GILLIARD Tallulah
LABARCHÈDE Mélody

Licence 3
MIASHS

Étude de Cas

Test du Chi-2 et analyse de la variance

19/11/2015

Première partie

Test du Chi-2 pour l'indépendance de variables qualitatives sous R

Les données dont nous allons nous servir sont issues du site de Gilles Hunault.
http://www.info.univ-angers.fr/~gh/Datasets/tips.htm
Dans cette étude, on veut établir un lien entre les pourboires, le sexe, l'aspect fumeur, le jour, le moment de la journée ainsi que d'autres variables.
Nous avons choisi deux variables qualitatives qui sont SMOKER qui prend la valeur 0 pour un non fumeur et la valeur 1 pour un fumeur, et la variable TIME qui prend la valeur 0 pour la journée et la valeur 1 pour la soirée.
On cherche à définir s'il y a une corrélation entre les variables SMOKE et TIME, de manière à savoir s'il y a plus ou moins de fumeur selon l'heure de la journée. Cela pourrait nous permettre ultérieurement d'établir des hypothèses sur le type de client (fumeur ou non) et les pourboires laissés.
On aurait aussi pu étudier l'influence de la variable SEXE sur le pourboire et établir des hypothèses sur le type de client (femme ou homme) et les pourboires laissés.

Dans notre cas, on peut supposer une corrélation entre la présence de personnes en soirée (plus de monde en soirée) et le nombre de pourboires selon le moment de la journée (plus de pourboires en soirée).

Le contexte se déroule dans un restaurant américain dans les années 80. Le restaurant comporte une zone fumeur et une zone non fumeur.

Affichage des données:

```
> data<-read.table("F:/Etude de cas/Lourme/SMOK.csv",sep=",",header=TRUE)
> data
```

Tableau de contingence:

```
> data[,1]->Smoker
> data[,2]->Time
> table(Smoker,Time) -> tableau2
> tableau2
```

On obtient le tableau suivant:

```
      Time
Smoker   0   1
     0  45 106
     1  23  70
```

Afin d'avoir davantage de clarté nous renommons les lignes et colonnes de ce tableau.

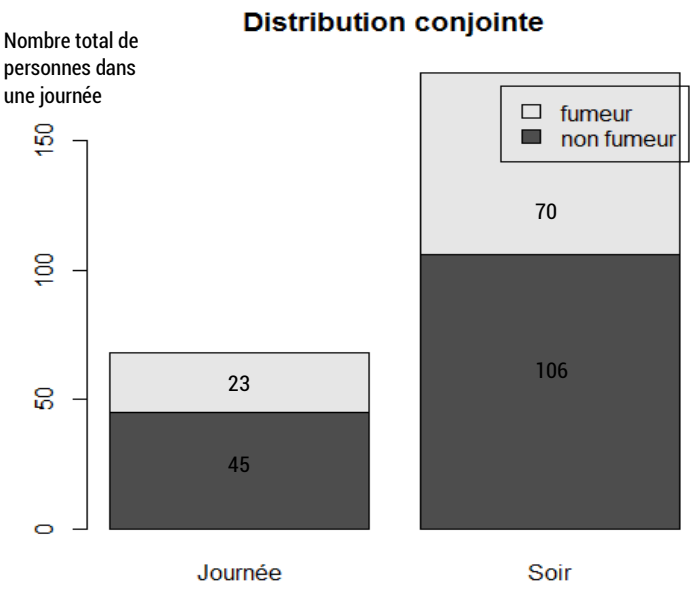
```
> rownames(tableau2)= c('Non fumeur', 'Fumeur')
> colnames(tableau2)= c('Journée', 'Soir')
> tableau2
```

```
      Time
Smoker        Journée  Soir
```

```
Non fumeur      45      106
Fumeur          23       70
```

Affichage du tableau de contingence de deux variables:

> barplot(tableau2, main= "Distribution conjointe", legend.text= c('non fumeur','fumeur'))

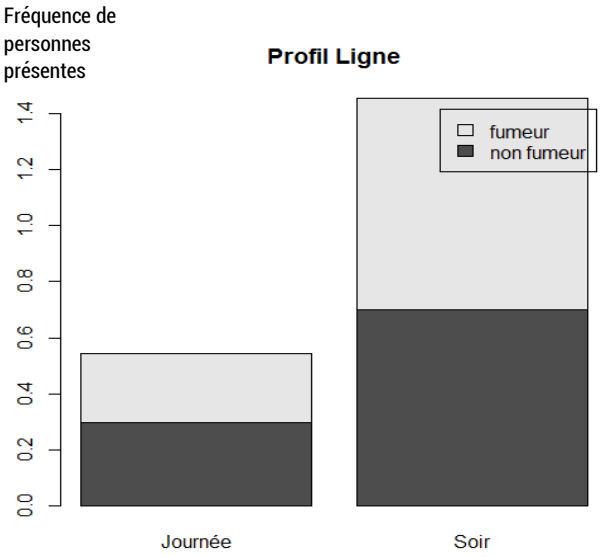


On voit qu'il y a plus de personnes présentes en soirée qu'en journée. Et on voit aussi qu'il y a plus de pourboires issus de non-fumeurs que de fumeurs.

Affichage du profil ligne:

(Profil ligne : on interprète le tableau en fonction des lignes, c'est-à-dire de la variable SMOKE)

> tableau3=prop.table(tableau2,margin=1)
> barplot(tableau3, main= "Profil Ligne", legend.text= c('non fumeur','fumeur'))

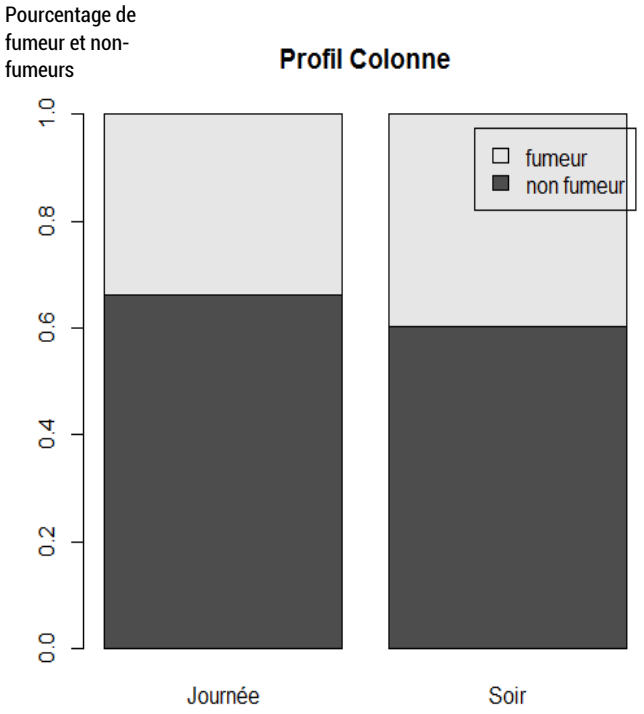


Peu importe que les gens fument ou non, il y a plus de gens le soir que le matin et donc plus de pourboires le soir que le matin.

Affichage du profil colonne:

(Profil colonne : on affiche le tableau en fonction des colonnes, c'est à dire)
Pour tt les gens du bar, totalité des personnes présentes

> tableau4=prop.table(tableau2,margin=2)
> barplot(tableau4, main= "Profil Colonne", legend.text= c('non fumeur','fumeur'))



Que ce soit le soir ou le matin, on a globalement la même proportion de pourboires donnés qu'on soit fumeur ou non. Les non-fumeurs donnent plus de pourboires dans les deux cas.

Tableau des effectifs théoriques

(Obtenues grâce au test du chi-2)

> mytest <- chisq.test(tableau2)
> summary(mytest)

```
          Length  Class  Mode
statistic 1       -none- numeric
parameter 1       -none- numeric
p.value   1       -none- numeric
method    1       -none- character
data.name 1       -none- character
```

```
observed  4     table  numeric
expected  4     -none- numeric
residuals 4     table  numeric
stdres    4     table  numeric
```

> *mytest$observed*

```
          Time
Smoker        Journée    Soir
 Non fumeur     45        106
 Fumeur         23         70
```

Le tableau des valeurs théoriques est :

> *mytest$expected*

```
          Time
Smoker            Journée       Soir
 Non fumeur     42.08197    108.91803
 Fumeur         25.91803     67.08197
```

Test du CHI 2

On fait un test du chi-2 pour tester la dépendance des variables SMOKE et TIME. On a pour hypothèse Ho : « les variables SMOKE et TIME sont dépendantes »

> *mytest <- chisq.test(tableau2)*
> *mytest*

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  tableau2
X-squared = 0.50537, df = 1, p-value = 0.4771
```

> *mytest$p.value*

```
[1] 0.4771486
```

On prend un seuil de 5%. p.value = 0,477 <0,05. On rejette Ho, il y a donc une indépendance des deux variables SMOKE et TIME.

Deuxième Partie

Analyse de la variance (ANOVA) à un facteur sous R

TP n°2 « ANOVA », exercice 7

Réalisez sous R une ANOVA à un facteur sur des données de votre choix en respectant les étapes/contraintes suivantes :

1- Choix d'un jeu de données.

> *data<-read.table("C:/Users/Utilisateur/Desktop/data.csv",sep=",",header= TRUE)*

Nous avons sélectionné un jeu de données appelé ici data qui provient du site de données de l'Université de Californie à Irvine.
https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

2- Description du facteur, de ses niveaux et de la variable à expliquer.

Dans cette étude, beaucoup de variables sont rassemblées et nous avons décidé d'étudier le facteur "Statut" composé des paramètres suivants: « Divorced » (divorcé), « Married AF spouse » (mariage militaire), « Married civ spouse » (mariage civil), « Married Spouse absent » (épouse absente), « Never married » (jamais marié), « Separated » (séparé) et « Widowed » (veuf).
La variable "Etude" sert à expliquer le nombre d'années d'études en fonction du statut. On cherche à comparer les moyenne des groupes du facteur (X) afin d'inférer une relation avec Y : le nombre d'années d'étude.

3- Enoncé des hypothèses en concurrence.

On considère dans un premier temps l'hypothèse nulle notée Ho qui indique qu'il n'y a pas de différence entre les moyennes des sept groupes.
La seconde hypothèse est l'alternative ou H1 qui définit qu'il existe une différence entre les différents groupes du facteur.
Ce qui permettrait d'inférer que X est la cause de Y.

4- Test d'ANOVA pour déterminer la présence/ l'absence d'effet significatif du facteur.

> *names(data)=c("Etude","Statut")*

> *tapply(Etude, Statut, var)*

```
        Divorced,     Married-AF-spouse,    Married-civ-spouse,
         5.185704              2.604743              7.384725
Married-spouse-absent,        Never-married,            Separated,
        10.640741              5.702673              6.055282
         Widowed,
         7.464002
```

Ci-dessus nous avons les diverses valeurs des variances pour chaque facteur.

> *tapply(Etude, Statut, length)*

```
          Divorced,     Married-AF-spouse,    Married-civ-spouse,
              4443                     23                  14976
Married-spouse-absent,        Never-married,              Separated,
               418                  10683                   1025
           Widowed,
               993
```

Ici, nous avons les différentes longueurs des facteurs étudiés.

Nous procédons à présent aux tests anova effectués de différentes manières.

> *oneway.test(Etude~Statut)*

```
        One-way analysis of means (not assuming equal variances)

data:  Etude and Statut
F = 66.323, num df = 6.00, denom df = 321.64, p-value < 2.2e-16
```

> *aov=lm(formula=Etude~Statut,data=data)*
> *summary(aov)*

```
Call:
lm(formula = Etude ~ Statut, data = data)

Residuals:
   Min     1Q  Median     3Q    Max
-9.3206 -1.3206 -0.3206  2.0375  6.9063

Coefficients:
                                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                     10.03038   0.03834      261.601   < 2e-16 ***
StatutMarried-AF-spouse,        0.14353    0.53429      0.269     0.788
StatutMarried-civ-spouse,       0.29026    0.04366      6.648     3.02e-11 ***
StatutMarried-spouse-absent,    -0.72177   0.13075      -5.520    3.41e-08 ***
StatutNever-married,            -0.06792   0.04562      -1.489    0.137
StatutSeparated,                -0.73673   0.08856      -8.319    < 2e-16 ***
StatutWidowed,                  -0.93673   0.08971      -10.442   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.556 on 32554 degrees of freedom
Multiple R-squared:  0.01334,   Adjusted R-squared:  0.01316
F-statistic: 73.35 on 6 and 32554 DF,  p-value: < 2.2e-16
```

> *aov.out = aov(Etude~Statut, data = data)*
> *summary(aov.out)*

```
              Df    SumSq     Mean Sq    F value     Pr(>F)
Statut         6    2875      479.1      73.35       <2e-16 ***
Residuals  32554   212636      6.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L'analyse des données ci-dessus indique que la différence entre les six groupes est significative (F = 73,35, ddl =6, p=< 2.2e-16). L'hypothèse H0 est rejetée. On peut donc conclure qu'il y a une influence du nombre d'années d'études sur le statut de la personne.

5- Test sur les contrastes en prenant un niveau du facteur pour référence.

> *summary.lm(aov.out)*

```
Call:
aov(formula = Etude ~ Statut, data = data)

Residuals:
   Min     1Q  Median     3Q    Max
-9.3206 -1.3206 -0.3206  2.0375  6.9063

Coefficients:
                                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                    10.03038    0.03834     261.601   < 2e-16 ***
StatutMarried-AF-spouse,        0.14353    0.53429       0.269     0.788
StatutMarried-civ-spouse,       0.29026    0.04366       6.648     3.02e-11 ***
StatutMarried-spouse-absent,   -0.72177    0.13075      -5.520     3.41e-08 ***
StatutNever-married,           -0.06792    0.04562      -1.489     0.137
StatutSeparated,               -0.73673    0.08856      -8.319     < 2e-16 ***
StatutWidowed,                 -0.93673    0.08971     -10.442    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.556 on 32554 degrees of freedom
Multiple R-squared:  0.01334,   Adjusted R-squared:  0.01316
F-statistic: 73.35 on 6 and 32554 DF,  p-value: < 2.2e-16
```

On constate que tous les facteurs hormis Married-AF-Spouse ont une probabilité inférieure à 0,05. Elle est donc moins significative que les autres et ne peut expliquer la relation.

6- Vérification des hypothèses du modèle.

On utilise dans un premier temps un test d'homogénéité de la variance.

> *bartlett.test(Etude~Statut, data = data)*

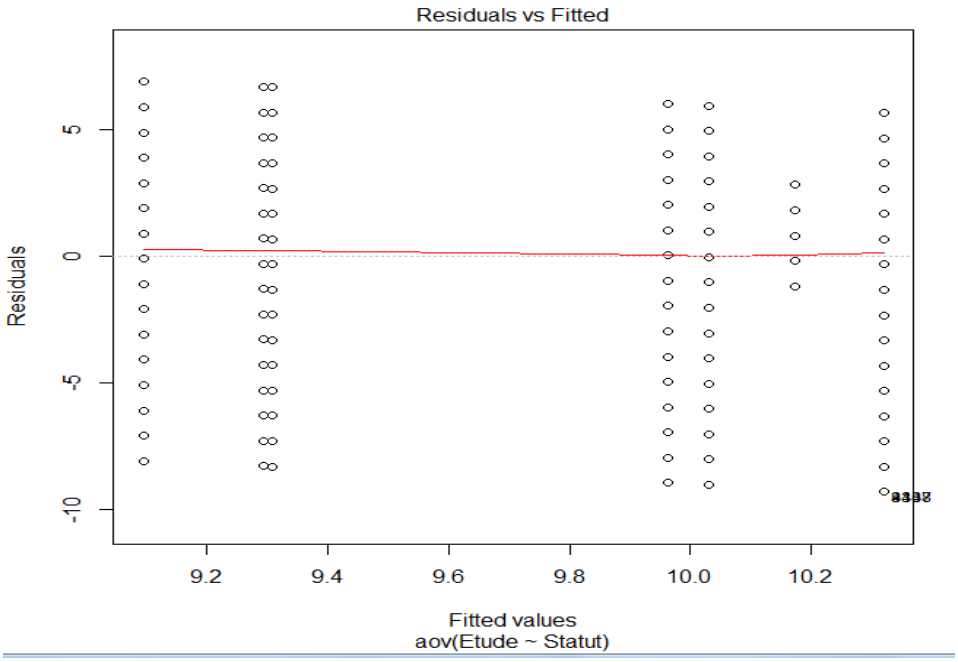
```
        Bartlett test of homogeneity of variances

data:  Etude by Statut
Bartlett's K-squared = 398.16, df = 6, p-value < 2.2e-16
```

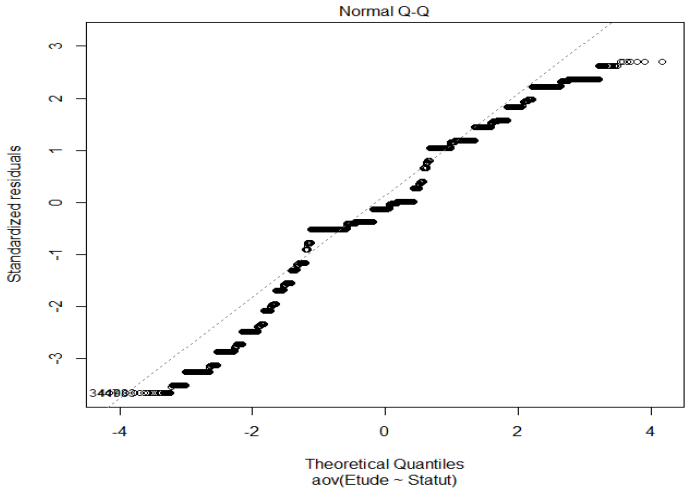
Il permet de confirmer ou d'infirmer l'hypothèse nulle. On voit ici que la p valeur est bien de < 2.2e-16 qui est inférieur à 0,05. L'hypothèse est bien infirmée.

On fait ensuite une représentation graphique.

> *plot(aov.out)*



Ce graphique montre une courbe quasi linéaire en y= 0, c'est la représentation entre les résidus et aov. Ce signifie que tous les résidus sont distribués de manière homogène. Les résidus sont dispersés et n'ont pas de structure précise ce qui confirme la normalité de ces derniers.



Ce graphique permet de comparer visuellement les deux distributions des échantillons. On remarque que les points sont globalement alignés, les deux échantillons ont donc des distributions similaires.