Cours de biostatistique ⋈ Illustrations dans 

Diversités génétiques et tests d'hypothèses

A.B. Dufour, J.R. Lobry & D. Chessel

31 mars 2008

Approche de la décision en statistique. Tests statistiques et loi de Mendel.

Exergue. *Un papa statisticien, à la psychologie rugueuse, veut apprendre les rudiments à son fils. Il place 99 pièces de 0.1€ et 1 pièce de 2€ sur une table et dit " Prend une pièce au hasard. Si tu l'as prise au hasard tu la gardes, sinon je te donne une claque ". Le petit prend la pièce de 2€ et une claque. " Bon, tu n'as pas tout compris. Je te bande les yeux. Prend une pièce au hasard. Si tu l'as prise au hasard tu la gardes, sinon je te donne une claque ". Le petit, qui se méfie, tire un peu sur le bandeau, voit la pièce de 2€ et en prend une autre. " Très bien, tu vois que quand tu ne vois pas, tu tires au hasard ".*
Comme quoi, il y a plusieurs façons de se faire avoir.

1 Approche de la décision en statistique

1.1 Une petite introduction

Les fondateurs de la théorie des tests d'**hypothèses** sont Jerzy Nyman et Ergon Pearson, fils de Karl. Neyman mentionne cependant qu'à sa connaissance, la première approche d'un test d'hypothèse est due à Laplace en 1812. La construction de la théorie des tests a eu lieu entre 1926 et 1933. Neyman et Pearson présentent leur point de vue du problème comme étant le choix entre deux décisions : accepter ou refuser une hypothèse privilégiée. Bien que cette approche a tendance à disparaître au profit des interprétations sur les intervalles de confiance et les probabilités p, c'est celle que nous présenterons dans ce document.

Notons également que l'idée de la **significativité** d'un test a tout d'abord été proposé par Fisher qui vit dans la valeur p un indice de la mesure de l'écart entre les données et l'hypothèse nulle : plus la valeur de p est petite, plus grand est l'écart. Il défendit $p = 0.05$ comme un seuil standard.

> 'If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05 ...'

Neyman et Pearson se sont opposés à cette approche subjective et ont proposé une approche de décision théorique liée aux résultats d'une expérimentation. Une fois définies un certain nombre de règles de décision, le résultat de l'analyse consiste simplement à accepter ou rejeter l'hypothèse nulle.

1.2 Le paradigme de la décision

Les tests ont pour but de vérifier, à partir de données observées dans un ou plusieurs échantillons, la validité d'hypothèses relatives à une ou plusieurs populations. Pour ce faire, il est nécessaire d'avoir une procédure "objective" de rejet ou d'acceptation d'une hypothèse.

Afin de bien comprendre la nature d'un test, nous allons réaliser une analogie avec le système judiciaire. Un prévenu est présenté devant un tribunal parce qu'il est soupçonné de meurtre. Deux hypothèses peuvent être émises : l'individu est innocent, l'individu est coupable. La première est appelée *hypothèse nulle* et notée H_0, la seconde *hypothèse alternative* et notée H_1. Nous avons d'une part la réalité : l'individu est innocent ou l'individu est coupable. Mais cette réalité est inconnue pour les membres du jury. Nous avons d'autre part, la décision que va prendre ce même jury. La situation devient alors la suivante.

		réalité inconnue	
		H_0	H_1
décision	H_0		erreur II
	H_1	erreur I	

Il y a quatre possibilités : deux sont favorables et deux sont défavorables. Les deux possibilités favorables sont les décisions du jury en accord avec la réalité : déclarer le prévenu innocent alors que dans la réalité, il est effectivement innocent : déclarer l'individu coupable alors que dans la réalité, il est effectivement coupable. Les deux possibilités défavorables sont deux erreurs judiciaires : déclarer l'individu innocent alors qu'il est coupable ; déclarer l'individu coupable alors qu'il est innocent. Si l'hypothèse H_1 est choisie quand l'hypothèse H_0 est vraie, alors on dit qu'une **erreur de première espèce** a été commise. Si l'hypothèse H_0 est choisie quand en fait l'hypothèse H_1 est vraie, alors on dit qu'une **erreur de deuxième espèce a été commise**.

Bien sûr, nous voudrions pouvoir prendre une décision en minimisant les probabilités de commettre des erreurs. Malheureusement, il y a un compromis entre les erreurs de première et de deuxième espèces qui empêche de minimiser en même temps les deux probabilités. La formulation selon Neyman-Pearson accordant une place privilégiée à l'hypothèse nulle H_0 revient à dire que le risque de première espèce est plus flagrant que le risque de deuxième espèce. Cette formulation impose une limite supérieure, tolérance associée à ce risque de première espèce, appelé le **niveau de signification**. Dans la description des résultats d'un test, beaucoup de statisticiens préfèrent la phrase "on a échoué à rejeter l'hypothèse nulle" plutôt que dire "on accepte l'hypothèse nulle". Pourquoi ? parce que choisir l'hypothèse H_0 ne signifie pas que H_0 est correcte mais seulement que le niveau d'évidence contre H_0 n'est pas suffisant pour garantir

son rejet au risque α. Dans notre analogie avec le système judiciaire, le jury rend un verdict de "non culpabilité" et non d'"innocence". L'acquittement ne signifie pas que le prévenu n'a pas commis le crime mais qu'il subsiste un doute raisonnable pour ne pas aller à l'encontre de l'innocence.

1.3 Le test du Chi-Deux

Construire une urne contenant 50 boules vertes, 50 rouges, 50 noires et 50 bleues. Puis réaliser un tirage aléatoire sans remise de 50 boules et compter la répartition des couleurs.

```
urne <- rep(x = c("V", "R", "N", "B"), times = c(50, 50, 50, 50))
echan <- sample(x = urne, size = 50, replace = FALSE)
table(factor(echan, levels = c("V", "R", "N", "B")))
```

```
 V  R  N  B
14 11 13 12
```

Recommencer l'expérience 1000 fois et stocker l'information dans la matrice `tableau`. Pour cela, créer la fonction `comptage()` ci-dessous :

```
comptage <- function(x) {
    echan <- sample(urne, 50, rep = F)
    comptes <- table(factor(echan, levels = c("V", "R", "N", "B")))
}
tableau <- t(apply(as.data.frame(1:1000), 1, comptage))
tableau[1:5, ]
```

```
     V  R  N  B
[1,] 13 14 12 11
[2,] 12 11 12 15
[3,] 16 10 15  9
[4,]  7 10 15 18
[5,] 15 12  8 15
```

Comme dans l'urne, le nombre de boules est le même quelle que soit la couleur, on a une distribution uniforme discrète où la probabilité d'apparition d'une couleur est $\frac{1}{4}$. Dans un échantillon de 50 boules, on attend donc en moyenne 12.5 boules vertes, 12.5 boules rouges, 12.5 boules noires et 12.5 boules bleues. Chaque expérience élémentaire donne un résultat autour de la moyenne. On attend (12.5, 12.5, 12.5, 12.5). On parle d'effectif théorique T_i. Le résultat obtenu dans un échantillon est par exemple (15, 9, 18, 8). On parle d'effectif observé O_i. On peut mesurer l'écart entre un effectif observé et le théorique par :

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - T_i)^2}{T_i}$$

Dans notre exemple, nous avons donc

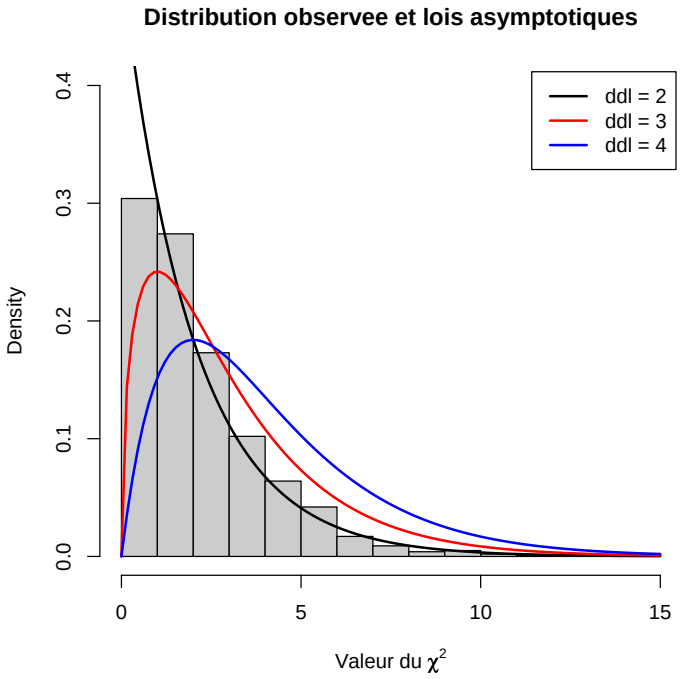
$$\chi^2 = \frac{(15-12.5)^2}{12.5} + \frac{(9-12.5)^2}{12.5} + \frac{(18-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5}$$

soit $\chi^2 = 6.16$.
Calculons les valeurs du chi-deux pour les 1000 échantillons réalisés.

```
result <- apply(tableau, 1, function(x) sum((x - rep(12.5, 4))^2/rep(12.5,
    4)))
```

Mesurer les écarts de cette manière, c'est utiliser la métrique du χ^2. Cette expérience montre que la distribution Chi-deux est une approximation (plus ou moins bonne) de la loi de la statistique χ^2. On a un théorème mathématique dit asymptotique (quand n le nombre de boules tirées tend vers l'infini on tend vers une loi Chi-deux). Regardons cette loi du Chi-deux, densité de probabilité à un paramètre : le degré de liberté (en anglais, degree of freedom *df*).

```
hist(result, nclass = 20, proba = TRUE, col = grey(0.8), main = "Distribution observee et lois asymptotiques",
    ylim = c(0, 0.4), xlab = expression(paste("Valeur du ", chi^2)))
x0 <- seq(0, 15, le = 100)
lines(x0, dchisq(x = x0, df = 2), lwd = 2)
lines(x0, dchisq(x = x0, df = 3), lwd = 2, col = "red")
lines(x0, dchisq(x = x0, df = 4), lwd = 2, col = "blue")
legend("topright", inset = 0.01, legend = paste("ddl =", 2:4), col = c("black",
    "red", "blue"), lwd = 2)
```



Tout ceci nous conduit donc à construire le test d'ajustement du Chi-Deux d'une distribution observée à une distribution théorique. Les hypothèses sont les suivantes :
H_0 La distribution observée suit la distribution théorique.
H_1 La distribution observée ne suit pas la distribution théorique.

`result` représente l'ensemble des valeurs possibles. On l'appelle statistique du test. Le résultat obtenu sur notre échantillon, χ^2 calculé, est une réalisation de la statistique du test. Sous l'hypothèse H_0, la statistique du test suit un Chi-deux à $\nu = k - c$ degrés de liberté où k est le nombre de modalités ou de classes et c la taille de l'échantillon et le nombre de paramètres estimés. Dans l'exemple, nous avons quatre couleurs. Le calcul de la probabilité associé à une classe se

fait sans aucune autre information. Les degrés de liberté sont donc $\nu = 4 - 1$ soit $\nu = 3$.
Lire attentivement la documentation de `chisq.test`.

```
chisq.test(x = c(15, 9, 18, 8), p = rep(0.25, 4))
```

```
        Chi-squared test for given probabilities
data:  c(15, 9, 18, 8)
X-squared = 5.52, df = 3, p-value = 0.1374
```

2 Entrons en génétique

2.1 Un peu d'histoire

FIG. 1 – G. Mendel

Il revient à Gregor Mendel (1822-1884), un moine tchèque, d'avoir établi les bases scientifiques de la génétique. Il a été précédé dans ses recherches par Joseph Gottlieb Koelreuter (1733-1806) et Carl Friedrich von Gaertner (1772-1850). Les combinaisons hybrides ont été formulées par Francis Galton, en 1875, et auparavant par Charles Naudin en 1860. Mais c'est Mendel qui en donne la formulation et la signification fondamentales et définitives. Une question le préoccupait profondément : comment les plantes héritent-elles de leurs traits distinctifs ? Dès 1856, il entreprend de faire pousser différents types de pois ensemble pour voir quels traits sont transmis à leurs successeurs. Les pois de Mendel proviennent généralement de semences de deux formes : rondes et plissées. Dans ses notes, il décrit les semences rondes comme ayant une cellule de type " A " et un pollen de type " A " et les semences plissées comme ayant une cellule de type " a " et un pollen de type " a ". Lorsqu'une cellule de type " A " est fécondée avec un pollen de type " a ", il en résulte une forme ronde. Cela est également vrai d'une cellule de type " a " fécondée avec un pollen de type " A ". La forme ronde est donc considérée comme dominante. La seule façon de produire une semence plissée est de féconder une cellule de type " a " avec un pollen de type " a ". L'article

de Mendel a reçu peu d'attention et son travail est demeuré sur les tablettes pendant presqu'un demi siècle avant d'être " redécouvert " en 1900.

Selon la croyance, Hugo de Vries (en 1907), Carl Correns et Eric Von Tschernak auraient, chacun de leur côté, " découvert " le rapport d'hybridité 3 :1 et ce ne fut qu'après la publication de leurs résultats que les trois ont " re-découvert " le travail antérieur de Mendel sur l'hybridation des plantes. Toutefois, on trouve des indices comme quoi les trois scientifiques auraient lu l'article de Mendel datant de 1865, que ce soit avant ou pendant leurs expériences. Peu importe qui a découvert quoi et quand, la **" loi de Mendel "** était la clé pour comprendre l'hérédité.

En 1908, à l'occasion d'un dîner, on demande à G.H. Hardy, professeur de mathématiques à Cambridge, s'il est possible d'énoncer une formule mathématique décrivant la proportion de gènes allélomorphes dominants à récessifs nécessaires pour permettre l'évolution. On prétend que le professeur aurait écrit la formule sur la manchette de sa chemise. Hardy considère son équation si élémentaire qu'il refuse de la publier dans une revue que pourraient lire ses collègues. Il ne la publie que dans une revue de biologie. Simultanément, mais sans le savoir, Wilheim Weinberg, un médecin allemand, publie la même solution au même problème. Cela donna naissance à **la loi de Hardy-Weinberg**.

2.2 Quelques définitions

Un **gène** est un élément du patrimoine génétique. Il occupe un emplacement bien précis sur un chromosome. Cet emplacement est appelé **locus**. Un gène se présente sous plusieurs états ou **allèles**. Le nombre d'allèles pour un gène varie de deux à plusieurs dizaines. Un organisme possédant un seul gène à chaque locus est dit **haploïde**. Un organisme possédant deux gènes à chaque locus est dit **diploïde**.
Un organisme diploïde ayant un seul allèle en un locus donné est dit **homozygote**. Un organisme diploïde ayant deux allèles différents en un locus donné est dit **hétérozygote**.
Le **génotype** est la constitution génétique d'un organisme à un ou des locus spécifiés. Le **phénotype** est défini par l'ensemble des caractères extérieurs d'un organisme de génotype spécifié.
Les cellules humaines possèdent, en principe, le même potentiel génétique : elles sont diploïdes. Elles possèdent deux fois un nombre de base de chromosomes à savoir 23.

La **fréquence allélique** est la fréquence d'un allèle dans une population. Prenons par exemple un système AB, composé de deux allèles A et B. Un échantillon est composé de :
- 135 individus du groupe [A]
- 244 individus du groupe [AB]
- 111 individus du groupe [B].

On peut s'interroger sur la fréquence de l'allèle A dans cet échantillon.

Phénotypes	[A]	[AB]	[B]
Génotypes	AA	AB	BB
Fréquences génotypiques	n_{AA}	n_{AB}	n_{BB}

La taille de l'échantillon est $n = 490$ soit $2n = 980$ gènes. En effet, par convention, on note n le nombre d'individus. Donc dans le cas d'organismes diploïdes, cela fait $2n$ gènes. La fréquence de l'allèle A dans cet échantillon est :

$$p = \frac{2n_{AA} + n_{AB}}{2n}; p = \frac{514}{980}.$$

p est l'*estimation ponctuelle* de la proportion d'individus possédant l'allèle A dans la population d'où a été extrait l'échantillon.

Nous pouvons également introduire la notion de diversité allélique. C'est la probabilité de tirer au hasard deux allèles différents à un locus dans une population. La probabilité de tirer deux fois le gène A est p^2. Si on note $q = 1 - p$, la probabilité de tirer deux fois le gène B est q^2. La probabilité de tirer deux fois le même gène est $p^2 + q^2$.
Dans le cas d'un gène possédant par exemple K allèles. La probabilité devient $\sum_{k=1}^{K} p_k^2$. La probabilité de tirer deux gènes différents est alors $1 - \sum_{k=1}^{K} p_k^2$.
Dans le cas de deux allèles cette fréquence est $1 - (p^2 + q^2)$ soit $2pq$ (que l'on retrouve facilement en remplaçant q par $1 - p$). La diversité allélique est $2\frac{514}{980}\frac{466}{980} = 0.4988$ C'est la fréquence attendue des hétérozygotes appelée **hétérozygotie**.

2.3 Les lois de Mendel

Les travaux de Mendel ont permis d'établir les lois statistiques de la transmission des caractères héréditaires à partir de lignées pures pour le(s) caractère(s) considéré(s).

Les lois de Mendel reposent sur les principes suivants :

1. Les caractères sont indépendants les uns des autres dans leur transmission ;
2. Mâle et femelle sont équivalents dans l'hybridation, ils déterminent, également malgré les faits de dominance, chaque caractère ;
3. Dans les cellules sexuelles, les deux composantes d'origine mâle et d'origine femelle, de chaque caractère, se dissocient et, dans la fécondation, les composantes de chaque origine s'unissent au hasard, pour chaque caractère.

1. *Loi d'uniformité*
La première génération est homogène (mêmes phénotype et génotype) pour un caractère donné. Si les individus présentent le phénotype de l'un des parents, on dit que le caractère transmis est dominant ; celui de l'autre parent étant récessif. Si les hétérozygotes ont un phénotype intermédiaire, on dit qu'il y a absence de dominance.

2. *Loi de disjonction*
Les deux allèles d'un gène se séparent lors de la formation des gamètes. Chaque

gamète, qui contient l'un ou l'autre des allèles, est pur. Les deux catégories de gamètes sont d'apparition **équiprobable**.

3. *Loi d'indépendance*
Les couples d'allèles suivent le processus indépendamment les uns des autres ; les individus de la première génération produisent des gamètes qui se répartissent en quatre catégories génotypiques équiprobables.

Pour ses expériences, Mendel choisit le pois comestible et s'intéressa à sept caractères dont la couleur (jaune ou verte) et la forme des graines (ronde ou ridée).

Expérience 1. Commençons par étudier un seul caractère : la forme du pois (ronde ou ridée). En croisant des lignées pures de pois ayant des graines rondes avec des lignées pures de pois ayant des graines plissées, il constata que les pois hybrides ainsi créés avaient des graines rondes. Il en conclut que la forme ronde était dominante sur la forme ridée tandis que la forme ridée était un caractère récessif. Cependant, en croisant les hybrides entre eux, il observa de nouveau le caractère récessif,forme ridée, chez une fraction des pois de la descendance.

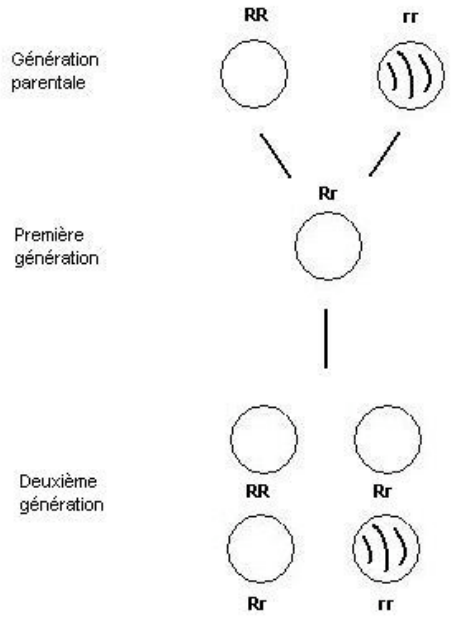


FIG. 2 – Première expérience de Mendel

Expérience 2. En croisant des lignées pures de pois ayant des graines jaunes et rondes (obtenues par autofécondation répétée) avec des lignées pures de pois

ayant des graines vertes et ridées, il constata que les pois hybrides ainsi créées avaient des graines jaunes et rondes. Il en conclut que la couleur jaune devait être dominante sur la couleur verte et la forme ronde dominante sur la forme ridée tandis que la couleur verte et la forme ridée étaient des caractères récessifs. Cependant, en croisant les hybrides entre eux, il observa de nouveau les caractères récessifs, graines de couleur verte et de forme ridée, chez une fraction des pois de la descendance.

Les résultats observés de Mendel sont dans le tableau ci-dessous.

Graines	Jaunes Rondes	Jaunes Ridées	Vertes Rondes	Vertes Ridées
Effectifs observés	315	101	108	32

Le caractère couleur est codé par un gène présentant deux formes allèles C et c, correspondant aux couleurs jaune et vert. Le jaune est dominant, le vert récessif. La forme, rond ou ridé, est portée par un autre gène à deux allèles R (dominant) et r (récessif). Le rond est dominant, le ridé est récessif. Si on croise deux individus dont le génotype est CcRr, on peut obtenir 16 génotypes équiprobables. Les descendants seront jaunes et ronds dans 9 cas sur 16, jaunes et ridés dans 3 cas sur 16, verts et ronds dans 3 cas sur 16, verts et ridés dans 1 cas sur 16. Si Mendel a raison, les effectifs théoriques devraient être pour le croisement des 556 pois.

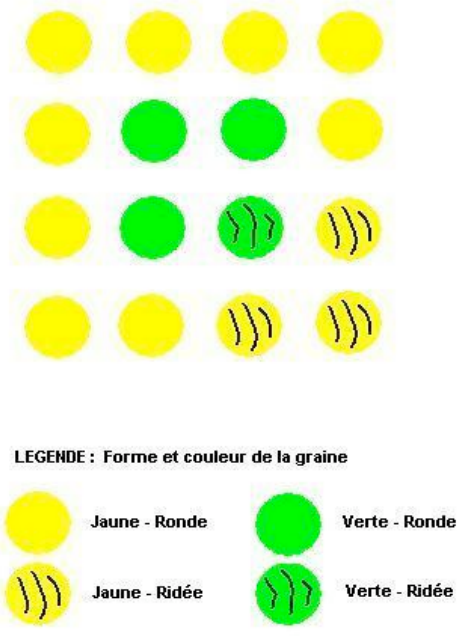


FIG. 3 – Deuxième expérience de Mendel

Graines	Jaunes Rondes	Jaunes Ridées	Vertes Rondes	Vertes Ridées
Effectifs théoriques	312.75	104.25	104.25	34.75

La procédure permettant de confronter les données observées avec les données théoriques conformes à Mendel est le test d'ajustement du Chi-deux que nous avons vu précédemment.

2.4 La loi de Hardy-Weinberg

Selon la loi de Hardy-Weinberg, les fréquences alléliques restent stables de génération en génération dans une population diploïde idéale et ne dépendent que des fréquences de la génération initiale. De plus, les fréquences génotypiques ne dépendent que des fréquences alléliques.
Une population idéale est une population fictive qui possède les propriétés suivantes :

1. Population de grande taille, idéalement 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. Pas de migration. Aucune copie allélique n'est apportée de l'extérieur.
4. Pas de mutation.
5. Pas de sélection.
6. Les générations sont séparées.

Cette population n'est définie ainsi que pour assurer la rigueur mathématique de la démonstration suivante.

Supposons que la taille de la population soit égale à N et considérons un locus à deux allèles A et a possédant respectivement des fréquences p et $q = 1-p$ à la génération t. Quelles vont être les fréquences des différents génotypes AA, Aa, aa à la génération t+1 ?
Pour qu'un individu soit AA, il faut qu'il ait reçu un allèle A de ses deux parents. Cet événement se réalise avec la probabilité $P(AA) = p^2$. Le raisonnement est identique pour le génotype aa soit $P(aa) = q^2$. Enfin, pour le génotype Aa, deux cas sont possibles : $P(Aa) = pq + qp; P(Aa) = 2pq$. Ainsi dans une population idéale, les proportions de Hardy-Weinberg sont données par :

AA	Aa	aa
p^2	$2pq$	q^2

Cette situation est généralisable à un locus avec plusieurs allèles A_1, A_2, ..., A_K. Les fréquences des homozygotes A_kA_k sont égales à p_k^2, celles des hétérozygotes A_iA_j à $2p_ip_j$.

La fréquence p' de l'allèle A à la génération $t+1$ est obtenue par simple comptage :

$$p' = \frac{2p^2N + 2pqN}{2N} \; soit \; p' = p^2 + pq$$

soit respectivement la proportion des allèles A portés par les homozygotes et les hétérozygotes, et donc :

$$p' = p(p+q) \; soit \; p' = p.$$

La fréquence de l'allèle A à la génération $t+1$ est donc identique à celle de la génération précédente et donc aussi à la génération initiale.

En conséquence, les relations de dominance entre allèles n'ont aucun effet sur l'évolution des fréquences alléliques. L'évolution étant définie par un changement de fréquences alléliques, une population diploïde idéale n'évolue pas.

Pour plus d'information, allez sur le site de Laurent Excoffier `http://anthro.unige.ch/evolution/GeneticDrift.html`[1] d'où est extrait en partie ce paragraphe et quelques exercices de la dernière partie.

3 Lien entre tests et génétique

3.1 Exercice Mendel

A partir des données des pois de Mendel, peut-on dire que les effectifs observés diffèrent des effectifs issus du modèle de Mendel $\left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$?

```
effobs <- c(315, 101, 108, 32)
probtheo <- c(9/16, 3/16, 3/16, 1/16)
```

```
        Chi-squared test for given probabilities
data:  effobs
X-squared = 0.47, df = 3, p-value = 0.9254
```

3.2 Exercice balsamines


source : wikipedia

On a effectué le croisement de balsamines blanches avec des balsamines pourpres [1]. En première génération, les fleurs sont toutes pourpres. On obtient en deuxième génération quatre catégories avec les effectifs suivants :

Couleurs	pourpre	rose	blanc lavande	blanc
Effectifs	1790	547	548	213

Peut-on accepter l'hypothèse de répartition mendélienne ?

```
effobs <- c(1790, 547, 548, 213)
probtheo <- c(9/16, 3/16, 3/16, 1/16)
```

```
        Chi-squared test for given probabilities
data:  effobs
X-squared = 7.0628, df = 3, p-value = 0.06992
```

[1] cette adresse est obsolète, à corriger

3.3 Exercice estérase

L'estérase Est 1 du lapin est une protéine monomérique codée par un seul gène à trois allèles E1, E2 et E3. Dans une population, on a trouvé [4] 72 homozygotes E1E1, 24 homozygotes E2E2, 15 homozygotes E3E3, 99 hétérozygotes E1E2, 57 hétérozygotes E1E3 et 33 hétérozygotes E2E3. Cette population suit-elle la loi de Hardy-Weinberg ?

Les effectifs observés sont : 72, 24,15, 99, 57, 33.

```
sum(c(72, 24, 15, 99, 57, 33))
[1] 300
(2 * 72 + 99 + 57)/600
[1] 0.5
(2 * 24 + 99 + 33)/600
[1] 0.3
(2 * 15 + 57 + 33)/600
[1] 0.2
```

Les probabilités sont donc :

$$\begin{array}{r|r} f(E_1E_1) & 0.50 * 0.50 = 0.25 \\ f(E_2E_2) & 0.3 * 0.3 = 0.09 \\ f(E_3E_3) & 0.2 * 0.2 = 0.04 \\ f(E_1E_2) & 2 * (0.5 * 0.3) = 0.30 \\ f(E_1E_3) & 2 * (0.5 * 0.2) = 0.2 \\ f(E_2E_3) & 2 * (0.3 * 0.2) = 0.12 \end{array}$$

```
chisq.test(c(72, 24, 15, 99, 57, 33), p = c(0.25, 0.09, 0.04, 0.3,
    0.2, 0.12))
        Chi-squared test for given probabilities
data:  c(72, 24, 15, 99, 57, 33)
X-squared = 2.5033, df = 5, p-value = 0.776
```

Ce résultat n'est pas correct et c'est important de comprendre pourquoi. On suppose en effet que les fréquences alléliques sont exactement connues, ce qui est faux.
Quand les probabilités théoriques sont inconnues le nombre de degré de liberté est égal aux nombres de classes moins le nombre de paramètres à estimer (ici 2 car la somme des trois fréquences alléliques vaut 1) moins 1. La *p-value* de l'exercice est donc :

```
1 - pchisq(2.503, df = 3)
[1] 0.4747492
```

A propos de ...
Depuis les années 1960, la variabilité des protéines est étudiée par électrophorèse. Les protéines sont des molécules chargées qui se déplacent dans un support poreux (gel d'agarose, d'amidon, de polyacrylamide, d'acétate de celllulose) lorsque celui-ci est soumis à un champ électrique. La vitesse de migration dépend de la charge globale de la protéine, de sa taille et de sa conformation. La mise en évidence de différents allèles d'un même gène est possible pour les enzymes grâce à la spécificité de la réaction enzyme-substrat visualisée par une réaction colorée. Si vous voulez en savoir plus, connectez vous au site de F. Fleury, Rubrique Cours, chapitre 2 : la variabilité génétique dans les populations naturelles, polymorphisme des protéines : `http://gen-net-pop.univ-lyon1.fr/`.

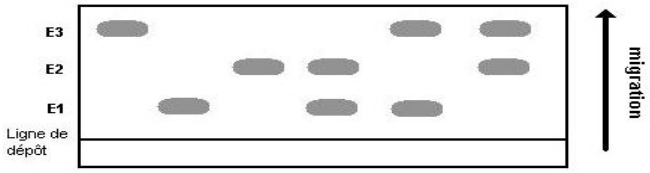


FIG. 4 – Gel de protéine

3.4 Exercice escargot


source : wikipedia

En 1951, le généticien français Maxime Lamotte échantillonne [5] deux formes génétiques de l'escargot des haies *Cepaea nemoralis* (Linnaeus, 1758) dans plusieurs populations d'Europe. Il trouve les proportions suivantes :

Localité	Jaune	Rose	Total
Stockholm, Suède	23	33	56
Dilbeek, Belgique	22	24	46
Niederbronn, France	50	15	65

Peut-on mettre en évidence une différence de formes génétiques de l'escargot entre les populations ?

Il s'agit d'un exemple montrant que le test du Chi-deux peut être utilisé non dans le cas d'un ajustement mais dans le cas d'une table de contingence. Quand on observe ces populations du nord au sud, il semble qu'il n'y ait pas de différence entre Suède et Belgique mais qu'il y ait une grosse différence entre Belgique et France. On réalise un test du Chi-Deux entre les deux échantillons. L'hypothèse posée est : les deux échantillons proviennent d'une même population.

```
table <- matrix(c(22, 50, 24, 15), nrow = 2)
table
```

```
     [,1] [,2]
[1,]   22   24
[2,]   50   15
```

```
chisq.test(table)
```

```
        Pearson's Chi-squared test with Yates' continuity correction
data:  table
X-squared = 8.7708, df = 1, p-value = 0.003061
```

3.5 Exercice mortel

Sur 10 individus issus de croisement de deux hétérozygotes Aa, on a observé 8 phénotype $[a]$ récessif. L'hypothèse que AA est un caractère léthal c'est-à-dire mortel avant la naissance semble-t-elle vérifiée ?

1. Soit S le nombre de sujets ayant le phénotype $[a]$. Donner la loi de S ?
2. Calculer les probabilités associées au croisement des hétérozygotes selon la loi de Mendel.
3. Ecrire les hypothèses H_0 et H_1.

Solution de l'exercice mortel

L'énoncé fait référence à la première expérience.

1. S est la variable "nombre de sujets ayant le phénotype [a]. S est une loi binomiale de paramètres $n = 10$ et p quelconque.
2. Les probabilités associées au croisement des hétérozygotes selon la loi de Mendel sont pour le phénotype [A] $\frac{3}{4}$ et pour le phénotype [a] $\frac{1}{4}$.
3. L'hypothèse H_0 est définie par le caractère léthal de AA. Avec l'équiprobabilité de Mendel, la probabilité du phénotype [a] passe donc à $\frac{1}{3}$. L'hypothèse H_1 consiste à dire que AA n'est pas léthal. Nous sommes alors strictement dans la loi de Mendel et on a $\frac{1}{4}$.

4 Conclure avec Fisher

Fisher [2], toujours le même, en 1918, a fourni une base permettant de réconcilier toute la variation génétique avec les principes Mendeliens. Il a identifié non seulement que toute la variation peut être divisée dans des causes génétiques et non-génétiques, mais que, alors que la corrélation parmi des enfants de mêmes parents est une expression de désaccord génétique (sélectionnable) additif, il y a également un résidu génétique. Ainsi, dans un papier simple, il y a presque 80 ans, Fisher a installé le modèle statistique de la transmission quantitative que nous employons toujours aujourd'hui.

Fisher a conclu que le changement évolutif le plus important est intervenu seulement dans de grandes populations, presque exclusivement par choix sur des lieux proches et indépendants. Dans un autre papier en 1922 [3], il s'est tourné vers l'étude de la variabilité dans les populations normales. A partir des conditions de Hardy-Weinberg, il a étudié les effets de la sélection, de la dominance, des taux de mutation, ...

Sir Ronald Aylmer Fisher a eu une influence considérable en biométrie comme en génétique. Il a beaucoup apporté à la science moderne. Il est considéré comme un des fondateurs de la statistique et des plans expérimentaux.

Références

[1] F. Couty, J. Debord, and D. Fredon. *Probabilités et Statistiques pour biologistes*. U-Flash. Armand Colin, Paris, 1990.

[2] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52 :399–433, 1918.

[3] R.A. Fisher. Darwinian evolution of mutations. *Eugenics Review*, 14 :31–34, 1922.

[4] F. Fleury. *Génétique des populations*. Centre National de la Promotion Rurale, Marmillat, BP 100, 63370 Lempdes, 1997.

[5] M. Lamotte. Recherche sur la structure génétique des populations naturelles de *Cepaea nemoralis*. *L. Bull. Biol. Fr.*, 35 :1–239, 1951.