Document d'accompagnement thématique

MINISTÈRE DE L'AGRICULTURE ET DE LA SOUVERAINETÉ ALIMENTAIRE
Liberté Égalité Fraternité

L'AVENTURE DU VIVANT

Inspection de l'enseignement agricole

Diplôme : BTSA ACD

Thème : Exemples d'utilisation des mathématiques dans des situations favorisant l'acquisition de capacités professionnelles

Commentaires, recommandations pédagogiques

L'enseignement des mathématiques doit contribuer, notamment en lien avec les disciplines professionnelles, à l'acquisition des capacités :

C51- Conduire une expérimentation factorielle

C52- Suivre une expérimentation système

L'enseignement des mathématiques vise à donner une assise scientifique permettant de développer l'esprit critique devant les résultats d'expérimentation ou encore de communiquer des résultats chiffrés sous une forme adaptée. L'enseignement s'appuie sur les acquis des étudiants pour développer de nouveaux outils mathématiques dans le but de répondre à des problématiques professionnelles. La mobilisation de ces outils dans le cadre de la résolution de problèmes concourt à la validation des capacités professionnelles susvisées.

L'enseignement des mathématiques est étroitement lié à l'enseignement des disciplines professionnelles. Sa mise en œuvre s'appuie sur les situations professionnelles enseignées. Les contextes doivent varier en fonction des situations techniques et provenir de documents issus de sources multiples : l'INSEE, AGRESTE, compte rendu des directeurs d'exploitation de l'établissement, documentations, résultats issus de projets.

La progression construite par le professeur de mathématiques devra être en lien direct avec celle proposée par les collègues de disciplines professionnelles.

La résolution de problèmes demande de mobiliser des techniques calculatoires. Les calculs, pour une grande partie, peuvent être délégués à un outil de calcul. Il ne s'agit pas ici de développer une virtuosité procédurale mais plutôt de se positionner comme observateur et de se questionner sur les processus mis en œuvre dans le domaine professionnel. La recherche de réponses amène naturellement à élaborer des démarches, à mener des calculs à l'aide d'un outil adapté, à s'assurer de la cohérence de résultats et à prendre des décisions.

L'institutionnalisation des notions, phase indispensable dans le processus d'apprentissage, a pour but d'expliciter les savoirs et les savoir-faire qui ont été mobilisés pendant la séance ou séquence, de donner des repères simples aux apprenants. Ce temps doit être court et synthétique. Les développements théoriques sont réduits à l'essentiel et toujours présentés dans un cadre accessible.

Des mathématiques transversales à tous les blocs de compétences.

L'acquisition des capacités professionnelles demande d'aborder de nouvelles notions qui s'appuient de façon implicite sur des connaissances mathématiques acquises dans les classes antérieures du collège et du lycée. Certaines difficultés d'apprentissage de ces nouveaux concepts proviennent d'un manque de maîtrise de ces prérequis. Il est indispensable d'y consacrer régulièrement du temps afin de réactiver et consolider ces savoirs sans entrer dans un schéma de révision. Le choix de réinvestir les notions transversales suivantes est décidé en fonction de la progression définie en cohérence avec les disciplines professionnelles :

- Proportion, pourcentage et proportionnalité,
- Sens des opérations, application de formule, représentation graphique de fonctions et exploitation graphique,
- Représentations de diagrammes statistiques pertinents, interprétation et utilisation d'indicateurs statistiques,
- Probabilités élémentaires, lien entre fréquences et probabilités, arbres de probabilités.

Afin que les élèves soient aguerris aux pratiques calculatoires élémentaires favorisant l'acquisition des capacités, des automatismes mathématiques doivent être développés par un travail régulier, afin d'obtenir une aisance suffisante, en s'appuyant préférentiellement sur des situations en lien avec les disciplines professionnelles.

Au-delà d'une pratique dans toutes les activités de la classe, il est aussi important d'entretenir ces automatismes par des rituels de début de séance, sous forme de « questions flash » privilégiant l'activité mentale avec un recours à des connaissances, des procédures, des méthodes et des stratégies fondamentales dans la pratique professionnelle. Cela ne doit pas faire l'objet d'un chapitre d'enseignement spécifique car les notions qui les sous-tendent ont été travaillées dans les classes antérieures. Cette pratique, propre à chaque enseignant, doit s'adapter aux besoins de la spécialité.

Les exemples ci-dessous ne sont pas exhaustifs mais donnent une orientation de ce qui peut être fait.

Parmi elles, certaines doivent être propices au calcul mental.

- Sens des opérations qui permet d'effectuer des calculs courants.
- Calculer une moyenne, une moyenne pondérée.
- Passer d'une proportion (1/2, 3/4, 1/5, …) à un pourcentage (50 %, 75 %, 20 %, …) et inversement.
- Calculer des pourcentages, d'un prix TTC à partir d'un prix HT et inversement, avec des taux de TVA différents.
- Lier augmentation et diminution en pourcentage avec coefficient multiplicateur et les utiliser en situation.
- Comparer en situation des proportions et des pourcentages.
- Appliquer des formules et déterminer la valeur numérique d'une grandeur connaissant les autres.
- Reconnaître graphiquement des fonctions de référence, en décrire les variations et les extremums.
- Lire graphiquement la pente d'une droite, la pente en un point de la représentation graphique d'une fonction, repérer les points d'inflexion et la concavité d'une courbe en lien avec les termes de la vie économique «'inflexion de la courbe de croissance », « accélération de la croissance » ou encore « accélération de la baisse »…
- Choisir une représentation graphique adaptée pour représenter des données, des proportions ou des pourcentages (graphique, diagramme circulaire, semi-circulaire, diagramme en bâton ou en barres, barres empilées, …).

- Inversement, interpréter des diagrammes et retrouver des données statistiques à partir de représentations.

Les outils numériques doivent être intégrés à l'enseignement des mathématiques. Ils apportent une plus-value permettant d'aborder de véritables problèmes issus des situations professionnelles. L'usage des outils numériques tels que le tableur, les logiciels de traitement de données statistiques, de sondage, de cartographie, … doit être pensé dans l'optique de résoudre des problèmes qui n'auraient pas été accessibles sans ses outils. La maîtrise des outils numériques n'est pas un but de l'enseignement des mathématiques. La calculatrice reste aussi un outil facilement mobilisable en classe. Cela n'est pas contradictoire avec une pratique du calcul mental régulière mais raisonnée, tant par la difficulté des questions posées que le contexte de sa pratique.

C51 – Conduire une expérimentation factorielle
C52 – Suivre une expérimentation système

L'expérimentation est une source de données pour l'agronomie. L'expérimentation factorielle est portée par une vision analytique des phénomènes. Elle est principalement réalisée dans le but de tester une hypothèse relative à l'effet d'un ou plusieurs facteurs dont on souhaite améliorer l'efficacité ou l'efficience dans un processus de production et est basée sur des comparaisons de modalités et des répétitions. Elle est mise en place par exemple pour choisir une ou des variétés, pour affiner l'efficacité d'un facteur (par exemple doses d'engrais).
L'expérimentation factorielle, appuyée par les analyses statistiques qui lui sont attachées (l'analyse de variance en premier lieu) permet d'apporter une réponse sur l'influence d'un facteur dans la production en classant statistiquement entre eux les modalités étudiées, toutes choses égales par ailleurs.

L'enseignement du module M5 doit s'appuyer sur des situations concrètes, des retours d'expérience ou des essais. Il s'agit de mettre en œuvre des outils statistiques d'aide à la décision, ou à la mesure de l'influence de facteurs. Le travail sur ce module étant conduit sur un temps long, il paraît donc essentiel de faire émerger les méthodes statistiques à partir de simulations. Le point de départ est la loi de Bernoulli et la loi binomiale. Le théorème central limite est le théorème sous-jacent. Il n'est pas nécessaire de l'énoncer mais par contre il est indispensable de l'illustrer pour diverses situations avec différentes lois. L'importance de la loi normale doit alors apparaître. Il ne s'agit pas ici de développer une grande technicité sur la loi normale mais plutôt de travailler sur la reconnaissance de la forme de la fonction densité de probabilité et la lecture graphique des paramètres. La symétrie de la courbe permet de dégager des propriétés simples. Les outils numériques ont dans leur grande majorité les lois normales implémentées, il est donc impératif de se séparer des tables de lois normales et du recours systématique au changement de variable. Le théorème central limite amène à s'interroger sur le passage du discret au continu et donc à développer la notion de loi continue, majoritairement inconnue des étudiants.

Le passage des lois de probabilité aux statistiques se fait naturellement en posant le problème de l'estimation. L'outil numérique permet d'effectuer un grand nombre de simulations en faisant varier les paramètres. Ceci débouche sur l'étude de la fluctuation d'échantillonnage des indicateurs d'une population donnée. La question inverse de remonter aux indicateurs de la population à partir d'un échantillon émerge alors. L'estimation ponctuelle et par intervalle de confiance de certains indicateurs d'une population est indispensable pour prendre des décisions. Sans entrer dans l'étude théoriques des biais et du calcul des variances des estimateurs, il peut être intéressant de soulever le problème de la qualité des estimateurs en montrant sur des exemples l'efficacité de deux estimateurs d'un même paramètre.

D'autres situations amènent à expliciter la corrélation entre deux grandeurs quantitatives. Encore une fois, la représentation graphique des données doit être le fil conducteur de la réflexion. Le choix du

modèle d'ajustement linéaire, quadratique, logarithmique et exponentiel doit être initié par l'observation de la forme du nuage de points. Ce choix peut alors être confirmé par le calcul d'un indicateur.

Par ailleurs, l'enseignement doit concourir à développer la capacité à repérer des situations de référence de mise en œuvre de tests statistiques. L'objectif est moins de faire apprendre un catalogue de tests statistiques que de faire comprendre la méthodologie des tests et la construction de règles de décision s'appuyant sur la fluctuation d'échantillonnage de certaines grandeurs obtenues en premier lieu par simulation. La connaissance de certaines lois de probabilité de grandeurs lors de la variabilité des échantillons est l'aboutissement d'un travail préparatoire effectué par des simulations. Les tests doivent être adaptés aux situations rencontrées par les étudiants. La diversité des situations permet d'ancrer l'utilisation des tests. Tous les calculs sont laissés à l'outil numérique. Même si son apprentissage peut être laborieux, le logiciel R fournit un grand nombre d'outils permettant de répondre à toutes les demandes et en particulier de travailler avec des données provenant de véritables expérimentations. Le travail est centré sur la reconnaissance des situations et le choix des méthodes.

Pour assurer la validité de l'analyse statistique, trois principes doivent impérativement être appliqués à la planification expérimentale :

- Randomisation : l'allocation des traitements (combinaison de modalités ou niveaux de facteurs) aux unités expérimentales (par exemple les parcelles) doit être faite par un tirage aléatoire,
- Répétitions : chaque traitement doit être affecté à plusieurs unités, afin de pouvoir estimer une erreur expérimentale,
- Contrôle de l'erreur : il faut réduire la part non contrôlée de l'expérience, donc diminuer l'erreur expérimentale.

L'erreur peut être technique. Mais il existe une autre source d'erreur. On appelle erreur unitaire l'erreur due à l'hétérogénéité des unités expérimentales, appelée aussi erreur d'hétérogénéité. On l'appelle aussi erreur de randomisation, car c'est la randomisation qui en fait une variable aléatoire dont on peut étudier la distribution. On appelle erreur expérimentale ou erreur résiduelle la somme de l'erreur unitaire et de l'erreur technique. La randomisation permet d'éviter tout biais plus ou moins conscient. En termes mathématiques, elle permet d'assurer que l'erreur unitaire est d'espérance nulle. Il est important de savoir que la randomisation n'annule pas, ni même ne diminue l'erreur. Elle permet d'éviter qu'elle soit systématique et de connaître suffisamment sa distribution pour assurer la validité du test des effets du traitement. La randomisation doit être faite indépendamment pour chaque essai. Aucun plan ne doit être réutilisé sans nouvelle randomisation.

Le rôle des répétitions est de permettre d'évaluer la part de l'erreur expérimentale dans la construction de l'observation. En effet, sans des répétitions de chaque traitement sur plusieurs unités, on ne peut pas distinguer l'effet dû au traitement de l'erreur expérimentale. Enfin, il est nécessaire d'avoir une erreur expérimentale aussi petite que possible, c'est-à-dire de contrôler l'erreur.

Un préalable à beaucoup de tests est la normalité des variables. Parfois, la situation impose de fait la normalité des grandeurs, d'autres fois il sera peut-être nécessaire de débuter par un test de normalité.

L'enseignement doit amener à questionner les points suivants :

- Protocole de mesure de grandeurs, de constitution d'échantillons, d'enquête. Identifier un prélèvement aléatoire simple. L'échantillonnage aléatoire simple correspond à des tirages successifs équiprobables et indépendants les uns des autres.
- Affectation de traitements (ou modalité d'un facteur) à une unité expérimentale par randomisation. Cas d'un facteur et de deux facteurs (dont l'un est contrôlé).
- Identification de situation modèle. Identifier une situation modélisée par une loi binomiale, une situation où le modèle de la loi normale est pertinent. Approcher la normalité avec une technique empirique et une méthode graphique (histogramme des fréquences, boxplot, droite de Henry). L'enseignant peut compléter cette approche par des tests de normalité tels que Shapiro-Wilk ou Kolmogorov-Smirnov.
- Estimation des indicateurs d'une population à partir d'échantillons.
- Mise en œuvre de tests statistiques permettant de répondre à une problématique à partir d'essais réalisés par les apprenants ou d'études publiées. Les tests à pratiquer sont à choisir

de préférence dans les tests de conformité d'une proportion, d'une moyenne, de comparaison de proportions, de moyennes, de variances, d'indépendance du χ_2 et d'analyse de la variance à un ou deux facteurs ainsi que le test de Newman-Keuls ou autres tests post-hoc adaptés à la situation étudiée.

- Communication de résultats d'études. Présenter des résultats sous forme synthétique. Choix du type de représentation (tableau, arbre, carte mentale, courbe, …). La construction des graphiques est réalisée à l'aide de logiciels. C'est la pertinence du choix qui guide les apprentissages et non la technique de construction.

Exemple 1 : Randomisation d'essais

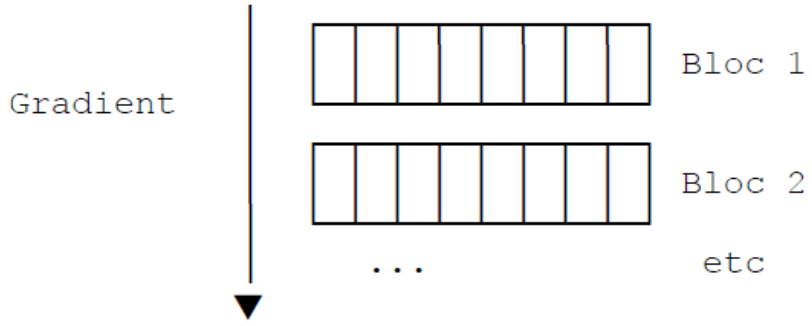
Rappelons que la randomisation répond au contrôle de l'erreur. Cet objectif doit être le fil conducteur lors de l'enseignement des différentes procédures de randomisation.

1er cas : Plan en randomisation totale

La situation de référence est la suivante. On considère un champ expérimental composé de N unités et un facteur composé de t modalités ou niveaux (les traitements). On note n le nombres d'unités expérimentales auxquelles est affecté chaque traitement (on a $N = nt$). Le dispositif est dit en randomisation totale si l'affectation des traitements se fait par un tirage aléatoire équiprobable parmi l'ensemble des unités expérimentales.
L'enseignement doit conduire à questionner la façon de réaliser cette randomisation totale. Quel modèle mathématique est sous-jacent ? Quelle modalité de mise en œuvre ? A la « main » avec par exemple un tirage dans une urne de papiers numérotés et évidemment avec l'outil numérique.

2e cas : Plan en blocs complets

La manière la plus courante de contrôler l'erreur d'hétérogénéité est de constituer des blocs. Par définition, un bloc est un groupe homogène d'unités expérimentales. C'est à dire que la variabilité du phénomène étudié doit être plus faible entre unités appartenant à un même bloc qu'entre unités appartenant à des blocs différents. Le regroupement des unités expérimentales en blocs résulte d'une information a priori dont on dispose sur l'espace physique de l'expérimentation. Le cas le plus classique se trouve lorsqu'il y a un gradient, induit par une pente, ou par une lisière brise-vent ou apportant de l'ombrage, ou bien par la proximité d'un cours d'eau…



Les blocs sont alors allongés perpendiculairement au gradient, les parcelles étant allongées dans le sens de ce gradient.
La procédure de randomisation consiste en un tirage aléatoire par bloc pour affecter les traitements aux unités expérimentales. On se limite ici au cas le plus simple où chaque traitement figure une fois et une seule dans chaque bloc. Donc un bloc comporte t traitements, si n est le nombre de blocs, le nombre total d'unités expérimentales (parcelles) est $N = n\,t$.
Ici encore, l'enseignement doit conduire à questionner la façon de réaliser cette procédure de randomisation. Comment faire à la main ? Comment faire avec un générateur de nombres pseudo-aléatoires ? L'outil numérique fournit aussi des fonctions pour choisir aléatoirement une permutation.

3e cas : Plans en carré latin

Ce type de dispositif permet un double contrôle de l'hétérogénéité, par opposition au plan en randomisation totale qui ne contrôle pas l'hétérogénéité et au plan en bloc qui permet un simple contrôle

de celle-ci. On considère simultanément des blocs dans le sens des lignes et des blocs dans le sens des colonnes.
Dans un plan en carré latin on impose, d'une part, que le nombre de répétitions soit égal au nombre de traitements (combinaison de modalités ou niveaux), d'autre part, que chaque traitement soit placé une fois et une seule dans chaque ligne et dans chaque colonne
Si on note t le nombre de traitements, on doit avoir $n = t$ nombre d'unités expérimentales élémentaires et $N = t^2$ d'unités.
La procédure de randomisation est la suivante : on commence par construire un carré latin arbitraire, puis on randomise les lignes, enfin on randomise les colonnes pour obtenir le dispositif définitif. À titre d'exemple, on considère un carré latin à cinq traitements (notés A, B, C, D et E) :

Carré latin arbitraire

	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

Randomisation des lignes

	1	2	3	4	5
3	C	D	E	A	B
2	B	C	D	E	A
1	A	B	C	D	E
5	E	A	B	C	D
4	D	E	A	B	C

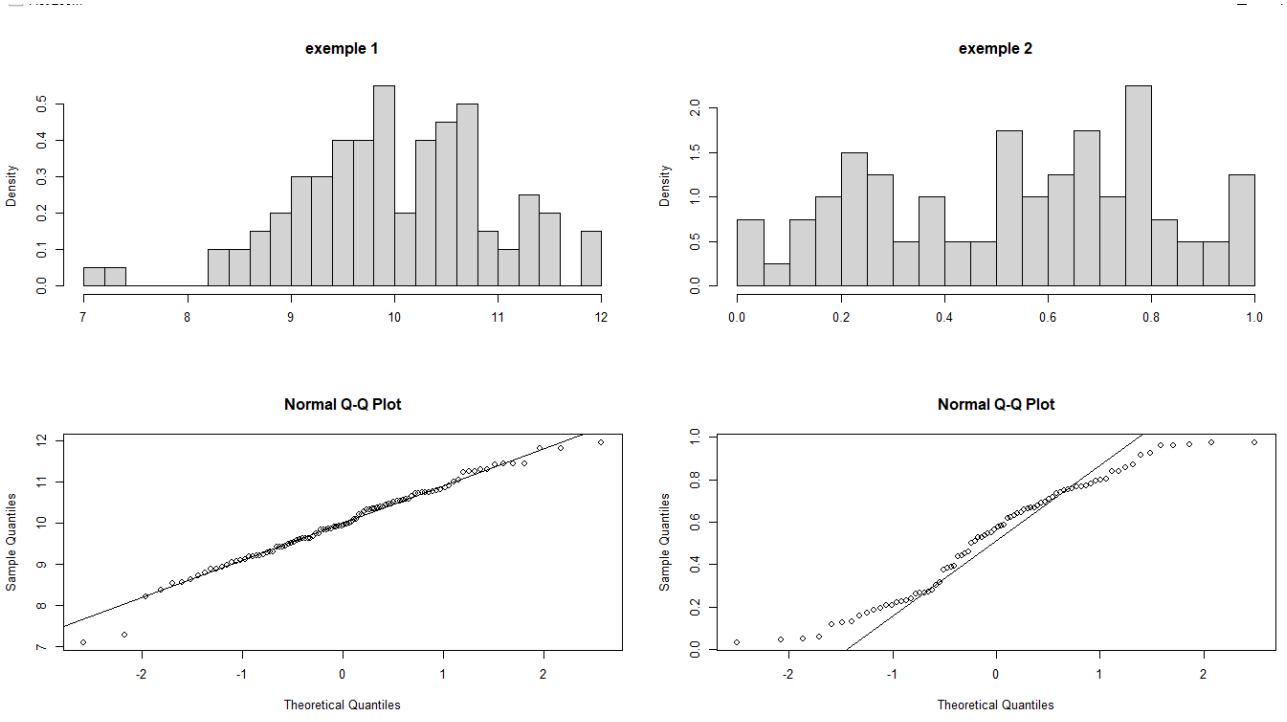
Randomisation des colonnes

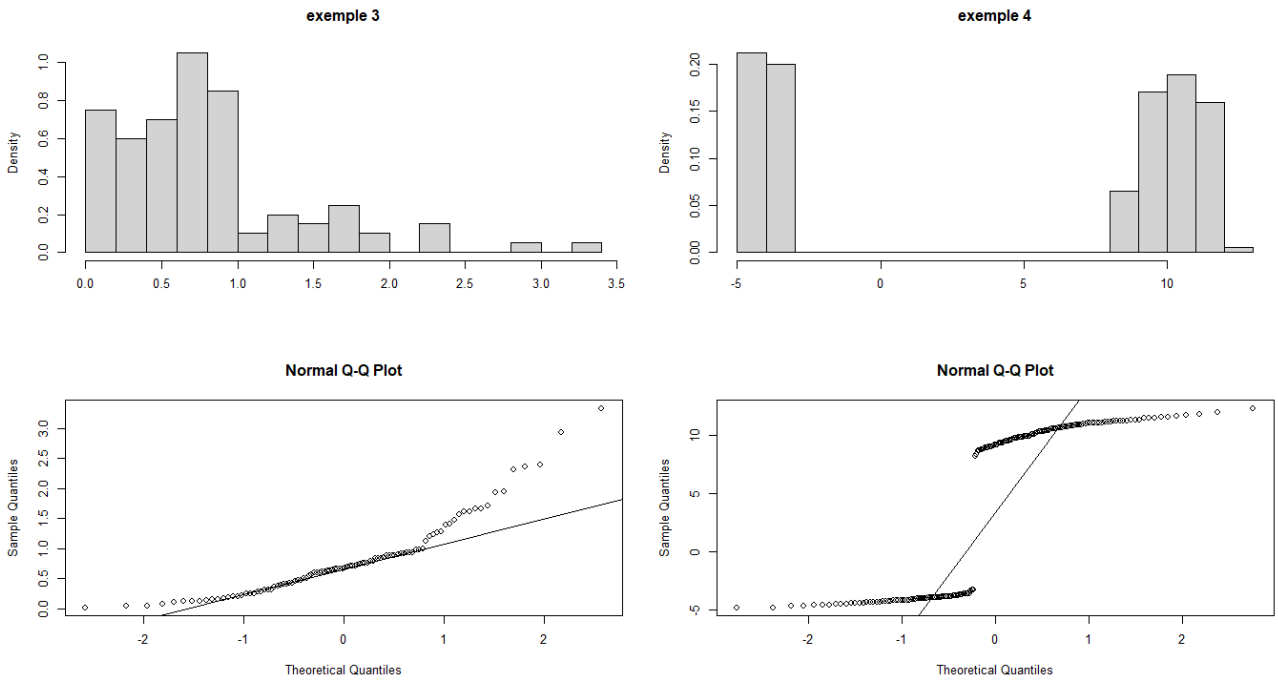
	5	2	1	4	3
3	B	D	C	A	E
2	A	C	B	E	D
1	E	B	A	D	C
5	D	A	E	C	B
4	C	E	D	B	A

L'enseignement doit amener à construire une procédure efficace dans la construction d'un carré latin. Il est possible d'y associer de l'algorithmique et de la programmation en Python ou en R.

Il existe d'autres plans d'expérience que l'on peut aborder si la situation s'y prête comme le split-plot ou le criss-cross. Au-delà de leur procédure de construction, il est important de montrer comment ils répondent au contrôle de l'erreur expérimentale.

Exemple 2 : Tester la normalité
Un préalable à beaucoup d'études est la normalité des grandeurs en jeu. Pour une première approche on peut s'appuyer sur la forme des histogrammes des échantillons et exposer la méthode de la droite de Henry. Tous les graphiques sont obtenus à l'aide de l'outil numérique. Par exemple, la commande qqnorm() du logiciel R permet de tracer le graphique quantile-quantile qui confronte les quantiles de la loi normale en abscisse et les quantiles empiriques de l'échantillon en ordonnée. La commande qqline() construit la droite joignant le couple des quantiles 0,25 et le couple des quantiles 0,75.





Cette approche graphique peut être complétée par le test de Shapiro-Wilk obtenu directement par la commande shapiro.test() du logiciel R. On n'entre pas dans les détails de ce test. Il s'agit de développer un questionnement sur l'hypothèse de normalité au regard de son importance dans les conclusions du théorème central limite et de la somme de variables aléatoires indépendantes suivant une loi normale.

On trouve pour les exemples ci-dessus :

Exemple 1	Exemple 2	Exemple 3	Exemple 4
W = 0,98982	W = 0,93316	W = 0,93141	W = 0,7231
p-value = 0,6503	p-value = 0.0004218	p-value = 5.993e-05	p-value < 2.2e-16

Pour le test de Kolmogorov-Smirnov, voir la commande ks.test() du logiciel R.

Exemple 3 : Analyse de variance (ANOVA) à 1 facteur

On se place dans la situation où pour chaque modalité ou niveau du facteur, le nombre de répétitions est identique. Un essai portant sur 3 variétés de pois chiche a été réalisé. Pour chaque variété 20 plants sont étudiés. On s'intéresse à la hauteur des plants au moment de la récolte. On se pose naturellement la question de savoir s'il y a une différence significative de hauteur entre les 3 variétés. La formation doit amener à s'approprier différents types de présentation des résultats, soit sous la forme d'un tableau à 2 colonnes comme ci-dessous.

Variété de pois chiche	Hauteur du plant en cm
Variété 1	35
Variété 1	38
…	…
Variété 2	32
…	…
Variété 3	37
…	…

Soit sous la forme d'un tableau croisé faisant apparaître le facteur (en ligne ou en colonne) et pouvant faire apparaître les moyennes pour chaque variété comme ci-dessous.

Variété de pois chiche	**Variété 1**	**Variété 2**	**Variété 3**
	35	32	37
	38	37	33
	…	…	
Moyenne			

Répondre à la question de savoir si la variété a une influence sur la hauteur des plants amène à construire un test d'hypothèse qui s'appuie sur un modèle mathématique. Pour une bonne compréhension des enjeux et des hypothèses de l'ANOVA, il peut être intéressant de faire apparaître le modèle mathématique derrière cette situation.

On peut donc considérer chaque échantillon (ici par variété de pois chiche) comme issu d'une variable aléatoire X_i pour $i = 1,2,3$ d'espérance m_i . On est donc amené à tester $H_0 : m_1 = m_2 = m_3$ contre $H_1 : \exists i, j\ m_i \neq m_j$

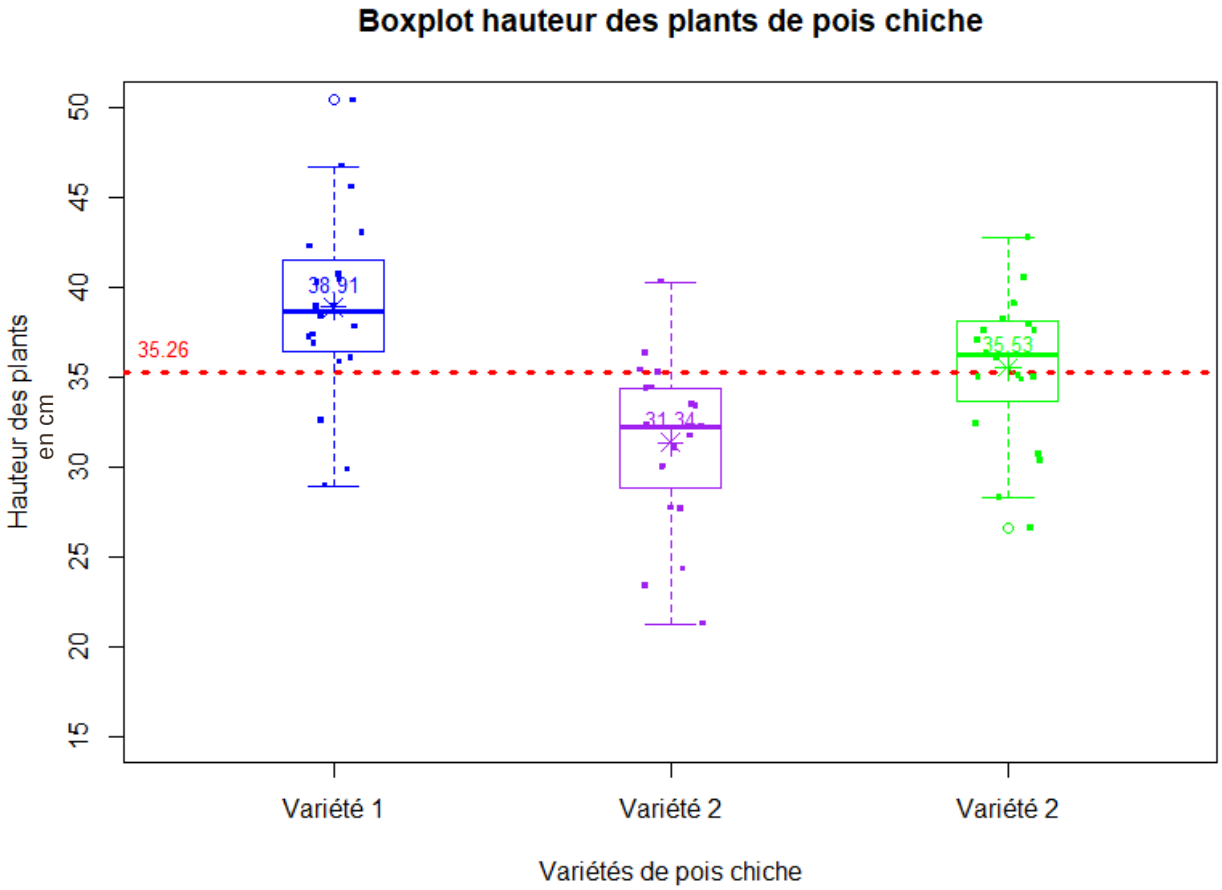
Pour réaliser ce test dans de bonnes conditions, des hypothèses sont à ajouter. Les X_i sont indépendantes et suivent des lois normales $N(m_i, \sigma)$. Le fait d'avoir une variance identique pour les 3 variables correspond à l'hypothèse que le facteur « variété » influe uniquement sur les moyennes.

Ce modèle peut être réécrit pour une meilleure compréhension sous la forme $X_i = m + \alpha_i + \varepsilon_i\ (i = 1,2,3)$ où m est la hauteur moyenne attendue des plants de pois chiche sans distinction de variété, α_i l'effet du niveau (ou modalité) i du facteur (variété) et ε_i la part de variabilité expérimentale. Les ε_i sont indépendantes et suivent la loi normale $N(0, \sigma)$. Dans cette écriture, le paramètre m s'appelle l'effet moyen général et les paramètres α_i sont appelés les effets principaux du facteur. Autrement dit, les α_i représentent le bonus ou le malus du niveau du facteur. Dans ce modèle, on impose comme contrainte $\sum \alpha_i = 0$ (dans le cas du même nombre d'unités expérimentales par modalités ou niveaux) qui dit uniquement que la moyenne des effets principaux du facteur est nulle. Le modèle sous cette forme permet d'appréhender la variabilité des résultats sous deux angles, celle due au facteur qui se voit dans α_i et celle due aux aléas de l'expérimentation (l'erreur expérimentale) qui se lit dans ε_i.

De cette réécriture du modèle, on est amené à tester $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ contre $H_1 : \exists i\ \alpha_i \neq 0$.

L'intérêt du modèle, au-delà de la compréhension de l'ANOVA, est aussi de pouvoir réaliser ses propres échantillons par simulation.

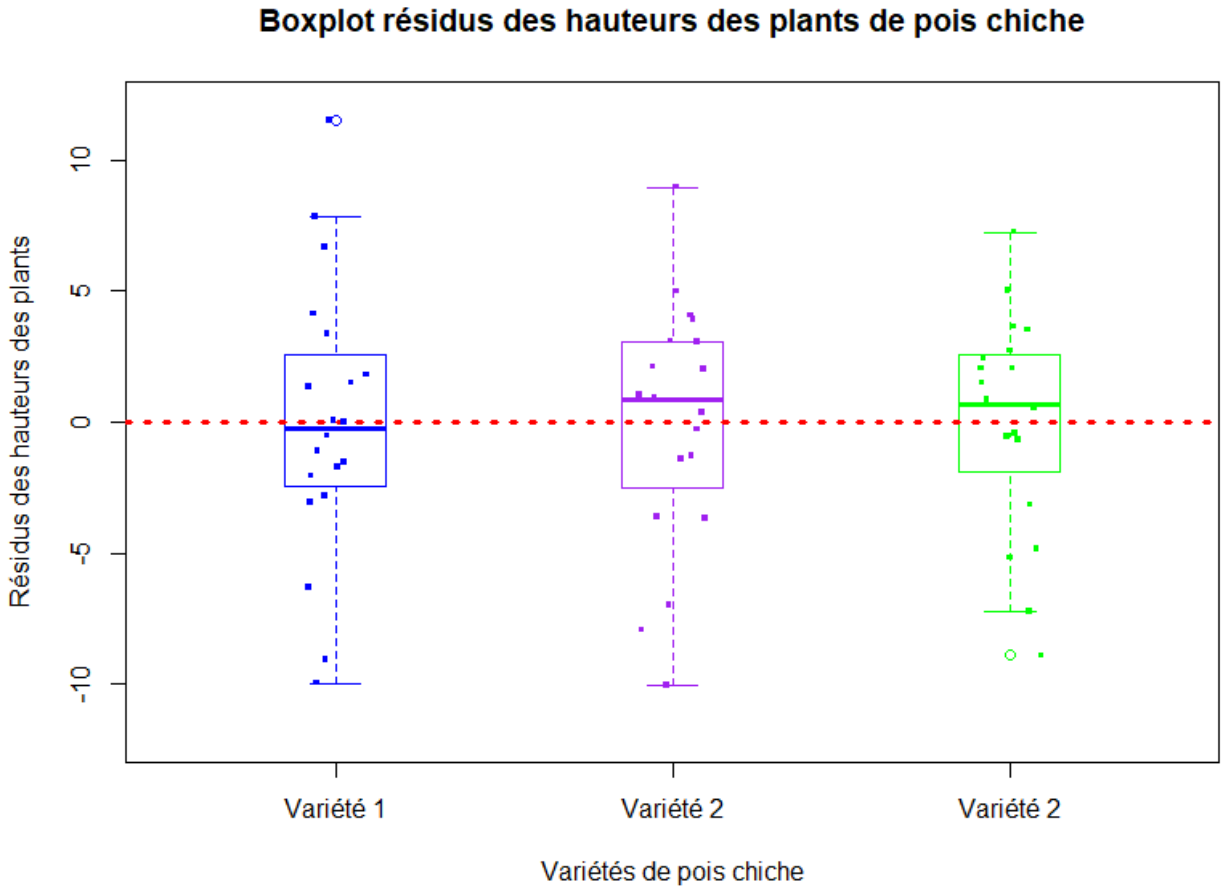
Pour vérifier ces hypothèses et se faire une première idée, la première étape indispensable est d'illustrer les données par une représentation appropriée faisant apparaitre la source de variabilité expliquée et la source de variabilité résiduelle. On obtient les boxplots ci-dessous. La ligne rouge matérialise la moyenne totale, les croix matérialisent les moyennes des trois échantillons correspondant aux trois environnements. Les écarts interquartiles permettent de se faire une idée de la dispersion et d'engager une discussion sur l'hypothèse d'homoscédasticité. D'un point de vue graphique, une dissymétrie remarquable des nuages de points par rapport à la médiane amène à s'interroger sur la normalité des échantillons. Cette observation peut être confirmée par le test de Shapiro-Wilk comme dans l'exemple1.



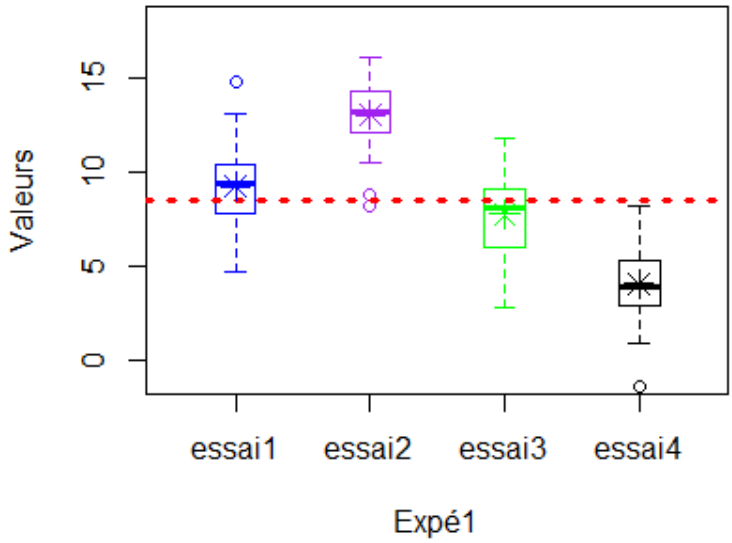
Ce travail n'est pas à négliger car il permet bien souvent d'émettre une conjecture à partir des représentations qui sera et qui devra être confirmée par un test. L'enseignement doit contribuer à développer le réflexe de commencer par des représentations. Les simulations permettent d'illustrer les situations variées.

Du point de vue du modèle $X_i = m + \alpha_i + \varepsilon_i\ (i = 1,2,3)$, les mesures sont des réalisations des variables aléatoires X_i. Si on note $(x_{i,j})$ pour $1 \leq i \leq 3$ et $1 \leq j \leq 20$ les mesures effectuées, l'estimation naturelle de m est $\bar{x} = \frac{1}{3\times 20}\sum_{i=1}^{3}\sum_{j=1}^{20} x_{i,j}$, celle des α_i est $\widehat{\alpha_i} = \bar{x_i} - \bar{x} = \frac{1}{20}\sum_{j=1}^{20} x_{i,j}$. On a donc $x_{i,j} = \bar{x_i} + e_{i,j}$ où $e_{i,j}$est la part variable autour de la valeur prédite que l'on nomme résidu. Ici, $\bar{x} \approx 35{,}26$ cm et $\widehat{\alpha_1} \approx 38{,}91 - 35{,}26 = 3{,}65$ cm

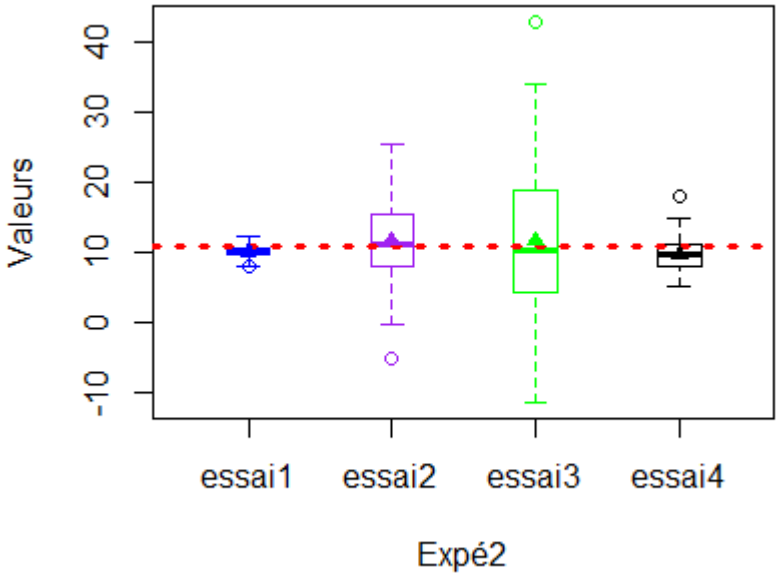
Afin de valider le modèle, on étudie la normalité et l'homoscédasticité soit des mesures brutes $x_{i,j}$ soit des résidus $e_{i,j}$. Le graphique des boxplots des résidus ci-dessous peut être plus simple à lire que celui des mesures brutes. Cependant, n'apparait plus la variabilité des moyennes.



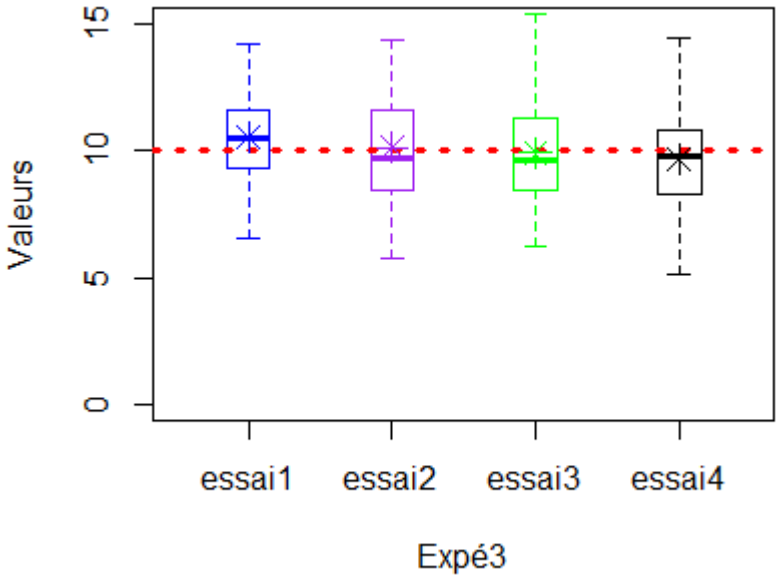
Dans les exemples suivants, nous avons pris le parti de présenter les échantillons dans leur globalité, mais on peut travailler uniquement sur les résidus pour la normalité et l'homoscédasticité. Ce premier exemple ci-dessous (Expé1) est obtenu en simulant des lois normales de même écart type mais de moyennes différentes. Ce premier exemple doit engager une discussion sur la variabilité des moyennes.



Le deuxième exemple (Expé2) est obtenu en simulant des lois normales de variances distinctes mais de même moyenne. Cet exemple doit engager une discussion sur l'homoscédasticité. La non homoscédasticité invite à s'interroger sur l'influence du facteur. Elle peut être confirmée par le test de Bartlett. Ce test peut très bien se réaliser sur les résidus plutôt que sur les mesures brutes. L'objectif de l'enseignement n'est pas d'apprendre un catalogue de test mais plutôt de développer une réflexion autour du traitement de données issues d'expérimentations, on peut suivant les cas se contenter de procédures graphiques pour vérifier les conditions d'homoscédasticité.



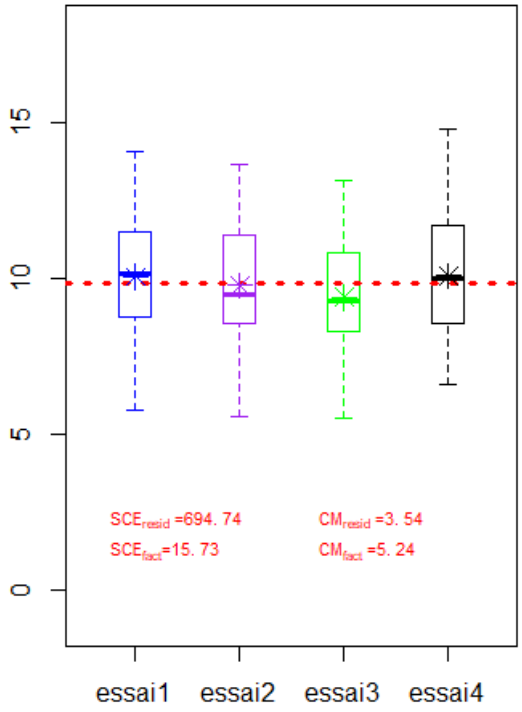
Ce troisième exemple (Expé3) est obtenu avec la même loi normale. La variété des simulations permet de développer le regard et l'interprétation de ces graphiques.

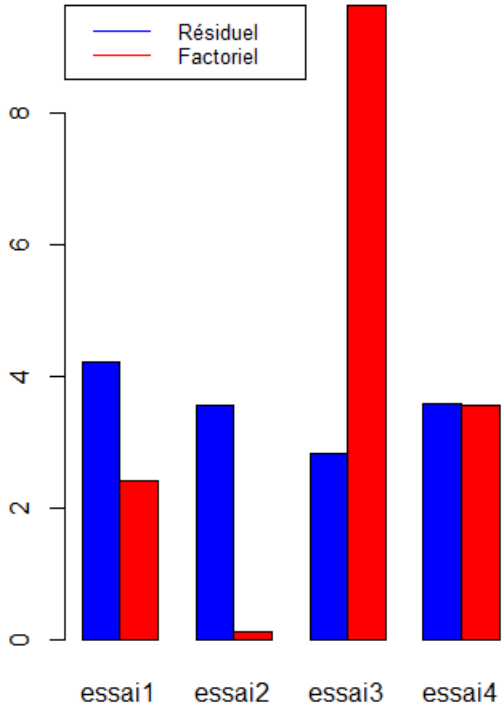


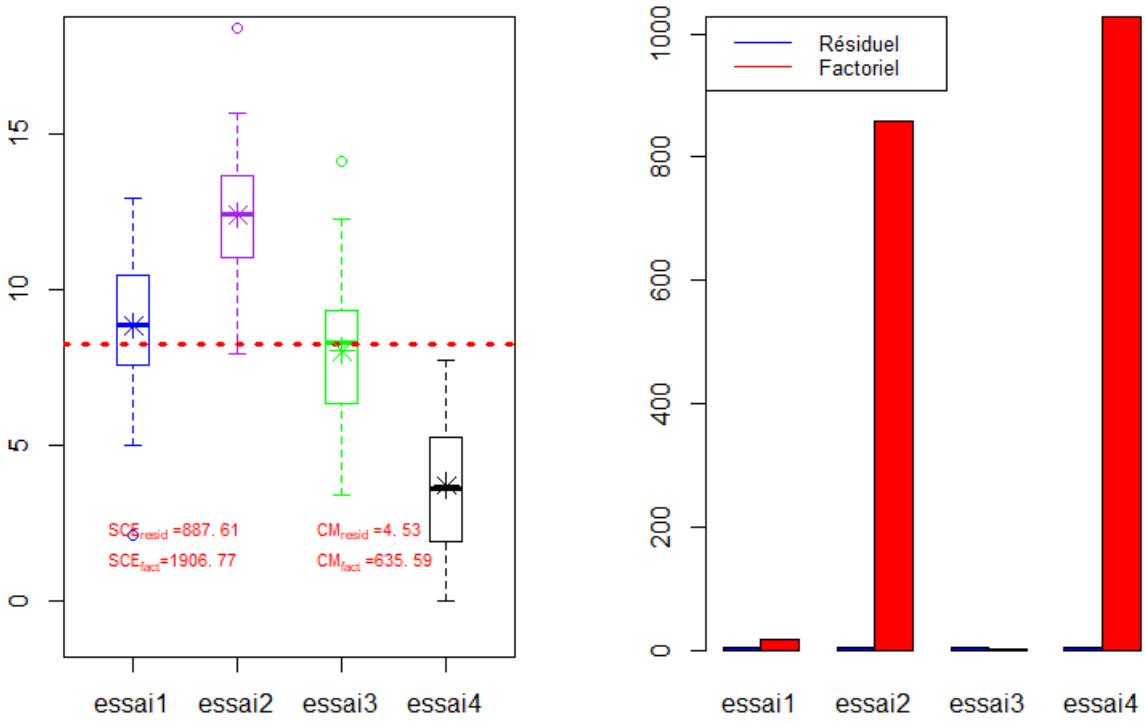
Cette approche par la simulation et les représentations doivent amener à la question de la variabilité des échantillons et des moyennes des échantillons et faire émerger l'équation de l'analyse de variance souvent présentée sous la forme :

$$SCE_{totale} = SCE_{factoriel} + SCE_{residuel} \ (SCE = \ somme\ des\ carrés\ des\ écarts).$$

Cette égalité n'est pas à démontrer mais doit être appréciée sur des exemples pour expliciter le comportement de SCE_{fact} et SCE_{resid} en fonction des paramètres (nombre de modalités du facteur, effectif). Ceci doit permettre d'arriver aux indicateurs que sont les carrés moyens souvent notés CM_{fact} et CM_{resid} qui apparaissent dans l'ANOVA. Il est alors intéressant d'illustrer de nouveau ces indicateurs via des simulations. Dans la suite nous nous plaçons dans le cas d'homoscédasticité. Le graphique de droite illustre les part de SCE_{fact}et de SCE_{resid} pour chaque essai.

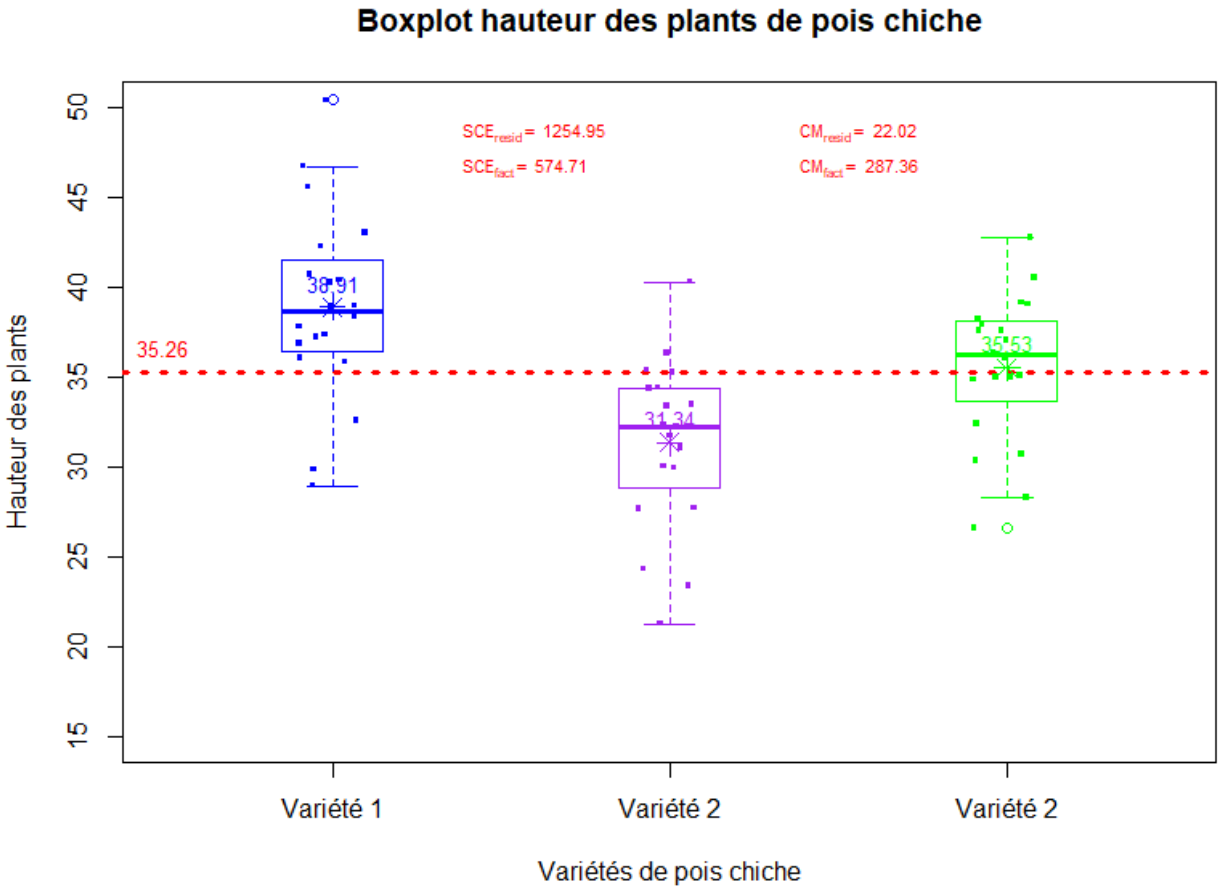






Les simulations et le calcul des CM_{fact} et CM_{resid} permettent de faire sentir qu'à variance constante dans les échantillons gaussiens simulés (ici l'écart type est égal à 2), CM_{resid} est plutôt stable alors que CM_{fact} varie fortement avec la variabilité des moyennes des échantillons. Cette constatation permet d'appréhender l'ANOVA et de considérer le rapport $\frac{CM_{fact}}{CM_{resid}}$. La discussion sur les degrés de liberté ne doit pas faire l'objet de justification théorique. On admet aussi l'utilisation de la loi de Fisher.

En reprenant l'exemple des plants de pois chiche, on obtient :



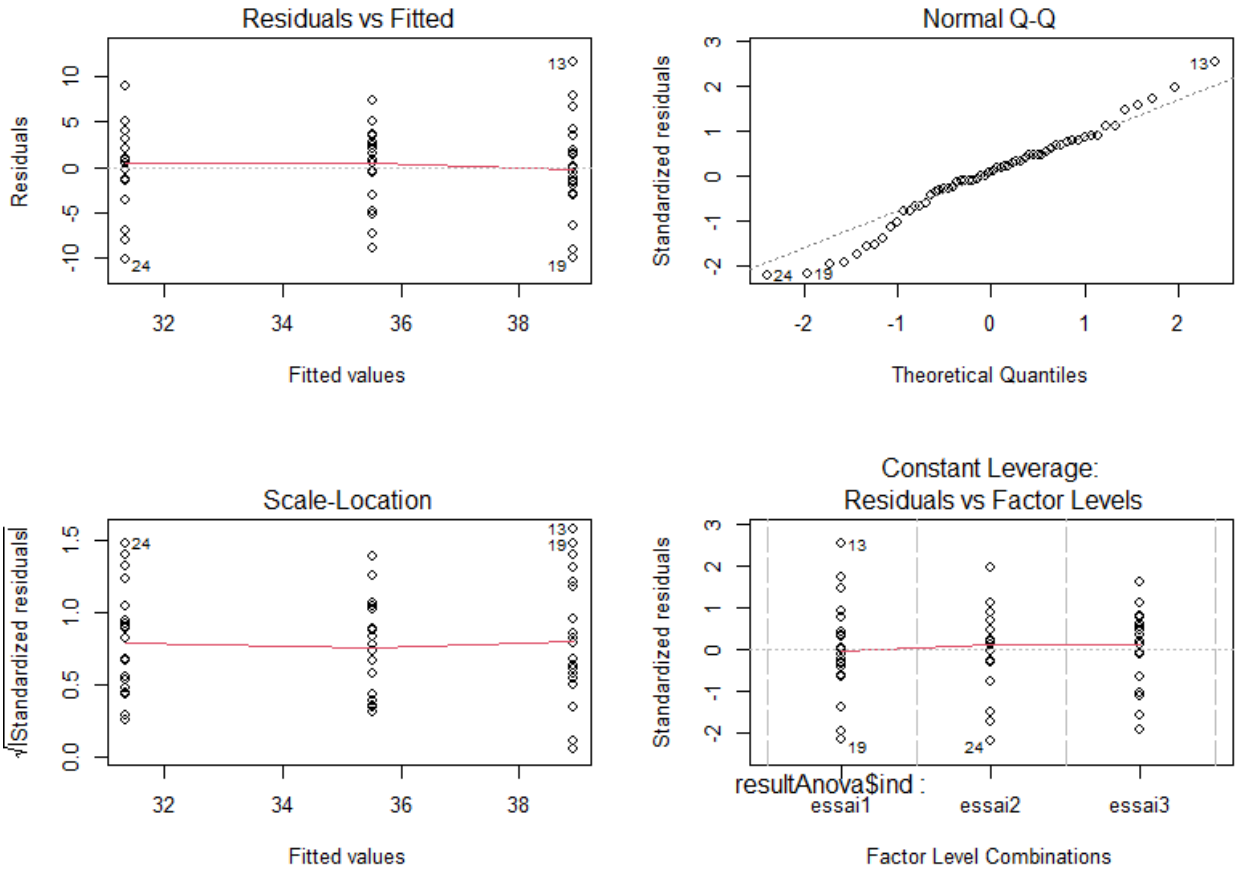
logiciel R donne le tableau d'analyse de variance suivant :

	Ddl	SC	CM	F value	P(>F)
Inter-groupes	2	574.7	287.36	13,05	2.15 e-05
Intra-groupes	57	1254.9	22.02		

L'hypothèse que le facteur n'a pas d'influence est alors rejetée.

Les conditions d'application de l'ANOVA, à savoir que le facteur n'influe que sur les moyennes des distributions (sous-entendu homoscédasticité) et que les échantillons sont gaussiens doivent être discutées et vérifiées. Pour la condition de normalité, voir l'exemple 2 « tester la normalité ».

Pour aller plus loin, le logiciel R permet de construire des graphiques diagnostiques. La normalité, l'indépendance et l'homoscédasticité se lisent sur les résidus.



Le 1er graphique en haut à gauche présente les résidus en fonction des moyennes des échantillons (ce qui revient à identifier les niveaux du facteur). On retrouve en abscisse les moyennes des échantillons 31.34, 35.53 et 38.91. Ceci nous ramène à la discussion sur la dispersion. Les points notés 24, 13 et 19 matérialisent des valeurs considérées aberrantes par l'algorithme dans R. Les nombres correspondent à l'indice de la valeur dans la liste afin de les retrouver facilement. On retrouve dans le 2e graphique (graphique quantile-quantile autour de la normalité des résidus standardisés) la discussion sur la normalité des relevés (voir exemple 2). Le 3e graphique en bas à gauche présente les racines carrées des résidus standardisés en fonction des moyennes. Ce graphique permet de lire l'homoscédasticité. La ligne rouge doit rester proche de l'horizontale pour avoir cette condition. Le dernier graphique sert à mesurer l'influence de certaines valeurs qui semblent aberrantes. On peut pour comprendre le rôle de ces graphiques effectuer une multitude de simulations avec des lois diverses.

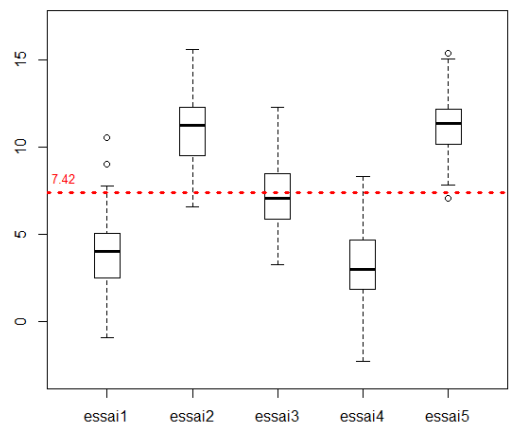
Lorsque l'ANOVA conduit à conclure à un effet significatif du facteur sur les moyennes, l'analyse n'est pas terminée. Nous savons qu'il y a au moins deux moyennes qui diffèrent l'une de l'autre, mais nous ne savons pas lesquelles. On est alors amené naturellement à vouloir effectuer des tests de comparaisons de deux moyennes. L'inflation du risque α doit être au cœur du débat. Plus les niveaux du facteur sont nombreux, plus il est nécessaire de réaliser de tests pour répondre à la question de savoir quelles sont les moyennes qui diffèrent, et plus le risque de conclure à tort à la significativité des différences est grand. Cette discussion pourrait aussi se faire plus tôt pour justifier l'utilisation de l'ANOVA au dépend de la multiplication de tests de comparaison de deux moyennes.

A titre d'exemple et pour comprendre le principe, il est possible de mettre en œuvre des tests de comparaisons de deux moyennes sur un facteur à trois niveaux et de discuter du risque α. Ceci amène naturellement à la correction de Bonferroni.

Sur ce principe, il existe plusieurs tests après une ANOVA (tests dits post-hoc). Il s'agit de procédures de comparaison multiple par étapes utilisées pour identifier des moyennes d'échantillonnage

significativement différentes les unes des autres. Suivant les situations et les disponibilités logicielles, on peut utiliser le test de Bonferroni, le test HSD de Tukey ou le test de Newman-Keuls.

Etudions un exemple à partir de simulation.



L'ANOVA réalisé sous R donne le tableau suivant qui indique une différence significative sur les moyennes.

```
                Df Sum Sq Mean Sq F value Pr(>F)
resultAnova$ind  4   2928   732.0   165.9 <2e-16
Residuals      245   1081     4.4
```

Le test de Bonferroni fournit le résultat ci-contre. Si on note H_0 : les deux essais ont même moyenne.
La comparaison de essai1 et essai4 donne une p-valeur de 1, de même pour la comparasion de essai2 et essai5.

```
        Pairwise comparisons using t tests with pooled SD

data:  resultAnova$values and resultAnova$ind

       essai1  essai2  essai3  essai4
essai2 < 2e-16 -       -       -
essai3 8.9e-13 < 2e-16 -       -
essai4 1       < 2e-16 2.7e-16 -
essai5 < 2e-16 1       < 2e-16 < 2e-16

P value adjustment method: bonferroni
```

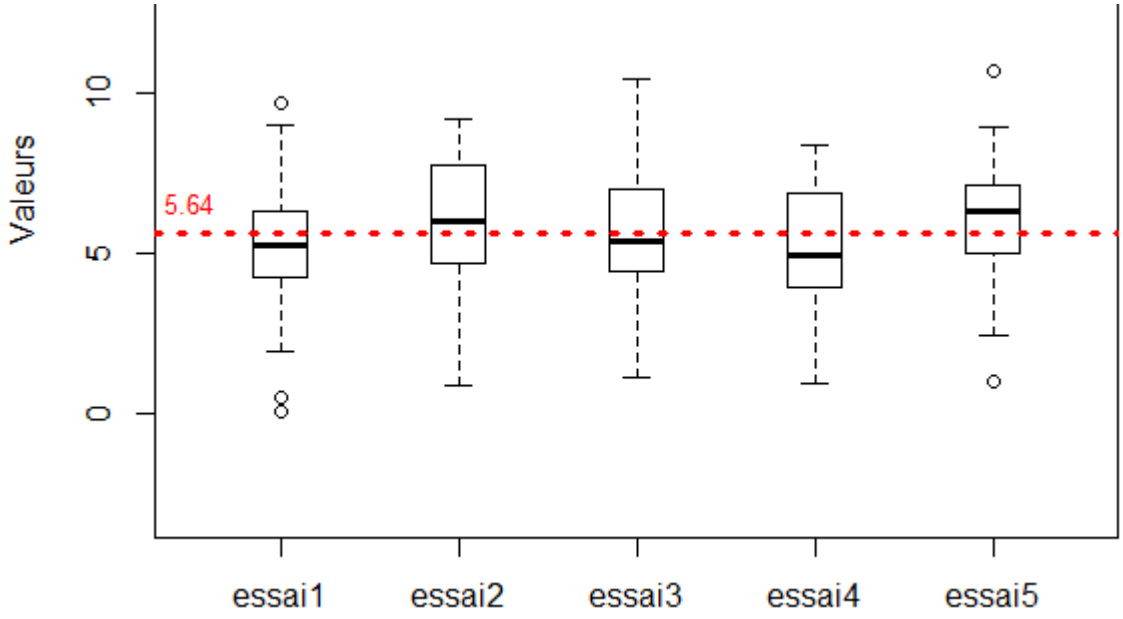
La comparaison de moyennes de Tukey fournit le tableau ci-contre . lwr et upr sont les bornes d'un intervalle de confiance à 95%. Si 0 n'est pas dedans, on rejette H_0 On retrouve la conclusion précédente avec des seuils de significativité différents du test de Bonferroni.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = resultAnova$values ~ resultAnova$in

$`resultAnova$ind`
                    diff        lwr       upr     p adj
essai1-essai4 0.5121242 -0.6425052 1.666754 0.7403799
essai3-essai4 3.8352026  2.6805731 4.989832 0.0000000
essai2-essai4 7.7640336  6.6094041 8.918663 0.0000000
essai5-essai4 8.0253168  6.8706874 9.179946 0.0000000
essai3-essai1 3.3230784  2.1684489 4.477708 0.0000000
essai2-essai1 7.2519094  6.0972799 8.406539 0.0000000
essai5-essai1 7.5131926  6.3585631 8.667822 0.0000000
essai2-essai3 3.9288310  2.7742015 5.083460 0.0000000
essai5-essai3 4.1901142  3.0354848 5.344744 0.0000000
essai5-essai2 0.2612832 -0.8933463 1.415913 0.9714809
```

Le test de Newman-Keuls fournit une classification en trois groupes. Le premier tableau donne les caractéristiques des différents essais (moyenne, écart type corrigé, valeur minimale et maximale). .

```
Student Newman Keuls Test
for resultAnova$values

Mean Square Error:  4.412942

resultAnova$ind,  means

       resultAnova.values      std  r        Min       Ma
essai1           3.902549 2.296074 50 -0.9022114 10.52817
essai2          11.154458 1.989246 50  6.5916313 15.59815
essai3           7.225627 2.222621 50  3.2717597 12.27443
essai4           3.390425 2.059994 50 -2.2406975  8.34942
essai5          11.415742 1.911029 50  7.0664491 15.37203

Alpha: 0.05 ; DF Error: 245

Critical Range
        2         3         4         5
0.8275473 0.9907101 1.0867973 1.1546295

Means with the same letter are not significantly differen

       resultAnova$values groups
essai5          11.415742      a
essai2          11.154458      a
essai3           7.225627      b
essai1           3.902549      c
essai4           3.390425      c
```

Il est important de montrer que l'ANOVA et les test Bonferroni, Turkey et SNK n'ont pas la même significativité. L'étudiant pourrait penser que réaliser l'ANOVA n'a pas d'intérêt et que l'on pourrait se contenter du test SNK. L'exemple suivant donne une situation où l'ANOVA rejette l'égalité des moyennes mais les tests Bonferroni, Turkey et SNK classent tous les essais dans la même catégorie.




```
                Df Sum Sq Mean Sq F value Pr(>F)
resultAnova$ind  4   37.5   9.373   2.513 0.0423 *
Residuals      245  913.8   3.730
---
```

```
Student Newman Keuls Test
```

```
        Pairwise comparisons using t tests with pooled

data:  resultAnova$values and resultAnova$ind

       essai1 essai2 essai3 essai4
essai2 0.22   -      -      -
essai3 1.00   1.00   -      -
essai4 1.00   0.16   1.00   -
essai5 0.54   1.00   1.00   0.40

P value adjustment method: bonferroni
```

```
Alpha: 0.05 ; DF Error: 245

Critical Range
        2         3         4         5
0.7608117 0.9108166 0.9991550 1.0615170

Means with the same letter are not significant

       resultAnova$values groups
essai2           6.096406      a
essai5           5.952473      a
essai3           5.785381      a
essai1           5.205697      a
essai4           5.156111      a
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = resultAnova$values ~ resultAnova$ind, data = resultAnova)

$`resultAnova$ind`
                    diff        lwr      upr     p adj
essai1-essai4 0.04958647 -1.0119305 1.111103 0.9999384
essai3-essai4 0.62927074 -0.4322463 1.690788 0.4802226
essai5-essai4 0.79636296 -0.2651540 1.857880 0.2402141
essai2-essai4 0.94029594 -0.1212211 2.001813 0.1097112
essai3-essai1 0.57968428 -0.4818327 1.641201 0.5629691
essai5-essai1 0.74677649 -0.3147405 1.808293 0.3024881
essai2-essai1 0.89070947 -0.1708075 1.952226 0.1464476
essai5-essai3 0.16709221 -0.8944248 1.228609 0.9926757
essai2-essai3 0.31102519 -0.7504918 1.372542 0.9288406
essai2-essai5 0.14393298 -0.9175840 1.205450 0.9958752
```

Exemple 4 : Analyse de variance (ANOVA) à 2 facteurs avec répétitions

On se place toujours dans le cas de plans d'expériences complets (toutes les combinaisons de facteurs sont présentes) équirépétés (le nombre de répétitions est le même pour toutes les combinaisons de facteurs). Dans ce type d'expérimentation, il s'agit d'étudier l'influence de deux facteurs sur une variable réponse.

Un essai a été mis en place afin d'étudier l'influence sur le rendement (variable réponse) de blé dur de la variété choisie et du fongicide utilisé (les deux facteurs). Chaque combinaison de facteur a été répétée 20 fois dans les mêmes conditions expérimentales. Les résultats peuvent être présentés de deux manières différentes. Le tableau de droite est le format utilisé par les logiciels de traitement statistique. Les rendements sont donnés en quintaux/hectare.

		variété		
		1	2	3
fongicide	1	52 …	55 …	51 …
	2	54 …	52 …	53 …

fongicide	variété	rendement
1	1	52
…	..	
1	2	55
…	…	
1	3	51
…	…	
2	1	54
…	…	
2	2	52
…	…	
2	3	53
…	…	

La situation étudiée est une version très simplifiée des études conduites sur le terrain. Nous avons choisi de restreindre le nombre de modalités pour faciliter l'appropriation des notions mathématiques mises en jeu.
Dans cette situation, on se pose naturellement deux questions :

- Y-a-t-il un effet de la variété sur le rendement ?
- Y-a-t-il un effet du fongicide sur le rendement ?

Mais le fait d'avoir des répétitions nous conduit à nous poser une troisième question :

- Y-a-t-il un effet d'interaction entre la variété et le fongicide sur le rendement ?

Cette notion d'interaction est fondamentale en agronomie et nécessite un travail d'appropriation par les étudiants mené conjointement par les enseignants de mathématiques et d'agronomie. Ce travail est nécessaire pour comprendre le modèle mathématique sur lequel s'appuie le traitement statistique de l'analyse de variance à deux facteurs mis en place ici. On dira qu'il y a interaction entre deux facteurs $F1$ et $F2$ sur une variable Y si l'effet de l'un diffère selon la modalité de l'autre. Dans la situation étudiée ici cela correspondrait, par exemple, au cas où l'un des fongicides serait particulièrement performant d'un point de vue du rendement sur l'une des variétés mais plutôt mauvais sur une autre.

Pour chaque combinaison de facteurs, on considère que l'échantillon de taille 20 est issu d'une variable aléatoire Y_{ij} (combinaison fongicide i avec la variété j pour $i = 1{,}2$ et $j = 1{,}2{,}3$).

Le modèle choisi se présente sous la forme :

$$Y_{ijk} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \text{ avec } i = 1{,}2 \text{ ; } j = 1{,}2{,}3 \text{ et } k = 1{,}2{,}\ldots{,}20$$

Ce modèle signifie que pour la combinaison de facteurs (i, j), le rendement attendu est le rendement moyen m, obtenu sans distinction des modalités des facteurs considérés, auquel s'ajoute l'effet α_i du fongicide i, l'effet β_j de la variété j , les interactions $\gamma_{i,j}$ entre les modalités i, j des deux facteurs et un

résidu aléatoire ϵ_{ijk} qui traduit la part de variabilité expérimentale. On retrouve des hypothèses semblables à celles déjà vues dans l'ANOVA à 1 facteur :

- les ε_{ij} sont des variables aléatoires indépendantes et suivent la loi normale $N(0,\sigma)$.
- Les résidus doivent être non corrélés entre eux : $cov(\epsilon_{ijk}, \epsilon_{rst}) = 0\ \forall (i,j,k) \neq (r,s,t)$.
- $\sum \alpha_i = 0$ et $\sum \beta_j = 0$
- $\forall j, \sum_i \gamma_{ij} = 0$ et $\forall i, \sum_j \gamma_{ij} = 0$

Avec cette écriture du modèle, pour répondre aux trois questions que l'on se pose, on est amené à réaliser trois tests statistiques :

- Pour évaluer l'effet du fongicide on teste l'hypothèse $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ contre $H_1 : \exists i\ \alpha_i \neq 0$.
- Pour évaluer l'effet de la variété on teste l'hypothèse $\mathrm{H}_0' : \beta_1 = \beta_2 = \beta_3 = 0$ contre $H_0' : \exists i\ \beta_i \neq 0$.
- Pour détecter la présence d'interactions entre les deux facteurs on teste l'hypothèse

$$\mathrm{H}_0'' : \gamma_{ij} = 0\ \forall (i,j) \text{ contre } \mathrm{H}_1'' : \exists (i,j) \text{tel que } \gamma_{ij} \neq 0.$$

Pour une valeur observée y_{ijk} du rendement, on note :

$$y_{ijk} = \hat{m} + \hat{\alpha_i} + \hat{\beta_j} + \hat{\gamma_{ij}} + e_{ijk} \text{ où}$$

- Le rendement moyen de référence est estimé par la moyenne de tous les rendements observés, c'est à dire $\hat{m} = \overline{y_{...}}$
- L'effet d'un fongicide est estimé par l'écart entre le rendement moyen lorsqu'il est employé et le rendement de référence c'est à dire $\hat{\alpha_i} = \overline{y_{i..}} - \overline{y_{...}}$
- L'effet variété est estimé par l'écart entre le rendement moyen de cette variété et le rendement de référence c'est à dire $\hat{\beta_j} = \overline{y_{.j.}} - \overline{y_{...}}$
- L'effet d'interaction entre le fongicide i et la variété j est estimé par :

$$\hat{\gamma_{ij}} = \overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}} \text{ qui peut s'écrire } \hat{\gamma_{ij}} = \left(\overline{y_{ij.}} - \overline{y_{j.}}\right) - \left(\overline{y_{i..}} - \overline{y_{...}}\right)$$

Cette dernière écriture permet de mieux relier ce terme à un terme d'interaction. L'écart $\left(\overline{y_{ij.}} - \overline{y_{j.}}\right)$ exprime l'effet du facteur $F1$ pour la modalité j de $F2$. Cet effet est ensuite comparé à l'effet général du facteur $F1$ $\left(\overline{y_{i..}} - \overline{y_{...}}\right)$. Les e_{ijk} expriment les écarts au modèle. Comme dans le cas d'un seul facteur, il s'agit de trouver les estimations qui ajustent au mieux le modèle c'est-à-dire qui minimisent la somme $\sum e_{ijk}^2$.

On introduit alors l'équation de l'analyse de la variance souvent donnée sous la forme :

$$SCE_{totale} = SCE_{fongicide} + SCE_{variete} + SCE_{interactions} + SCE_{residuel}\ .$$

Comme nous l'avons précisé au début, nous allons tester non pas une mais trois hypothèses récapitulées dans le tableau ci-dessous :

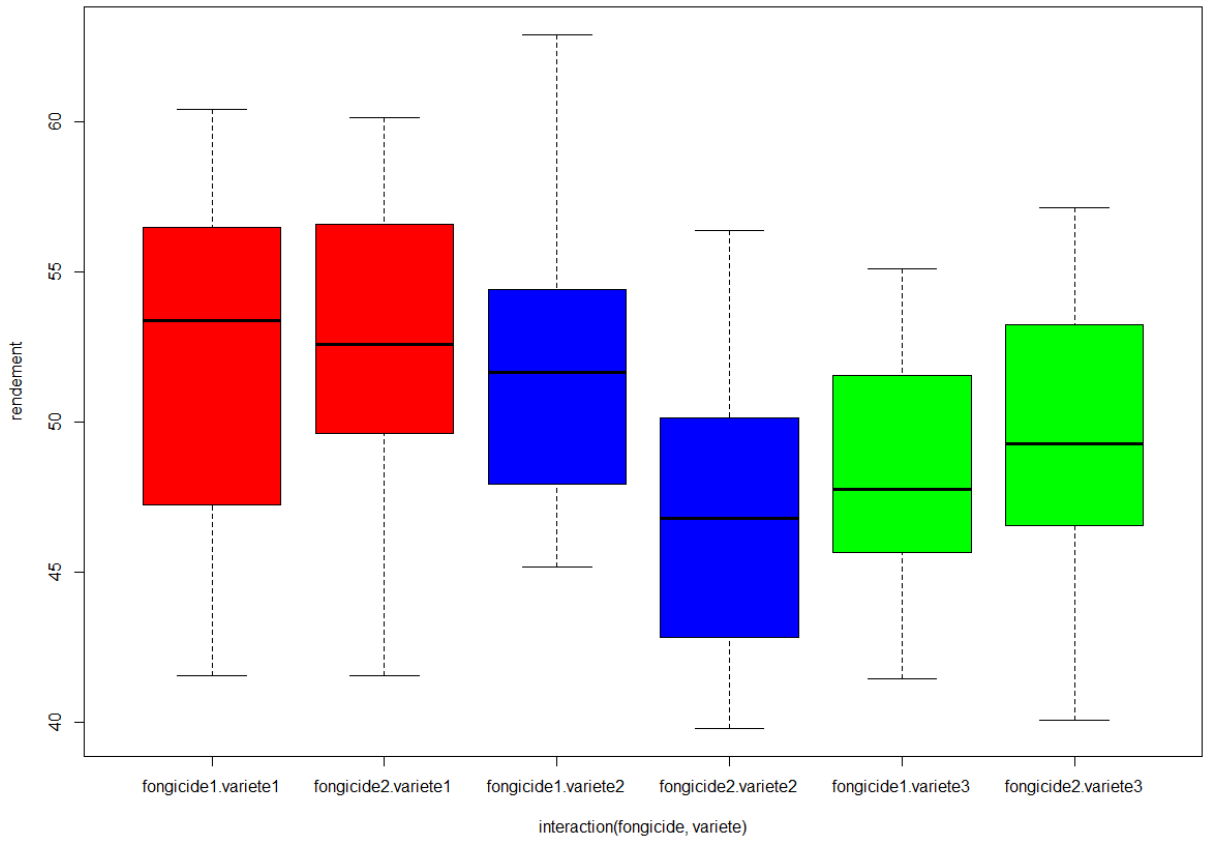
Effet testé	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
fongicide	$\alpha_i = 0\ \forall i$	$\exists i$ tel que $\alpha_i \neq 0$
variété	$\beta_i = 0\ \forall i$	$\exists i$ tel que $\beta_i \neq 0$
Interactions	$\gamma_{ij} = 0\ \forall i,j$	$\exists (i,j)$ tel que $\gamma_{ij} \neq 0$

Comme dans le cas de l'ANOVA à 1 facteur on considère les rapports $\frac{CM_{effet}}{CM_{residuel}}$, pour tester la significativité des effets. La discussion sur les degrés de liberté ne doit pas faire l'objet de justification théorique. On admet aussi l'utilisation de la loi de Fisher.

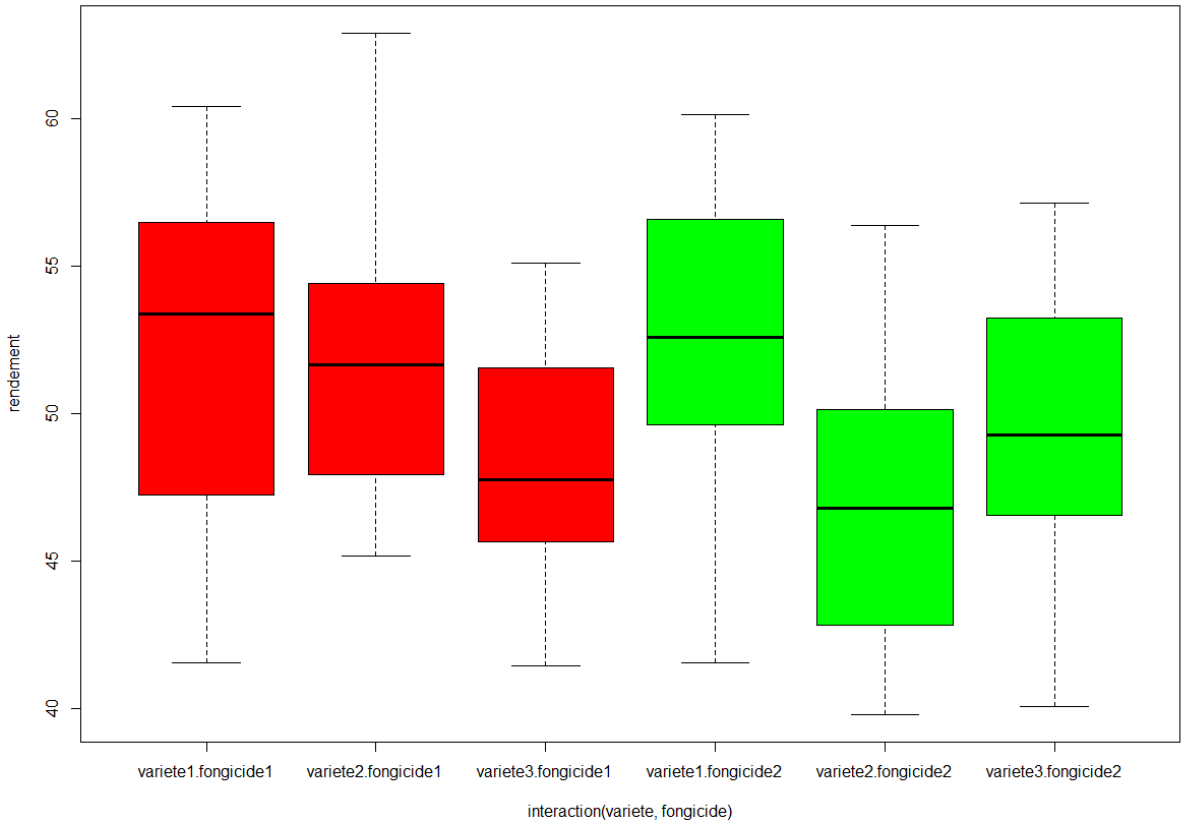
Reprenons l'exemple, que nous allons traiter à l'aide du logiciel R. Tout d'abord on importe les données que l'on suppose enregistrées dans un fichier au format csv.

Avant de lancer l'analyse de la variance, il est important d'avoir une approche graphique des données afin de se faire une première idée de la situation. On s'intéresse d'abord à l'effet des interactions avant d'envisager les effets individuels des facteurs.

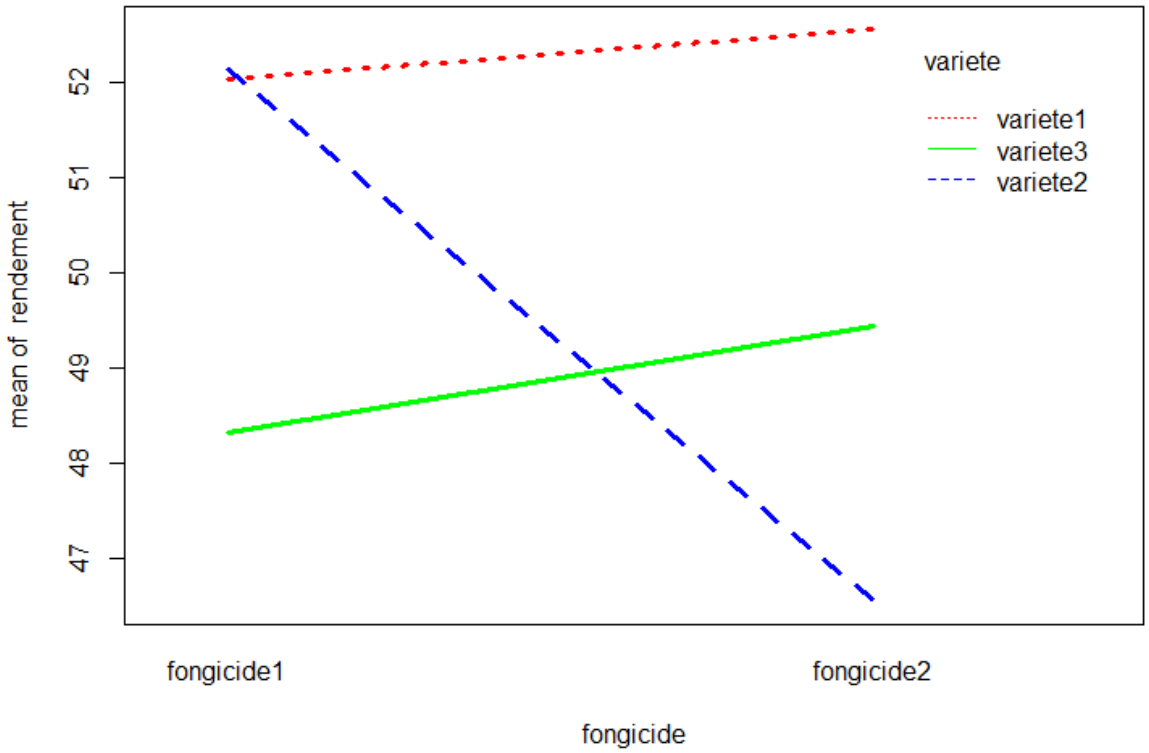
Un graphique de type boxplot dans lequel les 6 combinaisons de facteurs sont représentées permet d'avoir une première approche de l'effet des interactions. Dans notre exemple, on obtient les deux graphiques suivants selon que l'on choisisse d'organiser les boîtes par fongicide ou par variété :



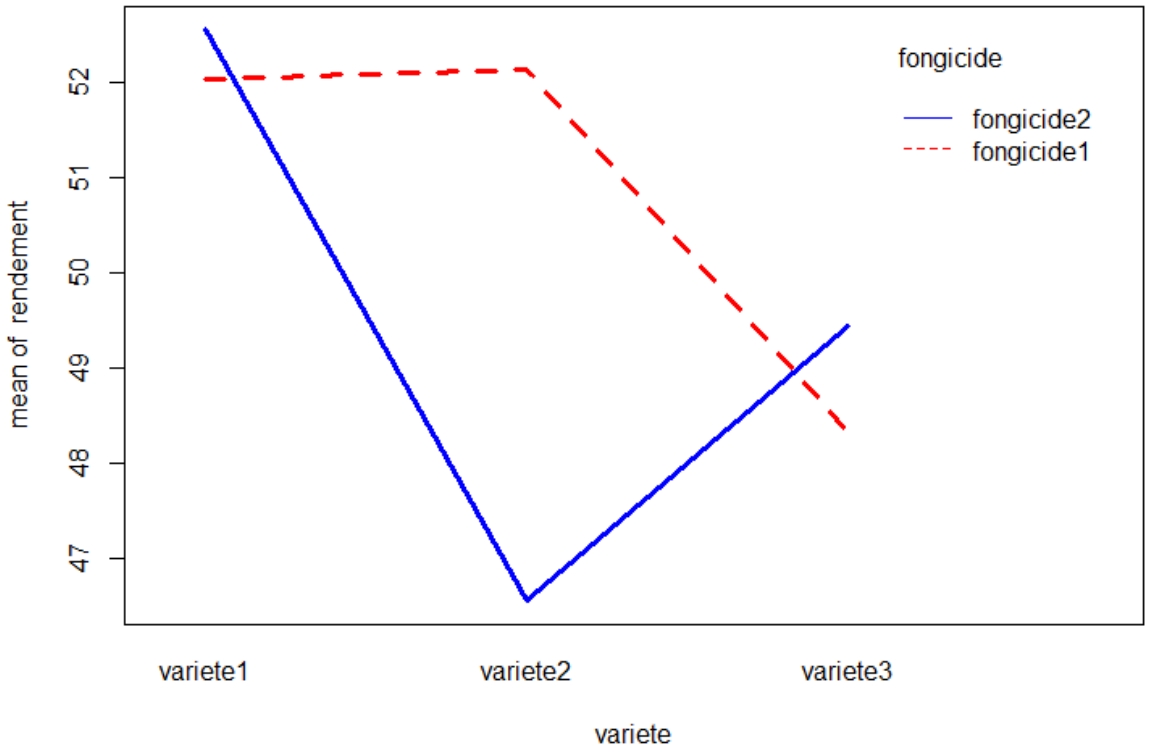
L'analyse de ce premier graphique permet d'identifier des interactions liées à la variété 2. En effet on constate que les deux boîtes bleues associées à cette variété sont positionnées entre elles bien plus séparées que les variétés 1 et 3. Plus précisément, alors que les fongicides ne semblent pas avoir un véritable effet sur les variétés 1 et 3, il semble qu'ils génèrent une réaction beaucoup plus forte sur la variété 2.



On observe ce même effet d'interaction sur ce deuxième graphique. Les boxplots de la variété 2 en vert n'ont pas un décalage dans le même ordre de grandeur.
Pour conforter notre conjecture, on peut alors décider de tracer les graphiques des interactions :



Ce premier graphique fait apparaître clairement une interaction (croisement des courbes) et plus précisément la responsabilité de la variété 2.



Dans ce deuxième graphique, la présence d'interaction est également mise en évidence par le croisement des courbes. L'influence de la variété 2 dans les interactions est peut-être moins lisible à première vue. De manière générale, on retient donc l'intérêt de générer tous les graphiques d'interactions possibles pour mieux analyser la situation.

Suite à cette première approche graphique des données, on s'attend à ce que l'ANOVA fasse apparaître un effet significatif des interactions.

À l'aide du logiciel R, on obtient le tableau d'ANOVA suivant :

```
                 Df Sum Sq Mean Sq F value  Pr(>F)
fongicide         1   52.3   52.33   2.385 0.12531
variete           2  276.0  138.01   6.289 0.00257 **
fongicide:variete  2  276.2  138.12   6.293 0.00256 **
Residuals       114 2501.8   21.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

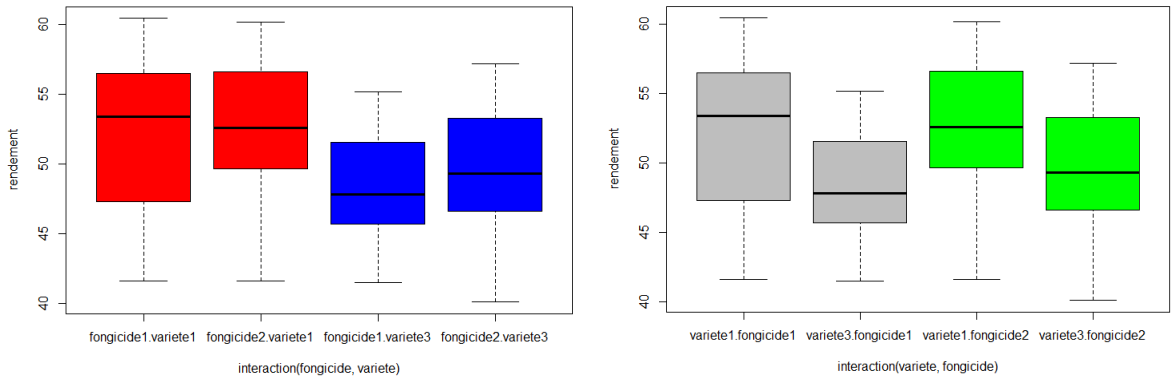
À la lecture du tableau, notre conjecture d'un effet significatif des interactions est confirmée. Cet effet est même très significatif ($p_{value} \approx 0{,}003$) donc très largement inférieur au seuil de 5%). On a un effet variétal différent selon les fongicides. L'interprétation seule des effets principaux n'a aucun sens. Il est en effet difficile de distinguer si les différences proviennent surtout de l'effet d'un facteur ou des interactions entre les facteurs. Notre étude prenant en compte les effets individuels de la variété et du fongicide s'arrête donc là. Si on souhaite exploiter les données de l'essai, on peut toutefois réaliser une ANOVA à un facteur en prenant comme facteur le traitement (combinaison d'un fongicide et d'une variété) à 6 modalités.

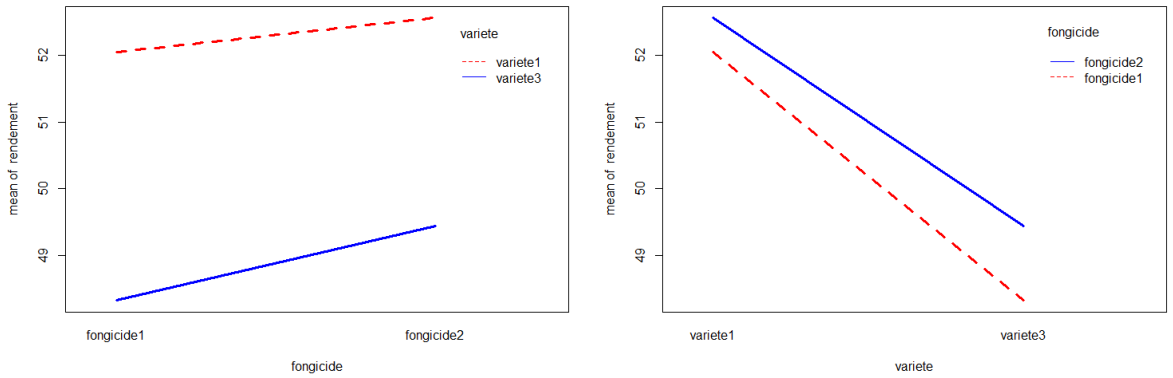
Dans notre cas précis, on pourrait choisir de retirer la variété 2 des données et relancer une ANOVA à 2 facteurs. Il convient avant de prendre cette décision d'échanger avec le technicien responsable de l'essai : la variété 2 a-t-elle rencontré des problèmes particuliers pendant l'expérimentation ? Est-il pertinent de l'extraire de l'analyse des données ? Tous ces choix sont intéressants à envisager dans le cadre de situations de formation intégrative mobilisant l'apport et la collaboration de l'enseignant d'agronomie et de l'enseignant de mathématiques.

La vérification des conditions d'application de l'ANOVA se fait selon les mêmes modalités que pour l'ANOVA à un facteur.

Nous faisons le choix de reprendre nos données et de retirer la variété 2 de nos modalités. Nous réalisons une nouvelle ANOVA à 2 facteurs.

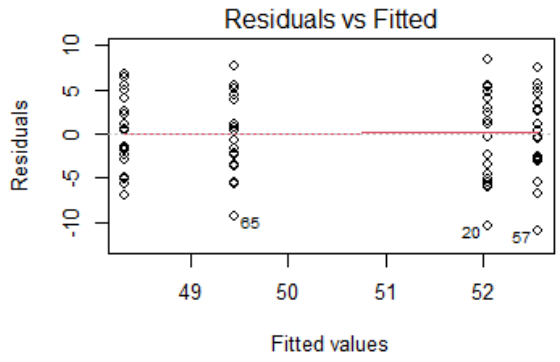
On obtient les graphiques ci-dessous :

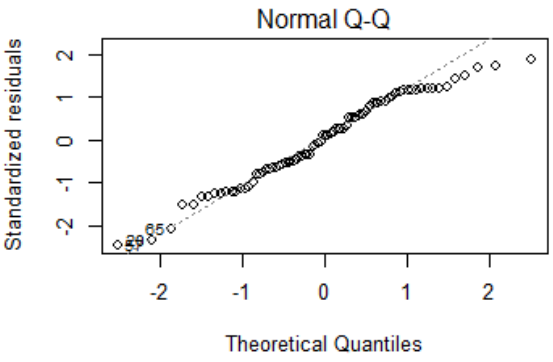


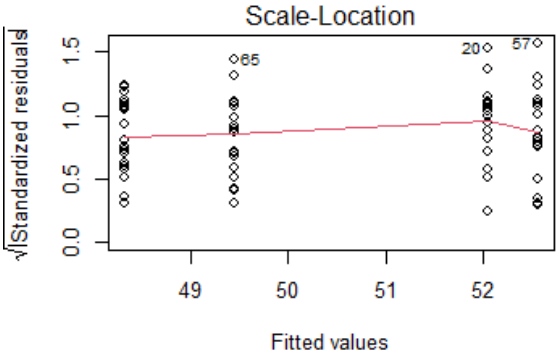


L'analyse de ces graphiques permet de conjecturer l'absence d'interaction et probablement un effet variété dont il reste à étudier la significativité à l'aide d'une ANOVA.

Une vérification des conditions d'application de l'ANOVA peut se faire par une approche graphique et être éventuellement complétée par des tests. Les graphiques obtenus mettent en évidence la normalité et l'homoscédasticité des résidus.







Le tableau de l'ANOVA permet alors de conclure.

```
                                                Df Sum Sq Mean Sq F value  Pr(>F)
donnees_ss_var2$fongicide                        1   13.4   13.37   0.628 0.43047
donnees_ss_var2$variete                          1  234.7  234.73  11.027 0.00138 **
donnees_ss_var2$fongicide:donnees_ss_var2$variete  1    1.8    1.85   0.087 0.76909
Residuals                                       76 1617.8   21.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comme attendu, l'hypothèse « présence d'interactions » est rejetée. On peut interpréter les effets individuels. Il existe ici un effet variétal très significatif.

Un test de Newman-Keuls permet alors de discriminer les modalités entre elles et de donner une classification en groupes homogènes pour chacun des facteurs. Dans notre exemple, n'ayant que deux modalités sur chacun des facteurs, ce test est inutile puisque les résultats de l'ANOVA nous permettent directement de discriminer les modalités entre elles. Dans un intérêt pédagogique nous donnons toutefois les lignes de commandes R associées à ces tests ainsi qu'une copie des résultats renvoyés par le logiciel R :

```
library(agricolae)
testsnk1=SNK.test(anova,'fongicide',group=TRUE)
testsnk1
testsnk2=SNK.test(anova,'variete',group=TRUE)
testsnk2
```

```
$groups
          rendement groups
variete1  52.30502      a
variete3  48.87916      b
```

```
$groups
           rendement groups
fongicide2  51.00094      a
fongicide1  50.18323      a
```

Exemple 5 : ANOVA à deux facteurs avec une seule répétition.

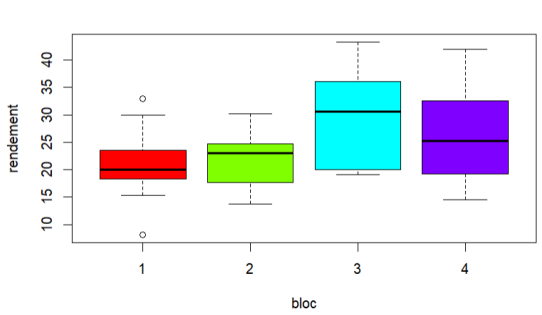
En expérimentation agronomique, les essais factoriels en blocs aléatoires complets sont assez fréquents. Dans ce type d'essai, on étudie l'effet d'un facteur principal et l'effet bloc sur une variable réponse. Le dispositif comporte plusieurs blocs de parcelles où toutes les modalités du facteur principal figurent une fois et une seule. Au sein d'un bloc, supposé homogène, les différentes modalités du facteur principal sont affectées au hasard. L'intérieur d'un bloc est le plus homogène possible. En revanche, au sein d'un même terrain, les blocs sont différents les uns des autres. Il est cependant souhaitable d'éviter des différences trop importantes, sources possibles d'interaction entre blocs et traitement. Le dispositif en bloc peut donc compenser des différences connues à l'avance, du moment qu'elles ne sont pas trop importantes.

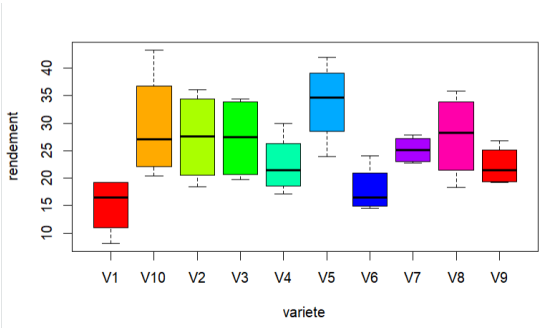
Le bloc est un facteur dit contrôlé prenant en compte l'hétérogénéité du milieu. Aussi, si les blocs sont correctement positionnés il n'y pas d'effet d'interaction entre les blocs et le facteur principal. Le modèle mathématique associé à ce type d'essai est donc de la forme $Y_{ijk} = m + \alpha_i + \beta_j + \epsilon_{ijk}$ en reprenant les notations de l'exemple 4. (α_i : effet du bloc, β_j : effet du facteur principal). Les contraintes sur les résidus et les paramètres sont les mêmes que pour celles de l'ANOVA à 2 facteurs.

On a réalisé un essai variétal pois chiche composé de 10 variétés (V1, V2, …V10) chez un agriculteur du Lauragais. Le plan est présenté et les données associées sont présentées ci-dessous :

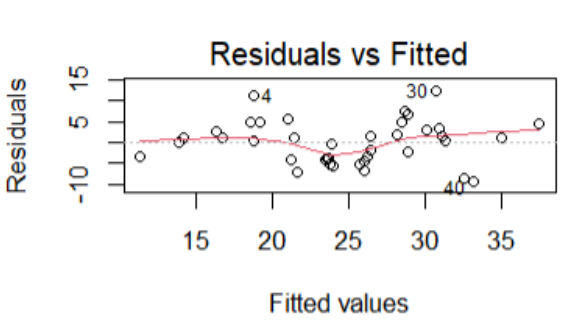
	Bloc 1	Bloc 2	Bloc 3	Bloc 4
V1	8,1	13,8	19,1	19,3
V2	18,5	22,7	36,1	32,6
V3	19,7	21,6	33,3	34,4
V4	29,9	17,2	20,0	22,8
V5	33,0	24,0	36,2	42,0
V6	15,3	17,7	24,1	14,6
V7	22,7	23,4	27,8	26,7
V8	18,3	24,7	35,8	31,9
V9	23,5	26,8	19,5	19,3
V10	20,4	30,2	43,3	23,9

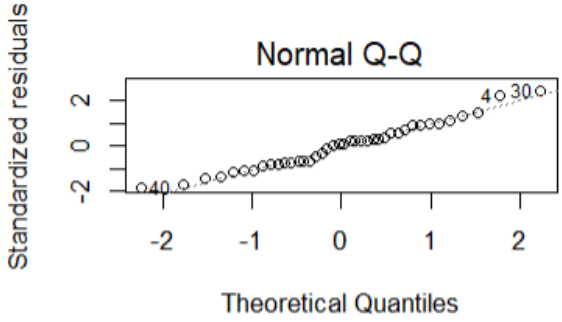
On extrait à l'aide du logiciel R plusieurs représentations graphiques qu'il convient d'analyser pour conjecturer des effets facteurs et bloc éventuels :

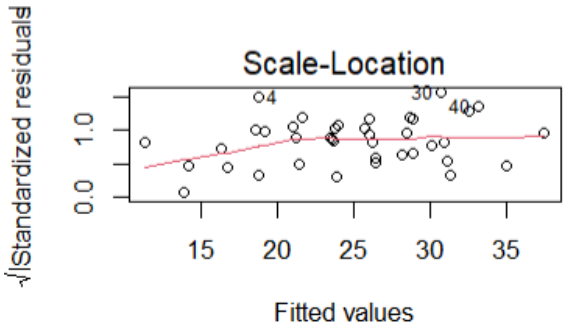


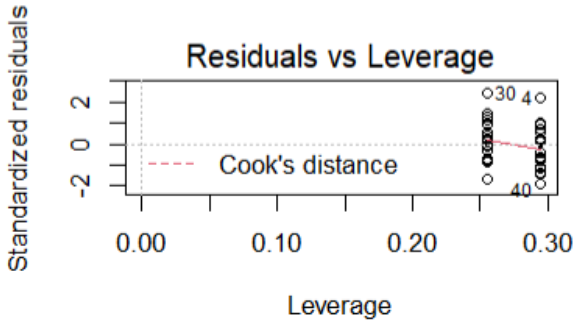


On s'assure ensuite que les conditions d'application de l'ANOVA sont vérifiées.









On peut effectuer l'ANOVA :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bloc	1	306	306.03	8.730	0.00616 **
variete	9	1111	123.45	3.521	0.00468 **
Residuals	29	1017	35.06		

L'effet bloc permet de s'assurer de la bonne prise en compte de l'hétérogénéité du terrain par le positionnement des blocs. En l'absence d'effet bloc on aurait pu traiter les données à l'aide d'une ANOVA à 1 facteur d'une seule répétition. L'ANOVA fait ici apparaître un effet variété significatif. Nous allons donc réaliser un test de Newman Keuls pour obtenir une classification en groupes homogènes de variétés.

	rendement	groups
V5	33.800	a
V10	29.450	ab
V8	27.675	abc
V2	27.475	abc
V3	27.250	abc
V7	25.150	abc
V4	22.475	abc
V9	22.275	abc
V6	17.925	bc
V1	15.075	c

Cette classification nous permet de dire que les variétés V5 et V10 ont un rendement significativement supérieur au rendement de la variété V1. On ne peut cependant pas discriminer les variétés V5 et V10 entre elles car toutes deux appartiennent au même groupe a. De manière générale, l'interprétation de ce type de classification ne peut se faire qu'en croisant les résultats statistiques avec les connaissances agronomiques. Là encore, une collaboration étroite entre l'enseignant de mathématiques et l'enseignant d'agronomie est fondamentale pour croiser les regards.

Annexe 1

Exemple 3

Les sorties R sont produites avec les lignes suivantes :

```
#############################################################
###Importation des données
donnees=read.csv2(file.choose(),header=TRUE)
head(donnees)
hauteur=donnees$hauteur
variete=donnees$variete
#############################################################
##    Première Boxplot
#############################################################
moy=mean(hauteur)
graphics.off()
boxplot(hauteur~variete,data=donnees,xlab='Variétés de pois chiche',
        ylab='Hauteur des plants',col='white',names=c('Variété 1','Variété 2','Variété 2'),
        border=c('blue','purple','green'),boxwex=0.3,ylim=c(15,50))
title(main='Boxplot hauteur des plants de pois chiche')
abline(moy,0,col='red',lty=3,lwd=3)
text(0.5,moy,round(moy,digits=2),col='red',cex=0.8,pos=3)
stripchart(hauteur~variete,data=donnees,vertical=T,method='jitter',add=T,
           cex=0.5,pch=15,col=c('blue','purple','green'))

########################################
##  Tableau de l'ANOVA
########################################
anova=aov(hauteur~variete,data=donnees)
tab.anova=summary(anova)
tab.anova
#########################################
#  Conditions d'application de l'ANOVA
#   Graphiques diagnostics
#########################################
par(mfrow=c(2,2))
plot(bouture.aov)
```

```
################################################
#
#   Test Post-Hoc
#
################################################
# D'abord récupérer le package agricolae
#chooseCRANmirror() permet de choisir le site de téléchargment des packages
#utils:::menuInstallPkgs() permet de choisir le package à télécharger et à installer
library(agricolae)# charge le package agricolae pour Newman-Keuls
TukeyHSD(anova,ordered=T)
pairwise.t.test(donnees$valeurs,donnees$facteur,p.adjust='bonferroni',alternative='two.sided')
SNK.test(anova,'facteur', console=T)
```

Exemple 4

Les sorties R sont produites avec les lignes suivantes :

```
###Importation des données
donnees=read.csv2(file.choose(),header=TRUE)
head(donnees)
rendement=donnees$valeurs
fongicide=donnees$fongicide
variete=donnees$variete

###Boxplot : premiere approche graphique :
graphics.off()
plot(rendement~ interaction(fongicide, variete), data=donnees, col = c("red","red","blue","blue",
                                                                       "green","green"))
plot(rendement~ interaction(variete,fongicide), data=donnees, col = c("red","red","red","green",
                                                                      "green","green"))
###Graphique des intéractions :
graphics.off()
interaction.plot(variete, fongicide, rendement,col=c('red','blue'), lwd=3)
interaction.plot(fongicide, variete, rendement,col=c('red','blue'), lwd=3)

### Tableau de l'ANOVA
anova = aov(rendement ~ fongicide*variete, data=donnees)
summary(anova)

###Approche graphique de la vérification des hypothèses d'utilisation de l'ANOVA
affichage = par(mfrow = c(2, 2))
plot(anova)
```

Exemple 5 :

Les sorties R sont produites avec les lignes suivantes :

```
###Importation des données
donnees=read.csv2(file.choose(),header=TRUE)
head(donnees)
rendement=donnees$rendement
variete=donnees$fvariete
bloc=donnees$bloc
###Boxplot : premiere approche graphique :
graphics.off()
boxplot(rendement~  variete, data=donnees,col=rainbow(9))
boxplot(rendement~ bloc, data=donnees, col = rainbow(4))
##########################################################
### Tableau de l'ANOVA
#
anova = aov(rendement ~ bloc+variete, data=donnees)
summary(anova)
#############################################################################
###Approche graphique de la vérification des hypothèses d'utilisation de l'ANOVA
affichage = par(mfrow = c(2, 2))
plot(anova)
#############################################################
###Test de Newman Keuls : classification en groupes homogènes
library(agricolae)
testsnk=SNK.test(anova,'variete',group=TRUE)
testsnk
```

Annexe 2

Introduire la distribution du χ^2 par simulation de l'adéquation à une loi. Par exemple, appuyons-nous sur l'expérience de Mendel sur la transmission de caractères sur les pois.

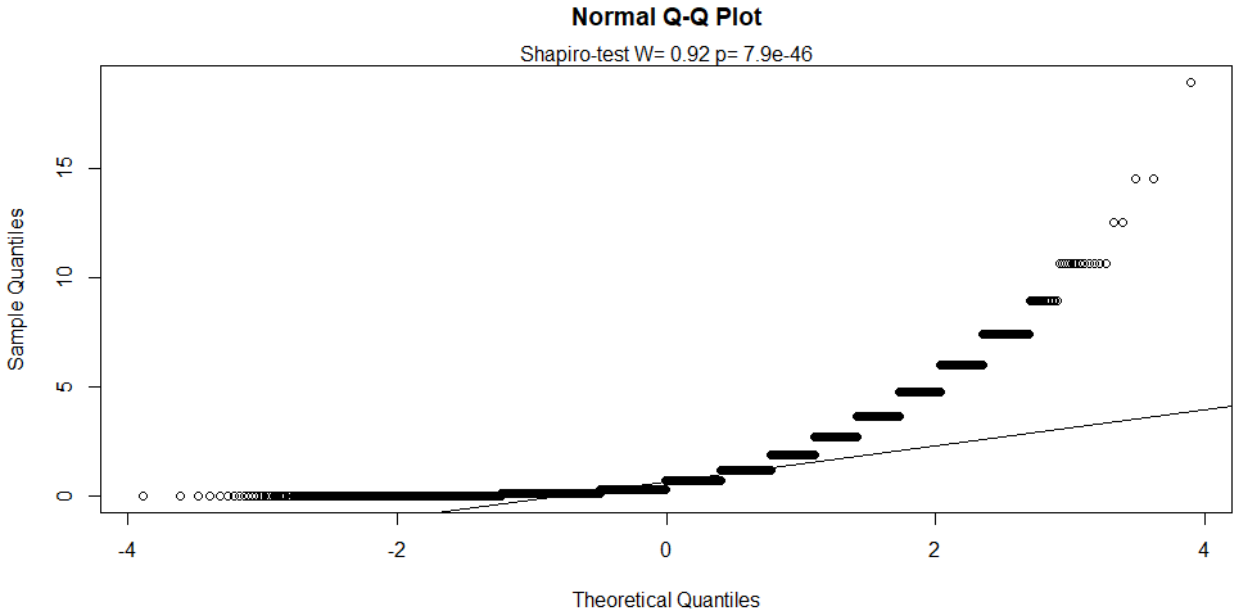
Pour comprendre la transmission d'un caractère d'une génération à l'autre, Mendel féconde artificiellement deux variétés de pois de lignée pure. L'un avec le caractère « graines lisses », l'autre avec le caractère « graines ridées ». La descendance obtenue (F1) ne possède que des graines lisses. Il poursuit l'expérience en réalisant l'autofécondation de la génération (F1). Il obtient la répartition suivante pour la génération (F2).

Caractère	**Graines ridées**	**Graines lisses**	**Total**
Effectifs	21	51	72

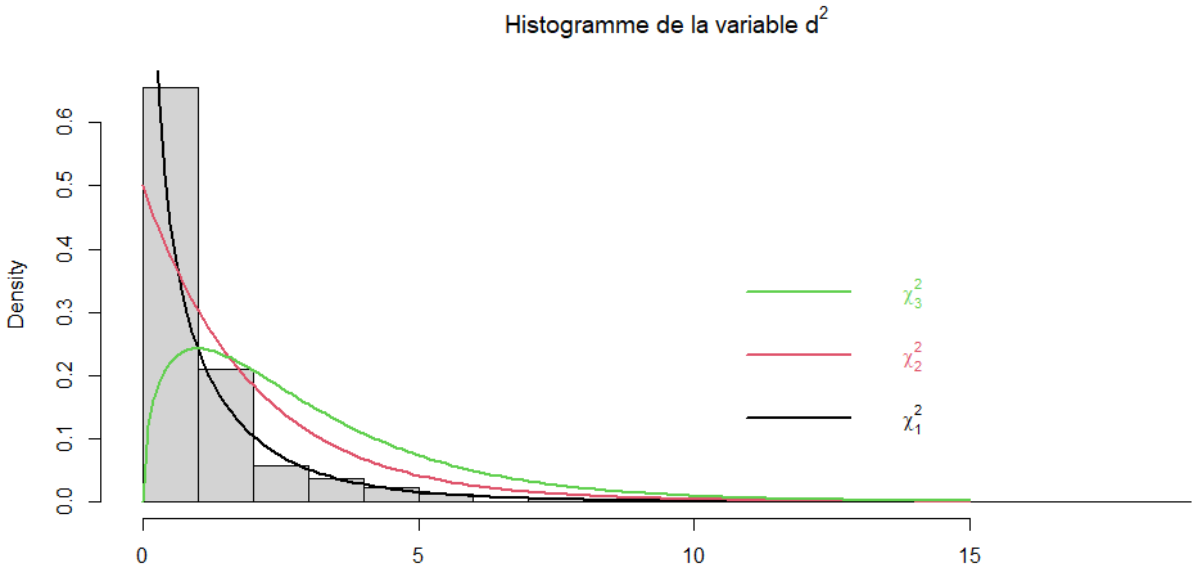
Ces résultats expérimentaux confirment-t-ils l'hypothèse de Mendel qui prévoit une répartition de 25% et 75% ?

On simule la variable statistique d^2 et on étudie sa répartition. En notant $O_{i,j}$ les effectifs observées et $E_{i,j}$ les effectifs attendus calculés à partir du modèle de Mendel.

$$d^2 = \sum_{i,j} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$



La forme de l'histogramme indique que la distribution de d^2 ne s'apparente pas à une loi normale, hypothèse qui pourra être confirmée avec un Q-Q Plot et un test de Shapiro-Wilk. On est donc amené à chercher une autre loi. Ce qui permet d'introduire les lois du χ^2.

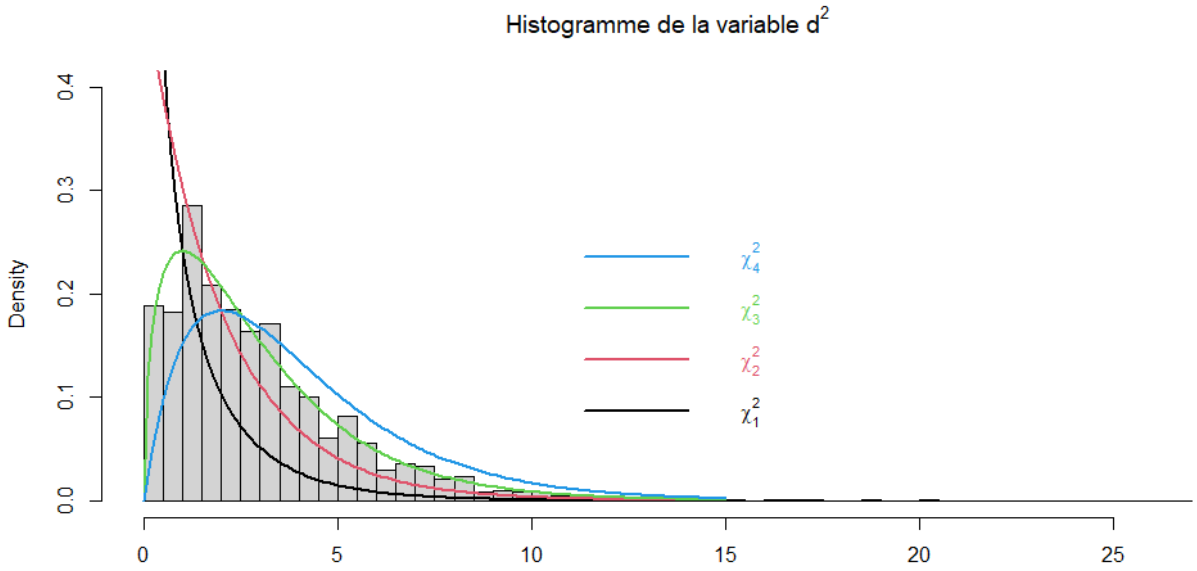


On peut alors s'intéresser à l'expérience de Mendel avec deux caractères exprimés par des gènes comportant deux allèles (l'un dominant A, B et l'autre récessif a, b) sur des chromosomes différents.
On obtient pour la génération (F2) le tableau suivant :

Caractères	ab	aB	Ab	AB	Total
Effectifs	3	15	13	33	64

Ces résultats expérimentaux confirment-t-ils l'hypothèse de Mendel qui prévoit la distribution $\left(\frac{1}{16}, \frac{3}{16}, \frac{3}{16}, \frac{9}{16}\right)$?

De la même manière, on obtient :



Application 14 :

On se propose de comparer les rendements moyens en q/ha de blé de variété "étoile de Choisy" pour une petite région agricole de la vallée de la Marne. On monte un plan d'expérience afin de tester si l'apport d'un autre élément fertilisant que l'azote a un effet significatif sur les rendements pour cette céréale et dans cette région.

Pour information, le plan d'expérience est représenté ci-dessous avec les rendements obtenus :

N	NK	NPK	N	NP
NK	NPK	NK	NP	N
NP	NP	N	NK	NPK
NPK	N	NP	NPK	NK

78	73,2	78,5	62,5	72,6
82,8	83,6	73	75,4	56,5
82,8	81,1	63,3	72,4	80,3
85,4	65,2	80,2	72,4	72,3

1. Déterminer des résumés paramétriques et graphiques de ces données.
2. Déterminer les résidus et les représenter.
3. A l'aide des graphiques diagnostiques, les conditions d'application d'une ANOVA semblent-elles vérifiées ?
4. Réaliser une cartographie des résidus à l'aide de R, en utilisant la fonction levelplot du package Lattice.