Fiche TD avec le logiciel R : tpRr2

L2 Biostatistiques-Bioinformatique
Tests statistiques avec R

A. Mary & J.R. Lobry

Nous vous proposons de mettre en œuvre sous R quelques tests d'hypothèses. Ce TP suppose que vous ayez déjà fait l'analyse exploratoire des données (`https://pbil.univ-lyon1.fr/R/pdf/tpRg2.pdf`) et que vous ayez les jeux de données `étudiants` et `bébés` dans votre environnement.

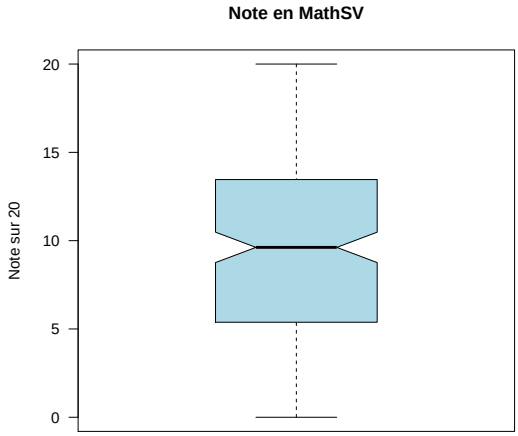
Table des matières

1 Intervalle de confiance d'une statistique

1.1 Restauration des jeux de données `étudiants` et `bébés`

Cette section vous permet de vérifier que vous arrivez à restaurer les jeux de données utilisés lors du précédent TP [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 :

```
with(étudiants, boxplot(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 % :

```
with(étudiants, median(note))
```

```
[1] 9.62
```

```
with(étudiants, quantile(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 :

```
with(étudiants, quantile(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 :

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 :

1.2 Intervalle de confiance de la médiane

Les encoches dans les boîtes à moustache donnent un intervalle de confiance à 95 % pour la valeur de la médiane. On peut récupérer les valeurs numériques correspondantes dans les valeurs de retour de la fonction `boxplot()` :

```
with(étudiants, boxplot(note)$conf)
```

```
         [,1]
[1,]  8.759289
[2,] 10.480711
```

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

```
         [,1]
[1,] 3.400014
[2,] 3.459986
```

Réponse :

Le calcul de l'intervalle de confiance de la médiane n'est pas au programme de la licence. Ce n'est pas grave, nous allons procéder par simulation. La fonction `sample(x, size)` permet de faire `size` tirages aléatoire dans l'urne `x` :

```
size <- nrow(étudiants) # le nombre total d'étudiants
with(étudiants, sample(note, size))
```

```
  [1] 12.31  1.92  3.08  5.38 11.54 11.54  2.31 12.31  5.77 13.08 13.08 11.92 13.85
 [14] 13.46  6.92  5.38 14.62  3.85  4.62 15.38 10.00  9.62 20.00 12.31  3.46 16.54
 [27]  8.46 17.69 16.15  9.62 13.08 12.31 11.54 16.92 17.31  0.00 17.69  9.62 13.46
 [40]  0.77  5.38  3.08 16.92  7.31  6.92  3.85  7.31 15.77  3.85 11.92 20.00 15.00
 [53]  8.46  6.54 14.23  6.92 14.62  9.23 11.15  8.85 11.54  7.31  3.46 10.00 10.38
 [66]  9.23 12.69  2.69 20.00 13.08  7.69 20.00 15.77 15.50 16.54  6.15  1.54  9.23
 [79]  3.85  5.77  5.00 17.69 16.92 13.46 13.08  1.92 16.15  6.15 10.77 17.69  7.31
 [92] 19.62  6.92  2.31  8.08 11.92  5.38 12.69 20.00 19.10 10.38 17.50 13.08 10.38
[105] 13.46  0.00 12.00 14.62  2.31  7.31 11.15  4.23 10.77 10.00  9.62 13.85  8.85
[118] 13.08  6.92 11.92  7.31 11.54  0.00 11.54  7.69  5.00  0.00  3.46 15.38 10.00
[131]  3.08  7.69 16.54 11.54 11.54  5.00 20.00 20.00  8.46 13.08  5.38  0.77 13.85
[144] 13.46  8.46  5.38  8.46 10.77 20.00  2.31 14.23 20.00 20.00 16.92  7.31 10.00
[157] 11.92  1.92 16.54 10.77  7.31  3.85  5.00 10.77 18.08  0.77  8.46  8.08  9.62
[170]  5.00  7.31  4.62 20.00  8.08  7.31  0.00  4.23  6.54  7.69 14.23  5.38 18.08
[183] 13.08  1.15  9.23  5.00 13.46 10.77 12.69  7.69  0.00 14.62  6.92  8.46  4.23
[196] 20.00 11.15  2.31  7.31  3.85  3.08  8.85 19.62  2.31  7.69  0.77 17.69  6.92
[209]  8.08  4.62 20.00  8.85 20.00 10.00  3.46 15.77  3.85  7.69 11.54 11.15
```

```
with(étudiants, median(sample(note, size)))
```
```
[1] 9.62
```

Nous ne sommes guère avancés puisque l'on retrouve exactement la même valeur pour la médiane. Par défaut, la fonction `sample()` fait des tirages dans l'urne sans remplacements, donc il est tout à fait logique que cela ne change rien ici, nous avons simplement permuté les valeurs. L'idée est que nous allons faire des tirages **avec** remplacement grâce à l'option `replace = TRUE` pour simuler la fluctuation d'échantillonnage :

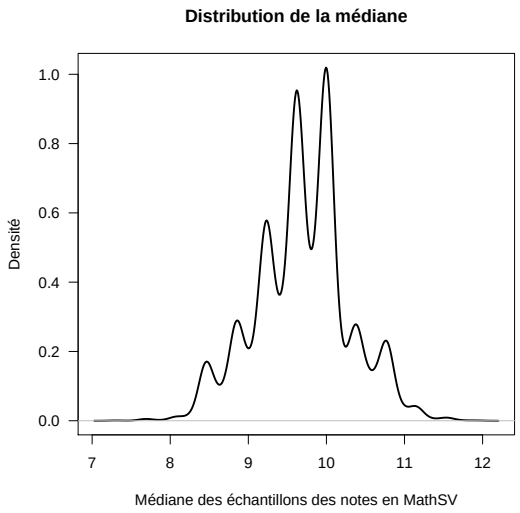
```
with(étudiants, median(sample(note, size, replace = TRUE)))
```
```
[1] 10
```

La fonction `replicate()` nous permet de répéter cette expérience, par exemple, pour effectuer 10 fois les tirages :

```
with(étudiants, replicate(10, median(sample(note, size, replace = TRUE))))
```
```
 [1] 10.000 10.000 10.190 10.000  9.425 10.000  9.425 10.000 10.770  9.230
```

Nous répétons maintenant un grand nombre de fois, 5000 fois, cette expérience et rangeons le résultat dans l'objet `x` pour représenter la distribution des valeurs de la médiane :

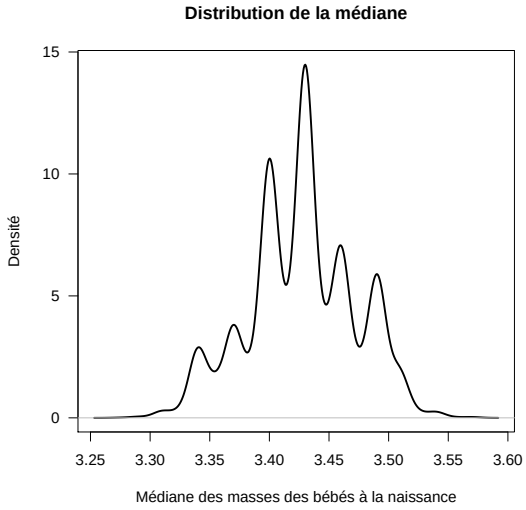
```
x <- with(étudiants, replicate(5000, median(sample(note, size, replace = TRUE))))
plot(density(x), lwd = 2, main = "Distribution de la médiane",
     xlab = "Médiane des échantillons des notes en MathSV",
     ylab = "Densité", las = 1)
```



Le caractère multimodal de la distribution vient du fait que nous avons un nombre relativement restreint de valeurs différentes possibles pour les notes (*cf.* TP précédent). Un intervalle de confiance à 95 % pour la valeur de la médiane s'obtient en calculant les quantiles à 2.5 % et 97.5 % de la distribution :

```
quantile(x, prob = c(0.025, 0.975))
```
```
2.5% 97.5%
8.46 10.77
```

Donnez le code R permettant en faisant 5000 simulations de représenter la distribution de la médiane de la masse des bébés à la naissance, vous devez obtenir un résultat proche du suivant :



Réponse :

Donnez le code R permettant à partir des 5000 simulations précédentes de donner un intervalle de confiance à 95 % pour la valeur de la médiane de la masse des bébés à la naissance :

```
2.5% 97.5%
3.34  3.51
```

Réponse :

1.3 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(étudiants) # Le nombre d'individus
mu <- mean(étudiants$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(étudiants$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(étudiants$note, conf.level = 1 - alpha)$conf.int
[1]  9.207226 10.645137
attr(,"conf.level")
[1] 0.95
```

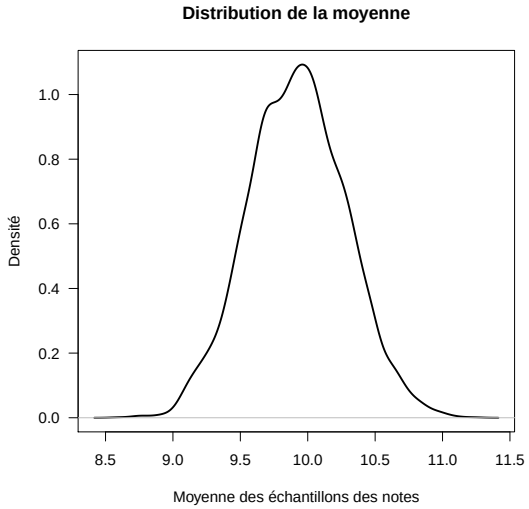
Donnez le code R permettant de calculer un intervalle de confiance à 95 % pour la moyenne de la masse des bébés à la naissance, vous devez obtenir le résultat suivant :

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

Réponse :

Tout comme nous l'avons fait pour le calcul de l'intervalle de confiance de la médiane, nous pouvons également procéder par simmulation pour le calcul de l'intervalle de confiance de la moyenne :

```
 x <- with(étudiants, replicate(5000, mean(sample(note, size, replace = TRUE))))
 plot(density(x), lwd = 2, main = "Distribution de la moyenne",
      xlab = "Moyenne des échantillons des notes",
      ylab = "Densité", las = 1)
```

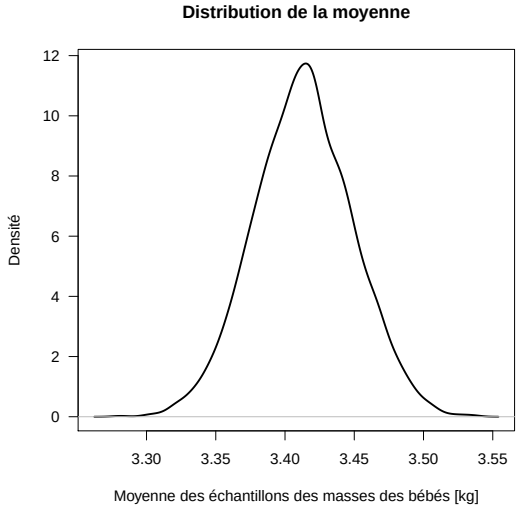


On peut calculer maintenant les quantiles à 2.5 % et 97.5 % pour la distribution de la moyenne :

```
quantile(x, prob = c(0.025, 0.975))
```
```
   2.5%    97.5%
9.21479 10.63633
```

On retrouve un résultat proche de ce qui avait été obtenu avec le calcul classique de l'intervalle de confiance pour la moyenne. Donnez le code R permettant de faire la même chose pour la masse des bébés à la naissance :

```
    2.5%    97.5%
3.344000 3.483503
```



Réponse :

1.4 Intervalle de confiance du coefficient de variation

Une statistique qui est souvent utilisée en sciences expérimentales pour apprécier la variabilité d'un jeu de données est le coefficient de variation, c_v, exprimant l'écart-type comme une fraction de la moyenne :

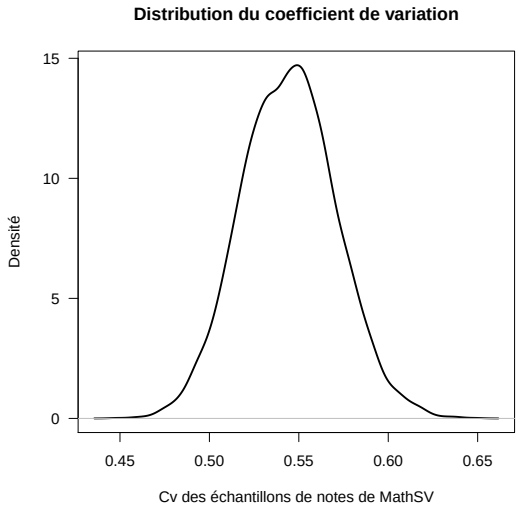
$$c_v = \frac{\sigma}{\mu}$$

Par exemple, pour calculer le coefficient de variation de la masse des bébés à la naissance :

```
with(bébés, sd(bwt)/mean(bwt))
```
```
[1] 0.1534691
```

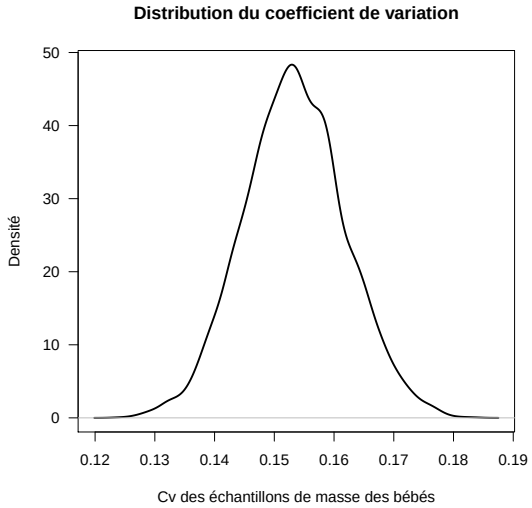
L'écart-type est donc de l'ordre de 15 % de la moyenne. Si on veut comparer la variabilité de la masse des bébés humains à la naissance avec celle, disons, de bébés éléphants, il sera plus facile de passer par le coefficient de variation. Le coefficient de variation est bien adapté quand on a des mesures strictement positives et une variabilité proportionnelle aux mesures.

```
cv <- function(z) sd(z)/mean(z) # Définition du coefficient de variation
with(étudiants, cv(note))
```
```
[1] 0.5451002
```
```
x <- with(étudiants, replicate(5000, cv(sample(note, size, replace = TRUE))))
plot(density(x), lwd = 2, main = "Distribution du coefficient de variation",
     xlab = "Cv des échantillons de notes de MathSV",
     ylab = "Densité", las = 1)
quantile(x, prob = c(0.025, 0.975))
```
```
    2.5%    97.5%
0.4935929 0.5964914
```



Pour les notes de MathSV l'écart-type représente donc de l'ordre de 54.5 % de la moyenne des notes, on a donc une plus grande variabilité relative que pour la masse des bébés à la naissance. Un intervalle de confiance à 95 % pour le coefficient de variation est [49.4 ; 59.5]. Faites la même étude pour la masse des bébés à la naissance :

```
    2.5%    97.5%
0.1371501 0.1703221
```



2 Test d'indépendance (quali-quali)

2.1 Le test d'indépendance du χ^2

Deux variables qualitatives ne sont pas indépendantes 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 `seq`, 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 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 :

```
with(étudiants, table(seq, ffsu))
```

```
   ffsu
seq 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 `seq` et `ffsu` ne sont 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 asymtotiquement la loi de probabilité du χ^2. La fonction `chisq.test()` permet de faire rouler le test. La dénomination `chisq` est une contraction anglaise pour « *chi squared* », soit « chi au carré ». Exécutez le test d'indépendance du χ^2 entre les variables `seq` et `ffsu` :

```
with(étudiants, chisq.test(table(seq, ffsu)))
```

```
        Pearson's Chi-squared test with Yates' continuity correction
data:  table(seq, 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 :

```
with(étudiants, chisq.test(table(seq, ffsu))$p.value)
```

```
[1] 0.00258064
```

C'est la probabilité, sous l'hypothèse nulle d'indépendance entre les deux variables, pour que la statistique du test soit supérieure à 9.0825. On la confronte avec le risque de première espèce choisi, ici $\alpha = 0.05$, pour prendre une décision :

```
alpha <- 0.05
valeur.p <- with(étudiants, chisq.test(table(seq, 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(parity, smoke)
X-squared = 0.075556, df = 1, p-value = 0.7834
```

Réponse :

Les variables `parity` et `smoke` sont-elles indépendantes ?

Réponse :

2.2 Effectifs théoriques inférieurs à 5 (hors programme)

Vous vous souvenez sans doute, les TD de MathSV de L1 ne sont pas si lointains, que pour que la loi de probabilité du χ^2 soit une bonne approximation de la statistique du test d'indépendance du χ^2, il est nécessaire que les effectifs théoriques soient tous supérieurs à 5. Que se passe-t-il en pratique sous R quand ce n'est pas le cas ? Testons l'indépendance entre le goût pour la physique et le fait d'être inscrit en FFSU de niveau 2 :

```
with(étudiants, chisq.test(table(ffsu, gPhys)))
```

```
        Pearson's Chi-squared test
data:  table(ffsu, gPhys)
X-squared = 4.579, df = 2, p-value = 0.1013
```

Vous devriez obtenir un message d'avis « *Chi-squared approximation may be incorrect* » vous signalant que la situation est problématique. Examinons les effectifs théoriques données par l'élément `expected` :

```
with(étudiants, chisq.test(table(ffsu, gPhys))$expected)
```

```
     gPhys
ffsu          -         +         0
  NON 25.704545 36.340909 132.95455
  OUI  3.295455  4.659091  17.04545
```

Effectivement, sous l'hypothèse d'indépendance entre les deux variables, il n'y a que peu d'étudiants attendus pour ceux qui n'aiment pas la physique et qui sont inscrits en FFSU de niveau 2. Que faire ? Nous allons procéder par simulation, dans le même esprit que ce que nous avons fait précédemment, sauf que c'est un peu plus compliqué ici puisque l'on veut tirer au hasard des tables de contingence ayant les mêmes totaux marginaux que pour nos données, mais la fonction `chisq.test()` sait très bien le faire :

```
with(étudiants, chisq.test(table(ffsu, gPhys),
      simulate.p.value = TRUE, B = 5000))
```

```
        Pearson's Chi-squared test with simulated p-value (based on 5000
        replicates)
data:  table(ffsu, gPhys)
X-squared = 4.579, df = NA, p-value = 0.09398
```

Avec un risque de première espèce de 5 %, les données ne nous permettent pas de rejeter l'hypothèse nulle de l'indépendance entre les deux variables étudiées ici.

3 Test de corrélation (quanti-quanti)

Nous avions vu lors du TP précédent (section 6.1) qu'il y avait une corrélation entre la note de MathSV et celle de BPSVT. La fonction `cor.test()` permet de tester si le coefficient de corrélation linéaire est significativement différent de zéro :

```
with(étudiants, cor.test(note, nBPSVT))
```

```
        Pearson's product-moment correlation
data:  note and 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
```

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:  bwt and 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 :

4 ANOVA1 (quanti-quali)

4.1 Analyse de la variance à un facteur

Dans l'analyse de la variance à un facteur, ANOVA1, on cherche à tester si les modalités d'une variable qualitative influent la valeur moyenne d'une variable quantitative. Vous avez déjà vu en première année comment tester s'il existe une différence significative de la valeur 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 :

```
with(étudiants, t.test(note[sexe == "F"], note[sexe == "M"]))
```

```
        Welch Two Sample t-test
data:  note[sexe == "F"] and note[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()` :

```
with(étudiants, var.test(note[sexe == "F"], note[sexe == "M"]))
```

```
        F test to compare two variances
data:  note[sexe == "F"] and note[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 :

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

```
        Two Sample t-test
data:  note[sexe == "F"] and note[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
```

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:  bwt[smoke == "NF"] and bwt[smoke == "F"]
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:  bwt[smoke == "NF"] and bwt[smoke == "F"]
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 :

Testons s'il y a une différence significative de la moyenne des notes des étudiants en MathSV entre les étudiants qui ont eu une mention `<12` au baccalauréat et ceux qui ont eu une mention `>14` :

```
with(étudiants, var.test(note[mBac == "<12"], note[mBac == ">14"]))
        F test to compare two variances
data:  note[mBac == "<12"] and note[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
```

```
with(étudiants, t.test(note[mBac == "<12"], note[mBac == ">14"], var.equal = TRUE))
        Two Sample t-test
data:  note[mBac == "<12"] and note[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 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, qui est une généralisation du test de comparaison de deux moyennes :

```
anova(lm(note~mBac, étudiants))
Analysis of Variance Table
Response: note
           Df Sum Sq Mean Sq F value    Pr(>F)
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
```

La valeur-p 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 :

```
summary(lm(note~mBac, étudiants))$r.squared
[1] 0.3079527
```

La variable `r.squared`, pour r^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.

Notez que dans le cas d'une variable à deux modalités on retrouve exactement le même résultat 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 :

```
with(étudiants, t.test(note[sexe == "F"], note[sexe == "M"], var.equal = TRUE))
        Two Sample t-test
data:  note[sexe == "F"] and note[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(note~sexe, étudiants))
Analysis of Variance Table
Response: note
           Df Sum Sq Mean Sq F value Pr(>F)
sexe        1   38.8  38.820   1.328 0.2504
Residuals 218 6372.7  29.233
```

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: bwt
            Df  Sum Sq Mean Sq F value    Pr(>F)
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 :

Donnez le code R permettant de calculer le pourcentage de la variabilité de la masse des bébés à la naissance pris en compte par le tabagisme des mères, vous devez obtenir le résultat suivant :

```
[1] 0.06102762
```

Réponse :

4.2 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 points, par rapport aux garçons. On conviendra qu'un écart de 0.01 points pour une note variant de 0 à 20 est quelque chose de parfaitement insignifiant. Simulez le tirage des notes dans une loi normale pour un échantillon d'un million de garçons et d'un million de filles, puis testez 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 = -6.9394, df = 2e+06, p-value = 3.938e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.012589761 -0.007044322
sample estimates:
mean of x mean of y
  9.99959  10.00941
```

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^2 est précieux :

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

```
      note sexe
1 11.392330    1
2  9.361899    1
3 12.282822    1
4 10.352408    1
5 11.107614    1
6  8.462940    1
```

```
tail(df)
```

```
            note sexe
1999995  9.307115    2
1999996  9.616905    2
1999997  9.853936    2
1999998 11.036264    2
1999999  7.698066    2
2000000  9.271055    2
```

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

```
[1] 2.40772e-05
```

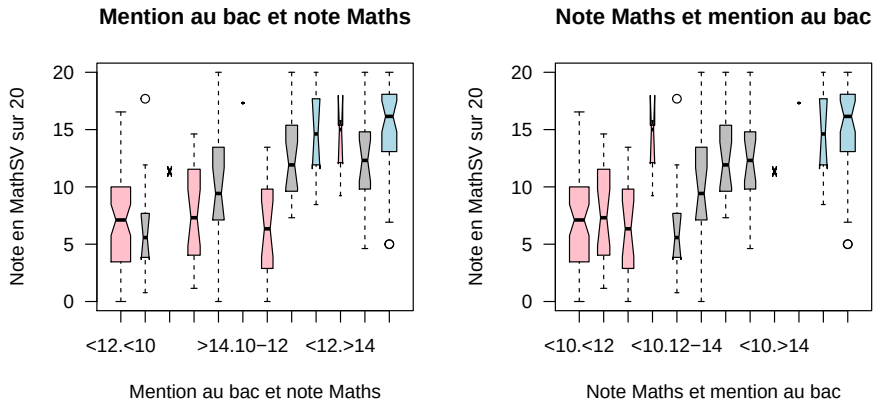
L'effet du sexe n'explique qu'une fraction négligeable de la variabilité des notes. Notre trouvaille n'en est pas une. Les tests d'hypothèse sont des garde-fous pour éviter de discuter d'effets quand les données ne sont pas suffisantes, ce ne sont pas des révélateurs miraculeux de la réalité.

5 ANOVA2 (quanti-quali-quali)

5.1 Deux facteurs sans interaction

Nous nous intéressons ici à la relation entre, d'une part, la mention au baccalauréat et de la note en mathématiques au baccalauréat, et, d'autre part, la note obtenue par les étudiants en MathSV :

```
par(mfrow = c(1, 2))
boxplot(note~mBac+nMaths, étudiants, main = "Mention au bac et note Maths",
        xlab = "Mention au bac et note Maths",
        ylab = "Note en MathSV sur 20",
        las = 1, col = c("pink", "grey", "lightblue"),
        varwidth = TRUE, notch = TRUE)
boxplot(note~nMaths+mBac, étudiants, main = "Note Maths et mention au bac",
        xlab = "Note Maths et mention au bac",
        ylab = "Note en MathSV sur 20",
        las = 1, col = rep(c("pink", "grey", "lightblue"), 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 et le résultat obtenu va dépendre de l'ordre d'introduction des facteurs :

```
 anova(lm(note~mBac+nMaths, étudiants))
Analysis of Variance Table
Response: note
           Df Sum Sq Mean Sq F value  Pr(>F)
mBac        2 1974.4  987.22 49.4710 < 2e-16 ***
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(note~nMaths+mBac, étudiants))
Analysis of Variance Table
Response: note
           Df Sum Sq Mean Sq F value    Pr(>F)
nMaths      3 1528.1  509.38  25.526 3.714e-14 ***
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 au baccalauréat, on a déjà une idée assez précise de la note en mathématiques au baccalauréat. Par exemple, les trois quarts de ceux qui ont eu une mention `>14` ont eu une note en mathématiques `>14`. Par conséquent, lorsque que l'on a pris en compte la mention au baccalauréat, la note en mathématiques n'est plus aussi informative que ça. L'examen du r^2 nous confirme que l'on ne gagne pas beaucoup au niveau de la variabilité expliquée par le modèle quand on ajoute l'effet de la note en mathématiques.

```
 summary(lm(note~mBac, étudiants))$r.squared
[1] 0.3079527
 summary(lm(note~mBac+nMaths, étudiants))$r.squared
[1] 0.333934
```

Nous allons nous intéresser à l'effet simultané du temps de gestation et du tabagisme des mères sur la masse des bébés à la naissance. 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 gestations en quatre classes d'effectifs voisins, nous voulons donc définir les bornes aux quartiles, et c'est ce que nous renvoie la fonction `quantile()` par défaut.

```
with(bébés, quantile(gestation))
```

```
  0%  25%  50%  75% 100%
 181  272  280  288  353
```

```
bébés$fgest <- with(bébés, cut(gestation, quantile(gestation), ordered_result = TRUE))
levels(bébés$fgest)
```

```
[1] "(181,272]" "(272,280]" "(280,288]" "(288,353]"
```

```
levels(bébés$fgest) <- 1:4 # Numéro des quartiles
levels(bébés$fgest)
```

```
[1] "1" "2" "3" "4"
```

```
table(bébés$fgest)
```

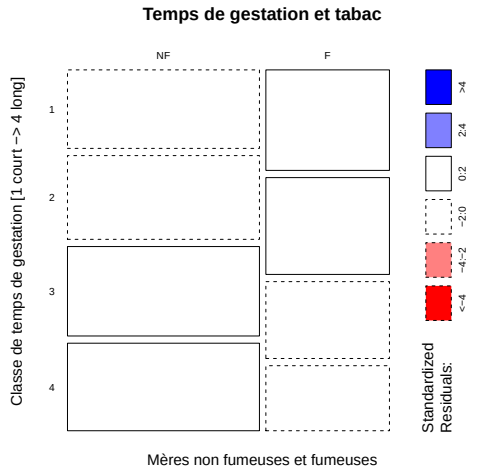
```
  1   2   3   4
301 307 292 272
```

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(smoke, fgest)
X-squared = 13, df = 3, p-value = 0.004637
```

Réponse :

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

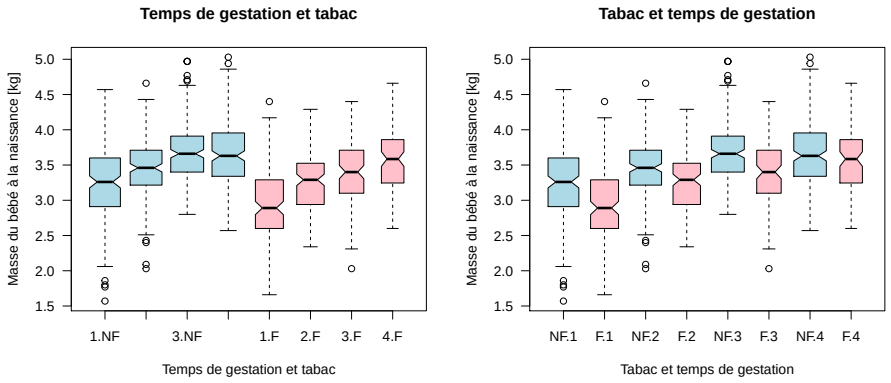


Réponse :

Il y a donc un excès de temps de gestation courts chez les mères fumeuses. Vous aviez montré à la fin de la section 4.1 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 soit un effet indirect du tabagisme sur le temps de gestation ? L'usage du tabac induirait un accouchement plus précoce, donc un temps de gestation plus court laissant moins de temps au bébé

pour augmenter sa masse corporelle. La seule solution pour y voir plus clair est d'étudier l'effet conjoint du temps de gestation et du tabagisme sur la masse des bébés à la naissance, on commence par représenter les données graphiquement :

```
par(mfrow = c(1, 2))
boxplot(bwt~fgest+smoke, bébés, main = "Temps de gestation et tabac",
        xlab = "Temps de gestation et tabac",
        ylab = "Masse du bébé à la naissance [kg]",
        las = 1, col = rep(c("lightblue", "pink"), each = 4),
        varwidth = TRUE, notch = TRUE)
boxplot(bwt~smoke+fgest, bébés, main = "Tabac et temps de gestation",
        xlab = "Tabac et temps de gestation",
        ylab = "Masse du bébé à la naissance [kg]",
        las = 1, 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. Sur le graphique de droite on voit que, quel que soit le temps de gestation, 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 de l'impact du tabagisme 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: bwt
           Df  Sum Sq Mean Sq F value    Pr(>F)
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: bwt
           Df  Sum Sq Mean Sq F value    Pr(>F)
fgest       3  56.736 18.9120  87.922 < 2.2e-16 ***
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 :

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 r^2 pour évaluer l'intensité de l'effet. Donnez le code 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
```

5.2 Deux facteurs avec interaction

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 et de leur interaction :

```
 anova(lm(note~mBac+nMaths+mBac:nMaths, étudiants))
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, étudiants))
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, étudiants))$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 16). Examinez, sur le graphique de droite, l'effet de la note de mathématiques par classe de mention au baccalauréat :

`<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 peu documentée.

`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.

`>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: bwt
              Df  Sum Sq Mean Sq F value    Pr(>F)
fgest          3  56.736 18.9120 88.3628 < 2.2e-16 ***
smoke          1  13.768 13.7679 64.3280  2.55e-15 ***
fgest: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 :

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 :

6 Pour aller plus loin (facultatif)

Vous avez vu lors de ces deux séances de TP ne 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, au lieu de suivre servilement et linéairement les exercices proposés dans les fiches de ces deux premières séances. L'idéal serait de traiter une problématique qui vous intéresse personnellement.

Nous vous proposons, à défaut, la problématique suivante : le jeu de données `étudiants` 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. Un facteur inutile est un facteur sur lequel l'étudiant ne peut pas jouer, par exemple son sexe.