Une introduction à l'analyse de variance et ses variantes

Jérôme Goudet
Département d'Ecologie & d'Evolution,
Biophore, Université de Lausanne,
CH-1015 Dorigny
Suisse.
E-mail :jerome.goudet@unil.ch

19 mars 2014

Table des matières

0.1 Préambule

Quelques références clefs pour ce cours :

- Whitlock and Schluter [2008] The analysis of Biological data. Un ouvrage récent et moderne, écrit par des biologistes et très bien illustré. Très vivement conseillé.
- Sokal and Rohlf [1981] Biometry. Une référence
- Zar [1984] Biostatistical analysis. Une autre référence, développant plus la notion de puissance de test.
- Dalgaard [2002] Introductory Statistics with R. Une excellente entrée en matière pour l'utilisation du logiciel R.

Chapitre 1

Introduction à l'analyse de variance

1.1 Pourquoi l'analyse de variance ? comparaison de la moyenne de plus de deux groupes

Vous vous rappelez sans doute de votre cours sur la comparaison de la moyenne de deux groupes, et du test t. Aujourd'hui, nous allons nous intéresser à la comparaison de la moyenne entre plus de deux groupes. Vous êtes peut-être surpris de voir que nous allons nous intéresser à la comparaison de moyennes (soit le test de l'hypothèse nulle $H_0 = \mu_1 = \mu_2 = \cdots = \mu_a$) quand a groupes sont à comparer, alors que le cours s'intitule introduction à l'analyse de variance.

Mais avant toutes choses, pourquoi devons-nous utiliser un nouveau test, alors qu'il serait possible de faire un test t pour chaque paire de groupes ($H_0^1 : \mu_1 = \mu_2; H_0^2 : \mu_1 = \mu_3 \cdots H_0^{a.(a-1)/2)} : \mu_{a-1} = \mu_a$) ?

Le problème, c'est justement le nombre de tests. A chaque test est associé un risque de première espèce α (l'erreur de type I), la probabilité de rejeter l'hypothèse nulle alors qu'elle est vrai. Ce seuil, en général fixé à 5% ($\alpha = 0.05$) signifie que nous prenons le risque 5 fois sur 100, ou 1 fois sur 20, de nous tromper, soit de dire que les deux moyennes sont différentes alors qu'elles sont égales. En multipliant le nombre de tests, on multiplie ce risque, puisque typiquement, si nous effectuons 20 tests comparant la moyenne de deux échantillons provenant d'une même population, nous rejetterons en espérance l'hypothèse d'égalité des moyennes une fois.

Exercice : vérifiez-le dans R en tapant les commandes suivantes :

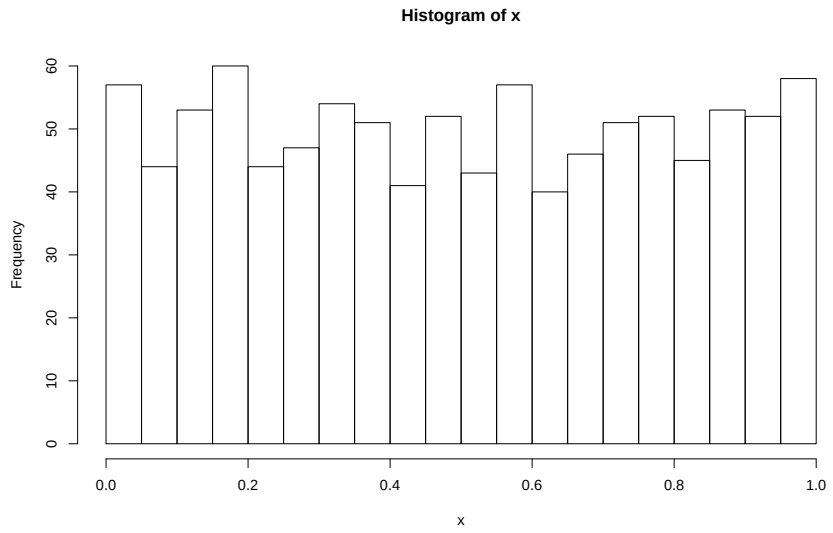
```
> set.seed(19)
> data<-matrix(rnorm(20000,18,2),nrow=20); #matrice 20 lignes et 1000 colonnes
> x<-vector(length=1000)
> for (i in 1:1000) (x[i]<-t.test(data[,i][1:10],data[,i][11:20])$p.value)
> head(x)
```

```
[1] 0.1882918 0.7764104 0.6202422 0.9012159 0.4074628 0.0726304

> sum(x<=0.05)/length(x) #proportion de proba associé au test <= à 0.05

[1] 0.057
```

La proportion de tests significatifs est donc de 0.057, ce qui signifie que 57 des 1000 tests que nous avons effectué sont significatifs, soit que les moyennes des 2 groupes ne sont pas égales.

Nous pouvons aussi représenter un histogramme des valeurs de x obtenues :

```
> hist(x,breaks=seq(0,1,0.05))
```

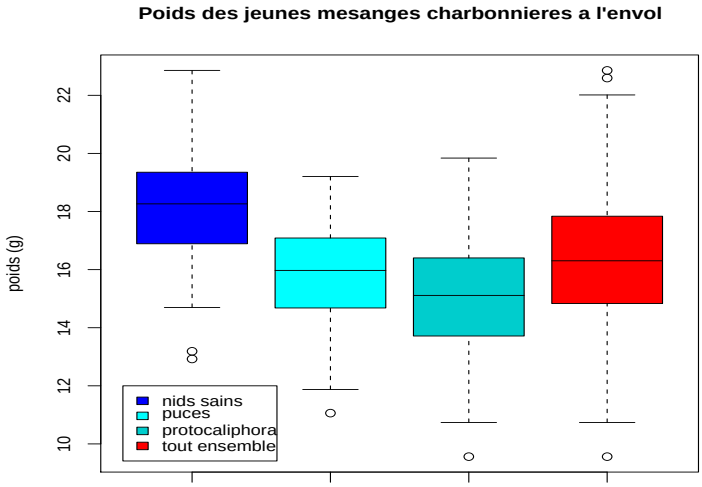


Si l'hypothèse nulle est vrai, la distribution des probabilités associée au test suit une loi uniforme $[0, 1]$. En effectuant tous ces tests, nous rejetterions donc faussement certaines de nos hypothèses. D'où la nécessité d'un autre type de test.

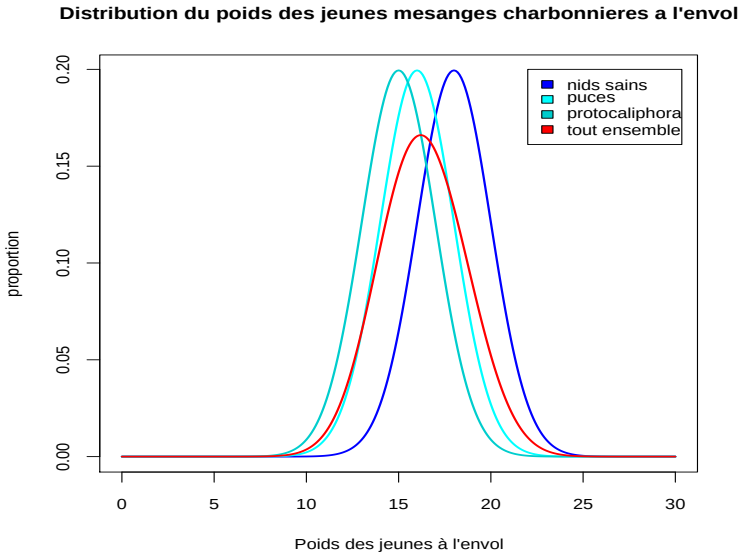
1.2 Un premier exemple

Nous allons nous intéresser au poids à l'envol des mésanges charbonnières (*Parus major*) des bois du Bourget et de Dorigny. Afin de voir quel était l'effet du parasitisme sur plusieurs aspects de la biologie de ces mésanges (dont le poids à l'envol des jeunes), 30 nids (C) ont été débarrassé de leurs parasites (en passant le matériel du nid dans un four à micro-ondes), 30 nids (T1), après avoir été passés au micro-onde, ont reçu un nombre fixe de puces (ces ecto-parasites sucent le sang des mésanges), et enfin 30 nids (T2) (aussi passés préalablement au micro-onde) ont reçu un nombre fixe de mouches (*Protocalliphora*, dont les larves sucent le sang des mésanges alors que les adultes sont détritivores).

Les nids ont ensuite été suivis journellement pendant toute la période d'incubation et d'élevage des jeunes, et le poids de ceux-ci à leur envol a été mesuré. La question que se posait les expérimentateurs, c'est bien sur, est-ce que le poids des jeunes diffère entre traitements ? On peut répondre à cette question en testant l'hypothèse nulle $H_0 : \mu_C = \mu_{T1} = \mu_{T2}$. Une première chose que nous pouvons faire est un box-plot des données :



Les 2 groupes parasités semblent en effet avoir des jeunes plus légers à l'envol que le groupe de contrôle. Notez aussi le boxplot pour toutes les données ensemble. Les moustaches sont plus longues que pour les 3 autres boxplots. La figure ci-dessous nous montre la distribution lissée des poids des jeunes à l'envol :



Nous voyons qu'elles ne sont pas complètement chevauchantes. La courbe rouge, qui représente la distribution de l'ensemble des observations, est plus large et moins haute que les autres. **Elle a une plus grande variance**. Si les 3 groupes provenaient d'une même population de poids, la distribution de l'ensemble des observations devraient être similaire à celle des différents groupes. Mais avant d'analyser cette situation, voyons un cas de figure particulier.

1.3 Une drôle d'expérience

Imaginons que nous avons pu échantillonner la totalité des jeunes mésanges charbonnières à l'envol et provenant de nids non parasités (nous aurions alors accès à la population entière, et non pas à un échantillon de cette population). Dans cette situation hypothétique et irréaliste, admettons que nous ayons trouvé que le poids moyen de jeunes à l'envol est de 18 grammes ($\mu = 18$ g) et l'écart-type de 2 grammes ($\sigma = 2$ g).

Nous allons maintenant effectuer un échantillonnage au hasard dans cette population, et prendre 7 groupes de 5 observations (par exemple, 7 biologistes qui pèsent chacun 5 jeunes). Les données sont représentées dans le tableau ci-dessous :

		g_1	g_2	g_3	g_4	g_5	g_6	g_7	*moyenne*
	n_1	16.4	19.7	19.9	17	13.8	16.4	20.1	
	n_2	20.4	23.4	15.1	16.9	17.8	16.4	16.8	
	n_3	15.4	19	21.4	19.3	18	16.9	17.7	
	n_4	13.5	16.7	19.6	17.1	18.2	15.5	20.3	
	n_5	17.7	16.2	15.9	17	19.6	20.5	20.1	
(a)	$\sum_i^n Y_{ij}$	83.4	95	91.9	87.3	87.4	85.7	95	89.39
(b)	$Y_{.j} = \frac{\sum_i^n Y_{ij}}{n}$	16.68	19	18.38	17.46	17.48	17.14	19	17.88
(c)	$\sum_i^n Y_{ij}^2$	1418	1838	1719	1529	1547	1484	1816	1621
(d)	$\sum_i^n (Y_{ij} - Y_{.j})^2$	26.71	32.98	29.83	4.25	18.93	15.13	10.64	19.78
(e)	$\frac{\sum_i^n (Y_{ij} - Y_{.j})^2}{(n-1)}$	6.677	8.245	7.457	1.063	4.732	3.783	2.66	4.95

Y_{ij} (ligne (a)) correspond à la ième observation du jième groupe, et $Y_{.j}$ (ligne (b)) correspond à la moyenne du groupe j. La moyenne de la dernière ligne (e) de cette table nous donne la moyenne des variances calculée sur les 7 groupes. Cette moyenne des variances intra-groupes est une estimation de la variance de la population, que nous appellerons la variance intra-groupes (within-groups en anglais) soit

$$s_w^2 = 4.95.$$

Notez que ce résultat n'est pas très éloigné de la valeur vrai de la variance $\sigma^2 = 4$.

Mais nous pourrions obtenir une autre estimation de la variance. En effet, nous pourrions estimer la variance de la moyenne des groupes (b), et vous avez vu précédemment (voir le polycopié du semestre d'automne, chapitre distribution d'un estimateur) comment cette dernière est reliée à la variance estimée de la population, soit $s^2(Y_{.j}) = \frac{s^2(Y_{ij})}{n}$, ou n est la taille de l'échantillon ayant servie à estimer la moyenne. La variance des moyennes s'estime comme suit :

$$s^2(Y_{.j}) = \frac{\sum_{j=1}^a (Y_{.j} - Y_{..})^2}{(a-1)} = 0.847$$

($Y_{..}$ correspond à la moyenne de toutes les observations, soit $Y_{..} = \frac{\sum_j^a \sum_i^n Y_{ij}}{an}$).

Il suffit alors de multiplier $s^2(Y_{.j})$ par le nombre d'observations dans chaque groupe pour obtenir une estimation de la variance que nous appellerons inter-groupes (between-groups en anglais), soit :

$$s_b^2 = n \times s^2(Y_{.j}) = 4.24$$

ce qui n'est pas très éloigné non plus de la valeur vrai de la variance.

Nous venons donc de voir qu'il n'existe pas une, mais deux manières différentes d'estimer la variance. Ces deux estimateurs, si les échantillons proviennent d'une même population, sont valides. Comment pourrait-on les comparer ?

Avant de répondre à cette question, notons encore qu'il existe bien évidemment une dernière manière d'estimer la variance, qui consiste tout simplement à considérer toutes les observation ensembles (et donc que les groupes 1 à 7 n'existent pas) :

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \widehat{Y_{..}})^2}{(an-1)} = 4.82$$

ce qui, à nouveau, n'est pas très éloigné de 4.

1.4 Test F pour comparer 2 variances

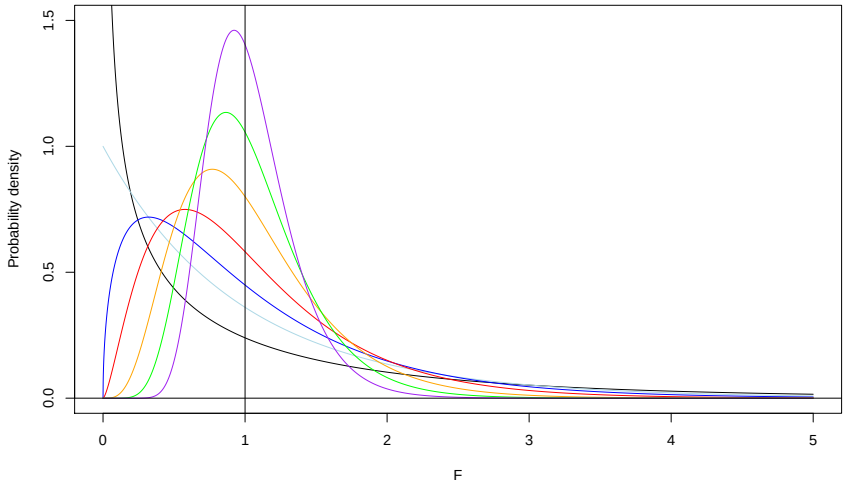
Nous nous intéressons ici à la comparaison de la variance de deux groupes d'observations. L'hypothèse nulle H_0 que nous souhaitons tester est $H_0 : \sigma_1^2 = \sigma_2^2$. Une manière de comparer ces deux variances serait de s'intéresser à leur différence, comme nous l'avons fait précédemment pour les moyennes. Il s'avère qu'il est plus pertinent de s'intéresser à leur ratio (ce qui équivaut à s'intéresser à la différence de leur logarithme).

On peut démontrer (nous ne le ferons pas, mais un exercice avec R nous permettra de le vérifier) *que le rapport de deux variances estimées* $F_s = {s_1}^2/{s_2}^2$ à partir de deux échantillons de taille n_1 et n_2 provenant de deux populations ayant la même variance σ^2 (mais pas nécessairement la même moyenne !) *suit une distribution* F à $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$ degrés de libertés. Cette distribution est derrière tous les tests d'hypothèses effectués en analyse de variance, aussi il n'est pas inutile de s'attarder un peu dessus.

Tout d'abord, si les deux variances estimées proviennent d'une même population, il est intuitif que leur rapport F_s doit être proche de 1. Ensuite, comme une variance est toujours positive, ce rapport sera aussi toujours positif. La forme exacte de la distribution va dépendre du nombre d'observations n_1 et n_2 ayant servi à estimer respectivement σ_1^2 et σ_2^2. La figure ci-dessous représente plusieurs distributions de F :

```
> x<-seq(0,5,0.001)
> plot(x,df(x,1,50),type="l",ylim=c(0,1.5),xlab="F",ylab="Probability density")
> lines(x,df(x,2,50),col="lightblue")
> lines(x,df(x,3,50),col="blue")
```

```
> lines(x,df(x,5,50),col="red")
> lines(x,df(x,10,50),col="orange")
> lines(x,df(x,20,50),col="green")
> lines(x,df(x,50,50),col="purple")
> abline(v=1)
> abline(h=0)
```



Lorsque les degrés de libertés du numérateur sont faibles, la distribution de F est en "L". Au fur et à mesure que les degrés de libertés augmentent, la distribution a son mode qui se rapproche de 1. Notons que **la forme de cette distribution ne tend jamais vers une distribution normale**, quels que soient les degrés de libertés au numérateur et dénominateur.

Nous nous intéresserons la plupart du temps à la valeur qui définit la limite des 5% (soit notre risque de première espèce α) supérieure de la distribution. Si la statistique calculée F_s est supérieure à la valeur de $F_{\alpha[\nu_2,\nu_1]}$, nous rejetterons l'hypothèse nulle d'égalité des variances. Dans le cas de figure qui nous préoccupe ci-dessus, on souhaite comparer les 2 estimations de la variance, celle obtenue à partir de la variance de la moyenne des groupes s_b^2, et celle obtenue en moyennant la variance intra groupes s_w^2. Le rapport à calculer est donc $F_s = s_b^2/s_w^2 = 4.24/4.95 = 0.856$. Nous pouvons d'ors et déjà dire que ces 2 variances ne sont pas différentes, puisque ce rapport est inférieur à 1, et que les 5% supérieurs de toutes les distributions de F correspondent à une valeur de F supérieure à 1. Les degrés de libertés pour notre exemple seront de $6 = 7 - 1$ pour s_b^2, puisque le calcul de cette variance était basé sur 7 groupes, et de $28 = (5-1) \times 7$ pour s_w^2. La probabilité associée a cette valeur de F est de

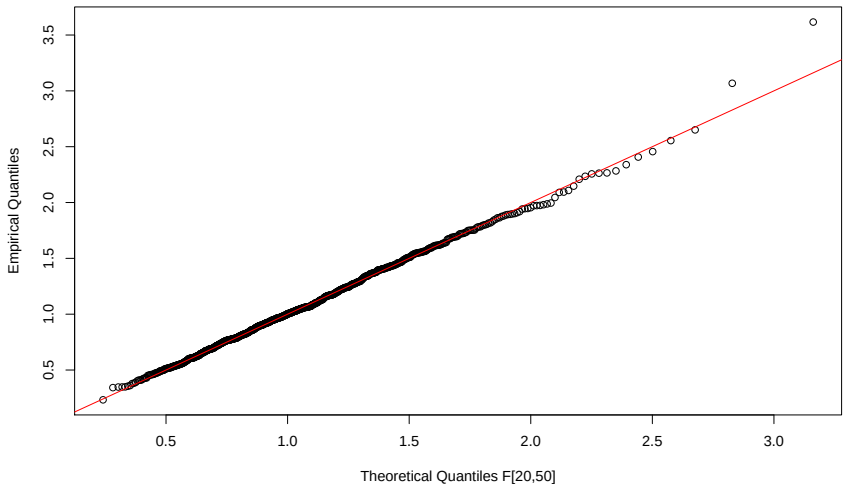
```
> pf(0.856,6,28,lower=FALSE)
```

```
[1] 0.5386607
```

Nous nous y attendions, puisque les 7 groupes proviennent d'une même population.

Notons encore ici que c'est l'égalité des variances qui est testée. Si les populations desquelles proviennent l'estimation des variances ont des moyennes différentes, le rapport des variances estimées suivra toujours une distribution de F (vous pouvez le vérifier à l'aide d'une petite simulation dans R) :

```
> set.seed(29)
> sdn<-4;sdd<-4;mn<-0;md<-100
> nrep=1000
> nnum<-21
> nden<-51
> num<-matrix(rnorm(nnum*nrep,mean=mn,sd=sdn),ncol=nrep)
> den<-matrix(rnorm(nden*nrep,mean=md,sd=sdd),ncol=nrep)
> myFs<-apply(num,2,var)/apply(den,2,var)
> qqplot(qf(ppoints(nrep),nnum-1,nden-1),myFs,
+ xlab=paste("Theoretical Quantiles F[",nnum-1,",",nden-1,"]",sep=""),
+ ylab="Empirical Quantiles")
> abline(c(0,1),col="red")
```



1.5 Différence de moyenne entre groupes

Supposons maintenant que les groupes 1, 2 et 3 correspondent en fait à des nids parasités. Soustrayons par exemple 2, 4 et 3 grammes respectivement à chacune

des observations de ces 3 groupes. Nous obtenons alors la table suivante (en rouge les chiffres modifiés par rapport aux observations initiales) :

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	*moyenne*
n_1	14.4	15.7	16.9	17	13.8	16.4	20.1	
n_2	18.4	19.4	12.1	16.9	17.8	16.4	16.8	
n_3	13.4	15	18.4	19.3	18	16.9	17.7	
n_4	11.5	12.7	17.6	17.1	18.2	15.5	20.3	
n_5	15.7	12.2	13.9	17	19.6	20.5	20.1	
$\sum_i^n Y_{ij}$	73.4	75	76.9	87.3	87.4	85.7	95	82.96
$Y_{.j} = \frac{\sum_i^n Y_{ij}}{n}$	14.68	15	15.38	17.46	17.48	17.14	19	16.59
$\sum_i^n Y_{ij}^2$	1104	1158	1213	1529	1547	1484	1816	1407
$\sum_i^n (Y_{ij} - Y_{.j})^2$	26.71	32.98	29.83	4.25	18.93	15.13	10.64	19.78
$\frac{\sum_i^n (Y_{ij} - Y_{.j})^2}{(n-1)}$	6.677	8.245	7.457	1.063	4.732	3.783	2.66	4.95

Notez que les deux dernières lignes, soit la somme des carrés des écarts à la moyenne du groupe et les variances intra-groupes, n'ont pas changées du tout. C'est normal, puisque nous avons seulement modifié la moyenne des 3 premiers groupes, mais pas leur variance. Calculons alors la variance inter groupe :

$$s_b^2 = n \times \frac{\sum_{j=1}^a (Y_{.j} - \overline{Y_{.j}})^2}{(a-1)} = 5 \times 15.30/(7-1) = 12.75$$

Cette estimation est maintenant fort éloignée de la variance de la population $\sigma^2 = 4$ (elle est plus de trois fois plus grande). Nous pouvons tester si elle diffère de la variance intra en prenant le rapport des deux estimations $F_s = 12.75/4.95 = 2.58$. La probabilité que le ratio de 2 estimateurs de la variance prenne une valeur aussi grande ou plus grande est de :

```
> pf(2.58,6,28,lower=FALSE)

[1] 0.04069785
```

et le seuil des 5% supérieur de la distribution de $F[6, 28]$ est atteint pour une valeur de F de :

```
> qf(0.95,6,28)
```

```
[1] 2.445259
```

Ces 2 estimateurs de la variance sont donc différents, l'estimateur s_b^2 est significativement plus grand que l'estimateur s_w^2. **Cette différence entre les 2 estimations de la variance provient de la différence de moyennes entre les groupes**.

C'est cette différence potentielle entre les 2 estimateurs de la variance que nous allons mettre à profit de le cadre de l'analyse de variance, ou ANOVA. Si les 2 estimateurs de la variance sont différents, s_b^2 étant supérieur à s_w^2, nous pourrons conclure qu'au moins un des groupes d'observations considérés a une moyenne qui diffère des autres.

1.6 Variances, Carrés Moyens et Somme des Carrés

Dans l'ANOVA, les variances sont couramment appelées Carrés Moyens (Mean Squares (MS) en anglais, notation qui sera utilisée par la suite), eux mêmes déduits de la Somme des Carrés (Sums of Squares (SS) en anglais, notation qui sera utilisée par la suite). Ces Sommes des Carrés correspondent à la somme des carrés des écarts à la moyenne. Nous allons tout de suite voir une propriété essentielle de ces sommes des carrés, que nous utiliserons continuellement en analyse de variance.

1.6.1 Décomposition de la Somme des Carrés

La somme des carrés globale s'écrit

$$SS_T = \sum_{j=1}^{a}\sum_{i=1}^{n}(Y_{ij} - Y_{..})^2$$

ou $Y_{..}$ signifie la moyenne des Y_{ij} sur l'ensemble des données, soit $\frac{\sum_j^a \sum_i^n Y_{ij}}{an}$. Afin d'alléger l'écriture dans ce qui suit, nous écrirons Y pour Y_{ij} et supprimerons les indices i et j dans l'écriture des sommes.

On peut ajouter et soustraire $Y_{.j}$, soit la moyenne de chaque groupe, au terme entre parenthèses sans en altérer la valeur :

$$SS_T = \sum^{a}\sum^{n}[(Y - Y_{.j}) + (Y_{.j} - Y_{..})]^2$$

On peut alors développer l'expression du carré entre crochets comme suit :

$$SS_T = \sum^{a}\sum^{n}(Y - Y_{.j})^2 + 2 \times \sum^{a}\sum^{n}(Y - Y_{.j})(Y_{.j} - Y_{..}) + \sum^{a}\sum^{n}(Y_{.j} - Y_{..})^2$$

Dans l'expression $\sum^a \sum^n (Y - Y_{.j})(Y_{.j} - Y_{..})$, le groupe $(Y_{.j} - Y_{..})$ ne varie pas selon l'indice i de la deuxième somme, on peut donc sortir cette expression de la somme sur n :

$$SS_T = \sum^a \sum^n (Y - Y_{.j})^2 + 2 \times \sum^a (Y_{.j} - Y_{..}) \sum^n (Y - Y_{.j}) + \sum^a \sum^n (Y_{.j} - Y_{..})^2$$

mais, par définition $\sum^n (Y - Y_{.j})$ est égal à 0, donc le terme central s'annule, et nous obtenons :

$$SS_T = \sum^a \sum^n (Y - Y_{.j})^2 + \sum^a \sum^n (Y_{.j} - Y_{..})^2$$

soit :

$$SS_T = SS_W + SS_B.$$

Notez bien que cette relation n'est valide que pour les sommes de carrés SS_X, pas pour les carrés moyens, et donc pas pour les variances.

Ces résultats sur la décomposition de la somme des carrés sont à la base du principe de l'analyse de variance. [1]

Finalement, en utilisant le fait que :

$$\begin{aligned}
\sum_i^n (Y_i - \bar{Y})^2 &= \sum_i Y_i^2 + \sum_i \bar{Y}^2 - 2 \sum_i Y_i \bar{Y} \\
&= \sum_i Y_i^2 + n\bar{Y}^2 - 2\bar{Y} \sum_i Y_i \\
&= \sum_i Y_i^2 + n\bar{Y}^2 - 2\bar{Y} n \bar{Y} \\
&= \sum_i Y_i^2 - n\bar{Y}^2
\end{aligned}$$

(où $\bar{Y} = \frac{\sum_i Y_i}{n}$ soit la moyenne des n observations Y_i), nous pouvons réécrire les 3 composantes que sont SS_T, SS_W et SS_B comme suit :

$$\begin{array}{lcccccc}
SS_T & = & \sum^a \sum^n Y^2 & & & - & \frac{1}{an} \left(\sum^a \sum^n Y \right)^2 \\
SS_B & = & & \frac{1}{n} \sum^a \left(\sum^n Y \right)^2 & - & \frac{1}{an} \left(\sum^a \sum^n Y \right)^2 & \\
SS_W & = & \sum^a \sum^n Y^2 & - \frac{1}{n} \sum^a \left(\sum^n Y \right)^2 & & &
\end{array}$$

Ce sont ces dernières expressions qui seront utilisées par la suite pour la colonne SS des tables d'analyses de variance.

1. Afin de bien les comprendre, (re)faites les démonstrations et vérifiez les en reprenant par exemple le jeu de données des 7 groupes de poids de jeunes mésanges à l'envol. Vous devez trouver $SS_T = 163.8817$, $SS_W = 138.468$ et $SS_B = 25.41371$

1.6.2 La table d'ANOVA

Les résultats de cette décomposition sont alors présentés sous la forme suivante, la table d'analyse de variance :

Source	*Df*	*SS*	*MS*	*F*
Entre groupes	$a-1$	$SS_B = \sum^a \frac{\left(\sum^{n_i} Y_{ij}\right)^2}{n_i} - \frac{\left(\sum^a \sum^{n_i} Y_{ij}\right)^2}{N}$	$\frac{SS_B}{a-1}$	$\frac{SS_B/(a-1)}{SS_W/(N-a)}$
Intra groupes	$\sum^a n_i - a = N - a$	$SS_W = \sum^a \sum^{n_i} Y_{ij}^2 - \sum^a \frac{\left(\sum^{n_i} Y_{ij}\right)^2}{n_i}$	$\frac{SS_W}{(N-a)}$	
Total	$\sum^a n_i - 1 = N - 1$	$SS_B + SS_W$		

ou $N = \sum^a (n_i)$. Le test consiste à comparer la variance inter-groupe $MS_B = SS_B/(a-1)$, quantité que nous avons appelée s_b^2 ci-dessus, à la variance intra-groupes $MS_W = SS_W/(N-a)$, quantité que nous avons appelée s_w^2 ci-dessus. Cette table présente les équations nécessaires pour effectuer une analyse de variance à un facteur quel que soit la taille des groupes, en particulier, il n'y a pas besoin que les groupes soient de tailles égales.

Si nous faisons ces calculs pour le jeu de données considéré ci-dessus(les 7 groupes de 5 mesures de poids de mésanges), nous obtenons la table d'ANOVA suivante :

```
> poids.mesanges<-c(
+ 16.4,19.7,19.9,17.0,13.8,16.4,20.1,
+ 20.4,23.4,15.1,16.9,17.8,16.4,16.8,
+ 15.4,19.0,21.4,19.3,18.0,16.9,17.7,
+ 13.5,16.7,19.6,17.1,18.2,15.5,20.3,
+ 17.7,16.2,15.9,17.0,19.6,20.5,20.1
+ )
> gr<-factor(rep(1:7,5))
> #anova(mod1<-lm(poids.mesanges~gr))
> ###les fonctions lm et aov sont équivalentes.
> ###Dans le cadre de l'analyse de variance seule, aov est préférable
> anova(mod1<-aov(poids.mesanges~gr))

Analysis of Variance Table
Response: poids.mesanges
          Df  Sum Sq Mean Sq F value Pr(>F)
gr         6  25.414  4.2356  0.8565 0.5383
Residuals 28 138.468  4.9453
```

Nous retrouvons bien les résultats précédents, aux erreurs d'arrondis prêt.

Chapitre 2

L'analyse de variance à 1 facteur

2.1 Introduction

Reprenons la table d'analyse de variance, vu dans le chapitre précédent :

Source	*Df*	*SS*	*MS*	*F*
Entre groupes	$a-1$	$SS_B = \sum^a \frac{\left(\sum^{n_i} Y_{ij}\right)^2}{n_i} - \frac{\left(\sum^a \sum^{n_i} Y_{ij}\right)^2}{N}$	$\frac{SS_B}{a-1}$	$\frac{SS_B/(a-1)}{SS_W/(N-a)}$
Intra groupes	$\sum^a n_i - a = N - a$	$SS_W = \sum^a \sum^{n_i} Y_{ij}^2 - \sum^a \frac{\left(\sum^{n_i} Y_{ij}\right)^2}{n_i}$	$\frac{SS_W}{(N-a)}$	
Total	$\sum^a n_i - 1 = N - 1$	$SS_B + SS_W$		

Si nous faisons ces calculs pour le premier jeu de données considéré ci-dessus, à savoir, les 7 groupes de mesures de poids de mésanges, nous avons vu que nous obtenons l'ANOVA suivante :

```
>anova(aov(poids.mesanges~gr))
Response: poids.mesanges
          Df  Sum Sq Mean Sq F value Pr(>F)
gr         6  25.414   4.236  0.8565 0.5383
Residuals 28 138.468   4.945
```

la même analyse effectuée sur le jeu de données modifié où les groupes 1, 2 et 3 ont vu 2, 4 et 3 grammes soustraits à chaque valeur donne :

```
>anova(aov(poids.mesange.2~gr))
Response: poids.mesanges.2
          Df  Sum Sq Mean Sq F value  Pr(>F)
```

```
gr         6  72.339  12.057  2.4615 0.04877 *
Residuals 28 137.148   4.898
```

une différence significative au seuil de 5%. La variance inter-groupe n'est pas la même que la variance intra-groupe, ce qui signifie qu'au moins un groupe a une moyenne qui diffère des autres.

2.2 Les modèles de l'analyse de variance : effet fixe et effet aléatoire

Jusqu'à présent, nous n'avons pas établi de liens entre l'analyse de variance et un modèle quelconque. En fait, il existe deux type d'analyse de variance, parce qu'il existe deux types de variables explicatives, ou prédicteurs.

2.2.1 Modèle à effet fixe (model I ANOVA)

Dans une ANOVA à effet fixe, soit toutes les valeurs possibles du prédicteur sont inclues dans l'analyse, soit on ne cherche à conclure que sur les valeurs observées du prédicteur. Quelques exemples devraient éclaircir ce point :

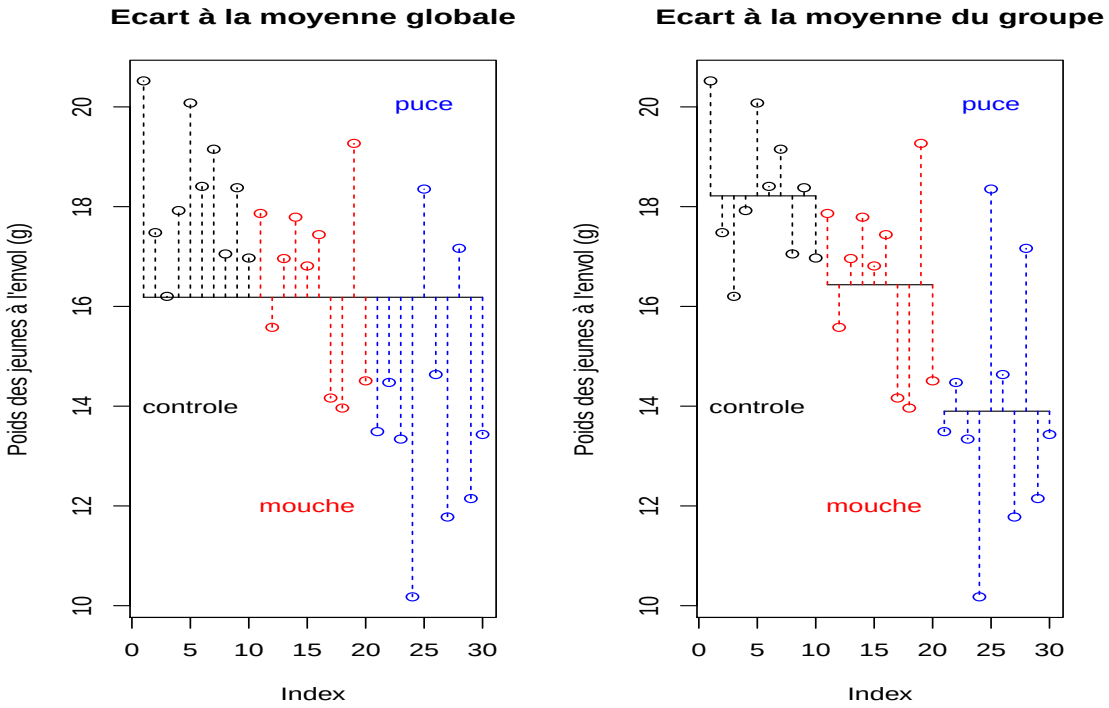
- Lorsque nous testons un placébo et deux médicaments, nous sommes dans le cadre de l'ANOVA à effet fixe.
- Lorsque nous cherchons à comparer un contrôle avec un ou plusieurs traitements (par exemple présence de différents types de parasites), nous sommes dans le cadre d'une ANOVA à effet fixe.
- Si nous souhaitons comparer différents types d'habitats que nous sommes à même de caractériser par un ou plusieurs descripteurs environnementaux (ensoleillement, température, type de végétation), nous sommes toujours dans le cadre des ANOVAs à effet fixe.

Le modèle de l'ANOVA à effet fixe s'écrit comme suit :

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2.1}$$

où Y_{ij} est la variable réponse , μ est la moyenne de la population, α_i correspond à la déviation de la moyenne du groupe i par rapport à la moyenne de la population et ϵ_{ij} représente l'écart des observations individuelles à la moyenne de leur groupe. Il est nécessaire que ces écarts proviennent d'une distribution normale de moyenne 0. Chaque observation pourra alors être décomposée en trois éléments : une moyenne générale μ, un effet α_i dû au groupe, et une déviation aléatoire ϵ_{ij}.

La figure suivante illustre cette décomposition pour le test de l'effet des parasites sur le poids des jeunes à l'envol :



Le but ultime de ces ANOVAs à effet fixe est de pouvoir conclure sur l'effet des différents traitements. Le traitement A diffère-t-il du contrôle ? Les traitements A et B sont-ils différents ? Pour ces modèles, on procède souvent à des comparaisons entre groupes, après avoir effectué l'ANOVA. Celle-ci n'est alors qu'un préalable, nécessaire mais pas suffisant, pour conclure sur les résultats de notre expérience.

2.2.2 Modèle à effet aléatoire (model II ANOVA)

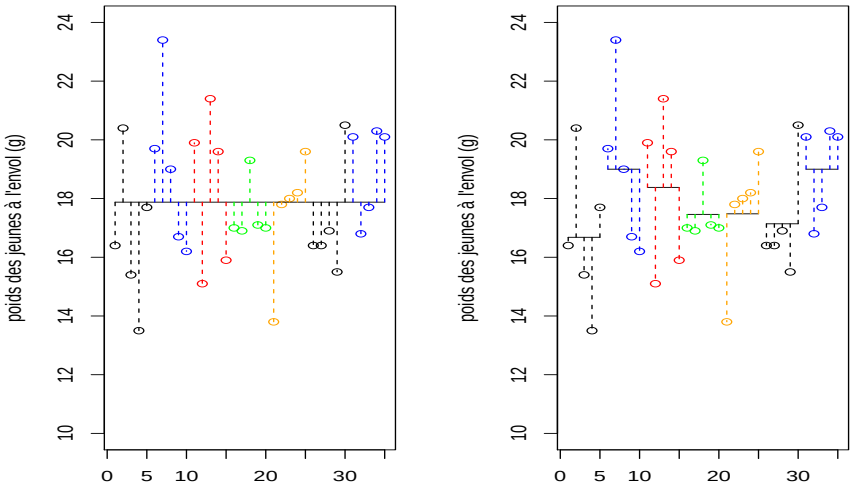
Dans un modèle à effet aléatoires par contre, les effets des groupes ne sont pas connus précisément. On suppose qu'ils ont un effet, mais on ne sait pas le quantifier. Des exemples d'effets aléatoires sont :

- l'effet de la parcelle pour une expérience agronomique
- l'effet de la forêt dans laquelle les mésanges de notre expérience sur les parasites ont été échantillonné. On ne sait pas dans quelle mesure la forêt du parc du Bourget, celle de Dorigny, ou celle d'Ecublens diffèrent, mais il est possible qu'elles diffèrent.
- Si plusieurs cages sont utilisées dans une expérience avec des souris, il est possible qu'un effet de la cage existe, mais on ne sait pas quantifier. Peut-être qu'elle était plus près de la porte de la pièce ou se fait l'expérience, peut être qu'elle était dans un endroit plus chaud.
- Si plusieurs descendants par parents sont observés, le génotype des parents sera un effet aléatoire

Dans tous ces cas de figure, nous supposons que la variable prédictive peut jouer un rôle, mais nous ne savons pas dans quelle mesure. En plus, les valeurs prises par le prédicteur ne représentent qu'un échantillon de toutes les valeurs possibles pour le prédicteur (typiquement, nous n'utiliserons pas tous les génotypes existants de souris pour notre expérience sur les souris, ni toutes les forêts de la région dans notre expérience sur les mésanges). Dans cette situation, nous sommes plus intéressés aux fluctuations dans les résultats qui peuvent être dues au facteur aléatoire. Le modèle de l'ANOVA aura donc une forme légèrement différente par rapport au modèle à effet fixe :

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \tag{2.2}$$

où, comme dans le cas de l'ANOVA à effet fixe, Y_{ij} est la variable réponse, μ est la moyenne générale de la population et ϵ_{ij} représente l'écart des observations individuelles à la moyenne de leur groupe. A_i par contre représente maintenant une variable aléatoire distribuée normalement, de moyenne 0 et de variance σ_A^2, indépendante de ϵ. Comme A_i est une variable aléatoire, il est futile de vouloir estimer la magnitude de ses effets pour l'un quelconque de ces groupes, ou la différence entre quelques groupes particuliers, mais nous pouvons estimer la variance entre groupes. Donc dans un modèle II, nous testerons pour la présence d'une variance entre groupes et nous estimerons sa magnitude s_A^2. Si nous prenons l'exemple des 7 groupes de mesures de poids à l'envol, nous obtenons la figure suivante :



Nous pouvons maintenant résumer la différence entre les modèles I et II par le tableau suivant qui donnent les carrés moyens estimés et attendus de ces 2 modèles, ainsi que le test F à effectuer :

Carrés moyens (MS)	E(MS) Modèle I	E(MS) Modèle II	F
MS_B	$\sigma^2_\epsilon + \sum^a n_i \frac{\alpha_i^2}{a-1}$	$\sigma^2_\epsilon + \sigma^2_A \frac{\left[(\sum^a n_i)^2 - \sum^a n_i^2\right]}{\sum^a n_i - a}$	$\frac{MS_B}{MS_W}$
MS_B si n_i égaux	$\sigma^2_\epsilon + n \sum^a \frac{\alpha_i^2}{a-1}$	$\sigma^2_\epsilon + n\sigma^2_A$	
MS_W	σ^2_ϵ	σ^2_ϵ	

2.2.3 Somme des carrés attendues ($E[SS]$)

L'ANOVA utilise les sommes des carrés et leurs corollaires les carrés moyens, pour estimer des quantités importantes, les composantes de la variance. La méthode présentée ci dessous est connue sous le nom de méthode des moments (methods of moments en anglais) [1]. Afin d'estimer les composantes de la variance, nous devons déterminer la valeur attendue des carrés moyens. C'est aussi grâce à l'expression de ces valeurs attendues que nous pourrons déterminer quel rapport de carrés moyens permettra de tester pour l'effet considéré.

Nous avons vu que la somme des carrés total SS_T, peut s'écrire $SS_B + SS_W$, soit la somme des carrés inter groupes et la somme des carrés intra groupes.

La valeur attendue de ces somme des carrés (expected sums of squares) s'obtient comme suit :

Pour la somme des carrés intra groupes, nous pouvons écrire :

$$E(SS_W) = \sum_{i=1}^{a} E\left[\sum_{j=1}^{n}(Y_{ij} - Y_{i.})^2\right] = a(n-1)\sigma_e^2$$

simplement parceque $\sum_j (Y_{ij} - Y_{i.})^2)/(n-1)$ est un estimateur sans biais de la variance intra groupe du groupe i, et que nous avons fait l'hypothèse que la variance au sein de ces groupes est σ_e^2. Et donc

$$E(MS_W) = \frac{E(SS_W)}{a(n-1)} = \sigma_e^2.$$

1. Cette méthode a l'avantage de donner des estimateurs non biaisés, mais a l'inconvénient de n'être valide que pour les plans expérimentaux équilibrés. Un petit déséquilibre n'est pas rhédibitoire, mais s'il est trop important, de multiples problèmes apparaissent. D'autres méthodes qui ne seront pas abordées dans le cadre de ce cours, existent. Il s'agit du Maximum de vraisemblance (Maximum Likelihood, ML en anglais, et du maximum de vraisemblance restreint, Restricted Maximum Likelihood, REML, en anglais)

Pour la somme des carrés inter-groupes, un raisonnement similaire nous amène à l'expression :

$$E(SS_B) = nE\left[\sum_{i=1}^{a}(Y_{i.} - Y_{..})^2\right] = n(a-1)\sigma^2_{Y_{i.}}$$

où $\sigma^2_{Y_{i.}}$ est la variance attendue de la moyenne des groupes, que nous supposons égale pour tous les groupes. Nous pouvons aller plus loin puisque la variance observée de la moyenne des groupes est une fonction de la variance vrai des moyennes des groupes $\mu + A_i$, mais aussi de leurs erreurs d'échantillonage, $\epsilon_{i.} = Y_{i.} - (\mu + A_i)$. Et donc, sous l'hypothèse que l'erreur est indépendante de la moyenne du groupe, nous pouvons écrire :

$$\sigma^2_{Y_{i.}} = \sigma^2(\mu + A_i) + \sigma^2(\epsilon_{i.})$$

Comme μ est une constante, le premier terme est simplement $\sigma^2(A_i)$, soit σ^2_a, alors que le second est la variance d'échantillonage attendue d'une moyenne de n observations, soit σ^2_e/n. Donc :

$$E(SS_B) = (a-1)\left[n\sigma^2_a + \sigma^2_e\right]$$

Finalement, l'espérance du carré moyen $E(MS_B)$ aura pour expression :

$$E(MS_B) = \frac{E(SS_B)}{(a-1)} = n\sigma^2_a + \sigma^2_e$$

Un raisonement similaire nous permettrait de déduire l'espérance des carrés moyens pour le modèle à effets fixes.

2.3 Conditions d'application de l'ANOVA

Comme tous les autres tests statistiques, l'ANOVA repose sur certaines hypothèses.

1. les observations doivent avoir été choisies aléatoirement
2. Les résidus sont indépendants.
3. La variance des observations (ou de leurs résidus) ne varie pas entre les traitements.
4. Les résidus (ϵ_{ij}) sont distribués normalement.

Les deux premieres sont essentielles quel que soit le type test appliqué, et font partie de la mise en place correct du plan d'expérience (voir le cours de design expérimental).

Comme les hypothèses implicites se réfèrent aux résidus, elles sont éprouvées a posteriori, quand l'ANOVA a été effectuée.

2.3.1 Choix aléatoire des observations

Toutes les ANOVAs ont comme prérequis que les observations soient choisies aléatoirement. Par exemple, dans une étude sur l'effet de 3 doses de médicaments sur le métabolisme de 5 rats (pour chaque dose), il est essentiel que ces rats aient été assignés aux différents traitements de manière aléatoire. Si nous choisissons les 5 plus jeunes, ou les 5 plus lourds pour l'un des traitements, nous ne pourrons pas conclure sur l'effet du traitement, qui serait alors confondu avec un effet âge ou poids. Notons que si les observations n'ont pas été attribuées au hasard, il est possible que cela se traduise par une variance non homogène entre les groupes.

2.3.2 Indépendance des résidus

Les résidus, soit la partie de chaque observation non-expliquée par le modèle, doivent être indépendants les uns des autres. Cela signifie que si nous rangeons les résidus dans un ordre quelconque, il ne doit pas rester de tendance (par exemple une longue série de valeurs positives, suivie d'une longue série de valeurs négatives, ou une alternance régulière de signe positif et négatif). Cette condition d'indépendance des résidus n'est pas toujours facile à remplir. Par exemple, des parcelles voisines dans l'espace auront tendances à avoir les mêmes types de sols et donc les mêmes rendements. Si les différents traitements ne sont pas alloués aléatoirement, une non-indépendance des résidus peut apparaître. La proximité des blocs peut être temporelle plutôt que spatiale. Par exemple, il est souvent impossible d'effectuer une expérience en serre en une seule fois, par manque de place. L'expérience est donc divisée en plusieurs blocs temporels. Si nous ne faisons pas attention à affecter les différents traitements aléatoirement dans ces blocs temporelles, nous risquons fort d'avoir une non-indépendance des résidus.

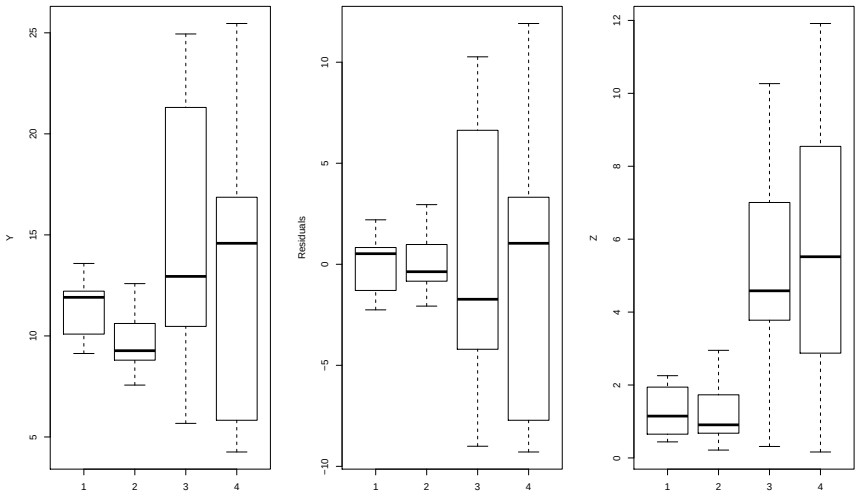
2.3.3 Homogénéité des variances=Homoscedasticité

En général, l'ANOVA est plus sensible à l'hétéroscedasticité qu'aux violations de l'hypothèse de normalité. *Un examen visuel des résidus en fonction des moyennes prédites pour chaque traitement* est le meilleur point de départ pour évaluer cette condition d'application (un boxplot par groupe fait très bien l'affaire). Si l'étendue des valeurs diffère considérablement entre traitements, il y a un problème potentiel. La règle d'usage veut que si on ne peut pas voir de différence de variance à l'oeil, alors les variances sont suffisamment homogènes pour ne pas affecter indûment l'ANOVA. Si vous percevez des différences à l'oeil, alors vous devriez effectuer un test statistique d'homogénéité des variances comme le *test de Levene*.

Test de Levene

Si les variances sont différentes entre groupes, alors l'analyse des valeurs absolues des résidus du modèle d'ANOVA devrait nous permettre de le détecter. Intuitivement, plus il y a de dispersion autour de la moyenne du groupe, et plus la moyenne des valeurs absolues des écarts à la moyenne (les résidus) sera grande. Donc un test de l'hypothèse nulle H_0 : "la moyenne des valeurs absolues des résidus $Z_{ij} = |Y_{ij} - Y_{i.}| = |\epsilon_{ij}|$ est la même dans tous les groupes" contre l'hypothèse alternative H_1 : "au moins une moyenne des valeurs absolues des résidus diffère des autres" devrait permettre de vérifier si les variances entre groupes sont homogènes ou non. La figure suivante illustre le principe de ce test

```
> set.seed(23)
> y1<-rnorm(20,10,2)
> y2<-rnorm(20,15,8)
> x<-factor(rep(1:4,each=10))
> y<-c(y1,y2)
> par(mfrow=c(1,3))
> boxplot(y~x,ylab="Y")
> resid<-y-rep(tapply(y,x,mean),each=10)
> boxplot(resid~x, ylab="Residuals")
> z<-abs(resid)
> boxplot(z~x,ylab="Z")
```



Le panneau de gauche représente les boxplots des observations pour les différents groupes, il est clair que les groupes 3 et 4 sont plus variables que les groupes 1 et 2. Ceci est amplifié sur le boxplot central des résidus en fonction des groupes. Le panneau de droite représente le boxplot des Z_{ij} par groupe. Les groupes 3 et

4 ont une médiane plus élevée que les groupes 1 et 2, comme attendu. Le test de Levene consiste alors à effectuer une ANOVA avec Z_{ij} comme variable réponse et les groupes comme variable explicative :

```
> anova(aov(z~x))

Analysis of Variance Table
Response: z
          Df Sum Sq Mean Sq F value    Pr(>F)
x          3 178.14  59.380  9.2366 0.0001148 ***
Residuals 36 231.44   6.429
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le résultat est hautement significatif, et nous sommes donc amenés à rejeter l'hypothèse nulle que les variances des différents groupes sont les mêmes. Dès lors, nous ne pouvons pas effectuer une ANOVA pour analyser les Y_{ij}, à moins de trouver une transformation qui élimine ou atténue ces différences (voir ci-dessous).

Notez encore qu'il existe une fonction `leveneTest` permettant d'effectuer directement ce test et des variations plus élaborées (utilisant la médiane plutot que la moyenne pour définir Z_{ij}, permettant d'éliminer les valeurs extrêmes, ...) dans le package `car`

Notons que d'autres tests d'homogénéité de variances existent, comme le test du $Fmax$ (qui requière le même nombre d'observations dans chaque groupe), le test de Box-Scheffé, ou le test de Fligner-Killeen (`fligner.test`). Le test de Bartlett (`bartlett.test`) est très puissant quand les données proviennent de populations normales, mais est très sensible aux écart à la normalité et n'est donc pas conseillé. Enfin, si nous n'avons que 2 groupes, le test le plus simple à effectuer est un test F du rapport de la plus grande à la plus petite des variances estimées, avec des degrés de libertés correspondant au nombre d'observations de chaque groupe moins une. Si vous ne vous souvenez pas du principe de ce test, revoyez le chapitre d'introduction à l'analyse de variance, ou les distributions F sont discutées.

2.3.4 Normalité des résidus

Comme les tests de t, les tests de F sur lesquels L'ANOVA repose sont relativement robustes aux déviations à la normalité (rappelons que la statistique d'intérêt pour l'ANOVA, comme pour les tests t, est une moyenne. Or la loi des grands nombres stipule que, quel que soit la distribution d'une série d'observations, leur moyenne tend vers une loi normale pour autant que le nombre d'observations ayant permis d'estimer ces moyennes soit suffisament grand. Il n'est donc pas très surprenant que pour un nombre suffisant d'observations par

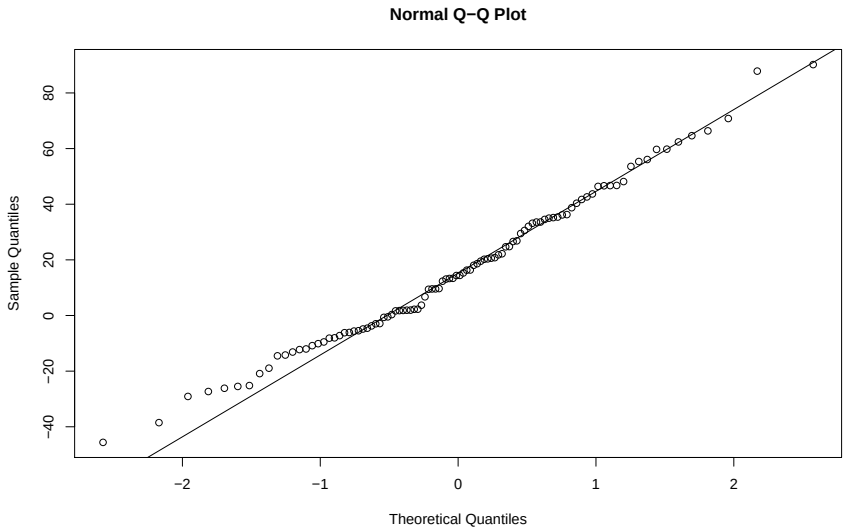
groupe, l'ANOVA soit assez peu sensible aux écarts à la normalité). Compte tenu de cette relative robustesse, un examen visuel d'un diagramme de la distribution empirique des quantiles des résidus par rapport aux quantiles d'une loi normale suffit (diagramme de probabilité, qqnorm dans R). Si le diagramme forme une droite, alors les résidus sont approximativement distribués selon la loi normale. On peut également éprouver la normalité des erreurs résiduelles par les tests de normalité (test de Kolmogorov-Smirnov (voir Chapitre 17 de Sokal and Rohlf [1981]) avec la correction de Lilliefors, ou bien avec le test de Wilk-Shapiro).

Test de Wilk-Shapiro

Sur un diagramme de probabilité normale (qqnorm dans R), une distribution normale apparaît comme une droite. La statistique de Wilks-Shapiro, `W`, mesure comment les données observées s'alignent sur cette droite (c'est en fait le carré du coefficient de corrélation entre les valeurs observées et leurs équivalents basés sur leurs fréquences cumulées relatives). Si W est près de 1, on peut alors présumer de la normalité des données. Les valeurs critiques de la statistique peuvent être retrouvées dans des tableaux spéciaux. Ce test est fastidieux à faire manuellement mais est considéré comme le meilleur pour les petits échantillons parce qu'il est très puissant. Heureusement, plusieurs logiciels statistiques, dont R, calculent cette statistique et donnent la probabilité qui lui est associée. Dans R, le test s'appelle `shapiro.test`.

L'exemple ci-dessous démontre l'utilisation de ce test dans R :

```
> set.seed(11)
> x<-rnorm(100,20,30)
> qqx<-qqnorm(x) #qqnorm des valeurs de x
> qqline(x)
> #lines(qqx$x,fitted(lm(qqx$y~qqx$x)))
> #droites des moindres carrées entre les quantiles empiriques et théoriques
```



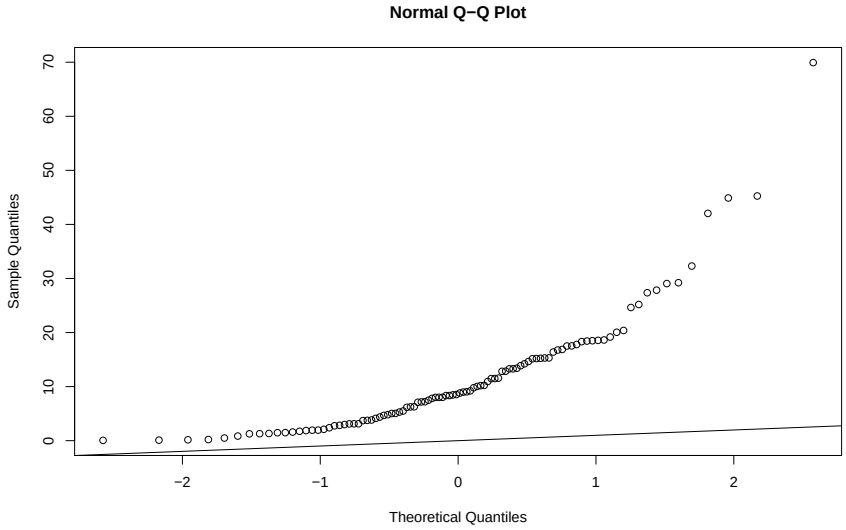
La droite des moindres carrées "colle" quasiment aux données observées, et l'interprétation graphique de cette figure nous inciterait à accepter la normalité de x. Ceci peut être vérifié par les commandes suivantes :

```
> summary(lm(qqx$y~qqx$x))$r.squared

[1] 0.9883395

> shapiro.test(x)

        Shapiro-Wilk normality test
data:  x
W = 0.9875, p-value = 0.473
```

Le texte de Shapiro-Wilk ne rejette pas (et donc accepte) la normalité comme nous pouvions nous y attendre. Notons que le statistique $W = 0.986$ du test de Shapiro-Wilk est très proche du carré du coefficient de corrélation $R^2 = 0.988$ entre valeurs prédites et observées de x. On appelle souvent ce R^2 le pourcentage de la variance expliquée par le modèle.

Si les données originales ne sont pas distribuées normalement, mais par exemple, proviennent d'une distribution exponentielle, nous obtenons les résultats suivants :

```
> set.seed(11)
> x<-rexp(100,1/10)
> qqx<-qqnorm(x)
> qqline(qqx$x)
> #lines(qqx$x,fitted(lm(qqx$y~qqx$x)))
```



Ici, il est clair que l'ajustement entre la droite des moindres carrées et les valeurs observées est mauvais. Nous le vérifions aisément avec les commandes suivantes :

```
> shapiro.test(x)

        Shapiro-Wilk normality test
data:  x
W = 0.8023, p-value = 2.921e-10
```

Le test de Shapiro-Wilk rejette très fortement la normalité des données, comme nous pouvions nous y attendre.

2.4 Equivalent non-paramétrique de l'ANOVA

Si l'une des conditions d'applications de l'analyse de variance n'est pas remplie, une solution consiste à utiliser un équivalent non paramétrique. Cet équivalent non paramétrique est, pour l'ANOVA à un facteur, le test de Kruskal-Wallis.

2.4.1 Test de Kruskal-Wallis

La statistique du test, H, est obtenue en calculant :

$$H = \frac{12}{\sum^a n_i(\sum^a n_i + 1)} \sum^a \frac{(\sum^{n_i} R)_i^2}{n_i} - 3(\sum^a n_i + 1)$$

ou a est le nombre de groupes, n_i est l'effectif du groupe i, et $(\sum^{n_i} R)_i$ est la somme des rangs pour les observations dans le groupe i.

Lorsqu'il y a moins de 6 traitements et que le nombre d'observations pour chaque traitement est faible, la statistique H doit être comparée aux valeurs critiques dans des tableaux pour cette statistique. Si les effectifs sont grands, ou si il y a de nombreux traitements qui sont comparés, alors la statistique H tend vers χ^2 avec $a-1$ degrés de liberté. Dans R, le test s'appelle `kruskal.test`.

Lorsque seuls deux groupes sont présents, le test de Wilcoxon, aussi appelé test de Mann-Whitney, est souvent utilisé (voir le cours du semestre d'automne). Ce dernier test possède aussi une variante pour des données appariées (il s'agit alors d'un équivalent non paramétrique du test t apparié). Le test de Wilcoxon est présent dans R, sous le nom `wilcox.test`.

2.4.2 Test de permutations pour l'ANOVA à 1 facteur

Si l'une des deux conditions d'application de l'ANOVA liées à la distribution de la statistique de test (soit l'homogénéité des variances et la normalité) n'est pas vérifiée, on peut utiliser l'ordinateur et sa puissance de calcul pour générer la distribution attendue de la statistique F sous l'hypothèse nulle, et étant donné les données de notre expérience. Un tel test est appelé test de permutations, car nous allons ré-échantillonner les données sans remise (voir le cours de design expérimental).

Ce test permet de s'affranchir des hypothèses de normalité des résidus et d'homogénéité des variances, sans perte de puissance. La probabilité associée au test sera estimée comme le nombre de permutations qui donnent une valeur de F au moins aussi grande que la valeur observée, cette dernière étant incluse dans la distribution de F générée par permutations. Si nous effectuons $n-1$ permutations, l'expression de p sera :

$$p = \frac{1 + \sum_i^{(n-1)} (F^* \geq F.obs)}{n}$$

où F^* représente la valeur de F obtenue quand les observations sont assignées aléatoirement à l'un des groupes et *F.obs* au F observé sur le jeu de données réel. Une valeur communément admise pour le nombre de permutations est 1000, permettant d'obtenir des probabilités associées au test de l'ordre de 10^{-3}.

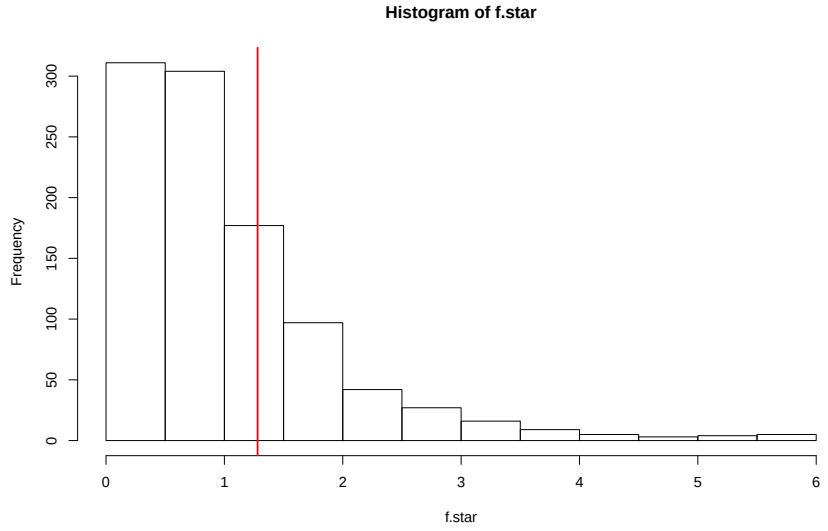
La manière la plus simple d'effectuer un tel test dans R est d'utiliser la fonction `sample`, qui par défaut donne une permutation du vecteur qui lui est passée comme argument :

```
> set.seed(11)
> sample(1:10)
```

```
 [1]  3  1  5  9  7  8  6  4  2 10
```

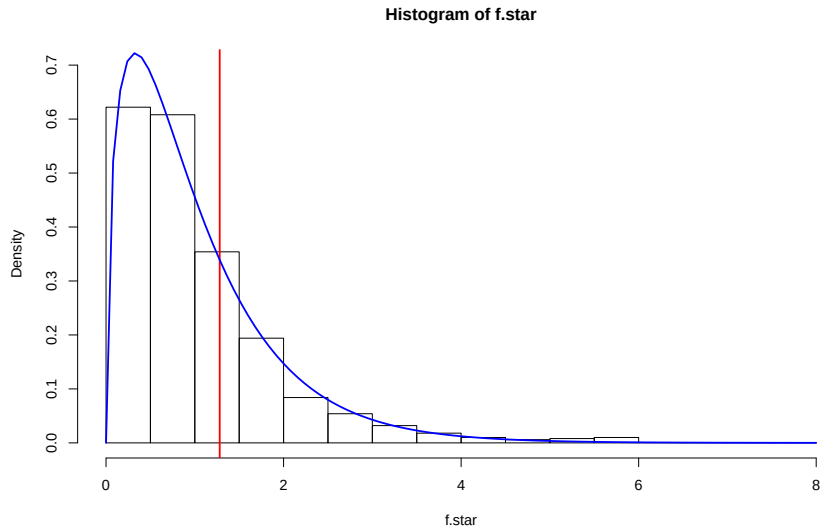
Pour vérifier que le test de permutation donne bien une probabilité équivalente à celle de l'ANOVA quand les conditions de l'ANOVA sont respectées, on peut l'effectuer sur un jeu de données fictif, généré grâce à la fonction `rnorm`.

```
> set.seed(11)
> d.norm<-rnorm(100,10,4) #100 nombres tirés d'une normale N(10,4)
> gr<-factor(rep(1:4,each=25)) #4 groupes de taille 25
> anova(aov(d.norm~gr))

Analysis of Variance Table
Response: d.norm
          Df  Sum Sq Mean Sq F value Pr(>F)
gr         3   50.97  16.990  1.2806 0.2855
Residuals 96 1273.65  13.267

> f.star<-vector(length=1000)
> fobs<-anova(aov(d.norm~gr))[1,4]
> for (i in 1:999) f.star[i]<-anova(aov(d.norm~sample(gr)))[1,4]
> f.star[1000]<-fobs
> sum(f.star>=f.star[1000])/1000

[1] 0.271

> hist(f.star)
> abline(v=f.star[1000],lwd=2,col="red")
```



Si les données ne proviennent pas d'une distribution normale, ou si les variances des différents groupes ne sont pas homogènes, ce test de permutations donnera des résultats valides contrairement à ceux de l'analyse de variance. Dans notre cas de figure, comme les données proviennent d'une distribution normale, la distribution de `f.star` devrait être très similaire à la distribution de $F_{[3,96]}$, ce que montre le graphique suivant :

```
> hist(f.star,freq=FALSE,ylim=c(0.0,0.7),breaks=seq(0,8,0.5))
> abline(v=f.star[1000],lwd=2,col="red")
> curve(df(x,3,96),0,8,add=TRUE,lwd=2,col="blue")
```



Notez aussi que ce test de permutations pourrait utiliser, en lieu et place de la statistique F, soit le carré moyen entre groupes, soit même la somme des carrées entre groupes. En effet, la seule quantité qui change lors de permutations, c'est $\sum_i (\sum_j^n Y_{ij})^2$, à savoir, la somme des carrées correspondant à l'agencement différent des observations au sein des groupes. Plus de détails sur les tests à base de ré-échantillonnage peuvent être trouvés dans l'excellent ouvrage de Brian Manly (Manly [1997]).

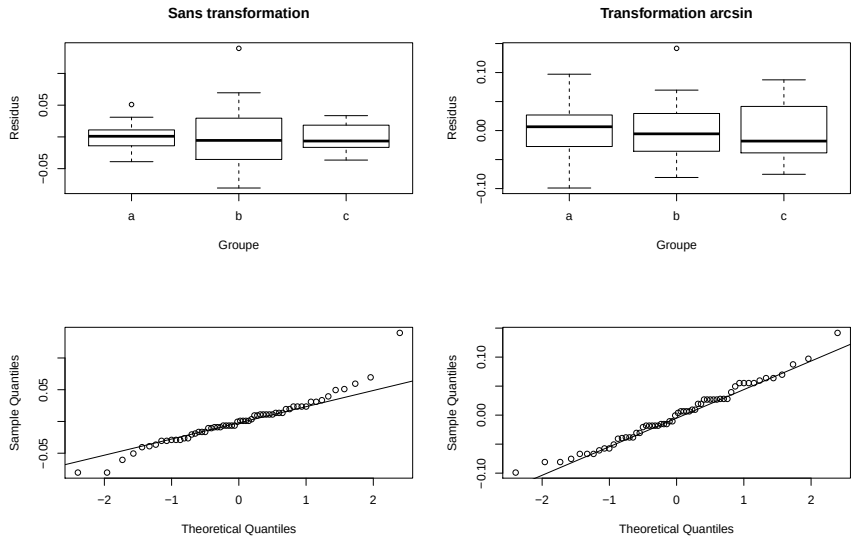
2.4.3 Transformations

Pour d'autres types d'ANOVA, il n'existe pas nécessairement d'équivalent non-paramétrique. Il devient alors nécessaire de transformer les données afin de se ramener à des conditions où l'ANOVA est applicable.

- Vous avez déjà vu l'une de ces transformations, c'est la transformation logarithme. Cette transformation est souvent utilisée dans le cadre de la régression, lorsque la dispersion de la variable réponse semble augmenter avec les valeurs de la variable prédictive. De manière similaire, dans le cadre de l'ANOVA, si nous voyons que la variance augmente avec la moyenne des différents groupes, c'est une transformation que nous pourrions appliquer afin de supprimer l'hétéroscedasticité des données.
- Une autre transformation couramment utilisée est la transformation $\arcsin(y^{1/2})$ qui s'applique souvent sur des proportions. En effet, les proportions suivent une distribution binomiale, et leur variance est donc une fonction de la

moyenne ($\sigma^2_{Binom(N,p)} = \frac{p(1-p)}{N}$). Plus cette moyenne est proche de 0.5, plus la variance est grande. La transformation arcsin permet de pallier à ce problème, et donc d'homogénéiser les variances. Dans l'exemple suivant, nous générons 3 fois 20 nombres aléatoires tirés de 3 distributions binomiales de paramètres 0.05, 0.5 et 0.95.

```
> set.seed(15)
> y1<-c(rbinom(20,100,0.05),rbinom(20,100,0.5),rbinom(20,100,0.95))/100
> a1<-factor(rep(letters[1:3],each=20))
> m1<-aov(y1~a1)
> m1t<-aov(asin(y1^0.5)~a1)
> par(mfrow=c(2,2))
> plot(m1$resid~a1,xlab="Groupe",ylab="Residus",main="Sans transformation")
> plot(m1t$resid~a1,xlab="Groupe",ylab="Residus",main="Transformation arcsin")
> qqnorm(m1$resid,main="");qqline(m1$resid)
> qqnorm(m1t$resid,main="");qqline(m1t$resid)
```



La colonne de gauche de la figure donne les variances par groupe et le qqplot des résidus du modèle sans transformation de la variable réponse, alors que la colonne de droite donne les mêmes graphiques mais pour l'analyse effectuée sur des données transformées par la transformation arcsin. Il est évident que la transformation arcsin a pour effet d'homogénéiser les variances et de réduire les écarts à la normalités des résidus.

– Une autre transformation courante est la transformation racine carrée. Si nous travaillons par exemple sur une mesure de la surface d'un organisme, la surface varie comme le carré de la taille, et donc la transformation racine carrée pourra permettre d'homogénéiser les variances.
– Comme il n'y a pas de raisons a priori de choisir une distribution plutôt qu'une autre, Box & Cox ont développé une procédure qui permet de trou-

ver la meilleur transformation vers la normalité, voir pour plus de détails Sokal and Rohlf [1981], chapitre 13. Une fonction implémentant cette transformation se trouve dans la bibliothèque MASS de R sous le nom de boxcox
- Enfin, la dernière transformation très courante est la transformation en rangs que nous avons déjà rencontré et qui nous ramène à des tests non-paramétriques.

2.5 Modèle à effet fixe : Comparaisons entre groupes

Dans le modèle à effet fixe, nous cherchons bien sur à pouvoir dire quels groupes diffèrent de quels autres. Deux cas de figures se présentent :
- Soit les comparaisons étaient planifiées a priori, avant de faire l'expérience. Par exemple, nous souhaitions comparer le groupe contrôle avec les groupes traités.
- Soit les comparaisons n'étaient pas planifiées. C'est seulement suite aux résultats de l'analyse que nous nous intéressons à la différence entre 2 groupes particuliers. Par exemple, la forêt de Dorigny donne les oiseaux les plus lourds, alors que celle d'Ecublens donne les oiseaux les plus légers. Cette différence est-elle significative ?

Nous voyons bien que dans le dernier cas de figure, il faudra être plus strict que dans le premier cas. Cela revient un peu à demander si la moyenne des 10% plus grands de cette classe est supérieure à la moyenne des 10% plus petits, plutôt que de regarder la différence de moyenne entre 2 groupes représentants chacun 10% de la population de cette classe mais tirés au hasard.

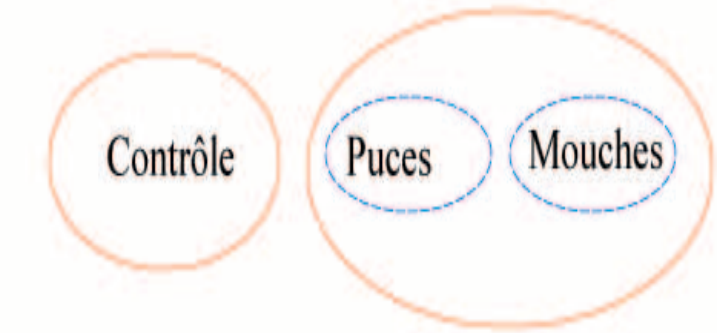
2.5.1 Comparaisons planifiées

Il s'agit de comparaisons prévus avant d'effectuer l'expérience et surtout avant d'avoir vu les données. On peut en prévoir autant qu'il y'a de degrés de libertés pour l'effet groupe, soit $(a-1)$ ou a est le nombre de groupes. Il faut que ces comparaisons soient orthogonales, c'est à dire indépendantes.

Reprenons l'exemple des mésanges déparasitées et parasitées (dans le fichier parus1.txt). Nous avons 3 groupes, donc 2 degrés de liberté, et seulement 2 comparaisons planifiées possible. Mais deux comparaisons planifiées semblent logique : nous pouvons d'abord comparer le groupe contrôle avec les 2 groupes parasités (la question étant : est ce que la présence d'un parasite, quel qu'il soit, affecte le poids à l'envol des jeunes ?). L'autre comparaison possible (la seule), c'est celle entre mouches et puces (Est-ce que le type de parasite affecte le poids à l'envol des jeunes ?). En effet, la comparaison entre contrôle et mouches ou entre contrôle et puces ne sont pas indépendantes de la première comparaison que nous avons faite entre contrôle et traités (La réponse à l'une ou l'autre des questions qui

serait posée par ces 2 comparaisons a déjà été en grande partie donnée par la première comparaison planifiée).

La manière la plus simple de s'assurer de l'indépendance des comparaisons est de les représenter sous forme de diagramme de Venn, ou les patatoïdes de même couleur entourent le ou les groupes à tester, et les patatoïdes de couleurs différentes représentent des comparaisons planifiées différentes. Si les cercles de couleur d'une des comparaisons planifiées se retrouvent dans plus d'un cercle d'une autre, les comparaisons ne sont pas indépendantes. Ce n'est pas le cas ici :



Nous sommes maintenant armés pour effectuer nos différents tests. Commençons par l'analyse globale :

```
> parus <- read.table("./data/parus1.txt",header=TRUE)
> anova(aov(alldat~treat,data=parus))

Analysis of Variance Table
Response: alldat
          Df Sum Sq Mean Sq F value    Pr(>F)
treat      2 94.133  47.066  12.728 0.0001277 ***
Residuals 27 99.843   3.698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comme le résultat est (très fortement) significatif, on peut s'attaquer aux comparaisons planifiées. Pour cela, il sera utile de créer des nouveaux facteurs (variables explicatives). Le premier correspond à la comparaison contrôle/traités, (1 contre 2&3) soit :

```
> cp1<-(parus$treat=="puce" | parus$treat=="mouche")+1
```

Pour le suivant, nous allons utiliser seulement une partie des données, celles des groupes 2 et 3. on aura simplement un `NA` pour les observations du groupe contrôle :

```
> cp2<-parus$treat
> cp2[cp2=="controle"]<-NA
```

Et nous pouvons maintenant calculer les carrés moyens nécessaires à nos comparaisons planifiés. Pour la première comparaison, il s'agira de comparer le groupe contrôle avec les 2 groupes traités, la quantité à calculer est :

$$SS_{cp1} = \frac{(\sum_{i=1}^{n} Y_{1i})^2}{n} + \frac{(\sum_{i=1}^{n}(Y_{2i} + Y_{3i}))^2}{2n} - \frac{(\sum_{i=1}^{n}(Y_{1i} + Y_{2i} + Y_{3i}))^2}{3n}$$

alors que pour la deuxième, ce sera :

$$SS_{cp2} = \frac{(\sum_{i=1}^{n} Y_{2i})^2}{n} + \frac{(\sum_{i=1}^{n} Y_{3i})^2}{n} - \frac{(\sum_{i=1}^{n}(Y_{2i} + Y_{3i}))^2}{2n}.$$

Pour tester les 2 effets, il faut diviser ces sommes des carrées par leurs degrés de libertés respectifs. Dans ce cas précis, il n'y a qu'un degré de liberté pour ces 2 comparaisons, donc $SS = MS$. La statistique F à calculer correspond à ces carrés moyens que l'on divise par le carré moyen de l'erreur, dont la meilleure estimation a été obtenue lors de l'analyse globale (meilleure, puisque basée sur le plus de données, et sur une échantillonnage aussi équilibré que possible).

Nous pouvons donc maintenant procéder à l'analyse :

```
> anova(aov(alldat~cp1,data=parus))

Analysis of Variance Table
Response: alldat
          Df  Sum Sq Mean Sq F value   Pr(>F)
cp1        1  61.996  61.996  13.153 0.001132 **
Residuals 28 131.980   4.714
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(aov(alldat~cp2,data=parus))

Analysis of Variance Table
Response: alldat
          Df Sum Sq Mean Sq F value  Pr(>F)
cp2        1 32.137  32.137  7.0093 0.01637 *
Residuals 18 82.527   4.585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les quantités qui nous intéressent dans ces 2 analyses sont les carrés moyens de `cp1` et `cp2` respectivement, qui seront testés par rapport au carré moyen des résidus du modèle initial (la variance intra-groupes calculée sur l'ensemble des données). Notez que nous n'utilisons pas directement le résultat de ces analyses, nous voulons simplement extraire les carrés moyens de chaque comparaison, et le

moyen le plus simple est de les extraire de l'analyse de variance correspondante. Nous aurions pu aussi écrire une petite fonction pour les extraire.

On extrait facilement les carrés moyens qui correspondent aux troisièmes colonnes des premières lignes des tables d'ANOVAs, mais vérifions auparavant que la somme des sommes des carrés (Les sommes des carrés sont localisées dans les deuxièmes colonnes des tables d'ANOVAs) de `cp1` et `cp2` est bien égale à la somme des carrés de `treat` de l'analyse initiale :

```
> m2<-anova(aov(alldat~cp2,data=parus))
> m1<-anova(aov(alldat~cp1,data=parus))
> m0<-anova(aov(alldat~treat,data=parus))
> m0[1,2] #la somme des carrés traitement

[1] 94.13266

> m1[1,2]+m2[1,2] # la somme de la somme des carrées cp1 & cp2

[1] 94.13266
```

Nous voyons clairement que ces 2 quantités sont égales, ce qui confirme bien que nos 2 comparaisons sont indépendantes et que nous avons effectué toute les comparaisons qu'il était possible de faire.

Nous pouvons maintenant calculer les 2 statistiques F de nos 2 comparaisons planifiées, et estimer directement la probabilité qui leur est associée :

```
> pf(m1[1,3]/m0[2,3],m1[1,1],m0[2,1],lower=F)

[1] 0.0003444914

> pf(m2[1,3]/m0[2,3],m2[1,1],m0[2,1],lower=F)

[1] 0.006524837
```

Les résultats de ces 2 tests sont très fortement significatifs. On peut donc conclure que

- La présence d'un parasite dans le nid, qu'il soit puce ou mouche, affecte fortement le poids à l'envol des jeunes. Les jeunes issus de nids non parasités sont plus lourds à l'envol.
- Les mouches ont un effet moindre sur le poids des jeunes que les puces.

On représentera généralement les résultats de l'analyse sous la forme suivante :

```
Response: alldat
              Df Sum Sq Mean Sq F value    Pr(>F)
treat          2 94.133  47.066  12.728 0.0001 ***
    cp1        1 61.996  61.996  16.765 0.0003 ***
    cp2        1 32.137  32.137  8.6905 0.0065 **

Residuals     27 99.843   3.698
```

Notez encore que l'intérêt de cette méthode est d'augmenter la puissance des tests associée aux comparaisons planifiées. En effet, comme l'erreur utilisée est celle de l'analyse initiale, elle a le nombre maximum de degré de liberté. Si nous avions utilisé les résultats de l'analyse pour `cp2` (m2), où seulement 20 observations ont été utilisées, nous aurions eu une analyse moins puissante (notion que nous définirons plus précisément plus tard), ce qui se traduit en général par une probabilité supérieure associée au test.

2.5.2 Comparaisons non planifiées

Pour ce type de comparaisons, il existe une quantité incroyable de tests qui ont tous des avantages et des inconvénients. Je présenterai ici la plus simple et la plus courante, appelée correction de Bonferroni.

Si nous effectuons k tests, la probabilité d'obtenir au moins un résultat significatif au seuil α, c'est un moins la probabilité qu'aucun test ne soit significatif au seuil α, soit $1-(1-\alpha)^k$. Si l'un au moins des tests est significatif, nous rejetterions notre hypothèse nulle de départ, à savoir, que toutes les moyennes sont identiques. Le seuil de rejet réalisé n'est donc pas α, mais : $\alpha' = 1-(1-\alpha)^k$. Si nous effectuons 20 tests au seuil de 5%, nous obtenons $\alpha' = 1-(1-0.05)^{20} = 0.64$. Nous aurions donc plus de 6 chances sur 10 (alors que nous souhaitions avoir 5 chances sur 100) d'obtenir au moins un résultat significatif.

Si nous prévoyons de faire k tests, le risque de première espèce α (l'erreur de type I) doit donc être ajusté de telle manière que globalement, nous ayons le seuil voulu. Soit $\alpha' = 1-(1-\alpha)^{(1/k)} \simeq \alpha/k$.

Ainsi, si dans une analyse avec 7 groupes nous désirons effectuer 20 tests pour un seuil global de $\alpha = 0.05$, le seuil de rejet des 20 différentes hypothèses nulles devra être ajusté à $\alpha' = 1-(1-0.05)^{(1/20)} = 0.0026$, ce qui est très inférieur à 0.05.

Signalons que cette procédure est très conservatrice (et donc qu'il sera très difficile de rejeter une fausse hypothèse nulle). Trouver des procédures de correction pour tests multiples qui soient moins conservatrices tout en respectant le seuil global de rejet que l'investigateur a prédéfini est l'un des domaines de recherches très actif en statistiques aujourd'hui. C'est notamment l'utilisation de plus en plus courante des méthodes à haut-débit (High Throuput ou HT en anglais) telles que génomique et séquençage haut débit, protéomique, métabolomique, transcriptomique, qui a rendu ce domaine de recherche en statistique extrêmement important.

Notons enfin que si, dans le cadre de comparaisons planifiées, nous souhaitons en effectuer qui ne soit pas indépendantes, il sera aussi nécessaire d'ajuster le risque de première espèce avec cette méthode.

False Discovery rate (FDR)

Supposons que nous ayons effectué m tests d'hypothèses. Les résultats peuvent être arrangés dans un tableau comme suit :

	Tests non-significatifs	Tests significatifs	Total
Hyp. nulles vrais	U	V	m_0
Hyp. nulles fausses	T	S	$m - m_0$
Total	$m - R$	R	m

Dans ce tableau, nous connaissons m et R, mais les autres éléments ne sont pas connus. V est le nombre d'erreurs de première espèce, et T est le nombre d'erreur de deuxième espèce. La correction de Bonferroni classique contrôle pour le "Family Wise Error Rate" ($FWER = Pr(V \geq 1)$, taux d'erreur sur l'ensemble de la famille), et contrôle que la probabilité de rejeter au moins une hypothèse nulle vrai soit inférieure à α, le risque de premiere espèce.

Mais une autre manière de corriger a été proposé dans le milieu des années 90. Il s'agit du "False Discovery Rate" (FDR, que l'on pourrait traduire par "taux de fausse découverte"). Ici, l'idée est de maintenir la proportion de faux positifs $\frac{V}{V+S} = \frac{V}{R}$ inférieure au seuil α que nous nous sommes fixé à priori. Il existe plusieurs méthodes pour estimer ce "FDR". La plus courante, quand les tests sont indépendants, consiste à trier les probabilités associées aux différents tests ("P-values") par ordre croissant $H_1, H_2, \cdots, H_k, \cdots, H_m$, puis à trouver le k le plus large pour lequel

$$P_{(k)} \leq \frac{k}{m}\alpha.$$

On rejettera alors toutes les hypothèses nulles $H_i, i = 1, \cdots, k$, et on acceptera les autres. De plus amples informations sur ce FDR et son utilisation vous seront donnés dans le cadre du cours de bio-informatique.

Comparaisons non-planifiées dans R

Outre la correction de Bonferroni, présentée en cours, il existe une multitude de procédures pour effectuer ces comparaisons. Elles sont présentées en détail dans le livre Biometry (Sokal and Rohlf [1981]) et dans une multitude d'autres ouvrages de statistiques. L'une de ces méthodes est la méthode de Tukey appelée "Honestly Significant Difference", et qui se trouve dans le package de base de R, sous le nom `TukeyHSD`. Cette méthode donne un intervalle de confiance (par défaut à 95%) pour la différence de moyenne entre les différents groupes, avec ajustement de la probabilité de rejet. Si l'intervalle de confiance n'est pas recouvrant avec 0, la différence entre les 2 groupes est significative au seuil fixé (1-conf.level). Pour notre cas :

```
> TukeyHSD(aov(alldat~treat,data=parus))

  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = alldat ~ treat, data = parus)

$treat
                     diff       lwr       upr     p adj
mouche-controle -1.781884 -3.914154  0.350386 0.1147447
puce-controle   -4.317101 -6.449371 -2.184831 0.0000834
puce-mouche     -2.535217 -4.667487 -0.402947 0.0173606
```

Et donc nous conclurions que le groupe puce et le groupe contrôle diffèrent, de même que le groupe puce et mouche, mais que le groupe mouche n'est pas différent du groupe contrôle. Notez l'utilisation de la fonction `aov` plutôt que `lm`. En effet, la fonction `TukeyHSD` n'accepte que le format de résultat provenant de la commande `aov`.

Nous aurions bien entendu aussi faire les 3 tests t pour chaque paire de groupes, en utilisant le seuil de rejet à $\alpha' \simeq \alpha/3$ soit 0.0167 :

```
> t.test(parus$alldat[1:10],parus$alldat[11:20],var.equal=T)

        Two Sample t-test
data:  parus$alldat[1:10] and parus$alldat[11:20]
t = 2.4826, df = 18, p-value = 0.02313
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2739514 3.2898166
sample estimates:
mean of x mean of y
 18.21635  16.43447

> t.test(parus$alldat[1:10],parus$alldat[21:30],var.equal=T)

        Two Sample t-test
data:  parus$alldat[1:10] and parus$alldat[21:30]
t = 4.8676, df = 18, p-value = 0.0001238
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.453771 6.180431
sample estimates:
mean of x mean of y
 18.21635  13.89925
```

```
> t.test(parus$alldat[11:20],parus$alldat[21:30],var.equal=T)

        Two Sample t-test
data:  parus$alldat[11:20] and parus$alldat[21:30]
t = 2.6475, df = 18, p-value = 0.01637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5234082 4.5470258
sample estimates:
mean of x mean of y
 16.43447  13.89925
```

Et nous arrivons, après ajustement du seuil de rejet à 0.017 plutôt que les 0.05 habituels, aux mêmes conclusions qu'avec `TukeyHSD`.

Sachez enfin que R possède une bibliothèque (library) particulière pour effectuer ce type de comparaisons, la bibliothèque `multcomp`.

2.6 Modèle à effet aléatoire : estimation des composantes de la variance

Nous avons vu précédemment que si l'hypothèse nulle est vrai, il existe 3 manières équivalentes d'estimer la variance : sur la base de la variance intra-groupe, sur la base de la variance des moyennes des groupes, et enfin en ignorant complètement l'appartenance aux différents groupes et en considérant toutes les observations ensemble. Si l'hypothèse nulle d'absence de différence entre groupes s'avère fausse par contre, nous nous intéresserons à la variance supplémentaire contribuée par la différence entre les groupes, qui correspond à la variance des A_i dans le modèle $Y_{ij} = \mu + A_i + \epsilon_{ij}$, soit σ_A^2. Cette variance supplémentaire s'estime à partir de la relation entre les carrés moyens et les carrés moyens attendus, donnés dans le tableau suivant :

Carré moyen	Modèle II
MS_B	$\sigma_\epsilon^2 + \sigma_A^2 \frac{\left[(\sum^a n_i)^2 - \sum^a n_i^2\right]}{\sum^a n_i - a}$
MS_B si n_i égaux	$\sigma_\epsilon^2 + n\sigma_A^2$
MS_W	σ_ϵ^2

La composante de la variance intra-groupes σ_ϵ^2 s'obtient simplement comme le carré moyen intra-groupes MS_W.

Pour le cas simple où chaque groupe est constitué du même nombre d'observations, la variance inter-groupes σ_A^2 s'estime comme :

$$\sigma_A^2 = \frac{MS_B - \sigma_\epsilon^2}{n}$$

Pour le cas où le nombre d'observations diffère entre les groupes, nous obtenons :

$$\sigma_A^2 = \left(MS_B - \sigma_\epsilon^2\right) \frac{\sum^a n_i - a}{[(\sum^a n_i)^2 - \sum^a n_i^2]}$$

La variance totale, incluant la variance intra et inter groupes s'estime enfin comme $\sigma_A^2 + \sigma_\epsilon^2$. Notez bien que si l'analyse de variance ne donne pas un résultat significatif (si la probabilité associée au test est supérieure au seuil α que nous nous étions fixés au préalable), cela signifie qu'il n'y a pas de variance supplémentaire due aux groupes, et donc la meilleure estimation de la variance totale est la moyenne des variances intra-groupes, soit $\sigma_\epsilon^2 = MS_W$.

2.6.1 Exemple : estimation de l'héritabilité d'un caractère

Le modèle d'ANOVA aléatoire est très fréquemment utilisé en *génétique quantitative*, afin d'estimer quelle part de la variance observée d'un caractère (le poids, la taille, le nombre de graines, l'aptitude au saut etc...) est dû à des effets génétiques ou à des effets environnementaux. Typiquement, des familles de demi frères (même père, différentes mères) sont élevés dans un environnement commun (il peut s'agir de plantes dans une serre ou d'animaux dans une même animalerie, ceci afin de minimiser les différences dues à des facteurs environnementaux qui seraient alors confondus avec un effet famille). La similarité génétique (la parenté, ou proportion attendue de gènes en commun) entre demi-frères est de 1/4. On estimera donc la composante génétique de la variance d'un trait comme 4 fois la covariance entre demi-frères, covariance qui s'estime comme la composante de la variance inter familiale σ_A^2. En effet, selon notre modèle d'ANOVA, nous aurons

$$Cov(DF) = Cov(Y_{ij}, Y_{ik}) = Cov((\mu + A_i + \epsilon_{ij}), (\mu + A_i + \epsilon_{ik}))$$

où i réfère le parent commun de la famille de demi-frère et j et k deux demis-frères différents. Ceci peut se réécrire, en utilisant les propriétés de la covariance :

$$Cov(DF) = Cov(A_i, A_i) + Cov(A_i, \epsilon_{ik}) + Cov(A_i, \epsilon_{ij}) + Cov(\epsilon_{i,j}, \epsilon_{i,k})$$

Mais, par définition $Cov(A_i, \epsilon_{ik})$, $Cov(A_i, \epsilon_{ij})$ et $Cov(\epsilon_{i,j}, \epsilon_{i,k})$ sont nulles, et l'expression pour la covariance entre demi frères se réduit alors à

$$Cov(DF) = Cov(A_i, A_i) = Var(A_i) = \sigma_A^2,$$

soit la variance entre famille de demi-frères. L'héritabilité d'un trait s'estime alors comme 4 fois le rapport entre la composante de la variance inter familiale σ_A^2 et la variance totale $\sigma_A^2 + \sigma_\epsilon^2$. Les généticiens, éleveurs et améliorateurs de plantes s'intéressent à l'héritabilité d'un trait parce que la sélection ne peut agir que sur un trait héritable, c'est à dire qui a une composante génétique transmissible. Ces notions seront approfondies dans le cadre du cours de biologie des populations en troisième année. Pour les impatients, deux excellentes références sont les ouvrages de Falconer and MacKay [1996] et de Lynch and Walsh [1998].

Chapitre 3

analyse de variance à 2 facteurs

3.1 Introduction

Il arrive fréquemment que l'effet de plusieurs facteurs intéresse le biologiste. L'ANOVA à deux critères de classification lui permet d'analyser les résultats d'une série d'expériences visant à tester si ces facteurs influencent la variable d'intérêt. Par exemple, on peut vouloir étudier si la température de l'eau et son pH influencent le taux de mortalité de jeunes truites d'élevage. On pourrait planifier une expérience dans laquelle on ferait varier la température, et une autre où on ferait varier le pH. On pourrait ensuite analyser les résultats de ces expériences par deux ANOVA. Cette approche, fort logique et naturelle, a cependant une faiblesse : elle ne permettrait pas de dire si l'effet de la température sur la mortalité dépend du pH ; ou si l'effet du pH sur la mortalité dépend de la température. Dans un cas comme celui-ci, un design expérimental d'ANOVA à deux critères de classification permettrait de répondre à cette question. Les modèles d'ANOVA à plusieurs facteurs de classification incluent des termes pour chaque facteur (effets principaux) et pour les interactions.

Supposons que nous ayons deux facteurs qui varient entre les traitements, A et B. Le facteur A varie entre deux niveaux (1 et 2) alors que le facteur B a cinq niveaux différents (I-V). Les données pourraient être regroupées de deux façons différentes :

	B				
A	I	II	III	IV	V
1	**	**	****	**	**
2	****	***	**	*****	**

ou bien :

	B									
	I		II		III		IV		V	
A	1	2	1	2	1	2	1	2	1	2
	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*		*	*	*
		*	*	*		*	*		*	*

La distinction est faite selon la nature des niveaux du facteur A. Si les niveaux 1 et 2 du facteur A sont les mêmes pour tous les niveaux du facteur B, alors il s'agit d'un design d'ANOVA factorielle à deux critères de classification. Par contre, si les niveaux du facteur A diffèrent entre les niveaux du facteur B, alors il s'agit d'une ANOVA hiérarchique à deux niveaux de classification. Par exemple, si un chercheur est intéressé à déterminer l'effet du sexe (facteur A) et de l'âge (facteur B) sur la taille des lézards, il s'agit d'un design à deux critères de classification. Par contre, si le même chercheur étudie l'effet de l'identité du technicien (facteur A) et de l'âge (facteur B) sur la taille des lézards et que chaque groupe d'âge de lézards (I à V) est mesuré par une paire différente de techniciens, il s'agit d'un design hiérarchique. Donc, quoique l'ANOVA hiérarchique à deux niveaux de classification semble faire intervenir deux facteurs, ce n'est pas vraiment le cas puisque le deuxième facteur représente seulement une autre source de variabilité qui n'est pas particulièrement intéressante (du moins dans la plupart des cas). L'ANOVA hiérarchique à deux niveaux de classification s'apparente plutôt à l'ANOVA à un seul critère de classification.

3.2 Analyse de variance hiérarchique

L'analyse de variance hiérarchique, où un facteur est niché (nested en anglais) dans l'autre, correspond à une situation expérimentale extrêmement fréquente : dès que nous sommes en présence de deux facteurs, et que les différents niveaux d'un des facteurs ne se retrouvent pas dans les différents niveaux de l'autre, nous sommes en présence d'une ANOVA hiérarchique. Le facteur niché est dans l'immense majorité des cas un facteur aléatoire, que nous pensons être susceptible d'ajouter de la variance. Voyons d'abord quelques exemples.

3.2.1 Exemples d'analyse hiérarchique

Revenons au poids de nos jeunes mésanges à l'envol provenant de nids que nous avons passés au micro-ondes puis manipulés en ajoutant des parasites. Nous avons considéré jusqu'à présent le poids des jeunes à l'envol, sans tenir compte du fait qu'ils proviennent du même nid ou pas. Or, deux jeunes provenant du même nid ont de fortes chances d'avoir un poids similaire, car ils proviennent très vraisemblablement des mêmes parents. En utilisant chaque jeune pour l'analyse, nous commettons l'un des péchés de l'apprenti statisticien, celui de la PSEUDO-RÉPLICATION, à savoir l'utilisation de données non-indépendantes. Une solution pourrait être de faire l'analyse non pas sur le poids individuel des jeunes, mais sur la moyenne du poids des jeunes à l'envol de chaque nid. Une solution plus élégante et amenant plus d'information consiste à effectuer une analyse hiérarchique, ou les nids seraient emboités dans les 3 traitements. En effet ici, chaque nid est affecté à un et un seul traitement. Nous pourrions ainsi estimer si l'effet nid à l'intérieur des traitements génère une variance additionnelle.

Imaginons que nous nous intéressions à la qualité des étudiants en mathématiques dans les différents gymnases de Suisse. On pourrait simplement analyser les notes des étudiants de chaque gymnase via une ANOVA à un facteur. Mais si plusieurs professeurs de mathématiques enseignent, il n'est pas impossible qu'il y ait un effet professeur sur les notes des élèves. Mais chaque professeur n'enseigne que dans un gymnase. On est donc à nouveau en présence d'un modèle hiérarchique, ou les professeurs de mathématiques sont nichés au sein des gymnases.

Très souvent, les animaux (souris, poissons...) qui sont utilisés pour une expérience sont logés dans des cages ou des aquariums, et l'ensemble des individus d'une cage ou d'un aquarium subit l'un des traitements de l'expérience. La véritable unité expérimentale, dans ce type d'expérience, est la cage ou l'aquarium. Si on souhaite estimer la variabilité au sein d'une cage ou d'un aquarium, nous devons alors effectuer une analyse hiérarchique, avec l'effet cage niché dans l'effet traitement.

Souvent, plusieurs mesures de même type sont prises sur un même individu lors d'une expérience avec plusieurs traitements. Par exemple, on fera peut être plusieurs comptages de globules blanc sur 3 lames contenant un échantillon san-

guin du même animal. Les différentes lames seront alors nichées dans le facteur animal, qui sera lui même niché dans le facteur traitement.

3.2.2 Modèles et hypothèses testées

Il existe deux types de modèles pour l'analyse de variance hiérarchique, le modèle mixte, et le modèle complètement aléatoire. Dans le modèle mixte, le facteur niché est aléatoire, alors que le facteur externe est fixe. Dans le modèle aléatoire, les 2 facteurs sont aléatoires.

Le facteur interne, ou niché, est aléatoire dans l'immense majorité des cas, bien que des exceptions existent.

Un exemple d'effet fixe niché est le suivant : si nous nous intéressions à la consommation d'essence de différents modèles automobiles. Ces modèles sont répartis entre marques (le facteur externe, par exemple Toyota, Volkswagen, Peugeot, Chrysler) et chaque marque possède plusieurs modèles (le facteur interne niché). Notons cependant que même dans cet exemple, les différents modèles pourraient être classifiés en catégories de véhicules (berline, monospace...), nous amenant alors à un modèles à 2 facteurs croisés plutôt qu'a un modèle hiérarchique.

Le modèle de l'analyse de variance pour un facteur fixe et un facteur aléatoire s'écrit comme suit :

$$Y_{ijk} = \mu + \alpha_i + B_{ij} + \epsilon_{ijk} \tag{3.1}$$

où μ est la moyenne générale (sur tous les niveaux des facteurs A et B) de la variable réponse Y, α_i correspond à l'écart à la moyenne générale du niveau i du facteur A, B_{ij} correspond à l'écart à la moyenne du niveau i du facteur A du groupe j du facteur B et ϵ_{ijk} correspond à l'erreur, soit la partie des variables Y_{ijk} non expliquée par le modèle.

Nous voulons tester les 2 hypothèses nulles suivantes :

- $H_0^\alpha : \alpha_1 = \alpha_2 = \ldots = \alpha_n$, c'est à dire que les moyennes de tous les groupes du facteur externe sont égales.
- $H_0^B : B_{11} = B_{12} = \ldots = B_{1k_1} = B_{21} = \ldots = B_{2k_2} = \ldots = B_{n1} = \ldots = B_{nk_n}$, c'est à dire que les moyennes de tous les groupes constitués par le niveau niché, une fois tenu compte des moyennes du facteur externe, sont égales. Cela revient à tester l'égalité des moyennes ajustées des sous-groupes, moyennes ajustées pour $\mu + \alpha_i$.

Les hypothèses alternatives correspondants à ces deux hypothèses nulles sont H_1^α : au moins un des groupes du facteur externe a une moyenne qui diffère des autres et H_1^B : au moins un des groupes constituant le facteur niché a une moyenne ajustée pour l'effet du facteur externe qui diffère des autres. Dans ce dernier cas cependant, nous ne sommes pas intéressés par la valeur précise des moyennes, mais plutôt par la variance additionnelle générée par la prise en compte de ce facteur. Pour le facteur externe par contre, c'est bien à l'effet de chacun des groupes sur la moyenne que nous nous intéressons.

L'expression de modèle pour 2 facteurs aléatoires est le suivant :

$$Y_{ijk} = \mu + A_i + B_{ij} + \epsilon_{ijk} \tag{3.2}$$

Le seul changement est la transformation de l'effet fixe α en un effet aléatoire A. Donc l'hypothèse nulle que nous testerons pour le niveau externe $H_0^A : A_1 = A_2 = \ldots = A_n$ n'aura pas pour but de calculer précisément les différences de moyennes entre chaque groupe mais plutôt de quantifier la variance supplémentaire générée par ce facteur.

Ensuite, le principe de l'analyse est le même que pour l'ANOVA à un facteur. Supposons que nous ayons b sous-groupes nichés dans a groupes, où chacun des sous-groupes est constitué de n observations. Il s'agit alors de décomposer la somme des carrés totale SS_T en une composante intra sous-groupes SS_W (la somme des carrés résiduels), en une somme des carrés inter sous-groupes, intra-groupes $SS_{S(G)}$, et en une somme des carrés inter-groupes SS_G. La table pour l'analyse de variance hiérarchique à un niveau se présente comme suit :

Source	DDL	SS	MS
Entre groupes	$(a-1)$	$\frac{\sum^a(\sum^b\sum^n Y_{ijk})^2}{nb} - \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	$\frac{SS_G}{a-1}$
Entre sous-groupes dans les groupes	$a(b-1)$	$\frac{\sum^a\sum^b(\sum^n Y_{ijk})^2}{n} - \frac{\sum^a(\sum^b\sum^n Y_{ijk})^2}{bn}$	$\frac{SS_{S(G)}}{a(b-1)}$
Intra sous-groupes	$ab(n-1)$	$\sum^a\sum^b\sum^n Y_{ijk}^2 - \frac{\sum^a\sum^b(\sum^n Y_{ijk})^2}{n}$	$\frac{SS_W}{ab(n-1)}$
Total	$abn-1$	$\sum^a\sum^b\sum^n Y_{ijk}^2 - \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	

Notez à nouveau que deux tests sont effectués dans cette analyse, un pour l'effet intra groupes, et qui sera le test d'un effet aléatoire, et un pour l'effet groupe, qui sera en fonction du type de données soit un effet fixe, soit un effet aléatoire. Afin de déterminer quel dénominateur est approprié pour le calcul des deux statistiques F correspondant aux deux tests, il est nécessaire d'obtenir l'expression des carrés moyens attendus. Ces carrés moyens attendus, dans le cadre de l'analyse de variance hiérarchique à un facteur, s'expriment comme suit :

Carrés moyens	Modèle aléatoire	Modèle mixte	Tests F
MS_G	$\sigma^2 + n\sigma^2_{S(G)} + nb\sigma^2_G$	$\sigma^2 + n\sigma^2_{S(G)} + nb\frac{\sum^a \alpha^2}{a-1}$	$MS_G/MS_{S(G)}$
$MS_{S(G)}$	$\sigma^2 + n\sigma^2_{S(G)}$	$\sigma^2 + n\sigma^2_{S(G)}$	$MS_{S(G)}/MS_W$
MS_W	σ^2	σ^2	

Pour les carrés moyens attendus intra sous-groupes, on reconnaît l'expression des carrés moyens attendus pour l'effet groupe de l'analyse de variance à un facteur aléatoire. Pour l'effet inter-groupe, le carré moyen attendu dépend du modèle sous-jacent. Notez encore que dans l'expression du carré moyen inter-groupe intervient non seulement l'erreur ou variance résiduelle, mais aussi la composante de la variance inter sous-groupe intra-groupe $\sigma^2_{S(G)}$. Le test adéquat pour l'effet inter-groupe sera donc $F_G = MS_G/MS_{S(G)}$, puisque cette quantité a une espérance de 1 si $\sigma^2_G = 0$.

3.2.3 Analyse de variance hiérarchique dans R

Un point important à garder en mémoire lorsque le logiciel R est utilisé concerne le type de test F que le logiciel effectue automatiquement. Par défaut, R va tester tous les effets d'une ANOVA contre l'erreur, quel que soit le modèle (I, II ou mixte). If faudra donc rester attentif à ce fait, et penser pour chaque analyse effectuée, à vérifier si le test effectué par R est correct, et, le cas échéant, le corriger.

Note : Il est possible de spécifier dans R le type d'erreur pour les différents effets à l'aide la commande `Error()` au sein de la formule du modèle. Il est cependant plus simple dans un premier temps de calculer soit même le rapport de carrés moyens adéquat. Notez encore que la bibliothèque `lme4` de R permet de définir tout type de modèle mixte et de structure d'erreur, son utilisation est par contre assez complexe.

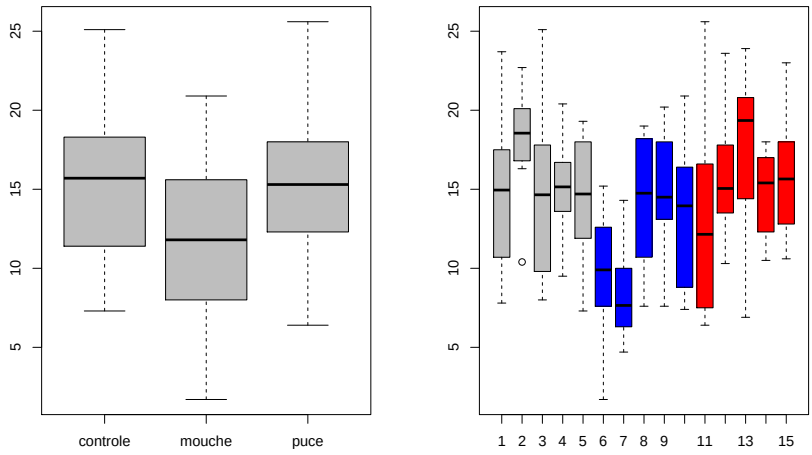
Nous allons reprendre comme exemple le poids de ces jeunes mésanges à l'envol. A nouveau, les nids (15 en tout) ont subi différents traitements (contrôle, mouche ou puce, 5 nids alloués à chaque traitement). Comme plusieurs jeunes (10) ont été pesés dans chaque nid, nous sommes typiquement en présence d'un modèle d'ANOVA hiérarchique, avec un effet fixe qui est le traitement, et un effet aléatoire qui est le nid, et l'effet nid qui est niché (sic !) dans l'effet traitement. En effet, les différents nids de chaque traitement n'ont rien à voir les uns avec les autres, et, même s'ils avaient été numérotés de 1 à 5 dans chaque traitement, le nid 1 du groupe contrôle n'aurait rien eu à voir avec le nid 1 du groupe mouche.

Afin d'éviter toute confusion, les nids dans le fichier de données (parus-hier.txt) ont été numérotés de 1 à 15. Je vous encourage vivement à procéder de cette manière (au moins au début) lorsque vous mettrez en place vos propres expériences, ceci évitera

de se mélanger entre modèles hiérarchiques et modèles croisés que nous verrons par la suite. Sachez cependant que la place en mémoire requise par l'ordinateur est plus importante si vous numérotez de cette manière.

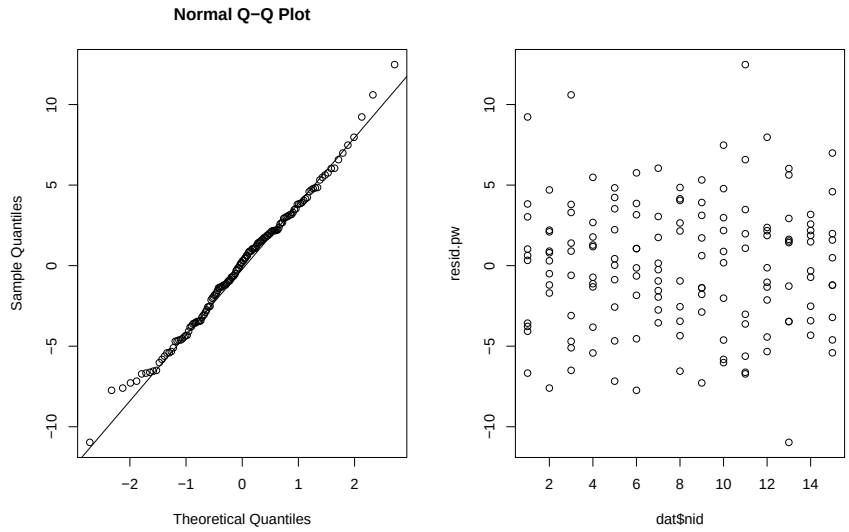
Une première chose à faire est de représenter les données. Un boxplot par traitement et par nid est représenté ci-dessous. Il semble que le poids des oisillons à l'envol diffère entre nids, mais aussi peut être entre traitements :

```
> dat<-read.table("./data/parus_hier.txt",header=TRUE)
> dat$parusw<-dat$data
> par(mfrow=c(1,2))
> with(dat,boxplot(parusw~trait,col="grey"))
> with(dat,boxplot(parusw~nid,col=rep(c("grey","blue","red"),each=5)))
```



Avant d'effectuer l'analyse, nous devons vérifier les conditions d'applications de l'ANOVA (ces conditions sont les mêmes que pour l'ANOVA à un facteur), et en particulier la normalité et l'homogénéité des variances :

```
> moy.nid<-with(dat,tapply(parusw,nid,mean))
> resid.pw<-with(dat,parusw-rep(moy.nid,each=10))
> par(mfrow=c(1,2))
> qqnorm(resid.pw)
> qqline(resid.pw)
> plot(dat$nid,resid.pw)
```



Aucune déviation franche à la normalité ou à l'homogénéité des variances n'est apparente sur ces 2 graphiques. Nous pouvons donc maintenant effectuer l'analyse. Afin de spécifier au logiciel que le facteur nid est niché dans le facteur traitement, on utilise l'opérateur "/" ou "%in%". La syntaxe sera donc la suivante

```
> dat$nid<-factor(dat$nid) #pourquoi cette ligne?
> anova(aov(parusw~trait/nid,data=dat))

Analysis of Variance Table
Response: parusw
           Df  Sum Sq Mean Sq F value    Pr(>F)
trait       2  375.94 187.971 10.6796 4.945e-05 ***
trait:nid  12  570.14  47.512  2.6994  0.002715 **
Residuals 135 2376.13  17.601
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(aov(parusw~trait+nid%in%trait,data=dat))

Analysis of Variance Table
Response: parusw
           Df  Sum Sq Mean Sq F value    Pr(>F)
trait       2  375.94 187.971 10.6796 4.945e-05 ***
trait:nid  12  570.14  47.512  2.6994  0.002715 **
Residuals 135 2376.13  17.601
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La syntaxe `trait/nid` est plus compacte que `trait+nid%in%trait`, aussi elle sera généralement préférée. Nous retrouvons bien nos trois lignes dans la table de l'ANOVA, correspondant respectivement à l'effet traitement, avec 2 degrés de libertés, l'effet nid à l'intérieur de traitement (nid(traitement) noté `trait:nid` par R), avec 12 degrés de liberté soit $3 \times (5-1)$, et enfin 135 degrés de liberté pour la ligne résidus, soit $3 \times 5 \times (10-1)$. Les carrés moyens pour l'effet traitement et nid(traitement) sont de 187.97 et 47.51 respectivement. Le test adéquat pour l'effet nid(traitement) (Notez que l'on commence à lire la table à partir du bas, c'est une règle pour la lecture de la table d'ANOVA) est $F = MS_{nid(trait)}/MS_{Residuals}$, soit 2.699, ce qui est bien la valeur affichée dans la table. Cet effet est très significatif ($p < 0.01$), donc il existe des différences importantes du poids moyen des jeunes à l'envol entre nids, et on peut estimer la composante de la variance du poids des jeunes à l'envol liée à cet effet :

$$\sigma^2_{Nid(Trait)} = (MS_{Nid(Trait)} - \sigma^2_e)/n = (47.51 - 17.60)/10 \simeq 3$$

La variance du poids à l'envol dû à l'effet nid est de l'ordre de $3gr^2$, alors que la variance du poids à l'intérieur d'un nid est estimée à $17.6gr^2$.

Nous continuons notre analyse en calculant la statistique F correspondant à l'effet traitement. Il s'agit de $MS_{Trait}/MS_{Nid(Trait)}$, soit $187.97/47.51 = 3.956$. Cette valeur est différente (2.5 fois plus basse !) de celle qui apparaît dans la table d'ANOVA produite par R (10.68). Comme expliqué plus haut, la valeur du test F donnée par R correspond toujours à la division par le carré moyen de l'erreur, ce qui ici est inadéquat. On peut calculer la probabilité associée à la valeur de F obtenue grâce à la commande suivante :

```
> m.n<-anova(aov(parusw~trait/nid,data=dat))
> pf(m.n[1,3]/m.n[2,3],m.n[1,1],m.n[2,1],lower=F)

[1] 0.04789789
```

Il y a donc un effet significatif du traitement puisque la probabilité associée au test est de 0.0479, mais l'effet est beaucoup moins fort que si nous avions pris les résultats de l'output de R pour argent comptant.

Une autre manière de faire l'analyse aurait été de prendre le poids moyen des jeunes à l'envol par nid, et donc de ne considérer que l'effet traitement, le poids moyen de chaque nid devenant alors la variable réponse :

```
> parusw.m<-tapply(dat$parusw,dat$nid,mean)
>   #tapply permet d'obtenir la valeur d'une fonction(ici la moyenne)
>   #pour chaque niveau d'une variable facteur, ici nid.
> trait.m<-factor(rep(1:3,each=5)) #plus que 15 obs, 5 nids par traitement
> anova(aov(parusw.m~trait.m))
```

```
Analysis of Variance Table
Response: parusw.m
          Df Sum Sq Mean Sq F value Pr(>F)
trait.m    2 37.594 18.7971  3.9563 0.0479 *
Residuals 12 57.014  4.7512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Et vous voyez que nous obtenons exactement le même résultat. Mais L'analyse hiérarchique donne une information supplémentaire, sur l'effet des nids : leur prise en compte permet il d'expliquer une part de la variance ? L'analyse hiérarchique sera donc préférée.

Enfin, nous aurions pu commettre le PÉCHÉ DE PSEUDO-RÉPLICATION et simplement ne pas considérer l'effet nid. Nous aurions alors obtenu les résultats suivants :

```
> anova(aov(parusw~trait,data=dat))
```

```
Analysis of Variance Table
Response: parusw
           Df  Sum Sq Mean Sq F value    Pr(>F)
trait       2  375.94 187.971  9.3786 0.0001468 ***
Residuals 147 2946.27  20.043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'effet traitement serait alors FAUSSEMENT apparu comme très hautement significatif. Il est important, lorsque vous planifiez vos expériences ou analysez les résultats de publications scientifiques, de toujours penser à ce problème de la pseudo-réplication : en multipliant dans l'analyse des observations non-indépendantes (les jeunes d'un même nid dans l'exemple ici), sans en tenir compte dans l'analyse des données, on peut créer des résultats faussement significatifs. L'analyse des degrés de liberté associés au dénominateur du test F permet très souvent de mettre le doigt sur ce type de problème.

3.2.4 Test par permutations pour l'ANOVA hiérarchique

Comme dans le cas de l'ANOVA à un facteur, il est possible d'utiliser la puissance de calcul de l'ordinateur pour simuler la distribution de la statistique F dans le cas de l'ANOVA à un facteur hiérarchique. Il faut par contre prendre garde à permuter le niveau d'observation adéquat. Si le test qui nous intéresse concerne les sous groupes à l'intérieur des groupes, alors on permutera les observations entre sous-groupes, mais en les gardant à l'intérieur de leur groupe respectif. Si le test qui nous intéresse concerne les groupes, les unités d'observations deviennent les

sous-groupes dans leur ensemble, qui seront permutés en entier entre les groupes. Les commandes R suivantes permettent d'effectuer ces tests et de représenter les distributions des deux hypothèses nulles générées par permutations. Nous commençons par analyser classiquement le jeu de données :

```
> set.seed(11)
> m.obs<-anova(aov(parusw~trait/nid,data=dat))
> p.F.trait<-pf(m.obs[1,3]/m.obs[2,3],m.obs[1,1],m.obs[2,1],lower=F)
> p.F.nid<-pf(m.obs[2,3]/m.obs[3,3],m.obs[2,1],m.obs[3,1],lower=F)
```

puis nous stockons les valeurs de F et sommes des carrés correspondant à l'effet traitement et à l'effet nid en dernière ligne de 2 matrices de `nperm` lignes et 2 colonnes (dans l'exemple, `nperm=1000`) :

```
> nperm<-1000
>   #stocke les valeurs de F
> f.star<-matrix(ncol=2,nrow=nperm)
> f.star[nperm,1]<-m.obs[1,3]/m.obs[2,3]
> f.star[nperm,2]<-m.obs[2,3]/m.obs[3,3]
>   #stocke les SS
> ss.star<-matrix(ncol=2,nrow=nperm)
> ss.star[nperm,1]<-m.obs[1,2]
> ss.star[nperm,2]<-m.obs[2,2]
```

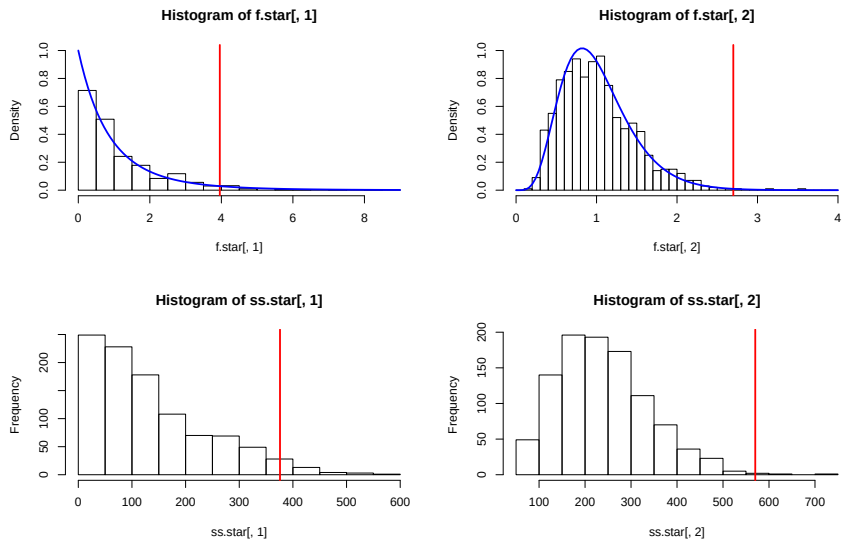
Ensuite, nous allons simplement permutter soit les observations entre nids au sein d'un traitement, soit les nids complets entre traitements, et recalculer les valeurs des statistiques F et SS sur ces données permuttées :

```
>   #nombre de nids affectés à chaque traitement
> trait.s<-rep(c("controle","mouche","puce"),each=5)
> for (i in 1:(nperm-1)){
+   #permutation de nids entre traitements
+ perm.trait<-sample(trait.s)
+   #rep permet de bouger un nid entier entre traitement
+ trait.star<-factor(rep(perm.trait,each=10))
+   #permutations des obs entre nids dans traitements
+ nid.star<-factor(unlist(with(dat,tapply(nid,trait,sample))))
+   #nids complets permuttés entre traitements
+ m1.star<-anova(aov(parusw~trait.star/nid,data=dat))
+   #obs permutés entre nids au sein d'un traitement
+ m2.star<-anova(aov(parusw~trait/nid.star,data=dat))
+
+ f.star[i,1]<-m1.star[1,3]/m1.star[2,3]
+ f.star[i,2]<-m2.star[2,3]/m2.star[3,3]
```

```
+
+ ss.star[i,1]<-m1.star[1,2]
+ ss.star[i,2]<-m2.star[2,2] }
> pss.trait<-sum(ss.star[,1]>=ss.star[nperm,1])/nperm
> pss.nid<-sum(ss.star[,2]>=ss.star[nperm,2])/nperm
> pf.trait<-sum(f.star[,1]>=f.star[nperm,1])/nperm
> pf.nid<-sum(f.star[,2]>=f.star[nperm,2])/nperm
```

Finalement, nous pouvons représenter graphiquement les résultats de ces tests :

```
> par(mfrow=c(2,2))
> hist(f.star[,1],freq=FALSE,breaks=seq(0,9,0.5),ylim=c(0,1))
> abline(v=f.star[nperm,1],lwd=2,col="red")
> curve(df(x,m.obs[1,1],m.obs[2,1]),type="l",add=TRUE,
+   xlim=c(0,9),col="blue",lwd=2)
> hist(f.star[,2],freq=FALSE,breaks=seq(0,4,0.1),ylim=c(0,1))
> abline(v=f.star[nperm,2],lwd=2,col="red")
> curve(df(x,m.obs[2,1],m.obs[3,1]),type="l",add=TRUE,
+  xlim=c(0,4),col="blue",lwd=2)
> hist(ss.star[,1]);abline(v=ss.star[nperm,1],lwd=2,col="red")
> hist(ss.star[,2]);abline(v=ss.star[nperm,2],lwd=2,col="red")
```



Les graphiques sur la première ligne correspondent aux distributions nulles des statistiques F, ceux de la deuxième ligne aux distributions nulles des sommes des carrés SS. Les panneaux de gauche correspondent au test de l'effet traitement, ceux de droite au test de l'effet nid. Sur chacun des panneaux, la droite rouge verticale correspond à la valeur observée de la statistique de test. Notez

bien que les distributions de F générées par permutations ressemblent très fortement aux distributions de F théoriques (les courbes bleues), avec respectivement $2, 12$ et $12, 135$ degrés de liberté [1]. Cette similitude est observée parceque les données expérimentales sont de variances homogènes et ont des résidus distribués normalement.

On peut vérifier que les probabilités associées aux 2 tests sont similaires par la méthode paramétrique et par les méthodes de permutations, soit basée sur la statistique F, soit basée sur les somes des carrés :

```
> c(p.F.trait,pf.trait,pss.trait)

[1] 0.04789789 0.03700000 0.03700000

> c(p.F.nid,pf.nid,pss.nid)

[1] 0.00271494 0.00400000 0.00400000
```

Les résultats entre tests paramétriques et tests de permutations sont quasiment identiques parce que les données proviennent de populations normales et de variances intra-nid homogènes. Si ça n'avait pas été le cas, seuls les tests basés sur les permutations auraient été valides. Notez encore que le test basé sur la statistique F donne la même $p-value$ que le test basé sur les sommes des carrés.

3.2.5 Le niveau niché est-il nécessaire ?

Il est important de garder en tête que le nombre de degrés de liberté du dénominateur du rapport F calculé va influencer grandement la puissance du test que nous effectuons, et plus ce nombre est élevé, plus le test sera puissant. Or, dans une ANOVA hiérarchique, lorsque nous testons pour le facteur A principal (non niché donc), les degrés de liberté du dénominateur du test F sont au nombre de $a(b-1)$. Si la composante de la variance liée à l'effet sous groupe est négligeable, on peut être tenté d'omettre le niveau sous-groupe de l'analyse, afin d'augmenter le nombre de degrés de liberté du dénominateur du test (qui passerait alors à $a(nb-1)$). Qu'est ce qu'on entend par négligeable ? Sokal and Rohlf [1981] (box 10.2), discutent ce point en détail. Une règle simple quoique conservatrice voudrait que si la probabilité associée au test du niveau sous-groupes est importante, supérieure à 50 ou 60%, alors il est conseillé d'omettre le niveau sous-groupe.

1. Un test de Kolmogorov Smirnov permettrait de le vérifier

3.3 ANOVA à 2 facteurs croisés et interaction

Lorsque les données peuvent être agencées sous forme d'un tableau à double entrée (voir début du chapitre), avec un des facteurs en colonnes et l'autre en lignes, nous sommes en présence d'une expérience à analyser à l'aide d'une ANOVA à 2 facteurs croisés. Dans ce modèle, nous serons intéressés non seulement par l'effet de chaque facteur pris isolément, mais aussi par L'INTERACTION entre ces 2 facteurs. Cette interaction apparaîtra comme un nouveau niveau dans la table d'analyse de variance.

3.3.1 Exemples d'analyse à 2 facteurs croisées

- L'exemple classique d'ANOVA à 2 facteurs croisés est celui d'une expérience sur le rendement chez le blé. Plusieurs variétés de blé existent, et on s'intéresse aux rendements de ces différentes variétés. Mais ces variétés sont appelées à être utilisées dans des conditions environnementales différentes, avec des agriculteurs utilisant beaucoup d'engrais et d'autres peu. Il est donc intéressant de mesurer comment réagit chaque variété à des quantités d'engrais différents. Nous souhaiterions en particulier savoir si ces variétés réagissent de la même manière à un apport donné d'engrais.
- Si nous revenons à nos mésanges parasités ou pas, imaginons que nous ayons travaillé sur non pas une, mais plusieurs forêts. Il est possible que les différentes forêts aient un effet sur le poids à l'envol, et que cet effet varie en fonction de la présence ou de l'absence de parasites
- Un agriculteur qui souhaite améliorer son cheptel fera des croisements entre ses animaux. Il verra dans la descendance des effets liés aux mâles, liés aux femelles, mais aussi liés à l'interaction entre les 2 (les effets liés à la dominance), pour autant que chaque femelle ait eu de la descendance avec plusieurs (tous) les mâles. Si chaque femelle n'est présenté qu'à un mâle, mais que chaque mâle couvre plusieurs femelles, nous sommes alors en présence d'une ANOVA hiérarchique, avec le facteur femelle niché dans le facteur mâle.

3.3.2 Modèles et hypothèses testés par le modèle ANOVA à 2 facteurs croisés

Nous allons bien entendu retrouver dans l'analyse de variance à 2 facteurs les différents type d'effets. Ils prennent ici une importance toute particulière puisque le type de test à effectuer va dépendre du type d'effets dans nos modèles. Comme nous avons 2 facteurs, ils peuvent correspondre tous 2 à des effets fixes (modèle I), à des effets aléatoires (modèle II) ou à un effet aléatoire et un effet fixe (modèle mixte, III).

Modèle à effets fixes (I)

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Modèle à effet aléatoire (II)

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$$

Modèle à effet mixte A fixe et B aléatoire (III)

$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \epsilon_{ijk}$$

Le principe de l'analyse est le même que dans le cas de l'ANOVA à 1 facteur et à 2 facteurs hiérarchiques. Il s'agira donc de décomposer en différents éléments la somme des carrés des écarts à la moyenne générale. La seule difficulté supplémentaire vient du fait que les 2 facteurs sont maintenant interchangeables, et donc on peut sommer d'abord sur le facteur A ou d'abord sur le facteur B. L'interaction, elle, se calculera comme :

$$\sum^{a}\sum^{b}\sum^{n}\left((Y_{ij.} - Y_{...}) - (Y_{i..} - Y_{...}) - (Y_{.j.} - Y_{...})\right)^2,$$

soit :

$$n\sum^{a}\sum^{b}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2.$$

Et nous obtenons les expressions suivantes pour les sommes des carrés et carrés moyens :

Source	DDL	SS	MS
A	$(a-1)$	$\frac{\sum^a(\sum^b\sum^n Y_{ijk})^2}{bn} - \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	$\frac{SS_A}{a-1}$
B	$(b-1)$	$\frac{\sum^b(\sum^a\sum^n Y_{ijk})^2}{an} - \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	$\frac{SS_B}{b-1}$
$A \times B$	$(a-1)(b-1)$	$\frac{\sum^a\sum^b(\sum^n Y_{ijk})^2}{n} - \frac{\sum^a(\sum^b\sum^n Y_{ijk})^2}{bn} - \frac{\sum^b(\sum^a\sum^n Y_{ijk})^2}{an} + \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	$\frac{SS_{A\times B}}{(a-1)(b-1)}$
Erreur	$ab(n-1)$	$\sum^a\sum^b\sum^n Y_{ijk}^2 - \frac{\sum^a\sum^b(\sum^n Y_{ijk})^2}{n}$	$\frac{SS_E}{ab(n-1)}$
Total	$abn-1$	$\sum^a\sum^b\sum^n Y_{ijk}^2 - \frac{(\sum^a\sum^b\sum^n Y_{ijk})^2}{abn}$	

La règle pour calculer le rapport F_s s'obtient à partir des carrés moyens attendus, donnés dans le tableau ci-dessous pour les 3 types de modèles possibles lors d'une ANOVA à 2 facteurs croisés :

Carrés moyens	modèle fixe	modèle aléatoire	modèle mixte (A fixe, B aléatoire)
MS_A	$\sigma^2 + \frac{nb}{a-1}\sum^a \alpha^2$	$\sigma^2 + n\sigma^2_{AB} + nb\sigma^2_A$	$\sigma^2 + n\sigma^2_{AB} + \frac{nb}{a-1}\sum^a \alpha^2$
MS_B	$\sigma^2 + \frac{na}{b-1}\sum^b \beta^2$	$\sigma^2 + n\sigma^2_{AB} + na\sigma^2_B$	$\sigma^2 + na\sigma^2_B$
$MS_{A\times B}$	$\sigma^2 + \frac{n}{(a-1)(b-1)}\sum^{a,b}(\alpha\beta)^2$	$\sigma^2 + n\sigma^2_{AB}$	$\sigma^2 + n\sigma^2_{AB}$
MS_E	σ^2	σ^2	σ^2

A partir de ce tableau, il est simple de définir quel test F est adéquat pour quel effet. Nous chercherons toujours à tester si la composante de la variance (modèle aléatoire) ou les écarts à la moyenne générale (effets fixes) diffèrent significativement de 0. Pour cela, il suffit de prendre le rapport des carrés moyens avec le numérateur du rapport correspondant au dénominateur plus l'effet du facteur en question. Nous voyons ainsi que pour le modèle à effets fixes, le carré moyen attendu des effets A et B est égale à la variance résiduelle σ^2 plus l'écart à la moyenne générale des différents niveaux du facteur considéré. Ce carré moyen attendu ne dépend pas de l'écart à la moyenne générale des sous groupes composant l'interaction. Le test F correspondant sera donc le rapport du carré moyen du facteur considéré (A ou B) divisé par le carré moyen de l'erreur.

Par contre, pour le modèle aléatoire, nous voyons que le carré moyen attendu des facteurs A et B dépend non seulement de la composante de la variance résiduelle σ^2 et de la composante de la variance dûe au facteur (σ^2_A ou σ^2_B), mais dépend aussi de la composante de la variance de l'interaction $\sigma^2_{A\times B}$. Le test F adéquat pour les facteurs A et B dans un modèle aléatoire sera donc le rapport entre le carré moyen du facteur considéré et celui de l'interaction. L'ensemble des tests à effectuer pour les différents cas de figure est résumé dans le tableau ci-après :

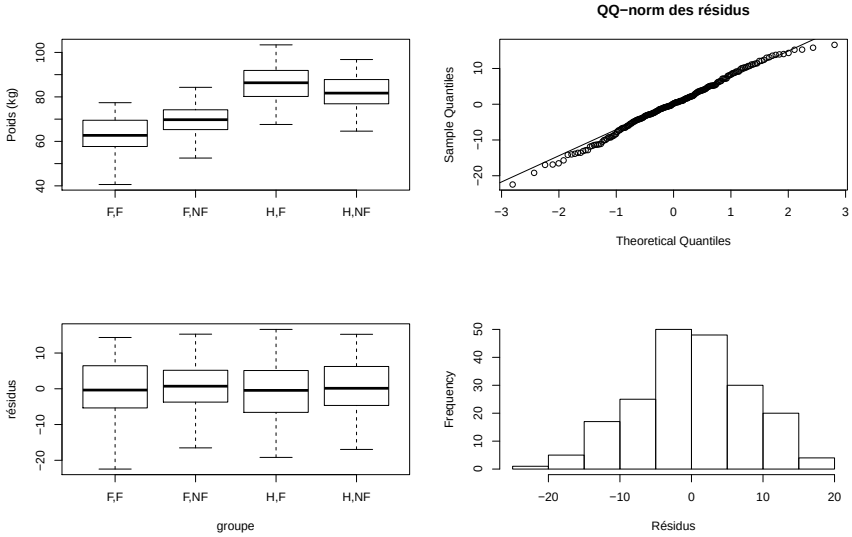
Source	Tests F à effectuer		
	modèle fixe	modèle aléatoire	modèle mixte A fixe, B aléatoire
A	$\frac{MS_A}{MS_E}$	$\frac{MS_A}{MS_{A\times B}}$	$\frac{MS_A}{MS_{A\times B}}$
B	$\frac{MS_B}{MS_E}$	$\frac{MS_B}{MS_{A\times B}}$	$\frac{MS_B}{MS_E}$
$A \times B$	$\frac{MS_{A\times B}}{MS_E}$	$\frac{MS_{A\times B}}{MS_E}$	$\frac{MS_{A\times B}}{MS_E}$

Test de l'interaction : Toujours $F = MS_{A\times B}/MS_E$
- Si A et B Fixes. Test pour A : $F_s = MS_A/MS_E$; Test pour B : $F_s = MS_B/MS_E$
- Si A et B aléatoire. Test pour A : $F_s = MS_A/MS_{A\times B}$; test pour B : $F_s = MS_B/MS_{A\times B}$
- Si A fixe et B aléatoire. Test pour A : $F_s = MS_A/MS_{A\times B}$; test pour B : $F_s = MS_B/MS_E$
- Si A aléatoire et B fixe. Test pour A : $F_s = MS_A/MS_E$; test pour B : $F_s = MS_B/MS_{A\times B}$

3.3.3 Interprétation de l'interaction

Prenons comme exemple de variable réponse le poids des adultes *Homo sapiens* en Suisse. Nous pourrions tenter de l'expliquer en fonction de 2 variables prédictives, d'une part le sexe des individus, et d'autre part s'ils fument ou pas. Clairement tous les individus mesurés pourront être associés avec l'une des 2 catégories de chacun de ces facteurs, et il n'existe pas d'autres catégories de sexe que homme et femme (H/F) ni d'autres catégories que celle des Fumeurs et celle des Non-Fumeurs (F/NF). Nous avons donc affaire à une ANOVA à deux facteurs fixes croisés. Les données et quelques résultats de l'analyse sont représentés ci dessous :

```
> dat<-read.table("./data/poids.txt",header=TRUE)
> gr<-factor(paste(dat$sexe,dat$fumeur,sep=","))
> m1<-aov(poids~sexe*fumeur,data=dat)
> #m1<-aov(poids~sexe+fumeur+sexe:fumeur,data=dat)
> par(mfrow=c(2,2))
> boxplot(poids~gr,data=dat,xlab="",ylab="Poids (kg)")
> qqnorm(m1$resid,main="QQ-norm des résidus");qqline(m1$resid)
> plot(gr,m1$resid,xlab="groupe",ylab="résidus",main="")
> hist(m1$resid,main="",xlab="Résidus")
```



Sur la première ligne de la figure sont donnés les boxplots par sexe et statut fumeur, ainsi que le normal quantile-quantile plot des résidus, qui ne montre pas d'écart à la normalité. Sur la deuxième ligne, le panneau de gauche nous indique que les variances des poids dans les 4 groupes sont à peu près similaires, et celui de droite est une autre manière de présenter la distribution des résidus, et nous montre qu'il n'y a pas d'écart drastique à leur normalité. Nous pouvons donc procéder à l'analyse de variance :

```
> anova(m1)

Analysis of Variance Table
Response: poids
             Df  Sum Sq Mean Sq  F value    Pr(>F)
sexe          1 16449.4 16449.4 266.2411 < 2.2e-16 ***
fumeur        1     6.8     6.8   0.1108    0.7396
sexe:fumeur   1  1569.1  1569.1  25.3970 1.057e-06 ***
Residuals   196 12109.6    61.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'interaction sexe :fumeur est clairement très significative *(Parce que nous sommes dans un modèle à 2 effets fixes, à ce point de l'analyse, les avis divergent sur la suite de l'interprétation. Certains statisticiens pensent que l'analyse de la table d'ANOVA lorsque les 2 effets sont fixes doit s'arrêter là. En effet, puisqu'il y'a interaction entre les facteurs "fumeur" et "sexe", il est difficile de statuer sur l'un sans spécifier pour quelle valeur de l'autre facteur. D'autres pensent qu'il est quand même possible d'interpréter les résultats pour les effets principaux)* de même que l'effet du sexe, par contre il n'y a pas d'effet du statut fumeur/non fumeur sur le poids. L'observation des boxplots nous montre que les femmes qui fument ont tendance à être plus légères que celles qui ne fument pas, alors que les hommes qui fument ont tendance à être plus lourds que

ceux qui ne fument pas. L'interaction, c'est exactement cela : l'effet des niveaux d'un facteur varie en fonction des niveaux de l'autre facteur. Dans notre cas de figure, cette interaction significative pourrait s'expliquer en supposant que les femmes qui fument utilisent le tabac comme coupe-faim, et sont donc attentives à leur poids, alors que les hommes fumeurs sont souvent des personnes qui ne prêtent que peu de cas à leur apparence et sont peu sportifs. On notera aussi sur le box-plot que les hommes, qu'ils fument ou qu'ils ne fument pas, ont un poids supérieur à celui des femmes, qu'elles fument ou qu'elles ne fument pas. La table d'ANOVA est claire aussi à ce propos : alors que MS de l'effet fumeur est négligeable (6.8), celui de l'effet sexe (16449.4) est très important. Dans notre cas de figure, nous pouvons donc sans crainte interpréter les résultats de l'effet principal lié au sexe : les hommes sont plus lourds que les femmes, qu'ils (qu'elles) fument ou qu'ils (qu'elles) ne fument pas.

D'une manière plus générale, une interaction significative signifie que les 2 facteurs ne sont pas indépendants, et donc que l'interprétation d'un des 2 facteurs ne peut pas se faire sans préciser quel niveau de l'autre facteur est considéré.

3.3.4 ANOVA à 2 facteurs croisés sans réplication

Lorsque nous n'avons qu'une observation pour chaque cellule de l'analyse à 2 facteurs croisés (soit une observation pour chaque combinaison des 2 facteurs), nous ne pouvons bien évidemment pas tester pour l'effet de l'interaction, puisque on ne peut pas calculer de variance sur la base d'une observation (il n'y aurait plus de degrés de libertés pour la ligne Résidus de la table d'ANOVA). Pour le modèle aléatoire, ça ne pose pas de problème particulier puisque les effets principaux sont testés par rapport à l'interaction. Pour les modèles fixes et mixtes par contre, comme il n'y a plus de degrés de libertés pour les résidus du modèle, on ne peut pas tester les effets fixes. Mais si on considère que les 2 facteurs sont indépendants (donc qu'il n'y a pas d'interactions, une hypothèse forte), il est possible de tester pour les effets principaux en utilisant l'interaction comme erreur. Ceci revient à spécifier un modèle où on omettrait l'interaction (`aov(resp~a+b)` au lieu de `aov(resp~a*b)`).

3.3.5 L'interaction est-elle nécessaire ?

Même dans les modèles avec réplication dans chaque cellule (soit une combinaison des différents facteurs), on peut se demander si il est nécessaire de garder l'interaction dans le modèle quand cette dernière n'est pas significative. En effet, que ce soit dans le modèle fixe ou dans le modèle aléatoire, supprimer l'interaction du modèle aura pour effet d'augmenter les degrés de liberté du dénominateur pour les 2 tests effectués. Si l'effet de l'interaction est loin d'être significatif (p-value plus grande que 60% par exemple), on pourra refaire une analyse sans le

terme d'interaction, en gagnant des degrés de libertés au dénominateur et donc en augmentant la puissance.

3.3.6 Plan expérimental non équilibré : Somme des carrés de type I, II et III

Dans l'analyse de variance à 2 facteurs croisés, si le nombre d'observations par cellule (soit une combinaison des différents facteurs) n'est pas égal, l'ordre dans lequel les 2 facteurs sont rentrés devient important. Si les sommes des carrés sont calculées en suivant l'ordre d'entrée des facteurs dans le modèle, on a affaire à des sommes des carrés appelés de type I. Les sommes des carrés de type II sont obtenus en comparant un modèle avec les deux effets principaux avec un modèle avec l'effet qui nous intéresse retiré. Enfin une somme des carrés de type III s'obtient en retirant au modèle complet (c'est à dire avec l'interaction) le facteur dont on essaye d'estimer la somme des carrés. Un exemple artificiel devrait clarifier les choses.

```
> set.seed(27)
> data<-rnorm(36)
> A<-factor(rep(c("A","B"),each=18)) #2 groupes de 18 observations
> B<-factor(rep(c("C","D"),18)) #idem, mais pas dans le même ordre
> anova(aov(data~A*B))

Analysis of Variance Table
Response: data
          Df Sum Sq Mean Sq F value Pr(>F)
A          1  0.555 0.55496  0.3845 0.5396
B          1  0.089 0.08916  0.0618 0.8053
A:B        1  0.006 0.00570  0.0040 0.9503
Residuals 32 46.191 1.44347

> anova(aov(data~B*A))

Analysis of Variance Table
Response: data
          Df Sum Sq Mean Sq F value Pr(>F)
B          1  0.089 0.08916  0.0618 0.8053
A          1  0.555 0.55496  0.3845 0.5396
B:A        1  0.006 0.00570  0.0040 0.9503
Residuals 32 46.191 1.44347
```

Que nous entrions le facteur A ou le facteur B d'abord ne change rien à la table d'ANOVA. Les choses seraient différentes si le nombre d'observations dans chaque cellule (9 ici) était différents. On peut le vérifier en supprimant quelques données :

```
> data[A=="A" & B=="C"][3:9]<-NA
> data[A=="B" & B=="D"][1:6]<-NA
> data[A=="A" & B=="C"]

[1]  1.9071626 -0.7645307         NA         NA         NA         NA         NA
[8]         NA         NA

> data[A=="B" & B=="C"]

[1] -0.02583971 -0.57488122  0.08706853 -2.99830401 -0.99707477  1.29534374
[7] -0.15204521  1.85821984 -0.39814797

> data[A=="A" & B=="D"]

[1]  1.14487689 -1.45743250  0.29524122  1.15741089  0.23784461  0.03482725
[7]  0.15801005 -1.06880297 -1.06858164

> data[A=="B" & B=="D"]

[1]         NA         NA         NA         NA         NA         NA -0.9589459
[8] -1.5747647  0.9686850

> (mab<-anova(aov(data~A*B)))

Analysis of Variance Table
Response: data
          Df  Sum Sq Mean Sq F value Pr(>F)
A          1  0.6697 0.66966  0.4311 0.5193
B          1  0.7748 0.77479  0.4988 0.4886
A:B        1  0.0997 0.09966  0.0642 0.8028
Residuals 19 29.5150 1.55342

> (mba<-anova(aov(data~B*A)))

Analysis of Variance Table
Response: data
          Df  Sum Sq Mean Sq F value Pr(>F)
B          1  0.0673 0.06728  0.0433 0.8374
A          1  1.3772 1.37717  0.8865 0.3582
B:A        1  0.0997 0.09966  0.0642 0.8028
Residuals 19 29.5150 1.55342
```

Et nous voyons clairement une différence entre les 2 modèles `mab` et `mba`. Ces 2 modèles nous donnent des sommes des carrés de type I.

Pour obtenir des sommes de carrés de type II, nous allons comparer le modèle sans interactions avec les 2 facteurs, à un modèle auquel nous aurons retiré le facteur d'intérêt. La fonction `anova` de R permet aussi de faire cette comparaison :

```
> m.noi.f<-aov(data~A+B)
> m.noi.A<-aov(data~A)
> m.noi.B<-aov(data~B)
> anova(m.noi.f,m.noi.A) #donne SS type II pour B

Analysis of Variance Table
Model 1: data ~ A + B
Model 2: data ~ A
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 29.615
2     21 30.390 -1  -0.77479 0.5232 0.4778

> anova(m.noi.f,m.noi.B) #donne SS type II pour A

Analysis of Variance Table
Model 1: data ~ A + B
Model 2: data ~ B
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 29.615
2     21 30.992 -1   -1.3772 0.9301 0.3464
```

Et nous voyons que la somme des carrés de type II correspond à la somme des carrés de type I quand le facteur d'intérêt est entré en 2ème position dans le modèle.

Une commande de R spécifique pour extraire ces sommes des carrés de type II et III est la commande `drop1`.

Notez encore que la plupart des auteurs s'accordent à dire que les sommes des carrés de type II sont souvent les plus adéquates. L'intérêt des sommes des carrés de type III reste un sujet d'âpres discussions entre spécialistes.

Ce problème d'ordre d'entrée des termes dans le modèles se retrouvera aussi dans le cadre de la régression multiple, et à nouveau les sommes des carrés de type II seront souvent plus adéquates que les sommes des carrés de type I (car on s'intéresse à l'effet marginale des variables explicatives, les autres sources de variation ayant déjà été prises en compte)

Ces complications liées à un plan expérimental non équilibré sont une raison de plus pour éviter au maximum ce type de plans expérimentaux. Notez enfn que pour les données très déséquilibrées incluant des facteurs fixes et aléatoires, on fera appel à d'autres méthodes que les moindres carrés pour estimer les paramètres du modèle. Les **modèles linéaires mixtes** sont l'objet de recherches actives aujourd'hui, un package R très utilisé dans ce cadre est le package *lme4*.

Chapitre 4

Analyse de covariance (ANCOVA)

4.1 Introduction

Nous sommes souvent confrontés à un problème du type : est ce que les groupes $a_1, a_2, \cdots, a_k$ diffèrent, une fois pris en compte une covariable B ? Lorsque la covariable est un facteur, c'est un cas typique d'analyse de variance à 2 facteurs, qui peuvent d'ailleurs être croisés ou hiérarchiques. C'est typiquement le cas de figure lorsque nous nous intéressons à l'effet de la présence ou pas de parasites sur le poids des oisillons à l'envol, en provenance de différents nids. Nous suspectons que les différents nids pourraient avoir un effet sur le poids des poussins, mais cet effet ne nous intéresse pas particulièrement, c'est plutôt un facteur de nuisance. Ce que nous souhaiterions, c'est pouvoir mesurer la différence de poids entre poussins issus de nids parasités ou non, indépendamment de leur nid de provenance. Nous ferions donc une analyse de variance hiérarchique, avec l'effet nid niché (dans l'effet traitement).

Souvent la covariable n'est pas factorielle, mais est continue. Dans l'exemple ci-dessus, nous pourrions penser qu'un des facteurs qui affecte le poids des poussins à l'envol est leur poids à la naissance. Un poussin lourd à la naissance sera certainement lourd à l'envol, quel que soit le traitement subi par son nid. Ce type de problèmes, où il existe un mélange de variables explicatives continues et factorielles s'analysent grâce à l'analyse de covariance ou ANCOVA (car ANalyse de COVAriance). Nous nous limiterons ici au cas où il y a 2 variables explicatives, une continue et une factorielle, mais le modèle s'étend facilement à plusieurs variables explicatives catégorielles et continues.

Le principe de l'analyse de l'ANCOVA est de tester pour une différence de moyenne entre groupes définis par la variable explicative factorielle (et donc en cela similaire à une ANOVA), UNE FOIS QUE CES MOYENNES ONT ÉTÉ AJUSTÉES pour des différences dues à la variable explicative continue, souvent appelée

covariable. Cet ajustement se fait à l'aide de régression linéaire. L'ANCOVA est donc une analyse hybride entre l'analyse de variance et la régression linéaire.

4.2 Exemples d'ANCOVA

- Les poids moyens des hommes et des femmes diffèrent, c'est connu. Mais hommes et femmes diffèrent pour d'autres caractères, comme la taille. Or taille et poids sont corrélés. Est-ce que, à taille égale, le poids des hommes et des femmes est différent ? L'ANCOVA nous permettra de répondre à cette question.
- On peut s'intéresser à l'effet de différents régimes alimentaires sur la taille des escargots à leur maturité sexuelle. Mais il est vraisemblable que la taille à l'éclosion joue aussi un rôle. On se posera donc la question : est ce que, à taille à l'éclosion égale, la taille à maturité est la même pour tout les régimes alimentaires.
- On peut s'intéresser à l'effet de l'EPO sur la capacité respiratoire d'une personne. Mais l'altitude à laquelle réside la personne aura vraisemblablement aussi un effet sur sa capacité respiratoire. On mesurera donc l'effet de l'EPO sur la capacité respiratoire après avoir pris en compte celle de l'altitude, en utilisant une ANCOVA.

4.3 Le modèle de l'ANCOVA

Le modèle de l'ANCOVA se présente comme suit :

$$Y_{ij} = \mu + \alpha_i + \beta_{\text{within}}(X_{ij} - \overline{X_i}) + \epsilon_{ij}$$

Y_{ij} représente l'observation j du $i$ème groupe, μ est la grande moyenne de la population, α_i est l'effet fixe pour le groupe i, $\beta_{\text{within}}(X_{ij}-\overline{X_i})$ est l'effet expliqué par la différence entre la covariable X_{ij} à sa moyenne $\overline{X_i}$ sur les j observations du groupe i (β_{within} est donc la pente commune de la régression de Y_{ij} sur les X_{ij} aux différents niveaux de l'effet fixe), enfin les ϵ_{ij} sont les résidus. L'ajustement pour la covariable se voit très bien si nous réecrivons le modèle comme suit :

$$Y_{ij} - \beta_{\text{within}}(X_{ij} - \overline{X_i}) = \mu + \alpha_i + \epsilon_{ij}$$

Le modèle de l'ANCOVA est donc bien un modèle d'ANOVA (la partie droite du modèle ci dessus), mais ou la valeur de la variable réponse est ajustée pour sa relation avec la covariable X. Il est important de rappeler ici que si la relation entre la variable réponse et la covariable n'est pas forte, le modèle d'ANCOVA n'est pas approprié.

Notez bien aussi que l'interaction ne fait pas partie du modèle. La première étape dans une ANCOVA (après avoir bien entendu vérifié les conditions d'applications, voir ci-dessous) sera donc de tester si l'interaction entre variable explicative et covariable est significative. C'est seulement dans le cas contraire que nous pourrons appliquer le modèle ci-dessus, ou nous forçons la même pente pour tous les niveaux de la variable explicative. Deux hypothèses nulles seront alors testées :

$H0^{Cov}$:la variable réponse est indépendante de la covariable (la prise en compte de la covariable ne permet pas d'expliquer la variance de la variable réponse). Notez que si $H0^{Cov}$ est vrai, il n'y a pas lieu de faire une analyse de covariance !

$H0^{A}$: les moyennes des différents niveaux du facteur A, ajustées pour l'effet de la covariable Cov, sont identiques. Ceci revient à dire que les droites de régression des différents niveaux du facteur, outre le fait qu'elles ont une même pente, ont aussi une même ordonnée à l'origine (intercept en anglais). C'est bien entendu cette hypothèse qui est la plus importante.

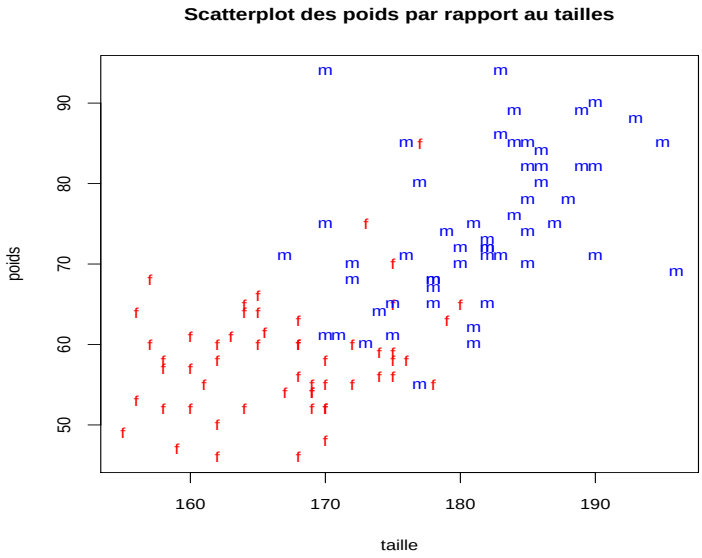
4.4 Conditions d'application de l'ANCOVA

L'ANCOVA étant hybride entre l'ANOVA et la régression, il faudra s'assurer que les conditions nécessaires pour effectuer une ANOVA et celles nécessaires pour effectuer une régression sont remplies (au fait, quelles sont ces conditions ?). On fera particulièrement attention à la distribution des résidus en fonction des valeurs ajustées (ou des valeurs de la covariable). Ces 2 types de graphiques sont souvent extrêmement informatif, nous aurons l'occasion d'y revenir dans le cadre des exercices.

4.5 L'ANCOVA dans R

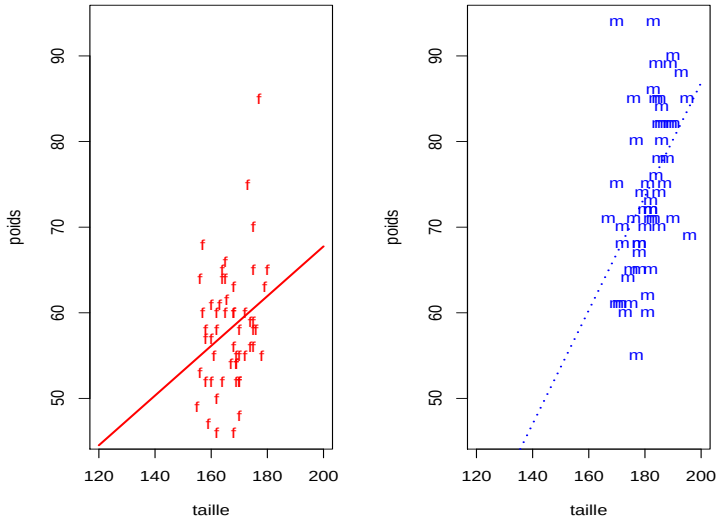
4.5.1 Le poids des hommes et des femmes

Nous allons à nouveau nous intéresser au poids des étudiants. Nous avons vu précédemment que le poids moyen des garçons est différent de celui des filles. Est ce que ce résultat pourrait être dû à une différence de taille ? Un "scatter-plot" des données nous permettra de nous en rendre compte.



De cette figure, il est clair que les garçons sont en général plus lourd, mais aussi plus grand que les filles. Cette différence de poids pourrait donc être simplement du a une différence de taille. Comment vérifier que ce n'est pas le cas ?

Nous pourrions effectuer une régression dans chacun des groupes (garçons / filles) et comparer les pentes et ordonnés à l'origine des 2 régressions.




```
> summary(reg.f)

Call: lm(formula = pf ~ tf, data = data)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6696    23.2593   0.416   0.6793
tf            0.2904     0.1392   2.086   0.0418 *

Residual standard error: 6.846 on 53 degrees of freedom
Multiple R-Squared:0.07588,    Adjusted R-squared: 0.05845
F-statistic: 4.352 on 1 and 53 DF,p-value: 0.04179
```

```
> summary(reg.m)
Call: lm(formula = pm ~ tm, data = data) Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.9855    31.9666  -1.439 0.156270
tm            0.6643     0.1762   3.770 0.000419 ***

Residual standard error: 8.533 on 52 degrees of freedom
Multiple R-Squared:0.2146,     Adjusted R-squared: 0.1995
F-statistic: 14.21 on 1 and 52 DF, p-value: 0.0004188
```

Les pentes des 2 régression sont-elles les mêmes ? C'est la qu'intervient l'ANCOVA. Plutôt que de régresser les tailles par groupe indépendamment l'une de l'autre, nous allons les comparer dans une seule et même analyse, l'analyse de covariance. La variable réponse est le poids, les 2 variables explicatives sont la taille, covariable continue, et le sexe, qui est factoriel à 2 catégories. Les 2 variables explicatives peuvent très bien avoir une interaction, soit que la pente de la régression pour un des niveaux du facteur diffère de l'autre.

Nous allons donc avant tout tester cette hypothèse $H_0^{A:X}$ =Les pentes des régressions de la variable réponse sur la covariable sont les mêmes pour les différents niveaux du facteur.

```
>ancov1<-lm(poids~taille*factor(sexe), data=data)

> anova(ancov1)
Analysis of Variance Table

Response: poids
                     Df Sum Sq Mean Sq  F value    Pr(>F)
taille                1 7180.3  7180.3 120.2367 < 2.2e-16 ***
factor(sexe)          1 1122.6  1122.6  18.7979 3.344e-05 ***
taille:factor(sexe)   1  166.4   166.4   2.7861   0.09807 .
Residuals           105 6270.4    59.7
```

Dans cette analyse, nous voyons que l'interaction n'est pas significative ($F = 2.7861, p = 0.098$). Nous ne pouvons pas rejeter l'hypothèse nulle que les pentes des régressions sont les mêmes, et donc nous l'acceptons. Nous pouvons dès lors continuer à interpréter les résultats. Notons que si l'interaction avait été significative, nous nous serions arrêtés là dans l'interprétation, et aurions conclus que les pentes de régression du poids sur la taille des hommes et des femmes diffèrent, sans pouvoir comparer les poids à tailles égales.

Comme l'interaction n'est pas significative, nous pouvons réajuster un modèle sans ce terme, afin de tester a/ si la pente commune des régressions est différente de 0 (on le suppose, sinon il n'y aurait même pas eu lieu de faire l'ANCOVA) et surtout b/ si les ordonnées à l'origine des régressions pour les différentes niveaux du facteur sont les mêmes. Cette dernière hypothèse revient à tester si, à taille égale, les hommes et les femmes ont le même poids.

$H_0^{X(\text{taille})}$: La pente commune de régression n'est pas différente de 0

$H_0^{A(\text{sexe})}$: Les ordonnées à l'origine des différentes niveau du facteur sont identiques.

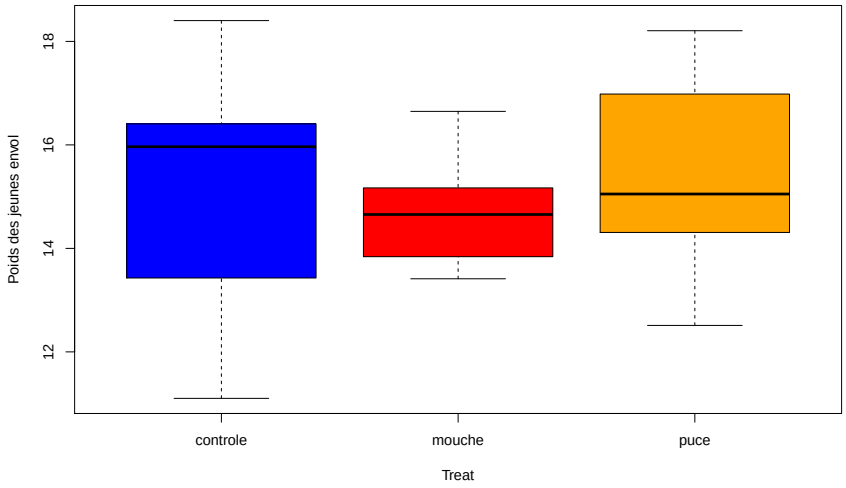
```
>ancov2<-lm(poids~taille+factor(sexe),data=data)

> anova(ancov2)
Analysis of Variance Table

Response: poids
              Df Sum Sq Mean Sq F value    Pr(>F)
taille         1 7180.3  7180.3 118.244 < 2.2e-16 ***
factor(sexe)   1 1122.6  1122.6  18.486 3.815e-05 ***
Residuals    106 6436.8    60.7
```

Il est clair que l'analyse rejette fortement ces 2 hypothèses : le poids dépend bien de la taille des individus (la pente de la régression est différente de 0), et surtout, les ordonnées à l'origine sont différentes. Nous pouvons donc conclure qu'a taille égale, le poids des hommes est différent de celui des femmes

4.5.2 Un exemple avec *Parus major*

Le jeu de données du poids des jeunes à l'envol en fonction des 3 traitements se trouve dans le fichier ancovmes.txt. Le box-plot suivant donne le poids des jeunes à l'envol en fonction du traitement qu'ils ont subi :

```
> dat.anc<-read.table("./data/ancovmes.txt",header=T)
> with(dat.anc,boxplot(j.env~treat, col=c("blue","red","orange"),
+ xlab="Treat",ylab="Poids des jeunes envol"))
```

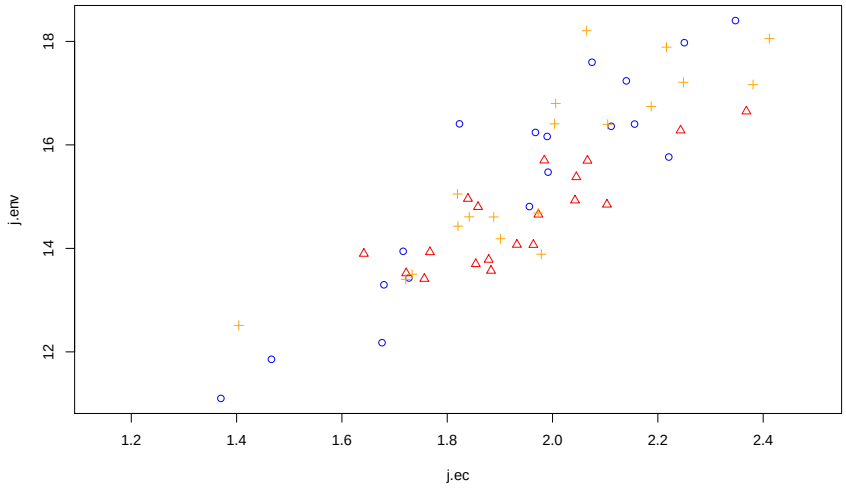


Il ne semble pas y avoir une énorme différence entre les groupes, et cette impression est confirmée par une analyse de variance à un facteur du poids à l'envol en fonction du traitement :

```
> anova(lm(j.env~treat,data=dat.anc))

Analysis of Variance Table
Response: j.env
          Df Sum Sq Mean Sq F value Pr(>F)
treat      2   8.75  4.3749   1.522 0.2277
Residuals 53 152.34  2.8744
```

Nous concluons donc à l'absence d'effet significatif des traitements. Cependant, les expérimentateurs ont relevé aussi le poids des jeunes à l'éclosion, et la relation entre le poids des jeunes à l'éclosion et le poids des jeunes à l'envol est très claire :

```
> coul<-rep(c("blue","red","orange"),table(dat.anc$treat))
> with(dat.anc,plot(j.env~j.ec,col=coul,pch=as.integer(treat)))
```



La taille à l'éclosion explique donc une part non négligeable de la variance du poids des jeunes à l'envol, et devrait peut-être permettre d'affiner notre première analyse. Nous pourrions effectuer la régression du poids à l'envol sur le poids à l'éclosion, et tester pour l'effet des traitements sur les résidus :

```
> m.reg<-lm(j.env~j.ec,data=dat.anc)
> anova(lm(m.reg$residuals~dat.anc$treat[!is.na(dat.anc$j.env)]))

Analysis of Variance Table
Response: m.reg$residuals
                                    Df Sum Sq Mean Sq F value   Pr(>F)
dat.anc$treat[!is.na(dat.anc$j.env)]  2  6.392  3.1960  5.1821 0.008799 **
Residuals                            53 32.687  0.6167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d'après vous que signifie `[!is.na(dat.anc$j.env)]` ? jetez un coup d'oeil au fichier de données, ça devrait vous mettre sur la piste)

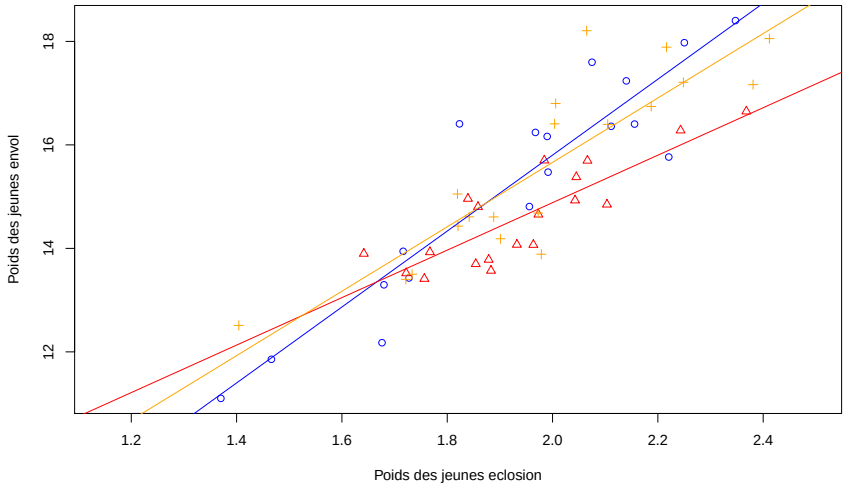
De cette analyse, il ressort clairement qu'une fois pris en compte l'effet de le poids à l'éclosion, le traitement à un effet sur le poids à l'envol. Cependant, il se pourrait que la relation entre poids à l'envol et poids à l'éclosion diffèrent entre les traitements. Pour s'en assurer, nous pourrions commencer par comparer les pentes des régressions à l'intérieur de chacun des groupes. Les résultats sont présentées ci-dessous :

```
> l.dat.anc<-split(dat.anc,dat.anc$treat)
> (coeff.anc<-lapply(l.dat.anc,fun<-function(x)
+ summary(lm(x$j.env~x$j.ec))$coeff))
```

```
$controle
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.127104  1.4609957 0.7714626 4.516767e-01
x$j.ec      7.337493  0.7515128 9.7636293 3.827642e-08
$mouche
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 5.708376  1.3843384 4.123541 7.096592e-04
x$j.ec      4.587562  0.7095233 6.465696 5.811894e-06

$puce
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.214004  1.6920920 1.899426 7.461142e-02
x$j.ec      6.223787  0.8465284 7.352130 1.130043e-06
```

Les intervalles de confiance à 95% des 3 pentes ne sont pas disjoints ($7.3 \pm 1.52, 4.6 \pm 1.42, 6.2 \pm 1.68$).

```
> with(dat.anc,plot(j.ec,j.env,pch=as.integer(treat),col=coul,
+ xlab="Poids des jeunes eclosion",ylab="Poids des jeunes envol"))
> l.coul<-c("blue","red","orange")
> for (i in 1:3) abline(coeff.anc[[i]][,1],col=l.coul[i])
```



L'ANCOVA avec interaction va nous permettre de tester si les pentes sont les mêmes ($H_0^{treat:pj.ec}$: les pentes des régressions du poids des jeunes à l'envol sur leur poids à l'éclosion sont les mêmes entre les 3 traitements) :

```
> anova(lm(j.env~j.ec*treat,data=dat.anc))
```

```
Analysis of Variance Table
Response: j.env
           Df  Sum Sq Mean Sq  F value  Pr(>F)
j.ec        1 122.013 122.013 205.7714 < 2e-16 ***
treat       2   6.394   3.197   5.3916 0.00758 **
j.ec:treat  2   3.038   1.519   2.5616 0.08728 .
Residuals  50  29.648   0.593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La probabilité associée à l'interaction est de 0.09, donc non-significative au seuil de 5%. Nous ne pouvons pas rejeter l'hypothèse que les pentes sont égales, donc nous l'acceptons.

L'analyse du modèle sans l'interaction nous donne alors :

```
> anova(mod.anco<-lm(j.env~j.ec+treat,data=dat.anc))

Analysis of Variance Table
Response: j.env
          Df  Sum Sq Mean Sq  F value    Pr(>F)
j.ec       1 122.013 122.013 194.1130 < 2.2e-16 ***
treat      2   6.394   3.197   5.0862  0.009606 **
Residuals 52  32.686   0.629
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Et nous en concluons qu'indépendamment d'un effet très significatif du poids à l'éclosion sur le poids à l'envol, il y a aussi un effet significatif du traitement.

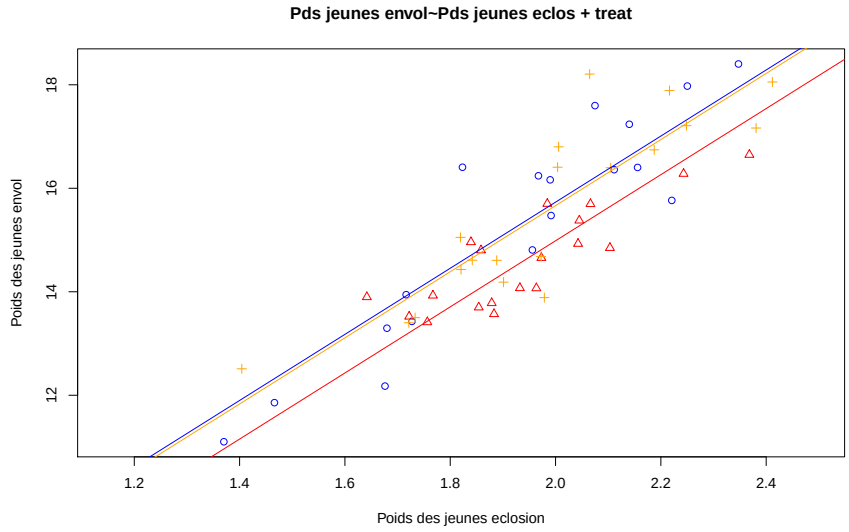
A l'aide de la commande summary, nous obtenons les sorties suivantes :

```
> summary(mod.anco)

Call:
lm(formula = j.env ~ j.ec + treat, data = dat.anc)
Residuals:
    Min      1Q  Median      3Q     Max
-1.6404 -0.4819 -0.1030  0.5575  2.1286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95754    0.91086   3.247  0.00204 **
j.ec         6.38701    0.46292  13.797  < 2e-16 ***
treatmouche -0.74608    0.26090  -2.860  0.00609 **
treatpuce   -0.06745    0.26219  -0.257  0.79798
```

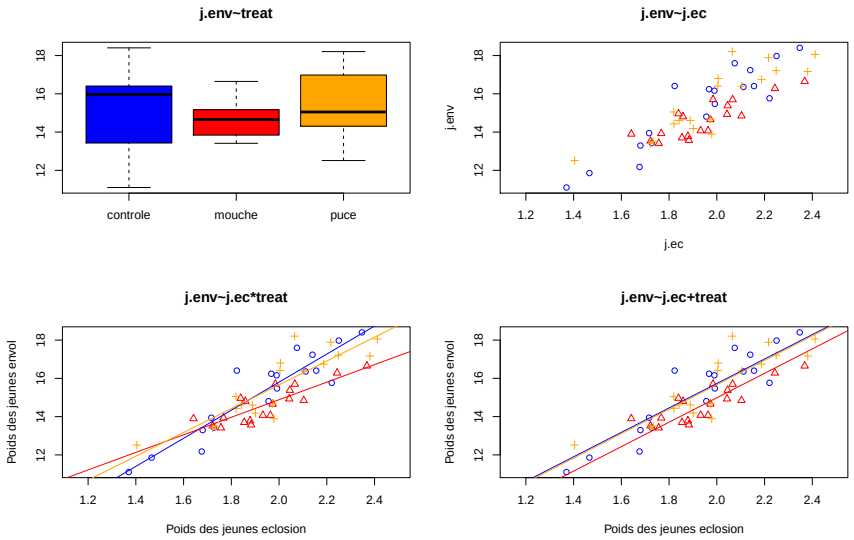
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7928 on 52 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.7971,        Adjusted R-squared: 0.7854
F-statistic:  68.1 on 3 and 52 DF,  p-value: < 2.2e-16
```

L'ordonnée à l'origine (OAO) du groupe contrôle vaut 2.958, et diffère significativement de 0. La pente de la régression générale est de 6.39, elle aussi très significativement différente de 0. Le groupe traité avec des mouches diffère significativement du groupe contrôle, la différence d'OAO entre les 2 étant de -0.746 ce qui signifie que les nids traités avec des mouches produisent des jeunes qui, à poids à l'éclosion égal, pèsent 0.746 grammes de moins que ceux du groupe contrôle. Le poids moyen des jeunes à l'envol des oisillons du groupe puce par contre ne diffère pas significativement de celui du groupe contrôle, toujours en contrôlant pour le poids à l'éclosion. En représentation graphique, nous obtenons :

```
> coeff.anco<-summary(mod.anco)$coeff
> with(dat.anc,plot(j.ec,j.env,pch=as.integer(treat),col=coul,
+ xlab="Poids des jeunes eclosion",ylab="Poids des jeunes envol"))
> l.coul<-c("blue","red","orange")
> abline(coeff.anco[1:2,1],col=l.coul[1])
> for (i in 3:4) abline(c(coeff.anco[1,1]+coeff.anco[i,1],
+ coeff.anco[2,1]),col=l.coul[i-1])
> title("Pds jeunes envol~Pds jeunes eclos + treat")
```



Finalement, nous pouvons résumer les diférentes étapes de l'analyse à l'aide de la figure suivante :


j.env~treat
12
14
16
18
controle
mouche
puce
j.env~j.ec
j.env
j.ec
1.2
1.4
1.6
1.8
2.0
2.2
2.4
j.env~j.ec*treat
Poids des jeunes envol
Poids des jeunes eclosion
j.env~j.ec+treat
Poids des jeunes envol
Poids des jeunes eclosion

Annexe A

Quelle méthode quand ?

Ce chapitre devrait vous servir comme guide dans le choix du test statistique à appliquer. Le principe du choix du test repose sur une série de règles simples : quel est le type de la variable réponse ? (quantitative et continue ? quantitative et discontinue ? ordonnée ? qualitative ?). Combien de variables explicatives (prédictives) ? de quel(s) type(s) ? Une fois que vous aurez répondu à ces trois questions, vous aurez (quasiment) cerné quel test devra être appliqué.

A.1 Une variable réponse quantitative continue

Il s'agit de tout type de variable qui peut prendre pour valeur un nombre réel quelconque. La taille, le poids sont des variables continues. Une variable ne pouvant prendre que des valeurs entières ne rentre normalement pas dans cette catégorie, mais si le nombre de valeurs possibles est suffisamment grand, on considérera que la variable est quasi-continue. C'est par exemple le cas du nombre d'oeuf pondu par un escargot (les pontes sont faites de quelques dizaines à quelques centaines d'oeufs) ou le nombre de graines produites par une fleur. Le nombre de poussins produit par un couple de mésange ne rentre par contre pas dans cette catégorie. Certaines variables réponse, comme une proportion, ne peuvent pas prendre une valeur réelle quelconque (une proportion est nécessairement comprise entre 0 et 1), mais elles peuvent souvent être transformées pour satisfaire à la définition (pour une proportion par exemple, on utilisera la transformation arc-tangente). Les variables réponses de ce type sont analysées grâce au **modèle linéaire** (`lm` dans R), cas particulier du modèle linéaire généralisé. Les différents cas de figure ci-dessous ne sont que des variations de ce modèle linéaire.

1. **Un échantillon**. Dans cette situation, plusieurs questions se posent :
   - Estimation des moments de la distribution de la population d'où provient l'échantillon (moyenne, écart-type, erreur-type) ; intervalle de confiance de la moyenne

- Test si la moyenne est égale à une valeur théorique : **test** t (`t.test` dans R) à un échantillon
- Test d'adéquation à une loi, c'est-à-dire, test si la distribution empirique est identique à une distribution théorique (par exemple, normale, poisson...) : **test de Kolmogorov-Smirnov** (`ks.test` dans R ; NOTE : si la distribution théorique que vous cherchez à ajuster est une normale de moyenne et variance inconnue, le test adéquat est le test de Shapiro `shapiro.test`), ou test de χ^2 (`chisq.test` dans R) . Pour ce dernier cas et si la distribution théorique est continue (par exemple normale) il sera nécessaire de définir des classes. D'autre part, il faut être prudent dans l'interprétation d'un test de χ^2 quand les attendus sont inférieurs à 5.

2. **Deux échantillons**. Ce cas de figure est équivalent à une variable explicative factorielle à 2 niveaux. Nous pourrions donc comparer les moyennes des 2 groupes grâce à une ANOVA. Il existe cependant des tests spécifiques pour la comparaison de 2 échantillons qu'il est nécessaire de connaître.
   - Comparaison de moyennes. **test** t (`t.test` dans R) si applicable (normalité des résidus). Si variance des deux groupes non homogènes (test F, `var.test` dans R), appliquer la correction de Welsh. Si résidus non normalement distribués, test non paramétrique de Wilcoxon (`wilcox.test` dans R, aussi connu sous le nom de Mann Whitney). *NOTE :* les données peuvent être appariées (pour chaque unité d'observation il y a 2 mesures). Dans ce cas, on effectue un test t ou Wilcoxon apparié, ce qui revient à tester si la différence entre les 2 mesures est différente de 0 (et donc à un test de t à un échantillon).
   - Comparaison de Variance. **test** F.
   - Comparaison de distributions empiriques. **Test de Kolmogorov-Smirnov** pour 2 échantillons (`ks.test` dans R).

   Pour les tests non appariés, la représentation des données la plus adéquate est sous la forme d'une colonne pour la variable réponse, et d'une autre pour une variable qualitative/catégorielle désignant le groupe auquel appartient l'observation.

3. **Une Variable explicative factorielle à plus de 2 niveaux**. Le test paramétrique adéquat est **l'analyse de variance (ANOVA)**. Si les variances ne sont pas homogènes (impression visuelle suffit souvent via boxplot des résidus, test à l'aide du test de Levene) ou les résidus ne sont pas distribués normalement (`qqnorm`, `shapiro.test`), on effectuera un test de Kruskal-Wallis (`kruskal.test`), ou bien on cherchera une transformation de la variable réponse qui normalise les résidus et homogénéise les variances, ou, mieux encore, un test par permutations. Rappelons que la condition la plus importante est l'homogénéité des variances. Un faible écart à la normalité n'est pas forcément rédhibitoire, puisque si les échantillons sont grands, la distribution de la moyenne des observations tendra vers une distribution

normale (théorème central limite).

4. **Une variable explicative quantitative**.
   - Nous utiliserons la **régression** si nous cherchons à prédire les valeurs de la variable réponse en fonction de celles de la variable explicative. Avant de valider le modèle de régression, il faudra décider quelle type de relation est attendue entre les 2 variables (linéaire ? quadratique ? logarithmique ?). Pour ce faire, rien ne remplace l'inspection visuelle des données (`plot` de la variable réponse en fonction de la variable explicative). En fonction de la forme de la relation, soit on transformera l'une des 2 variables, soit on ajustera une régression avec un terme quadratique ou logarithmique. On s'assurera aussi que les résidus de la variable réponse sont d'une part distribués normalement et d'autre part, distribués aléatoirement en fonction des valeurs ajustées (`plot` des résidus en fonction des valeurs ajustées). Notons ici qu'il n'y a pas d'équivalent non-paramétrique de la régression, mais que les permutations ou le bootstrap (re-échantillonnage avec remise) seront à nouveau utiles.
   - Si nous cherchons seulement si les 2 variables sont associées sans chercher de relation de causalité, nous utiliserons le coefficient de **corrélation** de Pearson, communément appelé coefficient de corrélation. Rappelons que ce coefficient mesure l'association *linéaire* entre les 2 variables. Pour pouvoir tester ce coefficient (`cor.test, method=pearson`), il faut que la distribution conjointe des 2 variables soit bivariée normale. Si ce n'est pas le cas, on pourra mesurer et tester (tests non-paramétriques) le **coefficient de correlation de rang** de Spearman (`cor.test, method=spearman`), ou de Kendall (`cor.test, method=kendall`). Ces 2 coefficients, calculés entre les rangs des 2 variables, mesurent si les rangs des 2 variables co-varient.
5. **Deux variables explicatives catégorielles**. **ANOVA à 2 facteurs, croisés ou hiérarchique**. Si les niveaux du premier facteur sont différents pour chaque niveau du deuxième facteur, nous avons affaire à une analyse hiérarchique, avec le premier facteur "niché" dans le second. Si les différents niveaux du premier facteur se retrouvent dans les différents niveaux du deuxième facteur, et vice-versa, nous avons affaire à un modèle croisé. Les facteurs seront fixes si l'analyse comprend tous les niveaux du facteur qui nous intéressent, aléatoires autrement.
6. **Une variable explicative catégorielle et une continue**. L'analyse à effectuer est alors l'**analyse de covariance ANCOVA**.
7. **Plus de deux variables explicatives, toutes catégorielles**. Ca reste le domaine de l'analyse de variance ANOVA. On parlera d'**ANOVA à plusieurs critères de classification**. Ces analyses pourront être hiérarchiques, croisées, ou un mélange des deux et avec des variables soient fixes,

soient aléatoires. Il devient dès lors assez difficile de savoir quel test appliquer pour chaque facteur !

8. **Plus de deux variables explicatives, toutes quantitatives**. Nous sommes alors dans le domaine de la **régression multiple** si nous cherchons à prédire la variable réponse en fonction des différentes variables explicatives, ou dans le domaine de la **corrélation partielle** si nous souhaitons simplement obtenir une mesure d'association entre les différentes variables explicatives et la variable réponse.
9. **Plus de deux variables explicatives, certaines quantitatives et d'autres catégorielles**. Dès que nous avons un mélange de variables explicatives quantitatives et catégorielles, nous revenons à l'**analyse de covariance**.

A.2 Une variable réponse qualitative/factorielle

Ce type de variable n'a pas été abordé dans ma partie du cours, et peu dans celle du Prof. Rousson. Elle s'analyse généralement, comme les variables réponses quantitatives discontinues, à l'aide du **modèle linéaire généralisé** (`glm` dans R). Une excellente introduction à ces modèles est le livre d'Annette Dobson (1990) *"An Introduction to Generalized Linear Models", Chapman & al.* Vous avez cependant vu un cas de figure ou un test simple (le test du χ^2) permet d'analyser l'association entre 2 variables qualitatives.

1. **Une variable explicative qualitative/factorielle**. Le test revient alors à analyser des **tables de contingence**. Il s'agit de tableaux ou les différentes catégories de la variable réponse sont croisées avec les différentes catégories de la variable explicative. Notons d'ailleurs que dans ce cas de figure, il n'y a pas vraiment de variable réponse ou explicative, les 2 variables sont sur un même pied, et ce que nous cherchons à mesurer, c'est dans quelle mesure ces 2 variables sont associées. Nous cherchons à tester si les valeurs d'une des variables sont indépendantes de celles de l'autre. On parle donc aussi de test d'indépendance des lignes et des colonnes dans une table de contingence. Ces tables de contingence s'obtiennent dans R grâce à la commande table. Les tests de l'indépendance des lignes et des colonnes sont le test du χ^2 (`chisq.test` dans R) ou le test exact de Fisher (`fisher.test` dans R). Notons ici aussi que le test du χ^2 permet de vérifier si la distribution des observations est conforme à une loi donnée (test de conformité à une loi, par exemple aux lois de Mendel ou au principe d'Hardy-Weinberg).
2. **Une variable explicative continue**. Il faut alors utiliser un **modèle linéaire généralisé**, avec une distribution sous-jacente qui dépend du nombre de catégories de la variable réponse. Des exemples classiques sont les analyses de survie (pour chaque observation, le résultat sera vivant ou

mort) et de sexe-ratio (mâle/femelle), pour une variable réponse binomiale ou la répartition de morphes distincts (variable réponse multinomiale).

3. **Plusieurs variables explicatives qualitatives/factorielles**. Il s'agit là d'analyser l'association entre plusieurs variables qualitatives. Nous avons affaire à des tables de contingences à 3 et plus de niveaux. Le modèle pour analyser ce type de données est le **modèle log-linéaire**, cas particulier du modèle linéaire généralisé.
4. **Plusieurs variables explicatives, certaines factorielles, d'autres continues**. A nouveau, il s'agit du **modèle linéaire généralisé**.

A.3 Une variable réponse quantitative discontinue

Le modèle sous-jacent à l'analyse de ce type de variable est aussi le **modèle linéaire généralisé** (`glm` dans R). Ce modèle n'a pas été abordé dans le cours. Notons juste ici qu'il permet d'analyser une variable réponse quand les résidus ont une distribution connue, non normale. Il permet en particulier de s'intéresser à des variables réponses de type binomiales et Poisson. Pour ces variables, il est parfois possible de les transformer pour les rendre continues et distribuées normalement, auquel cas nous pourrons utiliser le modèle linéaire (voir ci-dessus).

A.4 Plusieurs variables réponses

Ce type de problèmes n'a pas été abordé en cours. Pour analyser simultanément plusieurs variables réponses, nous ferons appel aux **analyses multivariées**, qui seront présentés dans un cours spécifique en 3ème année.

Bibliographie

R. A. Becker, J. M. Chambers, and A. R. Wilks. *The new S-Language ("the blue book").* Chapman and Hall, 1988.

M. J. Crawley. *Statistical Computing : an Introduction to Data Analysis using S-plus.* Wiley, 2002.

P. Dalgaard. *Introductory Statistics with R.* Springer, 2002.

A. Dobson. *An Introduction to Generalized Linear Models.* Chapman and Hall, 1990.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

D. Falconer and T. MacKay. *Introduction to Quantitative Genetics.* Prentice Hall, fourth edition, 1996.

A. Grafen and R. Hails. *Modern Statistics for the Life Sciences-Learn how to analyse your experiments.* Wiley, 2002.

M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer, first edition, 1998.

B. J. F. Manly. *Randomization and Monte Carlo methods in Biology.* Chapman and Hall, 1997.

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, 1989.

R. R. Sokal and F. J. Rohlf. *Biometry.* Freeman, 1981.

J. Verzani. *Using R for introductory statistics.* Chapman et Hall CRC, first edition, 2005.

M Whitlock and D Schluter. *The Analysis of Biological Data.* Roberts, first edition, 2008.

J. H. Zar. *Biostatistical Analysis.* Prenctice Hall, 1984.