Modélisation Statistique - TD7

Charles Vin

Chargement des données

Question 1

```r
df = read.csv("TD7_Tips.csv", sep=";")
head(df)
```

```
##   IDEN TOTBILL  TIP SEX SMOKER DAY TIME SIZE
## 1 R001   16.99 1.01   1      0   6    1    2
## 2 R002   10.34 1.66   0      0   6    1    3
## 3 R003   21.01 3.50   0      0   6    1    3
## 4 R004   23.68 3.31   0      0   6    1    2
## 5 R005   24.59 3.61   1      0   6    1    4
## 6 R006   25.29 4.71   0      0   6    1    4
```

Question 2

- `IDEN` : l'ID du client (chaîne de caractère).
- `TOTBILL`: L'addition en dollars. C'est une **variable quantitative**.
- `TIP` : Le pourboire en dollars. C'est une **variable quantitative**.
- `SEX`: : Indique le sexe de la personne qui a payé. 0 si c'est un homme et 1 si c'est une femme. C'est une **variable qualitative**.
- `SMOKER` : Indique la zone du restaurant où l'addition a été payé. 0 pour la zone non fumeur, 1 pour la zone fumeur. C'est une **variable qualitative**.
- `DAY` : Le jour de la semaine, en commençant par le Dimanche. Par exemple 3 correspond à mardi, 4 à mercredi, etc. C'est une **variable qualitative**.
- `TIME` : Le moment de la journée. 0 signifie en journée et 1 en soirée. C'est une **variable qualitative**.
- `SIZE` : Le nombre de convives. C'est une **variable quantitative**.

```r
cat("IDEN : String ? ", is.character(df$IDEN), "\n")
```

```
## IDEN : String ?  TRUE
```

```r
cat("TOTBILL : Numeric ? ", is.numeric(df$TOTBILL), "\n")
```

```
## TOTBILL : Numeric ?  TRUE
```

```r
cat("TIP : Numeric ? ", is.numeric(df$TIP), "\n")
```

```
## TIP : Numeric ?  TRUE
```

```r
cat("SEX : Factor ? ", is.factor(df$SEX), "\n")
```

```
## SEX : Factor ?  FALSE
```

```r
cat("SMOKER : Factor ? ", is.factor(df$SMOKER), "\n")
```

```
## SMOKER : Factor ?  FALSE
```

```r
cat("DAY : Factor ? ", is.factor(df$DAY), "\n")
```

```
## DAY : Factor ?  FALSE
```

```r
cat("TIME : Factor ? ", is.factor(df$TIME), "\n")
```

```
## TIME : Factor ?  FALSE
```

```r
cat("SIZE : Numeric ? ", is.numeric(df$SIZE), "\n")
```

```
## SIZE : Numeric ?  TRUE
```

Question 3

```r
df$SEX = as.factor(df$SEX)
df$SMOKER = as.factor(df$SMOKER)
df$DAY = as.factor(df$DAY)
df$TIME = as.factor(df$TIME)
```

Relation entre le montant de la facture et le montant du pourboire

Question 1

Observons la corrélation et une anova.

```r
cor(df$TIP,df$TOTBILL)
```

```
## [1] 0.6757341
```

Le montant de la facture et le montant du pourboire semble fortement corrélé.

```r
res = lm(df$TIP~df$TOTBILL)
res
```

```
##
## Call:
## lm(formula = df$TIP ~ df$TOTBILL)
##
## Coefficients:
## (Intercept)   df$TOTBILL
##      0.9203       0.1050
```

```r
anova(res)
```

```
## Analysis of Variance Table
##
## Response: df$TIP
##             Df Sum Sq Mean Sq F value    Pr(>F)
## df$TOTBILL   1 212.42 212.424  203.36 < 2.2e-16 ***
## Residuals  242 252.79   1.045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le modèle de régression linaire est extrêmement significative.

Question 2

Ces deux variables semble être linéaire entre elle. En observant le coefficient de la régression, on trouve que le pourboire représente en moyenne 10.5% de l'addition totale.

Question 3

Utilisons le rapport entre le pourboire donnée et l'addition totale pour obtenir un pourcentage et travailler dessus.

```
df$TIP_PERCENT = df$TIP / df$TOTBILL
```

Comparaison des comportements des clients

En fonction du moment de la journée

Question 1

Soit:

- $X_1, ..., X_{n_1}$ les pourboires des clients en journée. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_1, \sigma_1)$ avec m_1, σ_1 inconnu.
- $Y_1, ..., Y_{n_2}$ les pourboires des clients en soirée. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_2, \sigma_2)$ avec m_1, σ_1 inconnu.

Un employé affirme que les clients sont plus généreux en soirée qu'en journée. On vas donc tester la moyenne. En d'autre terme il pense que :

- $H_0 : m_1 = m_2$
- $H_1 : m_1 < m_2$

Pour trancher ces hypothèses, nous allons utiliser la statistique de test suivante $D = \bar{X}_{n_1} - \bar{Y}_{n_2}$.

- Sous $H_0, D \sim \mathcal{N}(0, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$
- Sous $H_1, D \sim \mathcal{N}(m, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$ avec m inconnu négatif. Ainsi sous H_1, D prend des valeurs plus petite que sous H_0. On en déduit la forme de la zone de rejet

$$\mathcal{R} = \{D \leq c\}$$

On utilisera le théorème de Student pour conclure après avoir vérifié les deux points suivants :

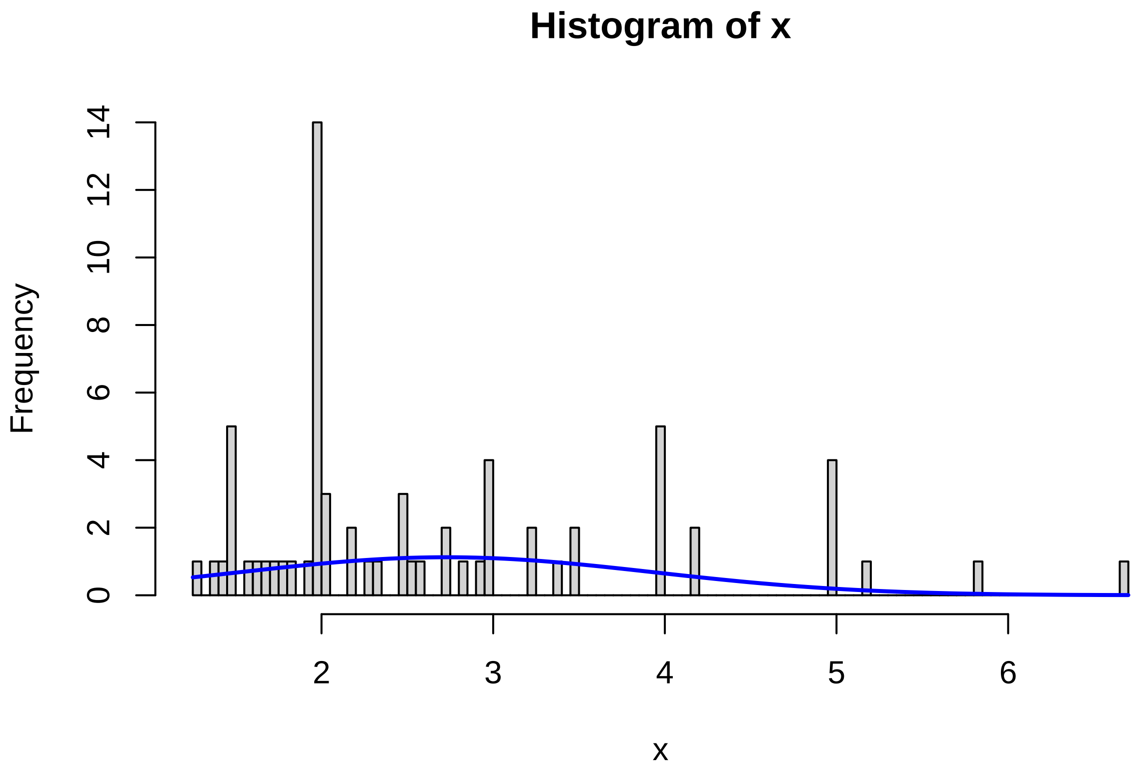
- La normalité des variables.
- L'éventuelle égalité des variances. En effet on peut penser que les variances des deux variables ne change pas, si c'est bien le cas, notre test pourrait gagner en puissance en utilisant cette information.

Question 2

Dans cette question nous allons tester les deux points précédents.

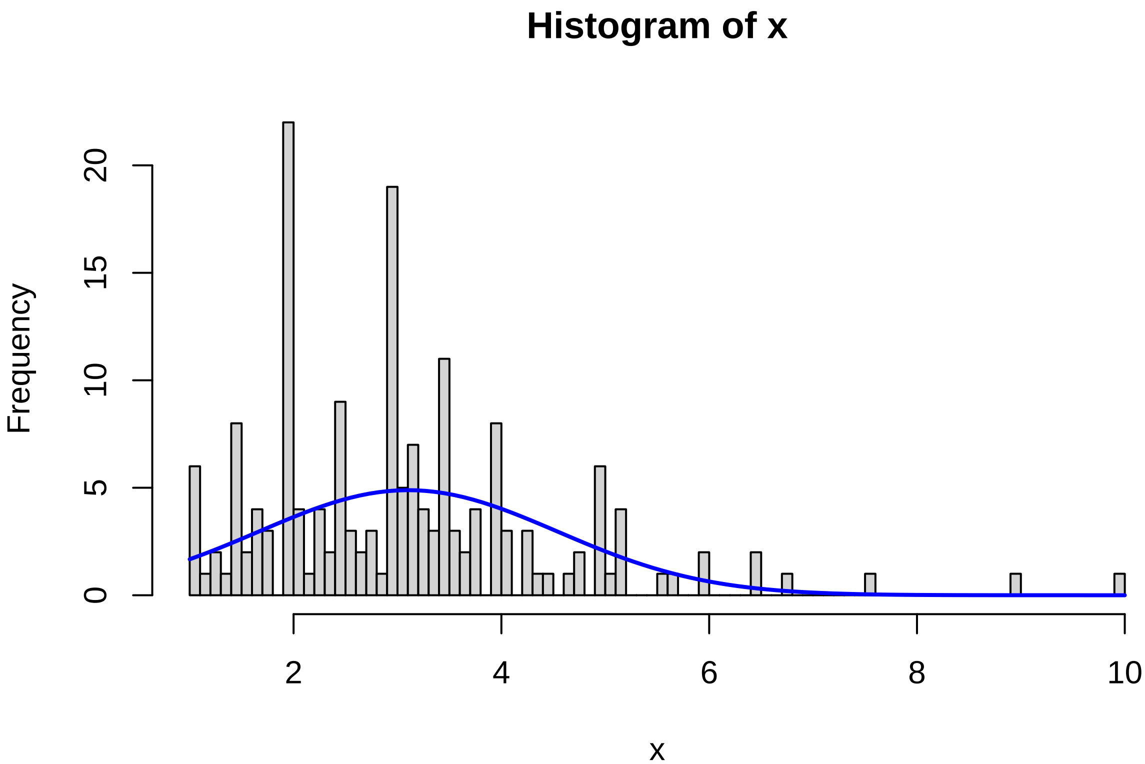
```
jour = df[df$TIME == 0, ]
nuit = df[df$TIME == 1, ]

x <- jour$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Normalité

```
x <- nuit$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Les données semblent moyennement gaussien au vu des graphiques, notamment pour le jour. Gardons néanmoins cette hypothèse car on sais que c'est le genre de chose suivant une gaussienne. Peut-être qu'avec plus de donnée on pourrait mieux le constater.

Égalité des variances Effectuons un test de Fisher. - $H_0 : \sigma_1 = \sigma_2$ - $H_0 : \sigma_1 \neq \sigma_2$

Avec la statistique de test suivante

$$D = \frac{V_{n_1}^X}{V_{n_2}^Y}$$

avec $V_{n_1}^X = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X}_{n_1})^2$ et $V_{n_2}^Y = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y}_{n_2})^2$

Zone de rejet :

Sous $H_0, \sigma^2 = \sigma_1^2 = \sigma_2^2$:

$$D \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Sous $H_1, \sigma_1^2 \neq \sigma_2^2$

$$D = \mathcal{F}(d_1, d_2) * \frac{\sigma_1^2}{\sigma_2^2}$$

Soit $h_{\alpha/2}$ le quantile $\frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$ et $h_{1-\alpha/2}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$.

Si $D > h_{1-\alpha/2}$ ou bien $D < h_{\alpha/2}$ on rejette H_0. Sinon on conserve H_0.

Effectuons ce test avec R pour obtenir la p-valeur.

```
var.test(jour$TIP, nuit$TIP)
```

```
##
##  F test to compare two variances
##
## data:  jour$TIP and nuit$TIP
## F = 0.70432, num df = 67, denom df = 175, p-value = 0.1001
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.4803123 1.0708931
## sample estimates:
## ratio of variances
##          0.7043156
```

On a une p valeur = 10%, on rejette $H_0 : \sigma_1 = \sigma_2$ avec 10% de chance de ce tromper. Nous effectuerons donc un test de Welsh sur les moyennes de gaussienne de variance inconnu.

Question 3

```r
t.test(jour$TIP, nuit$TIP, alternative="less", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  jour$TIP and nuit$TIP
## t = -2.0593, df = 144.07, p-value = 0.02063
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.07345444
## sample estimates:
## mean of x mean of y
##  2.728088  3.102670
```

La p-valeur est plus petite que 0.05, on rejette H_0 avec $\alpha = 5\%$. Le serveur semble avoir raison, les pourboires sont plus généreux en soirée.

Par curiosité, jetons un oeil au résultat du test avec les variances égales.

```r
t.test(jour$TIP, nuit$TIP, alternative="less", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  jour$TIP and nuit$TIP
## t = -1.9063, df = 242, p-value = 0.0289
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.05012406
## sample estimates:
## mean of x mean of y
##  2.728088  3.102670
```

On trouve une p-valeur proche.

En fonction du sexe du client

Question 1

Soit

- $X_1, ..., X_{n_1}$ les pourboires des clients masculins. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_1, \sigma_1)$ avec m_1, σ_1 inconnu.
- $Y_1, ..., Y_{n_2}$ les pourboires des clients féminins. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_2, \sigma_2)$ avec m_1, σ_1 inconnu.

Y'a-t-il égalité des pourboires en fonction du sexe ?

- $H_0 : m_1 = m_2$
- $H_1 : m_1 \neq m_2$

Pour trancher ces hypothèses, nous allons utiliser la statistique de test suivante $D = \bar{X}_{n_1} - \bar{Y}_{n_2}$.

- Sous $H_0, D \sim \mathcal{N}(0, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$
- Sous $H_1, D \sim \mathcal{N}(m, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$ avec m.

Ainsi sous H_1, D prend des valeurs plus extrême que sous H_0. On en déduit la forme de la zone de rejet

$$\mathcal{R} = \{D \leq c_1\} \cup \{D \geq c_2\}$$

On utilisera le théorème de Student pour conclure après avoir vérifié les deux points suivants :

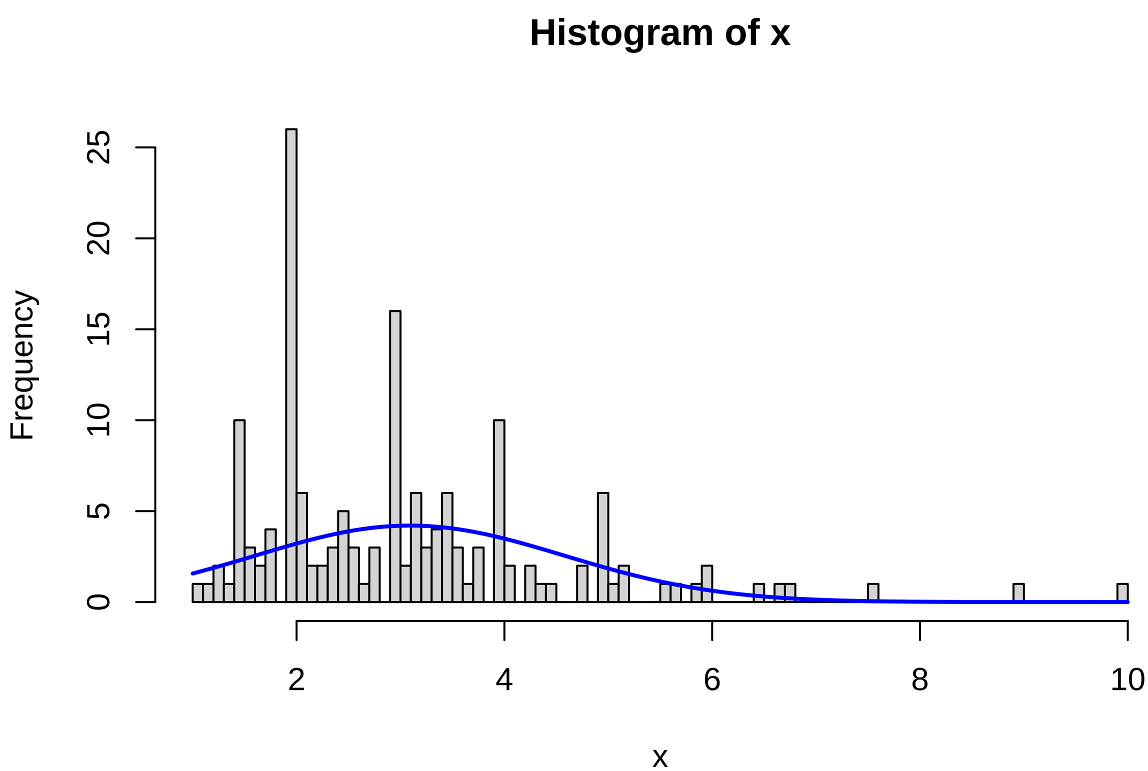
- La normalité des variables.
- L'éventuelle égalité des variances. En effet on peut penser que les variances des deux variables ne change pas, si c'est bien le cas, notre test pourrait gagner en puissance en utilisant cette information.

Question 2

Dans cette question nous allons tester les deux points précédents.

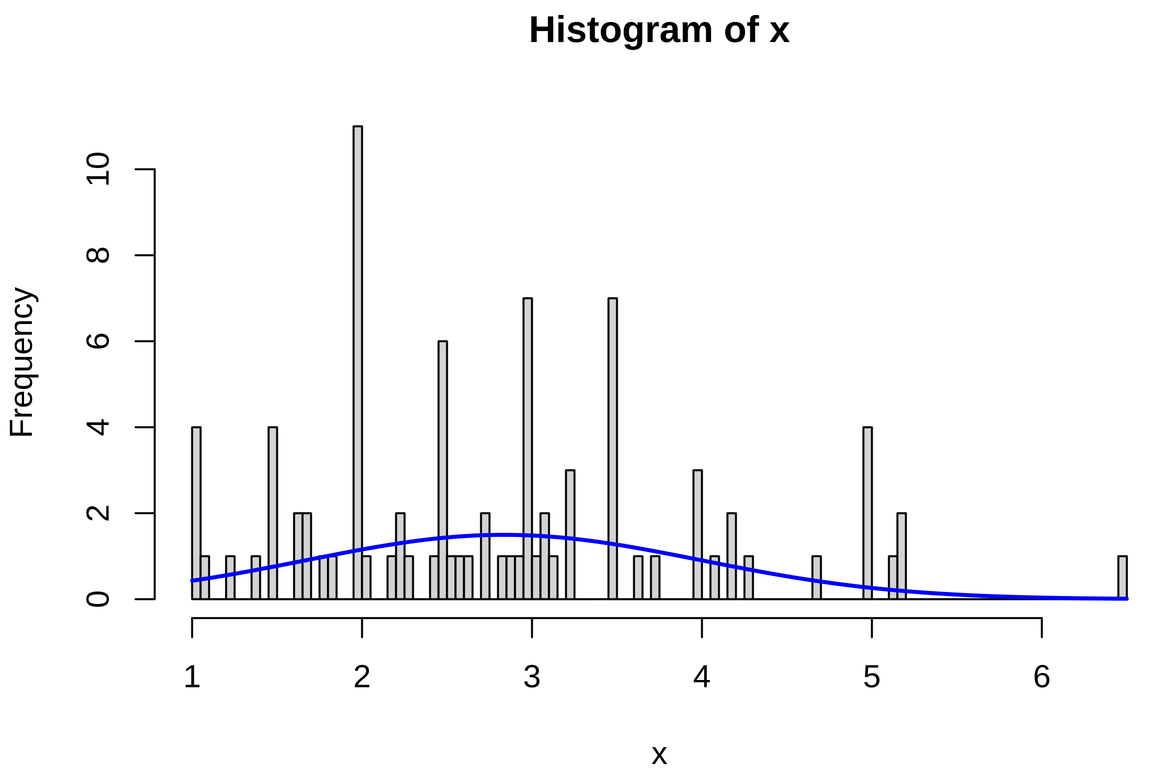
```
homme = df[df$SEX == 0, ]
femme = df[df$SEX == 1, ]

x <- homme$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit, mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Normalité

```
x <- femme$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Les données semblent gaussien ! Nous pouvons procéder au tests suivants.

Égalité des variances Effectuons un test de Fisher. - $H_0 : \sigma_1 = \sigma_2$ - $H_0 : \sigma_1 \neq \sigma_2$

Avec la statistique de test suivante

$$D = \frac{V_{n_1}^X}{V_{n_2}^Y}$$

avec $V_{n_1}^X = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X}_{n_1})^2$ et $V_{n_2}^Y = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y}_{n_2})^2$

Zone de rejet :

Sous $H_0, \sigma^2 = \sigma_1^2 = \sigma_2^2$:

$$D \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Sous $H_1, \sigma_1^2 \neq \sigma_2^2$

$$D = \mathcal{F}(d_1, d_2) * \frac{\sigma_1^2}{\sigma_2^2}$$

Soit $h_{\alpha/2}$ le quantile $\frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$ et $h_{1-\alpha/2}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$.

Si $D > h_{1-\alpha/2}$ ou bien $D < h_{\alpha/2}$ on rejette H_0. Sinon on conserve H_0.

Effectuons ce test avec R pour obtenir la p-valeur.

```
var.test(homme$TIP, femme$TIP)
```

```
##
##  F test to compare two variances
##
## data:  homme$TIP and femme$TIP
## F = 1.6493, num df = 156, denom df = 86, p-value = 0.01117
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.122793 2.373671
## sample estimates:
## ratio of variances
##           1.649345
```

On a une p valeur égal à 1%, on rejette $H_0 : \sigma_1 = \sigma_2$. Nous effectuerons donc un test de Welsh sur les moyennes de gaussienne de variance inconnu.

Question 3

```
t.test(homme$TIP, femme$TIP, alternative="less", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  homme$TIP and femme$TIP
## t = 1.4895, df = 215.71, p-value = 0.9311
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.5402706
## sample estimates:
## mean of x mean of y
##  3.089618  2.833448
```

La p-valeur est élevé. On conserve H_0, les pourboires sont les mêmes entre les hommes et les femmes.

En fonction de la zone fumeur ou non fumeur

Question 1

Soit

- $X_1, ..., X_{n_1}$ les pourboires des clients fumeur. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_1, \sigma_1)$ avec m_1, σ_1 inconnu.
- $Y_1, ..., Y_{n_2}$ les pourboires des clients non fumeur. Supposé Gaussien (voir question 2) $\sim \mathcal{N}(m_2, \sigma_2)$ avec m_1, σ_1 inconnu.

Y'a-t-il égalité des pourboires dans les deux zones fumeur et non fumeur du restaurant ?

- $H_0 : m_1 = m_2$
- $H_1 : m_1 \neq m_2$

Pour trancher ces hypothèses, nous allons utiliser la statistique de test suivante $D = \bar{X}_{n_1} - \bar{Y}_{n_2}$.

- Sous $H_0, D \sim \mathcal{N}(0, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$
- Sous $H_1, D \sim \mathcal{N}(m, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2})$ avec m.

Ainsi sous H_1, D prend des valeurs plus extreme que sous H_0. On en déduit la forme de la zone de rejet

$$\mathcal{R} = \{D \leq c_1\} \cup \{D \geq c_2\}$$

On utilisera le théorème de Student pour conclure après avoir vérifié les deux points suivants :

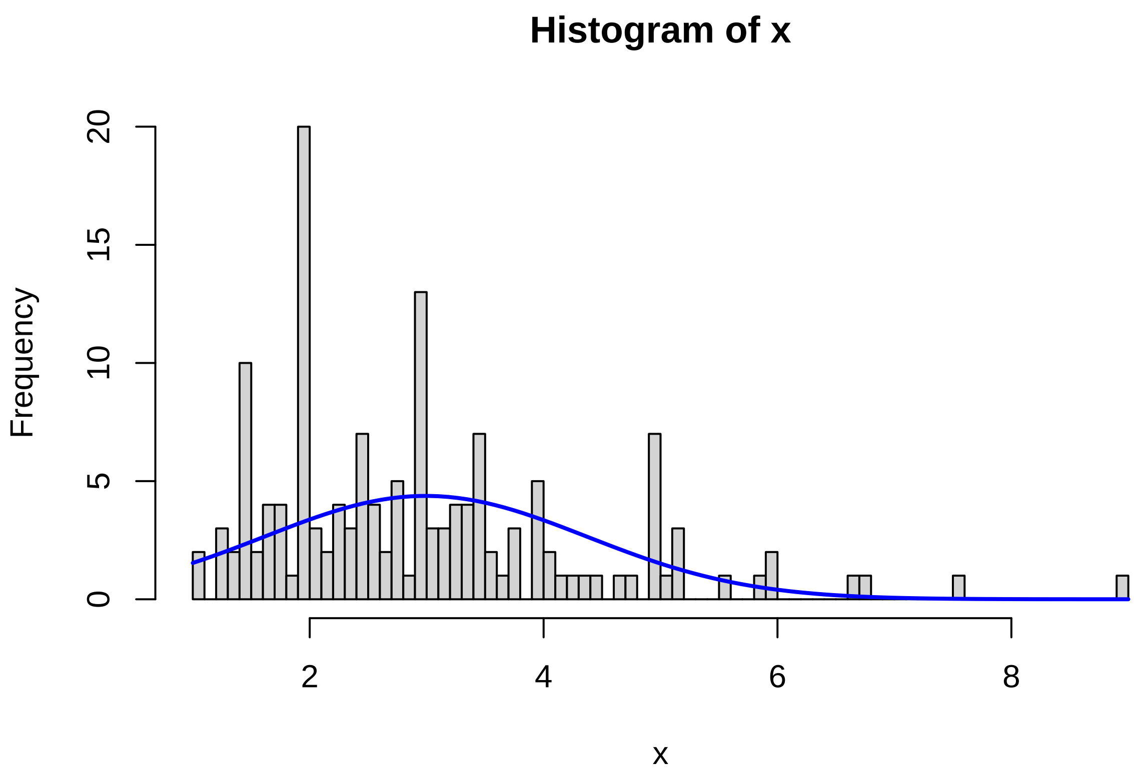
- La normalité des variables.
- L'éventuelle égalité des variances. En effet on peut penser que les variances des deux variables ne change pas, si c'est bien le cas, notre test pourrait gagner en puissance en utilisant cette information.

Question 2

Dans cette question nous allons tester les deux points précédents.

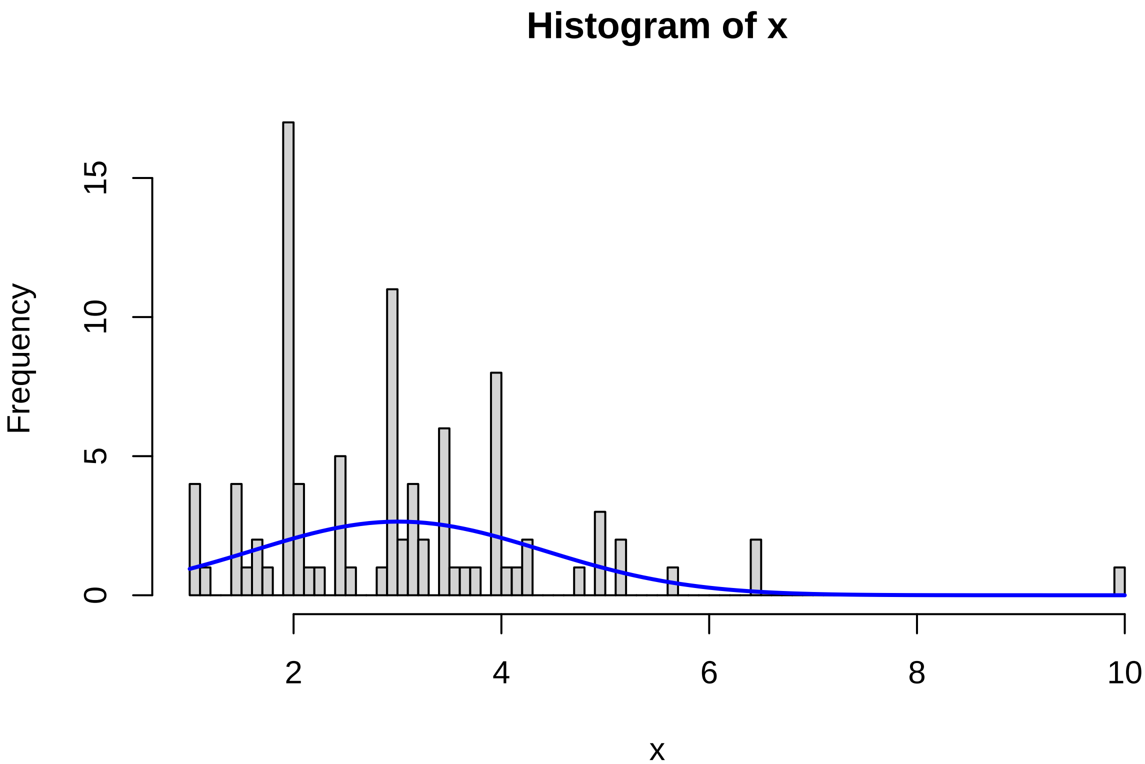
```
fumeur = df[df$SMOKER == 0, ]
non_fumeur = df[df$SMOKER == 1, ]

x <- fumeur$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Normalité

```
x <- non_fumeur$TIP
h<-hist(x, breaks=100)
xfit<-seq(min(x),max(x),length=100)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



Les données semblent bien gaussien ! Testons maintenant la variance.

Égalité des variances Effectuons un test de Fisher. - $H_0 : \sigma_1 = \sigma_2$ - $H_0 : \sigma_1 \neq \sigma_2$

Avec la statistique de test suivante

$$D = \frac{V^X_{n_1}}{V^Y_{n_2}}$$

avec $V^X_{n_1} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2$ et $V^Y_{n_2} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y}_{n_2})^2$

Zone de rejet :

Zone de rejet :

Sous $H_0, \sigma^2 = \sigma_1^2 = \sigma_2^2$:

$$D \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Sous $H_1, \sigma_1^2 \neq \sigma_2^2$

$$D = \mathcal{F}(d_1, d_2) * \frac{\sigma_1^2}{\sigma_2^2}$$

Soit $h_{\alpha/2}$ le quantile $\frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$ et $h_{1-\alpha/2}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une $\mathcal{F}(n_1 - 1, n_2 - 1)$.

Si $D > h_{1-\alpha/2}$ ou bien $D < h_{\alpha/2}$ on rejette H_0. Sinon on conserve H_0.

Effectuons ce test avec R pour obtenir la p-valeur.

```
var.test(fumeur$TIP, non_fumeur$TIP)
```

```
##
##  F test to compare two variances
##
## data:  fumeur$TIP and non_fumeur$TIP
## F = 0.96565, num df = 150, denom df = 92, p-value = 0.8403
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.6622073 1.3850964
## sample estimates:
## ratio of variances
##          0.9656543
```

Cette fois-ci, la p-valeur nous indique de conserver H_0. Nous considèreons alors que la variance est égale dans les deux groupe lors du test de Student.

Question 3

```r
t.test(fumeur$TIP, non_fumeur$TIP, alternative="less", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  fumeur$TIP and non_fumeur$TIP
## t = -0.092228, df = 242, p-value = 0.4633
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.2849092
## sample estimates:
## mean of x mean of y
##  2.991854  3.008710
```

La p-valeur est assez élevé, on conserve H_0. Les pourboires sont les mêmes dans les deux zones du restaurant.

En fonction du jour de la semaine

Question 1

```r
is.factor(df$DAY)
```

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

Question 2

On pose le modèle de régression linéaire suivant. - y_i^l le pourcentage que représente le pourboire sur l'addition totale pour le repas i le jour l.

$$Y_i^l = \mu_l + \epsilon_i^l \epsilon_i^l \sim \mathcal{N}(0, \sigma^2) iid.$$

On suppose que toutes les variables sont indépendante, intra et inter niveau.

```r
res = lm(df$TIP_PERCENT~df$DAY)
res # Ou est DAY3 ?
```

```
##
## Call:
## lm(formula = df$TIP_PERCENT ~ df$DAY)
##
## Coefficients:
## (Intercept)      df$DAY4      df$DAY5      df$DAY6
##    0.161276     0.008637    -0.008124     0.005622
```

Question 3

Lors d'un test d'analyse de la variance, on test

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- $H_1 : \exists l, m \in \{1, 2, 3, 4\} : \mu_l \neq \mu_m$.

On utilise alors la statistique de test

$$F = \frac{SCM/k-1}{SCR/n-k} = \frac{SCM/3}{SCR/240} \sim_{H_0} \mathcal{F}_{3,240}$$

Avec $SCM = \sum_{k}^{l=1} n_l(\bar{Y}_l - \bar{Y})^2, SCR = \sum_{k}^{l=1} \sum_{n_l}^{i=1} (Y_i^l - \bar{Y}_l)$ et $\bar{Y}_l$ la moyenne pour le niveau $l, \bar{Y}_l = \frac{1}{n_l} \sum_{n_l}^{i=1} X_i^l$

Dans la table d'ANOVA on retrouve dans la colonne

- `Df` : les degrés de liberté de la statistique de Fisher F
- `Sum Sq` : Somme des carrée des résidus si dans la colonne `Residuals` sinon somme des carrées du modèle
- `Mean Sq` : `Sum Sq/Df`. SCR et SCR divisé par leur degrés de liberté.
- `F value` : Valeur de la statistique de test F
- `Pr(>F)` : la p-value

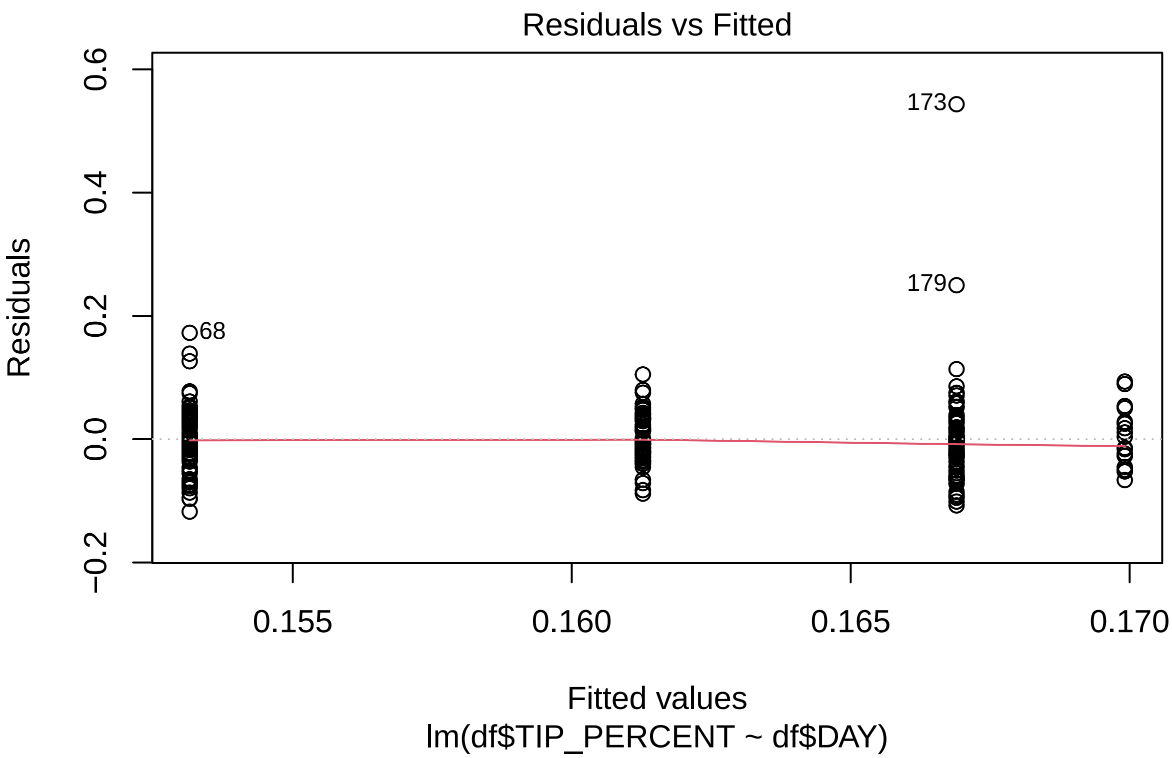
```
anova(res)
```

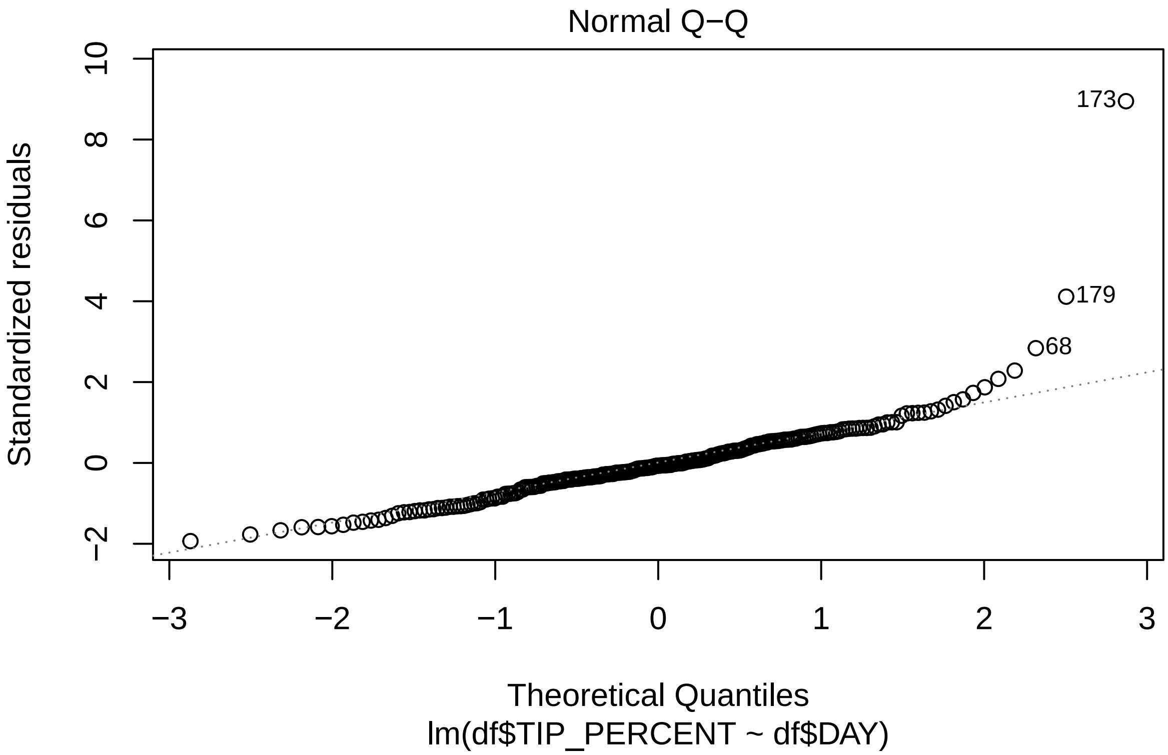
```
## Analysis of Variance Table
##
## Response: df$TIP_PERCENT
##            Df  Sum Sq   Mean Sq F value Pr(>F)
## df$DAY      3 0.00951 0.0031688   0.848 0.4688
## Residuals 240 0.89684 0.0037368
```

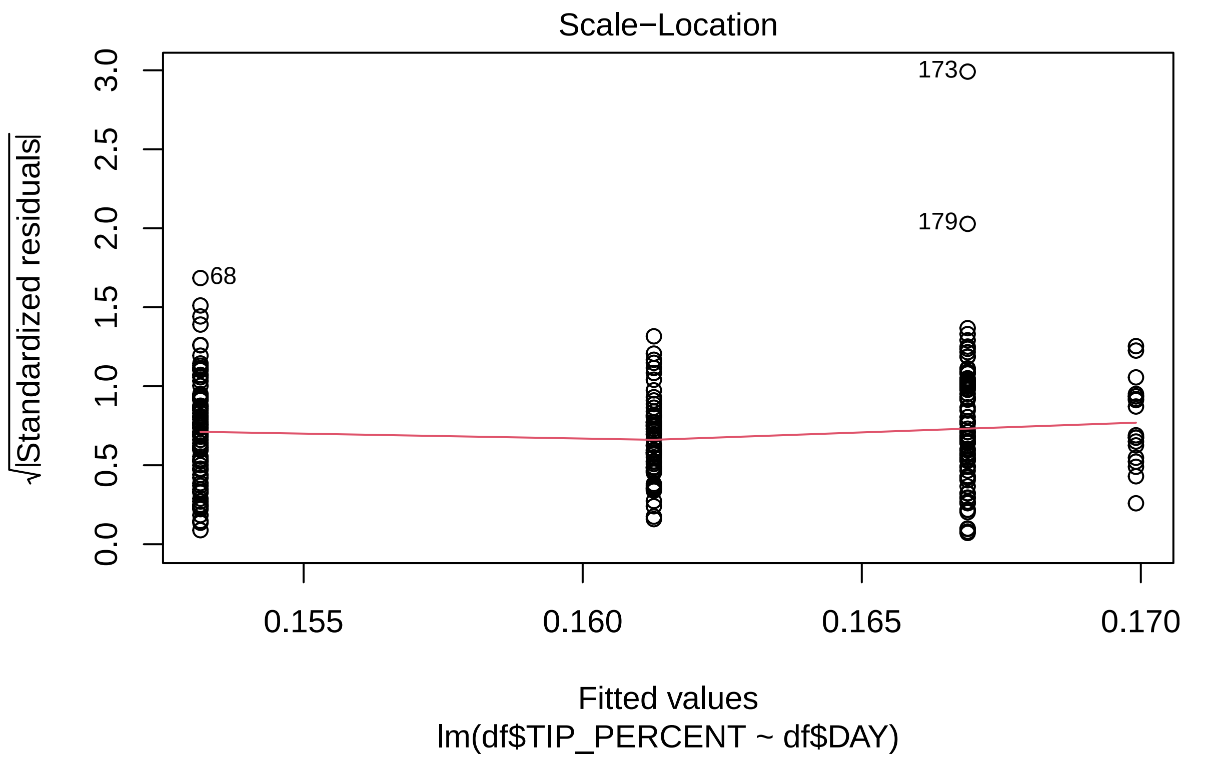
La p-value étant élevé, on conserve H_0. Le jour de la semaine n'a pas d'influence sur le montant des pourboires.

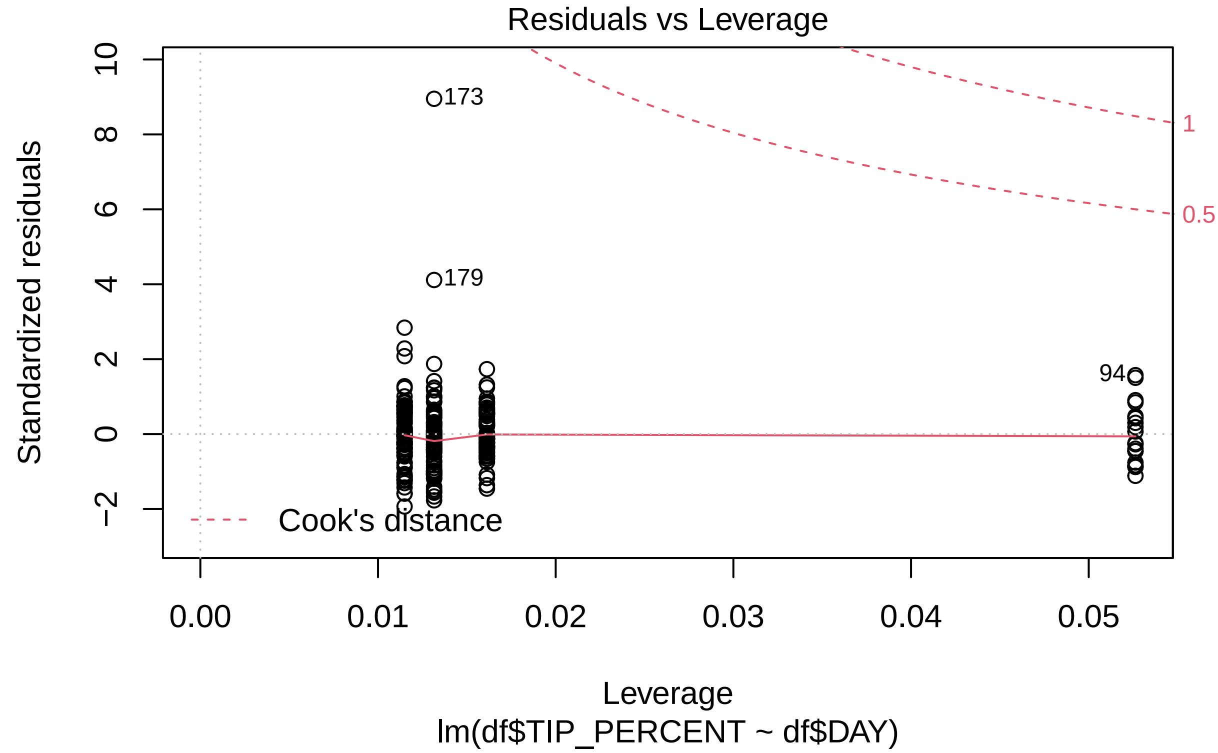
Question 4

```
plot(res)
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.2
173
179
68
0.155
0.160
0.165
0.170
Fitted values
lm(df$TIP_PERCENT ~ df$DAY)


Normal Q−Q
Standardized residuals
10
8
6
4
2
0
−2
173
179
68
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(df$TIP_PERCENT ~ df$DAY)





Le premier graphique `Residuals vs Fitted` permet de vérifier l'homoscédasticité. En effet les points semble dispersé de la même manière qu'importe le jour de la semaine.

Le `Normal Q-Q` plot montre que les points suivent plutôt bien les quantiles théorique d'une loi Normale. On valide l'hypothèse de normalité.

Le graphique de la distance de Cook nous permet de repérer les point influents. Le point 173 s'en rapproche mais n'est pas dans la zone 0.5. On peut donc le considérer comme un point normal.

Les point 68, 173 et 179 reviennent souvent dans en tant que points abérrant.

Question 5

Il faudrait vérfier les données sur ces 3 points abérrant et éventuellement refaire les tests précédants.