# Analyse de variance à 1 facteur

Le modèle d'analyse de variance à 1 facteur permet d'étudier l'effet d'un facteur qualitatif sur une variable quantitative Y.
Exemple : étudier l'effet de la variété de mais sur le rendement. Dans le but de comparer le rendement de trois variétés de mais numérotées 1, 2 et 3, on a cultivé chaque variété sur 5 parcelles et observé le rendement de chacune de ces 15 parcelles.

## Courte analyse descriptive des données

```
# charger les données :
data=read.table("rendement_mais.txt")

# vérfier que la variable variété est bien traitée par R comme une variable qualitative :
is.factor(data$variete)
        # si TRUE, c'est que oui, vous pouvez attacher les données et travailler avec data.
        # si FALSE, il faut transformer la variable variété en factor avant d'attacher les données.
data$variete=as.factor(data$variete)
attach(data)

# Réaliser une courte analyse descriptive des données, sous
  forme de "boite à moustaches" pour chaque variété. Commenter.
boxplot(rendement ~ variete)
points(tapply(rendement,variete,mean))  # affiche les rendements moyens par variété

> tapply(rendement,variete,mean)
   1    2    3
44.2 50.8 54.4
>  sqrt(tapply(rendement,variete,var))
       1        2        3
1.303840 1.643168 2.073644
```

## Le modèle d'analyse de variance à 1 facteur :

1ère paramétrisation :

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, I \text{ et } j = 1, \ldots, n_i \tag{1}$$

avec
$i$ : indice de la variété, $i = 1, \ldots, I = 3$
$j$ : numéro de parcelle au sein de la variété, $j = 1, \ldots, n_i = 5$
$Y_{ij}$ : rendement de la $j$ ème parcelle avec la variété $i$
$\mu_i$ = effet de la variété $i$
et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

1. Mettre ce modèle sous forme matricielle $Y = X\theta + \varepsilon$. Donner la matrice X. Quel est son rang ?
2. Calculer l'estimateur $\hat{\theta}$ des moindres carrés de $\theta$.
3. Application numétique : en déduire les estimations $\hat{\mu}_i$.

2ème paramétrisation : Voici la paramètrisation qu'utilise R (décomposition des effets).

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, I \text{ et } j = 1, \ldots, n_i \tag{2}$$

avec $\mu$ = effet global et $\alpha_i$ = l'effet spécifique de la variété $i$.

Mettre ce modèle sous forme matricielle $Y = X\theta + \varepsilon$. Donner la matrice X. Quel est son rang ?

Comme X n'est pas de plein rang, on va imposer des contraintes d'identifiabilité sur les paramètres ce qui va permettre d'obtenir un unique estimateur de $\theta$. En effet, on dit que le modèle avec la $2^{\text{ème}}$ paramétrisation n'est pas identifiable car pour $\mu_i$ donné, les paramètres $\mu$ et $\alpha_i$ ne sont pas définis de manière unique. La contrainte d'identifiabilité va garantir l'unicité de la décomposition $\mu_i = \mu + \alpha_i$.

Combien de contraintes sont nécessaires ?

R peut définir des contraintes de plusieurs manières. Voici l'une des plus commodes.
Les paramètres du modèle (2) sont $(\mu, \alpha_1, \alpha_2, \alpha_3)$ si $I = 3$. R va choisir la modalité $i = 1$ (variété 1) comme référence et prendre $\alpha_1 = 0$. Grace à cela tous les paramètres du modèle (2) seront définis de manière unique, i.e on peut écrire les paramètres $\mu, \alpha_i$ du modèle (2) de façon unique à partir des paramètres $\mu_i$ du modèle (1). En effet :

$$\begin{cases} \mu_1 & = \mu + \alpha_1 \\ \mu_2 & = \mu + \alpha_2 \\ \mu_3 & = \mu + \alpha_3 \\ \text{avec} & \text{la contrainte } \alpha_1 = 0 \end{cases} \iff \begin{cases} \mu & = \mu_1 \\ \alpha_1 & = 0 \\ \alpha_2 & = \mu_2 - \mu_1 \\ \alpha_3 & = \mu_3 - \mu_1 \end{cases} \tag{3}$$

Remarque : cette contrainte mathématique permet de calculer de façon unique les estimations de $\mu$ et $\alpha_i$. Tous les résultats des tests de Student ou de Fisher seront inchangés.

## Le modèle anova 1 sous R :

On rentre le modèle anova 1, en demandant à R de prendre la contrainte où la première variété est prise comme référence :

```
res=lm(rendement ~ variete, contrasts="contr.treatment", data)
```

Voici les résultats obtenus avec summary(res)

```
> summary(res)$coefficients
```

| | Estimate | ... |
|---|---|---|
| (Intercept) | $\hat{\mu}$ | ... |
| variete2 | $\hat{\alpha}_2$ | ... |
| variete3 | $\hat{\alpha}_3$ | ... |

D'après la contrainte, $\hat{\alpha}_1 = 0$. C'est pour cela que $\hat{\alpha}_1$ n'apparait pas dans ce tableau.

Lien avec la $1^{\text{ère}}$ paramétrisation : à partir de (3) et des estimations $\hat{\mu}_i = Y_{i\cdot}$ trouvés avec la première paramétrisation pour le modèle (1), vérifiez que les estimations données par R dans le tableau summary(res) sont cohérentes.

## Test de Fisher : y-a-t'il un effet du facteur ?

Test de $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I (= \mu)$ contre $H_1 : \exists i, l \in \{1, \ldots, I\}$ tel que $\mu_i \neq \mu_l$
Avec la $2^{\text{ème}}$ paramétrisation, ce test revient à tester :
$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$ contre $H_1 : \exists i \in \{1, \ldots, I\}$ tel que $\alpha_i \neq 0$
On obtient le résultat de ce test avec la ligne F-statistic de summary(res) ou avec le tableau anova(res).

A partir des données, tester l'effet de la variété sur le rendement.

On rejette $H_0$, il y a un effet significatif de la varieté sur le rendement : au moins 2 varietés n'ont pas le même effet sur le rendement. On peut alors se demander lesquelles, i.e quel <u>contraste</u> $\mu_i - \mu_l$ est différent de zéro.

## Test d'<u>un</u> contraste :

Soient $i$ et $l \in \{1, \ldots, I\}$ fixés. On souhaite tester $H_0 : \mu_i - \mu_l = 0$ contre $H_1 : \mu_i - \mu_l \neq 0$.
Cela correspond à un test de Student.

Comme on a choisi la contrainte où la variete 1 est choisie comme réference ($\alpha_1 = 0$), on trouve les résultats des tests de $H_0 : \mu_i - \mu_1 = 0$ dans summary(res) :

| | Estimate | Std. Error | t value | $Pr(>\|t\|)$ |
|---|---|---|---|---|
| (Intercept) | $\hat{\mu} = \hat{\mu}_1$ | .. | .. | ... |
| variete2 | $\hat{\alpha}_2 = \hat{\mu}_2 - \hat{\mu}_1$ | $\sqrt{\hat{\text{var}}(\hat{\mu}_2 - \hat{\mu}_1)}$ | $\frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\hat{\text{var}}(\hat{\mu}_2 - \hat{\mu}_1)}}$ | p value du test de $H_0 : \mu_2 - \mu_1 = 0$ |
| variete3 | $\hat{\alpha}_3 = \hat{\mu}_3 - \hat{\mu}_1$ | $\sqrt{\hat{\text{var}}(\hat{\mu}_3 - \hat{\mu}_1)}$ | $\frac{\hat{\mu}_3 - \hat{\mu}_1}{\sqrt{\hat{\text{var}}(\hat{\mu}_3 - \hat{\mu}_1)}}$ | p value du test de $H_0 : \mu_3 - \mu_1 = 0$ |

Cette table permet donc de comparer chaque varieté à la varieté de reférence, i.e varieté 2 contre varieté 1, et varieté 3 contre varieté 1.
Si on veut comparer à une autre variété, il suffit de changer la contrainte. Par exemple si on veut la variété 2 comme référence :

```
variete=relevel(variete, ref="2")
res2=lm(rendement~ variete, contrasts"contr.treatment")
```

La table summary(res2) permet de tester la variété 1 contre la variété 2 et la variété 3 contre la variété 2.

Tester toutes les variétés 2 à 2 (d'abord par un test bilatéral puis par un test unilatéral).

De façon plus automatique sous R :

```
> pairwise.t.test(rendement, variete, p.adjust="none")
```

Vous pouvez demander les pvalues pour un test unilatéral avec l'option *alternative* mais alors tous les tests sont faits "dans le même sens".

## Comparaison multiple

Si on veut tester tous les contrastes 2 par 2 ($\mu_1 - \mu_2, \mu_1 - \mu_3, \ldots, \mu_{(I-1)} - \mu_I$), il faut faire $C_I^2 = \frac{I(I-1)}{2}$ tests. Lorsque l'on effectue ces $\frac{I(I-1)}{2}$ simultanément, que contrôle t-on ?

Lorsque l'on teste un contraste au niveau 5% la probabilité de se tromper en disant à tort que le contraste est significativement différent de 0 (on parle de faux positif) est de 5%. Lorsque l'on fait $I(I-1)/2$ tests, la probabilité de faire au moins un faux positif (ie la probabilité alors que $H_0$ est vraie, de se tromper au moins une fois : c'est-à-dire de trouver au moins un contraste significativement différent de 0) devient largement supérieure à $\alpha = 5\%$. La probabilité de faire au moins un faux positif est appelée **family-wise error rate (FWER)** :

$$\begin{aligned} FWER = \mathbb{P}_{H_0}(\exists i, l \text{ t.q } \mu_i \neq \mu_l) &= \mathbb{P}_{H_0}(\mu_1 \neq \mu_2 \text{ ou } \mu_1 \neq \mu_3 \text{ ou } \ldots \text{ ou } \mu_{(I-1)} \neq \mu_I) && (4) \\ &\leq \mathbb{P}_{H_0}(\mu_1 \neq \mu_2) + \mathbb{P}_{H_0}(\mu_1 \neq \mu_3) + \ldots + \mathbb{P}_{H_0}(\mu_{(I-1)} \neq \mu_I) && (5) \\ &\leq \alpha + \alpha + \ldots + \alpha = \alpha \times I(I-1)/2 && (6) \end{aligned}$$

Par exemple, si $I = 10$, on a 45 tests à faire, et la probabilité de faire au moins un faux positif est contrôlée par $0.05 * 45 = 2.25$, c'est à dire par 1 (la majoration pour passer de (4) à (5) est brutale), autrement dit : on ne contrôle plus l'erreur globale que l'on peut commettre sur les $I(I-1)/2$.

Pour corriger ce problème, on va appliquer la **méthode de Bonferroni**, qui consite à tester chaque contraste au niveau $\alpha' = \frac{\alpha}{I(I-1)/2}$. Qu'est-ce-que cela donne sur notre exemple ?

De façon automatique sous R, on peut otbenir les pvalues ajustées (toujours bilatérales) qui sont les pvalues precedentes $\times I(I-1)/2$. On les compare donc directement à $\alpha$.

```
> pairwise.t.test(rendement, variete, p.adjust="bonf")
```

Remarque : lorsque le nombre de tests à faire est grand, la correction de Bonferroni est vraiment sévère (on dit conservative), et on risque de ne pas détecter de contraste signifiativement different de 0. Il existe des corrections mieux adaptées comme la correction de Holm par exemple (voir Littérature si besoin).

Une autre approche usuelle en tests multiples est le contrôle du **FDR (False Discovery Rate)**. L'idée est la suivante : rejeter 10 hypothèses nulles sur $\frac{I(I-1)}{2}$ tests (ie on déclare 10 contrastes significativement différents de 0) et faire 4 erreurs parmi ces 10 rejets, est moins acceptable que rejeter 30 hypothèses nulles et faire 7 erreurs parmi ces 30 rejets. Ce qui nous intéresse ici n'est pas la probabilité de faire au moins un faux positif, mais le taux (parmi les rejets) de faux positifs (appelé FDR : false discovery rate). Dans le 1er cas, on a un taux

de $\frac{4}{10} = 0.4$ et dans le second de $\frac{7}{30} = 0.23$. En 1995, **Benjamini et Hochberg** ont proposé une procédure permettant de contrôler le FDR : cette procédure correspond à un intermédiaire entre ne pas faire de correction (déclarer un contraste significativement significatif si la pvalue associée est inférieure à $\alpha$) et la correction de Bonferroni (déclarer un contraste significativement significatif si la pvalue associée est inférieure à $\frac{\alpha}{\frac{I(I-1)}{2}}$). Cette procédure est très utilisée en pratique lorsqu'il y a beaucoup de tests à faire parce qu'elle autorise plus de rejets tout en garantissant de ne pas faire trop d'erreurs.

```
> pairwise.t.test(rendement, variete, p.adjust="BH")  # ou p.adust="fdr"
```

Une autre procédure en tests multiples et souvent utilisée en Anova est la **méthode de Tuckey**. Elle permet de contrôler la probabilité de faire au moins un faux positifs, mais est moins convervative que Bonferroni :

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

## Et si on ne peut pas mettre en oeuvre un modèle d'analyse de variance à 1 facteur ?

Si vous avez peu de données, et que l'analyse des résidus montre qu'on ne peut pas supposer les résidus gaussiens, ou que l'on n'arrive pas à regler un problème d'hétéroscédasticité, on se tourne vers les tests non paramétriques. Rappel : Le test de Wilcoxon-Mann-Whitney pour comparer deux échantillons et le test de Kruskal-Wallis pour comparer plus de deux échantillons.

```
# y a t'il un effet de la variété sur le rendement ? Il y a 3 types de  variétés donc:
kruskal.test(rendement ~ variete)
# comparaisons 2 à 2 des variétés :
## sans correction :
pairwise.wilcox.test(rendement , variete, p.adjust.method="none")
## avec la correction de Bonferroni ou de BH:
pairwise.wilcox.test(rendement , variete, p.adjust.method="bonferroni")
pairwise.wilcox.test(rendement , variete, p.adjust.method="BH")
```

## Application

Travailler avec les données du fichier "maladie", qui donne la durée de vie (en jours), après détection d'une maladie, de patients auxquels on a administré de l'ascorbate (un antioxydant présent dans la vitamine C) en fonction de l'organe touché. La même dose a été administré à tous les patients. On s'interroge quant à l'effet de l'ascorbate sur l'allongement de la durée de vie selon le type de maladie.

1. Faire une courte analyse descriptive des données.
2. On considère un modèle d'analyse de variance pour tester si la durée de vie moyenne est la même pour tous les types de maladie. Effectuez l'analyse des résidus pour valider les hypothèses du modèle. Qu'en concluez-vous ?
3. On suggère les transformations suivantes pour la variable durée : inverse - logarithme - racine carree. Laquelle de ces transformations choisissez-vous et pourquoi ?
4. Vérifier la normalité des résidus avec le test de Shapiro-Wilk ($H_0$ : normalité) avec

   `shapiro.test(res$residuals)`

   et vérifier l'homoscédasticité de la variance avec le test de Bartlett ($H_0$ : tous les types $i$ de maladie ont la même variance ) avec

   `bartlett.test(res$residuals ~ type)`

   Remarques : le test de Bartlett nécessite d'avoir au moins 3 observations pour chaque type $i$ de cancer.
5. Pour la transformation que vous aurez choisie, testez l'hypothèse suivant laquelle l'ascorbate a le même effet en moyenne pour tous les types de maladie.
6. Quels sont les types de maladie qui diffèrent entre eux deux-à-deux (faire pour chaque test 2 à 2, un test unilateral) ? Utiliser d'abord la correction de Bonferroni, celle de BH, puis celle de Tuckey.