

---

## Tests du Khi2

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

---

Loi du Khi2 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 et tests de Monte-Carlo.

### Table des matières

<b>1</b>	<b>Loi du Khi2 et statistique du Khi2</b>	<b>2</b>
<b>2</b>	<b>La question des degrés de liberté</b>	<b>7</b>
<b>3</b>	<b>Test d'ajustement à une distribution connue</b>	<b>8</b>
3.1	Diagnostique[16] . . . . .	8
3.2	Balsamine[5] . . . . .	9
<b>4</b>	<b>Test d'ajustement à une distribution inconnue</b>	<b>9</b>
4.1	Estérase[9] . . . . .	9
4.2	Corégones [16] . . . . .	13
<b>5</b>	<b>Khi2 d'une table de contingence 2-2</b>	<b>14</b>
5.1	Trois modèles pour un test . . . . .	14
5.2	Le test exact de Fisher . . . . .	17
5.3	Le odds ratio . . . . .	19
<b>6</b>	<b>Tests de Monte-Carlo</b>	<b>20</b>
6.1	Quelle est la loi du plus petit? . . . . .	20
6.2	Le nombre de suites est-il gaussien? . . . . .	25
6.3	Exercice : le nombre de cases vides . . . . .	26
6.4	Trois œufs dans cinq cocons[12] . . . . .	26
<b>7</b>	<b>Exercices</b>	<b>26</b>
7.1	Guérison . . . . .	26
7.2	Anesthésie[10] . . . . .	27
7.3	Vaccin . . . . .	27
7.4	Bruit . . . . .	27
7.5	Latéralité . . . . .	27

7.6 Publications [14] . . . . . 28  
 7.7 Désastres miniers [13] . . . . . 28

**Références** **28**

## 1 Loi du Khi2 et statistique du Khi2

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

La loi de probabilité du  $\chi^2$  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  $\chi^2$ .

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" "R" "R" "R" "R" "R" "R" "R" "R" "R" "R"
[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] "B" "B" "B" "R" "B" "B" "B" "V" "V" "V"
sample(urne, 10, T)
[1] "B" "R" "B" "B" "B" "B" "B" "B" "R" "B"
```

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
2 2 1 5
```

On répète l'expérience 10 fois.

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

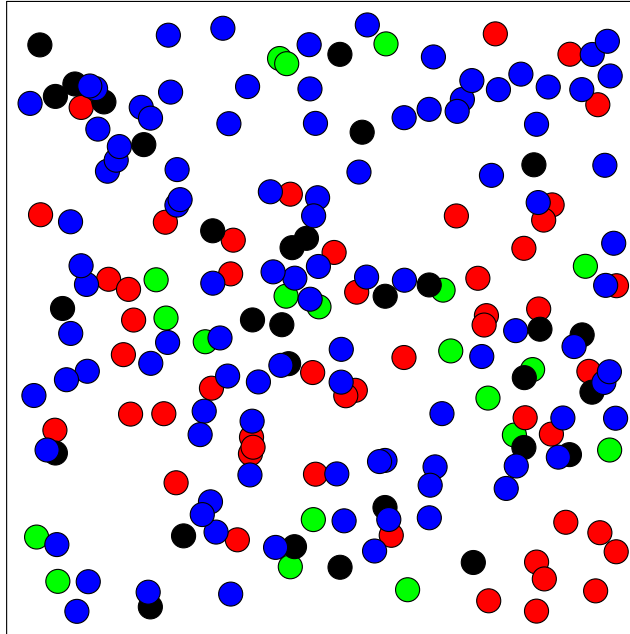


FIGURE 1 – L’urne utilisée pour notre expérience contient 200 boules de quatre couleurs différentes : 20 boules vertes, 50 rouges, 30 noires et 100 bleues.

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

On transpose pour avoir les expériences en ligne.

```
t(replicate(10, table(factor(sample(urne, 10, T), levels = couleurs))))
  V R N B
[1,] 0 2 2 6
[2,] 3 1 2 4
[3,] 1 5 0 4
[4,] 2 3 0 5
[5,] 3 1 4 2
[6,] 0 5 0 5
[7,] 0 4 0 6
[8,] 1 2 2 5
[9,] 1 1 1 7
[10,] 1 0 2 7
```

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
1.025 2.492 1.451 5.032
```

On a eu, en moyenne sur 1000 expériences, respectivement 1.025 vertes, 2.492 rouges, 1.451 noires et 5.032 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 :

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

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

TABLE 1 – Pourquoi introduit-on un dénominateur dans la statistique du Khi2 ? 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.

Par exemple :

$$\text{Khi2} = \frac{(1-1)^2}{1} + \frac{(6-2.5)^2}{2.5} + \frac{(1-1.5)^2}{1.5} + \frac{(2-5)^2}{5} = 6.867$$

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] 2.4
calculkhi2(exp[2, ])
[1] 4.2
```

Le premier écart vaut 2.4, le second vaut 4.2, etc.

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

```
resu <- apply(exp, 1, calculkhi2)
min(resu)
[1] 0.2667
max(resu)
```

[1] 18.87

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

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

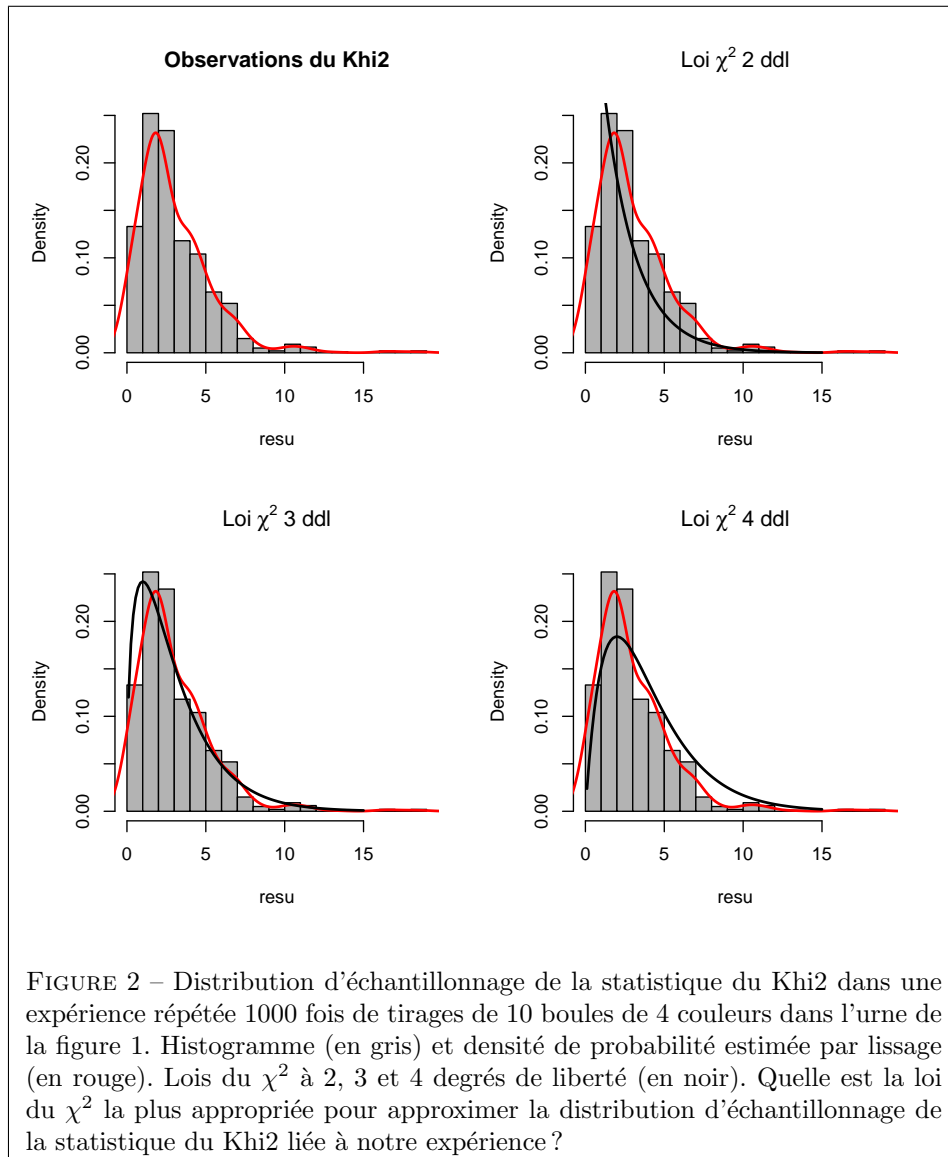
On donne le résultat de l'expérience la plus éloignée du modèle :

```
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.

```
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))
lines(density(resu, adj = 1.5), col = "red", lwd = 2)
hist(resu, pro = T, nclass = 20, main = expression(paste("Loi ",
  chi^2, " 2 ddl")), col = grey(0.7))
lines(density(resu, adj = 1.5), col = "red", lwd = 2)
lines(x0, dchisq(x0, 2), lwd = 2)
hist(resu, pro = T, nclass = 20, main = expression(paste("Loi ",
  chi^2, " 3 ddl")), col = grey(0.7))
lines(density(resu, adj = 1.5), col = "red", lwd = 2)
lines(x0, dchisq(x0, 3), lwd = 2)
hist(resu, pro = T, nclass = 20, main = expression(paste("Loi ",
  chi^2, " 4 ddl")), col = grey(0.7))
lines(density(resu, adj = 1.5), col = "red", lwd = 2)
lines(x0, dchisq(x0, 4), lwd = 2)
```

Cette expérience illustre que la distribution du  $\chi^2$  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  $\chi^2$ ). 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  $\chi^2$  existe en elle-même. La loi de la **statistique** du Khi2, est approximée par une loi du  $\chi^2$ . Le nombre de degrés de liberté est ici égal au nombre de classes moins un,  $4 - 1 = 3$ , c'est ce que nous avons vérifié expérimentalement dans la figure 2.



## 2 La question des degrés de liberté

Le test du  $\chi^2$  n'est pas né du jour au lendemain. Il a été proposé initialement par Karl Pearson en 1900 [15]. Mais un problème concernant les degrés de liberté a été soulevé par Yule et Greenwood en 1915 [17]. Il faut attendre 1922 pour que Fisher [8] 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. Considérez deux joueurs J1 et J2 qui tirent des boules rouges et/ou noires dans une urne et supposez que le résultat soit 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  $\mathbb{R}$ , l'argument lié à la correction de continuité de Yates<sup>1</sup>.

Le test de l'hypothèse  $A$  conduit au résultat suivant :

```
testA <- chisq.test(unlist(obs), correct = FALSE)
testA
      Chi-squared test for given probabilities
data:  unlist(obs)
X-squared = 7.692, 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 :

```
testI <- chisq.test(obs, correct = FALSE)
testI
      Pearson's Chi-squared test
data:  obs
X-squared = 7.692, 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  $A$

$$1\chi^2 = \sum \frac{(|o_i - t_i| - 0.5)^2}{t_i}$$

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 fautive alors que par ailleurs  $A \Rightarrow I$ . La raison est que le test du  $\chi^2$ , 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  $\chi^2$  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  $I$ ,  $\mathbb{R}$  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[16]

*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.

<i>assise</i>	<i>couchée</i>	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 = 5.143, 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

class(provi)
[1] "htest"
names(provi)
[1] "statistic" "parameter" "p.value"    "method"    "data.name" "observed"
[7] "expected"  "residuals"
```

Éditez l'objet, l'explorez, essayez d'identifier chacune de ses composantes.

### 3.2 Balsamine[5]

*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 :

Couleur	pourpre	rose	blanc lavande	blanc
Effectifs	1790	547	548	213

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

*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 = 7.063, df = 3, p-value = 0.06992
```

Discuter le résultat.

## 4 Test d'ajustement à une distribution inconnue

### 4.1 Estérase[9]

*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

	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, 0.3,
0.2, 0.12))
Chi-squared test for given probabilities
data: c(72, 24, 15, 99, 57, 33)
X-squared = 2.503, 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*.

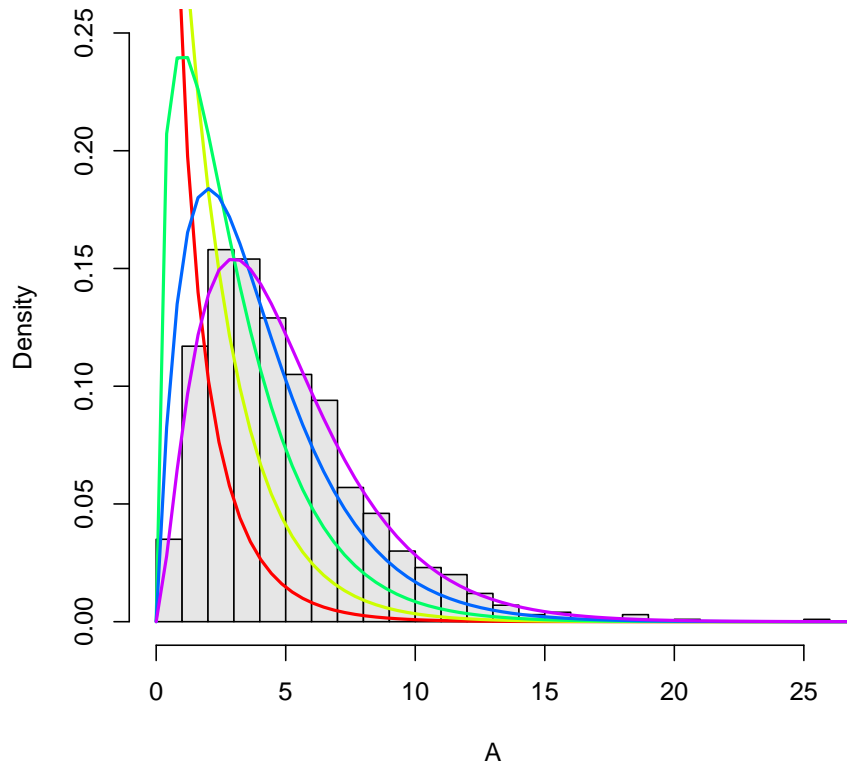
Supposons que ce soit vrai et faisons une simulation sur la loi de Hardy-Weinberg. La population est infinie, les individus sont diploïdes. Chaque individu observé est le résultat d'un tirage au hasard de deux copies du gène. On fait alors la simulation en tirant au hasard 300 copies, puis en tirant 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
 65  53  32  46  28  20  28  22  6
```

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.

```
theo <- 300 * c(0.25, 0.09, 0.04, 0.3, 0.2, 0.12)
simulelem <- fonction(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)
}
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 = 100)
palette(rainbow(5))
for (k in 1:5) lines(x0, dchisq(x0, df = k), lwd = 2, col = palette()[k])
```

### Histogram of A

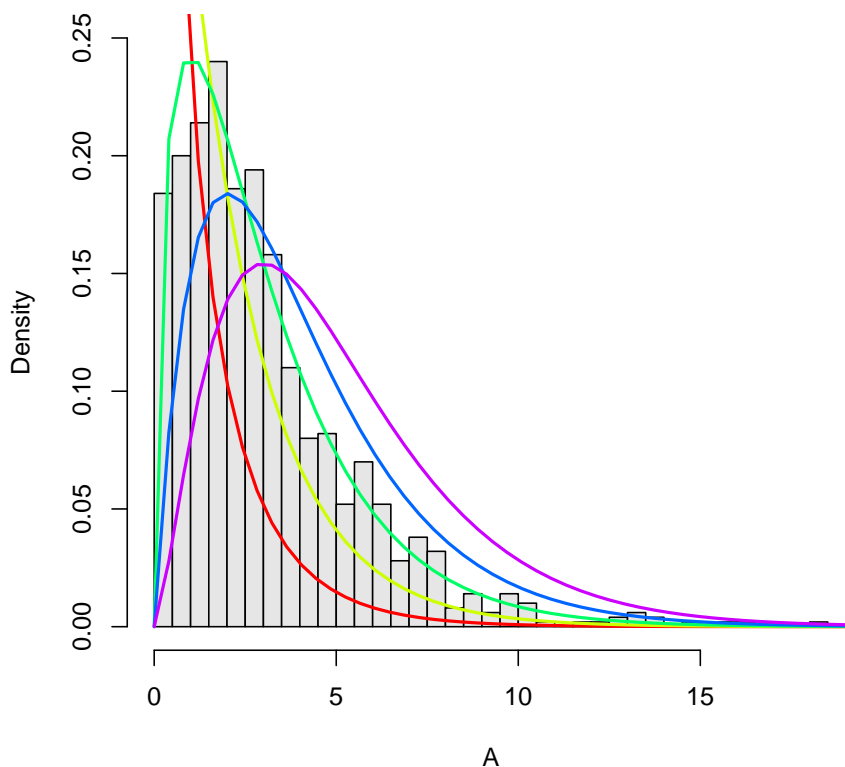


La loi de l'écart est clairement un  $\chi^2$  à 5 degrés de liberté (la figure représente les ddl de 1 à 5).

#### Discussion autour des degrés de liberté de la loi du $\chi^2$

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

### Histogram of A



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.

```

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 = 100)
palette(rainbow(5))

```

```
for (k in 1:5) lines(x0, dchisq(x0, df = k), lwd = 2, col = palette()[k])
```

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 (ici 2 car la somme des trois fréquences alléliques vaut 1) moins 1. La loi est un  $\chi^2$  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, 0.3, 0.2, 0.12))$statistic
```

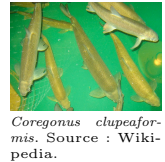
pour une loi à 3 ddl donc :

```
1 - pchisq(val, df = 3)
X-squared
0.4747
```

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

## 4.2 Corégones [16]

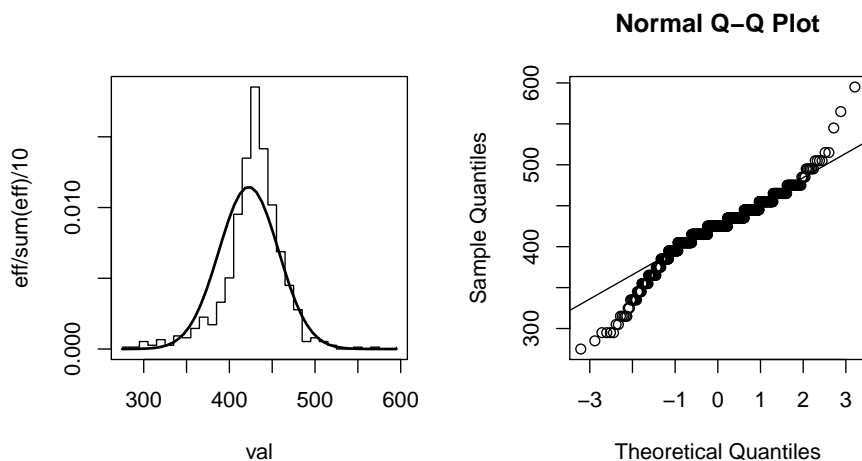
*Enoncé.* Dans le cadre d'une étude sur le cycle vital du grand corégone (*Coregonus clupeaformis*) du lac Nathalie, situé dans le territoire de la baie James, P. Dumont [7] a mesuré la longueur totale du corps (en mm) de 756 individus.



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



Graphiquement, on est très loin de la normalité. *Expliquer pourquoi  $\text{sum}(\text{eff})/10$ .* 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)
[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		Total
		Absence	Présence	
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.

$$\begin{array}{cccc}
 & 0 & 1 & \\
 0 & Z_{00} & Z_{01} & n - n_1 \\
 1 & Z_{10} & Z_{11} & n_1 \\
 & n - n_2 & n_2 & n
 \end{array}$$

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 :

$$P(Z_{11} = j) = \frac{\binom{n}{j, n_1 - j, n_2 - j, n - n_1 - n_2 + j}}{\binom{n}{n_1} \binom{n}{n_2}}$$

## 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.

		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 ?

$$\begin{array}{cccc}
 & 0 & 1 & \\
 A & A_0 & A_1 & n_A \\
 B & B_0 & B_1 & n_B \\
 & A_0 + B_0 & A_1 + B_1 & n_A + n_B
 \end{array}$$

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 :

$$P(A_1 = j / A_1 + B_1 = nk) = \frac{\binom{n_A}{j} \binom{n_B}{k-j}}{\binom{n_A + n_B}{k}}$$

### 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.

		Mâles		Total
		Absence	Présence	
Fem	Absence	49	9	58
	Présence	7	<b>11</b>	<b>18</b>
Total		56	<b>20</b>	<b>76</b>

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

$$\begin{matrix} & 0 & 1 \\ 0 & & \\ 1 & R & p \\ & q & n \end{matrix}$$

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 :

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

Ces trois modèles se résument en un seul par :

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

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))
      Pearson's Chi-squared test with Yates' continuity correction
data:  matrix(c(13, 6, 8, 11), nc = 2)
X-squared = 1.703, 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))
      Pearson's 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))
      Pearson's Chi-squared test with Yates' continuity correction
data:  matrix(c(49, 7, 11, 18), nc = 2)
X-squared = 20.29, df = 1, p-value = 6.665e-06
```

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

## 5.2 Le test exact de Fisher

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

Agresti (1990), p. 61f, Fisher's 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 women's 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)
      Pearson's 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 a 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] 0.01429 0.22857 0.51429 0.22857 0.01429
phyper(0:4, 4, 4, 4)
[1] 0.01429 0.24286 0.75714 0.98571 1.00000
```

```
1 - phyper(0:4, 4, 4, 4)
[1] 0.98571 0.75714 0.24286 0.01429 0.00000
```

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)
Pearson's 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] 0.1573
```

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 fautive. 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)
Pearson's 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")
```

```
Fisher's Exact Test for Count Data
data: TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3136      Inf
sample estimates:
odds ratio
 6.408
```

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

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

```
fisher.test(TeaTasting)

Fisher's 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:
 0.2117 621.9338
sample estimates:
odds ratio
 6.408
```

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))
```

### 5.3 Le odds ratio

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

	0	1		0	1		
$A$	$A_0$	$A_1$	$n_A$	0	$n_{00}$	$n_{01}$	$n - n_1$
$B$	$B_0$	$B_1$	$n_B$	1	$n_{10}$	$n_{11}$	$n_1$
	$A_0 + B_0$	$A_1 + B_1$	$n_A + n_B$		$n - n_2$	$n_2$	$n$

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  $\theta$  plus raffinée. Le concept de odds ratio est fondamental en épidémiologie et essais cliniques.

## 6 Tests de Monte-Carlo

Un papa statisticien, à la psychologie rugueuse, veut apprendre les rudiments à son fils. Il place 99 pièces de 50 centimes et 1 pièce de 5 euros 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 5 euros 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 5 euros 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.

Les plus simples des espaces probabilisés sont discrets : ils comptent un nombre fini d'événements. Quand ce nombre est grand, on connaît beaucoup de choses sur l'espace en tirant au hasard des éléments. On peut ainsi avoir une idée, plus ou moins précise, des lois des variables aléatoires associées à cet espace. La fonction `sample` est fondamentale.

Pour choisir  $m$  parmi  $n$  objets :

```
sample(c("a", "b", "c", "d"), 3, replace = F)
sample(c("a", "b", "c", "d"), 3, replace = F)
```

Pour permuter  $n$  objets :

```
sample(10, 10)
```

Pour placer  $p$  objets dans  $n$  cases :

```
sample(5, 8, replace = T)
```

A consommer sans modération.

### 6.1 Quelle est la loi du plus petit ?

Quelle est la loi du plus petit, de la somme, de la variance, du plus grand, de 10 entiers tirés au hasard parmi 100 sans remise ?

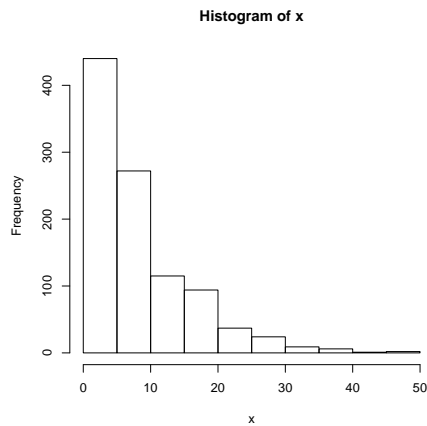
Implanter la fonction :

```
echa <- function(FUN, n = 1000) {
  x <- seq(0, le = n)
  for (i in 1:n) x[i] <- FUN(sample(1:100, 10, replace = F))
  return(x)
}
echa(min, 10)
echa(min, 10)
```

Utiliser cette fonction pour obtenir les lois de variables diverses. Pour le minimum de 10 entiers choisis parmi les 100 premiers :

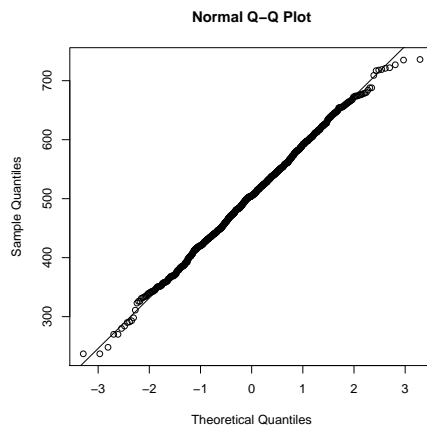
```
x <- echa(min, 1000)
hist(x, nclass = 10)
hist(x, nclass = 10, plot = F)
```

Examiner le détail de la sortie de la fonction `hist` quand on ne fait pas le dessin. C'est parfois fort utile.



Cette loi est très asymétrique. Par contre, la somme de 10 entiers choisis au hasard parmi 100 suit presque une loi normale.

```
x <- echa(sum, 1000)
qqnorm(x)
qqline(x)
```



Pour comprendre les `qqplot` (graphes quantiles-quantiles), il faut savoir qu'une loi de probabilités définit la probabilité de ne pas dépasser une valeur donnée. Pour la loi normale (de moyenne 0 et de variance 1), on a 2.5 chances sur 100 de ne pas dépasser  $-1.96$ , et on a 95 chances sur 100 d'être entre  $-1.96$  et  $1.96$  :

```
pnorm(-1.96)
[1] 0.025
pnorm(1.96)
[1] 0.975
```

Pour la même loi, la valeur qui a une probabilité donnée  $p$  de n'être pas dépassée est le quantile d'ordre  $p$  :

```
qnorm(0.025)
[1] -1.96
```

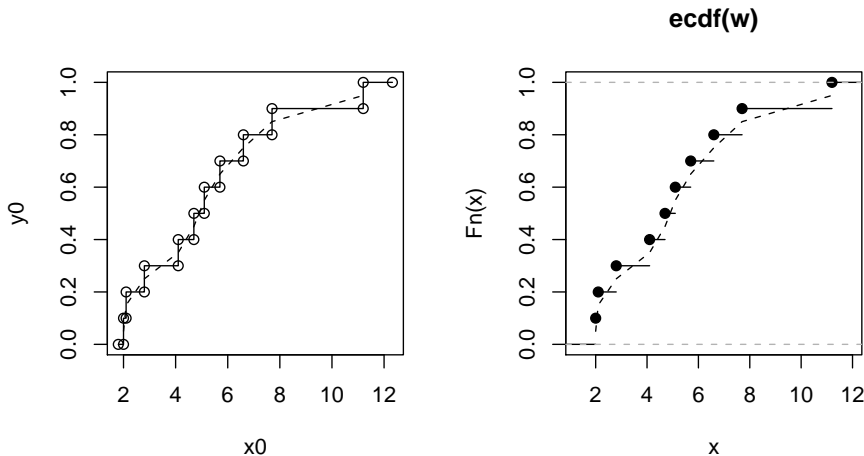
```
qnorm(0.975)
[1] 1.96
```

Considérons alors un échantillon d'observations :

```
w <- c(5.7, 4.1, 2.1, 5.1, 7.7, 4.7, 2, 2.8, 11.2, 6.6)
```

Rangeons ces valeurs par ordre croissant :

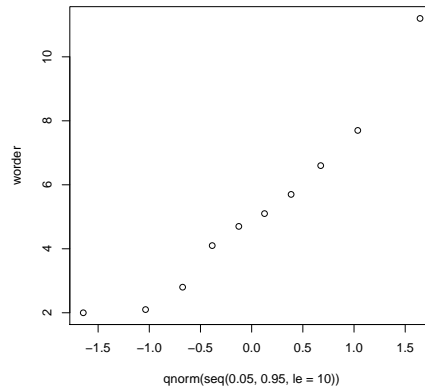
```
worder <- sort(w)
worder
[1] 2.0 2.1 2.8 4.1 4.7 5.1 5.7 6.6 7.7 11.2
x0 <- rep(worder, rep(2, 10))
x0 <- c(0.9 * worder[1], x0, 1.1 * worder[10])
x0
[1] 1.80 2.00 2.00 2.10 2.10 2.80 2.80 4.10 4.10 4.70 4.70 5.10 5.10
[14] 5.70 5.70 6.60 6.60 7.70 7.70 11.20 11.20 12.32
y0 <- rep(seq(0, 1, le = 11), rep(2, 11))
par(mfrow = c(1, 2))
plot(x0, y0, type = "o")
lines(worder, seq(0.05, 0.95, le = 10), lty = 2)
plot(ecdf(w))
lines(worder, seq(0.05, 0.95, le = 10), lty = 2)
```



Un autre procédé consiste à utiliser la fonction `ecdf`. La fonction en escalier qui monte de  $1/10$  à chaque valeur rencontrée est la fonction de répartition empirique. Quand on relie le milieu des "marches", on a le polygone des fréquences cumulées. Ceci montre que la première valeur estime grossièrement le quantile  $0.05$ , la seconde le quantile  $0.15$ , ... En général, les données sont rangées par ordre croissant et notées  $y_{(1)}, y_{(2)}, \dots, y_{(n)}$  (statistiques d'ordre).  $y_{(i)}$  est le quantile empirique pour  $p_i = (i - 0.5)/n$ . On trouvera des explications dans les ouvrages fondateurs de la graphique statistique [4, p. 193][11].

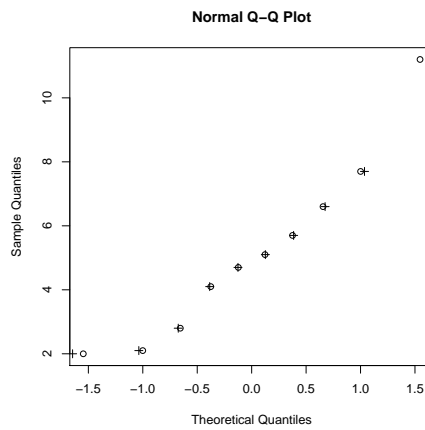
Le graphique quantile-quantile normal représente les couples de points  $(x_i, y_{(i)})$  où  $x_i$  est le quantile de la loi normale pour  $p_i = (i - 0.5)/n$ . Si l'échantillon suit une loi normale, ces points sont sur une droite. C'est l'équivalent du tracé de la droite de Henry sur papier gaussien de nos ancêtres. Ceci est obtenu directement par :

```
plot(qnorm(seq(0.05, 0.95, le = 10)), worder)
qqnorm(w)
```



Mais pas tout à fait !

```
qqnorm(w)
points(qnorm(seq(0.05, 0.95, le = 10)), worder, pch = 3)
```



En fait les  $p_i = (i - 0.5)/n$  :

```
seq(0.05, 0.95, le = 10)
[1] 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
```

ont été remplacées par :

```
ppoints(10)
[1] 0.06098 0.15854 0.25610 0.35366 0.45122 0.54878 0.64634 0.74390 0.84146 0.93902
```


ppoints(base) R Documentation  
 Ordinates for Probability Plotting  
 Description

Generates the sequence of ``probability'' points  $(1:m - a)/(m + (1-a)-a)$  where  $m$  is either  $n$ , if  $\text{length}(n)=1$ , or  $\text{length}(n)$ .

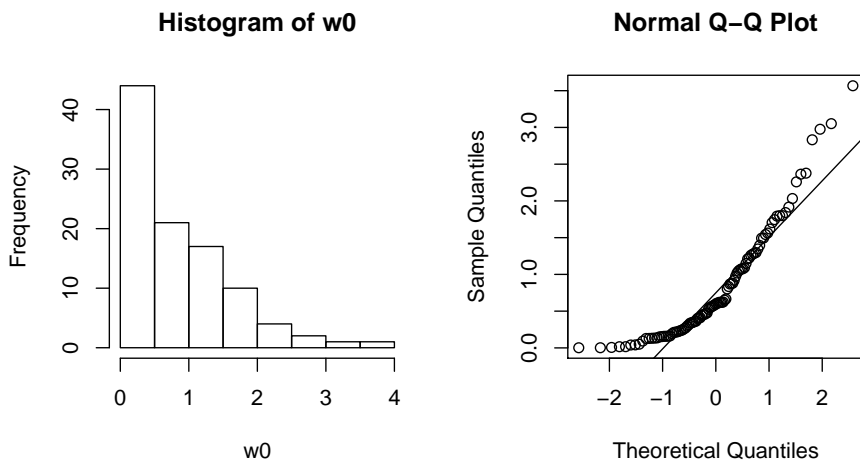
Usage

```
ppoints(n, a = ifelse(n <= 10, 3/8, 1/2))
```

```
ppoints(10)
[1] 0.06098 0.15854 0.25610 0.35366 0.45122 0.54878 0.64634 0.74390 0.84146 0.93902
((1:10) - 0.375)/(10 + 1 - 2 * 0.375)
[1] 0.06098 0.15854 0.25610 0.35366 0.45122 0.54878 0.64634 0.74390 0.84146 0.93902
```

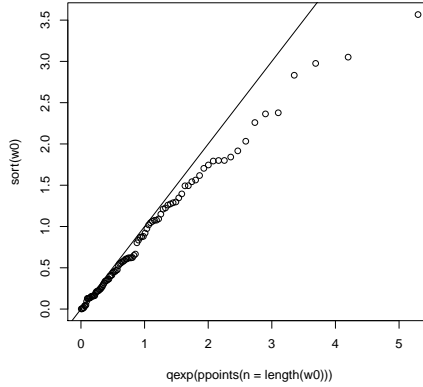
Ouf! La documentation est transparente. Il faut retenir de cette expérience les capacités considérables que  propose pour l'étude des distributions. Par exemple, générer un échantillon d'une loi exponentielle :

```
par(mfrow = c(1, 2))
w0 <- rexp(100)
hist(w0)
qqnorm(w0)
qqline(w0)
```



La distribution de gauche n'est pas gaussienne (et pour cause), elle est exponentielle :

```
qqplot(qexp(ppoints(n = length(w0))), sort(w0))
abline(0, 1)
```



## 6.2 Le nombre de suites est-il gaussien ?

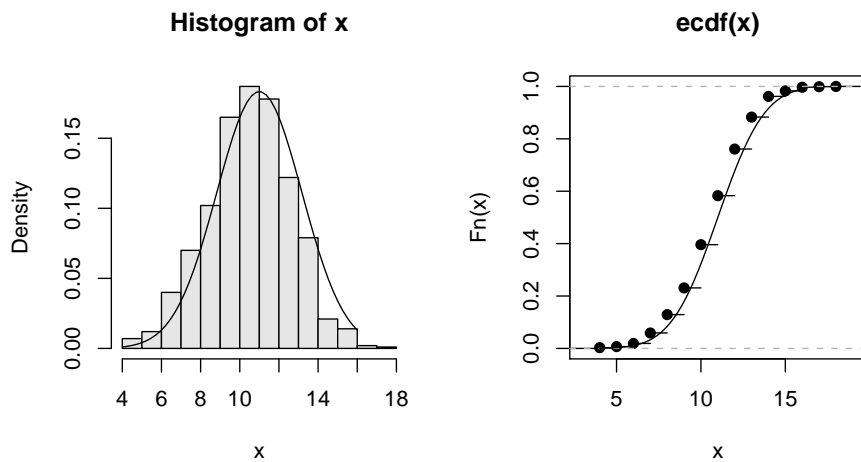
Le nombre de suites est réputé comme gaussien à partir de  $m \geq 10$  et  $n - m \geq 10$ .  
On vérifie ?

```
y <- rep(0, 10)
z <- c(1, 2, 4, 6, 7, 8)
y[z] <- 1
```

Observer ces objets et implanter la fonction :

```
echaNS <- function(necha = 1000, n = 100, m = 10) {
  x <- seq(0, le = necha)
  for (i in 1:necha) {
    y <- rep(0, le = n)
    z <- sample(1:n, m, replace = F)
    y[z] <- 1
    x[i] <- (sum(abs(diff(y)))) + 1
  }
  return(x)
}
echaNS(necha = 5, n = 10, m = 6)
[1] 7 7 6 6 6
```

Que fait-elle ? S'en servir :



```
x <- echaNS(1000, 20, 10)
m0 <- 2 * 10 * 10/20 + 1
v0 <- 2 * 10 * 10 * (2 * 10 * 10 - 20)/20/20/19
par(mfrow = c(1, 2))
hist(x, proba = T, col = grey(0.9))
lines(x0 <- seq(4, 16, le = 50), dnorm(x0, m0, sqrt(v0)))
plot(ecdf(x))
lines(x0, pnorm(x0, m0, sqrt(v0)))
```

C'est une approximation.

### 6.3 Exercice : le nombre de cases vides

Donner une approximation numérique de la loi du nombre de cases vides quand on place au hasard 36 boules dans 26 cases. Une réponse (*en attendant la votre !*)

x	12	13	14	15	16	17	18	19	20	21	22
	0.003	0.013	0.038	0.094	0.176	0.240	0.191	0.144	0.085	0.010	0.006

Si vous croyez que c'est un jeu stérile faites l'exercice suivant.

### 6.4 Trois œufs dans cinq cocons[12]

On propose à une femelle d'hyménoptère parasite 5 cocons de ses hôtes (teigne du poireau). En 24 heures, le parasite pond quelques œufs. L'expérience est répétée plusieurs jours avec plusieurs femelles. On retient l'ensemble des expériences au cours desquelles 3 œufs exactement ont été pondus. Il y en a eu 70. Dans 20 cas, les 3 œufs sont dans le même cocon. Dans 19 cas les 3 œufs sont dans 3 cocons différents. Dans 31 cas (les autres) 2 cocons sont parasités. Caractériser le comportement du parasite.

On peut avoir une solution mathématique. Il peut y avoir 2, 3 ou 4 cocons vides. Énumérer tous les cas possibles. Il y en a 125. 5 cas laissent 4 cocons vides.  $5 \cdot 4 \cdot 3$  cas laissent 3 cocons vides et  $5 \cdot 4 \cdot 3$  cas laissent 2 cocons vides. On a observé 19, 31, 20. On attendait 33.6, 33.6, 2.8 : c'est donc clair ! Il y a une tendance forte au surparasitisme.

```
Chi-squared test for given probabilities
data: c(19, 31, 20)
X-squared = 112.2, df = 2, p-value < 2.2e-16
```

Oui, mais quand on essaye un jeu où la probabilité de perdre est de une chance sur 25 et qu'on perd 20 fois sur 70, on n'a pas besoin de test !

On peut remplacer la distribution exacte par une approximation issue de la fonction précédente, par exemple (0.47783, 0.48175, 0.04042) (*en attendant la votre !*). Cela ne change rien à la conclusion biologique.

## 7 Exercices

### 7.1 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 ?

## 7.2 Anesthésie[10]

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	<i>Vivants</i>	<i>Décédés</i>
Halothane	53	8
Morphine	57	10

## 7.3 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 ?

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

## 7.4 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.

	<i>avec Succès</i>	<i>avec Échecs</i>
<i>sans Succès</i>	62	26
<i>sans Échecs</i>	7	23

Quelle conclusion tirez vous de cette expérience ?

## 7.5 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.

	Droitier	Gaucher
Tennis OUI	11	3
Tennis NON	34	8

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

## 7.6 Publications [14]

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).

	Oui	Non
<i>NEJM</i>	1	8
<i>Lancet</i>	10	4

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

## 7.7 Désastres miniers [13]

Les données sont page 4 dans dans l'ouvrage de Cox et Lewis (1969) traduit par J. Larrieu [6], l'original étant attribué à un article de *Biometrika* de 1952 [13]. 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] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical methods for data analysis*. Duxbury Press, Boston, 1983.
- [5] F. Couty, J. Debord, and D. Fredon. *Probabilités et Statistiques pour biologistes*. Armand Colin, Paris, 1990.

- [6] D.R. Cox and P.A.W. Lewis. *L'analyse statistique des séries d'évènements*. Traduction de Larrieu, J. Dunod, Paris, 1969.
- [7] P. Dumont. *Quelques aspects du cycle vital du grand coregone Coregonus Chupeaformis (Mitchill) de quatre lacs du territoire de la Baie James : les lacs Hélène, Nathalie, Yasinski et Alder*. 1977.
- [8] 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.
- [9] F. Fleury. *Génétique des populations*. Centre National de la Promotion Rurale, Marmillat, BP 100, 63370 Lempdes, 1997.
- [10] S.A. Glantz. *Introduction aux biostatistiques*. Mc Graw-Hill, 1998.
- [11] D. C. Hoaglin, F. Mosteller, and J.W. Tukey, editors. *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.
- [12] V. Labeyrie. Contribution à l'étude de la dynamique des populations d'insectes. influence de l'hôte *Acrolepia assectella* z. sur la multiplication d'un hyménoptère ichneumonidae (*Diadromus* sp.). *Entomophaga*, Mémoire hors-série 1, 1960.
- [13] B.A. Maguire, E.S. Pearson, and A.H.A. Wynn. The time intervals between industrial accidents. *Biometrika*, 39 :168–180, 1952.
- [14] 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.
- [15] 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.
- [16] B. Scherrer. *Biostatistique*. Gaëtan Morin éditeur, 1984.
- [17] G.U. Yule and M. Greenwood. The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Royal Society of Medicine Proceedings, Section of Epidemiology and State Medicine*, 8 :113–194, 1915.