Atelier théorique

Intervalle de confiance *vs.* intervalle de fluctuation
Illustration par 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 implique le calcul des proportions des différents génotypes (outils : arbre pondéré, conditionnement, indépendance), la démonstration de l'existence d'une probabilité stationnaire, et l'utilisation d'une loi binomiale. La seconde partie, consacrée aux statistiques, se base sur les données de Mendel et sur des simulations, et permet d'illustrer sur différents exemples les notions d'intervalle de fluctuation et d'intervalle de confiance.

1 Probabilités

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

On considère une population infinie de pois, dans laquelle le génotype RR est présent en proportion p_n, le génotype Rr en proportion q_n, et le génotype rr en proportion $r_n = 1 - p_n - q_n$.

Après croisement aléatoire des pois de la n-ième génération, quelles seront les proportions $p_{n+1}, q_{n+1}, r_{n+1}$ obtenues à la génération suivante ?

Méthode 1

On effectue le calcul à l'aide d'un arbre pondéré. Les premières branches correspondent aux 3 génotypes possibles pour le parent 1 (de la génération n), chaque embranchement générant à nouveau 3 branches correspondant aux génotypes possibles pour le parent 2 (de la génération n). Les dernières branches correspondent enfin aux génotypes possibles du descendant (de la génération n+1), et prennent en compte le fait que chaque gamète des parents peut être choisi avec probabilité $\frac{1}{2}$.

Méthode 2

Calculons par exemple p_{n+1}. On note $Gen1$ le génotype du premier parent, $Gam1$ l'allèle porté par le gamète du premier parent choisi pour la reproduction etc. Le choix des gamètes chez chacun des deux parents étant indépendant,

$$\begin{aligned} p_{n+1} &= P\left(\{Gam_1 = R\} \cap \{Gam_2 = R\}\right) \\ &= P\left(Gam_1 = R\right) P\left(Gam_2 = R\right) \\ &= P\left(Gam_1 = R\right)^2. \end{aligned}$$

D'après la formule des probabilités totales,

$$\begin{aligned} &P\left(Gam_1 = R\right) \\ &= P_{Gen_1=RR}\left(Gam_1 = R\right) P\left(Gen_1 = RR\right) + P_{Gen_1=Rr}\left(Gam_1 = R\right) P\left(Gen_1 = Rr\right) \\ &\quad + P_{Gen_1=rr}\left(Gam_1 = R\right) P\left(Gen_1 = rr\right) \\ &= P\left(Gen_1 = RR\right) + \frac{1}{2} P\left(Gen_1 = Rr\right) + 0 \\ &= p_n + \frac{1}{2} q_n. \end{aligned}$$

Ainsi

$$p_{n+1} = \big(p_n + \frac{1}{2}q_n\big)^2.$$

De même,

$$q_{n+1} = 2\big(p_n + \frac{1}{2}q_n\big)\big(r_n + \frac{1}{2}q_n\big), \quad r_{n+1} = \big(r_n + \frac{1}{2}q_n\big)^2.$$

Remarquons que $p_{n+1} + q_{n+1} + r_{n+1} = (p_n + q_n + r_n)^2 = 1$.

Question 2

Pour simplifier les notations, on appelle c la proportion initiale d'allèles de type R, c'est-à-dire $c = p_0 + \frac{1}{2}q_0$.

Montrer que pour toutes les générations suivantes, les proportions p_n, q_n, r_n sont constantes.

Puisque $p_0 + q_0 + r_0 = 1$, *on obtient* $r_0 + \frac{1}{2}q_0 = 1 - c$. *Ainsi*

$$\begin{cases} p_1 &= \big(p_0 + \frac{1}{2}q_0\big)^2 = c^2, \\ q_1 &= 2\left(p_0 + \frac{1}{2}q_0\right)\left(r_0 + \frac{1}{2}q_0\right) = 2c\,(1-c)\,, \\ r_1 &= \big(r_0 + \frac{1}{2}q_0\big)^2 = (1-c)^2\,. \end{cases}$$

Soit $n \geqslant 1$. *Supposons que* $p_n = c^2$, $q_n = 2c\,(1-c)$, $r_n = (1-c)^2$. *Alors*

$$p_{n+1} = \big(p_n + \frac{1}{2}q_n\big)^2 = \big(c^2 + c\,(1-c)\,\big)^2 = c^2 \quad \textit{etc.}$$

Question 3

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 obtient ainsi une génération initiale de pois (génération 0), tous ronds, de génotype Rr. Il compta alors le nombre de pois ronds obtenu à la génération 1 après croisement de ces pois ronds.

Supposons que l'on observe n pois parmi les pois de la génération 1 et que l'on note X_n le nombre de pois ronds.

Quelle est la loi de X_n ?

Il s'agit de la répétition de n *expériences identiques et indépendantes, valant* 1 *si le pois est rond et* 0 *s'il est ridé. L'allèle* R *étant dominant, la probabilité pour chacun des* n *pois d'être rond est* $p = p_1 + q_1$, *où* p_1 *et* q_1 *sont les proportions génotypiques à la génération issue des croisements de pois* Rr. *Or pour la génération initiale de pois de génotypes* Rr, *les proportions sont, par hypothèse,* $p_0 = r_0 = 0$ *et* $q_0 = 1$. *Ainsi* $p = \frac{1}{4}q_0^2 + 2\frac{1}{2}q_0\frac{1}{2}q_0 = \frac{3}{4}$, *et* X_n *suit donc la loi binomiale* $\mathcal{B}(n, \frac{3}{4})$.

2 Statistiques

Dans cette expérience, on observe n pois, chaque pois ayant une probabilité p d'être rond. On note X_n le nombre de pois ronds. En généralisant le résultat de la question 3, on sait que X_n suit la loi binomiale $\mathcal{B}(n, p)$. On s'intéresse à la fréquence $F_n = \frac{X_n}{n}$ des pois ronds observés.

2.1 Cas 1 : p est connu

2.1.1 Théorie

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.

Définition 1.

- Un **intervalle de fluctuation** au seuil $1-\alpha$ pour une variable aléatoire Y est un intervalle réel $[a,b]$ qui contient Y avec une probabilité supérieure à $1-\alpha$:

$$P\left(Y \in [a,b]\right) \geqslant 1-\alpha.$$

- 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. 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 2. Dans l'exemple étudié, où $X_n \sim \mathcal{B}(n,p)$, l'intervalle $[0,n]$ (resp. $[0,1]$) est un intervalle évident (au seuil 1) pour la variable aléatoire X_n (resp. 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 3. $u_{0,05} = 1,96$ et $u_{0,01} = 2,58$.

Remarque 4. On pratique cette approximation dès que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Remarque 5. 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 $p(1-p) \leqslant \frac{1}{4}$,

$$\left[p - u_{0,05}\sqrt{\frac{p(1-p)}{n}}; p + u_{0,05}\sqrt{\frac{p(1-p)}{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right].$$

2.1.2 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 5674 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 qu'elle 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 au 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.*

2.2 Cas 2 : p est inconnu

2.2.1 Théorie

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.

Définition 2.
Soit θ un paramètre inconnu de la loi d'une variable aléatoire Y, et soit $(Y_1, \ldots, Y_n)$ 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 $(Y_1, \ldots, Y_n)$, contenant θ avec une probabilité supérieure ou égale à $1-\alpha$:

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

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 $1-\alpha$ 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 6. Le lien avec la Définition 2 se fait en considérant n variables aléatoires indépendantes $(Y_1, \ldots, Y_n)$ 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 $(Y_1, \ldots, Y_n)$.

Remarque 7. 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. On peut donc par exemple dire que l'on "fait confiance" à l'intervalle $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ à 95%.

Il serait en revanche incorrect de conclure "la probabilité que p appartienne à l'intervalle numérique $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ est de 95%". En effet, p est un réel et non pas une variable aléatoire, donc soit appartient à l'intervalle soit ne lui appartient pas : cette probabilité est donc 0 ou 1.

2.2.2 Application

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

Un botaniste "inculte", qui ne connait 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 $[0,59; 0,79]$.

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.

3 Illustration numérique

Utilisation du logiciel Excel et de la fonction

```
SI(ALEA()<p;1;0)
```

pour simuler une variable aléatoire de Bernoulli de paramètre p.

Illustration 1 (intervalle de fluctuation)

On suppose ici pour tous les élèves le paramètre p connu, égal à $\frac{3}{4}$. Chaque élève simule plusieurs (par exemple 10) échantillons de taille $n = 100$ et relève les 10 fréquences f de 1 obtenues. Quelques fréquences f obtenues (de l'ordre de 90%, donc 270 parmi 300 fréquences

pour une classe de 30 élèves) devraient se trouver en dehors de l'intervalle de fluctuation de niveau 90% pour F_{100}, qui est (puisque $u_{0,1} = 1,64$) :

$$\left[\frac{3}{4} - 1,64\sqrt{\frac{3}{4}\frac{1}{4}\frac{1}{100}}; \frac{3}{4} + 1,64\sqrt{\frac{3}{4}\frac{1}{4}\frac{1}{100}}\right] = [0,68; 0,82]\,.$$

Illustration 2 (intervalle de confiance)

On suppose ici le paramètre $p \in [0,05 : 0,95]$ inconnu des élèves, choisi par l'enseignant. La restriction à l'intervalle $[0,05 : 0,95]$ est imposée par la Proposition 2 et ses conditions $np \geqslant 5$, $n(1-p \geqslant 5$, pour un échantillon de taille $n = 100$. Idéalement, le paramètre p est entré sur une case Excel par l'enseignant chez tous les élèves et immédiatement masqué. Les élèves peuvent alors simuler chacun 10 échantillons de taille 100, de loi $\mathcal{B}(p)$ (en utilisant `SI(ALEA()<A1;1;0)` si `A1` est la case masquée contenant p). Ils relèvent chacun les 10 fréquences f de 1 obtenue, et en déduisent 10 réalisations numériques de l'intervalle de confiance de niveau 95% pour le paramètre p inconnu :

$$[f - 0,1; f + 0,1]\,.$$

L'enseignant peut alors révéler la valeur réelle de p. Quelques intervalles obtenus (de l'ordre de 5% parmi les éléves, donc 15 parmi 300 intervalles pour une classe de 30 élèves) devraient ne pas contenir la vraie valeur de p.