Fiches de Biostatistique

3 - Pratique des tests élémentaires

D. Chessel & A.B. Dufour

Résumé

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.

Plan

1. Comparaison de la moyenne de deux échantillons

1.1 Le test t de comparaison de moyennes

Les exemples utilisés proviennent de l'ouvrage « Small data sets »[1].

Situation 1 - La variable mesurée est la longueur de la mâchoire inférieure (en mm) de 10 chacals mâles et 10 chacals femelles (*Canis Aureus*) conservées au British Museum[2]. La variable mesurée diffère t'elle entre les sexes dans cette espèce ?

```
mâles      120  107  110  116  114  111  113  117  114  112
femelles     110  111  107  108  110  105  107  106  111  111
```

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]. Cette variable dépend t'elle du type de cancer ?

```
Estomac       124  42   25    45 412  51    1112  46    103  876 146  340  396
Poumon       1235  24 1581  1166  40  727      3808 791  1804 3460 719
```

Test de comparaison de moyennes. Modèle gaussien (Rappel du cours de Deug 2).
$x_1, x_2, ..., x_{n_1}$ est un échantillon aléatoire simple d'une loi normale de moyenne μ_1 et de variance σ^2. $y_1, y_2, ..., y_{n_2}$ est un échantillon aléatoire simple d'une loi normale de moyenne μ_2 et de même variance σ^2.

$\widehat{\mu_1} = m_1 = \frac{x_1 + x_2 + ... + x_{n_1}}{n_1}$ est l'estimateur au maximum de vraisemblance de μ_1.

$\widehat{\mu_2} = m_2 = \frac{y_1 + y_2 + ... + y_{n_2}}{n_2}$ est l'estimateur au maximum de vraisemblance de μ_2.

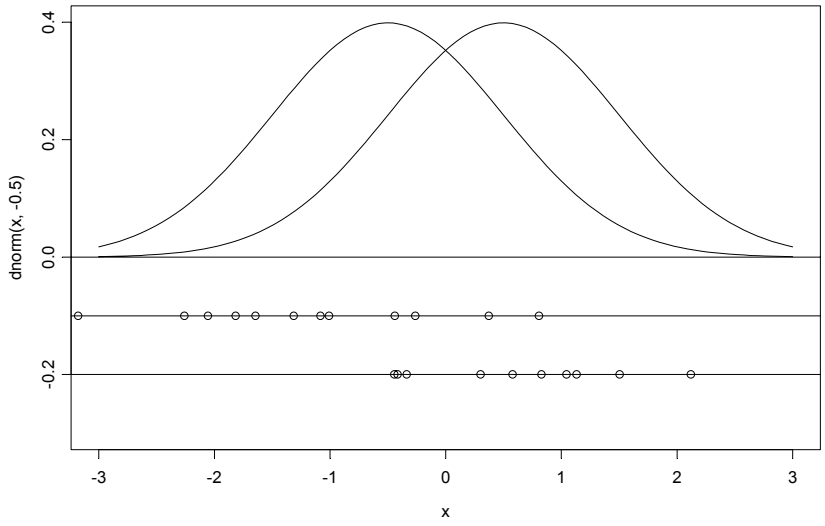
$\widehat{\sigma^2} = \frac{\sum_{=1}^{n_1}(x_i - m_1)^2 + \sum_{=1}^{n_2}(y_i - m_2)^2}{n_1 + n_2 - 2}$ est l'estimateur au maximum de vraisemblance de σ^2.

La différence des deux moyennes est une variable aléatoire. La valeur observée est $m_1 - m_2$. Sous l'hypothèse supplémentaire $\mu_1 = \mu_2 = \mu$, la moyenne de cette variable est nulle.

Sa variance est $\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Elle est estimée par $\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. La variable normalisée définie par :

$$t = \frac{m_1 - m_2}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

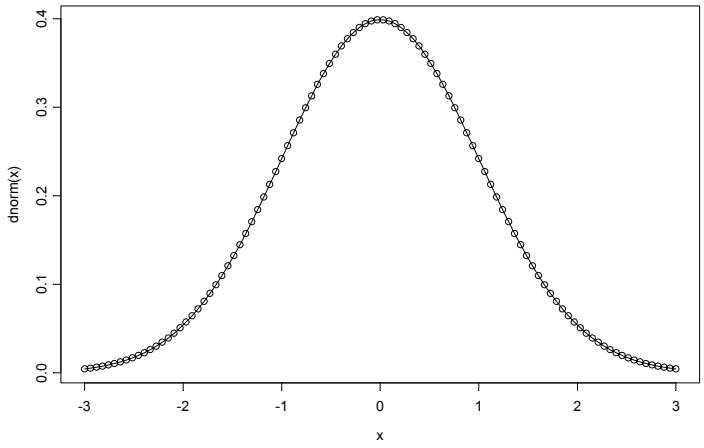
suit une loi T de Student à $n_1 + n_2 - 2$ degrés de liberté.




```
x <- seq(-3, 3, le = 100)
par(mfrow=c(1,1))
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))
points(y2, rep(-0.2,10))
```

La loi de la différence des moyennes est :

```
plot(x, dt(x,20), type = "l")
plot(x, dnorm(x), type = "o")
```



Si l'alternative est $\mu_1 > \mu_2$, les valeurs positives de $m_1 - m_2$ 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 $m_1 - m_2$ 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 $m_1 - m_2$ sont attendues et on rejette l'hypothèse avec le risque $2P(T > |t|)$.

Appliquée à la situation 1, cette procédure donne :

```
y1 <- c(120,107,110,116,114,111,113,117,114,112)
y2 <- c(110,111,107,108,110,105,107,106,111,111)
m1 <- mean(y1)
m2 <- mean(y2)
m1
[1] 113.4
m2
[1] 108.6
v <- (9*var(y1) + 9*var(y2))/18
v
[1] 9.489
t <- (m1-m2)/sqrt(v*(1/10 + 1/10))
t
[1] 3.484

tt0 <- t.test(y1, y2, var.eq = T)
tt0

          Two Sample t-test

data:  y1 and y2
t = 3.484, df = 18, p-value = 0.002647
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.906 7.694
sample estimates:
mean of x mean of y
    113.4     108.6
```

La moyenne des mâles est de 113.4 mm. La moyenne des femelles est de 108.6 mm. La différence normalisée est de 3.48 pour 18 degrés de liberté. La probabilité d'avoir une différence inférieure à -3.48 ou supérieure à 3.48 est :

```
1-pt(3.48,18)
[1] 0.001336
```

La probabilité de la zone de rejet est de 0.0026 et l'hypothèse nulle est rejetée au risque de première espèce de 3 pour 1000.

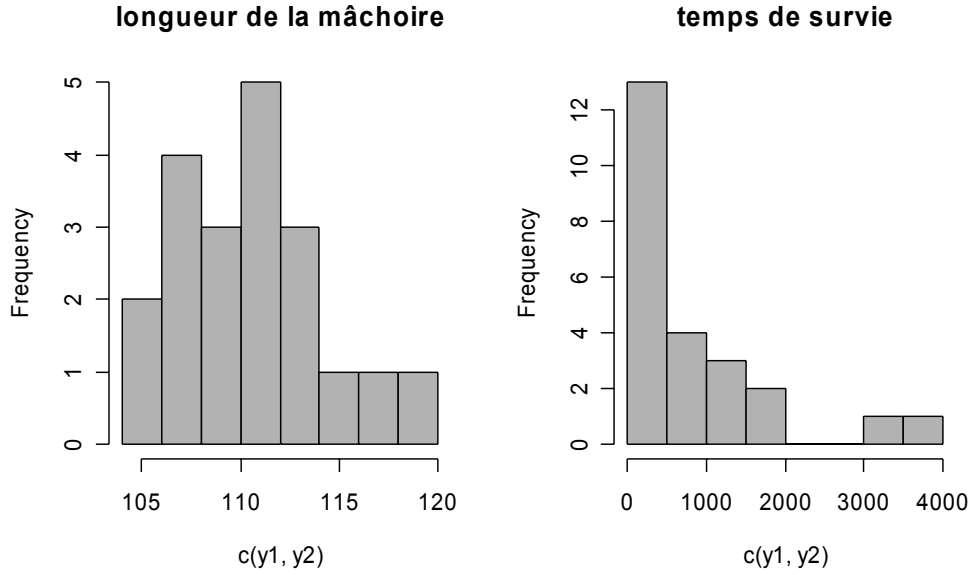
Appliquée à la situation 2, cette procédure donne :

```
y1 <- c(124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340, 396)
y2 <- c(1235, 24, 1581, 1166, 40, 727, 3808, 791, 1804, 3460, 719)
tt0 <- t.test(y1, y2, var.eq = T)
tt0

          Two Sample t-test

data:  y1 and y2
t = -3.101, df = 22, p-value = 0.005209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1852.1  -367.7
sample estimates:
mean of x mean of y
      286      1396
```

1.2 Le test de Wilcoxon (Mann-Whitney)

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

```
y1 <- c(120,107,110,116,114,111,113,117,114,112)
y2 <- c(110,111,107,108,110,105,107,106,111,111)
hist(c(y1,y2), nclass = 8, col=grey(0.7), main="longueur de la mâchoire")
y1 <- c(124,42,25,45,412,51,1112,46,103,876,146,340,396)
y2 <- c(1235,24,1581,1166,40,727,3808,791,1804,3460,719)
hist(c(y1,y2), nclass = 8, col=grey(0.7), main="temps de survie")
```



La première distribution est relativement symétrique, la deuxième ne l'est pas du tout. Dans l'ensemble des hypothèses nécessaires au test t, 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 simple. On réunit les deux échantillons :

```
z1 <- c(124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340, 396)
z2 <- c(1235, 24, 1581, 1166, 40, 727, 3808, 791, 1804, 3460, 719)
n1 <- length(z1)   # Il y a 13 individus dans le premier échantillon
n2 <- length(z2)   # 1l y a 11 individus dans le second échantillon
ztot <- c(z1, z2)
ztot
 [1]  124   42   25   45  412   51 1112   46  103  876  146  340  396 1235   24
[16] 1581 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. La plus petite donnée est située en 15ème position. Elle vaut 24. Son rang est 1. La plus grande donnée est située en 22ème position. Elle vaut 3460. Son rang est 23 :

```
rtot <- rank(ztot)
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. On appelle *m* l'effectif du premier groupe et *n* l'effectif total. Dans l'espace $\binom{n}{m}$, si $m \geq 10$ et $n - m \geq 10$ la somme des rangs *SR*, appelée statistique de Wilcoxon, suit approximativement une loi normale de moyenne et variance :

$$E(SR) = \frac{m(n+1)}{2} \qquad V(SR) = \frac{m(n-m)(n+1)}{12}$$

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.

Les logiciels utilisent la statistique *U* de Mann-Whitney : $U = SR - \frac{m(m+1)}{2}$. D'après ce qui précède, celle-ci suit approximativement une loi normale de moyenne et variance :

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

```
sr <- sum(r1)    # 122 est la somme des rangs de l'échantillon 1
esr <- (n1 * (n1 + n2 + 1))/2
esr
[1] 162.5
vsr <- (n1 * n2 * (n1 + n2 + 1))/12
vsr
[1] 297.9
t0 <- (sr - esr)/sqrt(vsr)
t0
[1] -2.346
2 * (pnorm(t0))
[1] 0.01895

u <- sr - (n1 * (n1+1))/2 #31 est la valeur de la statistique de Mann-Whitney
eu <- (n1 * n2)/2
eu
[1] 71.5
vu <- (n1 * n2 * (n1 + n2 + 1))/12
vu
[1] 297.9167
t1 <- (u - eu)/sqrt(vu)
t1
[1] -2.34643
2 * (pnorm(t1))
[1] 0.01895422

wilcox.test(z1, z2, exact = F, correct = F)

        Wilcoxon rank sum test

data:  z1 and z2
W = 31, p-value = 0.01895
alternative hypothesis: true mu is not equal to 0
```

Le logiciel donne également la valeur exacte :

```
wilcox.test(z1, z2, exact = T)

        Wilcoxon rank sum test

data:  z1 and z2
W = 31, p-value = 0.01836
alternative hypothesis: true mu is not equal to 0
```

L'approximation est très justifiée. On refera les calculs pour l'autre exemple sur la longueur des mâchoires des chacals..

```
y1 <- c(120, 107, 110, 116, 114, 111, 113, 117, 114, 112)
y2 <- c(110, 111, 107, 108, 110, 105, 107, 106, 111, 111)

n1 <- length(y1)      # le premier échantillon a 10 valeurs
n2 <- length(y2)      # le second échantillon a 10 valeurs

ytot <- c(y1, y2)     # les échantillons sont assemblés
120 107 110 116 114 111 113 117 114 112 110 111 107 108 110 105 107 106 111 111

rtot <- rank(ytot)
rtot # l'échantillon total est rangé. Noter les ex-aequos
 [1] 20.0  4.0  8.0 18.0 16.5 11.5 15.0 19.0 16.5 14.0  8.0 11.5  4.0  6.0  8.0
[16]  1.0  4.0  2.0 11.5 11.5

r1 <- rtot[1:n1]
r1    # On récupère les rangs du premier échantillon
 [1] 20.0  4.0  8.0 18.0 16.5 11.5 15.0 19.0 16.5 14.0

sr <- sum(r1)    # La somme des rangs est 142.5
u <- sr - (n1 * (n1+1))/2  # La valeur de la statistique de Mann-Whitney est 87.5

eu <- (n1 * n2)/2  # Moyenne attendue 50
vu <- (n1 * n2 * (n1 + n2 + 1))/12    # Variance attendue 175

t1 <- (u - eu)/sqrt(vu)            # Valeur normalisée de la statistique 2.834734
2 * (1 - pnorm(t1))                   # Probabilité de la zone de rejet 0.004586

wilcox.test(y1, y2, correct = F)

        Wilcoxon rank sum test

data:  y1 and y2
W = 87.5, p-value = 0.004301
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test.default(y1, y2, correct = F)

wilcox.test(y1, y2, correct = T)

        Wilcoxon rank sum test with continuity correction

data:  y1 and y2
W = 87.5, p-value = 0.004845
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test.default(y1, y2, correct = T)
```

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 donne un rejet à 0.003 et le test de Wilcoxon un rejet à 0.005. 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 et donne des résultats plus solides.

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 $(x_1, x_2, ...x_n)$ 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 $z_i = x_i - m$ est nulle contre l'hypothèse la moyenne des $z_i = x_i - m$ 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 7 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 $p = 1/2$.

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

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 7 parties sur 10, on a

$$P(X \geq x_{obs}) = P(X \geq 7) = 0.05469$$

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

$$P(X \geq x_{obs}) = P(X \geq 80) = 1.35 \times 10^{-10}$$

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 :

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

Pour 14 sur 20, on a exactement

$$P(X \geq x_{obs}) = P(X \geq 14) = 0.021$$

De manière approchée :

$$P(Y \geq y_{obs}) = P\left(Y \geq \frac{14-10}{\sqrt{5}}\right) = 0.036$$

Application : dans 5 villes des USA, le nombre de crimes pour 100 000 habitants était [4]:

```
          1     2     3     4     5
1960 10.1  10.6   8.2   4.9  11.5
1970  20.4  22.1  10.2   9.8  13.7
```

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. Dans 5 nouvelles villes, on a :

```
          6     7     8     9     10
1960 17.3  12.4  11.1   8.6   10.0
1970  20.4  22.1  10.2   9.8   13.7
```

On en est à 9 fois sur 10. Le seuil devient 0.0009766. Dans 5 nouvelles villes, on a :

```
          11     12     13     14     15
1960    9.1    7.9    4.5    8.1   17.7
1970   16.2    8.2   12.6   17.8   13.1
```

On passe à 13 fois sur 15 et un seuil de 0.0004883. 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
1960   7.9  4.5  8.1 17.7 11.0 10.8 12.5  8.9  4.4  6.4  3.8 14.2  6.6  6.2  3.3
1970    8.2 12.6 17.8 13.1 15.6 14.7 12.6  7.9 11.2 14.9 10.5 15.3 11.4  5.5  6.6
```

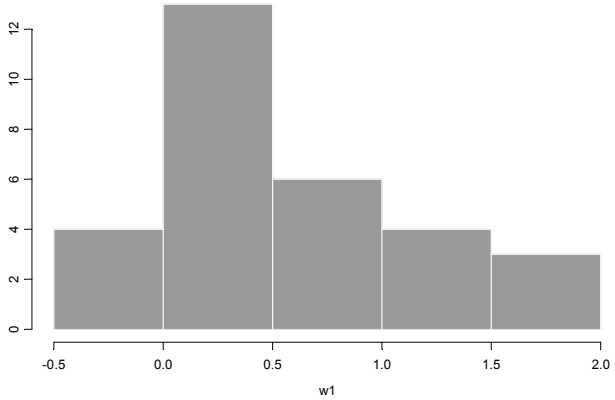
Remarque. L'ensemble des données est accessible de la manière suivante :

```
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%?

```
w0 <- (crim2 - crim1)/crim1
 [1]  1.01980  1.08491  0.24390  1.00000  0.19130  0.42775  0.24194  0.14414
 [9]  0.54651  0.84000 -0.11364  0.07692  0.19355  0.44444  0.78022  0.03797
[17]  1.80000  1.19753 -0.25989  0.41818  0.36111  0.00800 -0.11236  1.54545
[25]  1.32812  1.76316  0.07746  0.72727 -0.11290  1.00000
w1 <-  w0[order(w0)]
 [1] -0.25989 -0.11364 -0.11290 -0.11236  0.00800  0.03797  0.07692  0.07746
 [9]  0.14414  0.19130  0.19355  0.24194  0.24390  0.36111  0.41818  0.42775
[17]  0.44444  0.54651  0.72727  0.78022  0.84000  1.00000  1.00000  1.01980
[25]  1.08491  1.19753  1.32812  1.54545  1.76316  1.80000
```

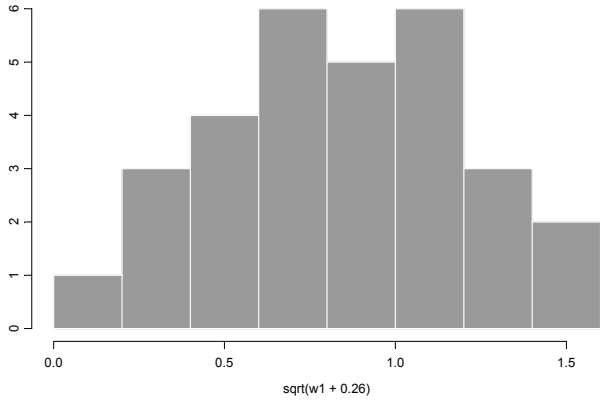


La symétrie n'est pas bonne et on ne peut pas aisément raisonner « une chance sur deux » d'être au dessus ou au dessous de la moyenne.

```
mean(w1)
[1] 0.5634 # La moyenne de l'accroissement est de 56%
median(w1)
[1] 0.423 # L'accroissement est <42% ou >42% avec « une chance sur deux »

sqrt(w1)
 [1]     NaN     NaN     NaN     NaN 0.08944 0.19487 0.27735 0.27832 0.37966
[10] 0.43738 0.43994 0.49187 0.49386 0.60093 0.64667 0.65402 0.66667 0.73926
[19] 0.85280 0.88330 0.91652 1.00000 1.00000 1.00985 1.04159 1.09432 1.15244
[28] 1.24316 1.32784 1.34164
Warning message:
NaNs produced in: sqrt(w1)

sqrt(w1 + 0.26)
 [1] 0.01063 0.38258 0.38353 0.38424 0.51769 0.54587 0.58045 0.58092 0.63572
[10] 0.67179 0.67346 0.70847 0.70986 0.78811 0.82352 0.82930 0.83931 0.89806
[19] 0.99362 1.01991 1.04881 1.12250 1.12250 1.13128 1.15970 1.20728 1.26021
[28] 1.34367 1.42238 1.43527
```

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

```
hist(sqrt(w1+0.26), col=grey(0.7))
```




```
median(sqrt(w1+0.26)) # 0.8264
mean(sqrt(w1+0.26)) # 0.841
```

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 de la variable transformée est :

```
> sqrt(0.5+0.26)
[1] 0.8718
```

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(1.26)
[1] 1.122
```

21 à gauche et 9 à droite. Bizarre ?

```
dbinom(0:30,30,0.5)
 [1] 9.313e-10 2.794e-08 4.051e-07 3.781e-06 2.552e-05 1.327e-04 5.530e-04
 [8] 1.896e-03 5.451e-03 1.332e-02 2.798e-02 5.088e-02 8.055e-02 1.115e-01
[15] 1.354e-01 1.445e-01 1.354e-01 1.115e-01 8.055e-02 5.088e-02 2.798e-02
[22] 1.332e-02 5.451e-03 1.896e-03 5.530e-04 1.327e-04 2.552e-05 3.781e-06
[29] 4.051e-07 2.794e-08 9.313e-10

pbinom(0:30,30,0.5)
 [1] 9.313e-10 2.887e-08 4.340e-07 4.215e-06 2.974e-05 1.625e-04 7.155e-04
 [8] 2.611e-03 8.062e-03 2.139e-02 4.937e-02 1.002e-01 1.808e-01 2.923e-01
[15] 4.278e-01 5.722e-01 7.077e-01 8.192e-01 8.998e-01 9.506e-01 9.786e-01
[22] 9.919e-01 9.974e-01 9.993e-01 9.998e-01 1.000e+00 1.000e+00 1.000e+00
[29] 1.000e+00 1.000e+00 1.000e+00

2 * (1 - sum(pbinom(21,30,0.5)))
[1] 0.01612
```

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] 0.1935
```

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] 0.7802
```

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

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ est un échantillon formé de *n couples*. (x_i, y_i) est un échantillon aléatoire simple d'une loi normale de moyenne μ_i et de variance σ^2. $z_i = x_i - y_i$ est la réalisation d'une variable aléatoire de moyenne 0 et de variance $2\sigma^2$. On veut comparer la moyenne de Z à 0.

Exemple [5] : on a mesuré la hauteur (en mètres) de 12 arbres selon deux méthodes différentes, avant et après la coupe de l'arbre.

```
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)
debout
 [1] 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
 [1] 21.7 26.3 26.8 28.1 26.2 27.3 29.5 32.0 30.9 32.3 32.3 31.7
dif <- debout - abattu
dif
 [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
```

Exercice : quelle information est elle fournie par ?

```
pbinom(0:12,12,1/2)
 [1] 0.0002441 0.0031738 0.0192871 0.0729980 0.1938477 0.3872070 0.6127930
 [8] 0.8061523 0.9270020 0.9807129 0.9968262 0.9997559 1.0000000

dbinom(0:12,12,1/2)
 [1] 0.0002441 0.0029297 0.0161133 0.0537109 0.1208496 0.1933594 0.2255859
 [8] 0.1933594 0.1208496 0.0537109 0.0161133 0.0029297 0.0002441
```

$\hat{\mu} = m = \frac{z_1 + z_2 + ... + z_n}{n}$ est l'estimateur au maximum de vraisemblance de la moyenne de Z. $\widehat{2\sigma^2} = \frac{\sum_{=1}^{n} (z_i - m)^2}{n-1}$ est l'estimateur au maximum de vraisemblance de $2\sigma^2$.

La moyenne des différences est une variable aléatoire. La valeur observée est $m_1 - m_2$. Sous l'hypothèse supplémentaire $\mu = 0$, la moyenne de cette variable est nulle.

Sa variance est $\frac{2\sigma^2}{n}$. Elle est estimée par $\frac{\widehat{2\sigma^2}}{n}$. La variable normalisée :

$$t = \frac{m_1 - m_2}{\sqrt{\frac{\sum_{=1}^{n} (z_i - m)^2}{n(n-1)}}}$$

suit une loi T de Student à $n-1$ degrés de liberté.

```
mean(debout-abattu)
[1] -1.075

var(debout-abattu)/12
[1] 0.1105

-1.075/sqrt(0.1105)
[1] -3.234

t.test(debout,abattu,paired=T)
```

```
   Paired t-Test

data:  debout and abattu
t = -3.234, df = 11, p-value = 0.008
alternative hypothesis: true mean of differences is not equal to 0
95 percent confidence interval:
 -1.8066 -0.3434
sample estimates:
 mean of x - y
       -1.075
```

Tout test fournit un intervalle de confiance. Discussion.

2.4 Le test de Wilcoxon sur données appariées

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ 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 ([6] 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] 0.1719
```

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 :

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

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 :

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

```
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  # somme totale des rangs 55 soit 8 + 47

moy <- (10 * 11)/4
var <- (10 * 11 * 21)/24

usom <- (som - moy)/sqrt(var)    # test sur la somme des rangs de signe -
usom
[1] -1.987624
pnorm(usom)
[1] 0.02343
2 * pnorm(usom)
[1] 0.04685

ucompsom <- (compsom - moy)/sqrt(var)    # test sur la somme des rangs de signe +
ucompsom
[1] 1.987624
1 - pnorm(ucompsom)
[1] 0.02343
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 mu 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 mu 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 mu 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 = 2.471, df = 9, p-value = 0.0355
alternative hypothesis: true mean of differences is not equal to 0
95 percent confidence interval:
```

```
 0.03802 0.86198
sample estimates:
 mean of x - y
          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 fausse 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) # Attention numéros de 1 à 11 pour les valeurs de 0 à 10
 [1] 0.0009766 0.0097656 0.0439453 0.1171875 0.2050781 0.2460938 0.2050781
 [8] 0.1171875 0.0439453 0.0097656 0.0009766
```

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

```
pbinom(0:10,10,0.5)
 [1] 0.0009766 0.0107422 0.0546875 0.1718750 0.3769531 0.6230469 0.8281250
 [8] 0.9453125 0.9892578 0.9990234 1.0000000
```

La probabilité d'en gagner au plus 3 est 0.1718.

```
1-pbinom(0:10,10,0.5)
 [1] 0.9990234 0.9892578 0.9453125 0.8281250 0.6230469 0.3769531 0.1718750
 [8] 0.0546875 0.0107422 0.0009766 0.0000000
```

La probabilité d'en gagner 4 ou plus est de 0.8282. Ils décident de faire un test au risque α=5%. La probabilité d'en gagner 8 ou plus vaut 0.055. 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 fausse ». Et elle est bien fausse. 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] 1.000e-010 9.000e-009 3.645e-007 8.748e-006 1.378e-004 1.488e-003 1.116e-002
 [8] 5.740e-002 1.937e-001 3.874e-001 3.487e-001
```

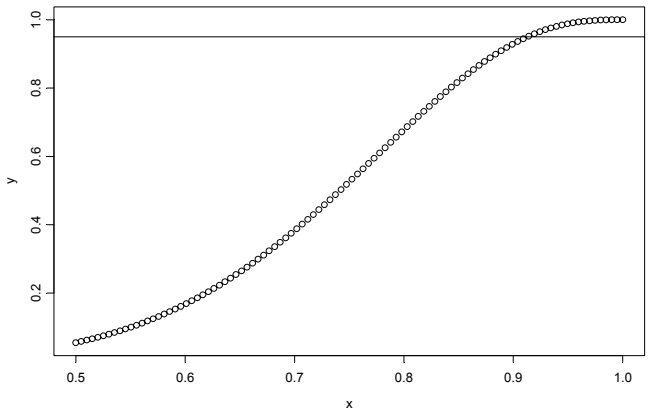
```
pbinom(0:10,10,0.9)
 [1] 1.000e-010 9.100e-009 3.736e-007 9.122e-006 1.469e-004 1.635e-003 1.280e-002
 [8] 7.019e-002 2.639e-001 6.513e-001 1.000e+000
1-pbinom(0:10,10,0.9)
 [1] 1.0000 1.0000 1.0000 1.0000 0.9999 0.9984 0.9872 0.9298 0.7361 0.3487 0.0000
```

Dans 35% des cas, A gagne 10 fois et le test est très significatif. Dans 39% des cas, A gagne 9 fois et le test est très significatif. Dans 19% des cas, A gagne 8 fois et le test est significatif. Dans 6 % des cas A, gagne 7 fois et le test n'est pas significatif. Dans 1% des cas, A gagne 6 fois et il l'est encore moins. Quand le test n'est pas significatif, B dit « H_0 est vraie » et se trompe. Quelle probabilité avait-il de se tromper ?

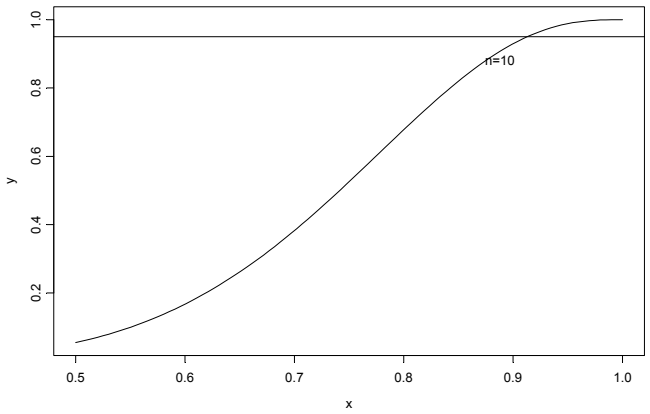
```
pbinom(7,10,0.9)
[1] 0.07019
```

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 :

```
x = seq(0.5, 1, le = 100)
y = 1 - pbinom(7,10,x)
plot(x,y)
abline(h = 0.95)
```



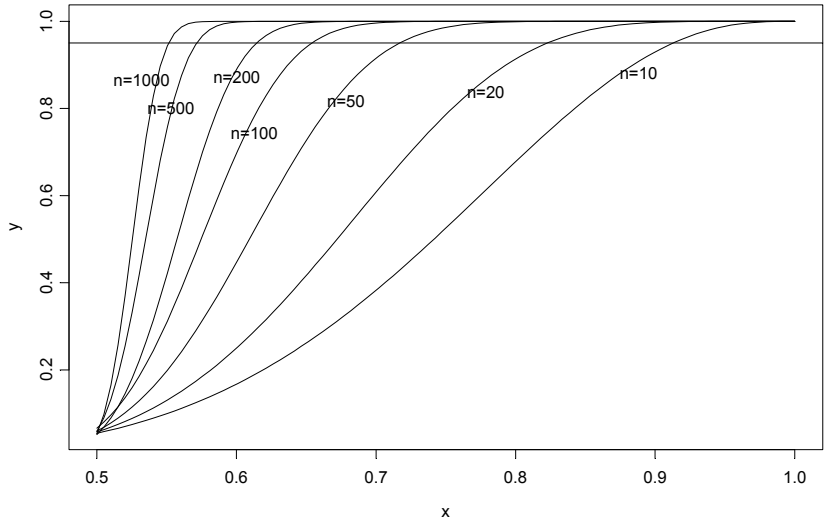
Ceci veut dire que, si le test est significatif, A peut dire qu'il est le plus fort avec une probabilité de se tromper de α=5% et que, s'il ne l'est pas, B ne peut pas dire qu'ils se valent. Il pourra juste dire que la puissance du test est suffisante pour dire que $p \leq 0.9$ mais pas que $p = 0.5$. Le graphe ci-dessus est celui de la puissance du test « une chance sur 2 » quand *n* vaut 10 et α=5%.



La puissance du test est une fonction de *n*. Il faut recommencer le raisonnement en entier pour $n = 20$. Ceci s'automatise par :

```
y = 1 - pbinom(qbinom(0.95,10,0.5) - 1, 10, x)
plot(x, y, type = "n")
lines(x,y)
text(locator(1), "n = 10")
abline(h=0.95)
```

On peut maintenant refaire cette opération pour 20, 50, 100, 200, 500 et 1000.



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.

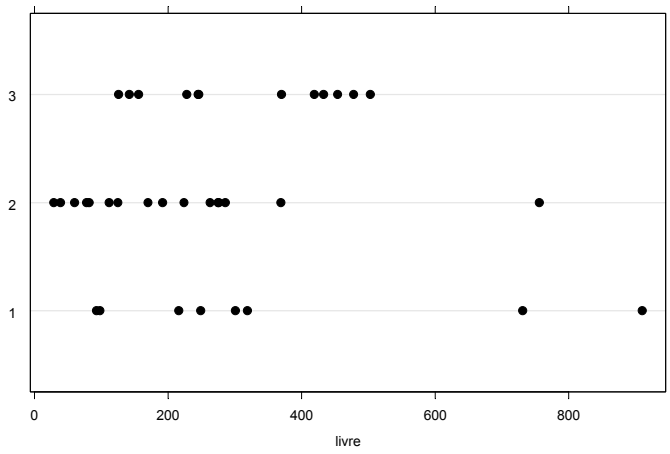
4. Comparaison de *s* échantillons

4.1 Test de Kruskal et Wallis

Dans la bibliothèque de Peter Sprent [6], il y a des livres de voyage, des ouvrages généraux et des livres de statistiques . On a trois échantillons.
Le premier est celui des livres de voyage. Ils ont respectivement 93, 98, 216, 249, 301, 319, 731 et 910 pages. 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. 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
 [1]  93  98 216 249 301 319 731 910  29  39  60  78  82 112 125 170 192 224 263
[20] 275 276 286 369 756 126 142 156 228 245 246 370 419 433 454 478 503
> group
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
```

Le « bon » dessin est :



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
[27] 12 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 :

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

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

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

```
sum(rank(livre))
[1] 666
36*37/2
[1] 666
sum(rank(livre)^2)
[1] 16206
36*37*(2*36+1)/6
[1] 16206
```

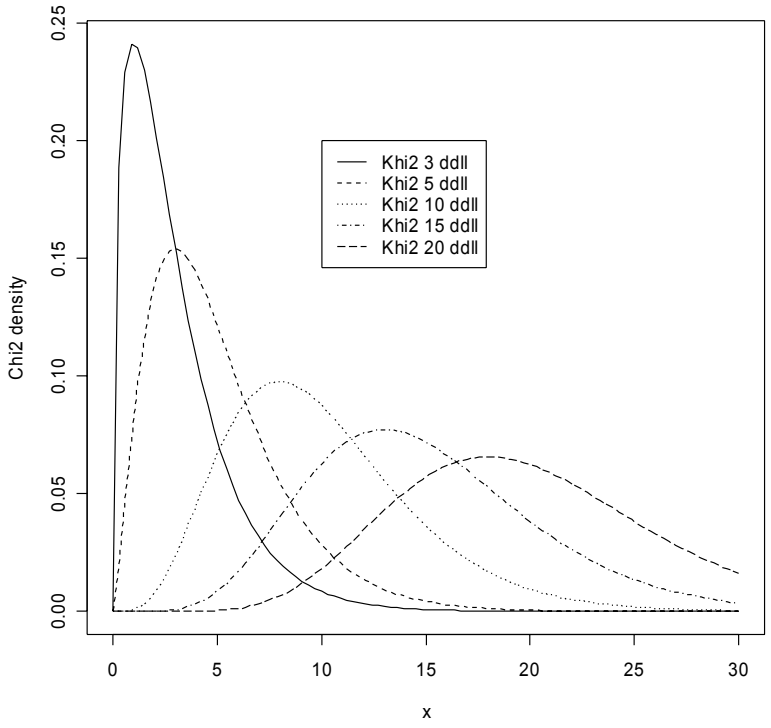
Si SR_j est la somme des rangs de l'échantillon j , la variable T définie par :

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

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
 (166^2/8)+(227^2/16)+(270^2/12)
[1] 12740
 (169^2/8)+(227^2/16)+(270^2/12)
[1] 12866
12866*12/(36*37)-3*37
[1] 4.91
```

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 ddll","Khi2 5 ddll","Khi2 10 ddll","Khi2 15 ddll","Khi2 20 ddll")
legend(10,0.2,l0, lty=1:5)
```

Les quantiles sont disponibles dans la table du Khi2. Conclure.

```
> kruskal.test(livre,groupe)

        Kruskal-Wallis rank sum test

data:  livre and groupe
Kruskal-Wallis chi-squared = 4.9071, df = 2, p-value = 0.08599
```

4.2 Comparer les variances

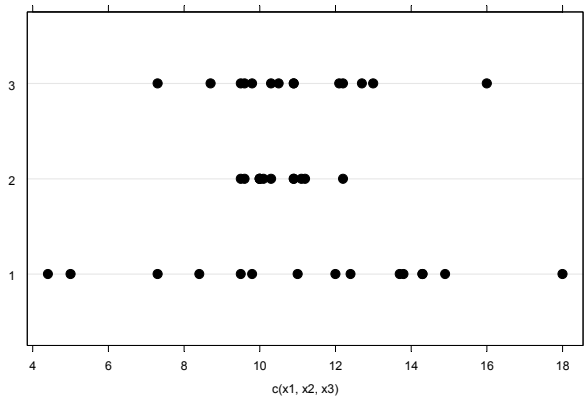
Dans trois groupes de TD, les notes de contrôle continu sont :

```
> x1
 [1] 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
 [1] 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
 [1] 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.

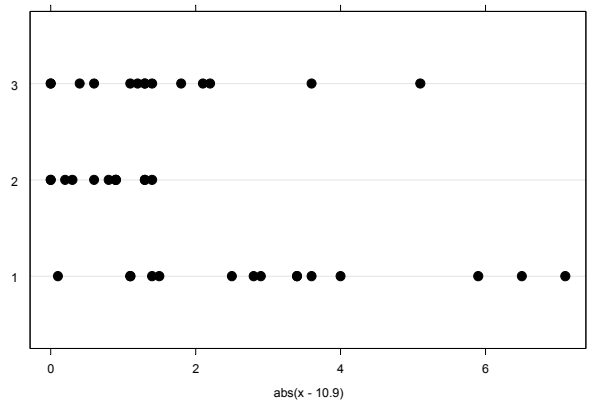
```
mean(x1)
[1] 11.25
mean(x2)
[1] 10.48
mean(x3)
[1] 10.96
gtd
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
[40] 3 3

kruskal.test(c(x1, x2, x3), gtd)

   Kruskal-Wallis rank sum test

data:  c(x1, x2, x3) and gtd
Kruskal-Wallis chi-square = 0.7204, df = 2, p-value = 0.6975
```

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



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

```
median(x)
[1] 10.9

kruskal.test(abs(x - 10.9), gtd)

   Kruskal-Wallis rank sum test

data:  abs(x - 10.9) and gtd
Kruskal-Wallis chi-square = 14.31, df = 2, p-value = 0.0008

dotplot(gtd ~ abs(x - 10.9), cex = 1.5)
```



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 :

```
739A684215      9  8  2  7  10 5  1  6  3  4
487196235A      4  7  8  1  9  6  3  2  5  10
76832149A5      6  5  4  7  10 2  1  3  8  9
673529A814      9  5  3  10 4  1  2  8  6  7
1235847A69      1  2  3  6  4  9  7  5  10 8
156327849A      1  5  4  8  2  3  6  7  9  10
26A5738149      8  1  6  9  4  2  5  7  10 3
6795328A14      9  6  5  10 4  1  2  7  3  8
76592138A4      6  5  7  10 3  2  1  8  4  9
5763914A28      6  9  4  7  1  3  2  10 5  8
6752314A98      6  4  5  7  3  1  2  10 9  8
756932A841      10 6  5  9  2  3  1  8  4  7
673521489A      6  5  3  7  4  1  2  8  9  10
72A5316894      6  2  5  10 4  7  1  8  9  3
67A1253489      4  5  7  8  6  1  2  9  10 3
6A73215489      6  5  4  8  7  1  3  9  10 2
37A2659814      9  4  1  10 6  5  2  8  7  3
879653124A      7  8  6  9  5  4  2  1  3  10
573869412A      8  9  3  7  1  5  2  4  6  10
463589721A      9  8  3  1  4  2  7  5  6  10
687532491A      9  6  5  7  4  1  3  2  8  10
675984312A      8  9  7  6  3  1  2  5  4  10
679583412A      8  9  6  7  4  1  2  5  3  10
637152849A      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, A-Bon Jovi. Les comparaisons de rangs portent sur les données de droite.

```
rock = read.table("http://pbil.univ-lyon1.fr/R/donnees/rock.txt")
matrock = as.matrix(rock)
neorock = matrix(rep(0,240), nrow = 24)
for (i in 1:24) (neorock[i,] = order(matrock[i,]))
```

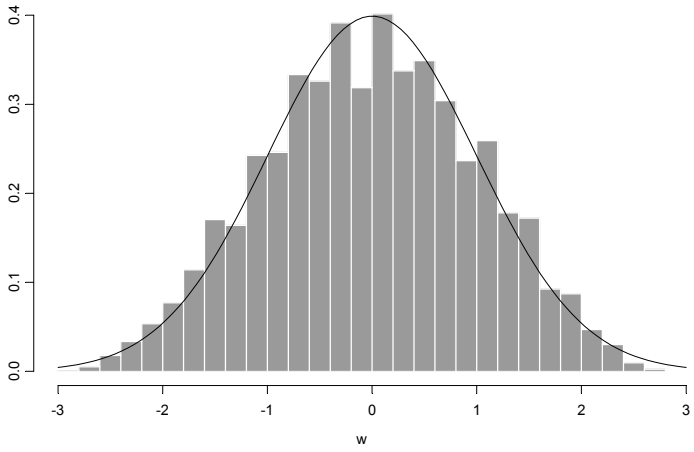
5.1 Corrélation de rang

```
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] 0.0303
```

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

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

```
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

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 :

```
554334721354454857854672 8
482415278816378578141544 6
261162155437226216212125 1
675826886665663681767416 7
143271643128111322628281 2
328658334781587443386867 3
716547417573832735473738 5
837783562242745164535353 4
```

Les dégustateurs sont de 5 catégories professionnelles: [1,5] Œnologues, [6,10] Restaurateurs, [11,15] Négociants, [16,20] Viticulteurs, [21,25] Organisateurs 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.

Les calculs se font avec un ordinateur. Les exercices sont destinés à comprendre les idées. Tous les documents personnels et les polycopes sont admis à l'examen.

Les calculs de ces fiches sont reproductibles avec le logiciel **R** développé aux AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. Diffusion libre sur Internet à **http://www.ci.tuwien.ac.at/R**.

7. Annexe : Tables statistiques

7.1 La Loi binomiale « une chance sur deux »

A la ligne étiquetée *i* dans la colonne *n* on trouve $1000 \times P(X \leq i)$ pour une loi binomiale de paramètres *n* et 1/2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	500	250	125	62	31	16	8	4	2	1	0	0	0	0
1	1000	750	500	312	187	109	63	35	20	11	6	3	2	1
2	0	1000	875	688	500	344	227	145	90	55	33	19	11	6
3	0	0	1000	938	813	656	500	363	254	172	113	73	46	29
4	0	0	0	1000	969	891	773	637	500	377	274	194	133	90
5	0	0	0	0	1000	984	938	855	746	623	500	387	291	212
6	0	0	0	0	0	1000	992	965	910	828	726	613	500	395
7	0	0	0	0	0	0	1000	996	980	945	887	806	709	605
8	0	0	0	0	0	0	0	1000	998	989	967	927	867	788
9	0	0	0	0	0	0	0	0	1000	999	994	981	954	910
10	0	0	0	0	0	0	0	0	0	1000	1000	997	989	971
11	0	0	0	0	0	0	0	0	0	0	1000	1000	998	994
12	0	0	0	0	0	0	0	0	0	0	0	1000	1000	999
13	0	0	0	0	0	0	0	0	0	0	0	0	1000	1000
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1000

	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	4	2	1	1	0	0	0	0	0	0	0	0	0	0
3	18	11	6	4	2	1	1	0	0	0	0	0	0	0
4	59	38	25	15	10	6	4	2	1	1	0	0	0	0
5	151	105	72	48	32	21	13	8	5	3	2	1	1	0
6	304	227	166	119	84	58	39	26	17	11	7	5	3	2
7	500	402	315	240	180	132	95	67	47	32	22	14	10	6
8	696	598	500	407	324	252	192	143	105	76	54	38	26	18
9	849	773	685	593	500	412	332	262	202	154	115	84	61	44
10	941	895	834	760	676	588	500	416	339	271	212	163	124	92
11	982	962	928	881	820	748	668	584	500	419	345	279	221	172
12	996	989	975	952	916	868	808	738	661	581	500	423	351	286
13	1000	998	994	985	968	942	905	857	798	729	655	577	500	425
14	1000	1000	999	996	990	979	961	933	895	846	788	721	649	575
15	1000	1000	1000	999	998	994	987	974	953	924	885	837	779	714
16	0	1000	1000	1000	1000	999	996	992	983	968	946	916	876	828
17	0	0	1000	1000	1000	1000	999	998	995	989	978	962	939	908
18	0	0	0	1000	1000	1000	1000	1000	999	997	993	986	974	956
19	0	0	0	0	1000	1000	1000	1000	1000	999	998	995	990	982
20	0	0	0	0	0	1000	1000	1000	1000	1000	1000	999	997	994
21	0	0	0	0	0	0	1000	1000	1000	1000	1000	1000	999	998
22	0	0	0	0	0	0	0	1000	1000	1000	1000	1000	1000	1000
23	0	0	0	0	0	0	0	0	1000	1000	1000	1000	1000	1000
24	0	0	0	0	0	0	0	0	0	1000	1000	1000	1000	1000
25	0	0	0	0	0	0	0	0	0	0	1000	1000	1000	1000
26	0	0	0	0	0	0	0	0	0	0	0	1000	1000	1000
27	0	0	0	0	0	0	0	0	0	0	0	0	1000	1000
28	0	0	0	0	0	0	0	0	0	0	0	0	0	1000

7.2 La loi normale

En face de la valeur x on trouve $P(X \leq x)$ pour une loi normale de moyenne 0 et de variance 1.

-6.0	9.866e-010	**-4.0**	3.167e-005	**-2.0**	2.275e-002
-5.9	1.818e-009	**-3.9**	4.810e-005	**-1.9**	2.872e-002
-5.8	3.316e-009	**-3.8**	7.235e-005	**-1.8**	3.593e-002
-5.7	5.990e-009	**-3.7**	1.078e-004	**-1.7**	4.457e-002
-5.6	1.072e-008	**-3.6**	1.591e-004	**-1.6**	5.480e-002
-5.5	1.899e-008	**-3.5**	2.326e-004	**-1.5**	6.681e-002
-5.4	3.332e-008	**-3.4**	3.369e-004	**-1.4**	8.076e-002
-5.3	5.790e-008	**-3.3**	4.834e-004	**-1.3**	9.680e-002
-5.2	9.964e-008	**-3.2**	6.871e-004	**-1.2**	1.151e-001
-5.1	1.698e-007	**-3.1**	9.676e-004	**-1.1**	1.357e-001
-5.0	2.867e-007	**-3.0**	1.350e-003	**-1.0**	1.587e-001
-4.9	4.792e-007	**-2.9**	1.866e-003	**-0.9**	1.841e-001
-4.8	7.933e-007	**-2.8**	2.555e-003	**-0.8**	2.119e-001
-4.7	1.301e-006	**-2.7**	3.467e-003	**-0.7**	2.420e-001
-4.6	2.112e-006	**-2.6**	4.661e-003	**-0.6**	2.743e-001
-4.5	3.398e-006	**-2.5**	6.210e-003	**-0.5**	3.085e-001
-4.4	5.413e-006	**-2.4**	8.198e-003	**-0.4**	3.446e-001
-4.3	8.540e-006	**-2.3**	1.072e-002	**-0.3**	3.821e-001
-4.2	1.335e-005	**-2.2**	1.390e-002	**-0.2**	4.207e-001
-4.1	2.066e-005	**-2.1**	1.786e-002	**-0.1**	4.602e-001
				0.0	5.000e-001

En face de la valeur p on trouve la valeur x telle que $P(X \leq x) = p$ pour une loi normale de moyenne 0 et de variance 1.

1.000e-010	-6.361
5.000e-010	-6.109
1.000e-009	-5.998
5.000e-009	-5.731
0.00000001	-5.612
0.00000005	-5.327
0.0000001	-5.199
0.0000005	-4.897
0.000001	-4.753
0.000005	-4.417
0.00001	-4.265
0.00005	-3.891
0.0001	-3.719
0.0005	-3.291
0.001	**-3.090**
0.005	-2.576
0.010	-2.326
0.025	-1.960
0.050	**-1.645**
0.100	-1.282
0.5	0.000

7.3 La loi de Wilcoxon

Un échantillon de taille *m* est mélangé avec un échantillon de taille *n* pour former un échantillon de taille *m+n*. La somme des rangs des éléments du premier échantillon S_m est la statistique de Wilcoxon. Si les deux échantillons proviennent de la même population S_m suit la loi de Wilcoxon $W_{m,n}$. Noter que si S_n est la somme des rangs du deuxième échantillon, on a toujours $S_m + S_n = 1 + 2 + ... + (m+n) = \frac{(m+n)(m+n+1)}{2}$.

Si S_m suit la loi $W_{m,n}$ alors S_n suit la loi $W_{m,n}$.

Quantiles 0.001

valeurs $x_{m,n}$ telles que $P(W_{m,n} \leq x_{m,n}) = 0.001$. *m* en ligne, *n* en colonne

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	6	6	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7
4	10	10	10	10	10	10	10	11	11	11	12	12	12	13	13	14	14	14
5	15	15	15	15	15	16	17	17	18	18	19	19	20	21	21	22	23	23
6	21	21	21	21	22	23	24	25	26	26	27	28	29	30	31	32	33	34
7	28	28	28	29	30	31	32	34	35	36	37	38	39	40	42	43	44	45
8	36	36	37	38	39	41	42	43	45	46	48	49	51	52	54	55	57	58
9	45	45	47	48	49	51	53	54	56	58	60	61	63	65	67	69	71	72
10	55	56	57	59	61	62	64	66	68	70	73	75	77	79	81	83	85	88
11	66	67	69	71	73	75	77	79	82	84	87	89	91	94	96	99	101	104
12	78	79	81	83	86	88	91	93	96	99	102	104	107	110	113	116	119	121
13	91	93	95	97	100	103	106	109	112	115	118	121	124	127	130	134	137	140
14	105	107	109	112	115	118	121	125	128	131	135	138	142	145	149	152	156	160
15	120	122	125	128	131	135	138	142	145	149	153	157	161	164	168	172	176	180
16	136	139	142	145	148	152	156	160	164	168	172	176	180	185	189	193	197	202
17	154	156	159	163	167	171	175	179	183	188	192	197	201	206	211	215	220	224
18	172	175	178	182	186	190	195	199	204	209	214	218	223	228	233	238	243	248
19	191	194	198	202	206	211	216	220	225	231	236	241	246	251	257	262	268	273
20	211	214	218	223	227	232	237	243	248	253	259	265	270	276	281	287	293	299

Quantiles 0.005

valeurs $x_{m,n}$ telles que $P(W_{m,n} \leq x_{m,n}) = 0.005$. *m* en ligne, *n* en colonne

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	6	6	6	6	6	7	7	7	8	8	8	9	9	9	9	10	10
4	10	10	10	11	11	12	12	13	13	14	14	15	16	16	17	17	18	19
5	15	15	16	17	17	18	19	20	21	22	23	23	24	25	26	27	28	29
6	21	22	23	24	25	26	27	28	29	31	32	33	34	35	37	38	39	40
7	28	29	30	32	33	35	36	38	39	41	42	44	45	47	48	50	51	53
8	36	38	39	41	43	44	46	48	50	52	54	55	57	59	61	63	65	67
9	46	47	49	51	53	55	57	59	62	64	66	68	70	73	75	77	79	82
10	56	58	60	62	65	67	69	72	74	77	80	82	85	87	90	93	95	98
11	67	69	72	74	77	80	83	85	88	91	94	97	100	103	106	109	112	115
12	80	82	85	88	91	94	97	100	103	106	110	113	116	120	123	126	130	133
13	93	95	99	102	105	109	112	116	119	123	126	130	134	137	141	145	149	152
14	107	110	113	117	121	124	128	132	136	140	144	148	152	156	160	164	169	173
15	123	126	129	133	137	141	145	150	154	158	163	167	172	176	181	185	190	194
16	139	142	146	150	155	159	164	168	173	178	182	187	192	197	202	207	211	216
17	156	160	164	169	173	178	183	188	193	198	203	208	214	219	224	229	235	240
18	174	178	183	188	193	198	203	209	214	219	225	230	236	242	247	253	259	264
19	194	198	203	208	213	219	224	230	236	242	248	254	260	265	272	278	284	290
20	214	219	224	229	235	241	247	253	259	265	271	278	284	290	297	303	310	316

Quantiles 0.01

valeurs $x_{m,n}$ telles que $P\left(W_{m,n} \leq x_{m,n}\right) = 0.01$. *m* en ligne, *n* en colonne

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	6	6	6	7	7	8	8	8	9	9	9	10	10	11	11	11	12
4	10	10	11	12	12	13	14	14	15	16	16	17	18	18	19	20	20	21
5	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
6	21	23	24	25	26	28	29	30	31	33	34	35	37	38	40	41	42	44
7	29	30	32	33	35	36	38	40	41	43	45	46	48	50	52	53	55	57
8	37	39	41	43	44	46	48	50	52	54	57	59	61	63	65	67	69	71
9	47	49	51	53	55	57	60	62	64	67	69	72	74	77	79	82	84	86
10	57	59	62	64	67	69	72	75	78	80	83	86	89	92	94	97	100	103
11	68	71	74	76	79	82	85	89	92	95	98	101	104	108	111	114	117	120
12	81	84	87	90	93	96	100	103	107	110	114	117	121	125	128	132	135	139
13	94	97	101	104	108	112	115	119	123	127	131	135	139	143	147	151	155	159
14	108	112	116	119	123	128	132	136	140	144	149	153	157	162	166	171	175	179
15	124	128	132	136	140	145	149	154	158	163	168	172	177	182	187	191	196	201
16	140	144	149	153	158	163	168	173	178	183	188	193	198	203	208	213	219	224
17	158	162	167	172	177	182	187	192	198	203	209	214	220	225	231	236	242	247
18	176	181	186	191	196	202	208	213	219	225	231	237	242	248	254	260	266	272
19	195	200	206	211	217	223	229	235	241	247	254	260	266	273	279	285	292	298
20	216	221	227	233	239	245	251	258	264	271	278	284	291	298	304	311	318	325

Quantiles 0.025

valeurs $x_{m,n}$ telles que $P\left(W_{m,n} \leq x_{m,n}\right) = 0.025$. *m* en ligne, *n* en colonne

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	6	7	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15
4	10	11	12	13	14	15	15	16	17	18	19	20	21	22	22	23	24	25
5	16	17	18	19	21	22	23	24	25	27	28	29	30	31	33	34	35	36
6	23	24	25	27	28	30	32	33	35	36	38	39	41	43	44	46	47	49
7	30	32	34	35	37	39	41	43	45	47	49	51	53	55	57	59	61	63
8	39	41	43	45	47	50	52	54	56	59	61	63	66	68	71	73	75	78
9	48	50	53	56	58	61	63	66	69	72	74	77	80	83	85	88	91	94
10	59	61	64	67	70	73	76	79	82	85	89	92	95	98	101	104	108	111
11	70	73	76	80	83	86	90	93	97	100	104	107	111	114	118	122	125	129
12	83	86	90	93	97	101	105	108	112	116	120	124	128	132	136	140	144	148
13	96	100	104	108	112	116	120	125	129	133	137	142	146	151	155	159	164	168
14	111	115	119	123	128	132	137	142	146	151	156	161	165	170	175	180	184	189
15	126	131	135	140	145	150	155	160	165	170	175	180	185	191	196	201	206	211
16	143	148	152	158	163	168	174	179	184	190	196	201	207	212	218	223	229	235
17	160	165	171	176	182	188	193	199	205	211	217	223	229	235	241	247	253	259
18	179	184	190	196	202	208	214	220	227	233	239	246	252	258	265	271	278	284
19	198	204	210	216	223	229	236	243	249	256	263	269	276	283	290	297	304	310
20	219	225	231	238	245	252	259	266	273	280	287	294	301	309	316	323	330	338

Quantiles 0.05

valeurs $x_{m,n}$ telles que $P\left(W_{m,n} \leq x_{m,n}\right) = 0.05$. *m* en ligne, *n* en colonne

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	7	8	9	9	10	10	11	12	12	13	14	14	15	16	16	17	18
4	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29
5	17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41
6	24	25	27	29	30	32	34	36	38	39	41	43	45	47	48	50	52	54
7	31	33	35	37	40	42	44	46	48	50	53	55	57	59	62	64	66	68
8	40	42	45	47	50	52	55	57	60	63	65	68	70	73	76	78	81	84
9	50	52	55	58	61	64	67	70	73	76	79	82	85	88	91	94	97	100
10	60	63	67	70	73	76	80	83	87	90	93	97	100	104	107	111	114	118
11	72	75	79	83	86	90	94	98	101	105	109	113	117	121	124	128	132	136
12	84	88	92	96	100	105	109	113	117	121	126	130	134	139	143	147	151	156
13	98	102	107	111	116	120	125	129	134	139	143	148	153	157	162	167	172	176
14	113	117	122	127	132	137	142	147	152	157	162	167	172	177	183	188	193	198
15	128	133	139	144	149	154	160	165	171	176	182	187	193	198	204	209	215	221
16	145	151	156	162	167	173	179	185	191	197	202	208	214	220	226	232	238	244
17	163	169	174	180	187	193	199	205	211	218	224	231	237	243	250	256	263	269
18	181	188	194	200	207	213	220	227	233	240	247	254	260	267	274	281	288	295
19	201	208	214	221	228	235	242	249	256	263	271	278	285	292	300	307	314	321
20	222	229	236	243	250	258	265	273	280	288	295	303	311	318	326	334	341	349

7.4 La loi de Student

A la ligne *m* et dans la colonne *p*, on trouve la valeur *x* telle que $P(X \leq x) = p$ pour une loi de Student à *m* degrés de liberté. La dernière ligne correspond à la loi normale de moyenne 0 et variance 1.

	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0	0.3249	0.7265	1.3764	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0	0.2887	0.6172	1.0607	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0	0.2767	0.5844	0.9785	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0	0.2707	0.5686	0.9410	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0	0.2672	0.5594	0.9195	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0	0.2648	0.5534	0.9057	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0	0.2632	0.5491	0.8960	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0	0.2619	0.5459	0.8889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0	0.2610	0.5435	0.8834	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0	0.2602	0.5415	0.8791	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.2596	0.5399	0.8755	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0	0.2590	0.5386	0.8726	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0	0.2586	0.5375	0.8702	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0	0.2582	0.5366	0.8681	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0	0.2579	0.5357	0.8662	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0	0.2576	0.5350	0.8647	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0	0.2573	0.5344	0.8633	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0	0.2571	0.5338	0.8620	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0	0.2569	0.5333	0.8610	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0	0.2567	0.5329	0.8600	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0	0.2566	0.5325	0.8591	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0	0.2564	0.5321	0.8583	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0	0.2563	0.5317	0.8575	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0	0.2562	0.5314	0.8569	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0	0.2561	0.5312	0.8562	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0	0.2560	0.5309	0.8557	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0	0.2559	0.5306	0.8551	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0	0.2558	0.5304	0.8546	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0	0.2557	0.5302	0.8542	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0	0.2556	0.5300	0.8538	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0	0.2550	0.5286	0.8507	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0	0.2545	0.5272	0.8477	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0	0.2542	0.5265	0.8461	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0	0.2540	0.5261	0.8452	1.290	1.660	1.984	2.364	2.626	3.174	3.390
200	0	0.2537	0.5252	0.8434	1.286	1.653	1.972	2.345	2.601	3.131	3.340
N01	0	0.2533	0.5244	0.8416	1.282	1.645	1.960	2.326	2.576	3.090	3.291

7.5 La loi du Khi2

A la ligne *m* et dans la colonne *p*, on trouve la valeur *x* telle que $P(X \leq x) = p$ pour une loi du Khi2 à *m* degrés de liberté. Pour $m \geq 30$ on utilise l'approximation donnée par la relation :

Si *X* suit une loi du Khi2 à *m* degrés de liberté, $\sqrt{2X} - \sqrt{2m-1}$ suit une loi normale de moyenne 0 et variance 1.

	0.001	0.005	0.01	0.05	0.1	0.9	0.95	0.99	0.995	0.999	0.9995	0.9999
1	0.0	0.0	0.0	0.0	0.0	2.7	3.8	6.6	7.9	10.8	12.1	15.1
2	0.0	0.0	0.0	0.1	0.2	4.6	6.0	9.2	10.6	13.8	15.2	18.4
3	0.0	0.1	0.1	0.4	0.6	6.3	7.8	11.3	12.8	16.3	17.7	21.1
4	0.1	0.2	0.3	0.7	1.1	7.8	9.5	13.3	14.9	18.5	20.0	23.5
5	0.2	0.4	0.6	1.1	1.6	9.2	11.1	15.1	16.7	20.5	22.1	25.7
6	0.4	0.7	0.9	1.6	2.2	10.6	12.6	16.8	18.5	22.5	24.1	27.9
7	0.6	1.0	1.2	2.2	2.8	12.0	14.1	18.5	20.3	24.3	26.0	29.9
8	0.9	1.3	1.6	2.7	3.5	13.4	15.5	20.1	22.0	26.1	27.9	31.8
9	1.2	1.7	2.1	3.3	4.2	14.7	16.9	21.7	23.6	27.9	29.7	33.7
10	1.5	2.2	2.6	3.9	4.9	16.0	18.3	23.2	25.2	29.6	31.4	35.6
11	1.8	2.6	3.1	4.6	5.6	17.3	19.7	24.7	26.8	31.3	33.1	37.4
12	2.2	3.1	3.6	5.2	6.3	18.5	21.0	26.2	28.3	32.9	34.8	39.1
13	2.6	3.6	4.1	5.9	7.0	19.8	22.4	27.7	29.8	34.5	36.5	40.9
14	3.0	4.1	4.7	6.6	7.8	21.1	23.7	29.1	31.3	36.1	38.1	42.6
15	3.5	4.6	5.2	7.3	8.5	22.3	25.0	30.6	32.8	37.7	39.7	44.3
16	3.9	5.1	5.8	8.0	9.3	23.5	26.3	32.0	34.3	39.3	41.3	45.9
17	4.4	5.7	6.4	8.7	10.1	24.8	27.6	33.4	35.7	40.8	42.9	47.6
18	4.9	6.3	7.0	9.4	10.9	26.0	28.9	34.8	37.2	42.3	44.4	49.2
19	5.4	6.8	7.6	10.1	11.7	27.2	30.1	36.2	38.6	43.8	46.0	50.8
20	5.9	7.4	8.3	10.9	12.4	28.4	31.4	37.6	40.0	45.3	47.5	52.4
21	6.4	8.0	8.9	11.6	13.2	29.6	32.7	38.9	41.4	46.8	49.0	54.0
22	7.0	8.6	9.5	12.3	14.0	30.8	33.9	40.3	42.8	48.3	50.5	55.5
23	7.5	9.3	10.2	13.1	14.8	32.0	35.2	41.6	44.2	49.7	52.0	57.1
24	8.1	9.9	10.9	13.8	15.7	33.2	36.4	43.0	45.6	51.2	53.5	58.6
25	8.6	10.5	11.5	14.6	16.5	34.4	37.7	44.3	46.9	52.6	54.9	60.1
26	9.2	11.2	12.2	15.4	17.3	35.6	38.9	45.6	48.3	54.1	56.4	61.7
27	9.8	11.8	12.9	16.2	18.1	36.7	40.1	47.0	49.6	55.5	57.9	63.2
28	10.4	12.5	13.6	16.9	18.9	37.9	41.3	48.3	51.0	56.9	59.3	64.7
29	11.0	13.1	14.3	17.7	19.8	39.1	42.6	49.6	52.3	58.3	60.7	66.2
30	11.6	13.8	15.0	18.5	20.6	40.3	43.8	50.9	53.7	59.7	62.2	67.6
31	12.2	14.5	15.7	19.3	21.4	41.4	45.0	52.2	55.0	61.1	63.6	69.1
32	12.8	15.1	16.4	20.1	22.3	42.6	46.2	53.5	56.3	62.5	65.0	70.6
33	13.4	15.8	17.1	20.9	23.1	43.7	47.4	54.8	57.6	63.9	66.4	72.0
34	14.1	16.5	17.8	21.7	24.0	44.9	48.6	56.1	59.0	65.2	67.8	73.5
35	14.7	17.2	18.5	22.5	24.8	46.1	49.8	57.3	60.3	66.6	69.2	74.9
36	15.3	17.9	19.2	23.3	25.6	47.2	51.0	58.6	61.6	68.0	70.6	76.4
37	16.0	18.6	20.0	24.1	26.5	48.4	52.2	59.9	62.9	69.3	72.0	77.8
38	16.6	19.3	20.7	24.9	27.3	49.5	53.4	61.2	64.2	70.7	73.4	79.2
39	17.3	20.0	21.4	25.7	28.2	50.7	54.6	62.4	65.5	72.1	74.7	80.6
40	17.9	20.7	22.2	26.5	29.1	51.8	55.8	63.7	66.8	73.4	76.1	82.1
41	18.6	21.4	22.9	27.3	29.9	52.9	56.9	65.0	68.1	74.7	77.5	83.5
42	19.2	22.1	23.7	28.1	30.8	54.1	58.1	66.2	69.3	76.1	78.8	84.9
43	19.9	22.9	24.4	29.0	31.6	55.2	59.3	67.5	70.6	77.4	80.2	86.3
44	20.6	23.6	25.1	29.8	32.5	56.4	60.5	68.7	71.9	78.7	81.5	87.7
45	21.3	24.3	25.9	30.6	33.4	57.5	61.7	70.0	73.2	80.1	82.9	89.1
46	21.9	25.0	26.7	31.4	34.2	58.6	62.8	71.2	74.4	81.4	84.2	90.5
47	22.6	25.8	27.4	32.3	35.1	59.8	64.0	72.4	75.7	82.7	85.6	91.8
48	23.3	26.5	28.2	33.1	35.9	60.9	65.2	73.7	77.0	84.0	86.9	93.2
49	24.0	27.2	28.9	33.9	36.8	62.0	66.3	74.9	78.2	85.4	88.2	94.6
50	24.7	28.0	29.7	34.8	37.7	63.2	67.5	76.2	79.5	86.7	89.6	96.0

7.6 Edition des tables dans R

Les fonctions qui suivent donnent les tables des pages qui précèdent. On peut s'en servir comme exercices.

```
> edit1
function(){
        a_matrix(0,15,14)
        for (i in 1:14) {
                a[1:(i+1),i]_round(1000*pbinom(0:i,i,1/2),digits=0)
        }
        a_as.data.frame(a)
        row.names(a)_0:14
        names(a)_1:14
        a_matrix(0,29,14);
        for (j in 1:14) {
                i_j+14
                a[1:(i+1),j]_round(1000*pbinom(0:i,i,1/2),digits=0)
        }
        a_as.data.frame(a)
        row.names(a)_0:28
        names(a)_15:28
        a
}
> edit1()
     15   16   17   18   19   20   21   22   23   24   25   26   27   28
0     0    0    0    0    0    0    0    0    0    0    0    0    0    0
1     0    0    0    0    0    0    0    0    0    0    0    0    0    0
2     4    2    1    1    0    0    0    0    0    0    0    0    0    0
. . .


> cbind.data.frame(seq(-6,0,by=0.1),pnorm(seq(-6,0,by=0.1)))

1                  -6.0                   9.866e-10
2                  -5.9                   1.818e-09
3                  -5.8                   3.316e-09
. . .

> edit2
function(){
        for(i in 10:1) {
                a0 <- 10^( - i)
                print(c(a0, qnorm(a0)))
                a0 <- 5 * (10^( - i))
                print(c(a0, qnorm(a0)))
        }
}

> edit2()
[1]  1.000e-10 -6.361e+00
[1]  5.00e-10 -6.11e+00
[1]  1.000e-09 -5.998e+00
[1]  5.000e-09 -5.731e+00
[1]  1.000e-08 -5.612e+00
. . .

> edit3
function(){
        a_matrix(0,36,11)
        qq_c(0.5,0.6,0.7,0.8,0.9,0.95,0.975,0.99,0.995,0.999,0.9995)
        for (i in 1:30) {
                a[i,]_qt(qq,i)
        }
        a[31,]_qt(qq,40)
        a[32,]_qt(qq,60)
        a[33,]_qt(qq,80)
        a[34,]_qt(qq,100)
        a[35,]_qt(qq,200)
```

```
        a[36,]_qnorm(qq)
        a_as.data.frame(a);row.names(a)_c(1:30,40,60,80,100,200,"N01");names(a)_qq
        a
}

> edit3()
           0.5    0.6    0.7    0.8   0.9  0.95  0.975   0.99  0.995   0.999  0.9995
1   -6.123e-17 0.3249 0.7265 1.3764 3.078 6.314 12.706 31.821 63.657 318.309 636.619
2    0.000e+00 0.2887 0.6172 1.0607 1.886 2.920  4.303  6.965  9.925  22.327  31.599
3    0.000e+00 0.2767 0.5844 0.9785 1.638 2.353  3.182  4.541  5.841  10.215  12.924
4    0.000e+00 0.2707 0.5686 0.9410 1.533 2.132  2.776  3.747  4.604   7.173   8.610
5    0.000e+00 0.2672 0.5594 0.9195 1.476 2.015  2.571  3.365  4.032   5.893   6.869
. . .

> edit4
function(){
        a_matrix(0,50,12)
        qq_c(0.001,0.005,0.01,0.05,0.10,0.90,0.95,0.99,0.995,0.999,0.9995,0.9999)
        for (j in 1:12) {
                a[,j]_round(qchisq(qq[j],1:50),digits=1)
        }
        a_as.data.frame(a,row.names=1:50)
        names(a)_qq
        a
}
> edit4()
   0.001 0.005 0.01 0.05  0.1  0.9 0.95 0.99 0.995 0.999 0.9995 0.9999
1    0.0   0.0  0.0  0.0  0.0  2.7  3.8  6.6   7.9  10.8   12.1   15.1
2    0.0   0.0  0.0  0.1  0.2  4.6  6.0  9.2  10.6  13.8   15.2   18.4
3    0.0   0.1  0.1  0.4  0.6  6.3  7.8 11.3  12.8  16.3   17.7   21.1
4    0.1   0.2  0.3  0.7  1.1  7.8  9.5 13.3  14.9  18.5   20.0   23.5
. . .


> edit5
function(){
        alpha_ 0.05
        a_matrix(0,20,20)
        for (i in 3:20) {
                for(j in 3:20){
                        a[i,j]_qwilcox(alpha,i,j)
                }
        }
        a_as.data.frame(a[3:20,3:20])
        row.names(a)_3:20;names(a)_3:20
        a
}
<environment: 02B3E198>
> edit5()
    3  4  5  6  7  8  9 10 11 12 13 14  15  16  17  18  19  20
3   0  1  2  3  3  4  4  5  6  6  7  8   9  10  10  11  12
4   1  2  3  4  5  6  7  8  9 10 11 12  13  15  16  17  18  19
5   2  3  5  6  7  9 10 12 13 14 16 17  19  20  21  23  24  26
. . .
```

1 Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994*) A handbook of small data sets*. Chapman & Hall, London. 1-458.

2 Manly, B.F.J. (1991) *Randomization and Monte Carlo methods in biology*. Chapman & Hall, London. 1-281.

3 Cameron, E. & Pauling, L. (1978) 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.

4 Mendenhall W., Ott W. & Larson R.F. (1974) Statistics: a tool for the social sciences. Duxbury Press, Boston.

5 Dagnelie, P. (1970) Théories et méthodes statistiques. Volume 2, Les méthodes de l'inférence statistique. Les presses agronomiques de Gembloux, Gembloux. 1-451.

6 Sprent, P. (1992) *Pratique des statistiques non paramétriques*. Traduction française de J.P. Ley. INRA Editions, Paris. 1-294.