Notions de statistiques

Pascal Bessonneau

2024-04-24

Contents

1 Statistiques descriptives et tests **5**

2 Statistiques descriptives **7**
2.1 Distribution 7
2.2 Indicateurs de position (mesures de tendance centrale) 10
2.3 Indicateur de dispersion 11
2.4 Les indicateurs de la loi normale 13
2.5 Quantiles, ex de la loi normale. 14
2.6 La médiane et les quantiles usuelles 18
2.7 Corrélations 19
2.8 Covariances 27

3 Tests **29**
3.1 Convergence vers la loi normale 29
3.2 Test Z 29
3.3 Test de Student 31
3.4 Erreur de Type I et II 33
3.5 Tests “paramétriques” et non paramétriques 34
3.6 Précautions à prendre quand on travaille avec des tests 34

Chapter 1

Statistiques descriptives et tests

Le but est de faire un peu de révision sur les statistiques descriptives et les tests statistiques.

Chapter 2

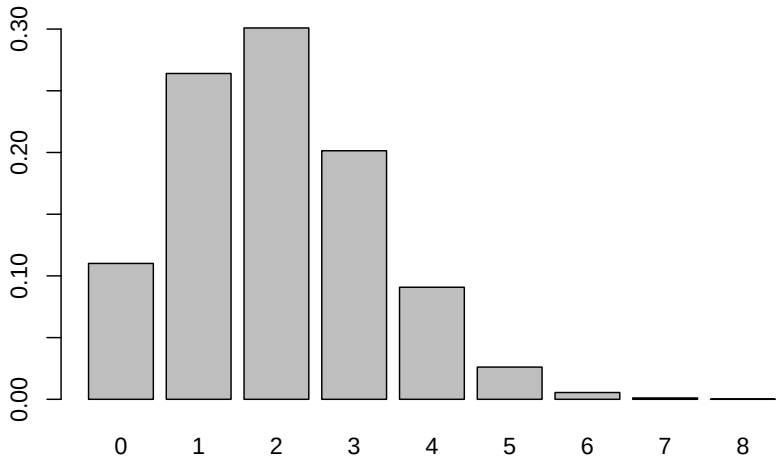
Statistiques descriptives

Le but de ce document est de rappeler les différents indicateurs utilisés dans les statistiques descriptives.

2.1 Distribution

Dans le cas qualitatif, on appele **distribution** la répartition des probabilités pour tous les points pour laquelle la distribution existe.

```
p <- rbinom(n=10000, size = 10,p = 0.2)
barplot(prop.table(table(p)))
```



Nous voyons pour une distribution binomiale la distribution, cela représente un lancer de dés avec ici, la somme de dix lancers, avec 0,2 la probabilité d'obtenir un 1 et 0,8 la probabilité d'obtenir un 0. Ici la probabilité d'obtenir un score de 3 est d'environ 20%.

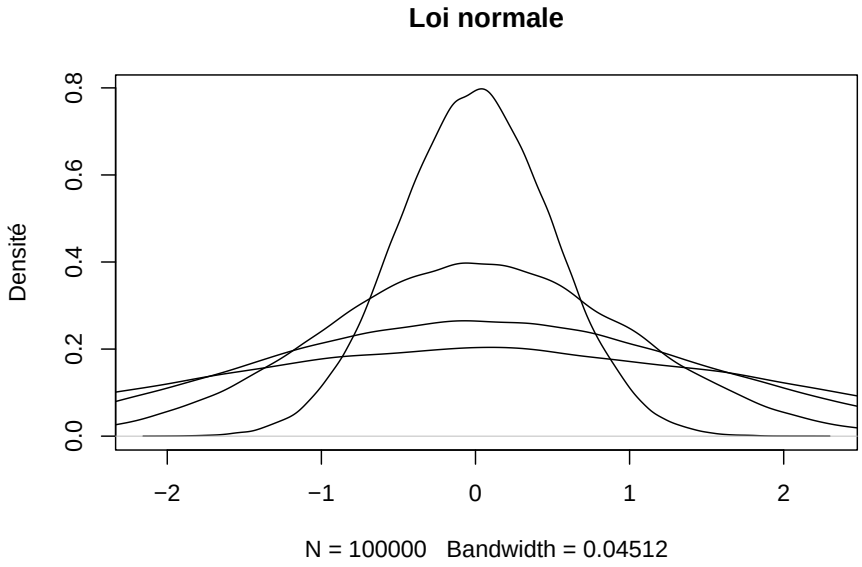
Dans le cas quantitatif, on appele **distribution** la répartition des probabilités pour tous les points pour laquelle la distribution existe. La plus connu est la loi normale:

```
exemple <- rnorm(1000000,0,1)
plot(density(exemple),main="Loi normale",ylab="Densité")
```

La loi normale prends deux paramètres : l'écart-type et la moyenne.

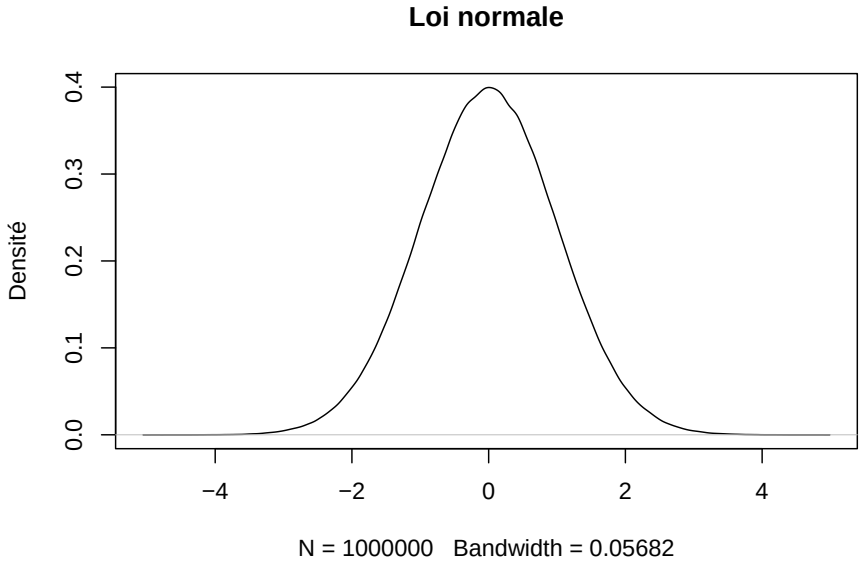
Si on fait varier l'écart-type : on a une forme plus aplatie ou inversement.

```
plot(density(rnorm(100000,0,0.5)),main="Loi normale",ylab="Densité")
lines(density(rnorm(100000,0,1)),main="Loi normale",ylab="Densité")
lines(density(rnorm(100000,0,1.5)),main="Loi normale",ylab="Densité")
lines(density(rnorm(100000,0,2)),main="Loi normale",ylab="Densité")
```



Si on fait varier la moyenne : la “pointe” de la courbe varie dans son positionnement.

```
exemple <- rnorm(1000000,0,1)
plot(density(exemple),main="Loi normale",ylab="Densité")
```



2.2 Indicateurs de position (mesures de tendance centrale)

La plupart des indicateurs statistiques s'inspire de la loi normale : Les indicateurs de position donnent une idée de où se répartissent les points.

Le premier indicateur est la moyenne arithmétique, elle a de nombreux prolongements en mathématiques, physique, etc.

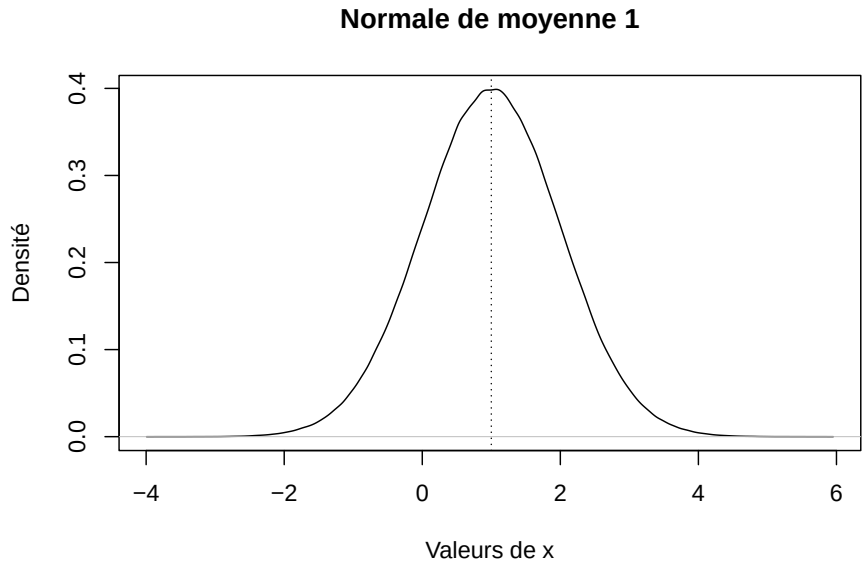
Il y a des indicateurs proches comme la moyenne géométrique, la moyenne harmonique, etc. La moyenne arithmétique est la plus utilisée.

2.2.1 Moyenne arithmétqiue

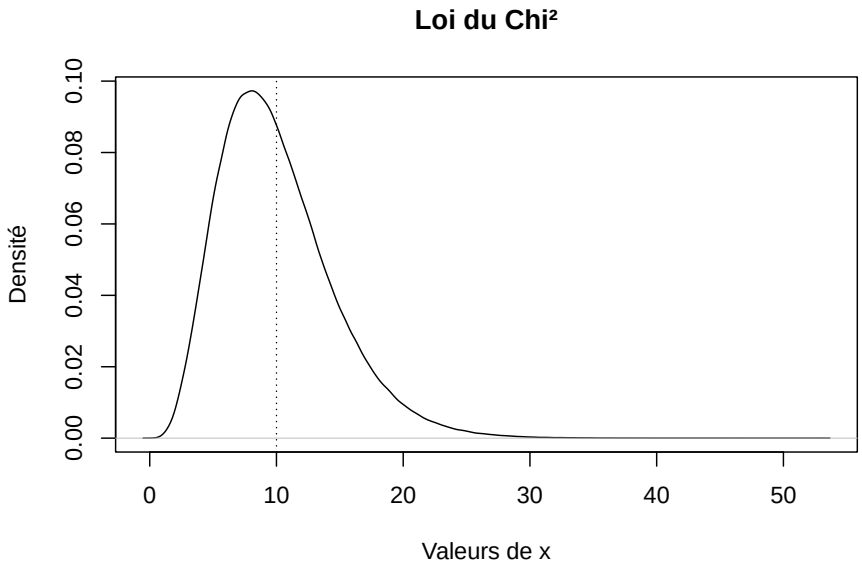
En statistiques, on retrouve cette moyenne dans l'ANOVA, la régression, le test de Student, etc.

C'est celle qui se retrouve dans la plupart des cas dans les formules.

```
y=rnorm(1000000,1,1)
plot(density(y),main="Normale de moyenne 1",xlab="Valeurs de x",ylab="Densité")
abline(v=1,lty=3)
```




```
y=rchisq(1000000,10)
plot(density(y),main="Loi du Chi²",xlab="Valeurs de x",ylab="Densité")
abline(v=mean(y),lty=3)
```



Dans des cas comme la loi normale, l'indicateur de position est centrale (ie. qu'il y a une symétrie). Mais souvent, la répartition des poids n'est pas symétrique et la moyenne est *décalée* brisant la symétrie comme dans l'illustration du Chi².

2.2.2 Autres moyennes

Elles sont peu utilisées.

2.3 Indicateur de dispersion

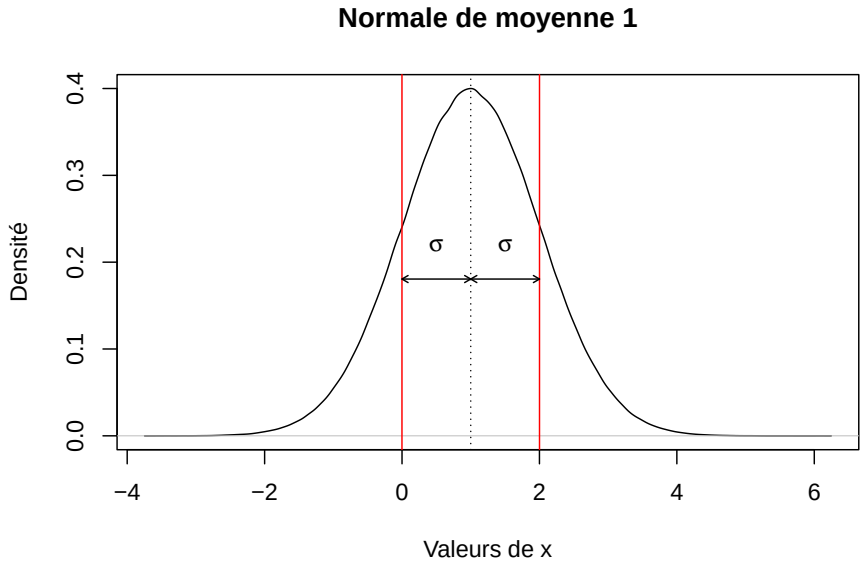
L'indicateur de dispersion le plus fréquent est l'écart-type. L'écart-type est le carré des écarts à la moyenne (arithmétique).

Il vient naturellement avec la loi normale normale où 50% des observations sont à 1 écart-type autour de la moyenne.

Plus l'écart-type est grand plus les écarts entre les valeurs observées et la moyenne sont grands. Les points sont plus dispersés.

```
y=rnorm(1000000,1,1)
plot(density(y),main="Normale de moyenne 1",xlab="Valeurs de x",ylab="Densité")
abline(v=1,lty=3)
abline(v=1-sd(y),lty=1,col="red")
abline(v=1+sd(y),lty=1,col="red")
text(1-sd(y)+0.5*sd(y),0.2+strheight("Ya"),expression(sigma))
text(1+0.5*sd(y),0.2+strheight("Ya"),expression(sigma))
```

```
arrows(
  1-sd(y),0.2-strheight("Ya"),
  1,0.2-strheight("Ya"),
  length=0.05,code=3
)
arrows(
  1,0.2-strheight("Ya"),
  1+sd(y),0.2-strheight("Ya"),
  length=0.05,code=3
)
```



Un écart-type et la moyenne sont de même unité : par exemple si on mesure le poids d'objets, on a par exemple moyenne et écart-type en **kg**.

Donc si on divise la moyenne par l'écart-type, on obtient un nombre sans unité ou plutôt avec comme unité un écart-type. Ce type d'opération de diviser par l'écart-type une ou des valeurs observées est appelé **réduire** une variable.

Cette opération est fréquemment associée au fait de soustraire par la moyenne avant de réduire. Donc les observations deviennent **centrées** autour de la moyenne. On appelle ces deux opérations centrer/réduire une variable soit **scale** en R.

Quand on analyse plusieurs variables d'unités très différentes en statistiques, avec d'un côté des chiffres très grands et de l'autre côté des petits par exemple, on est amené à centrer/réduire pour manipuler des chiffres de même ordre de

grandeur.

2.4 Les indicateurs de la loi normale

skweness et kurtosis.

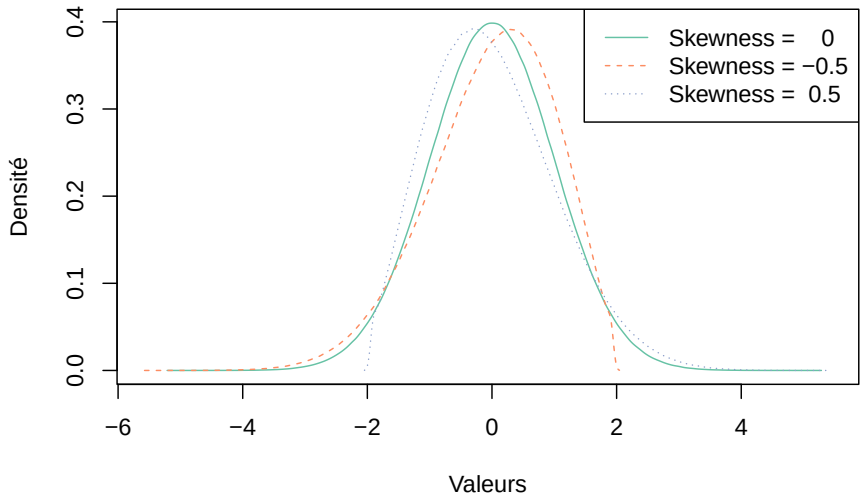
La kurtose est l'aplatissement de la distribution par rapport à la loi normale.

```
rr <- rnonnorm(1000000, mean = 0, sd = 1, skew = 0, kurt =  0)
r2 <- rnonnorm(1000000, mean = 0, sd = 1, skew = 0, kurt = -0.5)
r3 <- rnonnorm(1000000, mean = 0, sd = 1, skew = 0, kurt =  3)

d1 <- density(rr$dat)
d2 <- density(r2$dat)
d3 <- density(r3$dat)

plot(0,0,type="n",
     xlim=range(c(d1$x,d2$x,d3$x)),
     ylim=range(c(d1$y,d2$y,d3$y)),
     main = "",
     xlab="Valeurs",
     ylab="Densité"
       )
lines(d1,lty=1,col=brewer.pal(3,name = "Set2")[1])
lines(d2,lty=2,col=brewer.pal(3,name = "Set2")[2])
lines(d3,lty=3,col=brewer.pal(3,name = "Set2")[3])
legend("topright",c("Kurtose =    0","Kurtose = -0.5","Kurtose =  3"),
       lty=c(1,2,3),
       col=brewer.pal(3,name = "Set2"))
```

Le coefficient d'asymétrie ou skewness indique lui la symatrie par rapport à l'axe centrale de la distribution normale.

```
rr <- rnonnorm(9000000, mean = 0, sd = 1, skew =  0  , kurt = 0)
r2 <- rnonnorm(9000000, mean = 0, sd = 1, skew = -0.5, kurt = 0)
r3 <- rnonnorm(9000000, mean = 0, sd = 1, skew =  0.5, kurt = 0)

d1 <- density(rr$dat)
d2 <- density(r2$dat)
d3 <- density(r3$dat)

plot(0,0,type="n",
     xlim=range(c(d1$x,d2$x,d3$x)),
     ylim=range(c(d1$y,d2$y,d3$y)),
     main = "",
     xlab="Valeurs",
     ylab="Densité"
```

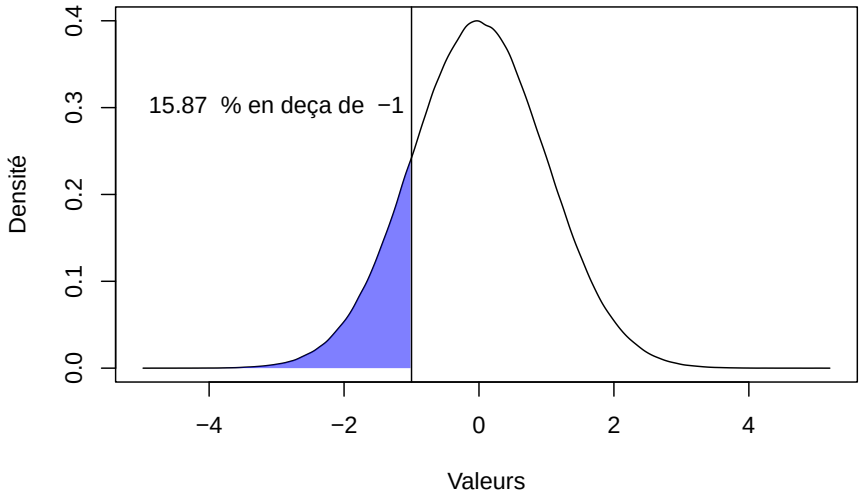
```
      )
lines(d1,lty=1,col=brewer.pal(3,name = "Set2")[1])
lines(d2,lty=2,col=brewer.pal(3,name = "Set2")[2])
lines(d3,lty=3,col=brewer.pal(3,name = "Set2")[3])
legend("topright",c("Skewness =    0","Skewness = -0.5","Skewness =  0.5"),
       lty=c(1,2,3),
       col=brewer.pal(3,name = "Set2"))
```



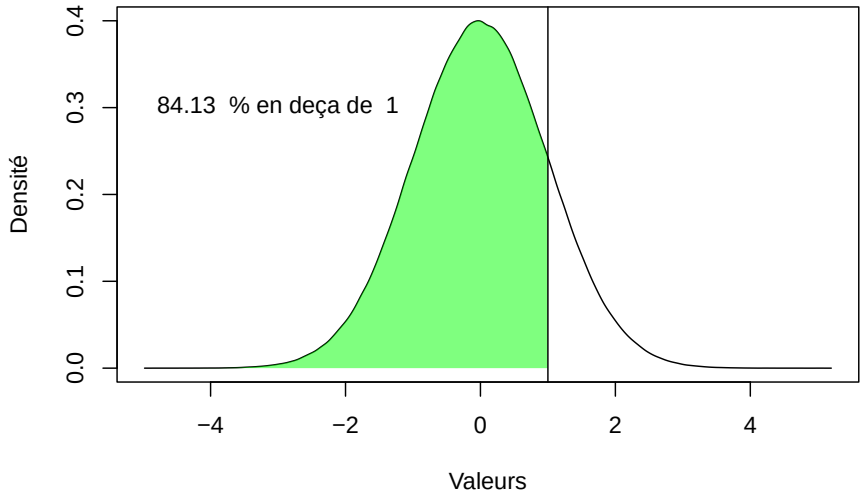
2.5 Quantiles, ex de la loi normale.

Les quantiles représentent la proportion d'individus qui se retrouvent en deça d'une valeur. Par exemple, pour une loi normale :

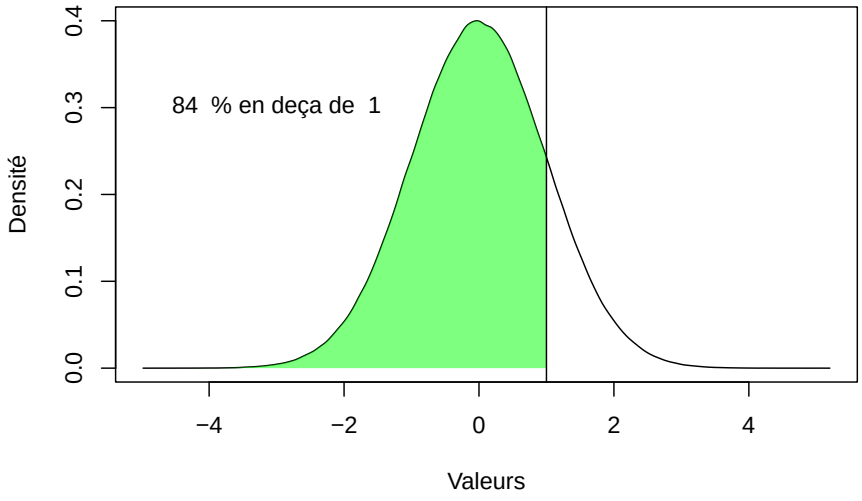
```
rr <- rnorm(1000000)
dd <- density(rr)

plot(0,0,type="n",main="",xlab="Valeurs",ylab="Densité",
     xlim=range(dd$x),ylim=range(dd$y))
lines(dd)
polygon(c(dd$x[dd$x < -1],rev(dd$x[dd$x < -1])),
        c(dd$y[dd$x < -1],rep(0,length(dd$y))[dd$x < -1]),
        col=rgb(0,0,1,0.5),border = NA)
abline(v=-1)
text(-3,0.3,paste(round(100*pnorm(-1),2)," % en deça de ",round(-1,2)))
```




```
plot(0,0,type="n",main="",xlab="Valeurs",ylab="Densité",
     xlim=range(dd$x),ylim=range(dd$y))
lines(dd)
polygon(c(dd$x[dd$x < 1],rev(dd$x[dd$x < 1])),
        c(dd$y[dd$x < 1],rep(0,length(dd$y))[dd$x < 1]),
        col=rgb(0,1,0,0.5),border = NA)
abline(v=1)
text(-3,0.3,paste(round(100*pnorm(1),2)," % en deça de ",round(1,2)))
```

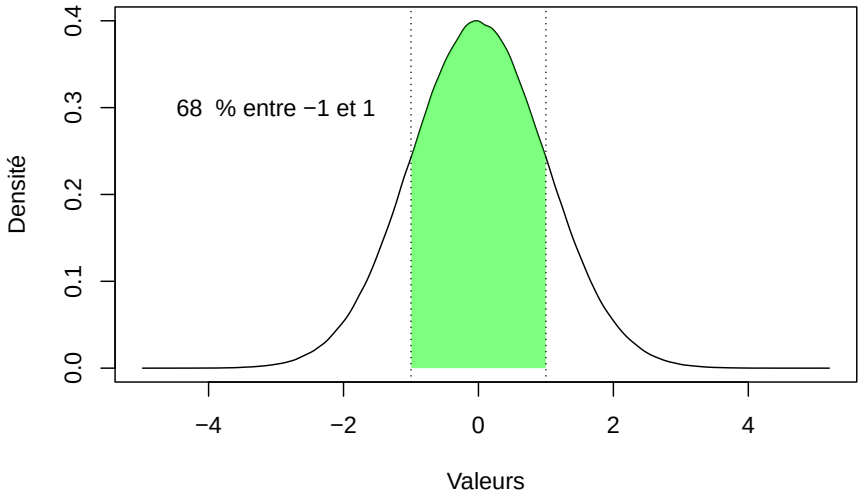



```
plot(0,0,type="n",main="",xlab="Valeurs",ylab="Densité",
     xlim=range(dd$x),ylim=range(dd$y))
lines(dd)
polygon(c(dd$x[dd$x < 1],rev(dd$x[dd$x < 1])),
        c(dd$y[dd$x < 1],rep(0,length(dd$y))[dd$x < 1]),
        col=rgb(0,1,0,0.5),border = NA)
abline(v=1)
text(-3,0.3,paste(100*round(pnorm(1),2)," % en deça de ",round(1,2)))
```



Soit entre la moyenne et les écart-types pour la loi normale, il y a 68 % des individus.

```
plot(0,0,type="n",main="",xlab="Valeurs",ylab="Densité",
     xlim=range(dd$x),ylim=range(dd$y))
lines(dd)
polygon(c(dd$x[dd$x > -1 & dd$x < 1],rev(dd$x[dd$x > -1 & dd$x < 1])),
        c(dd$y[dd$x > -1 & dd$x < 1],rep(0,length(dd$y))[dd$x > -1 & dd$x < 1]),
        col=rgb(0,1,0,0.5),border = NA)
abline(v=-1,lty=3)
abline(v= 1,lty=3)
text(-3,0.3,paste(100*round(pnorm(1) - pnorm(-1),2)," % entre -1 et 1"))
```



Entre les valeurs, les quantiles -2 et 2, il y a 95 % des individus.

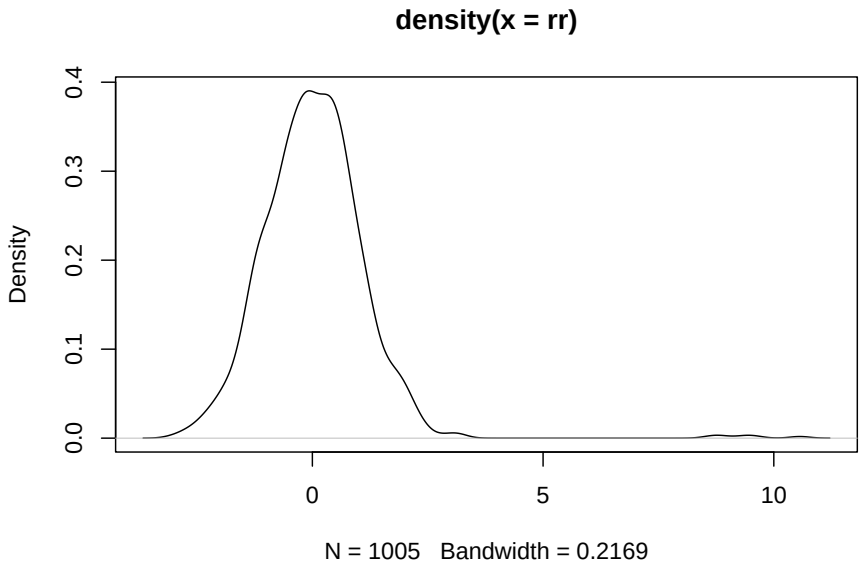
Les valeurs pour -2 et 2 sont respectivement -1,96 et 1,96.

2.6 La médiane et les quantiles usuelles

La médiane est le quantile le plus connu : il sépare 50% des individus à gauche et 50 % des individus à droite. Quand la fonction est symétrique, la médiane est égale à la moyenne.

La médiane est souvent utilisé comme indicatrice de tendance centrale dans le cas où la fonction est très asymétrique où s'il y a des valeurs extrêmes :

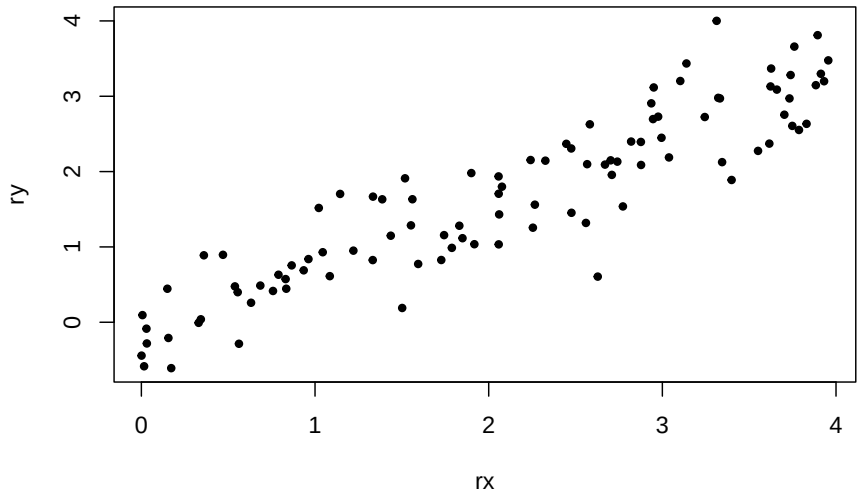
```
rr <- c(rnorm(1000),rnorm(5,10))
plot(density(rr))
```



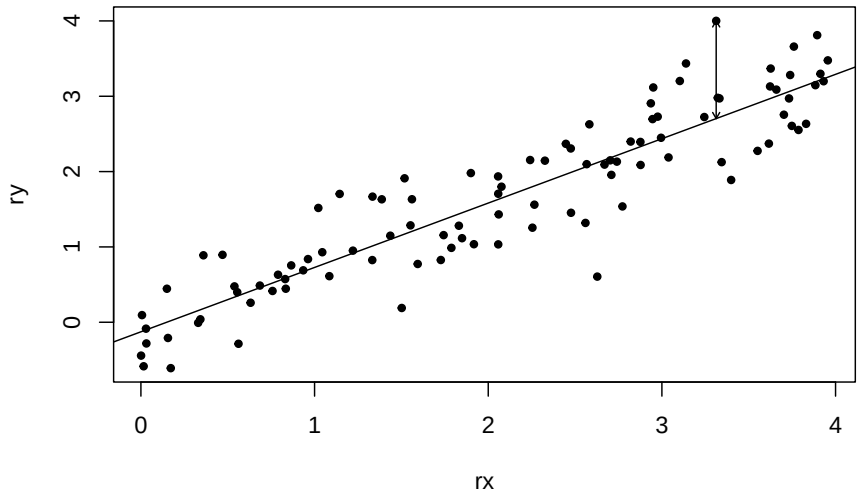
la moyenne est de 0.0364018 et l'écart-type de 1.181116. La moyenne sans les points extrêmes est de -0.0104208 tandis que la médiane est de -0.007937 avec et de -0.0146233.

Les autres quantiles quantiles les plus fréquents sont les quartiles : 0%, 25%, 50%, 75%, 100%.

```
set.seed(42)
rx <- runif(100,0,4)
ry <- 0.8*rx+rnorm(100,sd=0.5)
plot(rx,ry,pch=20)
```



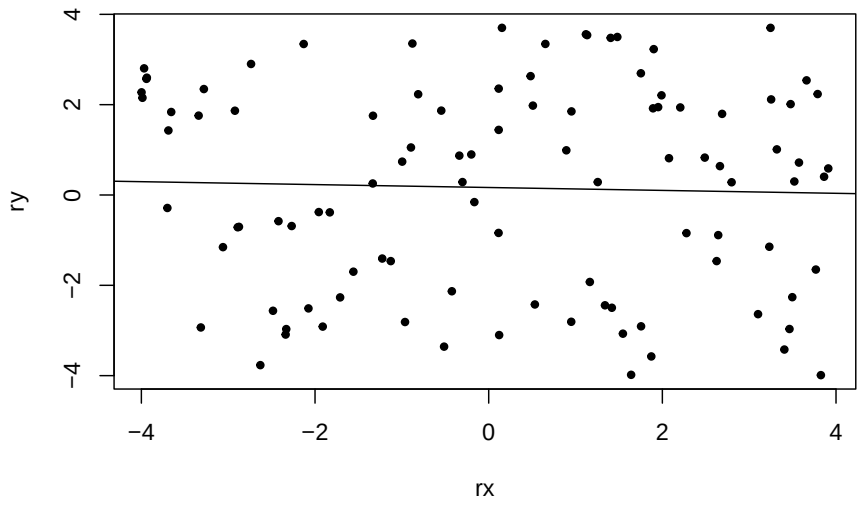

```
set.seed(42)
rx <- runif(100,0,4)
ry <- 0.8*rx+rnorm(100,sd=0.5)
plot(rx,ry,pch=20)
rr <- lm(ry~rx)
abline(rr)
pos <- which.max(ry)
  arrows(rx[pos],predict(rr)[pos],rx[pos],ry[pos],length=0.05,code = 3)
```



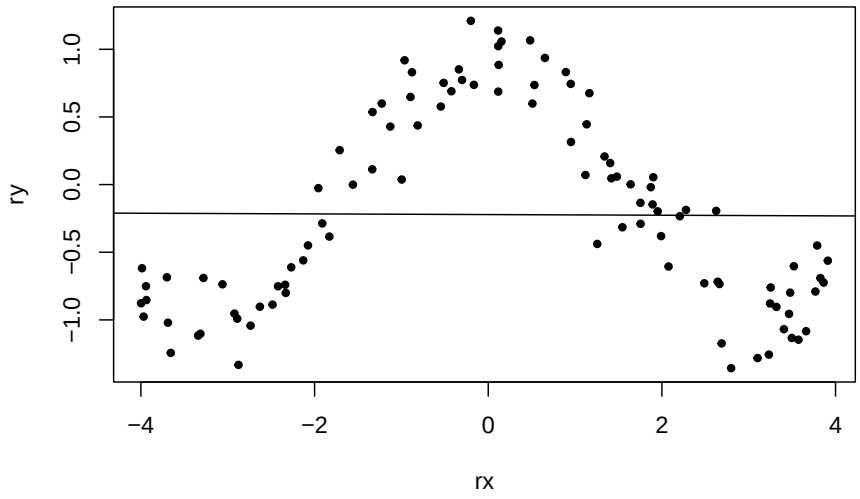

```
print(rr)
```

```
##
## Call:
## lm(formula = ry ~ rx)
##
## Coefficients:
## (Intercept)           rx
##     -0.1269       0.8545
```

```
set.seed(42)
rx <- runif(100,-4,4)
ry <- runif(100,-4,4)
plot(rx,ry,pch=20)
rr <- lm(ry~rx)
abline(rr)
```



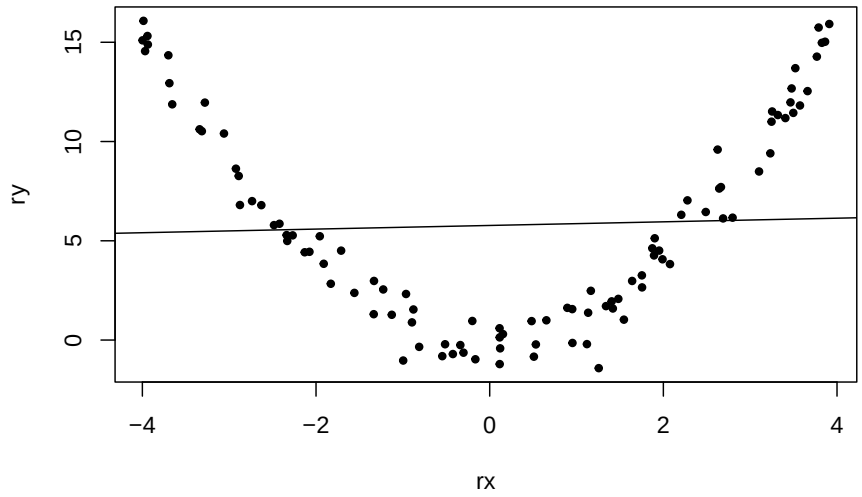

```
set.seed(42)
rx <- runif(100,-4,4)
ry <- cos(rx)+rnorm(100,sd=0.25)
plot(rx,ry,pch=20)
rr <- lm(ry~rx)
abline(rr)
```




```
print(rr)
```

```
##
## Call:
## lm(formula = ry ~ rx)
##
## Coefficients:
## (Intercept)           rx
##   -0.221598    -0.002439
```

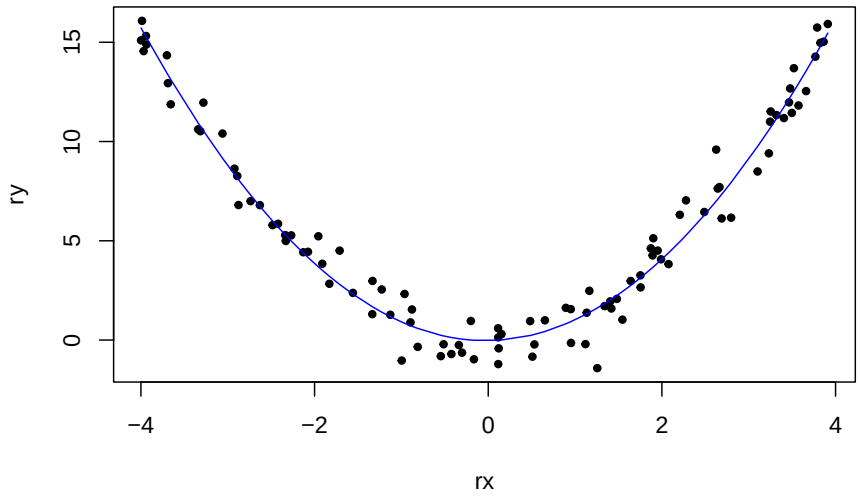
```
set.seed(42)
rx <- runif(100,-4,4)
ry <- rx^2+rnorm(100,sd=1)
plot(rx,ry,pch=20)
rr <- lm(ry~rx)
abline(rr)
```



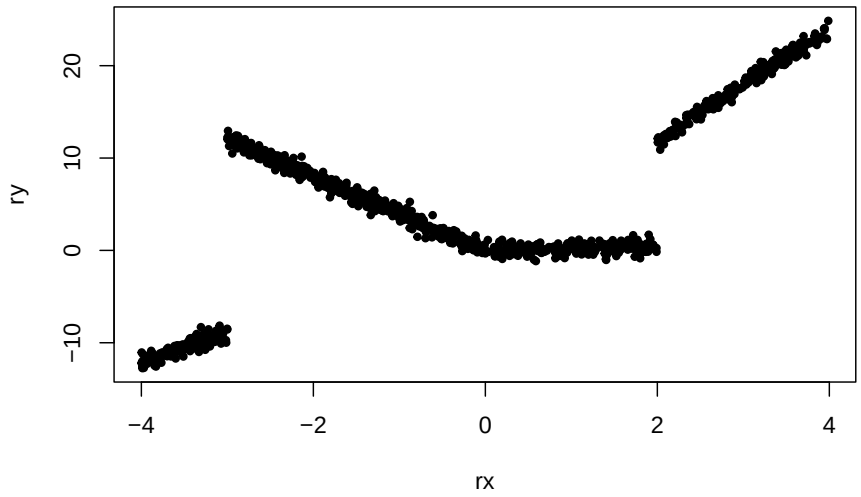

```
rr
```

```
##
## Call:
## lm(formula = ry ~ rx)
##
## Coefficients:
## (Intercept)           rx
##     5.77183      0.09141
```

```
a <- lm(ry ~ rx + I(rx^2))
plot(rx,ry,pch=20)
points(rx[order(rx)],predict(a)[order(rx)],col=rgb(0,0,1),type="l")
```

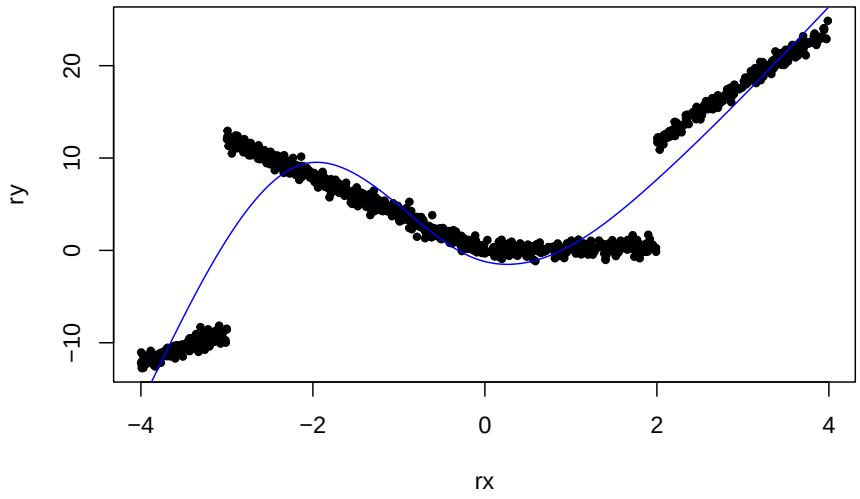



```
set.seed(42)
rx <- runif(1000,-4,4)
rx <- c(rx[rx >= -4 & rx < -3],rx[rx >= -3 & rx < 0],rx[rx>=0 & rx < 2],rx[rx >= 2 & rx <= 4])
ry <- c(3*rx[rx>= -4 & rx< -3],-4*rx[rx>= -3 & rx< 0],0.2*rx[rx >= 0 & rx<2],6*rx[rx>=2 & rx<=4])
ry <- ry+rnorm(length(ry),sd=0.5)
plot(rx,ry,pch=20)
```




```
rr <- lm(ry~rx)
```

```
library(splines)
a <- lm(ry ~ ns(rx, df = 4))
plot(rx,ry,pch=20)
points(rx[order(rx)],predict(a)[order(rx)],col=rgb(0,0,1),type="l")
```



2.8 Covariances

Les covariances sont les carrées des écarts entre les valeurs prises par deux variables aléatoires.

En fait ce sont les corrélations mais avec comme unités les unités naturelles des deux variables et non des écart-types comme unités.

Cela peut être très utile quand on calcule des corrélations entre des grands nombres et des petits nombre. Cela peut être difficile de *lire* les covariances donc on est amené à réduire les variables.

Par exemple :

Réduire les covariances équivaut à calculer la corrélation.

Chapter 3

Tests

3.1 Convergence vers la loi normale

Si on tire des échantillons aléatoirement, alors la moyenne de ces échantillons va converger vers une loi normale de moyenne *m* et d'écat-type *sigma*.

Central Limit Theorem

3.2 Test Z

Le test Z est le calcul de la position de la loi normale par rapport à ces quantiles.

On prend une variable que l'on centre/réduit si besoin. Alors si on moins de *xx* % de chances que la moyenne se trouve entre les valeurs (positives et négatives) des quantiles alors la différence entre la moyenne de la loi normale réduite est différente de m, une valeur théorique fixée.

Schématiquement on utilise les quantiles de la loi normale. Si la moyenne se distribue comme une loi normale alors, entre doit se trouver avec une confiance de 95% entre les deux quantiles qui *englobe* 95% des observations d'une loi normale.

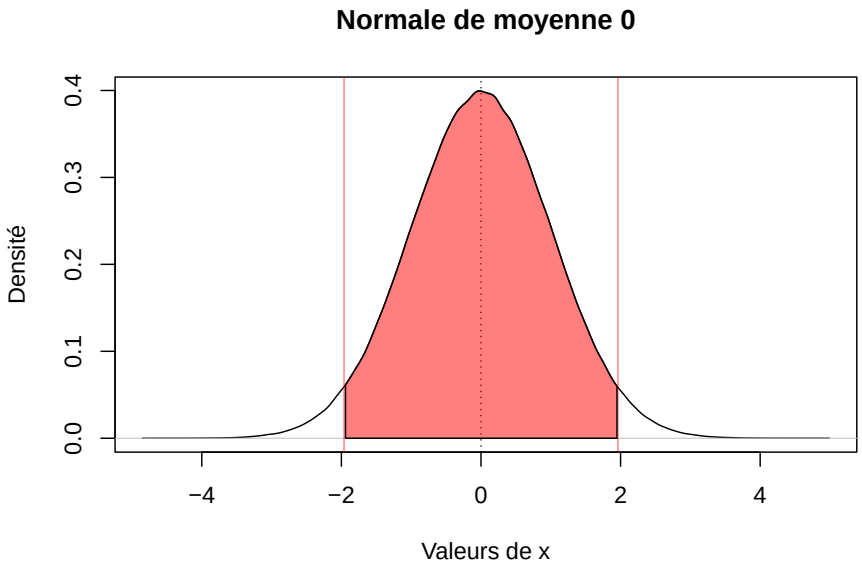
On a :

100 - 95 = 5%

Comme on tient compte des observations à gauche et à droite alors on a le quantile de 2,5% à gauche et 97,5% à droite.

```
y=rnorm(1000000,0,1)
dd=density(y)
plot(dd,main="Normale de moyenne 0",xlab="Valeurs de x",ylab="Densité")
abline(v=0,lty=3)
```

```
abline(v=0-1.96*sd(y),lty=1,col=rgb(1,0,0,0.5))
abline(v=0+1.96*sd(y),lty=1,col=rgb(1,0,0,0.5))

polygon(c(dd$x[dd$x>-1.96 & dd$x<1.96],rev(dd$x[dd$x> -1.96 & dd$x<1.96])),c(rep(0,leng
```

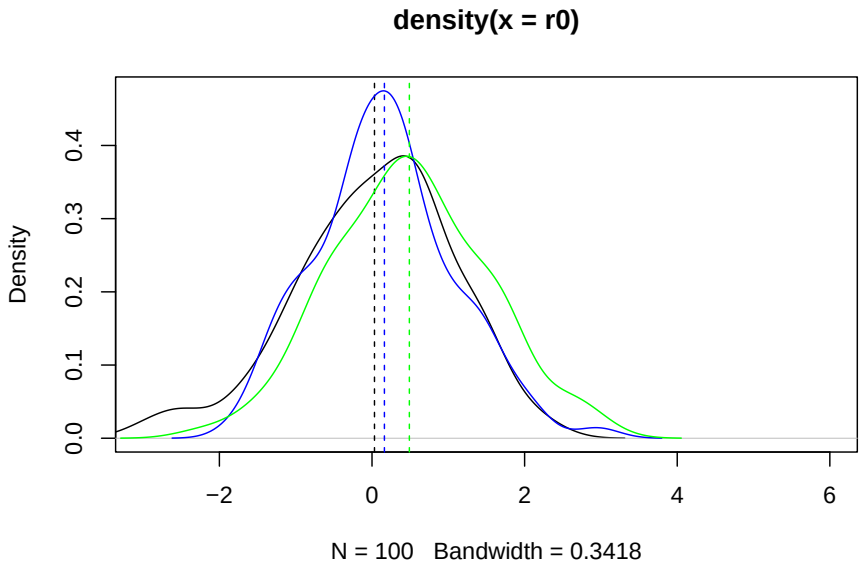


Le test est donc :

Si la distribution de la moyenne suit une loi normale centrée/réduite, sous l'hypothèse nulle que $m = m0$ alors si m est compris entre -1,96 et 1,96 alors m n'est pas différent de $m0$.

Sinon m est différent de $m0$ avec un seuil de 5%.

```
set.seed(42)
r0 <- rnorm(100)
r1 <- rnorm(100,mean=0.25)
r3 <- rnorm(100,mean=0.5)

plot(density(r0),xlim=c(-3,6),ylim=range(c(density(r0)$y,density(r1)$y,density(r3)$y)))
abline(v=mean(r0),lty=2)

points(density(r1),col="blue",type = "l")
abline(v=mean(r1),lty=2,col="blue")

points(density(r3),col="green",type = "l")
abline(v=mean(r3),lty=2,col="green")
```




```
z <- sqrt(length(r1))*(mean(r1)-0)/sd(r1)
z
```

```
## [1] 1.797402
```

```
abs(z) > 1.96
```

```
## [1] FALSE
```

```
z <- sqrt(length(r3))*(mean(r3)-0)/sd(r3)
z
```

```
## [1] 4.814414
```

```
abs(z) > 1.96
```

```
## [1] TRUE
```

Question : que se passe-t-il si on fait un test seulement à droite ou seulement à gauche ? C'est à dire si on s'intéresse non pas au cas où m est autour de $m0$ mais si m est plus grand que $m0$ ou plus petit que $m0$.

3.3 Test de Student

Le test de Student est similaire au test Z, il est même identique dans certains cas.

Les test de Student a deux fonctions :

- tester si une variable a une moyenne différente de m0.
- test si les moyenne de deux variables sont différentes

On reconnait le premier cas qui est le test Z. D'ailleurs ont retrouve la même valeur pour le test :

```
t.test(r1)
```

```
##
##  One Sample t-test
##
## data:  r1
## t = 1.7974, df = 99, p-value = 0.07532
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01689135  0.34192394
## sample estimates:
## mean of x
## 0.1625163
```

Le test de Student est proche de l'ANOVA. En fait les résultats sont identiques quand il y a l'argument *var.equal=TRUE*.

Il signifie que les variances des variables sont identiques.

Un exemple du test de Student, n'est pas sur deux variable séparée mais entre deux catégories d'individus qui forme chacun une population :

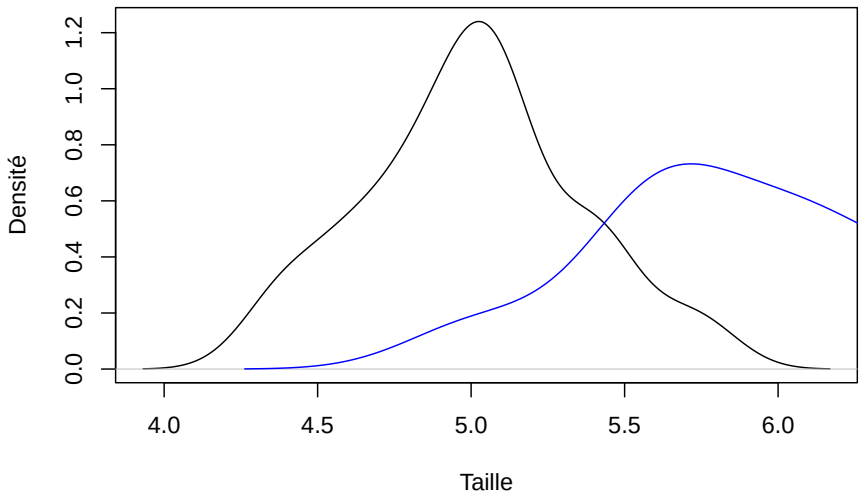
```
t.test(
  iris$Sepal.Length[iris$Species=="setosa"],
  iris$Sepal.Length[iris$Species=="versicolor"],
  var.equal = T
)
```

```
##
##  Two Sample t-test
##
## data:  iris$Sepal.Length[iris$Species == "setosa"] and iris$Sepal.Length[iris$Specie
## t = -10.521, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1054165 -0.7545835
## sample estimates:
## mean of x mean of y
##     5.006     5.936
```

Le résultat est assez intuitif si on regarde le graphique des deux distributions :

```
plot(density(iris$Sepal.Length[iris$Species=="setosa"]),type="l",main="",
     xlab="Taille",ylab="Densité")
```

```
points(density(iris$Sepal.Length[iris$Species=="versicolor"]),type="l",col="blue")
```



Pour résumer un test statistique est caractérisé :

- par des prérequis de travail (le plus souvent la normalité, égalité des variances, etc.).
- des hypothèses pour le test qui définir les conclusions que l'on peut tirer du test. A votre niveau les hypothèses sont le plus souvent binaires. Oui/non : H0 contre H1.
- le calcul de la valeur du test qui va être comparé aux "tables" de valeurs
- le calcul des degrés de liberté
- la conclusion du test : acceptation/rejet de H0 et idem pour H1.

Certaines statistiques dites "bayésiennes" fonctionnent différemment. Elles touchent surtout la façon de définir et de calculer les tests. La méthode décrite ici est appelé, à l'opposé de bayésiennes, la méthode fréquentiste.

3.4 Erreur de Type I et II

L'erreur de type I est : une erreur de type I survient dans un test d'hypothèse statistique lorsqu'une hypothèse nulle, qui est en réalité vraie, est rejetée par erreur. Les erreurs de type I sont également connues sous le nom de "faux

positifs", elles représentent la détection d'un effet positif alors qu'il n'existe aucun effet en réalité.

L'erreur de type II est : le risque de ne pas démontrer que deux groupes sont différents alors qu'ils le sont dans la réalité.

La puissance est 1 - l'erreur de type II.

Par exemple, dans le cadre d'une étude randomisée en double aveugle pour le développement d'un nouveau médicament, le risque de 2e espèce peut être la probabilité de conclure qu'un médicament n'est pas meilleur qu'un placebo alors qu'il l'est. Dans ce cas, la puissance du test serait la probabilité de conclure que le médicament est meilleur que le placébo, ce qui est vrai.

3.5 Tests "paramétriques" et non paramétriques

Dans le premier cas, tests paramétriques, ce sont les tests que l'on vient de voir. Il repose sur des hypothèses de distribution : en l'occurence ici que les données suivent une loi normale avec des paramètres : la moyenne et l'écart-type. Le test ne "fonctionne" donc que dans le cas où ces trois élements sont présents et corrects statistiquement.

Par exemple, si la distribution est très asymétrique, l'écart-type, la moyenne et la loi normale ne sont pas au rendez-vous alors il faut se tourner vers d'autres tests, en général des tests dits non-paramètriques.

Ces tests ne font pas d'hypothèse sur la distribution. Par exemple pour le test de Student, l'équivalent est le test de Wilcoxon-Mann-Whitney.

Ce dernier pour illustrer le propos est basé sur les rangs des observations plutôt que sur la valeur. Comme la médiane, cela rend le test plutôt robuste à l'asymétrie et aux valeurs extrèmes.

L'inconvénient de ces tests est que pour un type I donné la puissance est plus faible : on a des chances plus faibles de détecter un vrai positif qu'avec

3.6 Précautions à prendre quand on travaille avec des tests

- Ces préquis sont à **vérifier** avant de faire le test
- Il faut comparer le nombre de degrés de liberté avec le nombre d'observations. En effet il y a des "rules of thumb" qui définissent le nombre de degrés de liberté en fonction du nombre d'observations. Par exemple pour les analyses structurales il faut de 20 à 40 observations minimums par degré de liberté.

- Le choix est binaire. la p-value **ne donne pas de renseignements sur la "force" du test**.
- Le choix de la valeur seuil de la *p-value* doit être fait en amont et doit être contrôlé par des procédures statistiques si vous calculez de nombreux tests : cela s'appele correction de Bonferroni, Tukey, etc.