# –Texte–
# Loi de Hardy-Weinberg

## 1 La loi de Hardy-Weinberg

En 1908, un mathématicien anglais, G.H. Hardy, et un médecin allemand W. Weinberg ont formulé une loi, connue sous le nom de loi de Hardy-Weinberg, qui concerne les fréquences alléliques pour un gène pouvant s'exprimer sous la forme de deux allèles $A$ et $a$ dans une population diploïde[1] idéale. Nous dirons qu'une population est idéale lorsque

1. La population est de taille infinie.
2. Les individus s'y unissent aléatoirement, impliquant l'union aléatoire des gamètes. Il n'y a donc pas de choix du conjoint en fonction de son génotype. On dit alors que la population est panmictique.
3. Il n'y a pas de migration. Aucune copie allélique n'est apportée de l'extérieur.
4. Il n'y a pas de mutation.
5. Il n'y a pas de sélection.
6. Les générations sont séparées.

**Loi de Hardy-Weinberg.** Dans une population diploïde idéale, les fréquences alléliques d'un gène s'exprimant sous la forme de deux allèles $A$ et $a$ sont constantes au fil des générations. Plus précisément, si $p$ est la fréquence d'apparition de l'allèle $A$ (la fréquence de l'allèle $a$ étant $1-p$) à la génération $n$, alors c'est encore encore le cas à la génération $n+1$.

De plus, si $r$, $s$, $t$ sont les fréquences des génotypes $AA$, $Aa$ et $aa$ à la génération $n$ (avec $r+s+t=1$) alors les fréquences à la génération $n+1$ sont égales à $(r+s/2)^2$, $2(r+s/2)(t+s/2)$ et $(t+s/2)^2$ à la génération $n+1$ et restent inchangées aux générations suivantes.

*Remarque* 1.1. Cette loi est généralisable à un locus[2] avec plusieurs allèles $A_1$, $A_2$,..., $A_k$.

**Conséquences.**

1. Les relations de dominance entre allèles n'ont aucun effet sur l'évolution des fréquences alléliques.
2. La ségrégation mendélienne aléatoire des chromosomes préserve la variabilité génétique des populations.

[1] Diploïde : qui possède un double assortiment de chromosomes semblables.
[2] Locus : localisation précise d'un gène particulier sur un chromosome.

  

3. L'évolution étant définie par un changement des fréquences alléliques, une population diploïde idéale n'évolue pas.
4. Seules les violations des propriétés de la population idéale permettent le processus évolutif.

**Applicabilité de la loi de Hardy-Weinberg.**

Bien que les propriétés d'une population idéale apparaissent un peu surréalistes, la plupart des populations présentent des fréquences génotypiques en équilibre de Hardy-Weinberg pour une grande majorité des locus. Ceci est dû au fait que cet équilibre résulte avant tout de la ségrégation aléatoire des chromosomes qui a lieu à chaque génération.

Par contre, dans les populations naturelles, les fréquences alléliques varient constamment d'une génération à l'autre sous l'influence de divers facteurs (sélection, dérive génétique, etc...). Mais l'équilibre de Hardy-Weinberg est rétabli au début de chaque génération.

L'équilibre est avant tout perturbé si les gamètes ne sont pas produites aléatoirement (meiotic drive), ou si il y a choix du conjoint (consanguinité). Notez que la sélection naturelle n'affecte pas l'équilibre de Hardy-Weinberg parmi les nouveau-nés. Son effet ne devient perceptible que par la suite, au cours du développement.

L'objet de ce texte est de décrire le raisonnement permettant de décider si une population donnée est en équilibre de Hardy-Weinberg et si oui avec quel paramètre. Nous dirons dans la suite que la population est en équilibre de Hardy-Weinberg de paramètre $p \in ]0,1[$ si les trois génotypes $AA$, $Aa$ et $aa$ sont présents avec les fréquences

$$p_1 = p^2, \quad p_2 = 2p(1-p) \quad \text{et} \quad p_3 = (1-p)^2. \tag{1}$$

On effectue un tirage (avec remise) d'un échantillon de taille $n$ dans cette population et on note $(N_1, N_2, N_3)$ les effectifs observés des trois génotypes.

Lors d'une expérience menée sur une population de 100 individus, on a mesuré les données suivantes :

$$n = 100, \quad N_1 = 13, \quad N_2 = 49, \quad N_3 = 38.$$

## 2 L'échantillon est-il en équilibre de paramètre donné ?

La première question que l'on peut se poser est la suivante : est-il raisonnable de dire que la population observée est en équilibre Hardy-Weinberg pour un paramètre $p$ donné ? Pour répondre à cette question de manière statistique, on peut mettre en place un test du $\chi^2$ pour tester l'hypothèse

$H_0$ : « la population est en équilibre de Hardy-Weinberg de paramètre $p$ »,

contre l'hypothèse

$H_1$ : « la population n'est pas en équilibre de Hardy-Weinberg de paramètre $p$ ».

  

**Proposition 2.1.** *À niveau $\alpha \in ]0,1[$ fixé, l'ensemble des réels $p \in ]0,1[$ tels que le test précédent soit valide et conduise à accepter $H_0$ est un intervalle strictement inclus dans $]0,1[$.*

## 3 Estimation du paramètre de Hardy-Weinberg

On suppose à présent que la population considérée est en équilibre de Hardy-Weinberg de paramètre $p$ inconnu et l'on souhaite à estimer « du mieux possible » ce paramètre $p$. Le cadre est le suivant : on dispose d'une réalisation $(N_1, N_2, N_3)$ de loi multinomiale $\mathbb{P}_p$ de paramètres $n$ et $\nu_p = (p^2, 2p(2-p), (1-p)^2)$ sur $\{1,2,3\}$ (les nombres 1, 2 et 3 représentant les génotypes respectifs $AA$, $Aa$ et $aa$) avec $p \in ]0,1[$ inconnu[3].

*Remarque* 3.1 (Très très naïve). Une première approche naïve pourrait être la suivante : la v.a. $N_1$ suit une loi binomiale $B(n,p^2)$, donc $\hat{p} = \sqrt{N_1/n}$ converge presque sûrement vers $p$. La statistique $\hat{p}^2$ est un estimateur sans biais de $p^2$ mais $\hat{p}$ est biaisé pour $p$.

La théorie de l'estimation permet de trouver un estimateur plus performant. Il sera de fait à la fois l'estimateur du maximum de vraisemblance et l'estimateur sans biais de variance minimale. Écrivons la vraisemblance de l'échantillon, i.e. la densité de la v.a. $(N_1, N_2, N_3)$ par rapport à la mesure de comptage sur $\mathbb{N}^3$ :

$$L(p,X) = \frac{n!}{N_1!N_2!N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3} = \frac{n!2^{N_2}}{N_1!N_2!N_3!} p^{2N_1+N_2}(1-p)^{N_2+2N_3}.$$

Un estimateur du maximum de vraisemblance est une statistique $\hat{p}_{MV}$ telle que $L(\hat{p}_{MV}, N_1, N_2, N_3)$ est maximum (on pourra choisir de maximiser la fonction de log-vraisemblance $p \mapsto \log L(p, N_1, N_2, N_3)$).

**Proposition 3.2.** *L'estimateur du maximum de vraisemblance (il est ici unique) s'écrit*

$$\hat{p}_{MV} = \frac{2N_1 + N_2}{2n}.$$

On peut se demander en quoi ce nouvel estimateur est bon, voire optimal. Pour cela, on a recours à la notion d'exhaustivité.

**Définition 3.3.** On dira qu'une statistique $T$ à valeurs dans $\mathbb{R}$ (i.e. une variable aléatoire de la forme $T = \phi(N_1, N_2, N_3)$) est exhaustive pour $p$ si la vraisemblance se factorise de la façon suivante :

$$L(p, N_1, N_2, N_3) = h(N_1, N_2, N_3) g(p, T), \tag{2}$$

où $g$ (resp. $h$) est une application de $]0,1[\times\mathbb{R}$ (resp. $\mathbb{N}^3$) dans $\mathbb{R}_+$.

[3] La loi de $(N_1, N_2, N_3)$ est la loi du nombre de boules dans trois urnes 1, 2 et 3 lorsque l'on a répartit dans ces trois urnes $n$ boules indépendamment selon la loi $\nu_p$.

  

On dit de plus que la statistique $T$ est exhaustive complète pour $p$ si pour toute fonction mesurable bornée $f$,

$$\forall p \in ]0,1[, \quad \mathbb{E}_p(f(T)) = 0 \Longrightarrow \forall p \in ]0,1[, \quad f = 0 \ \mathbb{P}_p \ p.s.$$

**Proposition 3.4.** *La statistique $T = 2N_1 + N_2$ est exhaustive complète pour $p$.*

*Remarque* 3.5. La factorisation (2) assure que la loi de $(N_1, N_2, N_3)$ sachant $T$ ne dépend pas de $p$. L'exhaustivité de $T$ pour $p$ assure ainsi que toute « l'information pour l'estimation de $p$ » contenue dans les observations $(N_1, N_2, N_3)$ est contenue dans la variable aléatoire $T$, *i.e.*. Ainsi, on n'estimera pas mieux $p$ en connaissant $(N_1, N_2, N_3)$ qu'en connaissant seulement $T$.

On peut alors montrer le résultat suivant.

**Théorème 3.6** (Lehmann-Scheffé)**.** *Soit $\psi$ un estimateur sans biais de $p$. Si $\phi$ est une statistique exhaustive complète alors la statistique $\mathbb{E}(\psi|\phi)$ est $(\mathbb{P}_p)_{p\in]0,1[}$-presque sûrement l'unique estimateur sans biais de variance minimum.*

**Corollaire 3.7.** *L'estimateur du maximum de vraisemblance est l'unique estimateur sans biais de variance minimum.*

## 4 La population est-elle en équilibre ?

Pour finir, on souhaite utiliser l'estimateur du maximum de vraisemblance pour décider si oui ou non, il est raisonnable de considérer la population observée comme étant en équilibre de Hardy-Weinberg. Pour cela, on couple l'estimation de $p$ et le test du $\chi^2$. On peut montrer le théorème suivant :

**Théorème 4.1.** *Dans le modèle étudié ($p_1$, $p_2$ et $p_3$ s'expriment en fonction de $p$ d'après (1)), si $D_n$ désigne la statistique du $\chi^2$ :*

$$D_n(p) = \sum_{i=1}^{3} \frac{(N_i - np_i(p))^2}{np_i(p)},$$

*alors,*

$$D_n(\hat{p}_{MV}) = \sum_{i=1}^{3} \frac{(N_i - np_i(\hat{p}_{MV}))^2}{np_i(\hat{p}_{MV})} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(1).$$

*Remarque* 4.2. Le nombre de degrés de liberté dans le théorème 4.1 est égal à 3-1-1, c'est-à-dire le nombre de degrés de liberté d'un test du $\chi^2$ classique sur trois classes moins la « dimension » de $p$.

Ce résultat permet de tester l'hypothèse :

$$H_0 \ : \ \blacksquare \text{ la population est en équilibre de Hardy-Weinberg } \blacksquare,$$

contre l'hypothèse

$$H_1 \ : \ \blacksquare \text{ la population n'est pas en équilibre de Hardy-Weinberg } \blacksquare.$$

  

# 5 Suggestions

1. On pourra démontrer la loi de Hardy-Weinberg et expliquer en particulier le phénomène de stabilisation des fréquences génotypiques.
2. On pourra commenter les conditions d'applicabilité de la loi de Hardy-Weinberg.
3. On pourra démontrer la proposition 2.1 et déterminer l'intervalle en question à l'aide de l'outil informatique pour $\alpha \in \{0.1,\ 0.05,\ 0.01\}$. Que répond le test à la question : la population observée est-elle en équilibre de paramètre $p = 1/2$ ?
4. On pourra comparer, grâce à la simulation, les propriétés des estimateurs de $p$ proposés dans le texte ou de tout autre estimateur envisageable.
5. On pourra démontrer la proposition 3.2.
6. On pourra démontrer la proposition 3.4.
7. On pourra commenter la remarque 3.5.
8. On pourra commenter le corollaire 3.7 en donnant en particulier un intervalle de confiance pour $p$ construit à partir de $\hat{p}_{MV}$.
9. On pourra illustrer par la simulation le théorème 4.1.

 