APPROFONDISSEMENT POUR LE PROFESSEUR

Ce paragraphe participe à la formation des enseignants, nécessaire pour acquérir le recul permettant d'enseigner cette partie nouvelle du programme.

- FLUCTUATION DE LA FRÉQUENCE SELON LES ÉCHANTILLONS

On donne ici une justification de l'approche fréquentiste des probabilités par la loi des grands nombres.

On peut modéliser de nombreuses situations aléatoires à l'aide de « l'urne de Bernoulli », c'est-à-dire d'une urne comprenant deux sortes de boules. La problématique est alors résumée par la question suivante :

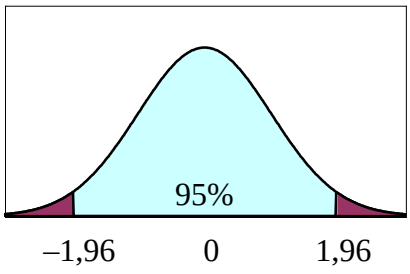
> « Combien faut-il tirer de boules dans une urne de Bernoulli pour pouvoir faire une estimation de sa composition avec une précision donnée a priori ? »

On considère une urne comprenant deux sortes de boules, noires et blanches, et où la proportion des boules noires est p. On effectue n tirages au hasard et avec remise (afin de ne pas modifier la structure de l'urne) dans cette urne. Le résultat est nommé échantillon aléatoire de taille n.
On note X la variable aléatoire correspondant au nombre de boules noires dans un échantillon aléatoire de taille n. Cette variable suit la loi binomiale de paramètres n et p dont l'espérance est $E(X) = np$ et l'écart type $\sigma(X) = \sqrt{n\,p\,(1-p)}$.
De manière à pouvoir comparer des tirages de tailles différentes, il est préférable, plutôt que de noter le nombre de boules noires obtenues, d'en considérer la fréquence. On introduit donc la variable aléatoire $F = \frac{1}{n} X$.

Pour cette variable aléatoire F, on a comme espérance $E(F) = \frac{1}{n} E(X) = \frac{1}{n} \times np = p$ et comme écart type $\sigma(F) = \frac{1}{n} \sigma(X) = \frac{1}{n} \sqrt{n\,p\,(1-p)} = \sqrt{\frac{p(1-p)}{n}}$.
L'interprétation de ces résultats est que si l'on prélève un grand nombre d'échantillons aléatoires de tailles n et que l'on considère la distribution des fréquences observées, ces fréquences ont pour moyenne p (la fréquence dans l'urne) et pour écart type $\sqrt{\frac{p(1-p)}{n}}$. C'est bien sûr cet indicateur de dispersion qui rend compte de la qualité d'un échantillon de taille n pour témoigner de la fréquence p dans l'urne : plus n est grand, plus l'information est précise, mais, ce qui est important c'est que le gain en qualité d'information est en « un sur racine de n ».
Pour préciser ceci, on peut recourir à la loi normale.

Pour n « assez grand » (dans la pratique, on prend n supérieur ou égal à 30, np et $n(1-p)$ supérieurs ou égaux à 5), la loi binomiale donne des résultats proches de ceux d'une loi normale.
On peut donc considérer que la variable aléatoire F suit approximativement une loi normale de moyenne p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$.

On sait que la loi normale a comme propriété qu'environ 95 % des observations se situent dans un intervalle de rayon deux écarts types autour de la moyenne (1,96 fois l'écart type pour être plus précis).



Si T suit la loi normale de moyenne 0 et d'écart type 1, on a :
$P(-\mathbf{1,96} \leq T \leq \mathbf{1,96}) = \frac{1}{\sqrt{2\pi}} \int_{-1,96}^{1,96} e^{-\frac{x^2}{2}}\,dx \approx \mathbf{0,95}.$

On peut donc vérifier qu'environ 95 % des échantillons aléatoires de taille n fournissent une fréquence comprise dans l'intervalle :

$$[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}\ ,\ p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\ ].$$

Ce résultat est très important car il mesure la variabilité « naturelle » des phénomènes aléatoires.
On peut donner une version simplifiée de cet intervalle, en le majorant.

La fonction $p \mapsto p(1-p)$ atteint son maximum pour $p = \frac{1}{2}$ donc, pour tout p, on a :

$p(1-p) \leq \frac{1}{4}$. On en déduit que $1{,}96\sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{\sqrt{n}}$.

Ainsi, l'intervalle $[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}\ ,\ p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\ ]$ est inclus dans l'intervalle

$[p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\ ]$.

On peut expérimenter, par simulation à l'aide des T.I.C., qu'environ plus de 95 % des échantillons de taille n fournissent une fréquence comprise dans l'intervalle de fluctuation :

$[p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\ ]$.

C'est ce qui figure dans le module 1.2 de première professionnelle. Il ne s'agit bien sûr que d'une expérimentation, aucune justification théorique n'est à apporter aux élèves.

Estimation de la valeur de p lorsque celle-ci est inconnue

D'après ce qui précède, on a, en termes de probabilité :

$$P(p - \frac{1}{\sqrt{n}} \leq F \leq p + \frac{1}{\sqrt{n}}) \geq 0{,}95$$

(sous réserve de l'approximation normale, c'est-à-dire $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$), où F est la variable aléatoire correspondant à la fréquence de boules noires observée sur un échantillon de taille n.
Ceci équivaut à (on soustrait $p + F$ et on multiplie par – 1) :

$$P(F - \frac{1}{\sqrt{n}} \leq p \leq F + \frac{1}{\sqrt{n}}) \geq 0{,}95.$$

Cette relation conduit à la notion d'intervalle de confiance, ou de « fourchette » (cette notion n'est pas au programme des baccalauréats professionnels).

Si, sur un échantillon aléatoire de taille n, on observe une fréquence f (autrement dit la variable aléatoire F prend la valeur f), on prendra comme fourchette pour p, au niveau de confiance de 95 %, l'intervalle :

$[f - \frac{1}{\sqrt{n}}\ ,\ f + \frac{1}{\sqrt{n}}\ ]$.

Dans plus de 95 % des cas, la fourchette recouvre effectivement la valeur p.

Une réponse à la question posée (combien faut-il tirer de boules dans une urne de Bernoulli pour pouvoir faire une estimation de sa composition avec une précision donnée a priori ?) est donc : en tirant n boules avec remise, on obtient une estimation de p par un intervalle d'amplitude $\frac{2}{\sqrt{n}}$, avec une confiance de plus de 95 %.
Si l'on tire n = 1 000 boules (avec remise) on a une estimation de p, à plus de 95 % de confiance, par un intervalle d'amplitude 6 %. Si par exemple le tirage de 1000 boules avec remise fournit une fréquence de boules noires égale à 0,47 , on peut estimer avec plus de 95 % de confiance, que la proportion p de boules noires dans l'urne est comprise entre 0,44 et 0,50. Les sondages, par exemple, sont souvent pratiqués sur un échantillon d'environ 1 000 personnes.

- Comment simuler un tirage dans une urne de Bernoulli à l'aide des T.I.C. ?

On atteint vite les limites de l'expérimentation physique. Si l'on veut étudier les lois du hasard, pour connaître la « variabilité naturelle » des échantillons de taille *n*, il faut un très grand nombre d'expériences. Un autre objectif est d'augmenter la taille des échantillons pour estimer des probabilités. La loi des grands nombres nécessite d'être expérimentée à l'aide des T.I.C., par simulation.

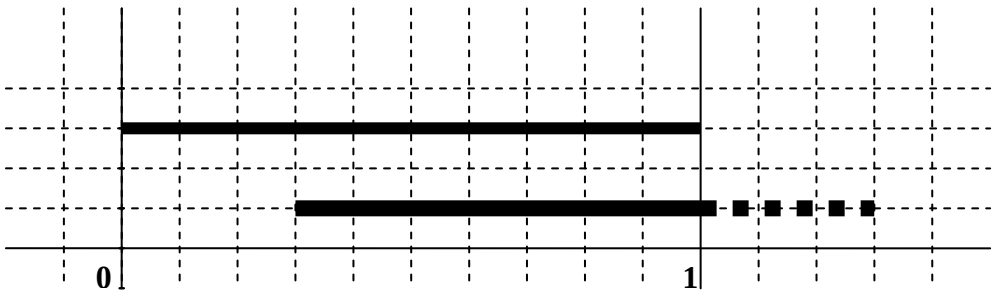
La plupart des calculatrices, même les modèles « collège », et les tableurs sont pourvus d'un générateur de nombres pseudo-aléatoires qui simule le tirage d'un nombre décimal au hasard dans l'intervalle [0, 1[(c'est-à-dire une réalisation d'une variable aléatoire de loi uniforme sur l'intervalle [0, 1[ ). Sur une calculatrice, ce générateur de nombres aléatoires correspond à la fonction « random » souvent notée rand ou Ran#. Sur un tableur, il correspond à la fonction ALEA(), introduite avec des parenthèses vides.

À partir de ce générateur de nombres aléatoires, on peut facilement simuler un tirage au hasard dans une urne bicolore.

Supposons que la proportion des boules dont la couleur nous intéresse soit dans l'urne de 30 %.

Sur un tableur, l'instruction =ALEA()+0,30 correspond au tirage au hasard d'un nombre de l'intervalle [0,30 ; 1,30[. En prenant la partie entière, le résultat vaut 0 si le nombre appartient à [0,30 ; 1[et 1 si le nombre appartient à [1 ; 1,30[.

Compte-tenu des longueurs respectives de ces intervalles, on a 70 % de chances d'avoir 0 et 30 % de chances d'avoir 1 (voir le graphique).



Sur une calculatrice, il suffit d'entrer l'instruction rand + 0.3 ou Ran# + 0.3 et de faire plusieurs fois ENTER ou EXE pour simuler des tirages avec remise dans cette urne, en ne tenant compte que de la partie avant la virgule (0 ou 1).

Sur un tableur, il suffit d'entrer dans une cellule la formule =ENT(ALEA()+0,3) puis d'approcher le pointeur de la souris du coin inférieur droit de la cellule. Lorsque le pointeur de la souris prend la forme d'une croix noire, on enfonce le bouton gauche puis on « glisse » vers le bas pour constituer un échantillon (on nomme « recopie » cette manipulation). On peut ensuite sélectionner l'échantillon (avec le pointeur en forme de croix blanche) puis le recopier vers la droite (avec le pointeur en forme de croix noire) pour constituer plusieurs échantillons.

- Éléments de bibliographie pour la partie statistique et probabilités

- CHAPUT (Brigitte) et HENRY (Michel) – ***Statistique au lycée volume 1*** – Brochure n° 156 APMEP 2005.
- CHAPUT (Brigitte) et HENRY (Michel) – ***Statistique au lycée volume 2*** – Brochure n° 167 APMEP 2007.
- DGESCO – Documents d'accompagnement des programmes de collège téléchargeables sur eduscol.education.fr :
  Organisation et gestion des données (janvier 2007) ;
  Probabilités (17 mars 2008).
- DOWEK (Gilles) – ***Peut-on croire les sondages ?*** – Le Pommier 2002.
- DROESBEKE – TASSI – ***Histoire de la statistique – « Que sais-je ? »*** n°2527 - PUF.
- DUTARTE (Philippe) – ***L'induction statistique au lycée illustrée par le tableur*** – Didier 2005.
- PIEDNOIR (Jean-Louis) et DUTARTE (Philippe) – ***Enseigner la statistique au lycée : des enjeux aux méthodes*** – Brochure n° 112 de la Commission inter-IREM lycées techniques – IREM de Paris-Nord 2001.
- ROBERT (Claudine) – ***Contes et décomptes de la statistique. Une initiation par l'exemple*** – Vuibert 2003.
- SAPORTA (Gilbert) – Probabilités***, analyse des données et statistiques*** –Ed. Technip.
- SCHWARTZ (Claudine) – ***Pratiques de la statistique*** – Vuibert 2006.
- WONNACOTT et WONNACOTT – ***Statistique*** – Ed. Economica.

Sur Internet : http://www.statistix.fr.