Simulation informatique et modélisation
Simulation d'un sondage

Michel Henry
CII *Statistique et probabilités*
IREM de Franche-Comté, Besançon

Abstract :

In this work, we put the question of the didactical status of computational simulations of random phenomenon. The computer's power as random numbers generator allows an experimental approach of sampling fluctuations in the class room, and asks for numerical models to simulate random experiences. In this setting, we can introduce new experimental methods to solve probability problems.

I – Le statut didactique de la simulation

1.1. Quelques questions didactiques posées par la simulation informatique

Dans cet atelier, nous posons la question du statut didactique de la simulation informatique d'un échantillonnage aléatoire dans une population statistique.

L'option expérimentaliste très marquée de la présentation des programmes de statistique et probabilités des lycées pose de nombreuses questions pour l'organisation de cet enseignement. On peut notamment s'interroger sur l'efficacité didactique des situations nées de la présence massive de l'ordinateur dans la classe :

- simulation informatique de données statistiques, approche expérimentale de leur gestion et compréhension de concepts de base : prélèvements aléatoires dans une population, échantillonnage, caractères, distributions de fréquences, fluctuations d'échantillonnage,
- apports de l'ordinateur comme générateur puissant de nombres au hasard (suites de chiffres pseudo-aléatoires) et introduction à la modélisation probabiliste d'expériences aléatoires,
- résolution expérimentale de problèmes et savoirs théoriques impliqués.

1.2. Remarques sur la simulation informatique

Une définition :

> La simulation est *l'expérimentation sur un modèle.* C'est une procédure de recherche scientifique qui consiste à réaliser une reproduction artificielle *(modèle)* du phénomène que l'on désire étudier, à observer le comportement de cette reproduction lorsque l'on fait varier expérimentalement les actions que l'on peut exercer sur celle-ci, et à en induire ce qui se passerait dans la réalité sous l'influence d'actions analogues. (Encyclopédie Universalis).

Les programmes proposent de multiplier les observations de l'aléatoire par la simulation informatique. Le document d'accompagnement du programme de première précise page 72 :

> « …modéliser consiste à associer un modèle à des données expérimentales, alors que simuler consiste à produire des données à partir d'un modèle prédéfini. On parlera de simulation d'une loi de probabilité… ».

L'ordinateur est-il un véritable générateur de hasard ? Non. Actuellement, les chiffres pseudo-aléatoires que l'ordinateur ou la calculatrice fournissent sont déterminés, dès lors qu'il ont commencé un calcul sur la base d'une initialisation qui détermine à chaque fois des suites différentes de chiffres, en principe équirépartis[18]. Mais la complexité de leur calcul rend impossible leur prévision par tout autre moyen humain. « *On a alors le phénomène fortuit »* selon l'expression de Henri Poincaré. Tout se passe donc comme si les chiffres fournis par la fonction ALEA de l'ordinateur ou RANDOM de la calculatrice étaient issus d'un tirage au hasard de boules numérotées de 0 à 9 dans une urne.

Mais, par exemple, en "simulant" un sondage sur Excel, ayant introduit la valeur p de la proportion à estimer dans l'ordinateur comme paramètre d'une loi de Bernoulli à simuler, que pouvons-nous réellement obtenir ? On peut seulement vérifier que le résultat de l'estimation de p issue d'un échantillonnage confirme que le fabricant de l'ordinateur et le concepteur du logiciel ont bien rempli leurs cahiers des charges.

Cependant, ne négligeons pas l'intérêt de l'outil informatique. Dans notre exemple, il permet une approche expérimentale des situations de sondages, et si la proportion p est cachée aux élèves, nous avons un outil de résolution de problèmes jouant le même rôle que les calculettes graphiques quand elles tracent des courbes représentatives de fonctions données.

Mais notre question est plus fondamentalement didactique. L'ordinateur nous permet d'abord de travailler rapidement sur de vastes séries statistiques, ce qui donne aux résumés statistiques (moyennes, écarts types, etc.) toute leur importance. Il nous permet ensuite une présentation animée du fonctionnement et de l'interaction des notions de fréquence et de probabilité abordées auparavant, soit théoriquement, soit au travers d'expériences pratiques en nombre trop limité pour pouvoir déboucher valablement sur une bonne compréhension de la loi des grands nombres.

Même si, dans son fonctionnement réel, l'ordinateur se limite à exhiber les effets sur les fréquences affichées du principe d'équirépartition des chiffres pseudo aléatoires qu'il génère, l'usage que nous en faisons dans la classe, comme « pseudo-générateur de hasard », permet de faire comprendre *en actes* ces notions de fréquence, de fluctuations d'échantillonnages et de probabilité.

Mais on ne peut se satisfaire de la seule exploitation de la puissance et la rapidité de l'ordinateur pour exhiber des nombres aléatoires, permettant de présenter aux élèves une grande richesse de nouvelles expériences aléatoires, car cela n'aurait qu'un intérêt limité. Son intérêt didactique, en tant qu'outil de simulation, tient plus essentiellement en ce qu'il nous oblige à analyser la situation aléatoire en jeu, à émettre des hypothèses de modèle, notamment sur la loi de probabilité idoine pour représenter l'intervention du hasard dans l'expérience réelle (par exemple pour un sondage, sur le choix de la valeur de la probabilité *p* de Bernoulli à implanter), et à traduire ces hypothèses en instructions informatiques pour que l'ordinateur nous permette de résoudre des problèmes éventuellement inaccessibles par le calcul a priori.

De ce point de vue, cela suppose la compréhension du processus de modélisation. Cela suppose aussi une connaissance théorique permettant d'interpréter les résultats obtenus expérimentalement, rapportés aux hypothèses de modèle introduites. Simulant ainsi une expérience aléatoire réelle, l'ordinateur, au-delà de l'exploration statistique, devient un outil didactique majeur pour l'apprentissage de la modélisation en statistique et probabilités.

---

[18] La production de chiffres aléatoires équirépartis fait l'objet de l'article de Bernard PARZYSZ : *Quelques questions à propos des tables et des générateurs aléatoires* dans le 1er volume de *Statistique au lycée,* brochure APMEP n° 156, p. 181.

1.3. Un modèle de base en probabilités : l'urne de Bernoulli

Les situations aléatoires les plus élémentaires de la réalité consistent en la réalisation ou non d'un événement fortuit à l'issue d'une expérience aléatoire. Toutes ces situations, d'un point de vue probabiliste, peuvent être représentées par un *modèle d'urne de Bernoulli*. C'est une urne fictive, parfaite, en ce sens que les boules qu'elle contient sont conçues comme rigoureusement identiques, ne différant que par la couleur, blanche ou noire par exemple. *Tirer une boule de cette urne* est une expérience de pensée qui repose sur l'*hypothèse de modèle* que toutes les boules ont « la même chance » d'être tirées (équiprobabilité postulée). Dans le vocabulaire de la réalité, l'urne de Bernoulli est un objet idéalisé, muni de propriétés théoriques implicites. Je le qualifie de " modèle pseudo concret ".

Un seul paramètre caractérise une urne de Bernoulli : la proportion *p* des boules blanches parmi l'ensemble des boules.

La formalisation de ce modèle (son expression dans le registre symbolique des mathématiques) consiste à considérer un ensemble fini Ω, abstrait (si on veut, on peut considérer que ses éléments représentent les tirages éventuels des différentes boules de l'urne de Bernoulli), sur lequel on considère la distribution uniforme (équiprobabilité) de la probabilité, notée P et appelée à ce niveau « loi de probabilité sur Ω ». On distingue ensuite une partie *A* de Ω (représentant les boules blanches) et on traduit l'hypothèse de modèle par la donnée Card (*A*) = *p*. Card (Ω). L'axiomatique du modèle probabiliste général conduit alors à P(*A*) = *p*. On a ainsi constitué un *modèle probabiliste* de la situation aléatoire donnée.

1.4. Processus de modélisation

Dans un processus de modélisation, je distingue trois étapes, pour l'analyse didactique. Chacune de ces étapes relève d'objectifs différents et donc de contrats didactiques différents.

Si l'on veut introduire en mathématiques une véritable démarche expérimentale, il convient de ne pas négliger la *première étape* de la modélisation au niveau de la situation concrète : l'observation d'une situation réelle et sa description en termes courants.

Cette description est déjà une sorte d'abstraction et de simplification de la réalité, dans la mesure où certains choix sont faits, pour ne retenir que ce qui semble pertinent de cette situation vis-à-vis du problème étudié. Cette description est d'ailleurs pilotée par ce que j'appelle *un regard théorique*, c'est-à-dire une connaissance s'appuyant sur des modèles généraux pré-construits, pour apprécier justement ce qui se révélera pertinent.

La démarche expérimentale consiste aussi à pouvoir agir sur la réalité, afin d'en étudier les évolutions et les invariants. Il faut donc pouvoir mettre en œuvre une expérimentation programmée par un *protocole expérimental*, c'est-à-dire l'ensemble des instructions à suivre pour réaliser cette expérience et éventuellement la reproduire.

De nouvelles compétences sont alors attendues, qui peuvent être objet de formation : savoir décrire une situation porteuse d'un problème (par exemple, l'évolution des files d'attente devant les caisses d'un supermarché), savoir mettre en œuvre un protocole expérimental, et recueillir les effets obtenus, savoir organiser les données recueillies, savoir lire une statistique (par exemple, pointer les files à intervalles réguliers).

Puis il s'agit de traduire cette description en un système simplifié et structuré : c'est le niveau du modèle pseudo-concret. Cela se traduit par l'appel à un modèle général dont les conditions de transfert sont maîtrisées. En didactique, nous appelons cela *contextualisation* d'un savoir ancien.

Dans l'exemple précédent, il faut dégager les hypothèses pertinentes pour décrire les arrivées des clients, notamment le nombre moyen d'arrivées par unité de temps. Cette construction est guidée par un premier niveau de connaissances théoriques du phénomène étudié (processus de Poisson) et par les outils mathématiques disponibles. Elle conduit à poser des hypothèses de modèle (indépendance des arrivées...).

On peut alors passer à la *deuxième étape* : la mathématisation ou formalisation du modèle. Cela suppose que les élèves soient capables de représenter le modèle dans la symbolique propre aux mathématiques. Puis ils doivent savoir interpréter la question posée en un problème purement mathématique et savoir faire appel aux outils mathématiques adaptés pour résoudre le problème abstrait (logique, notions ensemblistes, fonctions de variables réelles, intégration, raisonnement par récurrence...).

Enfin, il convient, en *troisième étape*, de pouvoir revenir à la question posée pour traduire dans les termes du modèle pseudo-concret, les résultats mathématiques obtenus, leur donner du sens pour dégager des réponses et relativiser ces réponses par rapport aux hypothèses de modèle. Il faut ensuite interpréter ces réponses pour apprécier leur validité et leur étendue dans la situation concrète (par exemple, décider de l'ouverture ou de la fermeture d'une caisse pour réguler les files d'attentes). Ces compétences peuvent faire l'objet de formation dans diverses disciplines. Elles prennent un aspect spécifique en mathématiques du fait du caractère particulièrement abstrait des outils que l'on désire mettre en œuvre.

Pour illustrer ces considérations, nous allons construire sur tableur une simulation d'un sondage (actualité brûlante…), sans perdre de vue la question didactique : quels apprentissages une telle simulation permet-elle d'accompagner ?

Avec les organisateurs du colloque, on a choisi pour cet atelier de présenter cette simulation[19], l'une des plus simples à mettre en œuvre, qui peut faire l'objet d'un thème d'étude en seconde.

II – Situations de sondages

La situation des sondages aléatoires simples est une application la plus élémentaire de la loi des grands nombres. Si, dans une population statistique, des éléments en proportion p présentent une certaine propriété *M*, un prélèvement aléatoire d'un échantillon de taille n dans cette population peut renseigner sur p (on suppose que la population est assez vaste pour pouvoir considérer ce prélèvement comme non exhaustif, i.e. « avec remises »). Par exemple, *M* peut être le choix préférentiel d'un consommateur ou d'un électeur.

Le fait pour chaque élément observé e_i de cet échantillon, prélevé au hasard, d'avoir la propriété *M* est un événement *E* de probabilité inconnue p. De manière perceptive, on sait que lorsqu'on répète cette expérience un grand nombre n de fois, la fréquence f_n de réalisations de *E* « tend à se stabiliser », et f_n peut être *observée* aussi proche de p que l'on veut, pourvu que n soit assez grand. Ce phénomène permet de proposer expérimentalement des encadrements possibles de p à partir des valeurs observées des f_n (fourchettes d'échantillonnage) avec d'assez grandes chances de ne pas se tromper. Un thème d'études au programme de seconde propose de vérifier la formule qui suit pour déterminer de telles fourchettes, contenant effectivement p dans 95 % des cas : $]f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}}[$.

III – Simulation d'un sondage

3.1 Données pour la simulation

En première étape, l'atelier commence par la réalisation de simulations d'un sondage aléatoire simple utilisant les propriétés de la fonction ALEA() d'un tableur. Cette fonction donne un ensemble de 15 chiffres aléatoires indépendants et équirépartis. Elle produit donc à chaque recalcul une observation d'une variable discrète uniforme sur [0 ; 1], avec probabilité $1/10^{15}$. On veut simuler des prélèvements d'échantillons de tailles variables n dans une population de taille N (où $n/N < 1/100$),

[19] Exposée dans l'article *Simulation d'un sondage. Fourchette d'échantillonnage et intervalles de confiance,* bulletin de l'APMEP n° 444, janvier-février 2003.

dans laquelle il y a une proportion p d'individus (statistiques) qui présentent le caractère *M*. Dans cette simulation, on peut mettre en évidence la variabilité des fréquences observées suivant les échantillons prélevés, de taille 100 d'abord puis de taille 1000 (fluctuations d'échantillonnage). Les fourchettes de sondages sont ensuite dégagées de multiples simulations puis calculées à l'aide de la formule ci-dessus.

Dans notre exemple, on a $N = 10^9$. On suppose que les N individus ont été numérotés. Le choix au hasard de l'un d'entre eux revient à produire par ALEA() un nombre entier x de 9 chiffres. Un nombre $p \in]0\ ;1[$ étant donné (que l'on peut implanter dans une cellule cachée du tableur), on considère que si $x \times 10^{-9} < p$, l'individu est de caractère *M*. La loi uniforme simulée par ALEA() permet de dire que le prélèvement au hasard d'un individu de caractère *M* est de probabilité p.

La fréquence f_n des individus de caractère *M* observés dans un échantillon de taille n donne une estimation de la proportion p. Bien sûr, on s'intéresse à la précision de cette estimation, c'est-à-dire à un encadrement *de confiance* de p à partir de cette valeur observée de f_n (fourchette de sondage), tel qu'un pourcentage *significatif* d'échantillons (95 %) fournissent une fourchette qui contient effectivement p. La formule « magique » $]f_n - \frac{1}{\sqrt{n}}\ ;\ f_n + \frac{1}{\sqrt{n}}[$ donne de telles fourchettes théoriques, pour des valeurs de p pas trop éloignées de 0,5 (entre 0,3 et 0,7) ; on peut facilement justifier ce résultat et l'étendre à toute valeur de p différente de 0 ou 1. En produisant un certain nombre d'échantillons, les élèves peuvent vérifier que pour environ 95 % d'entre eux l'encadrement de p précédent est vérifié.

3.2 Réalisation sous Excel

Le tableau qui suit est un extrait d'une feuille de calcul.

- La colonne A est obtenue à partir de la fonction « =ALEA() » d'Excel qui donne des chiffres pseudo-aléatoires équirépartis, sous la forme de décimales d'un nombre de [0, 1].

- La partie entière du produit de ces nombres par 10^6 permet de simuler le prélèvement au hasard d'éléments dans une population de taille 1 000 000 (colonne B : =ENT(A1*1000000)).

- Une valeur pour p ayant été introduite (cellule cachée F14), la comparaison de chacun de ces nombres avec $p \times 10^6$, traduite en 1 ou 0 (colonne C) par la fonction logique =SI(ENT(100*A1)<100F14;1;0) , simule l'observation des éléments prélevés relativement au caractère *M*, le choix 1 étant en proportion p dans la population.
- En totalisant par paquets de 100 la colonne C, on obtient les fréquences observées du 1 dans ces divers échantillons, 10 d'entre elles sont présentées en colonne D (« =SOMME(C1:C100)/100 », puis (C101:C200) etc.).

- Les bornes des fourchettes d'échantillonnage sont données dans les colonnes F et G (dans ce cas :]f – 0,1 ; f + 0,1[), à titre de comparaisons.

- Les 10 échantillons étant regroupés, on obtient un échantillon de taille 1000, avec une fourchette plus étroite (cellules F12 et G12), l'intervalle de confiance théorique étant calculé à partir de la formule indiquée en dessous (cellules F13 et G13), dont la démonstration fait l'objet du développement qui suit.

	A	B	C	D	E	F	G
1	0,49716356	497163	**0**	0,43	ECHANTILLONS, TAILLE N=100	0,53	0,53
2	0,85464871	854648	**0**	0,32		0,22	0,42
3	0,71631158	716311	**0**	0,36	fourchettes de sondages :	0,26	0,46

4	0,10738629	107386	**1**	0,31		0,21	0,41
5	0,85212683	852126	**0**	0,33	]f-1/√n ; f+1/√n[	0,23	0,43
6	0,94887578	948875	**0**	0,33		0,23	0,43
7	0,12033745	120337	**1**	0,38		0,28	0,48
8	0,04542526	45425	**1**	0,44		0,34	0,54
9	0,33275705	332757	**1**	0,28		0,18	0,38
10	0,21831516	218315	**1**	0,33		0,23	0,43
11	0,28453486	284534	**1**		ECHANTILLON DE TAILLE 1000, F =	**0,351**	
12	0,62820339	628203	**0**		fourchette simplifiée :	0,319	0,383
13	0,79683387	796833	**0**		intervalle de confiance :	0,321	0,381
14	0,91677761	916777	**0**		Probabilité théorique introduite, p =	0,37	
15	0,88421451	884214	**0**				
16	0,08202963	82029	**1**	A = 1000 nombres au hasard équirépartis			
17	0,82429676	824296	**0**	B = échantillon de 1000 personnes parmi 1 000 000			
18	0,62983241	629832	**0**	C = préférences de ces 1000 personnes			
19	0,79539564	795395	**0**	D = fréquences du choix 1 par échantillons de tailles 100			
20	0,79191754	791917	**0**	] F; G [: fourchettes de sondage simplifiées			
21	0,63348029	633480	**0**	et intervalle de confiance (niveau 0,95) :			
etc …	Jusqu'à 1000…	…	**…**	$]f-\frac{1,96\sqrt{f(1-f)}}{\sqrt{n}};f+\frac{1,96\sqrt{f(1-f)}}{\sqrt{n}}[$			

IV – Estimation de p par intervalle de confiance

4.1 Le modèle probabiliste

Introduisons le modèle probabiliste de cette situation de sondage. Le prélèvement au hasard d'un élément de la population est représenté par une variable aléatoire de Bernoulli X_0, variable parente de l'échantillonnage, qui prend la valeur 1 pour les éléments présentant la propriété *M* (*E* est réalisé), avec probabilité p, et 0 avec probabilité 1 – p sinon, d'espérance $E(X_0) = p.1+(1–p).0 = p$ et de variance $Var(X_0) = p\,(1–p)$.

Le prélèvement de l'échantillon des n éléments e_i est représenté par le vecteur aléatoire $X = (X_1, X_2, \ldots, X_n)$, où les X_i sont les répliques successives indépendantes de X_0. Ce sont des variables aléatoires de même loi de Bernoulli ***B***(1, p) que X (à condition de considérer que les prélèvements des e_i ne changent pas notablement la proportion p de ceux qui sont de modalité *M* dans la population). On va considérer de plus que les X_i sont indépendantes pour exprimer l'hypothèse que la réalisation ou non de l'événement *E* lors des prélèvements des premiers e_i n'a pas d'effet sur la probabilité de réalisation de *E* pour les suivants.

La moyenne $F_n = \frac{1}{n}\sum X_i$ dépend de l'aléa du prélèvement. C'est une variable aléatoire d'espérance $E(F_n) = \frac{1}{n}\sum E(X_i) = p$ et de variance $Var(F_n) = \frac{1}{n^2}\sum Var(X_i) = \frac{p(1-p)}{n}$ (l'indépendance des X_i permet l'additivité des variances).

Sa valeur f_n observée sur l'échantillon est la fréquence des éléments de modalité *M* dans l'échantillon (nombre des X_i prenant la valeur 1 divisé par n). En application de la loi des grands nombres (ici, le théorème de Bernoulli énoncé plus loin), elle est prise pour estimer la valeur de la proportion p inconnue (estimation ponctuelle).

L'écart $|f_n - p|$ dépend aussi de l'aléa du prélèvement. On ne peut donc espérer obtenir qu'un contrôle probabiliste a priori de $|F_n - p|$, majorant l'erreur que l'on fera en prenant pour p la valeur f_n de la fréquence de l'événement E dans l échantillon qui sera prélevé.

4.2 Le théorème de Bernoulli

Pour cela, on a un outil théorique simple : l'inégalité de Bienaymé-Tchebychev, résultat important de la théorie des probabilités :

Soit Y une variable aléatoire de loi quelconque mais dont l'espérance E(Y) et la variance Var(Y) existent. Alors pour tout $\varepsilon > 0$, la probabilité que Y s'écarte de E(Y) de plus que ε est contrôlée par la dispersion de Y. On a :

$$P(|Y - E(Y)| \geq \varepsilon) \leq \frac{Var(Y)}{\varepsilon^2}.$$

Appliquée à la fréquence F_n, cette inégalité donne pour tout $\varepsilon > 0$:

$$P(|F_n - p| \geq \varepsilon) \leq \frac{p\,(1-p)}{n\varepsilon^2},$$

probabilité qui tend vers 0 quand n tend vers l'infini. On dit que « la fréquence F_n tend vers p en probabilité ». C'est le théorème de Bernoulli, forme la plus simple de la loi (faible) des grands nombres.

4.3 Intervalle de confiance

En passant à l'événement contraire, cette inégalité s'écrit :

$$P(F_n - \varepsilon < p < F_n + \varepsilon) \geq 1 - \frac{p\,(1-p)}{n\varepsilon^2}$$

La valeur $\alpha = \frac{p\,(1-p)}{n\varepsilon^2}$ est la probabilité (appelée « le risque ») de se tromper en annonçant que p sera dans l'intervalle $]F_n - \varepsilon, F_n + \varepsilon[$, « intervalle de confiance pour p de niveau $1 - \alpha$ ».

En majorant grossièrement $p\,(1 - p)$ par 1/4, un majorant ε de l'écart $|F_n - p|$ étant donné, $\alpha = \frac{1}{4n\varepsilon^2}$ est aussi petit que l'on veut pourvu que n soit assez grand. On a donc a fortiori un intervalle de confiance simplifié de niveau $1 - \alpha$, aussi proche de 1 que l'on veut :

$$P\left(F_n - \frac{1}{2\sqrt{\alpha}\sqrt{n}} < p < F_n + \frac{1}{2\sqrt{\alpha}\sqrt{n}}\right) \geq 1 - \alpha.$$

Une fois l'échantillon prélevé, un niveau de confiance étant donné (par exemple 0,95), l'intervalle observé $\left]f_n - \frac{1}{2\sqrt{\alpha}\sqrt{n}} ; f_n + \frac{1}{2\sqrt{\alpha}\sqrt{n}}\right[$ est une interprétation théorique de *la fourchette de sondage* statistiquement obtenue si l'on répète un grand nombre de fois cet échantillonnage (pour environ 95 % des échantillons, on trouve effectivement p dans la fourchette). Au niveau de confiance $1 - \alpha$, l'approximation de l'estimation de p par f_n est majorée par $\frac{1}{2\sqrt{\alpha}\sqrt{n}}$ (pour = 0,05, cela donne une

demi fourchette de $\frac{2,23}{\sqrt{n}}$). On voit que plus le niveau de confiance souhaité sera proche de 1, moins bonne sera la précision de l'estimation proposée. On remarque aussi que celle-ci est en $1/\sqrt{n}$.

4.4 Approximation normale, le théorème de Moivre-Laplace

Mais, d'une part, la majoration de p (1 – p) par 1/4 peut paraître trop grossière. Le lien entre la précision souhaitée ε et la taille de l'échantillon n suppose le calcul exact de $P(]F_n - \varepsilon, F_n + \varepsilon[)$ et ce calcul met en jeu la loi binomiale ***B***(n, p) avec ses C_n^k hors d'atteinte pour n grand.

D'autre part, l'inégalité théorique de Bienaymé-Tchebychev, ne faisant pas intervenir la loi de F_n, est trop générale et ne peut être intéressante dans la pratique : si $\alpha = 0,05$, il faudrait 50 000 observations pour estimer p à 1 % près.

Le théorème de Moivre-Laplace (forme particulière d'un théorème puissant des probabilités, le *théorème limite central*) permet d'améliorer grandement la performance. Ce théorème dit que :

pour n > 50 et pour p pas trop voisin de 0 ou de 1 (ce qui n'est pas trop demander), on fait une erreur négligeable sur la valeur de la probabilité $P(|F_n - p| < \varepsilon)$ en considérant que F_n suit une loi normale ***N***$(p ; \sqrt{\frac{p(1-p)}{n}})$, ou encore que $\frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n}$ est normale, centrée réduite.

Mais, dans notre situation de sondages, p est inconnu. La variance $\frac{p(1-p)}{n}$ de la loi de F_n peut alors être obtenue en estimant ponctuellement p par la fréquence f_n. On fait alors une erreur négligeable (au second ordre près quand n est grand) dans le calcul des probabilités liées à F_n.

En réduisant donc F_n sous la forme $U = \frac{F_n - p}{\sqrt{F_n(1-F_n)}}\sqrt{n}$, la condition de confiance $P(|F_n - p| < \varepsilon) = 1 - \alpha$ s'écrit :

$$P(|U| < \varepsilon\frac{\sqrt{n}}{\sqrt{F_n(1-F_n)}}) = 1 - \alpha.$$

où l'on peut considérer que la loi de U est assez proche de celle d'une variable normale centrée réduite.

Si $u_{\alpha/2}$ désigne le *quantile* de cette loi tel que $P(|U| < u_{\alpha/2}) = 1 - \alpha$, on obtient l'intervalle de confiance pour p au niveau de confiance $1 - \alpha$:

$$]F_n - u_{\alpha/2}\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} ; F_n + u_{\alpha/2}\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}[.$$

Une fois l'échantillon prélevé, la fréquence observée de l'événement *E* étant f_n, l'encadrement proposé pour p est alors :

$$]f_n - u_{\alpha/2}\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}} ; f_n + u_{\alpha/2}\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}[.$$

Avec $\alpha = 0,05$, on a $u_{\alpha/2} = 1,96$ très voisin de 2. En majorant comme précédemment $f_n(1 - f_n)$ par 1/4 quand f_n n'est pas proche de 0 ou 1 ($0,3 < f_n < 0,7$), on perd un peu en précision

mais on simplifie notablement l'expression de la fourchette de sondage théorique : le demi écart $\varepsilon = u_{\alpha/2}\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}$ de l'encadrement de p est alors majoré par $\frac{1}{\sqrt{n}}$. On peut donc donner pour fourchette de sondage, au niveau de confiance $1 - \alpha = 0{,}95$, l'intervalle proposé dans le thème d'étude du programme de seconde :

$$]f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}} [.$$

Dans ces conditions, il suffit d'un échantillon de taille n > 10 000 (c'est encore cher) pour estimer p à 1 % près, avec une probabilité 0,95 de ne pas se tromper. Si on accepte une estimation de p à 3 % près, il suffit que n > 1 112, taille approximative des sondages les plus courants.

Le tableau suivant donne une idée des précisions (demi-écarts des fourchettes) obtenues pour différentes tailles d'échantillons et différents niveaux de confiance.

n α	500	800	1 000	2 000	10 000
0,1	3,7	2,9	2,6	1,8	0,8
0,05	4,4	3,5	3	2,2	1
0,01	5,7	4,5	4	2,9	1,3

Quelques références bibliographiques

Commission Inter-IREM Statistique et probabilités : *Statistique au Lycée***,** vol. 1, brochure APMEP n° 156, octobre 2005.

Commission Inter-IREM Statistique et probabilités : *Autour de la modélisation en probabilités*, Presses universitaires de Franche-Comté, 2001.

Dress, F. : *Probabilités, Statistique, rappels de cours, questions de réflexion, exercices d'entraînement*, Dunod, Paris, 1997.

Dutarte P. : *Pour une éducation à l'inférence statistique au lycée*, Repères-IREM n° 60, juillet 2005.

Dutarte P. : *L'induction statistique au lycée illustrée par le tableur*, Didier, 2005.

Dutarte P., Piednoir J.-L. : *Enseigner la statistique au lycée : des enjeux aux méthodes*, Commission inter-IREM Lycées techniques, brochure n° 112, 2001.

Groupe Probabilités & statistique de l'IREM de Besançon : *Lois continues, tests d'adéquation, une approche pour non spécialistes*, Presses universitaires de Franche-Comté, 2005.

Girard, J. C. & Henry, M., Modélisation et simulation en classe, quel statut didactique ? *Statistique au lycée*, vol. 1. CII Statistique et probabilités ed., brochure APMEP n° 156, juillet 2005, p. 147-159.

Henry, M., Simulation d'un sondage. Fourchette d'échantillonnage et intervalles de confiance, *Bulletin Vert* de l'APMEP n° 444, janvier-février 2003, p. 88-96.

Saporta, G. : *Probabilités, Analyse des données et Statistique*, Technip, Paris, 1990.

A paraître en octobre 2007, avec une partie sur les sondages :

Commission Inter-IREM Statistique et probabilités : *Statistique au Lycée***,** vol. 2, brochure APMEP n° 163.