Partiel 2019 _ Matisse Landais

Je certifie que ce travail (ce qui figure dans la copie et les documents déposés ce jour sous Chamilo) est le fruit de mon travail personnel exclusivement.

Exercice 2 : étude statistique globale

On a 3 jeux de données (groupe1, groupe2 et groupe3)

```
groupe1 <- c(15.3,9.3,13.3,14.0,7.3,7.3,16.0,16.7,13.3,12.7,14.7,12.0,9.3,11.3,8.7,16.0,8.7,11.3,12.0,1
groupe2 <- c(10.0,12.0,13.3,13.3,12.7,7.3,14.7,8.7,11.3,15.3,14.7,12.0,18.0,10.7,14.0,12.7,10.0,15.3,16
groupe3 <- c(7.3,11.3,4.0,10.7,10.0,11.7,17.3,7.3,13.7,11.3,13.3,10.0,16.0,11.7,8.7,8.7,12.0,13.3,10.0,9
```

A : calcul des moyenne, estimation générale

```
cat("Moyenne du groupe 1 : ", mean(groupe1), " et variance : ", sd(groupe1))
```

```
## Moyenne du groupe 1 :  12.64  et variance :  2.999379
```

```
cat("Moyenne du groupe 2 : ", mean(groupe2), " et variance : ", sd(groupe2))
```

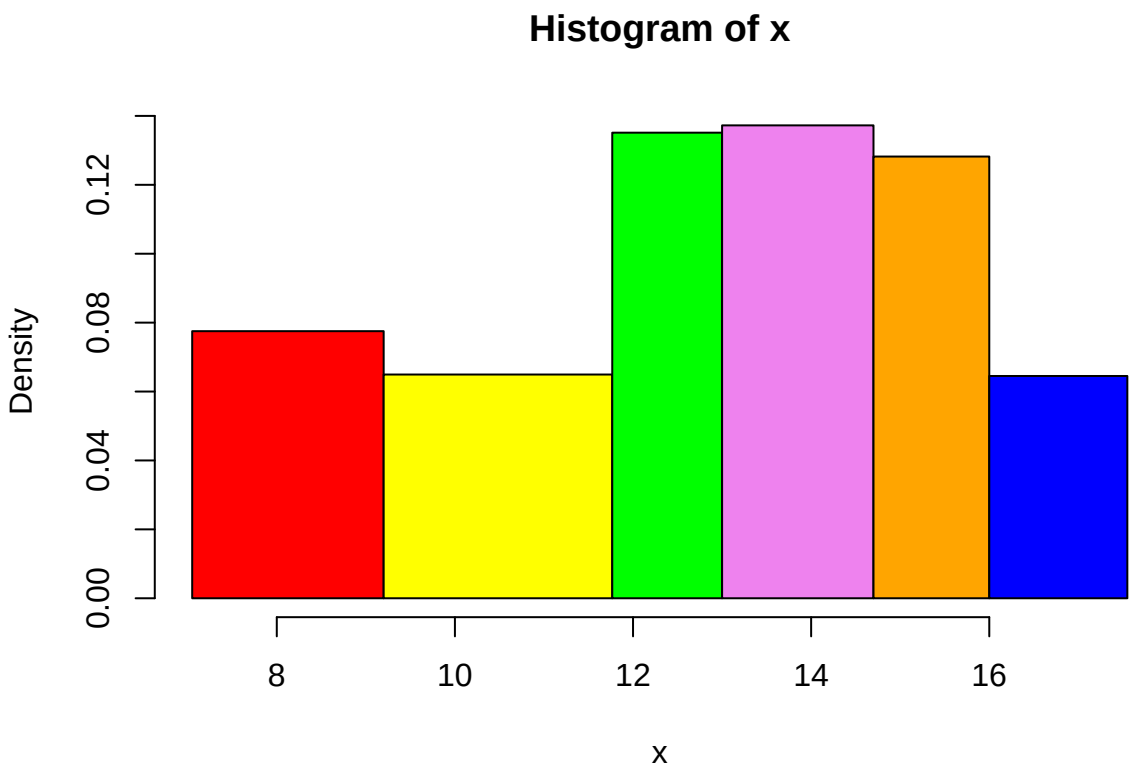
```
## Moyenne du groupe 2 :  12.57667  et variance :  2.75627
```

```
cat("Moyenne du groupe 3 : ", mean(groupe3), " et variance : ", sd(groupe3))
```

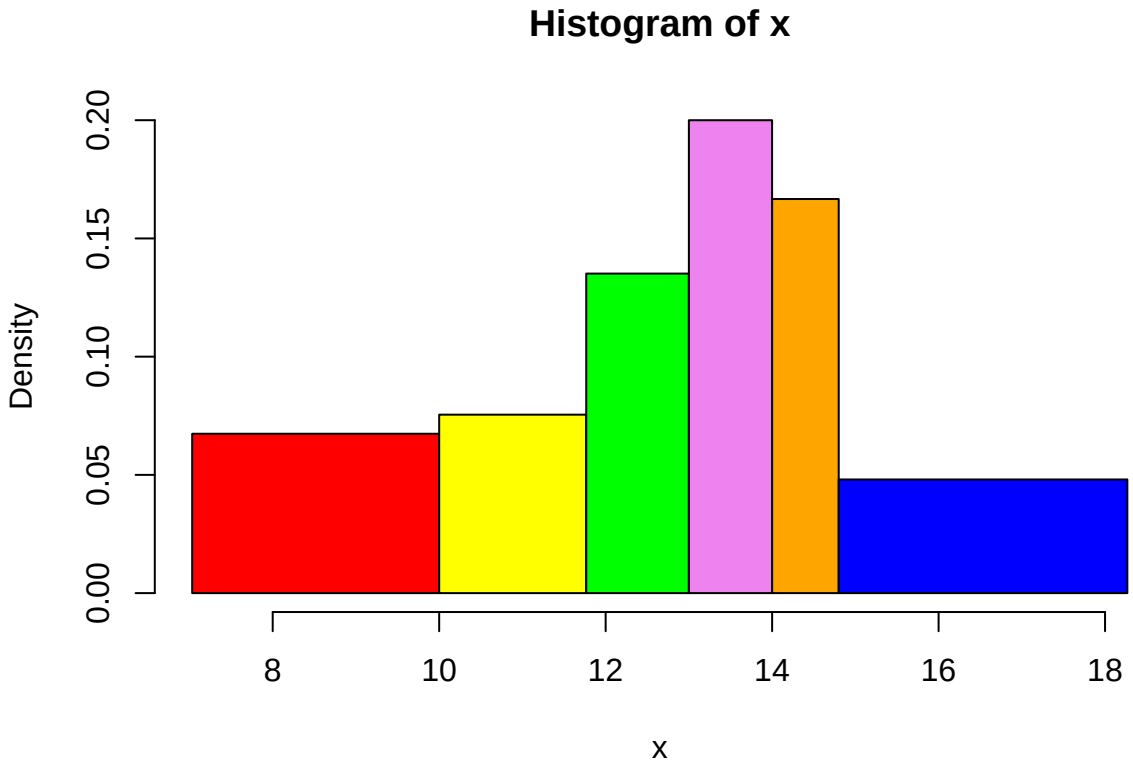
```
## Moyenne du groupe 3 :  11.28  et variance :  2.963386
```

La moyenne du groupe 3 est légèrement plus faible. On peut maintenant estimer la loi que pourrait suivre ces notes grâce à un graphe de probabilité et un histogramme.

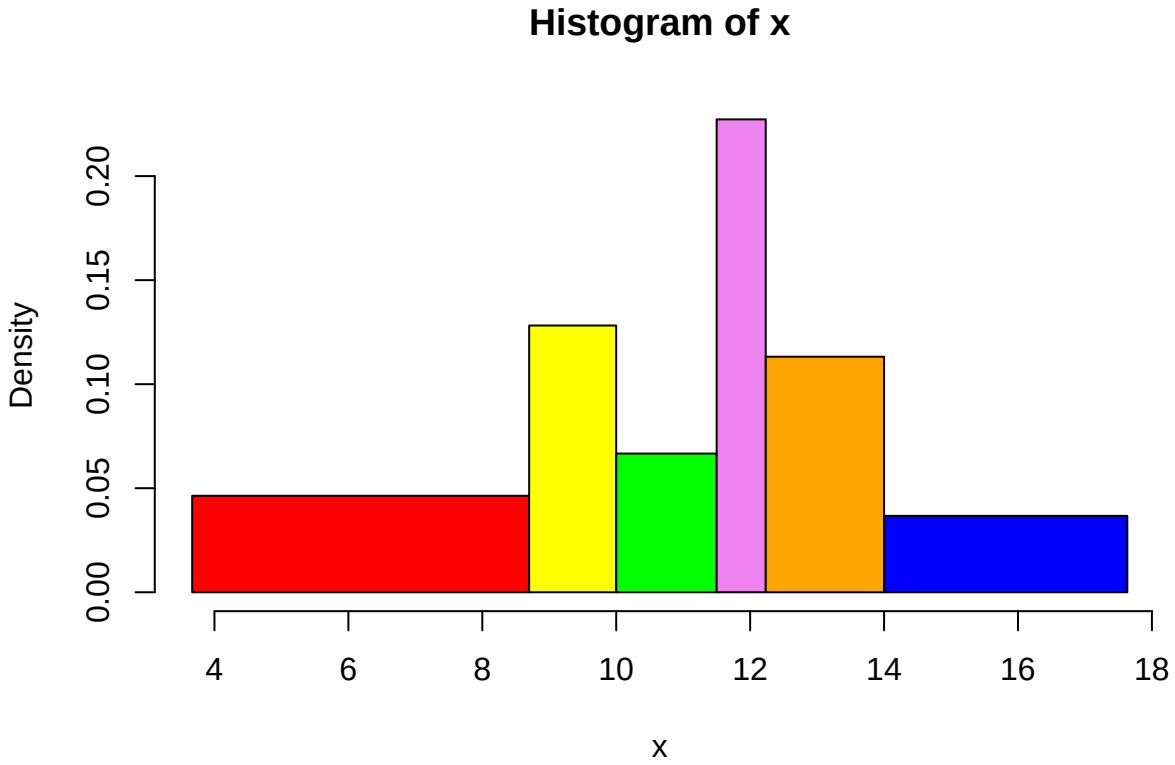
```
histoeff <- function(x, xlim=NULL, ...)
{
  sx <- sort(x)
  n <- length(x)
  k <- round(log(n)/log(2)+1)
  rangex <- max(x)-min(x)
  quantileVoulu <- quantile(x, seq(1,k-1)/k, max(x))

  breaks <- c(min(x)-0.025*rangex, quantileVoulu, max(x)+0.025*rangex)
  col <- 0
  if (is.null(xlim)) xlim<-c(breaks[1], breaks[k+1])
  colors = c("red", "yellow", "green", "violet", "orange", "blue", "pink", "cyan")

  hist(x, breaks=breaks, col=colors, xlim=xlim, probability=T)
}

histoeff(groupe1)
```




```
histoeff(groupe2)
```

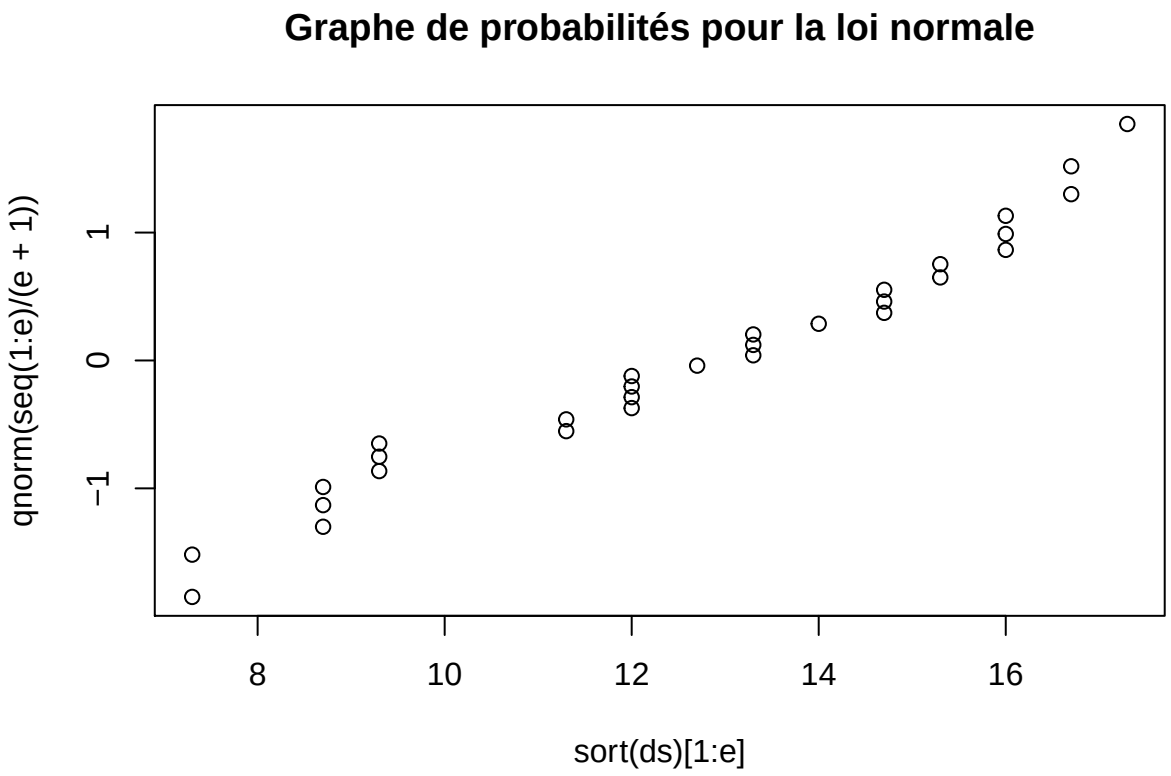



```
histoeff(groupe3)
```

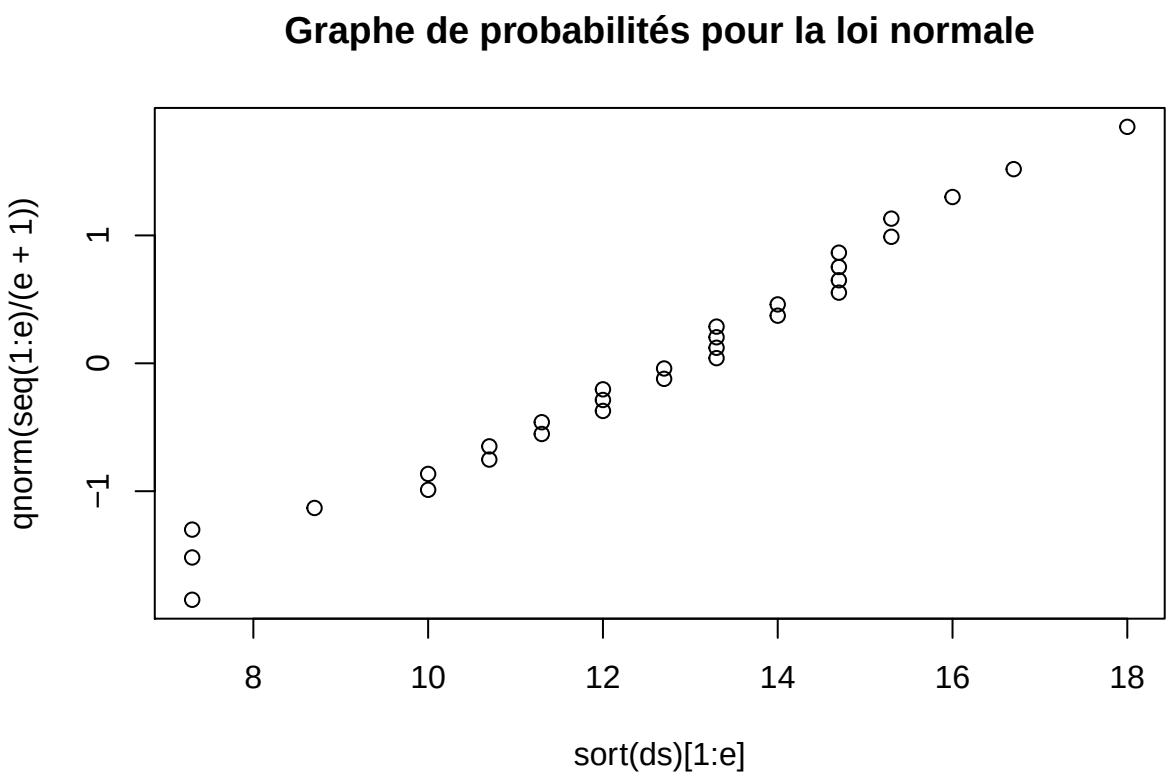


Cela ressemble à une loi normale. Pour être sur, on peut tracer le graphe de probabilité pour la loi normale avec les 3 groupes.

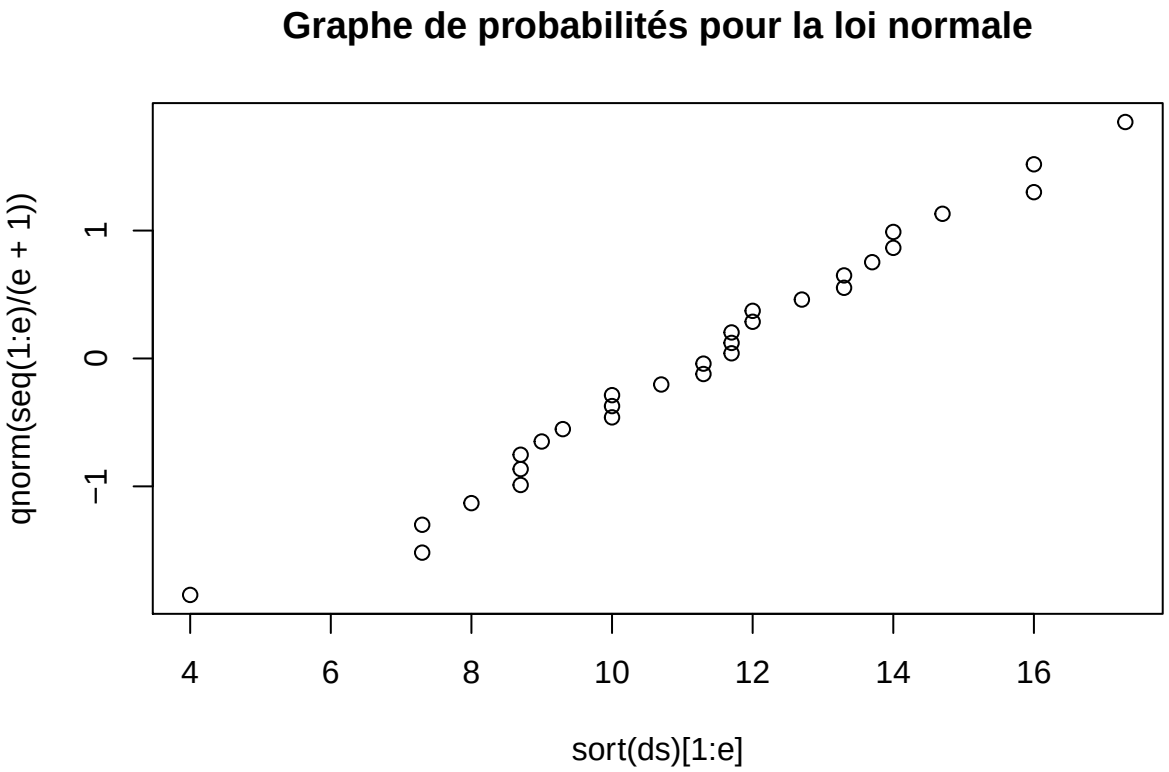
```
qqplotNormale <- function(ds)
{
  e = length(ds)
  plot(sort(ds)[1:e], qnorm(seq(1:e)/(e+1)), main="Graphe de probabilités pour la loi normale")
}
qqplotNormale(groupe1)
```




```
qqplotNormale(groupe2)
```




```
qqplotNormale(groupe3)
```



Les graphes de probabilités sont très semblables à une loi normale (voir les démonstrations sur les graphes de probabilités)

On cherche

$$P(X < 10)$$

tel que

$$X \sim N(\mu, \sigma^2)$$

(càd la densité sous la courbe). On test pour les 3 groupes puis on peut fusionner car d'après les résultats précédents, les données semblent suivre des lois normales.

```
dnorm(10,mean(groupe1), sd(groupe1))
```

```
## [1] 0.0902922
```

```
dnorm(10,mean(groupe2), sd(groupe2))
```

```
## [1] 0.09350147
```

```
dnorm(10,mean(groupe3), sd(groupe3))
```

```
## [1] 0.1226333
```

Résultat

On obtient une probabilité assez faible d'avoir une note inférieur à la moyenne. En effet, cette probabilité est respectivement de 9%, 9% et 12%. Le 3ème groupe semble avoir des notes plus faibles que les autres groupes, au vue de la moyenne. Vérifions ça avec des intervalles de confiance.

Intervalles de confiances pour les groupes

Précédemment, nous avons pu voir que les notes des groupes semblaient suivre des lois normales. Bien que la moyenne et l'écart type soient des indicateurs intéressants pour connaître l'étendu des notes, on peut réaliser des intervalles de confiance sur la moyenne pour savoir les valeurs qu'elle pourra prendre en général. On aurait aussi pu le faire sur la variance/l'écart type, pour savoir si les notes se resserrent autour de la moyenne ou non (les disparités de réussite sur la classe). Les intervalles de confiance sont intéressants ici car ils permettent, dans notre cas, de savoir si dans le cas où nous aurions beaucoup plus de groupes de TD, si il n'y aurait pas un déséquilibre en moyenne dans la notation par les intervenants/professeurs.

Simulations

On peut donc faire des intervalles de confiance au seuil 20% et 10% pour savoir si un maximum d'élèves valideront la matière ou non.

```
# à 20%
t.test(groupe1, conf.level=0.80)
```

```
##
##  One Sample t-test
##
## data:  groupe1
## t = 23.082, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 80 percent confidence interval:
##  11.92185 13.35815
## sample estimates:
## mean of x
##     12.64
```

```
t.test(groupe2, conf.level=0.80)
```

```
##
##  One Sample t-test
##
## data:  groupe2
## t = 24.992, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 80 percent confidence interval:
##  11.91672 13.23661
## sample estimates:
## mean of x
##  12.57667
```

```
t.test(groupe3, conf.level=0.80)
```

```
##
##  One Sample t-test
##
## data:  groupe3
## t = 20.849, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 80 percent confidence interval:
##  10.57046 11.98954
## sample estimates:
```

```
## mean of x
##     11.28
```

```
# à 10%
t.test(groupe1, conf.level=0.90)
```

```
##
##  One Sample t-test
##
## data:  groupe1
## t = 23.082, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  11.70954 13.57046
## sample estimates:
## mean of x
##     12.64
```

```
t.test(groupe2, conf.level=0.90)
```

```
##
##  One Sample t-test
##
## data:  groupe2
## t = 24.992, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  11.72163 13.43171
## sample estimates:
## mean of x
##  12.57667
```

```
t.test(groupe3, conf.level=0.90)
```

```
##
##  One Sample t-test
##
## data:  groupe3
## t = 20.849, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  10.36071 12.19929
## sample estimates:
## mean of x
##     11.28
```

Conclusion

Les résultats permettent donc de voir qu'il y a effectivement une différence entre les 3 groupes (le 3ème groupe ayant une notation qui semble plus sévère dans la moyenne) et l'étendu des moyennes des étudiants. Les étudiants du groupe 3 auront une moyenne entre 10.57 et 11.98 au seuil de 10%, ce qui est 1 point en dessous des deux autres groupes. La question reste à savoir si les étudiants souhaitent simplement valider ou si la note importe.

Imaginons maintenant que l'on réhausse les notes du groupe 3. Ajouter 1 à chaque étudiant revient à ajouter 1 à la moyenne (part linéarité). On peut ensuite refaire l'intervalle de confiance pour le groupe 3 et conclure.

```r
groupe3 <- groupe3 + 1
t.test(groupe3, conf.level=0.9)
```

```
##
##  One Sample t-test
##
## data:  groupe3
## t = 22.697, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  11.36071 13.19929
## sample estimates:
## mean of x
##     12.28
```

En ajoutant 1 à chaque élève du groupe 3, on obtient des moyennes et intervalles de confiance similaire aux deux autres groupes. Un outil plus puissant pour décider s'il faut ajuster ou pas les notes est le test d'hypothèses. Ici, ce serait un test composite avec >10 ou <= 10 (la note).

Cosa Nostra ... et le problème du dé

Le test du KhiDeux est un test d'adéquation. Il permet de savoir si au vu des résultats (ici les occurences / valeur du dé), les observations appartiennent ou non à une certaine loi. Ici, si le dé est équilibré, la loi que l'on veux tester (savoir si les données la suivent), est une loi uniforme. Donc l'hypothèse nulle est H0 : le dé est équilibré car cela correspond à la loi uniforme.

Le risque le plus important serai de décider H0 (càd le dé est équilibré => Don Pasquale ne meurt pas) alors que la vérité serait H1 ! (Don Pasquale est un tricheur).

- H0 : le dé est équilibré
- H1 : le dé n'est pas équilibré

Décision/Vérité	H0	H1
H0	$1-\alpha$	$1-\beta$
H1	α	β

Analyse du premier test

Avec la fonction chisq.test, on veut tester si le dé est effectivement équilibré. La p-valeur nous montre que si on veut rejeter H0, c'est à dire rejeter l'hypothèse comme quoi le dé est équilibré, est forte. Cela veut dire que l'on a 73% de chances de se tromper en disant que le dé n'est pas équilibré. Ce qui est beaucoup trop fort pour prendre une décision au vu des résultats. Heureusement, Claudio décide de prendre plus de valeurs pour essayer d'écarter au maximum l'aléatoire.

Analyse du second test

Ici, Claudio a multiplié les occurences par un facteur 5. C'est judicieux car l'aléatoire, même s'il réserve des surprises, peut être grandement mis de côté si on prend ces observations. La p-valeur est très faible, ce qui n'est pas une surprise car on prend beaucoup moins de risque en rejetant l'hypothèse nulle à tort. En effet, si on prend ces observations on a 1% de "chance" de se tromper en disant que le dé n'est pas équilibré. Le risque de tuer à tort Don Pasquale est de 1%, même si les observations ne sont pas réelles car il a simplement multiplié les occurences par 5.

Voir la généralisation mathématique sur ma feuille

Conclusion

Les deux tests sont des tests d'adéquation entre le jeu de données et une loi uniforme. Le premier, même s'il représente la réalité (l'homme de main n'allait pas lancer 300 fois le dé), est peu probant pour tirer une conclusion. On a de grandes "chances" de tuer Don Pasquale à tort. Le second est beaucoup plus dur à mettre en place mais on peut rapidement savoir si le dé est truqué. En l'occurence, si l'homme de main de Vito avait pu lancer le dé 300 fois et grâce aux resultats obtenus, Don Paquale ne serait plus là à l'heure qu'il est.