Journée de formation 15 janvier 2015

Modélisation d'une expérience aléatoire : les lois de Mendel

Sophie Pénisson, Université Paris-Est Créteil

Résumé. L'atelier a comme fil rouge les lois de Mendel, sur l'hérédité biologique. Il débute par la modélisation de l'expérience de Mendel (croisement de lignées pures de pois) et par sa simulation numérique. Le modèle probabiliste sous-jacent est une loi binomiale. La deuxième partie de l'atelier est consacrée aux statistiques liées à cette expérience, et permet d'illustrer sur différents exemples les notions d'intervalle de fluctuation et d'intervalle de confiance.

1 Modélisation

Les lois de Mendel concerne les principes de l'hérédité biologique, énoncées par le moine et botaniste autrichien Gregor Mendel (1822-1884) à partir d'expérience menées sur des pois. La forme du pois existe selon deux variantes : graine ronde (allèle dominant R), ou graine ridée (allèle récessif r).

Question 1

Une des expériences de Mendel consista à croiser des lignées pures de pois ayant des graines rondes (génotype RR) avec des lignées pures de pois ayant des graines ridées (génotype rr). Il obtint ainsi une première génération de pois, tous ronds, de génotype Rr. Il compta alors le nombre de pois ronds obtenu après croisement de ces descendants.

Écrire un algorithme simple modélisant l'expérience aléatoire consistant à faire n croisements de pois de génotype Rr, en n'utilisant que des générateurs de nombres aléatoires entre 0 et 1, sans faire intervenir de loi de probabilité.

Il s'agit d'itérer n fois l'expérience aléatoire consistant à choisir avec une probabilité $\frac{1}{2}$ l'allèle de chacun des deux parents (R ou r), de compiler le génotype correspondant (4 possibilités) et enfin d'écrire le phénotype qui en résulte (rond ou ridé).

La simulation du choix aléatoire de l'allèle chez chacun des parents peut se faire grâce à un générateur de nombre aléatoire dans $[0, 1]$. Par exemple `ALEA()` sous Excel ou `rand()` sous Scilab. L'événement `ALEA()<1/2` (ou `rand()<1/2`) se produisant avec probabilité $\frac{1}{2}$, on peut simuler le choix aléatoire d'un allèle par

`SI(ALEA()<0.5;1;0)` (Excel)

`1*(rand()<0.5)` (Scilab)

Afin de traduire la dominance de l'allèle R, il est pratique ici de coder le choix de l'allèle R chez un parent par 0. Le phénotype du descendant sera alors simplement obtenu par multiplication des deux allèles choisis : 0 codera pour le phénotype pois rond, 1 pour le phénotype pois ridé.

```
n=5000;
P=zeros(n,1);
for k=1:n,
A1=1*(rand()<0.5); \\choix de l'allèle 1
A2=1*(rand()<0.5);  \\choix de l'allèle 2
P(k)=A1*A2; \\phénotype du descendant issu du k-ème croisement
end,
X=n-sum(P); \\nombre de pois ronds
```

Question 2

Soit X_n le nombre de pois ronds obtenu après n croisements.

Quelle est la loi de X_n ? En déduire un algorithme plus direct modélisant l'expérience précédente, en n'utilisant toujours que des générateurs de nombres aléatoires entre 0 et 1.

Le nombre de pois ronds suit la loi binomiale $\mathcal{B}(n, \frac{3}{4})$, et correspond donc à la somme de n variables aléatoires indépendantes de même loi $\mathcal{B}(\frac{3}{4})$. Puisque l'événement `rand()<3/4` se produit avec probabilité $\frac{3}{4}$, en codant par 1 l'événement "phénotype pois rond" on obtient l'algorithme

```
n=5000;
P=zeros(n,1);
for k=1:n,
P(k)=1*(rand()<0.75);
end,
X=sum(P); \\nombre de pois ronds
```

ou

```
n=5000;
X=0;
for k=1:n,
X=X+1*(rand()<0.75);
end
```

Question 3

On suppose maintenant que les descendants ne sont pas nécessairement issus de lignées pures, et que la probabilité d'obtenir un pois rond lors d'un croisement est $p \in [0, 1]$. Soit X_n le nombre de pois ronds obtenu après n croisements.

Généraliser l'algorithme de la question 2 à ce cas.

```
n=5000;
X=0;
for k=1:n,
X=X+1*(rand()<p);
end
```

2 Intervalle de fluctuation

On observe n pois, chaque pois ayant une probabilité p d'être rond. On note X_n le nombre de pois ronds, et l'on s'intéresse à la fréquence $F_n = \frac{X_n}{n}$ des pois ronds observés. Dans ce premier cas de figure, on suppose que **le paramètre p est connu**, et l'on cherche des informations sur les valeurs possibles de F_n à partir de p.

2.1 Théorie

Définition 1. Un **intervalle de fluctuation asymptotique** au seuil $1-\alpha$ pour une variable aléatoire Y_n est un intervalle réel $[a_n, b_n]$ qui contient Y_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand :

$$\lim_{n\to\infty} P\left(Y_n \in [a_n, b_n]\right) \geqslant 1-\alpha.$$

Remarque 1. En anglais, on appelle *prediction interval* l'intervalle de fluctuation.

Remarque 2. L'intervalle de fluctuation permet de détecter un écart important par rapport à la valeur théorique pour une grandeur établie sur un échantillon. C'est un intervalle dans lequel la grandeur observée est censée se trouver avec une forte probabilité (souvent de l'ordre de 95 %, c'est-à-dire en choisissant $\alpha = 0.05$.). Le fait d'obtenir une valeur en dehors de cet intervalle s'interprète alors en mettant en cause la représentativité de l'échantillon ou la valeur théorique.

Remarque 3. Dans l'exemple étudié, où $X_n \sim \mathcal{B}(n,p)$, l'intervalle $[0,1]$ est un intervalle de fluctuation évident (au seuil 1) pour la variable aléatoire F_n.

Proposition 1.
Soit $X_n \sim \mathcal{B}(n,p)$ et $F_n = \frac{X_n}{n}$. Alors

$$\lim_{n\to\infty} P\left(F_n \in \left[p - u_\alpha\sqrt{\frac{p(1-p)}{n}}; p + u_\alpha\sqrt{\frac{p(1-p)}{n}}\right]\right) = 1-\alpha,$$

où u_α est tel que $P(|Z| \geqslant u_\alpha) = \alpha$, pour $Z \sim \mathcal{N}(0,1)$.

Remarque 4. $u_{0.05} = 1.96$ et $u_{0.01} = 2.58$.

Remarque 5. On pratique cette approximation dès que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Remarque 6. Dans le programme de Seconde, on fournit l'intervalle de fluctuation asymptotique simplifié, valable pour $n \geqslant 25$ et $0.2 \leqslant p \leqslant 0.8$:

$$P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0.95.$$

En effet, en utilisant le fait que $p(1-p) \leqslant \frac{1}{4}$,

$$\left[p - u_{0.05}\sqrt{\frac{p(1-p)}{n}}; p + u_{0.05}\sqrt{\frac{p(1-p)}{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right].$$

2.2 Illustration numérique

On suppose ici le paramètre p connu, égal à $\frac{3}{4}$. Pour un échantillon de taille $n = 5000$, l'intervalle de fluctuation asymptotique de niveau 90% pour F_{5000}, sachant que $u_{0.1} = 1.64$, est :

$$\left[\frac{3}{4} - 1.64\sqrt{\frac{\frac{3}{4}\frac{1}{4}}{5000}}; \frac{3}{4} + 1.64\sqrt{\frac{\frac{3}{4}\frac{1}{4}}{5000}}\right] = [0.740; 0.760].$$

Simuler 200 échantillons de taille 5000 de la loi $\mathcal{B}(\frac{3}{4})$ et relever à chaque fois la fréquence f de 1 obtenue. La plupart des fréquences observées devraient se trouver dans l'intervalle de fluctuation $[0.74; 0.76]$, mais certaines (de l'ordre de 10%) peuvent se trouver en dehors. Déterminer combien parmi les 200 fréquences sont en-dehors de l'intervalle.

```
p=3/4;
N=200; n=5000; F=zeros(N,1);
for i=1:N,
X=0;
for k=1:n, X=X+1*(rand()<p); end,
F(i)=X/n; \\i-ème fréquence observée
end,
v=find(F<0.74 | F>0.76); \\cherche les fréquences en-dehors de l'intervalle de fluctuation
length(v), \\nombre de fréquences en-dehors de l'intervalle de fluctuation
F(v), \\fréquences en-dehors de l'intervalle de fluctuation
```

3 Intervalle de confiance

On observe n pois, chaque pois ayant une probabilité p d'être rond. On note X_n le nombre de pois ronds, et l'on s'intéresse à la fréquence $F_n = \frac{X_n}{n}$ des pois ronds observés. Dans ce second cas de figure, **le paramètre p est inconnu**, et l'on cherche des informations sur les valeurs possibles de p à partir de l'observation d'un échantillon.

3.1 Théorie

Définition 2.
Soit θ un paramètre inconnu de la loi d'une variable aléatoire Y, et soit $(Y_1, \dots, Y_n)$ n variables aléatoires indépendantes et de même loi que Y. Un **intervalle de confiance** pour le paramètre θ à un niveau de confiance $1-\alpha$ est un intervalle aléatoire I_n déterminé à partir de $(Y_1, \dots, Y_n)$, contenant θ avec une probabilité supérieure ou égale à $1-\alpha$:

$$P(\theta \in I_n) \geqslant 1-\alpha.$$

Proposition 2.
Soit $X_n \sim \mathcal{B}(n,p)$ et $F_n = \frac{X_n}{n}$. On suppose que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$. Alors un intervalle de confiance de niveau 0.95 pour p est $I_n = [F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}]$:

$$P\left(p \in \left[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0.95.$$

Remarque 7. Le lien avec la Définition 2 se fait en considérant n variables aléatoires indépendantes $(Y_1, \dots, Y_n)$ de loi de Bernoulli $\mathcal{B}(p)$, avec p inconnu. Alors $X_n := \sum_{i=1}^n Y_i \sim \mathcal{B}(n,p)$, et l'intervalle aléatoire construit à partir de $F_n = \frac{X_n}{n}$ est donc bien construit à partir de $(Y_1, \dots, Y_n)$.

Remarque 8. Notez que, à la différence de l'intervalle de fluctuation, un intervalle de confiance est un intervalle **aléatoire**. Ainsi, en effectuant le tirage d'un échantillon, on obtient une réalisation de cet intervalle aléatoire qui fournit alors un intervalle **numérique** de la forme $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$. Si l'on fait un très grand nombre de tirages, on sait que théoriquement on devrait avoir pour au plus 5% d'entre eux des intervalles ne contenant pas la proportion inconnue p.

3.2 Illustration numérique

Simuler de manière aléatoire un paramètre $p \in [0,1]$, qui doit rester inconnu de l'utilisateur mais accessible à la fin de l'expérience. Simuler 200 échantillons de taille 5000 de la loi $\mathcal{B}(\frac{3}{4})$ et relever à chaque fois les valeurs numériques de l'intervalle de confiance de niveau 95% pour le paramètre p inconnu : $[f - \frac{1}{\sqrt{5000}}; f + \frac{1}{\sqrt{5000}}]$.

Ces 200 réalisations de l'intervalle de confiance (la plupart du temps, seule une réalisation de l'intervalle de confiance est disponible) seraient dans la pratique la conclusion de l'étude et la seule information disponible sur le paramètre p. Confronter ici ces informations avec la valeur réelle de p en la révélant a posteriori. La plupart des réalisations de l'intervalle de confiance contiennent effectivement le paramètre p mais certaines (de l'ordre de 5%) ne le contienne pas. Déterminer combien parmi les 200 intervalles ne contiennent pas p. Illustrer graphiquement l'appartenance de p à certains intervalles et la non-appartenance à d'autres.

```
p=rand();
N=200; n=5000;
I=zeros(N,2);
for i=1:N,
X=0;
for k=1:n, X=X+1*(rand()<p); end,
I(i,1)=X/n-1/sqrt(n); I(i,2)=X/n+1/sqrt(n); \\i-ème réalisation de l'intervalle de confiance
end,
p, \\révèle a posteriori la valeur de p
v=find(I(:,1)>p | I(:,2)<p); \\cherche les intervalles ne contenant pas p
length(v), \\nombre d'intervalles ne contenant pas p
I(v,:), \\intervalles ne contenant pas p
plot(I);
plot(ones(1,N)*p);
```

4 Exercices d'application

Situation 1 (prise de décision)

Mendel, qui vient d'échafauder sa théorie et sait que $p = \frac{3}{4}$, voudrait s'assurer que ses résultats d'observation d'un échantillon de taille 7324, obtenu par croisement de lignées pures, valident cette théorie. La règle de décision prise est la suivante : si la proportion de pois ronds observée est en dehors de l'intervalle de fluctuation asymptotique au seuil de 95%, alors il truquera ses résultats. Que déciderait-il s'il observe 5574 pois ronds ? S'il en observe 5474 (qui fut le résultat publié) ?

Intervalle de fluctuation asymptotique de niveau 95% *pour un échantillon de taille* $n = 7324$ *et pour une proportion* $p = \frac{3}{4}$ *:*

$$\left[\frac{3}{4} - 1.96\sqrt{\frac{3}{4}\frac{1}{4}\frac{1}{7324}}; \frac{3}{4} + 1.96\sqrt{\frac{3}{4}\frac{1}{4}\frac{1}{7324}}\right] = [0.740; 0.760]\,.$$

En ayant comme résultat 5674 *pois ronds parmi* 7324 *il observe une fréquence* $f = 0.775$*. Il devra donc truquer ses résultats. En revanche, s'il obtient* 5474 *parmi* 7324 *alors la fréquence est* $f = 0.747$*, et il conservera ses résultats.*

Situation 2 (précision pour un intervalle donné)

Un scientifique soupçonne Mendel d'avoir truqué ses résultats, car il trouve ses résultats "trop beaux pour être vrais". Il a connaissance des lois de l'hérédité, et aimerait savoir quelle était la probabilité avec un échantillon de taille 7324 d'observer une fréquence située dans un aussi petit intervalle que [0.745 ;0.755].

Il faut calculer α *tel que la longueur de l'intervalle* $2u_\alpha\sqrt{\frac{p(1-p)}{n}}$ *soit celle de l'intervalle* $[0.745; 0.755]$*, autrement dit* 0.01*. Puisque* $p = \frac{3}{4}$ *et* $n = 7324$*, on obtient* $u_\alpha = 0.988$*. D'après la table de la loi normale centrée réduite, on sait que* $\alpha \simeq 0.32$*, et donc que*

$$P\left(F_n \in [0.745; 0.755]\right) \simeq 0.68.$$

Situation 3 (taille minimale de l'échantillon pour une précision donnée)

Un étudiant en botanique un peu paresseux, qui sait que $p = \frac{3}{4}$ si l'on croise des lignées pures, aimerait fournir le moins d'effort possible. Il chercher à calculer le nombre minimum de croisements nécessaires pour avoir un intervalle de fluctuation d'amplitude 0.04 et de seuil 0.99.

La longueur de l'intervalle de fluctuation au seuil 0.99 *est* $2u_{0.01}\sqrt{\frac{p(1-p)}{n}}$*, il faut donc que* $n \geqslant 3120.2$*. Puisque* n *est un entier, il faut faire au moins* 3121 *croisements.*

Situation 4 (intervalle de confiance pour une taille d'échantillon donnée)

Un botaniste "inculte", qui ne connaît pas les lois de l'hérédité, a fait 100 croisements de lignées pures et observe 69 pois ronds, aimerait en déduire la proportion théorique p, à un niveau de confiance de 95%.

La fréquence observée est $f = 0.69$*, la taille de l'échantillon est* $n = 100$*, l'intervalle de confiance de niveau* 95% *pour* p *est donc* $[0.590; 0.790]$*.*

Situation 5 (taille de l'échantillon pour une amplitude fixée)

Un scientifique a à sa disposition différentes lignées qui ne sont pas nécessairement pures et ne sait pas quelle est la proportion théorique de pois ronds dans la descendance. Il aimerait connaître le nombre de croisements nécessaires pour obtenir un intervalle de confiance pour p d'amplitude 0.02 et de niveau 95%.

La longueur de l'intervalle de confiance de niveau 95% *est* $\frac{2}{\sqrt{n}}$*. Il faut donc faire au moins 10000 croisements.*