Statistique Inférentielle, M1 Info Lyon 2, 2021, TD2

Guillaume Metzler

Exercice 1 : Biostatistiques

On commence par charger le package nécessaire à l'import et à la lecture des données.

```
install.packages("data.table")
library(data.table)
```

Nous pouvons maintenant charger les données de l'étude et regarder de quoi il s'agit. L'option "header=TRUE" signifie que notre fichier comprend des noms pour les variables et que ces derniers seront utilisés comme nom de colonnes de votre jeu de données.

```
tryptone = read.table("Tryptone.dat.txt", header=TRUE)
head(tryptone)
```

```
##   Row Count1 Count2 Count3 Count4 Count5 Time Temp Conc
## 1   1      9      3     10     14     33   24   27  0.6
## 2   2     16     12     26     20     31   24   27  0.8
## 3   3     22     37     50     17     28   24   27  1.0
## 4   4     30     45     52     29     59   24   27  1.2
## 5   5     27     32     47     18     43   24   27  1.4
## 6   6     97     84    129    102     72   48   27  0.6
```

Ce jeu de données comporte 9 variables, la première est une sorte d'identifiant d'une expérience raélisée dans certaines conditions de "Températures" et de "concentrations" d'une certaine espèce chimique ou biologique. Les mesures sont également effectuées à des temps différents. Les variables "Count" vont compter le nombre de bactéries présentes pour les 5 souches de bactéries.

```
tryptone=rbind(data.frame(Souche=rep("Souche1",30),tryptone[,7:9],Count=tryptone[,2]),
               data.frame(Souche=rep("Souche2",30),tryptone[,7:9],Count=tryptone[,3]),
               data.frame(Souche=rep("Souche3",30),tryptone[,7:9],Count=tryptone[,4]),
               data.frame(Souche=rep("Souche4",30),tryptone[,7:9],Count=tryptone[,5]),
               data.frame(Souche=rep("Souche5",30),tryptone[,7:9],Count=tryptone[,6]))
```

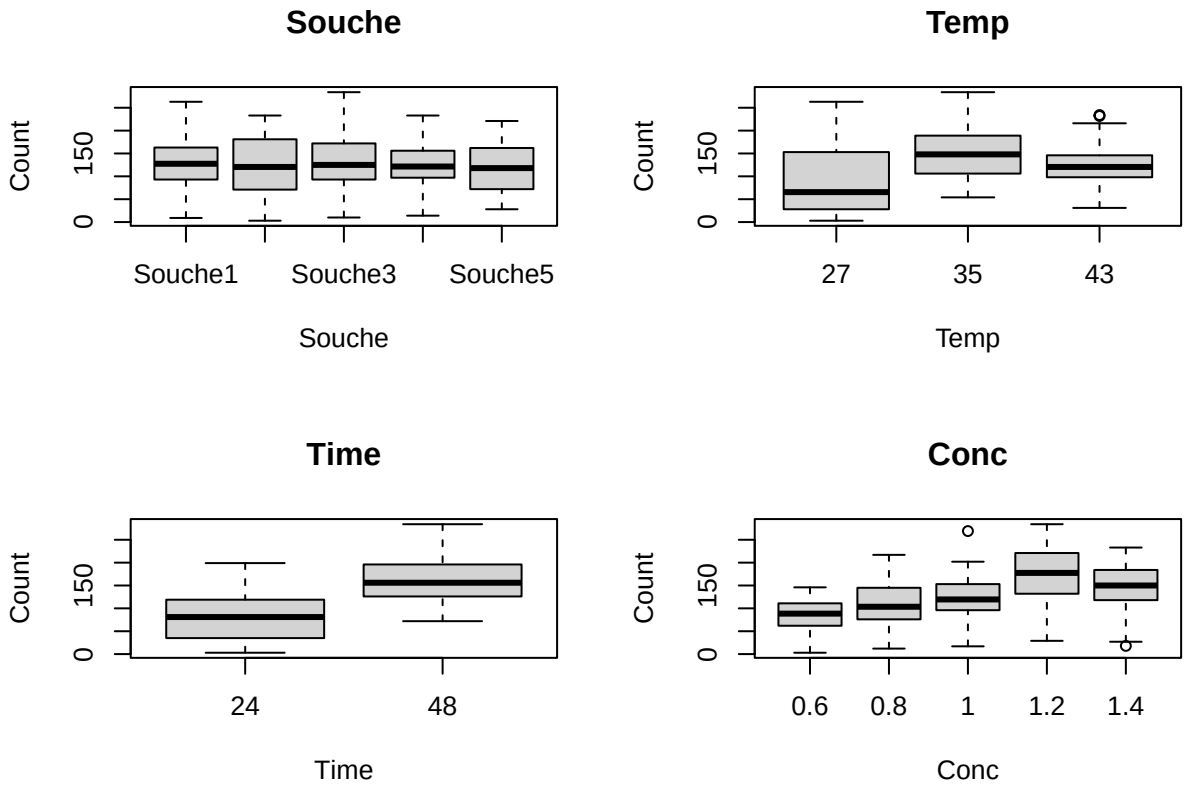
Nous aurions également pu utiliser la fonction *stack* de R

```
tryptone = read.table("Tryptone.dat.txt",header=TRUE)
tryptone=cbind(stack(tryptone[,2:6]),rep(tryptone$Time,5),rep(tryptone$Temp,5),rep(tryptone$Conc,5))
names(tryptone)=c('Count','Souche','Time','Temp','Conc')
levels(tryptone$Souche) = c("Souche1","Souche2","Souche3","Souche4","Souche5")
```

Lorsque l'on manipule des données à l'aide de R, il est souvent plus commode dans les mettre dans un format que l'on appelle data.frame et qui présente de nombreux avantages pour des études "graphiques" et "statistiques.

On pourra par exemple regarder, d'un point de vue graphique, si le nombre de bactéries varie d'une souche d'une souche à l'autre, dépend de la température, du temps de mesure ou encore de la concentration.

```
par(mfrow=c(2,2))
boxplot(Count~Souche, data = tryptone, main = 'Souche')
boxplot(Count~Temp, data = tryptone, main = 'Temp')
boxplot(Count~Time, data = tryptone, main = 'Time')
boxplot(Count~Conc, data = tryptone, main = 'Conc')
```



Pour les trois derniers graphes, l'étude d'un éventuel lien entre les facteurs "Time", "Conc" et "Temp" est faite sur le nombre total de bactéries, c'est-à-dire que l'on ne tient pas compte de la souche.\

Une première observation montre que le nombre de bactéries ne semble pas varier de façon signficative d'une souche à une autre. En revanche, les trois autrers facteurs semblent avoir un impact sur la numération du nombre de bactéries. Il nous faut employer des outils statistiques et notamment des tests afin de vérifier que l'impact de ces facteurs est réellement signicatif.

Influence du temps

On observe que le nombre de bactéries est plus faible à $t = 24$ qu'à $t = 48$, vérifions cela à l'aide d'un test statistique. \ Comme il s'agit d'étudier s'il y a une relation entre une variable quantitative et une variable qualitative, on va donc procéder à un test de Student où les hypothèses sont

H_0 : le nombre de cellules est indépendant du temps

v.s.

H_1 : le nombre de cellules est plus petit à t=24.

On indique cela avec l'alternative "less" car la modalité $t = 24$ est la première qui apparaît pour la variable "Time". On se rappelle que ce test est valable car nous avons des échantillons de tailles importantes (>30). Si cela n'avait pas été le cas, nous aurions dû vérifier que notre échantillon est bien gaussien à l'aide d'un test de Shapiro.\

```
t.test(Count~Time,data=tryptone,alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Count by Time
## t = -9.9741, df = 148, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 24 and group 48 is less than 0
## 95 percent confidence interval:
##       -Inf -67.11308
## sample estimates:
## mean in group 24 mean in group 48
##          82.2800         162.7467
```

On note que l'on effectue un test de Welch, ce qui signifie que l'on a fait les hypothèses que la variance de nos deux échantillons sont différentes. Nous aurions pu procéder, en amont, à un test Fisher afin de tester l'égalité des variances.

On observe que le nombre de bactéries semble plus important après 48 heures comme le confirme le test qui retourne une p-value de $2e^{-16}$. \

Dans le cadre de cours, nous serons toujours amenés à rejeter l'hypothèse nulle dans lorsque la $p-$value retournée par le test est inférieure à 0.05.\

Remarque : nous aurions également pu effectuer un test sur un seul échantillon en considérant l'échantillon "différence" entre la numération des cellules à $t = 24$ avec la numération des cellules à $t = 48$.

```
t.test(tryptone[tryptone$Time==24,"Count"]-tryptone[tryptone$Time==48,"Count"],alternative="less")
```

```
##
##  One Sample t-test
##
## data:  tryptone[tryptone$Time == 24, "Count"] - tryptone[tryptone$Time == 48, "Count"]
## t = -13.868, df = 74, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf -70.80193
## sample estimates:
## mean of x
## -80.46667
```

Le résultat reste inchangé.

Influence de la température

On doit à nouveau tester la dépendance entre une variable quantitative et une variable qualitative, cepndant, la variable qualitivative comporte cette fois-ci 3 modalités, on ne pourra donc plus effectuer un test de student, mais plutôt une analyse de variance (ANOVA). On va donc tester effectuer le test suivant

$$H_0 : \text{les deux variables sont indépendantes}$$

vs

$$H_1 : \text{la température a une incidence sur le nombre de cellules}$$

```
res = aov(Count~Temp,data=tryptone)
summary(res)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Temp          1  31329   31329   8.096 0.00507 **
## Residuals   148 572706    3870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La $p-$value est ici plus petite que 0.05, on en déduit que le facteur *Température* a bien une influence sur la numération du nombre de bactéries.\

Si on souhaite vérifier que ce test est valide, il faudrait à nouveau vérifier que nos échantillons sont gaussiens, ou, à défaut, que l'échantillon soit de taille suffisamment importante. En outre, il faudraitégalement vérifier que nos variances sont égales d'une population à une autre (test de Bartlett) et que nos résidus sont bien normalement distribués (test de shapiro)

Influence de la concentration

```
res = aov(Count~Conc,data=tryptone)
summary(res)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Conc          1 103491  103491    30.6 1.4e-07 ***
## Residuals   148 500545    3382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On doit à nouveau effectuer une ANOVA car la variable concentration prend un nombre restreint de valeurs (5 modalités), donc bien que continue d'un point de vue "physique", on va la considérer comme une variable catégorielle.\

La $p-$value de $1.4e^{-7}$ nous indique bien que le facteur "concentration" a une infleunce sur la numération des bactéries

Influence de l'origine de la souche

On peut faire de même pour les souches

```
res = aov(Count~Souche,data=tryptone)
summary(res)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Souche        4   4914    1228   0.297  0.879
## Residuals   145 599122    4132
```

Cette fois-ci, comme nous avons pu l'observer graphique, le facteur "souche" n'a aucune influence sur la numération.

Remarque : ici, nous avons simplement effectuer une analyse de variance à un seul facteur, mais nous aurions très bien pu faire de même avec 2 facteurs, en prenant en compte les intéractions (par exemple souche et concentration). Nous traiterons ces cas là dans les prochaines séances.

Exercice 2 : Statistiques pour le marketing

```
visa = read.table("VisaPremier.txt", header=TRUE, na.strings = ".")
head(visa)
```

```
##   matricul departem ptvente sexe age sitfamil anciente  csp codeqlt nbimpaye
## 1   148009       31       1 Shom  51     Fmar      238 Pcad       A        0
## 2   442153       82       6 Shom  52     Fmar      270 Pcad       A        0
## 3   552427       97       1 Shom  58     Fmar      139 Pcad       C        0
```

```
## 4   556005       40       1 Shom  27     Fcel       99 Psan       B        0
## 5   556686       65       1 Shom  49     Fsep       89 Pemp       A        0
## 6   642680       65       1 Shom  64     Fmar      216 Pcad       A        0
##   mtrejet nbopguic moycred3 aveparmo endette engagemt engagemc engagemm
## 1       0        0      115   701939       4   119216        0   119216
## 2       0        4    19579     8920       0        0        0        0
## 3       0        0       40     3402       0        0        0        0
## 4       0        0       17    76321       0        0        0        0
## 5       0        0      374   473350       0   209062    37859   171203
## 6       0        5       24    78462       0        0        0        0
##   nbcptvue moysold3 moycredi agemvt nbop mtfactur engageml nbvie  mtvie
## 1        2    35938      114     11   49   206016        0     1 152530
## 2        1   132468     4079     11   50    98500        0     0      0
## 3        1     1336       40     14    2     3394        0     0      0
## 4        1    12221       17     11   23        0        0     0      0
## 5        1    21187      208     11   49        0        0     1  21423
## 6        1     7154       24     11   67        0        0     0      0
##   nbeparmo mteparmo nbeparlo mteparlo nblivret mtlivret nbeparlt mteparlt
## 1        4   701939        2   520145        2   181794        0        0
## 2        3 19508920        2     8920        0        0        0        0
## 3        1     3402        0        0        1     3402        0        0
## 4        3    76321        1    46312        2    30009        0        0
## 5        5   473350        3   399999        2    73351        0        0
## 6        4    78462        1    34820        2     9803        1    33839
##   nbeparte mteparte nbbon    mtbon nbpaiecb nbcb nbcbptar avtscpte aveparfi
## 1        0        0     0        0       14    2        0  1303700   556967
## 2        0        0     1 19500000        5    2        0 19856243   133896
## 3        0        0     0        0        0    1        0   122745        0
## 4        0        0     0        0       14    2        0    83224        0
## 5        0        0     0        0       11    3        1   494773    21423
## 6        0        0     0        0       27    1        0    81218        0
##   cartevp sexer cartevpr nbjdebit
## 1    Coui     0        1        1
## 2    Coui     0        1        0
## 3    Coui     0        1        0
## 4    Coui     0        1        0
## 5    Coui     0        1       15
## 6    Coui     0        1        3
```

Pré-traitements

La première chose à faire lorsque l'on dispose d'un jeu de données est de vérifier que ce dernier soit "propre" afin que l'on puisse effectuer un traitement statistiques, *i.e.* il faudra par exemple vérifier que ce dernier ne contient pas de valeurs manquantes.

```
# Le jeu de données contient-il des valeurs manquantes ?
sum(is.na(visa))
```

```
## [1] 424
```

A priori le jeu de données contient beaucoup de valeurs manquantes. La façon dont on va remplacer les valeurs manquantes va dépendre de la nature des variables :

- **Quantitatives :** on remplace habituellement les valeurs manquantes par une valeur neutre comme la moyenne. Nous pourrions aussi utiliser un modèle de régression basé sur les autres variables pour traiter ces valeurs manquantes.

- **Qualitatives/Catégorielles :** on remplace généralement ces valeurs manquantes par le mode.

Il existe même des algorithmes dédiés spécialement au traitement des valeurs manquantes (fonction "mice" sous R), mais ce n'est pas le but ici. On commence par, identifier les variables catégorielles de notre jeu de données ainsi que les variables quantitatives. On va donc traiter deux data frame différentes pour cette phase de pré-traitements.

```
index_categ_features <- c(1,2,3,4,6,8,9,45,46,47)
visa_categ <- visa[,index_categ_features]
visa_quanti <- visa[,-index_categ_features]
```

On peut maintenant examiner les différentes variables. En commençant par les variables catégorielles

```
for (j in c(1:dim(visa_categ)[2])){

        nom_variable <- names(visa_categ)[j]
        if(sum(is.na(visa_categ[,j]))>0){
                print(paste(sprintf("La variable %s présente des valeurs manquantes",nom_variable)))
        }
}
```

```
## [1] "La variable departem présente des valeurs manquantes"
## [1] "La variable codeqlt présente des valeurs manquantes"
```

Donc deux features présentent des valeurs manquantes, les features "departem" et "codeqlt". On va remplacer ces valeurs manquantes par le mode. On va le faire en complétant la fonction qui précède.

```
for (j in c(1:dim(visa_categ)[2])){
        if(sum(is.na(visa_categ[,j]))>0){

                mode <- names(sort(table(visa_categ[,j]),decreasing=TRUE))[1]
                visa_categ[which(is.na(visa_categ[,j])),j] = mode
                visa_categ[,j] <- factor(visa_categ[,j])
        }
}
```

On peut appliquer la petite boucle qui précède pour vérifier que l'on n'a plus de valeurs manquantes ou avec :

```
summary(visa_categ)
```

```
##     matricul          departem      ptvente          sexe
##  Min.   : 113333   31     :698   Min.   :1.000   Length:1073
##  1st Qu.: 860436   65     :141   1st Qu.:1.000   Class :character
##  Median :1948586   32     : 69   Median :1.000   Mode  :character
##  Mean   :2489307   82     : 69   Mean   :1.664
##  3rd Qu.:3901594   75     : 14   3rd Qu.:2.000
##  Max.   :7589439   64     :  9   Max.   :7.000
##                    (Other): 73
##    sitfamil            csp            codeqlt   cartevp
##  Length:1073        Length:1073        A:207   Length:1073
##  Class :character   Class :character   B:436   Class :character
##  Mode  :character   Mode  :character   C:218   Mode  :character
##                                        D:168
##                                        E: 44
##
##
##      sexer           cartevpr
##  Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.3774   Mean   :0.3346
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

On remarque que la variable “ptvente” est considérée comme une variable numérique par R, il faut donc remédier à cela en précisant que les valeurs numériques font références à des catégories.

```
visa_categ$ptvente <- factor(visa_categ$ptvente)
```

Enfin, on peut remarquer que les variables “sexer”, “cartevpr” ne sont que des copies de “sexe” et “cartevp”, on peut donc les supprimer. Le matricule étant un identifiant, il n’est pas nécessaire à l’étude

```
visa_categ$matricul <- NULL
visa_categ$sexer <- NULL
visa_categ$cartevpr <- NULL
```

Traitons maintenant le cas des variables quantitatives, en remplacant les valeurs manquantes par la valeur moyenne.

```
for (j in c(1:dim(visa_quanti)[2])){
        if(sum(is.na(visa_quanti[,j]))>0){

                index <- which(is.na(visa_quanti[,j]))
                moy <- mean(visa_quanti[-index,j])
                visa_quanti[index,j] = moy
        }
}
```

On vérifie rapidement que le process ait bien fonctionné et on va en même temmps voir s’il n’y a pas d’autres variables que l’on pourrait exclure de l’étude.

```
summary(visa_quanti)
```

```
##       age           anciente        nbimpaye    mtrejet
##  Min.   :18.00   Min.   :  1.0   Min.   :0   Min.   :-51.00000
##  1st Qu.:33.00   1st Qu.: 45.0   1st Qu.:0   1st Qu.:  0.00000
##  Median :43.00   Median :136.0   Median :0   Median :  0.00000
##  Mean   :42.53   Mean   :157.1   Mean   :0   Mean   : -0.07269
##  3rd Qu.:52.00   3rd Qu.:216.0   3rd Qu.:0   3rd Qu.:  0.00000
##  Max.   :65.00   Max.   :870.0   Max.   :0   Max.   :  0.00000
##     nbopguic         moycred3          aveparmo         endette
##  Min.   : 0.000   Min.   :    0.00   Min.   :     0   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:    3.00   1st Qu.:     0   1st Qu.: 0.000
##  Median : 1.000   Median :   12.00   Median :  6017   Median : 0.000
##  Mean   : 1.505   Mean   :   47.63   Mean   : 57249   Mean   : 5.457
##  3rd Qu.: 2.000   3rd Qu.:   27.00   3rd Qu.: 57818   3rd Qu.: 6.000
##  Max.   :28.000   Max.   :19579.00   Max.   :970000   Max.   :99.000
##     engagemt          engagemc         engagemm          nbcptvue
##  Min.   :      0   Min.   :     0   Min.   :      0   Min.   :0.000
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:      0   1st Qu.:1.000
##  Median :      0   Median :     0   Median :      0   Median :1.000
##  Mean   :  77316   Mean   :  4199   Mean   :  20230   Mean   :1.028
##  3rd Qu.:  34927   3rd Qu.:     0   3rd Qu.:      0   3rd Qu.:1.000
##  Max.   :3472938   Max.   :500780   Max.   :1618242   Max.   :4.000
```

```
##     moysold3         moycredi           agemvt            nbop
##  Min.   :-70050   Min.   :   0.00   Min.   :  0.00   Min.   :  0
##  1st Qu.:   434   1st Qu.:   2.00   1st Qu.: 13.00   1st Qu.:  6
##  Median :  4371   Median :  11.00   Median : 13.00   Median : 25
##  Mean   : 10674   Mean   :  25.91   Mean   : 19.06   Mean   : 29
##  3rd Qu.: 11034   3rd Qu.:  24.00   3rd Qu.: 14.00   3rd Qu.: 43
##  Max.   :241827   Max.   :4079.00   Max.   :944.00   Max.   :262
##     mtfactur          engageml           nbvie             mtvie
##  Min.   :      0   Min.   :      0   Min.   : 0.0000   Min.   :      0
##  1st Qu.:      0   1st Qu.:      0   1st Qu.: 0.0000   1st Qu.:      0
##  Median :      0   Median :      0   Median : 0.0000   Median :      0
##  Mean   :  23379   Mean   :  52888   Mean   : 0.2395   Mean   :  35915
##  3rd Qu.:   3500   3rd Qu.:      0   3rd Qu.: 0.0000   3rd Qu.:      0
##  Max.   :1331530   Max.   :3472938   Max.   :13.0000   Max.   :5449561
##     nbeparmo        mteparmo           nbeparlo         mteparlo
##  Min.   :0.000   Min.   :       0   Min.   :0.0000   Min.   :     0
##  1st Qu.:0.000   1st Qu.:       0   1st Qu.:0.0000   1st Qu.:     0
##  Median :1.000   Median :    6017   Median :0.0000   Median :     0
##  Mean   :1.473   Mean   :   75442   Mean   :0.6524   Mean   : 32184
##  3rd Qu.:2.000   3rd Qu.:   58390   3rd Qu.:1.0000   3rd Qu.: 23854
##  Max.   :9.000   Max.   :19508920   Max.   :4.0000   Max.   :579603
##     nblivret         mtlivret         nbeparlt          mteparlt
##  Min.   :0.0000   Min.   :     0   Min.   :0.00000   Min.   :     0
##  1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.00000   1st Qu.:     0
##  Median :1.0000   Median :   127   Median :0.00000   Median :     0
##  Mean   :0.7586   Mean   : 20740   Mean   :0.05871   Mean   :  4325
##  3rd Qu.:1.0000   3rd Qu.: 15544   3rd Qu.:0.00000   3rd Qu.:     0
##  Max.   :4.0000   Max.   :970000   Max.   :6.00000   Max.   :559559
##     nbeparte           mteparte           nbbon              mtbon
##  Min.   :0.000000   Min.   :    0.00   Min.   :0.000000   Min.   :       0
##  1st Qu.:0.000000   1st Qu.:    0.00   1st Qu.:0.000000   1st Qu.:       0
##  Median :0.000000   Median :    0.00   Median :0.000000   Median :       0
##  Mean   :0.002796   Mean   :   19.71   Mean   :0.000932   Mean   :   18173
##  3rd Qu.:0.000000   3rd Qu.:    0.00   3rd Qu.:0.000000   3rd Qu.:       0
##  Max.   :1.000000   Max.   :21149.00   Max.   :1.000000   Max.   :19500000
##     nbpaiecb         nbcb          nbcbptar         avtscpte
##  Min.   : 0.0   Min.   :0.00   Min.   :0.0000   Min.   :       0
##  1st Qu.: 3.0   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:    3184
##  Median :11.5   Median :1.00   Median :0.0000   Median :   23993
##  Mean   :11.5   Mean   :1.07   Mean   :0.1361   Mean   :  146819
##  3rd Qu.:14.0   3rd Qu.:2.00   3rd Qu.:0.0000   3rd Qu.:  114807
##  Max.   :69.0   Max.   :5.00   Max.   :4.0000   Max.   :19856243
##     aveparfi          nbjdebit
##  Min.   :      0   Min.   :  0.00
##  1st Qu.:      0   1st Qu.:  0.00
##  Median :      0   Median :  0.00
##  Mean   :  50727   Mean   : 12.08
##  3rd Qu.:    500   3rd Qu.: 10.00
##  Max.   :7066619   Max.   :134.00
```

On remarque que la variable “nbimpaye” est constante, on peut donc la supprimer de notre étude

```
visa_quanti$nbimpaye <- NULL
```

Maintenant que les données sont prêtes, nous pouvons à nouveau regrouper notre base de données et démarrer

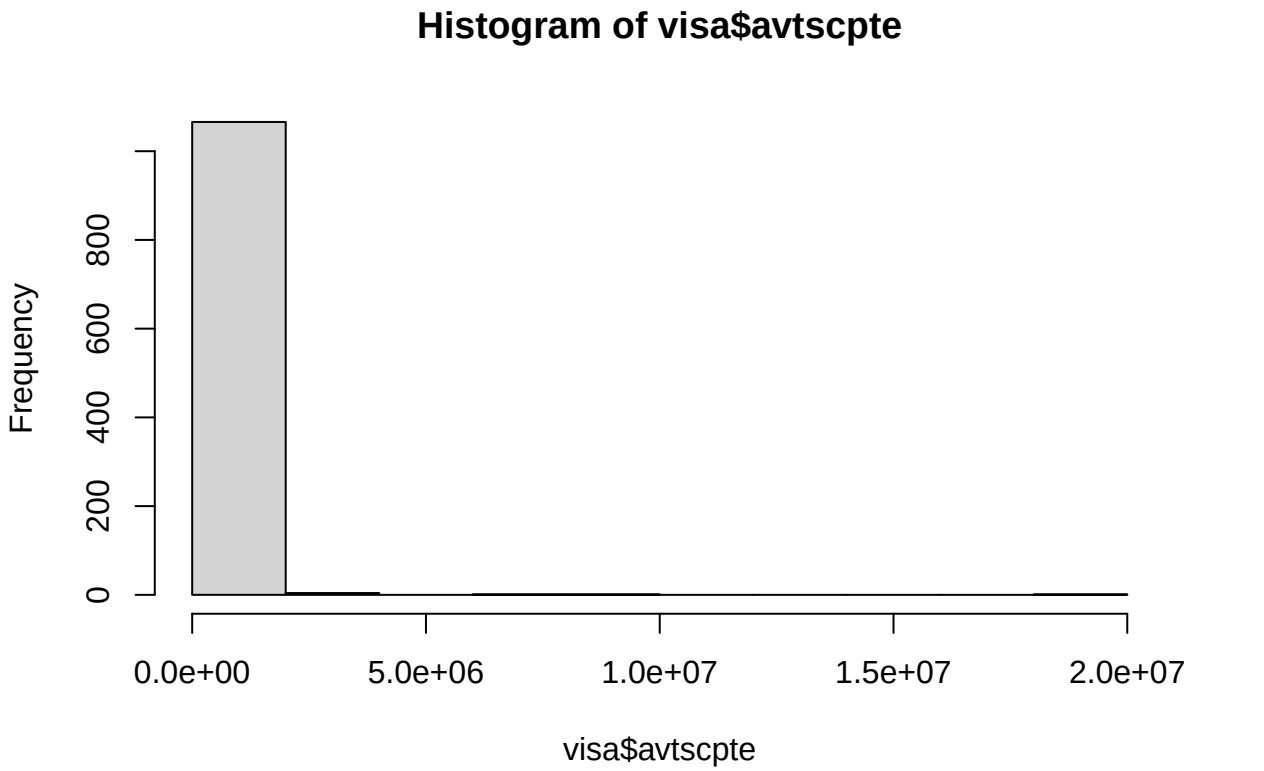
notre analyse

```
visa <- cbind(visa_categ,visa_quanti)
```

Analyse des données : étude de la corrélation

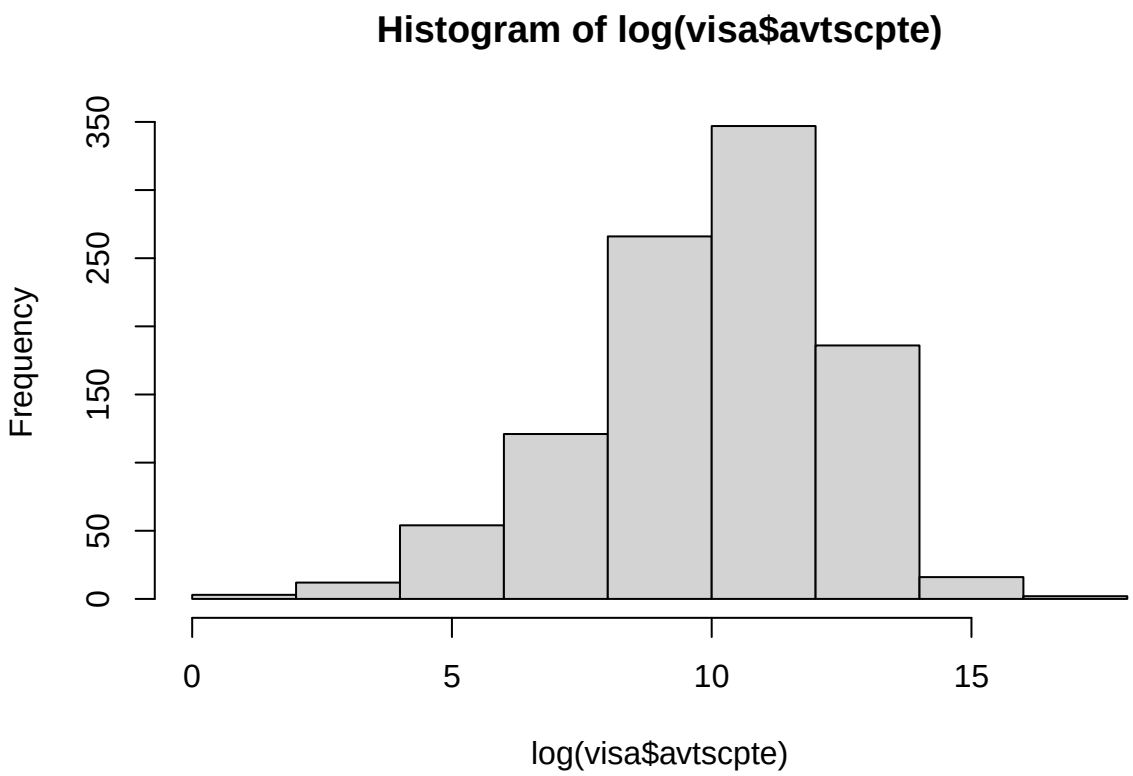
on va focaliser notre analyse sur les variables “avtscpte” et “nbpaiecb” et on va d’abord voire comment sont distribuées les valeurs de ces deux variables

```
hist(visa$avtscpte)
```



L’histogramme ne montre pas grand chose, si ce n’est que les valeurs des avoirs sont très étalées sur la droite. Peut-être qu’une transformation des données permettrait d’identifier plus facilement la distribution. Regardons ce qu’il se passe si on considère le log de nos données.

```
hist(log(visa$avtscpte))
```

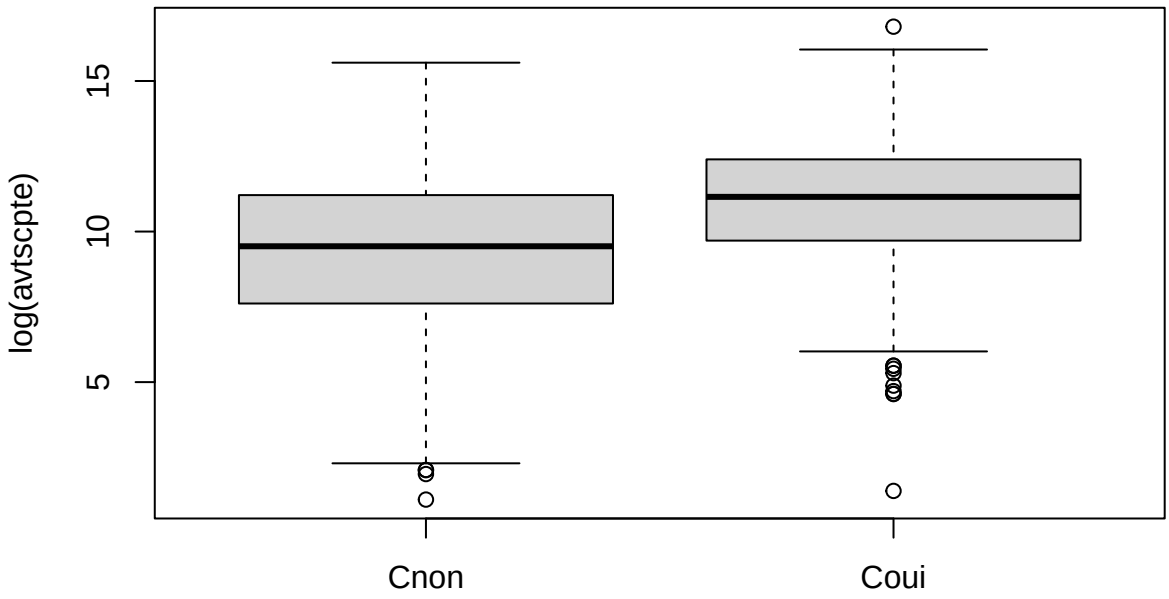


Nous pouvons maintenant essayer de voir s'il y a un lien éventuel entre le fait de posséder la carte Visa Premier et cette feature "avtscpte". Comme on veut étudier le lien entre une variable quanti et quali, on va donc utiliser un boxplot. Attention, étant donnée l'observation précédente, on va surtout considérer le log de la variable avtscpte

```
boxplot(log(avtscpte)~cartevp, data = visa)
```

```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out = z$out[z$group
## == : Une valeur extrême (-Inf) dans la boite de dispersion 1 n'est pas
## représentée
```

```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out = z$out[z$group
## == : Une valeur extrême (-Inf) dans la boite de dispersion 2 n'est pas
## représentée
```



A priori, il semblerait que les détenteurs d'une carte Visa Premier possèdent plus d'avoirs que les autres. On va donc procéder à un test statistique pour confirmer cette observation. Comme on veut étudier l'imapct d'une variable quali à deux modalités sur une variable quanti, on va procéder à un test de Student. On fera un test unilatéral inférieur, c'est-à-dire que l'on va préciser alternative = "less" dans les options du test.

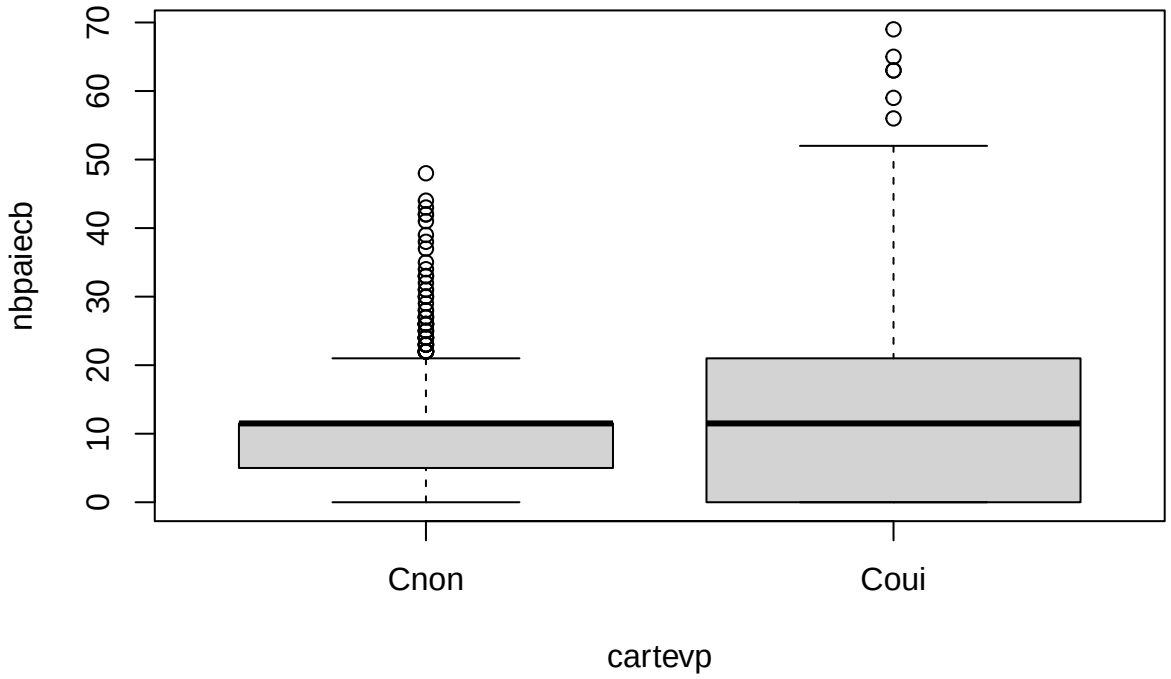
```
t.test(avtscpte~cartevp, data = visa, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  avtscpte by cartevp
## t = -3.3406, df = 377.16, p-value = 0.0004597
## alternative hypothesis: true difference in means between group Cnon and group Coui is less than 0
## 95 percent confidence interval:
##       -Inf -108806.2
## sample estimates:
## mean in group Cnon mean in group Coui
##           74931.98           289792.86
```

La $p-$value du test étant plus petite que 0.05, on peut donc rejeter l'hypothèse d'indépendance et dire que les deux variables sont liées. On peut même dire que le fait de posséder la carte Visa Premier implique des un montant d'avoirs supérieur à ceux ne possédant pas cette carte.\

Procédons de la même façon avec la variable "nbpaiecb". Nous pourrions refaire un histogramme des valeurs, mais on va de suite passer à l'analsye de corrélations

```
boxplot(nbpaiecb~cartevp, data = visa)
```



La réponse est un peu évidente en regardant ces deux boxplots. On peut voir que les médianes sont identiques, en revanche les distributions sont des différentes d'une modalité à une autre donc on peut espérer que la variable "cartevp" ait une influence sur la variable "nbpaiecb".

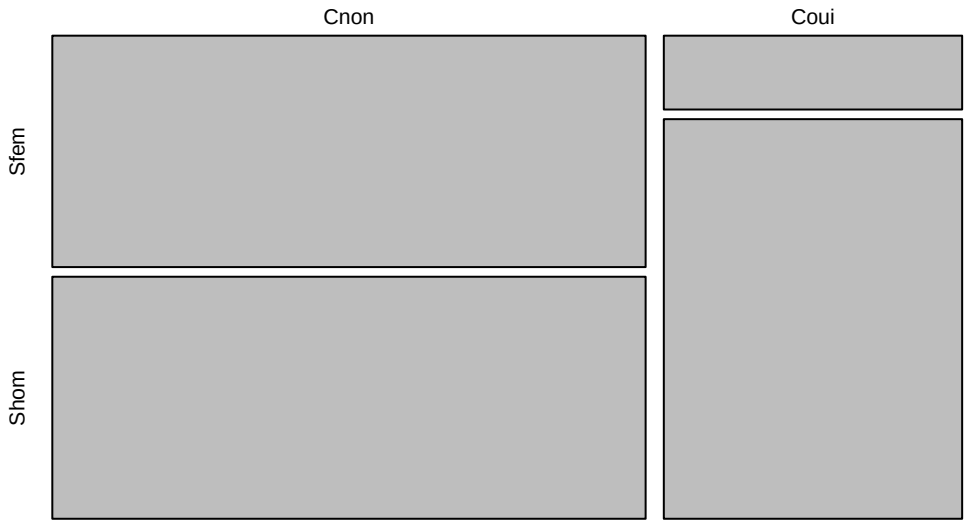
```
t.test(nbpaiecb~cartevp, data = visa)
```

```
##
##  Welch Two Sample t-test
##
## data:  nbpaiecb by cartevp
## t = -4.2469, df = 474.14, p-value = 2.61e-05
## alternative hypothesis: true difference in means between group Cnon and group Coui is not equal to 0
## 95 percent confidence interval:
##  -5.001106 -1.837140
## sample estimates:
## mean in group Cnon mean in group Coui
##           10.35667           13.77580
```

Ce test confirme bien que les deux variables sont liées. \

On finit cette étude en regardant si les variables "sexe" et "cartevp" sont éventuellement liées. Il s'agit de deux variables qualitatives cette fois-ci, on va donc représenter cela sous la forme d'un graphe un peu différent mais pas par un boxplot.

```
plot(table(visa$cartevp, visa$sexe))
```



Le graphique suggère effectivement que la proportion de femme ayant une carte vp n'est pas identique à la proportion de femme ne possédant pas de carte vp. On va confirmer cela à l'aide d'un test du Khi-deux (pour rappel, ce test est valable si le nombre d'individus dans chaque croisement est supérieur à 5)

```
chisq.test(visa$sexe,visa$cartevp)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  visa$sexe and visa$cartevp
## X-squared = 111.19, df = 1, p-value < 2.2e-16
```

A nouveau le sexe a bien une influence sur le fait d'être détenteur de la carte Visa Premier.