INTERVALLE DE CONFIANCE D'UNE PROPORTION

G.R.E.S.

Rappel de notations

Dans une population, le pourcentage des individus qui possèdent un caractère A est p.
On prélève dans cette population un échantillon aléatoire simple de taille n.
On appelle f le pourcentage d'individus possédant le caractère A dans l'échantillon et F la variable aléatoire d'échantillonnage correspondante.

Problème

Nous allons traiter le cas où n est "*grand*".*
Dans ce cas donc, la loi de probabilité de F est approximativement la loi normale de moyenne p et d'écart-type :

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

On note u le nombre tel que

$$\Phi(u) = 1 - \alpha/2$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

On a alors :

$$p\left(p - u\sqrt{\frac{p(1-p)}{n}} < F < p + u\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha \quad (1)$$

d'où

$$p\left(F - u\sqrt{\frac{p(1-p)}{n}} < p < F + u\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

Cette dernière égalité ne permet pas de construire un intervalle de confiance pour p, car celui-ci figure dans les trois membres de la double inégalité.

Comment faire ?

Dans certains ouvrages, on lit :

"On ne connaît pas p, mais on en connaît une estimation, c'est f. On remplace donc p par f dans les bornes de l'intervalle et l'on obtient un intervalle de confiance à (1 – α) avec la formule : (2)

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

Il y a là une sorte de tour de passe-passe un peu rapide. Essayons d'aller plus loin.

Une autre possibilité consiste à trouver un intervalle indépendant de p contenant, quel que soit p, l'intervalle :

$$\left[F - u\sqrt{\frac{p(1-p)}{n}} \ ; \ F + u\sqrt{\frac{p(1-p)}{n}}\right]$$

Pour cela, il suffit de remarquer que, pour tout p, $p(1-p)$ est inférieur à 1/4. On obtient donc l'intervalle de confiance aléatoire :

$$\left[F - \frac{u}{2\sqrt{u}} \ ; F + \frac{u}{2\sqrt{u}}\right]$$

on est sûr que :

$$p\left(F - \frac{u}{2\sqrt{u}} < p < F + \frac{u}{2\sqrt{u}}\right) \geqslant 1 - \alpha$$

mais la précision (c'est-à-dire l'amplitude de l'intervalle) n'est pas la meilleure.

Remarque : Ce résultat permet de comprendre la formule donnée dans le premier Thème de Statistique du **nou-**

(*) Que se passe-t-il pour les petits échantillons ? Les lois normales ne seraient plus d'un grand secours. Il faudrait donc travailler avec des lois binomiales.

veau programme de Seconde, applicable à la rentrée 2000. Dans l'explication de ce thème consacré aux "fourchettes" ou intervalles de confiance d'une proportion, on lit :

"... on incitera les élèves à connaître l'approximation usuelle de la fourchette au niveau de confiance 0,95, issue d'un sondage sur n individus (n > 30) dans le cas où la proportion observée $\hat{p}$ *est comprise entre 0,3 et 0,7, à savoir*

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

$\hat{p}$ représente la proportion constatée sur l'échantillon et n la taille de l'échantillon. On comprend donc la simplification du 1,96 (notre u pour 95%) avec le 2 du dénominateur.

Une autre solution consiste à résoudre de manière rigoureuse la double inéquation en p de la relation (1) : l'événement

$$-u < \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} < u$$

est égal à l'événement

$$\left(\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}\right)^2 < u^2$$

nous allons résoudre cette inéquation du second degré en p, qui peut s'écrire :

$$(n+u^2)p^2 - (2nF+u^2)p + nF^2 < 0.$$

Le coefficient du terme du second degré de ce trinôme du second degré en p est strictement positif, donc, si le discriminant de ce trinôme est positif, l'ensemble des valeurs de p solutions de l'inéquation est l'ensemble des valeurs comprises entre les deux racines du trinôme.

Calculons le discriminant :

$\Delta = (2nF+u^2)^2 - 4(n+u^2)nF^2$

$\Delta = 4n^2F^2 + 4nFu^2 + u^4 - 4n^2F^2 - 4nu^2F^2$

$\Delta = 4nu^2F(1-F) + u^4$

F prenant des valeurs comprises entre 0 et 1, cette quantité est strictement positive, donc le trinôme admet les deux racines suivantes :

$$\frac{(2nF+u^2) - \sqrt{4nu^2F(1-F)+u^4}}{2(n+u^2)} \text{ et}$$

$$\frac{(2nF+u^2) + \sqrt{4nu^2F(1-F)+u^4}}{2(n+u^2)}$$

d'où l'intervalle de confiance :

$$\left[\frac{(2nf+u^2) - \sqrt{4nu^2f(1-f)+u^4}}{2(n+u^2)} \ ; \ \frac{(2nf+u^2) + \sqrt{4nu^2f(1-f)+u^4}}{2(n+u^2)}\right]$$

en essayant de simplifier un peu ces expressions (division par $2n$ des numérateurs et dénominateurs et extraction de u des radicaux) on obtient : (3)

$$\left[\frac{f + \frac{u^2}{2n} - u\sqrt{\frac{f(1-f)}{n} + \frac{u^2}{4n^2}}}{1+\frac{u^2}{n}} \ ; \ \frac{f + \frac{u^2}{2n} + u\sqrt{\frac{f(1-f)}{n} + \frac{u^2}{4n^2}}}{1+\frac{u^2}{n}}\right]$$

On constate que, pour obtenir l'intervalle de confiance (2), on est amené à négliger certains termes(*).

Lorsqu'on utilise la formule :

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

on procède à deux approximations : l'approximation d'une loi binomiale par une loi normale, et le fait de négliger certains termes dans la formule (3).

C'est pour cette raison que, dans certains ouvrages(**), on préconise les conditions suivantes :

$$nf \geqslant 20 \text{ et } n(1-f) \geqslant 20. \quad (4)$$

Remarque : Le problème qui se pose

(*) Si on appelle $[a_n \ ; \ b_n]$ l'intervalle (2) et $[a'_n ; b'_n]$ l'intervalle (3), on démontre assez facilement que $(a_n - f)$ et $(a'_n - f)$ d'une part, $(b_n - f)$ et $(b'_n - f)$ d'autre part, sont équivalents lorsque n tend vers $+\infty$.

(**)P. DAGNELIE : *Statistique Théorique et Appliquée.*

ici est de nature très différente de celui qui se pose pour l'intervalle de confiance d'une moyenne lorsque l'écart type de la population n'est pas connu. Ce qui fait la spécificité du cas des proportions est le fait que p se trouve dans les trois membres de l'inégalité, il ne s'agit donc pas de remplacer, dans les deux membres extrêmes p par son estimation f ni $\frac{p(1-p)}{n}$ par son estimation, qui serait $\frac{f(1-f)}{n-1}$ (en utilisant un estimateur non biaisé), ni $\sqrt{\frac{p(1-p)}{n}}$ par une estimation.

Il est d'ailleurs à remarquer que, s'il est pratique de dire aux élèves pour une moyenne, lorsque l'écart type de la population n'est pas connu : « *on remplace σ par son estimation et on utilise une loi de Student* », ceci ne correspond pas à la réalité mathématique qui est derrière(*). De plus, S est un estimateur biaisé de σ.

(*) cf. Bulletin du GRES n° 7 : *"Résumé sur les lois de probabilité"*

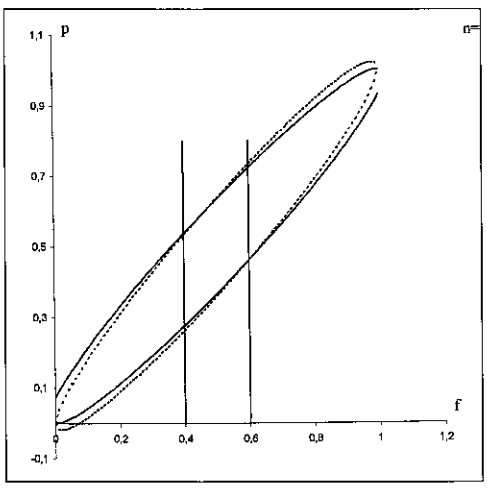
Exemple : déterminons les intervalles de confiance avec les formules (3) puis (2) pour un échantillon sur lequel on a constaté une proportion $f = 0,8$, le premier de taille 50, le deuxième de taille 100. On obtient :

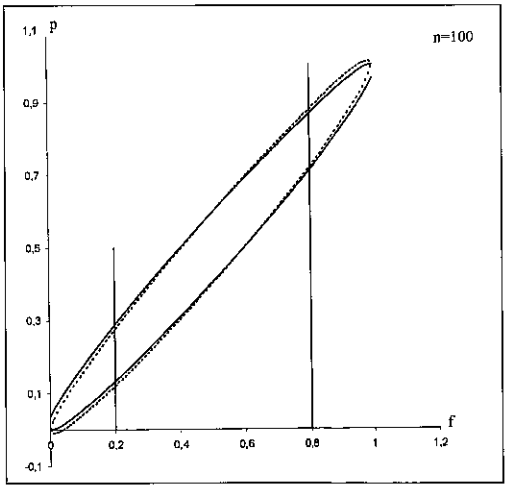
taille 50 : [0,669 6 ; 0,887 6]
et [0,689 1 ; 0,910 9]
pour l'approximation.

taille 100 : [0,711 2 ; 0,866 6]
et [0,721 6 ; 0,878 4]
pour l'approximation.

On constate une meilleure approximation pour $n = 100$.

A titre d'exemple, les deux graphiques qui suivent, réalisés avec Excel, représentent, pour $n = 50$ et $n = 100$, les courbes donnant p en fonction de f, en traits pleins, les bornes des intervalles de confiance à 95% calculées avec la formule (3), et en traits interrompus avec la formule (2).
Les traits verticaux correspondent aux conditions (4).





GREI : Groupe de **R**éflexion sur l'**E**nseignement de l'**I**nformatique
http://enfa.mip.educagri.fr/grei/liens.html
GRES : Groupe de **R**éflexion sur l'**E**nseignement des **S**tatistiques.
http://enfa.mip.educagri.fr/enfadraf/gres/gres.html