Université de Lille

L3 MIASHS

Nom : Belladjo

Prénom: Keroudine

modélisation statisiques

1) chargement de donnée

1. Charger la table de donn´ees Tips.csv.

data=read.table("Tips.csv", header = TRUE, sep = ";")

Code R:

```
> data=read.table("Tips.csv", header = TRUE, sep = ";")
> head(data)
  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
> 
```

2) Représentation des colonnes:
IDEN : est le numero ou l'indice' des clients du restaurant qui ont payé et donné le pourboire. (Variable qualitatives)
TOBILL : est le montant total de l'addition (la facture) exprimé en dollars. Cette colonne est intrinsèquement numériques de données (variable quantitatives) .
TIP : est le pourboire exprimé en dollars . Cette colonne est intrinsèquement numériques (valeurs quantitatives)
SIZE : donne le nombre de convives. Cette colonne est intrinsèquement numériques de données (variables quantitatives)
SEX : Indique le sexe des clients du restaurant qui ont payé et donné le pourboire. Ça vaut 0 ou 1 selon que le client est de sexe masculin ou féminin . C'est donc une variable facteur.
SMOKER : indique la zone fumeur ou non fumeur du restaurant. Ça vaut 0 ou 1 selon que le client est de fumeur ou non fumeur. C'est aussi une variable facteur.

DAY : indique le jour de la semaine. C'est également une variable facteur.
TIME : indique le moment de la journée (journée ou soirée) . Ça vaut 0 ou 1 selon qu'on est en journée ou en soirée. C'est aussi donc une variable facteur.

3) les variables:

```
> Iden= as.character(data$IDEN)
> Sex= as.factor(data$SEX)
> Day=as.factor(data$DAY)
> Facture=as.numeric(data$TOTBILL)
> Pourboire=as.numeric(data$TIP)
> Time= as.factor(data$TIME)
> SIZE= as.numeric(data$SIZE)
> Fumeur= as.factor(data$SMOKER)
>
```

2)Relation entre le montant de la facture et le montant du pourboire:

1) la corrélation est l'outils statistique
pour etudier le lien entre deux variables qualitatives

La corrélation entre la facture(TOTBILL) et le pourboire(TIP) :

```
> cor(Pourboire,Facture)
[1] 0.6757341
>
```

2) Il ya une relation entre le montant de la Facture (TOTBILL) et le montant du Pourboire (TIP) , elle est moyenne et vaut 0.6757341

3) data$TIPART = Pourboire*100/Facture
Cette nouvelle donnée donne le pourcentage des pourboire par rapport aux au montant de l'addition (la facture).

3) Comparaison des comportements des clients

1)Le test porte sur le test de comparaison des moyenne de deux échantillons des paramètre TIP et TIME

Notons:

M_1 la moyenne des pourboires qui ont été donné par les clients en journée et

M_2 la moyenne des pourboires qui ont été donné par les clients en soirée

Les hypothèses(Nulle et Alternatives):

H_0 : les clients sont autant généreux en soirée qu'en journée

H_1: les clients sont plus généreux en soirée qu'en journée

$\Longleftrightarrow$ $H_0 : M_1 = M_2$ contre $H_1 : M_1 > M_2$

On veut réaliser un Test de student a 2 échantillons :

Pour réaliser ce test il faut que les deux échantillons soient des gaussiens indépendants de variances égales. Les variances et la moyennes sont supposées inconnues.

Stat de test:

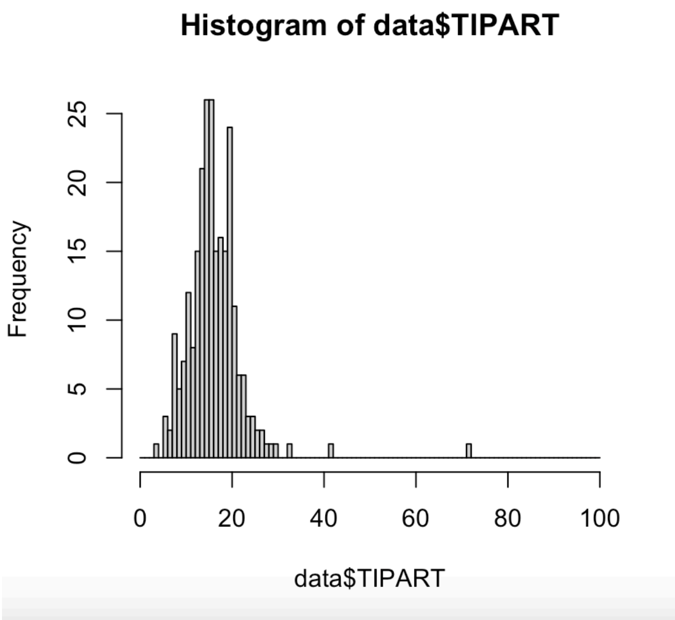
T est la statistique de test

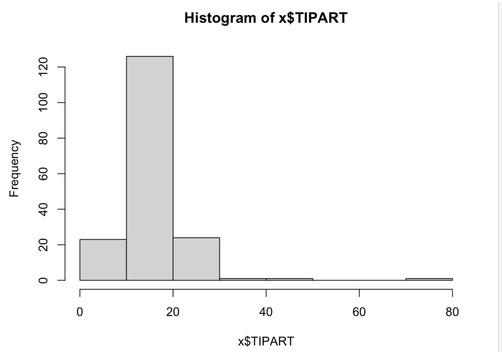
Sous H_0 T ~ $T_{(n1+n2-2)}$

2) Le test de Fisher

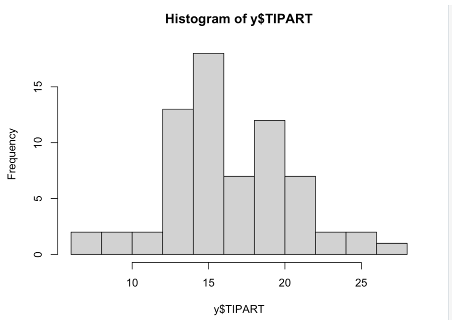
```
Erreur : > inattendu(e) dans >
> hist(data$TIPART,breaks=seq(0,100,1))
>
```

```
> x=data[data$TIME==1,]
> y=data[data$TIME==0,]
>
> hist(x$TIPART)
>
```






```
> hist(x$TIPART)
> hist(y$TIPART)
>
```



D’après ces histogrammes ci-desssus les données semble provenir des données gaussiennes. L’hypothèse du test de Fisher est donc acceptable.

```
> var.test(x$TIPART,y$TIPART)

        F test to compare two variances

data:  x$TIPART and y$TIPART
F = 2.8117, num df = 175, denom df = 67, p-value =
3.852e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.849205 4.122944
sample estimates:
ratio of variances
          2.811667
```

Interpretation du test:

La p-valeur est très faible on rejette alors H_0 , les variance non égale , on réalise un test de Welch

3) On utilise alors le test de Welch code R:

```
> x=data[data$TIME==1,]
> y=data[data$TIME==0,]
> t.test(x$TIPART,y$TIPART,var.equal = FALSE,"less")

        Welch Two Sample t-test

data:  x$TIPART and y$TIPART
t = -0.65404, df = 200.88, p-value = 0.2569
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 0.703772
sample estimates:
mean of x mean of y
 15.95178  16.41279

>
```

Conclusion du test:

P-valeurs = 0.26

Supposons α comme notre niveau du seuil de rejet

- Pour tout niveau $\alpha > 0.26$ (seuil de de rejet) on rejette H_0 au profit H_1

$\Longleftrightarrow$ les clients sont plus généreux en soirée qu'en journée

- Pour tout niveau $\alpha < 0.26$ on conserve H_0

$\Longleftrightarrow$ les clients sont autant généreux en soirée qu'en journée

3.2)En fonction du sexe du client

1)Le test porte sur le test de comparaison des moyenne de deux échantillons en des paramètre TIP et SEX .

Notons:

M_1 la moyenne des pourboires qui ont été donné par les clients de sexe masculin et

M_2 la moyenne des pourboires qui ont été donné par les clients de sexe féminin

Les hypothèses(Nulle et Alternatives):

H_0 : les homes sont autant généreux que les femmes

H_1: les hommes sont plus généreux que les femmes

$\Longleftrightarrow$ $H_0 : M_1 = M_2$ contre $H_1 : M_1 > M_2$

On veut réaliser un test de student a 2 échantillons :

Pour réaliser ce test il faut que les deux échantillons soient deux échantillons gaussiens indépendants de variances égales. Les variances et la moyennes sont supposées inconnues.

Stat de test:

T est la statistique de test

Sous H_0 , T ~ $T_{(n1+n2-2)}$

2) Le test de Fisher

```
> homme=data[data$SEX==0,]
> femme=data[data$SEX==1,]
> var.test(homme$TIPART,femme$TIPART)

        F test to compare two variances

data:  homme$TIPART and femme$TIPART
F = 1.4588, num df = 156, denom df = 86, p-value = 0.0542
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.993112 2.099515
sample estimates:
ratio of variances
          1.458847


>
```

Interpretation:

P-valeurs = 0.0542

Si on fixe α=0.05 (seuil de rejet)

On a $\alpha < 0.05$ (la p valeur grande) on on conserve alors H_0 , c'est à dire les variance sont égales

On applique alors le test de Student

3) On utilise le test de student :

```
> homme=data[data$SEX==0,]
> femme=data[data$SEX==1,]
> t.test(homme$TIPART,femme$TIPART,var.equal = TRUE,"less")

	Two Sample t-test

data:  homme$TIPART and femme$TIPART
t = -1.0834, df = 242, p-value = 0.1399
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.4632889
sample estimates:
mean of x mean of y 
 15.76505  16.64907 

>
```

Conclusion:

P-valeurs = 0.1399

- Pour tout niveau $\alpha > 0.1399$ (seuil de de rejet) on rejette H_0 au profit de H_1

$\Longleftrightarrow$ les hommes sont plus généreux que les femmes

- Pour tout niveau $\alpha < 0.1399$ on conserve H_0

$\Longleftrightarrow$ les homes sont autant généreux que les femmes

H_1: les hommes sont plus généreux que les femmes

3.3) En fonction de zone Fumeur/Non Fumeur

1) Le test porte sur le test de comparaison des moyenne de deux échantillons en des paramètre TIP et SMOKER .

Notons:

M_1 la moyenne des pourboires qui ont été donné par les clients Fumeurs et

M_2 la moyenne des pourboires qui ont été donné par les clients non fumeurs

Les hypothèses(Nulle et Alternatives):

H_0 : les clients sont autant généreux en étant fumeurs ou pas

H_1: les clients non fumeurs sont plus généreux que les non fumeurs

$\Longleftrightarrow$ $H_0 : M_1 = M_2$ contre $H_1 : M_1 > M_2$

On veut réaliser un test de student a 2 échantillons :

Pour réaliser ce test il faut que les deux échantillons soient des gaussiens indépendants de variances égales. Les variances et la moyennes sont supposées inconnues.

Stat de test:

T est la statistique de test

Sous H_0 T ~ $T_{(n1+n2-2)}$

2) Test de Fisher

```
> fumeur=data[data$SMOKER==0,]
> nonfumeur=data[data$SMOKER==1,]
> var.test(fumeur$TIPART,nonfumeur$TIPART)

        F test to compare two variances

data:  fumeur$TIPART and nonfumeur$TIPART
F = 0.21984, num df = 150, denom df = 92, p-value = 2.358e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1507563 0.3153274
sample estimates:
ratio of variances
         0.2198383
```

Conclusion du test:

La p-valeur est très faible on rejette alors H_0 , les variance ne sont donc pas égales , on réalise un test de Welch.

3) On utilise le test de Welch code R::

```
> t.test(fumeur$TIPART,nonfumeur$TIPART,var.equal = FALSE,"less")

	Welch Two Sample t-test

data:  fumeur$TIPART and nonfumeur$TIPART
t = -0.41123, df = 117.28, p-value = 0.3408
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.172548
sample estimates:
mean of x mean of y 
 15.93285  16.31960 
```

Interpretation:

P-valeurs = 0.34

Pour tout niveau $\alpha > 0.34$ (seuil de de rejet) on rejette H_0 au profit de

$\Longleftrightarrow$ les clients non fumeurs sont plus généreux que les non fumeurs

Pour tout niveau $\alpha < 0.34$ on conserve H_0

$\Longleftrightarrow$ les clients sont autant généreux en étant fumeurs ou pas

3.4) En fonction du jour de la semaine :

Question:

H_0 : Les pourboires sont égaux pour tous les jours de la semaines
H_1 : Les pourboires ne sont pas égaux pour les jours de la semaine

1) Vérifions que la classe de la colonne DAY est bien factor:

```
> Day=as.factor(data$DAY)
> str(Day)
 Factor w/ 4 levels "3","4","5","6": 4 4 4 4 4 4 4 4 4 4 ...
>
```

2. Réaliser une analyse de la variance à un facteur : le jour de la semaine.
Rappeler le modèle:

Le modele : Le principe de l’analyse de la variance est de déterminer, à l’aide d’un test statistique, si la part de dispersion imputable au facteur étudié ici le facteur DAY (les pourboires en fonction des jours), est significativement supérieure à la part résiduelle.

On suppose que toutes les variables sont indépendantes :

Effectivement ici nos variable sont indécents car c’est des pourboire de différents jours et de différentes personnes .

Les variables d’un même niveau sont indépendantes :

$\forall l \in \{1,2,\cdots,k\}$, $\forall$ i $\neq$ j, Y_i^l et Y_j^l sont indépendantes

Les variables de deux niveaux différents sont indépendantes:

$\forall l \neq$ m, $\forall$(i, j), Y_i^l et Y_j^m sont indépendantes.

- toutes les variables suivent une distribution normale

-l’espérance dépend du niveau k

- la variance est identique pour toutes les variables:

- $Y_i^l \sim N(\mu_l, \sigma^2)$, ∀ l∈{1,⋯,k}, ∀i∈{1,⋯,nl}

De manière équivalente, on pourra écrire que:

$Y_i^l = \mu_l + \varepsilon_l^i$ avec $\varepsilon_l^i \sim N(0,\sigma^2)$ et ind.

μ_l est l'espérance observée pour le niveau l du facteur :

En effet on peut réécrire le modèle de la manière suivante :

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$

3. l'ANOVA à un facteur est employée pour répondre à la question "est ce que les moyennes sont globalement différentes" comme dans notre cas ici.

un test statistique est employé pour déterminer si la variance factorielle est significativement supérieure à la variance résiduelle. Le test de validité globale du modèle permet de répondre à cette question. Il s'agit du test Fisher du rapport de ces deux variances.

Les hypothèses nulle et alternative de l'ANOVA à un facteur sont alors dans notre cas :
H_0 : Les pourboires sont égaux pour tous les jours de la semaines
H_1 : Les pourboires ne sont pas égaux pour les jours de la semaine

$\Longleftrightarrow H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ contre $H_0 : \exists(i, j)$ tel que $\mu_i \neq \mu_j$

la table d'analyse de la variance (code R):

```
> Day=as.factor(data$DAY)
> reg=lm(data$TIPART~Day)
> anova(reg)
Analysis of Variance Table

Response: data$TIPART
           Df Sum Sq Mean Sq F value Pr(>F)
Day         3   95.1  31.688   0.848 0.4688
Residuals 240 8968.4  37.368
>
```

La correspondance de chaque ligne et colonne:

Ligne 1:

- **DF : Nombre de degré de liberté (k-1)**

k-1= 3 ⇒ k= 4 qui est le nombre de modalité

- **Sum Sq : « somme des carrées » ici du modèle (SCM)**

SCM = 95.1

- **Mean Sq : Somme des carrées moyens ici du modèle**

$\frac{SCM}{K-1}$ **= 31.688**

- **F value : statistique de test**

$\frac{SCM/k-1}{SCR/n-k}$ **= 0.848**

- **Pr(>F): Probabilité critique :**

$pc = 0.4688$

Ligne 2:

- **DF : Nombre de degré de liberté (n-k)**

n-k= 240 ⇒ n= 244 qui est la taille de l'échantillon car on sait que k=4

- **Sum Sq : « somme des carrées » ici des résidus (SCR)**

SCM = 8968.4

- **Mean Sq : Somme des carrées moyens ici des résidus**

$\frac{SCR}{n-k}$ **= 37.368**

Interprétons la table et concluons:

$pc = 0.4688$ **(proba critique)**

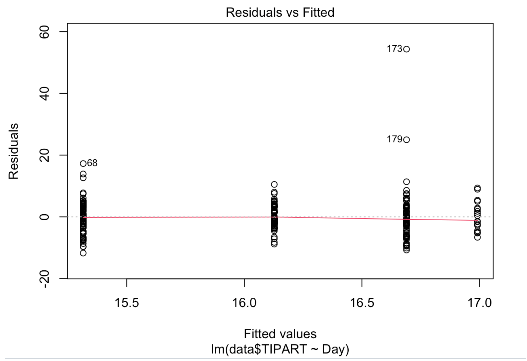
Fixons notre seuil de rejet $\alpha = 0.05$

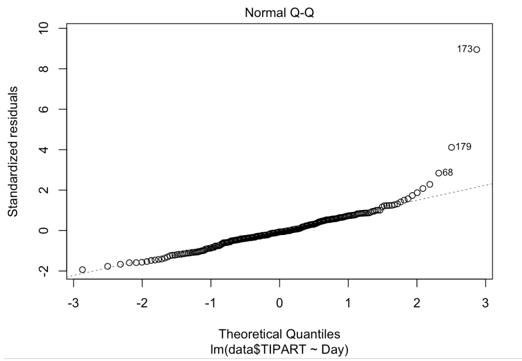
On a alors que $pc = 0.4688$ **>** $\alpha = 0.05$ **, on conserve alors** H_0

$\Longleftrightarrow$ La moyenne des pourboires sont égaux pour tous les jours de la semaines

4) Vérifier les hypothèses du modèle ainsi que la présence de points abberants ou influents avec plot :

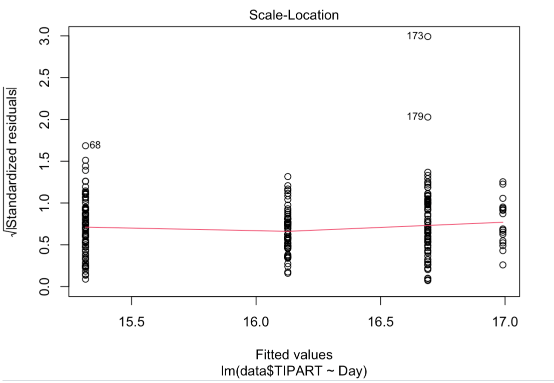
```
> plot(reg)
Tapez <Entrée> pour voir le graphique suivant :
Tapez <Entrée> pour voir le graphique suivant :
Tapez <Entrée> pour voir le graphique suivant :
Tapez <Entrée> pour voir le graphique suivant :
>
```

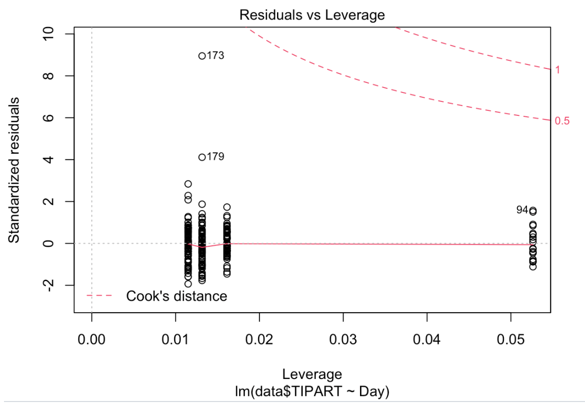




L'hypothèse de l'homoscédasticité est vérifiée car il y'a l'absence d'une forme en entonnoir et forme non linéaire des nuages des points mai il y'a quelques points aberrants comme le point 1730

L'hypothèse de la normalité est vérifié car les points sont à peu près alignés sur la première bissectrice.





5) Si vous deviez poursuivre l’étude, que feriez-vous ?
Validation des hypothèses du modèle:

Je valide le modèle car les condition sont vérifiées