M2 Écologie,Évolution,Biométrie 
UE Description Statistique des Structures Biologiques

Ségrégation sexuelle et statistique *ad hoc*

D. Chessel

Notes de cours cssb5

Commentaires sur un article de L. Conradt (1998 - Measuring the degree of sexual segregation in group-living animals,*Journal of Animal Ecology* 67:217-226) à partir de trois jeux de données sur le Cerf, le Chamois et le Chevreuil.

Table des matières

1 Introduction

On rend compte ici d'une discussion approfondie dans un groupe de travail animé par A. Loison, J.M. Gaillard et C. Bonnenfant. Ils mettent en commun des jeux de données qui permettent d'aborder un problème statistique par le biais de la pratique. Il s'agit de mesurer le mode de ségrégation des individus des deux sexes dans une structure en groupes.

Au programme des débats figure un article de L. Conradt [2] dont le sommaire est :

> 1. So far, no measure exists to quantify sexual segregation in animal populations. Studies on segregation have relied on ecological measures of overlap and association to estimate the degree of segregation.
> 2. However, existing ecological measures of overlap and association are stochastically related to sex ratio population density or group size. These stochastic relations can lead to confounding results, making the existing measures unsuitable for quantitative studies on segregation.
> 3. In the present paper a new measure of segregation : the 'segregation coefficient', is suggested, which is free of stochastic relations.
> 4. The segregation coefficient is suitable for quantitative studies on segregation. It also makes possible, for the first time, comparisons between the degree of social segregation and the degree of habitat or spatial segregation in a population of animals.

Les arguments biologiques justifiant cette étude sont indiscutables et on s'intéresse ici à la partie statistique. Trois jeux de données sur trois espèces différentes sont mis en commun.

1.1 Données Chamois

Le premier est un extrait aléatoire de données provenant d'un ensemble important proposé par Anne Loison et concerne le Chamois *Rupicapra rupicapra*.

Pour utiliser ces données :

```
library(ade4)
rup <- read.table("http://pbil.univ-lyon1.fr/R/donnees/mfdrupicapra.txt",
    h = T)
dim(rup)
```

```
[1] 265   3
```

```
names(rup)
```

```
[1] "mal" "fem" "mon"
```

```
summary(rup$mon)
```

```
m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
 16  14  20  30  21  19  44  27  19  23  21  11
```

L'objet contient un *data frame* avec 265 lignes (groupe d'individus) et 3 colonnes (variables). Chaque groupe donne un nombre de mâles, un nombre de femelles et le mois d'observation. Au total on a 358 observations de mâles et 1722 observations de femelles.

1.2 Données Chevreuil

Il s'agit d'un extrait systématique de deux années complètes des bases de données de Jean-Michel Gaillard et concerne le Chevreuil *Capreolus capreolus* (Roe Deer). Pour utiliser ces données :

```
cap <- read.table("http://pbil.univ-lyon1.fr/R/donnees/mfdcapreolus.txt",
    h = T)
dim(cap)
```

```
[1] 1214    3
```

```
names(cap)
```

```
[1] "mal" "fem" "mon"
```

```
summary(cap$mon)
```

```
m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
 54   9 116  44 132 209 126 156 187  66  82  33
```

L'objet contient un *data frame* avec 1214 lignes (groupe d'individus) et 3 colonnes (variables). Chaque groupe donne un nombre de mâles, un nombre de femelles et le mois d'observation. Au total on a 741 observations de mâles et 804 observations de femelles.

1.3 Données Cerf

Elles ont été préparées par Christophe Bonenfant avec l'accord de François Klein et concerne le Cerf élaphe ou cerf rouge *Cervus elaphus* (Red Deer). Pour les utiliser :

```
cer <- read.table("http://pbil.univ-lyon1.fr/R/donnees/mfdcervus.txt",
    h = T)
dim(cer)
```

```
[1] 677   3
```

```
names(cer)
```

```
[1] "mal" "fem" "mon"
```

```
summary(cer$mon)
```

```
m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
 64  50  77  59  45  49  43  31  98  60  50  51
```

L'objet contient un *data frame* avec 677 lignes (groupe d'individus) et 3 colonnes (variables). Chaque groupe donne un nombre de mâles, un nombre de femelles et le mois d'observation. Au total on a 544 observations de mâles et 1552 observations de femelles.

1.4 Pour éditer les données

On utilise la fonction `editmfd`, les tableaux étant du type mfd (mâles, femelles, dates) :

```
"splitmfd" <- function(mfd) {
    "loca1" <- function(x) {
        x <- t(x[, 1:2])
        dimnames(x) <- list(c("mal", "fem"), as.character(1:ncol(x)))
        x
    }
    l0 <- split(mfd, mfd$mon)
    l0 <- split(mfd, mfd$mon)
    lapply(l0, loca1)
}
splitmfd(rup)[[2]]
```

```
    1 2 3 4 5  6 7 8 9 10 11 12 13 14
mal 1 1 0 2 0  0 3 1 2  0  1  2  2  4
fem 5 3 5 0 5 10 1 4 0  2  6  1  0  0
```

```
splitmfd(cap)[[12]]
```

```
    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
mal 0 0 1 1 0 0 1 0 0  0  1  2  1  1  1  0  1  0  1  1  0  0  3  1  0  0  1  0  1  2
fem 1 1 3 0 1 1 1 1 1  1  1  2  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1
    31 32 33
mal  0  1  0
fem  1  0  1
```

```
splitmfd(cer)[[7]]
```

```
    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
mal 0 1 1 1 1 1 0 0 0  0  0  2  0  0  0  2  0  0  0  2  2  0  0  1  0  0  1  0  0  1
fem 3 2 2 2 2 2 4 1 1  4  1  0  3  3  1  0  2  2  1  0  2  2  1  0  2  1  2  1  1  1
    31 32 33 34 35 36 37 38 39 40 41 42 43
mal  1  0  0  1  1  1  1  1  1  0  2  3  2
fem  2  2  3  0  2  2  2  2  2  2  2  0  0
```

Dans les trois cas, on a 12 ensembles de groupes d'individus des deux sexes. Le seule contrainte préalable est que le groupe n'est pas vide ! Les 12 échantillons sont mensuels. La question est : peut-on mettre en évidence un effet saisonnier dans le mode de dispersion relatif des mâles et des femelles ?

2 Indices de base

On notera i le numéro d'un groupe, x_i le nombre de mâles dans ce groupe, y_i le nombre de femelles de ce groupe, $n_i = x_i + y_i$ le nombre total d'individus de ce groupe, p_i la proportion de mâles dans ce groupe, q_i la proportion de femelles dans ce groupe, p la proportion totale de mâles, X le nombre total de mâles, Y le nombre total de femelles. L'auteur cite plusieurs indices préexistants à son travail.

1. [4] L'indice de Dice est :

$$A = \frac{\sum_{i=1}^{k} x_i y_i}{p(1-p)\sum_{i=1}^{k} n_i(n_i - 1)}$$

2. [13] L'indice de Schoener est :

$$O = 1 - \frac{1}{2}\sum_{i=1}^{k} |\frac{x_i}{X} - \frac{y_i}{Y}|$$

3. [5] L'indice de Dixon est :

$$S = \log\left(\frac{\sum_i \frac{x_i(x_i-1)}{n_i-1}}{\sum_i \frac{x_i y_i}{n_i-1}} \frac{Y}{X-1}\right)$$

On aurait pu ajouter une quelconque des distances entre deux profils de fréquences, particulièrement celles qui dérivent des comparaisons de niches (classes de proies, classes de tailles, classes de gradient environnemental) et celles qui dérivent des comparaisons de fréquences alléliques. Elles sont nombreuses. Par exemple :

1. Le D2 (*Overlap index*) de Manly [9] est :

$$d_2 = 1 - \frac{\sum_{i=1}^{k} p_i q_i}{\sqrt{\sum p_i^2}\sqrt{\sum q_i^2}}$$

2. La distance de Rogers (un locus) [12] est :

$$d_3 = \sqrt{\frac{1}{2}\sum_{i=1}^{k}(p_i - q_i)^2}$$

3. La distance de Nei 1972 (un locus) [10] est :

$$d_4 = \ln \frac{\sum_{i=1}^{i} p_i q_i}{\sqrt{\sum_{i=1}^{k} p_i^2}\sqrt{\sum_{i=1}^{k} q_i^2}}$$

4. La distance d'Edwards (un locus) [6] est :

$$d_5 = \sqrt{1 - \sum_{i=1}^{k} \sqrt{p_1 q_i}}$$

L'intérêt de la remarque est de noter que l'essentiel de base est dans une table de contingence qui répartit les individus biologiques entre les deux sexes d'une part, entre les groupes d'autre part. Il est équivalent de dire :

> Comment la distribution entre groupes des mâles diffère-t-elle de la distribution entre groupes des femelles ?

ou

> Comment la distribution entre sexes diffère-t-elle d'un groupe à l'autre ?

Ces deux questions sont identiques et dans un cas on compare k distributions à 2 catégories, dans l'autre on compare 2 distributions de fréquences à k catégories. Il est donc troublant de lire *So far, no measure exists to quantify sexual segregation in animal populations* alors que d'entrée, on sait que la solution est dans le χ^2 de cette table de contingence (Pearson, 1900 [11]).

3 Indice ad hoc et χ^2 de contingence

L'auteur explicite ses exigences. Il faut (*no stochastic relation*) que l'indice qui décrit la ségrégation sexuelle, c'est-à-dire l'écart entre la distribution des mâles et celle des femelles –ou la variation du taux de mâles d'un groupe à l'autre– soit indépendant (au sens fonctionnel) du taux de mâle global, du mode de répartition global et de l'effectif global.

Cette exigence est assez ambiguë car elle fait l'impasse sur la partie inférentielle. Elle se veut descriptive. Il est clair que dire *la différence des moyennes de poids des deux échantillons est de 1 kg* n'a pas le même sens si l'échantillon est de taille 2 ou 2000 et si l'écart-type individuel est de 2 grammes ou 20 kg. De même, la ségrégation entre les sexes est de 0.4 n'a pas la même signification si on a observé 5 groupes ou 5000.

Le but est cependant d'introduire la mesure dans un plan d'observation (espace, temps, milieu) pour tester la variation de la ségrégation et on veut concilier la signification de la ségrégation en un point et la variation de cette ségrégation dans un ensemble, comme par exemple dans [3] ou dans [1].

L'indice de base de l'auteur est défini par :

$$SC = 1 - \frac{N}{XY} \sum_{i=1}^{k} \frac{x_i y_i}{n_i - 1}$$

Ceci exclue les groupes de **un seul** individu. L'astuce consiste à justifier la chose en disant qu'un individu isolé est séparé des individus du même sexe et des individus du sexe opposé, donc il ne peut participer à la mesure de la ségrégation entre sexes.

Curieux. Qui dira qu'il n'y a pas de ségrégation sexuelle dans le tableau suivant ?

	1	2	3
mâles	1	1	1
femelles	0	0	25

Oui, mais comment faire un test ? On y vient. Ce qui nous intéresse ici est la notion de statistique *ad hoc*. Sur la base d'un raisonnement sur lequel nous reviendrons, l'auteur pose que la ségrégation sexuelle s'exprime dans :

$$SC = 1 - \frac{N}{XY} \sum_{i=1}^{k} \frac{x_i y_i}{n_i - 1}$$

Ceci exclue les groupes de 1. En utilisant l'idée commune, on dira que la ségrégation sexuelle s'exprime par le χ^2. Or, quand on calcule le χ^2 d'une table de continence à deux lignes, on trouve un résultat très voisin. Comme chacun sait, naïvement :

$$\chi^2 = \sum \frac{(obs - calc)^2}{calc}$$

ce qui donne :

$$\chi^2 = \sum_{i=1}^{i=p} \frac{\left(x_i - \frac{X n_i}{N}\right)^2}{\frac{X n_i}{N}} + \sum_{i=1}^{i=p} \frac{\left(y_i - \frac{Y n_i}{N}\right)^2}{\frac{Y n_i}{N}}$$

où N est le nombre total d'individus et p est le nombre total de groupes. Or :

$$\left(x_i - \frac{Xn_i}{N}\right) + \left(y_i - \frac{Yn_i}{N}\right) = 0$$

Donc :

$$\chi^2 = \sum_{i=1}^{i=p} \left(-x_i + \frac{Xn_i}{N}\right)\left(y_i - \frac{Yn_i}{N}\right)\left(\frac{N}{X} + \frac{N}{Y}\right)\frac{1}{n_i}$$

De plus :

$$\sum_{i=1}^{i=p} \frac{1}{n_i}\left(x_i - \frac{Xn_i}{N}\right)\left(y_i - \frac{Yn_i}{N}\right) = \sum_{i=1}^{i=p} \frac{x_i y_i}{n_i} - \frac{XY}{N} - \frac{XY}{N} + -\frac{XY}{N}$$

D'où :

$$\frac{\chi^2}{N} = 1 - \frac{N}{XY}\sum_{i=1}^{i=p} \frac{x_i y_i}{n_i}$$

4 χ^2 de contingence et inférence

Pour remplacer le calcul du χ^2 par celui de l'indice de Conradt, il faudrait avoir de bonnes raisons. Après tout, si on ne veut pas utiliser les groupes de 1 individu, il n'y a qu'à les exclure. Pour le reste, diviser par $n_i - 1$ au lieu de n_i, ça n'introduit que des faiblesses inutiles. La plus forte touche à l'inférence.

D'un point de vue biologique, on vous dira : vous devez calculer les deux et montrer que le second est meilleur que le premier, ce qui en général n'a pas de sens. Un des calculs est un bricolage, l'autre un standard universel, les deux sont très proches numériquement mais pas dans l'ordre des idées. On le voit très bien en se posant la question : quelle est le signification statistique de l'indice, c'est-à-dire peut-on effectivement parler de ségrégation sexuelle. L'auteur fait l'impasse totale, comme si, pour la première fois, une mesure biologique était dépourvue de variabilité d'échantillonnage.

```
w <- matrix(c(1, 1, 1, 0, 0, 25), nrow = 2, byr = T)
w
```

```
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    0    0   25
```

```
chisq.test(w)
```

```
        Pearson's Chi-squared test

data:  w
X-squared = 17.9487, df = 2, p-value = 0.0001266
```

```
chisq.test(w, sim = T, B = 1e+05)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 1e+05
        replicates)

data:  w
X-squared = 17.9487, df = NA, p-value = 0.00782
```

Le χ^2 est une statistique calculée et c'est aussi une famille de lois de probabilités. Entre les deux, il y a un théorème d'approximation. Quand $N \rightarrow \infty$, sous l'hypothèse nulle, la loi de la statistique tend vers une loi χ^2. Donc, traduction pratique, quand N est assez grand, la statistique suit un χ^2. Quand N est-il assez grand ? Cette question a fait couler beaucoup d'encre et ne se pose plus. Le `sim=T,B=100000` installe la stratégie des tests de randomisation. Noter que la deuxième version considère 28 individus dont 25 sont femelles et 3 sont mâles, dont 26 sont '3', 1 est '2' et 1 est '1'. Avec cette configuration marginale imposée, il y a moins d'une chance sur 100 000 pour qu'il y ait deux mâles isolés. La robustesse du χ^2 est proverbiale. Nous intégrons donc les groupes de 1 et nous disposons d'un test bien connu depuis l'antiquité.

La version non paramétrique du test χ^2 est mis en œuvre dans la fonction `chisq.test`. Pour comprendre cette version, bien moins connue que l'autre, prenons un exemple. Soit trois groupes :

	1	2	3
mâles	4	3	2
femelles	3	2	1

Nous avons 15 individus.

```
w <- matrix(c(4, 3, 2, 3, 2, 1), nrow = 2, byrow = T)
dimnames(w) <- list(c("mal", "fem"), c("g1", "g2", "g3"))
site <- factor(c(rep(c("g1", "g2", "g3"), w[1, ]), rep(c("g1", "g2",
    "g3"), w[2, ])))
sexe <- factor(rep(c("mal", "fem"), rowSums(w)))
sise <- data.frame(site, sexe)
sise01 <- acm.disjonctif(sise)
cbind(sise, sise01)
```

```
   site sexe site.g1 site.g2 site.g3 sexe.fem sexe.mal
1    g1  mal       1       0       0        0        1
2    g1  mal       1       0       0        0        1
3    g1  mal       1       0       0        0        1
4    g1  mal       1       0       0        0        1
5    g2  mal       0       1       0        0        1
6    g2  mal       0       1       0        0        1
7    g2  mal       0       1       0        0        1
8    g3  mal       0       0       1        0        1
9    g3  mal       0       0       1        0        1
10   g1  fem       1       0       0        1        0
11   g1  fem       1       0       0        1        0
12   g1  fem       1       0       0        1        0
13   g2  fem       0       1       0        1        0
14   g2  fem       0       1       0        1        0
15   g3  fem       0       0       1        1        0
```

Dans ce tableau, on trouve exactement l'information de la table de contingence, mais exprimée au niveau des individus sous forme de deux variables qualitatives ou sous forme de deux paquets d'indicatrices des variables. La corrélation entre le facteur sexe et le facteur site est la mesure de la variation du sexe entre groupes. Cette corrélation serait de 1 si il n'y avait que des groupes entièrement de mâles ou entièrement de femelles.

Mais la corrélation entre deux variables qualitatives est la corrélation canonique, c'est-à-dire la première valeur propre de l'Analyse des Correspondances de la table de contingence [14] ou encore la corrélation canonique de l'analyse canonique des deux indicatrices, ou encore le maximum de la corrélation qu'on peut faire par un scoring des modalités qui induit un scoring des individus (dont on attribue la découverte à R.A. Fisher [7]) :

```
library(ade4)
chisq.test(w, sim = T, B = 1)$statistic/sum(w)
```

```
  X-squared
0.005291005
```

```
dudi.coa(as.data.frame(w), scan = F)$eig
```

```
[1] 0.005291005
```

```
cancor(sise01[1:3], sise01[4:5])$cor^2
```

```
[1] 0.005291005
```

Pour tester cette corrélation, il suffit de tester l'appariement aléatoire dans un individu de la modalité sexe et de la modalité site. On permute au hasard une des deux variables et on fait le χ^2 de la nouvelle table de contingence. Les marges sont conservées et on engendre ainsi des tables de contingence identiques pour les marges et basée sur l'indépendance des variables.

L'indice $\frac{\chi^2}{N}$ est un carré de corrélation. Il est toujours compris entre 0 et 1 (c'est faux pour l'indice de Conradt). Il est difficile de faire l'impasse sur tout ça. Il prend en compte les isolés dans une vision globale, ce qui n'est pas sans intérêt biologique.

En effet, comptons les groupes ne comportant qu'un seul individu :

```
nseul <- function(x) {
    x <- as.matrix(x)
    if (nrow(x) == 2)
        x <- t(x)
    mar <- apply(x, 1, sum)
    ng <- length(mar)
    n1 <- ng - sum(mar > 1)
    return(list(ng, n1, round(100 * n1/ng, 0)))
}
w1 <- matrix(unlist(lapply(splitmfd(rup), nseul)), nrow = 3)
dimnames(w1) = list(c("ngroupes", "nisoles", "taux"), as.character(levels(rup$mon)))
w1
```

```
         m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
ngroupes  16  14  20  30  21  19  44  27  19  23  21  11
nisoles    0   0   0   0   0   0   0   0   0   0   0   0
taux       0   0   0   0   0   0   0   0   0   0   0   0
```

```
w1 <- matrix(unlist(lapply(splitmfd(cap), nseul)), nrow = 3)
dimnames(w1) = list(c("ngroupes", "nisoles", "taux"), as.character(levels(cap$mon)))
w1
```

```
         m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
ngroupes  54   9 116  44 132 209 126 156 187  66  82  33
nisoles   28   3  70  34 113 180 106 130 157  53  66  19
taux      52  33  60  77  86  86  84  83  84  80  80  58
```

```
w1 <- matrix(unlist(lapply(splitmfd(cer), nseul)), nrow = 3)
dimnames(w1) = list(c("ngroupes", "nisoles", "taux"), as.character(levels(cer$mon)))
w1
```

```
         m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
ngroupes  64  50  77  59  45  49  43  31  98  60  50  51
nisoles    9  11   9  13  11  26  11   8  29  17  13  17
taux      14  22  12  22  24  53  26  26  30  28  26  33
```

Trois espèces : trois cas de figure. Pour la première, il n'y a jamais d'individus isolés, pour la seconde, pendant l'essentiel de l'année 85% des groupes sont des individus isolés, pour la troisième le taux de groupes de 1 varie entre 15% et 50%. Pour la seconde, l'indice de Conradt oblige à exclure une majorité d'observations : cette limitation est peu acceptable. Mais avec 85% de groupes à un seul individu, il n'est plus légitime de parier sur l'approximation de la loi χ^2 pour faire le test. Il faut accompagner le calcul de sa signification statistique dès qu'on en a besoin.

5 Le rôle de l'environnement

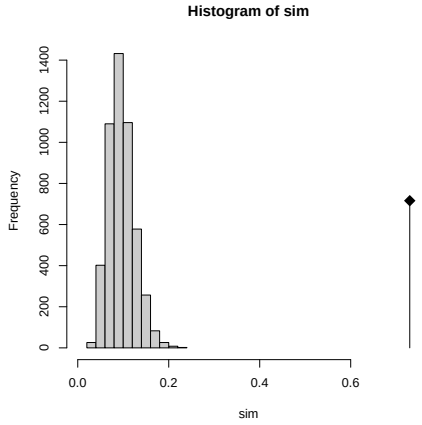
Observons enfin le rôle de l'environnement dans la pratique de la biostatistique. Nous n'avons pas besoin de réécrire la fonction de simulation. Il suffit de profiter de l'extraordinaire ouverture du système . La fonction IK est mise en place.

```
IK.randtest <- function(x, B = 5000) {
    x <- as.matrix(x)
    nr <- nrow(x)
    nc <- ncol(x)
    sr <- rowSums(x)
    sc <- colSums(x)
    n <- sum(x)
    E <- outer(sr, sc, "*")/n
    dimnames(E) <- dimnames(x)
    tmp <- .C("chisqsim", as.integer(nr), as.integer(nc), as.integer(sr),
        as.integer(sc), as.integer(n), as.integer(B), as.double(E),
        integer(nr * nc), double(n + 1), integer(nc), results = double(B),
        PACKAGE = "stats")
    obs <- sum(sort((x - E)^2/E, decreasing = TRUE))/n
    sim <- tmp$results/n
    return(as.randtest(sim, obs))
}
w <- as.data.frame(splitmfd(rup)[[3]])
w
```

```
    1 2 3 4 5 6 7 8  9 10 11 12 13 14 15 16 17 18 19 20
mal 11 1 0 9 1 2 0 2  0  0  1  0 14  3  0 16  0  1  0  0
fem 10 2 2 0 7 1 9 0 18  3  6  5  0  9 16  0 26  7 15  2
```

```
w.test <- IK.randtest(w)
plot(w.test)
quantile(w.test$sim)
```

```
        0%        25%        50%        75%       100%
0.03073241 0.07587654 0.09323938 0.11245025 0.24047411
```



Nous pouvons donc accompagner IK d'un intervalle de confiance estimé par simulation et utiliser la fonction finale :

```
IK <- function(x, conf.int = 0.95, B = 10000) {
    x <- as.matrix(x)
    nr <- nrow(x)
    nc <- ncol(x)
    sr <- rowSums(x)
    sc <- colSums(x)
    n <- sum(x)
    E <- outer(sr, sc, "*")/n
    dimnames(E) <- dimnames(x)
    tmp <- .C("chisqsim", as.integer(nr), as.integer(nc), as.integer(sr),
        as.integer(sc), as.integer(n), as.integer(B), as.double(E),
        integer(nr * nc), double(n + 1), integer(nc), results = double(B),
        PACKAGE = "stats")
    obs <- sum(sort((x - E)^2/E, decreasing = TRUE))/n
    sim <- tmp$results/n
    p0 <- (1 - conf.int)/2
    return(c(obs, quantile(sim, p0), quantile(sim, 1 - p0)))
}
IK(w)
```

```
                2.5%      97.5%
0.72983064 0.04769293 0.15813613
```

6 La signification biologique de IK

6.1 agrégation-ségrégation : un exemple

Quand on observe p groupes dans un échantillon donné, l'agrégation-ségrégation se mesure facilement. Donnons un exemple d'agrégation. Dans 82 pieds de pomme de terre J.M. Legay compte le nombre de doryphores [8] `http://pbil.univ-lyon1.fr/R/articles/arti014.pdf`. Les cas identiques sont regroupés et donne la distribution :

femelles	0	1	2	0	1	2	3	0	1	2	3
mâles	0	0	0	1	1	1	1	2	2	2	2
effectifs	34	13	3	6	9	6	2	1	4	3	1

```
w <- t(read.table("http://pbil.univ-lyon1.fr/R/donnees/doryphore.txt"))
w
```

```
     a  b c d e f g h i j k l
fem  0  1 2 3 0 1 2 3 0 1 2 3
mal  0  0 0 0 1 1 1 1 2 2 2 2
eff 34 13 3 0 6 9 6 2 1 4 3 1
```

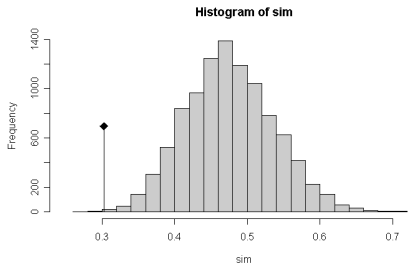
```
w <- w[, -1]
w <- cbind(rep(w[1, ], w[3, ]), rep(w[2, ], w[3, ]))
chisq.test(w)
```

```
        Pearson's Chi-squared test

data:  w
X-squared = 30.2742, df = 47, p-value = 0.9723
```

```
chisq.test(w, sim = T, B = 9999)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 9999
        replicates)

data:  w
X-squared = 30.2742, df = NA, p-value = 0.9995
```

```
plot(IK.randtest(w, B = 9999), nclass = 25)
```

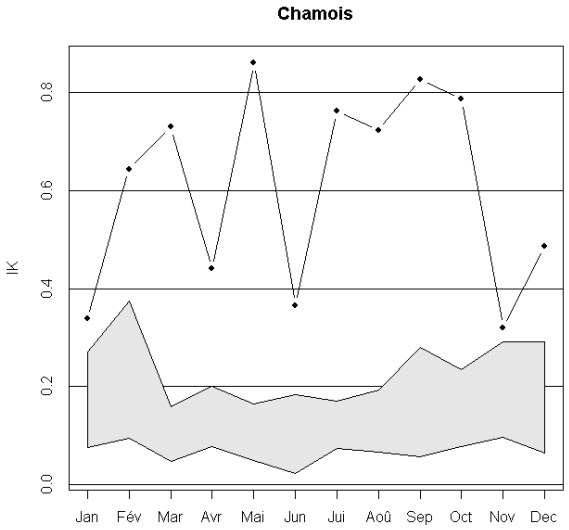


Il y a très significativement association des individus des deux sexes dans cette distribution. Si on peut donc mesurer la ségrégation, on voudra bien sûr mesurer son évolution dans le temps ou sa dépendance des contraintes environnementales. Ces questions sont caractéristiques de la pensée écologique et sont difficiles. En effet, la ségrégation est un indice de structure et non un indice de valeur. La densité, le poids à la naissance, ... dépend-il de x ou y, c'est une question de variations de valeur, le mode de dispersion, d'association, de ségrégation, ... dépend-il de x ou y, c'est une question de variations de structure et c'est beaucoup moins simple. Séparer la ségrégation sociale, intrinsèque au groupe et la ségrégation environnementale, liée à l'habitat du groupe voilà qui est plus sérieux.

Pour l'instant on se contentera de faire un bilan de l'usage de IK sur les trois jeux de données de l'introduction.

6.2 Le cas du Chamois

```
plot1 <- function(w, titre = "") {
    plot(1:12, w[, 1], ylim = range(w), axes = F, pch = 19, type = "n",
        ylab = "IK", xlab = "")
    title(main = titre)
    box()
    axis(1, 1:12, c("Jan", "Fév", "Mar", "Avr", "Mai", "Jun", "Jui",
        "Aoû", "Sep", "Oct", "Nov", "Dec"))
    axis(2, pretty(range(w)), tck = 1)
    polyx <- c(1:12, 12:1)
    polyy <- c(w[, 3], rev(w[, 2]))
    polygon(polyx, polyy, col = grey(0.9))
    points(w[, 1], ylim = range(w), axes = F, pch = 19, type = "b")
}
l1 <- splitmfd(rup)
w <- matrix(unlist(lapply(l1, IK)), ncol = 3, byrow = T)
plot1(w, "Chamois")
```

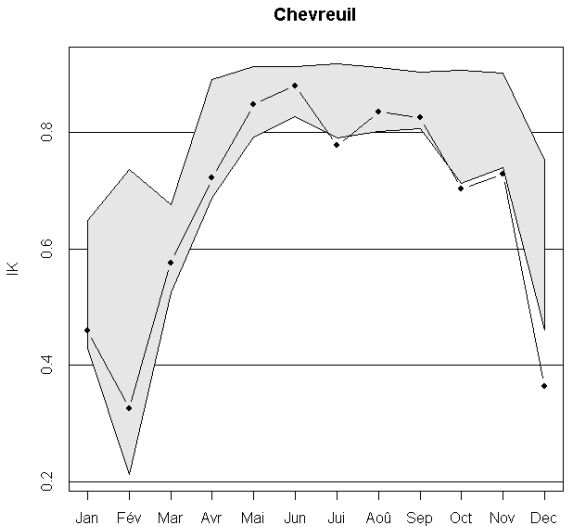


Le résultat est clair. Le point observé est systématiquement très significatif. La ségrégation sexuelle est forte et permanente. C'est le seul jeu de données sur lequel j'ai fait un échantillonnage aléatoire sévère. Ceci peut expliquer la variabilité d'échantillonnage de l'indice. Rien ne s'oppose à l'absence de modification de l'intensité de la ségrégation sexuelle.

6.3 Le cas du Chevreuil

```
l1 <- splitmfd(cap)
w <- matrix(unlist(lapply(l1, IK)), ncol = 3, byrow = T)
plot1(w, "Chevreuil")
```

Le résultat est totalement différent. L'indice et ses bornes de confiance varie fortement, ce qui est le signe d'une modification des modes d'agrégation. Cet effet étant supprimé, la ségrégation sexuelle est systématiquement nulle et très vraisemblablement négative. Le seuil du bas est significatif à 2.5%. Le total est



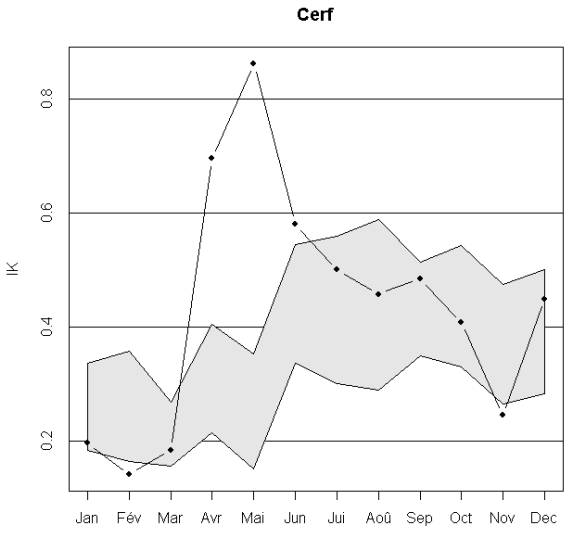
très significatif. Les rares rencontres entre individus privilégient des rencontres entre individus des deux sexes. On peut enfocer le clou si besoin est.

6.4 Le cas du Cerf

```
l1 <- splitmfd(cer)
w <- matrix(unlist(lapply(l1, IK)), ncol = 3, byrow = T)
plot1(w, "Cerf")
```

Le résultat est encore totalement différent. L'effet est très fortement saisonnier et on passe d'un système agrégé en hiver à un système fortement ségrégé au printemps, sans signification statistique en été-automne.

Le khi2 est une bonne mesure statistique : elle rend compte de la diversité biologique.


Cerf
IK
0.2
0.4
0.6
0.8
Jan
Fév
Mar
Avr
Mai
Jun
Jui
Aoû
Sep
Oct
Nov
Dec

Références

[1] C. Bonenfant, L.E. Loe, A. Mysterud, R. Langvatn, N.C. Stenseth, J.-M. Gaillard, and F. Klein. Multiple causes of sexual segregation in european red deer : enlightenments from varying breeding phenology at high and low latitude. *Proceedings of the Royal Society of London. Series B, Biological sciences*, 271 :883–892, 2004.

[2] L. Conradt. Measuring the degree of sexual segregation in group-living animals. *Journal of Animal Ecology*, 67 :217–226, 1998.

[3] L. Conradt. Social segregation is not a consequence of habitat segregation in red deer and feral soay sheep. *Animal Behaviour*, 57 :1151–1157, 1999.

[4] L. Dice. Measures of the amount of ecological association between species. *Ecology*, 15 :297–302, 1945.

[5] P. Dixon. Testing spatial segregation using a nearest neighbour contingency table. *Ethology*, 75 :1940–1948, 1994.

[6] A.W.F. Edwards. Distance between populations on the basis of gene frequencies. *Biometrics*, 27 :873–881, 1971.

[7] J.C. Gower. Fisher's optimal scores and multiple correspondence analysis. *Biometrics*, 46 :947–961, 1990.

[8] J.M. Legay and D. Chessel. Description et analyse de la répartition des insectes dans une population végétale. cas du doryphore sur pomme de terre. *Bulletin d'Ecologie*, 8 :23–34, 1977.

[9] B.F. Manly. *Multivariate Statistical Methods. A primer. Second edition.* Chapman & Hall, London, 1994.

[10] M. Nei. Genetic distances between populations. *The American Naturalist*, 106 :283–292, 1972.

[11] K. Pearson. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50 :157–175, 1900.

[12] J.S. Rogers. Measures of genetic similaritiy and genetic distances. In *Studies in Genetics VII*, volume 7213, pages 145–153. University of Texas Publications, 1972.

[13] T. Schoener. The anolis lizards of bimini : resource partitioning in a complex fauna. *Ecology*, 49(704-726), 1968.

[14] E.J. Williams. Use of scores for the analysis of association in contingency tables. *Biometrika*, 39 :274–289, 1952.