Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

5 - A PROPOS D'UN EXERCICE DE BAC : DE LA DIFFICULTÉ DE BIEN HABILLER LES ÉNONCÉS. UN PROBLÈME D'ESTIMATION PAR CAPTURE ET RE-CAPTURE

Michel HENRY

Voici l'énoncé d'un problème proposé dans un manuel scolaire en vue de la préparation du baccalauréat sous le titre : *Loi binomiale. Étude de fonctions puissances.*

Un étang contient un nombre N, inconnu mais défini de poissons.

L'objet du problème est de proposer une évaluation de N, basée sur des hypothèses bien définies.

1 - On pêche dans différents endroits de l'étang ; on en sort 20 poissons que l'on marque et que l'on remet vivants dans l'étang après les avoir marqués. Quelques jours plus tard, on effectue une nouvelle pêche dans des endroits variés du même étang en rejetant à l'eau les poissons pêchés, après avoir noté s'ils sont marqués ou non. On prend ainsi 50 poissons dont 4 sont marqués.

On suppose qu'entre les deux pêches, la population de l'étang n'a pas varié et que lors de la seconde pêche, à chaque coup il y a équiprobabilité de sortie pour chacun des N poissons de l'étang.

Avant la seconde pêche, on pouvait se poser le problème : quelle est la probabilité de sortir les poissons marqués sur les 50 pêchés ?

Répondez à cette question en donnant l'expression générale de P (X = k) où X désigne la variable aléatoire " nombre de poissons marqués que l'on peut sortir sur 50 pêchés ".

2 - f est la fonction qui, à un réel x supérieur à 20, associe :

$$f(x) = \left(\frac{20}{x}\right)^4 \cdot \left(1 - \frac{20}{x}\right)^{46}$$

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

Montrez que f a un maximum. Pour quelle valeur a de x ce maximum est-il atteint ?

3a - Si l'on admet que l'événement (X = 4) qui est réalisé correspond à l'événement de probabilité maximale parmi tous les événements possibles (X = k), k = 0, …, 50, quelle valeur doit-on attribuer à la population N de l'étang ?

3b - Cette hypothèse conduit à la même conclusion qu'une autre hypothèse très simple que l'on aurait pu faire pour évaluer N. Laquelle ?

Premières remarques et critiques sur l'habillage artificiel de problèmes de bac :

Dans cet exercice, la loi hypergéométrique et la loi binomiale interviennent. Mais il se pose une question : dans 3a il est question de maximum dans un ensemble qui n'est pas le même que l'ensemble évoqué au 2 et auquel l'énoncé semble faire référence.

P(X = 4) est maximale dans l'ensemble des P(X = k) pour un N donné. Cela ne signifie pas pour autant que P(X = 4) correspond au N qui rend P(X = 4) maximum lorsque cette fois l'on fait varier N. A première vue la question 3a semble incohérente avec ce qui précède (et la remarque est parfaitement justifiée), elle procède en fait d'une démarche correcte, comme nous allons le voir.

Voici un énoncé maladroitement habillé d'un modèle déjà là (le schéma de Bernoulli) pour " faire plus vrai ", alors que personne n'est dupe ! Il a alors l'inconvénient didactique de confondre modèle et réalité, ou plutôt de chercher à évacuer dans les implicites le rôle et le statut du modèle. Cela conduit à des contorsions de formulations qui feraient dresser les cheveux sur la tête d'enseignants de Français : *"poissons que l'on marque et que l'on remet vivants dans l'étang après les avoir marqués"*...

Comme l'auteur veut éviter certaines critiques de non adéquation du modèle binomial, il introduit des paraphrases comme : *"on effectue une nouvelle pêche dans des endroits variés du même étang"* sans que l'on discerne bien quelle hypothèse de modèle elle est censée induire chez les élèves.

Enfin le schéma binomial est mal séparé du modèle hypergéométrique : *"en rejetant à l'eau les poissons pêchés…"* sans que l'on sache si ce rejet se fait un à un ou après avoir pêché les 50. Remarquons qu'avec les données numériques ($^{50}/_{250}$) les deux modèles, binomial et

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

hypergéométrique, sont proches, mais ils sont aussi proches d'un modèle de Poisson (P majuscule) avec P = 8%, plus pratique. Heureusement, les élèves n'ont que le schéma binomial à leur programme, ils n'ont donc pas à avoir d'états d'âme sur le choix du meilleur modèle et cette phrase doit être interprétée comme une hypothèse de remise des poissons un par un après chaque tirage. Mais peut-on trouver satisfaisante une résolution qui doit plus aux effets de contrat qu'à une véritable maîtrise de la modélisation ?

Le caractère pseudo concret de l'énoncé conduit ensuite à des pseudo hypothèses de plus en plus artificielles : *"entre les deux pêches, la population n'a pas varié"*, pas de naissances donc, et pas de prédateurs, autres poissons, oiseaux ou pêcheurs braconniers …

De plus, *"lors de la seconde pêche, il y a équiprobabilité de sortie…"*, au moins, c'est dit, cela évite de se demander si cette hypothèse va de soi (et complètement irréalisable dans la pratique), mais alors pourquoi préciser que l'on pêche *"dans des endroits variés"* ?…

Ces incohérences d'énoncé du point de vue des formulations vont jusqu'à l'apothéose : *"quelle est la probabilité de sortir* ***les*** *poissons marqués sur les 50 pêchés"*. Les 20 différents poissons marqués, alors qu' il y a remise ? comment le vérifier, si le protocole expérimental ne le stipule pas ? On est de plus dans un problème (pas simple) totalement différent, où l'on chercherait la probabilité de trouver **exactement** les 20 poissons marqués, parmi les k poissons marqués figurant dans cet échantillon (avec remises) de 50 poissons.

Notre connaissance des problèmes de Terminale nous fait comprendre que ce n'est pas ce qu'a voulu dire le rédacteur de l'énoncé : il voulait sans doute demander la probabilité que « parmi les 50 poissons pêchés et remis un par un, on en trouve k marqués ». Mais les candidats au bac peuvent-ils le décrypter ainsi ? Heureusement, dans l'énoncé, la question est mathématisée à leur place : *"donner l'expression générale de P(X = k)…"* et, comme dans le formulaire (ou leur mémoire de la classe) il n'y a qu'une formule qui ressemble à cela, la seule réponse possible est formelle :

$$P(X = k) = C_{50}^{k} \left(\frac{20}{N}\right)^{k} \left(1 - \frac{20}{N}\right)^{50-k} .$$

On peut douter qu'il y ait un seul élève pour chercher à justifier l'emploi de la loi binomiale à partir des données aussi alambiquées de l'énoncé.

Alors, à quoi sert tout l'habillage, si c'est pour faire copier une formule ? Ne valait-il pas mieux la demander clairement ? Ou mieux, demander les

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

hypothèses nécessaires pour qu'un modèle d'urne par exemple conduise à l'étude d'une variable binomiale, dont on donnerait les probabilités élémentaires.

La question 2 tombe ensuite comme un cheveu sur la soupe. Pourquoi a-t-on fait k = 4 et pourquoi cette recherche de maximum ? Quels sont les élèves, habiles en analyse, qui seront passés par l'étude de ln f(x) ?

Considérons maintenant à la question 3a dont la critique est nettement plus intéressante. Celle-ci repose sur une propriété de maximum des probabilités binomiales dans un schéma de Bernoulli, que l'on trouve dans *"Ars Conjectandi"* de J. Bernoulli (cette propriété lui sert à sa démonstration sophistiquée de son *théorème d'or* : la loi des grands nombres), ainsi que sur la compréhension de la méthode du maximum de vraisemblance en statistique inférentielle. Mais auparavant, traduisons en termes de modèle d'urne le schéma des poissons pour préciser clairement les hypothèses.

Soit donc une urne de Bernoulli contenant N boules dont r blanches et s = N – r noires. On pose $p = \frac{r}{N}$, la proportion des boules blanches dans l'urne. L'hypothèse de modèle d'urne est l'équiprobabilité des boules (poissons) dans un tirage « au hasard » de l'une d'entre elles (pêche dans des lieux divers …). L'obtention d'une boule blanche est alors un événement de probabilité p. (Définition de la probabilité dans le modèle d'urne, définition de Laplace, ou interprétation fréquentiste via le phénomène de stabilisation de la fréquence induisant la notion de probabilité).

Soit le schéma de Bernoulli : n tirages successifs « avec remises » de boules de l'urne avec observation des couleurs obtenues. L'hypothèse de modèle dans ce schéma de Bernoulli est que la « remise » fournit la même urne de Bernoulli au tirage suivant (i.e. dotée des mêmes hypothèses), ce qui se traduit d'un point de vue probabiliste (axiomes) par le fait que les événements respectivement associés à deux tirages différents sont indépendants : la probabilité de leur conjonction est donc le produit des probabilités de chacun d'eux.

On s'intéresse au nombre X des boules blanches obtenues au cours de n tirages (n = 50 pour nous). On sait (le cours) que X suit une loi binomiale *B* (n,p). Le fait d'obtenir k boules blanches pour un k fixé de 0 à n est donc un événement associé au schéma de Bernoulli dont la probabilité est donnée par :

$$P(X = k) = C_{50}^{k}\, p^{k} (1 - p)^{n-k} .$$

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

Dans la suite, il sera plus simple d'utiliser les notations introduites sous la forme :

$$P(X = k) = \frac{n(n-1)\ldots(n-k+1)}{k!}\,\frac{r^k s^{n-k}}{N^n}\ .$$

Concernant les probabilités P(X = k), on a la propriété de maximum (que l'on peut qualifier « de Bernoulli ») suivante, qui sera démontrée ensuite :

Propriété de Bernoulli

Dans les hypothèses précédentes, la valeur m de k qui rend maximale la probabilité binomiale P(X = k) vérifie :
$\frac{r}{N}(n+1) - 1 \leq m \leq \frac{r}{N}(n+1)$.

Remarques

- Si $\frac{r}{N}(n+1)$ n'est pas entier, m est alors unique :

$$m = \text{Ent}\left[\frac{r}{N}(n+1)\right] \qquad \text{et} \quad P(X = m) = C_n^m \frac{r^m s^{n-m}}{N^n}$$

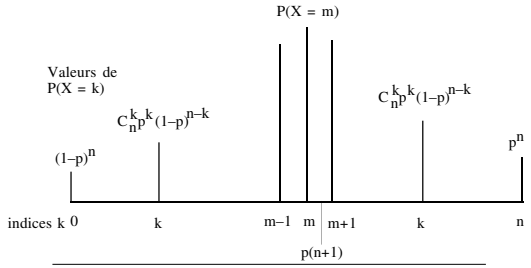
- Si $\frac{r}{N}(n+1)$ est entier, deux indices consécutifs donnent à cette probabilité une même valeur maximale égale à :

$$P\left[X = \frac{r}{N}(n+1) - 1\right] = P\left[X = \frac{r}{N}(n+1)\right].$$

Propriété de Bernoulli (suite)

Si l'on ordonne les probabilités binomiales selon les n+1 valeurs de k, de 0 à n, ces probabilités vont d'abord en croissant jusqu'à l'indice m tel que $\frac{m}{n+1}$ soit la valeur inférieure la plus voisine de $p = \frac{r}{N}$, puis décroissent.

On peut visualiser ce phénomène sur le schéma suivant :



Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

La démonstration est inspirée de celle de J. Bernoulli (prop. 2 et 3 de *Ars Conjectandi*). Elle repose sur l'étude du rapport $\frac{P(X = k)}{P(X = k - 1)}$ et de sa position par rapport à 1.

On a, pour $k > 0$: $\frac{P(X = k)}{P(X = k - 1)} = \frac{n - k + 1}{k} . \frac{r}{s}$, d'où $\frac{P(X = k)}{P(X = k - 1)} \geq 1$,

ce qui équivaut à : $\frac{r}{N} (n + 1) \geq k$ (rappelons que $r + s = n$).

Cela montre que si $0 < k \leq \frac{r}{N} (n + 1)$, la probabilité binomiale $P(X = k)$ va en croissant.

On a aussi : $\frac{P(X = k)}{P(X = k - 1)} = \frac{k + 1}{n - k} . \frac{s}{r} \geq 1 \Leftrightarrow k \geq \frac{r}{N} (n + 1) - 1$, ce qui montre que si

$\frac{r}{N} (n + 1) - 1 \leq k < n$, cette probabilité va en décroissant.

$P(X = k)$ passe donc par un maximum pour la ou les valeurs entières de k comprises entre $\frac{r}{N} (n + 1) - 1$ et $\frac{r}{N} (n + 1)$. Valeur unique notée m, si $\frac{r}{N} (n + 1)$ n'est pas entier : on a alors

$m = \text{Ent} [\frac{r}{N} (n + 1)]$ et le maximum vaut $P(X = m) = C_n^m p^m (1 - p)^{n-m}$.

Si $m = \frac{r}{N} (n + 1)$ est entier, la valeur commune $P(X = m-1) = P(X = m)$ est le maximum de la probabilité $P(X = k)$, ce que l'on vérifie en écrivant le rapport $\frac{P(X = m - 1)}{P(X = m)} = \frac{C_n^{m-1} r^{m-1} s^{n-m+1}}{C_n^m r^m s^{n-m}} = \frac{m}{n - m + 1} . \frac{s}{r}$, ce qui donne :

$$\frac{P(X = m - 1)}{P(X = m)} = \frac{\frac{r}{N} (n + 1)}{(n + 1)(1 - \frac{r}{N})} . \frac{s}{r} = \frac{r}{N - r} . \frac{s}{r} = 1$$

Revenons à l'exercice de « bac ».

On part du principe du maximum de vraisemblance :

> un événement observé à l'issue d'une expérience aléatoire est celui qui, parmi tous les événements comparables, avait la plus grande probabilité d'arriver.

On sait qu'en pratique cela n'est pas vrai, car à partir de la répétition de la même expérience, on observe différents résultats aléatoires qui ne sont donc pas tous de probabilité maximale. Mais, d'après la loi des grands

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

nombres, en moyenne, dans un (très) grand nombre de telles expériences, l'événement à probabilité maximale s'observe le plus souvent.

A partir de ce principe, à la base des méthodes bayésiennes, on infère que les paramètres inconnus qui contrôlent l'expérience aléatoire ont les valeurs qui rendent maximale la probabilité de l'événement observé. C'est la « méthode du maximum de vraisemblance ».

En fait, ces valeurs hypothétiques minimisent (il faut le montrer) le risque de se tromper en les retenant à la place d'autres éventuelles : c'est une inférence statistique, c'est à dire une hypothèse faite sur la valeur du ou des paramètres inconnus, avec une certaine précision (ou imprécision) qui peut être déterminée, et un risque de se tromper en donnant cette valeur, risque qui peut être aussi calculé ou contrôlé. Une telle inférence, sans ces deux contrôles - précision et fiabilité - n'a pas beaucoup de sens (surtout si on ne sait pas évaluer le coût d'une erreur), car elle conduit à une affirmation qui ne peut être appréciée, voire contestée.

Le raisonnement par maximum de vraisemblance, appliqué malgré cela au problème des poissons, nous dit : pour N fixé (inconnu), la valeur k = 4 est celle qui rend maximale la probabilité P(X = k) dans les conditions du schéma de Bernoulli. C'est à dire qu'avec les données numériques du problème, on pose

m = 4, i.e. Ent [$\frac{20}{N} - 51$] = 4, où N est un entier tel que :

$\frac{20}{N} - 51 - 1 < 4 < \frac{20}{N} - 51$, ce qui donne $204 < N < 255$.

Mais en appliquant à nouveau le même principe du maximum de vraisemblance, la valeur de N est celle qui rend maximale la probabilité

$$P(X = m) = C_{50}^{4}(\frac{20}{N})^{4}(1 - \frac{20}{N})^{46}$$

où m = 4, c'est à dire N = 250, selon l'étude des variations de cette fonction faite en 2ème question.

Ainsi, la question 3a est fondée. L'hypothèse de travail posée dans l'énoncé peut sembler surprenante a priori. Mais elle est pertinente du point de vue de l'application du principe du maximum de vraisemblance.

Cependant, elle aura sans doute eu un effet distracteur pour les bons candidats, celui de les lancer sur l'étude des variations de

$$P(X = k) = C_{n}^{k} p^{k} (1 - p)^{n-k}$$

quand k varie de 0 à n, pour vérifier à quelles conditions (sur N), P(X = m) est bien la valeur maximale de cette probabilité (objectif inaccessible sans

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X

Presses universitaires de Franche-Comté - Université de Franche-Comté
25030 Besançon cedex - France Tél. : (33) 03 81 66 59 70 - Fax : (33) 03 81 66 59 80
Mél : presses-ufc@univ-fcomte.fr - http : //presses-ufc.univ-fcomte.fr

indications à ce niveau), car on ne voit pas a priori pourquoi cela serait vrai pour m = 4 particulièrement. Par contre pour ceux qui ne se posent pas de question (est-ce le comportement attendu le jour du bac ?) il devenait « automathique » (pour reprendre le calembour de Stella Baruck) de renvoyer au maximum obtenu à la question précédente sans comprendre en quoi il concernait cette question 3a.

On peut enfin se demander comment l'attente exprimée par l'énoncé (formuler une hypothèse) peut être comprise par un candidat au bac ?

Le qualificatif de *"très simple"* n'apporte rien de plus et est au contraire un facteur déstabilisant et aggravant (tel le joueur de bridge qui interpelle son partenaire au moment où celui-ci va choisir sa carte par l'injonction : "joue bien !"). En effet, après une hypothèse complexe que l'élève de Terminale ne peut comprendre (celle du principe du maximum de vraisemblance) faute d'avoir déjà réfléchi à sa signification, on lui demande d'en fournir une autre pour obtenir le même résultat (forcément pas évident, puisque issu de 3a épaulé par l'étude des variations de f en 2). Comment peut-il s'attendre à ce que le simple « bon sens » suffise pour répondre à la question ?

Ce « bon sens » serait de faire l'hypothèse (qui relève aussi d'une certaine manière du principe du maximum de vraisemblance) que l'échantillon prélevé dans l'étang, dans les conditions d'un schéma de Bernoulli (ce qui dans l'habillage pseudo-concret de l'énoncé est moins que vraisemblable), est « représentatif » de la population du lac, c'est à dire composé proportionnellement à celle-ci :

$$\frac{4}{50} = \frac{20}{N} \text{ d'où } N = 250.$$

Résultat nécessaire (oui, car relevant de la même application du principe du maximum de vraisemblance !) ou magie des données numériques ? On retrouve miraculeusement la même valeur par ce raisonnement de proportionnalité (qui n'est en rien probabiliste) que dans l'étude des variations de la fonction f, qui n'a rien à voir, a priori, avec cette proportionnalité.

Édition intégrale en ligne de : *Autour de la modélisation en probabilités*
Commission inter-IREM *Statistique et Probabilités*, coordination Michel Henry
Presses universitaires de Franche-Comté, 2001, 262 pages, ISBN 2-84627-018-X