Haguenau

JcB

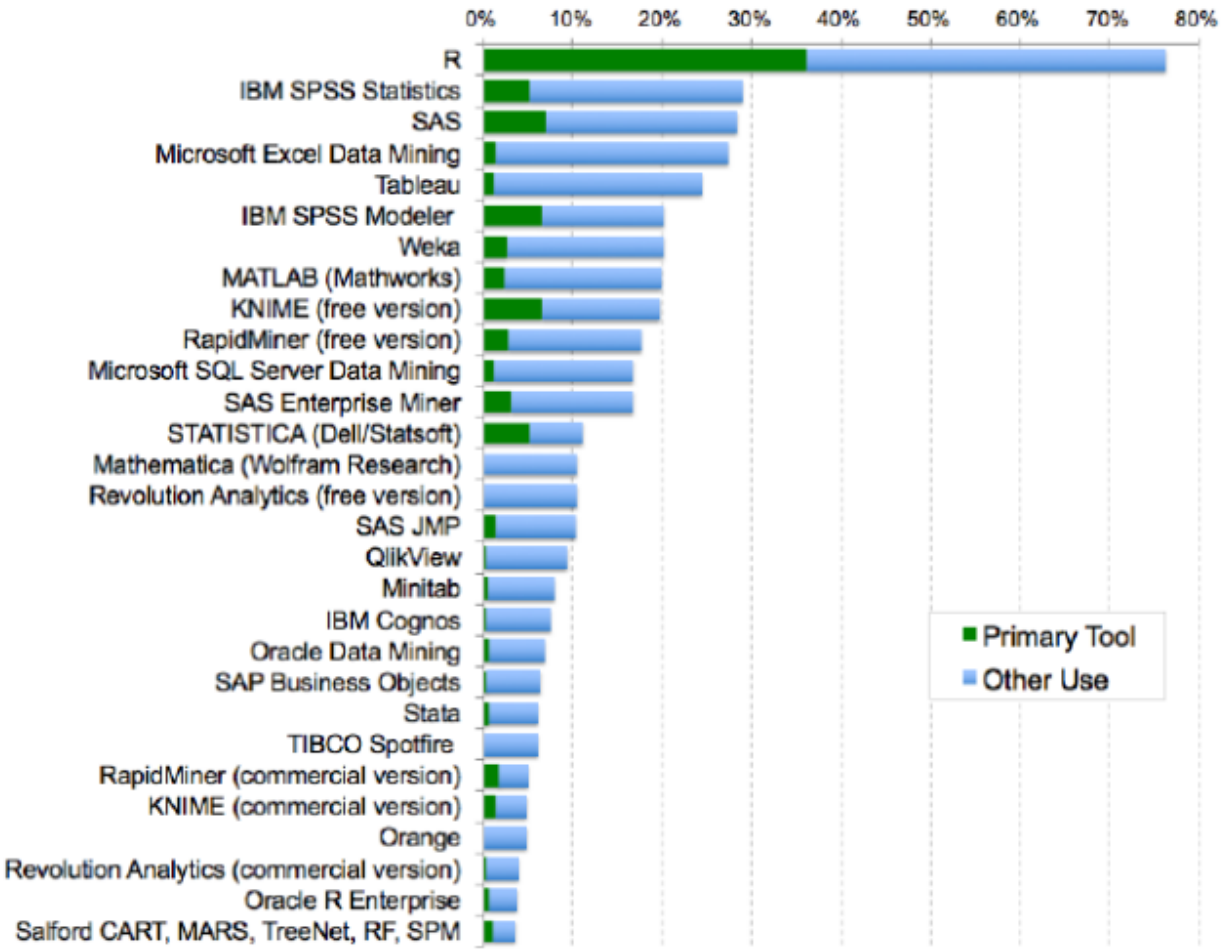
14/10/2015

Contents

1 Objectifs

- savoir utiliser un logiciel de statistiques
- savoir collecter correctement des données (tableur)
- transmettre les données au logiciel
- appliquer une démarche statistique

2 Le logiciel R parmis les autres logiciels statistiques



source:

3 Organiser son travail

- Démarrer RStudio
- Créer un nouveau _Projet___ dans un nouveau répertoire (directory): File/New project -> New directory
- Créer un sous répertoire **Data** qui servira à stocker les données
- Créer un sous répertoire **Cours_Stat_2015**

4 Savoir utiliser un logiciel de statistiques

- Utilisation de **R** (chercher *CRAN* The Comprehensive R Archive Network).
- c'est à la fois un language de programmation (on peut écrire ses propres routines) et un logiciel statistique.
- **R** est *libre*, *gratuit*, *multiplateforme*, *complet*, *évolutif* grâce à une énorme bibliothèque de fonctions appelées **Pakages** (environ 6000 à ce jour).
- On peut l'utiliser nativement ou par l'intermédiaire d'un IDE appelé **RStudio**.
- *RStudio* utilise le concept de **recherche reproductible** et permet de mettre en place une chaîne de production allant de la saisie des données à la production d'un document (mémoire, thèse, etc.) publiable.

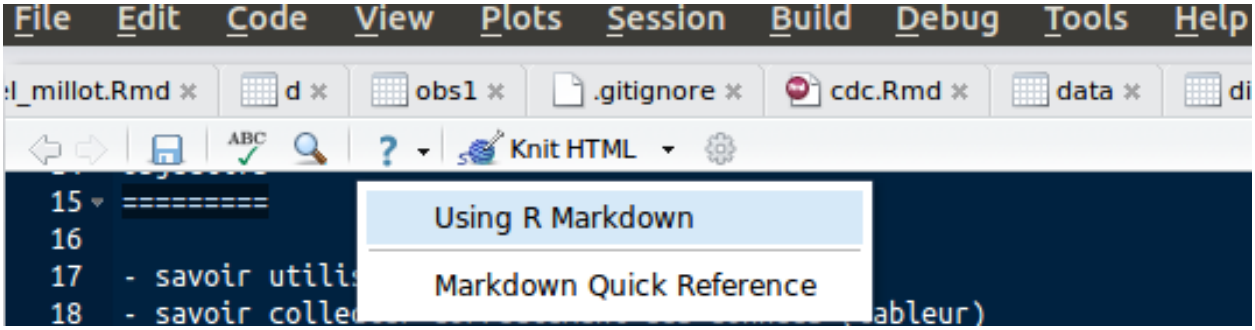
4.1 Caractéristiques de RStudio

4.1.1 4 fenêtres

- Ouvrir une nouvelle page de travail: file -> New File -> R Markdown...
- **Enregistrer** le document dans le dossier **Cours_Stat_2015**

4.1.2 RMarkdown

Un traitement de texte simple pour prendre des **notes**



4.1.3 les "Chunks"

Aujourd'hui, je crée mon premier programme R en fusionnant mon texte, mes calculs et graphiques dans le même document grace aux *chunks*:

```
print("Hello, R")
```

```
## [1] "Hello, R"
```

```
a <- 2 + 2
a
```

```
## [1] 4
```

Il semble que 2 + 2 fassent 4

4.1.4 Prouire un document

A partir de ce document écrit avec *markdown* je peut produire directement:

- un document **Html** pour un navigateur (site internet, blog. . .)
- un document **Pdf** non modifiable à distribuer
- document **Word** ou **Libre Office** modifiable
- un _diaporama___ à projeter
- un fichier des **graphiques** crées pouvant ˆetre copiés/collés dans n'importe quel document.

5 La démarche statistique

5.1 1. Collecte de données

et leur mise en forme pour être exploitées, c'est 80% du travail.

5.2 2. Statistiques descriptives

Nous allons générer des statistiques qui résument les données de façon concise, et d'évaluer les différents des moyens pour visualiser les données.

5.3 3. L'analyse exploratoire des données

Nous allons rechercher des modèles, les différences, et d'autres caractéritiques qui répondent aux questions nous sommes intéressés à. Dans le même temps, nous allons vérifier les incohérences et identifier limitations.

5.4 4. estimation

Nous allons utiliser les données à partir d'un échantillon pour estimer les caractéristiques de la population générale.

5.5 5. Tests d'hypothèses

Où l'on voit les effets apparents, comme une différence entre deux groupes, nous évaluerons si l'effet pourrait être dû au hasard.

6 La collecte des données

Application: analyse des friandises contenues dans un paquet de M&M's fabriquées à Haguenau.

Noter dans un tableur:

- nom
- couleur: ___R___ed, ___Y___ellow, ___G___reen, ___B___lue, ___M___aroon, ___B___lack
- nombre
- aspect: ___E___bréché, ___F___endu, ___P___arfait, ___N___on marqué

7 Manipulation de R

- dans **R** on stocke des données dans des conteneurs appelés **variables** que l'un désigne par un **nom**: n, x, tartampion, ...
- pour relier la variable **n** à une valeur, on utilise le symbole d'affectation "<-"

```
n <- 10
n * n
```

```
## [1] 100
```

```
b <- n * n / 5
```

7.1 Créer un vecteur de données

Un vecteur est un groupe de données créé avec l'opérateur de ___c___oncaténation

```
ages <- c(25, 18, 21, 21, 23, 22, 23, 18, 25, 19, 22, 22, 22, 22)
ages
```

```
##  [1] 25 18 21 21 23 22 23 18 25 19 22 22 22 22
```

```
n <- 1:10
```

7.2 Tableau de données

- Un tableau rectangulaire de données constitue un **dataframe**
- une feuille de tableur au format **.csv** est un exemple de *dataframe*

```
data <- data.frame(Seatbelts)
head(data)
```

```
##   DriversKilled drivers front rear   kms PetrolPrice VanKilled law
## 1           107    1687   867  269  9059   0.1029718        12   0
## 2            97    1508   825  265  7685   0.1023630         6   0
## 3           102    1507   806  319  9963   0.1020625        12   0
## 4            87    1385   814  407 10955   0.1008733         8   0
## 5           119    1632   991  454 11823   0.1010197        10   0
## 6           106    1511   945  427 12391   0.1005812        13   0
```

7.3 Paramètres statistiques de base

- variables quantitatives (je peux les additionner): age, poids, taille....
- variables qualitatives (je peux les dénombrer sans équivoque): sexe, statut marital, CSP, couleur des cheveux...
  - nominale: l'ordre n'a pas d'importance: sexe
  - ordinales: l'ordre est important: échelle de Likert
- ATTENTION: variables qualitatives qui se présentent comme des variables quantitatives: score de Glasgow

7.3.1 taille

```
n <- length(ages)
n
```

```
## [1] 14
```

7.3.2 Propotions et rapports [qual.]

```r
# on crée un vecteur de 12 hommes et 8 femmes avec la commande 'rep'ète et on vérifie avec la commande
sexe <- c(rep("H", 12), rep("F", 8))
sexe
```

```
##  [1] "H" "H" "H" "H" "H" "H" "H" "H" "H" "H" "H" "H" "F" "F" "F" "F" "F"
## [18] "F" "F" "F"
```

```r
ts <- table(sexe)
ts
```

```
## sexe
##  F  H
##  8 12
```

```r
hommes <- ts[2]
femmes <- ts[1]

rapport_de_masculinite <- hommes / femmes
rapport_de_masculinite
```

```
##   H
## 1.5
```

```r
sex_ratio <- hommes / (hommes + femmes)
sex_ratio
```

```
##   H
## 0.6
```

7.3.3 mode

Le mode identifie la valeur la plus fréquemment observée

```r
# pas de fonction, il faut en créer une
names(sort(-table(a)))[1]
```

```
## [1] "4"
```

7.3.4 moyenne (mean) [quant.]

```r
(25 + 18 + 21 + 21 + 23 + 22 + 23 + 18 + 25 + 19 + 22) / 11
```

```
## [1] 21.54545
```

```r
sum(ages) / length(ages)
```

```
## [1] 21.64286
```

```r
mean(ages)
```

```
## [1] 21.64286
```

7.3.5 Variance (variance) [quant.]

C'est la moyenne des écarts à la moyenne. Plus la variance est grande et plus l'effectif est dispersé.

```r
var(ages)
```

```
## [1] 4.708791
```

7.3.6 écart-type (standard déviation) [quant.]

C'est la racine carrée de la variance

```r
sd(ages)
```

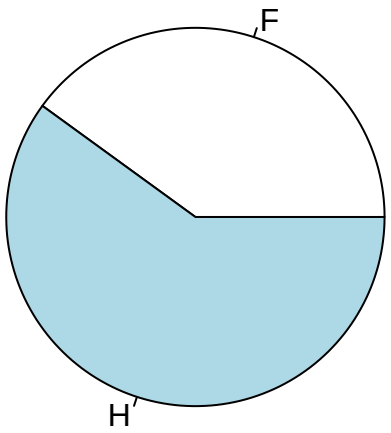
```
## [1] 2.169975
```

Si les données se distribuent selon un **loi normale**, alors 99% des données se situent dans l'intervalle défini par la moyenne +/- 3 fois l'écart-type.

7.4 Graphiques

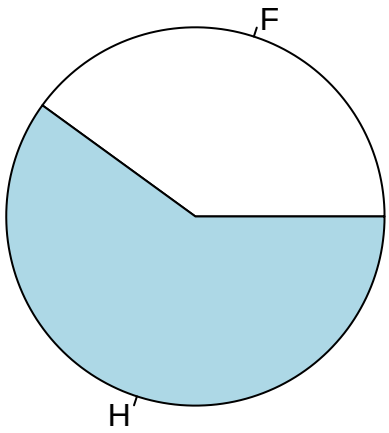
7.4.1 camemgerts (pie-chart)

```r
pie(ts)
```




```r
pie(ts, main = "Répartitions homme/femmes")
```

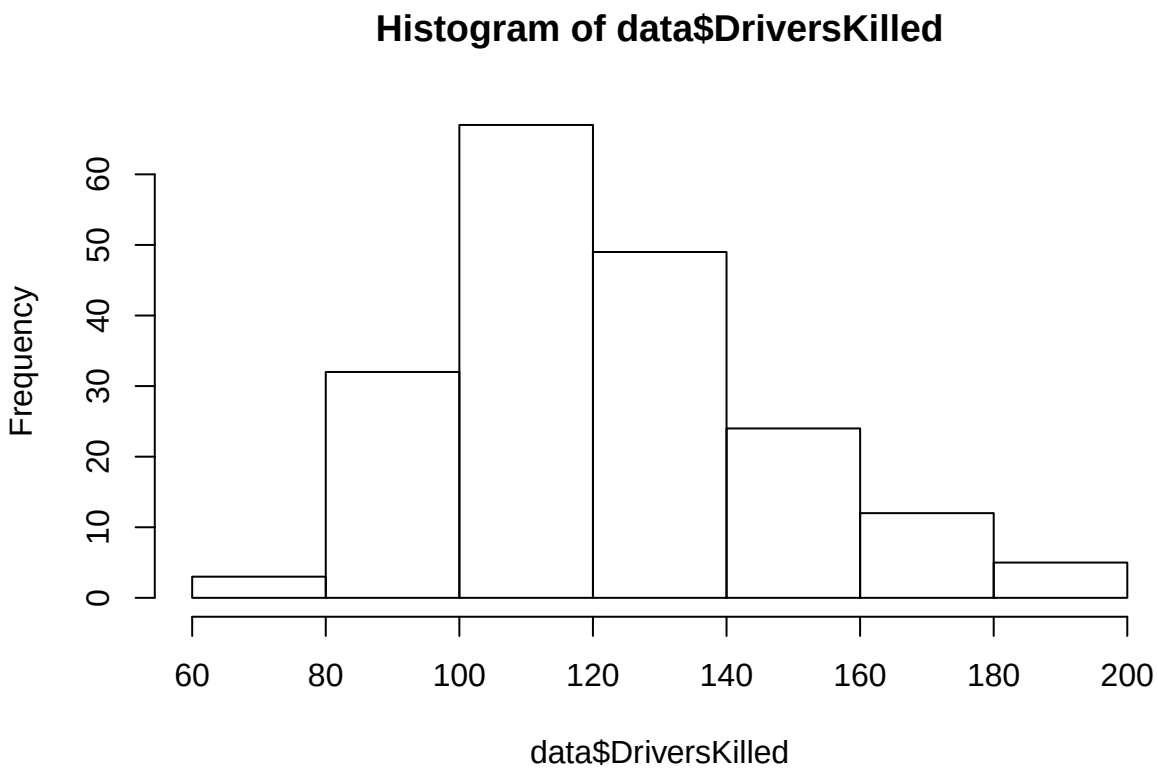
Répartitions homme/femmes



7.4.2 Histogramme [quant.]

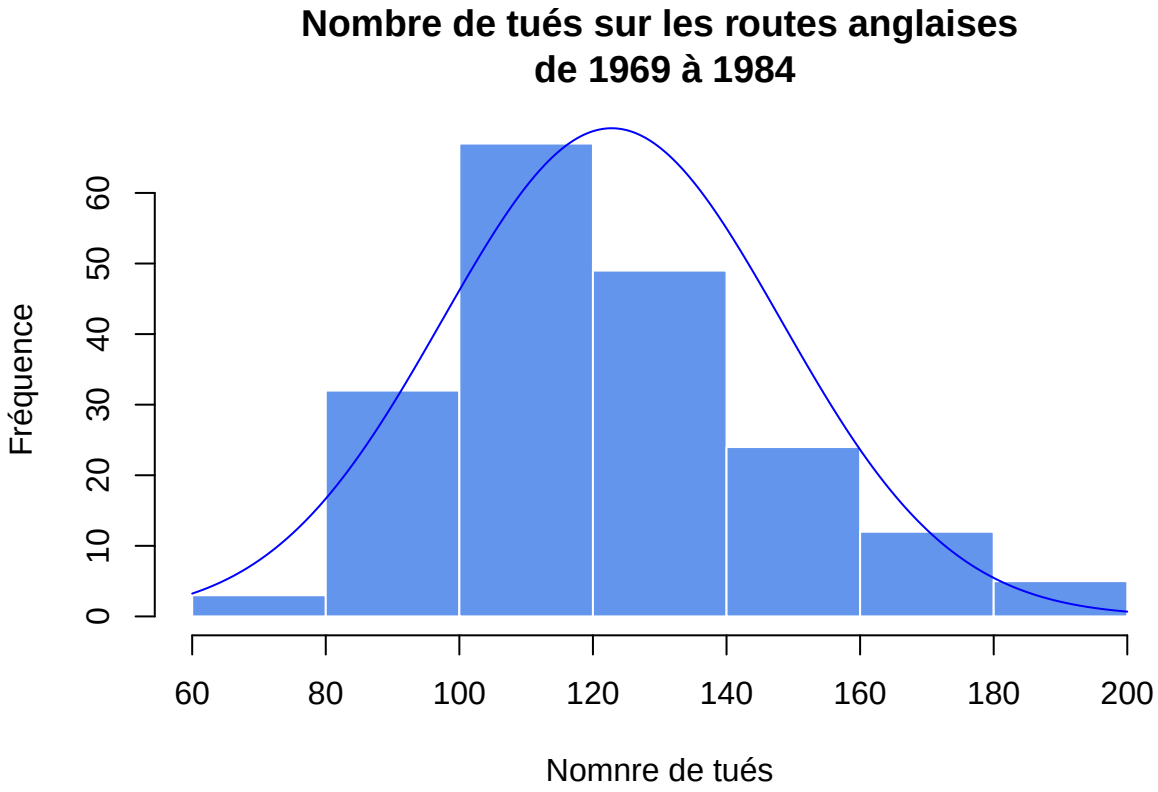
Conducteurs anglais tués par mois de janvier 1969 à décembre 1984.

```
data <- data.frame(Seatbelts)
data$an <- time(Seatbelts)
data$mois <- cycle(Seatbelts)
hist(data$DriversKilled)
```



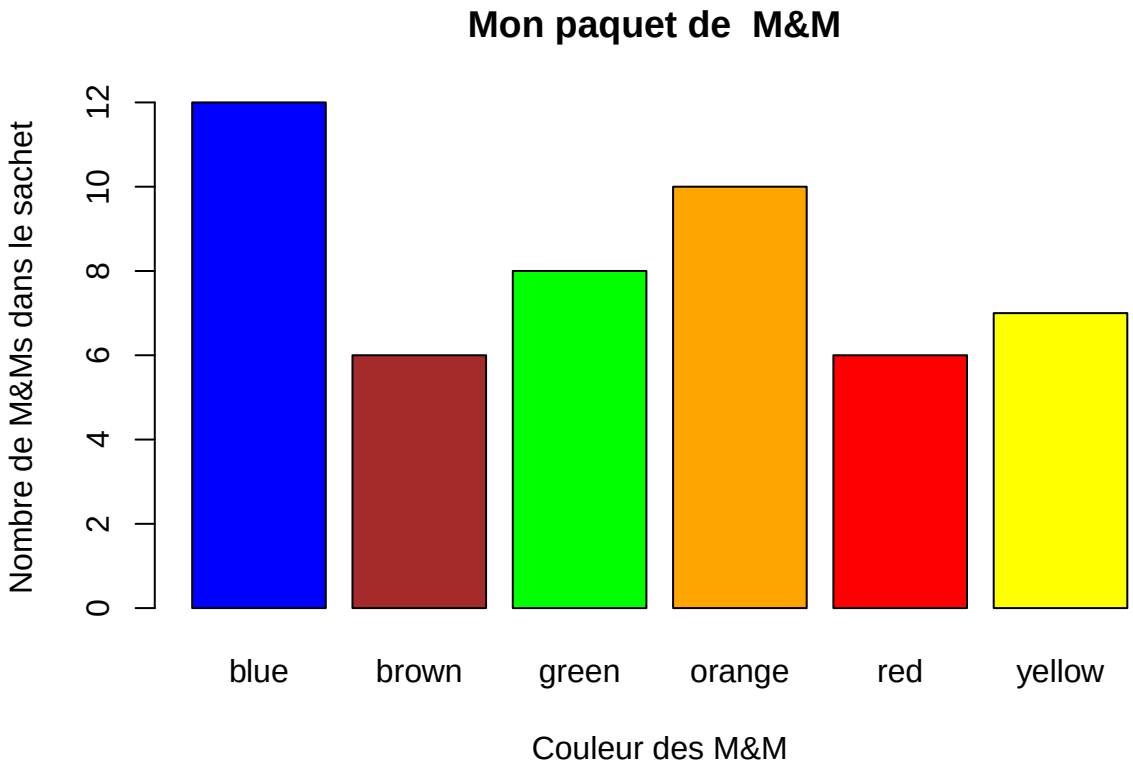
```r
hist(data$DriversKilled, ylab = "Fréquence", xlab = "Nomnre de tués", main = "Nombre de tués sur les rou

m <- mean(data$DriversKilled)
s <- sd(data$DriversKilled)
x <- seq(60, 200, 0.1)
lines(x, dnorm(x, m, s) * 4400, type = "l", col="blue")
```



7.4.3 Barplot [quant.]

```r
mm.counts <- c(12,6,8,10,6,7)
mm.colors <- c("blue","brown","green","orange","red","yellow")
names(mm.counts) <- mm.colors
barplot(mm.counts, main="Mon paquet de  M&M ",xlab="Couleur des M&M",ylab="Nombre de M&Ms dans le sache
```



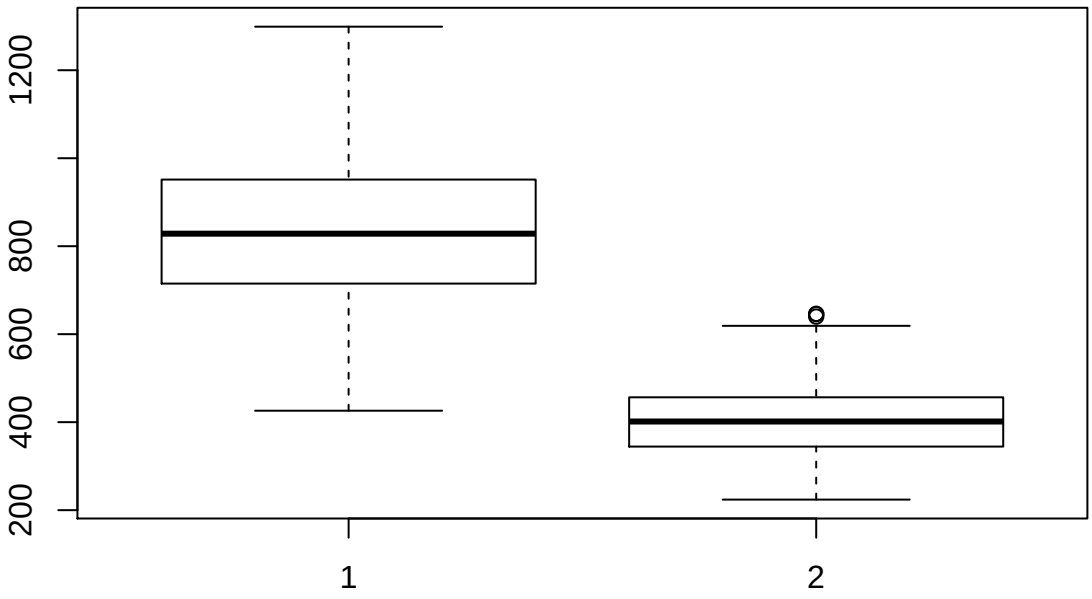
7.4.4 Boites à moustaches (Boxplot) [quant./qual.]

Une boxplot résume sur le même graphique 5 informations:

- minimum
- maximum
- 1er quartiles (25% des valeurs)
- médiane = 2éme quartile (50% des valeurs)
- 3ème quartile (75% des valeurs)

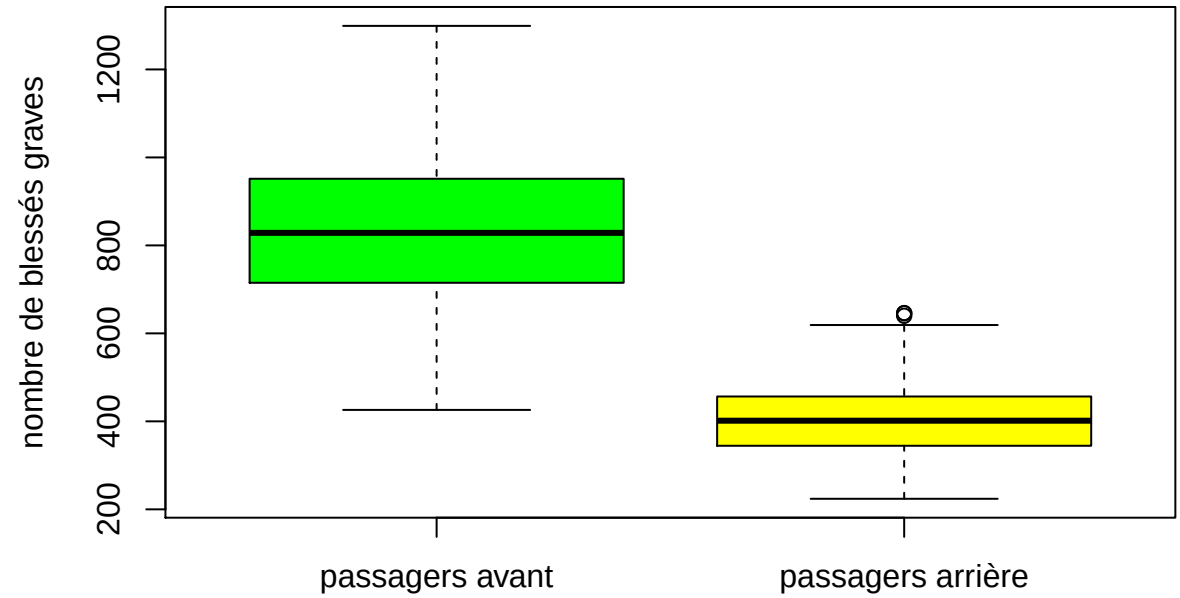
Exemple: comparaison du nombre mensuel de blessés graves selon le siège occuppé:

```
data <- data.frame(Seatbelts) # on récupère les données
boxplot(data$front, data$rear)
```



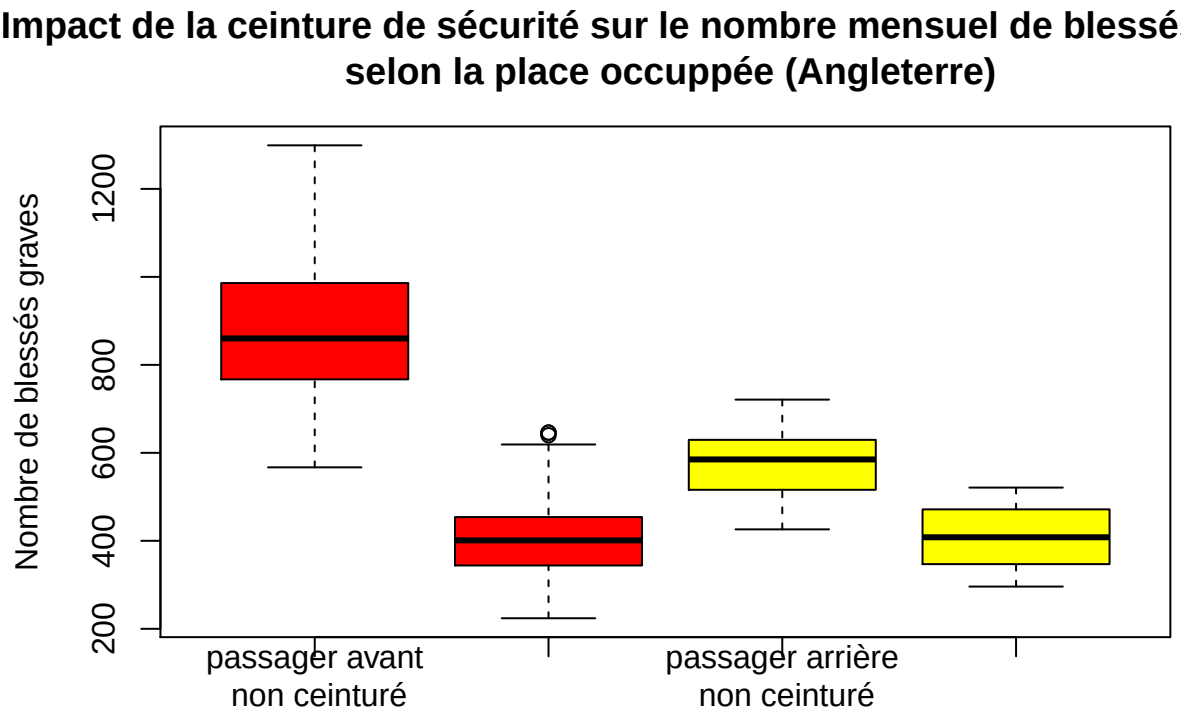

```
# avec habillage
boxplot(data$front, data$rear,
        names = c("passagers avant", "passagers arrière"),
        ylab = "nombre de blessés graves",
        main = "Nombre mensuel de blessés graves au cours des accidents de la voie publique\n en Anglet
        col = c("green", "yellow"))
```

ombre mensuel de blessés graves au cours des accidents de la voie pu
en Angleterre (1969−2004) selon le siège occuppé



Idem en prenant en compte le port de la ceinture de sécurité:

```
d1 <- data[data$law < 1,] #
d2 <- data[data$law > 0,]
boxplot(d1$front, d1$rear, d2$front, d2$rear, ylab = "Nombre de blessés graves", names = c("passager ava
```



7.5 Tests

7.5.1 Comparer deux moyennes (test de Student)

L'Angleterre à rendu obligatoire le port de la ceinture de sécurité sur les sièges avants le 31 décembre 1983. Cette mesure à t'elle eu un impact sur la mortalité routière ?

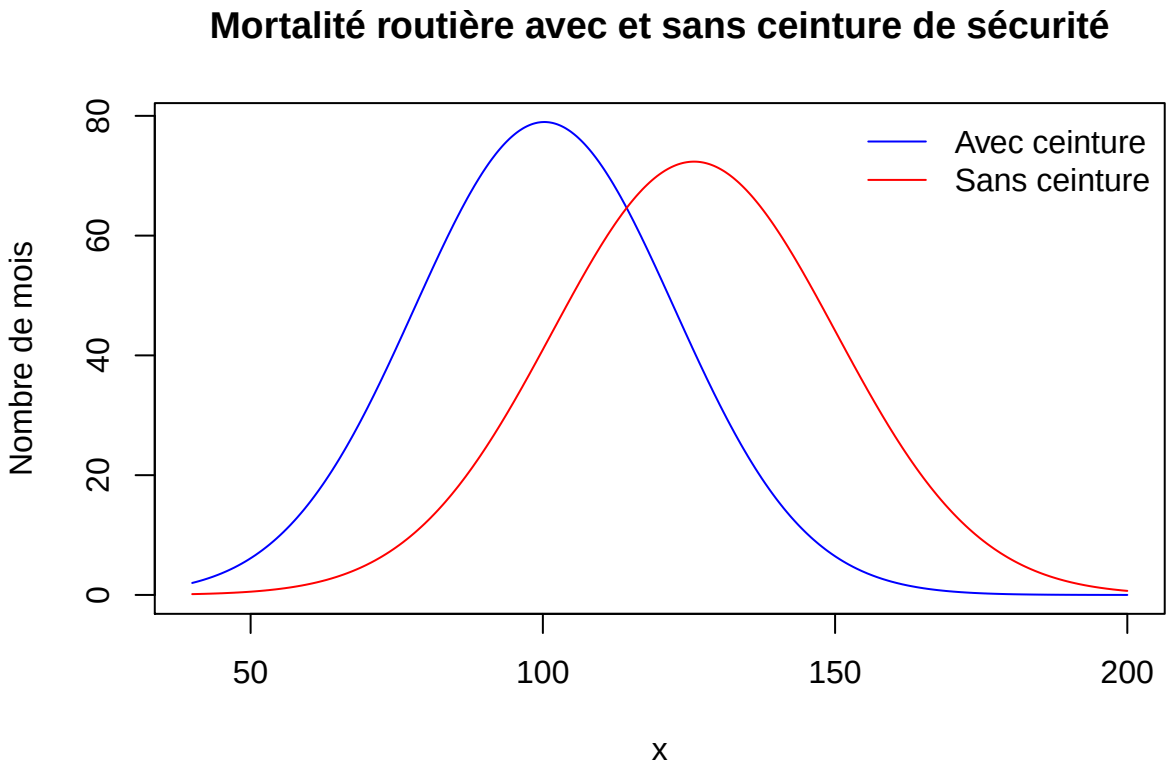
- Hypothèse neutre (ou nulle ou H0): il n'y a pas de différence de mortalité chez les conducteurs anglais selon qu'ils portent ou non une ceinture de sécurité.

```
data <- data.frame(Seatbelts) # on récupère les données
m <- tapply(data$DriversKilled, data$law, mean) # moyenne
s <- tapply(data$DriversKilled, data$law, sd) # écart-type
```

- mortalité mensuelle moyenne AVANT: 125.8698225 (+/- 24.2608758)
- mortalité mensuelle moyenne APRES: 100.2608696 (+/- 22.2286003)

Aspect graphique:

```
x <- seq(40, 200, 0.1)
plot(x, dnorm(x, m[2], s[2]) * 4400, type = "l", col="blue", ylab = "Nombre de mois", main = "Mortalité
lines(x, dnorm(x, m[1], s[1]) * 4400, type = "l", col="red")
legend("topright", legend = c("Avec ceinture", "Sans ceinture"), col = c("blue", "red"), lty = 1, bty =
```



test: on compare la mortalité mensuelle moyenne avant et après la promulgation de la loi ave le test de Student.

- Le test donne la probabilité (p) que la différence observée entre les deux moyennes soit due au hasard.
- C'est l'expérimentateur qui fixe le seuil à partir duquel on considère que ce n'est plus du hasard. De manière **consensuelle** (accord d'expert) cette limite est fixée à **0.05** ou **5%**.
- Si le résultat du test, $p < 0.05$, on considère que la différence n'est pas due au hasard et que l'hypothèse nulle doit être rejetée et par défaut on accepte l'hypothèse alternative: "le port de la ceinture de sécurité à un impact sur la mortalité des conducteurs"

```
# test
t.test(data$DriversKilled ~ data$law, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data$DriversKilled by data$law
## t = 4.7942, df = 190, p-value = 3.288e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.07239 36.14552
## sample estimates:
## mean in group 0 mean in group 1
##        125.8698        100.2609
```

Question: peut-on tirer les mêmes conclusions pour les blessés graves situés à l'avant (colonne 'front') ou situés à l'arrière (colonne 'rear') du véhicule ?

8 Transféréer les données

8.1 Passer du tableur à R

- format d'échange universel: .csv (comma separated values)
- Tableur -> Enregistrer sous -> TEXT CSV (.csv)
- ouvrir le fichier à partir de R avec **read.csv**

8.2 Travail collaboratif

- respectueux de la vie privée: **Framacalc** (+++) https://framacalc.org/_start

```
- libre, gratuit (dons à partir de 5€)
- permet de travailler à plusieurs sur le même tableur
- import direct à partir de __R__: https://framacalc.org/le_nom_de_mon_calc.csv
- exemple: https://framacalc.org/qKe5wD44QU
```

- Traitement de texte collaboratif: https://framapad.org/
- Organiser des réunions: http://framadate.org/
- Mind Mapping: http://framindmap.org/

8.3 Lecture du tableur Framasoft

- nécessite le package RCurl pour connexion sécurisée (Https)
- pour récupérer les données au format .csv, il suffit d'ajouter ".csv" au nom du tableur

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
jcb_url <- getURL("https://framacalc.org/qKe5wD44QU.csv")
jcb_data <- read.csv(textConnection(jcb_url), header = TRUE)
jcb_data
```

```
##   https...framacalc.org.le_nom_de_mon_calc.csv      X     X.1        X.2
## 1                                  Nom_Pseudo mms_ID Couleur       Etat
## 2                                         jcb      1     Red    Ebréché
## 3                                         jcb      2  Yellow      Fendu
## 4                                         jcb      3   Green    Parfait
## 5                                         jcb      4    Blue Non marqué
## 6                                         jcb      5  Maroon
## 7                                         jcb      6   Black
```

8.4 Organiser un questionnaire en ligne

- **Lime survey** (libre) LimeSurvey
- **Form** (propriétaire) google drive

9 Pour finir