DES IMAGES DANS DES PLAQUES DE CHOCOLAT

Expérimentation / Simulation / Modélisation

Claudine ROBERT
Irem de Grenoble

Préambule : *La situation que nous envisageons ci-dessous est, sous des appellations variées, classique (voir* [1, 2]*) ; elle peut être abordée sous plusieurs angles, à plusieurs niveaux. Nous nous attacherons d'une part à montrer comment elle peut servir à introduire quelques fondements de la statistique : fluctuation d'échantillonnage, loi des grands nombres, paramètres numériques (moyenne, médiane, déciles), risque, ainsi que la diversité des points de vue d'où on modélise une même expérience.*

D'autre part, nous mettrons l'accent sur la pratique de la simulation. La triade expérimentation-modélisation-simulation constitue la trame de la recherche, du développement et de l'enseignement du champ de l'aléatoire.

Une grande partie de l'article ci-dessous est publiée dans « graines de sciences 6 » (voir ([3]*). Par ailleurs, des expérimentations ont été faites en classe sur certaines parties par les membres du groupe statistique de l'Irem de Grenoble : Luc Bouttier, M. Gandit, C.Serret et d'autres calculs feront l'objet en 2005 d'une publication de l'Irem de Grenoble.*

Venons-en au thème traité :

Au vingtième siècle une marque de chocolat prestigieuse offrait des images : chaque plaque contenait une reproduction d'animal, fixée par une pointe de colle au papier argenté. Il y avait six sortes d'images différentes : tigre, éléphant, girafe, hippopotame, gazelle, tapir. On a dit que, sur les lieux d'achats, les images étaient réparties au hasard dans les plaques, et qu'aucune image n'était plus rare qu'une autre. La question qui va nous occuper est :

Combien de plaques permettent d'avoir toute la collection ?

Dans la première partie, des éléments de réponse sont donnés, sans utilisation d'outils théoriques ; cette partie constitue une introduction à la statistique, accessible dès la classe de seconde. Dans une deuxième partie, on donne des résultats théoriques : certains sont accessibles pour des élèves de première et terminale.

Voici quelques réponses entendues (une même personne donne souvent plusieurs de ces réponses dans des ordres variables) :

(1) — il manque des informations
(2) — il est impossible de répondre
(3) — si on fabrique 6000 plaques de chocolat, il faut en acheter 5 001.
(4) — au moins six
(5) — ça dépend
(6) — le nombre peut-être infini.

Quelle est la nature de ce qui manque pour répondre ? Recueillir et traiter des informations pertinentes est un des objectifs de cet article.

Le réponse (2) réfère à une habitude où la réponse à « combien » est un nombre. Or le nombre de plaques à acheter est variable (ce que dit la réponse (5)) et la réponse ne saurait donc être un nombre. Oublions les habitudes, ouvrons une porte pour répondre à la question posée, et… c'est la statistique qui entre par cette issue.

La réponse (3) est basée sur une interprétation particulière de la phrase « les images sont réparties au hasard dans les plaques et aucune image n'est plus commune qu'une autre ». Ceux qui la donnent pensent que sur 6000 plaques commercialisées, il y en a exactement 1000 de chaque sorte ; dans le cadre de cette interprétation bien particulière cette réponse est correcte [1]. Mais le sens donné à « répartie au hasard » est ici autre. A l'achat d'une plaque, on a des *chances égales* d'avoir chaque sorte d'images, soit *une chance sur 6* d'avoir l'image un, *une chance sur 6* d'avoir l'image deux,... , *une chance sur 6* d'avoir l'image six. C'est exactement comme pour les lancers de dés : on a une chance sur 6 d'observer chaque face.

Les réponses (4) et (5) sont à coup sûr *justes* ; mais elles sont loin d'épuiser la question. Nous reviendrons sur la réponse (6) à la fin de cet article.

On pourrait se demander pourquoi une image de tigre, de girafe… de tapir, plutôt qu'un écureuil, un chat, un âne, une truie, une libellule et un renard, et adresser cette question à des spécialistes de marketing. Mais le contenu des images ne nous aide en rien à répondre à la question posée ; nous allons donc nous en abstraire en les numérotant : image 1,…, image 6.

I. Avec des dés et avec un ordinateur

Le nombre de plaques pour avoir la collection complète est supérieur à 6, et variable. Comment savoir si l'ordre de grandeur de *la plupart* des résultats est la dizaine, la vingtaine, cinquante, cent ?

Quelques résultats d'expériences seraient donc les bienvenus, pour se faire une idée :

Définition de l'expérience : *On achète des plaques de chocolat jusqu'à avoir les six*

1 C'est le principe des chaussettes : quand on a 10 paires de chaussettes dans un tiroir, en sortant 11 chaussettes, on est sûr d'avoir au moins une paire.

sortes d'images (aucun échange n'est autorisé dans ce protocole expérimental). Le résultat d'une expérience est le nombre de plaques achetées.

Comment réaliser ces expériences ? (voir [4]).On ne sait même pas de quelle marque de chocolat il s'agit ! Mais on peut aisément voir que la marque et les lieux de ventes ne sont pas nécessaires pour répondre à la question.

En fait, le chocolat lui-même ne joue ici aucun rôle, et on va donc s'en abstraire : il suffit de mettre des images numérotées dans une boîte, faire des tirages de ces images avec remise — sans oublier de bien mélanger les images entre deux tirages — et de noter au bout de combien de tirages on possède la collection complète.

I-1. Une première simulation

Il y a six images … et six faces aux dés les plus usuellement commercialisés. On va se servir de cette constatation pour remplacer les tirages d'images dans une boîte par une autre expérience, à savoir :

On lance un dé jusqu'à avoir obtenu chacune des six faces au moins une fois. Le résultat de l'expérience est le nombre de lancers du dé.

Cette nouvelle expérience est une simulation de la première, construite à partir de l'hypothèse qu'à chaque achat, on a une chance sur six d'avoir chacune des images. Sous cette hypothèse, une expérience simulée avec un dé donne la même information qu'une expérience originelle.

Cinq personnes ont ainsi chacune simulé dix expériences et obtenu les résultats consignés dans le tableau 1 ci-dessous.

A la question initiale, notre expérimentateur A1 peut répondre par la liste des 10 nombres obtenus, ou en disant que sur 10 expériences simulées, le nombre moyen observé est 17,2 et les 10 résultats sont entre 8 et 36. Cette réponse donne plus d'information que la réponse (5).

Cependant, 10 expériences, c'est peu, et les quatre autres expérimentateurs donneront, par le même procédé des réponses sensiblement différentes (cf. tableau 2 page suivante).

A1	23	13	8	17	15	18	36	8	19	15
A2	8	21	13	19	29	6	28	15	8	14
A3	12	23	15	14	10	11	16	21	28	13
A4	14	43	23	31	17	8	16	10	21	11
A5	6	12	20	14	12	8	15	7	10	7

Tableau 1 : Cinq expérimentateurs simulent avec un dé dix expériences chacun.
Chaque résultat est le nombre de lancers d'un dé « jusqu'à ce que chaque face ait été obtenue au moins une fois ». La somme de tous les nombres du tableau est 791 : les 50 expériences faites ont donné lieu à 791 lancers de dés, soit une moyenne de 791/50 = 15,82 lancés par expérience.

	moyenne	minimum	maximum
A1	17,2	8	36
A2	16,1	6	29
A3	16,3	10	28
A4	19,4	8	43
A5	11 ,1	6	20

Tableau 2 : Chacun des 5 expérimentateurs résume ses 10 expériences par la moyenne, le maximum et le minimum.

Le résultat d'une seule expérience est variable, les moyennes et les valeurs extrêmes pour des échantillons [2] de 10 expériences varient aussi : on parle alors de *fluctuation d'échantillonnage*. Cette fluctuation d'échantillonnage est « responsable » de la disparité des lignes du tableau 2.

Des réponses de natures diverses vont être données à la question posée, qui décrivent des aspects différents du phénomène. Mais il n'est pas satisfaisant, même si on traite d'un phénomène variable, et qu'on se contente d'un seul type de réponse (moyenne, minimum, maximum) que celle-ci varie sensiblement d'un expérimentateur à l'autre. Aussi, nous allons regarder ce qui se passe avec plus de simulations.

Si les expérimentateurs se rencontrent, il vont regrouper leur données et répondre collectivement que sur 50 expériences simulées, la moyenne observée (ou empirique) est 15,8 et tous les résultats sont entre 6 et 43. Le phénomène de fluctuation d'échantillonnage persiste-t-il avec 50 données ? Pour le constater on pourrait par exemple lancer des dés jusqu'à obtenir deux échantillons de 50 expériences. Les statisticiens n'aiment pas particulièrement jouer aux cartes ou aux dés ; mais ils savent adroitement « faire parler les dés » au sens suivant : ils organisent des expériences avec des lancers de dés, ou des tirages de boules dans des urnes, pour en simuler d'autres plus coûteuses ou délicates ; cependant, on dispose aujourd'hui d'outils plus puissants que des dés pour percer quelques secrets de l'aléatoire.

I-2. Simulation automatisée

Les chercheurs de toutes les disciplines scientifiques utilisent aujourd'hui la simulation aléatoire, numérique, géométrique. On simule ainsi des essais nucléaires, la diffusion d'un médicament dans l'organisme, la croissance de certaines tumeurs, l'effet de contraintes sur divers matériaux, les phénomènes de turbulence dans le sillage des avions, etc.

Simuler une expérience, c'est simuler un modèle possible de cette expérience. On peut alors, en confrontant les résultats des simu-

2 On dira qu'une liste de N nombres est un échantillon si ces nombres sont les résultats d'une même expérience, ou d'une même simulation, reproduite N fois dans des conditions identiques, les résultats d'une expérience ou d'une simulation n'influant pas sur les autres.

lations (on parle aussi d'expériences virtuelles) et ceux des expériences, valider ou infirmer un modèle : si les données expérimentales ne sont pas *conformes*, en un sens à définir, aux données simulées, le modèle s'en trouvera invalidé.

Lorsqu'un modèle a été suffisamment validé par diverses procédures pour qu'on décide de l'adopter, au moins pour quelques temps, la simulation permet de faire des prévisions quantitatives pour lesquelles un calcul « exact » serait trop complexe, et d'expliquer certains phénomènes observés expérimentalement.

De plus, la simulation fait parfois apparaître des lois (quantités invariantes par exemple) et on cherche alors à démontrer au niveau théorique comment les ingrédients mis pour construire le modèle peuvent conduire à de telles lois : la simulation intervient ainsi aujourd'hui comme outil à la fois en recherche fondamentale et en recherche appliquée.

Revenons aux images. Nous ne parlerons plus de *chances*, mais de *probabilités* (ce sont des *chances théoriques*). La traduction probabiliste de l'hypothèse « les images ont des chances égales d'apparaître » peut être formulée ainsi : « chaque image a une probabilité (ou chance théorique) $1/6 \approx 0{,}17$ de sortir à l'ouverture d'une plaque. Les simulations proposées ci-dessous sont fondées sur cette hypothèse probabiliste.

Il suffit ici de disposer d'une calculatrice de poche (voir annexe) ou d'un *tableur* et d'utiliser une instruction du type *Alea*(1,6), qui renvoie un nombre choisi au hasard entre 1 et 6, les six nombres ayant des chances égales d'être obtenus. On peut alors écrire un programme autour de cette instruction, grâce auquel on obtient un millier de résultats en une fraction de seconde.

Le tableau 3 résume des résultats pour cinq échantillons de 1000 simulations. (voir [5])L'ampleur observée de la fluctuation d'échantillonnage de la moyenne y est bien moindre qu'avec des échantillons de taille 10. C'est un phénomène général : des moyennes calculées sur des échantillons de taille n sont d'autant moins dispersées que n est grand. C'est pourquoi il est intéressant de calculer des moyennes sur un grand nombre de résultats.

Par contre pour *l'étendue* (différence entre la plus grande valeur observée et la plus petite), l'ampleur de la fluctuation ne diminue pas. En effet, dans la mesure ou le résultat n'est pas limité supérieurement, lorsque le nombre d'expériences augmente, on a plus de chances d'observer un « grand » nombre d'achats.

	moyenne	**maximum**
A1	15,1	49
A2	14,6	59
A3	14,7	48
A4	14,7	57
A5	14,5	46

Tableau 3 : Chaque échantillon de 1000 simulations est résumé par la moyenne, le maximum ; le minimum ici vaut toujours 6. Ne pas obtenir 6 comme plus petite valeur pour 1000 simulations est un événement théoriquement possible mais tellement rare (voir tableau 5) *qu'en pratique* , on ne l'observera pas : il faudrait construire des séries de taille 1000, à raison de 100 par seconde, pendant une durée de l'ordre du siècle pour avoir de fortes chances de ne pas avoir 6 comme minimum sur une série de 1 000 simulations.

Allons plus loin : avec un échantillon de 10 000 simulations, une réponse à la question du début de cet article est :

Sur 10 000 expériences simulées, la moyenne du nombre d'achats est 14,7, le nombre minimum est 6 et le nombre maximum est 70.

Une réponse donnée à partir d'un autre échantillon de 10 000 simulations fera état d'une moyenne égale ou très voisine de celle-ci ; autrement dit, en observant les résultats par paquets de 10 000, la fluctuation devient quasiment indécelable.

La question posée n'admet pas un seul « format » de réponse (moyenne, minimum et maximum d'une série de mesures expérimentales ou issues de simulations) et résumer 10 000 nombres par leur moyenne, le minimum et le maximum est quand même restrictif. De plus, le minimum est quasiment constant et le maximum ne rend compte que d'une donnée éventuellement exceptionnelle, la plus grande. Il est intéressant de choisir des indicateurs qui rendent compte de la dispersion des résultats possibles et qui fluctuent de moins en moins à mesure que le nombre des données sur lesquels ils sont calculés augmente.

On peut ainsi considérer le premier et le neuvième décile qui indiquent où se situent les données observées quand on a éliminé les 10% plus petites et les 10% plus grandes. Ainsi, dire que dans une série de données numériques, le premier décile est 8 et le neuvième est 23 signifie que :

10% des valeurs observées sont ≤ 8,
80% des valeurs observées sont entre 8 et 23 .
10% sont > 23 .

Comme résumé d'une valeur centrale, on peut aussi donner la valeur médiane, qui partage la série en deux : environ 50% des valeurs sont inférieures, environ 50% supérieures. Une réponse à la question originelle est :

Sur une série de 10 000 *simulations, on observe que :*
— la moyenne est 14,7, *la médiane est* 13,
— le minimum est 6, *le maximum est* 70,
— le premier décile est 8 *et le neuvième décile est* 23.

Si plusieurs personnes donnent des réponses obtenues à partir de simulations de taille 10 000, les réponses, sauf le maximum, différeront de moins d'une unité.

Rien n'oblige, pour la réponse à la question, à rester dans le champ des lettres et des nombres. On peut aussi ajouter l'histogramme des 10 000 valeurs simulées. (voir figure 1 en haut de la page suivante)

I-3. Quelques remarques

• Lors des premières simulations avec un dé, on pouvait envisager d'observer la liste des résultats un par un ; on ne peut plus le faire pour 10 000 nombres. On doit passer de l'observation d'une liste à celle d'un histogramme, d'une moyenne, de déciles. On est amené à prendre le parti de regarder le phénomène à une autre échelle, et non plus résultat par résultat. Et le miracle des probabilités, c'est que ce qui est totalement désordonné à l'échelon d'une unité se stabilise à grande échelle.

En changeant l'échelle d'observation, on voit ainsi apparaître des régularités, dans certaines valeurs numériques (moyenne et

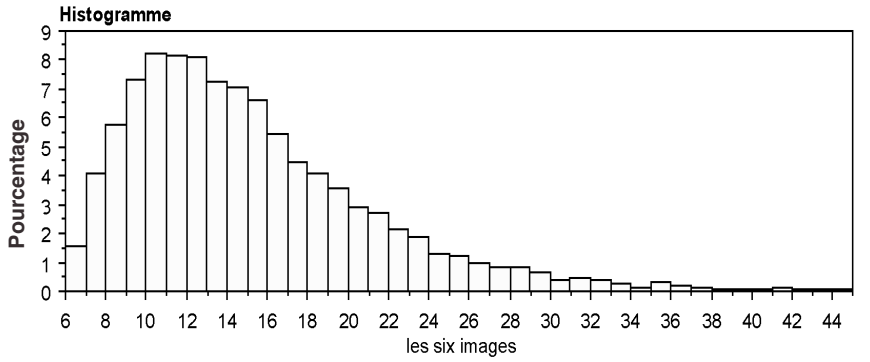

Figure 1 : 10 000 simulations pour six images. Les 10 plus grandes valeurs, qui vont de 46 à 70 ne sont pas prises en compte dans cet histogramme.

déciles par exemple) ou dans des formes (celle de l'histogramme, (voir figure 2 à la page suivante) : faire des statistiques, c'est ainsi changer constamment d'échelle d'observation.

• Il y a ici six images et, cela tombe bien, les dés usuels ont six faces. Mais avec les mêmes outils de simulations (tableurs ou calculatrices) on peut aussi résoudre diverses questions sur le nombre d'images distinctes à envisager :

(i) Combien de sortes d'images doit-on avoir pour qu'en moyenne, les amateurs de chocolat achètent 20 paquets pour les avoir toutes — sans faire d'échange entre eux ?

(ii) Combien de sortes d'images doit-on avoir pour que 50% des amateurs de chocolat aient la collection avec 10 plaques ou moins — sans faire d'échange entre eux ?

(iii) Quel est le plus grand nombre d'images distinctes possible pour qu'il y ait au plus 10% de mécontents, sachant qu'un mécontent est celui qui doit acheter 25 plaques ou plus pour avoir la collection ?

Considérons pour cela, pour des nombres différents d'images distinctes des séries de 10 000 expériences (tableau 4 ci-dessous).

Nombres d'images distinctes	4	5	6	7	8	9	10
Moyenne	8,4	11,5	14,7	18,3	21,7	25,4	29,5
médiane	7	10	13	16	20	23	27
9 ème décile	13	18	23	28	33	38	44

Tableau 4 : Pour un nombre n d'images distinctes, n = 4,… 10, 10 000 simulations ont été effectuées.

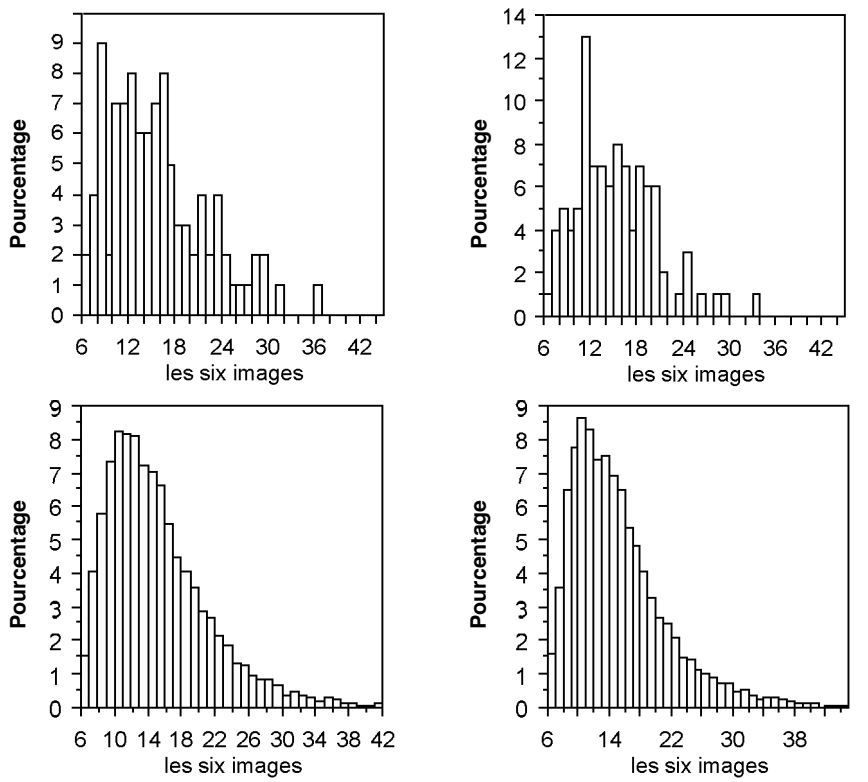

Figure 2 : Les deux premiers histogrammes sont tracés à partir 100 données chacuns : ils ne donnent pas l'impression d'une forme invariante ; les deux histogrammes du bas représentent deux séries distinctes de 10 000 simulations et on voit là se dessiner une forme invariante , celle d'un « histogramme théorique ».

En consultant le tableau 4, les réponses données au vu des simulations seront :

(i) 8 sortes d'images

(ii) 5 sortes d'images

(iii) 6 sortes d'images

On n'est pas « absolument sûr » des résultat produits ci-dessus, on n'est jamais à l'abri d'un échantillon aberrant. Le travail du statisticien est ici de quantifier ce risque. Néanmoins, dans la situation considérée où une erreur n'a pas de conséquence vraiment fâcheuse, des résultats obtenus sur 10 000 données sont suffisants pour prendre la décision commerciale du nombre d'images à fabriquer pour satisfaire à telle ou telle contrainte.

II. Des résultats théoriques

Dans ce paragraphe, nous allons donner d'autres réponses à la question initiale. Ces réponses, théoriques, sont issues de techniques très classiques du *calcul des probabilités*. Elles ne sont ni intrinsèquement meilleures ni moins bonnes que les précédentes : tout dépend de ce qu'on veut en faire. Dans l'exemple des plaques de chocolat, les réponses précédentes, basées sur des résultats de simulation, suffisent pour avoir une bonne idée du phénomène en jeu. Par contre, pour généraliser, ou élargir la question à d'autres situations où le risque et l'enjeu sont d'une autre nature, alors les réponses théoriques qu'on va donner ci-dessous peuvent s'avérer plus pertinentes.

L'intérêt ici est aussi de comparer résultats théoriques et estimations issues des simulations faites, afin de cerner la puissance et les limites de ce vaste champ qu'est la simulation aléatoire.

II-1. Premiers résultats en vrac

Nous donnons ici quelques résultats théoriques qui s'obtiennent aisément, pourvu que l'on regarde l'expérience sous plusieurs angles. Associer formules et tables de valeurs prises par les éléments calculables par ces formules permet, indépendamment de la question posée en début de ce texte, de se faire déjà une bonne idée du phénomène étudié et confirme certains résultats observés sur les simulations.

• On peut calculer la probabilité p_6 pour avoir toutes les images en un minimum d'achats, soit ici en 6 achats :

$$p_6 = \frac{5\times4\times3\times2}{6\times6\times6\times6\times6} \approx 0{,}015.$$

Les autres probabilités p_7 ,... p_n ,... d'avoir les 6 images distinctes en 7, ... , n,... achats sont difficiles à calculer. Quelques valeurs sont données dans la troisième ligne du tableau 10 ci-dessous.

On notera que si X est la variable aléatoire qui donne le nombre de plaques qu'il a fallu acheter pour avoir les 6 images , on a :

$$p_k = \text{Prob}(X = k)$$

Dans cette vision de l'expérience, on achète les plaques une par une et on s'arrête dès qu'on a les 6 images. En particulier, $X = k$ signifie qu'à l'achat $k - 1$, on avait 5 sortes d'images et que le $k$ème achat a fourni l'image manquante.

• La probabilité r_s que sur s expériences, le minimum soit 6 est donné par la formule :

$$r_s = 1 - (1 - p_6)^s$$

On peut voir quelques valeurs numériques dans le tableau 5. On voit que sur des échantillons de taille supérieure ou égale à 1000, la probabilité que le minimum soit 6 est très proche de 1 : on ne s'étonnera donc pas des résultats sur le minium observés dans le tableau 3.

s	10	50	100	200	500	1000
r_s	0,15	0,54	0,79	0,96	0,9996	$1 - 2 \times 10^{-7}$

Tableau 5 : probabilité pour que le minimum des résultat de s expériences soit 6.

• Notons q_n la probabilité d'avoir les 6 images lorsqu'on achète n plaques ; ici, on ne regarde pas les plaques les unes après les autres : on en achète n, puis on regarde si on a toute la collection.

Notons $A_i(n)$ l'événement « l'image i n'a pas été obtenue lors de n achats ». Alors :

$$q_n = 1 - \text{Prob}\,(A_1(n) \text{ ou } A_2(n) \text{ ou } \ldots \text{ ou } A_6(n))$$

$$\geq 1 - \sum_{i=1}^{6} \text{Prob}(A_i(n)) = 1 - 6 \times \left(\frac{5}{6}\right)^n = K_n$$

Le tableau 6 donne des valeurs du minorant K_n de q_n. On voit qu'avec 14 images, on a une probabilité supérieure à 0,5 d'avoir les 6 images et avec 25 images, elle est supérieure à 0,94 . Le tableau 10 complètera, à l'issue de calculs plus complexes, le tableau 6 en donnant non plus un minorant, mais une valeur approchée à 0,005 près de q_n .

En théorie, le nombre de plaques à acheter n'est pas borné, mais on peut déjà le borner par 50 avec une probabilité d'erreur inférieure à 1/1000. Dire que le résultat est entre 6 et 50 avec une probabilité supérieure à 0,999, c'est-à-dire diminuer l'ensemble des valeurs possibles au prix d'un risque *petit* (ici inférieur à 1/1000) est le genre de réponse que l'on cherche à donner en statistique.

• Notons Y_n la variable « compteur », celle qui compte le nombre d'images distinctes pour n achats. Elle prend ses valeurs dans {1, 2, 3, 4, 5, 6}. On a :

$$q_n = \text{Prob}(Y_n = 6).$$

On peut aussi écrire :

$$Y_n = Z_n^1 + \ldots . Z_n^6$$

où Z_n^i est une variable qui vaut 1 si l'image i est présente parmi les n images, et 0 sinon. Or l'espérance $E(Z_n^i)$ de Z_n^i s'écrit :

$$E(Z_n^i) = \text{Prob}(Z_n^i = 1) = 1 - (5/6)^n$$

Et l'espérance $E(Y_n)$ de Y_n est donc :

$$E(Y_n) = 6 \times (1 - (5/6)^n)$$

D'où le tableau 7. On voit ainsi que si un grand nombre de gens achètent dix plaques, en moyenne ils auront environ 5 images. Pour

n	10	11	12	13	14	15	20	25	50
K_n	0.03	0.19	0.33	0.44	0.53	0.61	0.84	0.94	0.999

Tableau 6 : La deuxième ligne donne un minorant de la probabilité d'avoir les 6 images en achetant n plaques.

n	5	8	9	10	12	13	15	20	25	50
$E(Y_n)$	3.59	4.6	4.83	5.03	5.33	5.44	5.61	5.67	5.94	5.999

Tableau 7 : La deuxième ligne donne la moyenne théorique, ou espérance, du nombre d'images différentes obtenues en achetant n plaques.

50 images, l'espérance de la variable aléatoire Y_n est très proche de la plus grande valeur qu'elle peut prendre, à savoir 6 : la probabilité qu'elle vaille 6 est donc très proche de 1.

• Revenons à la variable X qui, lorsqu'on achète les plaques les unes après les autres, donne le nombre d'achats nécessaires pour avoir toute la collection. Ecrivons :

$$X = 1 + X_2 + X_3 + \ldots + X_6 ,$$

où X_i désigne le nombre d'achats qui a permis de passer de $i-1$ à i images, $i = 2, \ldots, 6$ (voir tableau 8).

Lorsqu'on a obtenu $i-1$ images distinctes, la probabilité d'augmenter le compteur d'une unité au bout de k achats supplémentaires est :

$$\text{Prob}\,(X_i = k) = \left(\frac{i-1}{6}\right)^{k-1}\left(\frac{6-i+1}{6}\right)$$

La variable X_i suit une loi appelée loi géométrique de paramètre $((7-i)/6)$, son espérance est $6/(7-i)$. D'où le calcul exact de l'espérance de X_i, noté m_6 pour la situation envisagée avec 6 sortes d'images :

$$m_6 = 6\left(\frac{1}{6}+\frac{1}{5}+\frac{1}{4}+\frac{1}{3}+\frac{1}{2}+1\right) = 14{,}7$$

Un calcul analogue avec N sortes d'images donne :

$$m_N = N\left(\frac{1}{N}+\frac{1}{N-1}+\frac{1}{N-2}\ldots+\frac{1}{4}+\frac{1}{3}+\frac{1}{2}+1\right)$$

On constatera que les moyennes théoriques du tableau 9 sont proches des moyennes calculées sur 10000 simulations (tableau 4). C'est une conséquence de la loi des grands nombres, théorème qui énonce que les fréquences des résultats d'une expérience aléatoire "*convergent*" vers leurs probabilités.

II-2. Etude d'un compteur d'images

De nombreux autres résultats peuvent être démontrés à partir de l'idée suivante. On

Numéro d'achat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Numéro d'image	2	1	4	1	6	6	1	2	4	3	3	2	4	1	5
compteur	**1**	**2**	**3**	3	**4**	4	4	4	4	**5**	5	5	5	5	**6**

Tableau 8 : La valeur de X est ici 15. En troisième ligne, l'évolution du compteur du nombre d'images distinctes pour les paquets successifs d'images achetées. Les valeurs de $X_2, .., X_6$ sont respectivement 1, 1, 2, 5, 5.

Nombres d'images distinctes	4	5	6	7	8	9	10
Moyenne théorique	8,3	11,4	14,7	18,1	21,7	25,5	29,3

Tableau 9 : Calculs des moyennes théoriques lorsque le nombre d'images différentes vaut 4, 5, …, 10.

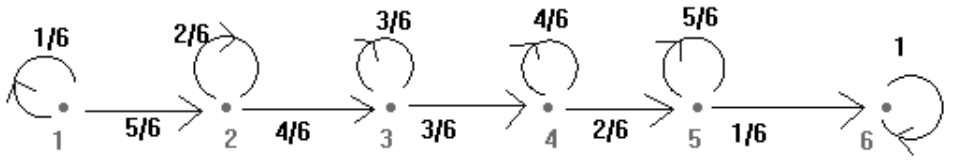

Figure 3 : Graphe fournissant les probabilités de passage du compteur d'une valeur à une autre. Ainsi, si on dispose de 4 images distinctes, à l'achat suivant on a 4 chances sur 6 (soit une probabilité 4/6) d'obtenir une des images qu'on a déjà, et 2 chances sur 6 (soit une probabilité 2/6) d'avoir une nouvelle image, le compteur prenant alors la valeur 5.

définit une suite $(Y_n)_n$ de *compteurs* qui, après chaque achat, comptent le nombre d'images distinctes. Quand le compteur prend la valeur 6, il garde définitivement cette valeur ; on suppose qu'on n'arrête pas pour autant d'acheter des plaques.

On peut associer au compteur ainsi défini le graphe de la figure 3 ci-dessus. On peut aussi associer à cette situation la matrice $\boldsymbol{T} = ((p_{ij})_{ij}$ dont le terme p_{ij} d'indices (i,j) est la probabilité, sachant qu'on a i images, de passer à l'achat suivant à j images ; cette probabilité ne dépend pas du nombre d'achats qui a conduit à avoir i images, soit :

$p_{i,j} = \text{Prob}(Y_{n+1} = j \,/\, Y_n = i)$, pour tout $n > 0$

Cette matrice s'écrit :

$$\mathbf{T} = \begin{pmatrix} 1/6 & 5/6 & 0 & 0 & 0 & 0 \\ 0 & 2/6 & 4/6 & 0 & 0 & 0 \\ 0 & 0 & 3/6 & 3/6 & 0 & 0 \\ 0 & 0 & 0 & 4/6 & 2/6 & 0 \\ 0 & 0 & 0 & 0 & 5/6 & 1/6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Notons $P_n(j)$ la probabilité que le compteur vaille j à l'achat n, soit :

$P_n(j) = \text{Prob}(Y_n = j)$, $j = 1, ..., 6$;

et avec le notations précédentes : $q_n = P_n(6)$.

La formule des probabilités composées permet alors de décomposer la probabilité que le compteur soit dans l'état j à l'étape n :

$\text{Prob}(Y_n = j) = \text{Prob}(Y_n = j \text{ et } Y_{n-1} = 1) + \ldots$
$\ldots + \text{Prob}(Y_n = j \text{ et } Y_{n-1} = 6).$

D'où :

$\text{Prob}(Y_n = j) =$
$\text{Prob}(Y_n = j \,/\, Y_{n-1} = 1) \times P_{n-1}(1) + \ldots$
$\ldots + \text{Prob}(Y_n = j \,/\, Y_{n-1} = 6) \times P_{n-1}(6)$

.

Ce qui s'écrit :

$P_n(j) = p_{1j} \times P_{n-1}(1) + \ldots +$
$p_{6j} \times P_{n-1}(6)$.

Notons $\boldsymbol{P}_n$ la matrice ligne :

$\boldsymbol{P}_n = (P_n(1)\ldots..,\ldots, P_n(6))$

de l'égalité ci-dessus , on déduit :

$\boldsymbol{P}_n = \boldsymbol{P}_{n-1} \times \mathbf{T}$

Il en découle :

$\boldsymbol{P}_n = \boldsymbol{P}_1 \times \mathbf{T}^{n-1}$
où $\boldsymbol{P}_1$ = (1,0, …,0) décrit l'état du compteur après le premier achat.

On peut donc calculer, avec un logiciel de calcul, des lois de probabilités $\boldsymbol{P}_n$:

$\boldsymbol{P}_{10} = \boldsymbol{P}_1 \times \mathbf{T}^9 =$
(0,0000 ; 0,0003 ; 0,0185 ; 0,2031 ; 0,5064 ; 0,2718)

$\boldsymbol{P}_{20} = \boldsymbol{P}_1 \times \mathbf{T}^{19} =$
(0,0000 ; 0,0000 ; 0,0000 ; 0,0045 ; 0,1475 ; 0,8480)

$\boldsymbol{P}_{50} = \boldsymbol{P}_1 \times \mathbf{T}^{49} =$
(0,0000 ; 0,0000 ; 0,0000 ; 0,0000 ; 0,0007 ; 0,9993)

On voit en particulier que la probabilité d'avoir 6 (resp. 5) images au bout de 20 achats vaut 0,8480 (resp. 0,1475), à 5×10^{-5} près.

D'autres résultats numériques sont donnés tableaux 10 et 11. Dans le tableau 10, on lit que si on achète 14 plaques, la probabilité d'avoir les 6 sortes d'images est 0,58 ; ou encore : on a une probabilité 0,58 qu'il faille 14 achats ou moins pour avoir les 6 images, et une probabilité 0,42 qu'il faille au moins 15 plaques.

Nous pouvons maintenant aussi faire le lien entre le point de vue « compteur » et le point de vue « nombre exact d'achats nécessaires pour avoir la collection complète des 6 images ». En effet, dire qu'à l'achat n le compteur vaut 6, c'est dire qu'il a fallu au plus n achats pour avoir les 6 images :

$$\text{Prob}\,(Y_n = 6) = \text{Prob}(X \leq n),$$

Le compteur Y_n vaut 6 soit parce qu'il valait déjà 6 à l'achat $n-1$ (la probabilité d'un tel événement est q_{n-1}), soit parce qu'

n	5	6	7	8	9	10	11	12	13	14	15
q_n	0	0,015	0,054	0,114	0,189	0,272	0,356	0,438	0,514	0,583	0,644
p_n	0	0,015	0,039	0,060	0,075	0,083	0,084	0,081	0,076	0,069	0,061
$1 - q_n$	1	0,985	0,946	0,886	0,811	0,728	0,644	0,562	0,486	0,417	0,456

Tableau 10 : En deuxième ligne, on a la probabilité q_n pour qu'en achetant un lot de n paquets, on ait les six sortes d'images. Par exemple, avec 13 achats, il y a 52 chances sur 100 pour qu'on ait les six images distinctes. En troisième ligne, on a la probabilité p_n qu'il faille exactement n achats pour avoir les 6 images distinctes. Les quantités q_n et p_n sont liées :

$$q_{13} = p_{13} + p_{12} + p_{11} + \ldots + p_6 = p_{13} + q_{12} \text{ et plus généralement } p_n = q_n - q_{n-1}.$$

La quatrième ligne donne la probabilité $1 - q_n$ pour que le résultat d'une expérience soit strictement supérieur à n. (on a ainsi environ une chance sur deux que le résultat dépasse 13).
La médiane de X est le plus le plus petit entier m tel la probabilité qu'il faille acheter m plaques ou moins est supérieure ou égale à 0,5. C'est donc 13. Le premier décile de X est par définition le plus petit entier d tel la probabilité qu'il faille acheter d plaques ou moins est supérieure ou égale à 0,10. Ici, le premier décile est 8.

n	20	21	22	23	25	50	100
q_n	0,85	0,87	0,89	0,91	0,94	0,999	$1 - 10^{-7}$

Tableau 11 : En deuxième ligne, on a la probabilité q_n pour qu'en achetant un lot de n paquets, on ait les six sortes d'images.
Le neuvième décile de X est le plus petit entier d' tel la probabilité qu'il faille acheter d' plaques ou moins est supérieure ou égale à 0,90. Ici, le neuvième décile est 23.

il est passé à la valeur 6 à l'achat n (la probabilité d'un tel événement est la probabilité p_n = Prob ($X = n$), définie au début de ce paragraphe). On a donc :

$$\text{Prob}(Y_n = 6) = \text{Prob}(Y_{n-1} = 6) + \text{Prob}\ (X = n)$$

Avec les notations choisies, on a donc la formule suivante qui relie le point de vue « compteur du nombre d'images distinctes » :

$$q_n = p_n + q_{n-1}\,.$$

Les tableaux 10 et 11 constituent aussi une réponse possible à la question initiale. On y voit que les premiers et neuvièmes déciles et la médiane de X sont respectivement 8, 23 et 13.

Le modèle du compteur est une *chaîne de Markov*. De tels modèles servent à étudier notamment les systèmes à nombre fini d'états, qui peuvent changer d'état à tout instant t discret (i.e. t est un entier positif), et où la probabilité de passer d'un état à un autre ne dépend pas de t (voir [1, 6, 7, 8]). De tels modèles ont de nombreuses propriétés, ce sont des objets mathématiques riches qui sont de plus actuellement beaucoup utilisés, par exemple en génomique ((voir [8]).).

II.3. Prendre des risques

Le nombre d'achats à faire pour avoir les six images n'est pas borné : la valeur de $1 - q_n = \text{Prob}(X > n)$ n'est nulle pour aucun n.

C'est la théorie. Bon, mais en pratique, on serait peu crédible de vivre comme si le nombre de plaques pouvait être aussi grand que l'on veut, ou comme si un singe pouvait par hasard dactylographier l'Odyssée. Pour n « grand », $\text{Prob}(X > n)$ est en effet si petite qu'on est conduit à la négliger : si on ne négligeait pas constamment les faibles probabilités, on ne prendrait jamais sa voiture, on ne descendrait jamais un escalier, on ne prendrait jamais un médicament, etc. et ce faisant, on augmenterait considérablement la probabilité d'autres événements fort néfastes !

Dans certains cas, il convient de connaître l'ordre de grandeur de la probabilité d'un événement dont on va en pratique considérer qu'il ne peut pas arriver. Le métier du statisticien est entre autre de les estimer. Ici, on sait faire les calculs, et un statisticien dira :

Le nombre de plaques à acheter
est inférieur à 50*, au risque* 1/1000.

où encore :

Le nombre de plaques à acheter
est inférieur à 100*, au risque* 10^{-7},

Il y a toujours un équilibre à trouver entre le risque consenti et le résultat à donner modulo ce risque ; mais le choix du niveau de risque relève d'une décision extérieure à l'étude statistique.

En guise de conclusion

Les enjeux, quantifiés en coûts des situations où des décisions réelles sont à prendre, n'apparaissent pas dans l'histoire racontée ci-dessus, mais celle-ci illustre cependant les articulations entre les différentes pratiques que sont l'expérimentation, la modélisation, la simulation .

La science de l'aléatoire est actuellement une des parties les plus visibles des sciences mathématiques. Toutes les sciences — et les medias — font appel à son langage, à ses méthodes et à ses résultats. Donc, tout le monde « devrait » connaître des éléments de statistique : première partie d'un discours maintenant bien rodé, dont la deuxième partie consiste le plus souvent à s'extasier sur les innombrables erreurs inévitables dans ce domaine.

L'injonction *« Faites de la statistique — mais vous ferez certaines des erreurs qui vous attendent inexorablement»* n'est pas de nature à donner envie de faire de la statistique.

Aussi, plutôt que de dresser une première typologie des erreurs à éviter, nous avons cherché à convaincre que la statistique n'est pas une promenade dans un champ de mines, mais une pensée à construire calmement, patiemment, en repassant de nombreuses fois sur les mêmes concepts et les mêmes idées, jusqu'à ce qu'ils deviennent *naturel*s et conduisent plus fermement la

Bibliographie :

[1] "An introduction to probability theory an its application". William Feller. (Voir à "Collector's problem"). Wiley, 1950.

[2] « En passant par hasard », G. Pages et C. Bouzitat, chapitre VI et VII, édition Vuibert.

[3] « Graines de Sciences 6 », *Le regard Statistique,* Claudine Robert page 71 à 101, édition le Pommier.

[4] http://www.mste.uiuc.edu/reese/cereal/cereal.html. Applet java qui simule des expériences avec des images d'animaux. Le programme Java est fourni.

[5] Un logiciel de simulations est téléchargeable sur le site http://perso.wanadoo.fr/jpq, dans la partie probabilités, sous le titre « images dans une boîte de céréales ». On peut simuler aisément 10 000 expériences, et aussi choisir le nombre d'images distinctes dont on dispose.

[6] « Modèles et algorithmes markoviens » . B. Ycart. Editions Springer 2002.

[7] « Chaînes de Markov. *Cours de Maîtrise d'Ingénierie mathématique, Université René Descartes, Paris, novembre 1999. Télechargeable à partir de :*

http://www.math-info.univ-paris5.fr/~ycart/polys/polys.html

[8] ADN, mots et modèles. S. Robin, F. Rodolphe, S.Schbath.