Table des matières

Préalable

Cela fait tout juste un an que la rédaction de la présente note de biométrie a débuté. L'idée d'écrire une telle note m'est venue d'un certain nombre de constats. En effet, les différentes consultations statistiques que j'ai effectuées, les discussions que j'ai eu avec des étudiants, chercheurs et enseignants-chercheurs m'ont permis de noter que très souvent, l'utilisateur des méthodes statistiques paramétriques se soucie très peu ou pas du tout de leurs conditions d'application. Ceci est généralement lié au fait que ces conditions sont souvent inconnues de l'utilisateur. Certains utilisateurs bien que connaissant les conditions d'application, ne les prennent pas en compte tout simplement parce qu'ils ne mesurent pas ou plutôt ne savent pas les conséquences liées à leur non-respect. D'autres utilisateurs par contre connaissent l'importance de ces conditions d'application mais ne savent pas comment les vérifier du moins en s'aidant de l'ordinateur.

J'ai alors décidé d'écrire une note de biométrie pour traiter ces différents aspects afin de sensibiliser la communauté scientifique sur l'importance du respect des conditions d'application des méthodes statistiques paramétriques et par la même occasion d'exposer en pratique la vérification de ces conditions avec les moyens informatiques.

Les préoccupations étant multiples, j'ai sollicité l'aide de mon collègue Sodjinou E. pour la réalisation de cette œuvre. La collaboration scientifique qui en est résultée a permis la rédaction de la présente note qui a été soumise à notre aîné dans le domaine, le professeur Fonton pour la touche finale. C'est le lieu pour moi de remercier les personnes qui nous ont aidés d'une manière ou d'une autre notamment le Professeur R. Palm de la Faculté Universitaire des Sciences Agronomiques de Gembloux (Belgique) dont les remarques et suggestions ont permis d'améliorer la qualité scientifique de l'ouvrage.

Mon souhait est que la présente note contribue au renforcement de l'excellence scientifique à travers l'amélioration de la qualité des résultats de travaux de recherche.

Glèlè Kakaï R.

1. Introduction

Les méthodes statistiques paramétriques nécessitent le respect des hypothèses de base faites lors de leur conception. La violation des conditions d'application de ces méthodes statistiques donne souvent lieu à de fausses interprétations des résultats obtenus puisque rien ne garantit la précision des méthodes en dehors de leurs hypothèses d'utilisation.

La méconnaissance par l'utilisateur des hypothèses d'utilisation de ces méthodes l'amène souvent à ignorer cette étape importante du traitement des données de recherche.

La présente note de biométrie a pour but essentiel de présenter les hypothèses d'utilisation des méthodes statistiques paramétriques courantes, le cadre théorique d'élaboration des méthodes statistiques paramétriques, les conséquences liées à la violation de ces conditions ainsi que les tests d'hypothèses utilisés pour leur vérification.

Après cette introduction (chapitre 1), nous abordons au chapitre 2 les principales méthodes statistiques paramétriques avec leurs conditions d'utilisation, le cadre théorique d'établissement des méthodes statistiques, l'importance du respect des conditions d'application et les conséquences pratiques de leur violation.

Le chapitre 3 aborde les principes sous-tendant les tests d'hypothèse pour la vérification de ces hypothèses avec à chaque étape, une présentation claire et illustrée de l'application de ces tests dans les logiciels statistiques Minitab, SPSS et SAS, afin d'aider le lecteur à mieux comprendre leur fondement et à pouvoir les appliquer sur ordinateur. Pour le logiciel, Minitab, la version 13 française est utilisée alors que dans le cas de SPSS, c'est la version française 10.1.3 qui est prise en compte. Quant au logiciel SAS, nous avons utilisé la version 9.1. Dans certains cas, nous avons eu recours à de la conception de procédures dans le langage Matlab pour les tests non disponibles dans les trois logiciels ci-dessus cités. La version du langage Matlab utilisée à cet effet est R2006a.

2. Conditions d'application des méthodes statistiques paramétriques

2.1. Présentation des principales méthodes avec leurs conditions d'application

2.1.1. Cas d'une variable

Les différents tests et leurs conditions d'utilisation sont présentés au tableau 1.

Tableau 1. Méthodes statistiques paramétriques pour une variable et conditions d'utilisation.

Méthodes statistiques paramétriques	Conditions d'application
Test de conformité d'une proportion Test d'égalité de 2 ou plusieurs proportions	- Echantillons aléatoires simples et indépendants.
Test de conformité d'une moyenne	- Echantillon aléatoire simple. - Echantillon tiré de population normale.
Test d'égalité de deux moyennes	- Echantillons aléatoires simples et indépendants. - Echantillons tirés de populations normales.
Test t pour données appariées	- Echantillons aléatoires simples et dépendants. - Echantillons tirés de populations normales.
Test de conformité d'un ou de deux écarts-types (ou variances)	- Echantillons aléatoires, simples et indépendants (ou non)[1]. - Echantillons tirés de populations normales.
Test de conformité du rapport de deux écarts-types ou de deux variances à une valeur théorique.	- Echantillons aléatoires et simples. - Echantillons tirés de populations normales.
Test d'égalité de plusieurs écarts-types ou de plusieurs variances (test de Bartlett, test de Levene, etc.).	- Echantillons aléatoires, simples et indépendants. - Echantillons tirés de populations normales ou non[2].
Analyse de la variance à p critères de classification.	- Echantillons aléatoires et indépendants. - Echantillons tirés de populations normales. - Egalité des variances des populations.

2.1.2. Cas de deux variables

[1] Les tests d'égalité de deux écarts-types ou de deux variances varient selon le caractère dépendant ou non des échantillons.

Dans le cas de deux variables considérées simultanément, le tableau 2 présente les méthodes statistiques paramétriques souvent utilisées et leurs conditions d'utilisation.

Tableau 2. Méthodes statistiques pour deux variables observées simultanément et conditions d'utilisation.

Méthodes statistiques paramétriques	**Conditions d'application**
Test d'indépendance ou test du Chi2.	- Echantillons aléatoires et simples.
Test de signification ou de conformité d'un coefficient de corrélation. Test d'égalité de deux coefficients de corrélation.	- Echantillons aléatoires et simples. - Echantillons tirés de populations normales bivariées. - Valeurs de variables connues sans erreurs de mesure.
Régression linéaire simple. Test d'égalité de deux coefficients de régression. Test de conformité d'un coefficient de régression.	- Normalité des résidus de régression. - Nullité de la moyenne des résidus. - Homogénéité des résidus de régression. - Indépendance des résidus de régression.

2.1.3. Méthodes multivariées

Les conditions d'application des méthodes statistiques multivariées sont présentées au tableau 3.

Tableau 3. Méthodes statistiques multivariées et conditions d'utilisation.

Méthodes statistiques paramétriques	**Conditions d'application**
Analyse en composantes principales (ACP)	- Aucune condition
Analyse factorielle des correspondances (AFC)	- Tableau de contingence.
La classification numérique	- Aucune condition
Analyse discriminante linéaire et quadratique	- Echantillons aléatoires et simples. - Echantillons tirés de populations multinormales. - Egalité ou non[1] des matrices de variances-covariances.
Analyse de la variance multivariée et analyse canonique discriminante	- Echantillons aléatoires et simples. - Echantillons tirés de populations multinormales. - Egalité des matrices de variances-covariances.
L'analyse de la corrélation canonique	- Echantillons aléatoires et simples. - Echantillons tirés de populations multinormales.

[1] L'analyse discriminante linéaire nécessite l'égalité des matrices de variances-covariances, ce qui est le contraire de l'analyse discriminante quadratique.

2.2. Cadre théorique d'établissement des méthodes statistiques paramétriques

Nous exposons dans ce paragraphe le fondement des méthodes statistiques paramétriques avec pour objectif d'expliquer et justifier l'origine des hypothèses d'utilisation de ces méthodes.

2.2.1. Les méthodes statistiques relatives à une ou plusieurs moyennes

Parmi ces méthodes, nous pouvons citer le test de conformité d'une moyenne, le test t d'égalité de deux moyennes et ses variantes ainsi que l'analyse de la variance.

2.2.1.1. Test de conformité d'une moyenne

Le test de conformité d'une moyenne permet de tester l'égalité de la moyenne inconnue m d'un caractère donnée d'une population à une valeur connue m_0 à partir d'un échantillon tiré de cette population. L'hypothèse nulle relative à ce test est :

$$\mathrm{H}_0 : m = m_0 .$$

Si la moyenne de la population était connue, il serait assez trivial de la comparer à la valeur m_0 et de décider si elles sont égales ou différentes. Puisque la moyenne de la population est inconnue, on considère un échantillon de cette population. La moyenne $\bar{x}$ du caractère, calculé à partir de l'échantillon est une estimation non biaisée de la moyenne m de la population. Lorsque les limites de variabilité admises de l'estimation de la moyenne de la population sont connues ou plus précisément lorsque les limites de confiance de la moyenne estimée de la population *peuvent être calculées*, il est alors facile de vérifier l'hypothèse nulle H_0. En effet, si la valeur m_0 est contenue dans l'intervalle de confiance de la moyenne estimée, on accepte l'hypothèse nulle H_0. Dans le cas contraire, cette hypothèse est rejetée. Pour déterminer ces limites de confiance de la moyenne estimée, il faut émettre l'hypothèse de variance minimum de cette estimation. Cette hypothèse n'est vérifiée que pour un nombre très limité de distributions théoriques dont la distribution normale. De plus, lorsque cette hypothèse est acceptée, il faut pouvoir trouver une méthode de calcul des limites de confiance. Le calcul des limites de confiance de cette estimation est assez complexe mais est facilité par la supposition du caractère normal de la population considérée. En d'autres termes, lorsque la population considérée suit une distribution normale et sa variance est connue, il est possible de déterminer de façon simple les limites de confiance $\bar{x}_1$ et $\bar{x}_2$ de la moyenne estimée à partir de la variable normale réduite (Dagnelie, 1998) :

$$\bar{x}_1 = \bar{x} - u_{1-\alpha/2}\sigma / \sqrt{n} \quad \text{et} \quad \bar{x}_2 = \bar{x} + u_{1-\alpha/2}\sigma / \sqrt{n} . \qquad (2.2.1)$$

Lorsque la variance σ de la population n'est pas connue, ce qui est courant en pratique, elle peut être estimée ($\hat{\sigma}$) à partir de l'échantillon considéré ; de ce fait la variable normale réduite est remplacée par la variable t de Student qui est asymptotiquement normale :

$$\bar{x}_1 = \bar{x} - t_{1-\alpha/2}\hat{\sigma} / \sqrt{n} \qquad \text{et} \qquad \bar{x}_2 = \bar{x} + t_{1-\alpha/2}\hat{\sigma} / \sqrt{n} . \qquad (2.2.2)$$

La détermination des limites de confiance de la moyenne estimée d'une population dans le cas du test de conformité d'une moyenne est donc subordonnée à l'hypothèse du caractère normal de la population. On peut alors comprendre que la normalité de la population est une condition importante à la réalisation du test. Lorsque la population considérée ne suit pas la distribution normale, les résultats du test peuvent être sensiblement biaisés.

2.2.1.2. Test d'égalité de deux moyennes et ses variantes

Supposons que l'on veuille comparer les moyennes inconnues de deux populations pour un caractère donné. L'hypothèse nulle de ce test est :

$$\mathrm{H}_0 : m_1 = m_2 .$$

Puisque les moyennes des deux populations ne sont pas connues, on considère un échantillon aléatoire et simple de chacune des deux populations à partir desquels on détermine les moyennes estimées $\hat{m}_1$ et $\hat{m}_2$ des deux populations. En supposant que ces deux populations considérées sont *normales* de moyennes m_1 et m_2 et d'écarts-types σ_1 et σ_2, les moyennes $\hat{m}_1$ et $\hat{m}_2$ des échantillons supposés indépendants, de tailles n_1 et n_2 sont des valeurs de deux variables aléatoires de moyennes m_1 et m_2 et d'écarts-types σ_1/n_1 et σ_2/n_2. Du fait des propriétés d'additivité et de linéarisation *des distributions normales*, la différence de moyennes $m_1 - m_2$ est une valeur d'une variable normale de moyenne $m_1 - m_2$ et d'écart-type $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. De ce fait, on peut déterminer comme dans le cas du test de conformité d'une moyenne, les limites de confiance de cette différence et vérifier si la valeur 0 est contenue dans cet intervalle de confiance de la différence. Dans le cas où les variances des populations seraient inconnues, on utilise la distribution t de Student qui est asymptotiquement normale.

Comme on peut donc le constater ci-dessus, la condition de normalité est essentielle pour l'application de cette méthode statistique dont le fondement théorique suppose que les populations considérées sont normales.

Par ailleurs, le fait de déterminer un écart-type commun pour la variable aléatoire normale relative à la différence entre les moyennes des deux populations suppose l'égalité des écart-types de ces populations pour obtenir une estimation non biaisée. Le biais de cette estimation est d'autant plus élevé que l'inégalité entre les écarts-types des populations est importante. Ceci amène à considérer une seconde condition à l'utilisation du test t classique de comparaison de deux moyennes à savoir, l'égalité des variances des populations. Néanmoins, dans le cas où l'égalité des variances des populations ne serait pas vérifiée, il est possible de déterminer de manière approchée l'estimateur non biaisé de l'écart-type commun des deux populations en déterminant de façon indépendante les sommes des carrés des écarts des deux populations. Cette méthode de détermination de l'écart-type commun des deux populations conduit à l'une des variantes du test t de Student, appelée *test de Welch*[1]. L'autre variante du test t de Student est le test t pour données appariées utilisé lorsque les échantillons considérés sont dépendants les uns des autres ; ce test nécessite seulement la condition de normalité des populations.

On peut donc noter que la condition de normalité est ici aussi nécessaire pour une application sans risque des tests t d'égalité de deux moyennes.

2.2.1.3. L'analyse de la variance univariée et multivariée

Considérons une population avec un caractère donné dont la moyenne est inconnue. On tire un échantillon de cette population et on calcule la moyenne du caractère qui est une estimation non biaisée de la moyenne de la population qui est de variance minimum (cf. paragraphe 2.2.1.1). Cette variance-seuil est assez complexe à calculer sauf si on suppose que la population considérée suit une distribution normale.

Sans perte de généralités, supposons que l'on veuille comparer les moyennes m_1 , m_2 , .., m_p de p populations pour un caractère donné. Pour ce faire, on considère l'hypothèse nulle :

$$\mathrm{H}_0 : m_1 = m_2 = \ldots m_p \, .$$

Pour vérifier cette hypothèse, on considère p échantillons tirés de façon aléatoire et indépendante des p populations et on calcule les moyennes estimées. L'hypothèse nulle H_0 établie ci-dessus suppose l'égalité des moyennes des populations ou encore que les échantillons considérés appartiennent à une même population du moins pour le caractère considéré. En

[1] En anglais : Welch's approximate t-test.

faisant une telle hypothèse, les observations des p échantillons sont considérées comme celles d'un seul échantillon tiré d'une même population. De ce fait, on calcule une variance factorielle de ladite (ou supposée) population résultant de la différence entre les échantillons. Si les échantillons appartenaient à une même population pour le caractère considéré, cette variance factorielle ne devrait pas dépasser la variance minimum admise d'une population qui ne peut par ailleurs, être approchée qu'en supposant le caractère normal des populations considérées comme on l'a notifié au début du paragraphe. Cette variabilité résiduelle est mesurée par une variance résiduelle issue de la différence entre les observations des p échantillons. Lorsque la variance factorielle dépasse la variance résiduelle admise pour une population, on conclut que les échantillons n'appartiennent en fait pas à une même population. On peut donc noter de ce qui précède que l'hypothèse nulle dans le cas d'une analyse de la variance peut être encore formulée de la façon suivante : *la variance observée entre les échantillons est purement aléatoire, c'est-à-dire n'est due qu'au hasard et non à des différences effectives entre les échantillons*. En réalité, puisque les deux types de variances sont estimés à partir d'échantillons tirés des p populations, leur comparaison ne peut pas se faire sur un seul jeu d'échantillons. De ce fait, en établissant la distribution du rapport des deux variances (la variance factorielle étant au numérateur), il est possible de déterminer une valeur-seuil de ce rapport, au-delà de laquelle l'hypothèse nulle de variance aléatoire sera rejetée. En considérant que les p populations suivent chacune une distribution normale de même variance et en utilisant les propriétés particulières des distributions normales, on peut montrer que chacun des deux types de variances (factorielle et résiduelle) suit à une constante près une distribution Chi-carré. Puisque le rapport de deux variables Chi-carré donne une variable F de Fisher-Snedecor, on peut admettre que le rapport des deux types de variances suit une distribution F et permet ainsi de déterminer dans le cas de l'hypothèse nulle, la valeur-seuil du rapport des deux variances et donc de réaliser le test d'analyse de la variance.

De ce qui précède, on note aisément que la distribution normale a servi de base à l'élaboration de l'analyse de la variance. De ce fait, la normalité et l'égalité des variances des populations ainsi que le caractère aléatoire et simple des échantillons sont les conditions d'application de cette méthode statistique. La condition d'égalité des variances est surtout nécessaire lors de la structuration des moyennes à la suite d'une analyse de la variance révélant une différence significative entre les moyennes des populations.

Lorsque plusieurs variables quantitatives sont observées de façon simultanée sur les mêmes objets, au lieu de réaliser une série d'analyses univariées indépendantes, l'analyse de la variance multivariée est plus indiquée puisqu'elle prend en compte les corrélations qui existent très souvent entre les variables étudiées. L'analyse de la variance multivariée est une extension *naturelle* de l'analyse de la variance univariée. Ainsi, les conditions d'application de l'analyse de la variance multivariée sont aussi des extensions naturelles des conditions d'application de l'analyse de la variance univariée : il faut que les échantillons soient aléatoires, simples et indépendants ; il faut en outre que les

populations considérées aient des distributions multinormales, de même matrice de variances-covariances. Ces conditions sont aussi nécessaires pour l'application de l'analyse canonique discriminante encore appelée analyse factorielle discriminante qui constitue un complément logique de l'analyse de la variance multivariée. En effet, cette méthode a pour but de décrire les différences liées aux facteurs étudiés, du moins lorsque ces différences existent.

2.2.2. Les méthodes statistiques relatives à une ou plusieurs variances

L'inférence statistique relative à une ou plusieurs variances ou écart-types prend en compte le test de conformité d'un écart-type et le test de comparaison de deux ou plusieurs écart-types ou variances.

2.2.2.1. Test de conformité d'une variance

L'objectif poursuivi ici est la comparaison de la variance σ d'une population donnée à une valeur σ_0 et l'hypothèse nulle est alors:

$$H_0: \sigma = \sigma_0$$

Puisque la variance de la population est généralement inconnue. On considère un échantillon tiré de cette population et on estime la variance théorique σ par la variance estimée $\hat{\sigma}$ à partir de l'échantillon. On détermine ensuite l'intervalle de confiance ou encore les limites de cette variance estimée. Lorsque la valeur σ_0 se trouve dans l'intervalle de confiance de la variance estimée $\hat{\sigma}$, on accepte l'hypothèse nulle et on conclut que la variance de la population est égale à σ_0 dans le cas contraire, on rejette l'hypothèse nulle.

Pour déterminer la variance estimée, on va utiliser une propriété donnée de la distribution d'échantillonnage de la variance qui stipule que (Dagnelie, 1998) *dans le cas d'un échantillonnage aléatoire et simple et quelle que soit la distribution de la population considérée, la variance pour un caractère donné est estimée sans biais lorsque la somme des carrés des écarts est divisée par le nombre de degrés de liberté*. De ce fait, on a :

$$\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ . \qquad (2.2.3)$$

La variance σ étant estimée sans biais, on va pouvoir déterminer les limites de confiance de cette estimation. Si l'estimation sans biais de la variance est obtenue quelle que soit la distribution de la population considérée, il n'en ait pas de même pour le calcul des limites de confiance de cette estimation. En effet, la détermination de l'intervalle de confiance de la variance estimée est assez complexe et ne conduit pas à une relation simple. Néanmoins, en supposant

que la population considérée suit une distribution normale du moins pour le caractère considéré, les observations faites pour ce caractère sont alors celle d'une variable aléatoire normale. La variance du caractère étant à une constante près, le carré de la transformation linéaire de la variable normale (cf. formule 2.2.3), elle suit une distribution Chi-carré de Pearson. De ce fait, les limites s_1 et s_2 de la variance estimée, en supposant un échantillonnage aléatoire et simple et surtout le caractère normal de la population considérée est alors:

$$s_1 = \mathrm{SCE}/\chi^2_{1-\alpha/2} \qquad \text{et} \qquad s_2 = \mathrm{SCE}/\chi^2_{\alpha/2} \tag{2.2.4}$$

On note alors de ce qui précède que le test de conformité d'une variance ou d'un écart-type est conçu en supposant le caractère normal de la population considérée. On peut donc comprendre que l'application de ce test à une population non normale peut conduire à des résultats erronés.

2.2.2.2. Test d'égalité de deux ou plusieurs variances

Pour comparer deux populations du point de vue de leurs variances σ_1 et σ_2, l'hypothèse nulle à considérer est :

$$\mathrm{H}_0 : \sigma_1 = \sigma_2 .$$

Si les variances de ces deux populations étaient connues, la comparaison serait simple et évidente. Mais puisque ce n'est pas le cas, en déterminant la distribution du rapport des deux variances, il est possible de déterminer la valeur-seuil de ce rapport au-delà de laquelle l'hypothèse nulle sera rejetée. La distribution de ce rapport est assez complexe et ne peut pas être établie quelle que soit la distribution des populations. Comme toujours, puisque la distribution normale est la plus simple et la plus intuitive, on suppose que les populations considérées sont normales et que les échantillons servant à estimer leurs variances sont tirés de façon aléatoire, simple et indépendante des populations. De ce fait, les deux variances sont les valeurs observées de deux variables aléatoires suivant chacune une distribution Chi-carré. Le carré de deux variables Chi-carré étant une variable F de Fisher-Snedecor, le rapport des deux variances suit une distribution F et sert donc à réaliser le test de comparaison des deux variances. On peut donc comprendre que l'une des conditions essentielles de ce test est la normalité des populations considérées puisque cette distribution a servi de base à l'élaboration du test.

Dans le cas de plus de deux populations, la plupart des tests utilisés notamment les tests de Hartley et de Bartlett suppose aussi la normalité des populations considérées. Le test de Levene par ailleurs, ne nécessite pas la condition de normalité pour son application comme on peut le noter par la suite.

2.2.3. Les méthodes statistiques relatives à la régression linéaire

Pour établir une relation permettant de prédire une variable donnée appelée variable dépendante en fonction d'une ou de plusieurs variables dites explicatives ou indépendantes, on peut utiliser la méthode de régression linéaire qui ajuste des observations au modèle linéaire.

Sans perte de généralités, considérons une variable aléatoire Y que l'on veut estimer à partir d'une autre variable aléatoire X en utilisant le modèle linéaire. De ce fait, pour toutes observations x et y des variables aléatoires X et Y, on peut écrire :

$$y = a + bx .$$

En considérant un échantillon bivarié tiré de la population considérée, les paramètres a et b peuvent être estimés par la méthode des moindres carrés. Lorsque l'ajustement est ainsi établi, on cherche à tester sa signification en d'autres termes, si les valeurs obtenues pour ces deux paramètres sont dues au hasard de l'échantillonnage, donc sont en réalité nulles ou au contraire, si elles sont différentes de zéro. De plus, on pourra étudier la distribution des écarts entre les valeurs réelles observées y de la variable aléatoire Y et les valeurs estimées $\hat{y}$ à partir de l'équation établie. Ces écarts sont appelés les résidus e :

$$e = y - \hat{y} .$$

Pour déterminer les limites de confiance de chaque estimation de la variable dépendante, il est nécessaire non seulement de connaître la distribution d'échantillonnage de la variance des résidus mais aussi celle des paramètres a et b. En supposant que les résidus e suivent une distribution normale, il est plus simple de calculer les limites de confiance de la variance résiduelle. En effet, dans de telles conditions, la variance résiduelle suit une distribution bien connue à savoir la distribution Chi-carré de Pearson. Dans le cas des paramètres a et b, il faut en plus admettre la constante de la variance résiduelle. Si cette condition n'est pas remplie, l'estimateur au sens des moindres carrés n'est plus de variance minimum et le calcul des limites de confiance ne serait plus précis (Palm, 1994).

On peut donc comprendre que les conditions d'application de la régression linéaire sont la normalité, l'homoscédasticité et l'indépendance des résidus.

2.2.4. Les méthodes d'analyse discriminante décisionnelle

L'analyse discriminante décisionnelle[1] est une méthode statistique d'affectation d'observations inconnues à un groupe, parmi deux ou plusieurs groupes connus *a priori* sur la base d'observations antérieures.

Considérons une situation où l'on cherche à affecter un individu i, caractérisé par un vecteur d'observations $\boldsymbol{x}_i$, au groupe le plus probable, parmi g groupes ou populations connues *a priori*. Le groupe le plus probable est celui pour lequel la probabilité théorique d'appartenance de l'individu i, connaissant son vecteur d'observations $\boldsymbol{x}_i$, est la plus élevée. Soit r, la règle de classement établie à cet effet, de sorte que la notation $r(\boldsymbol{x}_i)=k$ signifie que le vecteur d'observations $\boldsymbol{x}_i$ ou encore l'individu i est classé dans le $k^{ème}$ groupe ($k=1,...,g$). Elle peut alors être définie, pour des probabilités *a priori* égales de la manière suivante (Glèlè Kakaï et al., 2005) :

$$r(\boldsymbol{x}_i)=k \text{ si } f_k(\boldsymbol{x}_i) \geq f_j(\boldsymbol{x}_i),\ \forall\ j=1,\ldots k\text{-}1, k+1,\ldots,g. \qquad (2.2.5)$$

Cette règle ci-dessus définie est basée sur le calcul de la valeur des fonctions de densité de probabilité théoriques des individus dans chacune des g populations. De tout ce qui précède, la règle de classement, r, ne peut être appliquée que si les paramètres réels des populations sont connus et constitue ainsi « la règle idéale ». En pratique, les paramètres des populations ne sont pas connus. Les valeurs exactes des fonctions de densité de probabilité théoriques des individus dans les différents groupes ne peuvent donc plus être calculées. Beaucoup d'autres règles de classement sont alors conçues de sorte qu'elles peuvent être établies sur des échantillons représentatifs des g populations de départ, par l'estimation des fonctions de densité théoriques par exemple. Parmi celles-ci, nous pouvons citer la règle linéaire[2] et la règle quadratique[3].

Les règles linéaire et quadratique supposent la normalité des populations considérées. De ce fait, il est plus simple d'estimer les fonctions de densité de probabilité.

Dans le cas de la règle linéaire, pour l'individu i, de vecteur d'observations $\boldsymbol{x}_i$, on a alors (Dagnelie, 1998) :

$$\hat{f}_1(\boldsymbol{x}_i)=\frac{1}{2\pi\hat{\sigma}_1\hat{\sigma}_2\sqrt{(1-\hat{\rho}^2)}}\exp\left(-\frac{1}{2}\hat{d}_{1i}^2\right) \qquad (2.2.6)$$

et

[1] En anglais : discriminant analysis, predictive discriminant analysis.
[2] En anglais : linear discriminant analysis.
[3] En anglais : quadratic discriminant analysis.

$$\hat{f}_2(\boldsymbol{x}_i)=\frac{1}{2\pi\hat{\sigma}_1\hat{\sigma}_2\sqrt{(1-\hat{\rho}^2)}}\exp\left(-\frac{1}{2}\hat{d}_{2i}^2\right), \qquad (2.2.7)$$

avec :

$$\hat{d}_{1i}^2=[\boldsymbol{x}_i-\overline{\boldsymbol{x}}_1]'\hat{\boldsymbol{\Sigma}}^{-1}[\boldsymbol{x}_i-\overline{\boldsymbol{x}}_1] \quad \text{et} \quad \hat{d}_{2i}^2=[\boldsymbol{x}_i-\overline{\boldsymbol{x}}_2]'\hat{\boldsymbol{\Sigma}}^{-1}[\boldsymbol{x}_i-\overline{\boldsymbol{x}}_2].$$

On peut noter des expressions (2.2.6) et (2.2.7) que seules les distances $\hat{d}_{1i}$ et $\hat{d}_{2i}$ changent, toutes les autres composantes étant constantes d'une expression à l'autre. Ces deux distances prennent chacune en compte la matrice de variances-covariances groupée des deux échantillons, ce qui suppose une égalité des matrices de variances.

Ainsi, comme on peut le noter de ce qui précède, les conditions de multinormalité et d'égalité de matrice de variances-covariances doivent être remplies pour une performance optimale de la règle linéaire discriminante, du moins en espérance. Lorsque l'égalité des matrices de variances-covariances des populations considérées n'est pas acquise, l'utilisation de la règle quadratique discriminante est conseillée. Mais de récentes études, notamment celle de Glèlè Kakaï et Palm (2004), ont montré que la règle linéaire discriminante présente un taux d'erreur plus faible que la règle quadratique en cas d'hétéroscédasticité modérée.

2.3. Importance pratique du respect des conditions d'application

2.3.1. Importance de la normalité en inférence statistique

La normalité de la population dont est issu l'échantillon est l'une des conditions les plus importantes dans l'utilisation des méthodes paramétriques. Ainsi, de façon générale, en inférence statistique, le calcul de la probabilité associée à un test, de même que l'estimation et la détermination des limites de confiance d'une moyenne ou d'un écart-type se basent sur l'hypothèse de normalité des observations. On peut donc comprendre que lorsqu'une telle hypothèse (la normalité) n'est pas satisfaite, l'utilisation de la méthode statistique peut conduire à des résultats biaisés.

La propriété de normalité asymptotique de la distribution d'échantillonnage de la moyenne rend moins importante la condition de normalité pour de grands échantillons dans le cas des tests d'égalité de moyennes ou de vecteurs de moyennes. Malheureusement, il n'en est pas de même lorsqu'on s'intéresse à la structuration de moyennes après une analyse de la variance (tests de Newman et Keuls, de Dunnett, de Bonferroni, de Tukey, etc.). De plus, les tests d'égalité des variances, écarts-types ou matrices de variances-covariances sont nettement plus sensibles à la non-normalité des populations-parents.

2.3.2. Importance pratique de la condition d'homoscédasticité en inférence statistique

La condition d'homoscédasticité en inférence statistique concerne l'hypothèse d'égalité des variances ou écarts-types des échantillons dans le cas des tests univariés, et l'égalité des matrices de variances-covariances dans le cas des tests multivariés.

En inférence statistique à deux ou plusieurs dimensions, la condition d'homoscédasticité est nécessaire surtout en cas de structuration des vecteurs de moyennes avec l'analyse canonique discriminante puisqu'elle utilise l'estimation commune des matrices de variances-covariances des populations.

En inférence statistique à une dimension (tests de comparaison de moyennes, analyse de la variance etc.), l'hypothèse d'égalité des variances des échantillons prend toute son importance dans l'estimation et la détermination des limites de confiance ainsi que la détermination du nombre d'observations. En effet, pour ces différentes situations, c'est la variance estimée commune des différentes populations qui est utilisée. Ceci suppose que ces variances doivent être significativement égales pour garantir une bonne précision de calcul de la variance commune. Il en est de même en analyse discriminante linéaire où les matrices de variances-covariances doivent être significativement égales pour une estimation non biaisée de la matrice de variances-covariances groupée des populations multivariées considérées.

2.4. Conséquences pratiques du non-respect des conditions d'application

Pour illustrer les conséquences du non-respect des conditions d'application des méthodes statistiques paramétriques énoncées ci-dessus, nous considérons un exemple relatif au test t de comparaison de deux moyennes pour lequel les conditions d'application sont la normalité et l'égalité des variances des populations.

2.4.1. Non-normalité associée à une homoscédasticité

Considérons deux populations normales, l'une P_1, de moyenne 3 et d'écart-type 1 et l'autre, P_2, de moyenne 2 et d'écart-type 1. Considérons deux autres populations telle que la première, P_3 est normale de moyenne 3 et d'écart-type 2 et l'autre, P_4, de distribution Chi-carré à 2 degrés de liberté. Nous rappelons que la moyenne et l'écart-type d'une distribution Chi-carré à k degrés de liberté est respectivement k et $\sqrt{2k}$. On note que les populations P_1 et P_2 remplissent les conditions d'application du test t de Student (normalité et égalité des variances) et ont des moyennes différentes, la différence de moyennes étant égale à 1. Le couple de populations (P_3,P_4) ne remplit par contre que la condition d'égalité des variances ($\sigma_1=\sigma_2=2$). En effet, la population P_4 suit une distribution non-normale (Chi-carré à 2 degrés de liberté). Les moyennes de ces deux populations sont aussi différentes et la différence entre les deux moyennes est égale aussi à l'unité. Les fonctions de densité de probabilité de ces deux couples de populations sont présentées à la figure 1. Les distributions utilisées sont centrées et réduites.

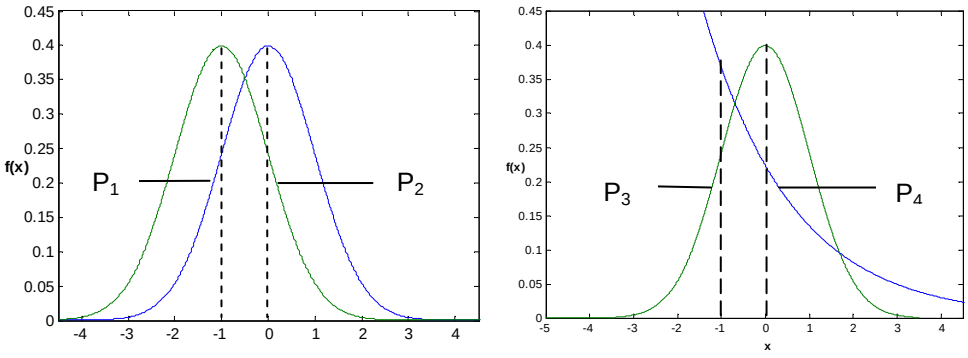


Figure 1. Fonctions de densité de probabilité des distributions considérées dans le cas d'égalité des variances mais de non-normalité.

Pour illustrer les conséquences de la non-normalité sur la précision des

résultats du test t d'égalité des moyennes, nous générons un couple d'échantillons de n observations de chacun des deux couples de populations (P_1,P_2) et (P_3,P_4) ci-dessus considérés et réaliser le test t d'égalité des moyennes des populations.

Les moyennes des couples de populations étant différentes dans les deux cas (3 et 2), on devrait s'attendre à ce que l'hypothèse d'égalité des moyennes soit rejetée, en d'autres termes, que la probabilité liée au test t d'égalité des moyennes à exécuter soit inférieure à 0,05. Nous rappelons que les conditions d'application de ce test sont respectées dans le premier cas (populations P_1 et P_2) alors que dans le second cas (populations P_3 et P_4), la condition de normalité n'est pas respectée, la distribution de P_4 n'étant pas normale.

Pour différentes valeurs de la taille commune n des échantillons tirés des couples de populations (P_1,P_2) et (P_3,P_4), la puissance du test d'égalité des moyennes est calculée dans les deux cas. La puissance du test, notée 1-β est la probabilité de rejeter l'hypothèse nulle (égalité des moyennes des populations) alors qu'elle est fausse. Elle est un critère de performance des tests inférentiels. Ainsi, puisque les moyennes des deux populations dans les deux cas sont différentes (hypothèse nulle fausse), on s'intéressera à la probabilité que le test rejette cette hypothèse nulle dans les deux cas et pour différentes tailles d'échantillons. Pour ce faire, 5000 couples d'échantillons de taille n sont générés à partir de chacun des couples de populations (P_1,P_2) et (P_3,P_4). Le test t d'égalité des moyennes est effectué dans les deux cas sur chacun des 5000 couples d'échantillons et le nombre N de fois que le test rejette l'hypothèse nulle d'égalité des moyennes est noté et permet d'estimer la puissance du test à partir de la relation :

$$\text{Puissance} = 1-\beta = \frac{N}{5000}.$$

Les valeurs de puissance du test d'égalité des moyennes calculées pour différentes tailles d'échantillons dans les deux cas sont représentées graphiquement à la figure 2.

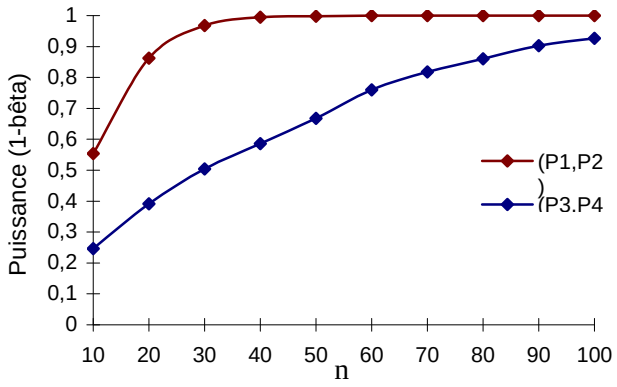


Figure 2. Non-normalité : puissance du test d'égalité des moyennes pour les couples de populations (P_1,P_2) et (P_3,P_4).

On note de cette figure que, de façon générale, la puissance du test d'égalité des moyennes augmente avec l'accroissement de la taille des échantillons. Par ailleurs, lorsque la condition de normalité est remplie (populations P_1 et P_2), la valeur maximale de la puissance est vite atteinte, notamment à partir de la taille *n* égale à 50, ce qui n'est pas le cas des populations P_3 et P_4 pour lesquelles cette condition n'est pas remplie. En effet, dans une telle situation, pour des tailles d'échantillons inférieures à 50, la puissance 1-β est au plus de 50 % ; en d'autres termes, une fois sur deux, le test t conduit à des conclusions erronées. De plus, pour une même taille d'échantillons *n*, la valeur de puissance du test lorsque la normalité est respectée est sensiblement plus élevée que dans le cas du non-respect de cette condition, la différence de valeur de puissance allant jusqu'à 0,48 pour certaines valeurs de *n*. En résumé, en cas de non-normalité, la probabilité de rejeter une fausse hypothèse est assez faible ; de ce fait, il sera relativement plus facile de conclure à une égalité de moyennes de populations alors qu'en réalité ces moyennes sont différentes.

De façon générale, les conséquences du non-respect de la condition de normalité en inférence statistique ont été abordées par nombre d'auteurs, notamment Dehlert (2000) du moins en ce qui concerne l'analyse de la variance. Pour étudier les conséquences du non-respect de la condition de normalité, cet auteur considère deux paramètres à savoir, les coefficients de symétrie et d'aplatissement dont ceux de Fisher sont notés γ_1 et γ_2 respectivement. Nous rappelons que pour la distribution normale, ces deux coefficients sont tous égaux à 0. Il ressort de son étude les conclusions présentées ci-après.

Pour des valeurs γ_1 (de la population-parent) supérieures à 0, le risque de première espèce α du test d'analyse de la variance se déplace à un niveau

supérieur à 5 % de sorte que l'utilisateur de la méthode conclura facilement à une différence significative alors qu'il en est rien. On parle de *test libéral*.

Pour des valeurs négatives de γ_1, le test d'analyse de la variance présente un risque réel α inférieur au risque nominal de 5 %. On parle de *test conservateur*.

Pour des valeurs γ_2 supérieures à 0, le test de l'analyse de la variance présente un risque réel inférieur à la valeur 5 % du risque nominal (*test conservateur*).

Pour des valeurs γ_2 inférieures à 0, le test d'analyse de la variance présente un risque réel supérieur à la valeur 5 % du risque nominal (*test libéral*).

Pour des valeurs positives des coefficients γ_1 et γ_2, le risque réel est souvent inférieur au risque nominal car γ_2 a une plus grande influence. Le test devient donc *conservateur*. Ceci rejoint les conclusions tirées plus haut en ce qui concerne le test t de Student.

Par ailleurs, l'effet du non-respect de la condition de normalité multivariée en inférence multivariée est étudié par nombre d'auteurs du moins en ce qui concerne l'analyse discriminante décisionnelle (Lachenbruch et al., 1973 ; Clarke et al., 1979 ; Bayne et al., 1983 ; Tomassone et al., 1988 ; Glèlè Kakaï et Palm, 2004 ; Glèlè Kakaï et Palm, 2005). Les conclusions de ces différentes études sont les suivantes : le taux d'erreur réel associé à la règle linéaire augmente avec la non-normalité des populations ; Il en est de même de la règle quadratique de classement. Par exemple, Glèlè Kakaï et Palm (2005) sont arrivés à la conclusion que le taux d'erreur réel de la règle linéaire est de 15,7 % lorsque la statistique *r* du test combiné de multinormalité de Rao-Ali et Ryan-Joiner présente des valeurs supérieures à 0,999 (normalité) alors que pour des valeurs de r inférieures à 0,85 (non-normalité), le taux d'erreur réel est de 22,2 %. Mais il est à noter que la règle linéaire présente encore de bonnes performances en cas de non-normalité modérée (Glèlè Kakaï et Palm, 2005).

2.4.2. Hétéroscédasticité associée à une normalité

Pour illustrer les conséquences du non-respect de la condition d'homoscédasticité (égalité des variances ou des matrices de variances-covariances), considérons que la population P_3 est normale de moyenne 3 et d'écart-type 1 alors que P_4 est normale de moyenne 2 et d'écart-type 2. Les populations P_1 et P_2 sont conservées et constituent la situation souhaitée de normalité et d'égalité des variances. On peut noter que les populations P_3 et P_4 ne remplissent que la condition de normalité mais présentent des écarts-types différents ($\sigma_1 = 1$; $\sigma_2 = 2$). Les moyennes des deux populations sont différentes, la différence de moyennes étant égale à l'unité. Les fonctions de densité de probabilité des deux couples de populations sont présentées à la figure 3.

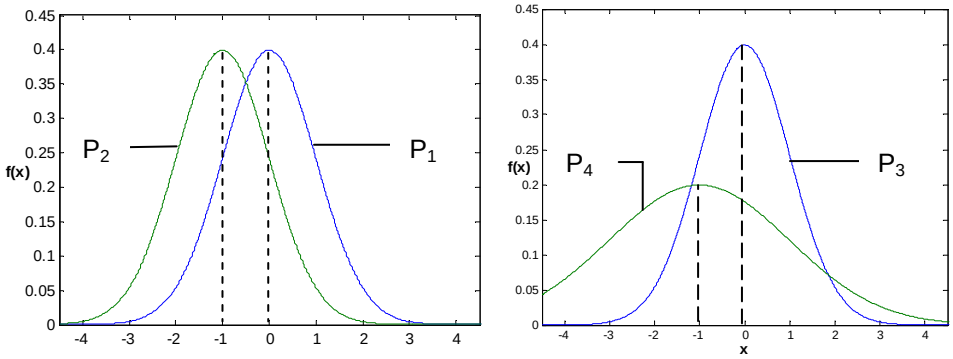


Figure 3. Fonctions de densité de probabilité des distributions considérées dans le cas de normalité et d'inégalité des variances.

Comme précédemment, la puissance du test t d'égalité de deux moyennes est calculée pour différentes tailles d'échantillons dans les deux cas et les résultats sont présentés sous forme graphique à la figure 4. On note de cette figure que, de façon générale, la puissance du test d'égalité des moyennes augmente ici aussi avec l'accroissement de la taille des échantillons. Par ailleurs, lorsque la condition d'égalité des variances est remplie (populations P_1 et P_2), la valeur maximale de la puissance est vite atteinte, notamment à partir de la taille n égale à 50. En situation d'inégalité des variances, la valeur maximale de la puissance 1-β est seulement atteinte lorsque la taille commune des échantillons est de 100. De plus, pour une même taille d'échantillons n, la valeur de puissance du test en cas d'égalité des variances est sensiblement plus élevée que dans le cas du non-respect de cette condition, la différence de valeur de puissance allant jusqu'à 0,38 pour certaines valeurs de n.

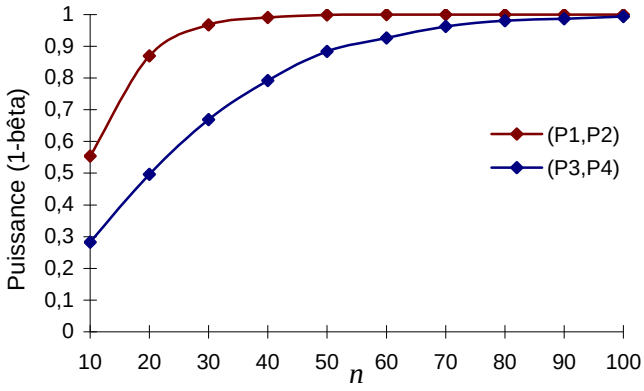


Figure 4. Hétéroscédasticité : puissance du test d'égalité des moyennes pour les couples de populations (P_1,P_2) et (P_3,P_4).

En résumé, en cas d'inégalité des variances, la probabilité de rejeter une fausse hypothèse est assez faible ; de ce fait, il sera ici aussi, relativement plus facile de conclure à une égalité de moyennes de populations alors qu'en réalité ces moyennes sont différentes. Mais notons que l'effet de la non-normalité sur la précision des résultats du test t d'égalité des moyennes semble plus important que celui de l'inégalité des variances du moins en ce qui concerne l'exemple considéré. Cela peut être liée au degré de non-normalité relativement élevé de la distribution Chi-carré à deux degrés de liberté utilisé dans l'exemple.

De façon générale, les conséquences de la non-homogénéité des variances en inférence statistique sont abordées par Dehlert (2000). Les résultats de son étude peuvent être résumés ci-dessous :

En cas d'égalité des tailles de deux échantillons, si le rapport des variances est inférieur à 5, le risque réel est de 50 % supérieur au risque nominal (5 %). De ce fait, les valeurs de probabilités sont sous-estimées conduisant à un *test libéral* (signification plus facile).

En cas d'inégalité des tailles avec un rapport des variances inférieur à 5 et si de plus, les plus grandes variances sont relatives aux plus grands échantillons, la valeur F de Fisher-Snedecor diminue et le risque réel est inférieur au risque nominal. De ce fait, les valeurs de probabilité sont surestimées : on parle de *test conservateur* car il serait difficile de rejeter l'hypothèse nulle.

En cas d'inégalité des tailles avec un rapport des variances inférieur à 5 et si de plus, les plus grandes variances sont relatives aux plus petits échantillons, le risque réel est supérieur de 400 % au risque nominal. Il y a augmentation de la valeur de F et les valeurs de probabilité sont largement sous-estimées. Le test est alors très *libéral* : il serait facile de rejeter l'hypothèse nulle.

Par ailleurs, en inférence multivariée, l'effet du non-respect de la condition d'homoscédasticité sur la performance de l'analyse discriminante linéaire est étudiée par Glèlè Kakaï et Palm (2005) qui sont arrivés à la conclusion que pour de faibles degrés d'hétéroscédasticité ($\hat{\Gamma}$ <1,2), la règle linéaire occasionne un taux d'erreur réel d'environ 16,2 % alors qu'en cas de forte hétéroscédasticité ($\hat{\Gamma}$ > 5), cette règle enregistre un taux d'erreur réel de 22,5 %. Dans cette étude, le paramètre d'hétéroscédasticité est défini pour k matrices de variances-covariances $\mathbf{\Sigma}_i$ (i =1,… k) par (Glèlè Kakaï et Palm, 2006) :

$$\Gamma = -\sum_{i=1}^{k} \ln (|\mathbf{\Sigma}_i| / |\mathbf{\Sigma}|), \qquad (2.4.1)$$

où $\mathbf{\Sigma}$ est la matrice de variances-covariances groupée des populations.

Mais il est à noter que la règle linéaire présente encore de bonnes performances en cas d'hétéroscédasticité modérée (Glèlè Kakaï et Palm, 2005).

2.4.3. Non-normalité et Hétéroscédasticité

L'effet du non-respect des deux conditions (normalité et égalité des variances) sur la précision des résultats du test t d'égalité des moyennes est illustré en considérant que la population P_3 est normale de moyenne 3 et d'écart-type 1 alors que P_4 suit une distribution Chi-carré à 2 degrés de liberté. Les populations P_1 et P_2 sont conservées et constituent la situation souhaitée de normalité et d'égalité des variances. On peut noter que les populations P_3 et P_4 ne remplissent ni la condition de normalité, ni la condition d'égalité des variances (homoscédasticité). Les fonctions de densité de probabilité de ces deux couples de populations sont présentées à la figure 5.

Comme au paragraphe 2.4.1, la puissance du test t d'égalité de deux moyennes est calculée pour différentes tailles d'échantillons dans les deux cas et les résultats sont présentés sous forme graphique à la figure 6.

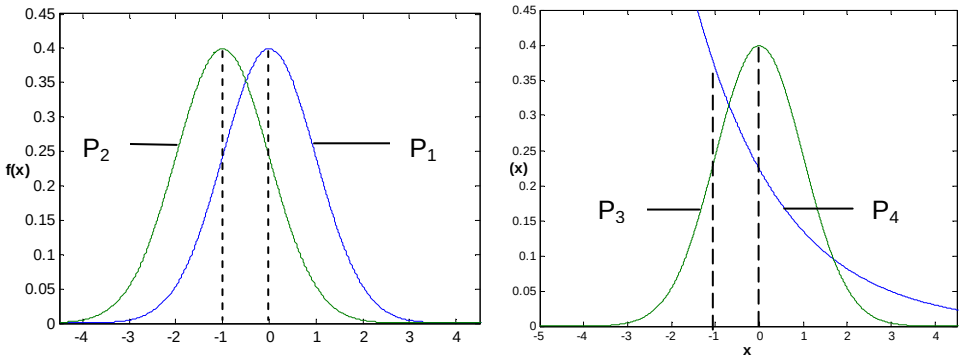


Figure 5. Fonctions de densité de probabilité des distributions considérées dans le cas de non-normalité et d'inégalité des variances.

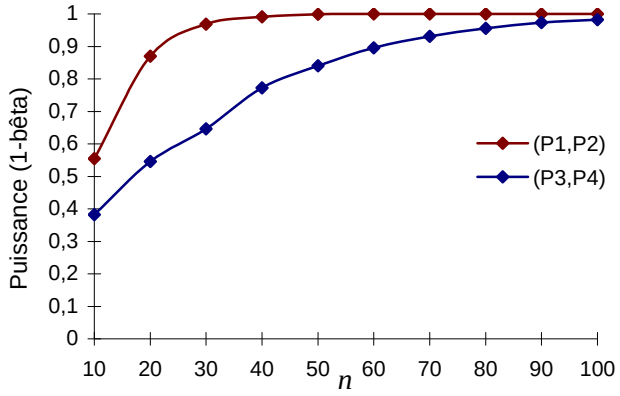


Figure 6. Non-normalité et hétéroscédasticité : puissance du test d'égalité des moyennes pour les couples de populations (P_1,P_2) et (P_3,P_4).

On note de cette figure que, de façon générale, la puissance du test d'égalité des moyennes augmente ici aussi avec l'accroissement de la taille des échantillons. Par ailleurs, lorsque les deux conditions sont remplies (populations P_1 et P_2), la valeur maximale de la puissance est vite atteinte, notamment à partir de la taille n égale à 50. En situation d'inégalité des variances, la valeur maximale de la puissance 1-β n'est pas atteinte même avec la taille maximale considérée qui est ici de 100.

De plus, pour une même taille d'échantillons n, la valeur de puissance du

test en cas d'égalité des variances est sensiblement plus élevée que dans le cas du non-respect de cette condition, la différence de valeur de puissance allant jusqu'à 0,34 pour certaines valeurs de n. Néanmoins, les conséquences du non-respect de ces deux conditions sur la précision des résultats du test t de Student semblent moins importantes que le seul non-respect de la normalité du moins pour l'exemple considéré.

2.5. Alternatives au non-respect des conditions d'application

La décision à prendre en cas de non-respect des conditions d'application d'une méthode statistique paramétrique pressentie pour des données collectées n'est pas aussi aisée. Le problème est très vaste et il n'y a pas de réponse toute faite. Nous proposons dans ce paragraphe des pistes de solutions à l'utilisateur des méthodes statistiques pour une décision conséquente.

2.5.1. Les transformations de variables

L'une des possibilités de traitement des données en cas de non-respect des conditions d'application est la transformation de variable. A ce titre, nous conseillons ici quelques familles de transformations de variable à savoir, la transformation de Box et Cox et la transformation angulaire.
Considérons une variable aléatoire X ne suivant pas une distribution normale. Soit Y la variable transformée de X dans le but de la rendre normale, la transformation de Box et Cox s'écrit :

$$Y = \begin{cases} (X^{\lambda}-1)/\lambda & \text{si } \lambda \neq 0 \\ LnX & \text{si } \lambda = 0 \end{cases} . \qquad 2.5.1.$$

Dans la formule ci-dessus, λ est une constante quelconque et peut être calculée lorsque les moyennes et variances des différents échantillons sont liées par une fonction puissance :

$$\sigma_X^2 = km_X^{\beta} , \qquad 2.5.2$$

avec $\beta = 2(1-\lambda)$ ou $\lambda = 1-\beta/2$.

La relation (2.5.2) peut être établie au moyen d'une transformation logarithmique suivie d'une régression linéaire. Les cas particuliers de la transformation de Box et Cox sont les transformations courantes à savoir la transformation logarithmique et la transformation racine-carré.

La transformation angulaire s'écrit :

$$Y = 2\arcsin\sqrt{X} ,$$

X pouvant aller de 0 à 1.

Cette transformation s'applique aux variables binomiales.

2.5.2. Les tests non paramétriques

Les méthodes statistiques non paramétriques sont conseillées en cas de non-respect des conditions d'applications malgré l'application de transformation linéaires. Nous présentons au tableau 4 une synthèse des méthodes non paramétriques.

Tableau 4. Tests non paramétriques correspondant aux tests paramétriques courants.

Test paramétrique	**Test non-paramétrique correspondant**	**Observations par rapport au test non paramétrique**
Test t à 1 échantillon	Test de Wilcoxon	Comparaison d'une médiane à une valeur connue (données de rangs)
Test t à 2 échantillons	Test de Mann-Whitney	Comparaison de deux médianes (données de rangs)
ANOVA à 1 critère	Test de Kruskal-Wallis	Comparaison de deux ou plusieurs médianes
ANOVA à 1 critère	Test de la médiane de Mood	Comparaison de deux ou plusieurs médianes
ANOVA à 2 critères	Tests de Friedman	Données appariées
Régression linéaire	Régression linéaire pondérée Régression non linéaire	
Analyse discriminante linéaire et quadratique	Analyse discriminante logistique ; Méthodes du noyau ; méthodes neuronales etc.	

On note du tableau 4 les méthodes statistiques non paramétriques relatives aux comparaisons de médiane à 1, 2 ou plus de deux échantillons ainsi que les méthodes multivariées non paramétriques. En ce qui concerne l'analyse de la variance à 1 critère, nous avons présenté deux méthodes à savoir le test de Kruskal-Wallis et le test de Mood qui est résistant vis-à-vis des valeurs aberrantes et des erreurs de données. Il est particulièrement adapté aux étapes préliminaires de l'analyse. Le test de la médiane de Mood est plus résistant que le test de Kruskal-Wallis vis-à-vis des valeurs aberrantes, mais il est moins puissant pour des données provenant de nombreuses distributions, y compris la loi normale (Minitab, 1996).

3. Tests d'hypothèses pour la vérification des conditions d'application

3.1. Introduction

Les pages antérieures ont abordé les conditions d'application des méthodes statistiques paramétriques courantes, leurs importances ainsi que les conséquences du non-respect des conditions. Ce chapitre présente le principe des méthodes de vérification de ces conditions ainsi que leur application dans les logiciels statistiques. Pour une présentation simple et claire de ce chapitre, nous commençons d'abord par présenter au tableau 5, les principales conditions d'application des méthodes statistiques, les tests d'hypothèses pour la vérification de ces conditions et la disponibilité de ces tests dans les logiciels statistiques MINITAB, SPSS et SAS.

Tableau 5. Récapitulatif des tests de vérification des conditions d'application et leur disponibilité dans les logiciels statistiques.

Condition d'application	**Tests disponibles**	**Disponibilité du test dans les logiciels Statistiques**		
		MINITAB	SPSS	SAS
Normalité univariée	Test de Ryan-Joiner	x	--	--
	Test de Shapiro-Wilk	--	x	x
	Test de Kolmogorov-Smirnov	x	x	x
	Anderson-Darling	x	--	x
Normalité multivariée	Test de Rao-Ali	--	--	--
	Test de Mardia	--	--	--
Homogénéité des variances	Test F	x	--	--
	Test de Bartlett	x	--	x
	Test de Hartley	--	--	--
	Test de Levene et ses variantes	x	x	x
Homogénéité des résidus de régression	Test de White	--	--	x
	Test de Breusch-Pagan	--	--	x
Egalité des matrices de variances-covariances	Test du rapport de vraisemblance	--	--	x
	Test M de Box	--	x	--

x : disponible dans le logiciel statistique ; -- : non disponible dans le logiciel.
Le test de Mardia, plus précisément le test d'aplatissement multivarié est disponible dans le logiciel SAS en utilisant la procédure CALIS (Proc Calis) avec spécification de l'option Kurtosis.

On note de ce tableau que tous les tests d'hypothèse de vérification des conditions d'application des méthodes statistiques ne sont pas disponibles dans les logiciels statistiques. De plus, les tests d'hypothèse pour la vérification de la condition de normalité multivariée (test de Rao-Ali et test de Mardia) ne sont disponibles dans aucun des logiciels statistiques utilisés dans la présente note. De ce fait, nous avons conçu dans le langage Matlab une procédure reprenant les principes des tests de Rao-Ali et de Mardia, et qui permet de vérifier cette condition de normalité multivariée. Notons par ailleurs que les tests d'hypothèse énumérés au tableau 5 ne sont pas exhaustifs, seuls les tests courants sont présentés.

3.2. Tests de normalité à une dimension

3.2.1. Méthode graphique de vérification de la normalité

Le contrôle de la normalité d'un nombre de séries de données peut se faire par l'examen préalable d'un histogramme de la série de données ou encore d'un diagramme de probabilité en portant en abscisses les observations. Les ordonnées sont déterminées de telle sorte que les fonctions de répartition $\mathrm{F}(x)$ apparaissent comme des droites. Ce type de représentation peut être appliqué aux séries statistiques en portant en abscisse les valeurs observées x_i classées par ordre croissant et en ordonnées les quantités :

$$N'(x_i)=(i-1/2)/n \quad , \tag{3.2.1}$$

avec $N'(x_i)$ la fréquence relative cumulée de l'observation i et n le nombre d'observations. La droite obtenue est parfois appelée droite de Henry.

Mais lorsqu'on souhaite utiliser en ordonnées, une échelle de quantiles de la variable normale réduite, on calcule des quantiles par la fonction inverse de la fonction de répartition $\varphi(u)$ de la distribution normale réduite, soit :

$$u_i=\varphi^{-1}[N'(x_i)] = \varphi^{-1}[(i-1/2)/n]\,. \tag{3.2.2}$$

Les quantités $(i-1/2)/n$ sont souvent remplacées par des valeurs $N'(x_i)=(i-3/8)/(n+1/4)$ qui permettent notamment d'obtenir, à partir des diagrammes de probabilités, des estimations plus correctes des écarts-types des populations considérées. Dans ce cas, on a :

$$u'_i=\varphi[(i-3/8)/(n+1/4)]\,. \tag{3.2.3}$$

Les valeurs u_i et u'_i ainsi définies sont généralement appelées

quantiles normaux ou scores normaux[1]. Notons que les valeurs u'_i de l'expression (3.2.3) peuvent être obtenues de façon automatique par la commande *Nscores* du logiciel Minitab (Minitab, 1996). Lorsque les observations sont tirées de populations normales, La relation entre les observations et leurs scores normaux est de type linéaire et la droite correspondante est appelée droite de HENRY.

Considérons des données relatives à la densité en *Acacia auriculiformis* de peuplements mélangés, présentées au tableau 6 (Fonton et al., 2002).

Tableau 6. Densité en pieds d'Acacia de peuplements mélangés

Peuplement	Densité
92/01	450
92/04	333
92/22	546
91/02	508
91/03	353
91/05	743
91/16	523
91/17	455
91/22	294
90/03	97
90/07	600
90/09	764

Nous allons faire un examen préalable de la normalité de cette série de données. Puisque les échantillons de densité de ces 3 peuplements ont des moyennes différentes, un test de normalité effectué sur une telle série de données aboutirait à de fausse conclusion. En effet, la différence entre les moyennes des 3 échantillons pourrait conduire à une non-normalité apparente. De ce fait, la normalité pourrait être examinée pour chacun des trois échantillons. On peut aussi réaliser le test de normalité sur l'ensemble des observations des trois échantillons mais en prenant soin de centrer chaque observation par la moyenne de son groupe ou peuplement d'appartenance afin d'éliminer la non-normalité apparente. Cette dernière option est choisie et les scores normaux relatifs aux observations centrées sont calculés en utilisant la formule (3.2.3) et présentés au tableau 7.

[1] En anglais : Normal score.

Tableau 7. Scores normaux des densités de pieds d'Acacia.

Peuplement	Densité en pieds d'Acacia	Données centrées	Scores normaux
92/01	450	7,0	-0,102
92/04	333	-110,0	-0,536
92/22	546	103,0	0,536
91/02	508	28,7	0,102
91/03	353	-126,3	-0,792
91/05	743	263,7	1,114
91/16	523	43,7	0,312
91/17	455	-24,3	-0,312
91/22	294	-185,3	-1,114
90/03	97	-390,0	-1,635
90/07	600	113,0	0,792
90/09	764	277,0	1,635

Les données centrées ainsi que les scores normaux présentés au tableau 7 ont permis d'établir le diagramme de probabilité présenté à la figure 7.

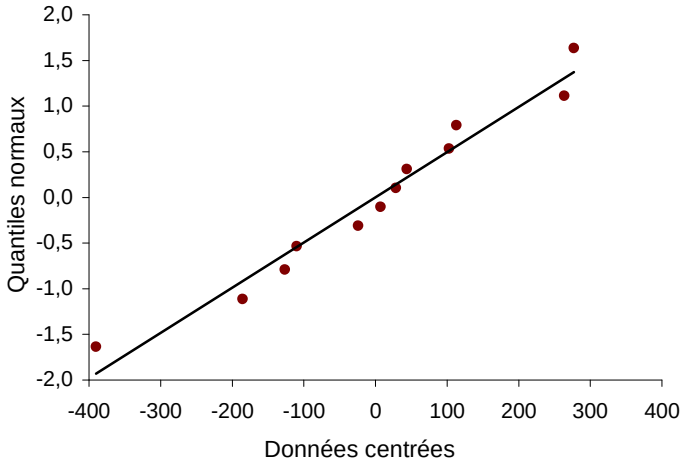


Figure 7. Densité des peuplements mélangés en Acacia : diagramme de probabilité.

Cette figure nous donne déjà une idée de la normalité de la série de données. En effet, les densités centrées des peuplements, représentées par les points sur la figure ne s'écartent pas trop de la droite de HENRY. On peut donc conclure à une normalité de la série d'observations sur base de l'examen graphique.

3.2.2. Méthodes paramétriques du test de normalité

Différents tests de normalité d'une série de données peuvent être considérés comme découlant plus ou moins directement des diagrammes de probabilité. Ils mesurent en fait le degré de linéarité des observations. Nous présentons ici les deux catégories de tests les plus utilisés à savoir les tests basés sur la distribution théorique de la série de données[1] et les tests basés sur la distribution empirique des séries de données[2]. Dans la première catégorie, l'un des tests les plus puissants et utilisés est le test de Shapiro-Wilk que nous présenterons dans la suite. Il sera suivi du test de Ryan-Joiner. Dans la deuxième catégorie de tests, nous pouvons citer les tests de Kolmogorov-Smirnov, de Cramer Von Miss et le test d'Anderson-Darling. Nous ne présenterons dans cette catégorie que le test de Kolmogorov-Smirnov.

3.2.2.1. Test de normalité de Shapiro-Wilk[3]

Le test de normalité de Shapiro-Wilk nécessite le calcul de la statistique W_{obs} dont l'expression est :

$$W_{obs}=\frac{\left[\sum_{i=1}^{n}(c_i\, x_i)\right]^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}. \qquad (3.2.4)$$

Les coefficients c_i sont donnés sous forme de tables fournies entre autres par Shapiro et Wilk (1965). Le symbole n représente le nombre d'observations et $\overline{x}$, la moyenne des observations x_i ($i = 1,\ldots,n$).

Le rejet de l'hypothèse de normalité intervient quand :

$$P(W \leq W_{obs}) \leq \alpha$$

La méthode de détermination des valeurs de probabilités p (la probabilité d'obtenir une valeur de W_n inférieure ou égale à W_{obs}) est assez complexe. Elle est fonction de l'effectif n des échantillons. Lorsque n =3, la distribution de probabilité de W_n est connue et est utilisée pour déterminer la probabilité. Pour n > 4, la transformation suivante est utilisée (SAS, 1999) :

---

[1] En anglais : Exact distribution function.
[2] En anglais : Empirical Distribution Function.
[3] En anglais : Shapiro-Wilk's test.

$$Z_n = \begin{cases} [\ln(1-W_n)-\mu]/\sigma & \text{si} \quad 4 \le n \le 11 \\ [\ln(1-W_n)-\mu]/\sigma & \text{si} \quad 12 \le n \le 2000 \end{cases} \qquad (3.2.5)$$

Les valeurs σ, γ et μ sont des fonctions de n et sont obtenues par des résultats de simulations. Les valeurs élevées de Z_n indiquent un écart élevé par rapport à la normalité et puisque la statistique Z_n est connue pour avoir une distribution normale, elle est utilisée pour calculer les valeurs de probabilité pour $n > 4$.

Ainsi, les valeurs critiques W_n sont données dans les mêmes tables que celles des coefficients c_i .

Selon Dagnelie (1998), la valeur W_n de Shapiro-Wilk n'est autre qu'un coefficient de détermination des couples (c_i , x_i). Ce coefficient est égal à 1 quand tous les points de coordonnées (c_i , x_i) sont strictement colinéaires ou encore lorsque tous les points de la figure 7 se retrouvent exactement sur la droite de HENRY.

Comme on peut le noter de la formule (3.2.5), le test de Shapiro-Wilk n'est pas conseillé lorsque la taille de l'échantillon considéré est supérieure à 2000. Dans un tel cas, le test de Kolmogorov-Smirnov est préconisé.

3.2.2.2. Test de normalité de Ryan-Joiner[1]

Les coefficients c_i sont comparables aux quantiles normaux u_i ou u'_i relatifs aux diagrammes de probabilités. Ainsi, Ryan et Joiner (1976) proposent de remplacer les valeurs c_i par les quantiles normaux dans l'expression (3.2.4).

La statistique du test de Ryan-Joiner est donc le coefficient de corrélation linéaire entre les données et les scores normaux. C'est le rapport entre la covariance des deux variables et le produit de leurs écarts-types respectifs, soit :

[1] En anglais : Ryan-Joiner's test.

$$\rho_{obs}=\frac{\mathrm{cov}(x,u)}{S_x.S_u}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(u_i-\bar{u})}{[\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum(u_i-\bar{u})^2]^{1/2}} \qquad (3.2.6)$$

La formule de calcul de la statistique, ρ_{th} , de Ryan-Joiner dépend de la taille *n* de l'échantillon considéré et est donnée ici pour un niveau de confiance de 0,95 :

$$\rho_{th} = 1{,}0063-(0{,}1288/\sqrt{n})-(0{,}6118/n)+(1{,}3505/n^2) \quad \text{si} \quad n < 50$$

$$\rho_{th} = 0{,}9995+(0{,}0178/\sqrt{n})-(1{,}7726/n)+(3{,}5582/n^{1{,}5}) \quad \text{si} \quad n > 50.$$

L'hypothèse de normalité est rejetée lorsque ρ_{obs} est inférieur à ρ_{th} .

L'application du test de normalité de Ryan-Joiner aux données du tableau 7 donne une valeur du coefficient de corrélation égale à 0,981. L'effectif des données du tableau 7, qui est de 12, étant inférieur à 50, la première formule de calcul de la statistique de Ryan-Joiner est utilisée et donne une valeur de 0,923. Puisque la valeur observée (0,981) est supérieure à la valeur critique, on accepte l'hypothèse nulle et on conclut au caractère normal des données du tableau 7.

3.2.2.3. Test de normalité de Kolmogorov-Smirnov[1]

Pour toute série d'observations $X_1,...,X_n$ de même fonction de distribution, une fonction de distribution empirique $f(x)$ peut être définie. Sous l'hypothèse nulle, $f(x)$ suit une distribution normale. Supposons que les observations soient classées par ordre croissant tel que : $X_{(1)}, \ldots, X_{(n)}$. La fonction de distribution empirique $f_n(x)$ est définie de la manière suivante :

$$f_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)} \\ \frac{1}{n} & \text{si} \quad X_{(i)} < X_{(i+1)}, \ i = 1,\ldots,n\text{-}1 \\ 1 & \text{si} \quad X_{(n)} \leq x \end{cases}$$

Notons que $f_n(x)$ est une fonction par intervalles qui prend un pas de hauteur $1/n$ à chaque observation. Cette fonction estime la valeur de la distribution $f(x)$.

[1] En anglais: Kolmogorov-Smirnov's test

A chaque valeur x, $f_n(x)$ est la proportion des observations inférieures ou égales à x, alors que $f(x)$ est la probabilité pour qu'une observation soit inférieure ou égale à x. La statistique EDF (Empirical Distribution Function) mesure la dissimilarité entre $f_n(x)$ et $f(x)$.

De façon générale, les tests EDF utilisent la fonction de répartition $U = F(x)$. Si $f(X)$ est la fonction de distribution de X, la variable aléatoire U est uniformément distribuée entre 0 et 1. Soient n observations $X_{(1)}$, ... ,$X_{(n)}$, les valeurs $U_{(i)} = f(X_{(i)})$ sont calculées.

La statistique D du test de Kolmogorov-Smirnov est définie de la façon suivante :

$$D = \sup_x \left| F_n(x) - F(x) \right| \qquad (3.2.7)$$

La statistique de Kolmogorov-Smirnov est basée sur la plus grande différence verticale entre $F(x)$ et $F_n(x)$. Elle est calculée en considérant le maximum de D^+ et D^-, où D^+ est la plus grande distance verticale entre la fonction de distribution empirique et la fonction exacte de distribution lorsque EDF est supérieure à la fonction de distribution, et D^- est la distance verticale la plus grande lorsque EDF est plus petite que la fonction de distribution.

$$D^+ = \max_i \left(\frac{i}{n} - U_{(i)}\right) ;$$

$$D^- = \max_i \left(U_{(i)} - \frac{i-1}{n}\right) ;$$

$$D = \max\left(D^+, D^-\right).$$

3.2.3. Application avec les logiciels statistiques

Dans ce paragraphe nous présentons l'application des tests de normalité dans les logiciels statistiques Minitab, SPSS et SAS. Nous séparons les tests de normalité à une dimension des tests de normalité à plusieurs dimensions.

3.2.3.1. Logiciel Minitab

Afin de mieux expliquer la procédure de réalisation du test de normalité à une dimension avec le logiciel statistique Minitab, nous reprenons les données du tableau 7. Le test de normalité appliqué à ces données s'exécute en sélectionnant « **Stat > Statistiques élémentaires > Test de normalité** » comme le montre la figure 8. Aussitôt sélectionné, la boîte de dialogue de la figure 9 s'affiche.

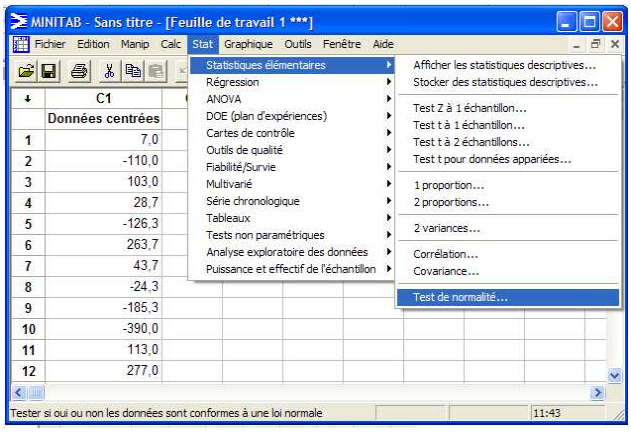


Figure 8. Procédure d'exécution du test de normalité à une dimension avec Minitab.

Dans cette boîte de dialogue (figure 9), la variable **données centrées** du tableau 7 est mise dans la fenêtre « **Variable** ». Dans la fenêtre « **Probabilités de référence** » on peut de manière facultative insérer une colonne contenant des probabilités à insérer sur le graphique de la courbe normale. Les valeurs contenues dans cette colonne doivent être comprises entre 0 et 1. La fenêtre « **Titre** » permet de donner un titre au graphique, mais ceci est aussi facultatif.

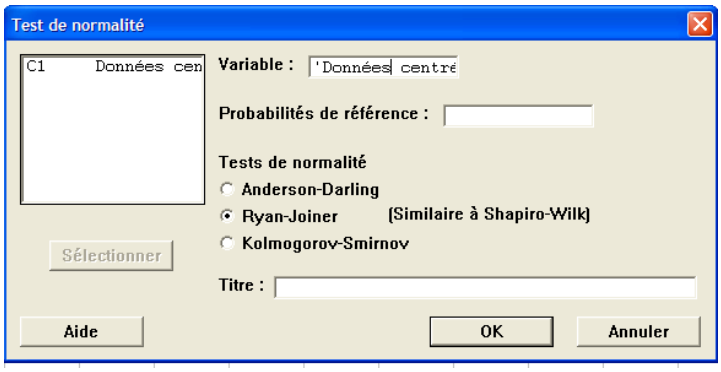


Figure 9. Procédure d'exécution du test de normalité à une dimension avec Minitab : boîte de dialogue 1.

De plus, la figure 9 montre que le logiciel Minitab permet de réaliser trois types de test de normalité, à savoir le test de Anderson-Darling, le test de Ryan-Joiner et le test de Kolmogorov-Smirnov. Comme indiqué dans le

paragraphe 3.2.2, les tests de Anderson-Darling et de Kolmogorov-Smirnov sont fondés sur un test EDF (fonction de répartition empirique), alors que celui de Ryan-Joiner se base sur un test de corrélation. Le test de Ryan-Joiner a été choisi (figure 9). On obtient les résultats de la figure 10.

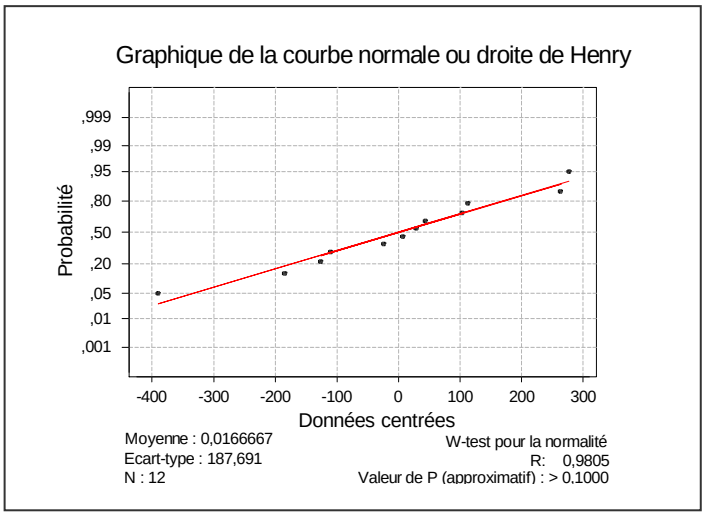


Figure 10. Densité des peuplements mélangés en Acacia : résultats du test de normalité avec Minitab.

Le graphique obtenu est le même que celui de la figure 7. En bas et à gauche du graphique, Minitab donne la statistique liée au test de Ryan-Joiner qui est égale à 0,981, la même valeur que celle obtenue au paragraphe 3.2.2.2. La valeur de la probabilité (Prob > 0,05) permet d'accepter l'hypothèse nulle de normalité des données du tableau 7. On peut reprendre les mêmes procédures en choisissant le test de Anderson-Darling ou de Kolmogorov-Smirnov.

3.2.3.2. Logiciel SPSS

Le logiciel SPSS propose deux tests de normalité à savoir le test de Kolmogorov-Smirnov et celui de Shapiro-Wilk. Il est utile de noter que la version 9 du logiciel SPSS réalise le test de Shapiro-Wilk uniquement pour des tailles d'échantillon inférieures à 50. Par contre, la version 10 du logiciel donne les résultats du même test quelle que soit la taille de l'échantillon.

Le test de normalité est réalisé en sélectionnant « **Analyse > Statistiques descriptives > Explorer… »**, comme le montre la figure 11. La boîte de dialogue de la figure 12 s'affiche. On sélectionne la variable **Densite** dans la fenêtre « **Dependent List** ».

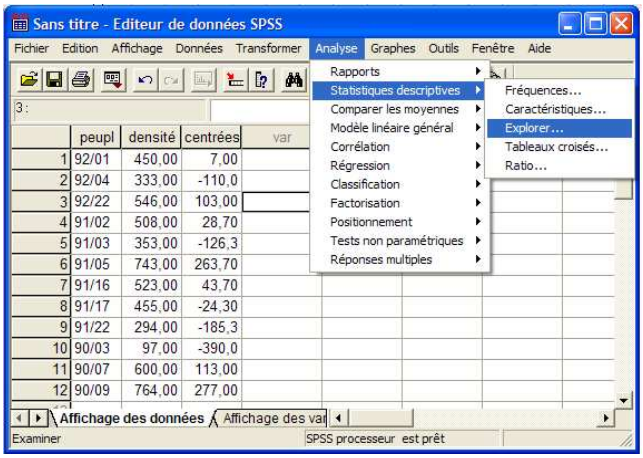


Figure 11. Procédure d'exécution du test de normalité à une dimension avec SPSS.

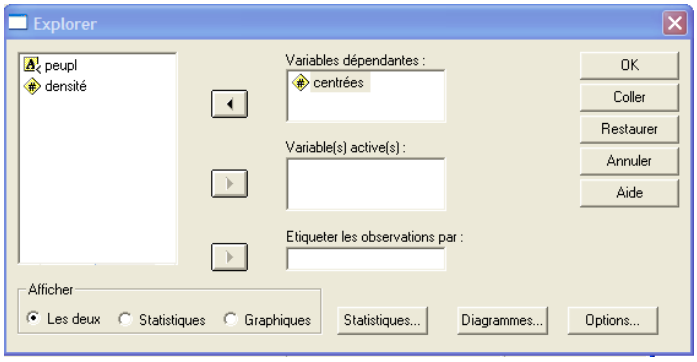


Figure 12. Procédure d'exécution du test de normalité à une dimension avec SPSS : boîte de dialogue 1.

Pour réaliser le test de normalité, il faut cliquer sur le bouton « **diagramme** », puis cocher « **Graphes de répartition gaussiens avec tests** » (cf. figure 13).

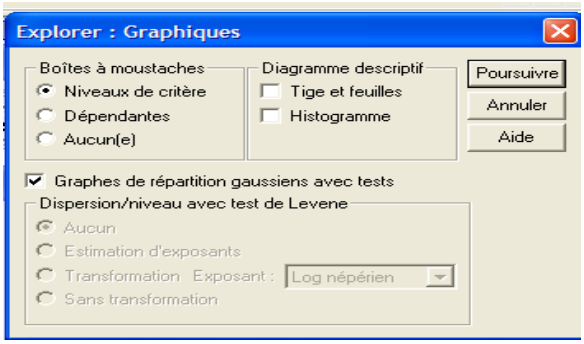


Figure 13. Procédure d'exécution du test de normalité à une dimension avec SPSS : boîte de dialogue 2.

On obtient les résultats de la figure 14. La droite de HENRY obtenue est la même que celle obtenue avec le logiciel Minitab. De plus, les résultats des deux tests de normalité permettent d'accepter l'hypothèse nulle de normalité de la population dont sont issues les données.

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistique	ddl	Signification	Statistique	ddl	Signification
CENTRÉES	,115	12	,200*	,965	12	,846

*. Il s'agit d'une borne inférieure de la signification réelle.

a. Correction de signification de Lilliefors

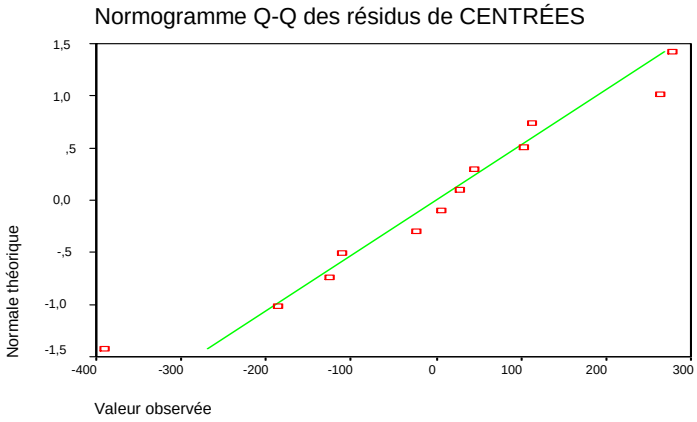


Figure 14. Densité des peuplements mélangés en Acacia: résultats du test de normalité avec SPSS.

Il est à noter que, dans le cas de la régression linéaire, le test de normalité est

intégré à la procédure de régression dans les logiciels Minitab et SPSS. En Minitab par exemple, la procédure d'exécution de la régression est illustrée à la figure 15.

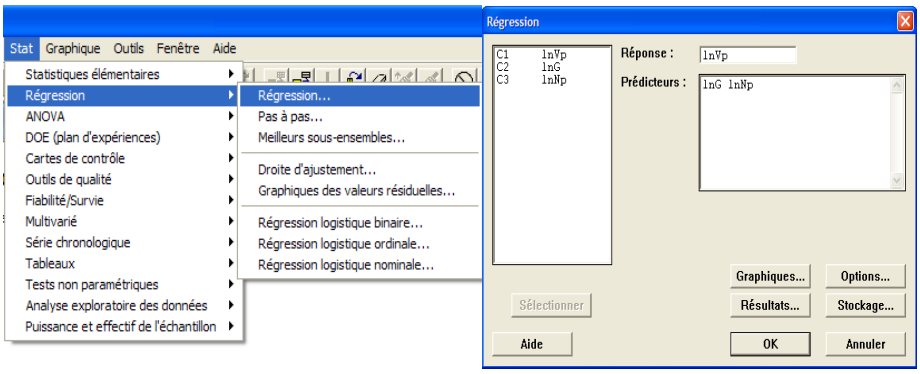


Figure 15. Procédure d'exécution de la régression avec Minitab.

On peut noter de cette figure que la régression est exécutée en sélectionnant « ***Stat > Régression > Régression… »***. On obtient la boîte de dialogue située du côté droit de la figure. Dans cette boîte de dialogue, la variable dépendante est mise dans la fenêtre « **Réponse** » alors que les variables indépendantes ou explicatives sont introduites dans la fenêtre « **Prédicteurs** ». Pour afficher les résultats de l'examen de la normalité des résidus dans les résultats de la régression linéaire, on clique sur la commande « **Graphiques** » de la boîte de dialogue de la figure 15. Dans le logiciel SPSS, on sélectionne « **Analyse > Regression > Linéaire… »**. Ensuite, on sélectionne la commande « **Diagrammes…** » et on coche « **Diagramme P-P gaussien** » dans la boîte de dialogue qui s'affiche (figure 16).

Figure 16. Procédure d'exécution du test de normalité des résidus de régression avec SPSS.

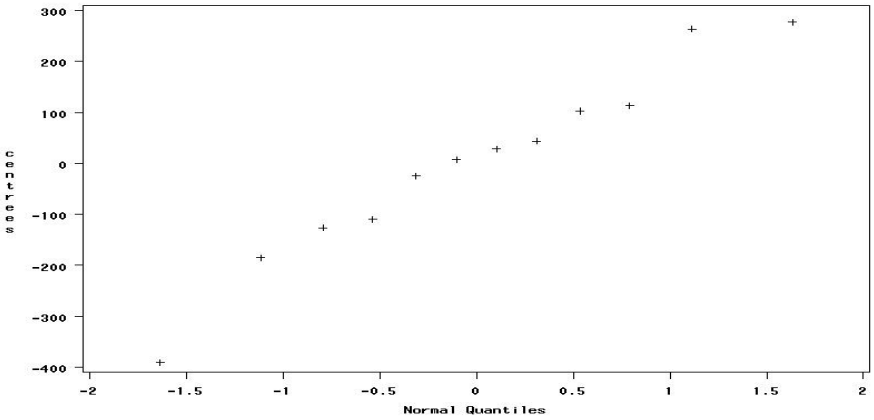
3.2.3.3. Logiciel SAS

Des tests de normalité à une dimension peuvent aussi être réalisés avec le logiciel statistique SAS. Ce dernier propose quatre types de test à savoir, les tests de Shapiro-Wilk, de Kolmogorov-Smirnov, de Cramer-von Mises et de Anderson-Darling. La réalisation de ces tests se fait à travers la procédure « **univariate** », avec utilisation de l'option « **normal** » ou « **normaltest** ». L'utilisation de l'option « **qqplot** » dans la procédure permet d'obtenir le diagramme de probabilité. Pour illustrer la procédure de réalisation des tests de normalité dans SAS, reprenons les données du tableau 7. La figure 17 donne la procédure utilisée à ce propos.

```
DATA dens;
Input centrees;
Cards;
 7
-110
 103
 28.7
-126.3
 263.7
 43.7
-24.3
-185.3
-390
 113
 277
;
Proc univariate  Normaltest;
Qqplot centrees/Normal;
Run;
```

Figure 17. Programme SAS pour la réalisation du test de normalité sur les données du tableau 7.

On note de la figure 17 que le logiciel SAS donne, aux arrondis près, les mêmes résultats que SPSS et Minitab. Contrairement à ce qui est observé avec les logiciels SPSS et Minitab, les axes du graphique produit par le logiciel SAS sont inversés (cf. figures 10, 14 et 18). Dans tous les cas, la normalité des données est acceptée.



Tests for Normality

Test		--Statistic---		-----p Value
Shapiro-Wilk	W	0.964567	Pr < W	0.8465
Kolmogorov-Smirnov	D	0.115125	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.027879	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.210303	Pr > A-Sq	>0.2500

Figure 18. Densité des peuplements mélangés en Acacia : résultats du test de normalité avec SAS.

3.3. Tests de normalité à plusieurs dimensions

Nous abordons essentiellement deux tests de normalité multivariés à savoir le test de Rao-Ali et le test de Mardia.

3.3.1. Le test de Rao-Ali

Supposons que l'on veuille comparer deux types de pâturage sur la base de leurs poids en graminée et légumineuse (tableau 8). Une manière judicieuse de faire une telle comparaison est d'appliquer l'analyse de la variance multivariée qui prend en une fois les deux variables considérées. Une des conditions d'application de ce test est le caractère multinomial des données. Nous proposons d'utiliser la méthode de Rao et Ali pour ce faire. Ces deux auteurs proposent de transformer toutes les variables dont on veut vérifier le caractère normal en une seule variable.

Tableau 8. Poids en graminées et légumineuses de divers types de pâturage.

Type Pâture	Graminées	Légumineuses
1	120	315
1	450	30
1	757	0
1	212	120
1	185	244
1	451	52
2	598	164
2	599	1203
2	0	219
2	5855	2616
2	5520	211
2	2540	696

De façon générale, considérons un échantillon global, représentatif de g groupes, de nombre p de variables (dans le cas présent, la valeur de g est de 2 et le nombre de variables est de 2). L'effectif global, N, des données est obtenu par la formule :

$$N = \sum_{j=1}^{g} n_j ,$$

n_j, étant l'effectif du groupe j. Dans le cas présent, la valeur de N est de 12.

Les différentes étapes de la procédure de Rao et Ali sont les suivantes :

- chaque observation i est centrée et réduite par le vecteur de moyennes et la racine-carrée de la matrice de variances-covariances de son groupe d'appartenance. Ainsi, toute observation i du groupe k (k=1,…, g), représentée par le vecteur $\boldsymbol{x}_{ki}$, est centrée et réduite, respectivement par le vecteur de moyennes, $\overline{\boldsymbol{x}}_k$ et la racine-carrée de la matrice de variances-covariances, $\hat{\boldsymbol{\Sigma}}_k^{1/2}$, du groupe k pour obtenir un nouveau vecteur d'observations $\boldsymbol{y}_{ki}$,

$$\boldsymbol{y}_{ki} = \hat{\boldsymbol{\Sigma}}_k^{-1/2}(\boldsymbol{x}_{ki} - \overline{\boldsymbol{x}}_k) \ ;$$

- toutes les pN composantes de tous les vecteurs $\boldsymbol{y}_{ki}$ (i = 1,… N, k=1,…, g) de tous les g groupes sont considérées comme des observations d'un nouvel échantillon univarié, d'effectif pN, sur lequel tout test de normalité univarié peut être appliqué (tests de Ryan-Joiner, d'Anderson-Darling, de Kolmogorov-Smirnov, etc.).

L'application de la méthode de Rao et Ali aux données du tableau 8, avec utilisation du test de Ryan-Joiner donne une valeur du coefficient de corrélation, ρ_{obs}, égale à 0,898. Cette valeur observée étant inférieure à la valeur critique, ρ_{th}, de Ryan-Joiner (0,928), on rejette l'hypothèse de multinormalité des données.

3.3.2. Le test de Mardia[1]

Ce test est encore appelé tests des coefficients de symétrie et d'aplatissement de Mardia. Soient $\mathbf{X}_1$, ..., $\mathbf{X}_n$ un échantillon aléatoire composé de *n* vecteurs-lignes $\mathbf{X}$. Soit p le nombre de variables de chaque vecteur-ligne, n, le nombre d'observations et $\boldsymbol{\mu}$, le vecteur de moyennes.

Une mesure de la symétrie multivariée[2] de cette série de vecteurs est donnée par l'expression canonique (Mardia, 1980) :

[1] En anglais : Mardia's test.
[2] En anglais : Multivariate skewness

$$b_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} [(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{S}^{-1}(\mathbf{X}_j - \boldsymbol{\mu})]^3 \qquad (3.3.1)$$

Dans l'expression (3.3.1), $\mathbf{S}$ est la matrice de variances-covariances de l'échantillon.

Mardia (1980) a montré que sous l'hypothèse nulle de multinormalité, $(n/6)b_1$ a asymptotiquement une distribution Chi-Carré à $p(p+1)(p+2)/6$ degrés de liberté.

Une mesure de l'aplatissement multivarié[1] est donnée par l'expression canonique (Mardia, 1980) :

$$b_2 = \frac{1}{n}\sum_{i=1}^{n} [(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{S}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})]^2 \qquad (3.3.2)$$

Sous l'hypothèse nulle de multinormalité, b_2 suit asymptotiquement une distribution normale de moyenne $p(p+2)$ et de variance $8p(p+2)/n$ (Mardia, 1980).

Dans le cas des données du tableau 8, la valeur de p est égale à 2 et n est égale à 12. La moyenne et la matrice de variances-covariances de l'échantillon donnent respectivement :

$$\mathbf{S} = \begin{pmatrix} 4371700 & 97790 \\ 97790 & 56360 \end{pmatrix} \quad ; \quad \boldsymbol{\mu} = [\,1440{,}6\,;\,489{,}2\,]\,.$$

On a : $b_1 = 6{,}249$ et $b_2 = 11{,}495$.

Pour le test de symétrie multivariée, la valeur $\chi^2_{obs} = (n/6)b_1$ donne 12,499. La probabilité associée à cette valeur, suivant la distribution χ^2 à 4 degrés de liberté ($p(p+1)(p+2)/6 = 4$) est égale à 0,014. On rejette donc l'hypothèse de symétrie à deux dimensions des données du tableau 8.

Pour le test d'aplatissement multivarié, la valeur U_{obs} est de 11,495 et la probabilité associée à cette valeur suivant la distribution normale de moyenne 8

[1] En anglais : Multivariate kurtosis.

($p(p+2)$ =8) et de variance 5,33 ($8p(p+2)/n$ =5,33) est égale à 0,065. De ce fait, on accepte au seuil de 5 % l'hypothèse d'aplatissement normal des données du tableau 8.

En conclusion, l'hypothèse de symétrie multivariée étant rejetée, les données du tableau 8 ne proviennent pas de populations multinormales malgré l'acceptation de l'hypothèse d'aplatissement normal.

3.3.3. Application avec le langage Matlab

3.3.3.1. Conception d'une Fonction « **normalite** » dans le langage Matlab

Le test de normalité à plusieurs dimensions n'est pas disponible dans les logiciels statistiques SAS, Minitab et SPSS. Pour permettre à l'utilisateur de vérifier cette condition de multinormalité des populations, nous avons conçu une fonction dans le langage Matlab appelée « **normalite** ». Pour des informations sur le langage Matlab, consulter la note de Akossou et al. (2001). La fonction **normalite** conçue dans le langage Matlab peut être utilisée pour exécuter le test de Mardia et le test de Rao-Ali associé au test de Ryan-Joiner (cf. paragraphes 3.3.1 et 3.3.2). La figure 19 présente le programme conçu.

```
function [Test1,R1,note,NB1,Test2,R2,Note,NB2]=normalite(X)

      % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
      % Fonction MATLAB permettant la réalisation des tests   %
      % de normalité multivariée de Ryan-Joiner et de Rao-Ali %
      % X est la matrice de données dont on veut tester la    %
      % normalité. Les sorties de la fonction donnent les     %
      % résultats des tests.                                  %
      % Auteur: GLELE K. Romain. Date: 04/05/2006.            %
      % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
Test1=['                    TEST DE MARDIA'];
Test2=['                    TEST DE RAO-ALI'];
[n,p]=size(X); c=cov(X);mu=mean(X);ic=inv(c);
for i=1:n
   for j=1:n
   bo1(i,j)=(X(i,:)-mu)*ic*(X(j,:)-mu)';
   end
   bo2(i,1)=(X(i,:)-mu)*ic*(X(i,:)-mu)';
end
b1=(1/n^2)*sum(sum(bo1.^3));b2=(1/n)*sum(bo2.^2);
pb1=(n/6)*b1;ddl=p*(p+1)*(p+2)/6;p1=1-
cdf('chi2',pb1,ddl);m=p*(p+2);v=8*p*(p+2)/n;
bp2=(b2-m)/sqrt(v); p2=1-(normcdf(bp2,0,1)); R1=[b1,b2,p1,p2];
note1=['1ère valeur: coefficient de symétrie multivariée
'];
note2=['2ème valeur: coefficient d''aplatissement multivarié
'];
note3=['3ème valeur: probabilité liée à l''hypothèse de symétrie normal
'];
note4=['4ème valeur: probabilité liée à l''hypothèse d''aplatissement
normal      '];
note=[note1;note2;note3;note4];
if p1<=0.05 | p2<=0.05
   NB1=['hypothèse de multinormalité rejetée'];
else
   NB1=['hypothèse de multinormalité acceptée'];
end
s=cov(X);m=mean(X);f=inv(sqrtm(s));yn=(X-repmat(m,n,1))*f;
y=sort(yn(:));wa1=(1:n*p)';Nw1=(wa1-(3/8))/((n*p)+0.25);
xw1=norminv(Nw1,0,1);rw1=corrcoef(y,xw1);
rw1=rw1(2,1);rwe(1,1)=rw1;
if n<=50
   rl05=1.0063-(0.1288/sqrt(n))-(0.6118/n)+(1.3505/(n^2));
end
if n>50
   rl05=0.999494+(0.0177805/sqrt(n))-(1.77265/n)+(3.55823/(n^1.5));
end
note1=['1ère valeur: coefficient de corrélation au sens de Rao-Ali'];
note2=['2ème valeur: coefficient de corrélation-seuil             '];
Note=[note1;note2];R2=[rwe,rl05];
if rwe<rl05
   NB2=['hypothèse de multinormalité rejetée'];
end
if rwe>=rl05
   NB2=['hypothèse de multinormalité acceptée'];
end
```

Figure 19. Fonction Matlab pour la réalisation de tests de normalité multivariée.

3.3.3.2. Lecture des données dans le langage Matlab

Lorsque la matrice de données à faire lire par le langage Matlab n'est pas très large, on peut la copier directement dans le logiciel comme le montre la figure 20. Il s'agira de copier la matrice de données dans la fenêtre « **MATLAB**» en la mettant entre crochets et en lui affectant un nom. Dans le cas présent, la matrice est notée X. Après cela, on **valide** la lecture des données en appuyant le bouton correspondant du clavier de l'ordinateur.

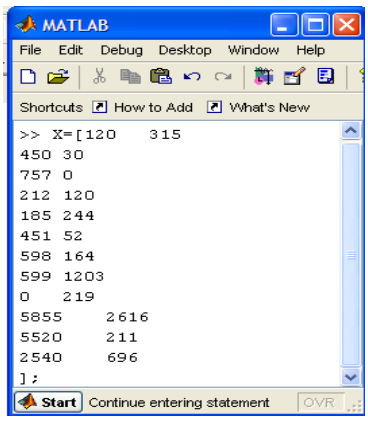


Figure 20. Processus de lecture de la matrice X dans Matlab.

Lorsque la matrice des données a une taille très importante, il est plus aisé de la transférer du logiciel Excel dans le langage Matlab. Il existe en effet, une telle procédure de transfert des données.

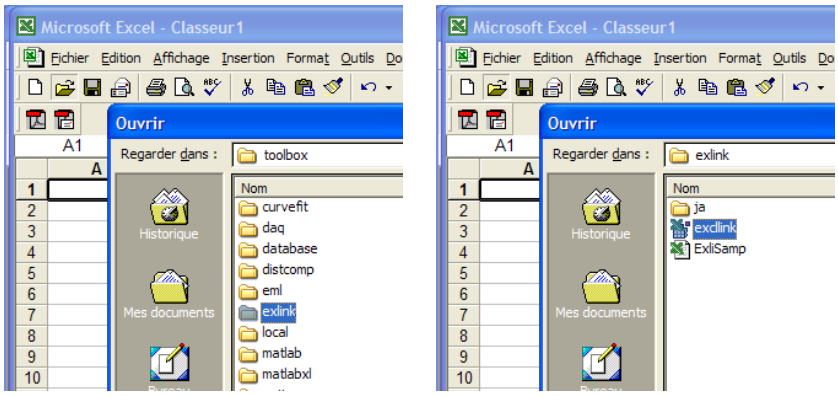


Figure 21. Processus d'activation de la procédure excllink à partir de Excel.

Après avoir saisi les données dans une feuille de calcul Excel, on active la procédure **excllink** du langage Matlab à partir du logiciel Excel. Il s'agira d'ouvrir le fichier excllink qui se trouve dans le répertoire « **C:\Program Files\Matlab\R2006a\Toolbox\exlink\excllink** » si le logiciel Matlab est installé sur un disque dur « C:\ » comme le montre la figure 21. Dans le cas où le logiciel est installé sur un autre disque « D:\ » par exemple, alors le fichier excllink à ouvrir à partir de Excel se trouve dans le répertoire « **D:\Program Files\Matlab\R2006a\Toolbox\exlink\excllink** ».

Lors de l'ouverture du fichier excllink, il peut arriver que l'ordinateur demande s'il faut activer ou désactiver les macros (figure 22). Dans un tel cas, on accepte d'activer les macros.

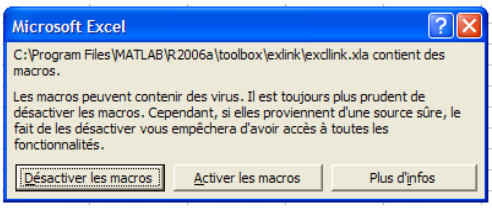


Figure 22. Activation des macros lors de l'ouverture du fichier excllink.

Dès l'ouverture du fichier excllink dans le logiciel Excel, quatre nouveaux menus s'ajoutent à la barre de menu Excel à savoir : **Startmatlab, putmatrix**, **getmatrix** et **evalstring**. Après avoir sélectionné les données[1] à transférer dans le langage Matlab (figure 23), on clique sur le menu **putmatrix** et la boîte de dialogue située dans la partie gauche dans la figure 23 apparaît.

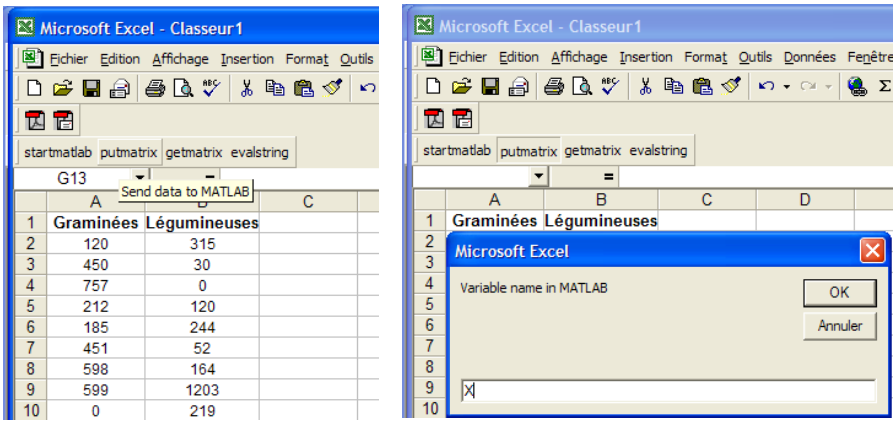


Figure 23. Transfert des données dans le langage Matlab.

[1] Sélectionner uniquement les données et non les données avec les noms des colonnes.

Il faut spécifier un nom pour la matrice des données ainsi transférées dans le langage Matlab. Dans le présent exemple, la matrice de données est notée X. Les données sont ainsi transférées.

3.3.3.3. Enregistrement de la fonction **Normalite** dans Matlab\R2006a\work

Une fois la matrice de données transférée ou copiée dans Matlab, on copie la fonction **normalite** de la figure 19 dans la fenêtre **Editor** et on enregistre le fichier dans le sous-répertoire work du répertoire **R2006a\Matlab** situé dans C:\Program Files (« **Matlab\R2006a\work** ») comme le montre la figure 24.

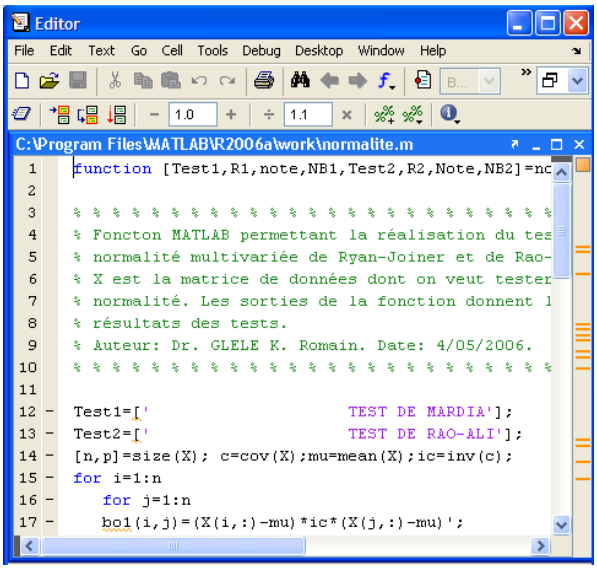


Figure 24. Copie et enregistrement de la fonction **normalite** dans Matlab\work.

3.3.3.4. Exécution de la fonction **Normalite**

On copie dans la fenêtre **Matlab**, le titre de la fonction **Normalite** c'est-à-dire la 1ère ligne de la fonction sans le nom « function » comme le montre la figure 25 et on valide en tapant sur la touche « ↵ » du clavier de l'ordinateur.

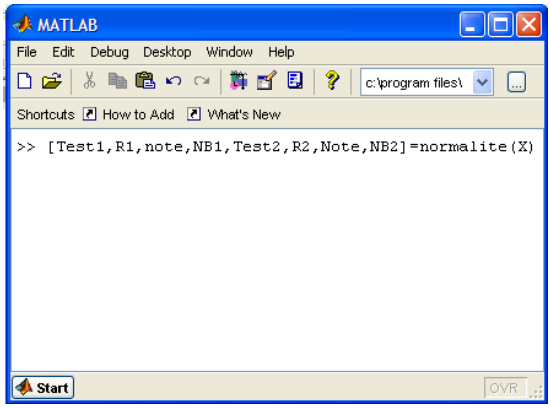


Figure 24. Exécution de la fonction **Normalite**.

Il est à noter que dans le cas où la matrice de données est copiée ou transférée dans le langage Matlab sous un nom différent, par exemple « A » au lieu de « X », l'appel de la fonction se fait en remplaçant « X » par « A » comme le montre la figure 26. Il est aussi utile de noter que le langage Matlab est sensible à la casse des lettres. En d'autres termes, une matrice de données notée « x » (minuscule) est différente d'une matrice de données notée « X » (majuscules). De ce fait, il est important de tenir compte de la casse du nom de la matrice de données copiée ou transférée dans Matlab lors de l'appel de la fonction **normalite**. L'application de la fonction **normalite** aux données du tableau 8 donne les résultats de la figure 27.

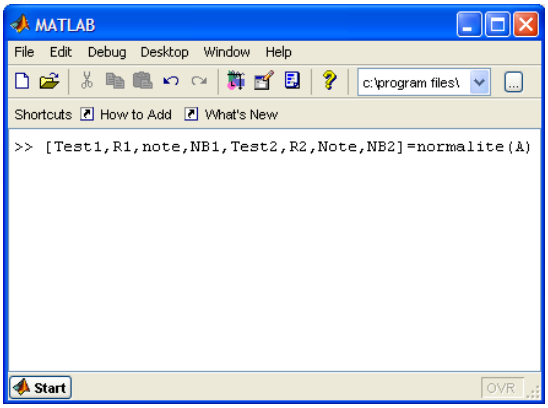


Figure 26. Appel de la fonction **normalite** avec changement de nom de la matrice de données.

```
» [Test1,R1,note,NB1,Test2,R2,Note,NB2]=normalite(X)

Test1 =
                    TEST DE MARDIA

R1 =
    6.2499   11.4949    0.0140    0.0651

note =
1ère valeur: coefficient de symétrie multivariée
2ème valeur: coefficient d'aplatissement multivarié
3ème valeur: probabilité liée à l'hypothèse de symétrie normal
4ème valeur: probabilité liée à l'hypothèse d'aplatissement
normal

NB1 =
hypothèse de multinormalité rejetée

Test2 =
                    TEST DE RAO-ALI

R2 =
    0.8985    0.9275

Note =
1ère valeur: coefficient de corrélation au sens de Rao-Ali
2ème valeur: coefficient de corrélation-seuil

NB2 =
hypothèse de multinormalité rejetée
```

Figure 27. Sortie du langage Matlab après exécution de la fonction **normalite** sur les données du tableau 8.

Ces résultats indiquent que les probabilités liées aux tests de symétrie et d'aplatissement multivariés de Mardia sont respectivement égales à 0,014 et 0,0651. En d'autres termes, l'hypothèse de normalité multivariée est rejetée. Ces résultats montrent également que le test de Rao-Ali et Ryan-Joiner effectué donne une corrélation des observations avec les scores normaux égale à 0,899. La corrélation-seuil étant de 0,9275, ce test indique aussi le rejet de l'hypothèse nulle de multinormalité. De ce fait, les données de la matrice X (tableau 8) ne proviennent pas de populations multinormales.

3.4. Tests d'homoscédasticité à une dimension

3.4.1. Tests d'égalité des variances

3.4.1.1. Comparaison de deux populations

- *Echantillons indépendants*

Supposons que l'on veuille comparer les peuplements de 1991 et 1992 du tableau 8, du point de vue du diamètre moyen des arbres des placettes échantillonnées. Pour ce faire, nous devons vérifier au préalable l'hypothèse d'égalité des variances diamétriques des deux peuplements. Les données relatives à ces deux peuplements sont présentées au tableau 9.

Tableau 9. Diamètre moyen (cm) des peuplements de 1991 et 1992.

Diamètre moyen des peuplements de 1991	Diamètre moyen des peuplements de 1992
11,09	14,60
11,29	13,01
13,40	13,67
9,55	12,68
10,15	13,13
14,44	11,81
15,24	12,12
16,28	-
14,63	-
12,00	-
10,87	-
14,04	-
12,71	-
12,96	-
14,06	-
13,71	-
11,60	-
11,62	-
12,37	-
12,76	-
11,61	-
12,06	-
12,05	-
10,14	-
$\hat{\sigma}_1^2 = 2,889$	$\hat{\sigma}_2^2 = 0,886$
$n_1 = 24$	$n_2 = 7$

On note du tableau 9 que nous disposons de 24 observations relatives aux peuplements de 1991 et 7 observations en ce qui concerne les peuplements de 1992. Les deux séries d'observations sont indépendantes de variances respectives $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$. Puisque le test d'égalité des variances nécessite l'hypothèse de normalité des populations, le test de normalité de Ryan-Joiner a été effectué. Les probabilités obtenues pour les données de chacun des deux peuplements sont toutes supérieures à 0,1 et permettent de conclure à la normalité des deux populations.

L'hypothèse nulle du test d'égalité des variances est :

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

Les symboles σ_1^2 et σ_2^2 représentent les variances des peuplements desquelles sont issus les échantillons. Pour effectuer le test, on calcule la quantité :

$$F_c = \frac{\hat{\sigma}^2_{sup}}{\hat{\sigma}^2_{inf}} . \qquad (3.4.1)$$

On rejette l'hypothèse nulle si :

$F_c > F_{1-\alpha/2}(ddl_1 ; ddl_2)$ ou $P(F_c > F) \leq \alpha/2$ avec $ddl_1 = n_1 - 1$ et $ddl2 = n_2 - 1$.

Les symboles ddl_1 et ddl_2 sont les nombres de degrés de liberté.

Dans le cas des données du tableau 9, $F_c = \frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2}$ = 3,258, ddl_1 =23 et ddl_2 =6.

La valeur $F_{1-\alpha/2}(23 ; 6)$ est égale à 5,128 et $P(F_c > F) = 0,146$. De ce fait, on accepte l'hypothèse nulle pour un niveau de confiance de 0,95 et on conclut à l'égalité des variances diamétriques des deux peuplements.

- *Echantillons non indépendants*

Supposons que l'on veuille comparer pour chacun des 11 peuplements du tableau 6, le volume estimé à partir d'une équation donnée et le volume calculé à partir d'une autre équation. Le tableau 10 présente les deux séries de volumes obtenues à partir des deux équations.

Il est aisé de noter de ce tableau que, pour chaque peuplement, nous disposons de deux valeurs de volume en bois : volume estimé1 et volume estimé2. Ainsi, les deux échantillons constitués respectivement des volumes-peuplements estimés de deux façons sont dépendants l'un de l'autre. En d'autres termes, ils sont constitués de volumes provenant d'une même série de peuplements. Dans un tel cas, l'hypothèse d'égalité des variances-populations

(hypothèse nulle), nécessaire à la comparaison des moyennes (par un test t par paires) peut être vérifiée en calculant la quantité :

$$t_{obs}=\frac{\left|SCE_1-SCE_2\right|\sqrt{n-2}}{2\sqrt{SCE_1SCE_2-SPE^2}}, \qquad (3.4.2)$$

SCE_1 et SCE_2 étant les sommes des carrés des écarts relatives respectivement aux volumes-peuplements estimés à partir des deux équations ($\hat{V}_1$ et $\hat{V}_2$). On les calcule de la façon suivante :

$$SCE_1=\sum_{i=1}^{n}(\hat{V}_1-\overline{\hat{V}_1})^2 \quad \text{et} \quad SCE_2=\sum_{i=1}^{n}(\hat{V}_2-\overline{\hat{V}_2})^2 .$$

Le symbole SPE représente la somme des produits des écarts entre volumes-peuplements estimés et calculés :

$$SPE=\sum_{i=1}^{n}(\hat{V}_1-\overline{\hat{V}_1})(\hat{V}_2-\overline{\hat{V}_2}) .$$

Tableau 10. Volumes-peuplements estimés de deux manières.

Peuplement	Volume calculé ($\hat{V}_1$)	Volume estimé ($\hat{V}_2$)
92/01	68,033	69,803
92/04	57,685	53,851
92/22	63,561	61,991
91/02	49,749	49,698
91/03	29,400	28,910
91/05	50,907	51,827
91/16	69,616	73,151
91/17	48,618	48,309
91/22	33,549	33,999
90/03	36,966	36,787
90/07	41,554	43,561
90/09	80,319	77,148
Variance	248,763	243,332

Dans le cas des données du tableau 10, nous avons :

SCE_1 = 2736,40 ; SCE_2 = 2676,70 ; SPE = 2682,42 et t_{obs} = 0,263.

Nous rejèterons l'hypothèse nulle d'égalité des variances lorsque :

$$t_{obs} \geq t_{1-\alpha/2} \quad \text{avec } n\text{-2 degrés de liberté.}$$

Dans le cas des données du tableau 10, pour un risque α = 0,05, la valeur $t_{0,975}$ (10) = 2,228. On constate que $t_{obs} < t_{0,975}$ (10) ; on accepte donc l'hypothèse nulle d'égalité des variances liées aux volumes-peuplements. La probabilité liée à ce test est en réalité de 0,601.

3.4.1.2. Comparaison de plus de deux populations

- *Echantillons indépendants*

Supposons, à titre didactique qu'on veuille comparer les peuplements de 90, 91 et 92 du point de vue de leur densité en *Acacia auriculiformis* (tableau 7). Une analyse de la variance à un critère de classification permettra de faire une telle comparaison.

Nous allons au préalable vérifier l'hypothèse d'égalité des variances des peuplements, l'hypothèse de normalité des populations étant déjà acceptée à l'aide du test de Ryan-Joiner (Paragraphe 3.2.2.2). Il est aisé de constater qu'ici, nous disposons de plus de deux populations (3 peuplements).

Pour réaliser ce test, l'hypothèse nulle peut s'écrire de la façon suivante :

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2 \qquad (3.4.3)$$

L'expression (3.4.3) signifie que les variances des p populations (ici, p =3) sont égales.

Nous proposons les trois méthodes les plus utilisées dans la réalisation de ce test à savoir le test de Bartlett, le test de Hartley et le test de Levene. Le test de Bartlett[1] s'applique à des échantillons d'effectifs égaux ou inégaux mais issus de populations normales, ce qui nécessite au préalable un test de normalité. Le test de Hartley[2] s'applique à des échantillons d'effectifs égaux issus de populations normales. Le test de Levene[3] s'applique aux échantillons de distributions continues non nécessairement normales. Rappelons que ces différents tests peuvent aussi être appliqués dans le cas de comparaison de variances de deux populations.

---

[1] En anglais : Bartlett's test.
[2] En anglais : Hartley's test.
[3] En anglais : Levene's test.

Test de Bartlett

Nous avons déjà vérifié la condition de normalité des données du tableau 7, nous pouvons donc appliquer le test de Bartlett.

Pour réaliser ce test, on calcule la statistique χ^2_{obs} donnée par la formule suivante :

$$\chi^2_{obs} = \frac{(n.-p)\ln\hat{\sigma}^2 - \sum_{i=1}^{p}\left[(n_i-1)\ln\hat{\sigma}_i^2\right]}{1+\frac{1}{3(p-1)}\left[\left(\sum_{i=1}^{p}\frac{1}{n_i-1}\right)-\frac{1}{n.-p}\right]} \quad . \qquad (3.4.4)$$

Dans l'expression (3.4.4), $\hat{\sigma}^2 = SCE/(n. - p)$ avec :

$$SCE = \sum_{i=1}^{p} SCE_i \ ; \ \hat{\sigma}_i^2 = SCE_i/(n_i - 1) .$$

Le symbole $n.$ représente l'effectif total : $n. = \sum_{i=1}^{p} n_i$.

On rejette l'hypothèse nulle d'égalité des variances lorsque : $\chi^2_{obs} \geq \chi^2_{1-\alpha}$ ($p-1$ degrés de liberté). De façon plus exacte, la probabilité associée à ce test est de la forme :

$$P(\chi^2 \geq \chi^2_{obs}) .$$

Lorsque la valeur de cette probabilité est inférieure à 0,05, on rejette l'hypothèse d'égalité des variances des peuplements. Pour les données du tableau 7, nous avons :

$\hat{\sigma}_1^2$ = 11379 ; $\hat{\sigma}_2^2$ = 24630 ; $\hat{\sigma}_3^2$ = 120799 ; $\hat{\sigma}^2$ = 43056 ; p =3 ;

SCE = 387510 ; $n.$ = 12.

$$\chi^2_{obs} = 2{,}870 \quad \text{avec} \quad P(\chi^2 \geq \chi^2_{obs}) = 0{,}238.$$

La probabilité associée au test étant supérieure à 0,05, on accepte l'hypothèse nulle d'égalité des variances en densité des trois peuplements. On pouvait toutefois comparer la valeur χ^2_{obs} à celle de $\chi^2_{1-\alpha}$ (2 degrés de liberté) qui est de 5,99. La valeur de α étant prise égale à 0,05, on constate que χ^2_{obs}

est inférieur à 5,99.

Test de Hartley

Le test de Hartley est conçu pour les échantillons d'effectifs égaux et est basé sur la statistique :

$$H_{obs} = \frac{\hat{\sigma}^2_{max}}{\hat{\sigma}^2_{min}}. \qquad (3.4.5)$$

Les symboles $\hat{\sigma}^2_{max}$ et $\hat{\sigma}^2_{min}$ dans l'expression (3.4.5) représentent respectivement la plus grande et la plus petite des variances des p échantillons. On rejette l'hypothèse nulle lorsque $H_{obs} \geq H_{1-\alpha}(ddl = n-1)$. Les valeurs particulières H_{obs} sont disponibles sous forme de tables qu'on peut trouver par exemple dans Dagnelie (1998).

Lorsque les effectifs des différents échantillons sont inégaux, sans être trop différents les uns des autres, il est possible d'utiliser ce test de façon approchée en prenant comme valeur critique la valeur $H_{1-\alpha}$ qui correspond à la moyenne des nombres de degré de liberté (Dagnelie, 1998).

Bien que le test de Hartley ne soit pas approprié aux données du tableau 7, du fait de l'inégalité marquée entre les effectifs des échantillons, nous allons appliquer ce test à ces données pour des raisons *strictement didactiques*. La valeur H_{obs} donne 5,325. La moyenne arrondie des nombres de degré de liberté est de 3. La valeur $H_{1-\alpha}$ est égale à 27,8 pour un risque α = 0,05. Puisque $H_{obs} < H_{1-\alpha}$, on accepte l'hypothèse nulle d'égalité des variances des densités des 3 peuplements. Néanmoins, il est à noter que ce résultat est biaisé du fait de l'inégalité des effectifs des échantillons.

Test de Levene

Le test de Levene (1960) est basé sur une analyse de la variance effectuée sur les écarts absolus par rapport à la moyenne de chaque échantillon. La statistique du test s'obtient comme suit :

$$L = \frac{(N-p)\sum_{i=1}^{p} n_i (\overline{V_{i.}} - \overline{V_{..}})^2}{(p-1)\sum_{i=1}^{p}\sum_{j=1}^{n} (V_{ij} - \overline{V_{i.}})^2} \qquad (3.4.6)$$

où

$V_{ij} = \left| X_{ij} - \overline{X}_i \right|$ et $\overline{V}..$ est la moyenne des $\overline{V}_{ij.}$; $\overline{V}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} V_{ij}$ et avec $i = 1,..., p$; $j = 1,..., n$, et $\overline{X}_i$, la moyenne des $\{X_{i1},..., X_{in_i}\}$.

La variable L suit une distribution F de Fisher-Snedecor de p -1 et N - p degrés de liberté. On rejette l'hypothèse nulle d'égalité des variances des populations lorsque :

$$L \geq F_{1-\alpha/2} \quad \text{ou} \quad P(F \geq L) \leq \alpha/2 .$$

Il est utile de noter que des modifications ultérieures ont été apportées au test original de Levene. Brown et Forsythe (1974) proposent de remplacer la moyenne $\overline{X}_i$ des observations de chaque échantillon par leur médiane, $\tilde{X}_i$. D'autres modifications comme celle de O'Brien (1979) ont aussi été apportées au test original de Levene mais le test de Brown et Forsythe (1974) paraît le plus précis parmi les autres modifications apportées au test original de Levene car, en utilisant la médiane plutôt que la moyenne de l'échantillon, les tests sont plus robustes pour des échantillons plus petits et la procédure est asymptotiquement indépendante de la distribution (Olejnik et Algina, 1987).

Dans le cas des données du tableau 7, l'utilisation du test de Brown et Forsythe (1974) donne les résultats ci-dessous:

N = 11, p = 3, $\sum_{i=1}^{p} n_i (\tilde{V}_{i.} - \tilde{V}_{..})^2$ = 37958 ; $\sum_{i=1}^{p} \sum_{j=1}^{n_i} (V_{ij} - \tilde{V}_{i.})^2$ = 187300

et

L = 0,912. Pour α=0,05, $F_{1-\alpha/2}(2 , 8)$ = 6,060 ; $P(F \geq L)$ = 0,436.

On constate que $L < F_{1-\alpha/2}(2 , 8)$ ou encore $P(F \geq L) > 0{,}025$.

On accepte donc l'hypothèse nulle d'égalité des variances des densités des trois peuplements.

- *Echantillons non indépendants*

Dans un tel cas, la méthode de Levene peut encore être utilisée (Dagnelie, 1998).

3.4.1.3. Test d'homogénéité des résidus de régression

L'une des conditions d'application de la régression linéaire est l'homogénéité des résidus de régression. En d'autres termes, l'équation de régression établie sera validée lorsque la variance conditionnelle des résidus est constante quelle que soit l'observation considérée. Lorsque cette hypothèse n'est pas vérifiée, on parle d'hétéroscédasticité des résidus et l'équation de régression établie peut conduire à des estimations biaisées.

Plusieurs méthodes de vérification de l'homogénéité des résidus sont proposées dans la littérature. On peut citer entre autres le test de White (1980) et le test de Breusch-Pagan (1979). Le test de Breusch-Pagan est plus spécifique à la vérification de l'hypothèse d'homogénéité que celui de White (1980) qui est plus général en détectant des formes d'anomalie des résidus autres que l'hétéroscédasticité (non-normalité par exemple). De ce fait, le test de White peut rejeter l'hypothèse d'homogénéité en l'absence d'hétéroscédasticité des résidus lorsque que le modèle de régression est imprécis dans un autre sens (Thursby, 1982). Le test de Breusch-Pagan est présenté ci-dessous. Nous proposons la version modifiée de ce test qui est moins sensible à une non-normalité que le test originel (Greene, 1993).

Pour Breusch-Pagan (1979), la variance résiduelle α_i de l'individu i de vecteur d'observations $\mathbf{z}_i$ peut être écrite sous la forme :

$$\sigma_i^2 = \sigma^2(\alpha_0 + \boldsymbol{\alpha}'\mathbf{z}_i) \qquad (3.4.7)$$

avec α_0 l'ordonnée à l'origine ; $\boldsymbol{\alpha}'$ le transposé du vecteur de coefficients de régression partiels et σ^2 une constante réelle.

L'hypothèse nulle du test est :

$$H_0 : \boldsymbol{\alpha}' = \mathbf{0} \ .$$

La statistique de ce test est :

$$bp = \frac{1}{v}(\mathbf{u} - \bar{u}\mathbf{i})'Z(Z'Z)^{-1}Z'(\mathbf{u} - \bar{u}\mathbf{i}) \qquad (3.4.8)$$

Dans l'expression (3.4.8), $\mathbf{u} = [e_1^2, e_2^2, ..., e_n^2]$ avec e_i le résidu de l'observation i. $\mathbf{i}$ est le vecteur-colonne unitaire ($n \times 1$), $\bar{u}$ est la moyenne du vecteur $\mathbf{u}$ et Z, la matrice des observations. Par ailleurs, v est une constante calculée à l'aide de la formule :

$$v = \frac{1}{n}\sum_{i=1}^{n}(e_1^2 - \frac{\mathbf{e'e}}{n})^2 .$$

Dans l'expression ci-dessus, $\mathbf{e}$ est le vecteur de résidus de régression.

Sous l'hypothèse nulle d'homogénéité des variances résiduelles, bp suit une distribution Chi-carré à p degrés de liberté. La probabilité associée à cette hypothèse est :

$$P(\chi^2 \geq bp) .$$

Lorsque la valeur de cette probabilité est inférieure à 0,05, on rejette l'hypothèse d'homogénéité des résidus.

Pour illustrer ce test, considérons un exemple relatif à l'établissement d'une équation permettant de prédire le volume de peuplement de *Acacia auriculiformis* en fonction de certains paramètres notamment la surface terrière (G), la densité du peuplement (N_p). Les données ayant servi à l'établissement de cette équation sont présentées au tableau 11. Dans ce tableau, les peuplements sont désignés comme auparavant par l'année de plantation et le numéro d'ordre d'installation. Ainsi, le premier peuplement 92/01 est installé en 1992 avec le numéro d'ordre 01.

L'équation obtenue de cet ajustement est la suivante :

$$\ln(Vp) = 3{,}344 + 1{,}239\ln(G) - 0{,}296\ln(Np) \qquad (3.4.9)$$

$$R^2 = 0{,}99$$

Pour vérifier l'homogénéité des résidus de cette équation de régression, nous avons appliqué le test de Breusch-Pagan aux résidus de régression présentés dans le tableau 11. La valeur de bp obtenue est égale à 1,78 avec une probabilité de 0,41. En d'autres termes, on accepte l'hypothèse d'homogénéité des résidus de la régression.

Tableau 11. Données ayant servi à la construction du tarif de cubage-peuplement.

peuplement	lnVp(m^3/ha)	lnG(m^2/ha)	lnNp(t/ha)	Résidus de la régression
92/01	4,220	2,310	6,623	-0,027
92/04	4,055	2,108	6,654	0,068
92/22	4,152	2,240	6,731	0,024
91/02	3,907	2,037	6,628	0,000
91/03	3,381	1,379	5,704	0,016
91/05	3,930	2,041	6,503	-0,019
91/16	4,243	2,349	6,628	-0,050
91/17	3,884	2,006	6,594	0,006
91/22	3,513	1,593	6,052	-0,014
90/03	3,610	1,379	4,890	0,004
90/07	3,727	1,858	6,324	-0,048
90/09	4,386	2,353	6,465	0,039

3.4.2. Application avec les logiciels statistiques

Plusieurs logiciels statistiques offrent des possibilités de réalisation des tests d'égalité des variances ou de matrices de variances-covariances (tests d'homoscédasticité). Certains de ces tests sont intégrés aux modules d'analyse de la variance des logiciels statistiques. Les autres peuvent être exécutés indépendamment de l'analyse de la variance. Nous présentons ici l'application de ces tests aux données des tableaux 9 et 11 en utilisant les logiciels Minitab, SPSS et SAS.

3.4.2.1. Logiciel Minitab

Nous reprenons l'exemple du tableau 9 afin d'exposer la procédure de réalisation du test d'égalité des variances de deux populations avec le logiciel Minitab. Pour ce faire, les données peuvent être saisies de deux différentes manières :

- les deux échantillons peuvent être mis dans une seule colonne et les indices d'identification des deux populations dans une autre colonne ;

- dans le second cas, chaque échantillon peut être placé dans deux colonnes distinctes. Les échantillons peuvent ne pas être de même effectif.

Utilisons la première manière d'enregistrement des données et désignons le peuplement de 1991 par l'indice 1 et celui de 1992 par l'indice 2. Le test d'égalité des variances des deux peuplements se réalise en sélectionnant : « **Stat > Statistiques élémentaires > 2 variances** » (cf. figure 28). Dans la boîte de dialogue qui s'affiche (figure 29), on introduit la colonne contenant les

échantillons dans la fenêtre « **Echantillons** » et la colonne des indices d'identification des échantillons dans la fenêtre « **Indices** ». Dans le cas où les deux échantillons sont placés dans deux colonnes différentes (deuxième possibilité), il faut plutôt sélectionner l'option « **Echantillons dans plusieurs colonnes** ». Ensuite, il faut insérer la colonne contenant les données du premier échantillon dans la fenêtre « **Premier** » et celle contenant les données du second échantillon dans la fenêtre « **Deuxième** ».

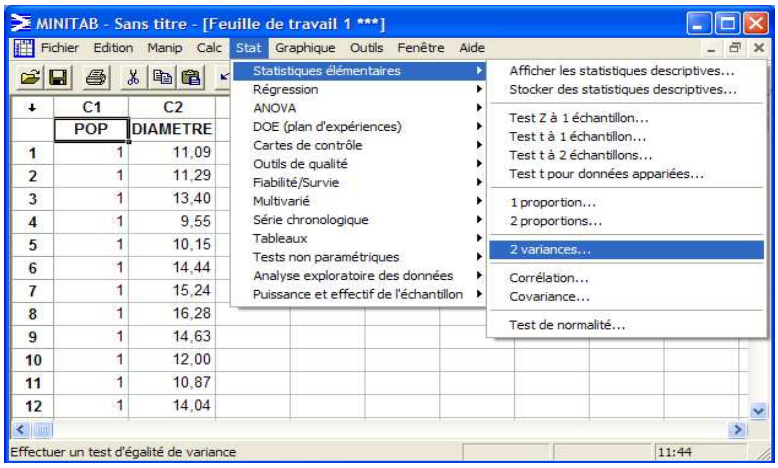


Figure 28. Procédure d'exécution du test d'égalité de deux variances avec le logiciel Minitab.

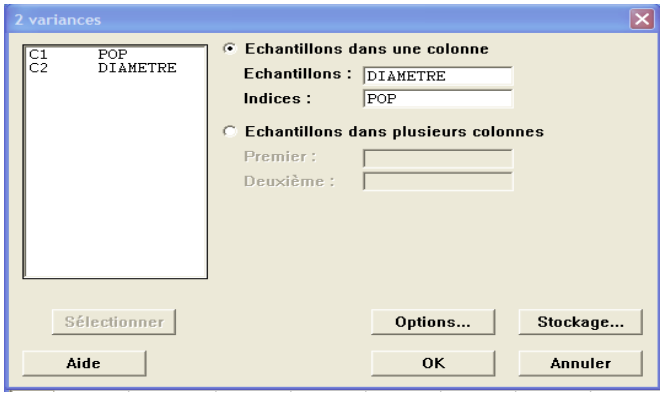


Figure 29. Procédure d'exécution du test d'égalité de deux variances avec le logiciel Minitab: boîte de dialogue 1.

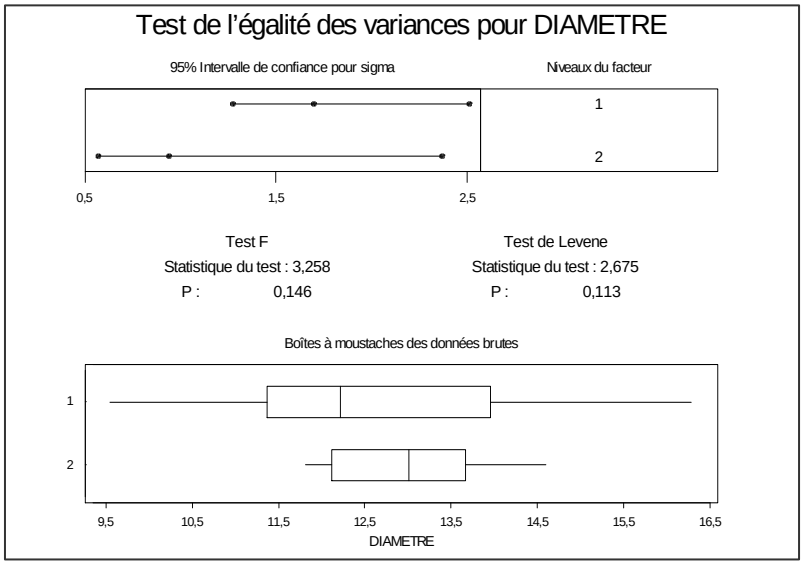


Figure 30. Diamètre moyen (cm) des peuplements de 1991 et 1992 : résultats du test d'égalité de variances avec Minitab.

Quelle que soit la procédure utilisée, on obtient les résultats de la figure 30 dans le cas des données du tableau 9. Comme le montre cette figure, Minitab fournit deux types de tests à savoir le test F et le test modifié de Levene ou test de Brown et Forsythe (1974). Pour ce qui est du test F, les résultats fournis par Minitab sont, aux arrondis près, les mêmes que ceux obtenus manuellement au paragraphe 3.4.1.1. Le test F indique que l'hypothèse nulle doit être acceptée au seuil de 5 % et on peut conclure à l'égalité des variances diamétriques des deux peuplements. Les résultats du test de Levene indiquent aussi une acceptation de l'hypothèse d'égalité des variances diamétriques.

Pour plus de deux populations, Minitab offre une autre procédure pour le test d'égalité des variances. Pour illustrer cette procédure, reprenons l'exemple du tableau 6 relatif à la densité en *Acacia auriculiformis* de peuplements mélangés. L'exécution du test se fait en sélectionnant « **Stat > ANOVA > Test de l'égalité des variances** » comme le montre la figure 31. On obtient la boîte de dialogue de la figure 32. Pour l'exécution de la procédure, les échantillons doivent être mis dans une seule colonne et les indices identifiant les populations dans une autre colonne. Ces indices prennent seulement en compte les deux derniers chiffres de l'année de plantation comme le montre la figure 33. La colonne contenant les échantillons doit être placée dans la fenêtre « **Réponse** » et les indices dans la fenêtre « **Facteurs** ». Dans cette dernière fenêtre, on a la possibilité de spécifier jusqu'à neuf facteurs. Dans le cas de notre exemple, l'exécution de la procédure donne les résultats de la figure 33.

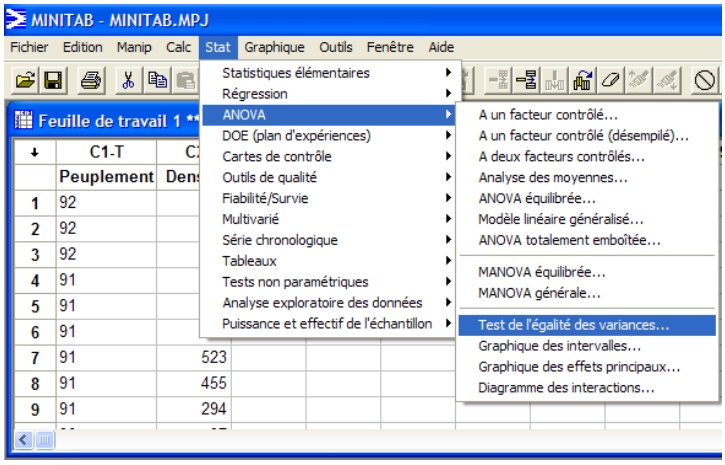


Figure 31. Procédure d'exécution du test d'égalité de plus de deux variances avec le logiciel Minitab.

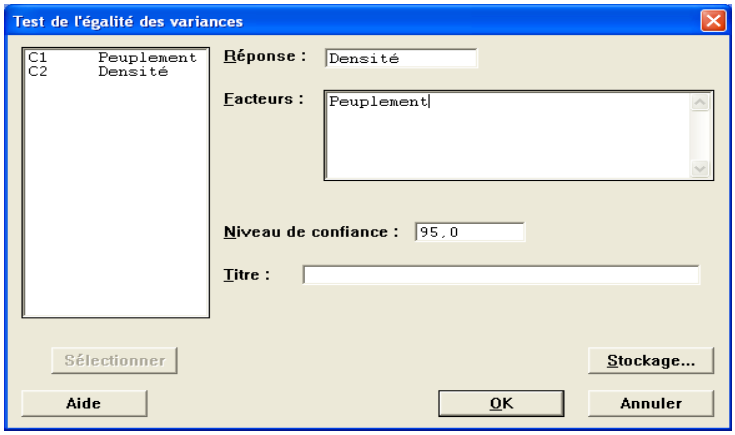


Figure 32. Procédure d'exécution du test d'égalité de plus de deux variances avec le logiciel Minitab : boîte de dialogue.

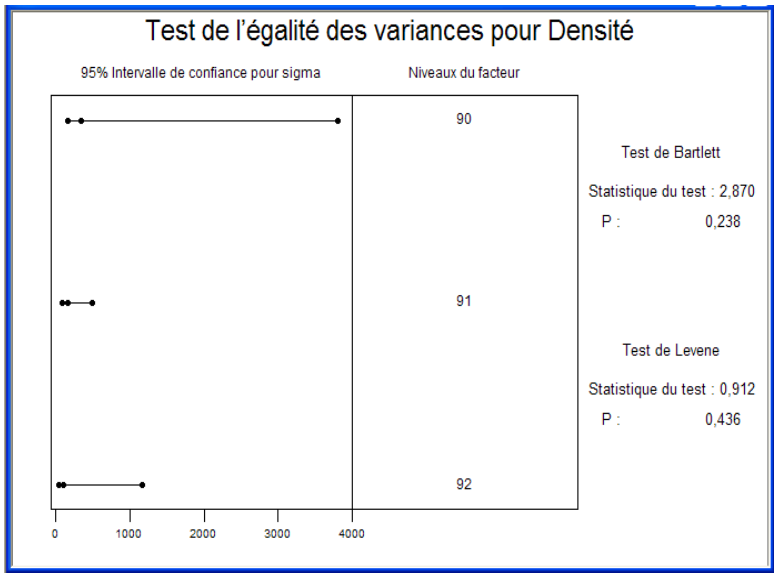


Figure 33. Densité en pieds d'Acacia de peuplements mélangés : résultats du test d'égalité des variances avec Minitab.

Les résultats de la figure 33 montrent que Minitab propose deux types de tests d'égalité des variances de plus de deux populations : le test de Bartlett et le test modifié de Levene (Brown et Forsythe, 1974). Les statistiques de Bartlett et de Levene et les probabilités associées sont identiques aux résultats obtenus au paragraphe 3.4.1.2.

Il est à noter que la procédure d'exécution du test d'égalité des variances de plus de deux populations peut également être utilisée pour tester l'égalité des variances de deux populations. Dans ce cas, le test de Bartlett est automatiquement remplacé par un test F. Rappelons également que le test de Bartlett est utilisé lorsque les données sont issues de lois normales alors que le test de Levene peut être utilisé même si la distribution ne suit pas une loi normale. Autrement dit, le test de Levene est robuste en cas d'écart par rapport à la normalité.

3.4.2.2. Logiciel SPSS

Le logiciel SPSS propose plusieurs procédures pour le test de l'égalité des variances de deux ou plusieurs populations. Mais contrairement à Minitab qui propose le test de Levene et de Bartlett, SPSS ne fournit que le test de Levene. Nous allons présenter trois procédures pour la réalisation du test, en nous basant sur les mêmes données que celles utilisées dans le paragraphe 3.4.1.1.

La première procédure est celle qui est réalise en sélectionnant **« Analyse > Statistiques descriptives > Explorer…** » (cf. figure 11). On obtient alors la boîte de dialogue de la figure 12. On insère ensuite la variable dépendante (ici, Diamètre) dans la fenêtre « **Variables dépendantes :** » puis la colonne contenant les indices d'identification des populations dans la fenêtre « **Variable(s) active(s) :** » (cf. figure 34).

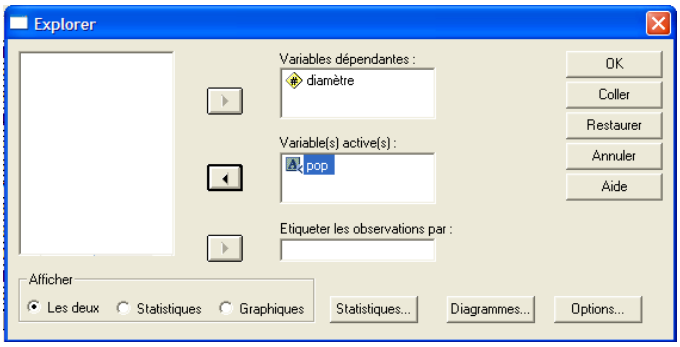


Figure 34. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue 1.

Pour exécuter les tests d'égalité des variances dans le logiciel SPSS, il faut cliquer sur la commande « **Diagramme…** » de la figure 34, puis cocher « **Estimation d'exposants** » (cf. figure 35).

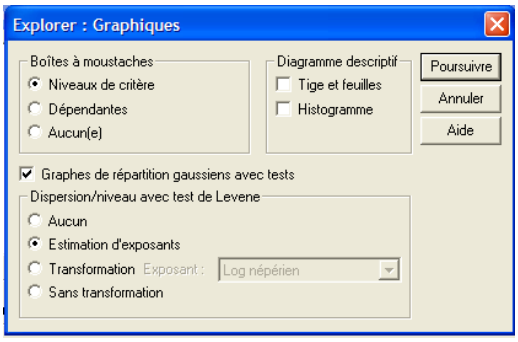


Figure 35. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue 2.

On obtient les résultats de la figure 36. Ces résultats montrent que SPSS fournit quatre types de test de Levene. La méthode basée sur les médianes donne les mêmes résultats que le test de Levene fournit par Minitab, c'est-à-dire le test de Brown-Forsythe.

Test d'homogénéité de la variance

		Statistique de Levene	ddl1	ddl2	Signification
DIAMÈTRE	Basé sur la moyenne	3,149	1	29	,086
	Basé sur la médiane	2,675	1	29	,113
	Basé sur la médiane et avec ddl ajusté	2,675	1	26,287	,114
	Basé sur la moyenne tronquée	3,061	1	29	,091

Figure 36. Diamètre moyen (cm) des peuplements de 1991 et 1992 : résultats du test d'égalité de variances avec SPSS.

La deuxième procédure de réalisation du test d'égalité des variances avec SPSS est celle qui s'effectue en sélectionnant **Analyse > Comparer les moyennes > ANOVA à 1 facteur…>** (cf. figure 37). On obtient la boîte de dialogue de la figure 38, dans laquelle il faut insérer la variable dépendante « **diamètre** » dans la fenêtre « **Variables dépendantes** » puis la colonne « **pop** » contenant les indices d'identification des populations dans la fenêtre « **Critère :** ».

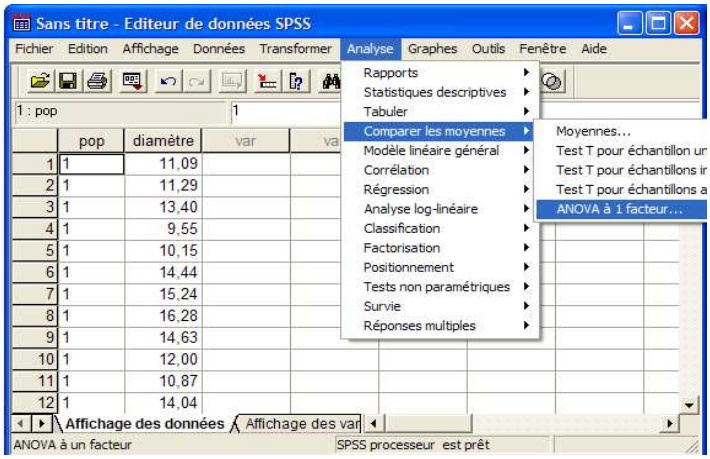


Figure 37. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS (seconde procédure).

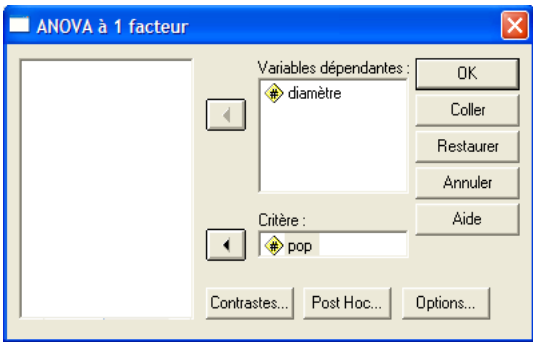


Figure 38. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue de la seconde procédure.

Pour réaliser le test d'égalité des variances en SPSS, il faut cliquer sur la commande « **Options…** » de la figure 38, puis cocher « **Test d'égalité des variances** » (cf. figure 39).

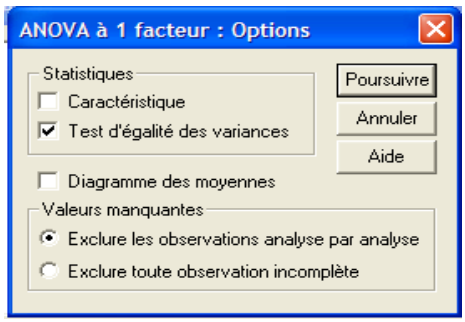


Figure 39. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue de la commande « Options ».

L'exécution donne les résultats présentés à la figure 40. L'observation de cette figure permet de noter que les résultats fournis par cette seconde procédure correspondent à ceux de la méthode basée sur la moyenne (figure 36). Ce sont les résultats fournis par défaut par le logiciel SPSS lorsqu'on choisit d'utiliser la deuxième ou la troisième procédure qui s'effectue en sélectionnant **Analyse > Modèle linéaire général > univariée…>** (cf. figure 41, 42 et 43). Cette troisième procédure permet d'obtenir les résultats présentés à la figure 44.

Test d'homogénéité des variances

DIAMÈTRE

Statistique de Levene	ddl1=	ddl2	Signification
3,149	1	29	,086

Figure 40. Diamètre moyen des peuplements de 1991 et 1992 : résultats du test d'égalité des variances avec SPSS (seconde procédure).

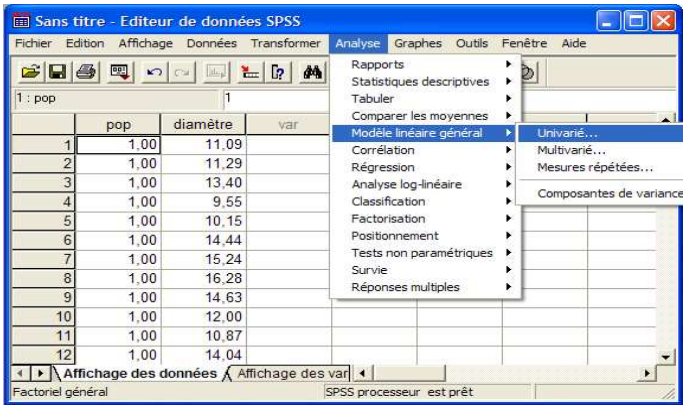


Figure 41. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS (troisième procédure).

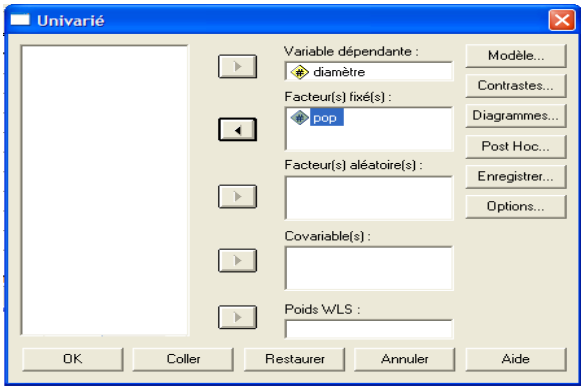


Figure 42. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue. de la troisième procédure.

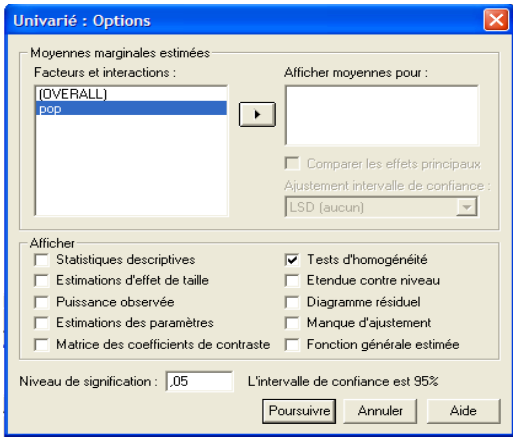


Figure 43. Procédure d'exécution du test d'égalité des variances avec le logiciel SPSS : boîte de dialogue de la commande « Options ».

Test d'égalité des variances des erreurs de Levene[a]

Variable dépendante: DIAMÈTRE

F	ddl1	ddl2	Signification
3,149	1	29	,086

Teste l'hypothèse nulle que la variance des erreurs de la variable dépendante est égale sur les différents groupes.

a. Plan : Intercept+POP

Figure 44. Diamètre moyen des peuplements de 1991 et 1992 : résultats du test d'égalité de variances avec SPSS (troisième procédure).

Il est à noter que l'une ou l'autre de ces trois procédures peut être utilisée pour la réalisation du test d'égalité des variances, que l'on ait deux ou plus de deux populations. La première procédure présente le test original de Levene, le test de Brown-Forsythe avec ou sans ajustement de nombre de degrés de liberté ; la deuxième et la troisième procédures présentent le test original de Levene. Une des limites de la première procédure est qu'elle ne réalise le test de Levene que lorsque chacune des populations contient plus de trois observations. Dans le cas des données du tableau 6, l'utilisation de la deuxième procédure d'exécution du test d'égalité des variances donne les résultats de la figure 45.

Test d'homogénéité des variances

DENSITÉ

Statistique de Levene	ddl1=	ddl2	Signification
2,953	2	9	,103

Figure 45. Densité en pieds d'Acacia des 3 peuplements : résultats du test d'égalité de variance avec SPSS (2nde procédure).

3.4.2.3. Logiciel SAS

Tout comme Minitab et SPSS, le logiciel SAS peut aussi être utilisé pour le test d'égalité des variances de deux ou plus de deux populations. Quatre types de test d'égalité des variances sont proposés par SAS à savoir le test de Bartlett, le test de Brown-Forsythe, le test de Levene et le test de O'Brien. De manière pratique, ces différents tests s'exécutent à travers l'utilisation de la procédure PROC GLM liée à l'analyse de la variance. Dans cette procédure, on spécifie la commande MEANS puis l'option HOVTEST. A titre d'illustration, reprenons l'exemple du tableau 6 et exécutons les trois tests avec le logiciel SAS. La figure 46 donne la procédure de réalisation des tests.

```
Data Dens;
Input Pop Densite;
Cards;
92      450
92      333
92      546
91      508
91      353
91      743
91      523
91      455
91      294
90      97
90      600
90      764
;
Proc glm data = Dens;
Class Pop;
Model Densite = Pop;
Means Pop / Hovtest=Bartlett Hovtest=BF
Hovtest=Levene(Type=ABS);
Run;
```

Figure 46. Procédure SAS d'exécution des tests d'égalité des variances.

Nous présentons à la figure 46, trois des quatre tests à savoir le test de Bartlett, le test de Brown et Forsythe (1974) et le test originel de Levene. Les résultats de l'exécution du programme sont présentés à la figure 47 et sont identiques à ceux obtenus plus haut au paragraphe 3.4.1.2. Que l'on ait deux ou plus de deux populations, la procédure présentée à la figure 46 peut être utilisée.

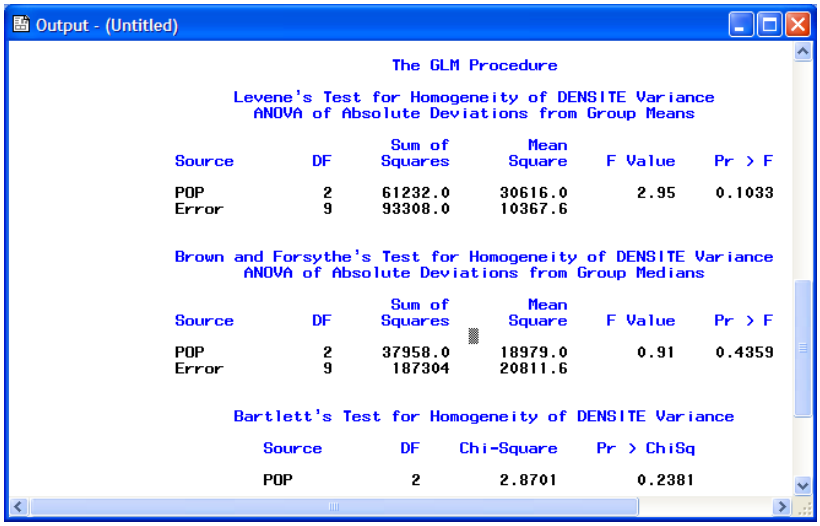


Figure 47. Densité des peuplements 1990, 1991 et 1992 : résultats des tests d'égalité de variances avec SAS.

On peut noter de cette figure que la probabilité liée au test de Levene est la même que celle présentée à la figure 45. Les résultats des tests de Brown-Forsythe et de Bartlett sont identiques à ceux obtenus au paragraphe 3.4.1.2. et à la figure 33.

Pour le test de Breusch-Pagan, la procédure SAS utilisée est présentée à la figure 48.

```
data acacia;
input lnVp  lnG  lnNp;
cards;
4.22   2.310    6.623
4.055  2.108    6.654
4.152  2.240    6.731
3.907  2.037    6.628
3.381  1.379    5.704
3.930  2.041    6.503
4.243  2.349    6.628
3.884  2.006    6.594
3.513  1.593    6.052
3.610  1.379    4.89
3.727  1.858    6.324
4.386  2.353    6.465
;
proc model data=acacia;
fit lnVp / white breusch = (1 lnG lnNp);
run;
```

Figure 48. Procédure SAS de réalisation des tests de White et Breusch-Pagan d'homogénéité des résidus.

Nous ne présentons pas les résultats de l'application de la procédure de la figure 48 ; la version 9.1 du logiciel SAS que nous avons utilisée ne comporte pas cette procédure.

3.5. Tests d'égalité des matrices de variances-covariances

Dans le présent paragraphe et pour des raisons didactiques, nous comparons les deux types de pâturages du tableau 8 du point de vue du poids moyen en graminées et en légumineuses. Pour effectuer statistiquement une telle comparaison, la méthode appropriée est l'analyse de la variance multivairiée[1], suivie le cas échéant de l'analyse canonique discriminante[2]. L'une des conditions d'application de ces deux méthodes est l'égalité des matrices de variances-covariances ou homoscédasticité[3]. Nous décrivons dans cette note deux méthodes de vérification de l'homoscédasticité : le test d'homoscédasticité du rapport de vraisemblance[4] et le test M de Box[5] qui utilisent globalement les mêmes principes.

3.5.1. Test d'homoscédasticité du rapport de vraisemblance

Considérons g populations dans lesquelles sont tirés g échantillons d'effectifs respectifs $n_i \ (i=1,...,g)$. L'effectif global étant $n=\sum_{i=1}^{g} n_i$.

Le test d'homoscédasticité du rapport de vraisemblance sous l'hypothèse de normalité des données a pour hypothèse nulle :

$$\mathrm{H}_0 : \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \ (i=1,...,\ g). \tag{3.5.1}$$

Le test est basé sur la statistique :

$$\Gamma = \sum_{i=1}^{g} n_i \ln[|\hat{\boldsymbol{\Sigma}}| / |\hat{\boldsymbol{\Sigma}}_i|]. \tag{3.5.2}$$

Dans l'expression (3.5.2), $\hat{\boldsymbol{\Sigma}}_{\mathrm{i}}$ est la matrice de variance-covariance de l'échantillon tiré de la population $G_i \ (i=1,...,g)$ et $\hat{\boldsymbol{\Sigma}}=\sum_{i=1}^{g}(n_i/n)\hat{\boldsymbol{\Sigma}}_i$.

Sous l'hypothèse H_0, Γ suit asymptotiquement une distribution Chi-carré à $\frac{1}{2}(g-1)p(p+1)$ degrés de liberté (McLachlan, 1992).

Lorsque les effectifs n_i sont inégaux, le test basé sur Γ est biaisé et de ce fait, on utilise en pratique la statistique modifiée (McLachlan, 1992) :

---

[1] En anglais : Multivariate analysis of variance (MANOVA).
[2] En anglais : Canonical discriminant analysis.
[3] En anglais : Homoscedasticity.
[4] En anglais : Likelihood ratio test for homoscedasticity.
[5] En anglais : Box's M test.

$$\Gamma^{*}=-\sum_{i=1}^{g}(n_i-1)\ln[|\mathbf{S}_i|/|\mathbf{S}|], \qquad (3.5.3)$$

où

$$\hat{\mathbf{S}}_i=n_i\hat{\mathbf{\Sigma}}_i/(n_i-1)$$

est l'estimation non biaisée de $\mathbf{\Sigma}_i$ $(i=1,...,\ g)$ et $\hat{\mathbf{S}}=n\hat{\mathbf{\Sigma}}/(n-g)$, l'estimation non biaisée de la matrice de variances-covariances commune $\mathbf{\Sigma}$ sous l'hypothèse d'homoscédasticité.

Greenstreet et Connor (1974) ont montré comment la statistique Γ ou Γ^{*} peut être modifiée par un facteur multiplicatif constant pour donner un test de même puissance que précédemment mais avec des valeurs de Γ ou Γ^{*} plus faibles.

Dans le cas de Γ^{*}, le facteur multiplicatif constant est défini par :

$$C=1-\left[\left\{(\sum_{i=1}^{g}\frac{1}{n_i-1})-\frac{1}{n-g}\right\}(2p^2+3p-1)/6(g-1)(p+1)\right].$$

Il est utile de noter que Layard (1974) a démontré la non-robustesse de ce test par rapport à une non-normalité.

En pratique, l'hypothèse d'homoscédasticité est rejetée lorsque $\Gamma^{*} \geq \chi^2_{1-\alpha}$ ou $P(\chi^2 \geq \Gamma^{*}) \leq \alpha$.

Pour les données du tableau 8, nous avons les résultats suivants : g =2, n_1 = 6, n_2 =6 ; p = 2.

$$\hat{\mathbf{S}}_1 = \begin{pmatrix} 6{,}8180 & -3{,}1327 \\ -3{,}1327 & 1{,}9259 \end{pmatrix} \qquad \hat{\mathbf{S}}_2 = \begin{pmatrix} 8125800 & 1488000 \\ 1488000 & 1090600 \end{pmatrix}$$

$$\hat{\mathbf{S}} = \begin{pmatrix} 4097000 & 72830 \\ 72830 & 55490 \end{pmatrix}.$$

Γ^{*} = 36,142.

En appliquant le facteur multiplicatif (C = 0,7833), la valeur de Γ^* devient 28,311. Le nombre de degré de liberté étant égale à 3, la probabilité correspondante à l'hypothèse nulle dans le cas d'utilisation de la valeur modifiée de Γ^* est inférieure à 0,0001 (0,0000031). On rejette donc l'hypothèse nulle et on conclut à l'hétéroscédasticité du modèle. En d'autres termes, les matrices de variances-covariances des données relatives aux deux pâturages sont significativement inégales.

3.5.2. Test M de Box

Le test de Box est une généralisation du test de Bartlett dans le cas de données multivariées. Ce test est basé sur l'hypothèse de multinormalité des données. Ainsi, sous une telle hypothèse, la statistique M de Box a pour expression :

$$M = (N-g)\ln\left|\boldsymbol{\Sigma}\right| - \sum_{i=1}^{g}(n_i - 1)\ln\left|\boldsymbol{\Sigma}_i\right| . \qquad (3.5.4)$$

Le symbole $\boldsymbol{\Sigma}_i$ représente la matrice de variances-covariances du groupe ou population i et $\boldsymbol{\Sigma} = \frac{1}{(N-g)}\sum_{i=1}^{g}(n_i - 1)\boldsymbol{\Sigma}_i$ représente la matrice de variances-covariances inter-classes combinée.

Soient e_1 et e_2, deux réels dont les expressions sont les suivantes :

$$e_1 = \left[\left(\sum_{i=1}^{g} 1/(n_i - 1)\right) - (1/(N-g)) \right] \frac{2p^2 + 3p - 1}{6(g-1)(p+1)}$$

et

$$e_2 = \left[\left(\sum_{i=1}^{g} 1/(n_i - 1)^2\right) - (1/(N-g)^2) \right] \frac{(p-1)(p+2)}{6(g-1)} .$$

Soient t_1, t_2 et b, trois réels tels que :

$$t_1 = (g-1)p(p+1)/2 \quad ; \quad t_2 = (t_1 + 2)/\left|e_2 - e_1^2\right| \quad \text{et}$$

$$t_1/(1 - e_1 - t_1/t_2) \quad \text{si} \quad e_2 > e_1^2$$

$$t_2/(1 - e_1 + 2/t_2) \quad \text{si} \quad e_2 < e_1^2$$

Le rapport :

$$\mathrm{F} = \begin{cases} M/b & \text{si} \quad e_2 > e_1^2 \\ t_2 M/t_1(b - M) & \text{si} \quad e_2 < e_1^2 \end{cases} \qquad (3.5.5)$$

suit une distribution de Fisher-Snedecor à t_1 et t_2 degrés de liberté. Dans le cas où $e_1^2 - e_2$ est proche de 0, on utilise la statistique $(1-e_1)M$ qui est approximativement une χ^2 à t_1 degrés de liberté (Saporta, 1990).

Dans le cas des données du tableau 8, en supposant la multinormalité des données, nous avons :

e_1^2 = 0,0469 e_2 = 0,0466 ; $e_2 < e_1^2$ donc $b = t_2/(1-e_1+2/t_2)$ = 22975 ;
M = 36,142. F = 9,453 ; t_1 =3 et t_2 = 17999,9.

La probabilité $P(F \geq 17999)$ pour 3 et 17999 degrés de liberté est égale à 0,000. On rejette alors l'hypothèse d'égalité des matrices de variances-covariances des deux groupes de peuplements.

Par ailleurs, puisque $e_1^2 - e_2$ est assez proche de 0, on devrait utiliser la statistique $(1-e_1)M$ qui donne la valeur 28,311 du paragraphe 3.5.1 et la probabilité correspondante est égale à 0,000. Dans tous les cas, on rejette l'hypothèse nulle d'égalité des matrices de variances-covariances.

3.5.3. Applications avec les logiciels statistiques

La version 13Fr du logiciel Minitab utilisé ne comporte aucun test d'égalité de matrices de variances-covariances. De ce fait, l'application des tests se fera dans les logiciels SAS et SPSS. Le logiciel SAS prend en compte le test d'homoscédasticité du rapport de vraisemblance alors que SPSS pend en compte le test M de Box.

3.5.3.1. Logiciel SPSS

En SPSS, le test M de Box est intégré à la procédure d'analyse de la variance multivariée. Pour réaliser ce test, on sélectionnant **« Analyse > Modèle linéaire général > Multivariée…** » (cf. figure 49). On obtient alors la boîte de dialogue de la figure 50 dans laquelle on insère les variables dépendantes (ici, graminée et legum) dans la fenêtre « **Variables dépendantes :** » puis la colonne contenant les indices d'identification des pâturages dans la fenêtre « **Facteur(s) fixé(s) : »** . En cliquant sur le bouton « **option…** », on obtient la boîte de dialogue de la figure 51 dans laquelle on coche « **Tests d'homogénéité** ».

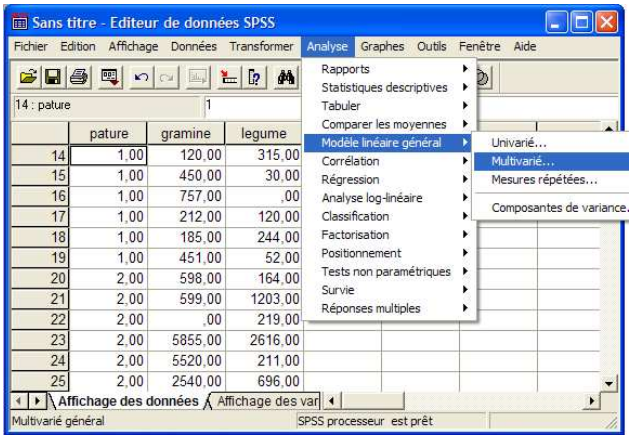


Figure 49. Procédure d'exécution du test M de Box en SPSS.

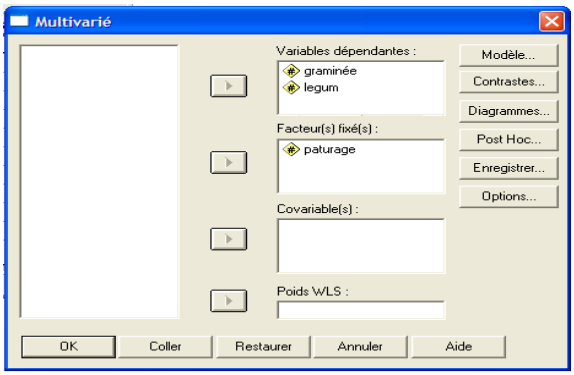


Figure 50. Procédure d'exécution du test M de Box en SPSS : boîte de dialogue 1.

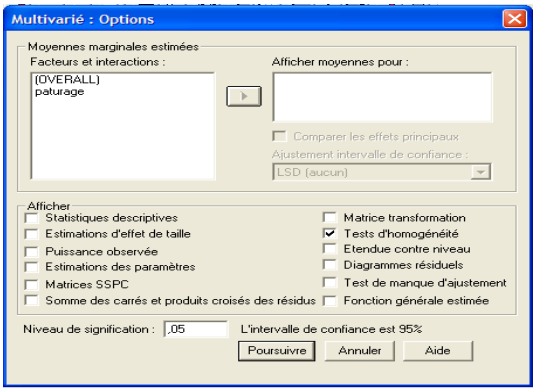


Figure 51. Procédure d'exécution du test M de Box en SPSS : boîte de dialogue 2.

Les résultats obtenus et présentés à la figure 52 sont identiques à ceux obtenus au paragraphe 3.5.2. De ce fait, on rejette l'hypothèse d'égalité des matrices de variances-covariances des deux populations.

Test d'égalité des matrices de covariance de Box[a]

M de Box	36,142
F	9,435
ddl1	3
ddl2	18000,000
Signification	,000

Teste l'hypothèse nulle selon laquelle les matrices de covariances observées des variables dépendantes sont égales sur l'ensemble des groupes.

a. Plan : Intercept+PATURAGE

Figure 52. Test M de Box : résultats obtenus avec SPSS.

3.5.3.2. Logiciel SAS

Dans le logiciel SAS, le test du rapport de vraisemblance pour la comparaison des matrices de variances-covariances ou test généralisé de Bartlett est intégré à la procédure d'analyse discriminante. Les résultats de ce test permettent en effet à l'utilisateur de choisir l'analyse discriminante linéaire en cas d'acceptation de l'hypothèse d'égalité des matrices de variances-

covariances et d'utiliser l'analyse discriminante quadratique dans le cas contraire, pour autant que les populations soient multinormales. La procédure à utiliser est appliquée aux données du tableau 8 et est présentée à la figure 53.

```
Data paturage;
Input pature gramine legumineuse;
Cards;
1    120   315
1    450   30
1    757   0
1    212   120
1    185   244
1    451   52
2    598   164
2    599   1203
2    0     219
2    5855 2616
2    5520 211
2    2540 696
;
Proc discrim Data=paturage Method=Normal OUT=Sortie All
Pool=test;
Class pature;
Run;
```

Figure 53. Procédure SAS d'exécution du test généralisé de Bartlett pour la comparaison de deux matrices de variances-covariances.

Les résultats obtenus de l'exécution de la procédure ci-dessus sont présentés à la figure 54 et sont identiques à ceux du paragraphe 3.5.1. Le logiciel présente d'abord la méthodologie de réalisation du test généralisé de Bartlett pour la comparaison des matrices de variances-covariances. Cette méthodologie est différente de celle présentée au paragraphe 3.5.1 mais elles conduisent aux mêmes résultats.

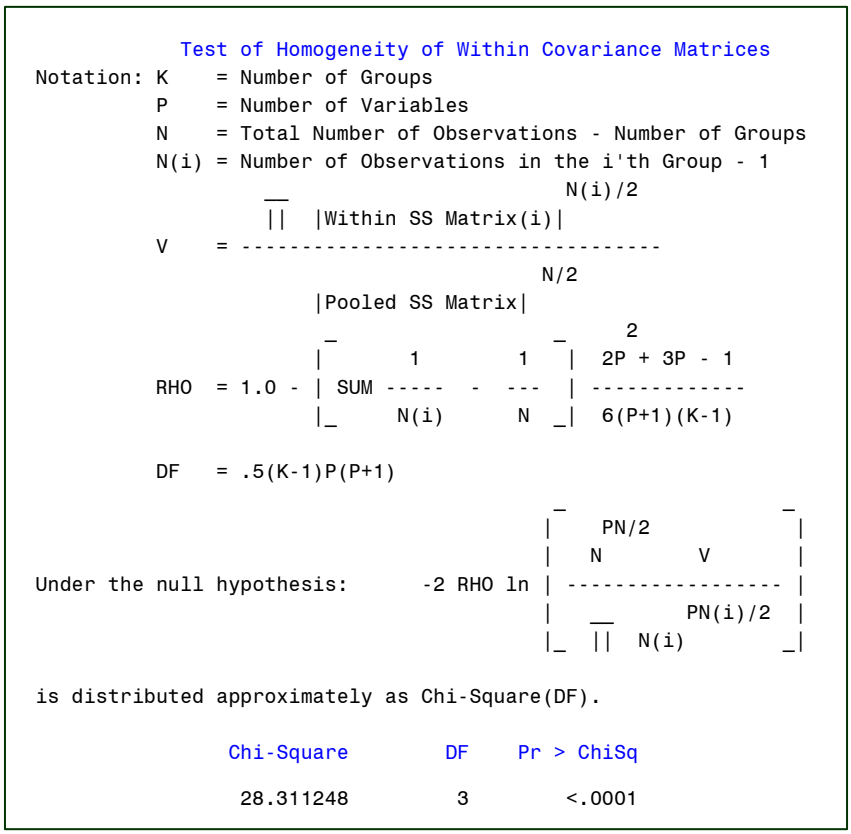


Figure 54. Test du rapport de vraisemblance pour la comparaison des matrices de variances-covariances : résultats obtenus avec SAS.

4. Conclusion

La vérification des conditions d'application est une étape importante dans l'utilisation des méthodes statistiques paramétriques. Lorsque ces conditions ne sont pas respectées, il est possible que les résultats obtenus soit biaisés. Nous avons exposé dans la présente note, le cadre théorique d'établissement des méthodes pour montrer l'importance des hypothèses sous-tendant leur application ainsi que les conséquence du non-respect de ces hypothèses. Dans le cas d'une non-normalité des populations, le risque de première espèce peut être modifié de sorte que le test inférentiel effectué est soit plus libéral ou plutôt conservateur. De la même manière, le non-respect de la condition d'homoscédasticité affecte le risque nominal et surtout la précision des tests de structuration de moyennes en analyse de la variance. De même, le non-respect de l'hypothèse de normalité multivariée peut affecter l'analyse de la variance multivariée ainsi que la qualité des règles de classement paramétriques établies en analyse discriminante. Par ailleurs, une forte hétéroscédasticité peut affecter spécifiquement la règle linéaire de classement.

Nous avons aussi abordé, entre autres, les méthodes de vérification du respect des hypothèses d'utilisation des tests statistiques paramétriques. Pour la normalité univariée, les méthodes de vérification abordées sont le test de Ryan-Joiner, le test de Shapiro-Wilk et le test de Kolmogorov-Smirnov alors que la normalité multivariée est vérifiée à l'aide des tests de Mardia et de Rao-Ali. L'égalité des variances est vérifiée avec les tests de Hartley, de Bartlett, de Levene et le test modifié de Breusch-Pagan pour la vérification de l'hypothèse d'homogénéité des résidus. L'égalité des matrices de variances-covariances est vérifiée par le test généralisé du rapport de vraisemblance de Bartlett et le test M de Box.

L'application de ces différentes méthodes de vérification des tests paramétriques sur ordinateur est abordée avec notamment, les logiciels SAS, Minitab, SPSS et Matlab. Ceci donne à l'utilisateur un éventail de possibilités d'exécution de ces méthodes dans les logiciels statistiques. Il y trouve ainsi un moyen rapide pour éviter une utilisation abusive des méthodes statistiques courantes.

Enfin, il est à noter que toutes les conditions d'application des méthodes statistiques paramétriques ne sont pas exposées dans cette note de biométrie comme par exemple le test de parallélisme des droites de régression nécessaire à l'exécution de l'analyse de la covariance. L'objectif poursuivi a été de présenter les conditions d'application des méthodes les plus courantes dans le traitement des données.

5. Références bibliographiques

Akossou A.J.Y., Fonton N.H., Claustriaux J.J. (2001). Introduction à la programmation avec Matlab sous windows. *Notes Biom. Info*. Bibliothèque Nationale.

Bayne C. K., Beauchamp, J. J., Kane, V. E., McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Comput. Stat. Data Anal.*, 1: 257-273.

Breusch T.S. and Pagan, A.R., (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47 (5), 1287--1294.

Brown M.B., Forsythe, A.B. (1974), Robust Tests for Equality of Variances. *Journal of the American Statistical Association,* 69 : 364 -367.

Clarke, W. R., Lachenbruch, P. A., Broffit, B. (1979). How nonnormality affects the quadratic discrimination function. *Communications in Statistics-Theory and Methods*, A8, 1285-1301.

Dagnelie P. (1998). *Statistique théorique et appliquée* vol. 1 & 2. Paris, De Boeck et Larcier.

Dehlert G.W. (2000). *A first course in Design and Analysis of Experiments*, Freeman and company, NY, USA.

Fonton N. H., Glèlè Kakaï R., Rondeux J. (2002). Etude dendrométrique de *Acacia auriculiformis (Cunn A.)* en mélange sur du vertisol au Bénin. *Biotechnol. Agron. Soc. Environ.* 6 (1) : *29-37.*

Glèlè Kakaï R**.**, Palm. R. (2006). Methodological contribution to control heteroscedasticity in discriminant analysis studies *Global Journal of Pure and Applied Sciences*, 12 (1), 107-110.

Glèlè Kakaï R**.**, Palm. R. (2005). Minimal Error Rate Of Linear, Quadratic And Logistic Rules In Discriminant Analysis. *Global Journal Of Mathematical Sciences.* 4 (1 , 2), 89-93.

Glèlè Kakaï R**.**, Palm R. (2004). Performance relative des règles linéaire, quadratique et logistique en analyse discriminante. *35ème journées françaises de statistique,* 24 – 28 mai 2004*,* Montpellier, France (CD Rom).Website : http://www.agro-montpellier.fr/sfds/CD/textes/glele1.pdf

Glèlè Kakaï R**.**, Palm R., Kokode G. (2005). L'analyse discriminante décisionnelle : aspects théoriques et applications sur ordinateur. *Notes tech. Biom*. Bibliothèque Nationale, Bénin.

Greenstreet R.L. et Connor R.J. (1974). Power of tests for equality of covariance matrices. *Technometrics,* 16 : 27-30.

McLachlan G. J. (1992). *Discriminant analysis and statistical pattern recognition*, Wiley, New York.

Lachenbruch P. A., Sneeringer C., Revo L. T. (1973). Robustess of the linear and quadratique discriminant function to certain types of non-normality. *Comm. Stat.*, **1**, 39-57.

Layard M.W.J. (1974). A Monte Carlo comparison of tests for equality of covariance matrices. *Biometrika*, 16 : 461-465.

Levene H. (1960). Contributions to Probability and Statistics, pp.278-292.

Stanford University Press, CA.

Mardia K. V. (1980). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika*, 57(3): 519-530

Minitab (1996). *Minitab for Windows Release 11*. Minitab, Inc., State College, PA, USA.

O'Brien R.G. (1979). A General ANOVA Method for Robust Tests of Additive Models for Variances. *Journal of the American Statistical Association,* 74 : 877 –880.

Olejnik S.F., Algina J. (1987). Type I Error Rates and Power Estimates of Selected Parametric and Non-parametric Tests of Scale. *Journal of Educational Statistics*, 12 : 45 -61.

Owen D.B. (1962). *Handbook of statistical tables*. Reading, Addison-Wesley.

Palm R. (1994). La régression linéaire pondérée : principes et application. *Notes Stat. Inform.* 94/4 (Gembloux).

Ryan T.A., Joiner B.L. (1976). Normal Probability Plots and Tests for Normality. Technical Report, Statistics Department, The Pennsylvania State University.

Saporta G. (1990). *Probabilités analyse des données et statistique*. Technip, Paris.

SAS Institute Inc. (1999). SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc.

Shapiro S.S, Wilk M.B. (1965). An analysis of variance test for normality. *Biometrika*, 52(3) : 591-599.

Tomassone R., Donzart M., Daudin J. J., Masson J. P. (1988). *Discrimination et classement*, Masson, Paris.

Thursby J. (1982). Misspecification, Heteroscedasticity, and the Chow and Goldfield-Quandt Test. *Review of Economics and Statistics*, 64: 314-321.

White H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48 (4): 817-838.