L3 Pro "Biotechnologies végétales et création variétale"
TP 4 : plus de deux échantillons

D. Chessel, A-B. Dufour, J.R. Lobry, M. Royer & M. Bailly-Bechet

Automne 2012

Dans la pratique, on a parfois non pas deux échantillons à comparer, mais 3 ou plus. Une façon de faire serait de faire toutes les comparaisons possibles 2 à 2, mais cette manière de faire mène à des résultats parfois contradictoires : par exemple l'échantillon A et l'échantillon B ne sont pas significatiement différents, B et C non plus, mais A et C le sont. Ceci vient d'un problème avec la procédure statistique de comparaisons de 2 échantillons, qui n'est pas faite pour être appliquée en série.

1 Données de comptage : le test du χ^2

Si on dispose de données de comptage sur 2 échantillons, on va utiliser un test de comparaison de proportions (`prop.test`) pour savoir si les proportions observées sont les mêmes dans les deux populations. Mais on peut avoir des données avec plus de deux échantillons ; par exemple dans le jeu de données suivant on a 12 modalités mesurées, pour deux *facteurs*, le type de vaccin et la réaction au vaccin :

Réaction	*Légère*	*Moyenne*	*Ulcération*	*Abcès*
Vaccin A	13	158	8	7
Vaccin B	30	133	6	3
Vaccin C	9	129	10	6

Mesure de la réaction locale au point d'injection chez 500 patients.

Sur plus de deux échantillons, on va effectuer un test du χ^2. Les hypothèses qu'il teste sont les suivantes :

H_0 : la répartition des comptages dans les colonnes du tableau *est la même* des lignes.
H_1 : la répartition des comptages dans les colonnes du tableau *est différente en fonction des lignes.*

Dans certains contextes on préfère formuler ces hypothèses de la manière suivante :

H_0 : les facteurs *sont* indépendants.
H_1 : les facteurs *ne sont pas* indépendants.
On teste ces hypothèses avec un χ^2. Pour cela on va mettre en mémoire les valeurs :

```
vac<-matrix(c(13,158,8,7,30,133,6,3,9,129,10,6),byrow=T,ncol=4)
vac
```

```
     [,1] [,2] [,3] [,4]
[1,]   13  158    8    7
[2,]   30  133    6    3
[3,]    9  129   10    6
```

On rentre les chiffres de la matrice dans l'ordre, en indiquant à R qu'on les rentre en ligne (`byrow=T`) et qu'il y a 4 colonnes pour qu'il sache quand aller à la ligne. Pour s'y retrouver, on va nommer les choses :

```
rownames(vac)<-c("Vaccin A", "Vaccin B", "Vaccin C")
colnames(vac)<-c("Leger","Moyen","Ulcere","Abces")
vac
```

```
         Leger Moyen Ulcere Abces
Vaccin A    13   158      8     7
Vaccin B    30   133      6     3
Vaccin C     9   129     10     6
```

Puis faire le test :

```
chisq.test(vac)
```

```
        Pearson's Chi-squared test

data:  vac
X-squared = 17.6026, df = 6, p-value = 0.007306
```

Ce test contient peu d'éléments, mais on va comme avant s'intéresser surtout à la p-valeur pour conclure. Ici, la p-valeur étant assez faible, on rejette H_0 et on accepte H_1 : la réaction dépend du vaccin. Bien. Mais peut-on en savoir plus ? Pour comprendre ce qui fait, dans ces données, que la p-valeur est petite, il faut comprendre ce que fait ce test. En pratique, il calcule ce que devraient valoir les comptages si H_0 était vraie ; si c'état le cas, les comptes dans chaque case devraient être proportionnels au nombre de cas totaux dans la ligne et dans la colonne. R nous dit ce que devraient valoir ces valeurs :

```
restest<-chisq.test(vac)
restest$expected
```

```
            Leger    Moyen  Ulcere  Abces
Vaccin A 18.89062 152.5781 8.71875 5.8125
Vaccin B 17.46875 141.0938 8.06250 5.3750
Vaccin C 15.64062 126.3281 7.21875 4.8125
```

Et oui, on peut affecter le résultat d'un test dans une variable, et aller en chercher les morceaux qui nous plaisent par après. Ici, on voit que les comptages moyens que l'on aurait du obtenir si on n'avait aucun effet du choix du vaccin sur la réaction. On peut donc comparer à l'oeil les valeurs *théoriques* et les valeurs *observées*, pour voir lesquelles sont les plus différentes.

Cependant, il est parfois difficile de comparer des valeurs qui sont aussi différentes. Une façon plus classique de faire consiste à demander à R de calculer les résidus (pour ceux qui en ont déjà entendu parler, l'écart entre valeurs théoriques et valeurs observées, le tout divisé par les valeurs théoriques), qui nous diront à quel point chaque case a une importance dans le rejet final de H_0 :

```
restest$residuals
```

```
             Leger      Moyen     Ulcere      Abces
Vaccin A -1.355309  0.4389382 -0.2434169  0.4925521
Vaccin B  2.998220 -0.6813902 -0.7263720 -1.0244113
Vaccin C -1.679121  0.2377202  1.0351637  0.5413127
```

Les valeurs négatives indiquent que l'on voit moins de cas qu'attendu, les valeurs positives plus de cas. Ce qui va nous intéresser sont les plus grosses valeurs absolues : ici le 2.99 en "leger" pour le vaccin B nous indique que beaucoup plus de vaccins B qu'attendu n'ont qu'un effet léger (ce qui est biologiquement plutot intéressant), alors que les autres grosses valeurs sont les deux autres de la colonne "léger", qui sont négatives : les vaccins A et C ont tendance à provoquer de plus grosses réactions que le vaccin B.

Le test lui même ne donne qu'une réponse globale, il faut regarder les éléments dans le détail pour donner un sens biologique aux conclusions.

1.1 Latéralité manuelle

On connaît pour 119 étudiants en Activités Physiques et Sportives (APS) et 88 étudiants en Biologie la main préférentielle d'écriture :

	droite	gauche	total
APS	101	18	119
Biologie	81	7	88
total	182	25	207

Faites le test du χ^2 sur ces données. Que concluez-vous sur le lien entre latéralité des étudiants et choix de la filière universitaire ?

Ici, on n'a que 2 échantillons à 2 modalités chacun. Dans ce cas, on pourrait faire un test de comparaison de proportions pour poser la même question que le test du χ^2. Faites-le et comparez vos résultats : les deux tests donnent-ils la même réponse ?

1.2 Des iris de toutes les couleurs

Récupérez le jeu de données suivant :

```
iris <- read.table("http://pbil.univ-lyon1.fr/members/mbailly/iris_color.dat",
header = T)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species  Color
1          5.8         4.0          1.2         0.2  setosa purple
2          5.7         4.4          1.5         0.4  setosa    red
3          5.7         3.8          1.7         0.3  setosa    red
4          5.5         4.2          1.4         0.2  setosa purple
5          5.5         3.5          1.3         0.2  setosa purple
6          5.4         3.9          1.7         0.4  setosa   blue
```

C'est un jeu de données qui contient les caractéristiques des pétales et des sépales, ainsi que la couleur et l'espèce pour différents iris. Si on veut tester le lien entre l'espèce et la couleur, il faut d'abord compter tous les différents cas ; ceci est fait automatiquement avec la fonction `table` :

```
table(iris$Species, iris$Color)
```

```
            blue orange purple red
  setosa       15     12     13  10
  versicolor   12     10     17  11
  virginica     7     11     17  15
```

À vous de faire le test du χ^2 sur ces données pour conclure si l'espèce a une influence sur la couleur dans ces données.

1.3 Le retour des mères fumeuses

Récupérez le jeu de données du TP précédent, sur les données concernant les caractéristiques des mères et des bébés :

```
baby<-read.table("http://pbil.univ-lyon1.fr/R/donnees/TP_bioinfo_L2_baby.txt",header=T)
head(baby)
```

```
   bwt gestation parity age height weight smoke tension
1 3.43       284  FALSE  27  155.0  45.30 FALSE    16.1
2 3.23       282  FALSE  33  160.0  61.16 FALSE    12.7
3 3.66       279  FALSE  28  160.0  52.09  TRUE    14.8
4 3.51        NA  FALSE  36  172.5  86.07 FALSE    12.8
5 3.09       282  FALSE  23  167.5  56.62  TRUE    13.3
6 3.89       286  FALSE  25  155.0  42.13 FALSE    13.3
```

Y a-t-il un lien entre la parité (le fait que ce soit le premier accouchement ou non) et le fait que la mère fume ?

1.4 Et les conditions d'application ?

Pour utiliser le test du χ^2, ou le test de comparaison de proportions, il faut que les effectifs soient "assez grands". Une manière de vérifier si le test du χ^2 est approprié consiste à vérifier si les valeurs *théoriques* sont toutes supérieures à 5 ; si ce n'est pas le cas (ce qui arrive souvent quand on a de nombreuses modalités ou des échantillons tous petits), il est possible que le test commette des erreurs. Il faut alors appliquer un test non paramétrique, dit test de Fisher. Par exemple pour les données de vaccination au début du TP, toutes les données théoriques sauf une sont supérieures à 5 ; on peut donc probablement considérer que le χ^2 est une bonne approximation. Si, pour l'exemple, nous faisions un test de Fisher, on aurait :

```
fisher.test(vac)
```

```
        Fisher's Exact Test for Count Data

data:  vac
p-value = 0.009734
alternative hypothesis: two.sided
```

Vérifiez dans les tests que vous avez fait a) que soit les conditions d'application du χ^2 étaient bien respectées, b) si ce n'est pas le cas comment le résultat change avec un test de Fisher.

2 Comparaison de moyennes : ANOVA

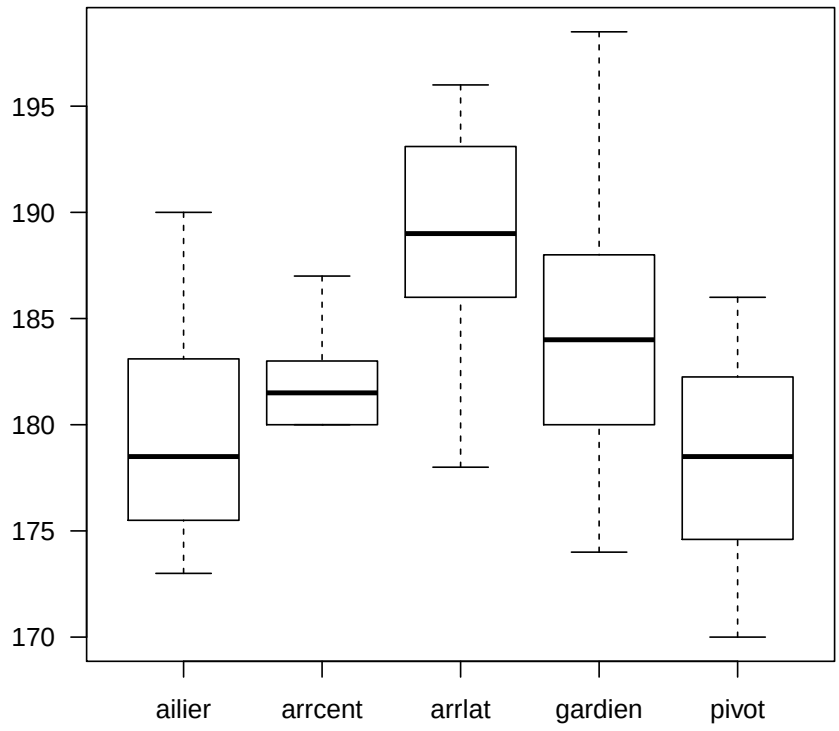
Dans le cas des données quantitatives – et non pas des comptages, souvent associés à des données qualitatives ou quantitatives discrètes – on peut se retrouver à vouloir comparer la moyenne de plusieurs échantillons à la fois. Prenons par exemple des données morphométriques dans un club de handball :

```
handball <- read.table("http://pbil.univ-lyon1.fr/R/donnees/handball.txt",h=T)
head(handball)
```

```
    TAD   TAA DBI   ENV HAU  LMS  LMI  EMP   POSTE NIVEAU
1 184.0  95.7  34 186.0 234 76.0 88.3 20.8 gardien    N1A
2 188.0  96.4  37 198.0 242 80.5 91.6 24.8  arrlat    N1A
3 190.0  97.3  36 193.0 240 78.5 92.7 22.7  ailier    N1A
4 185.0 103.6  35 186.0 229 75.5 84.0 25.2   pivot    N1A
5 176.0  93.8  35 184.0 227 74.5 82.2 21.8  ailier    N1A
6 183.5 100.0  37 193.5 238 78.2 83.5 22.6 gardien    N1A
```

Imaginons que l'on veuille comparer la taille des joueurs (`TAD`) en fonction de leur poste (`POSTE`). Comme souvent, mieux vaut commencer par un graphique :

```
boxplot(handball$TAD~handball$POSTE,las=1)
```



Visuellement, on a l'impression que les tailles sont assez variables d'un poste à l'autre, avec de petits ailiers et pivots, et de grands arrières et gardiens. Mais comment savoir si cet effet est simplement du à l'échantillonnage ou à une réelle influence de la taille des joueurs sur leur poste ? On va faire un test global, qui analyse simultanément toutes les modalités. Ce test s'appelle "Analyse de la variance", en anglais "ANalysis O VAriance" ou ANOVA, et comme ici on a un seul facteur de variabilité (le poste) on parle d'ANOVA1. Les hypothèses de ce test sont les suivantes :

H_0 : il *n'y a pas* de différence entre les moyennes des tailles aux différentes postes
H_1 : *au moins un poste* est caractérisé par une taille moyenne différente des autres
L'idée clef consiste à décomposer la variabilité totale dans les données en deux éléments :
- la variabilité *inter-groupes*, qui prend en compte les écarts entre les joueurs aux différents postes – sur le graphe les écarts entre les médianes (lignes noires) des différents groupes.
- la variabilité *intra-groupe*, qui prend en compte les écarts entre les joueurs au même poste – sur le graphe la largeur de boîtes.

Globalement, la procédure statistique va calculer ces deux éléments et les comparer :
- Si la variabilité inter-groupes est beaucoup plus grande que l'intra-groupe, la p-valeur va être faible et on pourra conclure que H_1 est vraie : il y a trop d'écart entre les groupes pour que le hasard seul explique les différences.

– Si au contraire la variabilité intra-groupe est plus grande que l'inter-groupes, la p-valeur sera élevée et on concluera que les faibles écarts entre groupes sont dus au hasard, et donc on conservera H_0.

Sous R, cela donne :

```
result<-aov(handball$TAD~handball$POSTE)
summary(result)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)
handball$POSTE  4   1104  275.88   9.175 7.82e-06 ***
Residuals      59   1774   30.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment lire ce tableau de résultats ? Ce qui nous intéresse est sur la ligne `handball$POSTE` :

Df : le nombre de degrés de liberté, soit le nombre de catégories moins une.

Sum Sq : un calcul intermédiaire.

Mean Sq : la variabilité inter-groupes (la variabilité intra-groupes est située juste au dessous, sur la ligne `Residuals`)

F value : le rapport de la variabilité inter sur la variabilité intra, ici $\frac{275.8}{30.07}$

Pr(>F) : la p-valeur associée au test.

Que peut-on conclure ici ?

2.1 Exercices

Sur le même jeu de données :

1. Vérifiez si le niveau de jeu (`NIVEAU`) a une influence sur la taille de l'empan (`EMP` ; n'oubliez pas le graphique)
2. Décomposez la hauteur bras levés (`TAU`) en 3 modalités : < 225, > 240 et le reste. Analysez ensuite l'influence de cette nouvelle variable qualitative à 3 modalités sur la taille des joueurs.

2.2 Conditions d'application

Pour appliquer une analyse de variance (ceci restera vrai au paragraphe suivant), il faut respecter comme conditions :

– L'indépendance des données dans les différents sous-groupes, qui doit être respectée comme dans les cas à deux échantillons.

– La normalité de chaque sous-échantillon, à vérifier avec un test de Shapiro ou une analyse graphique avec `qqnorm` (voir TP précédent).

– L'homoscédasticité entre les échantillons : elle se vérifie non pas avec plein de `var.test`, mais ici aussi avec un test global, le test de Bartlett. Les hypothèses de ce test, dans notre exemple précédent, sont :

H_0 : il *n'y a pas* de différence entre les variances des tailles aux différentes postes
H_1 : *au moins un poste* est caractérisé par une variance différente des autres

```
bartlett.test(handball$TAD~handball$POSTE)
```

```
        Bartlett test of homogeneity of variances

data:  handball$TAD by handball$POSTE
Bartlett's K-squared = 4.565, df = 4, p-value = 0.3349
```

Vérifiez dans les analyses précédentes que ces conditions étaient bien respectées. Si elle ne le sont pas, une alternative non paramétrique à l'ANOVA 1, qui teste les mêmes hypothèses mais qui est moins puissante, est le test de Kruskal-Wallis (qui est une généralisation du test de Wilcoxon Mann-Whitney dans les cas à plus de 2 échantillons). Par exemple, sur les données précédentes :

```
kruskal.test(handball$TAD~handball$POSTE)
```

```
        Kruskal-Wallis rank sum test

data:  handball$TAD by handball$POSTE
Kruskal-Wallis chi-squared = 23.4333, df = 4, p-value = 0.0001037
```

Rappelez-vous que les tests non paramétriques sont moins puissants que les tests paramétriques, et que vous avez intérêt à utiliser ces derniers si leurs conditions d'application sont réunies.

3 Plans expérimentaux plus complexes : ANOVA(s) 2

On peut vouloir généraliser la procédure précédente, et vouloir expliquer une variable d'intérêt par plusieurs facteurs expérimentaux. Par exemple, reprenons le jeu de données des mères et de leurs bébés :

```
head(baby)
```

```
   bwt gestation parity age height weight smoke tension
1 3.43       284  FALSE  27  155.0  45.30 FALSE    16.1
2 3.23       282  FALSE  33  160.0  61.16 FALSE    12.7
3 3.66       279  FALSE  28  160.0  52.09  TRUE    14.8
4 3.51        NA  FALSE  36  172.5  86.07 FALSE    12.8
5 3.09       282  FALSE  23  167.5  56.62  TRUE    13.3
6 3.89       286  FALSE  25  155.0  42.13 FALSE    13.3
```

Une question que l'on pourrait se poser est : le fait de fumer a-t-il une influence sur le poids du bébé à la naissance. Pour le vérifier, on fait un test de comparaison de deux moyennes, puisque l'on a deux modalités : fumer ou non.

```
 shapiro.test(baby$bwt[baby$smoke==1])
```

```
        Shapiro-Wilk normality test

data:  baby$bwt[baby$smoke == 1]
W = 0.997, p-value = 0.5196
```

```
 shapiro.test(baby$bwt[baby$smoke==0])
```

```
        Shapiro-Wilk normality test

data:  baby$bwt[baby$smoke == 0]
W = 0.9882, p-value = 1.057e-05
```

```
 wilcox.test(baby$bwt[baby$smoke==0],baby$bwt[baby$smoke==1])
```

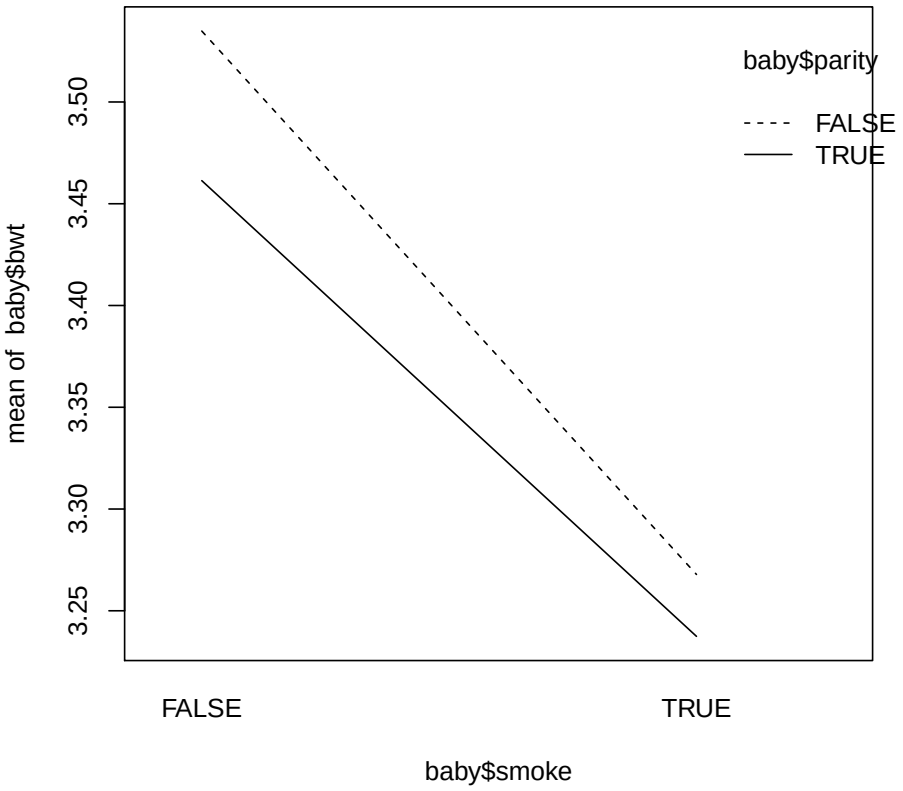
```
        Wilcoxon rank sum test with continuity correction

data:  baby$bwt[baby$smoke == 0] and baby$bwt[baby$smoke == 1]
W = 231918, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Vérifiez que vous savez pourquoi on fait ces tests là et pas d'autres. Il semble que le poids des bébés dépende du fait que la mère fume ou pas. Dépend-il aussi du fait que ce soit la première grossesse (`parity`) ? Faites la bonne série de tests pour le vérifier.

Une autre question, plus compliquée, que l'on peut se poser, est de savoir comment, *pris ensemble*, ces deux éléments influent sur le poinds des bébés à la naissance. L'effet des deux éléments est-il simplement la somme des effets de chacun, ou a-t-on des subtilités ? Si les mères arrêtent de fumer après leur première grossesse, il est certain que prendre en compte les deux effets indépendamment va rater des choses intéressantes... Dans ce cas, on va vouloir procéder à un type d'ANOVA 2 (car on a deux facteurs d'intérêt). Il en existe de nombreux types, en fonction du fait que les facteurs que l'on étudie sont *fixes* – c-à-d provoque un effet constant sur la variable d'intérêt – ou *aléatoires* – c-à-d. provoquent une augmentation de la variabilité de la variable étudiée, sans modifier sa valeur moyenne, et encore du fait que l'on dispose ou non de *répétitions* pour chaque condition expérimentale. Ici on ne va traiter qu'un exemple de modèle fixe avec répétitions : on suppose que le fait que les mères fument, ou que ce soit leur première grossesse, a un effet constant sur le poids du bébé à la naissance, et on a plusieurs bébés pour chaque cas. On commence par essayer de représenter les données de manière intelligente :

```
interaction.plot(baby$smoke, baby$parity, baby$bwt)
```



Essayez de bien comprendre ce qui est représenté sur ce graphe. Que se passe-t-il pour le poids des bébés quand les mères fument ? Quand c'est leur première grossesse ? Imaginez maintenant que l'on voie une croix sur ce graphe, par exemple que la ligne pointillée aille de en bas à gauche vers en haut à droite. Qu'est ce que cela voudrait dire ? Dans ce cas on parle d'*interaction* entre les deux facteurs : l'influence de chaque facteur n'est pas constante mais dépend de l'autre.

La procédure d'ANOVA 2 fixe avec répétitions va tester plusieurs ensembles d'hypothèses :

H_0 : le premier facteur *n'a pas* d'effet sur la variable mesurée
H_1 : le premier facteur *a* un effet sur la variable mesurée

H'_0 : le deuxième facteur *n'a pas* d'effet sur la variable mesurée
H'_1 : le deuxième facteur *a* un effet sur la variable mesurée

H''_0 : il *n'y a pas* d'interaction entre les facteurs
H''_1 : il *y a* d'interaction entre les facteurs

En pratique, cela va se faire de manière très proche à l'ANOVA1 – dans ce cas précis :

```
res<-aov(baby$bwt~baby$smoke*baby$parity)
summary(res)
```

```
                        Df Sum Sq Mean Sq F value Pr(>F)
baby$smoke               1  19.10  19.103  74.982 <2e-16 ***
baby$parity              1   0.75   0.755   2.962 0.0855 .
baby$smoke:baby$parity   1   0.10   0.103   0.405 0.5248
Residuals             1222 311.33   0.255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
10 observations deleted due to missingness
```

L'interprétation se fait ligne à ligne comme pour une ANOVA 1 : la ligne `baby$smoke` indique le résultat pour le facteur `smoke`, la ligne `baby$parity` pour la grossesse, et la ligne `baby$smoke:baby$parity` pour l'interaction entre les deux facteurs. Ici, on conclut donc que le fait de fumer affecte significativement le poids du bébé, que le fait que ce soit la première grossesse ou non n'a pas d'effet, et qu'il n'y a pas d'interaction – donc pas de changement de l'effet du fait de fumer en fonction de la grossesse, une confirmation que cet effet est bien *fixe*.

L'analyse des plans d'expérience à plus de un facteur est complexe. Avant de vous lancer dans une expérience de la sorte, vérifiez que vous avez une idée des facteurs à analyser (fixes, aléatoires, avec répétitions ou non) et posez la question à un biostatisticien avant de commencer !

Pour ceux et celles qui seraient intéressés par plus de documentation sur les questions statistiques abordées dans ce TP, d'autres exemples, des approfondissements, je conseille le site d'une collègue, avec cours et fiches de TP sous R :
`http://www3.vet-lyon.fr/ens/biostat/accueil.html`