Analyse de la variance à 2 facteurs

Le modèle d'analyse de variance à 2 facteurs permet d'étudier l'effet de deux facteurs qualitatifs (facteurs A et B) sur une variable quantitative Y.

Exemple : Une expérience est destinée à étudier l'adaptation de deux variétés de moutarde à la sécheresse. On a un dispositif expérimental permettant d'étudier 2 variétés de moutarde, notées A et B, et l'effet de 2 intensités lunimeuses notées X et Y. Un indicateur de l'adaptation à la sécheresse est l'appartition de racines courtes tubérisées. Le tableau suivant indique le nombre observé de ces racines :

	variété A	variété B
Luminosité X	78	137
	79	85
	44	302
	77	315
Luminosité Y	64	64
	96	67
	30	102
	20	47

- Charger les données (fichier anova2.txt)
- Commenter le plan d'expérience : complet ? équilibré ?
- Vérifier que les variables variete et lux sont bien traitées par R comme des variables qualitatives.
- Réaliser une courte analyse descriptive des données, sous forme de boite à moustaches (pour chaque variété, puis pour chaque luminosité).

Le modèle d'analyse de variance à 2 facteurs :

1ère paramétrisation :

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, I \ , j = 1, \dots, J \text{ et } k = 1, \dots, n_{ij}$$

avec μ_{ij} = effet (inconnu) de la combinaison "modalité i du facteur A $\times$ modalité j du facteur B", et $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

2ème paramétrisation : décomposition des effets :
On décompose μ_{ij} afin de faire apparaître les effets des facteurs A et B.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, I, \ j = 1, \dots, J \text{ et } k = 1, \dots, n_{ij} \tag{1}$$

avec
μ = effet global
α_i = l'effet spécifique de la modalité i du facteur A (=effet de la variété i),
β_j = l'effet spécifique de la modalité j du facteur B (=effet de la luminosité j),
γ_{ij} = terme d'intéraction donnant l'effet spécifique de la combinaison variété i $\times$ luminosité j,
et ε_{ijk} $i.i.d \sim \mathcal{N}(0, \sigma^2)$.

Remarque : Le modèle (1) avec le terme d'intéraction est appelé Modèle complet d'Anova 2. La présence du terme d'interaction n'est pas systématique. Lorsque $n_{ij} = 1 \ \forall i, j$, c'est-à-dire lorsqu'il n'y a pas de répétition, on ne peut pas faire apparaitre le terme d'interaction dans le modèle. En effet comme $k = 1$, le terme d'interaction γ_{ij} est confondu avec l'erreur rédisuelle ε_{ij} puisque ces deux termes portent les mêmes indices. On est donc amené à négliger le terme d'interaction. Cela ne signifie pas qu'il n'existe pas, mais que nous n'avons pas assez de données pour l'estimer. Si cette interaction existe effectivement, les tests des effets des facteurs A et B seront moins puissants.
Ecrire le modèle sous forme matricielle : $Y = X\theta + \varepsilon$ (prendre $I = 2, J = 2, n_{ij} = 4$).

Donner la matrice X. Quel est le rang de la matrice X ?
La matrice X n'étant pas de plein rang, on pose des contraintes d'itentifiabilité sur les paramètres ce qui va permettre d'obtenir un unique estimateur de θ. Voici une manière commode de définir les contraintes : les modalités 1 de chaque facteur sont choisies comme référence

$\alpha_1 = \beta_1 = 0$, $\gamma_{i1} = 0 \ \ \forall i \in \{1, \ldots, I\}$ et $\gamma_{1j} = 0 \ \ \forall j \in \{1, \ldots, J\}$

Le modèle Anova 2 sous R :

En fonction de si oui ou non on peut modèliser un terme d'interaction :

```
# modèle anova 2 complet ie avec le terme d'interaction
res=lm(RC~variete*lux,contrasts=c("contr.treatment", "contr.treatment"),data)

# modèle anova 2 sans le terme d'interaction
res=lm(RC~variete+lux,contrasts=c("contr.treatment", "contr.treatment"),data)
```

1. Rentrer le modèle sous R (ici peut-on modéliser un terme d'interaction ?)
2. Effectuer l'analyse des résidus. Si elle ne vous satisfait pas, essayer de transformer les données (avec log).
3. Vous pouvez vérifier par un test certains points de l'analyse des résidus.
   - le test de Shapiro pour la normalité
   - l'homoscédasticité de la variance avec le test de Bartlett. Vous pouvez vérifier que la variance ne dépend ni de la variété, ni de la luminosité avec ceci :

   ```
   bartlett.test(res$residuals ~ variete)
   bartlett.test(res$residuals ~ lux)
   ```
4. Une fois que vous avez un modèle qui vous convient, regarder la table summary(res). Vous semble-t-elle cohérente avec la contrainte ?

Tests de Fisher :

1. Existence d'un effet d'au moins un des 2 facteurs ?

La ligne F-statistic de summary(res) donne la statistique du test de Fisher, les ddl associés et la p-value de :

$$H_0 : y_{ijk} = \mu + \varepsilon_{ijk} \quad \text{contre } H_1 : y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

Si on rejette ce test, il existe un effet d'au moins un des facteurs A ou/et B.
Est-ce-qu'au moins un des facteurs variété ou luminosité est significatif ?

2. Tests de chacun des facteurs :

Lorsque $n_{ij} > 1$ on peut considérer le modèle complet d'Anova 2, c'est-à-dire en prenant en compte un terme d'interaction.

Dans un premier temps, on teste si le terme d'interaction entre les 2 facteurs est significatif :

$$H_0 : \forall (i,j) \ \ \gamma_{ij} = 0 \text{ contre } H_1 : \exists (i,j) \ \ t.q \ \ \gamma_{ij} \neq 0$$

Cela revient à tester : $H_0 : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{contre } Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$

Puis, on teste l' effet de chaque facteur.

Avant de donner les résultats obtenus avec R, on définit les modèles suivants. Les dimensions de chacuns de ces modèles ont été obtenus en écrivant le modèle sous forme matricielle et en regardant le rang de la matrice X associée (les colonnes linéairement indépendantes).

Modèle M_0 :	$y_{ijk} = \mu + \varepsilon_{ijk}$	$\dim(M_0) = 1$
Modèle M_A :	$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$	$\dim(M_A) = I$
Modèle M_B :	$y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$	$\dim(M_B) = J$
Modèle M_{A+B} :	$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$	$\dim(M_{A+B}) = I + J - 1$
Modèle $M_{complet}$:	$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$	$\dim(M_{complet}) = IJ$

	DF	Sum Sq	F value	
Facteur A	$I-1$	$R(\alpha/\mu) = \|\widehat{Y}_{M_A} - \widehat{Y}_{M_0}\|^2$	$\frac{R(\alpha/\mu)/(I-1)}{SCR_{complet}/(n-IJ)}$	(a)
Facteur B	$J-1$	$R(\beta/\mu,\alpha) = \|\widehat{Y}_{M_{A+B}} - \widehat{Y}_{M_A}\|^2$	$\frac{R(\beta/\mu,\alpha)/(J-1)}{SCR_{complet}/(n-IJ)}$	(b)
Interaction	$IJ-(I+J-1)$	$R(\gamma/\mu,\alpha,\beta) = \|\widehat{Y}_{M_{complet}} - \widehat{Y}_{M_{A+B}}\|^2$	$\frac{R(\gamma/\mu,\alpha,\beta)/(IJ-(I+J-1))}{SCR_{complet}/(n-IJ)}$	(c)
Residuals	$n-IJ$	$SCR_{complet} = \|Y - \widehat{Y}_{M_{complet}}\|^2$		

TABLE 1 – Table obtenue avec anova(res)

La TABLE 1 est la table d'analyse de variance donnée par **anova(res)** : elle donne dans la colonne "Sum Sq" ce que l'on appelle "les sommes des carrés de type I ".

Cette table permet de tester :
(c) : l'effet de l'interaction : modèle M_{A+B} contre modèle $M_{complet}$
(a) : l'effet du facteur A : modèle M_0 contre modèle M_A
(b) : l'effet du facteur B après prise en compte du facteur A : modèle M_A contre modèle M_{A+B}

Inconvénient de la table donnée par anova(res) : on teste l'effet du facteur A sans s'ajuster à l'effet du facteur B. On risque alors de conclure qu'il existe un effet significatif du facteur A alors qu'il est peut-être uniquement dû à l'effet du facteur B.
On va donc plutôt utiliser la table d'analyse de variance donnée par "**Anova(res)**" (cf TABLE 2) : elle donne dans la colonne "Sum Sq" ce que l'on appelle "les sommes des carrés de type II". Pour utiliser "Anova(res)", il faut charger le package "car" en tapant library(car). Chez vous le package est à installer si besoin avec install.packages("car").

	DF	Sum Sq	F value	
Facteur A	$I-1$	$R(\alpha/\mu,\beta) = \|\widehat{Y}_{M_{A+B}} - \widehat{Y}_{M_B}\|^2$	$\frac{R(\alpha/\mu,\beta)/(I-1)}{SCR_{complet}/(n-IJ)}$	(a')
Facteur B	$J-1$	$R(\beta/\mu,\alpha) = \|\widehat{Y}_{M_{A+B}} - \widehat{Y}_{M_A}\|^2$	$\frac{R(\beta/\mu,\alpha)/(J-1)}{SCR_{complet}/(n-IJ)}$	
Interaction	$IJ-(I+J-1)$	$R(\gamma/\mu,\alpha,\beta) = \|\widehat{Y}_{M_{complet}} - \widehat{Y}_{P_{A+B}}\|^2$	$\frac{R(\gamma/\mu,\alpha,\beta)/(IJ-(I+J-1))}{SCR_{complet}/(n-IJ)}$	
Residuals	$n-IJ$	$SCR_{complet} = \|Y - \widehat{Y}_{M_{complet}}\|^2$		

TABLE 2 – Table obtenue avec Anova(res)

Cette table TAB 2 permet de tester les effets de chaque facteur en s'ajustant à l'autre facteur :
(a') : test de l'effet du facteur A après prise en compte du facteur B : modèle M_B contre modèle M_{A+B}.

Cas d'un plan d'expérience équilibré :
Lorsque le plan d'expérience est équilibré, les estimations des paramètres sont les mêmes quelque soit le modèle. Par exemple, l'estimation de α_i est la même dans le modèle M_A, M_{A+B} ou $M_{complet}$. En conséquence, le test de M_0 contre M_A est équivalent au test de M_B contre M_{A+B}. Les sommes des carrés de type I et II dont identiques et les tables TAB 1 et TAB 2 sont donc les mêmes.

Application :

1. Tester l'existence d'un terme d'interaction entre les facteurs variété et luminosité.
2. Tester l'existence d'un effet variété.
3. Tester l'existence d'un effet luminosité.
   Remarque : Pour les questions 2, 3 et 4), quelle table d'analyse de variance utilisez-vous ? anova(res) ou Anova(res) ? Justifier.
4. Quel modèle garder ?
5. Quelle est la variété qui s'adapte le mieux à la sécheresse ?

6. Dans le cas où on dispose de données pour 10 variétés, si on souhaite tester simultanément toutes les variétés 2 par 2, que doit-on faire ?
   Remarque : pairwise.t.test ne permet pas de tester les variétés 2 à 2 dans un modèle anova 2 (se place automatiquement dans un modèle anova 1). On peut utiliser la méthode de Tukey.
   Sous R :

```
# le code nécessite d'utiliser la fonction aov plutôt que lm
# donne les p-values pour les tests bilatéraux.
comp.Tukey = TukeyHSD(aov(log(RC)~variete + lux))
# on peut visualiser les intervalles de confiance simultanées avec :
par(mfrow=c(1,2))
plot(comp.Tukey)
```

7. Est-ce-que les plants soumis à la luminosité X s'adapte mieux à la sécheresse que ceux soumis à la luminosité Y ?

Et si on ne peut pas mettre en oeuvre un modèle d'analyse de variance à 2 facteurs ?

On se tourne vers la version non paramétrique de l'anova 2 et le test de Ray-Hare-Scheirer (besoin de répétitions dans le plan d'expérience) :

```
# y a t-il un effet de la variété et/ou de la luminosité  ?
library(rcompanion)
scheirerRayHare(RC~variete + lux, data=donnees)
```

Pour ensuite comparer les variétés 2 à 2 ou les luminosités 2 à 2, on peut utiliser le test de Wilcoxon-Mann-Whitney (comme en Anova 1)

```
pairwise.wilcox.test(RC , variete, p.adjust.method="bonferroni")
pairwise.wilcox.test(RC , lux, p.adjust.method="bonferroni")
```

Compléments

- Modèles à plus de deux facteurs ?
  En théorie oui, mais il faut beaucoup de données si on a beaucoup de paramètres à estimer !
- Une variable explicative peut parfois être traitée comme quantitative ou qualitative ($x = dose$ par exemple avec comme valeurs 1 : 5). En la considérant comme quantitative, on réduit le nombre de paramètres à estimer (1 en régression simple, 5 en anova à un facteur).
- Facteurs hiérarchisés.
  Certains facteurs peuvent être emboités, les niveaux d'un facteur étant alors conditionnées par un autre facteur.
- Modèles à effets aléatoires et modèles mixtes.
  Dans ce module, on a étudié les modèles linéaires à effet fixes. Par exemple en analyse de variance à 1 facteur, les différents niveaux du facteur (les I types de variétés) étaient fixés par l'expérimentateur et on voulait comparer ces I types de variétés. Pafois les I types de variétés ne sont pas les seuls qui intéressent l'expérimentateur mais sont I niveaux pris au hasard parmi un nombre de niveaux très élevé. Le choix des I niveaux est donc aléatoire ce qui doit se traduire dans le modèle en utilisant un facteur à effet aléatoire.

A vous...

On veut étudier l'effet de deux facteurs qualitatifs "niveau de la fertilisation" et "rotation de la culture" sur le rendement de colza. (fichier fertilisation.rotation.txt) On compare :

— $I = 2$ niveaux de fertilisation (1 pour faible et 2 pour fort)
— $J = 4$ types de rotation (A, B, C et D)

Chaque combinaison "niveau de la fertilisation" et "rotation de la culture" est appliquée sur plusieurs parcelles. Au total on a $n = 25$ parcelles.
Les données sont dans le fichier fertilisation.rotation.txt.