RÉPUBLIQUE DU SÉNÉGAL

Un peuple- un But- une Foi

****

Agence Nationale de la Statistique et de la Démographie

ANSD

Agence Nationale de la Statistique et de la Démographie

Ecole Nationale de la Statistique et de l'Analyse Ecónomique Pierre Ndiaye

ensae

Analyse de la variance sur mesures repétées

Analyse des effets de la variété et du type de fertilisant sur la croissance des plantes

Rédigé par :

BIDIAS ASSALA JULIEN PARFAIT

Élève ingénieur statisticien économiste

Sous la supervision de :

M. CARLOS AKAKPOVI

Enseignant à l'ENSAE

Avant-propos

Ce travail a été réalisé dans le cadre de la validation du **cours d'ANOVA sur mesures répétées**. L'objectif principal de celui-ci était de mettre en pratique les méthodes d'analyses économétriques entre variables qualitatives et variables quantitatives apprises au cours d'Analyse de la variance. Mais pour le cas d'espèce il s'agit particulièrement de l'analyse de la variance à mesures répétées.

Le jeu de données utilisé pour cette analyse était constitué de 253 observations. L'analyse a permis de mettre en évidence des effets similaires entre les groupes (modalité de chaque facteur) avec une p valeur supérieur à 0,05.

Le présent rapport est structuré en plusieurs sections, commençant par un rappel sur la théorie de l'analyse de la variance à mesures répétées, enchainant ensuite avec la présentation des données, les analyses descriptives et enfin la modélisation avec présentation des résultats plus interprétation.

Table des matières

1 Théorie sur l'Analyse de la variance à mesures repétées 7
1.1 Présentation des données 7
1.2 Plan équilibré 7

2 Présentation du problème, des données et analyse descriptive des différentes variables 11
2.1 Présentation du problème 11
2.2 Présentation des données et méthodologie 11
2.3 Analyse descriptive 13
2.4 Hypothèses de base nécessaires à la modélisation 20
2.4.1 Hypothèses de normalité 20
2.4.2 Hypothèse de sphéricité 23

3 Modélisation, résultats et interprétation 24
3.1 Anova à un facteur à mesures repétées pour le facteur Variété 24
3.2 Anova à un facteur à mesures repétées pour le facteur Fertilisation 24
3.3 Analyse de la variance à deux facteurs à mesures répétées avec interaction 25

Liste des figures

1 Hauteur moyenne des plantes suivant le fertilisant et pour chaque variété 14
2 Taille moyenne des pantes en fonction de chaque fertilisant 15
3 Taille des plantes en fonction des variétés de fruits 15
4 Box Plot de la hauteur pour chaque période 16
5 Box Plot de la Hauteur des plantes pour chaque modalité du facteur Fertilisation 16
6 Box Plot de la Hauteur des plantes pour chaque type d'engrais 17
7 Box Plot de la Hauteur des plantes pour chaque Variété 17
8 Box-plot de l'effet de la variété pour chaque mesure-période 18
9 Box-plot de l'effet des fertilisants pour chaque mesure (période) 18
10 Alignement de Hauteurs moyennes-cas du fertilisant 19
11 Alignement de Hauteurs moyennes-Cas de la variété 20
12 QQplot : la droite d'Henry 22

Liste des tableaux

1 Données . . . 7
2 Données . . . 8
3 Degrés de liberté . . . 9
4 Tableau d'analyse de la variance-Rappel . . . 9
5 Base partielle-Début . . . 11
6 Base partielle-Fin . . . 12
7 Tendances centrales . . . 13
8 Hauteur Moyenne pour chaque fertilisant pour différentes mesures . . . 13
9 Hauteur Moyenne pour chaque variété pour différentes mesures . . . 13
10 Hauteur moyenne pour chaque modalité des facteurs . . . 14
11 Tableau d'identification des valeurs aberrantes . . . 19
12 Log de la hauteur pour les 5 premières lignes . . . 21
13 Normalité croisée des modalités des facteurs . . . 23
14 Degrés de liberté . . . 25
15 Tableau d'analyse de la variance-Rappel . . . 26
16 Analyse de la variance . . . 26

Contexte d'application et objectif

L'analyse de variance (ANOVA) à deux facteurs à mesures répétées avec interaction est une méthode statistique utilisée pour analyser les données issues d'une expérience avec au moins deux variables indépendantes (facteurs) et une variable dépendante mesurée à plusieurs reprises (d'où les termes : mesures répétées). L'objectif étant d'étudier l'interaction entre ces deux facteurs ceci afin d'analyser l'influence conjointe des facteurs sur la variable dépendante.

En effet, les mesures sont répétées pour chaque combinaison des niveaux des deux facteurs. L'idée de **répétition** étant mise en exergue pour inférer sur les effets des facteurs et leur interaction car elles permettent de contrôler les sources de variation entre les sujets considérés. D'autre part, l'interaction signifie que l'effet d'un facteur dépend de la valeur de l'autre facteur. Laquelle interaction sera significative lorsque les effets des deux facteurs ne seront pas constants dans toutes les combinaisons des niveaux des facteurs. L'Anova à mesures répétées cherche donc à répondre à la question de savoir quelle est l'influence des facteurs (qualitatifs) sur une variable (quantitative) compte tenu des mesures de cette dernière sur différentes périodes ? Il existe donc à cet effet une pléthore de méthodes d'ajustement pour effectuer de l'ANOVA à plusieurs facteurs à mesures répétées avec ou sans interaction. On distingue :
Les méthodes de type Greenhouse-Geisser, qui permettent de corriger les estimations des degrés de liberté et de F-tests en cas de non-sphéricité des données ;
Les méthodes de type Huynh-Feldt sont également couramment utilisées pour corriger les estimations des degrés de liberté et de F-tests.

1 Théorie sur l'Analyse de la variance à mesures repétées

Nous allons dès à présent expliquer succinctement l'anova à mesures répétées en partant de la présentation des données, de l'écriture du modèle,les hypothèses de base et enfin nous présenterons l'inférence du modèle via le tableau d'analyse de la variance.

1.1 Présentation des données

Pour mieux comprendre l'Anova à mesures répétées, il convient de préciser que celle-ci peut se faire dans un plan équilibré comme dans un plan déséquilibré. De manière générale, toute expérimentation comportant au moins un facteur et comportant un nombre identique de répétitions dans chacune des modalités des facteurs est un plan équilibré. Dans le cas contraire, on parle d'un plan **déséquilibré**. Nous présenterons succinctement l'analyse de la variance à deux facteurs à mesures répétées. Laquelle permettra selon la décision du test de prolonger sur l'analyse de la variance à un facteur à mesures répétées en cas d'absence d'interaction ou à faire des comparaisons par paires dans le cas contraire.

1.2 Plan équilibré

Nous allons nous placer dans le cadre de l'anova à deux facteurs à mesures répétées avec des effets fixes. le cas des effets aléatoires se déduisant facilement car pour ce dernier il suffit juste de considérer que les facteurs sont des variables aléatoires suivant une loi normale. Nous palerons brièvement du cas des facteurs emboîtés.

Considérons donc le tableau ci-dessous

Table 1: Données

Facteur A	Facteur B				
	B_1	...	B_j	...	B_J
A_1	$Y_{1,1,1} \dots Y_{1,1,K}$	...	$Y_{1,j,1} \dots Y_{1,j,K}$	...	$Y_{1,J,1} \dots Y_{1,J,K}$
⋮	⋮	⋮	⋮	⋮	⋮
A_i	$Y_{i,1,1} \dots Y_{i,1,K}$	...	$Y_{i,j,1} \dots Y_{i,j,K}$	...	$Y_{i,J,1} \dots Y_{i,J,K}$
⋮	⋮	⋮	⋮	⋮	⋮
A_I	$Y_{I,1,1} \dots Y_{I,1,K}$	...	$Y_{I,j,1} \dots Y_{I,j,K}$	...	$Y_{I,J,1} \dots Y_{I,J,K}$

Pour lequel nous postulons le modèle suivant :

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

Avec $i = 1, 2, ..., I$ le nombre de modalités du facteur α, $j = 1, 2, ..., J$ le nombre de modalités du facteur β et $k = 1, 2, ..., K$ le nombre de périodes ou nombre de répétitions. Où Y_{ijk} est la valeur prise par la réponse Y dans les conditions (A_i, B_j) lors du $k^{ème}$ essai. μ désigne le facteur commun, α le premier facteur pour I modalités, β le second

facteur pour J modalités et ε est l'erreur. Tel que susmentionné, nous voulons étudier l'interaction entre les facteurs afin de voir leur effet sur la variable Y.

Pour ce faire, nous partons donc des hypothèses de l'anova : $\forall(i,j,k)\ 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K\ \mathcal{L}(\varepsilon_{ijk}) = \mathcal{N}(0,\sigma^2)$, et $cov(\varepsilon_{ijk}, \varepsilon_{lmn}) = 0$ si $ijk \neq lmn$ avec $1 \leq i,l \leq I$, $1 \leq j,m \leq J$ et $1 \leq k,n \leq K$. Nous supposerons qu'elles sont bien remplies pour la suite. Par ailleurs, certaines métriques utilisées pour résoudre notre modèle dans la pratique sont robustes. Ce qui fait souvent que nous puissions grâce aux logiciels ajuster nos modèles. A ces hypothèses, l'on pourrait ajouter celles de normalité sur les α_i et les β_j et dans ce cas l'on parlera d'un modèle d'anova à deux facteurs à effets aléatoires.

Par ailleurs, la notion d'analyse de la variance à mesures répétées à facteurs emboîtés quant-à-elle fait référence au cas où les modalités d'un facteur dépend des modalités de l'autre facteur.

Les contraintes suivantes sont établies pour pour pouvoir estimer les paramètres du modèle et passer à la décomposition de la variance.

$\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_i = 0$, $\sum_{i=1}^{I} (\alpha\beta)_{ij} = 0\ \forall j \in \{1,...,J\}$ et $\sum_{j=1}^{J} (\alpha\beta)_{ij} = 0\ \forall i \in \{1,...,I\}$.

Les estimateurs μ, α_1, ..., α_I, β_1, ..., β_J des paramètres μ, α_1, ..., α_I, β_1, ..., β_J sont donnés par les formules suivantes :

$\hat{\mu} = \bar{Y}$, $\hat{\alpha}_i = Y_{i\cdot} - \hat{\mu}\ 1 \leq i \leq I$, $\hat{\beta}_j = Y_{\cdot j} - \hat{\mu}$, $1 \leq j \leq J$.

Ainsi, nous donnons le tableau suivant qui est celui des réalisations de Y.

Table 2: Données

Facteur A	Facteur B				
	B_1	...	B_j	...	B_J
A_1	$y_{1,1,1} \cdots y_{1,1,K}$	...	$y_{1,j,1} \cdots y_{1,j,K}$	...	$y_{1,J,1} \cdots y_{1,J,K}$
⋮	⋮	⋮	⋮	⋮	⋮
A_i	$y_{i,1,1} \cdots y_{i,1,K}$	...	$y_{i,j,1} \cdots y_{i,j,K}$	...	$y_{i,J,1} \cdots y_{i,J,K}$
⋮	⋮	⋮	⋮	⋮	⋮
A_I	$y_{I,1,1} \cdots y_{I,1,K}$	...	$y_{I,j,1} \cdots y_{I,j,K}$	...	$y_{I,J,1} \cdots y_{I,J,K}$

La décomposition de la variance se fait donc ainsi qu'il suit :

$$sc_{TOT} = sc_A + sc_B + sc_{AB} + sc_R$$

La variation due au facteur A observée sur la liste de données y est définie par :

$$sc_A = JK \sum_{i=1}^{I} (y_{i..} - y_{...})^2$$

La variation due au facteur B observée sur la liste de données y est définie par :

$$sc_B = IK \sum_{j=1}^{J} (y_{.j.} - y_{...})^2$$

La variation due à l'interaction des facteurs A et B observée est :

$$sc_{AB} = K \sum_{j=1}^{J} \sum_{i=1}^{I} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$$

La variation résiduelle observée sur la liste de données y est définie par :

$$sc_R = \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{k=1}^{K} (y_{ijk} - y_{ij.})^2$$

La relation fondamentale de l'ANOVA reste valable lorsqu'elle est évaluée sur la liste de données y. Nous introduisons les degrés de liberté (ddl) associés à chaque ligne du tableau de l'ANOVA :

Table 3: Degrés de liberté

Source	Degrés de liberté
Facteur A	$n_A = I - 1$
Facteur B	$n_B = J - 1$
Interaction AB	$n_{AB} = (I-1)(J-1)$
Résiduelle	$n_R = IJ(K-1)$
Totale	$n_{TOT} = IJK - 1$

D'où le tableau d'analyse de la variance suivant :

Table 4: Tableau d'analyse de la variance-Rappel

Source	Variation	Ddl	Carré Moyen	F	Décision
Facteur A	sc_A	n_A	$s_A^2 = \frac{sc_A}{n_A}$	$f_A = \frac{s_A^2}{s_R^2}$	$\mathcal{H}_0'$ ou $\mathcal{H}_1'$
Facteur B	sc_B	n_B	$s_B^2 = \frac{sc_B}{n_B}$	$f_B = \frac{s_B^2}{s_R^2}$	$\mathcal{H}_0''$ ou $\mathcal{H}_1''$
Interaction	sc_{AB}	n_{AB}	$s_{AB}^2 = \frac{sc_{AB}}{n_{AB}}$	$f_{AB} = \frac{s_{AB}^2}{s_R^2}$	$\mathcal{H}_0'''$ ou $\mathcal{H}_1'''$
Résiduelle	sc_R	n_R	$s_R^2 = \frac{sc_R}{n_R}$		
Totale	sc_{TOT}	n_{TOT}			

Test d'hypothèse sur le premier facteur

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = ... = \alpha_I = 0$ contre l'hypothèse

H_1: il existe i_0 appartenant à $1, 2, 3, ..., I$ tel que $\alpha_{i_0} \neq 0$

Sous l'hypothèse nulle (H_0) représente l'absence d'effet du facteur α et lorsque les conditions de validité du modèle sont respectées, $F_{(\alpha,obs)}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $I - 1$ degrés de liberté. Si la valeur de la statistique calculée est supérieure à la valeur tabulée alors on accepte l'hypothèse nulle d'absence d'effet de groupe des modalités du facteur sur l'explication de la variable dépendante.

Test d'hypothèse sur le second facteur

H_0: $\beta_1 = \beta_2 = \beta_3 = ... = \beta_J = 0$ contre l'hypothèse

H_1: il existe j_0 appartenant à $1, 2, 3, ..., J$ tel que $\beta_{i_0} \neq 0$

L'hypothèse nulle (H_0) représente l'absence d'effet du facteur β et lorsque les conditions de validité du modèle sont respectées, $F_{\beta,obs}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $(J - 1)$ degré de liberté. On conclut comme précedemment.

Pour l'intéraction nous avons :

H_0: $(\alpha\beta)_{11} = (\alpha\beta)_{12} = ... = (\alpha\beta)_{ij} = 0$

H_1: il existe i_0,j_0 tel que $(\alpha\beta)_{i_0 j_0} \neq 0$

Sous l'hypothèse nulle H_0 représente l'absence d'effet du facteur et lorsque les conditions de validité du modèle sont respectées, $F_{(\alpha\beta),obs}$ est la réalisation d'une variable aléatoire qui suit une loi de Fisher à $(I - 1)(J - 1)$ degré de liberté. Si la statistique calculée de Fisher est supérieure à la valeur tabulée alors on acceptera l'hypothése nulle d'absence d'effet de groupe des modalités d'un facteur croisé avec d'autres. C'est-à-dire que les facteurs ont les mêmes effets pris ensemble.

2 Présentation du problème, des données et analyse descriptive des différentes variables

2.1 Présentation du problème

Nous voulons étudier l'influence de deux facteurs que sont : **le fertilisant et la variété de fruits** sur la croissance des plantes. Pour cela, nous disposons du jeu de données suivant ayant en colonne le numéro, les répétitions (identificateur), le temps ("time"), le fertilisant qui a trois modalités qui sont en fait les trois types de terreau qui ont été utilisés :

- Terreau de Manguier (Ma)
- Terreau de Caïlcédrat (Ca)
- Terreau d'Anacarde (An)

La variété de fruit qui a deux modalités :

- Variété1 (Dg)
- Variété2 (Ds)

Les plantes ayant été mesurées à quatre périodes (4 répétitions de mesure). L'objectif étant de voir s'il existe une combinaison ou un traitement (type de terreau X variété) qui favorise la croissance ou non des plantes.

2.2 Présentation des données et méthodologie

C-dessous la base :

Table 5: Base partielle-Début

No	Repetition	Time	Fertilisation	Varietes	Hauteur
1	1	T1	Ma	Dg	43
2	1	T1	Ma	Dg	42
3	1	T1	Ma	Dg	40
4	1	T1	Ma	Dg	46
5	1	T1	Ma	Dg	42

Affichons aussi es 5 dernières lignes :

Table 6: Base partielle-Fin

No	Répétition	Time	Fertilisation	Variétés	Hauteur
249	4	T4	An	Ds	19.0
250	4	T4	An	Ds	43.3
251	4	T4	An	Ds	45.0
252	4	T4	An	Ds	46.0
253	4	T4	An	Ds	19.5

Pour apporter des éclairages à notre problème, nous allons effectuer une analyse de la variance à deux facteurs à mesures répétées.

Le modèle s'écrit donc comme suit :

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

Où μ représente l'effet commun des différents facteurs ;

α_i représente le facteur associé au fertilisant. Avec i la i-ème modalité de ce dernier (Ma, Ca, An). On a $i = 1, ..., I$ avec I le nombre de modalités. $I = 3$;

β_j représente le facteur associé à la variété. Avec j la j-ème modalité (Ds,Dg). on a $j = 1, ..., J$ avec $J = 2$ le nombre de modalités ;

$(\alpha\beta)_{ij}$ est l'interaction entre les modalités des deux facteurs que sont le fertilisant et la variété.

k étant le nombre de répétitions. $k = 1, ..., K(i, j)$. avec $K(i, j) = 4$

Y_{ijk} est la hauteur de la plante pour différentes périodes et ce pour chaque modalité des deux facteurs.

ε_{ijk} représente l'erreur pour différentes périodes et ce pour chaque modalité des deux facteurs. Elle capte tout ce que la partie déterministe du modèle $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ ne peut expliquer.

Enfin nous postulons le modèle général avec interaction. Mais pour mieux le comprendre nous partirons du modèle avec un facteur à mesures répétées et ensuite à deux facteurs à mesures répétées avec interaction.

Informations sur la base de données

```
tibble [253 x 6] (S3: tbl_df/tbl/data.frame)
 No           : num [1:253] 1 2 3 4 5 6 7 8 9 10 ...
 Repetition   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 Time         : Factor w/ 4 levels "T1","T2","T3",..: 1 1 1 1 1 1 1 1 1 1 ...
 Fertilisation: Factor w/ 3 levels "An","Ca","Ma": 3 3 3 3 3 3 3 3 3 3 ...
 Varietes     : Factor w/ 2 levels "Dg","Ds": 1 1 1 1 1 1 1 1 1 1 ...
 Hauteur      : num [1:253] 43 42 40 46 42 40 32.5 44 39.3 42 ...
 NULL
```

Les variables Répetition, Time, Fertisation et Varietes sont de type facteur. La Hauteur ici notre variable dépendante est de type numérique. Ce qui semble correct car elle est

par nature quantitative. Nous avons donc 253 lignes pour 6 colonnes. En effet nous sommes dans **un plan déséquilibré** compte tenu du fait que pour différentes mesures nous n'avons pas le mêmes nombre d'individus suivant les modalités des facteurs.

2.3 Analyse descriptive

Elle sera basée sur la hauteur moyenne des plantes pour chaque modalité du facteur considéré afin de voir numériquement comment celles-ci sont réparties. Ensuite, nous ferons des représentations graphiques pour apprécier ces résultats généraux. Nous ferons donc aussi des box-plots.

Tendances centrales

Table 7: Tendances centrales

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
No	1	253	127.00	73.18	127	127.00	93.40	1.0	253.0	252.0	0.00	-1.21	4.60
Repetition*	2	253	2.38	1.14	2	2.34	1.48	1.0	4.0	3.0	0.12	-1.40	0.07
Time*	3	253	2.38	1.14	2	2.34	1.48	1.0	4.0	3.0	0.12	-1.40	0.07
Fertilisation*	4	253	2.09	0.84	2	2.12	1.48	1.0	3.0	2.0	-0.18	-1.57	0.05
Varietes*	5	253	1.51	0.50	2	1.51	0.00	1.0	2.0	1.0	-0.04	-2.01	0.03
Hauteur	6	253	34.27	8.16	36	34.94	6.67	8.5	56.2	47.7	-0.72	0.39	0.51

Pour mieux apprécier les statistiques ci-dessous, nous ferons des moyennes groupées :

Table 8: Hauteur Moyenne pour chaque fertilisant pour différentes mesures

	x
An	33.52949
Ca	35.76438
Ma	33.76471

Table 9: Hauteur Moyenne pour chaque variété pour différentes mesures

	x
Dg	36.59113
Ds	32.03721

D'après ces deux tableaux, on peut dire que prit sans interaction (au sein de chaque groupe) les hauteurs moyennes pour chaque modalité des facteurs sont très proches. Comme pour

dire que les différents types d'engrais ont des effets similaires sur la croissance des plantes ou les types de variétés aussi. On ne peut dire indépendamment des facteurs. Mais cela reste valable du point de vue des moyennes. Toutefois, nous confirmerons cela à l'étape de la modélisation.

Donnons dès à présent le tableau des moyennes croisées de chaque facteur :

Table 10: Hauteur moyenne pour chaque modalité des facteurs

Fertilisation	Varietes	variable	n	mean	sd
An	Dg	Hauteur	43	35.640	8.517
An	Ds	Hauteur	35	30.937	9.706
Ca	Dg	Hauteur	35	38.206	6.844
Ca	Ds	Hauteur	38	33.516	7.366
Ma	Dg	Hauteur	46	36.252	6.454
Ma	Ds	Hauteur	56	31.721	7.968

A la lecture de ce tableau, quand on prend la colonne des moyennes, on peut faire le même constat que précédemment. Aussi, la colonne des tailles d'échantillon montre bien que le plan est déséquilibré. Par ailleurs, quand on prend la combinaison (Ca, Dg) on voit que la moyenne est de 38,2. Ce qui semble légèrement supérieur aux autres couples. On pourrait penser à un effet de groupe. Toutefois, grâce aux tests nous pourrons vérifier cela. Calculons dès à présent la moyenne des hauteurs moyennes et la médiane aussi.

```
moyenne = 34.37867 et médiane = 34.57800
```

Elles sont presque identique. Il y a une présomption de symétrie de la distribution des Hauteurs moyennes des plantes suivant chaque modalité des facteurs prit deux à deux. D'où l'idée encore d'une potentielle présence d'effets similaires des facteurs sur la variable dépendante. On peut apprécier tout ceci via des graphiques croisés. Aussi des graphiques indépendants.

Visualisation graphique du croisement

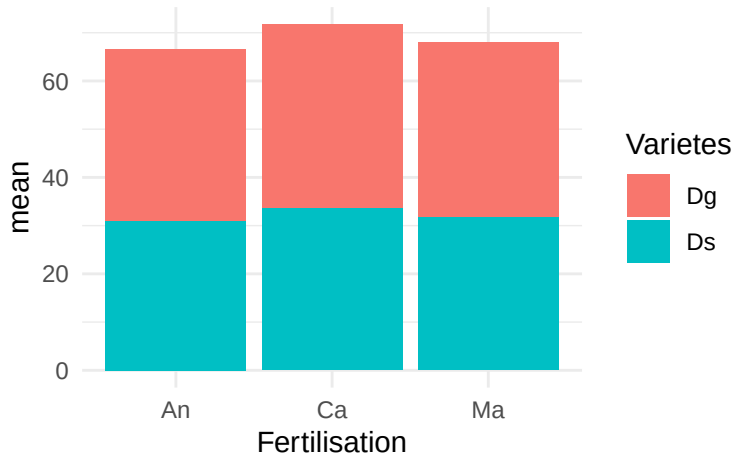


Figure 1: Hauteur moyenne des plantes suivant le fertilisant et pour chaque variété

Visualisation de la taille des plantes en fonction du type d'engrais

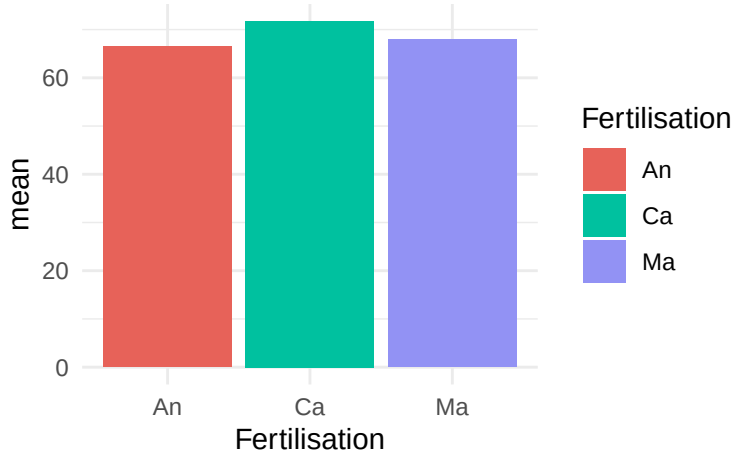


Figure 2: Taille moyenne des pantes en fonction de chaque fertilisant

A ce niveau, le constat semble pareil de part l'observation de ce graphique.

Visualisation de la taille des plantes en fonction des variétés de fruits

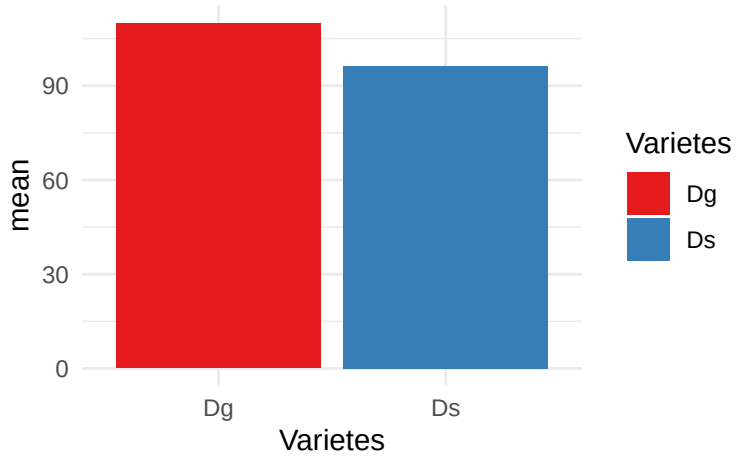


Figure 3: Taille des plantes en fonction des variétés de fruits

Jusqu'ici les constats restent valables. Il apparaît nécessaire de spécifier qu'il s'agit à ce niveau de constats. Lesquels seront infirmés ou confirmés plus bas.

Box-plot

Commençons par les box-plots de la hauteur moyenne suivant les différentes périodes :

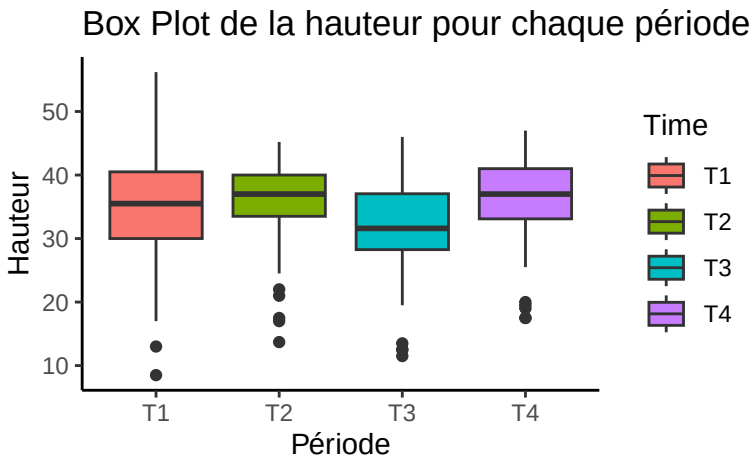


Figure 4: Box Plot de la hauteur pour chaque période

Ensuite nous représentons les box-plots de la hauteur moyenne pour chaque type d'engrais.

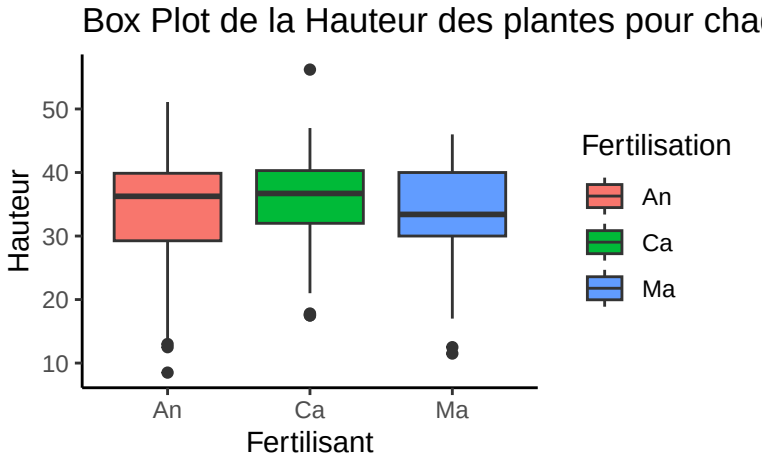


Figure 5: Box Plot de la Hauteur des plantes pour chaque modalité du facteur Fertilisation

Box-plots de la hauteur moyenne pour les différentes variétés.

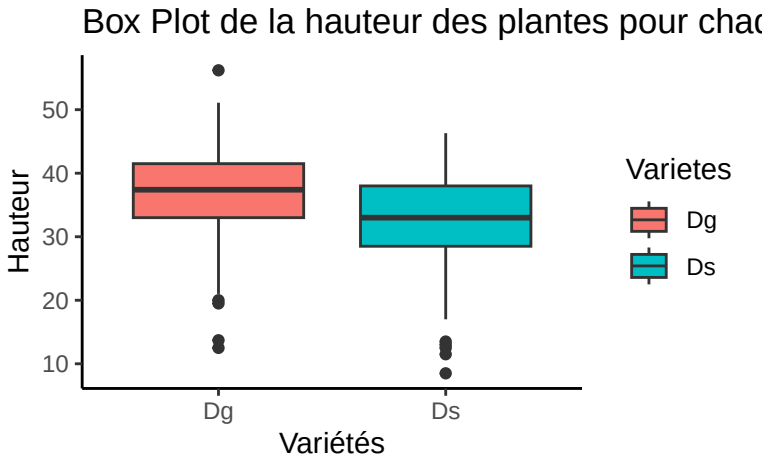


Figure 6: Box Plot de la Hauteur des plantes pour chaque type d'engrais

On voit qu'il y a des valeurs aberrantes et que les moyennes suivant les variétés sont presque proche. Enfin, nous représentons les box-plots pour chaque modalité des différents facteurs.

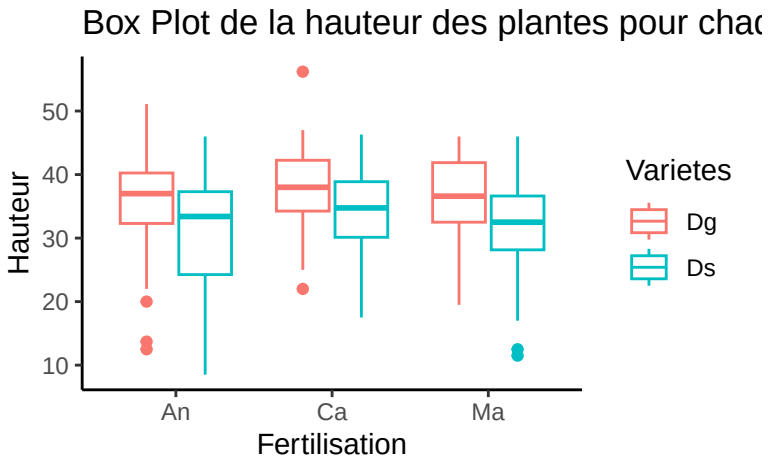


Figure 7: Box Plot de la Hauteur des plantes pour chaque Variété

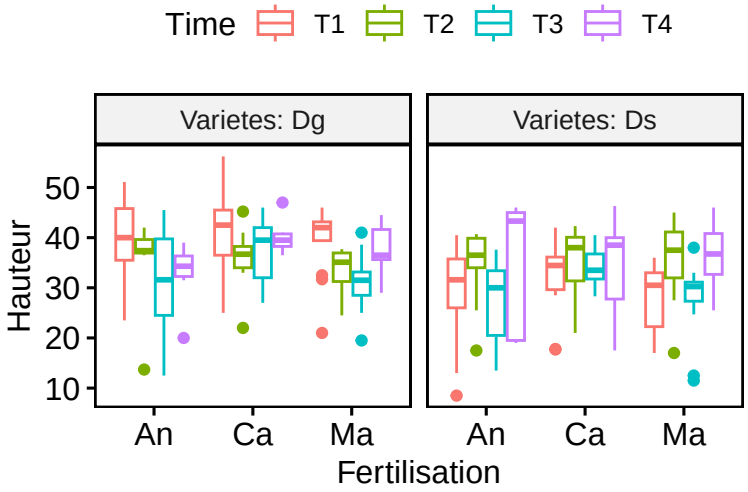


Figure 8: Box-plot de l'effet de la variété pour chaque mesure-période

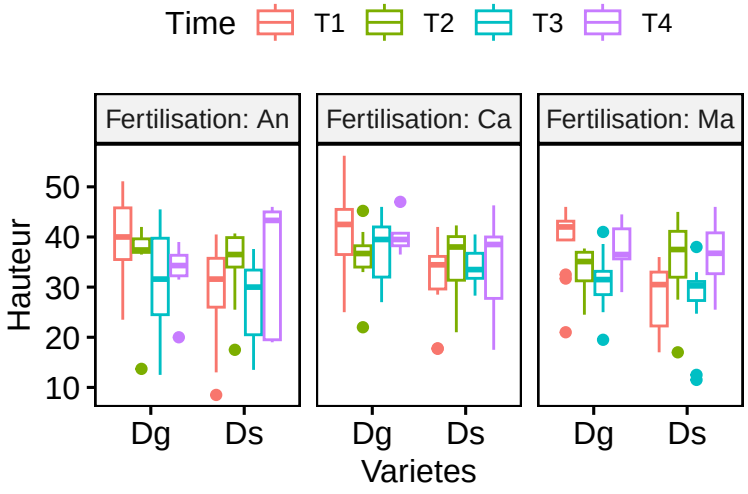


Figure 9: Box-plot de l'effet des fertilisants pour chaque mesure (période)

L'on constate la présence de valeurs aberrantes sur les box-plots ci-dessus. Lesquelles peuvent biaiser l'interprétation des résultats. Des transformations seraient donc appropriées ou la suppression de celles-ci. Mais toutefois, nous pouvons d'abord essayer voir si numériquement elles sont assez élevées. Ensuite, conclure sur la décision à prendre.

Identification des valeurs aberrantes

Table 11: Tableau d'identification des valeurs aberrantes

Time	No	Repetition	Fertilisation	Varietes	Hauteur	is.outlier	is.extreme
T1	190	1	An	Ds	8.5	TRUE	FALSE
T1	191	1	An	Ds	13.0	TRUE	FALSE
T2	44	2	Ma	Ds	17.0	TRUE	FALSE
T2	133	2	Ca	Dg	22.0	TRUE	FALSE
T2	141	2	Ca	Ds	21.0	TRUE	FALSE
T2	201	2	An	Dg	13.7	TRUE	TRUE
T2	219	2	An	Ds	17.5	TRUE	FALSE
T3	61	3	Ma	Ds	12.5	TRUE	FALSE
T3	63	3	Ma	Ds	11.5	TRUE	FALSE
T3	228	3	An	Dg	12.5	TRUE	FALSE
T3	234	3	An	Ds	13.5	TRUE	FALSE
T4	169	4	Ca	Ds	17.5	TRUE	FALSE
T4	172	4	Ca	Ds	17.5	TRUE	FALSE
T4	245	4	An	Dg	20.0	TRUE	FALSE
T4	249	4	An	Ds	19.0	TRUE	FALSE
T4	253	4	An	Ds	19.5	TRUE	FALSE

Nous avons bien des valeurs aberrantes dans notre jeu de données. Mais elles sont pas nombreuses et tellement élevées. On peut faire un nuage de points des facteurs pour chaque hauteur des plantes.

Pour les différents types d'engrais

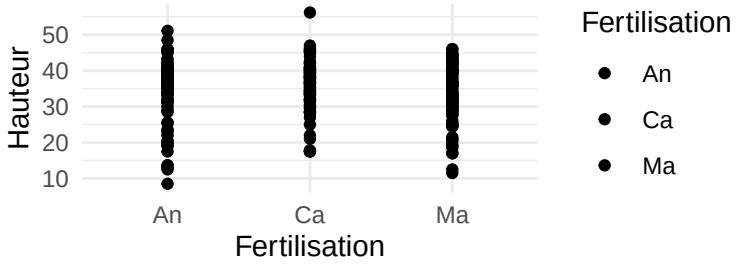


Figure 10: Alignement de Hauteurs moyennes-cas du fertilisant

Pour les différentes variétés

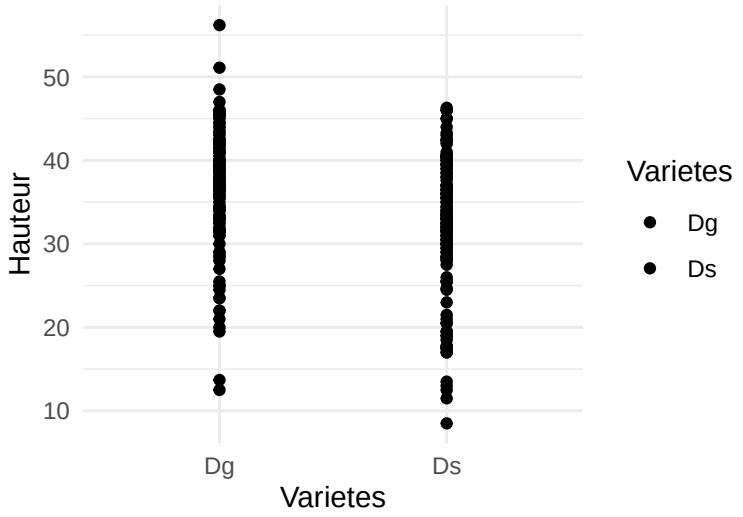


Figure 11: Alignement de Hauteurs moyennes-Cas de la variété

On peut dire que ces valeurs aberrantes ne sont pas tellement élevées. Pour le voir on regarde juste les points les plus hauts sur chaque graphique. En conclusion, ce n'est pas la peine de traiter les valeurs aberrantes, sinon on risque modifier la structure des données.

2.4 Hypothèses de base nécessaires à la modélisation

2.4.1 Hypothèses de normalité

Elle est faite sur la variable cible pour chaque modalité du facteur considéré suivant la période de mesure. L'objectif étant de pouvoir construire des tests pour une interprétation fiable. Commençons par faire une observation graphique via un qqplot.

Hypothèse de normalité de la variable cible prise individuellement

```
Shapiro-Wilk normality test

data:  datas$Hauteur
W = 0.95877, p-value = 1.252e-06
```

On constate que notre variable cible y n'est pas distribuée suivant une loi normale car la p-value étant inférieure au seuil de 5%. Combinée à la présence de valeurs manquantes, on peut voir s'il y a des transformations sur nos données qui pourront aux mieux nous aider.

Transformation log-linéaire

Table 12: Log de la hauteur pour les 5 premières lignes

x
3.761200
3.737670
3.688880
3.828641
3.737670

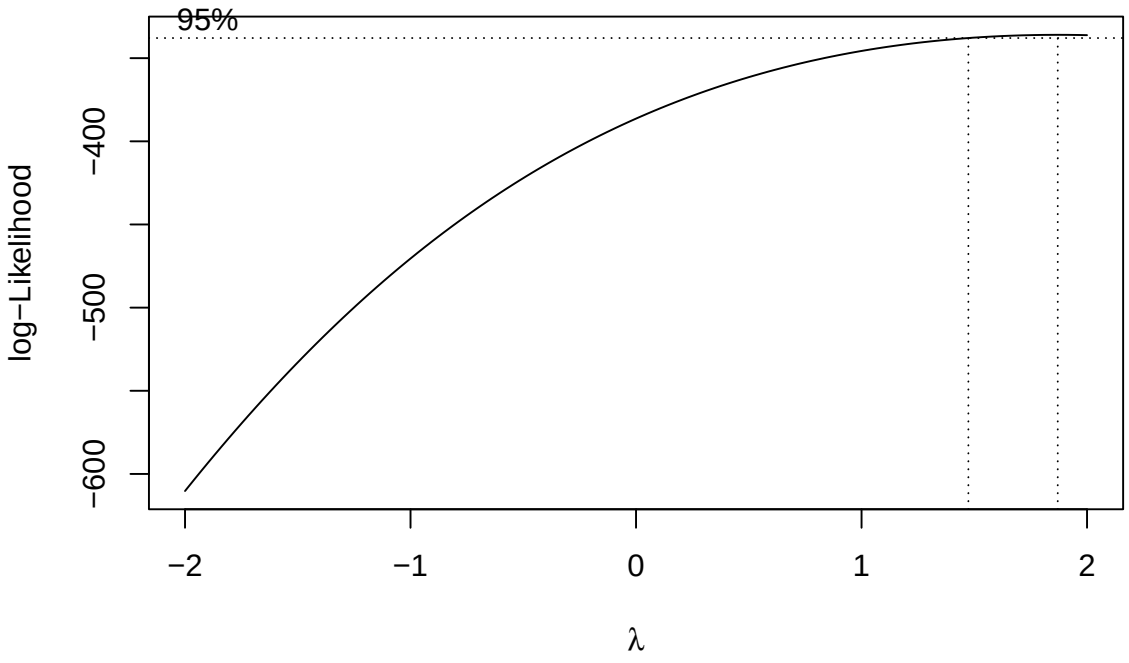
```
Shapiro-Wilk normality test

data:  datas$log_hauteur
W = 0.85978, p-value = 2.052e-14
```

La conclusion reste la même. On ne peut donc utiliser une transformation log-linéaire.

Transformation Box-cox

La transformation de Box-Cox peut être utilisée pour rendre les données plus symétriques et plus conformes à une distribution normale. Cela peut faciliter l'analyse statistique, en particulier lorsque des tests statistiques qui supposent une distribution normale sont utilisés. Le choix optimal de la valeur de lambda dépend de la distribution initiale de la variable et peut être déterminé en utilisant une procédure de recherche d'optimisation sur la fonction de vraisemblance de l'échantillon ici notre jeu de données.



Notre paramètre optimal λ est donc entre $]1,5; 1,99[$. Prenons $\lambda = 1,95$ on fait donc une transformation de la hauteur de la forme :

$$datasHauteur_{bc} = (datasHauteur^{\lambda} - 1)/\lambda$$

```
  Shapiro-Wilk normality test

data:  datas$Hauteur_bc
W = 0.98773, p-value = 0.02957
```

Ainsi, au seuil de 2% nous avons l'hypothèse de normalité de la distribution. Les données sont très robustes au changement d'échelle. Nous les utiliserons ainsi, tout en considérant comme robuste les test de type III et les métriques qui seront utilisées plus-tard.

Normalité croisée

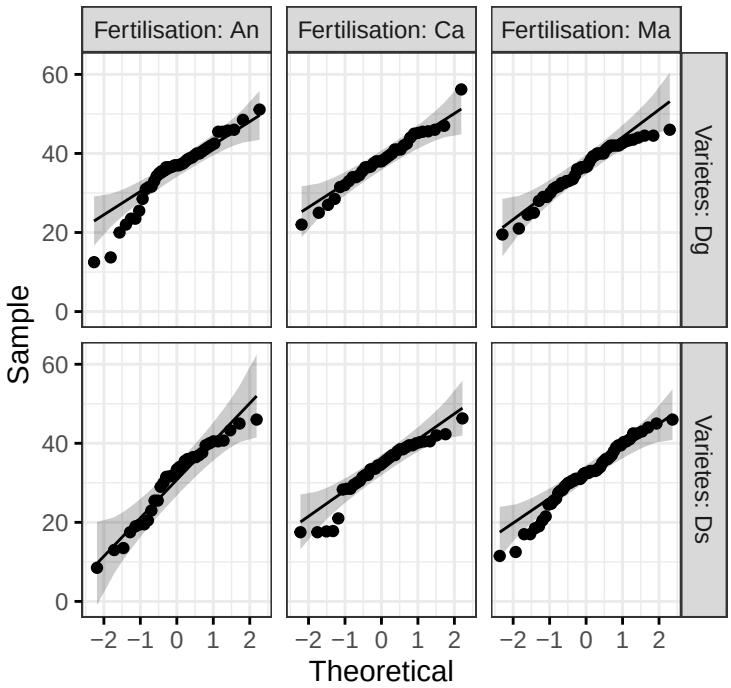


Figure 12: QQplot : la droite d'Henry

L'alignement de ces points sur la droite **(droite d'Henry)** montre que l'hypothèse de normalité est vérifiée. Toutefois, elle semble moins plausible pour le croisement entre les modalités An et Dg ; Ma et Dg ; Ca et Ds. Le classique test de shapiro-wilk sera donc utilisé pour soutenir cette idée.

Test de normalité

Table 13: Normalité croisée des modalités des facteurs

Fertilisation	Varietes	variable	statistic	p
An	Dg	Hauteur	0.9271465	0.0092944
Ca	Dg	Hauteur	0.9787729	0.7188226
Ma	Dg	Hauteur	0.9456286	0.0318978
An	Ds	Hauteur	0.9448721	0.0788482
Ca	Ds	Hauteur	0.9069730	0.0040346
Ma	Ds	Hauteur	0.9645802	0.0988511

Ces résultats semblent confirmer ceux précédemment édictés. Ainsi nous devons procéder à des transformations de type logarithmiques, box-cox, etc. Mais cela risque modifier la structure des données. Aussi, cette hypothèse n'est vérifiée que pour 3 croisements sur 6. Cependant, les métriques qui seront effectuées pour l'Anova plus tard sont robustes. Donc elles pourront palier à ces problèmes.

2.4.2 Hypothèse de sphéricité

L'hypothèse de sphéricité est une hypothèses importante dans les analyses de variance à mesures répétées. Elle concerne la structure de la matrice de variance-covariance des différences entre les mesures répétées de la variable dépendante (ou variable cible) pour chaque combinaison de niveaux des facteurs.

Plus précisément, l'hypothèse de sphéricité stipule que la variance des différences entre chaque paire de mesures répétées est la même pour toutes les paires. Cela signifie que la covariance entre les différences entre deux mesures répétées doit être identique pour toutes les paires de mesures répétées.

Si cette hypothèse est respectée, les résultats d'une ANOVA à mesures répétées peuvent être interprétés de manière fiable. Cependant, si elle n'est pas respectée, cela peut entraîner des erreurs de type I (faux positifs) ou des résultats erronés dans l'analyse statistique.

Il existe plusieurs tests pour évaluer l'hypothèse de sphéricité, tels que le test de Mauchly ou le test de Huitfeldt-Gabriel. Si l'hypothèse de sphéricité n'est pas respectée, des corrections peuvent être appliquées, telles que l'utilisation de l'ajustement de Greenhouse-Geisser ou de l'ajustement de Huynh-Feldt pour ajuster les degrés de liberté et obtenir des p-values plus précises.

Elle sera automatiquement vérifiée lors du test d'analyse de la variance via les fonctions **Anova_test** ou **aov_car** de Rstudio. Elle est déjà intégrée à l'intérieur de ces dernières.

Test d'homogénéité de la variance

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group  23  0.9193 0.5728
      229
```

On a une p-valeur supérieure au seuil de 5% donc on accepte l'hypothèse d'homogénéité des variances.

3 Modélisation, résultats et interprétation

Nous allons tout d'abord effectuer une analyse de la variance à un facteur pour chaque facteur (variété et fertilisant) pour déterminer si chaque facteur a un effet significatif sur la taille des plantes. Si l'un ou l'autre des facteurs n'est pas significatif, cela signifie qu'il n'y a pas de différence significative entre les modalités de ce facteur.

3.1 Anova à un facteur à mesures repétées pour le facteur Variété

Nous postulons le modèle suivant au vue des hypothèses ci-dessus vérifiées :

$$Y_{jk} = \mu + \beta_j + +\varepsilon_{jk}$$

Pour le cas d'espèce, on peut utiliser la fonction aov qui nous fournit directement les résultats de l'estimation et des tests statistiques du modèle ci-dessous. Les autres fonctions aboutissant à la même conclusion quant au test de significativité.

```
Error: Repetition
          Df Sum Sq Mean Sq F value Pr(>F)
Varietes   1   61.8    61.8   0.166  0.723
Residuals  2  744.3   372.2

Error: Repetition:Varietes
          Df Sum Sq Mean Sq F value Pr(>F)
Varietes   1   1290  1289.7   3.807  0.146
Residuals  3   1016   338.8

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 245  13682   55.85
```

On constate donc que la p-value (0.146) est supérieur au seuil de 5%. On a donc confiance à 95% que le facteur **variété** n'est pas significatif. cela signifie qu'il n'y a pas de différence significative entre les modalités de ce facteur. Pour chaque modalité de ce facteur, les effets sont les mêmes sur la hauteur des plantes pour chaque mesure. Autrement dit, la hauteur des plantes n'est pas forcément l'effet du type de variété.

3.2 Anova à un facteur à mesures repétées pour le facteur Fertilisation

Soit le modèle ci-dessous

$$Y_{ik} = \mu + \alpha_i + + \varepsilon_{ik}$$

On utilise également la fonction aov :

```
Error: Repetition
             Df Sum Sq Mean Sq F value Pr(>F)
Fertilisation  2  334.0   167.0   0.354  0.765
Residuals      1  472.2   472.2

Error: Repetition:Fertilisation
             Df Sum Sq Mean Sq F value Pr(>F)
Fertilisation  2  266.4  133.22   2.042  0.211
Residuals      6  391.4   65.23

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 241  15330   63.61
```

La p-value étant également supérieure au seuil, il n'y a donc pas de différence significative entre les modalités du facteur fertilisation sur la croissance des plantes. Pour chaque modalité de ce facteur, les effets sont les mêmes sur la hauteur des plantes pour différentes mesures.

3.3 Analyse de la variance à deux facteurs à mesures répétées avec interaction

Au regard des précédents résultats, on peut donc effectuer une ANOVA à deux facteurs à mesures répétées afin de voir les effets principaux de la variété et du fertilisant, ainsi que l'interaction entre les deux facteurs. L'interaction est importante car elle indique si l'effet de la variété sur la taille des plantes dépend du fertilisant utilisé (et vice versa).

Soit le modèle suivant :

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

Nous avons les degrés de libertés ci-dessous :

Table 14: Degrés de liberté

Source	Degrés de liberté
Facteur A	$n_A = I - 1$
Facteur B	$n_B = J - 1$
Interaction AB	$n_{AB} = (I-1)(J-1)$
Résiduelle	$n_R = IJ(K-1)$
Totale	$n_{TOT} = IJK - 1$

Table 15: Tableau d'analyse de la variance-Rappel

Source	Variation	Ddl	Carré Moyen	F	Décision
Facteur A	sc_A	n_A	$s_A^2 = \frac{sc_A}{n_A}$	$f_A = \frac{s_A^2}{s_R^2}$	$\mathcal{H}_0'$ ou $\mathcal{H}_1'$
Facteur B	sc_B	n_B	$s_B^2 = \frac{sc_B}{n_B}$	$f_B = \frac{s_B^2}{s_R^2}$	$\mathcal{H}_0''$ ou $\mathcal{H}_1''$
Interaction	sc_{AB}	n_{AB}	$s_{AB}^2 = \frac{sc_{AB}}{n_{AB}}$	$f_{AB} = \frac{s_{AB}^2}{s_R^2}$	$\mathcal{H}_0'''$ ou $\mathcal{H}_1'''$
Résiduelle	sc_R	n_R	$s_R^2 = \frac{sc_R}{n_R}$		
Totale	sc_{TOT}	n_{TOT}			

Le tableau d'analyse de la variance se présente comme suit :

En faisant une application numérique, on obtient :

Table 16: Analyse de la variance

Effect	df	MSE	F	ges	p.value
Fertilisation	1.79, 5.36	6.40	3.31	.152	.117
Varietes	1, 3	29.64	3.22	.311	.171
Fertilisation:Varietes	1.54, 4.63	4.57	0.31	.010	.698

Le tableau ci-dessus est celui de l'analyse de la variance pour les deux facteurs et avec interaction entre les deux. Fort est de constater que l'interaction entre les deux n'est pas significative, effectuer des comparaisons post-hoc pour déterminer les différences significatives entre les modalités des deux facteurs pour chaque niveau du facteur restant n'est donc pas nécessaire. On peut dire en conclusion que l'effet de la variété sur la taille des plantes ne dépend pas du fertilisant utilisé et vice versa.

Comme mentionné dans les sections précédentes, l'hypothèse de sphéricité sera automatiquement vérifiée lors du calcul du test ANOVA en utilisant la fonction R anova_test() ou aov_car (ci-dessus). Le test de Mauchly est utilisé en interne pour évaluer l'hypothèse de sphéricité. Aussi, la correction de sphéricité de Greenhouse-Geisser est automatiquement appliquée aux facteurs qui violent l'hypothèse de sphéricité.

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                      Sum Sq num Df Error SS den Df   F value    Pr(>F)
(Intercept)          28299.0      1   67.421      3 1259.2147 4.921e-05 ***
Fertilisation           37.9      2   34.329      6    3.3137    0.1073
Varietes                95.5      1   88.926      3    3.2221    0.1705
Fertilisation:Varietes   2.2      2   21.131      6    0.3068    0.7467
```

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


Mauchly Tests for Sphericity

                      Test statistic p-value
Fertilisation                0.88154 0.88154
Fertilisation:Varietes       0.70351 0.70351


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                      GG eps Pr(>F[GG])
Fertilisation         0.89408     0.1174
Fertilisation:Varietes 0.77131     0.6983

                        HF eps Pr(>F[HF])
Fertilisation         2.125981  0.1072770
Fertilisation:Varietes 1.430837  0.7466771
```

Mauchly Tests for Sphericity représente le test de sphéricité de Mauchly qui est donc significatif. L'hypothèse de sphéricité est bien vérifiée. Ensuite, nous avons la correction de Greenhouse-Geisser et les degrés de libertés.

Validation des résultats

Normalité des résidus

```
Data was changed during ANOVA calculation. Thus, residuals cannot be added to origin
residuals(..., append = TRUE) will return data and residuals.

Shapiro-Wilk normality test

data:  residus
W = 0.91606, p-value = 0.04783
```

Au seuil de 1% et de 2%, de 3 et de 4 les résidus suivent une loi normale

conclusion

La présente étude avait pour objectif d'étudier l'influence du type de terreau et de la variété sur la croissance des plantes. Il en ressort que quelque la soit la combinaison des facteurs : à un facteur, deux facteurs avec ou sans interaction, il y a absence d'effets de groupe des modalités prises croisement pour les deux facteurs. Et prises individuellement. En d'autres, termes elles ont les mêmes effets sur la croissance des plantes. Comme susmentionné ci-haut : l'effet de la variété sur la taille des plantes ne pas dépend du fertilisant utilisé et vice versa. Certaines hypothèses étant violées pour un certain seuil, il convient de pousser l'analyse plus loin pour conforter les résultats. Par exemple en considérant de l'Anova non paramétrique.