

# Pratique des tests élémentaires

A.B. Dufour & D. Chessel

31 mars 2015

La fiche met en évidence le raisonnement commun à tous les tests statistiques utilisés dans des conditions variées. Sont abordées les comparaisons de moyennes, les données appariées et les statistiques de rang.

## Table des matières

<b>1</b>	<b>Comparaison de la moyenne de deux échantillons</b>	<b>2</b>
1.1	Le test t de comparaison de moyennes - Rappel . . . . .	2
1.2	Le test de Wilcoxon (Mann-Whitney) . . . . .	5
<b>2</b>	<b>Comparaison de données appariées</b>	<b>8</b>
2.1	Le test une chance sur deux . . . . .	9
2.2	L'intervalle de confiance du test "une chance sur deux" . . . . .	11
2.3	Le test t sur données appariées . . . . .	14
2.4	Le test de Wilcoxon sur données appariées . . . . .	15
<b>3</b>	<b>Puissance d'un test</b>	<b>17</b>
<b>4</b>	<b>Comparaison de s échantillons</b>	<b>20</b>
4.1	Test de Kruskal et Wallis . . . . .	20
4.2	Comparer les variances . . . . .	22
<b>5</b>	<b>Comparaison de rangements</b>	<b>24</b>
5.1	Corrélation de rang . . . . .	25
5.2	Concordance de Friedman . . . . .	25
<b>6</b>	<b>Conclusion</b>	<b>26</b>
	<b>Références</b>	<b>27</b>

# 1 Comparaison de la moyenne de deux échantillons

## 1.1 Le test t de comparaison de moyennes - Rappel

*Situation 1* - La variable mesurée est le poids du cerveau (en grammes) de 10 hommes et de 10 femmes (d'après R.J. Glastone, 1905). La variable mesurée diffère-t-elle entre les deux sexes ?

```
males <- c(1381,1349,1258,1248,1355,1335,1416,1475,1421,1383)
females <- c(1055,1305,1155,1120,1252,1208,1154,1197,1229,1212)
```

*Situation 2* - La variable mesurée est le temps de survie (en jours) de patients atteints d'un cancer et traités avec un médicament donné ([3], [1]). Cette variable dépend-t-elle du type de cancer ?

```
estomac <- c(124,42,25,45,412,51,1112,46,103,876,146,340,396)
poumon <- c(1235,24,1581,1166,40,727,3808,791,1804,3460,719)
```

$x_1, x_2, \dots, x_{n_1}$  est un échantillon aléatoire simple. Chaque  $x_i$  est la réalisation d'une variable aléatoire  $X_i$  normale de moyenne  $\mu_1$  et de variance  $\sigma^2$ . Les variables aléatoires  $X_i$  sont indépendantes entre elles, deux à deux. On parle de variables i.i.d.

$y_1, y_2, \dots, y_{n_2}$  est un échantillon aléatoire simple. Chaque  $y_i$  est la réalisation d'une variable aléatoire  $Y_i$  normale de moyenne  $\mu_2$  et de même variance  $\sigma^2$ . Les variables  $Y_i$  sont également indépendantes entre elles, deux à deux. Les  $X_i$  et les  $Y_i$  sont indépendantes.

La différence des deux moyennes est une variable aléatoire  $\bar{X} - \bar{Y}$ .

$E(\bar{X}) = \mu_1$ ;  $E(\bar{Y}) = \mu_2$ .

L'espérance de la variable  $\bar{X} - \bar{Y}$  est donc  $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$ .

$V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y})$ .

$V(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$

La variable normalisée de la différence des moyennes est :

$$Z = \frac{(\bar{X} - \bar{Y}) - E(\bar{X} - \bar{Y})}{\sqrt{V(\bar{X} - \bar{Y})}}$$

soit

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Sous l'hypothèse  $H_0$ ,  $\mu_1 = \mu_2$ , la moyenne de la variable aléatoire  $\bar{X} - \bar{Y}$  est nulle.

De plus, la variance inconnue est estimée par  $\widehat{\sigma^2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$ .

La variable normalisée définie par :

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\widehat{\sigma^2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

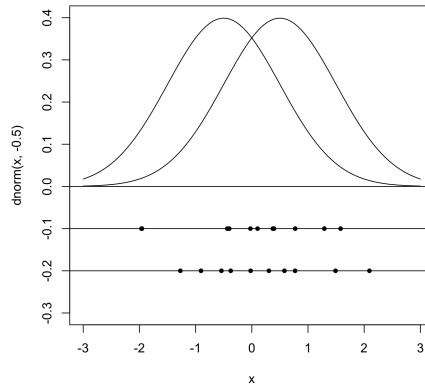
suit une loi T de Student à  $n_1 + n_2 - 2$  degrés de liberté.

Elle prend la valeur particulière :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

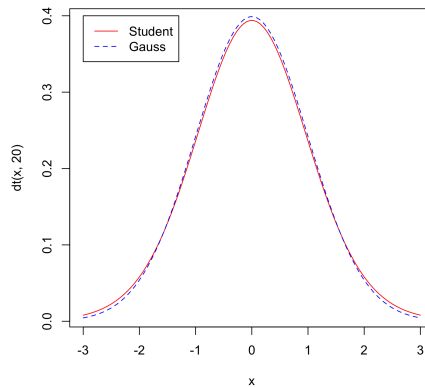
où  $\bar{x} = \frac{x_1 + x_2 + \dots + x_{n_1}}{n_1}$  et  $\bar{y} = \frac{y_1 + y_2 + \dots + y_{n_2}}{n_2}$  sont les moyennes des échantillons.

```
x <- seq(-3, 3, le = 100)
plot(x, dnorm(x,-0.5), type = "l", ylim = c(-0.3,0.4))
lines(x, dnorm(x,0.5), type = "l")
y1 <- rnorm(12,-0.5)
y2 <- rnorm(10,0.5)
abline(h = c(0,-0.1,-0.2))
points(y1, rep(-0.1,12),pch=20)
points(y2, rep(-0.2,10),pch=20)
```



La loi de la différence des moyennes est :

```
plot(x, dt(x,20), type="l", col="red")
lines(x, dnorm(x), type="l", col="blue", lty=2)
legend(-3,0.4,lty=1:2, col=c("red","blue"),legend=c("Student","Gauss"))
```



Si l'alternative est  $\mu_1 > \mu_2$ , les valeurs positives de  $\bar{x}_1 - \bar{x}_2$  sont attendues et on rejette l'hypothèse avec le risque  $P(T > t)$ . Si l'alternative est  $\mu_1 < \mu_2$ , les valeurs négatives de  $\bar{x}_1 - \bar{x}_2$  sont attendues et on rejette l'hypothèse avec le risque  $P(T < t)$ . Si l'alternative est  $\mu_1 \neq \mu_2$ , les valeurs positives ou négatives de  $\bar{x}_1 - \bar{x}_2$  sont attendues et on rejette l'hypothèse avec le risque  $2P(T > |t|)$ .

Appliquée à la situation 1, cette procédure donne :

```
(m1 <- mean(males))
[1] 1362.1
(m2 <- mean(females))
[1] 1188.7
n1 <- length(males)
n2 <- length(females)
varcom <- ((n1-1)*var(males) + (n2-1)*var(females))/(n1+n2-2)
varcom
[1] 4989.056
t <- (m1-m2)/sqrt(varcom*(1/n1 + 1/n2))
t
[1] 5.489401
```

La moyenne du poids du cerveau chez les hommes est de 1362.1g. La moyenne du poids du cerveau chez les femmes est de 1188.7g. La différence normalisée est de 5.4894 pour 18 degrés de liberté. La probabilité d'avoir une différence inférieure à -5.4894 ou supérieure à 5.4894 est :

```
if (t<0) pt(t,18) else (1-pt(t,18))
[1] 1.629934e-05
pcrit <- if (t<0) pt(t,18) else (1-pt(t,18))
2*pcrit
[1] 3.259868e-05
```

La probabilité de la zone de rejet est de 0 et l'hypothèse nulle est rejetée.

Ces résultats se retrouvent sous  par la formulation suivante :

```
tt0 <- t.test(males, females, var.eq = T)
tt0
      Two Sample t-test
data:  males and females
t = 5.4894, df = 18, p-value = 3.26e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 107.0358 239.7642
sample estimates:
mean of x mean of y
 1362.1    1188.7
```

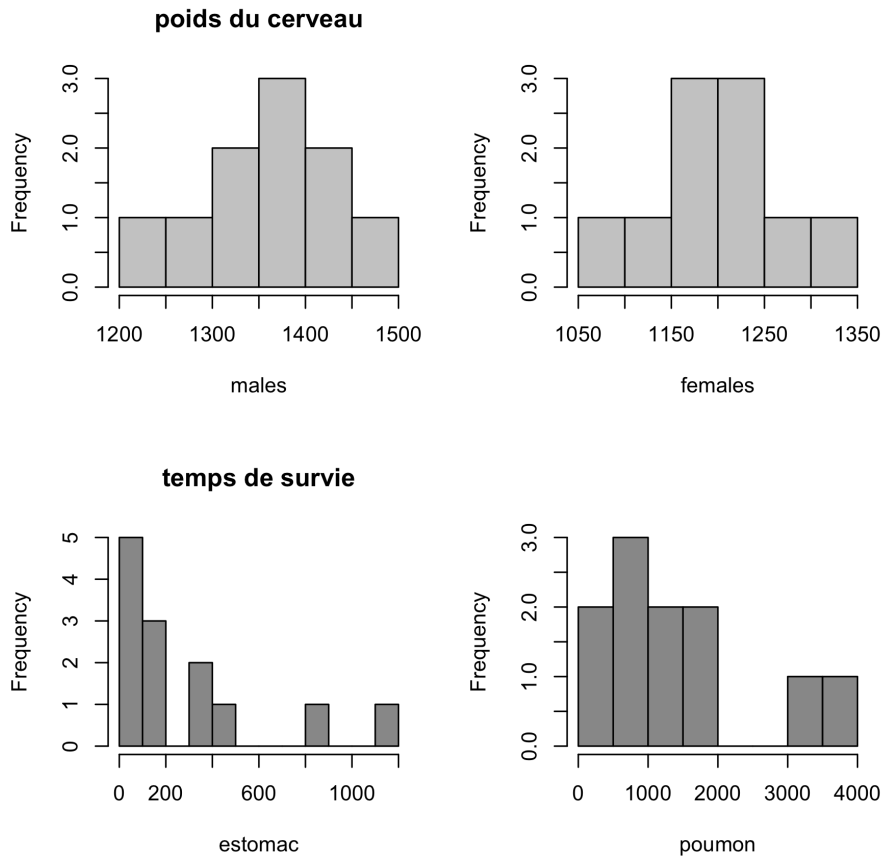
Appliquée à la situation 2, cette procédure donne :

```
tt0 <- t.test(estomac, poumon, var.eq = T)
tt0
      Two Sample t-test
data:  estomac and poumon
t = -3.1013, df = 22, p-value = 0.005209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1852.1210 -367.6971
sample estimates:
mean of x mean of y
 286.000  1395.909
```

## 1.2 Le test de Wilcoxon (Mann-Whitney)

On pourrait croire que les deux situations précédentes sont identiques. Il n'en est rien :

```
par(mfrow=c(2,2))
hist(males,nclass = 8, col=grey(0.8), main="poids du cerveau", xlab="males")
hist(females,nclass = 8, col=grey(0.8), main="", xlab="females")
hist(estomac, nclass = 8, col=grey(0.6), main="temps de survie", xlab="estomac")
hist(poumon, nclass = 8, col=grey(0.6), main="", xlab="poumon")
par(mfrow=c(1,1))
```



La première distribution est relativement symétrique, la deuxième ne l'est pas du tout. Dans l'ensemble des hypothèses nécessaires au test de Student, celui de la normalité peut être globalement invalide. Rejeter l'hypothèse d'égalité des moyennes n'a pas de sens. Il existe des stratégies utilisables quelle que soit la forme de variation des données. On les appelle *libre de distribution*. Le plus simple est le test de Wilcoxon (on dit aussi Mann-Whitney).

Le raisonnement est le suivant. On réunit les deux échantillons :

```
n1 <- length(estomac)
n1
```

```
[1] 13
n2 <- length(poumon)
survie <- c(estomac, poumon)
survie
[1] 124 42 25 45 412 51 1112 46 103 876 146 340 396 1235 24 1581
[17] 1166 40 727 3808 791 1804 3460 719
```

Le rang d'une donnée est le numéro d'ordre de cette donnée quand elles sont rangées par ordre croissant.

```
which.min(survie)
[1] 15
survie[which.min(survie)]
[1] 24
which.max(survie)
[1] 20
survie[which.max(survie)]
[1] 3808
```

La plus petite donnée est située en quinzième position. Elle vaut 24. Son rang est 1.

La plus grande donnée est située en vingtième. Elle vaut 3808. Son rang est 24 :

```
rtot <- rank(survie)
rtot
[1] 9 4 2 5 13 7 18 6 8 17 10 11 12 20 1 21 19 3 15 24 16 22 23 14
```

Les individus du premier groupe ont les rangs :

```
r1 <- rtot[1:n1]
r1
[1] 9 4 2 5 13 7 18 6 8 17 10 11 12
```

Les individus du deuxième groupe ont les rangs :

```
r2 <- rtot[(n1 + 1):(n1 + n2)]
r2
[1] 20 1 21 19 3 15 24 16 22 23 14
```

Si les individus des deux groupes proviennent de la même population, les rangs des individus du premier groupe sont tirés au hasard dans l'ensemble des 24 premiers entiers. Si les moyennes des deux échantillons ne sont pas égales, les rangs du premier groupe auront tendance à être trop grands ou trop petits. La statistique utilisée est **la somme des rangs**.

Evidemment, le raisonnement est symétrique sur les deux groupes. Par convention, on raisonne sur le premier échantillon. On appelle  $m$  l'effectif de cet échantillon et  $n$  l'effectif total. L'effectif du second groupe est par déduction  $n - m$ . Les valeurs  $x_i$  des deux échantillons réunis sont classées par ordre croissant  $x_{(i)}$  notées  $r_i$ , réalisations de variables aléatoires  $R_i$  de loi uniforme (sous réserve qu'il n'y ait pas d'ex-aequo, pour simplifier la démonstration), d'espérance  $E(R_i) = \frac{n+1}{2}$  et de variance  $V(R_i) = \frac{n^2-1}{12}$ .

La statistique du test est la somme des rangs  $SR$  du premier groupe et est appelée statistique de Wilcoxon.

$$SR = \sum_{i=1}^m R_i$$

$$E(SR) = E(\sum_{i=1}^m R_i) = \sum_{i=1}^m E(R_i) \text{ soit } \boxed{E(SR) = \frac{m(n+1)}{2}}$$

$$V(SR) = V(\sum_{i=1}^m R_i) = \sum_{i=1}^m V(R_i) + \sum_{i \neq j} cov(R_i, R_j)$$

On démontre que  $cov(R_i, R_j) = -\frac{n+1}{12}$

$$V(SR) = mV(R_i) + m(m-1)cov(R_i, R_j) \text{ soit } \boxed{V(SR) = \frac{m(n+1)(n-m)}{12}}$$

La statistique de Wilcoxon  $SR$  suit approximativement une loi normale de moyenne  $E(SR)$  et variance  $V(SR)$  ( $m \geq 10$  et  $n - m \geq 10$ ). Il existe des tables donnant les seuils de signification pour les petits effectifs et les bons logiciels donnent la loi exacte connue s'il n'y a pas d'ex aequo.

Souvent, les logiciels utilisent la statistique  $U$  de Mann-Whitney :

$$U = SR - \frac{m(m+1)}{2}$$

D'après ce qui précède et sous les mêmes conditions, celle-ci suit approximativement une loi normale de moyenne et variance :

$$E(U) = \frac{m(n-m)}{2} \quad ; \quad V(U) = \frac{m(n-m)(n+1)}{12}$$

#### A. Calculs à la main sous

- *Statistique de Wilcoxon sr*

```
(sr <- sum(r1))
[1] 122
(esr <- (n1 * (n1 + n2 + 1))/2)
[1] 162.5
(vsr <- (n1 * n2 * (n1 + n2 + 1))/12)
[1] 297.9167
(t0 <- (sr - esr)/sqrt(vsr))
[1] -2.34643
2 * (pnorm(t0))
[1] 0.01895422
```

- *Statistique de Mann-Whitney u*

```
(u <- sr - (n1 * (n1+1))/2)
[1] 31
(eu <- (n1 * n2)/2)
[1] 71.5
(vu <- (n1 * n2 * (n1 + n2 + 1))/12)
[1] 297.9167
(t1 <- (u - eu)/sqrt(vu))
[1] -2.34643
2 * (pnorm(t1))
[1] 0.01895422
```

B. La procédure sous  est :

```
wilcox.test(estomac, poumon, exact = F, correct = F)
      Wilcoxon rank sum test
data:  estomac and poumon
W = 31, p-value = 0.01895
alternative hypothesis: true location shift is not equal to 0
```

Le logiciel donne également la valeur exacte :

```
wilcox.test(estomac, poumon, exact = T)
      Wilcoxon rank sum test
data:  estomac and poumon
W = 31, p-value = 0.01836
alternative hypothesis: true location shift is not equal to 0
```

L'approximation est très justifiée. On referra les calculs pour l'autre exemple sur le poids du cerveau chez l'homme.

```
n1 <- length(males)
n2 <- length(females)
longueur <- c(males, females)
rtot <- rank(longueur)
rtot
[1] 16 14 11  9 15 13 18 20 19 17  1 12  4  2 10  6  3  5  8  7
r1 <- rtot[1:n1]
r1
[1] 16 14 11  9 15 13 18 20 19 17
sr <- sum(r1)
u <- sr - (n1 * (n1+1))/2
eu <- (n1 * n2)/2
vu <- (n1 * n2 * (n1 + n2 + 1))/12
t1 <- (u - eu)/sqrt(vu)
2 * (1 - pnorm(t1))
[1] 0.0003810585
wilcox.test(males, females)
      Wilcoxon rank sum test
data:  males and females
W = 97, p-value = 7.578e-05
alternative hypothesis: true location shift is not equal to 0
```

Le logiciel (professionnel) utilise une correction pour les ex æquo et une correction supplémentaire de continuité. Ce sont des raffinements. On s'en tiendra à la version simple.

Dans la situation 1 (normalité acceptable pour des données morphométriques), le test t et le test de Wilcoxon donnent tous deux un rejet à 0.000. Dans la situation 2 (normalité non acceptable pour des données de temps d'attente), le test t donne un rejet à 0.005 et le test de Wilcoxon un rejet à 0.02. Le test non paramétrique est utilisable dans tous les cas.

## 2 Comparaison de données appariées

La situation des données appariées s'exprime parfaitement en terme de jumeaux. Pour savoir si le produit A permet une croissance plus rapide, on prend  $n$  paires de jumeaux et à l'un des deux, on donne le produit A. L'observation du résultat  $(x_i, y_i)$  ( $x$  avec A,  $y$  sans A) pour la paire  $i$  donne une idée de l'effet de A "toutes choses égales par ailleurs".



La situation des données appariées est celle d'un seul échantillon. Un individu de l'échantillon est un couple. Tester l'hypothèse "A n'a pas d'effet" contre l'hypothèse "A a un effet positif", c'est tester l'hypothèse "la moyenne des  $z_i = x_i - y_i$  est nulle" contre l'hypothèse "la moyenne des  $z_i = x_i - y_i$  est positive".

Il s'agit d'un cas particulier. Un autre cas est celui d'une observation d'un échantillon  $(x_1, x_2, \dots, x_n)$  dont on se demande si la moyenne de la population dont il est extrait dépasse une valeur  $m$  connue. Ceci se ramène à tester l'hypothèse la moyenne des différences est nulle contre l'hypothèse la moyenne des différences est positive.

Nous allons utiliser cette situation pour approfondir la notion de tests statistiques.

## 2.1 Le test une chance sur deux

Un étudiant A soutient qu'il est aussi fort que B aux échecs. Si A gagne une partie contre B, cela ne prouve pas grand chose. S'il en gagne 2 sur 3 non plus. S'il en gagne 8 sur 10, on a une indication. S'il en gagne 99 sur 100, on n'est plus dans la statistique.

La statistique permet d'économiser le travail pour prendre une décision. Le test "une chance sur 2" est simple.  $X$  suit une loi binomiale de paramètres  $n$  et  $p = 1/2$ .

$$P(X = j) = \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} = \frac{n(n-1)\dots(n-j+1)}{j!} \times \frac{1}{2^n}$$

Le niveau de signification du test contre l'hypothèse  $p > 1/2$  est donné par  $P(X \geq x_{obs})$ . Par exemple, si A gagne 8 parties sur 10, on a

$$P(X \geq x_{obs}) = P(X \geq 8) = 0.05469$$

8 sur 10 n'élimine pas le doute sur l'hypothèse "A et B sont de même force aux échecs". Mais 80 sur 100 donne :

$$P(X \geq x_{obs}) = P(X \geq 80) = 5.58 \times 10^{-10}$$

L'intuition qu'il y a longtemps qu'on sait que A est plus fort que B est confirmée. Ce test utilise l'approximation normale pour des valeurs de  $n$  supérieures à 20 :

$$Y = \frac{X - n/2}{\sqrt{n/4}} \rightsquigarrow \mathcal{N}(0, 1)$$

Pour 14 sur 20, on a exactement

$$P(X \geq x_{obs}) = P(X \geq 14) = 0.05766$$

De manière approchée :

$$P(Y \geq y_{obs}) = P\left(Y \geq \frac{14 - 10}{\sqrt{5}}\right) = 0.03682$$

**Exemple.** Dans 5 villes des USA, le nombre de crimes pour 100 000 habitants était entre 1960 et 1970 ([3], dataset 167 - Murder Rates) :

	1	2	3	4	5
<b>1960</b>	10.1	10.6	8.2	4.9	11.5
<b>1970</b>	20.4	22.1	10.2	9.8	13.7

```
dbinom(5,5,0.5)
[1] 0.03125
```

$$P(X \geq 5) = 0.03125$$

La tendance est-elle à l'augmentation ? Il y a augmentation 5 fois sur 5. L'hypothèse " une fois sur deux " est rejetée au seuil de 1/32 (0.0312).

Dans 5 nouvelles villes, on a :

	6	7	8	9	10
<b>1960</b>	17.3	12.4	17.7	8.6	10.0
<b>1970</b>	24.7	15.4	13.1	13.3	18.4

```
sum(dbinom(9:10,10,0.5))
[1] 0.01074219
```

On en est à 9 fois sur 10.

$$P(X \geq 9) = 0.01074$$

Le seuil devient 0.01074.

Dans 5 nouvelles villes, on a :

	11	12	13	14	15
<b>1960</b>	9.1	7.9	4.5	8.1	4.4
<b>1970</b>	16.2	8.2	12.6	17.8	3.9

```
sum(dbinom(13:15,15,0.5))
[1] 0.003692627
```

On passe à 13 fois sur 15.

$$P(X \geq 13) = 0.003693$$

et un seuil de 0.003693.

L'hypothèse de l'absence d'augmentation est grossièrement fausse. Si l'hypothèse est fausse, le risque d'erreur diminue avec le nombre d'échantillons.

Conclure avec le reste des données disponibles :

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
<b>1960</b>	13	9.3	11.7	11.1	11.0	10.8	12.5	8.9	4.4	6.4	3.8	14.2	6.6	6.2	3.3
<b>1970</b>	14	11.1	16.9	12.7	15.6	14.7	12.6	7.9	11.2	14.9	10.5	15.3	11.4	5.5	6.6

Remarque. L'ensemble des données est accessible de la manière suivante :

```
crime <- read.table("http://pbil.univ-lyon1.fr/R/donnees/crimi.txt")
```

## 2.2 L'intervalle de confiance du test "une chance sur deux"

Peut-on admettre un taux d'augmentation de la criminalité de 50% ?

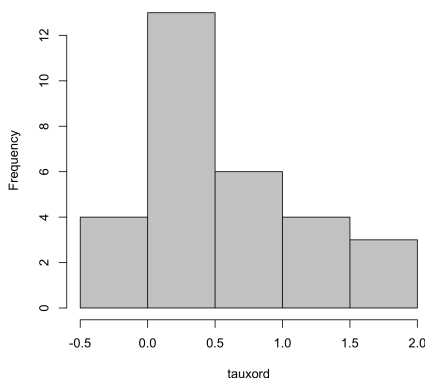
```

crim60 <- crime$V1
crim70 <- crime$V2
taux <- (crim70-crim60)/crim60
taux
[1] 1.01980198 1.08490566 0.24390244 1.00000000 0.19130435 0.42774566
[7] 0.24193548 0.14414414 0.54651163 0.84000000 -0.11363636 0.07692308
[13] 0.19354839 0.44444444 0.78021978 0.03797468 1.80000000 1.19753086
[19] -0.25988701 0.41818182 0.36111111 0.00800000 -0.11235955 1.54545455
[25] 1.32812500 1.76315789 0.07746479 0.72727273 -0.11290323 1.00000000

tauxord <- sort(taux)
tauxord
[1] -0.25988701 -0.11363636 -0.11290323 -0.11235955 0.00800000 0.03797468
[7] 0.07692308 0.07746479 0.14414414 0.19130435 0.19354839 0.24193548
[13] 0.24390244 0.36111111 0.41818182 0.42774566 0.44444444 0.54651163
[19] 0.72727273 0.78021978 0.84000000 1.00000000 1.00000000 1.01980198
[25] 1.08490566 1.19753086 1.32812500 1.54545455 1.76315789 1.80000000

hist(tauxord, col=grey(0.8), main="")

```



La symétrie n'est pas bonne et on ne peut pas aisément raisonner "une chance sur deux" d'être au dessus ou au dessous de la moyenne.

```

mean(tauxord)
[1] 0.5633625

median(tauxord)
[1] 0.4229637

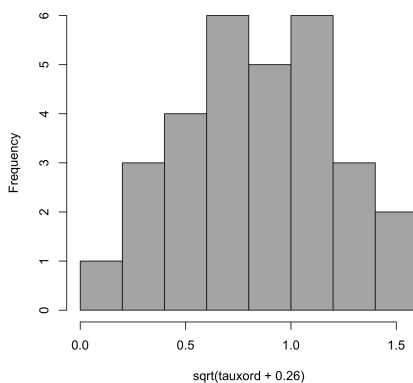
sqrt(tauxord)
[1]      NaN      NaN      NaN      NaN 0.08944272 0.19487094 0.27735010
[8] 0.27832497 0.37966320 0.43738352 0.43994135 0.49186938 0.49386480 0.60092521
[15] 0.64666979 0.65402268 0.66666667 0.73926425 0.85280287 0.88330050 0.91651514
[22] 1.00000000 1.00000000 1.00985245 1.04158805 1.09431753 1.15244306 1.24316312
[29] 1.32783956 1.34164079

sqrt(tauxord + 0.26)
[1] 0.01062988 0.38257501 0.38353197 0.38424009 0.51768716 0.54587057 0.58045075
[8] 0.58091720 0.63572332 0.67179189 0.67346001 0.70847405 0.70986086 0.78810603
[15] 0.82351795 0.82930433 0.83931189 0.89805992 0.99361599 1.01991165 1.04880885
[22] 1.12249722 1.12249722 1.13128333 1.15970068 1.20728243 1.26020832 1.34367204
[29] 1.42237755 1.43527001

```

Observons que le changement de variable  $\sqrt{x + 0.26}$  rend la distribution symétrique :

```
hist(sqrt(tauxord+0.26), col=grey(0.7), main="")
```



```
median(sqrt(tauxord+0.26))
[1] 0.8264111
mean(sqrt(tauxord+0.26))
[1] 0.8410213
neotaux <- sqrt(tauxord+0.26)
```

On repose la question : peut-on admettre un taux d'augmentation de la criminalité de 50% ? Si la médiane réelle est 0.5, la médiane théorique de la variable transformée est :

```
sqrt(0.5+0.26)
[1] 0.8717798
medtheo <- sqrt(0.5+0.26)
compt <- 0
for (i in 1:30) if (neotaux[i] < medtheo) {compt = compt + 1}
compt
[1] 17
```

17 valeurs à gauche et 13 à droite pour "une chance sur deux". Rien à redire.

Peut-on admettre un taux d'augmentation de la criminalité de 100% ?

```
sqrt(1+0.26)
[1] 1.122497
medtheo <- sqrt(1+0.26)
compt <- 0
for (i in 1:30) if (neotaux[i] < medtheo) {compt = compt + 1}
compt
[1] 21
```

21 valeurs à gauche et 9 à droite pour "une chance sur deux". Bizarre ?

```
dbinom(0:30,30,0.5)
[1] 9.313226e-10 2.793968e-08 4.051253e-07 3.781170e-06 2.552290e-05 1.327191e-04
[7] 5.529961e-04 1.895986e-03 5.450961e-03 1.332457e-02 2.798160e-02 5.087564e-02
[13] 8.055309e-02 1.115351e-01 1.354354e-01 1.444644e-01 1.354354e-01 1.115351e-01
[19] 8.055309e-02 5.087564e-02 2.798160e-02 1.332457e-02 5.450961e-03 1.895986e-03
[25] 5.529961e-04 1.327191e-04 2.552290e-05 3.781170e-06 4.051253e-07 2.793968e-08
[31] 9.313226e-10
```

```

pbinom(0:30,30,0.5)
[1] 9.313226e-10 2.887100e-08 4.339963e-07 4.215166e-06 2.973806e-05 1.624571e-04
[7] 7.154532e-04 2.611440e-03 8.062401e-03 2.138697e-02 4.936857e-02 1.002442e-01
[13] 1.807973e-01 2.923324e-01 4.277678e-01 5.722322e-01 7.076676e-01 8.192027e-01
[19] 8.997558e-01 9.506314e-01 9.786130e-01 9.919376e-01 9.973886e-01 9.992845e-01
[25] 9.998375e-01 9.999703e-01 9.999958e-01 9.999996e-01 1.000000e+00 1.000000e+00
[31] 1.000000e+00
2 * (1 - pbinom(21,30,0.5))
[1] 0.0161248

```

C'est significatif au seuil de 0.016.

Faisons le test pour  $x$  au risque de première espèce de 10%.

1. calculer la variable transformée  $y = \sqrt{x + 0.26}$
2. calculer la fréquence à gauche  $g = Nval < y$  et à droite  $d = Nval > y$
3. rejeter l'hypothèse si à gauche on a " 11 au plus " ou si à droite on a " 20 au moins ".

On rejette si  $g$  vaut au plus 11, donc si  $y$  vaut au plus 0.67346 , donc si  $x$  vaut au plus :

```

0.67346 * 0.67346 - 0.26
[1] 0.1935484

```

On rejette si  $d$  vaut au moins 20, donc si  $y$  vaut au moins 1.01991 , donc si  $x$  vaut au moins :

```

1.01991 * 1.01991 - 0.26
[1] 0.7802164

```

L'intervalle de confiance de l'augmentation du taux de criminalité est [19%,78%] au seuil de 90%.

Pour calculer un intervalle de confiance d'une statistique :

- 1- Trouver un test qui permet de rejeter l'hypothèse nulle au risque  $p$  donné.
- 2- Trouver toutes les valeurs de la statistique qui permettent au test de rejeter l'hypothèse nulle au risque  $p$ .
- 3- L'intervalle de confiance au seuil  $1-p$  est l'ensemble des autres valeurs.

Il y a autant d'intervalles de confiance que de tests valides. Pour travailler avec le test " une chance sur deux ", il faut des distributions symétriques. Si ce n'est pas le cas, on fait un changement de variable et on travaille sur la médiane.

**Un test ne sert pas seulement à démontrer une évidence** (le taux de criminalité augmente) **mais aussi à estimer un paramètre** (le taux de criminalité a augmenté d'une fraction comprise entre 19% et 78%).

### 2.3 Le test t sur données appariées

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  est un échantillon formé de  $n$  couples.  $(x_i, y_i)$  est un échantillon aléatoire simple, réalisation de variables aléatoires normales  $X_i$  et  $Y_i$ , indépendantes, de même moyenne  $\mu$  et de même variance  $\sigma^2$ .

$z_i = x_i - y_i$  est la réalisation de la variable aléatoire  $Z_i = X_i - Y_i$  de moyenne 0 et de variance  $2\sigma^2$ .

$$E(Z_i) = E(X_i - Y_i) = E(X_i) - E(Y_i) = \mu - \mu = 0$$

$$V(Z_i) = V(X_i - Y_i) = V(X_i) + V(Y_i) = \sigma^2 + \sigma^2 = 2\sigma^2$$

On veut comparer la moyenne  $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$  à 0.

$\hat{\mu} = \bar{z} = \frac{z_1 + z_2 + \dots + z_n}{n}$  est l'estimateur au maximum de vraisemblance de la moyenne de  $Z$ .

$\widehat{2\sigma^2} = \frac{\sum_{i=1}^n (z_i - m)^2}{n-1}$  est l'estimateur au maximum de vraisemblance de  $2\sigma^2$ .

La moyenne des différences est une variable aléatoire  $\bar{Z}$ .

$$E(\bar{Z}) = \mu \text{ et } V(\bar{Z}) = \frac{2\sigma^2}{n}.$$

La variable normalisée de la moyenne des différences est  $\frac{\bar{Z} - E(\bar{Z})}{\sqrt{V(\bar{Z})}}$ .

Sous l'hypothèse  $H_0$ , la moyenne des différences est nulle :  $\mu = 0$ . La variable normalisée suit une loi de Student à  $n - 1$  degrés de liberté.

Sa variance  $\frac{2\sigma^2}{n}$  est estimée par  $\widehat{\frac{2\sigma^2}{n}}$ .

Une valeur particulière de la variable normalisée, sous l'hypothèse  $H_0$  est :

$$t = \frac{\bar{z}}{\sqrt{\frac{\sum_{i=1}^n (z_i - m)^2}{n(n-1)}}}$$

**Exemple.** On a mesuré la hauteur (en mètres) de 12 arbres selon deux méthodes différentes, avant et après la coupe de l'arbre (Dagnelie [2]).

```
debout <- c(20.4, 25.4, 25.6, 25.6, 26.6, 28.6, 28.7, 29.0, 29.8, 30.5, 30.9, 31.1)
abattu <- c(21.7, 26.3, 26.8, 28.1, 26.2, 27.3, 29.5, 32.0, 30.9, 32.3, 32.3, 31.7)
diflong <- debout - abattu
diflong
```

[1] -1.3 -0.9 -1.2 -2.5 0.4 1.3 -0.8 -3.0 -1.1 -1.8 -1.4 -0.6

```
mean(debout-abattu)
[1] -1.075
var(debout-abattu)/12
[1] 0.1104735
mean(debout-abattu)/sqrt(var(debout-abattu)/12)
[1] -3.234294
t.test(debout, abattu, paired=T)
```

```

Paired t-test
data: debout and abattu
t = -3.2343, df = 11, p-value = 0.007954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.8065536 -0.3434464
sample estimates:
mean of the differences
 -1.075
    
```

Tout test fournit un intervalle de confiance. Discussion.

## 2.4 Le test de Wilcoxon sur données appariées

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  est un échantillon formé de  $n$  couples.  $(x_i, y_i)$  est un échantillon aléatoire simple d'une loi quelconque  $L_i$ .  $z_i = x_i - y_i$  est la réalisation d'une variable aléatoire de médiane 0. L'écart  $z_i$  mesuré sur le couple  $i$  a deux propriétés, son signe et sa valeur absolue. Il se pourrait que le signe soit + environ une fois sur deux mais que les différences dans un sens soient plus faibles en valeur absolue que les différences dans l'autre sens.

Par exemple, des échantillons de crème prélevés dans 10 laiteries sont divisés en deux parties. L'une est envoyée au laboratoire 1 et l'autre au laboratoire 2 pour en compter les bactéries ([4] p. 97). Les résultats (en milliers de bactéries par ml) sont :

```

lab1 <- c(11.7,12.1,13.3,15.1,15.9,15.3,11.9,16.2,15.1,13.8)
lab2 <- c(10.9,11.9,13.4,15.4,14.8,14.8,12.3,15.0,14.2,13.2)
    
```

Les différences des résultats sont :

```

lab1 - lab2
[1] 0.8 0.2 -0.1 -0.3 1.1 0.5 -0.4 1.2 0.9 0.6
    
```

3 valeurs négatives sur 10 n'invalident pas l'hypothèse "une chance sur deux" :

```

pbinom(3,10,0.5)
[1] 0.171875
    
```

On peut améliorer le raisonnement. Si les différences en valeur absolue sont rangées par ordre croissant, chacune d'entre elles, *quel que soit son rang*, a une chance sur deux d'être négative. Le rang 1 a une chance sur deux de porter le signe -. Le rang 2 a une chance sur deux de porter le signe -. Le rang  $n$  a une chance sur deux de porter le signe -. La somme des rangs qui portent le signe - vaut en moyenne :

$$\frac{1}{2} + 2\frac{1}{2} + \dots + n\frac{1}{2} = \frac{n(n+1)}{4}$$

La variance de la somme des rangs qui portent le signe - est  $\frac{n(n+1)(2n+1)}{24}$ . On utilise l'approximation de la loi normale pour la variable :

$$\frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

```

rank(abs(lab1 - lab2))
[1] 7 2 1 3 9 5 4 10 8 6
    
```

```

rank(abs(lab1 - lab2))[lab1 < lab2]
[1] 1 3 4
som <- sum(rank(abs(lab1 - lab2))[lab1 < lab2])
som
[1] 8
compsom <- sum(rank(abs(lab1 - lab2))[lab1 > lab2])
compsom
[1] 47
10 * 11 / 2
[1] 55
moy <- (10 * 11)/4
var <- (10 * 11 * 21)/24
usom <- (som - moy)/sqrt(var)
usom
[1] -1.987624
pnorm(usom)
[1] 0.02342664
2 * pnorm(usom)
[1] 0.04685328
ucompsom <- (compsom - moy)/sqrt(var)
ucompsom
[1] 1.987624
1 - pnorm(ucompsom)
[1] 0.02342664
2 * (1 - pnorm(ucompsom))
[1] 0.04685328
wilcox.test(lab1, lab2, paired = T, correct = F, exact = F)
      Wilcoxon signed rank test
data:  lab1 and lab2
V = 47, p-value = 0.04685
alternative hypothesis: true location shift is not equal to 0

```

Il y a une correction possible pour les petits échantillons :

```

wilcox.test(lab1, lab2, paired = T, correct = T, exact = F)
      Wilcoxon signed rank test with continuity correction
data:  lab1 and lab2
V = 47, p-value = 0.05279
alternative hypothesis: true location shift is not equal to 0

```

On peut calculer la loi exacte pour les petits échantillons :

```

wilcox.test(lab1, lab2, paired = T, correct = T, exact = T)
      Wilcoxon signed rank test
data:  lab1 and lab2
V = 47, p-value = 0.04883
alternative hypothesis: true location shift is not equal to 0

```

On s'en tiendra à la forme la plus simple. On obtient un résultat voisin avec le test t :

```

t.test(lab1, lab2, paired = T)
      Paired t-test
data:  lab1 and lab2
t = 2.4709, df = 9, p-value = 0.03552
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03802232 0.86197768
sample estimates:
mean of the differences
      0.45

```

**Remarque** : on peut faire le test sur la somme des rangs qui portent le signe + et évidemment prendre la même décision.



### 3 Puissance d'un test

En général, quand un test n'est pas significatif, on dit : " Le test ne permet pas de rejeter l'hypothèse nulle ". Le risque de première espèce (probabilité de rejeter l'hypothèse quand elle est vraie) est toujours connu. La probabilité d'accepter l'hypothèse quand elle est fautive l'est rarement. Dans les cas simples on peut la calculer. C'est une fonction du caractère plus ou moins faux de l'hypothèse nulle.

Un étudiant A soutient qu'il est plus fort que B aux dames. Supposons que ça soit vrai. A et B décident de tester l'hypothèse  $H_0$  " A et B sont égaux aux dames " à l'aide de 10 parties. Ils se mettent d'accord sur la procédure. S'ils sont de même force, ils ont une chance sur deux de gagner chaque partie. On dira que A est plus fort s'il gagne un nombre anormal de parties.

```
dbinom(0:10,10,0.5)
```

```
[1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000 0.2050781250 0.2460937500
[7] 0.2050781250 0.1171875000 0.0439453125 0.0097656250 0.0009765625
```

La probabilité de gagner 3 parties est  $P(X = 3) = \binom{10}{3} \left(\frac{1}{2}\right)^{10} = \frac{10 \times 9 \times 8}{3 \times 2} \frac{1}{2^{10}} = \frac{720}{6 \times 1024} = 0.1171875$

```
pbinom(0:10,10,0.5)
```

```
[1] 0.0009765625 0.0107421875 0.0546875000 0.1718750000 0.3769531250 0.6230468750
[7] 0.8281250000 0.9453125000 0.9892578125 0.9990234375 1.0000000000
```

La probabilité d'en gagner au plus 3 est  $P(X \leq 3) = 0.1719$ .

```
1 - pbinom(0:10,10,0.5)
```

```
[1] 0.9990234375 0.9892578125 0.9453125000 0.8281250000 0.6230468750 0.3769531250
[7] 0.1718750000 0.0546875000 0.0107421875 0.0009765625 0.0000000000
```

```
pbinom(0:10,10,0.5, lower.tail=F)
```

```
[1] 0.9990234375 0.9892578125 0.9453125000 0.8281250000 0.6230468750 0.3769531250
[7] 0.1718750000 0.0546875000 0.0107421875 0.0009765625 0.0000000000
```

La probabilité d'en gagner 4 ou plus est de 0.8281.

$$P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3)$$

Ils décident de faire un test au risque  $\alpha = 0.05$ . La probabilité d'en gagner 8 ou plus vaut 0.0547. Donc si A gagne 8, 9 ou 10 parties, on rejettera l'hypothèse  $H_0$ . Sinon ?

Ce qui va se passer dépend de la façon dont A est plus fort que B.

Ceci peut se mesurer par la probabilité  $p$  réelle que A a de gagner une partie contre B. Si A est vraiment très fort  $p = 1$ . Ils jouent 10 parties, A gagne 10 fois et le test donne "  $H_0$  est fautive ". Et elle est bien fautive. B ne dit plus rien. Si A est grandement plus fort que B mais si sa domination n'est pas écrasante, on aura  $p = 0.9$ . Ils jouent 10 fois :

```
dbinom(0:10,10,0.9)
```

```
[1] 0.0000000001 0.0000000090 0.0000003645 0.0000087480 0.0001377810 0.0014880348
[7] 0.0111602610 0.0573956280 0.1937102445 0.3874204890 0.3486784401
```

★ Dans 34.87% des cas, A gagne 10 fois et le test est très significatif :

$$P_{H_0}(X \geq 10) = 0.00097.$$

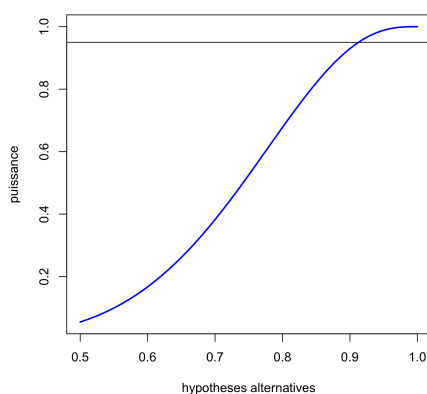
- ★ Dans 38.74% des cas, A gagne 9 fois et le test est très significatif :  
 $P_{H_0}(X \geq 9) = 0.01074$ .
- ★ Dans 19.37% des cas, A gagne 8 fois et le test est significatif :  
 $P_{H_0}(X \geq 8) = 0.05469$ .
- ★ Dans 5.74% des cas A, gagne 7 fois et le test n'est pas significatif :  
 $P_{H_0}(X \geq 7) = 0.1718$ .
- ★ Dans 1.12% des cas, A gagne 6 fois et il l'est encore moins :  
 $P_{H_0}(X \geq 6) = 0.3770$ .

Quand le test n'est pas significatif, B dit "  $H_0$  est vraie " et se trompe. Quelle probabilité avait-il de se tromper c'est-à-dire  $P_{H_1}(X < 8)$  ?

```
pbinom(7,10,0.9)
[1] 0.07019083
```

Il peut dire " j'accepte l'hypothèse nulle ", j'ai 7 chances sur 100 de me tromper ... dans le cas seulement où  $p = 0.9$  ... dans le cas seulement où le test est à 5%. On appelle puissance du test *la probabilité d'erreur quand on accepte l'hypothèse nulle*. C'est une fonction du risque de première espèce et de la vraie alternative. On peut donc tracer la fonction :

```
hypo <- seq(0.5, 1, le = 100)
puiss <- 1 - pbinom(7,10,hypo)
plot(hypo,puiss, xlab="hypothèses alternatives", ylab="puissance", type="l", col="blue", lwd=2)
abline(h=0.95)
```

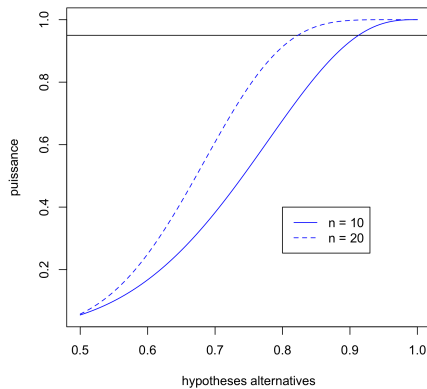


Ceci veut dire que, si le test est significatif, A peut dire qu'il est le plus fort avec une probabilité de se tromper de  $\alpha=5\%$  et que, s'il ne l'est pas, B ne peut pas dire qu'ils se valent. Il pourra juste dire que la puissance du test est suffisante pour dire que  $p \leq 0.9$  mais pas que  $p = 0.5$ . Le graphe ci-dessus est celui de la puissance du test " une chance sur 2 " quand  $n$  vaut 10 et  $\alpha=5\%$ . La puissance du test est une fonction de  $n$ . Il faut recommencer le raisonnement en entier pour  $n = 20$ . Ceci s'automatise par :

```

puissance <- fonction(n, trait, couleur) {
hypo <- seq(0.5, 1, le = 100)
puiss <- 1 - pbinom(qbinom(0.95,n,0.5) - 1, n, hypo)
lines(hypo, puiss, lty=trait, col=couleur)
}
plot(hypo,puiss, xlab="hypothèses alternatives", ylab="puissance", type="n")
abline(h = 0.95)
puissance(10,1,"blue")
puissance(20,2,"blue")
legend(0.8,0.4,lty=c(1,2),legend=c("n = 10","n = 20"), col="blue")

```

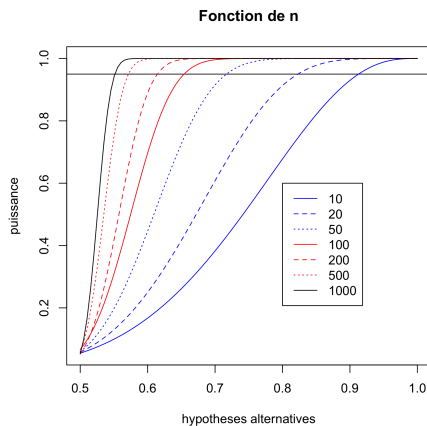


On peut maintenant refaire cette opération pour 50, 100, 200, 500 et 1000.

```

plot(hypo,puiss, xlab="hypothèses alternatives", ylab="puissance", type="n", main = "Fonction de n")
abline(h = 0.95)
effectif <- c(10,20,50,100,200,500,1000)
trait <- c(rep(1:3,2),1)
couleur <- rep(c("blue","red","black"),c(3,3,1))
for (i in 1:7) puissance(effectif[i],trait[i],couleur[i])
legend(0.8,0.6,lty=trait, col=couleur,legend=c(effectif))

```



En faisant 20 parties, on pourra mettre en évidence  $p < 0.83$ ; en faisant 100 parties  $p < 0.63$ ; en faisant 1000 parties  $p < 0.53$ . On n'aura jamais d'argument pour  $p = 0.5$ .

**En statistique, on n'accepte JAMAIS une hypothèse. On discute au mieux de celles qu'on peut refuser avec un risque donné.**

## 4 Comparaison de s échantillons

### 4.1 Test de Kruskal et Wallis

Dans la bibliothèque de Peter Sprent <sup>6</sup>, il y a des livres de voyage, des ouvrages généraux et des livres de statistiques . On a trois échantillons.

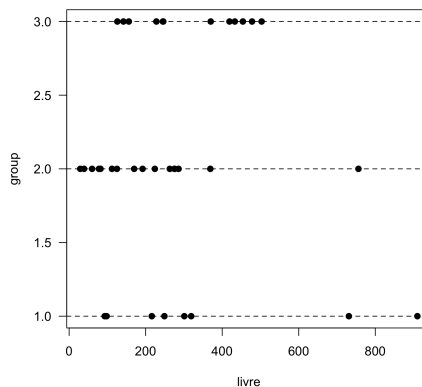
1. Le premier est celui des livres de voyage. Ils ont respectivement 93, 98, 216, 249, 301, 319, 731 et 910 pages.
2. Le second est celui des livres généraux. Ils ont 29, 39, 60, 78, 82, 112, 125, 170, 192, 224, 263, 275, 276, 286, 369 et 756 pages.
3. Le troisième est celui des livres de statistiques. Ils ont 126, 142, 156, 228, 245, 246, 370, 419, 433, 454, 478 et 503 pages.

La question est “ Ces échantillons proviennent-ils d’une même population ou au contraire un au moins des trois échantillons présentent une originalité en ce qui concerne la taille moyenne? ”.

```
livre <- c(93,98,216,249,301,319,731,910,29,39,60,78,82,112,125,170,192,224,263,
          275,276,286,369,756,126,142,156,228,245,246,370,419,433,454,478,503)
group <- factor(rep(1:3,c(8,16,12)))
```

Le ”bon” dessin est :

```
plot(livre, group, pch=20, las=1, cex=1.5)
abline(h=1, lty=2)
abline(h=2, lty=2)
abline(h=3, lty=2)
```



Les données complètes sont rangées :

```
rank(livre)
[1] 6 7 15 20 25 26 34 36 1 2 3 4 5 8 9 13 14 16 21 22 23 24 27 35 10 11 12
[28] 17 18 19 28 29 30 31 32 33
rank(livre)[group==1]
[1] 6 7 15 20 25 26 34 36
rank(livre)[group==2]
[1] 1 2 3 4 5 8 9 13 14 16 21 22 23 24 27 35
```

```
rank(livre)[group==3]
[1] 10 11 12 17 18 19 28 29 30 31 32 33
```

Si  $n$  est le nombre total d'échantillons, la somme des rangs vaut toujours :

$$SR_{tot} = 1 + \dots + n = \frac{n(n+1)}{2}$$

```
sum(rank(livre))
[1] 666
36*37/2
[1] 666
```

Si  $n$  est le nombre total d'échantillons, la somme des carrés des rangs vaut toujours :

$$SCR_{tot} = 1^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

```
36*37*(2*36+1)/6
[1] 16206
```

Si  $SR_j$  est la somme des rangs de l'échantillon  $j$ , la variable  $T$  définie par :

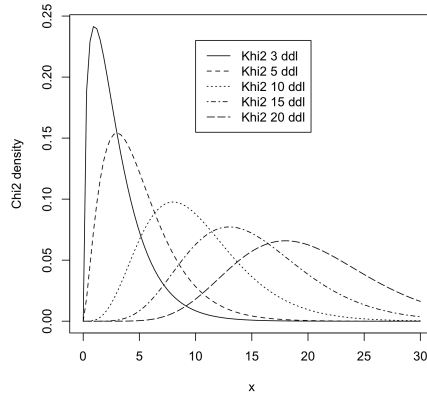
$$T = \frac{12 \sum_{j=1}^s \frac{SR_j^2}{n_j}}{n(n+1)} - 3(n+1)$$


suit une loi Khi2 à  $(s-1)$  degrés de liberté. Cela suffit pour exécuter le test de Kruskal-Wallis.

```
s1 <- sum(rank(livre)[group==1])
s2 <- sum(rank(livre)[group==2])
s3 <- sum(rank(livre)[group==3])
c(s1,s2,s3)
[1] 169 227 270
(s1^2/length(s1))+(s2^2/length(s2))+(s3^2/length(s3))
[1] 152990
(s1^2/length(s1))+(s2^2/length(s2))+(s3^2/length(s3)) -> partiel
partiel*12/(36*37)-3*37
[1] 1267.288
```

La loi Khi2 à  $m$  ddl est définie dans la théorie comme celle de la somme de  $m$  carrés de lois normales indépendantes. Les densités de probabilité ont des formes typiques :

```
x0 <- seq(0,30,le=100)
y1 <- dchisq(x0,3)
y2 <- dchisq(x0,5)
y3 <- dchisq(x0,10)
y4 <- dchisq(x0,15)
y5 <- dchisq(x0,20)
plot(x0, y1, type="n", xlab="x", ylab="Chi2 density")
lines(x0, y1, lty = 1)
lines(x0, y2, lty = 2)
lines(x0, y3, lty = 3)
lines(x0, y4, lty = 4)
lines(x0, y5, lty = 5)
l0 = c("Khi2 3 ddl", "Khi2 5 ddl", "Khi2 10 ddl", "Khi2 15 ddl", "Khi2 20 ddl")
legend(10,0.23,10, lty=1:5)
```



Les quantiles sont disponibles dans la table du Khi2. Conclure.  
La procédure du test sous  est donnée par :

```
kruskal.test(livre~group)
      Kruskal-Wallis rank sum test
data:  livre by group
Kruskal-Wallis chi-squared = 4.9071, df = 2, p-value = 0.08599
```

## 4.2 Comparer les variances

Dans trois groupes de TD, les notes de contrôle continu sont :

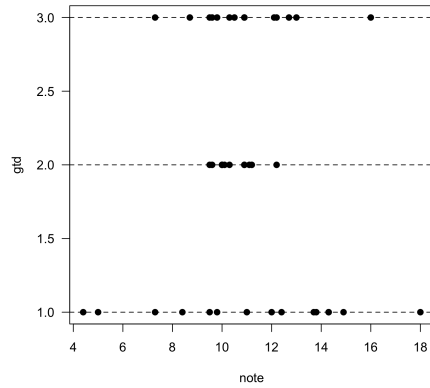
```
x1 <- c(14.9,12.0,9.5,7.3,8.4,9.8,11.0,13.8,14.3,5.0,4.4,14.3,13.7,18.0,12.4)
x2 <- c(10.9,10.1,10.0,12.2,10.0,11.1,10.3,9.5,9.6,10.0,10.9,11.2)
x3 <- c(13.0,12.1,8.7,10.9,12.7,9.5,10.5,12.2,16.0,10.3,9.6,10.9,7.3,9.8)
```

Les enseignants se réunissent pour vérifier qu'il n'y a pas de différences sensibles entre leur notation.

```
mean(x1)
[1] 11.25333
mean(x2)
[1] 10.48333
mean(x3)
[1] 10.96429
gtd <- rep(1:3, c(15,12,14))
note <- c(x1, x2, x3)
kruskal.test(note, gtd)
      Kruskal-Wallis rank sum test
data:  note and gtd
Kruskal-Wallis chi-squared = 0.7204, df = 2, p-value = 0.6975
```

Tout va bien. "Pas du tout" dit le représentant des étudiants. Faites donc le bon dessin :

```
plot(note, gtd, pch=20, las=1, cex=1.5)
abline(h=1, lty=2)
abline(h=2, lty=2)
abline(h=3, lty=2)
```

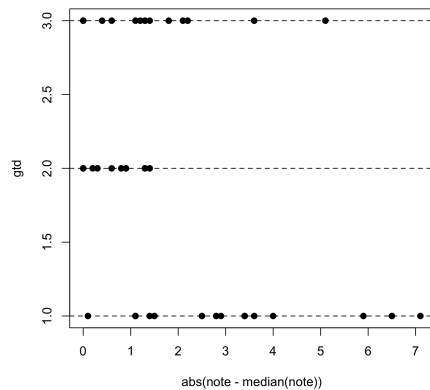


- ★ Vous voyez bien que les amplitudes de notation diffèrent grandement entre les groupes.
- ★ Mais non, c'est le hasard.
- ★ Impossible!
- ★ Alors, prouvez le!

```

median(note)
[1] 10.9
kruskal.test(abs(note - median(note))~gtd)
      Kruskal-Wallis rank sum test
data: abs(note - median(note)) by gtd
Kruskal-Wallis chi-squared = 14.3197, df = 2, p-value = 0.0007772
plot(abs(note - median(note)), gtd, pch=20, cex = 1.5)
abline(h=1, lty=2)
abline(h=2, lty=2)
abline(h=3, lty=2)

```



**Moralité** : un test peut en cacher un autre.

## 5 Comparaison de rangements

La comparaison des produits alimentaires est souvent basée sur la dégustation. On demande à un expert de classer du premier au dernier  $n$  produits. Quand deux experts font la même étude, se pose la question de la cohérence de leur jugement. Quand  $p$  experts forment un jury la mesure de cette cohérence est essentielle!

Par exemple, on a demandé à 24 étudiants de ranger par ordre de préférence 10 groupes de musique. La personne n°1 (première ligne) préfère le 7, ensuite le 3, ensuite le 9, ensuite le 10, ... enfin le 5 :

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

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	7	3	9	10	6	8	4	2	1	5
2	4	8	7	1	9	6	2	3	5	10
3	7	6	8	3	2	1	4	9	10	5
4	6	7	3	5	2	9	10	8	1	4
5	1	2	3	5	8	4	7	10	6	9
6	1	5	6	3	2	7	8	4	9	10
7	2	6	10	5	7	3	8	1	4	9
8	6	7	9	5	3	2	8	10	1	4
9	7	6	5	9	2	1	3	8	10	4
10	5	7	6	3	9	1	4	10	2	8
11	6	7	5	2	3	1	4	10	9	8
12	7	5	6	9	3	2	10	8	4	1
13	6	7	3	5	2	1	4	8	9	10
14	7	2	10	5	3	1	6	8	9	4
15	6	7	10	1	2	5	3	4	8	9
16	6	10	7	3	2	1	5	4	8	9
17	3	7	10	2	6	5	9	8	1	4
18	8	7	9	6	5	3	1	2	4	10
19	5	7	3	8	6	9	4	1	2	10
20	4	6	3	5	8	9	7	2	1	10
21	6	8	7	5	3	2	4	9	1	10
22	6	7	5	9	8	4	3	1	2	10
23	6	7	9	5	8	3	4	1	2	10
24	6	3	7	1	5	2	8	4	9	10

```
matrock <- as.matrix(rock)
neorock = matrix(rep(0,240), nrow = 24)
for (i in 1:24) (neorock[i,] = order(matrock[i,]))
neorock
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	9	8	2	7	10	5	1	6	3	4
[2,]	4	7	8	1	9	6	3	2	5	10
[3,]	6	5	4	7	10	2	1	3	8	9
[4,]	9	5	3	10	4	1	2	8	6	7
[5,]	1	2	3	6	4	9	7	5	10	8
[6,]	1	5	4	8	2	3	6	7	9	10
[7,]	8	1	6	9	4	2	5	7	10	3
[8,]	9	6	5	10	4	1	2	7	3	8
[9,]	6	5	7	10	3	2	1	8	4	9
[10,]	6	9	4	7	1	3	2	10	5	8
[11,]	6	4	5	7	3	1	2	10	9	8
[12,]	10	6	5	9	2	3	1	8	4	7
[13,]	6	5	3	7	4	1	2	8	9	10
[14,]	6	2	5	10	4	7	1	8	9	3
[15,]	4	5	7	8	6	1	2	9	10	3
[16,]	6	5	4	8	7	1	3	9	10	2
[17,]	9	4	1	10	6	5	2	8	7	3
[18,]	7	8	6	9	5	4	2	1	3	10
[19,]	8	9	3	7	1	5	2	4	6	10
[20,]	9	8	3	1	4	2	7	5	6	10
[21,]	9	6	5	7	4	1	3	2	8	10
[22,]	8	9	7	6	3	1	2	5	4	10
[23,]	8	9	6	7	4	1	2	5	3	10
[24,]	4	6	2	8	5	1	3	7	9	10

Le code des groupes est 1-Metallica, 2-Guns n' Roses, 3-Nirvana, 4-AC/DC, 5-Noir Désir, 6-U2, 7-Pink Floyd, 8-Led Zeppelin, 9-Deep Purple, 10-Bon Jovi. Les comparaisons de rangs portent sur les données de droite.



## 5.1 Corrélation de rang

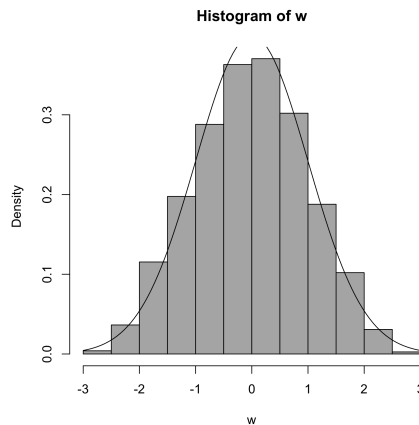
Prenons par exemple les classements des deux premiers étudiants.

```
c1 <- c(9, 8, 2, 7, 10, 5, 1, 6, 3, 4)
c2 <- c(4, 7, 8, 1, 9, 6, 3, 2, 5, 10)
cor(c1, c2)
[1] 0.03030303
```

$$r = \frac{\frac{1}{n} \sum_{i=1}^n r_i s_i - \frac{(n+1)^2}{4}}{\frac{n^2-1}{12}}$$

Cette valeur se teste comme un cas particulier du chapitre 2, § 4.4 :

```
w = rep(0,10000)
n = 10
for (i in 1:10000) {
  a0 = sum(1:n)*sample(n)
  a0 = a0-(n*(n+1)*(n+1)/4)
  a0 = a0/sqrt(n*n*(n+1)*(n*n-1)/12/12)
  w[i] = a0
}
x0 = seq(-3,3,le=100)
hist(w, proba=T, nclass=20, col=grey(0.7))
lines(x0,dnorm(x0))
```



L'approximation est satisfaisante à partir de 10.

## 5.2 Concordance de Friedman

```
apply(neorock,2,sum)
[1] 159 139 108 179 109 68 64 152 160 182
```

Discuter. Les différences sont-elles extraordinaires ?  $R_j$  est la somme des rangs obtenus par le produit  $j$ . Il y a  $n$  produits et  $p$  juges.

La variable  $Q = \frac{12}{n(n+1)p} \sum_{j=1}^n R_j^2 - 3p(n+1)$  suit une loi Khi2 à  $n-1$  degrés de liberté.

```
sum(apply(neorock,2,sum)^2)
```

```
[1] 190736
      sum(apply(neorock,2,sum)^2)*12/10/11/24-3*11*24
[1] 74.98182
      friedman.test(neorock)
      Friedman rank sum test
data: neorock
Friedman chi-squared = 74.9818, df = 9, p-value = 1.593e-12
```

Les préférences des étudiants sont très marquées et le “ jury ” est cohérent. Pour un autre exemple, 25 juges classent par ordre de préférence les 8 bouteilles finalistes du concours de la foire de Mâcon :

```
library(ade4)
data(macon)
macon
  a b c d e f g h i j k l m n o p q r s t u v w x y
A 5 5 4 3 3 4 7 2 1 3 5 4 4 5 4 8 5 7 8 5 4 6 7 2 8
B 4 8 2 4 1 5 2 7 8 8 1 6 3 7 8 5 7 8 1 4 1 5 4 4 6
C 2 6 1 1 6 2 1 5 5 4 3 7 2 2 6 2 1 6 2 1 2 1 2 5 1
D 6 7 5 8 2 6 8 8 6 6 6 5 6 6 3 6 8 1 7 6 7 4 1 6 7
E 1 4 3 2 7 1 6 4 3 1 2 8 1 1 1 3 2 2 6 2 8 2 8 1 2
F 3 2 8 6 5 8 3 3 4 7 8 1 5 8 7 4 4 3 3 8 6 8 6 7 3
G 7 1 6 5 4 7 4 1 7 5 7 3 8 3 2 7 3 5 4 7 3 7 3 8 5
H 8 3 7 7 8 3 5 6 2 2 4 2 7 4 5 1 6 4 5 3 5 3 5 3 4
```

Les dégustateurs sont de 5 catégories professionnelles : [1,5] Œnologues, [6,10] Restaurateurs, [11,15] Négociants, [16,20] Viticulteurs, [21,25] Organismes du concours (Données non publiées de la Société d’Œnologie).

## 6 Conclusion

La statistique n’est pas un ensemble de recettes de “ cuisine ” mais une manière d’extraire l’information des résultats de l’expérience. C’est la dernière partie de l’expérience. A retenir les idées essentielles :

**Inférence** : inférer c’est parler de la population dont on a extrait un échantillon. Le calcul des probabilités parle de l’échantillon à partir de la population, la statistique inférentielle parle de la population à partir de l’échantillon. Les animaux capturés sont un échantillon des animaux capturables, les personnes interrogées sont un échantillon des personnes interrogeables, les résultats de l’expérience sont un échantillon des résultats des expériences réalisables.

**Vraisemblance** : La probabilité d’observer le résultat sous une hypothèse  $H$  arbitraire est la vraisemblance de cette hypothèse pour cette observation.

**Estimation** : Estimer un paramètre, c’est donner les valeurs qui ne sont pas contradictoire avec les résultats. Estimer au maximum de vraisemblance, c’est choisir l’hypothèse qui donne à l’observation la plus grande vraisemblance.

**Test** : Tester une hypothèse, c’est décider au vu du résultat si elle est vraie ou si elle est fausse. Si le résultat est invraisemblable sous l’hypothèse choisie, on la rejette. Si le résultat est vraisemblable sous l’hypothèse choisie, on ne la rejette pas. Dans tous les cas, on risque de se tromper. Dire que l’hypothèse est fausse alors qu’elle est vraie, c’est faire une erreur de première espèce.

**Ajustement** : pour étudier l'écart entre une distribution théorique et une distribution observée, on utilise un test Khi2 d'ajustement (2.2.2, 2.2.3, 2.4.1). Pour étudier l'écart entre une fonction de répartition et son modèle, on utilise le test de Kolmogorov-Smirnov (2.3.1) ou le test de Cramer (2.3.2). Pour tester l'autocorrélation dans une suite binaire on utilise le nombre de suites (2.4.3). Sur les mêmes données, on peut avoir des tests significatifs et d'autres qui ne le sont pas. Ils ne portent pas sur la même alternative. Il y a de nombreux tests : ils ont en commun le même raisonnement.

**Alternative** : faire un test statistique, c'est choisir une hypothèse nulle, une statistique et une zone de rejet peu probable ( $p$ ) quand l'hypothèse nulle est vraie et probable quand une hypothèse alternative précisée est vraie. La valeur calculée tombe dans la zone de rejet, on rejette l'hypothèse nulle au profit de l'alternative. Si l'hypothèse nulle est fausse, tant mieux. Si elle est vraie, on a commis une erreur de première espèce. La probabilité de se tromper est  $p$ .

**Non paramétrique** : on peut comparer deux moyennes par un test t si les données sont normales. Sinon on utilise un test de Wilcoxon. Les tests "une chance sur deux", les tests de Wilcoxon, le test de Friedman sont non paramétriques. Ils ne supposent pas d'hypothèses particulières. Seul le raisonnement qu'on utilise quand on s'en sert permet leur usage dans des conditions très variées.

## Références

- [1] E. Cameron and L. Pauling. Supplemental ascorbate in the supportive treatment of cancer : re-evaluation of prolongation of survival times in terminal human cancer. *Proceeding of the National Academy of Sciences of the USA*, 75 :4538–4542, 1978.
- [2] P. Dagnelie. *Thies et modes statistiques. Applications agronomiques. II Les modes de l'infnce statistique*. Editions L. Duculot, Gembloux, 1970.
- [3] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski, editors. *A handbook of small data sets*. Chapman & Hall, London, 1984.
- [4] P. Sprent. *Pratique des statistiques non paramiques*. Editions INRA, Techniques et Pratiques, Paris, 1992.