Document d'accompagnement thématique

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

ENSEIGNEMENT AGRICOLE
L'AVENTURE DU VIVANT
LES MÉTIERS GRANDEUR NATURE

Inspection de l'enseignement agricole

Diplôme : BTSA Métiers du Végétal
Alimentation, Ornement et Environnement

Thème : Enseignement de mathématiques

Commentaires, recommandations pédagogiques

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

C51- Organiser l'environnement de production d'un milieu

C61- Gérer des équipes de travail

C62- Gérer l'activité de production du système de production

C82- Produire des références

L'enseignement des mathématiques vise à :

- donner une assise scientifique permettant d'appréhender l'environnement physique et spatial du végétal, l'espace professionnel d'un point de vue géométrique et dimensionnel,
- raisonner l'organisation des tâches et la gestion des stocks,
- développer l'esprit critique devant les résultats d'expérimentation ,
- communiquer des résultats chiffrés sous une forme adaptée.

L'enseignant veille à s'appuyer sur les acquis des élèves 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 fortement 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 chefs d'exploitation de l'établissement, documentations, résultats issus de projets.

La progression construite par le professeur de mathématiques est 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 technique 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 amènera naturellement à élaborer des démarches, mener des calculs à l'aide d'un outil adapté, s'assurer de la cohérence de résultats et 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 simple.

Les situations développées dans ce document ne sont pas exhaustives mais illustrent l'esprit dans lequel l'enseignement des mathématiques doit être mis en œuvre.

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 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 choisie, 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.
- Calcul de pourcentages, calcul de 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 la « diminution d'une augmentation » ou « la diminution d'une 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 ces 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 Organiser l'environnement de production d'un milieu

L'enseignement des mathématiques contribue à développer des compétences permettant de s'approprier l'environnement physique et spatial du végétal. L'objectif, ici, est de prendre conscience des dimensions de l'espace cultivé. L'enseignement doit être ouvert sur l'extérieur afin de concilier l'étude d'une représentation, qui peut être un plan, une maquette, une image obtenue à partir d'un logiciel de cartographie ou d'une photographie prise par un drone, avec la réalité du terrain. La formation doit amener les apprenants à se construire un ensemble de modèles, de représentations et des méthodes permettant d'évaluer rapidement sur le terrain les dimensions d'une parcelle dans une unité adaptée, la densité de plantation ou encore le volume d'une serre, d'un tunnel…

- Déterminer le périmètre, encadrer l'aire de parcelles sur un plan, sur une vue aérienne en s'appuyant sur la connaissance des périmètres et des aires des figures usuelles du plan (triangle, disque, carré, rectangle, trapèze, …). L'usage de découpage et de recollement de régions (méthode dite du tangram) permet d'estimer des surfaces sans recours systématique à la mesure et aux formules d'aires. Appliquer un agrandissement ou une réduction. Utiliser une échelle, changer d'unité. Effets d'un agrandissement et d'une réduction sur les dimensions (périmètres et aires).
- Construire des références de surface permettant de donner un ordre de grandeur d'une parcelle dans une unité adaptée lorsque l'on se trouve sur le terrain. Par exemple, un terrain de football ou de rugby est proche d'un hectare, un terrain de basket est proche de 4 ares, …Ce travail peut être effectué en extérieur, sur des photos ou tout support rendant compte d'une situation vécue.
- Construire une référence de solides (prisme droit, pyramide, cylindre, cône...) et de leur volume respectif afin d'estimer ou de donner un ordre de grandeur, sur le terrain, du volume d'un contenant en lien avec le domaine professionnel (serre, tunnel, …).
- Calculer ou estimer sur le terrain une densité de plantation et plus généralement une population (bio agresseur). Estimer la masse ou le volume d'une récolte (fraises, pommes, …) sur une parcelle. L'estimation de ces quantités dans certains cas s'effectue par sondage. On fait alors le lien avec la fluctuation d'échantillonnage.

C61 - Gérer des équipes de travail

L'enseignement de mathématiques peut apporter son appui à d'autres disciplines dans la construction d'outils de planification tels que les graphes de PERT (*program evaluation and review technology)* ou les diagrammes de Gantt. En lien avec les SESG, l'enseignement des mathématiques permet de développer des méthodes pour représenter graphiquement ces outils.

Construire un graphe de PERT complété par un diagramme de Gantt en lien avec la situation. Développer une méthodologie de construction de ces objets. Les graphes de PERT sont des outils de planification de projet dont l'objectif est la recherche d'un ordonnancement minimisant la durée totale. Ils permettent d'apporter une réponse aux questions (Q) suivantes :

- Quel est le temps nécessaire pour réaliser l'ensemble du projet ?
- À quelle date doit débuter chaque tâche ?
- Quelles sont les tâches critiques ?

La méthode PERT s'appuie sur la conception d'un graphe auquel on ajoute un tableau de synthèse des dates des tâches. La méthode permet d'apporter des réponses aux questions (Q) ci-dessus. La conception du graphe et du tableau de synthèse s'appuie sur l'application d'algorithmes que l'on peut présenter en langage naturel.

Étude d'un exemple :

On considère un projet constitué des tâches suivantes dont la durée en jours est précisée entre parenthèses : A(2), B(8), C(5), D(2), E(6), F(5) et G(3). On suppose de plus que C ne peut pas débuter tant que A n'est pas réalisée, que D et E ne peuvent pas débuter tant que B n'est pas été réalisée, que F ne peut pas débuter tant que E n'est pas réalisée et G ne peut pas débuter tant que A et D ne sont pas réalisées. La méthode débute par la construction du tableau dit d'antériorité. Celui-ci permet de déterminer le nombre d'étapes et de construire la structure du graphe.

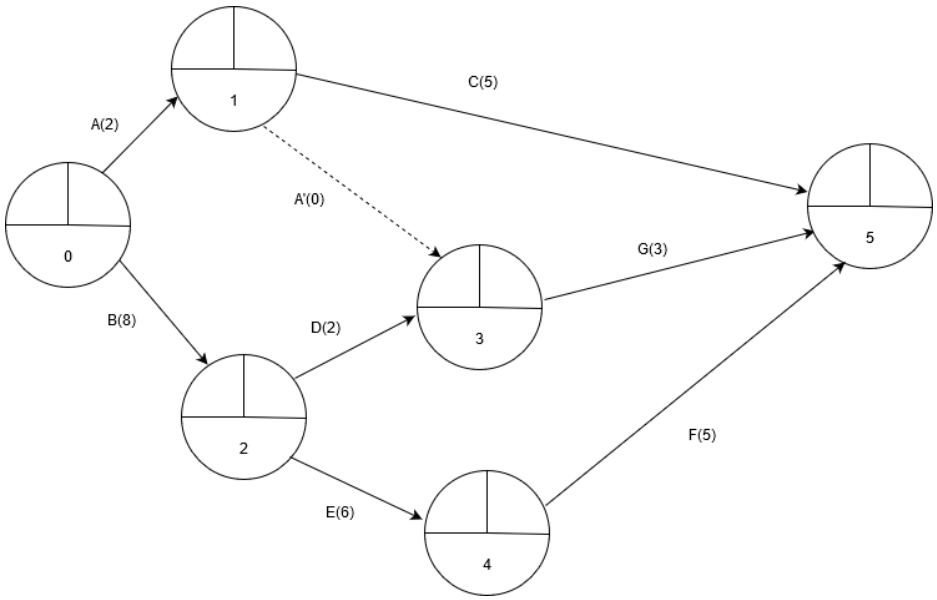
→déclenche→	A	B	C	D	E	F	G
A			X				X
B				X	X		
C							
D							X
E						X	
F							
G							

On cherche alors les colonnes vides, ce qui correspond aux tâches n'ayant pas de prédécesseurs dans le projet. Ici, A et B. Ce sont les tâches de niveau 1. On élimine les colonnes et lignes A et B du tableau et on réitère, ce qui nous donne les tâches de niveau 2. On poursuit ceci jusqu'à la fin.

→précède→	A	B	C	D	E	F	G
A			X				X
B				X	X		
C							
D							X
E						X	
F							
G							

Ceci amène au graphe ci-dessous. Les tâches sont représentées par des arêtes orientées, une arête par tâche. Le fait que G soit déclenché lorsque A est terminé oblige à créer une tâche fictive A' de durée 0 pour

les antécédents de la tâche G. Les nœuds sont numérotés simplement pour pouvoir les nommer. Les quarts de cercle des nœuds serviront aux calculs des dates de début au plus tôt et des dates de fin au plus tard.

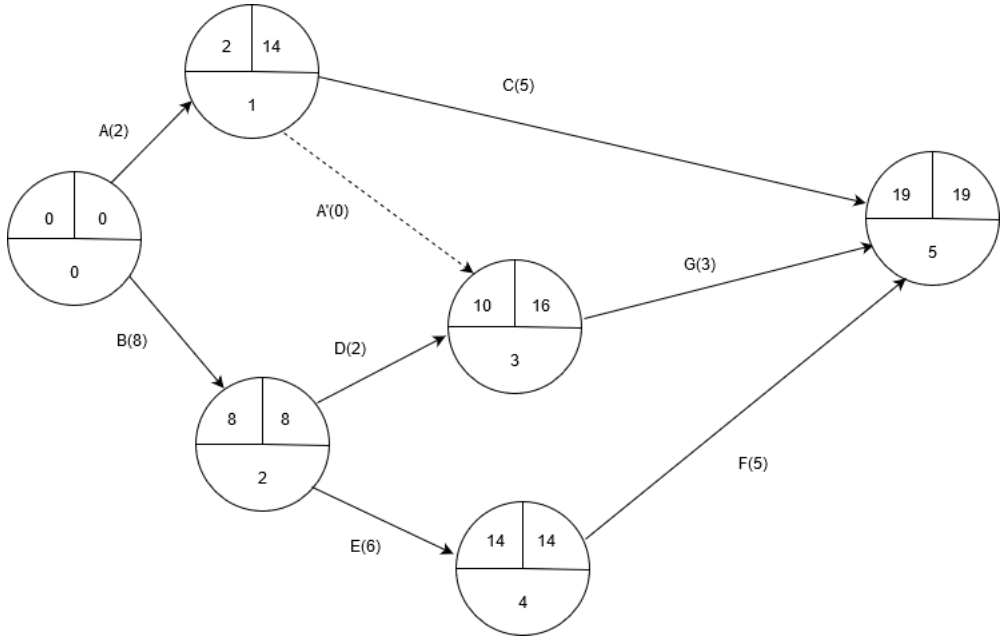


La méthode PERT a pour but de planifier la durée d'un projet ; pour cela des calculs doivent être menés à partir du graphe afin d'en déduire des renseignements sur son exécutabilité. On complète le tableau de synthèse suivant en s'appuyant sur le graphe. Les calculs sont menés en appliquant des algorithmes de cheminement dans un graphe.
On débute en remplissant les dates de début au plus tôt, en appliquant la règle suivante. La date de début au plus tôt d'une tâche se calcule en prenant le maximum des dates de début au plus tôt des tâches la déclenchant augmentées de leur durée. L'algorithme est initialisé par la mise à zéro des dates de début au plus tôt des tâches initiales.
Par exemple, les tâches A et B débutent au temps 0. Les tâches D et E étant déclenchées par la tâche B qui a une date au plus tôt de 0 et d'une durée de 8, ces deux tâches ont donc une date au plus tôt de 8. On utilise la partie en haut à gauche des nœuds du graphe pour calculer ces dates.

Une fois les dates de début au plus tôt complètement remplies, on complète les dates de fin au plus tôt en ajoutant simplement la durée des tâches dans le tableau de synthèse (cf ci-dessous). On poursuit alors avec les dates de fin au plus tard. La date de fin au plus tard d'une tâche se calcule en prenant le minimum des dates de fin au plus tard des tâches qu'elle déclenche déduites de leur durée. L'algorithme est initialisé avec la plus grande date de fin au plus tôt des dernières tâches (ou d'une date supérieure).

Tâches	Durée	Début au plus tôt	Début au plus tard	Fin au plus tôt	Fin au plus tard
A	2	0	12	2	14
B	8	0	0	8	8
C	5	2	14	7	19
D	2	8	14	10	16
E	6	8	8	14	14
F	5	14	14	19	19
G	3	10	16	13	19

On obtient les dates de fin au plus tard en complétant la partie en haut à droite des nœuds du graphe puis les dates de début au plus tard en retranchant les durées des tâches dans le tableau de synthèse. L'outil tableur permet d'automatiser les calculs.



On peut alors donner les quelques définitions suivantes.

- La **marge libre** d'une tâche correspond au retard admissible sur une tâche qui n'entraîne pas de modification des calendriers des tâches suivantes. Elle est égale à la date de début au plus tôt de la tâche suivante (ou de la date de fin de projet s'il n'existe pas de tâche suivante) moins la durée de la tâche moins la date de début au plus tôt de la tâche.
- La **marge totale** correspond au retard admissible du début d'une tâche qui n'entraîne aucun recul de la date de fin du projet, mais qui consomme les marges libres des opérations suivantes. Elle est égale à la date de début au plus tard moins la date de début au plus tôt.
  La marge libre est inférieure ou égale à la marge totale.
- Une tâche est **critique** lorsque sa marge totale est nulle. Un chemin allant du début à la fin du projet est appelé **chemin critique** s'il est constitué uniquement de tâches critiques. C'est un chemin dont la succession des tâches donne la durée d'exécution la plus longue du projet et fournit le délai d'achèvement le plus court. Si l'on prend du retard sur la réalisation de ces tâches, la durée globale du projet est allongée.

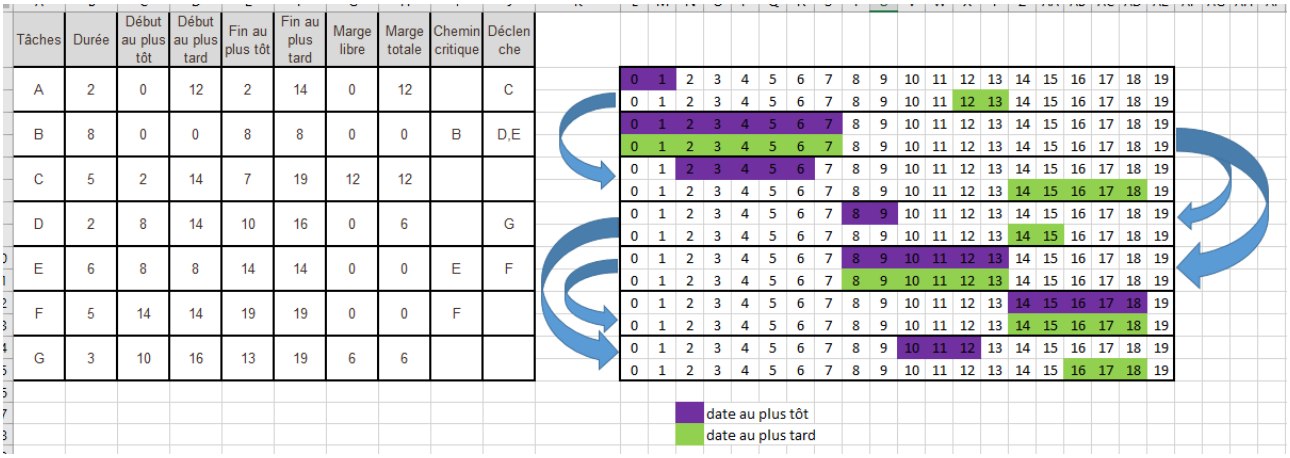
Tableau de synthèse

Tâches	Durée	Début au plus tôt	Début au plus tard	Fin au plus tôt	Fin au plus tard	Marge libre	Marge totale	Chemin critique
A	2	0	12	2	14	0	12	
B	8	0	0	8	8	0	0	B
C	5	2	14	7	19	12	12	
D	2	8	14	10	16	0	6	
E	6	8	8	14	14	0	0	E
F	5	14	14	19	19	0	0	F
G	3	10	16	13	19	6	6	

Ici, B-E-F est un chemin critique.

On pourra se référer à :
http://lycees.ac-rouen.fr/modeste-leroy/spip/IMG/pdf/_PLANIFICATION_et_Ordonnancement-2.pdf
http://www.unit.eu/cours/EnsROtice/module_avance_thg_voo6/co/module_avance_thg_13.html
Le graphe de PERT peut être complété par le diagramme de Gantt. Le diagramme de Gantt est un graphique qui consiste à placer les tâches chronologiquement en fonction des contraintes techniques de succession

(contraintes d'antériorités). L'axe horizontal des abscisses représente le temps et l'axe vertical des ordonnées les tâches. On y représente chaque tâche par un segment dont la longueur est proportionnelle à sa durée. L'origine du segment est calée sur la date de début au plus tôt de l'opération (diagramme de jalonnement au plus tôt) et l'extrémité du segment représente la fin de la tâche. On pourra à cet effet automatiser le diagramme de Gantt des dates au plus tôt dans un tableur à partir de formules et de mises en forme conditionnelle.



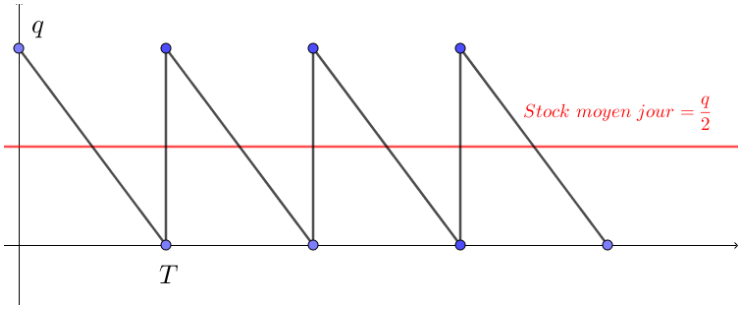
Tâches	Durée	Début au plus tôt	Début au plus tard	Fin au plus tôt	Fin au plus tard	Marge libre	Marge totale	Chemin critique	Déclenche
A	2	0	12	2	14	0	12		C
B	8	0	0	8	8	0	0	B	D,E
C	5	2	14	7	19	12	12		
D	2	8	14	10	16	0	6		G
E	6	8	8	14	14	0	0	E	F
F	5	14	14	19	19	0	0	F	
G	3	10	16	13	19	6	6		

C62 – Gérer l'activité de production du système de production

Les contraintes de gestion des stocks amènent à étudier le modèle de Wilson de gestion des stocks et l'analyse ABC (méthode de Pareto) de classification des stocks.

- ***Modèle de Wilson pour optimiser les coûts***

Le stockage de produits et le réapprovisionnement induisent des coûts de stockage ainsi que des frais de livraison. Il faut donc jouer sur ces paramètres et trouver une optimisation. Le modèle de Wilson est un modèle très simplifié qui permet de comprendre comment déterminer cet optimum et comprendre le fait que des logiciels peuvent être paramétrés. La détermination de cet optimum en situation professionnelle est à réaliser à l'aide de logiciels professionnels propres à chaque option.
On ne tient pas compte du stock de sécurité, mais de la quantité *q* de réapprovisionnement commandée et consommée régulièrement sur une période T, le **stock moyen** par jour peut être évalué à $\frac{q}{2}$.



Un exemple pour comprendre

Une exploitation horticole utilise 10 000 sacs de terreau par année dans la confection de pots de plantes et de potées fleuries et pour satisfaire les commandes et les besoins de ses clients. Chaque sac s'achète au prix de 7 €, le coût de passation de commande (ensemble des coûts supportés par l'entreprise lors de l'achat d'une commande) a été estimé par le chef d'entreprise à 20 €, le coût du stockage entreposage à l'abri, amortissement du transpalette…) se chiffre à 20% du prix unitaire d'un sac.
Le chef d'exploitation vous demande s'il pourrait améliorer sa rentabilité, en changeant sa méthode de commande. L'objectif est de déterminer le nombre de commandes et la quantité optimale *q* à commander pour optimiser les coûts.

Si l'on n'effectue sur l'année qu'une commande de 10 000 sacs, le coût total se décompose en trois parties :

- le coût de passation de commande sur la période, soit 20 €.
- Le stock moyen sur la période est $\frac{10000}{2} = 5000$, donc le coût de stockage sur la période est 5000× 7 × 20% = 7 000€
- Le coût de l'ensemble des sacs de 10 000 × 7 = 70 000 €

On obtient un coût total de 70 000+ 20 + 7 000= 77 020 €

Si on répartit 16 commandes de 625 sacs dans l'année :

- le coût de passation par commande est de 20 € , donc 16 × 20 = 320 €
- Le stock moyen par commande est $\frac{625}{2} = 312{,}5$, donc le coût de stockage sur l'année est 312,5×7 × 20 % = 437,5 €
- Le coût d'achat de l'ensemble des sacs reste identique soit 70 000 €

Le coût total dans ce cas est 70 000 + 437,5 + 320 = 70 757,5 €, soit un gain de 6262,5 €.

Ces deux exemples montrent qu'il ne revient pas au même de faire une ou plusieurs commandes.
Une automatisation sur Excel peut être faite avec une recherche de minimum avec le solveur.

De façon plus générale, si l'on consomme 10 000 sacs sur l'année et que l'on effectue des commandes de *q* sacs par commande, il faudra réaliser $\frac{10000}{q}$ commandes.

- le coût de passation sur la période annuelle est de $20 \times \frac{10\,000}{q} = \frac{200\,000}{q}$.
- Le stock moyen jour sur la période est $\frac{q}{2}$, donc le coût de stockage sur la période annuelle est

$$\frac{q}{2} \times 7 \times 20\ \% \ = 0{,}7\ q$$

- Le coût d'achat de l'ensemble des 10000 sacs de 10 000 × 7 = 70 000 €
- Le coût annuel s'exprime en fonction de la quantité q à commander par :

$$C(q) = 70\,000 + \frac{200\,000}{q} + 0{,}7q$$

La représentation graphique suggère l'existence **d'une quantité optimale** à commander que l'on peut déterminer graphiquement, à l'aide de la calculatrice, par l'étude des variations de la fonction C

$C'(q) = 0{,}7 - \frac{200000}{q^2} = 0$ pour $q \approx 534{,}53$, ce qui correspond à environ $\frac{10000}{535} \approx 19$ commandes et un coût annuel est de 70748,33 €.

Pour poursuivre la réflexion :
Le chef d'entreprise pourra ajouter les contraintes suivantes dans son cahier des charges :

- La place de stockage dans des conditions optimales ;
- l'occupation des espaces aux différents usages de l'entreprise ;
- la sécurité des personnes dans le poste de travail (stabilité des palettes empilées), manipulation des palettes, manipulation des sacs en sécurité par les personnes… ;
- la disponibilité des moyens de déplacement des sacs ;
- …

La généralisation doit se faire à l'aide de données issues de situations en lien avec les sciences et techniques professionnelles par la répétition de telles démarches. Pour la compréhension de la démarche, il faut renouveler plusieurs calculs de quantités optimales dans des situations concrètes. La connaissance de la formule générale donnant la valeur optimale de q que l'on trouve dans beaucoup de littérature n'est pas un attendu. Cela peut toutefois expliquer comment se paramètrent des logiciels permettant de donner les quantités optimales minimisant les coûts.

C82 – Produire des références

L'enseignement du module M8 s'appuyant 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 développer des méthodes statistiques à partir de simulations.

La mise en œuvre d'essais amène à considérer le schéma général d'une expérimentation en lien avec les disciplines techniques. Le vocabulaire associé est indispensable à la communication entre les différents acteurs des essais.

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 continue et donc de développer la notion de loi continue, majoritairement inconnue des étudiants.

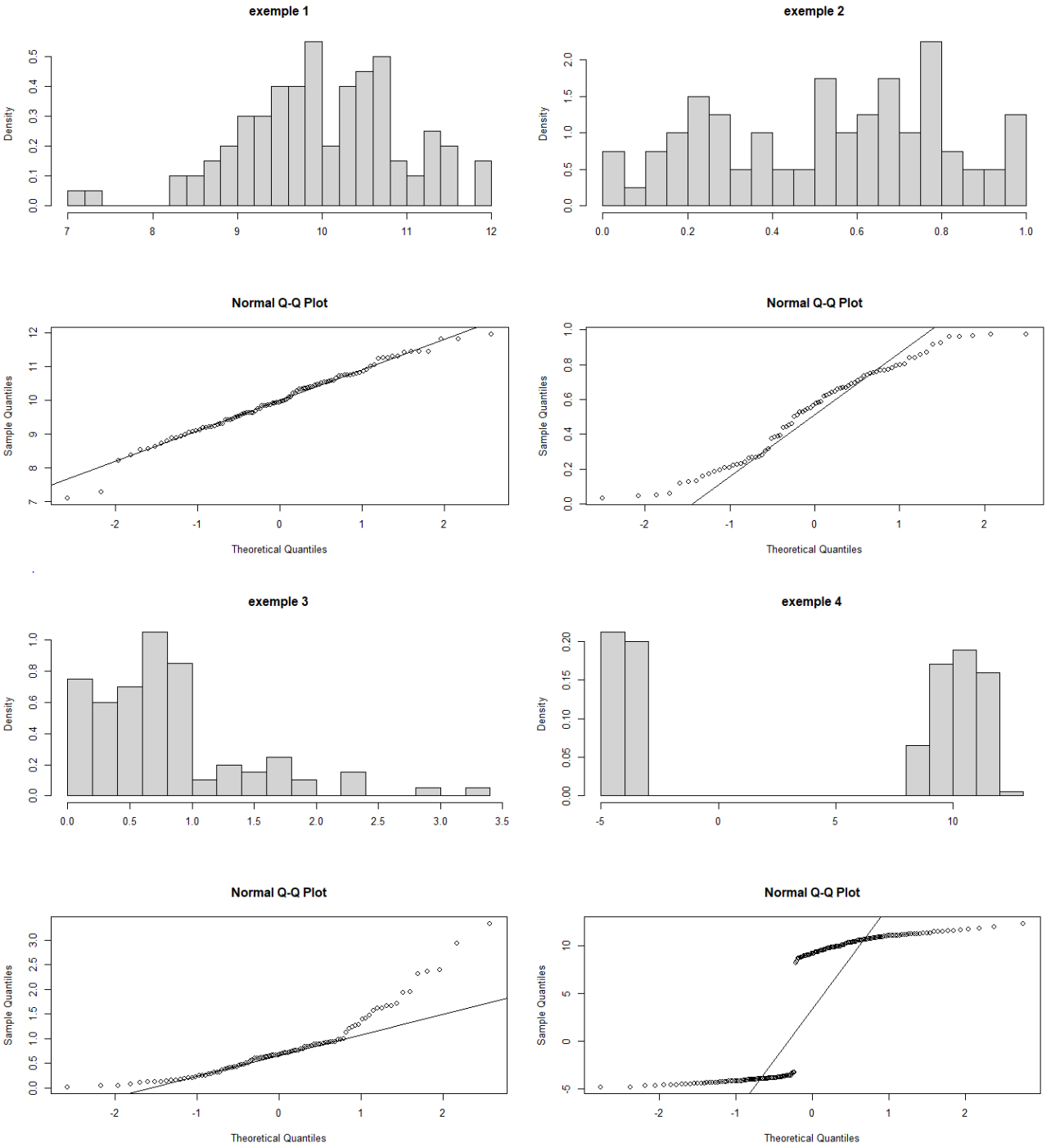
L'enseignement doit concourir à développer la capacité à repérer des situations de référence de mise en œuvre de tests statistiques. L'objectif est moins de faire apprendre un catalogue de tests statistiques que de faire comprendre la méthodologie des tests et la construction de règles de décision s'appuyant sur la fluctuation d'échantillonnage de certaines grandeurs obtenue 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 élèves, l'enseignant veille à avoir des exemples suffisamment diversifiés. Tous les calculs sont laissés à l'outil numérique. Même si son apprentissage requiert un certain investissement, 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.

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 peut être nécessaire de débuter par un test de normalité.

- 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.
- 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.
- À partir d'essais réalisés par les apprenants ou d'études publiées, mettre en œuvre des tests statistiques permettant de répondre à une problématique. Les tests à pratiquer sont à choisir de préférence dans les tests de conformité d'une proportion, d'une moyenne, de comparaison d'une proportion, d'une moyenne, d'une variance, d'indépendance et d'analyse de la variance à un facteur ainsi que le test de Newman-Keuls.
- Réaliser l'ajustement entre deux grandeurs observées. L'approche graphique est à privilégier. La forme des nuages doit être reliée aux fonctions usuelles (fonction affine, carrée, exponentielle, …), pour en dégager un type d'ajustement (linéaire, polynomial, exponentiel, …). Les calculs sont laissés à l'outil numérique. Le travail se porte sur le choix de l'ajustement.
- 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 : 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. Voir les codes en annexe1.



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

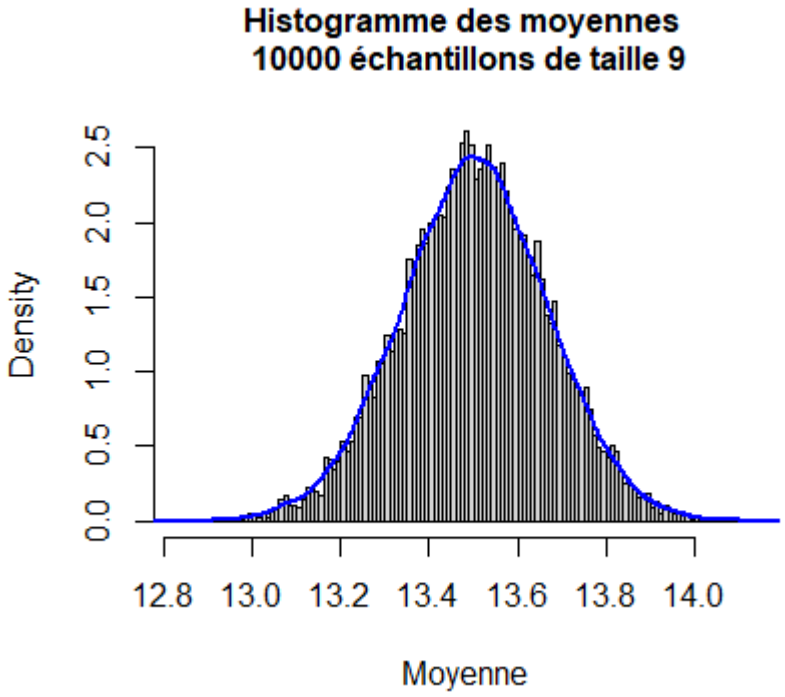
Pour le test de Kolmogorov-Smirnov, voir la commande `ks.test()` du logiciel R.

Exemple 2 : Introduire le test de conformité d'une moyenne

Le cahier des charges du label rouge n° LA 04/96 « Pommes » impose une teneur en sucre de 13,5° brix minimum.
On décide de mettre en place un contrôle de la teneur en sucre en prélevant des échantillons de taille $n = 9$.
On admet que la distribution de la teneur en sucre est gaussienne et d'écart type $\sigma = 0{,}5$° brix.
Un échantillon est prélevé et on obtient par exemple les valeurs :

13,2 13,5 13,7 13,3 13,3 13,6 13,5 13,6 13,5

On simule par exemple 10000 échantillons de taille $n = 9$ d'une loi normale de paramètres $\mu = 13{,}5$ et $\sigma = 0{,}5$ et on étudie la distribution de la moyenne des volumes de chaque échantillon. L'outil numérique permet d'obtenir un graphique du type ci-dessous. En s'appuyant sur ce graphique, on peut alors s'interroger sur la distribution d'échantillonnage d'une moyenne en particulier la normalité de la distribution puis amener le questionnement de la conformité d'une moyenne. La mise en place du test de conformité d'une moyenne et l'utilisation de la loi normale sont l'aboutissement de l'étude. On discute de l'unilatéralité ou la bilatéralité du test.



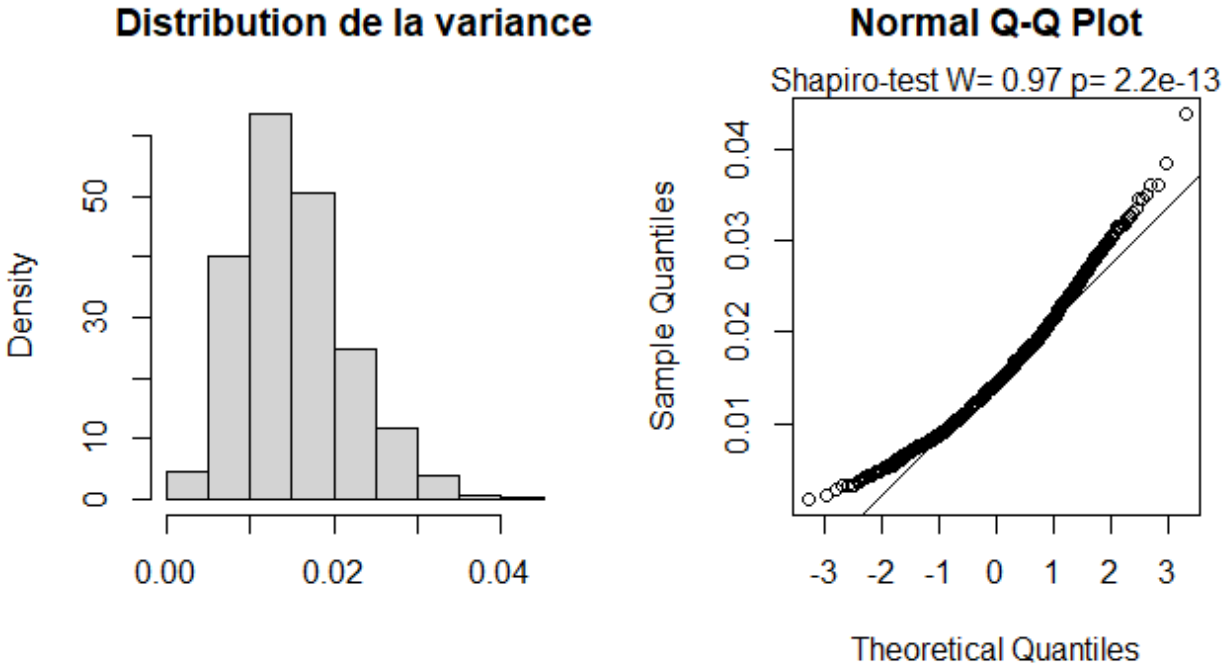
Exemple 3 : Introduire une nouvelle variable statistique

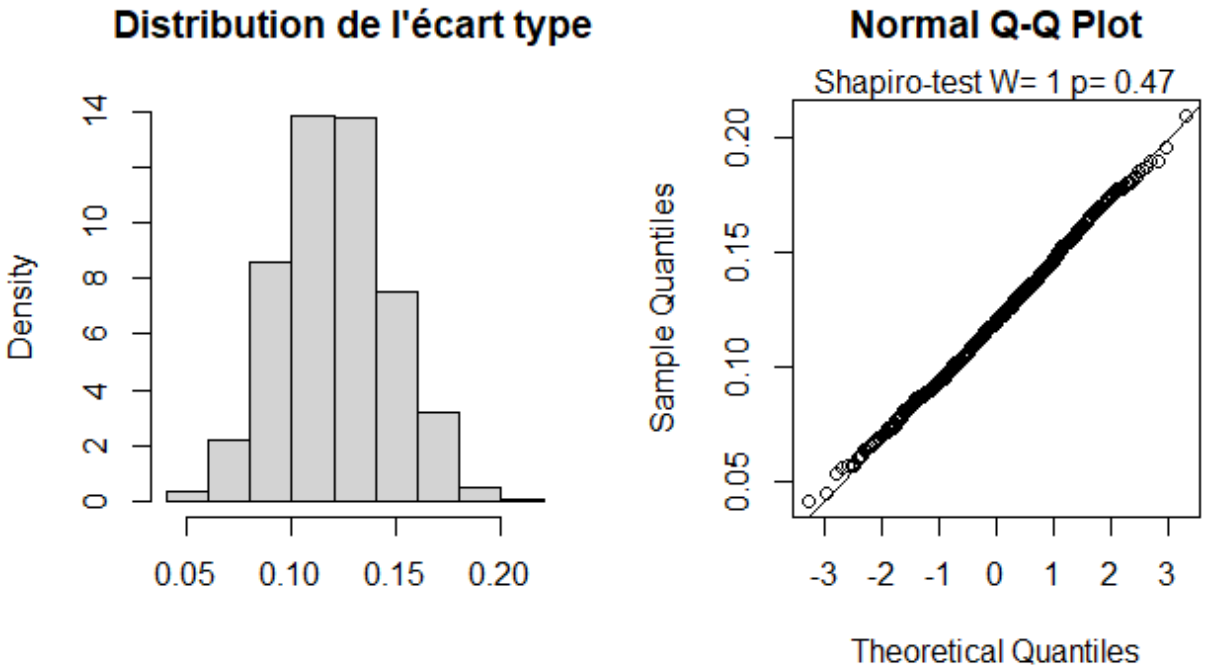
Le pH du substrat a une influence sur le développement des plantes. Les plantes neutrophiles supportent des substrats dont le pH est entre 6 et 6,5. Pour l'exemple, on choisit 6,2. On mesure à plusieurs endroits de l'exploitation le pH des substrats. On souhaite vérifier la stabilité du pH et on s'intéresse donc à la variabilité des résultats. On admet que l'on obtient des mesures avec une résolution de 0,01 pH.
Les pH d'un échantillon de 12 prélèvements sont les suivants :

Échantillon 0 : 6,26 6,33 6,19 6,21 6,25 6,19 6,05 6,14 6,47 6,13 6,21 6,44

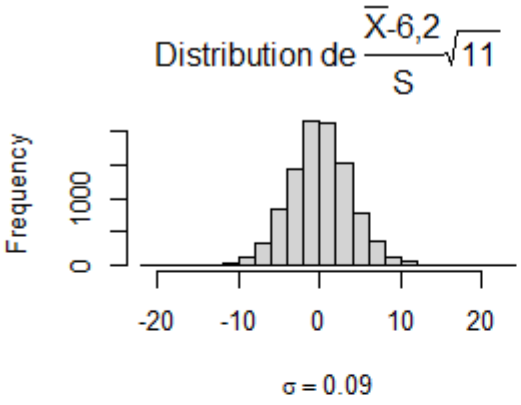
Après avoir posé la question de la normalité des valeurs de l'échantillon, l'estimation de l'écart type est au cœur des discussions. On est amené à étudier graphiquement les distributions des variances et des écarts types pour des échantillons gaussiens de taille 12 obtenus par simulation de la loi normale de moyenne 6,2

et d'écart type égale à l'écart type de l'échantillon 0. On peut au passage se poser la question de la normalité de ces distributions.

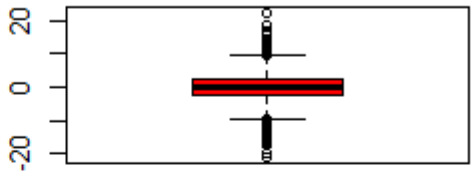




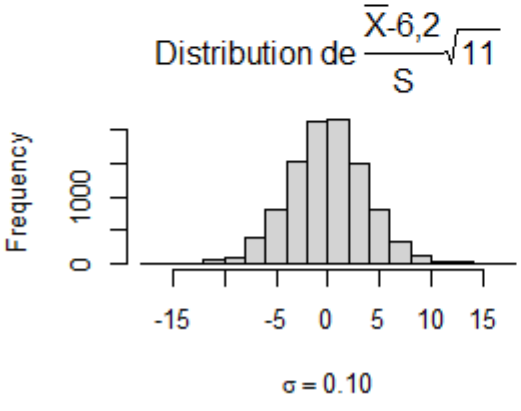
La variabilité de l'écart type des échantillons et le choix arbitraire de la valeur de l'écart type de l'échantillon 0 plutôt qu'une autre valeur amène à considérer une autre variable statistique $\frac{\bar{X}-6,2}{S}\sqrt{11}$ où chaque échantillon simulé a la même importance dans le choix de l'écart type. On peut constater que si l'on fait varier l'écart type autour de la valeur de l'échantillon 0, la loi de cette variable reste stable, ce qui peut se constater en observant les histogrammes et les diagrammes en boîte ci-dessous.



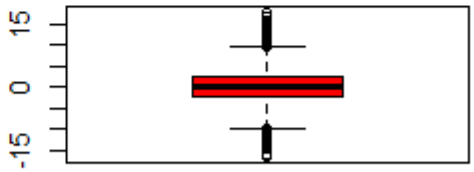
σ = 0.09



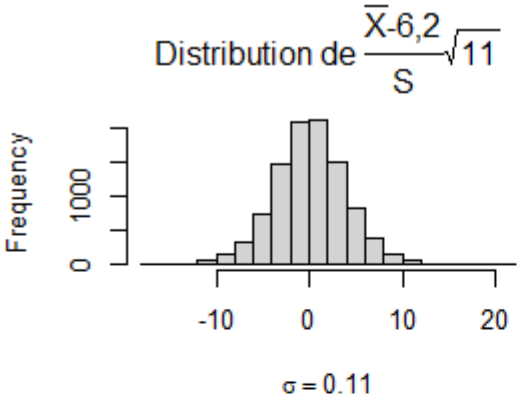
σ = 0.09



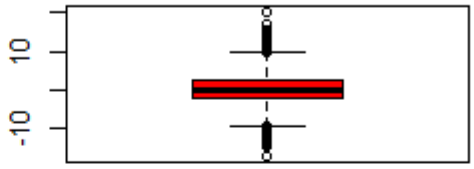
σ = 0.10



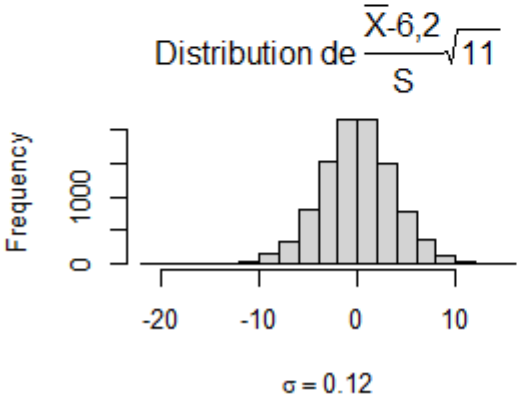
σ = 0.10



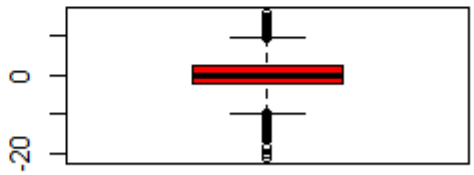
σ = 0.11



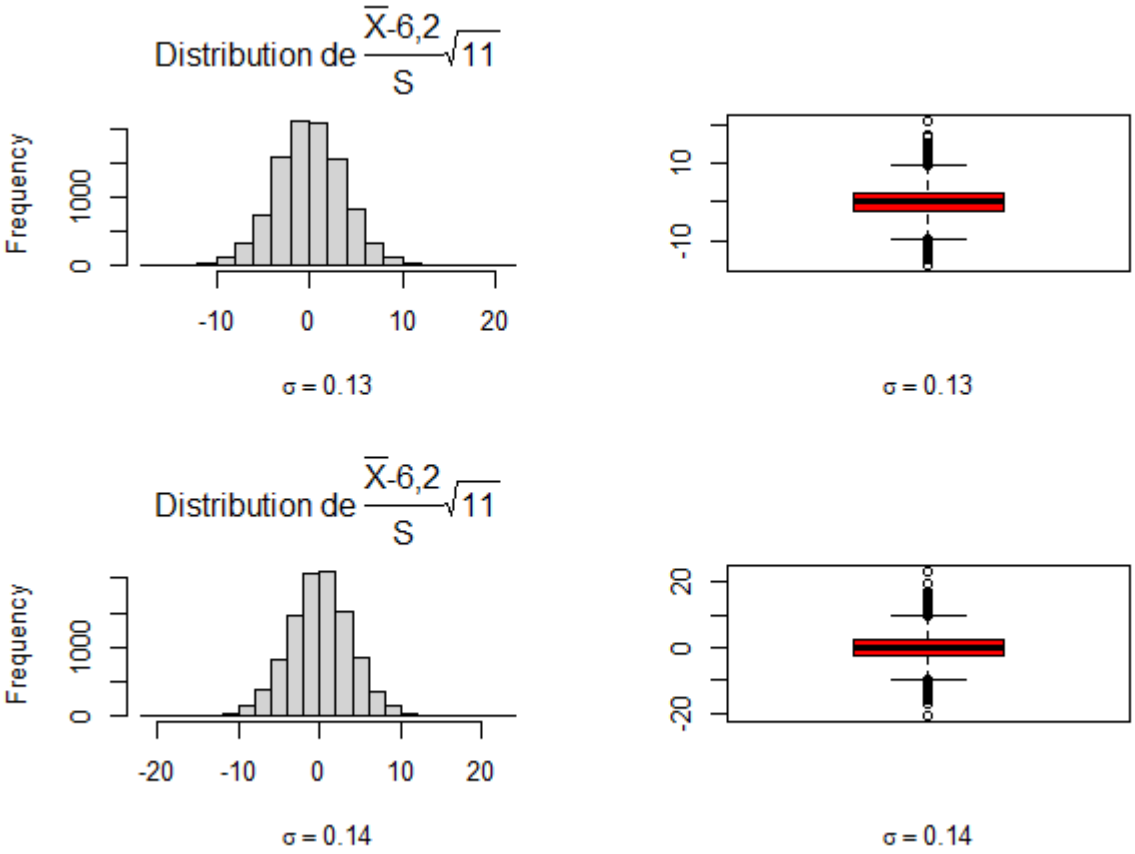
σ = 0.11



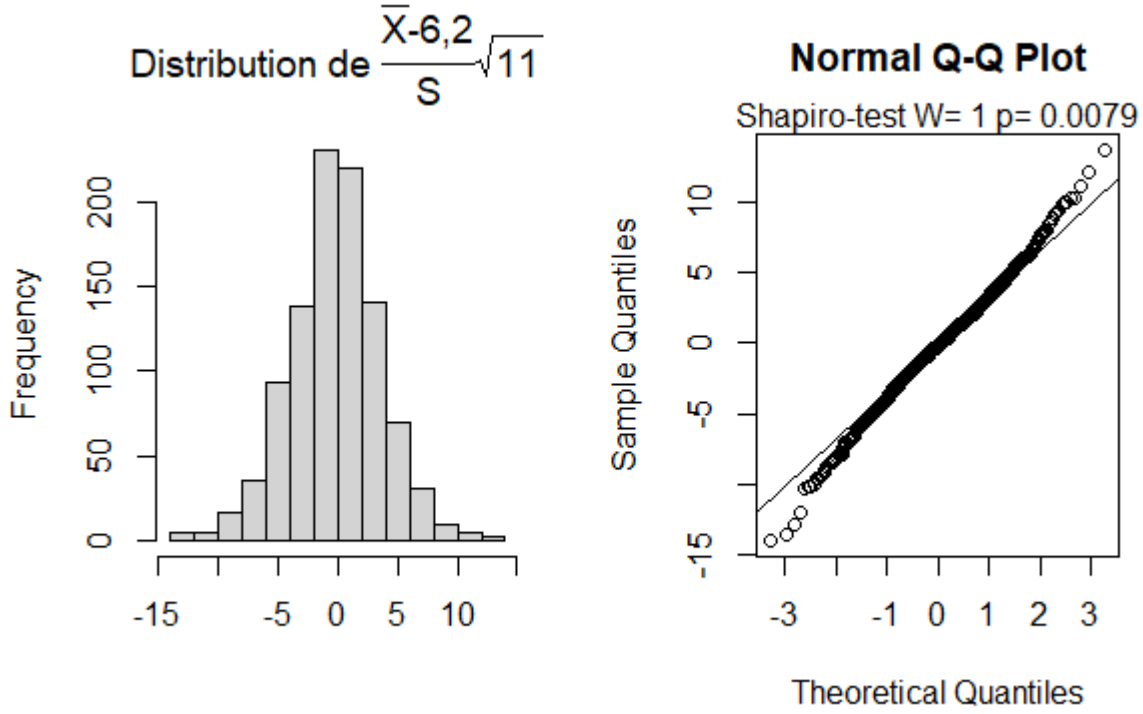
σ = 0.12



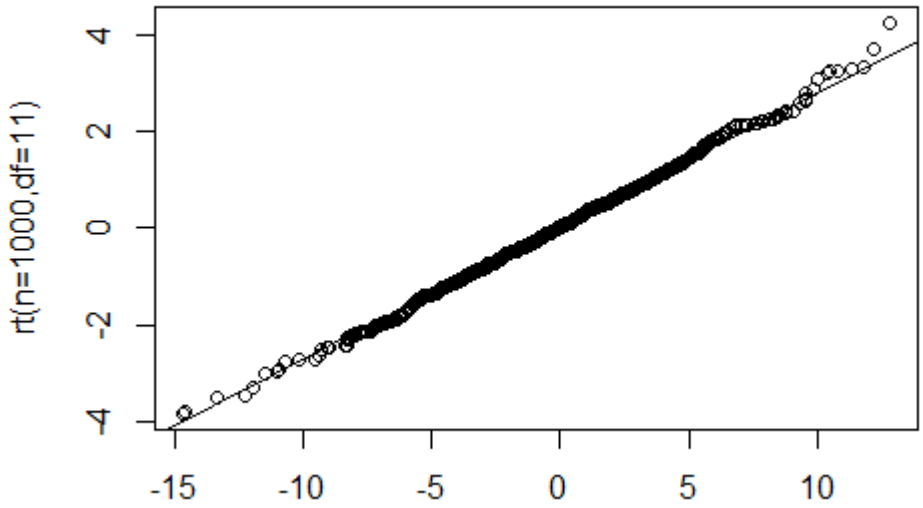
σ = 0.12



Le rejet de la loi normale pour cette nouvelle variable statistique induira l'introduction de la loi de Student.



On peut pour confirmer construire le graphique quantile-quantile entre $\frac{\bar{X}-6,2}{S}\sqrt{11}$ et la loi de Student T_{11}.



La mise en place du test avec la loi de Student conclut l'exemple. On peut discuter de l'unilatéralité ou de la bilatéralité du test.

Exemple 4 : Introduire le test d'indépendance du Khi2

Afin de comparer l'influence de deux paillages sur la culture des fraisiers, on paille 91 placettes d'une surface de 1 m² avec des aiguilles de pins et protège 107 autres placettes de même dimension avec de la bâche plastique.
Sur ces placettes sont plantées des fraisiers « Mara des bois » qui ont reçu la même quantité de compost.
Les mesures ont porté sur la masse des fraises récoltées sur chaque placette. Selon les cas, les quantités sont qualifiées de très satisfaisante, satisfaisante ou faible. On cherche à évaluer l'influence du type de paillage.
Le traitement de l'exemple est effectué avec le logiciel R.

Production / Type de paillage	Faible	Satisfaisante	Très satisfaisante	Total
Aiguilles de pin	13	51	27	91
Bâche plastique	35	46	26	107
Total	48	97	53	198

Le tableau ci-dessus peut être saisi dans R via l'instruction

```
>tableau=data.frame(faible=c(13,35),satisf=c(51,46),tressatisf=c(27,26),
                    row.names=c('Aiguilles','Bâche))
```

Dans le cas où les variables seraient indépendantes, on s'attend à obtenir le tableau des effectifs suivants.

Production / Type de paillage	Faible	Satisfaisante	Très satisfaisante	Total
Aiguilles de pin	22.06061	44.58081	24.35859	91
Bâche plastique	25.93939	52.41919	28.64141	107
Total	48	97	53	198

Celui-ci s'obtient en récupérant le résultat du test dans une variable que l'on va nommer khi puis l'affichage des effectifs attendus.

>khi=chisq.test(tableau)
>khi$expected

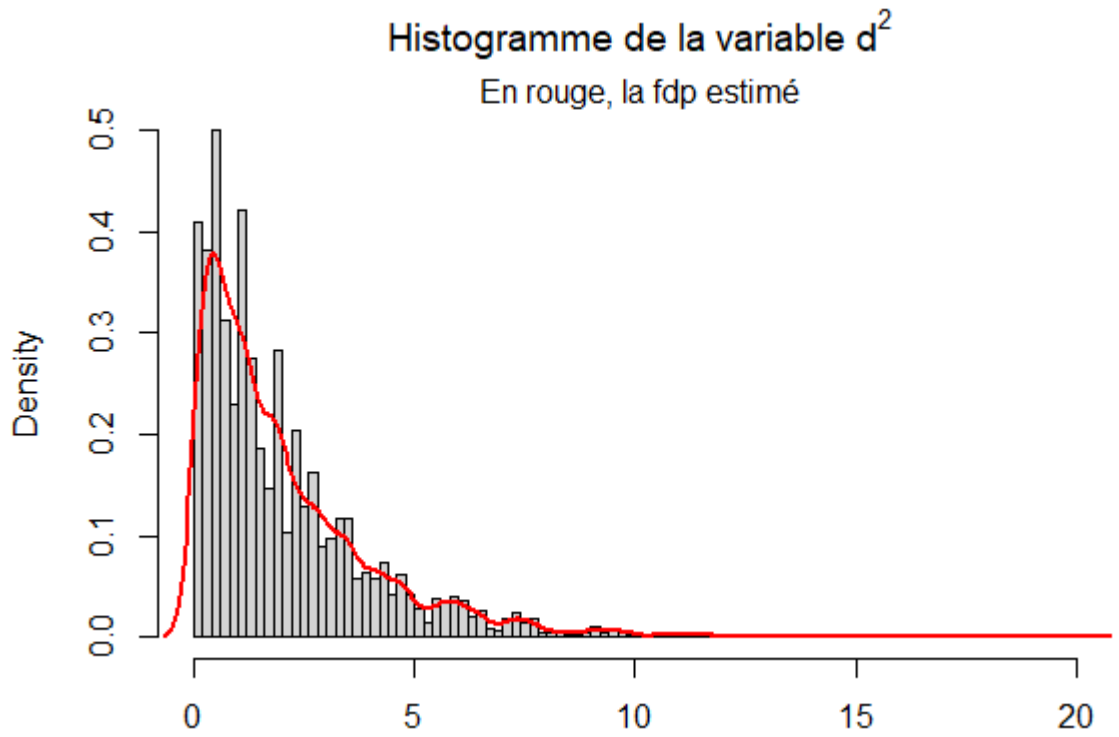
Dans le cas de l'indépendance « parfaite », le tableau des effectifs attendus devrait être identique au tableau des effectifs observés. La mesure de l'indépendance amène donc à mesurer l'écart existant entre les deux tableaux via la formule :

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

où les $O_{i,j}$ correspondent aux effectifs observés et les $E_{i,j}$ aux effectifs attendus sous l'hypothèse d'indépendance.
Pour obtenir une idée de la distribution de la variable d^2, on simule des tableaux dont les lignes et les colonnes sont indépendantes et ayant les mêmes marges que le tableau initiale. Ce type de simulation étant difficile à mettre en œuvre, on peut plutôt pour faire émerger la loi du χ^2 s'intéresser à l'adéquation à une loi. (cf annexe2). Pour les plus curieux, on peut consulter l'algorithme RCont dû à Boyett[1] sur la génération des tableaux de contingence de marges en ligne et en colonne fixées.
Les lignes de commande[2] ci-dessous permettent d'obtenir l'histogramme où la variable informatique d contient le calcul de d^2 pour 1000 simulations de tableaux, ainsi que la fonction de densité de probabilité estimée.

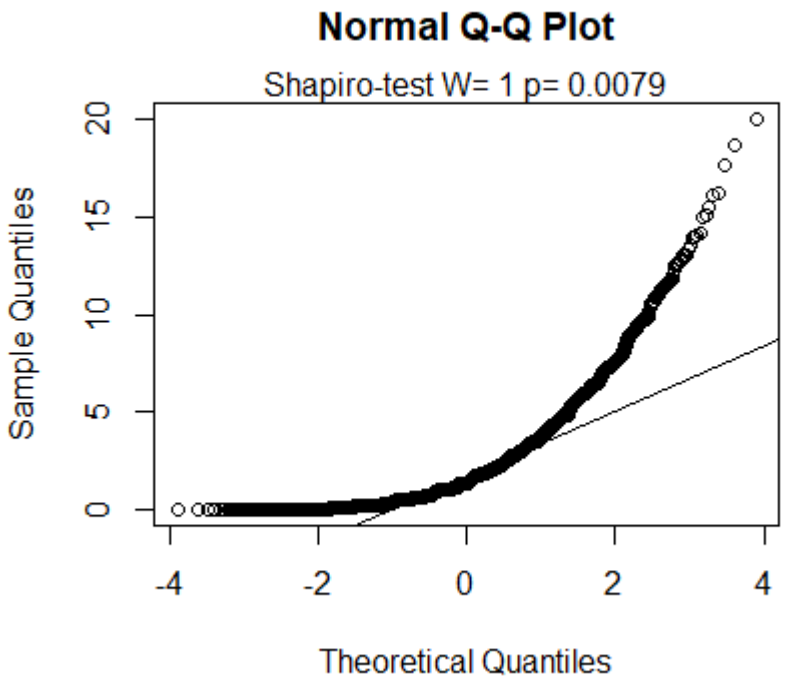
>hist(d,breaks=100,prob=T)

>lines(density(d),col='red',lwd=2)

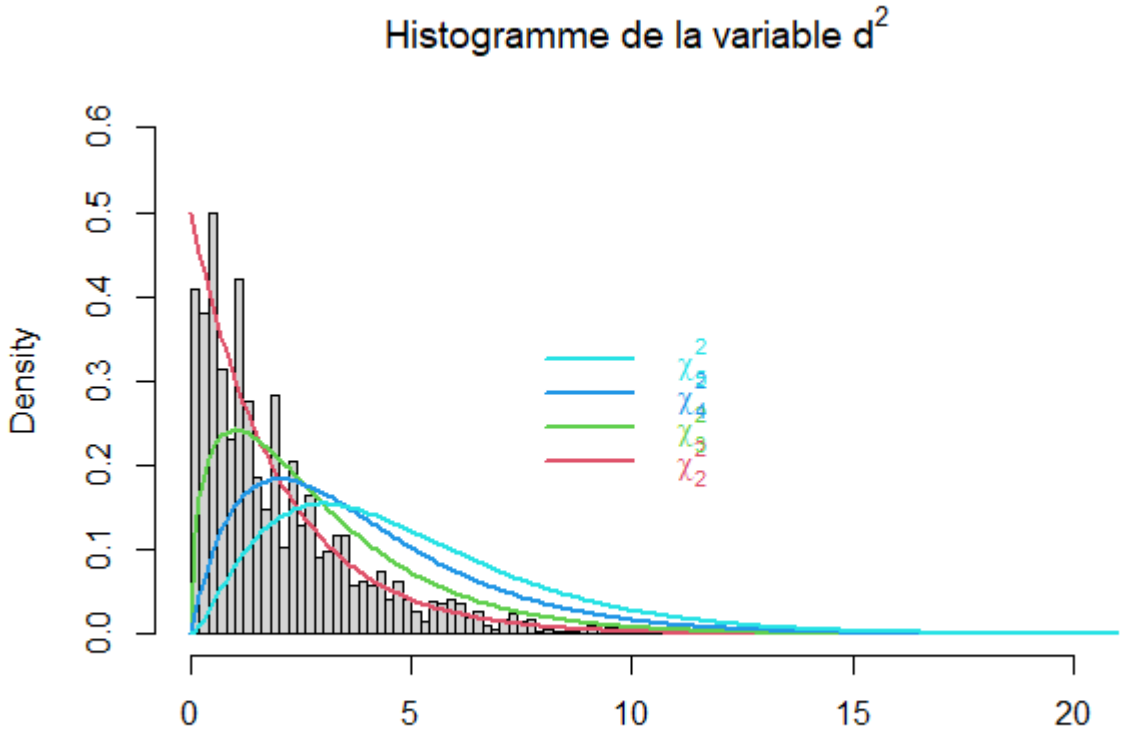


On peut alors tester la normalité de cette distribution. Avec les commande qqnorm(d), qqline(d) et shapiro.test(d)

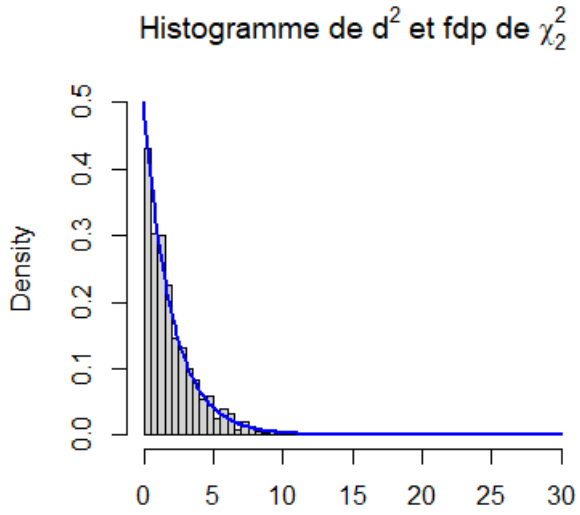
[1] James M. Boyett *Journal of the Royal Statistical Society. Series C (Applied Statistics)*Vol. 28, No. 3 (1979), pp. 329-332
[2] Voir annexe1 pour davantage de commandes.

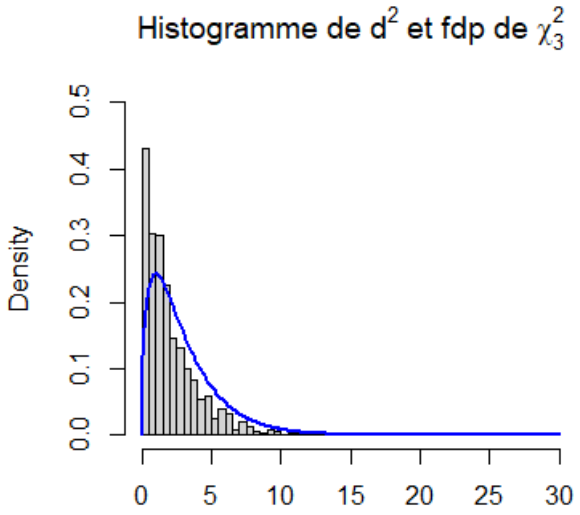


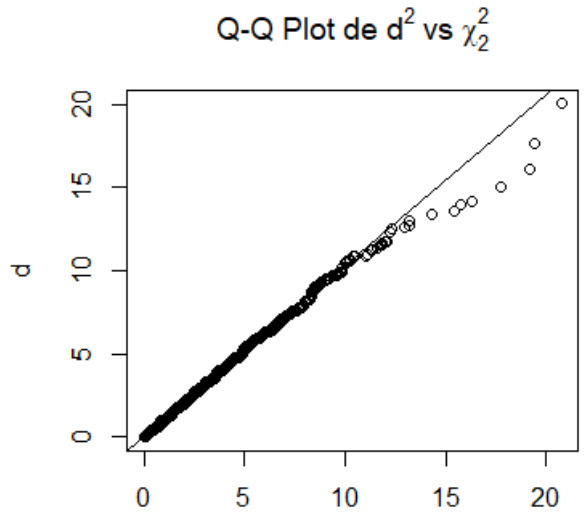
On rejette donc la normalité, ce qui induit la recherche d'une loi approchant d^2.
La forme de l'histogramme incite à essayer une loi du Khi2 avec un certain degré de liberté.

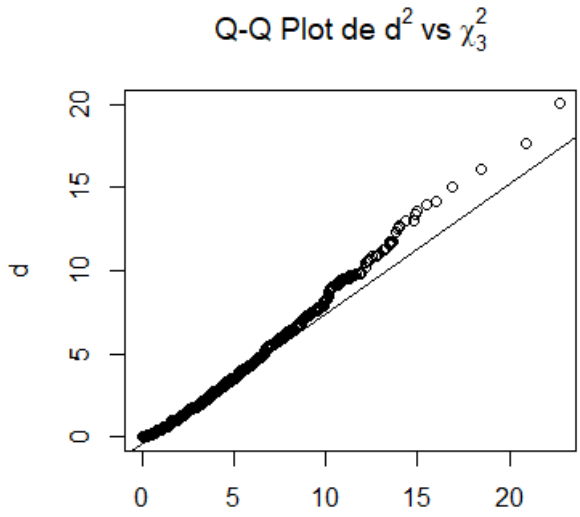


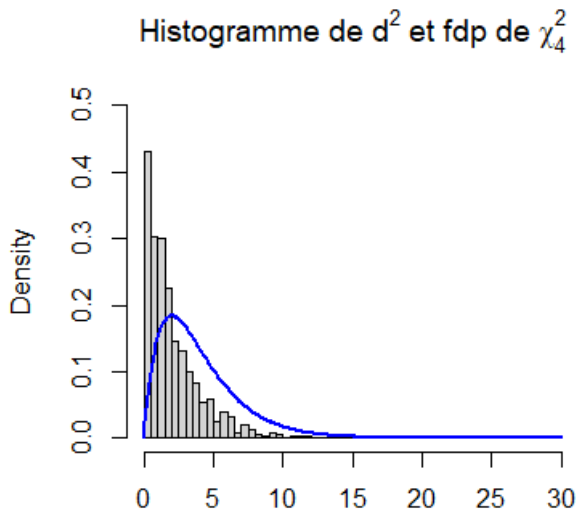
Etudions alors cas par cas.

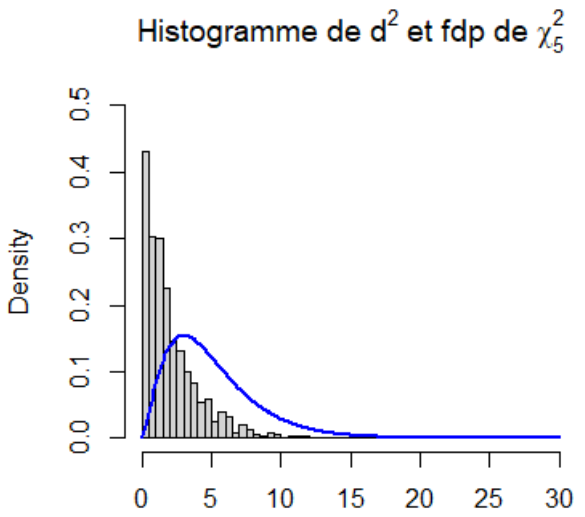


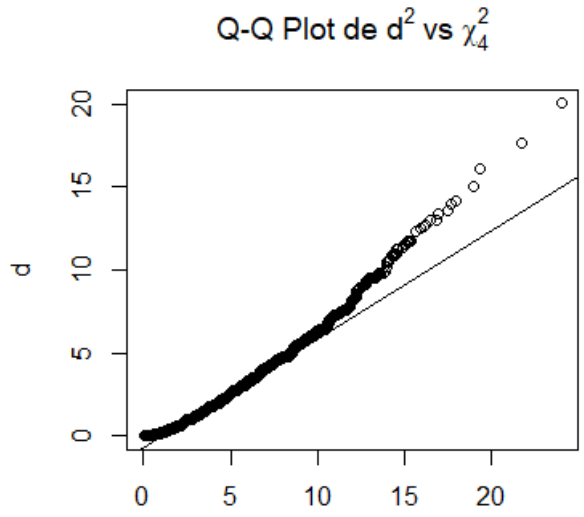


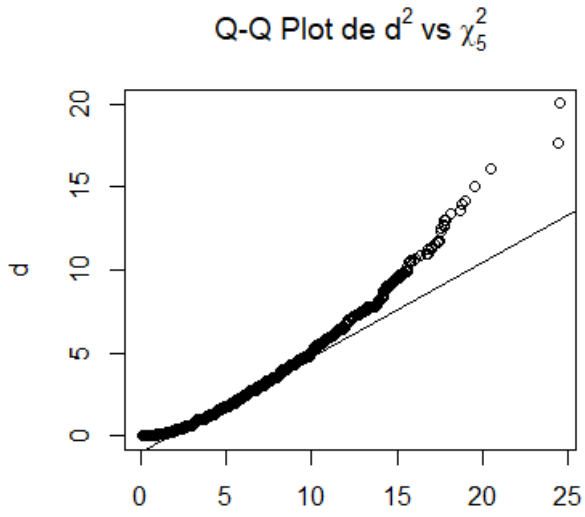






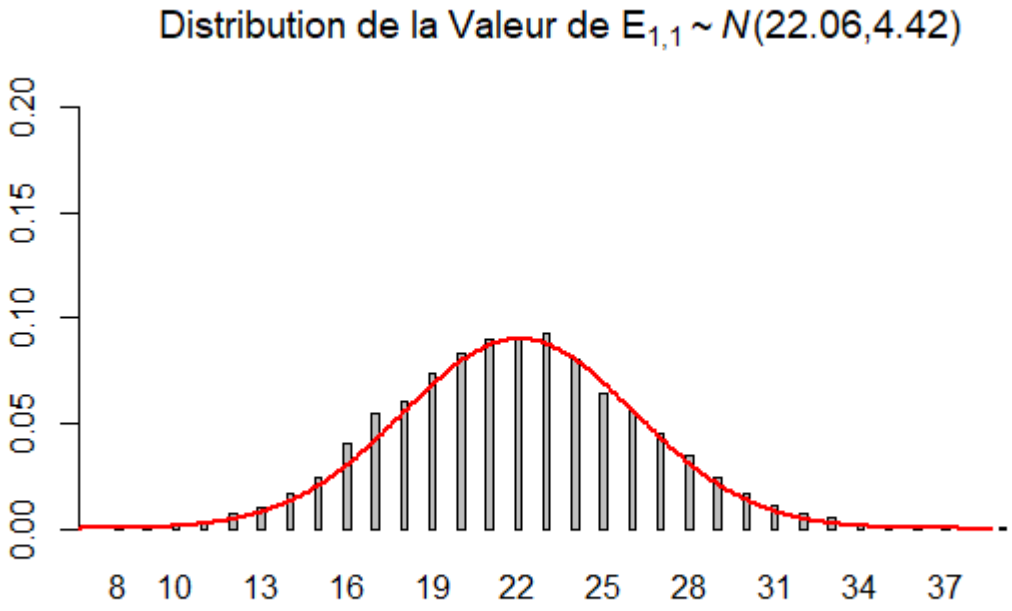






On peut alors émettre la conjecture que d^2 semble suivre une loi du χ^2 à 2 degrés de liberté. Cette approche par simulation est complétée par la mise en place du test du χ^2 puis l'institutionnalisation de la loi asymptotique de d^2.

Dans le cas du rejet de l'indépendance, l'étude se poursuit par l'analyse des écarts à l'indépendance. On peut à cet effet s'intéresser à une unique cellule du tableau. Par exemple, choisissons la cellule correspondant à un paillage de type Aiguilles de Pin et une production faible. Sous l'hypothèse d'indépendance, la valeur attendue est $E_{1,1} \approx 22{,}06$ et la fréquence attendue est environ égale à 0,1114. On peut alors par simulation étudier la distribution des valeurs de cette cellule et en déduire « l'écart » à la valeur attendue. D'après le théorème central limite, la distribution des valeurs de cette cellule suit approximativement une loi normale de moyenne $E_{1,1} \approx 198 \times 0{,}1114 \approx 22{,}06$ et d'écart type $\sqrt{198 \times 0{,}1114(1-0{,}1114)} \approx \sqrt{E_{1,1} \times (1-0{,}1114)} \approx 4{,}42$.



La valeur observée est 13, qui sous les conditions de l'indépendance, a peu de chance d'être observée, de probabilité inférieure à 0,05.

On peut alors considérer les résidus de Pearson standardisées égaux à $\frac{O_{i,j}-E_{i,i}}{\sqrt{E_{i,j}}}$ que l'on retrouve dans l'expression de la variable statistique d^2. Les résidus de Pearson standardisées vont suivre approximativement une loi normale centrée et d'écart type inférieur à 1. Ils mesurent l'attraction (positif) ou la répulsion (négatif) par rapport à la valeur attendue. Les résidus, correspondant à des écarts statistiquement significatifs et dont la contribution à d^2est forte, sont ceux dont la valeur est supérieure à 2 ou inférieure à -2, qui correspondent à au moins deux écarts types dans la loi normale centrée réduite.

Le tableau des résidus de Pearson s'obtient sous R avec la commande >khi$residuals où khi est la variable informatique contenant le tableau des effectifs.

Production / Type de paillage	Faible	Satisfaisante	Très satisfaisante
Aiguilles de pin	-1.929072	0.9614050	0.5351930
Bâche plastique	1.779006	-0.8866153	-0.4935592

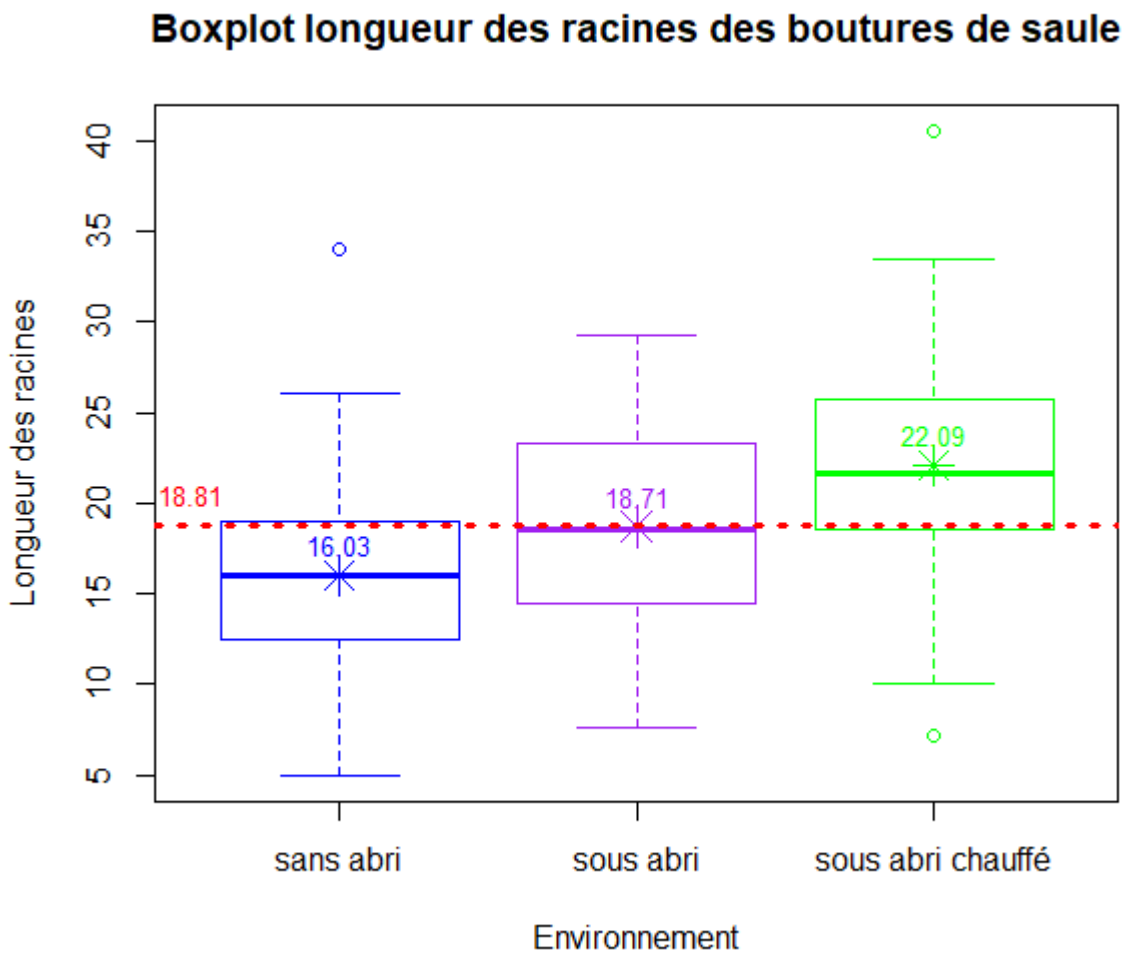
Le rejet de l'indépendance peut s'expliquer, par exemple, par une sous-représentation de la valeur observée du paillage aux aiguilles de pin et d'une faible production et par la surreprésentation de la valeur observée du paillage bâche plastique et d'une faible production par rapport aux valeurs attendues.

Exemple 5 : Analyse de variance (ANOVA à 1 facteur)

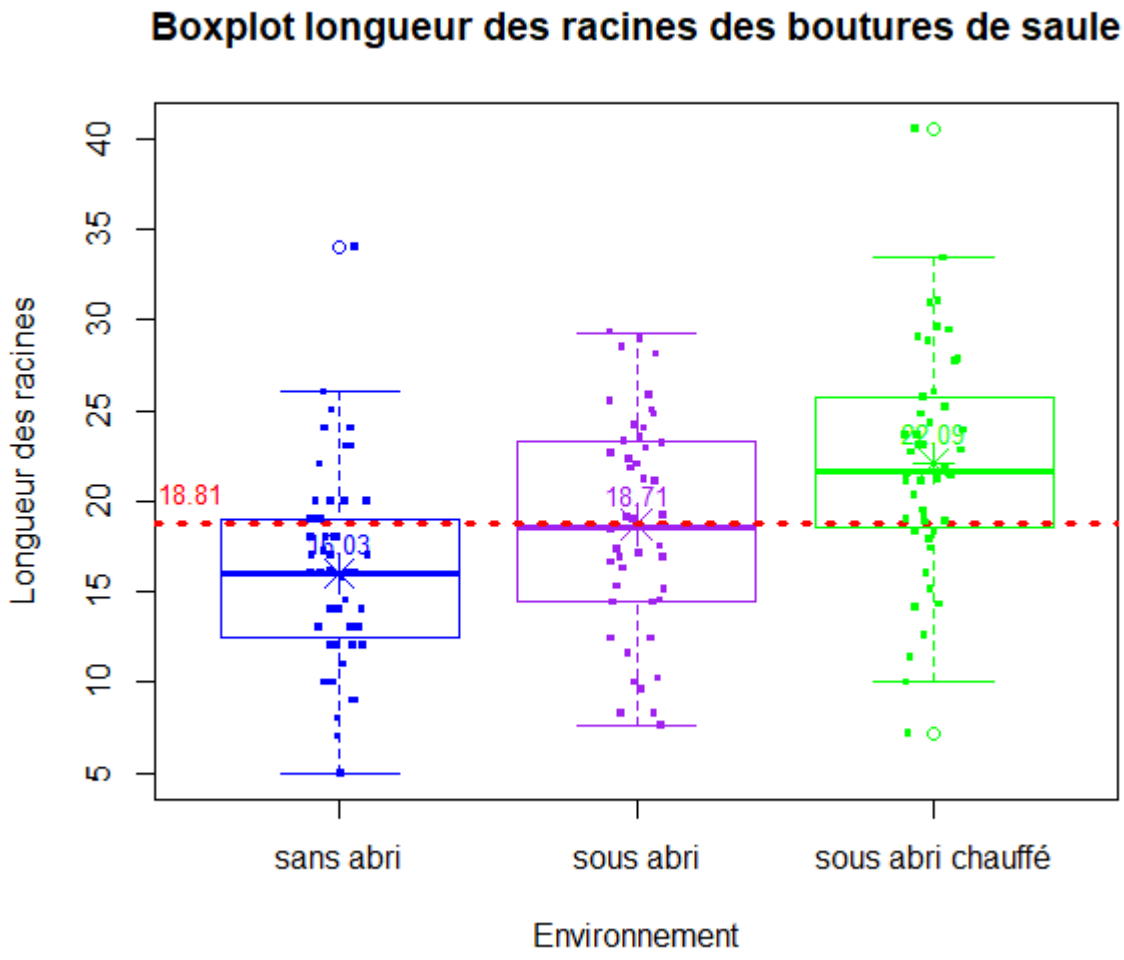
Une expérimentation pour mesurer l'influence de trois environnements sur le développement racinaire de boutures de saule a été menée. 150 boutures de saule ont été plantées en novembre 2017 dans trois environnements différents :

- Environnement 1 : 50 boutures en pleine terre.
- Environnement 2 : 50 boutures en pleine terre sous abri.
- Environnement 3 : 50 boutures en pleine terre sous abri chauffé.

Ces boutures ont été arrachées en novembre 2018 et leur longueur racinaire a été mesurée. Les résultats sont donnés en cm. On cherche à mesurer l'influence du facteur environnement sur la moyenne des longueurs des racines.
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.

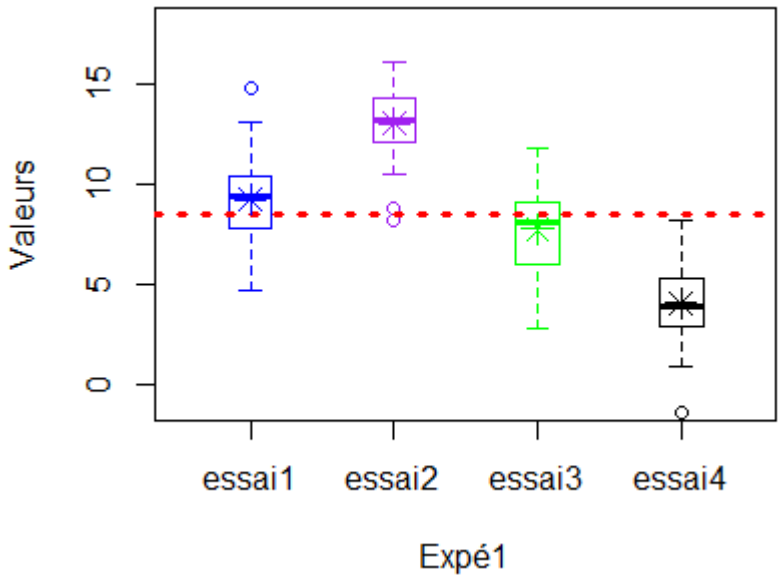


Même si ceci alourdit la représentation, on peut ajouter les points représentant les mesures avec un effet aléatoire sur la largeur pour les rendre lisibles.

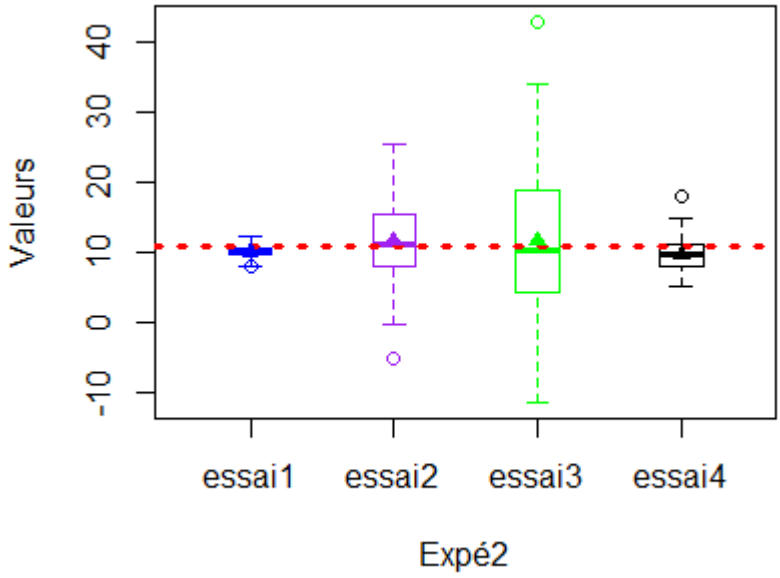


Ce travail n'est pas à négliger car il permet bien souvent d'émettre une conjecture à partir des représentations qui sera et qui devra être confirmée par un test. L'enseignement doit contribuer à développer le réflexe de commencer par des représentations. Les simulations permettent d'illustrer les situations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

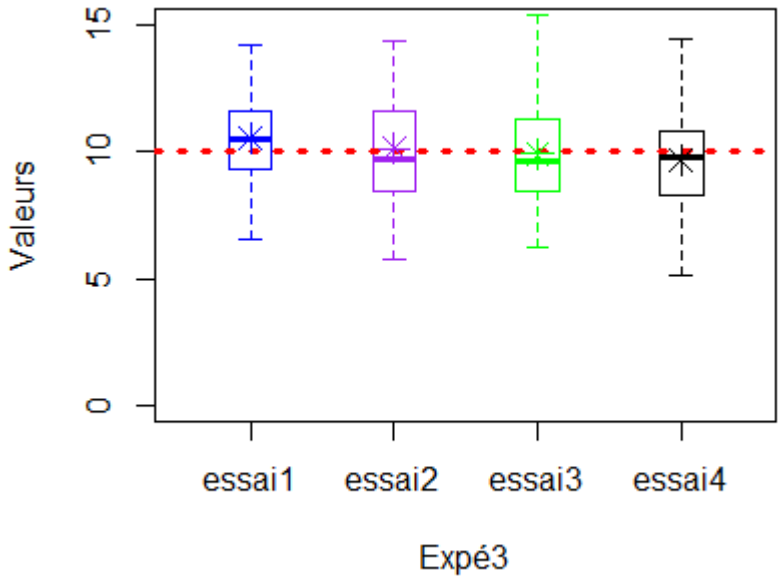
Ce premier exemple ci-dessous (Expé1) est obtenu en simulant des lois normales de même écart type mais de moyennes différentes. Il doit engager une discussion sur la variabilité des moyennes.



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. 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.

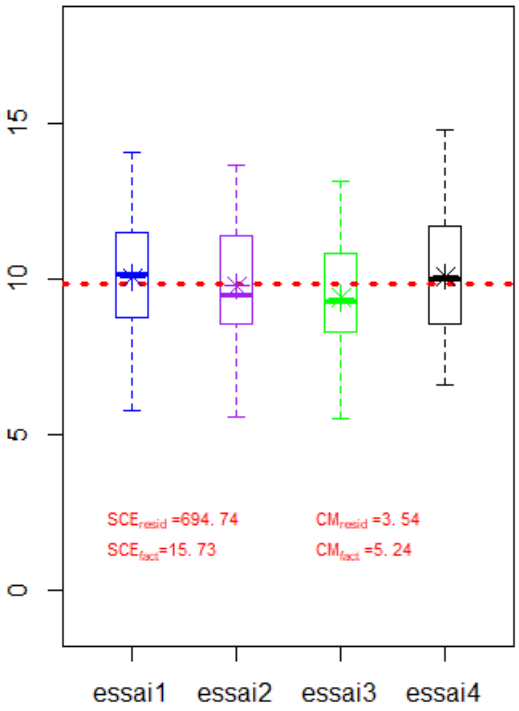


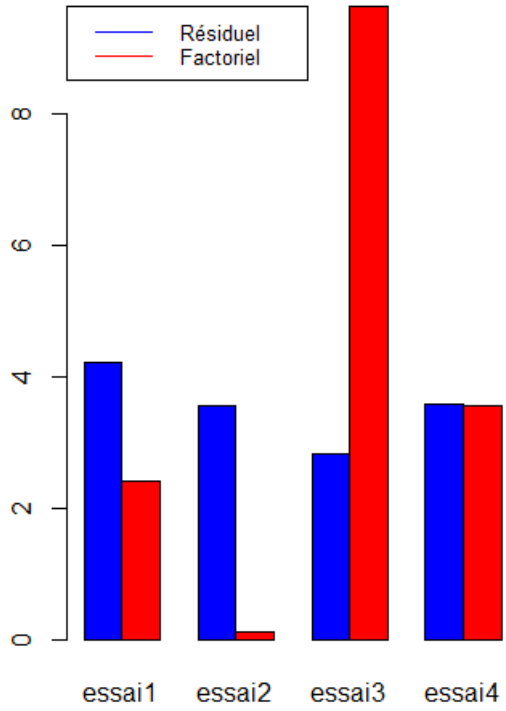
Ce troisième exemple (Expé3) est obtenu avec la même loi normale. La variété des simulations permet de développer le regard et l'interprétation de ces graphiques.

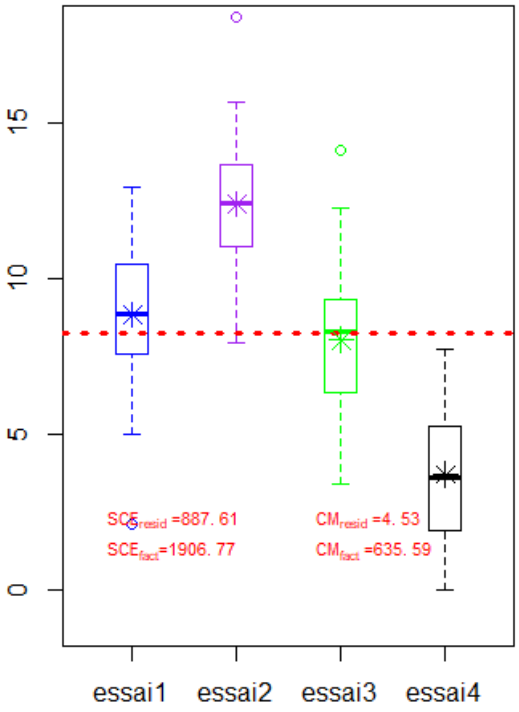


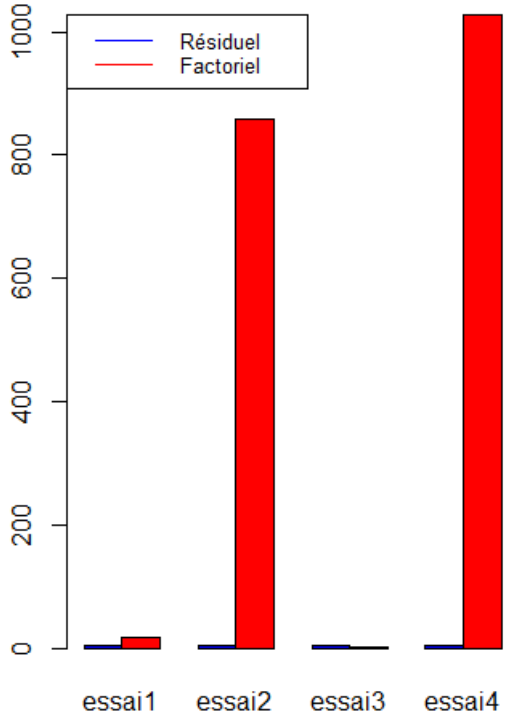
Cette approche par la simulation et les représentations doivent amener à la question de la variabilité des échantillons et des moyennes des échantillons, et faire émerger l'équation de l'analyse de variance souvent présentée sous la forme $SCE_{totale} = SCE_{factoriel} + SCE_{residuel}$ ($SCE = $ *somme 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 permet 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.



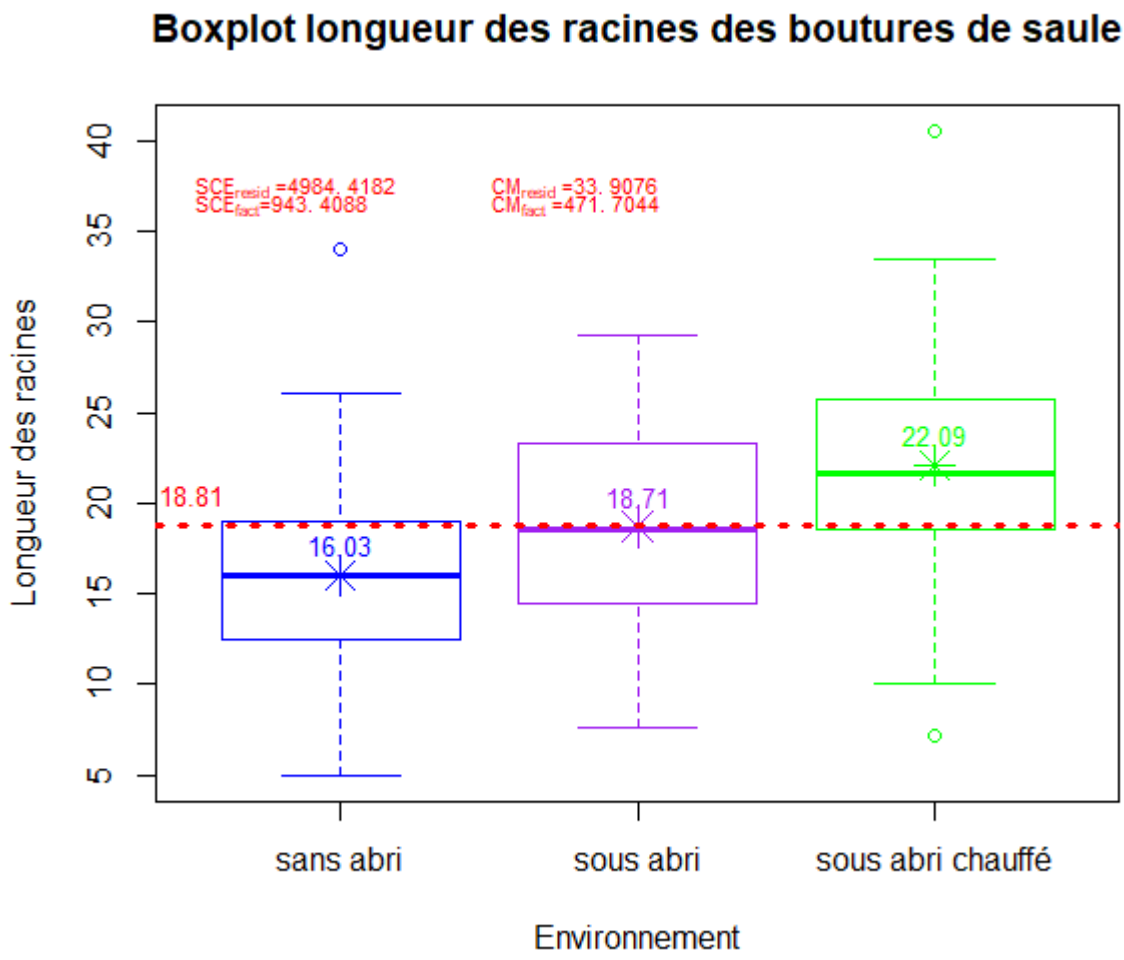






Les simulations et le calcul des CM_{fact} et CM_{resid} permettent de mettre en évidence qu'à variance constante dans les échantillons gaussiens simulés (ici l'écart type est égal à 2), CM_{resid} est plutôt stable alors que CM_{fact} varie fortement avec la variabilité des moyennes des échantillons. 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'expérimentation sur les boutures de saule, on obtient :



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

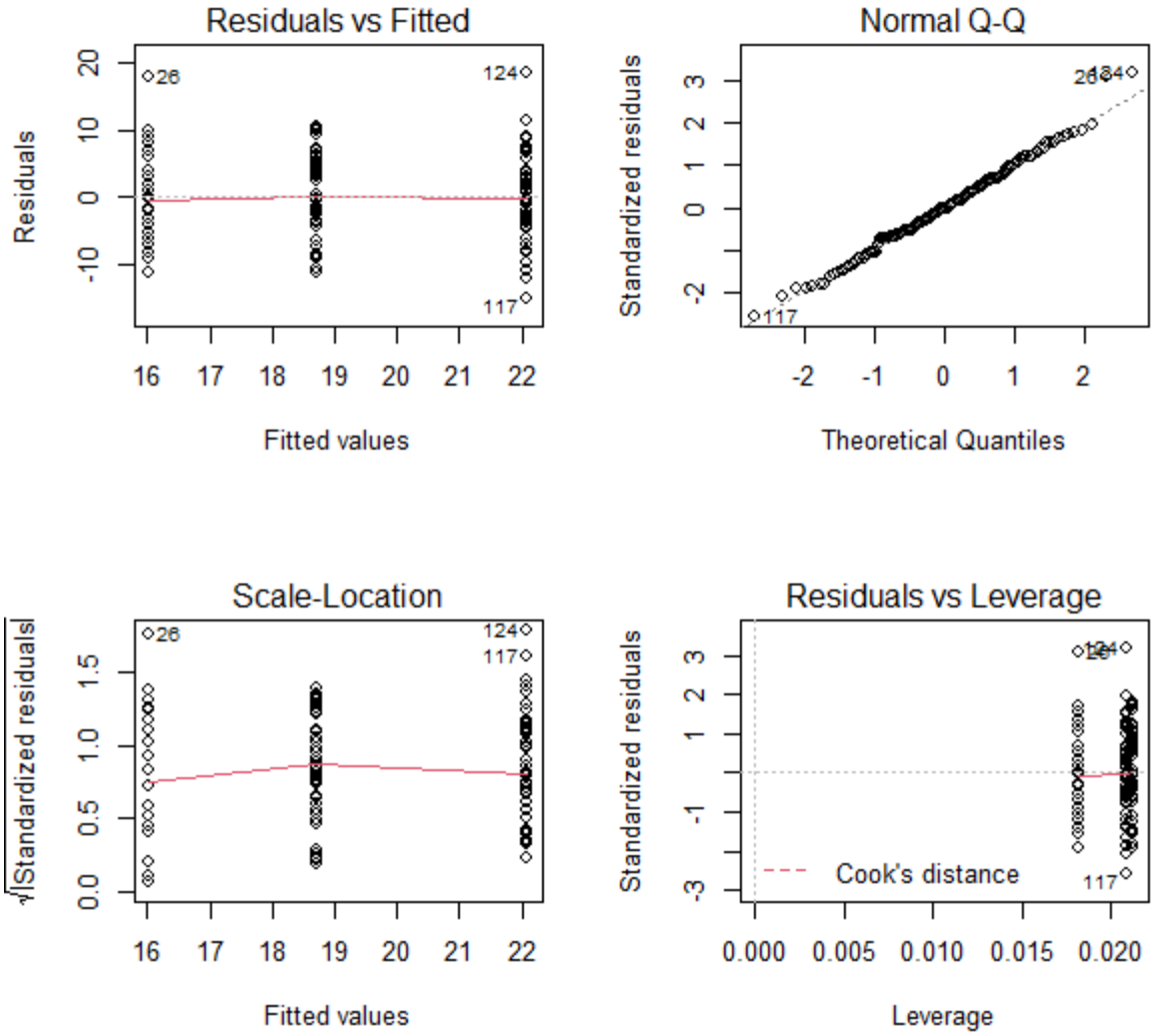
	Ddl	SC	CM	F value	P(>F)
Inter-groupes	2	943	471,7	13,91	2.93 e-06
Intra-groupes	147	4984	33,9		

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 sont discutées et vérifiées. Pour la condition de normalité, voir l'exemple 1 « 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.

Pour l'expérimentation sur les boutures on obtient :



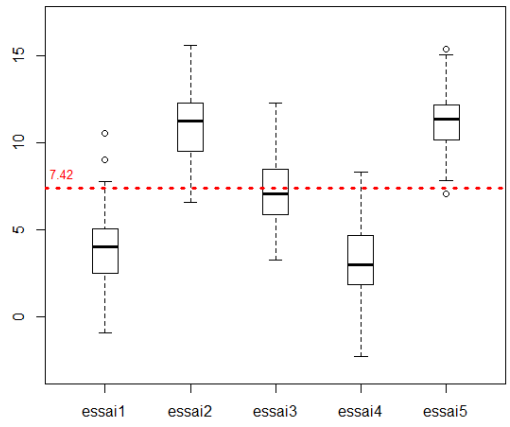
Le 1er graphique en haut à gauche présente les résidus en fonction de la moyenne. On retrouve en abscisse les moyennes des échantillons 16.03, 18.71 et 22.09. Ceci nous ramène à la discussion sur la dispersion. Les points notés 26, 117 et 124 matérialisent des valeurs considérées aberrantes par l'algorithme 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 1). 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 simulation 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.

À 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.

Étudions 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
```

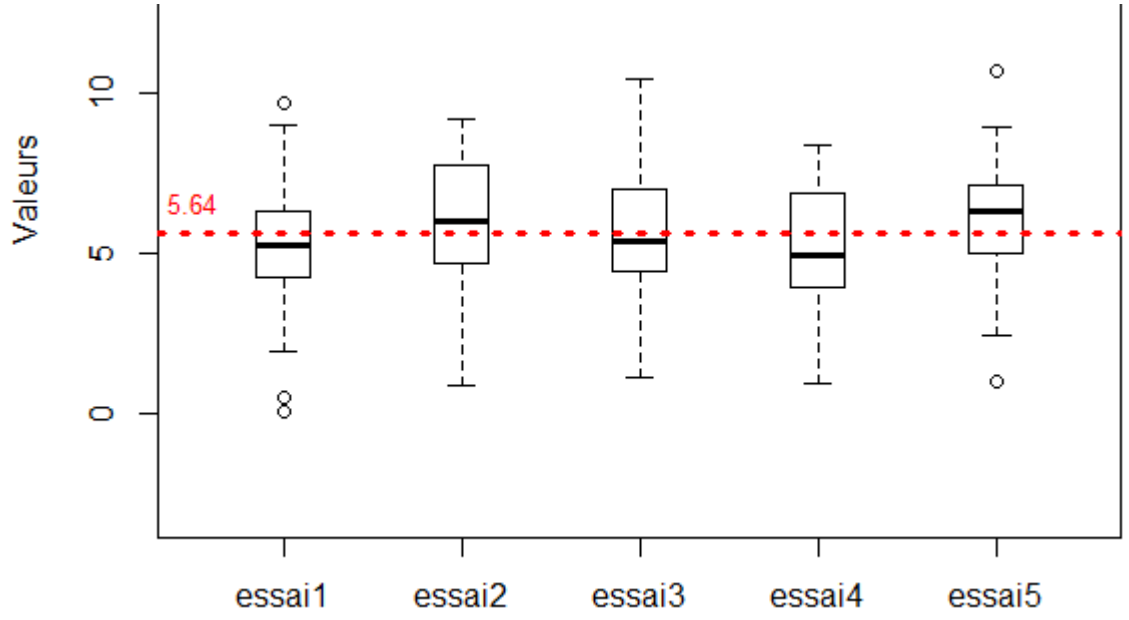
La comparaison de moyennes de Tukey fournit le tableau ci-contre . lwr et upr sont les bornes d'un intervalle de confiance à 95%. Si 0 n'est pas dedans, on rejette H_0 On retrouve la conclusion précédente avec des seuils de significativité différents du test de Bonferroni.

```
  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       Max
essai1           3.902549 2.296074 50 -0.9022114 10.528178
essai2          11.154458 1.989246 50  6.5916313 15.598150
essai3           7.225627 2.222621 50  3.2717597 12.274430
essai4           3.390425 2.059994 50 -2.2406975  8.349426
essai5          11.415742 1.911029 50  7.0664491 15.372037

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 different.

       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 tests Bonferroni, Turkey et SNK n'ont pas la même significativité. L'étudiant pourrait penser que de 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 SD

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 significantly

       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
```

Annexe 1

Exemple 1

Les deux premiers graphiques ont été produits avec le code suivant :

```
normale1=rnorm(n=100, mean=10, sd=1)
uniforme=runif(n=80,min=0,max=1)

par(mfrow=c(2,2))
hist(normale1,prob=TRUE,main='exemple 1',xlab='',breaks=20)
hist(uniforme,prob=TRUE,main='exemple 2',xlab='',breaks=20)
#
qqnorm(normale1)
qqline(normale1)
#
qqnorm(uniforme)
qqline(uniforme)
```

Exemple 2

Le graphique a été produit avec le code suivant :

```
graphics.off()
donnees=c()
ecart=c()

for (i in 1:10000){
valeurs=rnorm(n=9, mean=13.5, sd=0.5)
donnees[i]=mean(valeurs)
ecart[i]=sd(valeurs)
}
mean(donnees)
mean(ecart)
#par(mfrow=c(1,2))


hist(donnees,prob=TRUE,breaks=101,main='',xlab='')
title(main=paste('Histogramme des moyennes','\n','10000 échantillons de taille 9'),
      cex.main=1,xlab='Moyenne')
lines(density(donnees),col="blue",lwd=2)
```

Exemple 3

Le premier graphique a été produit avec le code suivant :

```
moyenne=c()
s=c()

for (i in 1:1000){
  a=rnorm(n=12,mean=6.2,sd=0.123)
  moyenne[i]=mean(a)
  s[i]=sd(a)
}
t=(moyenne-6.2)/(s/12)#s est déjà corrigé
par(mfrow=c(1,2))

hist(s^2,prob=T,main='Distribution de la variance',xlab='')
qqnorm(s^2)
qqline(s^2)

shapiro.test(moyenne)
a=shapiro.test(s^2)
W=round(a$statistic[['W']],2)
pval=format(a$p.value,digits=2)
mtext(text=paste('Shapiro-test','W=',W,'p=',pval))
```

Les graphiques contenant les « boxplot » ont été produits avec le code suivant :

```
library(latex2exp)#Ecrire en latex dans les titres des graphiques
moyenne=c()
s=c()
sigma=0.123

for (i in 1:1000){
  a=rnorm(n=12,mean=6.2,sd=sigma)
  moyenne[i]=mean(a)
  s[i]=sd(a)
}
t=(moyenne-6.2)/(s/12)#s est déjà corrigé

par(mfrow=c(2,2))
hist(t,breaks=20,
       main='',
       xlab=TeX(sprintf(r'($\\sigma = %.2f$)', sigma[i])),main.cex=0.5)
  title(main=TeX(r'(Distribution de $\\frac{\\bar{X}-6,2}{S}\\sqrt{11}$)'))

boxplot(t, col="red",xlab=TeX(sprintf(r'($\\sigma = %.2f$)', sigma[i])))
```

Exemple 4

Les lignes suivantes permettent d'obtenir la légende du graphique avec les différentes lois du χ^2 :

```
graphics.off()
library(latex2exp)#Ecrire en latex dans les titres de graphiques

x=seq(0,30,0.1)
hist(d,breaks=100,prob=T,main=TeX(r'(Histogramme de la variable $d^2$)'),xlab='',ylim=c(0,0.6))
lines(density(d),col='red',lwd=2)
for (i in 2:5){
  df=i
  lines(x,dchisq(x,df),col=i,lwd=2)
  legend(x=7,y=0.2+0.04*i,col=i,text.col=i,legend=(TeX(sprintf(r'($\chi_{%i}^{2}$)',df))),
         cex=1,bty='n',lwd=2)
}
```

Exemple 5

Les deux lignes suivantes permettent d'importer des données d'un fichier au format CSV

```
setwd("Chemin du dossier de travail")
expSaule=read.csv2(file="données.csv")# le fichier de données est au format CSV
```

Les deux premiers graphiques ont été composés avec les lignes suivantes :

```
boxplot(expSaule$long~expSaule$env,data=expSaule,xlab='Environnement',ylab='Longueur des racines',col='white',
        border=c('blue','purple','green'),names=c('sans abri','sous abri','sous abri chauffé'))
title(main='Boxplot longueur des racines des boutures de saule')

moyExp=by(expSaule$long,expSaule$env,mean)
for (i in 1:3){
  points(i,moyExp[[i]],lwd=1,pch=8,col=c('blue','purple','green','black')[i],cex=2)
  text(i,moyExp[[i]],round(moyExp[[i]],digits=2),pos=3,col=c('blue','purple','green')[i],cex=0.8)
}
moy=mean(expSaule$long)
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(expSaule$long~expSaule$env,vertical=T,method='jitter',add=T, pch=15,cex=0.5,
           col=c('blue','purple','green'))
```

Les graphiques diagnostiques sont produits avec les commandes suivantes :

```
bouture.aov=aov(expSaule$long~expSaule$env)# ANOVA
par(mfrow=c(2,2))# Grille des graphiques 2x2
plot(bouture.aov)# Affichage des 4 graphiques diagnostiques
```

Les tests dits post-hoc sont réalisés de la manière suivante :

```
# D'abord récupérer le pakage 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(bouture.aov,ordered=T)
pairwise.t.test(expSaule$long,expSaule$env, p.adjust='bonferroni',alternative='two.sided')
SNK.test(bouture.aov,'expSaule$env', console=T)
```

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.

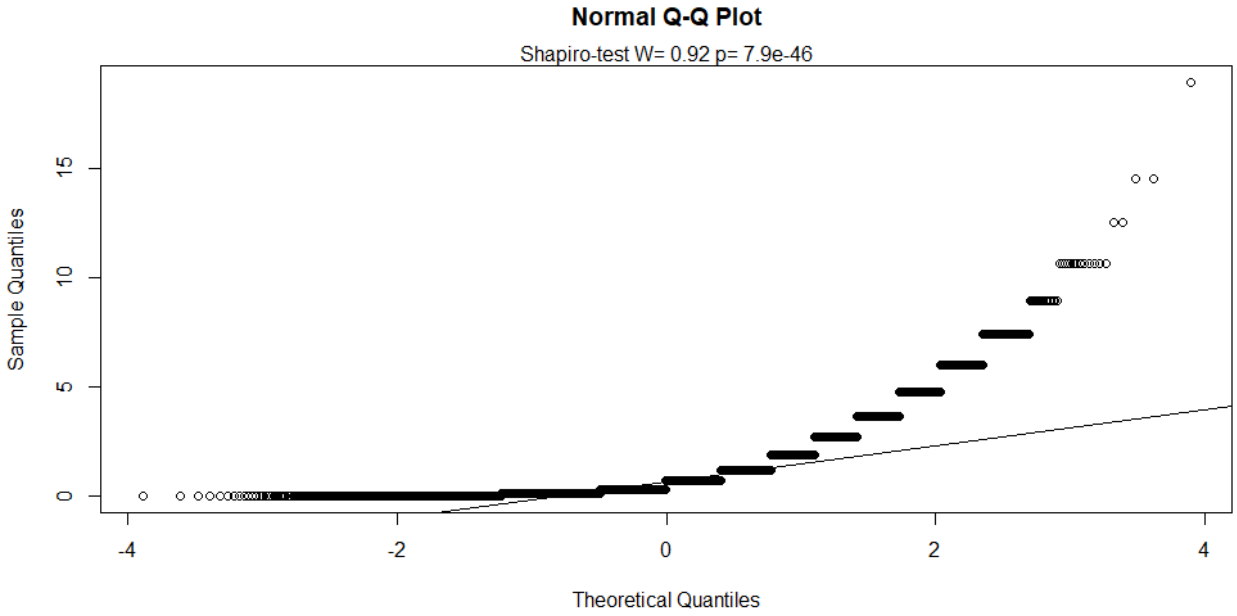
Pour comprendre la transmission d'un caractère d'une génération à l'autre, Mendel féconde artificiellement deux variétés de pois de lignée pure. L'un avec le caractère « graines lisses », l'autre avec le caractère « graines ridées ». La descendance obtenue (F1) ne possède que des graines lisses. Il poursuit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

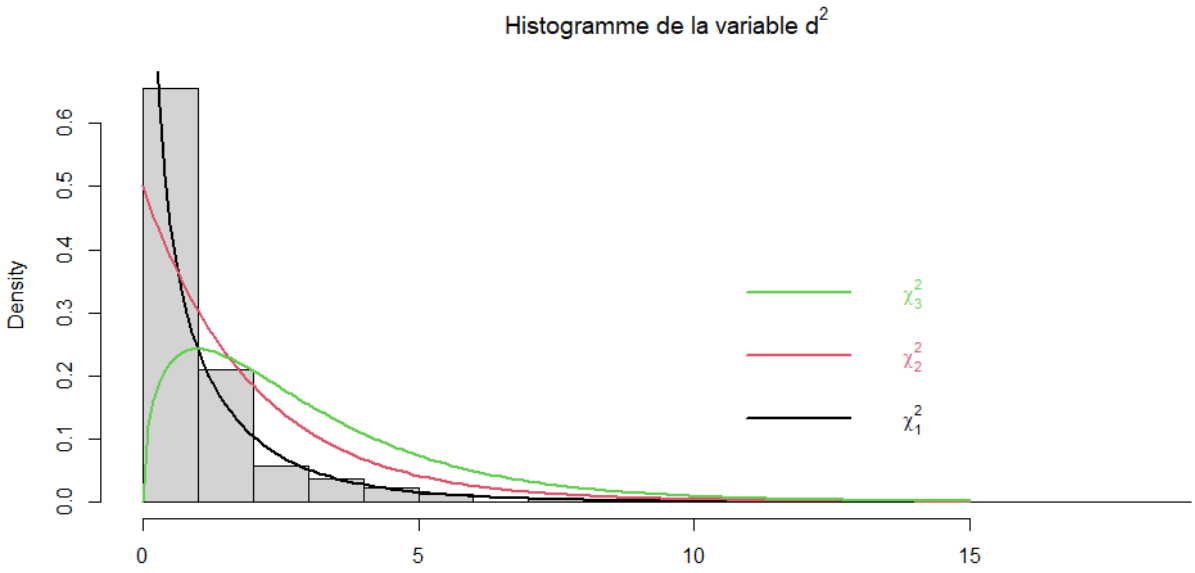
Ces résultats expérimentaux confirment-t-ils l'hypothèse de Mendel qui prévoit une répartition de 25% et 75% ?

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 peut 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.

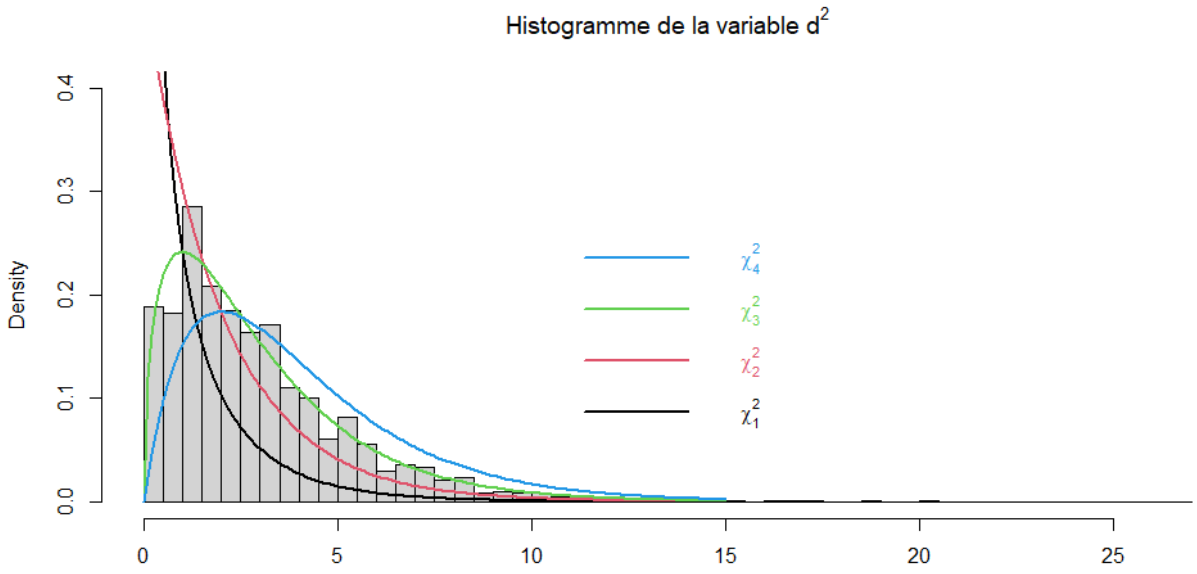


On peut alors s'intéresser à l'expérience de Mendel avec deux caractères exprimés par des gènes comportant deux allèles (l'un dominant A, B et l'autre récessif a, b) sur des chromosomes différents.
On obtient pour la génération (F2) le tableau suivant :

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 :


Histogramme de la variable d^2
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
χ^2_4
χ^2_3
χ^2_2
χ^2_1