

NOTES DE STATISTIQUE ET D'INFORMATIQUE

2012/1

TABLES DE CONTINGENCE À TROIS
DIMENSIONS : ASPECTS THÉORIQUES,
APPLICATIONS ET ANALOGIE AVEC
L'ANALYSE DE LA VARIANCE À TROIS
CRITÈRES DE CLASSIFICATION

J. J. CLAUSTRIAUX, R. PALM, S. FERRANDIS-
VALLTERRA, Y. BROSTAUX et V. PLANCHON

Université de Liège – Gembloux Agro-Bio Tech
*Unité de Statistique, Informatique et Mathématique
appliquées à la bioingénierie*

GEMBLOUX

(Belgique)

**TABLES DE CONTINGENCE À TROIS DIMENSIONS :
ASPECTS THÉORIQUES, APPLICATIONS ET ANALOGIE
AVEC L'ANALYSE DE LA VARIANCE À TROIS
CRITÈRES DE CLASSIFICATION**

J. J. CLAUSTRIAUX^{*}, R. PALM[†], S. FERRANDIS-VALLTERRA[‡],
Y. BROSTAUX[§] et V. PLANCHON[¶]

RÉSUMÉ

Les notions d'indépendance mises en évidence lors de l'étude d'une table de contingence, à trois dimensions au moins, ne semblent pas être évidentes à comprendre en pratique.

C'est pourquoi, des aspects théoriques faisant référence au modèle log-linéaire sont détaillés et illustrés, et une analogie avec l'analyse de la variance à trois critères de classification, modèle fixe est établie.

SUMMARY

For contingency table, concepts of independence are not so easy to understand in statistical practice. So, theoretical aspects of the log-linear model are described and applied to a three levels example; an analogy with analysis of variance is also established.

1. INTRODUCTION

Cette note a essentiellement un caractère didactique. En effet, une longue pratique de l'enseignement des méthodes statistiques et de la consultation statistique a mis en évidence la difficulté pour un étudiant et pour un chercheur

^{*}Professeur ordinaire à l'Université de Liège, Gembloux Agro-Bio Tech.

[†]Professeur à l'Université de Liège, Gembloux Agro-Bio Tech.

[‡]Assistant à l'Université de Liège, Gembloux Agro-Bio Tech.

[§]Chef de Travaux et Chargé de cours à l'Université de Liège, Gembloux Agro-Bio Tech.

[¶]Attachée scientifique, Centre wallon de Recherches agronomiques.

de comprendre et d'interpréter les différents types d'indépendance qui peuvent être mis en évidence dans une table de contingence, alors que l'un et l'autre maîtrisent suffisamment les concepts de l'analyse de la variance, en particulier les notions d'interaction et de hiérarchisation des facteurs analysés.

Dès lors, après cette introduction (paragraphe 1), des notions théoriques permettant l'analyse statistique d'une table de contingence sont introduites (paragraphe 2), d'une part, pour deux critères, et, d'autre part, pour trois critères.

Ensuite, une application est présentée (paragraphe 3) et la correspondance entre les notions d'indépendance (table de contingence) et d'absence d'interaction (analyse de la variance) est établie.

Une brève conclusion clôture cette publication (paragraphe 4).

En terminant cette introduction, que Madame D. MARCHAL soit remerciée pour sa contribution à la mise au point de la publication.

2. ASPECTS THÉORIQUES

2.1. Modèle log-linéaire pour deux critères indépendants

1° L'analyse d'une table de contingence peut se réaliser en se référant au modèle log-linéaire. Comme le fait DAGNELIE [2011], une présentation de ce modèle dans le cas d'une table à deux dimensions ($i : 1, \dots, p; j : 1, \dots, q$) permet de mieux comprendre son usage. Signalons aussi qu'une ancienne publication de FIENBERG [1970] est très intéressante à consulter si besoin pour étendre ces notions à une table à trois dimensions. De nombreux éléments théoriques sont aussi repris dans les contributions de SIMAR [1979] et ROLIN [1979].

2° En cas d'indépendance entre deux événements A et B ou par extension entre deux niveaux i et j de deux critères de classification, on peut écrire :

$$P(A \text{ et } B) = P(A) P(B)$$

ou

$$P_{ij} = P_i \cdot P_j$$

ou encore

$$nP_{ij} = (nP_i \cdot nP_j)/n,$$

sachant que :

$$n = \sum_{i=1}^p \sum_{j=1}^q n_{ij}.$$

Par transformation logarithmique, l'avant-dernière formule s'écrit :

$$\log(nP_{ij}) = \log(nP_{i.}) + \log(nP_{.j}) - \log n.$$

Ajoutons et soustrayons à cette équation les termes :

$$\frac{1}{q} \sum_{j=1}^q \log(nP_{.j}), \quad \frac{1}{p} \sum_{i=1}^p \log(nP_{i.}) \quad \text{et} \quad \log n.$$

L'équation se transforme alors de la manière suivante :

$$\begin{aligned} \log(nP_{ij}) &= \log(nP_{i.}) + \log(nP_{.j}) - \log n \\ &+ \frac{1}{q} \sum_{j=1}^q \log(nP_{.j}) - \frac{1}{q} \frac{p}{p} \sum_{j=1}^q \log(nP_{.j}) \\ &+ \frac{1}{p} \sum_{i=1}^p \log(nP_{i.}) - \frac{1}{p} \frac{q}{q} \sum_{i=1}^p \log(nP_{i.}) \\ &+ \frac{pq}{pq} \log n - \log n \end{aligned}$$

$$\begin{aligned} \log(nP_{ij}) &= \log(nP_{i.}) + \log(nP_{.j}) \\ &+ \frac{1}{q} \sum_{j=1}^q \log(nP_{.j}) + \frac{1}{p} \sum_{i=1}^p \log(nP_{i.}) - 2 \log n \\ &- \left[\frac{1}{pq} \left(q \sum_{i=1}^p \log(nP_{i.}) + p \sum_{j=1}^q \log(nP_{.j}) - pq \log n \right) \right] \end{aligned}$$

$$\begin{aligned} \log(nP_{ij}) &= \frac{q}{q} \log(nP_{i.}) + \frac{1}{q} \sum_{j=1}^q \log(nP_{.j}) - \frac{q}{q} \log n \\ &+ \frac{p}{p} \log(nP_{.j}) + \frac{1}{p} \sum_{i=1}^p \log(nP_{i.}) - \frac{p}{p} \log n \\ &- \left[\frac{1}{pq} \left(q \sum_{i=1}^p \log(nP_{i.}) + p \sum_{j=1}^q \log(nP_{.j}) - pq \log n \right) \right] \end{aligned}$$

$$\begin{aligned}
\log(nP_{ij}) &= \frac{1}{q} \sum_{j=1}^q [(\log(nP_{i.}) + \log(nP_{.j}) - \log n)] \\
&\quad + \frac{1}{p} \sum_{i=1}^p [(\log(nP_{i.}) + \log(nP_{.j}) - \log n)] \\
&\quad - \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q [(\log(nP_{i.}) + \log(nP_{.j}) - \log n)] \\
\log(nP_{ij}) &= \frac{1}{q} \sum_{j=1}^q \log(nP_{ij}) + \frac{1}{p} \sum_{i=1}^p \log(nP_{ij}) \\
&\quad - \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q \log(nP_{ij}).
\end{aligned}$$

Notons encore comme suit les différentes composantes :

$$[1] = \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q \log(nP_{ij})$$

ou moyenne arithmétique générale des logarithmes des fréquences attendues,

$$[1] + [A]_i = \frac{1}{q} \sum_{j=1}^q \log(nP_{ij})$$

ou moyenne arithmétique des logarithmes des fréquences attendues pour le niveau i du premier critère noté A , non ajustée, par rapport à la « moyenne générale »,

$$[1] + [B]_j = \frac{1}{p} \sum_{i=1}^p \log(nP_{ij})$$

ou moyenne arithmétique des logarithmes des fréquences attendues pour le niveau j du second critère noté B , non ajustée par rapport à la « moyenne générale ».

Alors, on aboutit à l'équation du modèle non saturé qui s'écrit comme suit :

$$\log(nP_{ij}) = [1] + [A]_i + [B]_j$$

où :

$$[A]_i = \frac{1}{q} \sum_{j=1}^q \log(nP_{ij}) - [1]$$

est l'écart de l'« effet i du critère A » par rapport à la « moyenne générale » et

$$[B]_j = \frac{1}{p} \sum_{i=1}^p \log(nP_{ij}) - [1]$$

est l'écart de l'« effet j du critère B » par rapport à la « moyenne générale ».

Comme en analyse de la variance, il en résulte que :

$$\sum_{i=1}^p [A]_i = \sum_{j=1}^q [B]_j = 0.$$

2.2. Modèle log-linéaire pour deux critères non indépendants et tests d'ajustement

1° Pour étudier explicitement le lien entre les deux critères de classification, il faut introduire un terme d'interaction dans le modèle, à savoir :

$$[AB]_{ij} = \log(nP_{ij}) - [1],$$

sachant que :

$$\sum_{i=1}^p \sum_{j=1}^q [AB]_{ij} = 0.$$

Le modèle s'écrit alors comme suit :

$$\log(nP_{ij}) = [1] + [A]_i + [B]_j + [AB]_{ij}.$$

Ce modèle est saturé car le nombre de paramètres à estimer est égal au nombre de données disponibles, c'est-à-dire aussi qu'il n'y a aucune variation résiduelle.

Pour comprendre simplement le qualificatif saturé, considérons une table 2×2 . Disposant de quatre données, l'ajustement du modèle nécessite quatre estimations, à savoir $[1]$, $[A_1]$ ou $[A_2]$, $[B_1]$ ou $[B_2]$, et l'un des quatre termes $[AB_{ij}]$, les trois autres termes étant fixés suite aux estimations précédentes qui toutes tiennent compte des propriétés de la nullité des sommes.

Signalons encore qu'à deux dimensions, la notion d'indépendance et celle d'absence d'interaction, rencontrée en analyse de la variance, se confondent. Leurs différences apparaîtront par la suite.

2° Pour une table à deux critères, on pourrait débiter l'inférence en testant l'égalité des fréquences marginales relatives à chaque critère, c'est-à-dire en testant l'adéquation des ajustements aux modèles suivants :

$$\log(nP_{ij}) = [1] + [A]_i,$$

$$\log(nP_{ij}) = [1] + [B]_j,$$

les degrés de liberté pour ces tests d'égalité de proportions étant respectivement égaux à $p - 1$ et $q - 1$.

En pratique, lorsque l'utilisateur dispose d'une telle table, il s'intéresse essentiellement à l'étude de l'indépendance entre les deux critères de classification et il effectue le test de l'hypothèse nulle reprise ci-après :

$$H_0 : A \text{ II } B.$$

Le symbole II est utilisé pour identifier la notion d'indépendance entre les deux critères.

3° Le test proprement dit se réalise en mesurant les écarts entre les fréquences observées n_{ij} et les fréquences attendues :

$$n\hat{P}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n},$$

par l'intermédiaire du test de WILKS (critère du rapport de vraisemblance) :

$$G_{obs}^2 = 2 \sum_{i=1}^p \sum_{j=1}^q \left[n_{ij} \log_e \left(\frac{n_{ij}}{n\hat{P}_{ij}} \right) \right].$$

ou du test χ^2 de PEARSON (critère de RAO) :

$$\chi_{obs}^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - n\hat{P}_{ij})^2}{n\hat{P}_{ij}}.$$

Le rejet de l'hypothèse nulle (RH_0) s'effectue comme suit :

$$RH_0 \text{ si } G_{obs}^2 \text{ ou } \chi_{obs}^2 \geq \chi_{1-\alpha}^2,$$

le nombre de degrés de liberté associé à la variable χ^2 de PEARSON étant égal à :

$$pq - [1 + (p - 1) + (q - 1)] = (p - 1)(q - 1).$$

Le non rejet de l'hypothèse nulle signifie que le modèle retenu est le modèle non saturé suivant :

$$\log(nP_{ij}) = [1] + [A]_i + [B]_j.$$

Signalons que l'approximation à la distribution χ^2 est satisfaisante si les fréquences attendues sont au moins égales à 5. Par ailleurs, sous l'hypothèse du modèle non saturé, les deux tests sont asymptotiquement équivalents, sachant que la statistique G_{obs}^2 est préférée à la statistique χ_{obs}^2 car, en cas de décompositions, le G^2 total est égal à la somme des écarts G^2 calculés pour tester les différentes hypothèses successives à trois critères (voir les propriétés ci-dessus), ce qui n'est pas nécessairement le cas pour une décomposition de la valeur χ^2 totale en χ^2 partielles. Ces écarts peuvent être utiles pour une construction progressive d'un modèle, comme c'est le cas en régression multiple pour évaluer l'apport successif des différentes variables explicatives.

2.3. Modèles log-linéaire pour trois critères

1° Une table de contingence à trois dimensions ($i : 1, \dots, p; j : 1, \dots, q; k : 1, \dots, r$) peut tout d'abord s'analyser sous l'angle de l'indépendance complète ou mutuelle entre les trois critères.

Le modèle log-linéaire sous-jacent, l'hypothèse nulle associée, la méthode d'estimation des fréquences attendues et les degrés de liberté (dl) correspondant à la quantité G^2 étendue à trois dimensions et calculée pour tester l'hypothèse nulle, sont successivement présentés :

$$\begin{aligned} \log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k \\ H_0 : A \amalg B \amalg C \\ n\hat{P}_{ijk} &= (n_{i..}n_{.j.}n_{..k})/n \\ n &= \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r n_{ijk} \\ dl &= pqr - [1 + (p-1) + (q-1) + (r-1)] = pqr - p - q - r + 2 \\ G_{obs}^2 &= 2 \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r \left[n_{ijk} \log_e \left(\frac{n_{ijk}}{n\hat{P}_{ijk}} \right) \right] \\ RH_0 &\text{ si } G_{obs}^2 \geq \chi_{1-\alpha}^2. \end{aligned}$$

Intuitivement, le nombre de degrés de liberté repris ci-dessus représente l'ensemble des « liens » possibles entre les trois critères.

Dès lors, en cas de rejet de cette hypothèse nulle d'indépendance complète, on peut prendre trois chemins différents pour tenter de comprendre le lien entre les facteurs ; ils correspondent à l'étude des indépendances partielles de type 1, 2 ou 3. Elles sont successivement décrites ci-après en donnant des informations analogues au cas précédent.

Type 1

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [BC]_{jk} \\ H_0^1 &: A \amalg (B, C) \\ n\hat{P}_{ijk} &= (n_{i..}n_{.jk})/n \\ dl &= (p-1)(qr-1)\end{aligned}$$

Type 2

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [AC]_{ik} \\ H_0^2 &: B \amalg (A, C) \\ n\hat{P}_{ijk} &= (n_{.j.}n_{i.k})/n \\ dl &= (q-1)(pr-1)\end{aligned}$$

Type 3

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [AB]_{ij} \\ H_0^3 &: C \amalg (A, B) \\ n\hat{P}_{ijk} &= (n_{..k}n_{ij.})/n \\ dl &= (r-1)(pq-1)\end{aligned}$$

Comme on le constate au niveau de l'écriture de chaque hypothèse et des degrés de liberté, on étudie chaque fois l'indépendance entre un critère et l'association des deux autres, ce qui correspond à l'analyse d'une table à deux dimensions.

Par ailleurs, les degrés de liberté sont chaque fois inférieurs à ceux relatifs à l'indépendance complète entre les trois critères. Les différences sont égales à :

$$\begin{aligned}(q-1)(r-1) & \text{ pour le type 1,} \\ (p-1)(r-1) & \text{ pour le type 2,} \\ (p-1)(q-1) & \text{ pour le type 3.}\end{aligned}$$

Cela résulte des propriétés qui lient les différents types d'indépendance stochastique, à savoir :

$$\begin{aligned}\text{si } & A \amalg B \amalg C, \text{ alors :} \\ & A \amalg (B, C) \quad \text{et } B \amalg C, \\ \text{ou } & B \amalg (A, C) \quad \text{et } A \amalg C, \\ \text{ou } & C \amalg (A, B) \quad \text{et } A \amalg B.\end{aligned}$$

2° Le rejet de l'hypothèse nulle relative à une indépendance partielle peut chaque fois ouvrir deux nouvelles pistes pour étudier les indépendances conditionnelles. Elles sont résumées comme suit.

Type 1a

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [AC]_{ik} + [BC]_{jk} \\ H_0^{1a} &: (A \amalg B)|C \\ n\hat{P}_{ijk} &= (n_{i.k}n_{.jk})/n_{..k} \\ dl &= (p-1)(q-1)r\end{aligned}$$

Type 1b

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [AB]_{ij} + [BC]_{jk} \\ H_0^{1b} &: (A \amalg C)|B \\ n\hat{P}_{ijk} &= (n_{ij.}n_{.jk})/n_{.j.} \\ dl &= (p-1)(r-1)q\end{aligned}$$

Type 2a = type 1a

Type 2b

$$\begin{aligned}\log(nP_{ijk}) &= [1] + [A]_i + [B]_j + [C]_k + [AB]_{ij} + [AC]_{ik} \\ H_0^{2b} &: (B \amalg C)|A \\ n\hat{P}_{ijk} &= (n_{ij.}n_{i.k})/n_{i..} \\ dl &= (q-1)(r-1)p\end{aligned}$$

Type 3a = type 1b

Type 3b = type 2b

Par conséquent, dans le cas des indépendances conditionnelles, la table à trois dimensions est étudiée par analyse de tables à deux dimensions, établies pour chaque niveau du troisième critère choisi, c'est-à-dire de façon hiérarchisée par rapport à ce critère.

A nouveau, les degrés de liberté associés sont inférieurs aux degrés de liberté utilisés pour l'étude des indépendances partielles; les différences correspondent à l'une des quantités suivantes :

$$\begin{aligned} (p-1)(r-1) & \text{ pour les types 1a et 3b,} \\ (p-1)(q-1) & \text{ pour les types 1b et 2b,} \\ (q-1)(r-1) & \text{ pour les types 2a et 3a.} \end{aligned}$$

en vertu des propriétés suivantes :

$$\begin{aligned} \text{si } A \text{ II } (B, C), \text{ alors :} \\ & (A \text{ II } B) \mid C \quad \text{et } A \text{ II } C, \\ \text{ou } (A \text{ II } C) \mid B \quad & \text{et } A \text{ II } B, \\ \\ \text{si } B \text{ II } (A, C), \text{ alors :} \\ & (B \text{ II } A) \mid C \quad \text{et } B \text{ II } C, \\ \text{ou } (B \text{ II } C) \mid A \quad & \text{et } B \text{ II } A, \\ \\ \text{si } C \text{ II } (A, B), \text{ alors :} \\ & (C \text{ II } A) \mid B \quad \text{et } C \text{ II } B, \\ \text{ou } (C \text{ II } B) \mid A \quad & \text{et } C \text{ II } A, \end{aligned}$$

3° Si tous les types d'indépendances sont rejetés, la notion d'interaction entre les trois critères est alors introduite dans le modèle :

$$\log(nP_{ijk}) = [1] + [A]_i + [B]_j + [C]_k + [AB]_{ij} + [AC]_{ik} + [BC]_{jk} + [ABC]_{ijk}.$$

L'estimation des nP_{ijk} se réalise en appliquant l'algorithme itératif suivant, avec arrêt dès que la convergence souhaitée est atteinte, c'est-à-dire que les différences entre les fréquences attendues calculées pour deux étapes successives sont considérées comme nulles ou négligeables.

Initiation de l'algorithme

Calculer $n_{ij.}, n_{i.k}, n_{.jk}$, fixer toutes les fréquences attendues à l'unité, à savoir $(n\hat{P}_{ijk})_s = 1$, et calculer les totaux marginaux $(n\hat{P}_{ij.})_s$; fixer l'indice $s = 0$.

Itérations

Etape 1 : calculer les fréquences attendues comme suit :

$$(n\hat{P}_{ijk})_1 = (n\hat{P}_{ijk})_s \frac{n_{ij.}}{(n\hat{P}_{ij.})_s}$$

et calculer les totaux marginaux $(n\hat{P}_{i.k})_1$;

Etape 2 : calculer les fréquences attendues comme suit :

$$(n\hat{P}_{ijk})_2 = (n\hat{P}_{ijk})_1 \frac{n_{i.k}}{(n\hat{P}_{i.k})_1}$$

et les totaux marginaux $(n\hat{P}_{.jk})_2$;

Etape 3 : calculer les fréquences attendues comme suit :

$$(n\hat{P}_{ijk})_3 = (n\hat{P}_{ijk})_2 \frac{n_{.jk}}{(n\hat{P}_{.jk})_2}$$

et les totaux marginaux $(n\hat{P}_{ij.})_3$;

Etape 4 : poser $s = s + 1$, $(n\hat{P}_{ij.})_s = (n\hat{P}_{ij.})_3$, $(n\hat{P}_{ijk})_s = (n\hat{P}_{ijk})_3$, et retour à l'étape 1.

4° Pour tester l'absence d'interaction, la quantité G_{obs}^2 est calculée en utilisant les fréquences attendues déterminées par cette procédure itérative.

Quant aux degrés de liberté de la variable χ^2 , ils valent $(p-1)(q-1)(r-1)$.

5° Pour être complet, signalons encore qu'à partir d'une table de contingence à trois dimensions, d'autres hypothèses plus simples peuvent aussi être testées. Ainsi, une table de contingence à trois dimensions peut être considérée comme une table à une dimension pour laquelle on veut tester l'égalité des pqr proportions (ajustement à une distribution uniforme) :

$$\begin{aligned} \log(nP_{ijk}) &= [1] \\ H_0 : P_{111} &= \dots = P_{ijk} = \dots = P_{pqr} \\ nP_{ijk} &= n/pqr \\ dl &= (pqr - 1). \end{aligned}$$

On peut aussi s'intéresser à chaque critère séparément ; à nouveau il s'agit de tests d'égalité de proportions qui, en principe, sont réalisés si les hypothèses d'indépendance et de l'interaction des critères n'ont pas été rejetées :

- pour le critère A uniquement :

$$\begin{aligned} \log(nP_{ijk}) &= [1] + [A]_i \\ H_0 : P_{1..} &= \dots = P_{i..} = \dots = P_{p..} \\ nP_{i..} &= n/p \\ dl &= (p - 1); \end{aligned}$$

- pour le critère B uniquement :

$$\begin{aligned} \log(nP_{ijk}) &= [1] + [B]_j \\ H_0 : P_{.1.} &= \dots = P_{.j.} = \dots = P_{.q.} \\ nP_{.j.} &= n/q \\ dl &= (q - 1); \end{aligned}$$

- pour le critère C uniquement :

$$\begin{aligned} \log(nP_{ijk}) &= [1] + [C]_k \\ H_0 : P_{.1} &= \dots = P_{..k} = \dots = P_{..r} \\ nP_{..k} &= n/r \\ dl &= (r - 1). \end{aligned}$$

3. APPLICATION

3.1. Données

Les données de l'exemple sont extraites d'une recherche menée par GILLET [1993]. Celle-ci s'inscrivait dans le cadre d'un travail sur la modélisation du développement végétatif et génératif du pommier, financé par l'Institut pour l'Encouragement de la Recherche Scientifique dans l'Industrie et l'Agriculture. Elles figurent également dans l'ouvrage de DAGNELIE [2011].

Plus précisément, des nombres de rameaux ont été observés sur deux variétés de pommiers (V1 : Jonagold ; V2 : Cox), situés à différents niveaux de croissance (R1 : premier ordre ; R2 : deuxième ordre ; R3 : troisième ordre) et qui se distinguent par la présence (F1) ou l'absence (F2) de fleurs. Les arbres étant greffés, les rameaux de premier ordre sont les rameaux directement insérés sur le porte-greffe, ce qui correspond finalement au tronc de l'arbre, les rameaux de deuxième ordre sont ceux qui ont pris naissance sur les rameaux de premier ordre, ce qui correspond aux branches charpentières, et les rameaux de troisième ordre sont ceux qui ont pris naissance sur les rameaux de deuxième ordre. Signalons encore que d'autres informations sur l'objet de la recherche sont exposées au fur et à mesure de l'analyse statistique des données ; elles seront utiles à la compréhension du cheminement de cette dernière.

Le tableau 1 rassemble, d'une part, la description des combinaisons des critères étudiés (cellules i, j, k), ainsi que les fréquences observées et, d'autre part, les fréquences attendues pour l'étude de l'interaction (paragraphe 3.2, 6⁰).

Tableau 1 – Pour deux variétés (lettre V ; indice i), nombres de rameaux situés à trois niveaux de croissance (lettre R ; indice j), fleuris et non fleuris (lettre F ; indice k) et fréquences attendues pour l'étude de l'interaction.

Cellules (i, j, k)	Fréquences observées	Fréquences attendues
1 1 1	8	13,03
1 1 2	14	8,97
1 2 1	102	107,58
1 2 2	186	180,41
1 3 1	84	73,39
1 3 2	79	89,61
2 1 1	16	10,97
2 1 2	3	8,03
2 2 1	58	52,42
2 2 2	88	93,59
2 3 1	22	32,61
2 3 2	53	42,39
Totaux	713	713

3.2. Analyse statistique

1° Si on appliquait l'analyse de la variance à trois critères de classification, modèle croisé fixe, échantillon unitaire, on obtiendrait la décomposition des degrés de liberté en fonction des sources de variation données dans le tableau 2. C'est à partir de ce tableau que l'analogie entre la décomposition des indépendances stochastiques et l'analyse de la variance est établie.

2° Pour débiter l'analyse de la table de contingence à trois dimensions, calculons les valeurs G^2 (tableau 3) pour l'ensemble des douze combinaisons des critères pris distinctement (Totaux), pour chaque critère pris séparément (V , F et R) et en rassemblant en une seule source de variation (Autres) tout ce qui concerne les notions de dépendance et d'interaction. La valeur G_{obs}^2 reprise au tableau 3 sur la ligne « Totaux » permet de tester l'égalité des douze proportions P_{ijk} . Les valeurs de G_{obs}^2 correspondant à V , F et R permettent de tester des proportions pour chaque critère (paragraphe 2.3, 5⁰). Enfin, la valeur G_{obs}^2 pour la source de variation « Autres » permet de tester l'indépendance des trois critères (paragraphe 2.3, 1⁰).

Au niveau inférentiel, l'hypothèse nulle (source de variation « Autres ») est

Tableau 2 – Tableau d’analyse de la variance.

Sources de variation	Degrés de liberté
Variétés (V)	1
Floraison ou pas (F)	1
Types de rameaux (R)	2
$V \times F$	1
$V \times R$	2
$F \times R$	2
$V \times F \times R$	2
Total	11

Tableau 3 – Table de contingence : 1^{re} partie.

Sources de variation	Degrés de liberté	G_{obs}^2	$\chi_{1-\alpha}^2$ $\alpha = 0,05$
V	1	77,56	3,84
F	1	24,95	3,84
R	2	379,26	5,99
Autres	7	34,21	14,07
Totaux	11	515,98	19,68

rejetée. Par ailleurs, l’examen des sources de variation et des degrés de liberté figurant dans les tableaux 2 et 3 fait bien apparaître les correspondances pour V , F et R .

3° Pour la décomposition de la source de variation « Autres » (7 degrés de liberté), nous devons choisir un des trois chemins pour tester une indépendance partielle; ce choix est fonction des réponses à apporter à l’étude.

Pour la recherche, les trois critères de classification des rameaux ne sont pas équivalents, l’objectif essentiel étant de mieux comprendre le lien éventuel entre l’architecture du pommier et la floraison, passage obligé pour produire des pommes. Par ailleurs, l’expression du critère présence ou absence de fleurs doit être considérée comme aléatoire c’est-à-dire non déterminée *a priori*. Par contre, pour étudier l’influence de l’architecture de l’arbre, il faut bien entendu disposer de variétés et de rameaux à différents niveaux de croissance.

Dès lors, l'indépendance partielle choisie est la suivante :

$$H_0 : F \text{ II } (V, R),$$

c'est-à-dire qu'on analyse la table à deux dimensions construite en fonction de F et de l'association V et R , ou encore qu'on teste la validité du modèle :

$$\log(nP_{ijk}) = [1] + [V]_i + [R]_j + [F]_k + [VR]_{ij}.$$

Tenant compte des propriétés, on obtient aussi l'analyse relative à l'étude de l'indépendance entre V et R , à savoir $H_0 : V \text{ II } R$. Le tableau 4 présente les résultats ; seule la première hypothèse d'indépendance est rejetée.

On peut conclure, d'une part, que la présence ou l'absence de fleurs n'est pas indépendante à la fois des variétés et des niveaux de croissance des rameaux et, d'autre part, que les proportions de rameaux observés, pour chaque niveau de croissance, sont équivalentes quelles que soient les variétés.

Cette seconde conclusion nécessite quelques commentaires. En effet, elle n'aurait aucun sens si le chercheur avait choisi *a priori* les nombres de rameaux à observer pour chaque variété et pour chaque niveau de croissance. Dans ce cas-ci, il est évident que tout pommier planté dispose d'un rameau de niveau 1 et, ainsi, les fréquences totales $V1R1$ (22) et $V2R1$ (19) sont prédéterminées. Dès lors, ce qu'il importe d'étudier, c'est la table partielle $V1R2$ (288), $V1R3$ (163), $V2R2$ (146), $V2R3$ (75), qui conduit à la conclusion analogue au niveau de la charpente des pommiers.

Tableau 4 – Table de contingence : 2^e partie.

Sources de variation	Degrés de liberté	G_{obs}^2	$\chi_{1-\alpha}^2$ $\alpha = 0,05$
$F \text{ II } (V, R)$	5	30,89	11,07
$V \text{ II } R$	2	3,37	5,99
Autres	7	34,21	14,07

Au niveau des analogies, l'examen des tableaux 2 et 4 montre les correspondances suivantes :

$$\begin{aligned} [(V \times F) + (F \times R) + (V \times F \times R)] \text{ et } F \text{ II } (V, R) & : 5 \text{ degrés de liberté;} \\ (V \times R) \text{ et } (V \text{ II } R) & : 2 \text{ degrés de liberté.} \end{aligned}$$

4^e Poursuivons l'analyse en choisissant d'étudier l'indépendance conditionnelle $(F \text{ II } R)|V$ et, par conséquent, l'indépendance marginale résultante $(V \text{ II } F)$.

En effet, pratiquement, il semble assez cohérent d'étudier le lien éventuel entre les critères F et R , conditionnellement à V , au lieu de V et F , conditionnellement à R , l'expérimentateur souhaitant surtout s'intéresser "au comportement" des variétés.

Le modèle étudié s'écrit comme suit :

$$\log(nP_{ijk}) = [1] + [V]_i + [R]_j + [F]_k + [VR]_{ij} + [VF]_{ik}.$$

L'examen du tableau 5 indique qu'on rejette la première hypothèse nulle ($H_0 : (F \amalg R)|V$), mais pas la deuxième ($H_0 : V \amalg F$); celle-ci signifie que globalement la présence ou non de boutons est équivalente pour les deux variétés, si on ne tient pas compte des niveaux de croissance. Précisons que dans le cadre de cette recherche, les arbres étaient en forme libre, sans aucune opération de taille.

Pour cette étape, les correspondances entre les tableaux 2 et 5 s'inscrivent comme suit :

$$\begin{aligned} [(F * R) + (V * F * R)] \text{ et } (F \amalg R)|V & : 4 \text{ degrés de liberté;} \\ (V * F) \text{ et } (V \amalg F) & : 1 \text{ degré de liberté.} \end{aligned}$$

Tableau 5 – Table de contingence : 3^e partie.

Sources de variation	Degrés de liberté	G_{obs}^2	$\chi_{1-\alpha}^2$ $\alpha = 0,05$
$(F \amalg R) V$	4	30,83	9,49
$V \amalg F$	1	0,07	3,84
$F \amalg (V, R)$	5	30,89	11,07

5^o Le rejet de l'hypothèse nulle $H_0 : (F \amalg R)|V$ ouvre alors le chemin à l'étude de l'indépendance entre F et R en distinguant les deux variétés (tableau 6).

Comme le montrent les résultats repris au tableau 6, l'indépendance entre les critères F et R est rejetée pour les deux variétés. Tenant compte de la remarque relative aux rameaux de niveau 1 formulée ci-dessus au 4^o, si on considère uniquement les niveaux de croissance $R2$ et $R3$, l'hypothèse d'indépendance est rejetée pour la variété 1 uniquement ($V1 : G_{obs}^2 = 11,11; V2 : G_{obs}^2 = 2,36; \chi_{1-\alpha}^2 = 5,99$), ce qui signifie que le nombre de rameaux de niveau 2 et 3 fleuris ou non dépend de l'architecture de l'arbre.

6^o Enfin, la valeur G_{obs}^2 associée à l'hypothèse d'absence d'interaction peut être calculée au départ des fréquences attendues estimées (tableau 1 : troisième colonne) par la procédure itérative (9 itérations) décrite au paragraphe 2.3, 3^o.

Tableau 6 – Table de contingence : 4^e partie.

Sources de variation	Degrés de liberté	G_{obs}^2	$\chi_{1-\alpha}^2$ $\alpha = 0,05$
$(F \amalg R) V1$	2	11,31	5,99
$(F \amalg R) V2$	2	19,51	5,99
$(F \amalg R) V$	4	30,83	9,49

Cette fois, le modèle est saturé et il est le suivant :

$$\log(nP_{ijk}) = [1] + [V]_i + [R]_j + [F]_k + [VR]_{ij} + [VF]_{ik} + [RF]_{jk} + [VRF]_{ijk}.$$

La valeur G_{obs}^2 vaut 21,35. L'hypothèse d'absence d'interaction est rejetée ($\chi_{0,95}^2 = 5,99$) ; les deux degrés de liberté correspondent à ceux de l'interaction $V \times F \times R$ du tableau 2.

Au niveau de l'interprétation, cette information n'apporte rien de particulier.

Une remarque importante doit être formulée avant de clôturer cet exemple. Si le chemin suivi pour son analyse se base sur les décompositions progressives du paragraphe 2, en pratique, il convient tout d'abord de définir le chemin, ensuite de réaliser tous les calculs, y compris les tests, et enfin de choisir le chemin inverse pour l'interprétation, c'est-à-dire d'analyser en premier lieu l'interaction entre les trois facteurs et, ensuite si besoin, de l'interpréter en utilisant les éléments précédents, comme c'est la règle en analyse de la variance à plusieurs critères de classification.

3.3. Utilisation des logiciels

Pour l'étude des indépendances, les résultats précédents (valeurs G^2) peuvent être obtenus en utilisant le logiciel SAS, plus particulièrement la procédure CATMOD [SAS Institute, 1994].

En supposant que les informations du tableau 1 sont enregistrés dans une table SAS, intitulée T1, selon quatre variables notées V, R, F pour les indices et Eff pour les fréquences observées, la procédure générale est la suivante :

```
PROC CATMOD DATA=LSAS.T1 ;
WEIGHT EFF ;
MODEL T = _RESPONSE_ ;
REPEATED/ _RESPONSE_ = M ;
RUN ;
```

Le tableau 7 décrit les instructions à encoder pour les différentes tables (T) et les modèles (M) à considérer en fonction des types d'indépendance choisis.

Tableau 7 – Instructions SAS.

Types d'indépendance	T	M
$V \amalg R$	$V * R$	$V \ R \ V * R$
$V \amalg F$	$V * F$	$V \ F \ V * F$
$R \amalg F$	$R * F$	$R \ F \ R * F$
$V \amalg R \amalg F$	$V * R * F$	$V \ R \ F$
$V \amalg (R, F)$	$V * R * F$	$V \ R \ F \ R * F$
$R \amalg (V, F)$	$V * R * F$	$V \ R \ F \ V * F$
$F \amalg (V, R)$	$V * R * F$	$V \ R \ F \ V * R$
$(R \amalg F) V$	$V * R * F$	$V \ R \ F \ V * R \ V * F$
$(V \amalg F) R$	$V * R * F$	$V \ R \ F \ V * R \ R * F$
$(V \amalg R) F$	$V * R * F$	$V \ R \ F \ V * F \ R * F$
Interaction	$V * R * F$	$V \ R \ F \ V * R \ V * F \ R * F$

Par ailleurs, une table à trois dimensions peut toujours se décomposer en tables à deux dimensions pour lesquelles on peut ensuite tester l'indépendance entre les deux critères associés. Pour ces cas, d'autres procédures sont disponibles à la fois dans le logiciel SAS (procédure `FREQ`) et dans le logiciel Minitab (commande `CHISQUARE`). Toutefois, il faut noter que les valeurs critiques sont calculées par la méthode du χ^2 minimum.

Enfin, dans la mesure où le critère F a été privilégié dans cet exemple, l'analyse aurait pu se réaliser par l'intermédiaire de la régression logistique binaire (commande `BLOGISTIC`). En choisissant la fonction de lien logit, celle-ci conduirait à des résultats analogues à ceux obtenus par le modèle log-linéaire : on vérifierait que la proportion des rameaux fleuris dépend de la combinaison variété - niveau de croissance, le terme d'interaction étant significatif.

4. EN GUISE DE CONCLUSION

1^o Comme on a pu le constater, la compréhension et la mise en oeuvre de la théorie relative aux tables de contingence à plus de deux dimensions n'est pas si évidente qu'il n'y paraît. Il en est de même pour l'interprétation des résultats. Une fois de plus, l'accent a été mis sur la nécessité de choisir judicieusement un chemin d'analyse correspondant au problème posé.

Signalons aussi que la surabondance d'informations, résultant de l'usage d'un logiciel spécialisé et dépassant largement le nombre de données analysées, ne simplifie pas cette interprétation, surtout lorsque l'analyse d'une table de contingence revêt un caractère irrégulier.

2^o Au niveau théorique, au moins deux remarques doivent être formulées en relation avec l'application.

D'une part, il n'a pas été tenu compte de la qualité des critères, en particulier pour les types de rameaux, facteur qui pourrait être considéré comme quantitatif et nécessiter une méthode d'analyse spécifique.

D'autre part, les 713 observations ne proviennent pas de 713 rameaux différents et choisis aléatoirement ; les données ne sont donc pas indépendantes.

A priori et de façon plus générale, les conséquences du non respect de cette condition d'application sur la robustesse de la méthode d'analyse ne sont pas encore connues.

BIBLIOGRAPHIE

- DAGNELIE P. [2011]. *Statistique théorique et appliquée. Inférence statistique à une et à deux dimensions*. Bruxelles, De Boeck, 736 p.
- GILLET M. [1993]. *Contribution à la modélisation de la croissance et du développement du pommier*. Gembloux, Faculté des Sciences agronomiques, 91 p.
- FIENBERG S. E. [1970]. The analysis of multidimensional contingency tables. *Ecology*, 51, 419-433.
- ROLIN J. M. [1979]. Modèles log-linéaires. *In. Gerard G., Rolin J.M. Analyse des données discrètes*. Louvain-la-Neuve, Université Catholique, 139-181.
- SAS Institute [1994]. *SAS/STAT-User's guide*, version 6, vol 1, 405-517.
- SIMAR L. [1979]. Tables de contingence à plusieurs critères. *In. Gerard G., Rolin J.M. Analyse des données discrètes*. Louvain-la-Neuve, Université Catholique, 115-137.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de l'Université de Liège – Gembloux Agro-Bio Tech et de l'Unité Systèmes agraires, Territoire et Technologies de l'Information du Centre wallon de Recherches agronomiques (Gembloux - Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

*Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLoux (Belgique)
E-mail : sima.gembloux@ulg.ac.be*

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

<http://www.gembloux.ulg.ac.be/si/>

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

- CHARLES C. [2008]. Introduction à OCTAVE. *Notes Stat. Inform.* (Gembloux) 2008/3, 19 p.
- PALM R., BROSTAUx Y. [2009]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 2009/1, 17 p.
- CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/1, 22 p.
- CHARLES C. [2011]. Introduction aux applications des ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/2, 35 p.
- PALM R., BROSTAUx Y. et CLAUSTRIAUX J. J. [2011]. Macros Minitab pour le choix d'une transformation pour la normalisation de variables. *Notes Stat. Inform.* (Gembloux) 2011/3, 15 p.
- PALM R., BROSTAUx Y. [2011]. La régression logistique avec Minitab. *Notes Stat. Inform.* (Gembloux) 2011/4, 15 p.
- PALM R., BROSTAUx Y., CLAUSTRIAUX J. J. [2011]. Inférence statistique et critères de qualité de l'ajustement en régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2011/5, 32 p.