L'intervalle de confiance pour une proportion et son enseignement dans le secondaire

Turlot, Jean-Christophe [1], et Petiot, Jean-François [2]

[1] Université de Pau et des Pays de l'Adour, Institut Universitaire de Technologie (IUT) de Pau, UMR 5142, Pau, France
[2] Université de Bretagne-Sud (UBS), Institut Universitaire de Technologie (IUT) de Vannes, UMR 6205, Vannes, France
jean-christophe.turlot@univ-pau.fr

Résumé : La rénovation des programmes de probabilité et de statistique dans l'enseignement secondaire propose d'introduire la notion d'intervalle de confiance pour une proportion dans les classes de terminales. Une capacité attendue est d'estimer une proportion inconnue par l'intervalle :

$$IC = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right]$$

où *f* est la fréquence observée sur un échantillon de taille *n.*
Les commentaires donnés dans le BO [Bulletin officiel] sont les suivants : cet intervalle contient la proportion dans au moins 95% des cas pour *n* grand, ce qui peut être vérifié par simulation. Les conditions d'applications données sont : $n > 30, \quad np > 5, \quad n(1-p) > 5$.

L'objet de cette présentation est de montrer que cette expression non standard est critiquable en ce sens qu'elle ne peut être expliquée dans le cadre de la démarche de la statistique inférentielle classique qui vise à la recherche d'un intervalle de confiance le plus précis possible. De l'équivalence entre un intervalle de confiance $IC(X)$ de niveau η pour une proportion p et une famille de régions d'acceptation $\{A(p_0);0 < p_0 < 1\}$ de niveau η pour le problème de test $H_0 : p = p_0$ pour tout p_0, on déduit les propriétés de l'intervalle $[0,1]$ de confiance dans le cadre de la théorie des tests. L'élève qui appliquera la formule ci-dessus dans le cadre d'un sondage a bien des chances de trouver un intervalle recouvrant des valeurs de p hors de l'intervalle $[0,1]$. Le professeur pourra être également surpris lors des simulations préconisées. En effet, La probabilité de recouvrement de la vraie valeur du paramètre lorsque n est modéré ou même assez grand ($30 \le n \le 100$) peut être éloignée de la valeur attendue en faisant appel à la loi normale. Ceci est encore vrai lorsque l'on utilise l'intervalle de confiance standard de Wald enseigné en première année de STID [STatistique et Informatique Décisionnelle] :

$$IC_S = \left[\hat{p} - \phi(1-\alpha/2)\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}, \hat{p} + \phi(1-\alpha/2)\frac{\hat{p}(1-\hat{p})}{\sqrt{n}} \right]$$

de niveau de confiance $\eta = 1-\alpha$, où ϕ représente la fdr [fonction de répartition] de la loi normale standardisée.
En effet, sous les conditions standard $n > 30, \quad np > 5, \quad n(1-p) > 5$, le biais peut être élevé : par exemple, si $p = 0.2$, il faut que n dépasse *118* pour être sûr que l'intervalle de confiance de niveau nominal $\eta = 0.95$ soit en réalité d'un

niveau de confiance dépassant $\eta = 0.93$. Cela est dû au fait que cet intervalle n'est pas correctement centré d'une part, qu'il y a des oscillations dues au fait que l'on approxime une loi discrète par une loi normale d'autre part. En suivant L.D. Brown, T.T. Caland, et A. Dasgupta (Annals of Statistics, 2002), on montre comment un développement d'Edgeworth à l'ordre 2 accélère la vitesse de convergence vers la loi normale en éliminant les deux sources de biais citées ci-dessus.

Des statistiques autour de l'IMC et de la météo de la 6ème à la terminale

Garat, Philippe [1,2], Girod, Florent [1], Jacquemoud, Damien [1], et Letué, Frédérique [1, 2]

[1] IREM de Grenoble, Institut de Recherche sur l'Enseignement des Mathématiques, Grenoble, France

[2] Université Pierre Mendès-France (UPMF), Institut universitaire de Technologie 2 (IUT 2), département STID STatistique et Informatique décisionnelle, Laboratoire Jean Kuntzmann (LJK), Grenoble, France

Frederique.Letue@iut2.upmf-grenoble.fr

Résumé : Les programmes de lycées publiés récemment en France et qui seront appliqués dans leur totalité à la rentrée scolaire 2012 font la part belle aux probabilités et à la statistique, laissant parfois démunis les enseignants de mathématiques, peu formés dans ce domaine. Nous présentons ici des activités de statistique proposées à des élèves de collège ou lycée autour de deux thèmes cités comme thèmes de convergence dans l'introduction commune des programmes de mathématiques du collège (BO [Bulletin officiel] spécial n° 6 du 28 août 2008 [1]) : météorologie et climatologie (thème 4) et santé (thème 5). Nous traitons ces thèmes à travers la notion d'IMC ([Indice de Masse Corporelle] ; voir définition en [2]), dont des courbes sont présentes dans les carnets de santé des enfants et de données météorologiques de certaines stations météorologiques (Embrun, Besançon).

Nous montrons comment nous déclinons ces thèmes selon le niveau des élèves et les notions du programme. Le but des activités est de faire passer la notion d'adéquation à une loi sur des exemples concrets, sans passer par la théorie formelle des tests statistiques. Sur le thème de l'IMC, il s'agit de savoir si les IMC d'élèves (recueillis par sondage anonyme dans le collège ou de toute autre manière) sont bien conformes aux courbes du carnet de santé ou de l'INPES [3]. Sur le thème de la météo, il s'agit de regarder si les données météorologiques recueillies dans une station donnée sur les quelques dernières années (par exemple 2007-2010) sont conformes aux "normales saisonnières" calculées sur 30 ans (par exemple 1961-1990, voir [4]). Nous nous appuyons essentiellement sur les outils de statistique descriptive au collège et sur la loi binomiale au lycée. Au final, ces activités auront permis de travailler les notions suivantes :

Au collège :

- lecture et interprétation de données présentées sous forme de tableaux, effectifs, proportionnalité, fréquences, ... ;
- diagrammes en bâtons, histogrammes ;
- médianes, quartiles ;