Test de comparaisons de moyennes

On a souvent besoin de comparer les moyennes de deux groupes d'observations représentant des populations différentes pour savoir si les populations diffèrent par leurs positions. Dans ces situations, l'hypothèse nulle sera 'il n'y a pas de différence entre les moyennes des deux populations ", soit en symboles, $H_0: \mu_1 = \mu_2$. L'hypothèse alternative est $H_1: \mu_1 \neq \mu_2$ *c.à.d.*, $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$.

1. Echantillons indépendants

Pour vérifier l'hypothèse qui précède, on prélève au hasard des échantillons de chaque population, puis on calcule la moyenne et l'écart-type de chaque échantillon. Notons $\bar{x}_1$ la moyenne et s_1 l'écart-type d'un échantillon de taille n_1 de la première population, $\bar{x}_2$ et s_2 la moyenne et l'écart-type d'un échantillon de taille n_2 de la seconde population. Dans ce contexte, on peut utiliser le critère de test suivant,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

où $\bar{x}_1 = \frac{\sum x_{1i}}{n_1}$, $\bar{x}_2 = \frac{\sum x_{2i}}{n_2}$

s^2 est la variance groupée donnée par

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_1^2 = \frac{\sum x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n_1}}{n_1 - 1} \quad \text{et} \quad s_2^2 = \frac{\sum x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n_2}}{n_2 - 1}$$

Le critère de test t suit une loi de Student avec $n_1 + n_2 - 2$ degrés de liberté. Dans ce cas particulier, le degré de liberté est un paramètre associé à la distribution de t qui gouverne la forme de la distribution. Le concept de degré de liberté est mathématiquement assez obscur, mais d'une manière générale, il peut être considéré comme le nombre d'observations indépendantes dans un ensemble de données, ou comme le nombre de comparaisons indépendantes pouvant être faites à propos d'un ensemble de paramètres.

Ce test repose sur des hypothèses précises, à savoir: i) Les variables entrant en jeu sont continues (ii) La population-mère des échantillons prélevés suit une loi de distribution normale (iii) Les échantillons sont prélevés de manière indépendante (iv) Les variances des deux populations dans lesquelles on prélève les échantillons sont homogènes (égales). L'homogénéité de deux variances peut être testée à l'aide du test *F*.

La procédure ci-dessus n'est pas applicable si les variances des deux populations ne sont pas égales. Dans ce cas, on adoptera une méthode légèrement différente :

*Etape 1. Calculer la valeur du critère de test *t* à l'aide de la formule suivante

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{\sqrt{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]}}$$

*Etape 2. Comparer la valeur de *t* ainsi obtenue avec la valeur pondérée (*t'*) donnée par la table, au niveau de probabilité voulu. La valeur tabulaire pondérée de *t* se calcule comme suit.

$$t' = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

où $w_1 = \frac{s_1^2}{n_1}$, $w_2 = \frac{s_2^2}{n_2}$,

t_1 et t_2 sont les valeurs tabulaires de *t* données par la loi de Student avec $\left(n_1 - 1\right)$ et $\left(n_2 - 1\right)$ degrés de liberté respectivement, au niveau de probabilité voulu.

2. Echantillons appariés

Lorsqu'on compare les moyennes de deux groupes d'observations, il arrive que les groupes soient appariés, au lieu d'être indépendants. C'est par exemple le cas, lorsque l'on compare l'état d'un ensemble d'individus avant et après un traitement, ou les propriétés de la partie basse et de la partie haute des tiges de bambous etc... Dans de telles situations, deux ensembles d'observations sont extraits d'un seul ensemble d'unités expérimentales. Les observations peuvent aussi être appariées pour d'autres raisons, notamment lorsqu'elles portent sur des paires de boutures de tiges issues de plantes-mères différentes et sur les membres d'une paire soumise à deux traitements différents, dans le but de comparer l'effet des deux traitements sur les boutures. On notera que les observations obtenues à partir de ces paires peuvent être corrélées. Le test statistique utilisé pour comparer des moyennes d'échantillons appariés est généralement appelé *test jumelé t*.

Soient $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, les n observations appariées. Supposons que les observations concernant la variable x proviennent d'une population de moyenne μ_1 et celles qui concernent la variable y d'une population de moyenne μ_2. L'hypothèse à vérifier est $H_0: \mu_1 = \mu_2$. Si on forme les différences $d_i = x_i - y_i$ pour $i = 1, 2, \ldots, n$ (on peut considérer qu'elles appartiennent à une population normale de moyenne zéro et de variance connue), on pourra utiliser le critère de test suivant :

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

où $s_d^2 = \frac{1}{n-1}\left(\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right)$

Le critère de test t suit une loi de Student t avec $n - 1$degrés de liberté. La valeur de t ainsi obtenue est donc comparable à la valeur tabulaire de t correspondant à $n - 1$ degrés de liberté, au niveau de probabilité souhaité.

3. Un seul échantillon

Le test de Student cas d'un seul échantillon est aussi appelé test de conformité, ce test a pour but de vérifier si notre échantillon provient bien d'une population avec la moyenne spécifiée, μ_0, ou s'il y a une différence significative entre la moyenne de l'échantillon et la moyenne présumée de la population.

Exemple: Une usine veut vérifier le bon fonctionnement de ces machines car l'usure des machines peut impliquer une déviation aux normes imposées.

Nous tirons aléatoirement un certain nombre d'éléments de la production, nous calculons la moyenne et nous comparons celle-ci avec la norme imposée. Les hypothèses à tester sont :

- hypothèse nulle : $H_0 : \mu = \mu_0$
- hypothèse alternative : $H_1 : \mu \neq \mu_0$

Ce test repose sur des hypothèses précises, le caractère de l'échantillon étant supposé aléatoire, l'hypothèse de normalité de la variable *x* doit être vérifiée (par exemple) avec le test de Kolmogorov-Smirnov si n < 30.

Soit *x* une variable aléatoire distribuée selon une loi normale, la variable aléatoire définie ci-dessus suit une loi de Student avec *n* - 1 degrés de liberté.

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

où μ_0 est la moyenne de la population spécifiée par H_0, $\bar{x}$ est la moyenne de l'échantillon, s^2 est la variance de l'échantillon et n est la taille de l'échantillon

On compare la valeur calculée de *t* avec la valeur critique appropriée de *t* avec *n* - 1 degrés de liberté. On rejette H_0 si la valeur absolue de *t* est supérieure à cette valeur critique. Les valeurs critiques pour différents degrés de liberté et différents seuils de signification sont données par la table de Student.

Test de comparaison de variances

On a souvent besoin de vérifier si deux échantillons aléatoires indépendants proviennent de populations de même variance. Supposons que le premier échantillon de n_1 observations ait pour variance s_1^2 et que le second échantillon de n_2 observations ait pour variance s_2^2, et que les deux échantillons proviennent de populations distribuées normalement. L'hypothèse nulle à tester est: "les deux échantillons sont indépendants et prélevés au hasard dans des populations normalement distribuées de même variance ", soit en symboles :

$$H_0: \sigma_1^2 = \sigma_2^2$$

où σ_1^2, σ_2^2 sont les variances de deux populations dans lesquelles sont prélevés les deux échantillons. L'hypothèse alternative est la suivante :

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Le critère statistique utilisé pour tester l'hypothèse nulle est

$$F = \frac{s_1^2}{s_2^2}$$

où s_1^2 est le plus grand carré moyen

Dans l'hypothèse nulle, on peut montrer que le critère statistique suit une distribution de F avec $(n_1 - 1, n_2 - 1)$ degrés de liberté. La règle de décision est la suivante: si la valeur calculée du critère statistique est inférieure à la valeur critique de la distribution de F, au seuil de signification voulu, on accepte l'hypothèse nulle, à savoir que les deux échantillons sont prélevés dans des populations de même variance. Dans le cas contraire, l'hypothèse nulle est rejetée.

Test de proportions

Lorsque les observations consistent à classer les individus dans des catégories particulières, comme 'malade' ou 'en bonne santé', 'mort' ou 'vivant' etc…, les données sont généralement résumées en termes de proportions. Il peut alors être intéressant de comparer les proportions de l'incidence d'un caractère dans deux populations. L'hypothèse nulle à formuler dans de telles situations est $H_0: P_1 = P_2$, alors que l'hypothèse alternative est $H_1: P_1 \neq P_2$ (ou $P_1 > P_2$ ou $P_1 < P_2$), où P_1 et P_2 sont des proportions représentant les deux populations. Pour tester cette hypothèse, on prélève deux échantillons indépendants de grande taille, par exemple n_1 et n_2, dans les deux populations. On obtient ainsi deux échantillons de proportions respectives p_1 et p_2. Le critère statistique utilisé est le suivant :

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

où $q_1 = 1 - p_1$, $q_2 = 1 - p_2$. Cette statistique suit une loi de distribution normale standard.

Test de la validité de l'ajustement

Les tests d'hypothèses ont parfois pour but de vérifier si la population dont provient un échantillon suit une loi de distribution de probabilité déterminée. La distribution escomptée peut être basée sur un modèle théorique (loi normale, binomiale ou de Poisson) ou sur un schéma particulier, en raison de facteurs techniques. Il peut par exemple être intéressant de vérifier si une variable comme la hauteur des arbres suit une loi normale de distribution. Un spécialiste de l'amélioration génétique des arbres peut avoir besoin de savoir s'il existe une déviation significative entre les rapports de ségrégation relatifs à un caractère, tels qu'ils sont observés, et ceux de Mendel. Dans de telles situations, on est amené à vérifier la correspondance entre les fréquences observées et théoriques. Ce type de test a reçu le nom de test de la validité de l'ajustement.

Pour appliquer le test de la validité de l'ajustement, on utilise uniquement les fréquences réelles observées, à l'exclusion des pourcentages ou proportions. De plus, il est indispensable que les observations faites sur un même échantillon ne se chevauchent pas et soient indépendantes. Les fréquences attendues dans chaque catégorie doivent de préférence être supérieures à 5. Le nombre total d'observations doit être élevé, en général supérieur à 50.

Dans les tests de la validité de l'ajustement, l'hypothèse nulle est " il n'y a pas de discordance entre la distribution observée et la distribution théorique ", ou " la distribution observée est ajustée à la distribution théorique ". Le critère de test utilisé est le suivant :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

où O_i = fréquence observée dans la $i$ème classe,
E_i = fréquence attendue dans la $i$ème classe.
k = nombre de catégories ou classes.

La statistique χ^2 de l'équation ci-dessus suit une distribution de χ^2 avec k-1 degrés de liberté. Si les fréquences attendues sont dérivées de paramètres estimés dans l'échantillon, les degrés de libertés sont au nombre de (k-p-1) (où p est le nombre de paramètres estimés). Si, par exemple, on veut tester la normalité d'une distribution, une estimation de μ et σ^2 à partir de l'échantillon sera donnée par x et s^2. Les degrés de liberté se réduisent donc à (k-2-1).

Les fréquences escomptées peuvent être calculées d'après la fonction de probabilité de la distribution théorique appropriée à la situation, ou obtenues par dérivation, en prenant pour base la théorie scientifique que l'on compte tester, par exemple la loi de Mendel sur l'hérédité. Dans le cas où il n'existe pas de théorie bien définie, on supposera que toutes les classes se retrouvent avec la même fréquence dans la population. Par exemple, l'hypothèse de départ peut être que le nombre d'insectes pris au piège à différents moments d'une journée, ou le nombre de fois où l'on voit un animal dans différents habitats etc… sont égaux et soumettre ces fréquences au test statistique. Dans ces situations, la fréquence attendue est donnée par la formule :

$$E = \frac{\text{Total des fréquences observées}}{\text{Nombre des groupes}} = \frac{n}{k}$$

Corrélation

Dans beaucoup de systèmes naturels, les changements d'un attribut s'accompagnent de variations d'un autre attribut, et il existe une relation définie entre les deux. En d'autres termes, il existe une corrélation entre les deux variables. Par exemple, plusieurs propriétés des sols, comme la teneur en azote, la teneur en carbone organique ou le pH, sont corrélées et varient de façon concomitante. On a observé une forte corrélation entre plusieurs caractéristiques morphométriques d'un arbre. Dans de telles situations, il peut être intéressant pour un chercheur de mesurer l'importance de cette relation. Si (x_i,y_i); $i = 1, ..., n,$ est un ensemble d'observations appariées effectuées sur n unités d'échantillonnage indépendantes, une mesure de la relation linéaire entre deux variables est donnée par la quantité suivante, appelée coefficient de corrélation linéaire de Pearson, ou simplement coefficient de corrélation.

$$r = \frac{\text{Covariance de x et y}}{\sqrt{(\text{Variance de x})(\text{Variance de y})}} = \frac{\text{Cov(x, y)}}{\sqrt{(\text{V(x)})(\text{V(y)})}}$$

où $Cov\ (x,y) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)$

$$V(x) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

$$V(y) = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)$$

Ce paramètre statistique indique à la fois la direction et le degré de la relation existant entre deux caractères quantitatifs x et y. La valeur de r peut varier de –1 à +1, sans atteindre ces valeurs. Si la valeur de r est nulle, cela signifie qu'il n'y a pas de relation linéaire entre les deux variables concernées (il peut toutefois y avoir une relation non-linéaire). La relation linéaire est forte lorsque la valeur de r approche –1 ou +1. Une valeur négative de r indique que si la valeur d'une variable augmente, celle de l'autre diminue. Au contraire, une valeur positive indique une relation directe, c'est à dire que l'augmentation de la valeur d'une variable est associée à une augmentation de la valeur de l'autre. Un changement d'origine, d'échelle, ou d'origine et d'échelle est sans incidence sur le coefficient de corrélation. Lorsque l'on ajoute ou soustrait un terme constant aux valeurs d'une variable, on dit que l'on a changé d'origine, alors que lorsque l'on multiplie ou divise par un terme constant les valeurs d'une variable, on parle de changement d'échelle.

Test de signification du coefficient de corrélation.

La signification d'une valeur du coefficient de corrélation calculée à partir d'un échantillon doit être testée pour confirmer l'existence d'une relation entre les deux variables, dans la population considérée. En général, on définit l'hypothèse nulle comme $H_0 : \rho = 0$ alors que l'hypothèse alternative est $H_1 : \rho \neq 0$.

Pour *n* relativement petit, l'hypothèse nulle ($\rho = 0$) peut être testée à l'aide du critère statistique

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Ce critère statistique suit une distribution de Student *t* avec *n*-2 degrés de liberté.

Régression

Le coefficient de corrélation mesure le degré de la relation entre deux variables qui varient de façon concomitante, avec des effets qui se renforcent mutuellement. Dans certains cas, les changements relatifs à une variable sont provoqués par les variations d'une variable connexe, sans qu'il y ait de dépendance mutuelle. En d'autres termes, une variable est considérée comme dépendante des variations de l'autre variable, dans la mesure où elles dépendent de facteurs externes. Une telle relation entre deux variables est appelée régression. Lorsque ces relations sont exprimées sous forme mathématique, il est possible d'estimer la valeur d'une variable d'après la valeur de l'autre. Par exemple, le rendement de conversion photosynthétique et le coefficient de transpiration des arbres dépendent de conditions atmosphériques comme la température ou l'humidité, sans pour autant que l'on s'attende généralement à une relation inverse. Toutefois certaines variables sont souvent déclarées indépendantes uniquement au sens statistique, même dans des situations où des effets inverses sont concevables. Par exemple, dans une équation servant à estimer le volume, le volume es arbres est souvent considéré comme dépendant du diamètre à hauteur d'homme, même si le diamètre ne saurait être considéré comme indépendant des effets du volume des arbres au sens physique. C'est pourquoi, dans le contexte de la régression, les variables indépendantes sont souvent appelées variables exogènes (explicative), et la variable dépendante variable endogène (expliquée).

La variable dépendante est habituellement notée y et la variable indépendante x. Dans le cas où il n'y a que deux variables en jeu, la relation fonctionnelle est appelée régression simple. Si la relation entre les deux variables est linéaire, on parle de régression linéaire simple ; dans le cas contraire, la régression est dite non-linéaire. Lorsqu'une variable dépend d'au moins 2 variables indépendantes, la relation fonctionnelle entre la variable dépendante et l'ensemble des variables indépendantes est une régression multiple. Dans un souci de simplification, on se limitera ici à examiner le cas d'une régression linéaire simple.

Régression linéaire simple

La régression linéaire simple de y en x dans la population peut s'exprimer comme

$$y = \alpha + \beta x + \varepsilon$$

où α et β sont des paramètres**,** appelés aussi coefficients de régression, et ε est une déviation aléatoire pouvant dériver de la relation attendue. Si la valeur moyenne de ε est zéro, l'équation ci-dessus représente une droite de pente β et d'ordonnée à l'origine α. Autrement dit, α est la valeur présumée de y lorsque x prend la valeur zéro et β représente la variation attendue de y correspondant à une variation unitaire de la variable x. La pente d'une droite de régression linéaire peut être positive, négative ou nulle, selon la relation entre y et x.

En pratique, les valeurs de α et β doivent être estimées à partir d'observations des variables y et x effectuées sur un échantillon. Par exemple, pour estimer les paramètres d'une équation de régression proposée liant la température atmosphérique et le taux de transpiration des arbres, un certain nombre d'observations appariées sur la température et le taux de transpiration sont effectuées sur plusieurs arbres, à différents moments de la journée. Notons (x_i, y_i); $i = 1, 2, \ldots, n$ ces couples de valeurs, n étant le nombre de d'observations appariées indépendantes. Les valeurs de α et β sont estimées par la méthode des moindres carrés de sorte que la somme des carrés des différences entre les valeurs observées et prévues soit minimale. Le processus d'estimation repose sur les hypothèses suivantes: i) Les valeurs de x sont non aléatoires ou fixes ; ii) Pour tout x, la variance de y est la même ; iii) Les valeurs de y observées pour différentes valeurs de x sont complètement indépendantes. Si l'une de ces hypothèses n'est pas vérifiée, il faut apporter les changements voulus. Pour les tests d'hypothèses se référant à des paramètres, une hypothèse additionnelle de normalité des erreurs est nécessaire.

En effet, les valeurs de α et β s'obtiennent grâce à la formule,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

L'équation $\hat{y} = \hat{\alpha} + \hat{\beta}x$ représente la droite de régression ajustée, qui peut être utilisée pour estimer la valeur moyenne de la variable dépendante, y, associée à une valeur particulière de la variable indépendante, x. En général, il est plus sûr de limiter ces estimations à la fourchette des valeurs de x dans les données.

ANOVA 1 et 2

1. ANOVA 1

Il existe deux sources de variation entre les n observations tirées d'un essai en randomisation totale. L'une est la variation due aux traitements et l'autre est l'erreur expérimentale. Leur taille relative indique si la différence observée entre les traitements est réelle ou si elle est due au hasard. La différence due au traitement est " réelle " si elle dépasse dans une mesure significative l'erreur expérimentale.

Nous allons voir ci-dessous les étapes de l'analyse de variance des données provenant d'une expérimentation relative à un dispositif complètement randomisé (DCR) comportant un nombre de répétitions non uniforme. Les formules peuvent être adaptées facilement en cas de répétitions égales, de sorte qu'elles ne sont pas décrites à part.

*Etape 1. Regrouper les données par traitements et calculer les totaux des traitements (T_i) et le total général (G).

*Etape 2. Dresser un Tableau d'analyse de variance, suivant le modèle du Tableau 1.

Tableau 1. Schéma de l'analyse de variance d'un DCR, avec répétitions inégales

Source de variation	Degrés de liberté (df)	Somme des carrés (SS)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	Valeur calculée de F
Traitement	t - 1	SST	MST	$\frac{MST}{MSE}$
Erreur	n - t	SSE	MSE	
Total	n - 1	$SSTO$		

*Etape 3. Avec les totaux des traitements (T_i) et le total général (G), calculer comme suit le facteur de correction et les différentes sommes des carrés.

$$C.\ F. = \frac{G^2}{n}$$

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r_i} y_{ij}^2 - C.\ F.$$

$$SST = \sum_{i=1}^{t} \frac{T_i^2}{r_i} - C.\ F.$$

$$SSE = SSTO - SST$$

*Etape 4. Entrer toutes les valeurs des sommes des carrés dans le tableau d'analyse de la variance et calculer les carrés moyens et la valeur de *F*.

*Etape 5. Prendre les valeurs tabulaires de *F*, avec f_1 et f_2 degrés de liberté, où f_1 = df du traitement = $(t - 1)$ et $f_2 = df$ de l'erreur = $(n - t)$, respectivement.

*Etape 6. Comparer la valeur calculée de *F* de l'Etape 4 avec la valeur tabulée de *F* de l'Etape 5, et déterminez si la différence entre les traitements est significative, d'après les règles ci-après :

i) Si la valeur calculée de *F* est supérieure à sa valeur tabulaire au seuil de signification de 5%, la variation due aux traitements est dite *significative,* ce qui est généralement indiqué par un astérisque au-dessus de la valeur calculée de F, dans l'analyse de variance.

ii) Si la valeur calculée de *F* est inférieure ou égale à la valeur tabulaire de *F* au seuil de signification de 5%, la variation due aux traitements est dite non significative, ce qui est indiqué par la mention ns au-dessus de la valeur calculée de *F* (ou par l'absence d'indication au-dessus de cette valeur).

Une valeur non significative de *F* dans l'analyse de variance indique que l'expérience n'a pas réussi à détecter de différence entre les traitements. Elle ne prouve en aucun cas que tous les traitements sont les mêmes car la non détection d'une différence entre les traitements, attestée par une valeur non significative du critère *F*, pourrait s'expliquer par une différence nulle ou minime, ou par une erreur expérimentale importante, ou encore par ces deux facteurs. Ainsi, dans tous les cas où la valeur de *F* n'est pas significative, le chercheur devrait examiner l'ampleur de l'erreur expérimentale et les différences numériques entre les moyennes des traitements. Si ces deux valeurs sont grandes, il est conseillé de refaire l'essai et de tenter de réduire l'erreur expérimentale pour que les éventuelles différences entre les traitements puissent être détectées. En revanche, si les deux valeurs sont petites, les différences entre les traitements sont probablement trop faibles pour avoir une signification économique, si bien qu'il n'est pas nécessaire de faire de nouveaux essais.

On notera qu'une valeur significative de *F* confirme l'existence de quelques différences entre les traitements testés, mais ne précise pas pour quelle(s) paire(s) de traitements spécifiques la différence est significative. Ces informations s'obtiennent grâce aux procédures de comparaison des moyennes.

*Etape 7. Calculer comme suit la moyenne générale et le coefficient de variation (*cv*):

$$\text{Moyenne générale} = \frac{G}{n}$$

$$cv = \frac{\sqrt{MSE}}{\text{Moyenne générale}}(100)$$

Le *cv* affecte le degré de précision des comparaisons entre les traitements et donne une bonne indication de la fiabilité de l'expérience. C'est une expression de l'erreur expérimentale totale, en pourcentage de la moyenne totale ; Ainsi, plus la valeur de *cv* est grande, moins l'expérience est fiable. Le *cv* varie considérablement suivant le type d'expérience, la plante cultivée, et les caractères mesurés. Toutefois, un chercheur expérimenté peut relativement bien juger de l'acceptabilité d'une valeur spécifique du *cv* pour un type d'expérience donné. Les résultats d'expériences donnant un *cv* supérieur à 30% sont sujets à caution.

**** Comparaison des traitements***

Dans le domaine de la recherche végétale, l'une des procédures les plus couramment employées, pour les comparaisons appariées est le test de Newman-Keuls (Δ_{NK}). D'autres méthodes, comme le test de la plus petite différence significative (PPDS), le test de Duncan et le test de la différence raisonnablement significative sont employées. Le test de Newman-Keuls fournit une valeur unique qui, à un niveau de signification déterminé, marque la limite entre la différence significative et non significative entre une paire de moyennes de traitements quelconque.

Deux traitements présentent donc des différences significatives à un seuil de signification prescrit si leur différence est supérieure à la valeur calculée de Δ_{NK}. Dans le cas contraire, leurs différences sont considérées comme non significatives.

Ce calcul nécessite l'utilisation de la table NK à 3 entrées comportant:

1) risque globale de première espèce α
2) le nombre de degrés de liberté (k) avec lesquels est estimée la variance de population
3) le nombre de moyennes à comparer (i)

La table fournit alors la valeur $q_{1}^{\ i,k}{}_{-\alpha}$

Chaque amplitude est alors comparée à:

$$q_{1-\alpha}^{i,k}\sqrt{\frac{\hat{\sigma}^2}{n}}$$

2. ANOVA 2

Dans toute expérience, une ou plusieurs variables de réponse peuvent être affectées par un certain nombre de facteurs dans le système global, dont certains sont maîtrisés ou maintenus aux niveaux voulus dans l'expérience. Une expérience dans laquelle les traitements sont constitués de toutes les combinaisons possibles de deux ou plusieurs facteurs, aux niveaux sélectionnés, est appelé plan d'expérience factoriel. Par exemple, une expérience sur l'enracinement des boutures englobant deux facteurs, mesurés à deux niveaux – par exemple deux hormones à deux dosages différents – est une expérience factorielle 2 x 2 ou 2^2. Les traitements sont constitués des quatre combinaisons possibles de chacun des deux facteurs, aux deux niveaux considérés.

	Combinaison des traitements	
Numéro du traitement	Hormone	Dose (ppm)
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

On utilise parfois l'expression *expérience factorielle complète* lorsque les traitements comprennent toutes les combinaisons des niveaux sélectionnés des facteurs, mais l'expression *expérience factorielle fractionnée* ne s'applique que le test ne porte que sur une fraction de toutes les combinaisons. Toutefois, pour simplifier, les expériences factorielles complètes seront, tout au long de ce manuel, appelées simplement expériences factorielles. On notera que le terme *factoriel* se réfère au mode de constitution spécifique des traitements et n'a rien à voir avec le plan décrivant le dispositif expérimental. Par exemple, si l'expérience factorielle 2^2 dont nous avons parlé plus haut fait partie d'un plan d'expérience en blocs aléatoires complets, l'expérience devrait être définie par l'expression *expérience factorielle* 2^2 *dans un plan en blocs aléatoires complets.*

Dans un plan d'expérience factoriel, le nombre total de traitements est égal au produit du nombre de niveaux de chaque facteur; dans l'exemple factoriel 2^2 , le nombre de traitements est égal à 2 x 2 = 4, dans une expérience factorielle 2^3, le nombre de traitements est 2 x 2 x 2 = 8. Le nombre de traitements augmente rapidement avec le nombre de facteurs ou avec les niveaux de chaque facteur. Pour une expérience factorielle comprenant 5 clones, 4 espacements et 3 méthodes de désherbage, le nombre total de traitements sera 5 x 4 x 3 = 60. On évitera donc le recours inconsidéré aux expériences factorielles en raison de leur ampleur, de leur complexité et de leur coût. De plus, il est peu raisonnable de se lancer dans une expérience de grande ampleur au début d'un travail de recherche, alors qu'il est possible, avec plusieurs petits essais préliminaires, d'obtenir des résultats prometteurs. Imaginons par exemple qu'un généticien ait fait venir 30 nouveaux clones d'un pays voisin et veuille voir comment ils réagissent à l'environnement local. Etant donné que normalement les conditions de l'environnement varient en fonction de plusieurs facteurs, tels que la fertilité du sol, le degré d'humidité, etc. l'idéal serait de tester les 30 clones dans le cadre d'une expérience factorielle englobant d'autres variables, telles que engrais, niveau d'humidité et densité de population.

Le problème est que l'expérience devient alors extrêmement vaste du fait de l'adjonction d'autres facteurs que les clones. Même si l'on incluait qu'un seul facteur, comme l'azote ou l'engrais, à trois dosages différents, le nombre de traitements passerait de 30 à 90. Une expérience de cette ampleur pose divers types de problèmes, notamment pour obtenir des financements ou une surface expérimentale adéquate, ou pour contrôler l'hétérogénéité du sol etc. Pour faciliter les choses, il est donc préférable de commencer par tester les 30 clones dans une expérience à un facteur, puis de sélectionner sur la base des résultats obtenus un petit nombre de clones à soumettre à un examen plus détaillé. Par exemple la première expérience à un facteur peut montrer que seuls cinq clones ont des performances suffisamment remarquables pour justifier des tests plus approfondis. Ces cinq clones pourraient ensuite être insérés dans une expérience factorielle avec trois dosages d'azote, ce qui donnerait un expérience à quinze traitements , alors qu'il en faudrait 90 dans une expérience factorielle avec 30 clones.

Dans la majorité des plans d'expérience factoriels, les traitements sont trop nombreux pour qu'un plan en blocs aléatoires puisse être efficace. Certains types de plans ont cependant été spécifiquement mis au point pour des expériences factorielles de grande envergure.

**** Analyse de variance***

Tout plan en blocs complets pour des expériences à un facteur est applicable à un plan d'expérience factoriel. Les procédures de randomisation et de représentation schématique de chaque plan peuvent être appliquées directement, en ignorant simplement la composition factorielle des traitements et en faisant comme s'il n'existait pas de relation entre les traitements. Pour l'analyse de variance, les calculs examinés pour chaque plan sont aussi directement applicables. Toutefois, des étapes de calcul doivent être ajoutées pour répartir les sommes des carrés des traitements entre les composantes factorielles correspondant aux effets principaux des facteurs individuels et à leurs interactions.

Nous allons décrire les différentes étapes de la procédure d'analyse de la variance d'une expérience à deux facteurs, avec deux niveaux (Facteur A) et trois niveaux (facteur B), à trois répétitions. La liste des six combinaisons factorielles des traitements figure dans le Tableau 2, le dispositif expérimental est illustré à la Figure 1.

Tableau 2. Les combinaisons factorielles (2 x 3) des traitements, avec deux niveaux (Facteur A) et trois niveaux (Facteur B).

Facteur B	Facteur A	
	(a_1)	(a_2)
(b_1)	a_1b_1	a_2b_1
(b_2)	a_1b_2	a_2b_2
(b_3)	a_1b_3	a_2b_3

Figure 1. Schéma-type d'un plan d'expérience factoriel 2×3, avec deux niveaux (facteur A) et trois niveaux (facteur B) et 3 répétitions.

a_2b_3	a_2b_3	a_1b_2
a_1b_3	a_1b_2	a_1b_1
a_1b_2	a_1b_3	a_2b_2
a_2b_1	a_2b_1	a_1b_3
a_1b_1	a_2b_2	a_2b_1
a_2b_2	a_1b_1	a_2b_3

*Etape 1. Soit *r* le nombre de répétitions, *a* le nombre de niveaux du facteur A, et *b* le nombre de niveaux du facteur B. Dresser le tableau de l'analyse de variance:

Tableau 3. Représentation schématique de l'analyse de variance d'une expérience factorielle avec deux niveaux du facteur A, trois niveauxs du facteur B et trois répétitions.

Source de variation	Degrés de liberté (*df*)	Somme des carrés (*SS*)	Carré moyen $\left(MS = \frac{SS}{df}\right)$	*F* calculé
Répétition	r-1	*SSR*	MSR	
Traitement	ab- 1	*SST*	MST	$\frac{MST}{MSE}$
A	a- 1	*SSA*	MSA	$\frac{MSA}{MSE}$
B	b- 1	*SSB*	MSB	$\frac{MSB}{MSE}$
AB	(a-1)(b-1)	*SSAB*	MSAB	$\frac{MSAB}{MSE}$
Erreur	(r-1)(ab-1)	*SSE*	MSE	
Total	rab -1	*SSTO*		

*Etape 2. Calculer les totaux des traitements (T_{ij}), les totaux des répétitions (R_k), le total général (G) et calculer *SSTO*, *SSR*, *SST* et *SSE*. Notons y_{ijk} l'observation correspondant au $i^{ème}$ niveau du facteur A et au $j^{ème}$ niveau du facteur B dans la $k^{ième}$ répétition.

$$C.F. = \frac{G^2}{rab}$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}^2 - C.F.$$

$$SSR = \frac{\sum_{k=1}^{r} R_k^2}{ab} - C.F.$$

$$SST = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij}^2}{r} - C.F.$$

$$SSE = SSTO - SSR - SST$$

*Etape 3. Construire le tableau à double entrée des totaux facteur A x facteur B, avec le calcul des totaux du facteur A et les totaux du facteur B.

*Etape 4. Calculer les trois composantes factorielles de la somme des carrés des traitements:

$$SSA = \frac{\sum_{i=1}^{b} A_i^2}{rb} - C.F.$$

$$SSB = \frac{\sum_{j=1}^{b} B_j^2}{ra} - C.F.$$

$$SSAB = SST - SSA - SSB$$

*Etape 5. Calculer le carré moyen de chaque source de variation en divisant chaque somme des carrés par les degrés de liberté qui lui sont associés et obtenir les valeur du rapport *F* pour les trois composantes factorielles, selon le schéma du Tableau 4.

*Etape 6. Entrer toutes les valeurs obtenues durant les Etapes 3 à 5, dans l'analyse de variance de l'Etape 2 en suivant les indications du Tableau 4.

*Etape 7. Comparer chaque valeur calculée de *F* avec la valeur tabulaire de *F*,
avec $f_1 = df$ du *MS* du numérateur et $f_2 = df$ du *MS* du dénominateur, au seuil de signification voulu.

*Etape 8. Calculer le coefficient de variation:

$$\text{cv} = \frac{\sqrt{\text{Erreur MS}}}{\text{Moyenne générale}} \times 100$$

**** Comparaison des traitements***

Les moyennes des traitements sont comparées selon la méthode décrite pour la section ANOVA 1.