Fiche TD avec le logiciel R : tdr22

Graphes quantiles-quantiles

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

Histogrammes, fonctions de répartition, droite de Henri (ou de Henry) et `qqnorm()`. Mélanges de distributions

1 Introduction

V. M. Nigon pose par e-mail des questions fort claires dont la première concerne les mélanges de lois normales :

> *Pour toutes sortes de problèmes, nous sommes amenés à utiliser la droite de Henry. Nous le faisons par la méthode graphique, à la main. Beaucoup de gens disent qu'il existe des programmes informatiques qui permettent de faire le travail par ordinateur. Jusqu'à présent, je ne suis pas parvenu à découvrir les logiciels appropriés. Il s'avère que toutes les personnes que j'ai consultées m'orientent sur des voies qui ne correspondent pas à la droite de Henry (transformations probits ou autres, destinées à tester la normalité). En particulier, que faire lorsque le graphique en droite de Henry semble se résoudre en deux droites. Comment calculer les distributions normales correspondant à chacune des droites ? Etc.*

Dans R on trouve les outils nécessaires. Cette fiche donne des illustrations.

2 Un mélange observé

On demande aux étudiants d'un amphi d'indiquer leur sexe (h/f), leur poids (en kg) et leur taille (en cm). Les données sont les mêmes que celles de la fiche tdr12.pdf.

```
stp <- read.table("http://pbil.univ-lyon1.fr/R/donnees/t3var.txt", header = TRUE)
stp[1:5, ]
```

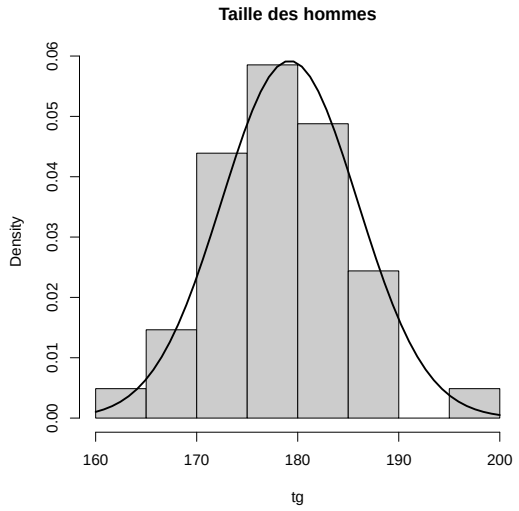
```
  sexe poi tai
1    h  60 170
2    f  57 169
3    f  51 172
4    f  55 174
5    f  50 168
```

On sépare la taille des hommes et la taille des femmes :

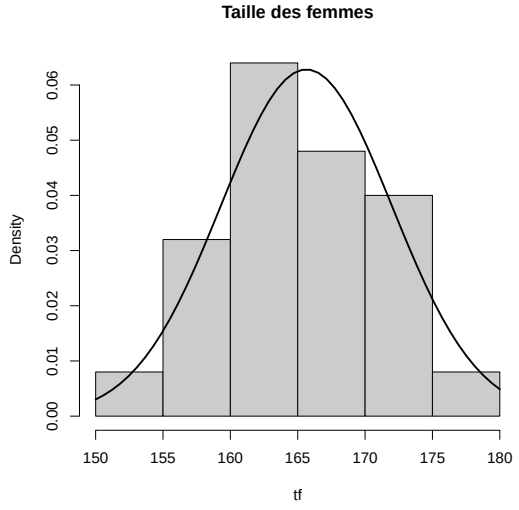
```
tg <- stp[stp$sex=="h","tai"]
tg
```

```
 [1] 170 189 175 164 175 184 178 179 182 174 172 185 178 180 189 200 178 178 175 180
[21] 169 173 182 183 184 181 180 178 178 168 171 180 174 175 182 181 188 182 189 178
[41] 186
```

```
hist(tg, proba = TRUE, col = grey(0.8), main = "Taille des hommes")
provi <- seq(160, 200, length=50)
lines(provi,dnorm(provi,mean(tg), sd(tg)), lwd = 2)
```

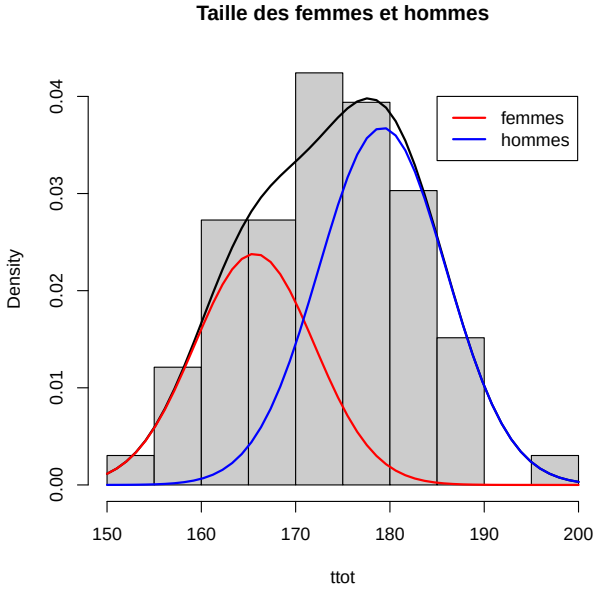


Faire la même représentation graphique pour les femmes :



Représenter le mélange des deux distributions :

```
ttot <- stp[,"tai"]
hist(ttot, proba = TRUE, nclass=10, col = grey(0.8), main = "Taille des femmes et hommes")
provi <- seq(150,200,length=50)
x1 <- dnorm(provi,mean(tf),sd(tf))
x2 <- dnorm(provi,mean(tg),sd(tg))
pf <- length(tf)/length(ttot)
x3 <- pf*x1 + (1-pf)*x2
lines(provi,x3, lwd = 2)
lines(provi, pf*x1, col = "red", lwd = 2)
lines(provi, (1-pf)*x2, col = "blue", lwd = 2)
legend(185, 0.04, c("femmes","hommes"), col = c("red","blue"), lty = 1, lwd = 2)
```



Les questions qui se posent sont : comment détecter un mélange, comment estimer un mélange, quelles contraintes prévoir pour réussir ces opérations ?

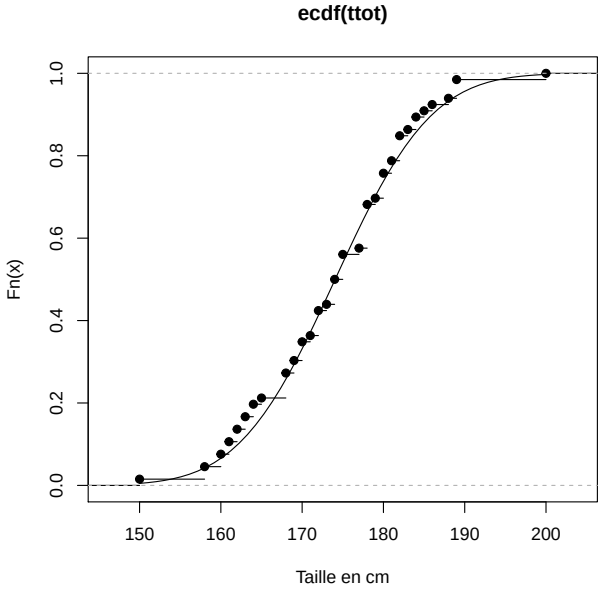
3 Droite de Henri et `qqnorm()`

3.1 Normalité et fonction de répartition empirique

Proposé par Martin Maechler (maechler@stat.math.ethz.ch) dans la bibliothèque de base `stats`. Consulter la documentation de la fonction `ecdf` avec `?ecdf` :

> `The e.c.d.f. (empirical cumulative distribution function) Fn is a step function with jump 1/n at each observation (possibly with multiple jumps at one place if there are ties). Missing values are ignored.`

```
plot(ecdf(ttot), xlab = "Taille en cm")
lines (provi, pnorm(provi,mean(ttot),sd(ttot)))
```



Pour tester la normalité en utilisant la fonction de répartition :

```
ks.test(ttot,pnorm,mean(ttot), sd(ttot))
```

```
        One-sample Kolmogorov-Smirnov test
data:  ttot
D = 0.088092, p-value = 0.685
alternative hypothesis: two-sided
```

Attention, la documentation de la fonction (`?ks.test`) nous dit :

> Exact p-values are only available for the two-sided two-sample test with no ties. In that case, if 'exact = NULL' (the default) an exact p-value is computed if the product of the sample sizes is less than 10000. Otherwise, asymptotic distributions are used whose approximations may be inaccurate in small samples.

Ce qui convient nous est suggéré par :

> See Also:
>
> 'shapiro.test' which performs the Shapiro-Wilk test for normality.

Consulter la documentation de cette fonction (`?shapiro.test`), puis effectuer le test [9] :

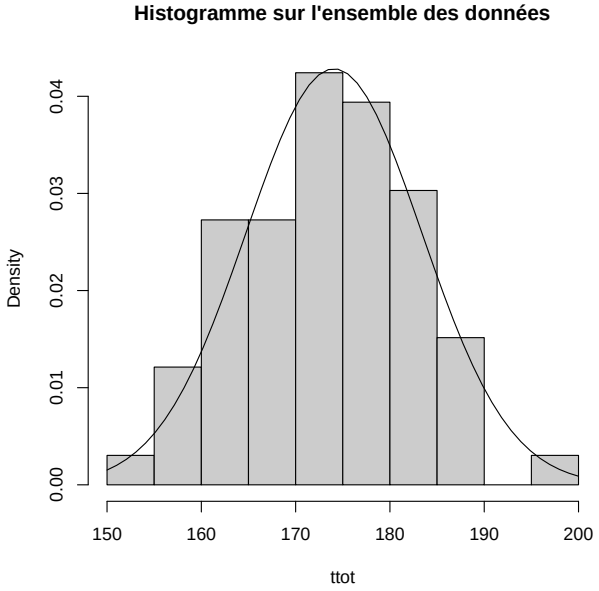
```
shapiro.test(ttot)
```

```
        Shapiro-Wilk normality test
data:  ttot
W = 0.98767, p-value = 0.7585
```

Rien à faire : on ne voit pas le mélange de lois normales sur l'exemple.

3.2 Normalité et histogramme

```
hist(ttot, proba = TRUE, col = grey(0.8), main = "Histogramme sur l'ensemble des données")
lines (provi, dnorm(provi, mean(ttot), sd(ttot)))
```



Pour tester la normalité en utilisant le test du χ^2 :

```
hist(ttot,plot=FALSE, breaks = seq(150,200, by =10))
```

```
$breaks
[1] 150 160 170 180 190 200
$counts
[1]  5 18 27 15  1

$density
[1] 0.007575758 0.027272727 0.040909091 0.022727273 0.001515152

$mids
[1] 155 165 175 185 195

$xname
[1] "ttot"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

Mettre ces paramètres dans une liste `w` :

```
is.list(w)
```

```
[1] TRUE
```

```
names(w)
```

```
[1] "breaks"   "counts"   "density"  "mids"     "xname"    "equidist"
```

```
w$breaks
```

```
[1] 150 160 170 180 190 200
```

```
w$counts
```

```
[1]  5 18 27 15  1
```

```
sum(w$counts)
```

```
[1] 66
```

```
nrow(stp)
```

```
[1] 66
```

```
br0 <- w$breaks[-5]
obs <- w$counts[-5]
obs
```

```
[1]  5 18 27 15
```

```
obs[4] <- obs[4]+1
obs
```

```
[1]  5 18 27 16
```

On a les classes et les effectifs. Il reste à calculer les probabilités :

```
mean(ttot)
```

```
[1] 174.0606
```

```
sd(ttot)
```

```
[1] 9.313148
```

```
pnorm(br0, mean(ttot), sd(ttot))
```

```
[1] 0.00489004 0.06555251 0.33141551 0.73817946 0.99732563
```

```
t0 <- pnorm(br0, mean(ttot), sd(ttot))
t0[1] <- 0
t0[5] <- 1
t0
```

```
[1] 0.00000000 0.06555251 0.33141551 0.73817946 1.00000000
```

Expliquer.

```
t0 <- diff(t0)
sum(t0)
```

```
[1] 1
```

Consulter la documentation (`?chisq.test`), puis effectuer le test.

```
chisq.test(obs,p=t0)
```

```
        Chi-squared test for given probabilities
data:  obs
X-squared = 0.21227, df = 3, p-value = 0.9756
```

```
1-pchisq(0.2416,df = 1)
```

```
[1] 0.6230529
```

```
Warning message:
Chi-squared approximation may be incorrect in: chisq.test(obs, p = t0, sim = TRUE)
```

Pourquoi *may be incorrect* ?

```
sum((obs-66*t0)^2/(66*t0))
```

```
[1] 0.2122668
```

Expliquez ce que l'on a retrouvé ici.
On peut procéder à l'envers (et c'est meilleur) :

```
ttot
```

```
 [1] 170 169 172 174 168 161 162 189 160 175 165 164 175 184 178 158 164 179 182 174
[21] 158 163 172 185 170 178 180 189 172 174 200 178 178 168 170 160 163 168 172 175
[41] 180 162 177 169 173 182 183 184 181 180 178 178 168 161 171 180 174 175 182 181
[61] 188 182 189 178 150 186
```

```
wnorm <- qnorm(seq(from = 0, to = 1, by = 0.1))
wnorm
```

```
 [1]       -Inf -1.2815516 -0.8416212 -0.5244005 -0.2533471  0.0000000  0.2533471
 [8]  0.5244005  0.8416212  1.2815516        Inf
```

```
 br0 <- wnorm*sd(ttot)+mean(ttot)
 br0
```

```
 [1]     -Inf 162.1253 166.2225 169.1768 171.7011 174.0606 176.4201 178.9444 181.8987
[10] 185.9959      Inf
```

```
 br0[1] <- -1000
 br0[11] <- 1000
 hist(ttot, br0, plot = FALSE)$counts
```

```
 [1] 9 5 6 4 9 4 8 7 8 6
```

```
 w1 <- hist(ttot,br0, plot = FALSE)$counts
 sum(w1)
```

```
[1] 66
```

```
 chisq.test(w1,p=rep(0.1,10))
```

```
        Chi-squared test for given probabilities
data:  w1
X-squared = 4.9091, df = 9, p-value = 0.8422
```

```
 1-pchisq(4.909,7)
```

```
[1] 0.6710675
```

Il n'y a rien à faire !

3.3 Droite de Henri (ou de Henry)

C'est une pratique tombée en désuétude qui se réalisait sur un papier spécial dit papier gausso-arithmétique qui par anamorphose transformait la sigmoïde de la fonction de répartition de loi normale en une droite. D'après Funkhouser [4], l'intérêt d'une tel papier fut soulignée par Francis Galton en 1899 [5] et introduit aux États-Unis d'Amérique en 1914 par Hazen [7] sous le nom de « *arithmetic probability paper* ». Un papier similaire fut introduit indépendamment en Angleterre par Dufton en 1930 sous le nom de « *permille paper* » [2], ce même auteur fit d'ailleurs remarquer [3] qu'il y avait une erreur dans l'article de Hazen [7]. En France, la tradition remonte encore plus loin [1], le terme de droite de Henri fait référence aux cours donnés en 1894 par le colonel Henri à l'école d'artillerie Fontainebleau [8]. Il a introduit un papier gausso-arithémtique pour vérifier la normalité de la portée de tir des canons [1]. Le terme de « droite de Henry » ou « droite de Henri » perdure en France [2] .

3.4 Papier gausso-arithmétique

Essayons de tracer une feuille de papier gausso-arithmétique. L'ordonnée est celui d'un papier millimétré ordinaire. L'abscisse par contre porte des lignes verticales étiquetées 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 98, 99, 99.9, 99.99. Chaque étiquette a_1 est placée à la valeur qui a la probabilité $\frac{a_1}{100}$ de ne pas être dépassée, c'est à dire le quantile de la loi normale :

```
 a1 <- c(0.01,  0.05,  0.1,  0.2,  0.5,  1,  2,  5, 10, 20, 30, 40,
 50, 60, 70, 80, 90, 95, 98, 99, 99.9, 99.99)
 a1
```

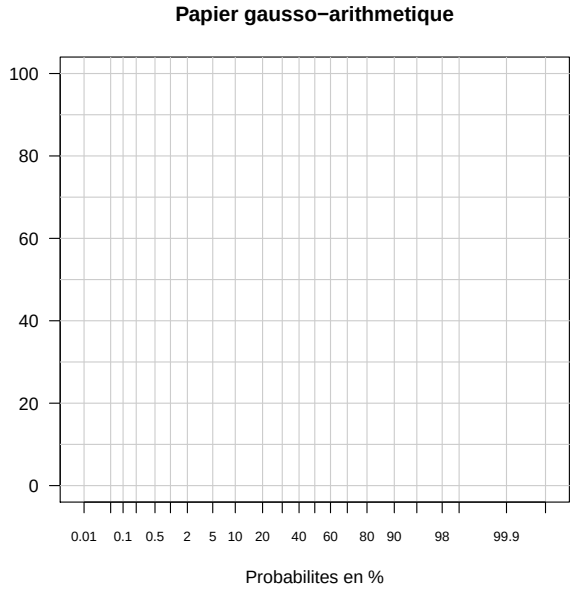
```
 [1]  0.01  0.05  0.10  0.20  0.50  1.00  2.00  5.00 10.00 20.00 30.00 40.00 50.00
[14] 60.00 70.00 80.00 90.00 95.00 98.00 99.00 99.90 99.99
```

```
 qnorm(a1/100)
```

1. Voir aussi [6] pour la solution de Lhoste à l'épineux problème de l'auto-corrélation temporelle des conditions de tir.
2. L'orthographe des noms de famille n'a pas toujours été stable...

```
 [1] -3.7190165 -3.2905267 -3.0902323 -2.8781617 -2.5758293 -2.3263479 -2.0537489
 [8] -1.6448536 -1.2815516 -0.8416212 -0.5244005 -0.2533471  0.0000000  0.2533471
[15]  0.5244005  0.8416212  1.2815516  1.6448536  2.0537489  2.3263479  3.0902323
[22]  3.7190165
```

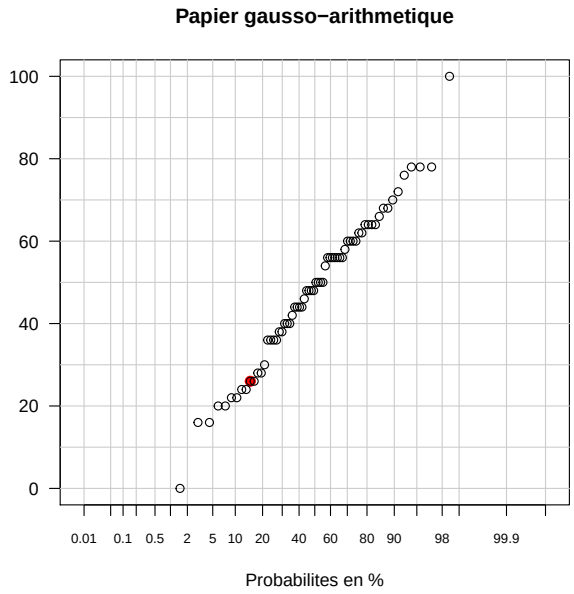
```
plot(0,0,xlim=c(-3.8,3.8),ylim=c(0,100),type="n", las = 1, xaxt = "n",
xlab = "Probabilites en %", ylab = "",
main = "Papier gausso-arithmetique")
abline(v=qnorm(a1/100), col = grey(0.8))
abline(h=seq(0,100,by=10), col = grey(0.8))
axis(side = 1, at = qnorm(a1/100), label = a1, cex.axis = 0.75)
```



Pour placer un point d'un échantillon rangé par ordre croissant $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ on portait en abscisse $100\frac{i}{n+1}$ (en échelle probabilité) et en ordonnée $\frac{x_{(i)}-min}{max-min}$.

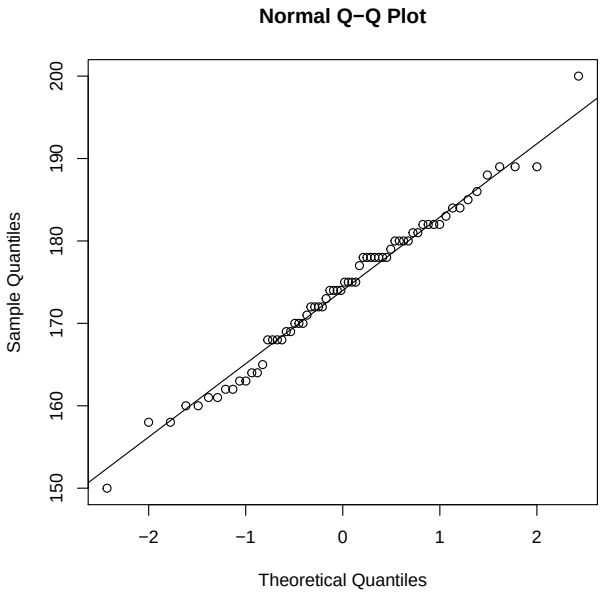
Exercice.

1. Prenons par exemple $i = 10$. Calculer l'abscisse et l'ordonnée associées. Placer le point en rouge sur le graphique. Puis représenter l'ensemble des valeurs.
2. Placer alors "à vue" avec une règle une droite dans le nuage. Récupérer avec $x = 50$ la moyenne et avec $x = 0.16$ et $x = 0.84$ le double de l'écart-type.



Ce dessin est basé sur le quantile de la loi normale d'un côté et le quantile observé de l'autre. On a exactement le même avec la fonction `qqnorm()` :

```
qqnorm(ttot)
qqline(ttot)
```



Ce graphe s'appelle quantile-quantile car il confronte les quantiles de la loi normale (en abscisse) et les quantiles empiriques de l'échantillon (en ordonnée).

La droite joint le couple des quantiles 0.25 et le couple des quantiles 0.75. Pour suivre exactement :

```
print(qqnorm(ttot))
qnorm((0.5/66)+(0:65)/66)
```

Observer que les 66 observations définissent 65 intervalles égaux auxquels on ajoute une moitié à gauche et une moitié à droite et qu'on utilise les quantiles théoriques de $\frac{2i-1}{2n}$ au lieu de $\frac{i}{n+1}$. Donc R est un logiciel libre qui trace les droites de Henri, mais les droites de Henri ne montrent pas facilement les mélanges de lois normales.

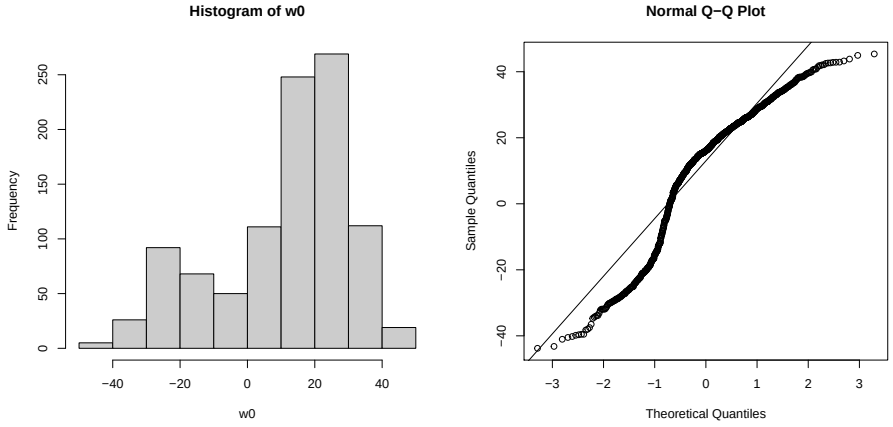
4 qqnorm et mélanges

On écrit une petite fonction pour simuler des mélanges :

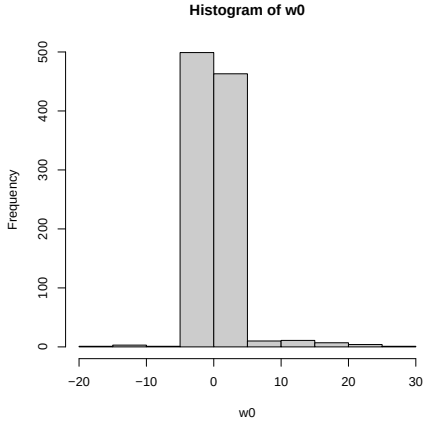
```
simulmixnor <- function(n = 100, p= 0.5, m1 = -1, sd1 = 1, m2 = 2, sd2 = 2)
{
      n1 <- rbinom(1,n,p)
      x1 <- rnorm(n1, m = m1, sd = sd1)
      x2 <- rnorm(n-n1, m = m2, sd = sd2)
      c(x1,x2)
}
```

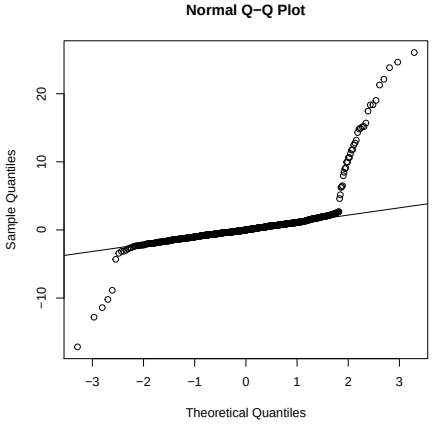
Puis on expérimente un petit peu :

```
par(mfrow=c(1,2))
w0 <- simulmixnor(1000,0.25,-20,10,20,10)
hist(w0, col = grey(0.8))
qqnorm(w0)
qqline(w0)
```

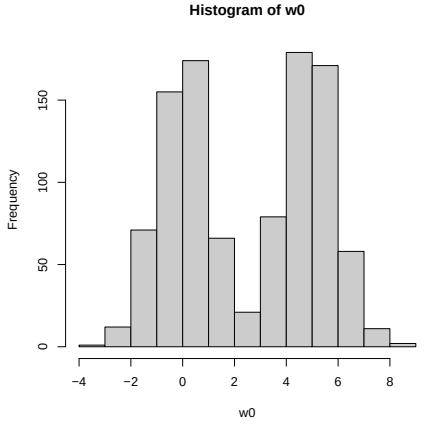


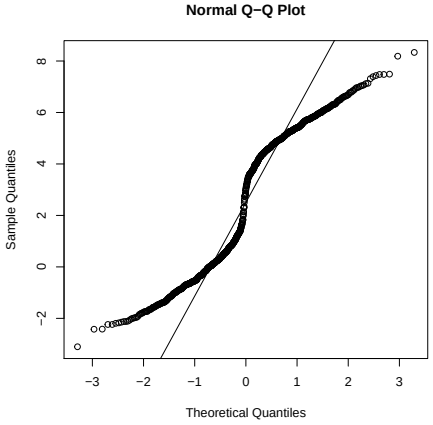

```
par(mfrow=c(1,2))
w0 <- simulmixnor(1000,0.95,0,1,10,10)
hist(w0, col = grey(0.8))
qqnorm(w0)
qqline(w0)
```



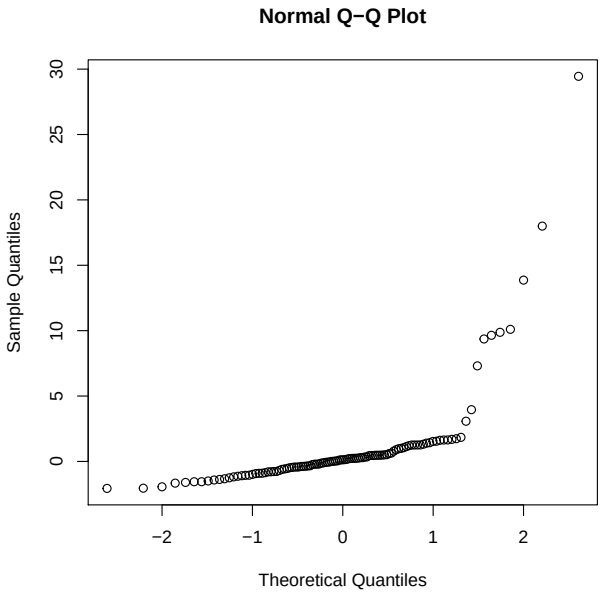



```
par(mfrow=c(1,2))
w0 <- simulmixnor(1000,0.5,0,1,5,1)
hist(w0, col = grey(0.8))
qqnorm(w0)
qqline(w0)
```






```
par(mfrow=c(1,1))
qqnorm(c(rnorm(100,0,1),rexp(10,0.1)))
```



On a plus facilement l'impression de deux droites de Henri en ajoutant un échantillon d'une population d'un autre type qu'avec un mélange de lois normales.

Références

[1] P. Crépel. Henri et la droite de Henry. *Matapli*, 36 :19–22, 1993.

[2] A.F. Dufton. Graphic statistics : permille paper. *Philosophical Magazine*, 10 :566, 1930.

[3] A.F. Dufton. Graphic statistics. *Science*, 79 :564–565, 1934.

[4] H.G. Funkhouser. Historical development of the graphical representation of statistical data. *Osiris*, 3 :269–404, 1937.

[5] F. Galton. A geometric determination of the median value of a system of normal variants from two of its centiles. *Nature*, 61 :102–104, 1899.

[6] N. Hadjadji Seddik-Ameur. Les tests de normalité de Lhoste. *Math. & Sci. hum., Mathematics and Social Sciences*, 41(162) :19–43, 2003.

[7] A. Hazen. Storage to be provided in impounding reservoir for municipal water supply. *Transactions American Society of Civil Engineers*, 77 :1529–1669, 1914.

[8] P. Henri. Probabilité du tir. facsimile of lectures notes from 1894. *Mémorial de l'Artillerie Française*, 5(2) :294–447, 1926.

[9] P. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31 :115–124, 1982.