amplitude), est égale à : $\frac{2}{\sqrt{n}}$.

Plus n est grand, plus l'amplitude est faible.

Plus le niveau de confiance est élevé, plus l'intervalle de confiance est large.

• Dans certains domaines d'application, on utilise l'intervalle de confiance au niveau 95 % donné par :

$$\mathrm{I} = \left[f-1,96\sqrt{\frac{f(1-f)}{n}}\,;\, f+1,96\sqrt{\frac{f(1-f)}{n}}\right].$$

Il n'est pas possible de justifier cet intervalle dans le cadre du programme.

Il s'agit d'une formule plus précise.

En l'absence d'indication, on utilisera plutôt l'intervalle de confiance $\mathrm{I} = \left[f-\frac{1}{\sqrt{n}}\,;\, f+\frac{1}{\sqrt{n}}\right]$.

L'énoncé dira toujours s'il faut utiliser l'autre formule.

On dira toujours dans les énoncés laquelle utiliser.

• Il existe des formules donnant des intervalles de confiance pour d'autres niveaux. Par exemple, l'intervalle de confiance au niveau 99 % est $\left[f - \frac{1,29}{\sqrt{n}} \,;\, f + \frac{1,29}{\sqrt{n}} \right]$ (formule à savoir).

• Les intervalles de confiance sont utilisés lors d'élections (on parle alors de fourchette de sondage).

• À chaque tirage d'un échantillon, on obtient une valeur différente de *f* et donc un intervalle de confiance différent (cf. simulation d'un sondage sur tableur permettant d'observer différentes fourchettes de sondages).

Deux échantillons différents donneront des intervalles de confiance différents.

VI. Comparaison de deux proportions à l'aide d'intervalles de confiance

1°) Exemple : essai thérapeutique

Pour déceler la présence d'une certaine maladie M chez un individu, on peut utiliser deux tests différents.
Afin de comparer ces tests, deux séries d'observations indépendantes ont été faites.

> Sur 300 personnes présentant effectivement la maladie M, le premier test a décelé la présence de la maladie pour 243 de ces malades.
>
> Sur 300 autres malades, le second test a décelé la présence de la maladie dans 152 cas.

Peut-on admettre que les deux tests ont un pouvoir de détection sensiblement égal ?

1. Soit p_1 la proportion de personnes dont le premier test décèle la présence de la maladie et p_2 la proportion de personnes dont le deuxième test décèle la présence de la maladie.

Déterminer l'intervalle de confiance I_1 au niveau 95 % de la proportion p_1.
Déterminer l'intervalle de confiance I_2 au niveau 95 % de la proportion p_2.

2. Pour comparer ces deux tests, on va étudier l'intersection des deux intervalles trouvés précédemment.

a) Justifier que ces deux intervalles sont disjoints.

b) En déduire au seuil 95 % que la différence observée entre les deux tests est significative.
On peut dire que les deux tests ont un pouvoir de détection différent, au seuil 95 %.

1. On calcule les fréquences dans chacun des deux échantillons.

$f_1 = \frac{243}{300} = 0,81$ donc l'intervalle de confiance de p_1 au niveau de confiance 95 % est : $\mathrm{I}_1 = [0,752\,;\,0,868]$.

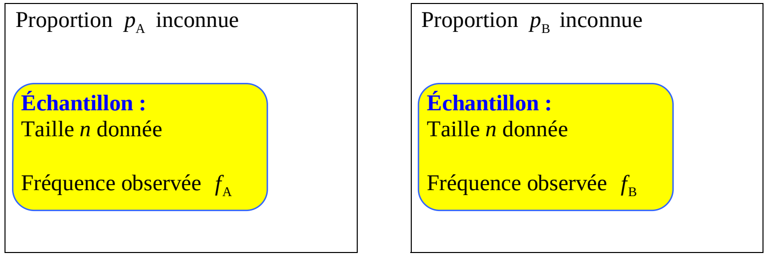
$f_2 = \frac{152}{300}$ donc l'intervalle de confiance de p_1 au niveau de confiance 95 % est : $\mathrm{I}_2 = [0,449\,;\,0,564]$.

2.
a) On a : $\mathrm{I}_1 \cap \mathrm{I}_2 = \varnothing$.

b) I_1 et I_2 sont disjoints donc la différence observée entre les deux tests est significative au seuil de 95 %.

2°) Prise de décision lors de la comparaison de deux proportions

Cas général



Population A **Population B**

La différence entre les deux fréquences observées f_A et f_B est considérée comme significative quand les deux intervalles de confiance à 95 % I_A et I_B sont disjoints.

On juge alors que les deux proportions p_A et p_B sont différentes (avec un petit risque d'erreur).
Dans le cas contraire, on juge que les deux proportions p_A et p_B sont égales (avec un petit risque d'erreur).

Appendice

Notations :

p : proportion dans la population

f : fréquence dans l'échantillon

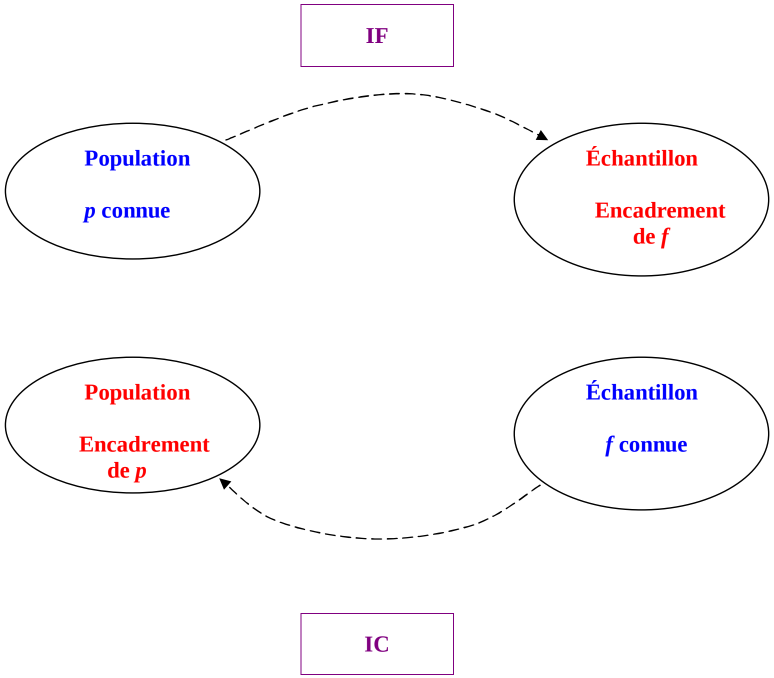
On notera que l'on parle de **proportion** pour la population et de **fréquence** pour l'échantillon.

Intervalle de fluctuation (IF) et intervalle de confiance (IC)

Ne pas confondre IF et IC.

→ L'IF se calcule avec les paramètres vrais de la loi ; il est fixe.

→ L'IC se calcule avec les valeurs observées dans l'échantillon ; il dépend de l'échantillon.



Ces deux schémas montrent bien la différence de démarche :

- prise de décision : de la population vers l'échantillon ;

- estimation : de l'échantillon vers la population.

Les mots qui marchent ensemble :

intervalle de fluctuation → seuil

intervalle de confiance → niveau

Quelques noms importants des statistiques au XX^{e} siècle

Ce sont **Jerzy Neyman** (1894-1988) et **Egon Pearson** (1895-1980) qui ont les premiers utilisé la méthode de prise de décision sur un échantillon décrite dans ce chapitre.

Edwin Bidwell Wilson (1879-1964) est un mathématicien américain, qui a le premier, en 1927, introduit la notion d'intervalle de confiance.

Appendice : Des votes de paille aux e-sondage

Les sondages d'opinion en France n'utilisent pas de tirage aléatoire. Ils sont effectués selon la méthode des « quotas » : des quotas proportionnels à la composition de la population (âge, sexe, catégorie socio-professionnelle, habitat…) sont déterminés puis un nombre correspondant de personnes interrogées.
On peut estimer la fiabilité d'un sondage de taille 1 000 selon la méthode des quotas à celle d'un échantillonnage aléatoire « simple ». Aux États-Unis en revanche, un institut comme Gallup pratique un échantillonnage aléatoire « stratifié » depuis 1952.

Appendice

Programmes permettant de calculer les bornes d'un IF asymptotique et d'un IC au niveau 95 % (permet d'éviter de se tromper dans les calculs)

Afin de gagner du temps, il est vivement recommandé de rentrer des programmes permettant de trouver les bornes de l'intervalle de fluctuation asymptotique.

Plutôt que de prendre 1,96 pour valeur approchée de $u_{0,95}$, il est possible d'utiliser la commande « FracNormale » de manière à obtenir les bornes avec davantage de précision.

Programme donnant l'intervalle de fluctuation asymptotique à un seuil donné par Alexandre Séveno (élève de TS1) le 6-6-2016

: Prompt P,N,F
: $\text{FracNormale}\left(F/2+1/2,0,1\right) \rightarrow U$
: $P-U*\sqrt{\ }\left(\left(P\left(1-P\right)\right)/N\right) \rightarrow X$
: $P+U*\sqrt{\ }\left(\left(P\left(1-P\right)\right)/N\right) \rightarrow Y$
: {1,2 → dim([A]
: X → [A](1,1
: Y → [A](1,2
: Pause [A]

Commentaires :

- P correspond à la probabilité.
- N correspond à la taille de l'échantillon.
- F correspond au seuil de l'intervalle (exemple : pour 95 % on rentre 0,95).
- Pour taper [A], il faut impérativement utiliser la touche $\boxed{\text{matrice}}$ de la calculatrice.

L'utilisation de la matrice ne se justifie ici que pour avoir un affichage qui se rapproche le plus de celui d'un intervalle.

Note de Loïs Balandreaud le 24-5-2017

À propos du U : précision mathématique excellente

En médecine, les intervalles de fluctuations (IF) sont aussi appelés **intervalles de pari** (IP)