Fiche TD avec le logiciel : tdr32

Tests du Khi2

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

Loi du χ^2 et statistique du khi2. Test d'ajustement à une distribution connue. Test d'ajustement à une distribution inconnue. La question des degrés de liberté. Khi2 d'une table de contingence. Tests exacts.

Table des matières

1 Loi du χ^2 et statistique du Khi2

Les termes "Khi-carré", "Khi2", "Chi-carré", "Chi-squared", "Chi2", "χ^2" ou "variable de Pearson" sont utilisés et sont équivalents. Cette situation induit une confusion assez gênante pour le débutant. On utilisera ici χ^2 pour désigner une loi de probabilité et Khi2 pour parler d'une statistique calculée sur des observations.

La loi de probabilité du χ^2 est la somme de carrés de lois normales centrées réduites indépendantes. Si Z_i désigne une loi normale centrée réduite, on note $\chi^2 = \sum_{i=1}^{p} Z_i^2$. p est le nombre de lois normales centrées réduites sommées et est appelé 'degré de liberté' du χ^2.

La statistique du Khi2 est une mesure de l'écart entre une distribution de probabilité et un tirage observé. La situation typique est celle du tirage de boules de couleurs différentes dans une urne. Le nombre de boules contenu dans l'urne est connu ainsi que la couleur de chaque boule. On construit, par exemple, l'urne de la figure 1 contenant 20 boules vertes, 50 rouges, 30 noires et 100 bleues.

```
couleurs <- c("V","R","N","B")
composition <- c(20,50,30,100)
urne <- rep(x = couleurs, times = composition)
urne
```

```
  [1] "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V" "V"
 [21] "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R"
 [41] "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R"
 [61] "R" "R" "R" "R" "R" "R" "R" "R" "R" "R" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
 [81] "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
[101] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[121] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[141] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[161] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[181] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
```

On tire au hasard 10 boules avec remise et on recommence autant de fois que l'on veut :

```
sample(x = urne, size = 10, replace = TRUE)
```

```
[1] "V" "B" "N" "R" "N" "B" "B" "N" "B" "B"
```

```
sample(urne, 10, T)
```

```
[1] "B" "B" "B" "N" "B" "R" "N" "V" "N" "R"
```

On compte le nombre de boules de chaque couleur en faisant apparaître les couleurs éventuellement absentes et on recommence autant de fois que l'on veut.

```
table(factor(sample(urne, 10, T), levels = couleurs))
```

```
V R N B
0 5 0 5
```

On répéte l'expérience 10 fois.

```
replicate(10, table(factor(sample(urne, 10, T), levels = couleurs)))
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
V    1    0    1    1    2    1    2    0    0     1
R    1    2    1    1    3    1    1    4    3     2
N    1    2    3    2    2    3    1    0    0     2
B    7    6    5    6    3    5    6    6    7     5
```

On transpose pour avoir les expériences en ligne.

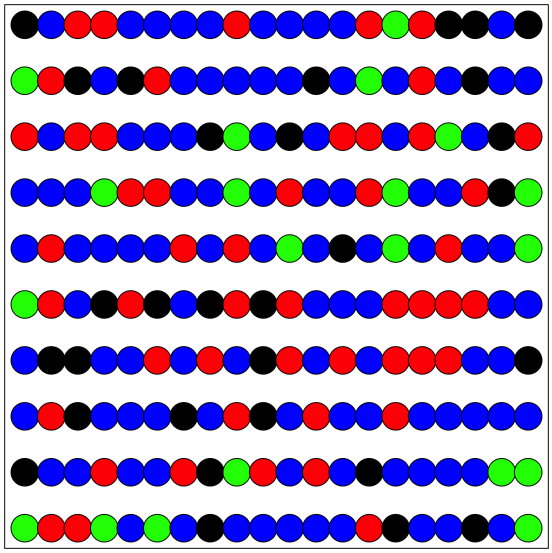

FIGURE 1 : L'urne utilisée pour notre expérience contient 200 boules de quatre couleurs différentes : 20 boules vertes, 50 rouges, 30 noires et 100 bleues.

```
t(replicate(10, table(factor(sample(urne, 10, T), levels = couleurs))))
```

```
      V R N B
 [1,] 0 3 2 5
 [2,] 0 2 4 4
 [3,] 1 5 1 3
 [4,] 2 1 2 5
 [5,] 2 6 0 2
 [6,] 0 2 3 5
 [7,] 3 2 2 3
 [8,] 1 3 1 5
 [9,] 2 1 4 3
[10,] 0 4 1 5
```

On répète l'expérience 1000 fois et on place le résultat dans l'objet `exp`.

```
exp <- t(replicate(1000, table(factor(sample(urne, 10, T), levels = couleurs))))
dim(exp)
```

```
[1] 1000    4
```

On attendait, en moyenne, respectivement 1 verte, 2.5 rouges, 1.5 noires et 5 bleues.

```
attendu <- 10*composition/sum(composition)
names(attendu) <- couleurs
attendu
```

```
  V   R   N   B
1.0 2.5 1.5 5.0
```

On calcule les moyennes par colonne de nos 1000 expériences :

```
colMeans(exp)
```

```
    V     R     N     B
1.027 2.488 1.445 5.040
```

On a eu, en moyenne sur 1000 expériences, respectivement 1.027 vertes, 2.488 rouges, 1.445 noires et 5.04 bleues.

Chaque expérience élémentaire donne un résultat autour de la moyenne. L'attendu (1, 2.5, 1.5, 5) représente l'information théorique (t_i). On peut mesurer l'écart entre un échantillon observé (o_i) et cet attendu à l'aide de la statistique du Khi2 :

$$\text{Khi2} = \sum_{i=1}^{p} \frac{(o_i - t_i)^2}{t_i}$$

Pourquoi introduit-on un dénominateur dans la statistique du Khi2 ?

On prend l'exemple suivant. Je joue à pile ou face 10 fois. Je gagne 1 fois, l'écart est de 4. Je joue à pile ou face 100 fois. Je gagne 46 fois, l'écart est le même. Oui, mais gagner 1 fois sur 10 n'est pas la même chose que gagner 46 fois sur 100. Dans le premier cas, je suis en train de me faire avoir. Diviser le carré de l'écart par les effectifs théoriques, c'est utiliser la statistique du Khi2. Cette statistique est anormalement élevée dans la première situation mais pas dans la deuxième.

théorique	observé	écart	écart au carré	écart au carré/théorique
5	1	4	16	3.2
50	46	4	16	0.32

On définit une fonction pour calculer la statistique du Khi2 :

```
calculkhi2 <- function (observe) sum((observe-attendu)^2/attendu)
```

On calcule la statistique du Khi2 pour la première expérience, pour la deuxième expérience, etc.

```
calculkhi2(exp[1,])
```
```
[1] 4.4
```
```
calculkhi2(exp[2,])
```
```
[1] 5.066667
```

Le premier écart vaut 4.4, le second vaut 5.067, etc.

On calcule ces 1000 valeurs et on donne leur minimum et maximum :

```
resu <- apply(exp,1,calculkhi2)
min(resu)
```
```
[1] 0.2666667
```
```
max(resu)
```
```
[1] 18.86667
```

On donne le résultat de l'expérience la plus proche du modèle :

```
exp[which.min(resu), ]
```
```
V R N B
1 2 2 5
```

On donne le résultat de l'expérience la plus éloignée du modèle :

```
exp[which.max(resu), ]
```
```
V R N B
5 1 2 2
```

La question est : quelle est la distribution de l'écart ?

L'essentiel est dans la figure 2.

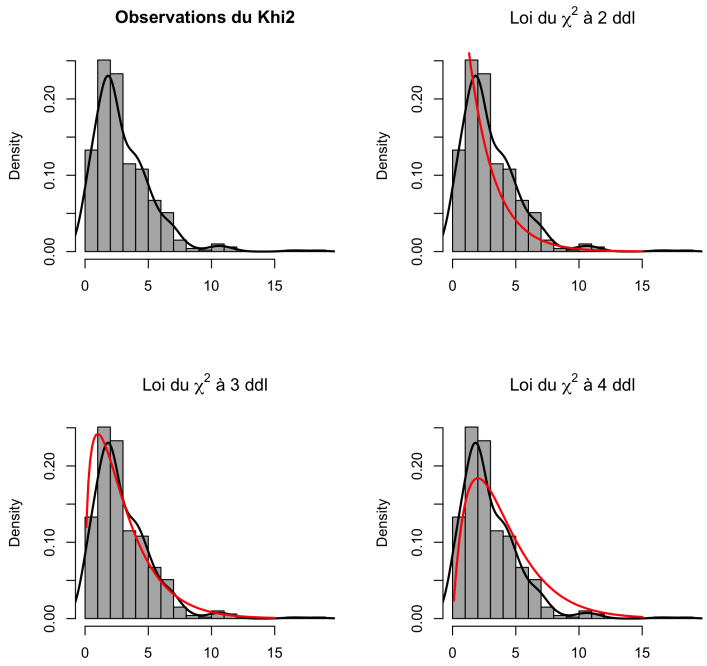

FIGURE 2 : Distribution d'échantillonnage de la statistique du Khi2 dans une expérience répétée 1000 fois de tirages de 10 boules de 4 couleurs dans l'urne de la figure 1. Histogramme (en gris) et densité de probabilité estimée par lissage (en noir). Lois du χ^2 à 2, 3 et 4 degrés de liberté (en rouge).

```
x0<-seq(from = 0.1, to = 15, length = 100)
par(mfrow=c(2,2))
hist(resu,pro=T,nclass=20,main="Observations du Khi2", col = grey(0.7), xlab="")
lines(density(resu,adj=1.5),col="black",lwd=2)
hist(resu,pro=T,nclass=20, main=expression(paste("Loi du ", chi^2," à 2 ddl")),
col = grey(0.7), xlab="")
lines(density(resu,adj=1.5),col="black",lwd=2)
lines(x0,dchisq(x0,2),col="red",lwd=2)
hist(resu,pro=T,nclass=20, main=expression(paste("Loi du ", chi^2," à 3 ddl")),
col = grey(0.7), xlab="")
lines(density(resu,adj=1.5),col="black",lwd=2)
lines(x0,dchisq(x0,3),col="red",lwd=2)
hist(resu,pro=T,nclass=20, main=expression(paste("Loi du ", chi^2," à 4 ddl")),
col = grey(0.7), xlab="")
lines(density(resu,adj=1.5),col="black",lwd=2)
lines(x0,dchisq(x0,4),col="red",lwd=2)
```

Quelle est la loi du χ^2 la plus appropriée pour approximer la distribution d'échantillonnage de la statistique du Khi2 liée à notre expérience ?

Cette expérience illustre que la distribution du χ^2 est une approximation (plus ou moins bonne) de la loi de la statistique du Khi2. On a un théorème mathématique dit *asymptotique* (quand n le nombre de boules tirées tend vers l'infini, on tend vers une loi du χ^2). Mais on s'en sert pour n limité (ici 10). Volontairement,

on est loin des conditions habituellement requises pour accepter l'approximation (5 boules en moyenne de chaque couleur). La **loi** du χ^2 existe en elle-même. La loi de la **statistique** du Khi2, est approximée par une loi du χ^2. Le nombre de degrés de liberté est ici égal au nombre de classes moins un, 4 - 1 = 3, c'est ce que nous avons vérifé expérimentalement dans la figure 2.

2 La question des degrés de liberté

Le test du χ^2 n'est pas né du jour au lendemain. Il a été proposé initialement par Karl Pearson en 1900 [12]. Mais un problème concernant les degrés de liberté a été soulevé par Yule et Greenwood en 1915 [15]. Il faut attendre 1922 pour que Fisher [7] résolve ce problème. Mais il faudra attendre encore longtemps avant que la solution de Fisher ne s'impose. On peut encore trouver en 1947 des statisticiens non encore convaincus (*e.g.* [3]). Davis Baird [2] propose un exemple très simple pour comprendre pourquoi la solution de Fisher a rencontré tant de résistance. On considère deux joueurs J1 et J2 qui tirent des boules rouges et/ou noires dans une urne et on suppose que le résultat est le suivant :

```
obs <- data.frame(list(boule.rouge = c(18,8), boule.noire = c(8,18)))
row.names(obs) <- c("J1","J2")
obs
```

```
   boule.rouge boule.noire
J1          18           8
J2           8          18
```

On veut tester deux hypothèses :

1. Hypothèse A. Tous les couples de résultats (joueur $\times$ couleur) ont même probabilité : (J1,rouge) ; (J1,noir) ; (J2,rouge) ; (J2,noir).
2. Hypothèse I. Il y a indépendance entre l'identité du joueur et la couleur des boules.

Pour reproduire exactement les résultats de Davis Baird, on n'utilisera pas, dans la fonction `chisq.test` de R, l'argument lié à la correction de continuité de Yates[1].
Le test de l'hypothèse A conduit au résultat suivant :

```
chisq.test(unlist(obs), correct = FALSE) -> testA
testA
```

```
        Chi-squared test for given probabilities
data:  unlist(obs)
X-squared = 7.6923, df = 3, p-value = 0.05282
```

La statistique du Khi2 vaut donc ici 7.692 et la p-value est de 0.0528. Donc, avec un risque de première espèce $\alpha = 0.05$, les données ne nous permettent pas de rejeter l'hypothèse A.

Le test de l'hypothèse I conduit au résultat suivant :

```
chisq.test(obs, correct = FALSE) -> testI
testI
```

[1] $\chi^2 = \sum \frac{(|o_i - t_i| - 0.5)^2}{t_i}$ voir `http://pbil.univ-lyon1.fr/R/pdf/qro.pdf`.

```
        Pearsons Chi-squared test
data:  obs
X-squared = 7.6923, df = 1, p-value = 0.005546
```

La statistique du Khi2 n'a pas bougé et vaut toujours 7.692, mais la p-value est maintenant de 0.0055. Donc, avec un risque de première espèce $\alpha = 0.05$, les données sont en contradiction avec l'hypothèse I. On rejette l'hypothèse I.

En résumé, le *même* jeu de données nous conduit à conclure que l'hypothèse A est vraie et que l'hypothèse I est fausse. Le problème ici est que l'hypothèse A implique logiquement l'hypothèse I : s'il est vrai que tous les événements sont équiprobables, alors, automatiquement, on aura aussi l'indépendance entre les deux variables.

Le paradoxe apparent est donc de décider à partir des mêmes observations que l'hypothèse A est vraie et que l'hypothèse I est fausse alors que par ailleurs $A \Rightarrow I$. La raison est que le test du χ^2, comme tous les tests statistiques modernes, tient compte du niveau de généralité de l'hypothèse testée. L'hypothèse A, avec un seul paramètre estimé $((18+8+18+8)/4 = 13)$, est beaucoup plus générale que l'hypothèse I avec 3 paramètres estimés (18, 8, 18) : il est beaucoup plus facile de s'ajuster aux données sous I que sous A, le test du χ^2 en tient compte en pénalisant plus l'hypothèse I. C'est la notion des degrés de liberté.

Pour comprendre l'importance de cette notion, il suffit d'envisager l'hypothèse extrême suivante :
Hypothèse T : les résultats possibles pour les couples de résultats (joueur $\times$ couleur) sont égaux aux valeurs observés.

C'est une hypothèse qui conduira toujours à une statistique du Khi2 nulle puisque l'ajustement sera toujours parfait quelles que soient les observations. Il comporte 4 paramètres, un pour chaque cas, et correspond à un degré de liberté égal à 0. L'ajustement est parfait, mais le modèle est strictement non informatif puisqu'il se contente de répéter le jeu de données.

Le degré de liberté est là pour aider à trouver un compromis entre le degré de généralité d'une hypothèse et la qualité de l'ajustement. Pour les hypothèses courantes A et I, R est capable de déterminer seul le degré de liberté correspondant, *mais ce n'est pas toujours le cas*, comme nous le verrons plus loin.

3 Test d'ajustement à une distribution connue

Lire attentivement la documentation de `chisq.test`.

3.1 Diagnostique[13]

Enoncé. Le rétrécissement des artères et des veines sous claviaires au niveau de l'articulation du bras engendre chez des patients des démangeaisons pouvant nécessiter des interruptions de travail. Le diagnostic du syndrome peut être posé grâce à l'angiographie (c'est-à-dire la radiographie des vaisseaux après injection d'un liquide opaque aux rayons X) effectuée sur des patients en position assise ou

couchée. Pour tester la position la plus efficace, on a relevé la présence (positif) ou l'absence (négatif) de détection de la maladie chez 112 patients.

Position		
assise	*couchée*	Effectif
positif	positif	59
positif	négatif	8
négatif	positif	20
négatif	négatif	25

La position couchée améliore-t-elle la détection de rétrécissement des artères ?

Solution. Les positif-positif et les négatif-négatif n'apportent aucune information. Si les deux positions sont équivalentes, un résultat contradictoire est, en théorie, une fois sur deux positif-négatif et une fois sur deux négatif-positif.

```
 chisq.test(c(8,20),p=c(0.5,0.5))
        Chi-squared test for given probabilities
data:  c(8, 20)
X-squared = 5.1429, df = 1, p-value = 0.02334
```

Un test est un objet :

```
 provi <- chisq.test(c(8,20),p=c(0.5,0.5))
 summary(provi)
          Length Class  Mode
statistic 1      -none- numeric
parameter 1      -none- numeric
p.value   1      -none- numeric
method    1      -none- character
data.name 1      -none- character
observed  2      -none- numeric
expected  2      -none- numeric
residuals 2      -none- numeric
stdres    2      -none- numeric
 class(provi)
[1] "htest"
 names(provi)
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals" "stdres"
```

Éditer et explorer l'objet, essayer d'identifier chacune de ses composantes.

3.2 Balsamine[4]

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

pourpre	rose	blanc lavande	blanc
1790	547	548	213

Peut-on accepter l'hypothèse de répartition mendélienne $\left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$?

Solution.

```
 chisq.test(c(1790,547,548,213),p=c(9/16,3/16,3/16,1/16))
        Chi-squared test for given probabilities
data:  c(1790, 547, 548, 213)
X-squared = 7.0628, df = 3, p-value = 0.06992
```

Discuter le résultat.

4 Test d'ajustement à une distribution inconnue

4.1 Estérase[8]

Enoncé. 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 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 ?

Solution. On a 300 individus soient deux fois plus d'allèles.

- ⋆ La fréquences allélique de E1 est (2*72+99+57)/600 = 0.5
- ⋆ La fréquences allélique de E2 est (2*24+99+33)/600 = 0.3
- ⋆ La fréquences allélique de E3 est (2*15+57+33)/600 = 0.2

Les probabilités estimées des génotypes sont

Génotype	E1E1	E2E2	E3E3	E1E2	E1E3	E2E3
proba	0.25	0.09	0.04	0.30	0.2	0.12

```
chisq.test(c(72, 24, 15, 99, 57, 33),p=c(0.25, 0.09, 0.04, 0.30, 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. Pourquoi ?

On suppose, en effet, que les fréquences alléliques sont exactement connues, *ce qui est faux.*

Supposons que ce soit vrai et faisons une simulation sur la loi de Hardy-Weinberg. La population est infinie, les individus sont diploïdes. Chaque individu observé est le résultat d'un tirage au hasard de deux copies du gène. On fait alors la simulation en tirant au hasard 300 copies, puis en tirant au hasard 300 autres et en les appariant tels quels :

```
w1 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
w2 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
w <- table(paste(w1,w2,sep=""))
w
E1E1 E1E2 E1E3 E2E1 E2E2 E2E3 E3E1 E3E2 E3E3
  78   50   27   51   33   13   27   18    3
```

Il faut encore regrouper, par exemple, les types E1E2 et E2E1 pour avoir une réalisation de l'échantillon puis comparer le résultat au théorique puis refaire cela 1000 fois.

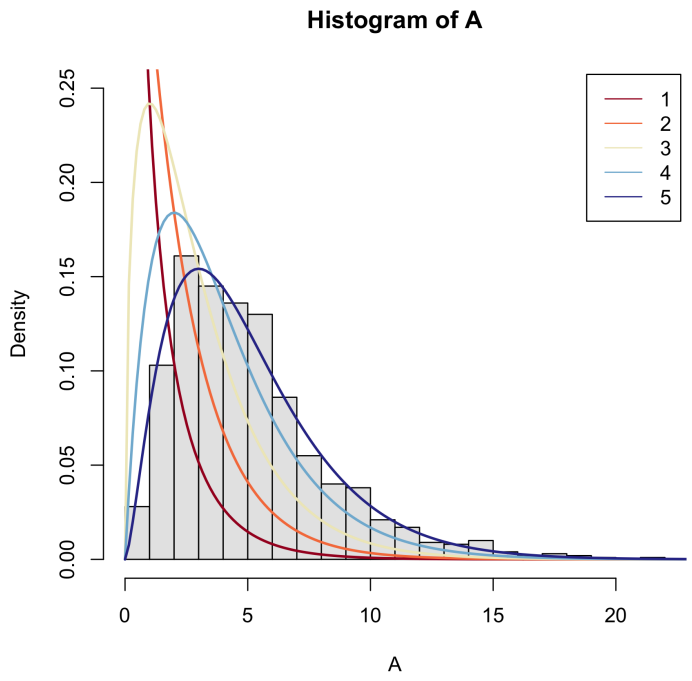

FIGURE 3 : Distribution simulée - proportions connues - et représentation des lois du $\chi 2$ pour les 5 premiers degrés de liberté

```
theo <- 300*c(0.25, 0.09, 0.04, 0.30, 0.2, 0.12)
simulelem <- function (k) {
    w1 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
    w2 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
    w <- table(paste(w1,w2,sep=""))
    if (length(w)!=9) return(NA)
    obs <- c(w[1],w[5],w[9],w[2]+w[4],w[3]+w[7],w[6]+w[8])
    res <- sum(((theo-obs)^2)/theo)
}
library(RColorBrewer)
colddl <- colorRampPalette(brewer.pal(10, "RdYlBu"))(5)
A <- sapply(1:1000, simulelem)
hist(A,proba=T,nclass=30,ylim=c(0,0.25), col=grey(0.9))
x0 <- seq(0,40,le=255)
for (k in 1:5) lines(x0, dchisq(x0,df=k), lwd=2, col=colddl[k])
legend("topright", inset = 0.01, leg = 1:5, lty = 1, col = colddl[1:5])
```

La loi de l'écart est clairement un χ^2 à 5 degrés de liberté.

Discussion autour des degrés de liberté de la loi du χ^2

Quand les probabilités théoriques sont exactement connues, le nombre de degrés de liberté est égal au nombre de classes moins 1. Mais si elles sont estimées, ceci est faux.

En faisant une estimation, on ne trouvera pas, évidemment, les bonnes valeurs $0.5, 0.3, 0.2$. On les estimera et dans chaque expérience, elles seront différentes.

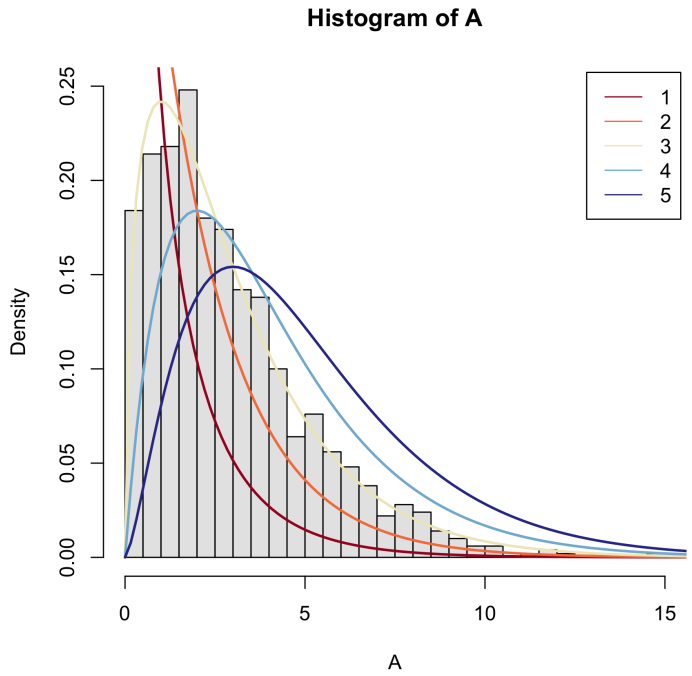

FIGURE 4 : Distribution simulée - proportions estimées - et représentation des lois du $\chi 2$ pour les 5 premiers degrés de liberté

```
simulcomp <- function (k) {
    w1 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
    w2 <- sample(c("E1","E2","E3"),300,p=c(0.5,0.3,0.2),rep=T)
    w <- table(paste(w1,w2,sep=""))
    if (length(w)!=9) return(NA)
    obs <- c(w[1],w[5],w[9],w[2]+w[4],w[3]+w[7],w[6]+w[8])
    fa <- 0
    fa[1] <- (2*obs[1]+obs[4]+obs[5])/600
    fa[2] <- (2*obs[2]+obs[4]+obs[6])/600
    fa[3] <- (2*obs[3]+obs[5]+obs[6])/600
    theo[1] <- fa[1]*fa[1] ; theo[2] <- fa[2]*fa[2] ; theo[3] <- fa[3]*fa[3]
    theo[4] <- 2*fa[1]*fa[2] ; theo[5] <- 2*fa[1]*fa[3]; theo[6] <- 2*fa[2]*fa[3]
    theo <- 300*theo
    res <- sum(((theo-obs)^2)/theo)
    return(res)
}
A <- sapply(1:1000, simulcomp)
hist(A,proba=T,nclass=30,ylim=c(0,0.25), col=grey(0.9))
x0 <- seq(0,40,le=255)
for (k in 1:5) lines(x0, dchisq(x0,df=k), lwd=2, col=colddl[k] )
legend("topright", inset = 0.01, leg = 1:5, lty = 1, col = colddl[1:5])
```

Quand les probabilités théoriques sont inconnues, le nombre de degrés de liberté est égal au nombre de classes moins le nombre de paramètres à estimer moins 1. Dans cet exemple, le nombre de paramètres à estimer vaut 2 car la somme des trois fréquences alléliques vaut 1. La loi est donc un χ^2 à $6 - 2 - 1 = 3$ degrés de liberté.

A retenir :

Au maximum de vraisemblance, l'estimation est plus vraisemblable que la vraie valeur et l'écart au modèle estimé est toujours plus petit que l'écart au vrai modèle (qui lui n'existe que dans les simulations et les hypothèses !)

La *p-value* de l'exercice est donc celle de la statistique du Khi2 :

```
val <- chisq.test(c(72, 24, 15, 99, 57, 33),p=c(0.25, 0.09, 0.04, 0.30, 0.2, 0.12))$statistic
```

pour une loi à 3 `ddl` donc :

```
1-pchisq(val,df=3)
```
```
X-squared
 0.474689
```

L'hypothèse d'Hardy-Weinberg n'est pas rejetée.

Coregonus clupeaformis. Source : Wikipedia.

4.2 Corégones [13]

Enoncé. Dans le cadre d'une étude sur le cycle vital du grand corégone (*Coregonus clupeaformis*) du lac Nathalie, situé dans le territoire de la baie James, P. Dumont [6] a mesuré la longueur totale du corps (en mm) de 756 individus.

Centre de classes	275	285	295	305	315	325	335	345	355	365	375
Effectifs	1	1	4	2	5	2	7	6	11	17	13
Centre de classes	385	395	405	415	425	435	445	455	465	475	485
Effectifs	25	38	72	102	140	107	77	52	34	21	4
Centre de classes	495	505	515	525	535	545	555	565	575	585	595
Effectifs	6	4	2	0	0	1	0	1	0	0	1

Les longueurs totales se distribuent-elles selon une loi normale ?

```
eff <- c(1,1,4,2,5,2,7,6,11,17,13,25,38,72,102,140,107,77,52,34,21,4,6,4,2,0,0,1,0,1,0,0,1)
val <- seq(275,595,by=10)
x <- rep(val,eff)
m0 <- mean(x)
sd0 <- sd(x)
par(mfrow=c(1,2))
plot(val,eff/sum(eff)/10,type="s")
lines(x0 <- seq(275,595,le=50), dnorm(x0,m0,sd0), lwd=2)
qqnorm(x, pch=20) ; qqline(x)
```

Graphiquement, on est très loin de la normalité. *Expliquer pourquoi* `sum(eff)/10`.

Le test d'une telle hypothèse n'a de sens que didactique. On peut tester l'écart entre histogramme et densité par un Khi2. Préparer des bornes, compter les valeurs, calculer les probabilités et la statistique. Compléter par le test de l'écart entre fonctions de répartition.

```
br0 <- c(270,340,380,seq(390,460,by=10),500,600)
obs <- table(cut(x,br0))
theo <- rep(0,12)
theo[1] <- 756*pnorm(340,m0,sd0)
theo[12] <- 756*(1-pnorm(500,m0,sd0))
theo[2:11] <- 756*diff(pnorm(br0[2:12],m0,sd0))
1-pchisq(sum(((obs-theo)^2)/theo),df=9)
```

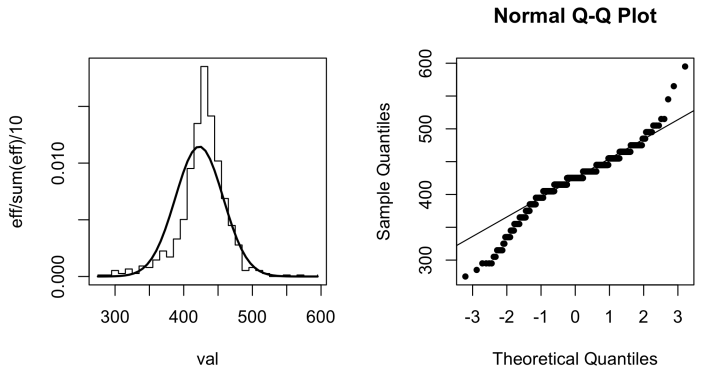

FIGURE 5 : Distribution de la longueur totale du corps (en mm) de 756 corégones et comparaison à la normalité

```
[1] 0
 shapiro.test(x)
        Shapiro-Wilk normality test
data:  x
W = 0.9344, p-value < 2.2e-16
```

Statistiquement, on est aussi très loin de la normalité.

5 Khi2 d'une table de contingence 2-2

5.1 Trois modèles pour un test

Modèle 1

Dans une formation végétale homogène, on choisit 38 points au hasard. En chaque point, on écoute pendant un temps fixé les oiseaux présents. On repère la pie *Pica pica* 17 fois et la corneille *Corvus corone* 19 fois. Sachant que les deux espèces ont été enregistrées simultanément 11 fois, peut-on affirmer que la présence d'une espèce influence la présence de l'autre ?

		Corneille		
		Absence	Présence	Total
Pie	Absence	13	8	21
	Présence	6	11	17
	Total	19	19	38

Quand on fixe une valeur (11) et si les marges sont fixées, toutes les autres valeurs sont définies. On peut dire que concrètement : il y a un degré de liberté.

	0	1	
0	Z_{00}	Z_{01}	$n - n_1$
1	Z_{10}	Z_{11}	n_1
	$n - n_2$	n_2	n

On a deux variables de Bernoulli indépendantes et la loi du couple. Sachant que $Z_{01} + Z_{11} = n_2$ et $Z_{10} + Z_{11} = n_1$, on démontre que :

$$P\left(Z_{11} = j\right) = \frac{\left(\begin{array}{c} n \\ j, n_1 - j, n_2 - j, n - n_1 - n_2 + j \end{array} \right)}{\left(\begin{array}{c} n \\ n_1 \end{array} \right) \left(\begin{array}{c} n \\ n_2 \end{array} \right)}$$

Modèle 2

Sur le même versant d'une montagne, on considère une station haute et une station basse. Dans chacune des stations, on choisit au hasard 32 pieds de faux alpha *Lygeum spartum.* Ces plantes sont les refuges hivernaux de la Punaise des céréales *Blissus leucopterus.* Dans chaque pied, on note la présence ou l'absence de l'insecte dans la plante. En haut, 13 pieds sur 32 sont occupés. En bas, 3 pieds sur 32 sont occupés.

		Insecte		
		Absence	Présence	Total
Station	Haute	19	13	32
	Basse	29	3	32
	Total	48	16	64

La probabilité pour qu'une plante soit occupée est-elle une fonction de l'altitude ?

	0	1	
A	A_0	A_1	n_A
B	B_0	B_1	n_B
	$A_0 + B_0$	$A_1 + B_1$	$n_A + n_B$

Quand on fixe la valeur du nombre total de pieds occupés (16), toutes les autres valeurs sont définies. On peut dire que concrètement, il y a un degré de liberté. On a deux variables hypergéométriques indépendantes et la loi de l'une sachant la somme des deux. On démontre que :

$$P\left(A_1 = j / A_1 + B_1 = nk\right) = \frac{\left(\begin{array}{c} n_A \\ j \end{array} \right) \left(\begin{array}{c} n_B \\ k - j \end{array} \right)}{\left(\begin{array}{c} n_A + n_B \\ k \end{array} \right)}$$

Modèle 3

Sur les 76 pieds d'une rangée de plants de pomme de terre, on a noté la présence de doryphores *Leptinotarsa decemlineata* des deux sexes.

		Mâle		
		Absence	Présence	Total
Femelle	Absence	49	9	58
	Présence	7	11	18
	Total	56	20	76

Soit n le nombre de plantes, p d'entre elles contiennent une marque blanche et q d'entre elles contiennent une marque noire.

	0	1	
0			
1		R	p
		q	n

On note :

$$(x)_k = x\,(x-1)\cdots(x-k+1)$$

Si on énumère toutes les configurations possibles supposées équiprobables, la variable aléatoire R, nombre de plantes contenant les deux marques, a pour loi :

$$P\,(R=j;p,q,n) = \frac{(p)_j\,(q)_j}{j!}\,\frac{(n-p)!\,(n-q)!}{n!\,(n-p-q+j)!}$$

Conclusion.

Ces trois modèles se résument en un seul par :

$$P\left(Configuration\begin{array}{|c|c|}\hline A & B\\ \hline C & D\\ \hline\end{array}\right) = \frac{(A+B)!\,(A+C)!\,(B+D)!\,(C+D)!}{A!B!C!D!\,(A+B+C+D)!}$$

Ils permettent tous de faire le test du Khi2 *pour des raisons différentes*. En conclusion, un test peut correspondre à plusieurs modèles.

Solutions des trois modèles *via* le test du Khi2

- **Modèle 1**

```
chisq.test(matrix(c(13,6,8,11),nc=2))
```

```
        Pearsons Chi-squared test with Yates continuity correction
data:  matrix(c(13, 6, 8, 11), nc = 2)
X-squared = 1.7031, df = 1, p-value = 0.1919
```

On ne peut pas dire que les deux espèces d'oiseaux sont associées ou séparées plus souvent que le veut le hasard.

- **Modèle 2**

```
chisq.test(matrix(c(19,29,13,3),nc=2))
```

```
        Pearsons Chi-squared test with Yates continuity correction
data:  matrix(c(19, 29, 13, 3), nc = 2)
X-squared = 6.75, df = 1, p-value = 0.009375
```

La proportion de plantes occupées est significativement différente entre les deux stations.

- **Modèle 3**

```
chisq.test(matrix(c(49,7,11,18),nc=2))
```

```
        Pearsons Chi-squared test with Yates continuity correction
data:  matrix(c(49, 7, 11, 18), nc = 2)
X-squared = 20.2872, df = 1, p-value = 6.665e-06
```

Les plantes contenant des individus des deux sexes sont significativement plus nombreuses que dans le modèle aléatoire.

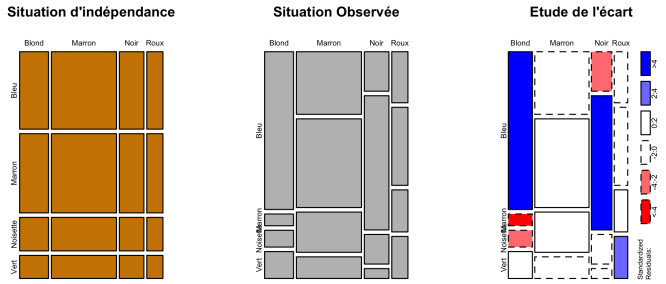

FIGURE 6 : Représentations en forme de mosaïque

5.2 Une représentation graphique de la table de contingence

Une fois admis que la statistique du Khi-Deux permet de rassembler des situations biologiques de nature très différente, on se pose la question de la représentation des données. La méthode commune est le graphe en mosaïque proposée par M. Friendly.
On étudie la relation entre la couleur des yeux et la couleur des cheveux de 592 étudiants en statistique [14].

```
couleur <- read.table("http://pbil.univ-lyon1.fr/R/donnees/snee74.txt", header=TRUE)
reschevyeux <- table(couleur$cheveux,couleur$yeux)
```

Pour une table de contingence donnée, les fréquences attendues sous l'indépendance peuvent être représentées par des rectangles dont la largeur est proportionnelle à la fréquence totale pour chaque colonne $n_{.j}$ et la hauteur proportionnelle à la fréquence totale pour chaque ligne $n_{i.}$.

```
apply(reschevyeux,2,sum)
```

```
   Bleu   Marron Noisette     Vert
    215      220       93       64
```

```
apply(reschevyeux,1,sum)
```

```
Blond Marron   Noir   Roux
  127    286    108     71
```

```
reschi2 <- chisq.test(reschevyeux)
reschi2$expected
```

```
            Bleu    Marron Noisette      Vert
Blond   46.12331  47.19595 19.95101 13.729730
Marron 103.86824 106.28378 44.92905 30.918919
Noir    39.22297  40.13514 16.96622 11.675676
Roux    25.78547  26.38514 11.15372  7.675676
```

```
par(mfrow=c(1,3))
mosaicplot(reschi2$expected, main="Situation dindépendance", col="orange3")
mosaicplot(reschevyeux, main="Situation Observée")
mosaicplot(reschevyeux, shade=TRUE, main="Etude de lécart")
```

S'il y avait indépendance entre les deux variables, les distributions conditionnelles seraient égales aux distributions marginales. Les mosaïques seraient toutes

alignées (figure 6, mosaïque de gauche, en orange).

Une autre représentation se construit autour des résidus standardisés :

$$d_{ij} = \frac{n_{ij} - \frac{n_{i.} n_{.j}}{n}}{\sqrt{\frac{n_{i.} n_{.j}}{n}}}$$

Deux couleurs sont superposées sur la mosaïque (figure 6, mosaïque de droite).

- ⋆ Le bleu signifie une sur-représentativité des effectifs observés par rapport aux effectifs théoriques.
- ⋆ Le rouge signifie une sous-représentativité des effectifs observés par rapport aux effectifs théoriques.

5.3 Le test exact de Fisher

Trouver dans la documentation de `fisher.test` cet admirable exemple détaillé p. 40 dans [1] :

```
Agresti (1990), p. 61f, Fishers Tea Drinker
A British woman claimed to be able to distinguish whether milk or
tea was added to the cup first. To test, she was given 8 cups of
tea, in four of which milk was added first. The null hypothesis
is that there is no association between the true order of pouring
and the womens guess, the alternative that there is a positive
association (that the odds ratio is greater than 1).
```

```
TeaTasting <- matrix(c(3, 1, 1, 3), nr = 2,
                     dimnames = list(Guess = c("Milk", "Tea"), Truth = c("Milk", "Tea")))
TeaTasting
```

```
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3
```

```
chisq.test(TeaTasting)
```

```
        Pearsons Chi-squared test with Yates continuity correction
data:  TeaTasting
X-squared = 0.5, df = 1, p-value = 0.4795
```

Suite à la réalisation du test, le message d'avis attire l'attention. En effet, sachant qu'il y a 8 tasses de thé et 4 dans chacune des catégories, il n'y a que 5 résultats possibles à cette expérience (0, 1, 2, 3 4). La dame doit choisir 4 tasses sur les 8 dans lesquelles elle pense que le lait a été mis en premier. Il y 4 tasses sur les 8 dans lesquelles le lait a été effectivement mis en premier. Elle peut se tromper complètement et ne trouver aucune bonne tasse, elle peut en trouver 1, 2, 3 (ce qui s'est passé) ou 4. La loi du nombre de bons choix sous l'hypothèse nulle (elle tire au hasard !) est hypergéométrique (tirage de 4 boules dans une urne qui contient 4 blanches et 4 noires).

```
dhyper(0:4,4,4,4)
```

```
[1] 0.01428571 0.22857143 0.51428571 0.22857143 0.01428571
```

```
phyper(0:4,4,4,4)
```

```
[1] 0.01428571 0.24285714 0.75714286 0.98571429 1.00000000
```

```
1-phyper(0:4,4,4,4)
```

```
[1] 0.98571429 0.75714286 0.24285714 0.01428571 0.00000000
```

La probabilité d'avoir un résultat au moins aussi bon que le résultat observé est la *p-value* du test de l'hypothèse nulle contre l'alternative "elle peut reconnaître si le lait est placé en premier". Le Khi2 ne peut donc prendre qu'un nombre très limité de valeurs, en fait 3 seulement :

```
 chisq.test(matrix(c(4,0,0,4),nc=2))$statistic
X-squared
      4.5
```

Pour retrouver les valeurs attendues du calcul manuel, enlever la correction de continuité :

```
 chisq.test(matrix(c(4,0,0,4),nc=2),correct=F)$statistic
X-squared
        8
 chisq.test(matrix(c(3,1,1,3),nc=2),correct=F)$statistic
X-squared
        2
 chisq.test(matrix(c(2,2,2,2),nc=2),correct=F)$statistic
X-squared
        0
 chisq.test(matrix(c(1,3,3,1),nc=2),correct=F)$statistic
X-squared
        2
 chisq.test(matrix(c(0,4,4,0),nc=2),correct=F)$statistic
X-squared
        8
 chisq.test(matrix(c(3,1,1,3),nc=2),cor=F)

        Pearsons Chi-squared test

data:  matrix(c(3, 1, 1, 3), nc = 2)
X-squared = 2, df = 1, p-value = 0.1573

 1-pchisq(2,1)
[1] 0.1572992
```

Quelques remarques s'imposent.

1. Le test Khi2 s'occupe de l'écart en plus ou en moins autour du théorique (ici 2 tasses repérées). La vraie *p-value* est :

$$0.01429+0.22857+0.22857+0.01429 = 0.4857$$

   L'approximation par le Khi2 est donc complètement fausse. Le problème est corrigé par la correction de continuité de Yates qui est bien meilleure :

```
 chisq.test(matrix(c(3,1,1,3),nc=2),cor=T)

        Pearsons Chi-squared test with Yates continuity correction

data:  matrix(c(3, 1, 1, 3), nc = 2)
X-squared = 0.5, df = 1, p-value = 0.4795
```

2. On peut vouloir faire un test unilatéral ce que ne peut faire le test Khi2. On utilise alors le test exact :

```
 fisher.test(TeaTasting, alternative = "greater")

        Fishers Exact Test for Count Data

data:  TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

La *p-value* = 0.2429 vient de 0.22857 + 0.01429.

3. Si on n'a pas besoin d'un test unilatéral, le test exact donnera le seuil exact du test Khi2 :

```
fisher.test(TeaTasting)
```

```
        Fishers Exact Test for Count Data
data:  TeaTasting
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

La *p-value* vient de 0.01429 + 0.22857 + 0.22857 + 0.01429.

Etudier les résultats du test de Fisher dans les trois modèles développés précédemment :

```
fisher.test(matrix(c(13,6,8,11),nc=2))
fisher.test(matrix(c(19,29,13,3),nc=2))
fisher.test(matrix(c(49,7,11,18),nc=2))
```

5.4 Le odds ratio

On voit apparaître le terme `odds ratio`. Que signifie-t-il ?

	0	1	
A	A_0	A_1	n_A
B	B_0	B_1	n_B
	$A_0 + B_0$	$A_1 + B_1$	$n_A + n_B$

	0	1	
0	n_{00}	n_{01}	$n - n_1$
1	n_{10}	n_{11}	n_1
	$n - n_2$	n_2	n

On reprend le cas de deux échantillons A et B. 0 est l'échec (perdu). 1 est le succès (gagné). Prenons un seul échantillon. La probabilité de gagner est p. La probabilité de perdre est $1 - p$. En terme de course de chevaux, si un cheval a une probabilité de gagner de 1 chance sur 4 (et une probabilité de perdre de 3 chances sur 4) sa cote est de 3 contre 1. La cote (*odds*) est :

$$odds = \frac{p}{1-p} \Leftrightarrow p = \frac{odds}{odds + 1}$$

Quand on a deux échantillons, le rapport des cotes (*odds ratio*) est :

$$\theta = \frac{odds_A}{odds_B} = \frac{p_A/1 - p_A}{p_B/1 - p_B} \Rightarrow \hat{\theta} = \frac{A_0 B_1}{A_1 B_0} = \frac{n_{00} n_{11}}{n_{01} n_{10}}$$

Quand les p sont petits, les 1 - p sont voisins de 1 et le rapport est voisin du rapport des probabilités qu'on appelle *le risque relatif*. La fonction `fisher.test` ne donne pas le résultat de ce calcul simple mais une estimation non biaisée de θ plus raffinée. Le concept de odds ratio est fondamental en épidémiologie et essais cliniques.

6 Exercices

6.1 Guérison

Sur deux groupes de malades, on essaie deux thérapies. Dans le premier groupe, constitué de 200 individus, on observe 72% de guérisons et dans le second groupe, constitué de 100 individus, on observe 88% de guérisons. Qu'en concluez-vous ?

6.2 Anesthésie[9]

Existe-t-il une relation entre le type d'anesthésie (halothane ou morphine) et la mortalité à la suite d'une opération à cœur ouvert ?

Anesthésie	*Vivants*	*Décédés*
Halothane	53	8
Morphine	57	10

6.3 Vaccin

Pour trois modes de préparation d'un vaccin (A,B,C) on mesure la réaction locale au point d'injection chez 500 patients. On obtient le tableau ci-dessous. Conclusion ?

Réaction	Légère	Moyenne	Ulcération	Abcès
Vaccin A	13	158	8	1
Vaccin B	30	133	6	1
Vaccin C	9	129	10	2

6.4 Bruit

On fait passer un test de réactivité visuelle à un groupe de 118 sujets. Chaque sujet passe le test dans deux conditions différentes : avec ou sans bruit dans la salle. Le test se présente comme suit. Deux lampes sont placées à droite et à gauche du sujet. Chaque lampe s'allume de façon aléatoire, avec un temps d'attente variable (entre 0.2s et 0.5s). Le sujet est assis les mains sur les genoux. Dès qu'une lampe s'allume, il doit frapper une plaque située en dessous de la lampe correspondante. On considère qu'un sujet a réussi le test lorsqu'il a réalisé la bonne association " lumière, frappe " au moins 7 fois sur 10.

		Avec Bruit	
		Succès	Échecs
Sans Bruit	Succès	62	26
	Échecs	7	23

Quelle conclusion tirez vous de cette expérience ?

6.5 Latéralité

On étudie la latéralité chez 56 étudiants de l'UFR STAPS. Le caractère gaucher ou droitier est défini classiquement selon la main d'écriture. Les étudiants sont divisés en deux groupes : les joueurs de tennis et les autres. Les données sont consignées dans le tableau ci-dessous.

Latéralité	Droite	Gauche
Tennis	11	3
Autres sports	34	8

Peut-on conclure que la présence de gauchers est plus grande chez les joueurs de tennis ?

6.6 Publications [11]

Pour savoir si les chercheurs se rendent compte ou non s'ils utilisent ou non un test exact de Fisher unilatéral ou bilatéral, W.P. McKinney et ses collègues ont examiné l'utilisation du test exact de Fisher dans des articles de deux grandes revues médicales–*New England Journal of Medicine (NEJM)* et *Lancet*–afin de vérifier si les auteurs notaient correctement le type de test utilisé (Oui-Non).

Utilisation du test	Oui	Non
NEJM	1	8
Lancet	10	4

Existe-t-il une différence dans la présentation correcte du test exact de Fisher entre ces deux journaux ?

6.7 Désastres miniers [10]

Les données sont page 4 dans dans l'ouvrage de Cox et Lewis (1969) traduit par J. Larrieu [5], l'original étant attribué à un article de *Biometrika* de 1952 [10]. La statistique donne la date en jours de 110 désastres survenus dans les mines de charbon en Grande-Bretagne pendant la période 1875-1951. Un désastre implique la mort de dix hommes ou plus. Faire l'analyse de cette série en utilisant le nombre de catastrophes par périodes de temps d'une durée donnée. Utiliser le fichier `http://pbil.univ-lyon1.fr/R/donnees/mine.txt` par :

```
scan("http://pbil.univ-lyon1.fr/R/donnees/mine.txt")
```

```
  [1]     0   378   414   429   460   675   686   823   827   842   914  1010  1134
 [14]  1184  1304  1507  1683  1738  1831  1890  2205  2264  2325  2326  2339  2528
 [27]  2873  2893  2974  3260  3374  3482  3670  3903  3931  3953  4014  4092  4191
 [40]  4517  4792  4846  5063  5176  5208  5231  5382  5743  6055  6409  6467  6742
 [53]  6820  6837  8042  8686  9153 10024 10072 10195 10652 11150 11199 11330 11512
 [66] 11767 11962 12186 12752 13142 13214 13442 13713 13921 14438 16051 16105 16431
 [79] 17743 18091 18836 19053 19173 19448 19468 19534 19825 19829 20198 20536 20872
 [92] 20891 21220 21550 21862 22033 22178 22253 22617 22654 22673 22829 22876 23005
[105] 24635 24664 24881 24888 24906 26263
```

Références

[1] A. Agresti. *An introduction to categorical data analysis.* John Wiley, New York, 1996.

[2] D. Baird. The fisher/pearson chi-squared controversy : a turning point for inductive inference. *The British Journal for the Philosophy of Science*, 34 :105–118, 1983.

[3] A. Bowley. *Elements of Statistics, 6th ed.* Charles Scibner's Sons, New York, USA, 1947.

[4] F. Couty, J. Debord, and D. Fredon. *Probabilitt Statistiques pour biologistes.* Armand Colin, Paris, 1990.

[5] D.R. Cox and P.A.W. Lewis. *L'analyse statistique des ses d'nements.* Traduction de Larrieu, J. Dunod, Paris, 1969.

[6] P. Dumont. *Quelques aspects du cycle vital du grand coregone Coregonus Clupeaformis (Mitchill) de quatre lacs du territoire de la Baie James : les lacs Hne, Nathalie, Yasinski et Alder.* 1977.

[7] R.A. Fisher. On the interpretation of chi square from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, LXXXV(1) :87–94, 1922.

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

[9] S.A. Glantz. *Introduction aux biostatistiques.* Mc Graw-Hill, 1998.

[10] B.A. Maguire, E.S. Pearson, and A.H.A. Wynn. The time intervals between industrial accidents. *Biometrika*, 39 :168–180, 1952.

[11] W. P. McKinney, M. J. Young, A. Hartz, and M. B. Lee. The inexact use of fisher's exact test in six major medical journals. *The Journal of the American Medical Association*, 261(23) :3430–3433, 1989.

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

[13] B. Scherrer. *Biostatistique.* Gan Morin teur, 1984.

[14] R.D. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28(1) :9–12, 1974.

[15] G.U. Yule and M. Greenwood. The statistics of anti-typhoid and anti-cholera innoculations, and the interpretation of such statistics in general. *Royal Society of Medicine Proceedings, Section of Epidemiology and State Medicine*, 8 :113–194, 1915.