Fiche TD avec le logiciel R : tpRr4

BIO2049L Statistique pour la Biologie et Bioinformatique
Tests statistiques avec R

I. Amat, D. Fouchet, L. Keurinck, V. Lacroix, J.R. Lobry, A. Mary, L. Nicvert, M.C. Venner & S. Venner

Nous vous proposons de mettre en oeuvre sous R quelques tests d'hypothèse. Il est nécessaire d'avoir déjà fait l'analyse exploratoire des données sur la base du fascicule 1 (`https://pbil.univ-lyon1.fr/R/pdf/tpRg4.pdf`) et que vous ayez les jeux de données `etudiants` et `bebes` dans votre environnement `Rstudio`.

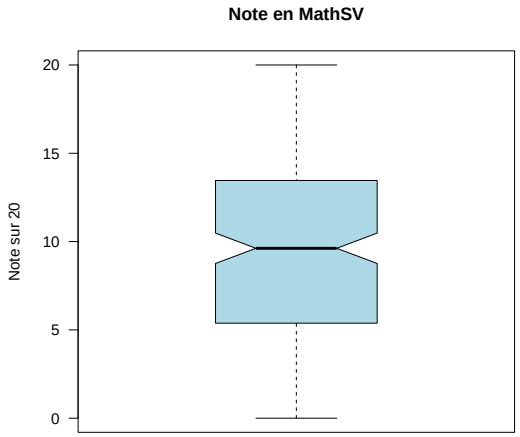
Contents

1 Intervalle de confiance d'une statistique

1.1 Restauration des jeux de données `etudiants` et `bebes`

Cette section vous permet de vérifier que vous arrivez à restaurer les jeux de données utilisés lors du précédent TD[1] . Restaurez votre environnement de travail en cliquant sur la première icône sous l'onglet « *Environment* ». Donnez la distribution des notes de MathSV à l'aide d'une boîte à moustache :

```
boxplot(etudiants$note, col = "lightblue",
                        main = "Note en MathSV",
                        las = 1, notch = TRUE,
                        ylab = "Note sur 20")
```



Nous allons commencer par quantifier les indications données par ce graphique pour vérifier que l'on retrouve bien les mêmes valeurs. L'indicateur de tendance centrale utilisé dans les boîtes à moustaches est la médiane. Il est calculé par la fonction `median()` ou bien par la fonction `quantile()` avec une probabilité de 50 % :

```
median(etudiants$note)
```

```
[1] 9.62
```

```
###
quantile(etudiants$note, prob = 0.5)
```

```
 50%
9.62
```

Les extrémités de la boîte sont données par le premier et le troisième quartiles, il y a donc la moitié des étudiants qui ont eu une note comprise entre ces deux valeurs :

```
quantile(etudiants$note, prob = c(0.25, 0.75))
```

[1] Si vous n'arrivez pas à restaurer les données de la séance précédente, collez dans votre console la commande suivante :
`load(url("https://pbil.univ-lyon1.fr/R/donnees/tpRg2/finTP1.rda"))`
Si vous avez des problèmes pour gérer des noms avec des accents, collez dans votre console la commande suivante :
`load(url("https://pbil.univ-lyon1.fr/R/donnees/tpRg2/finTP1bis.rda"))`

```
 25%   75%
5.38 13.46
```

Donnez le code R permettant de calculer la masse médiane des bébés à la naissance, vous devez obtenir le résultat suivant :

```
[1] 3.43
```

Réponse 1 :

Donnez le code R permettant de calculer le premier et le troisième quartile de la masse des bébés à la naissance, vous devez obtenir le résultat suivant :

```
 25%  75%
3.09 3.74
```

Réponse 2:

1.2 Intervalle de confiance de la moyenne

Vous avez vu en première année qu'un intervalle de confiance pour la moyenne était défini par :

$$\hat{\mu} \pm \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$$

où n est la taille de l'échantillon, $\hat{\mu}$ l'estimateur de la moyenne de la population, égal à la moyenne de l'échantillon, $\bar{x}$, calculé par la fonction `mean()`, $\hat{\sigma}$ l'estimateur non biaisé de l'écart-type calculé par la fonction `sd()` ($\hat{\sigma^2} = \frac{n}{n-1} s^2$), et t la fonction de répartition de la loi de Student à $n-1$ degrés de liberté donnant le quantile $1-\frac{\alpha}{2}$ calculé par la fonction `qt()`. On peut faire les calculs « à la main » comme pendant les TD de L1 en mathSV :

```
alpha <- 0.05
###
n <- nrow(etudiants) # Le nombre d'individus
###
mu <- mean(etudiants$note)
###
qt(1 - alpha/2, n - 1)
```

```
[1] 1.970855
```

```
# Notez que l'approximation par la loi normale est bonne ici :
###
qnorm(1 - alpha/2)
```

```
[1] 1.959964
```

```
###
delta <- sd(etudiants$note)*qt(1 - alpha/2, n - 1)/sqrt(n)
###
delta
```

```
[1] 0.7189554
```

```
mu - delta
```

```
[1] 9.207226
```

```
mu + delta
```

```
[1] 10.64514
```

Une façon équivalente et plus compacte de faire ce calcul sous R est de récupérer les valeurs de retour de la fonction `t.test()` :

```
t.test(etudiants$note, conf.level = 1 - alpha)$conf.int
```

```
[1]  9.207226 10.645137
attr(,"conf.level")
[1] 0.95
```

Donnez le code R permettant de calculer un intervalle de confiance à 95 % pour la moyenne de la masse des bébés à la naissance, vous devez obtenir le résultat suivant :

```
[1] 3.383282 3.443299
attr(,"conf.level")
[1] 0.95
```

Réponse 3:

2 Test d'indépendance du χ^2 (quali-quali)

Deux variables qualitatives ne sont pas indépendantes en probabilité quand la connaissance de la réalisation d'une modalité de l'une influence la probabilité de réalisation d'une modalité de l'autre variable. Prenons un exemple concret avec `sequence`, la séquence de l'étudiant, et `ffsu` indiquant s'il est inscrit en sport de niveau 2. Lors des inscriptions pédagogiques, les sportifs de niveau 2 sont presque toujours inscrits en séquence 3 pour la transversale (=l'enseignement de leur spécialité sportive) afin de les rendre disponibles pour les créneaux d'entraînement et de compétition qui leur sont réservés. On se retrouve donc avec un excès d'étudiants sportifs en séquence 2 pour MathSV :

```
table(etudiants$sequence, etudiants$ffsu)
```

```
   NON OUI
2 135  25
3  60   0
```

Si on sait qu'un étudiant est sportif de niveau 2 alors il y a une probabilité très forte qu'il soit inscrit en séquence 2. Dans l'autre sens, si on sait qu'un étudiant est en séquence 3 il y a une probabilité très faible qu'il soit sportif de niveau 2. La connaissance de la réalisation de la modalité d'une variable influence la probabilité de réalisation d'une modalité de l'autre de l'autre variable. Les variables `sequence` et `ffsu` ne sont donc pas indépendantes. En pratique, les situations rencontrées se sont pas toujours aussi tranchées, et on a besoin d'un test pour décider si deux variables sont indépendantes. On utilise généralement à cet effet le test d'indépendance dit « du χ^2 » parce que sa statistique suit asymptotiquement la loi de probabilité du χ^2. La fonction R `chisq.test()` permet de faire rouler le test. La dénomination `chisq` est une contraction anglaise pour « *chi squared* », soit « chi au carré ». Exécutons le test d'indépendance du χ^2 entre les variables `sequence` et `ffsu` :

```
chisq.test(table(etudiants$sequence, etudiants$ffsu))
```

```
        Pearson's Chi-squared test with Yates' continuity correction
data:  table(etudiants$sequence, etudiants$ffsu)
X-squared = 9.0825, df = 1, p-value = 0.002581
```

Le résultat du test s'interprête en examinant la valeur de `p-value`, soit la « valeur p » pour « valeur de probabilité ». On peut l'extraire avec l'opérateur `$` de la façon suivante :

```
chisq.test(table(etudiants$sequence, etudiants$ffsu))$p.value
```
```
[1] 0.00258064
```

La probabilité critique (ou *p-value*) est la probabilité, sous l'hypothèse nulle d'indépendance entre les deux variables, pour que la statistique du test (soit χ^2_{obs}) soit égale ou supérieure à 9.0825.

Dit autrement, plus la *p-value* est proche de zéro (une probabilité peut prendre des valeurs comprises entre 0 et 1), plus il est improbable d'obtenir une telle répartition des étudiants dans un scénario où la séquence et le statut de sportif seraient indépendants entre eux.

On confronte la *p-value* avec le risque de première espèce choisi, ici $\alpha = 0.05$, pour prendre une décision :

```
alpha <- 0.05
###
valeur.p <- chisq.test(table(etudiants$sequence, etudiants$ffsu))$p.value
###
if(valeur.p < alpha){
  print("Je rejette l'hypothèse nulle")
} else {
  print("Je ne peux pas rejeter l'hypothèse nulle")
}
```
```
[1] "Je rejette l'hypothèse nulle"
```

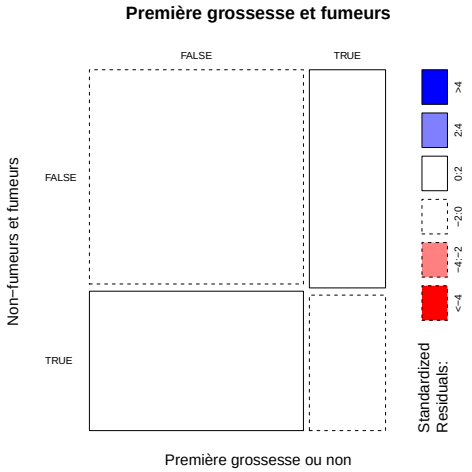
La statistique du test est un indice d'écart entre la distribution observée et la distribution théorique. Sous l'hypothèse nulle il y a une probabilité très faible que d'observer un écart aussi grand par hasard. C'est pourquoi, avec un risque de première espèce $\alpha = 0.05$, on décide de rejeter l'hypothèse nulle : il n'y a pas indépendance entre la séquence des étudiants et leur inscription en sport de niveau 2. Donnez le code R permettant de tester s'il y a indépendance entre la parité des mères et leur consommation de tabac, vous devez obtenir le résultat suivant :

```
        Pearson's Chi-squared test with Yates' continuity correction
data:  table(bebes$parity, bebes$smoke)
X-squared = 0.075556, df = 1, p-value = 0.7834
```

Réponse 4:

Les variables `parity` et `smoke` sont-elles indépendantes ? Faites le test, et pour conclure, aidez-vous également du graphique obtenu ci-dessous à l'aide de la fonction `mosaicplot()` déjà vue dans le fascicule 1.

```
mosaicplot(table(bebes$parity, bebes$smoke),
                          main = "Première grossesse et fumeurs",
                          xlab = "Première grossesse ou non",
                          ylab = "Non-fumeurs et fumeurs",
                          las = 1, shade = T)
```



Réponse 5:

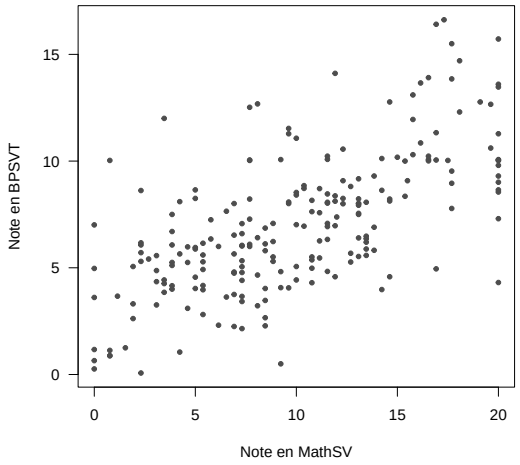
3 Corrélation et régression linéaire simple (quanti-quanti)

3.1 Test de Corrélation

Nous avions vu lors du TD précédent (*Exploration Graphique d'une Jeu de Données*) qu'il semblait y avoir une corrélation positive entre la note de MathSV et celle de BPSVT[2] (section 7.2). Le nuage de points semble en effet s'étirer selon un axe de pente positive:

```
plot(etudiants$note, etudiants$nBPSVT, las = 1,
     xlab = "Note en MathSV",
     ylab = "Note en BPSVT",
       pch = 20,
      col=grey(0.3))
```

[2]Variable `note`= note obtenue dans l'UE MathSV
Variable `nBPSVT`=note obtenue dans l'UE « Bases de Physique pour les Sciences de la Vie et de la Terre »



La fonction `cor.test()` permet de tester si le coefficient de corrélation linéaire est significativement différent de zéro :

```
 cor(etudiants$note, etudiants$nBPSVT)
[1] 0.6534518
 cor.test(etudiants$note, etudiants$nBPSVT)
        Pearson's product-moment correlation
data:  etudiants$note and etudiants$nBPSVT
t = 12.746, df = 218, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5704895 0.7232137
sample estimates:
      cor
0.6534518
```

Rappelez la formule du coefficient de corrélation linéaire appliqué à deux variables X et Y, puis indiquer l'hypothèse nulle et interprétez la sortie du test ci-dessus .

Réponse 6:

Donnez le code R permettant de tester si le coefficient de corrélation linéaire entre la masse des bébés à la naissance et le temps de gestation est significativement différent de zéro, vous devez obtenir le résultat suivant :

```
        Pearson's product-moment correlation
data:  bebes$bwt and bebes$gestation
t = 15.749, df = 1171, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3696967 0.4642158
sample estimates:
      cor
0.4180872
```

Réponse 7:

3.2 La Régression Linéaire Simple

Dans la section précédente, vous avez conclu que les 2 variables quantitatives `note` et `nBPSVT` étaient linéairement dépendantes l'une de l'autre. Cela légitime la recherche du modèle linéaire décrivant le mieux la relation entre ces variables, en estimant les paramètres de la droite de régression simple. Dans le cas où les mesures sont issues d'un échantillon aléatoire dans la population pour les 2 variables, la construction de cette droite impose de choisir au préalable (*i*) quelle variable sera prédite par le modèle (variable appelée 'dépendante') et (*ii*) quelle variable sera explicative (ou encore appelée variable 'indépendante').
Dans le travail mené jusqu'ici, on a cherché à comprendre la note obtenue en MathSV: la variable `note` sera donc la variable prédite, et la note obtenue en BPSVT sera utilisée comme variable indépendante. Une fois ce modèle paramétré, on va en principe pouvoir estimer pour tout individu i sa note en MathSV à partir de sa note connue en BPSVT.

```
 class(etudiants$note)
[1] "numeric"
 class(etudiants$nBPSVT)
[1] "numeric"
 lm(note~nBPSVT, data=etudiants)
Call:
lm(formula = note ~ nBPSVT, data = etudiants)
Coefficients:
(Intercept)       nBPSVT
      2.224        1.088
 model<-lm(note~nBPSVT, data=etudiants)
 anova(model)
Analysis of Variance Table
Response: note
           Df Sum Sq Mean Sq F value    Pr(>F)
nBPSVT      1 2737.7 2737.72  162.45 < 2.2e-16 ***
Residuals 218 3673.8   16.85
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
 summary(model)
Call:
lm(formula = note ~ nBPSVT, data = etudiants)
Residuals:
     Min       1Q   Median       3Q      Max
-12.3668  -2.7694   0.1248   2.6551  13.0868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.22370    0.66468   3.345 0.000967 ***
nBPSVT       1.08805    0.08537  12.746  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.105 on 218 degrees of freedom
Multiple R-squared:  0.427,        Adjusted R-squared:  0.4244
F-statistic: 162.5 on 1 and 218 DF,  p-value: < 2.2e-16
```

Examinez attentivement les commandes ci-dessus, et analysez leurs résultats: qu'apportent les commandes lm() ? anova()? summary()? .

Réponse 8:

La fonction `lm()` exécutera des calculs différents selon la nature des variables indépendantes choisies: ici il s'agit de la régression linéaire car la variable `nBPSVT` est identifiée comme quantitative et continue. De nombreuses fonctions dans R conduisent aussi à des calculs différents selon les arguments introduits.

```
r2 <- summary(model)$r.squared
r2
```

```
[1] 0.4269992
```

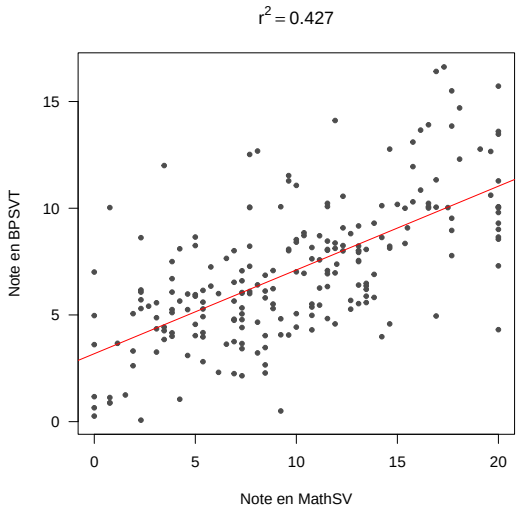
```
r2<-signif(r2,3)
r2
```

```
[1] 0.427
```

```
cor(etudiants$note, etudiants$nBPSVT)^2
```

```
[1] 0.4269992
```

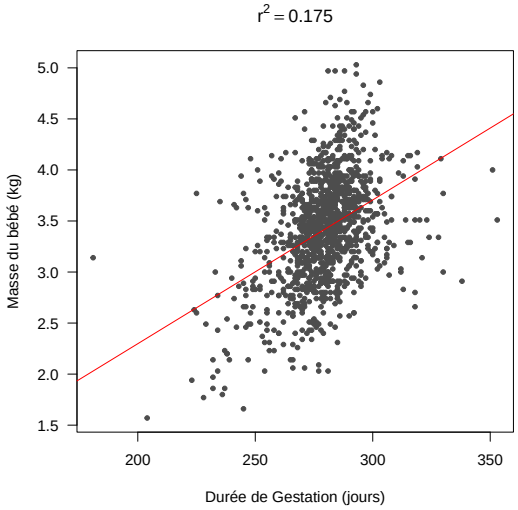
```
plot(etudiants$note, etudiants$nBPSVT, las = 1,
    main = bquote(r^2 == .(r2)),
    xlab = "Note en MathSV",
    ylab = "Note en BPSVT",
    pch = 20,
    col=grey(0.3))
#tracé de la droite de régression
abline(lm(nBPSVT~note, etudiants), col = "red")
```



Dans la figure ci-dessus, le modèle de régression linéaire simple a été ajusté aux données d'échantillonnage et la droite ajoutée dans le graphique. Que signifie r^2? Comment le calcule-t-on (2 manières possibles)? Comment interprète-t-on sa valeur?

Réponse 9:

Donnez le code R permettant de prédire la masse corporelle du bébé à la naissance en fonction de la durée de sa gestation, vous devez obtenir le résultat suivant :



4 Tests de comparaison de moyennes

4.1 Comparer 2 moyennes observées: `t.test`

Vous avez déjà vu en première année comment tester s'il existe une différence significative de la valeur moyenne d'une variable quantitative pour des individus issus de deux populations différentes. Par exemple, on aimerait tester s'il existe une différence pour la note de MathSV entre les étudiants de sexe mâle ou femelle. La fonction `t.test()` va faire rouler le test pour vous :

```
t.test(etudiants$note[etudiants$sexe == "F"], etudiants$note[etudiants$sexe == "M"])
```

```
        Welch Two Sample t-test
data:  etudiants$note[etudiants$sexe == "F"] and etudiants$note[etudiants$sexe == "M"]
t = 1.1789, df = 207.9, p-value = 0.2398
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5734648  2.2793614
sample estimates:
mean of x mean of y
10.278992  9.426044
```

La fonction `t.test()` est conservative et suppose qu'il y a hétéroscédasticité. On peut tester l'égalité des variances avec la fonction `var.test()` :

```
tapply(etudiants$note, etudiants$sexe, length)
```

```
  F   M
129  91
```

```
tapply(etudiants$note, etudiants$sexe, var)
```

```
       F        M
32.32243 24.83814
```

```
var.test(etudiants$note[etudiants$sexe == "F"], etudiants$note[etudiants$sexe == "M"])
```

```
        F test to compare two variances
data:  etudiants$note[etudiants$sexe == "F"] and etudiants$note[etudiants$sexe == "M"]
F = 1.3013, num df = 128, denom df = 90, p-value = 0.1844
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8811653 1.8970637
sample estimates:
ratio of variances
          1.301323
```

Comme on ne peut pas rejeter l'hypothèse nulle d'égalité des variances, on peut rejouer le test de comparaison de moyennes avec l'option `var.equal = TRUE` pour indiquer que l'on considère qu'il y a homoscédasticité :

```
t.test(etudiants$note[etudiants$sexe == "F"], etudiants$note[etudiants$sexe == "M"], var.equal = TRUE)
```

```
        Two Sample t-test
data:  etudiants$note[etudiants$sexe == "F"] and etudiants$note[etudiants$sexe == "M"]
t = 1.1524, df = 218, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6058499  2.3117465
sample estimates:
mean of x mean of y
10.278992  9.426044
```

Analysez les résultats de la commande ci-dessus: quelle est l'hypothèse nulle H_0? L'hypothèse alternative? Comment est obtenue la valeur de la statistique observée `t`? Quelle est votre conclusion statistique?

Réponse 10:

Donnez le code R pour tester si l'usage du tabac chez les mères influence la masse des bébés à la naissance, vous devez obtenir les résultats suivants :

```
        F test to compare two variances
data:  bebes$bwt[bebes$smoke == "FALSE"] and bebes$bwt[bebes$smoke == "TRUE"]
F = 0.90836, num df = 713, denom df = 458, p-value = 0.2525
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7681258 1.0709175
sample estimates:
ratio of variances
         0.9083552

        Two Sample t-test
data:  bebes$bwt[bebes$smoke == "FALSE"] and bebes$bwt[bebes$smoke == "TRUE"]
t = 8.724, df = 1171, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2054351 0.3246489
sample estimates:
mean of x mean of y
 3.517003  3.251961
```

Réponse 11:

Interprétez les résultats des commandes qui précèdent.

Réponse 12:

4.2 Comparer plus de 2 moyennes observées: l'Analyse de la Variance à un Facteur Contrôlé (ANOVA 1)

On souhaite maintenant savoir s'il y a une différence significative de la moyenne obtenue à l'UE MathSV entre les étudiants selon la mention qu'ils ont eue au bac.
Dans un premier temps, on peut être tenté de comparer la moyenne des notes de MathSV entre les étudiants ayant eu une mention `<12` au baccalauréat et ceux qui ont eu une mention `>14` :

```
 var.test(etudiants$note[etudiants$mBac == "<12"], etudiants$note[etudiants$mBac == ">14"])

        F test to compare two variances

data:  etudiants$note[etudiants$mBac == "<12"] and etudiants$note[etudiants$mBac == ">14"]
F = 0.96771, num df = 103, denom df = 45, p-value = 0.8707
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5724011 1.5570282
sample estimates:
ratio of variances
         0.9677106

 ###
 t.test(etudiants$note[etudiants$mBac == "<12"], etudiants$note[etudiants$mBac == ">14"], var.equal = TRUE)

        Two Sample t-test

data:  etudiants$note[etudiants$mBac == "<12"] and etudiants$note[etudiants$mBac == ">14"]
t = -10.21, df = 148, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.094609 -6.144998
sample estimates:
mean of x mean of y
 7.079327 14.699130
```

Le problème ici est que pour être complet il faudrait aussi comparer les mentions `<12` aux mentions `12-14` et les mentions `12-14` aux mentions `>14`. Pour une variable catégorielle (ou Facteur) ayant n modalités cela ferait $\frac{n(n-1)}{2}$ tests à effectuer. Ce n'est pas la bonne méthode quand on veut vérifier l'influence *globale* d'un facteur (ici la mention au baccalauréat) sur une variable quantitative (ici la note en MathSV). On utilise pour cela l'analyse de la variance à un facteur contrôlé, qui est une généralisation du test de comparaison de deux moyennes :

```
 levels(etudiants$mBac)
[1] "<12"   "12-14" ">14"
 bartlett.test(etudiants$note~etudiants$mBac)

        Bartlett test of homogeneity of variances

data:  etudiants$note by etudiants$mBac
Bartlett's K-squared = 3.6984, df = 2, p-value = 0.1574

 lm(etudiants$note~etudiants$mBac)

Call:
lm(formula = etudiants$note ~ etudiants$mBac)
Coefficients:
        (Intercept)  etudiants$mBac12-14    etudiants$mBac>14
              7.079                3.940                7.620

 anova(lm(etudiants$note~etudiants$mBac))
Analysis of Variance Table
Response: etudiants$note
               Df Sum Sq Mean Sq F value    Pr(>F)
etudiants$mBac   2 1974.4  987.22  48.281 < 2.2e-16 ***
Residuals      217 4437.1   20.45
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Que représente la commande `bartlett.test`, et quelle est votre analyse des résultats de celle-ci?

Réponse 13:

La probabilité critique (*p-value*) associée à la mention au baccalauréat est tellement faible ici que R ne peut en calculer qu'un majorant ($< 2.10^{-16}$). Nous avons mis en évidence un effet très significatif de la mention au baccalauréat sur la note au contrôle terminal de MathSV. La fonction `summary()` nous permet de quantifier cet effet :

```
eta2<-summary(lm(etudiants$note~etudiants$mBac))$r.squared
eta2
```
```
[1] 0.3079527
```

La variable `r.squared`, pour η^2, donne la fraction de la variabilité totale prise en compte par le modèle. Autrement dit, on a ici de l'ordre de 30 % de la variabilité de la note en MathSV qui est expliquée par la mention au baccalauréat. Soit encore environ 70% qui reste à expliquer...

On retrouve la fonction `lm()` déjà utilisée dans le cadre de la régression linéaire, mais ici, elle fournit un modèle ANOVA car la variable `mBac` est un facteur et reconnue comme tel.

Notez que dans le cas d'un facteur à deux modalités on retrouve exactement le même résultat, en termes de probabilité critique, qu'avec la fonction `t.test()` dans le cas où on suppose l'homoscédasticité. Par exemple, intéressons nous à l'effet du sexe des étudiants sur la note en MathSV :

```
t.test(etudiants$note[etudiants$sexe == "F"], etudiants$note[etudiants$sexe == "M"], var.equal = TRUE)
```
```
        Two Sample t-test
data:  etudiants$note[etudiants$sexe == "F"] and etudiants$note[etudiants$sexe == "M"]
t = 1.1524, df = 218, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6058499  2.3117465
sample estimates:
mean of x mean of y
10.278992  9.426044
```
```
anova(lm(etudiants$note~etudiants$sexe))
```
```
Analysis of Variance Table
Response: etudiants$note
                Df Sum Sq Mean Sq F value Pr(>F)
etudiants$sexe   1   38.8  38.820   1.328 0.2504
Residuals      218 6372.7  29.233
```

Il existe une relation d'égalité stricte entre les lois de proba $F_{1,n}$ et $(t_n)^2$.

Comme cette dernière notation est plus compacte, et similaire à celle que l'on utilise pour construire les boîtes à moustache, c'est celle que l'on utilise le plus souvent, même pour un simple test t.
Donnez le code R compact permettant de tester l'influence du tabagisme des mères sur la masse des bébés à la naissance, vous devez obtenir le résultat suivant :

```
Analysis of Variance Table
Response: bebes$bwt
              Df  Sum Sq Mean Sq F value    Pr(>F)
bebes$smoke    1  19.626 19.6265  76.108 < 2.2e-16 ***
Residuals   1171 301.973  0.2579
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Réponse 14:

Donnez le code R permettant de calculer la part de la variabilité de la masse des bébés à la naissance qui est expliquée par le tabagisme des mères, vous devez obtenir le résultat suivant :

```
[1] 0.06102762
```

Réponse 15:

Que pouvez-vous conclure (et ne pas conclure) à l'issue de ces analyses?

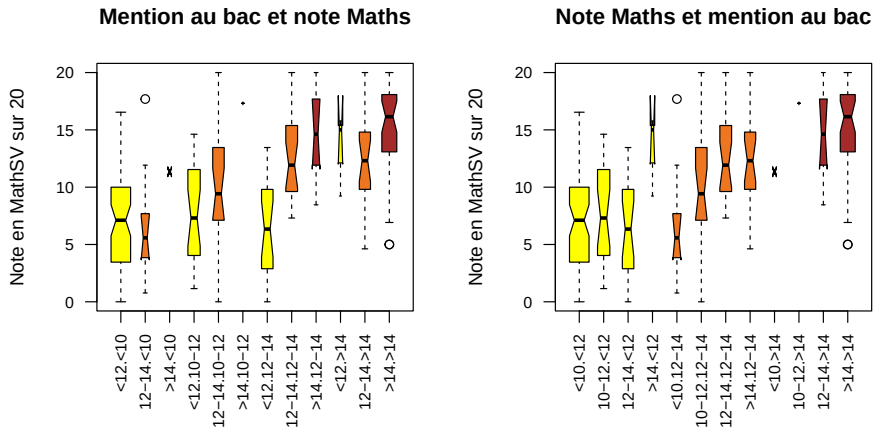
Réponse 16:

4.3 Comparer plus de 2 moyennes observées: l'Analyse de la Variance à deux Facteurs Contrôlés (ANOVA 2)

4.3.1 Deux facteurs sans interaction: des effets additifs

Nous nous attachons ici à comprendre comment la note obtenue par les étudiants en MathSV varie en moyenne, d'une part, par leur mention au baccalauréat et, d'autre part, par leur note obtenue en mathématiques au baccalauréat :

```
levels(etudiants$mBac) # 3 modalités pour la mention au bac
[1] "<12"   "12-14" ">14"
levels(etudiants$nMaths) # 4 modalités pour la note en Maths
[1] "<10"   "10-12" "12-14" ">14"
par(mfrow = c(1, 2))   #pour lire deux graphiques sur une même ligne
###
boxplot(etudiants$note~etudiants$mBac+etudiants$nMaths, main = "Mention au bac et note Maths",
        xlab = "",
        ylab = "Note en MathSV sur 20",
        las = 2, cex.axis = 0.8,  col = c("yellow", "chocolate2", "brown"),
        varwidth = TRUE, notch = TRUE)
###
boxplot(etudiants$note~etudiants$nMaths+etudiants$mBac, etudiants, main = "Note Maths et mention au bac",
        xlab = "",
        ylab = "Note en MathSV sur 20",
        las = 2, cex.axis = 0.8 , col = rep(c("yellow", "chocolate2", "brown"), each = 4),
        varwidth = TRUE, notch = TRUE)
```



Pour prendre en compte simultanément l'effet de deux facteurs on utilise le caractère d'addition `+`, par exemple `mBac+nMaths`. Mais **attention**, sauf cas très particulier, cette opération, contrairement à l'addition, n'est pas commutative dans l'écriture du modèle ANOVA, c'est-à-dire que le résultat obtenu va dépendre de l'ordre d'introduction des deux facteurs :

```
anova(lm(etudiants$note~etudiants$mBac+etudiants$nMaths ))
Analysis of Variance Table
Response: etudiants$note
                Df Sum Sq Mean Sq F value  Pr(>F)
etudiants$mBac   2 1974.4  987.22 49.4710 < 2e-16 ***
etudiants$nMaths   3  166.6   55.53  2.7825 0.04191 *
Residuals        214 4270.5   19.96
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
###
anova(lm(etudiants$note~etudiants$nMaths+etudiants$mBac))
Analysis of Variance Table
Response: etudiants$note
                Df Sum Sq Mean Sq F value    Pr(>F)
etudiants$nMaths   3 1528.1  509.38  25.526 3.714e-14 ***
etudiants$mBac     2  612.9  306.44  15.356 5.863e-07 ***
Residuals        214 4270.5   19.96
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Quand on connaît la mention obtenue au baccalauréat (première colonne de la table de contingence ci-dessous), on a déjà une idée assez précise de la note en mathématiques au baccalauréat (première ligne de la table). Ainsi, environ les trois quarts de ceux qui ont eu une mention `>14` ont eu une note en mathématiques `>14`:

```
table(etudiants$mBac,etudiants$nMaths )
       <10 10-12 12-14 >14
 <12    58    23    20   3
 12-14  10    20    21  19
 >14     2     1     9  34
34/46
[1] 0.7391304
```

Par conséquent, dans l'écriture du modèle, une fois que l'on a pris en compte la mention au baccalauréat (variable `mBac` introduite en premier), la note en mathématiques (variable `nMaths` introduite en second) n'est plus aussi informative que ça. L'examen des rapports η^2 nous confirme que l'on ne gagne pas beaucoup (moins de 3%) au niveau de la variabilité expliquée par le modèle quand on ajoute l'effet de la note en mathématiques:

```
eta1<-summary(lm(etudiants$note~etudiants$mBac ))$r.squared
eta1
[1] 0.3079527
eta2<-summary(lm(etudiants$note~etudiants$mBac+etudiants$nMaths))$r.squared
eta2
[1] 0.333934
eta2-eta1
[1] 0.02598137
```

Mais revenons à nos bébés. Nous allons nous intéresser à l'effet simultané de deux facteurs, le temps de gestation et le tabagisme de la mère, sur la masse des bébés à la naissance. Vérifions si nous avons bien affaire à deux facteurs:

```
 class(bebes$smoke)
[1] "logical"
 class(bebes$gestation)
[1] "integer"
 is.numeric(bebes$gestation)
[1] TRUE
```

La variable `gestation` est une variable quantitative, on ne peut pas l'utiliser telle quelle ici. Dans un premier temps nous allons utiliser la fonction `cut()` pour regrouper les temps de gestation en quatre classes aux effectifs équilibrés et suffisants, et créer une nouvelle variable de type facteur (`fgest`). Nous voulons donc connaître les bornes aux quartiles de la variable `gestation` avec la fonction `quantile()`:

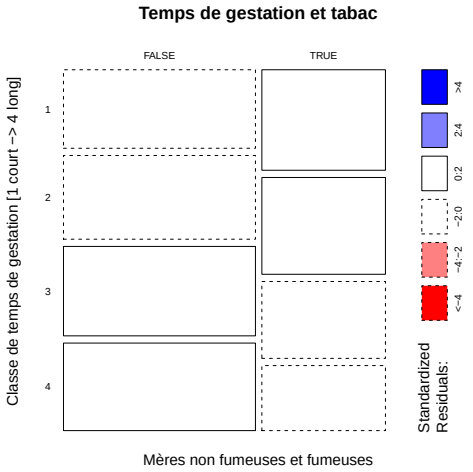
```
 quantile(bebes$gestation)
  0%  25%  50%  75% 100%
 181  272  280  288  353
 ###
 bebes$fgest <- cut(bebes$gestation, quantile(bebes$gestation), ordered_result = TRUE)
 ###
 class(bebes$fgest)
[1] "ordered" "factor"
 levels(bebes$fgest)
[1] "(181,272]" "(272,280]" "(280,288]" "(288,353]"
 ###
 levels(bebes$fgest) <- 1:4 # Numéro des quartiles (1: durée la plus courte, 4: la plus longue)
 ###
 levels(bebes$fgest)
[1] "1" "2" "3" "4"
 table(bebes$fgest)
  1   2   3   4
301 307 292 272
```

Donnez le code R permettant de tester s'il y a indépendance entre le temps de gestation et le tabagisme des mères, vous devez obtenir le résultat suivant :

```
        Pearson's Chi-squared test
data:  table(bebes$smoke, bebes$fgest)
X-squared = 13, df = 3, p-value = 0.004637
```

Réponse 17:

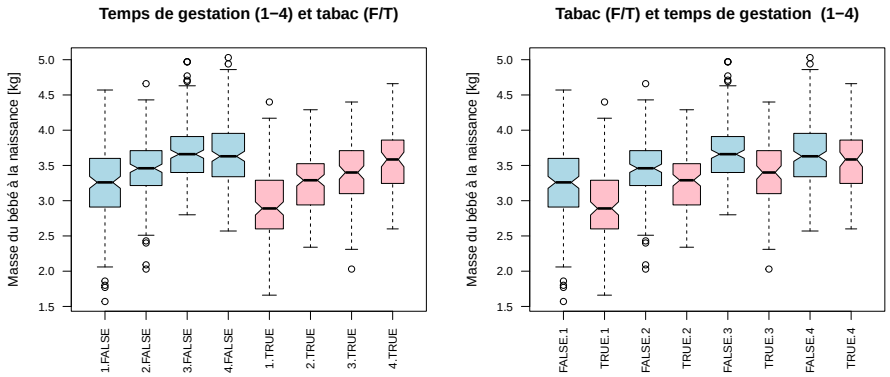
Donnez le code R produisant le graphique permettant d'interpréter ce résultat, vous devez obtenir ceci :



Réponse 18:

Il y a donc un excès de temps de gestation courts chez les mères fumeuses. Vous aviez montré à la fin de la section 4.2 qu'il y avait un effet significatif du tabagisme des mères sur la masse des bébés à la naissance, expliquant de l'ordre de 6 % de la variabilité. Se pourrait-il que ce résultat soit un effet indirect du tabagisme sur le temps de gestation ? L'effet du tabac pourrait ainsi se limiter à induire un accouchement plus précoce, donc un temps de gestation plus court laissant moins de temps au bébé pour augmenter sa masse corporelle. Mais il pourrait aussi diminuer, en moyenne, la masse corporelle des bébés à la naissance, à durée de gestation identique. La seule solution pour y voir plus clair est de démêler les effets conjoints du temps de gestation et du tabagisme sur la masse des bébés à la naissance, en commençant par représenter les données graphiquement :

```
par(mfrow = c(1, 2))
###
boxplot(bebes$bwt~bebes$fgest+bebes$smoke, main = "Temps de gestation (1-4) et tabac (F/T)",
        xlab = "",
        ylab = "Masse du bébé à la naissance [kg]",
        las = 2, cex.axis = 0.8 , col = rep(c("lightblue", "pink"), each = 4),
        varwidth = TRUE, notch = TRUE)
###
boxplot(bebes$bwt~bebes$smoke+bebes$fgest,  main = "Tabac (F/T) et temps de gestation  (1-4)",
        xlab = "",
        ylab = "Masse du bébé à la naissance [kg]",
        las = 2, cex.axis = 0.8 , col = c("lightblue", "pink"),
        varwidth = TRUE, notch = TRUE)
```



On constate sur le graphique de gauche que, quel que soit l'usage du tabac des mères, la masse des bébés à la naissance augmente avec le temps de gestation, c'était attendu, non? Sur le graphique de droite on voit que, à durée de gestation égale, la masse des bébés à la naissance diminue avec la consommation de tabac des mères. On peut voir des choses plus subtiles, on y renviendra lorsque l'on étudiera la notion d'interaction.

Si notre problématique est l'étude statistique de l'impact du tabagisme des mères sur la masse corporelle des bébés alors le temps de gestation joue le rôle de ce que l'on appelle un « facteur de nuisance ». C'est un facteur qui a un effet majeur sur la variable étudiée et dont la non prise en compte risque de fausser les résultats. La démarche consiste alors à introduire en premier ce facteur de nuisance pour voir si la variabilité résiduelle est expliquée par notre facteur d'intérêt. Donnez le code R illustrant cette démarche, vous devez obtenir le résultat suivant :

```
Analysis of Variance Table
Response: bebes$bwt
             Df  Sum Sq Mean Sq F value    Pr(>F)
bebes$fgest   3  56.736 18.9120  83.422 < 2.2e-16 ***
Residuals  1168 264.789  0.2267
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Analysis of Variance Table
Response: bebes$bwt
             Df  Sum Sq Mean Sq F value    Pr(>F)
bebes$fgest   3  56.736 18.9120  87.922 < 2.2e-16 ***
bebes$smoke   1  13.768 13.7679  64.007 2.969e-15 ***
Residuals  1167 251.021  0.2151
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Réponse 19:

Nous avons donc un effet très significatif du tabagisme sur la masse des bébés à la naissance, même quand on a corrigé pour l'effet du temps de gestation. Nous pouvons donc nous intéresser aux η^2 pour évaluer l'intensité de l'effet. Donnez le code R illustrant cette démarche, vous devez obtenir le résultat suivant :

```
[1] "Temps de gestation seul :"
[1] 0.1764588
[1] "Temps de gestation et tabagisme :"
[1] 0.2192793
```

4.3.2 Deux facteurs avec interaction: effets additifs et non additifs

On dit qu'il y a interaction entre deux facteurs sur une variable lorsque l'effet d'un facteur sur la variable est modifié par la modalité de l'autre facteur. L'interaction entre deux facteurs se note par exemple `mBac:nMaths` pour l'interaction entre la mention au baccalauréat et la note en mathématiques au baccalauréat. On dispose également de la notation compacte `mBac*nMaths` pour signifier que l'on veut l'effet des deux facteurs (additifs) et de leur interaction :

```
 anova(lm(note~mBac+nMaths+mBac:nMaths, etudiants))
Analysis of Variance Table
Response: note
             Df Sum Sq Mean Sq F value  Pr(>F)
mBac          2 1974.4  987.22 51.4487 < 2e-16 ***
nMaths        3  166.6   55.53  2.8938 0.03632 *
mBac:nMaths   6  279.3   46.55  2.4258 0.02745 *
Residuals   208 3991.2   19.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
 ###
 anova(lm(note~mBac*nMaths, etudiants))
Analysis of Variance Table
Response: note
             Df Sum Sq Mean Sq F value  Pr(>F)
mBac          2 1974.4  987.22 51.4487 < 2e-16 ***
nMaths        3  166.6   55.53  2.8938 0.03632 *
mBac:nMaths   6  279.3   46.55  2.4258 0.02745 *
Residuals   208 3991.2   19.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
 ###
 summary(lm(note~mBac*nMaths, etudiants))$r.squared
[1] 0.3774947
```

Nous avons donc une interaction significative entre les deux facteurs. Pour comprendre ce que cela signifie le plus simple est d'examiner avec soin la représentation graphique déjà effectuée des données (*cf.* page 15). Examinez, sur le graphique de droite, l'effet de la note de mathématiques par classe de mention au baccalauréat :

Mention `<12` Pour ces étudiants la note en MathSV n'est quasiment pas modulée par la note en mathématiques au baccalauréat : elle reste autour de 7.5/20, sauf éventuellement pour ceux qui ont eu une note `>14` dont la médiane monte à 15, mais cette catégorie est très marginale.

Mention `12-14` Pour ces étudiants on observe un fort effet de la note en mathématiques puisque la note en MathSV va passer de 5/20 à 12/20 au fil des modalités croissantes.

Mention `>14` Pour ces étudiants on observe peu d'effet pour les deux catégories bien documentées : la note médiane en MathSV est de l'ordre de 15/20.

Il y a interaction entre les deux deux facteurs dans le sens où la note en mathématique au baccalauréat n'a d'effet sur la note en MathSV que pour pour les étudiants ayant eu une mention intermédiaire au baccalauréat.
Donnez le code R permettant de tester l'effet de la cigarette, du temps de gestation, et de l'interaction de ces deux facteurs sur la masse des bébés à la naissance, vous devez obtenir le résultat suivant :

```
Analysis of Variance Table
Response: bebes$bwt
                        Df  Sum Sq Mean Sq F value    Pr(>F)
bebes$fgest              3  56.736 18.9120 88.3628 < 2.2e-16 ***
bebes$smoke              1  13.768 13.7679 64.3280  2.55e-15 ***
bebes$fgest:bebes$smoke  3   1.895  0.6315  2.9506   0.03175 *
Residuals             1164 249.127  0.2140
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
[1] 0.2251717
```

Réponse 20:

Comme l'interaction entre les deux facteurs est significative ici, en examinant attentivement le graphique page 18, expliquez en quoi consiste l'interaction entre ces deux facteurs.

Réponse 21:

5 De la signification du significatif

Le très très très gros piège de l'interprétation du résultat des tests d'hypothèse est de croire que quand on a un résultat statistiquement significatif cela veut forcément dire que l'on a trouvé quelque chose de pertinent. Supposez, par exemple, que dans une population d'étudiants les filles aient en moyenne une note légèrement supérieure, de 0.01 point, par rapport aux garçons. On conviendra qu'un écart de 0.01 point pour une note variant de 0 à 20 est quelque chose de parfaitement insignifiant. Simulons un tirage aléatoire des notes dans une loi normale pour un échantillon d'un million de garçons et un autre tirage dans la loi normale pour un million de filles, puis testons s'il y a une différence significative :

```
garçons <- rnorm(10^6, mean = 10,    sd = 1)
###
filles  <- rnorm(10^6, mean = 10.01, sd = 1)
###
t.test(garçons, filles)
        Welch Two Sample t-test
data:  garçons and filles
t = -7.6996, df = 2e+06, p-value = 1.365e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.013656837 -0.008114784
sample estimates:
mean of x mean of y
 10.00024  10.01112
```

❀ Renseignez-vous sur la fonction `rnorm()`, bien pratique pour construire un raisonnement statistique et accéder à des distributions d'échantillonnage d'une loi normale.

On a un résultat qui est statistiquement très significatif mais sans aucune pertinence. Avec des tailles d'échantillons aussi élevées le test de comparaison de moyenne est extrêment puissant et rejettera l'hypothèse nulle même pour des différences epsilonesques. Ne croyez pas que cela ne soit qu'un exemple d'école : avec l'avènement du « *big data* » en biologie il est très facile de se faire piéger. C'est ici que l'examen du `r.squared` est précieux :

```
df <- data.frame(note = c(garçons, filles), sexe = gl(2, 10^6))
###
head(df)
```

```
      note sexe
1 11.447261    1
2 11.163094    1
3  9.231702    1
4 10.070042    1
5 10.293556    1
6  7.841780    1
```

```
tail(df)
```

```
            note sexe
1999995  8.254773    2
1999996  9.113651    2
1999997  9.227423    2
1999998  8.209804    2
1999999  8.231362    2
2000000 10.393468    2
```

```
###
anova(lm(note~sexe, df))
```

```
Analysis of Variance Table
Response: note
             Df  Sum Sq Mean Sq F value    Pr(>F)
sexe      1e+00      59  59.250  59.284 1.365e-14 ***
Residuals 2e+06 1998868   0.999
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
summary(lm(note~sexe, df))$r.squared
```

```
[1] 2.964111e-05
```

L'effet du sexe n'explique qu'une fraction négligeable de la variabilité des notes. Notre trouvaille n'en est donc pas une. Les tests d'hypothèse sont des garde-fous pour éviter de discuter d'effets quand les données ne sont pas suffisantes, ce ne sont pas des révélateurs miraculeux de la réalité.

6 Pour aller plus loin (facultatif)

Vous avez vu lors de ces deux séances de TD de nombreux outils qui couvrent une très large fraction des besoins pour les analyses statistiques de base. Avant de vous lancer dans l'apprentissage d'outils plus sophistiqués, nous vous conseillons de commencer par vous approprier ces bases. La seule solution est d'utliser R pour résoudre une problématique donnée, l'idéal étant de traiter une problématique qui vous intéresse personnellement.

Nous vous proposons, à défaut, la problématique suivante : le jeu de données `etudiants` contient des données réelles, même s'il a été anonymisé et censuré *ad usum delphini* pour vous faciliter la tâche (pas de données manquantes,

que des étudiants ayant eu un baccalauréat série S pour faciliter l'interprétation), il n'en reste pas moins utilisable pour faire des inférences. Votre objectif est de donner des conseils utiles et argumentés à destination des étudiants de L1 qui souhaitent avoir une bonne note en MathSV. Vous devez donc hiérarchiser les facteurs explicatifs de la note en MathSV, étudier leurs interactions, et présenter ce de façon compréhensible et convaincante pour un étudiant de L1. Dans le cadre de cette problématique vous devez distinguer les facteurs utiles des facteurs inutiles. Un facteur utile est un facteur sur lequel l'étudiant de L1 peut jouer, par exemple décider de s'inscrire en FFSU de niveau 2, ou encore d'être ou non assidu aux TD. Un facteur inutile est un facteur sur lequel l'étudiant ne peut pas jouer, par exemple son sexe.