Corrigé devoir R cmdo

Soren Harnois-Leblanc

26/07/2020

Application 1 : jeu de données ToothGrowth

Indiquer à quel endroit le corrigé sera sauvegardé sur mon ordinateur.

```
setwd("C:\\Users\\soren\\OneDrive - Universite de Montreal\\CMDO\\CIÉ\\")
```

Spécifier au début du script les packages qui nous seront nécessaires.

```
library(ggplot2)
```

Ouvrir le jeu de données.

```
data("ToothGrowth")
?ToothGrowth

## starting httpd help server ... done
```

Question 1

Vérifier la structure du jeu de données. Combien y a-t-il de sujets ? Combien y a-t-il de variables et quel est le type de chacune ?

```
dim(ToothGrowth)

## [1] 60  3

str(ToothGrowth)

## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

60 sujets sont inclus dans le jeu de données. Il y a 3 variables: len = numérique, supp = facteur, dose = numérique.

Question 2

Créer deux groupes de sujet selon type de supplément reçu. Générer les statistiques descriptives de longueur des dents par groupe, ainsi que l'histogramme.

```r
# crée les groupes
oj <- subset(ToothGrowth, ToothGrowth$supp=="OJ")
head(oj,10)
```

```
##     len supp dose
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
```

```r
tail(oj,10)
```

```
##     len supp dose
## 51 25.5   OJ    2
## 52 26.4   OJ    2
## 53 22.4   OJ    2
## 54 24.5   OJ    2
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```r
vc <- subset(ToothGrowth, ToothGrowth$supp=="VC")
head(vc,10)
```

```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

```r
tail(vc,10)
```

```
##     len supp dose
## 21 23.6   VC    2
## 22 18.5   VC    2
## 23 33.9   VC    2
## 24 25.5   VC    2
```

```
## 25 26.4   VC    2
## 26 32.5   VC    2
## 27 26.7   VC    2
## 28 21.5   VC    2
## 29 23.3   VC    2
## 30 29.5   VC    2
```

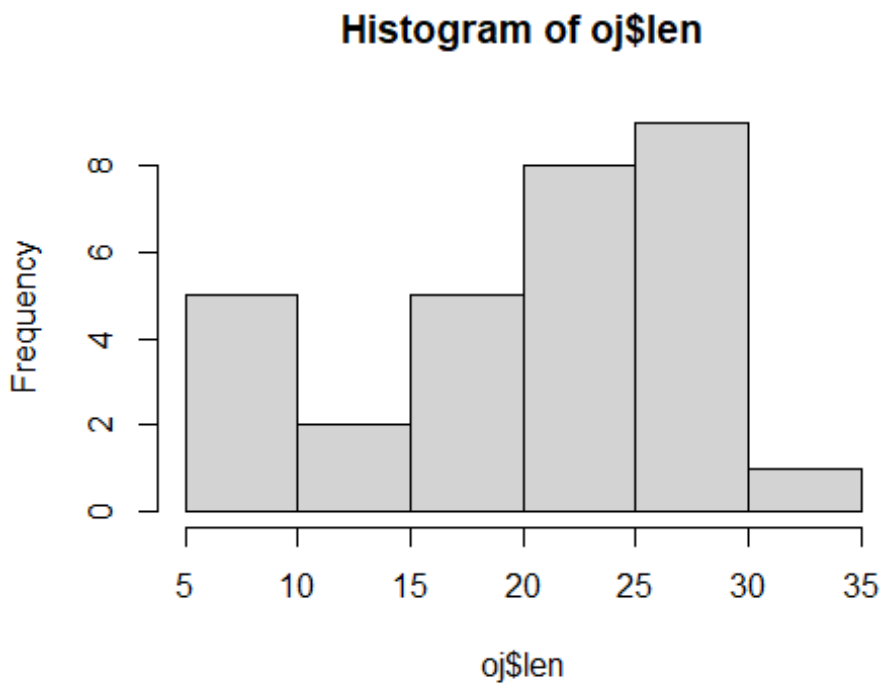
```
# statistiques descriptives groupe oj
summary(oj)
```

```
##       len        supp         dose
##  Min.   : 8.20   OJ:30   Min.   :0.500
##  1st Qu.:15.53   VC: 0   1st Qu.:0.500
##  Median :22.70           Median :1.000
##  Mean   :20.66           Mean   :1.167
##  3rd Qu.:25.73           3rd Qu.:2.000
##  Max.   :30.90           Max.   :2.000
```

```
sd(oj$len)
```

```
## [1] 6.605561
```

```
# histogramme groupe oj
hist(oj$len)
```



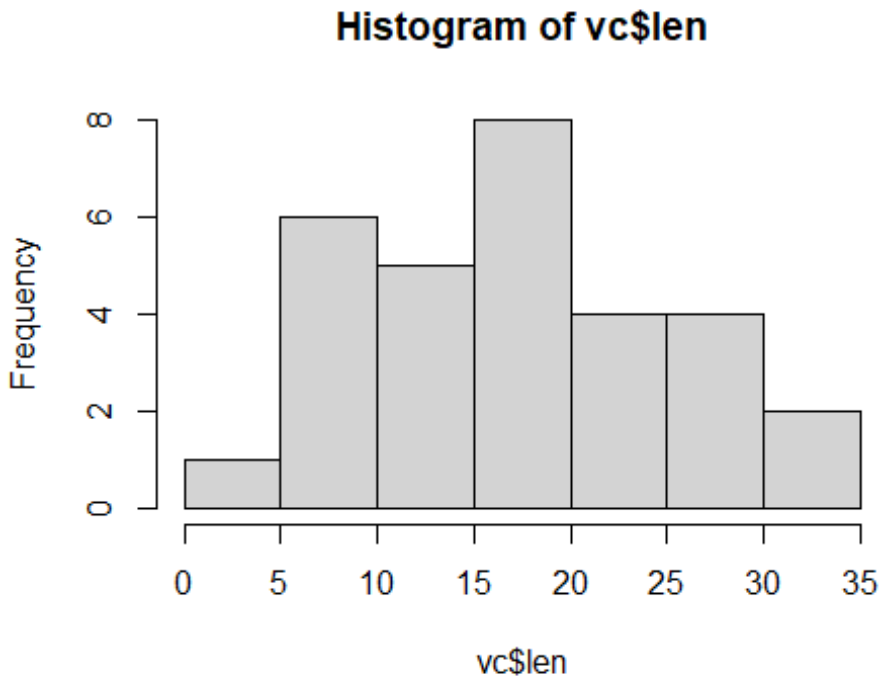

```
# statistiques descriptives groupe vc
summary(vc)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ: 0   Min.   :0.500
##  1st Qu.:11.20   VC:30   1st Qu.:0.500
##  Median :16.50           Median :1.000
##  Mean   :16.96           Mean   :1.167
##  3rd Qu.:23.10           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
sd(vc$len)
```

```
## [1] 8.266029
```

```
# histogramme groupe vc
hist(vc$len)
```



Note: s'il y avait des valeurs manquantes de la variable len, sd nous retournerait NA. Il faudrait spécifier sd(oj$len, na.rm=T).

Groupe OJ a une longueur de dents médiane de 22,70 mm (1er quart 15,53 - 3e quart 25,73). La médiane de longueur de dents est de 16,50 mm (11,20 - 23,10) dans le groupe VC. La variable n'est pas distribuée normalement dans les deux groupes.

Question 3

Comparaison de la longueur des dents entre le groupe avec supplément OJ vs supplément VC. On utilisera le test de Wilcoxon.

```
wilcox.test(len ~ supp, data=ToothGrowth, paired=F, conf.int=T)
```

```
## Warning in wilcox.test.default(x = c(15.2, 21.5, 17.6, 9.7, 14.5, 10, 8.2,
:
## cannot compute exact p-value with ties

## Warning in wilcox.test.default(x = c(15.2, 21.5, 17.6, 9.7, 14.5, 10, 8.2,
:
## cannot compute exact confidence intervals with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  len by supp
## W = 575.5, p-value = 0.06449
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.1000097  8.5000168
## sample estimates:
## difference in location
##               4.000014

# utilise par défaut two-sided test, pas besoin de spécifier.

# si on veut vérifier avec le test T de Student
t.test(len ~ supp, data=ToothGrowth, paired=F)

##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333

# quelle est la différence de moyenne ?
mean(ToothGrowth$len[ToothGrowth$supp=="OJ"])-mean(ToothGrowth$len[ToothGrowt
h$supp=="VC"])

## [1] 3.7
```

Les sujets avec supplément de jus d'orange ont une longueur des dents moyenne 3,7 mm (95% CI: -0.2 ; 7.6) plus élevée que les sujets avec supplément de vitamine C, toutefois, la différence n'est pas statistiquemment significative (p=0.064).

Question 4

On effectue une régression linéaire entre la dose reçue et la longueur des dents, sans considérer le type de supplément reçu pour le moment.

```r
# je mets la dose en facteur pour comparer 0.5 / 1 / 2
ToothGrowth$dose_f <- as.factor(ToothGrowth$dose)

# vérifier que ça a bien fonctionné
table(ToothGrowth$dose_f)
```

```
##
## 0.5   1   2
##  20  20  20
```

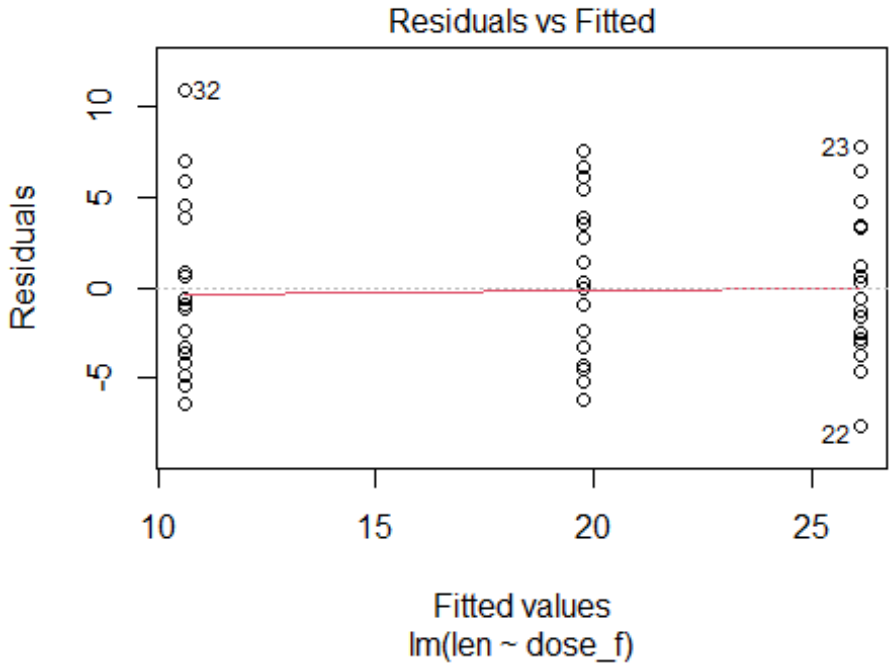
```r
# modèle de régression linéaire
mod <- lm(len ~ dose_f, data=ToothGrowth)

# demandons le sommaire, les intervalles de confiance des coefficients bêta e
stimés et 4 graphiques pour examiner la distribution des résidus.
summary(mod)
```

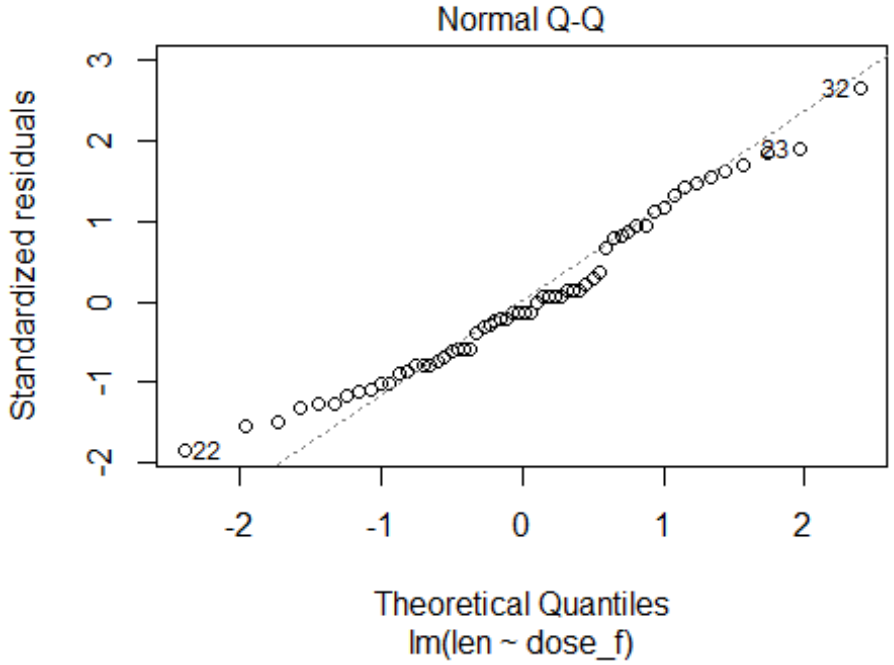
```
##
## Call:
## lm(formula = len ~ dose_f, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6000 -3.2350 -0.6025  3.3250 10.8950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6050     0.9486  11.180 5.39e-16 ***
## dose_f1       9.1300     1.3415   6.806 6.70e-09 ***
## dose_f2      15.4950     1.3415  11.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.242 on 57 degrees of freedom
## Multiple R-squared:  0.7029, Adjusted R-squared:  0.6924
## F-statistic: 67.42 on 2 and 57 DF,  p-value: 9.533e-16
```

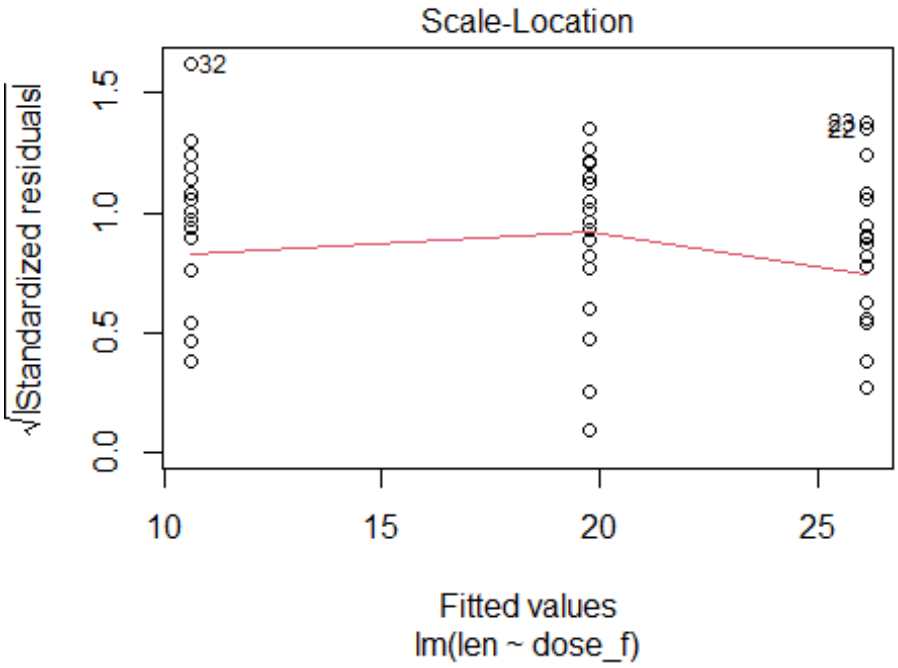
```r
confint(mod)
```

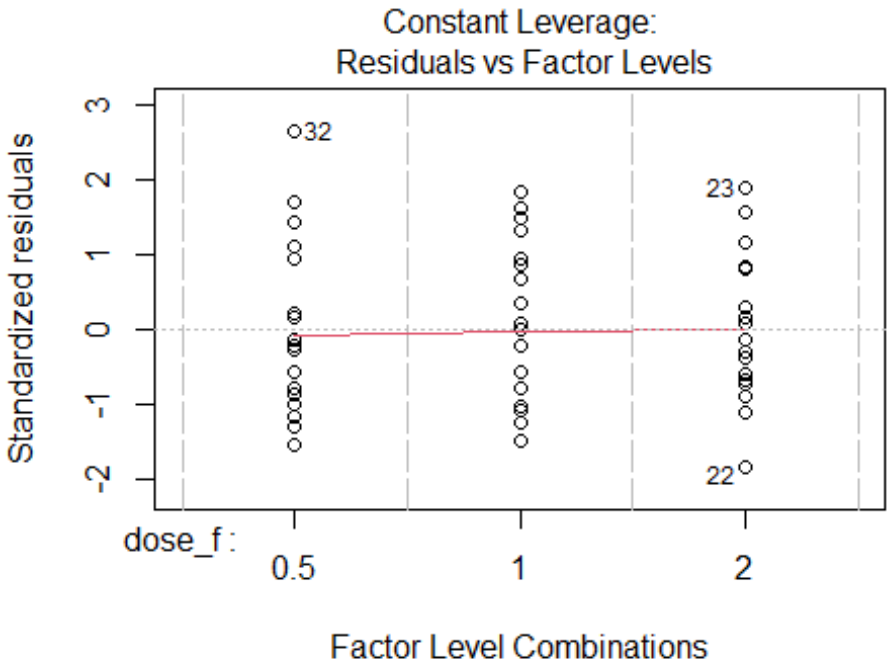
```
##                 2.5 %   97.5 %
## (Intercept)  8.705503 12.50450
## dose_f1      6.443705 11.81629
## dose_f2     12.808705 18.18129
```

```r
plot(mod)
```


Residuals vs Fitted
Residuals
Fitted values
lm(len ~ dose_f)
32
23
22
10
15
20
25
-5
0
5
10


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(len ~ dose_f)
32
23
22
-2
-1
0
1
2
3



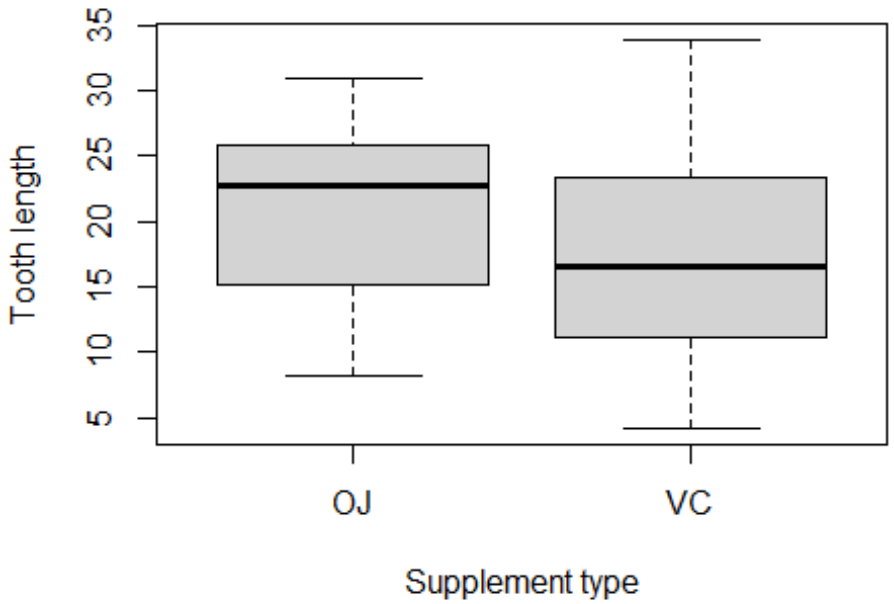


On observe une association positive entre la dose administrée et la longueur des dents, selon les coefficients bêta significatifs de la dose 1 et celui de la dose 2, en comparaison à la dose 0.5.

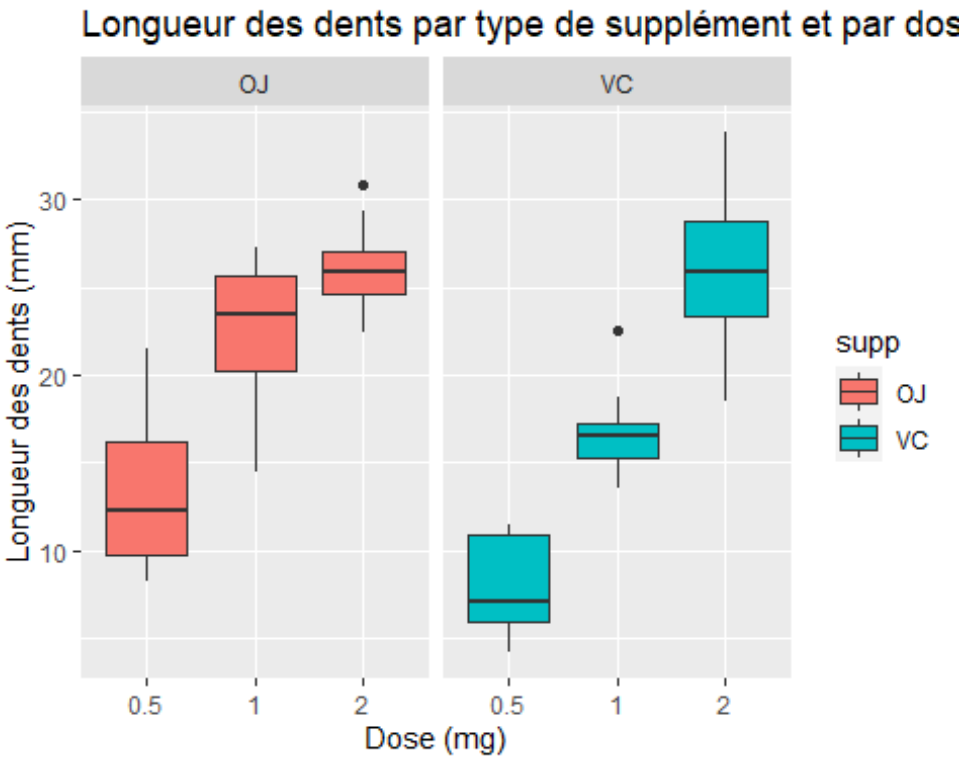
Question 5

On génère des boxplots pour mieux comprendre l'effet du type de supplément et de la dose sur la longueur des dents. En premier, on produit un boxplot par type de supplément. En deuxième, on réalise un boxplot par sous-groupe de combinaison supplément et dose. Le package ggplot2 sera utile pour faire les 6 boxplots.

```
# Boxplot de longueur des dents par type de supplément
boxplot(len ~ supp, data=ToothGrowth, xlab="Supplement type", ylab="Tooth len
gth")
```




```
# Boxplot de longueur des dents par combinaison supplément-dose
ggplot(ToothGrowth, aes(x=dose_f, y=len, fill=supp)) +
  geom_boxplot()+ facet_grid(.~supp)+ labs(x="X (binned)") +
  scale_x_discrete("Dose (mg)") + scale_y_continuous("Longueur des dents (mm)
") +
  ggtitle("Longueur des dents par type de supplément et par dose administrée"
)
```



Les graphiques nous indiquent que plus la dose administrée est élevée, plus la longueur des dents est augmentée, sans différence notable entre le type de supplément reçu.

Question 6

Nous comparons les moyennes de longueur des dents entre les sous-groupes de combinaison supplément-dose à l'aide du test de Kruskal-Wallis. On juge que le nombre de sujets par sous-groupes, ainsi que la distribution anormale de longueur des dents observée à la question 2, ne permettent pas de rencontrer les présuppositions requises pour le test d'ANOVA.

```
# Créons la variable sg, qui indiquera les sous-groupes supplément-dose. La v
ariable sera un facteur. J'assigne une étiquette "abc" à chaque niveau plutôt
qu'un nombre pour m'aider à identifier les sous-groupes.
ToothGrowth$sg <- NA
ToothGrowth$sg [ToothGrowth$supp=="OJ" & ToothGrowth$dose_f=="0.5"] <- "OJ_0.
5"
ToothGrowth$sg [ToothGrowth$supp=="OJ" & ToothGrowth$dose_f=="1"] <- "OJ_1"
ToothGrowth$sg [ToothGrowth$supp=="OJ" & ToothGrowth$dose_f=="2"] <- "OJ_2"
ToothGrowth$sg [ToothGrowth$supp=="VC" & ToothGrowth$dose_f=="0.5"] <- "VC_0.
5"
ToothGrowth$sg [ToothGrowth$supp=="VC" & ToothGrowth$dose_f=="1"] <- "VC_1"
ToothGrowth$sg [ToothGrowth$supp=="VC" & ToothGrowth$dose_f=="2"] <- "VC_2"

# est-ce que j'ai bien catégorisé ? Je regarde le tableau de la variable sg a
```

```
insi que la ligne de données de quelques sujets au hasard.
table(ToothGrowth$sg)

##
## OJ_0.5   OJ_1   OJ_2 VC_0.5   VC_1   VC_2
##     10     10     10     10     10     10

ToothGrowth[12,]

##     len supp dose dose_f   sg
## 12 16.5   VC    1      1 VC_1

ToothGrowth[24,]

##     len supp dose dose_f   sg
## 24 25.5   VC    2      2 VC_2

ToothGrowth[36,]

##    len supp dose dose_f     sg
## 36  10   OJ  0.5    0.5 OJ_0.5

# test de Kruskal Wallis
kruskal.test(len ~ sg, data = ToothGrowth)

##
##  Kruskal-Wallis rank sum test
##
## data:  len by sg
## Kruskal-Wallis chi-squared = 45.806, df = 5, p-value = 9.948e-09
```

La distribution de longueur des dents diffère entre au moins 2 sous-groupes (p = 9.9 x 10^-9). Il est approprié de vérifier quels sont les sous-groupes qui diffèrent avec le test de Wilcoxon avec correction pour la multiplicité des comparaisons. On choisit de corriger avec la méthode de false discovery rate, mais d'autres méthodes auraient pu être spécifiées, voir ?p.adjust

```
pairwise.wilcox.test(ToothGrowth$len, ToothGrowth$sg, p.adjust.method = "fdr"
)

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correct
ion
##
## data:  ToothGrowth$len and ToothGrowth$sg
##
##        OJ_0.5  OJ_1    OJ_2    VC_0.5  VC_1
## OJ_1   0.00248 -       -       -       -
## OJ_2   0.00054 0.09407 -       -       -
## VC_0.5 0.03162 0.00054 0.00054 -       -
## VC_1   0.09348 0.00605 0.00060 0.00054 -
## VC_2   0.00060 0.16121 1.00000 0.00054 0.00082
```

```
##
## P value adjustment method: fdr
```

On observe des différences significatives entre les sous-groupes:

- OJ_1 vs OJ_0.5
- OJ_2 vs OJ_0.5
- VC_0.5 vs OJ_0.5, OJ_1, OJ_2
- VC_1 vs OJ_1, OJ_2, VC_0.5
- VC_2 vs OJ_0.5, VC_0.5, VC_1

La dose semble être le facteur le plus important sur la longueur des dents. De plus, aux doses 0.5 et 1, le jus d'orange permet d'avoir une longueur des dents plus élevée que la vitamine C (se référer aux boxplots de la question précédente).

```
#et si je veux vérifier avec la méthode de Bonferroni au lieu de fdr
pairwise.wilcox.test(ToothGrowth$len, ToothGrowth$sg, p.adjust.method = "bonf
erroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correct
ion
##
## data:  ToothGrowth$len and ToothGrowth$sg
##
##        OJ_0.5 OJ_1   OJ_2   VC_0.5 VC_1
## OJ_1   0.0223 -      -      -      -
## OJ_2   0.0027 1.0000 -      -      -
## VC_0.5 0.3478 0.0027 0.0027 -      -
## VC_1   1.0000 0.0605 0.0036 0.0027 -
## VC_2   0.0042 1.0000 1.0000 0.0027 0.0065
##
## P value adjustment method: bonferroni
```

La correction de Bonferroni semble beaucoup plus conservatrice que la méthode du false discovery rate.

Ceci conclut l'application 1 du devoir.