REPUBLIQUE DU CAMEROUN
Paix – Travail – Patrie
****

UNIVERSITE DE YAOUNDE I
ECOLE NORMALE SUPERIEURE
DEPARTEMENT DE Mathematiques
****

REPUBLIC OF CAMEROUN
Peace – Work – Fatherland
****

UNIVERSITY OF YAOUNDE I
HIGHER TEACHER TRAINING COLLEGE
DEPARTMENT OF Mathematics
****

Indice de gini et evaluation de la regularite de la reproduction chez le palmier a huile

Mémoire de D.I.P.E.S II de Mathematiques

Par :

DJOU MBOU Ferdinand Georges
Licencie en Mathematiques et applications

Sous la direction
NKAGUE NKAMBA Leontine
Chargee de cours

Année Académique
2015-2016

♠ Dédicace ♠

Je dédie ce travail à

ma feue mère : PIEBENG Lydie

♠ Remerciements ♠

Tout seul, on ne peut arriver à bâtir.

J'adresse ici mes sincères remerciements et ma profonde gratitude à tous ceux qui m'ont porté jusque là par leur amour, leur amitié, leurs enseignements, leurs conseils, leurs encouragements, leurs aides et leurs reproches. Ma profonde reconnaissance va tout d'abord :

♡ Au souverain Seigneur dieu , pour son amour, sa miséricorde, sa bonté, pour la protection qu'il m'a accordé durant toutes ces deux années. Merci Seigneur pour tous ces dons.

Mes remerciements s'adressent également de façon particulière :

♡ À mes encadreurs, docteur **Leontine NKAGUE NKAMBA** et le docteur **David CROS** qui ont de bon gré accepté de diriger ce travail, malgré leurs occupations, ont soutenu mes efforts jusqu'au bout. Vous vous êtes révélé réellement présents et surtout ouverts et n'avez ménagé aucun effort pour que ce travail puisse être effectué dans les délais. Merci grandement.

♡ **À tous les enseignants de mathématiques de l'école normale supérieure de Yaoundé** , pour les enseignements et le suivi qu'ils m'ont apportés durant ces deux années passées à l'école normale. Merci beaucoup.

♡ Docteur **Isidore Seraphin NGONGO** pour les encouragéments et le soutien accordés.

♡ À tous mes frères et soeurs : **Gerard PONSONG, Ghislain DOUANLA, Rinos DJOUDA, Viviane MBOU,Gael MBOU, Blandine DJAFIA, Darios KUETE, Aline YONTA, Sylviane YONTA, Christian YONTA, Camel ADJOU, Valdes FOMEKONG, Sylvestre TANEKEM, Mme Aristide LONTSI**, je n'oublie pas l'amour et le réconfort que vous m'avez toujours témoigné. Merci à vous tous.

♡ À ma grande soeur **Mme Marie YONTA** et son époux **Mr Martin YONTA** pour L'encadrement depuis ma naissance, la grande attention que vous m'accordez et pour votre soutient sans cesse, merci pour vos conseils et vos encouragements.

♡ À **Mme Odette FOLEM** pour son amour son encadrement social durant ces deux années

♠ Déclaration sur l'honneur ♠

Le présent document est une œuvre originale du candidat et n'a été soumis nulle part ailleurs en partie ou en totalité, pour une autre évaluation académique. Les contributions externes ont été dûment mentionnées et recensées en bibliographie.

Signature du candidat

DJOU MBOU Ferdinand Georges

♠ Table des matières ♠

♠ Résumé ♠

Le palmier à huile produit des régimes tout au long de l'année. Cette production est marquée par une saison de pointe (période d'abondance de récolte des régimes). Le but de cette étude est de pouvoir améliorer la régularité de la production le long de l'année. Notre étude porte sur des données issues d'un ensemble de croisements de palmiers à huile. Nous quantifierons la régularité de la production par l'indice de Gini de chaque parcelle élémentaire. Nous déterminerons le nombre d'individus nécessaire pour que l'indice de Gini soit représentatif de la répartition de la production à l'échelle de la plantation. Nous avons ainsi montré qu'il fallait au moins 10 palmiers pour obtenir un indice de Gini représentatif. Nous avons ensuite vérifié que, dans ces conditions, l'indice de Gini était un bon indicateur pour mesurer la régularité de la production tout au long de l'année, en notant sur quelques parcelles extrêmes la bonne correspondance entre les valeurs de l'indice de Gini et les profils de production sur l'année. Nous avons aussi mis en évidence une grande variabilité dans l'indice de Gini, ce qui laisse espérer des possibilités d'amélioration génétique. Enfin, par des tests statistiques nous avons vu que l'indice de Gini décroit légèrement mais significativement après 3 ans avant de se stabiliser, et qu'il existe une légère mais significative corrélation négative entre l'indice de Gini et la production annuelle de régimes, ce qui est favorable à la filière.

Mots clés : Indice de Gini, régularité, palmier à huile, production de régimes et mesure d'inégalités

♠ Abstract ♠

The oil palm bears bunches all year round. This growth is characterised by a peak season (period of abundant harvest). The aim of this study is to improve the regularity of production along the year. Our study is based on data obtained from a set of crosses of oil palm individuals. We will measure the regularity of production using the Gini index of each plot. We will determine the number of individuals necessary to obtain a value of the Gini index representative at the the scale of the plantation. Thus we proved that at least 10 palm individuals were required to obtain a relevant Gini index. We then verified that under these conditions, the Gini index was a good indicator to measure the regularity of production throughout the year, as on extreme plots there was a the good relationship between the values of the Gini index and production profiles in a year. We also found a high variability in the Gini index, which gives hope on the possibility of genetic improvement. Finally, through statistical tests we saw that the Gini index decreases slowly but significantly after 3 years before stabililizing and that there is a small but significant negative correlation between Gini index and annual production of oil palm bunches, which is favourable for the sector.

Keys words : Gini index, regularity, oil palm trees, oil palm bunches, measures of inequality

♠ Table des figures ♠

♠ Introduction ♠

La production de régimes chez le palmier à huile suit des cycles annuels avec une période marquée par un pic de production où le rendement est plus élevé. La régularité de la production sur l'année a une grande importance économique. Elle détermine tout d'abord la répartition le long de l'année de la charge de travail (récolte, transport et traitement des régimes). Une charge de travail trop grande à une période donnée peut avoir plusieurs impacts négatifs. En effet, elle peut amener le producteur, en particulier les petits producteurs, à récolter trop tard une partie des régimes ce qui accroit le pourcentage de fruits détachés. Ceci génère des coûts additionnels liés au ramassage de ces fruits, et/ou réduit le rendement. Dans tous les cas, cela accroit l'acidité de l'huile, ce qui en réduit la qualité. Pour les petits producteurs qui commercialisent leurs régimes à une structure d'extraction d'huile, la répartition de la production de régimes conditionne la régularité des revenus. Pour les industriels, elle détermine la capacité de l'usine de traitement des régimes (qui doit être suffisamment grande pour traiter tous les régimes livrés en période de pic de production), et donc la taille de l'investissement. Le développement de variétés de palmier à huile avec une production de régimes régulière tout au long de l'année serait donc très favorable à la filière. Cependant, aucune étude de génétique n'a été publiée à ce sujet. Un tel travail nécessite tout d'abord un indicateur numérique permettant de quantifier la régularité de la production sur l'année.

L'objectif de l'étude présente ici est d'évaluer le potentiel de l'indice de Gini, suggéré par Cros et al (2013), comme mesure de la régularité de la production de régimes chez le palmier à huile. Dans une première partie, Notre travail consistera à calculer l'indice de Gini représentatif d'un champ de palmiers à huile pour évaluer la régularité de sa production tout au long de l'année. Pour y arriver, nous calculerons l'indice de Gini de toutes les parcelles disponibles et nous sélectionnons des parcelles ayant des indices de Gini contrastés forts, médians, ou faibles. A partir de là, on identifie un nombre seuil d'individus par parcelle qui permet de calculer un indice de Gini représentatif de ce qui serait observé sur une grande superficie. Par la suite, on calcule uniquement l'indice de Gini sur les parcelles

dont le nombre de palmier est supérieur ou égal à ce seuil. On vérifiera alors que l'indice de Gini reflète efficacement la répartition en mettant en relation le profil de production et l'indice de Gini. Enfin on va mesurer la variabilité de l'indice de Gini au sein de la population étudiée.

Dans le premier chapitre nous présenterons la revue de la littérature, qui se subdivisera en deux grandes parties à savoir : la présentation du palmier à huile et les outils mathématiques nécessaires pour notre étude. Au deuxième chapitre nous présenterons les matériels et méthodes utilisés, avec en particulier une application numérique détaillant le calcul de l'indice de Gini pour une parcelle choisie à titre d'exemple. Enfin au chapitre trois nous présenterons les résultats et interprétations des analyses faites sur l'ensemble des données.

⋆ ⋆ Chapitre Un ⋆ ⋆

Revue de la littérature

1.1 Le palmier à huile

Le palmier à huile pousse dans les régions situées autour de l'Équateur. C'est une herbe géante (voir figure1.1) tropicale avec une production de régimes qui peut, dans de bonnes conditions, s'étaler sur toute l'année. Originaire d'Afrique, le palmier à huile est désormais principalement cultivé en Indonésie et en Malaisie, principaux pays producteurs au monde.

1.1.1 Filière du palmier à huile

Le palmier à huile est aujourd'hui la première plante oléagineuse au monde en termes de production. La production d'huile de palme à été multipliée par 3,8 entre 1990 et 2010 et elle dépasse aujourd'hui 55 Mt (USDA, 2014). On s'attend à ce qu'elle continue d'augmenter très fortement car la demande devrait se situer entre 120 et 156 Mt en 2050 (Corley, 2009). L'Asie du Sud-est réalise aujourd'hui l'essentiel de la production.

Le palmier à huile est aussi la première plante oléagineuse pour le rendement à l'hectare. Sa surface cultivée représente seulement 7% des surfaces mondiales en oléagineux mais réalise environ 39% de la production. Le rendement moyen du palmier à huile dans le monde atteint presque 4 tonnes d'huile par hectare et par an, soit environ 10 fois plus que le soja et quatre fois plus que le colza. Dans les environnements favorables, les plantations les plus performantes produisent plus de 6 t/ha sur plusieurs dizaines de milliers d'hectares et les meilleurs croisements évalués en essais génétiques dépassent 10 t/ha. Par ailleurs, l'huile de palme est l'huile végétale la moins couteuse à produire, avec des coûts de production inférieurs de 20% à ceux du soja, et elle peut se substituer à la plupart des autres huiles végétales (Fonds français pour l'alimentation et la santé, 2012).

L'huile de palme est utilisée à 80% dans l'alimentation humaine (huile de table, huile de friture, margarine, etc.) et à 20% dans l'industrie (savonnerie, cosmétiques, lubrifiants, etc.). Environ 1% de

l'huile de palme est utilisée pour produire du biodiesel. L'huile de palme brute ou raffinée contient quasiment 100% de lipides sous forme principalement de triglycérides, constitués d'un glycérol auquel sont fixés trois acides gras. La part des acides gras saturés, acide palmitique en tête, est d'environ 50%. De ce fait, son intérêt nutritionnel fait débat (Fonds français pour l'alimentation et la santé, 2012). Cependant, la présence d'huile de palme dans un régime alimentaire équilibré ne semble pas poser de problème de santé. Les principaux consommateurs sont des pays émergents. Dans certains pays, en particulier en Afrique, l'huile de palme est la principale source de corps gras dans le régime alimentaire. Elle joue alors un rôle majeur dans les apports lipidiques, énergétiques et vitaminiques. Le palmier à huile est un fort enjeu de développement pour de nombreux pays du Sud. Quand il est correctement planifié par les gouvernements et mis en œuvre par les planteurs, le développement du palmier à huile se traduit par un fort développement économique des régions concernées et par une importante réduction de la pauvreté rurale. Son exploitation repose sur des systèmes de culture très diversifiés allant de l'exploitation familiale de quelques hectares au périmètre agroindustriel de plusieurs dizaines (voir centaines) de milliers d'hectares. Plus de la moitié de l'huile de palme produite aujourd'hui provient de petites exploitations, au nombre d'environ trois millions. La culture du palmier à huile est capable de générer des revenus élevés et stables.

En 2009-2010, la production mondiale d'huiles végétales commercialisées dépassait 138 millions de tonnes (MT), soit une croissance moyenne de 4,7% pendant la décennie précédente. Ces quantités ne prennent pas en compte la production auto consommée. L'huile de palme occupe la première place avec 47,5 MT soit (34 %) de consommation, devant l'huile de soja avec 37,9 MT soit (27%) de consommation. L'huile de colza se classe en troisième position avec 22,1 MT soit (16%) précédant l'huile de tournesol avec 11,3 MT soit 8% (figure1.2). Il faut noter que plus de 45% de la production mondiale d'huiles végétales comestibles provient d'Asie. Au Cameroun, la production d'huile de palme brute est assurée à hauteur de plus de 60% par les planteurs industriels et le reste par les plantations villageoises. La production totale d'huile de palme brute au Cameroun est estimée à plus ou moins 210 000 tonnes. Le segment des agro-industries est dominé par cinq acteurs majeurs. Leur production estimée pour l'année 2008 se cumule à hauteur de 145 000 tonnes. Elle se répartit comme suit :

Dénomination sociale	Localisation	Production/an (Tonnes)
SOCAPALM	MBONGO, NKAPA, KIENKE, ESEKA	83 000
CDC	LIMBE, IDENAU	18 000
SPFS	APOUH (EDEA)	15 000
SAFACAM	DIZANGUE (EDEA)	12 000
PAMOL	LOBE	16 000

Source : SNPHC (Syndicat National des Producteurs d'huile de palme)

FIGURE 1.1 – Palmier à huile

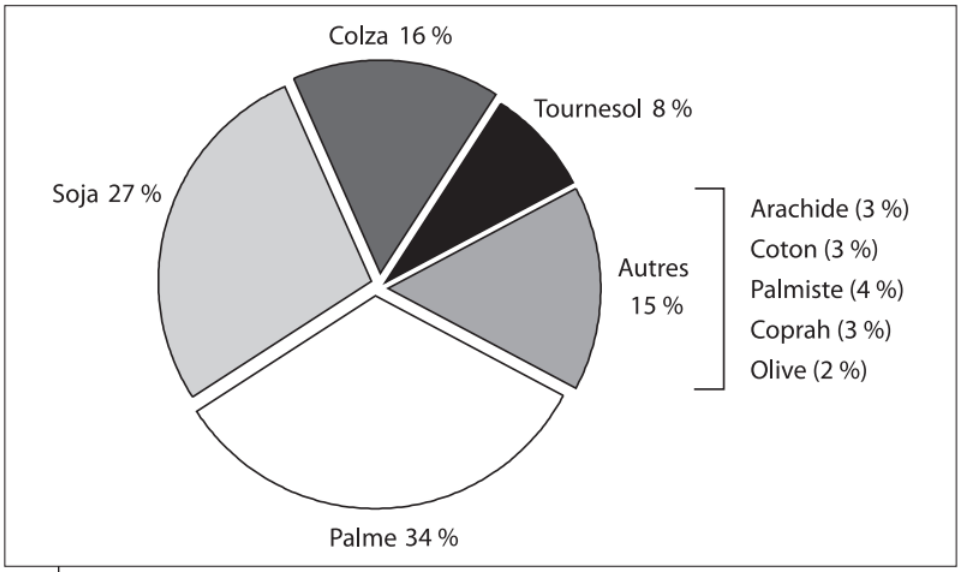


FIGURE 1.2 – Consommation mondiale d'huile végétale (2009-2010)

On peut observer la production de l'huile végétale en 2012 dans le diagramme suivant.

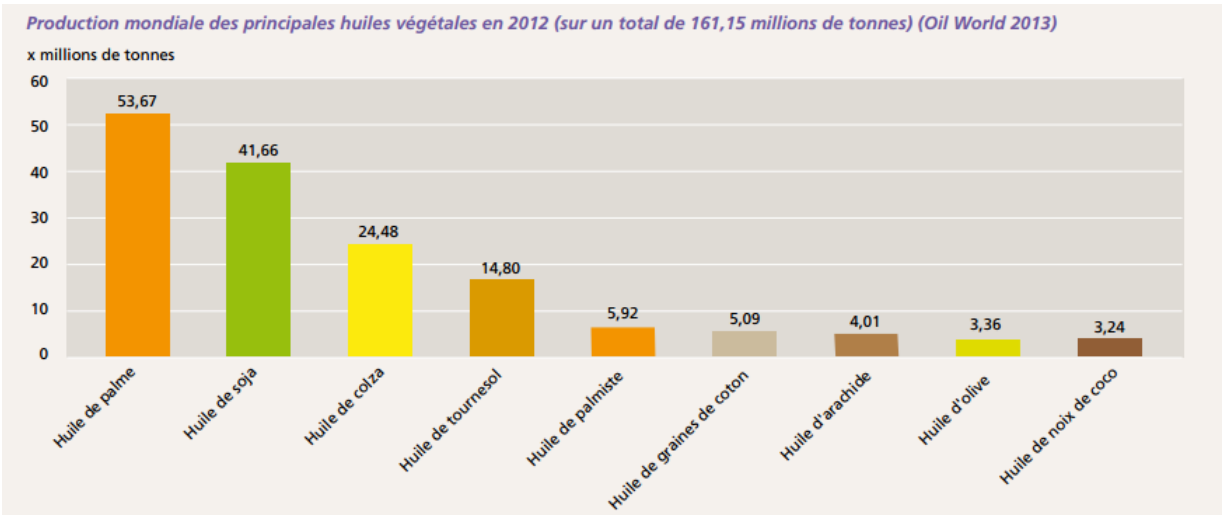


FIGURE 1.3 – Production mondiale des principales huiles végétales en 2012

1.1.2 Caractéristiques générales de la plante

Le palmier à huile est une herbe géante qui produit tout au long de l'année des feuilles, entourant le bourgeon végétatif pour former la couronne. Les feuilles mesurent 6 à 9 mètres et sont composées de plus de 300 folioles. A l'aisselle de chaque feuille se trouve une inflorescence dont le devenir dépend des conditions environnementales au cours de son développement (en particulier du bilan hydrique) et des cycles sexuels endogènes du palmier. Une inflorescence pourra avorter ou devenir mâle ou femelle. Une

fois fécondées, les inflorescences femelles évoluent normalement en régimes. Un régime est constitué d'un rachis (ou pédoncule) portant des épillets, sur lesquels se trouvent les drupes (fruits à noyaux). Un régime pèse entre 5 et 50 kg et contient 500 à 4 000 drupes, selon l'âge du palmier, sa population d'origine, son environnement, etc. Un fruit pèse entre 10 et 30 g et se compose généralement d'une amande (faite d'un embryon et d'albumen), d'un endocarpe ligneux (coque), de mésocarpe (pulpe) et d'un exocarpe (peau) (figure1.5). La pulpe des fruits fournit l'huile de palme et l'amande l'huile de palmiste, dont il ne sera pas question ici.
Chez le palmier à huile coexistent trois types, définis par la morphologie interne de leurs fruits :

- **Le dura :** il s'agit du type prépondérant dans la nature (>90%). Ses fruits possèdent une coque épaisse (de 2.5 à 7 mm) et, par conséquent, un pourcentage de pulpe assez faible (figure1.4).
- **Le pisifera :** il est très rare dans la nature (<5%). Ses fruits sont dépourvus de coque et sa pulpe renferme des fibres lignifiées qui, lors d'une coupe transversale du fruit, forment un anneau autour de l'amande. Les pisifera sont généralement improductifs car leurs régimes avortent avant maturité. Les fruits de pisifera sont donc très rares mais lorsqu'ils existent ils possèdent un pourcentage de pulpe très élevé (figure1.4).
- **Le tenera :** il est très rare dans la nature (<5%). Ses fruits possèdent une coque de faible épaisseur (<2 mm), et un anneau de fibres lignifiées dans la pulpe, autour du noyau. (figure1.4)

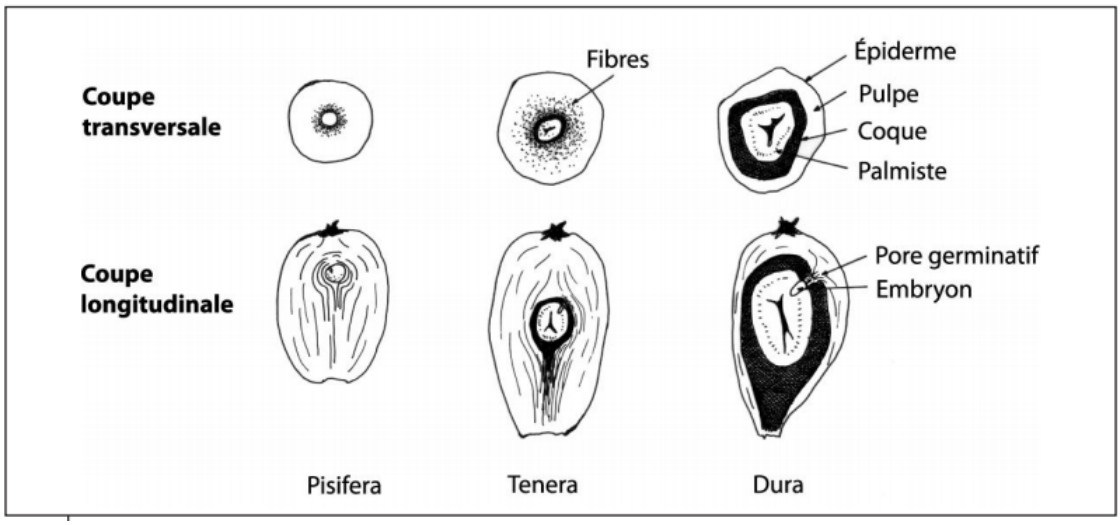


FIGURE 1.4 – Type de fruits produits par le palmier à huile

FIGURE 1.5 – Régime de fruit de type tenera

1.1.3 Les populations de palmier à huile

Bien que l'utilisation du palmier à huile par les populations d'Afrique subsaharienne soit ancestrale, cette espèce n'a pas subi une domestication marquée et il n'existe pas de types distincts « sauvage » et « cultivé ».

Le palmier à huile a été introduit en Asie du Sud-est en 1848, avec quatre individus dura plantés dans le jardin botanique de Bogor (Java, Indonésie) à des fins ornementales (Corley et Tinker, 2003). Leur origine exacte reste inconnue mais la population asiatique qui en a découlé (Deli) a des caractéristiques proches de celles des populations d'Afrique centrale.

Les principales populations créées en Afrique sont La Mé (Côte d'Ivoire) et Yangambi (RDC). On trouve aussi les populations Yocoboué (Côte d'Ivoire), Sibiti (République du Congo), Ekona (Cameroun), WAIFOR (Nigéria) et Pobè (Bénin) (Nouy et al, 1999) mais elles sont beaucoup moins utilisées. La population La Mé trouve son origine dans les prospections faites dans la région de Bingerville dans les années 1920. Ceci a abouti à la sélection de 19 individus choisis car leurs fruits possédaient des proportions équilibrées entre le mésocarpe (60%), l'amande (20%) et la coque (20%). La population Yangambi est issue de plantations faites dans les années 1920 à partir de 10 à 20 tenera en pollinisation libre, incluant Djongo (« le meilleur ») du jardin botanique d'Eala et des tenera de Yawenda, Ngazi et Isangi.

En Asie, les quatre palmiers de 1848 ont donné naissance à la population Deli, dans laquelle on distingue aujourd'hui plusieurs sous populations, principalement Marihat Baris en Indonésie, SOCFIN en Indonésie et Malaisie, Serdang Avenue, Ulu Remis (ou Guthrie), Johor Labis et Elmina (dont les Dumpy) en Malaisie. Les premières activités connues de sélection de la population Deli pour le

rendement en huile à partir d'observations rigoureuses datent des années 1910-1930, selon les sociétés de plantation (Corley et Tinker, 2003 ; Cochard, 2008). Les détails concernant cette période sont incertains (caractères sélectionnés, intensité de sélection, etc).

Les populations de palmier à huile peuvent se répartir en deux groupes A et B selon les caractéristiques de production de leurs régimes (Gascon et de Berchoux, 1964). Le groupe A produit des régimes plus gros que le groupe B mais le groupe B produit un plus grand nombre de régimes. Le groupe A est composé des populations Deli et Angola, le groupe B des autres populations africaines. On peut à nouveau faire des distinctions entre populations du groupe B sur la base du phénotype, avec La Mé caractérisé par un nombre très faible de régimes et Yangambi par des régimes relativement gros.

Les palmiers à huile distribués aux planteurs par les producteurs de semences sont des tenera hybrides entre A et B, en général des croisements **Deli x La Mé** en Afrique.

1.1.4 Répartition de la production

1. **La sexualité du palmier à huile, *Elaeis guineensis* Jacq.**

   On connaît chez les plantes des fleurs mâles, des fleurs femelles ou des fleurs bisexuées, sur la même plante ou sur des plantes différentes, la maturité sexuelle apparaissant simultanément pour les deux sexes ou à des périodes différentes. Dans tous les cas, le but est de perpétuer l'espèce. Dans la grande majorité des plantes, de l'initiation florale à la graine, le processus est annuel et peut être fortement influencé par certains paramètres climatiques. On connaît l'importance de certaines pluies sur l'apparition des fruits à noyaux ainsi que les ravages de certaines gelées. Lorsqu'elles réalisent qu'elles vont mourir, certaines plantes utilisent leurs dernières énergies pour produire le maximum de graines ; ce phénomène est bien connu des planteurs dans les plantations d'*Hevea brasiliensis*. L'objectif est, plus que jamais, la pérennité de l'espèce.

   Chez le palmier à huile, entre l'initiation florale et la maturité du régime de graines, il se passe plus de trois années pendant lesquelles la plante subit les aléas du milieu et pendant lesquelles les stades de la maturation sexuelle s'alignent sur ceux du développement anatomique de la feuille. Les stades de développement d'une inflorescence chez le palmier à huile sont les suivants :

   - **l'initiation florale**, l'apparition du bourgeon axillaire, la formation des pseudo-folioles et l'apparition de la spathe extérieure qui emballera l'inflorescence (entre 35 et 45 mois avant la maturation du régime) ;
   - **la différenciation sexuelle** apparaît entre 25 et 30 mois avant la maturation du régime ;

- **l'élongation de la feuille** est assez rapide entre les 16ème et 19ème mois avant la maturation du régime ; on observe également pendant cette période l'élongation assez rapide des spathes internes et externes ;
- **l'élongation de l'inflorescence**, qu'elle soit mâle ou femelle, a lieu environ 13 mois avant la maturité du régime et est suivie ou non par une période d'avortements vers 10 semaines avant cette maturité ;
- **la floraison des fleurs femelles** et leur fécondation par le pollen des fleurs mâles d'un autre individu se déroulent environ six mois avant la maturité des régimes.

2. **Cycles de production chez le palmier à huile**

Etant donné que l'émission de nouvelles feuilles se fait potentiellement toute l'année, et que chacune de ces feuilles peut porter une inflorescence femelle, la production de régimes du palmier à huile peut-être permanente, avec une récolte tous les 15 jours. Cependant, il existe des cycles d'émission et de développement des feuilles, ainsi que des cycles de développement des inflorescences. On observe d'ailleurs au niveau des individus des périodes plus ou moins longues pendant lesquelles toutes les inflorescences qui apparaissent successivement sont du même sexe. Ces cycles ont deux origines : internes (cycles endogènes) ou externes à la plante (cycles exogènes liés aux conditions environnementales). Le principal facteur environnemental affectant le développement des inflorescences est la pluviométrie. Nouy et al (1996) ont comparé plusieurs croisements dans des environnements présentant des déficits hydriques variables. Ils ont noté que, lorsque le déficit hydrique augmente, la récolte des régimes est de plus en plus groupée sur des périodes de plus en plus courtes. Plusieurs études sur ce sujet ont été réalisées au Bénin, où la culture du palmier à huile est marquée par des saisons très contrastées, avec une grande saison sèche très prononcée (Olivin, 1966 ; Nouy et al, 1996 ; Cros et al, 2013). Les saisons sèches conduisent à une relative concentration des apparitions puis des ouvertures de feuilles. Elles réduisent surtout le sexe-ratio (% inflorescences femelles/ total des inflorescences) et provoquent des avortements d'ébauches florales. Ces deux mécanismes, cumulés aux variations de durée entre l'ouverture des feuilles et la floraison, font apparaitre des périodes improductives et, par conséquent, un pic de floraisons. Par la suite, les variations de la durée de développement des régimes amplifient encore la concentration des récoltes. Enfin, les variations au cours de l'année de la charge en régimes ainsi créées viennent elles aussi contribuer à ces cycles : une forte charge en régimes à une période donnée aboutit à une réduction du sexe ratio et à une augmentation des avortements. Il existe cependant une variabilité relativement forte entre croisements dans le

profil de répartition de la production (Nouy et al, 1996 ; Cros et al, 2013), qui s'explique par des différences de réponse selon le croisement face aux variations du bilan hydrique et de la charge en régimes.

3. **Pratiques agronomiques résultant des cycles de production du palmier à huile.** Compte tenu de la biologie du développement des inflorescences du palmier à huile, la production de régimes est étalée, au moins dans de bonnes conditions climatiques, sur toute l'année. En conséquence, une plantation de palmiers à huile est visitée tous les 10-15 jours pour récolter les régimes matures. La castration des jeunes palmiers dès l'apparition des premières inflorescences durant une période de 6 mois à un an, soit des inflorescences mâles seules ou de toutes les inflorescences, a pour objectifs de réduire les besoins énergétiques du palmier et de favoriser la production de matières sèches végétales, tout en induisant de longs cycles femelles en début de production. Ceci rend le jeune palmier plus vigoureux, et les premiers régimes récoltés sont plus gros et plus facilement usinables. Cette importante récolte redevient normale après un certain temps quand on retrouve la succession des cycles mâles et femelles. La sélection de très bons producteurs à très haut sexe-ratio engendre parfois des situations où les fécondations sont déficientes. Dans ce cas (notamment en Malaisie) on a recours à la fécondation assistée avec du pollen récolté sur des palmiers à longs cycles mâles.

1.2 Outils mathématiques

Pour mener à bien notre étude, nous aurons besoin des outils mathématiques en statistique descriptive, et en analyse combinatoire.

1.2.1 Analyse combinatoire

- Le nombre d'applications d'un ensemble à p éléments vers un ensemble à n éléments est n^p.
- Le nombre de permutations d'un ensemble à n éléments de bijections de cet ensemble dans lui-même est $n!$.
- Le nombre d'arrangements d'injections d'un ensemble à p éléments dans un ensemble à n éléments est $A_n^p = \dfrac{n!}{(n-p)!}$
- Le nombre de combinaisons ou sous-ensembles à p éléments dans un ensemble à n éléments ($\geq p$) est $C_n^p = \dfrac{n!}{p!(n-p)!}$

Tirages

Considérons l'ensemble **E** comme étant une urne contenant n boules numérotées de 1 à n, chacune des boules s'interprètant comme un élément de **E**. On tire p boules de **E**. On a les cas suivants :

- **Tirages successifs avec remise :** On tire au hasard une boule dans l'urne puis on la remet dans l'urne avant d'effectuer le tirage suivant. Si on effectue ainsi p tirages avec remise, le résultat global s'interprète comme une $p-liste$. Il y a donc n^p tirages avec remise (de p éléments) possibles.
- **Tirages successifs sans remise :** On tire au hasard une boule dans l'urne que l'on conserve, la boule n'est donc pas remise dans l'urne qui contient ainsi après chaque tirage une boule de moins. Si on effectue ainsi **p** tirages sans remise ($p \leq n$), le résultat global s'interprète comme une $p-liste$ d'éléments 2 à 2 distincts ou encore comme un arrangement de p éléments de **E**. Il y a donc A_n^p tirages sans remise (de p éléments) possibles.
- **Tirages simultanés :** On tire simultanément p boules de l'urne (et non plus successivement, cela revient à dire que l'ordre du tirage des boules est sans importance). Un tel tirage s'interprète comme un sous ensemble de **E** et donc comme une combinaison de p éléments de **E**. Il y a donc C_n^p tirages simultanés (de p éléments) possibles.

1.2.2 Quelques notions de statistique

Données quantitatives d'un caractère

Définition 1.1. Soit $x = (x_1, ..., x_n)$ une suite de données quantitatives. **L'étendue** est la différence des valeurs extrêmes.

Pour $x \in \mathbb{R}$ **l'effectif cumulé** de x est le cardinal $k(x) = \text{card}\{i, x_i \leq x\}$.

la fréquence cumulée $F(x)$ est $F(x) = \dfrac{k(x)}{n}$.

La médiane x_{med} divise les données en deux parties égales, celles qui lui sont plus petites et celles qui lui sont plus grandes. On a $F(x_{med}) = \dfrac{1}{2}$. On définit aussi les trois quartiles q_1, $q_2 = x_{med}$ et q_3 par $F(q_i) = \dfrac{i}{4}$.

La moyenne $\overline{x}$ de x est : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{n}(x_1 + x_2 + ... + x_n)$

On appelle variance empirique le nombre positif

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

où $\overline{x^2}$ désigne la moyenne de la suite des carrés : $x^2 = \left(x_1^2 + x_2^2 + ... + x_n^2\right)$ et $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

L'écart-type est : $\sigma = \sqrt{V}$ (Yadolah Dodge, 2007)

Remarque 1.2.1. La variance et l'écart-type mesurent comment la population se disperse par rapport à la moyenne.

Remarque 1.2.2. Si, au lieu de donner la valeur x_i du caractère de chaque individu i on donne pour chaque valeur c_j prise par le caractère le nombre n_j d'individus alors on définit la fréquence de la valeur c_j par $f_j = \frac{n_j}{n}$ et la fréquence cumulée F_j de la valeur c_j comme la somme des fréquences des valeurs inférieures ou égales à c_j. La moyenne $\overline{x}$ se calcule par la formule

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{d} n_j c_j = \frac{1}{n}\left(n_1 c_1 + n_2 c_2 + ... + n_d c_d\right) = \left(f_1 c_1 + f_2 c_2 + ... + f_d c_d\right)$$

où d est le nombre de valeurs différentes prises par le caractère. La variance empirique est donnée par les formules :

$$V = \frac{1}{n}\sum_{j=1}^{d} n_j \left(c_j - \overline{x}\right)^2 = \sum_{j=1}^{d} f_j \left(c_j - \overline{x}\right)^2 = \left(\frac{1}{n}\sum_{j=1}^{d} n_j c_j^2\right) - \overline{x}^2 = \left(\sum_{j=1}^{d} f_j c_j^2\right) - \overline{x}^2$$

En pratique on utilise les formules avec les c_i.

Remarque 1.2.3. On divise souvent l'étendue en intervalles appelés classes. On peut s'intéresser à l'effectif n_j, à l'effectif cumulé N_j, à la fréquence f_j et à la fréquence cumulée F_j d'une classe $[a_j, a_{j+1}[$. L'effectif cumulé N_j est égal à $n_1 + ... + n_j$. La fréquence f_j est égale à $\frac{n_j}{n}$, la fréquence cumulée Fj est la somme $f_1 + ... + f_j$. On calcule la moyenne et la variance en utilisant les formules précédentes avec pour n_j l'effectif de la classe et $c_j = \frac{a_j + a_{j+1}}{2}$.
Pour calculer avec précision la médiane on repère d'abord la classe médiane. C'est la classe dont la fréquence cumulée $F_{j_{med}}$ est la première à passer $\frac{1}{2}$. On calcule alors x_{med} par interpolation linéaire en posant :

$$x_{med} = a_{j_{med}} + \left(\frac{a_{j_{med}+1} - a_{j_{med}}}{F_{j_{med}} - F_{j_{med}-1}}\right)\left(\frac{1}{2} - F_{j_{med}-1}\right)$$

On fait de même pour les quartiles en remplaçant $\frac{1}{2}$ par $\frac{1}{4}$ (q_1) ou $\frac{3}{4}$ (q_3)

Données quantitatives de deux caractères

Définition 1.2. Soit $(x, y) = \left((x_1, y_1), \cdots, (x_n, y_n)\right)$ une suite de données quantitatives couplées.

– **La covariance**

La covariance est : $cov\,(x,y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y}) = \dfrac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{xy}$

– **Le coefficient de corrélation**

L'indicateur de liaison approprié dans le cas de deux variables quantitatives est la corrélation. Il est défini comme le rapport entre la covariance des deux variables et le produit de leurs écarts-types respectifs. Ce rapport est :

$$r = \frac{cov\,(x,y)}{\sigma_x\sigma_y} \in [-1,1]$$

– **Régression linéaire**

Les paramètres de régression pour un modèle en ligne droite ($Y = aX + b$) sont calculées par la méthode des moindres carrés. Les formules suivantes donnent la pente a et l'ordonnée à l'origine b de la droite :
$a = \dfrac{cov(x,y)}{V(x)}$ et $b = \overline{y} - a\overline{x}$ car cette droite passe par le point moyen $G = (\overline{x},\overline{y})$
Ainsi la droite de régression linéaire est :

$$Y = \frac{cov(x,y)}{V(x)}X + \overline{y} - \frac{cov(x,y)}{V(x)}\overline{x}$$

1.2.3 Indice de Gini et courbe de Lorenz

Définition 1.3. on considère une série statistique de forme générale $(x_i, n_i)_{1\leq i\leq p}$

– la fréquence cumulée croissante $p_i = \sum_{j\leq i} f_j$ où f_j est la fréquence de chaque modalité x_i
– le coefficient q_i est le rapport entre la masse de la modalité cumulée divisée par la masse de la modalité totale $M = \sum_{j\leq p} n_jx_j$: $\qquad q_i = \dfrac{\sum_{j\leq p} n_jx_j}{M}$

Définition 1.4. L'indice de Gini est le double de l'aire compris entre la courbe de Lorentz et la première bissectrice noté le plus souvent **G** (dasgupta et al, 1973).

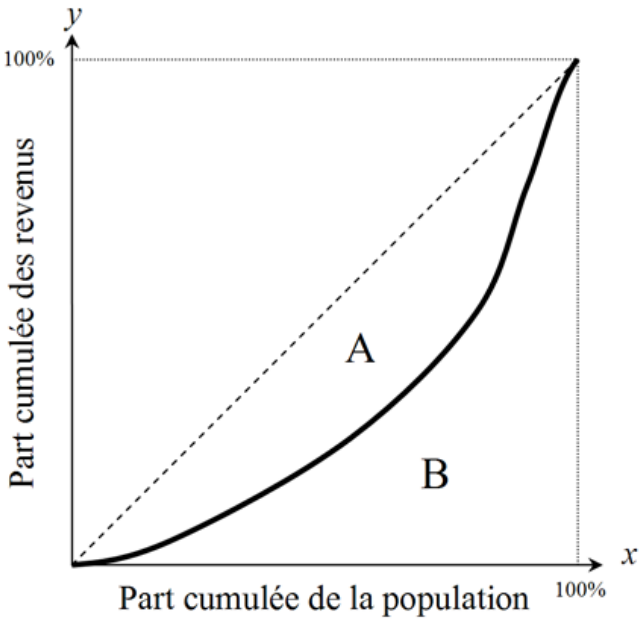


FIGURE 1.6 – Courbe de Lorenz

Le coefficient de Gini est égal : à la différence entre 1 et le double de l'intégrale de la fonction représentée par la courbe de Lorenz.

L'indice de Gini (G) est une mesure statistique de la dispersion d'une distribution dans une population donnée, développée par le statisticien italien **Corrado Gini**. Le coefficient de Gini est un nombre variant de 0 à 1, où 0 signifie l'égalité parfaite et 1 signifie une inégalité parfaite.

Le plus souvent, G est utilisé pour évaluer les inégalités des revenus dans un pays à différentes époques ou entre différents pays à la même époque.

Durant les dernières années, G a été utilisé dans d'autres domaines que économique : biologie (Graczyk et Piotr P, 2007), environnement (Druckman and Jackson, 2008 ; Groves-Kirkby et al., 2009) ou astrophysique (Lisker et Thorsten, 2008).

En pratique, on ne dispose pas de la courbe de Lorenz, mais du revenu par « tranches » de la population. Pour n tranches, le coefficient s'obtient par la formule de **Brown** :

$$G = 1 - \sum_{k=0}^{n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k)$$

où X est la part cumulée de la population, et Y la part cumulée du revenu.

Pour n personnes ayant des revenus y_i, pour i allant de 1 à n, indicés par ordre croissant ($y_i \leq y_{i+1}$) :

$$G = \frac{2\Sigma_{i=1}^{n} i y_i}{n\Sigma_{i=1}^{n} y_i} - \frac{n+1}{n} \tag{1.1}$$

Lorsque G est basée sur la courbe de Lorenz de la répartition des revenus , il peut être interprété comme l'écart de revenu attendu entre deux individus choisis au hasard dans la population (Sen , 1973).

La définition classique de G apparaît dans la notation de la théorie de la différence moyenne relative :

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\overline{x}}$$

où x est une valeur observée, n est le nombre de valeurs observées, et $\overline{x}$ représente la valeur moyenne. Si les valeurs de x sont d'abord placées dans l'ordre croissant, de sorte que chaque x est de rang i, la partie des comparaisons ci-dessus peuvent être évités et le calcul est plus rapide par cette formule :

$$G = \frac{2}{n^2\overline{x}} \sum_{i=1}^{n} i\,(x_i - \overline{x})$$

qui est ensuite

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)\, x_i}{n \sum_{i=1}^{n} x_i} \tag{1.2}$$

où x est une valeur observée, n est le nombre de valeurs observées et i est le rang des valeurs dans l'ordre croissant. Notez que les valeurs non nulles uniquement positives sont utilisées.

G est une mesure de l'inégalité , définie comme la moyenne des différences absolues entre toutes les paires d'individus pour une certaine mesure. La valeur minimale est 0 lorsque toutes les mesures sont égales et le maximum théorique est de 1 pour un ensemble infiniment grand d'observations où toutes les mesures , mais un a une valeur de 0, ce qui est l'inégalité ultime (Stuart et Ord , 1994).

Définition 1.5. La courbe de Lorenz va nous permettre de mesurer l'indice de Gini graphiquement. Elle est représentée à l'aide des points de coordonnées (p_i, q_i) avec $p_0 = q_0 = 0$.

La courbe de Lorenz est la courbe qui relie la suite des points représentant en ordonnée la progression de la fonction "cumul des ressources d'une classe donnée" et en abscisse la progression de la fonction "cumul des effectif des classes"

Remarque 1.2.4. La courbe de Lorenz se trouve en dessous de la première bissectrice

Exemple 1.2.1. Soit le salaire des employés d'une entreprise organisée en classes dans le tableau suivant :

Salaire mensuel	$[500; 1500[$	$[1500; 2500[$	$[2500; 5500[$
Effectif de la classe (n_i)	50	125	25

On s'intéresse à la répartition du salaire (que l'on note x) sur la population des salariés. Il nous faut donc calculer :

1. les fréquences cumulées pour la variable de salaire x,
2. les fréquences cumulées pour la masse salariale.

I. Calcul des fréquences cumulées pour la variable x :

1. On calcule les fréquences (ou pourcentage) de chaque classe : pour la classe i, $f_i = \frac{n_i}{n}$, où n est l'effectif total (200 salariés).
2. On calcule les fréquences cumulées : $F(x)$.

On obtient :

Salaire mensuel	$[500; 1500[$	$[1500; 2500[$	$[2500; 5500[$
Effectif de la classe (n_i)	50	125	25
Fréquence de la classe i (f_i)	$0,25$	$0,625$	$0,125$
Fréquences cumulées ($F(x)$)	$0,25$	$0,875$	1

II. Calcul des fréquences cumulées pour la masse salariale :

1. Comme les salaires sont donnés par classe, on ne peut pas calculer exactement la masse salariale d'une classe (la somme de tous les salaires des personnes de cette classe). On fait l'approximation suivante :
   - On calcule le centre de la classe : pour la classe $[500; 1500[$, le centre (milieu de l'intervalle) est $x_1 = (500 + 1500)/2 = 1000$,
   - On approche la masse salariale de la classe i en faisant comme si les n_i personnes de cette classe gagnaient toutes x_i. La masse salariale de la classe est environ $n_i x_i$.
2. On calcule le pourcentage de la masse salariale de chaque classe par rapport à la masse salariale totale (ici $n_1x_1 + n_2x_2 + n_3x_3$). Pour la classe i : $g_i = \frac{n_i x_i}{n_1x_1 + n_2x_2 + n_3x_3}$.
3. On calcule les fréquences cumulées : $F(nx)$.

On obtient :

Salaire mensuel	[500; 1500[	[1500; 2500[	[2500; 5500[
Effectif de la classe (n_i)	50	125	25
Centre de la classe x_i	1000	2000	4000
Masse salariale de la classe ($n_i x_i$)	50000	250000	100000
Pourcentage de la masse salariale (g_i)	$0,125$	$0,625$	$0,25$
Fréquences cumulées de la masse salariale ($F(nx)$)	$0,125$	$0,75$	1

III. Courbe de Lorenz La courbe de Lorenz représente les fréquences cumulées de la masse salariale $F(nx)$ en fonction des fréquences cumulées pour la variable salaire F(x).

On obtient la courbe suivante :

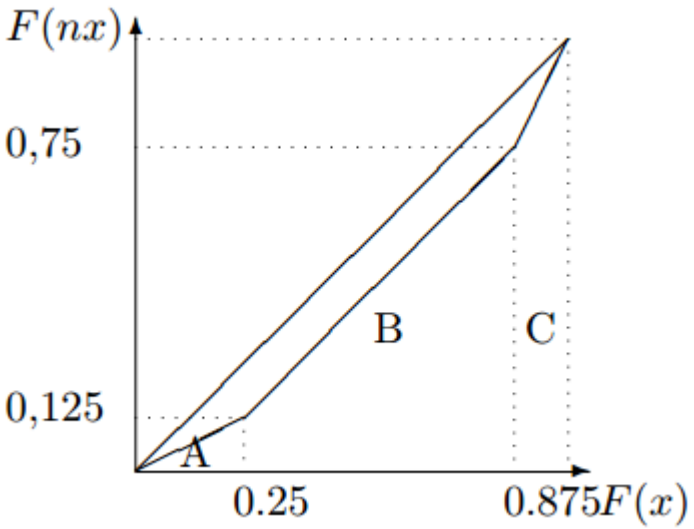


FIGURE 1.7 – Courbe de Lorenz de l'exemple choisi

IV. Calcul de l'indice de Gini :

L'indice de Gini est 2 fois l'aire entre la courbe de Lorentz et la première bissectrice. En regardant la figure, on voit que l'aire entre la courbe de Lorentz et la première bissectrice est égale à :

- L'aire du triangle de côté 1 est : $\dfrac{1 \times 1}{2} = \dfrac{1}{2}$
- Moins l'aire du triangle A qui est égale à : $\dfrac{0,25 \times 0,125}{2} = 0,0156$
- Moins l'aire du trapèze B. L'aire d'un trapèze étant donnée par la formule : $\dfrac{hauteur(grande \quad base + petite \quad base)}{2}$.
  l'aire du trapèze B est donc : $\dfrac{(0,875 - 0,25)\,(0,125 + 0,75)}{2} = 0,2734$

– Moins l'aire du trapèze C qui est égale à : $\dfrac{(1-0,875)\,(0,75+1)}{2} = 0,1094$

Au final, on a l'indice de Gini qui est égale à :

$G = 2 \times \left(\dfrac{1}{2} - (0,0156 + 0,2734 + 0,1094)\right) = 0,2032$

Ici, comme $0,2032$ est faible, l'indice de Gini est faible donc les salaires sont assez bien répartis sur l'ensemble des salariés.

Remarque 1.2.5. A partir de la courbe de Lorenz on peut déduire si les revenus sont bien distribués dans la population étudiée. Ainsi plus la courbe de Lorenz sera proche de la première bissectrice plus les revenus seront bien distribués à la population car dans ce cas l'indice de Gini sera proche de zéro. De même plus la courbe de Lorenz sera loin de la première bissectrice moins les revenus sont bien distribués car dans ce cas l'indice de Gini est proche de un.

1.2.4 L'analyse de variance

L'analyse de variance ou l'ANOVA (ANalysis Of VAriance) correspond à un modèle linéaire gaussien dans lequel toutes les variables explicatives sont qualitatives. Dans ce contexte, elles sont appelées facteurs et leurs modalités sont appelées niveaux. Ces niveaux sont supposés choisis par l'utilisateur, de sorte que l'on parle souvent de facteurs contrôlés. La variable aléatoire réponse est toujours quantitative et supposée gaussienne.

L'ANOVA permet d'étudier la modification de la moyenne μ du phénomène étudie Y selon l'influence d'un ou de plusieurs facteurs d'expérience qualitatifs. Dans le cas où la moyenne n'est influencée que par un seul facteur, il s'agit d'une **analyse de la variance à un facteur**.

On suppose que le nombre de modalités (niveaux) d'un facteur est noté I et que $Y \backsim \mathcal{N}(\mu_i, \sigma^2)$ sur chaque sous population i définie par les modalités. L'objectif est de tester l'égalité des moyennes de ces I populations, à savoir de tester l'hypothèse nulle

$$\mathcal{H}_0 : \mu_1 = \mu_2 = ... = \mu_I$$

contre l'assertion d'intérêt

$$\mathcal{H}_1 : \exists\ \mu_i \neq \mu_{i'}\ ;\ i, j \in I$$

Pour chaque sous population on dispose d'une sous population i, on dispose d'un échantillon de n_i observations de la variable quantitative Y : $y_{i,1},\ y_{i,2},\ ...\ , y_{i,n_i}$

Le modèle **d'analyse de variance à un facteur** s'écrit :

$Y_{ik} = \mu_i + \epsilon_{ij} \quad pour \quad k = 1, ..., n_i \quad et \quad i = 1, ..., I \quad$ avec $\epsilon_{ij} \backsim \mathcal{N}(0, \sigma^2)$.

On peut aussi écrire $\mu_i = \mu + \alpha_i$ pour $i = 1, ..., I$. Dans ce cadre là, μ est appelé **l'effet moyen du facteur** et $\alpha_i = \mu_i - \mu$ est appelé **l'effet différentiel** du niveau i du facteur. Le modèle ci dessous peut donc aussi s'écrire sous la forme :

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ij} \quad pour \quad k = 1, ..., n_i \quad et \quad i = 1, ..., I$$

Pour tester la significativité du facteur F (ou du modèle envisagé), il est fréquent de résumer la construction de la statistique du test de Fisher (à J-1 et n-J degré de liberté (d.d.l.)) au sein d'un tableau, appelé tableau d'analyse de la variance, qui se présente sous la forme ci-dessous.

Source de variation	somme des carrées(SS)	d.d.l	carrés moyens (MS)	valeur-p
Facteur(F)	SSF	$I-1$	$MSF = \dfrac{SSF}{I-1}$	$\dfrac{MSF}{MSE}$
Erreur(E)	SSE	$n-I$	$MSE = \dfrac{SSF}{n-I} = \widehat{\sigma}^2$	–
Total(T)	SST	$n-1$	–	–

Où **valeur-p** désigne la valeur de la statistique de Fisher.

Le modèle **d'analyse de variance à deux facteurs** s'écrit :

$Y_{ik} = \mu_{ij} + \epsilon_{ijk} \quad pour \quad k = 1, ..., n_{ij} \quad , \quad i = 1, ..., I et j = 1, ..., J \quad$ avec $\quad \epsilon_{ijk} \backsim \mathcal{N}(0, \sigma^2)$.

Dans ce modèle, les paramètres réels μ_{11}, μ_{11}, ... ,μ_{I1}, ... ,μ_{1J}, ... ,μ_{IJ} sont inconnus ainsi que la variance σ^2. On décompose μ_{ij} de façon à faire apparaitre les effets des facteurs A et B et de leurs interactions :

$\mu_{ij} = \mu_{\bullet\bullet} + \alpha_i^A + \alpha_j^B + \beta_{ij}$

Où

- $\mu_{\bullet\bullet} = \dfrac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$: effet moyen général.
- $\mu_{i\bullet} = \dfrac{1}{J} \sum_{j=1}^{J} \mu_{ij}$: effet du niveau i du facteur A.
- $\alpha_i^A = \mu_{i\bullet} - \mu_{\bullet\bullet}$: effet différentiel du niveau i du facteur A.
- $\mu_{\bullet j} = \dfrac{1}{I} \sum_{i=1}^{I} \mu_{ij}$: effet du niveau j du facteur B.
- $\alpha_j^B = \mu_{\bullet j} - \mu_{\bullet\bullet}$: effet différentiel du niveau j du facteur B.
- $\beta_{ij} = \mu_{ij} - \mu_{\bullet\bullet} - \alpha_i^A - \alpha_j^B$: effet d'interaction du niveau i du facteur A et du niveau j du facteur B.

Comme pour le modèle d'ANOVA à un facteur, on obtient le tableau d'analyse de la variance suivant, pour deux facteurs F_1 et F_2 :

Source de variation	somme des carrées(SS)	d.d.l	carrés moyens (MS)	valeur-p
F_1	SSF_1	$I-1$	$MSF_1 = \frac{SSF_1}{I-1}$	$\frac{MSF_1}{MSE}$
F_2	SSF_2	$J-1$	$MSF_2 = \frac{SSF_2}{J-1}$	$\frac{MSF_2}{MSE}$
$F_1 * F_2$	SSF_{12}	$(I-1)(J-1)$	$MSF_{12} = \frac{SSF_{12}}{(I-1)(J-1)}$	$\frac{MSF_{12}}{MSE}$
Erreur(E)	SSE	$n-IJ$	$MSE = \frac{SSF}{n-IJ} = \widehat{\sigma}^2$	–
Total(T)	SST	$n-1$	–	–

Où $F_{12} = F_1 * F_2$ désigne le terme d'interaction entre les facteurs F_1 et F_2.

Chapitre Deux

Matériels et méthodes

2.1 Matériels et dispositif expérimental

Pour ce travail nous utilisons les données fournies par **PalmElit** (www.palmelit.com), une compagnie privée d'amélioration génétique et de commercialisation de semences sélectionnées de palmiers à huile.

Elles ont été collectées dans deux sites d'expérimentation proches, tous deux situés en Indonésie et appartenant à la SOCFINDO (Aek Loba et Aek Kwasan). Elles portent sur un ensemble de palmiers tenera appartenant à des croisements, essentiellement de type hybride entre le groupe A et le groupe B.

2.1.1 Dispositif expérimental

Les différents croisements sont plantés dans des essais, selon des dispositifs expérimentaux en blocs de Fisher (blocs complets). Un bloc complet contient tous les croisements de l'essai. Au sein d'un bloc complet, les individus d'un même croisement sont plantés ensemble en une "parcelle élémentaire", qui constitue l'unité de base de l'expérimentation. La production mensuelle des régimes a été notée pour tout individu hybride de 3 ans à 6 ans. En détails les données sont :

- Nombre d'essais= 48
- Nombre de croisements par essai : moyenne= 20 ; écart-type= 7.31
- Nombre de croisements **AxB** : 730
- Nombre de parents **A** : 229
- Nombre de parents **B** : 216
- Nombre d'individus par croisements : moyen= 68 ; écart-type= 64.7
- Nombre total d'individus : 49840
- Nombre de parcelles élémentaires : 151

– Nombre d'individus par parcelles élémentaires : moyenne= 9.2 ; écart-type= 3.4. La plus grande partie des parcelles a un nombre d'arbres compris entre **10 et 12**. Le nombre d'arbre minimal par parcelle est **1** et le nombre d'arbre maximal par parcelle est **26**. La distribution des nombres d'individus par parcelle élémentaire est sur l'histogramme suivant :

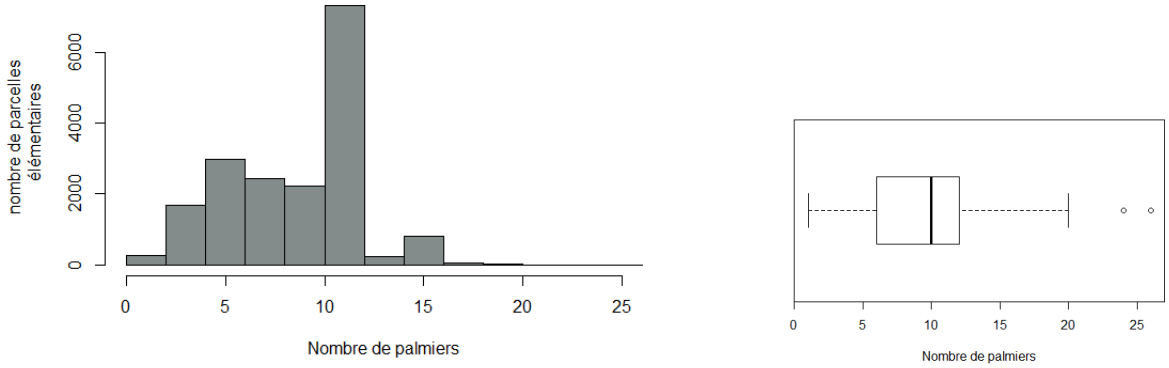


FIGURE 2.1 – Nombre d'individus par parcelle élémentaire

2.1.2 Aperçu des données

STATION	PARCELLE	LIGNE	ARBRE	camp	pr1	pr2	pr3	pr4	pr5	pr6	pr7	pr8	pr9	pr10	pr11	pr12
AL	41A	2	4	2000	6	8	29	24	17	28	27	0	28	27	16	0
AL	41A	2	4	2001	11	10	0	9	19	53	32	22	19	0	14	11
AL	41A	2	5	1998	9	5	11	3	8	6	5	17	6	3	15	20
AL	41A	2	5	1999	5	10	10	17	14	8	15	25	0	31	18	29
AL	41A	2	5	2000	0	17	14	0	0	0	0	0	51	17	40	38
AL	41A	2	5	2001	17	8	21	15	8	20	0	10	0	40	24	75
AL	41A	2	8	1998	1	7	5	6	10	9	8	13	6	5	13	11
AL	41A	2	8	1999	8	11	10	9	0	9	20	12	13	10	21	32
AL	41A	2	8	2000	9	17	30	22	29	24	12	6	13	25	24	21
AL	41A	2	8	2001	7	16	18	40	18	16	17	27	29	21	17	40
AL	41A	2	11	1998	7	2	7	0	10	8	12	8	8	7	7	15
AL	41A	2	11	1999	10	10	9	21	24	18	13	10	15	21	28	26
AL	41A	2	11	2000	9	18	16	24	26	15	23	17	15	15	22	22
AL	41A	2	11	2001	8	13	20	26	21	38	10	26	20	11	34	44

APLANT	NUMESSAI	NUMPAREXP	REPET	MERE_RECOD	PERE_RECOD
1995	1	1	1	D63980	P64216
1995	1	1	1	D63980	P64216
1995	1	1	1	D63980	P64216
1995	1	1	1	D63980	P64216
1995	1	1	1	D63980	P64216
1995	1	1	1	D63980	P64216
1995	1	6	1	D63955	T62965
1995	1	6	1	D63955	T62965
1995	1	6	1	D63955	T62965
1995	1	6	1	D63955	T62965
1995	1	11	1	D63969	T62965
1995	1	11	1	D63969	T62965
1995	1	11	1	D63969	T62965
1995	1	11	1	D63969	T62965

Dans le tableau de données, un enregistrement (ou ligne du tableau) est une année d'observation sur un palmier. Les palmiers sont plantés en ligne dans des parcelles. Chaque palmier est identifié de manière unique en concaténant le nom de la parcelle sur laquelle il se trouve (colonne PARCELLE du tableau des données), sa ligne dans la parcelle (LIGNE) et sa position le long de la ligne (ARBRE).

- **STATION :** cette colonne présente le site d'expérimentation, **AL** pour **Aek Loba** et **AK** pour **Aek Kwasan**
- **PARCELLE :** identifiant de la parcelle.
- **ARBRE :** numéro de l'arbre sur sa ligne.
- **camp :** année d'observation de chaque palmier.
- **nr1, nr2, nr3, ..., nr12 :** présente le nombre de régimes produit le mois de **janvier, février, mars, ..., décembre**
- **pr1, pr2, pr3, ..., pr12 :** présente le poids total de régime produit le mois de **janvier, février, mars, ..., décembre**
- **APLANT :** année de plantation de l'individu.
- **NUMESSAI :** numéro de l'essai.
- **NUMPAREXP :** identifiant de la parcelle élémentaire.
- **REPET :** identifiant du bloc complet dans l'essai.
- **BLOC :** identifiant du bloc incomplet (dans certains essais les blocs complets ont été subdivisés en blocs incomplets pour mieux tenir compte de l'hétérogénéité du terrain).
- **MERE_RECOD :** nom de la mère.
- **PERE_RECOD :** nom du père

Ce tableau de données contient, en plus des données concernant des croisements hybrides A x B, des données sur d'autres types de croisements (A x A, etc), qui ne sont pas représentatif du matériel commercial existant dans les plantations. Afin de les éliminer de nos analyses, nous utilisons un autre fichier qui donne le groupe **A** ou **B** auquel appartiennent les parents des croisements étudiés. Après la lecture des données, nous déterminons le groupe du père et le groupe de la mère puis on élimine les croisements qui ne sont pas **AxB**.

On ajoute à nos données la colonne appelé **âge** des palmiers (différence entre l'année d'observation et l'année de plantation).

2.2 Méthodes

Pour notre étude le traitement des données est fait par le logiciel **R**.

Nous présentons les manipulations qui ont été faites sur les données et les principales fonctions et packages de R utilisés pour ce travail.

2.2.1 Présentation de quelques fonctions de R utilisées

- **La fonction merge du package de base de R :** permet de fusionner deux dataframes (tableaux de données) par colonnes communes ou par lignes communes.
- **Les fonctions cbind et rbind** permettent respectivement d'ajouter des colonnes ou des lignes soit à un dataframe soit à une matrice. Elles prennent en arguments des séquences de vecteurs, des matrices ou d'un dataframe.
- **La fonction paste du package de base R :** permet de concaténer des vecteurs après les avoir convertir en caractère.
- **La fonction aggregate** permet de découper un data.frame en sous populations suivant un facteur (spécifié par le paramètre by) et d'appliquer une fonction donnée sur chacune de ces sous-populations.
- **La fonction stack** permet d'empiler dans un seul vecteur les valeurs de certaines colonnes d'un dataframe. Cette fonction renvoie un dataframe dont la première colonne contient le vecteur ainsi empilé, et dont la deuxième colonne contient un facteur indiquant l'origine de chaque observation. La fonction **unstack** effectue l'opération inverse. Cette fonction apparait particulièrement utile en analyse de la variance (ANOVA).
- **la fonction apply :** fonction très utilisée qui applique une fonction donnée (fournie comme valeur du paramètre FUN) aux lignes (MARGIN=1) ou bien aux colonnes (MARGIN=2) d'une matrice ou d'un data.frame.
- **La fonction unlist :** permet de transformer une list en un vecteur.
- **La fonction sample** permet de simuler des tirages au hasard. elle prend essentiellement deux arguments : le premier est un vecteur de valeurs parmi lesquelles le tirage est effectué, le second est le nombre de valeurs à tirer autrement dit la taille de l'échantillon à fabriquer. Un argument optionnel appelé **replace** permet d'identifier si le tirage se fait avec ou sans remise. C'est une valeur logique (TRUE ou FALSE) : s'il est fixé à la valeur TRUE, le tirage se fait avec remise. Par défaut, les tirages se font sans remise : le nombre de valeurs demandées doit donc être inférieur dans ce cas au nombre de valeur disponible.
- **La fonction plot :** permet de dessiner des diagrammes de dispersion aussi bien que des courbes représentatives de fonctions dans le plan. Elle prend essentiellement deux arguments qui sont des vecteurs numériques de même longueur représentant respectivement les abscisses et les ordonnées des points.
- **La fonction Gini** du package DescTools permet de calculer le coefficient de Gini. Voir annexe

II

2.2.2 Exemple numérique de calcul de l'indice de Gini d'une parcelle

L'approche géométrique du calcul de l'indice de Gini est adaptée dans des situations où les données ne sont pas nombreuses. Plus les données sont nombreuses, plus l'évaluation de l'indice de Gini devient difficile.

Dans cette partie de notre travail nous allons utiliser la formule explicite de l'équation1.2 pour illustrer le calcul de l'indice de Gini d'une parcelle.

De nos données présentées ci-dessus, on extrait les données de production d'une parcelle élémentaire de **12 palmiers**, pour une seule année.

Palmiers	pr1	pr2	pr3	pr4	pr5	pr6	pr7	pr8	pr9	pr10	pr11	pr12
1	7	17	6	22	15	22	19	7	24	19	24	18
2	17	32	9	21	12	41	19	21	11	24	9	15
3	18	13	29	28	27	14	0	30	9	21	40	19
4	18	21	21	17	16	23	31	22	19	24	33	18
5	18	15	15	14	13	15	28	18	18	21	18	17
6	6	10	13	7	19	20	13	23	31	15	16	5
7	6	12	20	18	28	7	42	23	28	20	34	8
8	27	20	22	17	15	19	27	20	18	24	7	0
9	18	22	20	23	25	26	20	22	20	10	29	17
10	10	35	16	29	27	8	21	30	9	30	25	22
11	40	29	8	14	13	23	7	0	0	0	0	27
12	24	10	15	25	14	17	6	29	16	22	12	13

TABLE 2.1 – production mensuelle de douze palmiers d'une parcelle

Où les lignes du tableau représentent la production mensuelle d'un palmier observé un an.

Nous ferons le calcul en deux étapes :

première étape : calcul de la production moyenne des individus de la parcelle élémentaire pour chaque mois.

Pour le mois de janvier on somme le poids total de la production de chaque palmier puis on divise par douze, car la parcelle a douze palmiers. On fait de même pour les autres mois, et on a alors le résultat dans le tableau suivant :

	pr1	pr2	pr3	pr4	pr5	pr6	pr7	pr8	pr9	pr10	pr11	pr12
Total	209	236	194	235	224	235	233	245	203	230	247	179
Moyenne	17.42	19.67	16.17	19.58	18.67	19.58	19.42	20.42	16.92	19.17	20.58	14.92

TABLE 2.2 – Production moyenne mensuelle sur une parcelle élémentaire

deuxième étape : calcul de l'indice de Gini

– **courbe de Lorenz :**

Déterminons les coordonnées des points $(p_i; q_i)$

où $p_i = \dfrac{i}{n}$ avec $i = 1, 2, 3, \cdots, 12$ et $n = 12$ $q_i = \dfrac{\sum_{j \leq i} pr_j}{\sum_{i=1}^{12} pr_i}$ avec $i, j = 1, 2, 3, \cdots, 12$

Ces points nous permettent de construire la courbe de Lorenz. On a :

$\sum_{i=1}^{12} pr_i = 222.5$, le point $(p_0, q_0) = (0, 0)$

$p_1 = \dfrac{1}{12} = 0.083$; $q_1 = \dfrac{pr_1}{222.5} = \dfrac{17.42}{222.5} = 0.0783$ on a donc le point $(p_1, q_1) = (0, 083, 0, 0783)$

$p_2 = \dfrac{2}{12} = 0.167$; $q_2 = \dfrac{pr_1 + pr_2}{222.5} = \dfrac{17.42 + 19.67}{222.5} = 0.166$ on a donc le point $(p_2, q_2) = (0.167, 0.166)$

$p_3 = \dfrac{3}{12} = 0.25$; $q_3 = \dfrac{pr_1 + pr_2 + pr_3}{222.5} = \dfrac{17.42 + 19.67 + 16.167}{222.5} = 0.24$ on a donc le point $(p_3, q_3) = (0.25, 0.24)$

De façon similaire on a les autres points et $(p_{12}, q_{12}) = (1, 1)$

On a donc la courbe de Lorenz suivante :

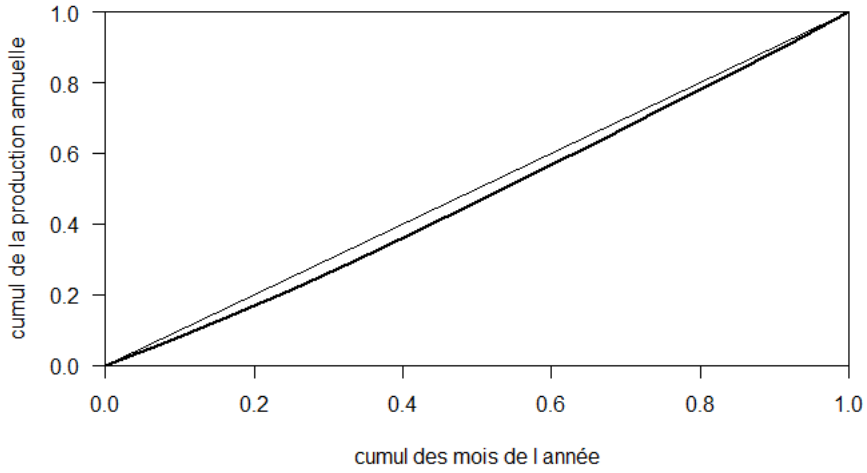


FIGURE 2.2 – Courbe de Lorenz de la production d'une parcelle élémentaire pour une année

Interprétation 2.0.1. La courbe de Lorenz de cette parcelle montre que l'indice de Gini est proche de zéro, car la courbe est assez proche de la première bissectrice. Ceci sera donc vérifié par calcul pour mieux apprécier la valeur de l'indice de Gini.

– **Calcul de l'indice de Gini de la parcelle**

Utilisons la formule de l'équation1.2. Pour cela intéressons nous aux productions mensuelles moyennes de la parcelle élémentaire. Rangeons par ordre croissant les productions mensuelles puis pour tout i calculons $2i-n-1$ et $(2i-n-1)\,pr_i = (2i-13)\,pr_i$ car $n=12$. On a le tableau suivant :

	pr1	pr2	pr3	pr4	pr5	pr6	pr7	pr8	pr9	pr10	pr11	pr12
Moyenne	17.42	19.67	16.17	19.58	18.67	19.58	19.42	20.42	16.92	19.17	20.58	14.92
MOC	14.92	16.17	16.92	17.42	18.67	19.17	19.42	19.58	19.58	19.67	20.42	20.58
$2i-13$	-11	-9	-7	-5	-3	-1	1	3	5	7	9	11
$(2i-13)\,pr_i$	-164.1	-145.5	-118.4	-87.1	-56	-19.2	19.42	58.75	97.9	137.7	183.8	226.4

Dans ce tableau, MOC signifie moyenne de production par ordre croissante.

L'équation1.2 dévient :

$$G = \frac{\sum_{i=1}^{n} (2i-13)\,pr_i}{n\sum_{i=1}^{n} pr_i}$$

D'après le tableau on a : $\sum_{i=1}^{12} pr_i = 222,5$ et $\sum_{i=1}^{n} (2i-13)\, pr_i = 133.7$

Ainsi $G = \dfrac{133.7}{12 \times 222.5} = 0.0501$

G est très proche de zéro ce qui qui confirme l'interprétation de la courbe de Lorenz.

conclusion G étant très proche de zéro, il y a donc régularité de la production sur cette parcelle. On observe cela avec le diagramme suivant :

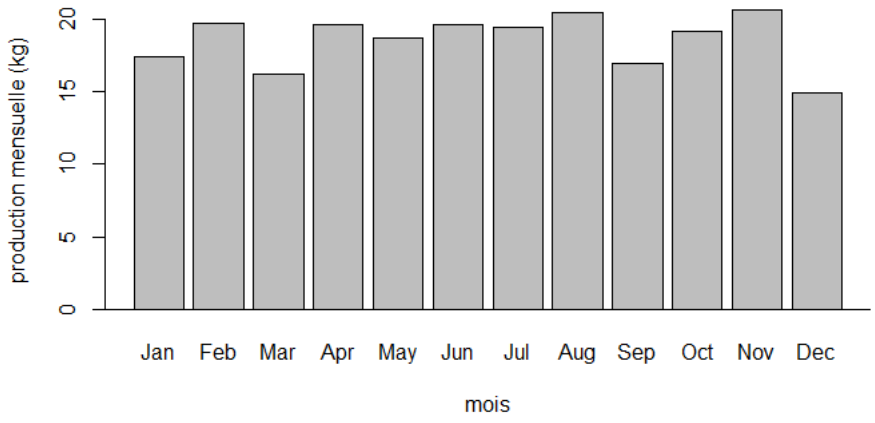


FIGURE 2.3 – Répartition de la production mensuelle d'une parcelle élémentaire pour une année

2.2.3 Définition du nombre d'individus seuil pour obtenir un indice de Gini représentatif d'un croisement

Les indices de Gini calculés ici doivent être représentatifs de ce qu'observerait un producteur ayant planté les croisements étudiés ici sur une grande superficie. On peut donc se demander combien d'individus d'un même croisement doivent être considérés simultanément pour que l'indice de Gini obtenu soit représentatif de la valeur du croisement. En effet, le profil de répartition de la production d'un croisement sur l'année dépend à la fois du profil de répartition des individus qui composent ce croisement et de leur synchronisation. Considérons le cas théorique où chaque palmier produirait l'ensemble de ses régimes sur un seul mois de l'année, et aurait donc un indice de Gini de 1. Dans cette situation, on a deux possibilités extrêmes au niveau du croisement. Si tous les palmiers sont parfaitement synchronisés, le croisement aura lui aussi un indice de Gini de 1 ; si bien que l'indice de Gini d'un individu au sein du croisement est représentatif de la valeur du croisement. Par contre, si les

palmiers ne sont pas du tout synchronisés entre eux, c'est-à-dire qu'ils ont tous un mois de production différent, la répartition de la production au niveau du croisement sera parfaitement lisse sur l'année (G=0). Dans ce cas, l'indice de Gini d'un individu au sein du croisement n'est pas représentatif de la valeur du croisement ; et il faudrait considérer conjointement les données de 12 palmiers pour avoir un indice de Gini traduisant effectivement le comportement du croisement. Dans la pratique, les palmiers d'un même croisement ne sont pas complètement désynchronisés, car leurs cycles dépendent en partie des conditions environnementales auxquelles ils sont tous soumis. On va donc utiliser nos données pour voir s'il existe un seuil minimum d'individus à considérer pour que l'indice de Gini calculé soit représentatif du croisement. Pour cela, on va rechercher des parcelles élémentaires ayant un assez grand nombre d'individus et des indices de Gini contrastés (lorsqu'ils sont calculés sur l'ensemble des individus disponibles). Une fois de telles parcelles élémentaires identifiées, on recalculera leur indice de Gini en boucle en considérant un nombre croissant d'individus. Cette méthode permettra de vérifier si l'indice de Gini évolue avec le nombre d'individus utilisés pour faire le calcul et, le cas échéant, de déterminer le seuil à partir duquel il devient stable, c'est à dire représentatif du croisement dans son ensemble.

Dans le détail, nous suivons les étapes suivantes :

1. On calcule pour chaque parcelle élémentaire ayant 12 palmiers la production moyenne pour chaque mois, en considérant l'ensemble des palmiers.
2. On se sert de ces productions mensuelles moyennes pour calculer l'indice de Gini des parcelles élementaires.
3. On ordonne les parcelles élémentaires par ordre croissant d'indice de Gini.
4. On sélectionne les 5 meilleures parcelles, 5 intermédiaires et les 5 dernières parcelles par rapport à l'indice de Gini.
5. Une fois ce jeu de 15 parcelles identifiées, on reprend le calcul de l'indice de Gini en considérant un seul individu, puis deux, trois, etc. jusqu'à 12. Pour chaque niveau de nombre d'individus, on réalise des tirages au sort sans remise pour choisir aléatoirement les individus qui entreront dans le calcul. Par ailleurs, afin de s'affranchir d'un éventuel biais qui serait lié aux individus effectivement tirés au sort, on effectue 12 répétitions pour chaque niveau de nombre d'individus. Ainsi, pour le niveau "3 individus par croisement" par exemple, on calculera 12 fois l'indice de Gini sur trois palmiers, en tirant au hasard trois individus à chaque fois. La valeur de l'indice de Gini pour les différents nombres d'individus considérés sera la moyenne des 12 valeurs obtenues

avec les 12 répétitions (sauf pour le niveau "12 individus par croisement", où un seul indice de Gini peut être calculé).

6. A partir du graphique de l'indice de Gini exprimé en fonction du nombre d'individus utilisés pour le calculer, on détermine le nombre de palmiers nécessaires pour le calcul d'un indice de Gini représentatif du croisement.

Chapitre Trois

Résultats et discussion

3.1 Calcul de l'indice de Gini et détermination du seuil pour obtenir un indice de Gini représentatif

Pour calculer l'indice de Gini, nous créons d'abord un identifiant qui permet de repérer de manière unique chaque parcelle élémentaire à un âge donné. Pour cela, on ajoute une colonne dans nos données appelée **PARCELLE_elementaire_Gini**, qui est une concaténation entre les colonnes NUMPAREXP, NUMESSAI, et âge.

Par la fonction **table** appliquée sur la colonne **PARCELLE_elementaire_Gini** on a le nombre de palmiers par parcelle élémentaire. Leur distribution est représentée sur la figure 2.1

La plus grande partie des parcelles a un nombre d'individus compris entre **10 et 12**. Pour déterminer le seuil nécessaire pour calculer l'indice de Gini, nous considérons les parcelles avec 12 palmiers, car ce sont les plus nombreuses et qu'un nombre de 12 permet de faire varier de manière significative le nombre d'individus (1 à 12) utilisés pour ce calcul.

Nous calculons donc l'indice de Gini sur les parcelles élémentaires ayant 12 palmiers (calcul à chaque âge). Pour cela, avec la fonction **aggregate**, on calcule la production moyenne mensuelle de régimes par parcelle élémentaire. On a alors un nouveau dataframe appelé **moyenne_poids_par_parcelle**. A ceci on ajoute la colonne **indice de Gini**, calculée par la fonction **Gini** du package **Desctools**.

Nous avons donc la répartition de l'indice de Gini par l'histogramme suivant :

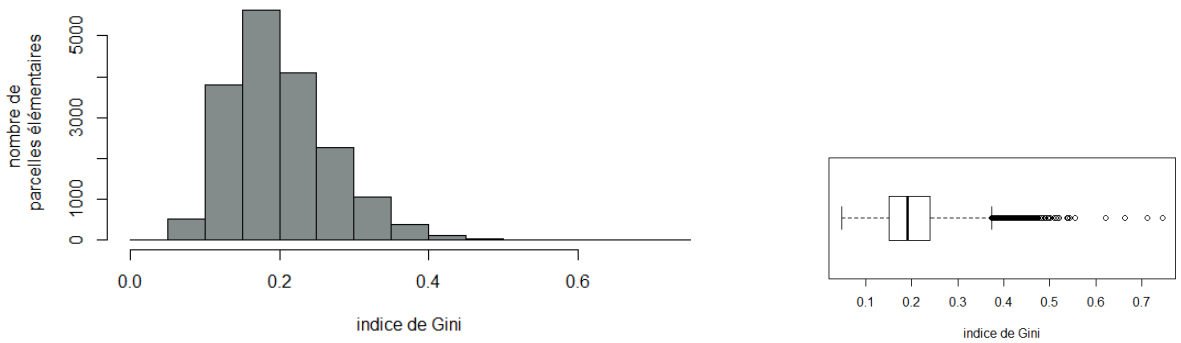


FIGURE 3.1 – Distribution des valeurs d'indice de Gini

La plus petite valeur de l'indice de Gini est : 0.048, la plus grande est : 0.746, la médiane est : 0.19 (l'indice de Gini de la moitié des parcelles élémentaires est donc inférieur ou égal à 0.19). Il y a donc une variabilité importante en terme d'indice de Gini entre parcelles élémentaires.

Ordonnons les parcelles élémentaires par rapport à l'indice de Gini grâce à la fonction **order** afin de sélectionner les 5 parcelles élémentaires donc l'indice de Gini est le plus faible, 5 parcelles élémentaires dont l'indice de Gini est centré autour de la médiane et les 5 parcelles élémentaires donc l'indice de Gini est le plus fort.

Sur ces 15 parcelles, on applique la procédure décrite dans "matériel et méthodes" pour calculer l'évolution de l'indice de Gini en fonction du nombre de palmiers utilisés pour faire le calcul. Le résultat est présenté sur le graphique suivant, obtenu grâce à la fonction **ggplot** du package ggplot2.

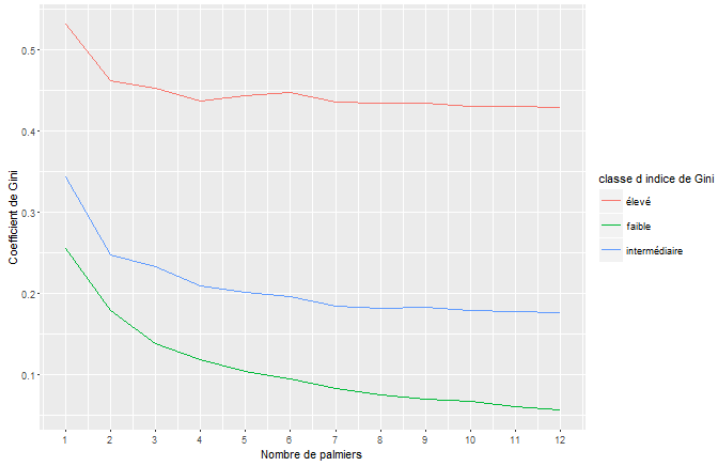


FIGURE 3.2 – Evolution de l'indice de Gini en fonction du nombre de palmiers utilisés pour faire le calcul

Il y a une décroissance de l'indice de Gini en fonction du nombre de palmiers. Pour les parcelles

dont l'indice de Gini est élevé, on observe qu'il y a une stabilité de l'indice de Gini à partir de 6 palmiers. Pour celles ayant un indice de Gini intermédiaire, on observe une stabilité à partir de 8 palmiers. Pour les parcelles élémentaires ayant un indice de Gini faible, on observe une stabilité à partir de 10 palmiers. Etant donné que le nombre de palmiers nécessaires par parcelle élémentaire pour avoir un indice de Gini représentatif doit être adapté à toutes les valeurs de Gini, on doit donc choisir comme seuil le plus grand nombre trouvé ici, c'est-à-dire 10 palmiers. Le seuil retenu est donc de 10 palmiers.

On pourra se reporter en annexe pour voir le graphique donnant les courbes pour chacune des 15 parcelles utilisées pour ce calcul (l'annexe I), ainsi que le tableau qui a permis de tracer la figure 3.2.

3.2 Pertinence de l'indice de Gini pour quantifier la production

De nos données de départ, nous allons extraire les parcelles élémentaires dont le nombre d'individus est supérieur ou égale au nombre seuil qui est 10. On constitue ainsi un nouveau tableau de données avec lequel on va continuer notre étude.

Pour montrer que l'indice de Gini est un bon indicateur de mesure d'inégalité nous présentons les statistiques des parcelles élémentaires ayant l'indice de Gini le plus faible, un indice de Gini médian et celle ayant l'indice de Gini le plus fort.

La parcelle élémentaire dont l'indice de Gini est le plus petit doit être celle là qui a la meilleure répartition de la production tout au long de l'année.

Le plus petit indice de Gini du nouveau tableau est : 0.05

La courbe de Lorenz de la parcelle correspondante est :

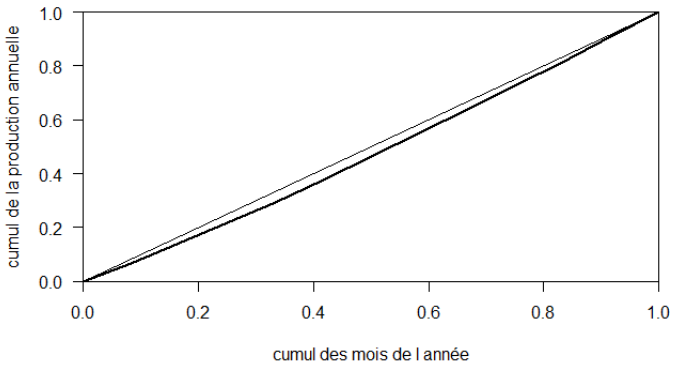


FIGURE 3.3 – Courbe de Lorenz de la parcelle élémentaire ayant l'indice de Gini le plus faible

Interprétation 3.0.2. On observe que la courbe de Lorenz est très proche de la première bissectrice ce qui permet de conclure que la répartition de la production des régimes est bonne tout au long de l'année, c'est-à-dire faite de façon régulière. La régularité de la production dans cette parcelle élémentaire peut s'observer avec le diagramme suivant :

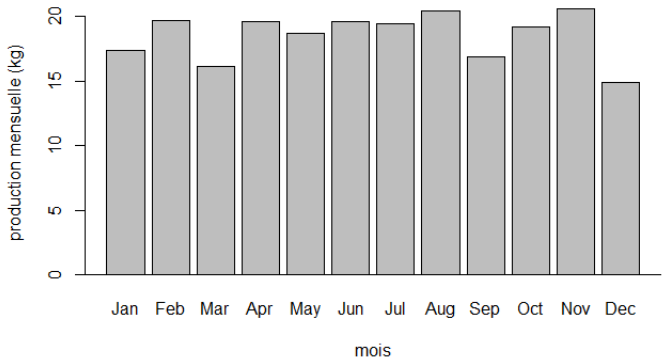


FIGURE 3.4 – Répartition de la production mensuelle de la parcelle élémentaire ayant l'indice de Gini le plus faible

Il y a une très bonne répartition dans cette parcelle élémentaire car la production du mois de janvier est en moyenne 17.42 kg, celle de février est 19.67 kg, celle de mars est 16.2 kg, celle du mois d'avril est 19.6 kg, celle de mai est 18.67 kg... : la récolte chaque mois de l'année ne varie presque pas d'un mois à l'autre.

Parcelle élémentaire ayant un indice de Gini médian

L'indice de Gini de cette parcelle élémentaire est : 0.179. Observons cela avec la courbe de Lorenz.

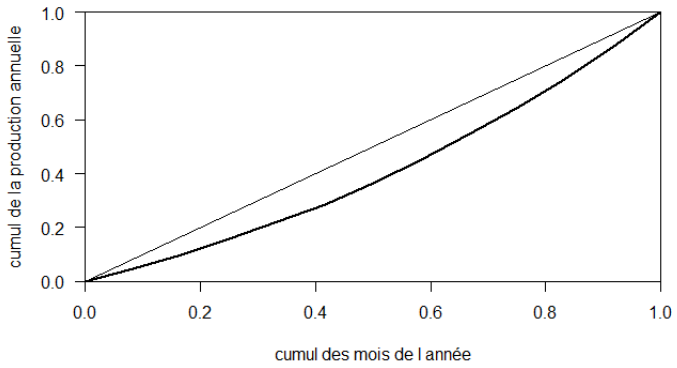


FIGURE 3.5 – Courbe de Lorenz de la parcelle élémentaire ayant un indice de Gini médian

Interprétation 3.0.3. Cette courbe est plus écartée de la première bissectrice que celle de la figure 3.3, donc la répartition de la production tout au long de l'année n'est plus aussi régulière que celle de la parcelle précédente. Observons cette répartition dans le diagramme suivant :

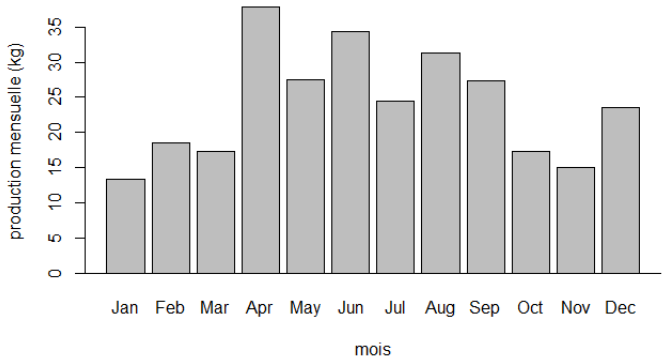


FIGURE 3.6 – Répartition de la production mensuelle de la parcelle élémentaire ayant un indice de Gini médian

Interprétation 3.0.4. L'étendue des productions moyennes mensuelles de cette parcelle élémentaire est : $37.8 - 13.4 = 24.4$ qui est la différence de production du mois d'avril et du mois de janvier. Cette étendue permet de conclure que la production sur cette parcelle élémentaire est moins régulière que sur la parcelle précédente. La récolte au mois de janvier est 13.4 kg, celle de février est 18.6 kg, celle de mars est 17.3 kg et subitement celle d'avril est 37.8 kg, qui est le double de la production du mois de février.

Parcelle élémentaire ayant un indice de Gini fort

L'indice de Gini de cette parcelle élémentaire est : 0.492. Observons la courbe de Lorenz associée à la production de cette parcelle élémentaire.

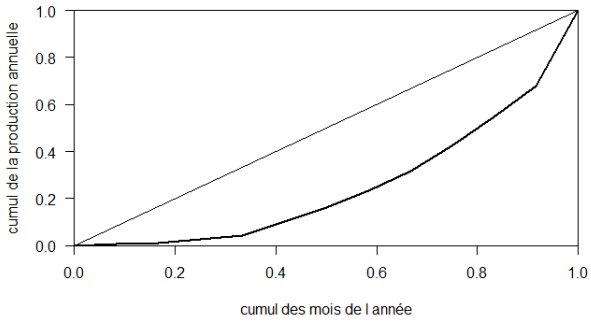


FIGURE 3.7 – Courbe de Lorenz de la parcelle élémentaire ayant l'indice de Gini le plus fort

Interprétation 3.0.5. cette courbe de Lorenz est très écartée de la première bissectrice ce qui permet de conclure que la production n'est pas bien répartie tout au long de l'année par rapport au parcelles précédentes. Observons cette répartition avec le diagramme suivant :

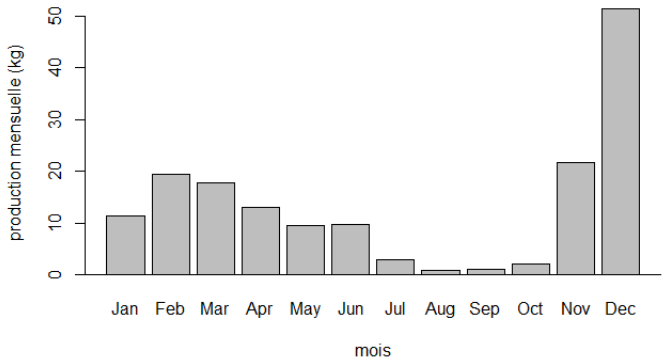


FIGURE 3.8 – Répartition de la production de la parcelle élémentaire ayant l'indice de Gini le plus fort

Interprétation 3.0.6. L'étendue de production moyenne de cette parcelle élémentaire est : $51.36 - 0.82 = 50.54$kg qui est la différence de production du mois de décembre et le mois d'aout. Cette étendue permet de conclure qu'il y a une mauvaise régularité de la production sur cette parcelle élémentaire car au mois d'aout, de septembre et octobre, il n'y a quasiment pas de récolte et subitement au mois de décembre on récolte 51.36 kg.

Conclusion

De ces différents résultats obtenus nous déduisons que l'indice de Gini est un bon indicateur pour mesurer la régularité de la production, car la valeur de l'indice de Gini d'une parcelle élémentaire reflète bien le diagramme représentant la répartition annuelle de la production annuelle associée.

3.3 Variabilité de l'indice de Gini

Dans la perspective d'une amélioration génétique de la répartition de la production, il est intéressant d'évaluer la variabilité de l'indice de Gini, afin de voir s'il pourrait être possible de trouver des croisements avec un indice de Gini faible. On observe la variabilité de l'indice de Gini sur les parcelles élémentaires dont le nombre de palmiers est supérieur ou égal à 10 dans la figure suivante :

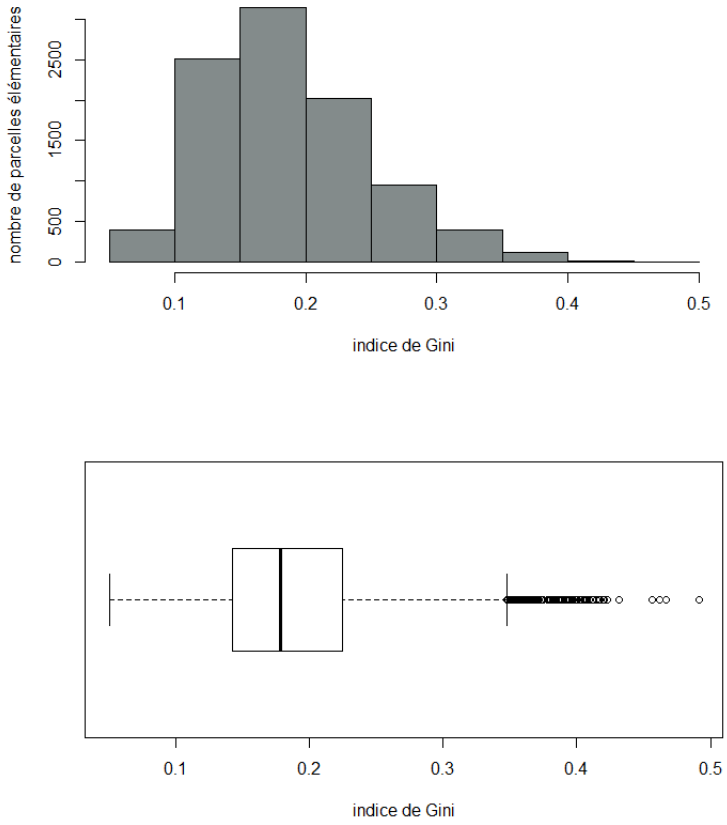


FIGURE 3.9 – Variabilité de l'indice de Gini, calculé sur les parcelles élémentaires dont le nombre de palmiers est supérieur ou égal à 10, et aux différents âges

Interprétation 3.0.7. Il existe effectivement une petite proportion de parcelles élémentaires présentant un faible indice de Gini.

Maintenant, nous allons étudier si l'indice de Gini évolue avec l'âge et s'il est corrélé avec la production annuelle de régimes.

Pour évaluer si l'indice de Gini est affecté par l'âge des palmiers, nous ferons d'abord un test d'ANOVA à deux facteurs, âge et parcelle élémentaire. Le test nous donne le résultat dans le tableau suivant :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	0.67	0.67	234.29	2e-16
parcelle	2949	17.18	0.01	2.03	2e-16
Résidus	6613	19.01	0.001		

TABLE 3.1 – Influence de l'âge sur l'indice de Gini

Ce tableau permet de conclure que l'âge a un effet hautement significatif sur l'indice de Gini car la P-value associée est très faible.

La matrice de corrélation de Pearson entre les différents âges est :

	3	4	5	6
3	1.00	0.26	0.19	-0.02
4	0.26	1.00	0.46	0.11
5	0.19	0.46	1.00	0.33
6	-0.02	0.11	0.33	1.00

Les valeurs de corrélation sont en général faibles ou nulles, à l'exception de la valeur intermédiaire 0.46 entre 4 et 5 ans. Ceci montre que les indices de Gini aux différents âges ne sont pas corrélés entre eux. Pour une parcelle élémentaire donnée, on ne peut donc pas prédire l'indice de Gini d'une année à l'autre.

Par le test de Tukey réalisé à l'issue de l'ANOVA on observe une décroissance de l'indice de Gini avec l'âge.

On a :

	âge	indice de Gini	M
1	3	0.21	a
2	4	0.18	b
3	5	0.18	b
4	6	0.18	b

Ce tableau nous permet d'obtenir le graphique suivant :

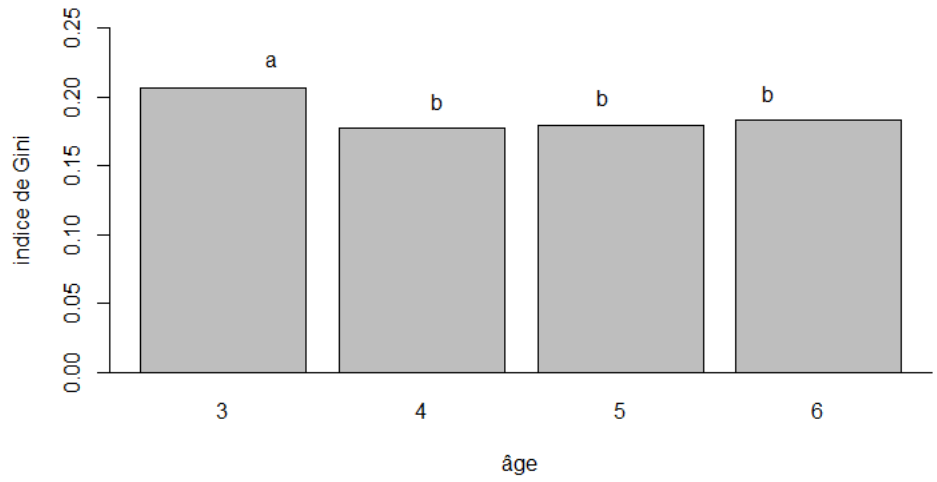


FIGURE 3.10 – Evolution de l'indice de Gini avec l'âge.

Les moyennes ayant la même lettre ne sont pas significativement différentes au seuil $\alpha = 0.001$

Interprétation 3.0.8. L'indice de Gini décroit avec l'âge, avec un effet significatif de l'âge sur l'indice de Gini entre 3 ans et les autres années. Autrement dit, l'indice de Gini décroit significativement entre 3 et 4 ans puis reste stable.

Influence de la production annuelle sur l'indice de Gini

On recherche maintenant s'il existe une corrélation linéaire entre la production annuelle des régimes et l'indice de Gini des parcelles élémentaires.

Après le test statistique on a le tableau de régression linéaire suivant :

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.2479	0.0022	111.96	2e-16
production annuelle	-0.0004	0.00001	-28.25	2e-16

Les résultats présentés dans ce tableau permettent de conclure que la production annuelle a un effet hautement significatif sur l'indice de Gini (P-value très faible).
Bien que la relation soit significative, le coefficient de corrélation de Pearson est faible, avec une valeur de -0.277. L'indice de Gini et la production annuelle sont donc significativement mais faiblement corrélés, et ce de manière négative.
Ainsi, nous avons le nuage de points suivant :

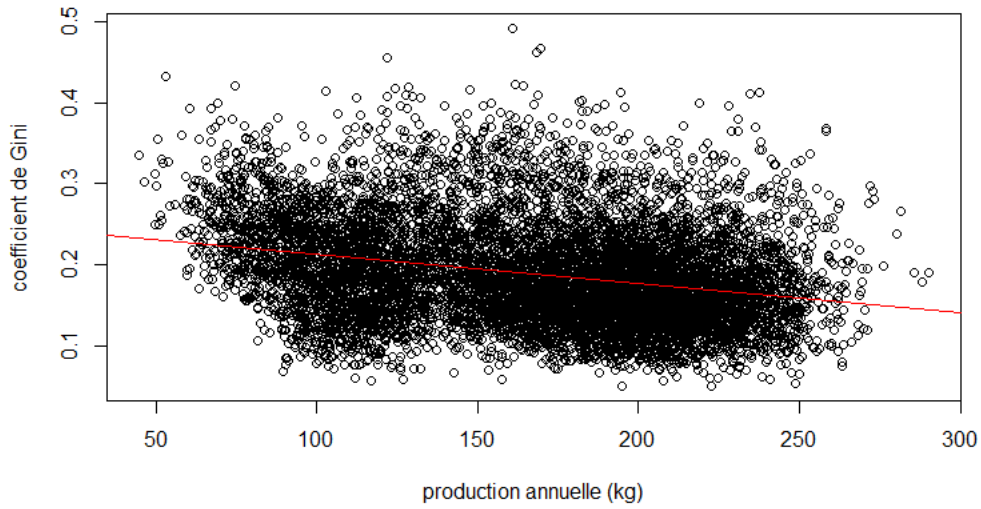


FIGURE 3.11 – Nuage des points entre l'indice de Gini et la production annuelle de régimes

Interprétation 3.0.9. Par le nuage des points et la droite de régression linéaire, on voit à nouveau qu'il y a corrélation entre l'indice de Gini et la production annuelle mais avec une faible pente (descendante). Ce que confirme le coefficient de corrélation.

3.4 Discussion

1. L'indice de Gini est l'indice le plus connu pour quantifier les inégalités. Il existe cependant de nombreux autres indices qui auraient pu aussi être utilisés ; tels que :

les indices d'entropie, le coefficient de variation et son carré, l'indice de Herfindahl, les rapports des quantiles etc.

Par exemple l'indice de Herfindahl appliqué à la répartition sur l'année de la production des régimes chez le palmier à huile serait la somme des carrés de la production mensuelle (exprimée en pourcentage par rapport à la production annuelle). Il donnerait plus de poids aux mois très productifs, et varierait entre $\frac{1}{12}$ et 1.

Tout comme l'indice de Gini, les indices cités plus haut sont invariants à l'échelle de mesure, ce qui est indispensable pour pouvoir comparer des croisements ayant des production annuelles variables. Par ailleurs, des indices non bornés, tels que le coefficient de variation, aurait peut être pu se trouver plus adaptés aux analyses telles que le modèle linéaire mixte, qui serait utilisé pour les études génétiques et qui nécessite que les données suivent une loi normale.

Cependant nous avons fait une rapide vérification en calculant avec R ces différents indices alternatifs, (grâce aux fonctions du package DescTools) et nous avons constaté que l'indice dont le résultat s'approche le plus d'une loi normale est Gini. (voir figure3.12 pour la distribution de l'indice de Herfindahl et du coefficient de variation)

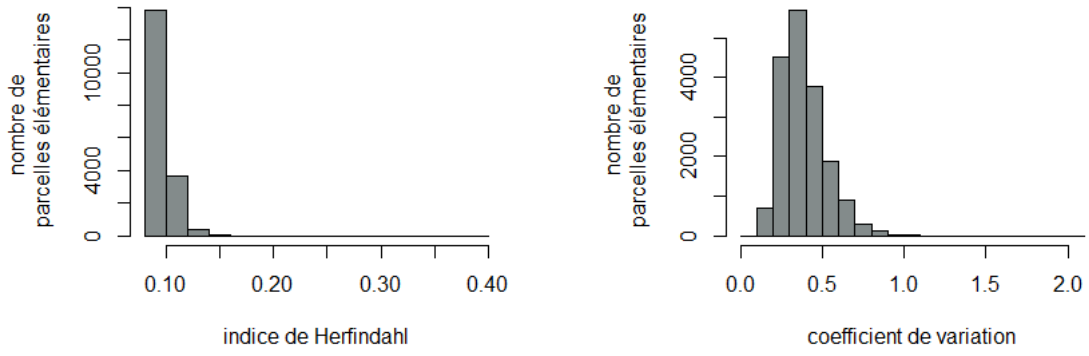


FIGURE 3.12 – distribution de l'indice de Herfindahl et du coefficient de variation

2. Dans nos résultats on a observé une faible corrélation avec une pente négative entre l'indice de Gini et la production annuelle. Ceci semble être favorable pour la culture du palmier à huile et l'amélioration génétique, qui laisse à croire que si on augmente la production annuelle dans les parcelles alors l'indice de Gini serait d'avantage faible.

3. Par ailleurs on s'interroge sur le fait qu'il y ait absence de corrélation entre les indices de Gini

obtenus aux différents âges. Cela peut-être lié au fait que toutes les parcelles n'ont pas été plantées toutes la même année alors que les conditions climatiques varient en fonction des années, et conditionnent la répartition. Donc il faudrait faire une analyse qui tienne en compte du dispositif expérimental. Pour cela, un modèle linéaire mixte semble adapté, car il pourrait gérer conjointement les effets associés au dispositif expérimental, et les effets aléatoires liés à la valeur génétique des croisements présents dans le dispositif. Par ailleurs, l'indice de Gini décroit entre 3 ans et 4 ans puis reste stable jusqu'à 6 ans, mais les palmiers étant encore jeunes il faudrait vérifier si l'indice de Gini reste toujours stable à l'âge adulte.

⋆ ⋆ Chapitre Quatre ⋆ ⋆

Implication pédagogique

Dans le cadre de notre travail, nous avons beaucoup utilisé les notions de statistiques. Cette notion est très enseignée dans les lycées camerounais, dans le nouveau programme établi par le MINESEC (ministère des enseignements secondaires). Cette notion commence en classe de cinquième mais les classes où cette notion est dense sont les classes de première et terminale.

Pour permettre à nos apprenants de mieux assimiler leur leçon en statistique, nous proposons des travaux pratiques (TP) pour les classes de première C et terminale D avec le logiciel sine qua non. Ce TP peut bien être fait avec le logiciel R mais la compréhension ou les procédures de calcul ne seront pas faciles pour les élèves.

Ce TP suppose que toutes les notions ont été faites en cours, il vise à consolider les acquis ou à susciter chez d'autres élèves une envie de mieux chercher à comprendre leurs cours.

4.1 Fiche des exercices

Données statistiques regroupées en classes

Exercice 1

Les notes de mathématiques d'une classe de première scientifique sont récapitulées dans le tableau ci-dessous.

Notes	[0 ;5[	[5 ;7[	[7 ;10[	[10 ;12[	[12 ;14[	[14 ;16[
Effectifs	6	8	10	3	1	1

1. Calculer la moyenne et l'écart-type de cette série statistique.
2. Construire un histogramme de cette série.
3. Construire le polygone des effectifs cumulés croissants et décroissants.

4. Par calcul déterminer la médiane de cette série statistique.

Séries statistiques à caractère double

Exercice 2

On a relevé dans un centre hospitalier le nombre d'accouchements et le nombre de césariennes effectuées par jour. Les résultats sont consignés dans le tableau suivant.

Nombre d'accouchements (x_i)	3	5	4	7	4	6	8
Nombre de césariennes (y_i)	1	2	4	3	1	5	3

1. Déterminer le nombre moyen d'accouchement et de césariennes de cette série statistique.
2. Représenter le nuage des points de cette série.
3. Calculer la covariance de x et y et le coefficient de corrélation linéaire.
4. Par la méthode des moindres carrés, déterminer l'équation de la droite d'ajustement affine de y en x

4.2 Solution guidée par le logiciel sine qua non

On procède d'abord au démarrage du logiciel.

Après le démarrage du logiciel l'interface se présente comme suit :

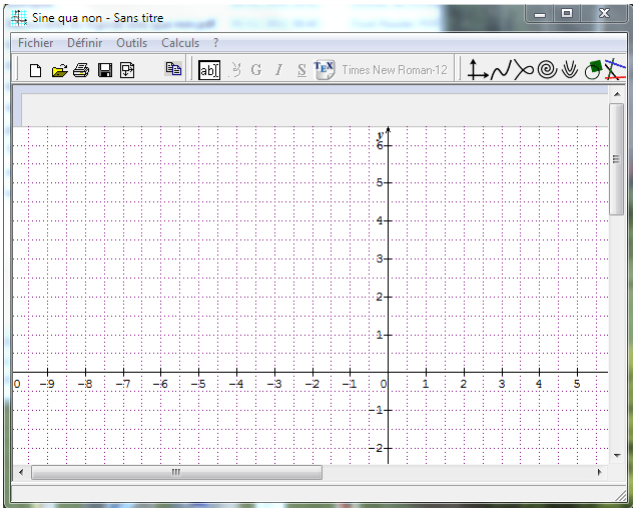


FIGURE 4.1 – interface de sine qua non

Dans la barre de tâche, on clique sur l'onglet qui permet de définir une série statistique simple (vaiables non numériques, valeurs isolées, valeurs regroupées en classes). Ensuite on clique sur "valeurs groupées en classes" on a ceci en figure

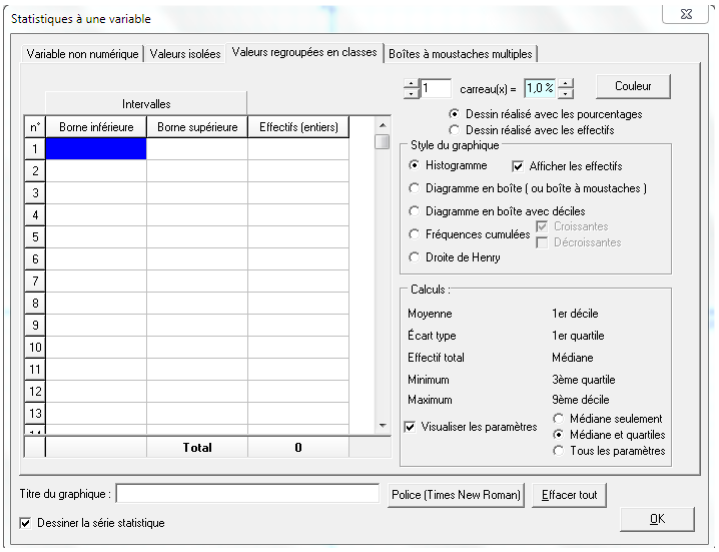


FIGURE 4.2 – valeurs regroupées en classes

Faisons entrer les données de notre série statistique tel qu'indique le tableau affiché à la figure4.2. Une fois les données entrées on a :

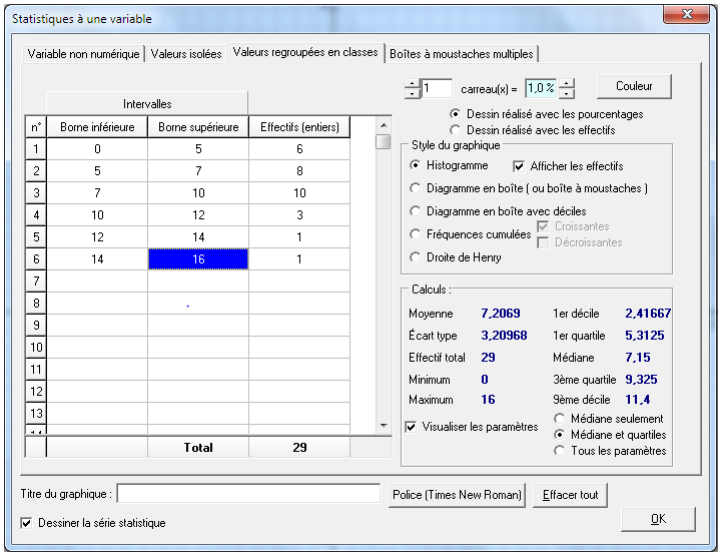


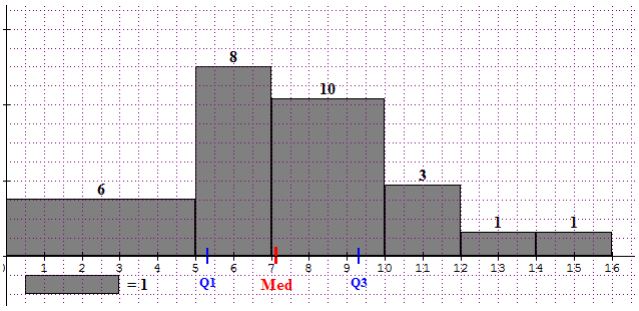
FIGURE 4.3 – Données insérées dans sine qua non

Après ceci nous avons immédiatement les résultats comme la moyenne, l'écart-type, l'effectif total, la médiane, les quartiles (premier et troisième). Nous avons également l'histogramme et le polygone des effectifs cumulés de cette série statistique, pour visualiser il suffit de choisir le point devant

"histogramme" ou "effectifs cumulés" puis "OK"

Tous les résultats obtenus ci-dessus (sauf les diagrammes) avec le logiciel sine qua non peuvent également s'obtenir à l'aide d'une calculatrice non programmable autorisée dans les salles des examens officiels. Mais toute fois on doit rappeler aux élèves de ne pas oublier les formules qui permettent de calculer les différents paramètres (moyenne, écart-type, etc) et la méthode de résolution d'un exercice de statistique. Tout résultat donné sans au préalable une formule n'est pas valide, et est supposé comme une tricherie.

Alors pour l'exercice 1 nous avons une moyenne de : 7.21, un écart-type de : 3.21. l'histogramme de cette série est :



On fait de même pour avoir le polygone des effectifs cumulés.

On suivra la même procédure pour avoir les résultats de l'exercice 2 (Séries statistiques à caractère double).

4.3 Intérêt didactique

Ce TP s'inscrit dans le cadre des enseignements avec les TIC (les Technologies de l'Information et de la Communication) qui a pour but de renforcer le développement de l'apprentissage scolaire.

L'avantage de faire ce TP est que les élèves trouveront un intérêt à connaitre ou à découvrir le logiciel sine qua non pour résourdre un exercice de statistique. Ce qui les emmènera à étudier encore plus leur leçon.

♠ Conclusion ♠

Dans notre travail, nous avons étudié la régularité de la répartition de la production du palmier à huile sur l'année par le biais de l'indice de Gini. La répartition de la production sur une plantation résultant entre autres de la synchronisation entre les palmiers qui la constituent, nous avons tout d'abord déterminé le nombre d'individus nécessaire pour que l'indice de Gini soit représentatif de la répartition de la production à l'échelle de la plantation. Nous avons ainsi montré qu'il fallait au moins 10 palmiers pour obtenir un indice de Gini représentatif. Nous avons ensuite vérifié que, dans ces conditions, l'indice de Gini était un bon indicateur pour mesurer la régularité de la production tout au long de l'année, en notant sur quelques parcelles extrêmes la bonne correspondance entre les valeurs de l'indice de Gini et les profils de production sur l'année. Nous avons aussi mis en évidence une grande variabilité dans l'indice de Gini, ce qui laisse espérer des possibilités d'amélioration génétique. Enfin, par des tests statistiques nous avons vu que l'indice de Gini décroit légèrement mais significativement après 3 ans avant de se stabiliser, et qu'il existe une légère mais significative corrélation négative entre l'indice de Gini et la production annuelle de régimes, ce qui est favorable à la filière. A la suite de ce travail, une étude de génétique visant à comprendre la transmission de ce caractère des parents aux enfants est nécessaire, afin que les sélectionneurs puissent développer des croisements avec une production plus régulièrement répartie sur l'année. Aussi, la poursuite de l'étude à l'âge adulte sera utile.

♠ Bibliographie ♠

[1] R Corley and P Tinker. Selection and breeding. *The oil palm. 4th ed. Oxford : Blackwell Science Ltd Blackwell Publishing*, pages 133–200, 2003.

[2] RHV Corley. How much palm oil do we need ? *Environmental Science & Policy*, 12(2) :134–139, 2009.

[3] David Cros, Albert Flori, Léifi Nodichao, Alphonse Omoré, and Bruno Nouy. Differential response to water balance and bunch load generates diversity of bunch production profiles among oil palm crosses (elaeis guineensis). *Tropical plant biology*, 6(1) :26–36, 2013.

[4] Partha Dasgupta, Amartya Sen, and David Starrett. Notes on the measurement of inequality. *Journal of economic theory*, 6(2) :180–187, 1973.

[5] Pierre Lafaye De Micheaux, Rémy Drouilhet, and Benoît Liquet. *Le logiciel R : Maitriser le langage-Effectuer des analyses statistiques*. Springer Science & Business Media, 2011.

[6] Yadolah Dodge. *Statistique : dictionnaire encyclopédique*. Springer Science & Business Media, 2007.

[7] A Druckman and T Jackson. Measuring resource inequalities : The concepts and methodology for an area-based gini coefficient. *Ecological economics*, 65(2) :242–252, 2008.

[8] JP Gascon and C De Berchoux. Caractéristiques de la production d'elaeis guineensis (jacq.) de diverses origines et leurs croisements. *Application à la sélection du palmier à huile. Oléagineux*, 19(2) :75–84, 1964.

[9] Piotr P Graczyk. Gini coefficient : a new way to express selectivity of kinase inhibitors against a family of kinases. *Journal of medicinal chemistry*, 50(23) :5773–5779, 2007.

[10] Chris J Groves-Kirkby, Anthony R Denman, and Paul S Phillips. Lorenz curve and gini coefficient : novel tools for analysing seasonal variation of environmental radon gas. *Journal of environmental management*, 90(8) :2480–2487, 2009.

[11] Thorsten Lisker. Is the gini coefficient a stable measure of galaxy structure ? *The Astrophysical Journal Supplement Series*, 179(2) :319, 2008.

[12] B Nouy, L Baudouin, N Djégui, and A Omoré. Le palmier à huile en conditions hydriques limitantes. *Plant. Rech. Dév*, pages 31–40, 1999.

[13] B Nouy, A Omore, and F Potier. Oil palm production cycles in different ecologies : consequences for breeding. In *International Palm Oil Congress, Competitiveness for the 21st Century. PORIM Kuala Lumpur, Malaysia*, pages 62–75, 1996.

[14] Wikipédia. Coefficient de gini — wikipédia, l'encyclopédie libre, 2016. [En ligne ; Page disponible le 31-mai-2016].

[15] Wikipedia. Diversity index — wikipedia, the free encyclopedia, 2016. [Online ; accessed 31-May-2016].

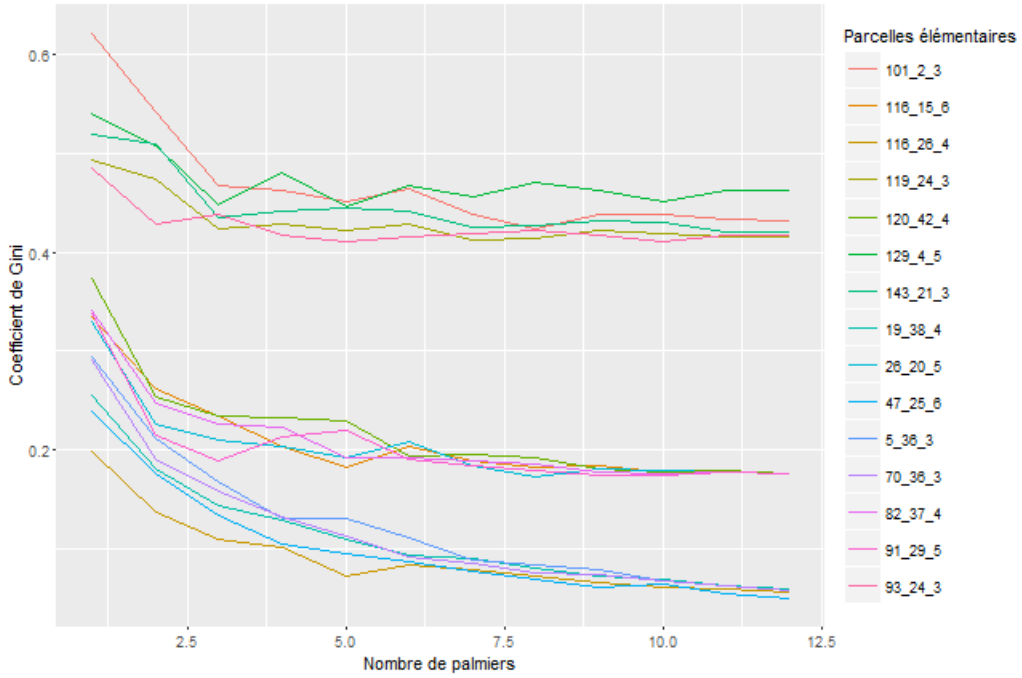
♠ Annexe ♠

Annexe I

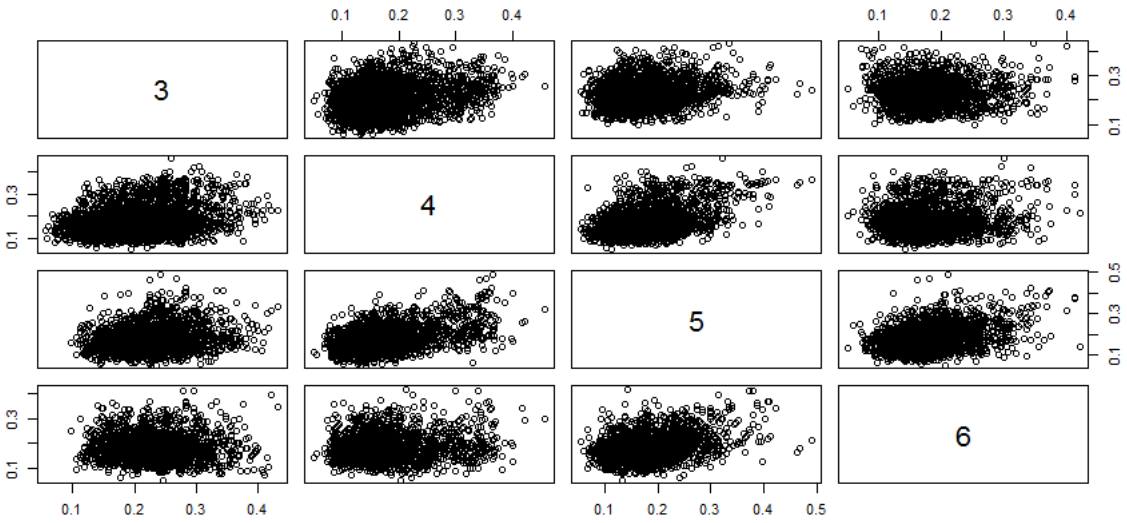
Ce tableau nous a permis de faire le graphique pour déterminer le nombre d'arbre seuil

	Classe d'indice de Gini	Nombre de palmiers	Indice de Gini moyen
1	élevé	1.00	0.53
2	faible	1.00	0.26
3	intermédiaire	1.00	0.34
4	élevé	2.00	0.46
5	faible	2.00	0.18
6	intermédiaire	2.00	0.25
7	élevé	3.00	0.45
8	faible	3.00	0.14
9	intermédiaire	3.00	0.23
10	élevé	4.00	0.44
11	faible	4.00	0.12
12	intermédiaire	4.00	0.21
13	élevé	5.00	0.44
14	faible	5.00	0.10
15	intermédiaire	5.00	0.20
16	élevé	6.00	0.45
17	faible	6.00	0.09
18	intermédiaire	6.00	0.20
19	élevé	7.00	0.44
20	faible	7.00	0.08
21	intermédiaire	7.00	0.18
22	élevé	8.00	0.43
23	faible	8.00	0.08
24	intermédiaire	8.00	0.18
25	élevé	9.00	0.43
26	faible	9.00	0.07
27	intermédiaire	9.00	0.18
28	élevé	10.00	0.43
29	faible	10.00	0.07
30	intermédiaire	10.00	0.18
31	élevé	11.00	0.43
32	faible	11.00	0.06
33	intermédiaire	11.00	0.18
34	élevé	12.00	0.43
35	faible	12.00	0.06
36	intermédiaire	12.00	0.18

Ce graphique permet de déterminer le nombre d'arbre seuil qu'il faut par parcelle élémentaire pour avoir un indice de Gini représentatif.



Annexe II



Script R de la fonction Gini du package DescTools

Le script R suivant programme la formule de l'indice de Gini de l'équation1.1

function (x, n = rep(1, length(x)), unbiased = TRUE, conf.level = NA,

```
R = 1000, type = "bca", na.rm = FALSE) {
x <- rep(x, n)
if (na.rm)
x <- na.omit(x)
if (any(is.na(x)) || any(x < 0))
return(NA_real)
i.gini <- function(x, unbiased = TRUE){
n <- length(x)
x <- sort(x)
res <- 2 * sum(x * 1 :n)/(n * sum(x)) - 1 - (1/n)
if (unbiased)
res <- n/(n - 1) * res
return(pmax(0, res))
}
if (is.na(conf.level)) {
res <- i.gini(x, unbiased = unbiased)
}
else {
boot.gini <- boot(x, function(x, d) i.gini(x[d], unbiased = unbiased),
R = R)
ci <- boot.ci(boot.gini, conf = conf.level, type = type)
res <- c(gini = boot.gini$t0, lwr.ci = ci[[4]][4], upr.ci = ci[[4]][5])
}
return(res)
}
```