République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud Mammeri, Tizi-Ouzou

Faculté des Sciences
Département de Mathématiques

Mémoire de fin d'études
Master Académique

Spécialité : Probabilités et Statistiques

Thème

Le modèle DIRICHLET MULTINOMIAL GÉNÉRALISÉ pour données catégorielles

---

Présenté par : M^{lle} ABDERRAHMANI Amira

Président : BOUALAM Karima, (UMMTO) MCB

Examinateur : BEDOUHENE Kahina, (UMMTO) MCB

Encadrant : MEHIRI Mohamed, (UMMTO) MAA

Promotion : 2024/2025

Table des matières

Remerciements

Ce mémoire marque l'aboutissement d'un parcours riche en apprentissages, en efforts et en émotions. Je tiens ici à exprimer toute ma gratitude envers ceux qui m'ont soutenue, guidée et encouragée.

Je tiens tout d'abord à remercier chaleureusement mon encadrant, **M.MEHIRI** , pour son accompagnement rigoureux, ses conseils éclairés et sa disponibilité tout au long de ce travail ; ainsi que **Mme BOUALAM**, pour pour l'honneur qu'elle m'a fait en acceptant de présider le jury .

Je remercie égalment **Mme BEDOUHANE**, d'avoir accepté de faire partie du jury à titre d'examinatrice .

Je remercie également l'ensemble de mes enseignants du **département de mathématiques** pour la qualité de leur enseignement, leur exigence et leur engagement, qui ont largement contribué à ma formation et à mon développement intellectuel.

Liée à **ma chère mère**, pour son amour inépuisable, ses sacrifices silencieux, ses prières constantes et sa foi indéfectible en moi, ainsi qu'à **ma belle-mère**, pour sa tendresse, ses prières et son soutien réconfortant tout au long de ce parcours.

Mes remerciements vont également à **mon mari**, pour sa patience, sa compréhension et son soutien indéfectibles tout au long de ce parcours.

Je remercie de tout cœur **mes sœurs**, pour leur présence affectueuse, leurs encouragements et leur soutien au quotidien.

Enfin, un grand merci à **mes amies**, pour leur aide précieuse, leur bonne humeur, et les moments de partage et de motivation qui ont rendu ce chemin plus agréable.

À toutes et à tous, merci du fond du cœur.

Introduction Générale

L'étude des données catégorielles occupe une place centrale dans de nombreux domaines d'application, notamment en médecine, en sciences sociales, en marketing ou en ingénierie. Ces données, souvent représentées sous forme de fréquences/effectifs ou de proportions, nécessitent des modèles statistiques adaptés permettant de rendre compte de leur structure et de leur variabilité.

Le modèle **multinomial** est un modèle de base largement utilisé pour modéliser ce type de données. Toutefois, dans de nombreux cas, ce modèle s'avère **insuffisant**, notamment lorsqu'il existe une **hétérogénéité** entre les unités ou une **dépendance** temporelle entre les observations. Cette limitation se traduit souvent par une **surdispersion**, que le modèle multinomial classique ne peut expliquer.

C'est dans ce contexte que s'inscrit le modèle Dirichlet-Multinomial Généralisé (**DMG**). Ce modèle étend le modèle multinomial en intégrant une structure de dépendance entre les observations répétées, tout en permettant une variabilité accrue des proportions à travers les unités. Le modèle DMG s'impose ainsi comme une solution naturelle pour l'analyse de données catégorielles longitudinales.

Ce mémoire a pour objectif d'étudier la structure et les propriétés du modèle Dirichlet-Multinomial généralisé, puis de l'appliquer à un exemple de données issues du domaine médical. Il s'organise comme suit :

- Le **premier** chapitre est consacré aux notions de base : (certaines) lois de probabilité usuelles, propriétés des données catégorielles et rappels d'inférence statistique.
- Le **deuxième** chapitre présente, en détail, le modèle Dirichlet-Multinomial Généralisé, ses propriétés théoriques, l'estimation de ses paramètres, ainsi que des tests d'hypothèses associées.
- Le **troisième** chapitre est dédié à une application réelle du modèle DMG à des données médicales -simulées, illustrant son utilité dans l'analyse de données longitudinales présentant de la dépendance et de la surdispersion.

À travers cette étude, nous cherchons à démontrer la **pertinence** et la puissance du modèle DMG dans un cadre pratique, et à mettre en évidence les enjeux liés à l'analyse correcte de données catégorielles complexes.

Chapitre 1 :

Notions de base et Rappels

1.1 Introduction

L'analyse statistique des données catégorielles occupe une place centrale dans de nombreux domaines, tels que les sciences sociales, le marketing, la biologie ou encore l'économie. Ces données, qui représentent des variables qualitatives prenant un nombre limité de modalités, nécessitent des modèles adaptés pour être étudiées correctement.

Parmi les modèles classiques, le modèle *multinomial* constitue une généralisation naturelle du modèle binomial à plus de deux modalités. Il permet de modéliser la probabilité d'observer une répartition d'effectifs parmi plusieurs catégories mutuellement exclusives, lors d'une série d'expériences répétées de manière identique et indépendantes.

Ce chapitre présente les fondements théoriques nécessaires à la compréhension du modèle multinomial. Nous y introduisons d'abord la définition formelle du modèle, avant d'en examiner les principales propriétés (espérance, variance, covariance, estimation des paramètres). Enfin, une section sera consacrée à la nature des données catégorielles, à leurs différentes représentations, et à quelques outils d'analyses utilisées dans la pratique.

Ces rappels sont essentiels pour aborder, dans les chapitres suivants, la généralisation du modèle multinomial via le modèle Dirichlet-Multinomial, ainsi que son application sur des données réelles.

1.2 Données catégorielles

1.2.1 Nature des données catégorielles

En statistique, on distingue généralement deux grandes familles de données : les données quantitatives, qui sont numériques et mesurent des grandeurs (taille, poids, revenu ...), et les données qualitatives, aussi appelées catégorielles, qui décrivent des attributs non numériques.

Les données catégorielles sont donc des observations qui prennent la forme de modalités distinctes, consistant en un nombre fini ou dénombrable de catégories. Ces catégories servent à classifier les individus ou les objets en fonction d'une ou de plusieurs caractéristiques. Contrairement aux données numériques, ces données n'ont pas de valeur arithmétique directe (on ne peut pas les additionner ou calculer une moyenne entre elles sans transformation).

Les données catégorielles sont omniprésentes dans de nombreux domaines :

En marketing : type de produit préféré, marque achetée, canal de communication utilisé.

En santé : présence ou non d'une maladie, groupe sanguin, catégorie de traitement, réponse favorable ou défavorable à un traitement.

En sociologie : Genre, statut matrimonial, origine ethnique.

En sciences politiques : intention de vote, orientation idéologique, obédience politique.

Types de variables catégorielles

Il existe deux grands types de variables catégorielles, selon la nature des relations entre les modalités :

1. Variables **Nominales** :

   Ce sont des variables dont les modalités n'ont pas d'ordre naturel. Les catégories sont mutuellement exclusives mais équivalentes d'un point de vue statistique.

   Exemple : Genre (homme, femme)

2. Variables **Ordinales** :

   Ce type de variable catégorielle présente un ordre naturel ou une hiérarchie entre les modalités. Cependant, l'intervalle entre les catégories n'est pas nécessairement constant, ni interprétable de manière numérique.

   Exemple : Niveau de satisfaction : Faible "<" Moyen "<" Élevé

Les variables ordinales permettent de comparer les modalités (on peut dire qu'une modalité est « supérieure » à une autre), mais on ne peut ni calculer une moyenne, ni faire des opérations arithmétiques directes sans transformation.

1.2.2 Représentation des données catégorielles

Comme les données catégorielles ne sont pas numériques, on utilise des outils spécifiques pour les représenter visuellement et les intégrer dans des modèles statistiques.

a) Tableaux de contingence : Ils permettent de croiser deux (ou plusieurs) variables catégorielles et d'afficher les effectifs pour chaque combinaison possible des modalités -appelés profils réponses. Cela facilite l'analyse des dépendances et des associations.

b) Diagrammes en barres : Utilisés pour représenter les fréquences absolues ou relatives des modalités. Chaque catégorie est représentée par une barre dont la hauteur correspond au nombre d'observations.

c) Diagrammes circulaires : Ils sont parfois utilisés pour montrer la part relative de chaque modalité, mais sont moins lisibles que les diagrammes en barres.

d) Codage pour analyse statistique : Avant d'utiliser les données catégorielles dans un modèle (comme la régression logistique ou multinomiale), il est souvent nécessaire de les transformer :

e) Codage binaire (dummy variables) : chaque modalité devient une variable binaire (0/1).

1.3 Loi de Bernoulli b(p)

Définition 1.3.1 *La distribution de Bernoulli ou loi de Bernoulli est une distribution de probabilité discrète à deux issues (complémentaires).*

On considère une épreuve aléatoire et un événement A lié à cette épreuve tel que P(A) = p. On effectue une fois cette épreuve et on désigne par X la variable aléatoire indiquant la réalisation ou non de A, définie par :

$$X = \begin{cases} 1 & \textit{si } A \textit{ est réalisé} \\ 0 & \textit{si } A \textit{ n'est pas réalisé,} \end{cases}$$

avec :

$$P(X=1)=p \quad , \quad P(X=0)=1-p,$$

ou, de manière équivalente,

$$P(X = x) = p^x(1-p)^{1-x} \ , \ x \in \{0, 1\}.$$

L'espérance et la variance

L'espérance de la variable de Bernoulli est : $E(X) = p$, car par définition :

$$E(X) = \sum_{i=1}^{n} x_i \cdot p_i = \sum_{i=1}^{2} x_i \cdot p_i = 0(1-p) + 1 \cdot p = p.$$

La variance de la variable de Bernoulli est : $V(X) = p(1-p)$, car par définition :

$$V(X) = E(X^2) - E(X)^2 = \sum_{i=1}^{2} x_i^2 \cdot p_i - E(X)^2 = 0(1-p) + 1 \cdot p - p^2 = p(1-p).$$

Exemple

On lance un dé une fois et on considère par exemple comme succès "obtenir un six" et comme échec "ne pas obtenir un six".

1.4 Loi binomiale $\mathcal{B}(n, p)$

Définition 1.4.1 *On répète n fois de maniére indépendante l'épreuve de Bernoulli (expérience aléatoire à deux issues). La probabilité de voir se réaliser A est p. Soit X le nombre de réalisations de l'événement A au cours des n épreuves. La variable X prend les valeurs 0, 1, 2, ..., n. il est clair que X peut être regardée comme la somme de n variables X_i de Bernoulli indépendantes de même paramètre p, soit*

$$X = X_1 + X_2 + ... + X_n.$$

On définit donc

$$P(X = k) = C_n^k p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k \in \{0, \ldots, K\}.$$

Espérance et variance

L'espérance d'une variable binomiale est : E(X)=np,
car : $E(X) = E(X_1) + E(X_2) + ... + E(X_n) = np.$

La variance d'une variable binomiale est : $V(X) = np(1-p)$,
car : $V(X) = V(X_1) + V(X_2) + ... + V(X_n) = np(1-p)$

Exemple

L'effectif d'une section L2 mathémathiques est de cent quatre vingts (180) étudiants, la probabilité qu'un étudiant soit admis en L3 et de $\frac{1}{3}$.

Soit X la variable aléatoire qui compte le nombre d'étudiants admis en L3. Pour déterminer la loi de X, posons

Ω = "Ensemble des 180 étudiants" = $\{e_1, e_2, ..., e_{180}\}$.

À chaque étudiant e_i, on associe une variable aléatoire X_i, $\forall i$, $X_i \to b(p)$ (loi de Bernoulli de paramètre p) : $p = P(x_i = 1) = \frac{1}{3}$.

De même, les variables aléatoires $X_1, ..., X_n$ sont indépendantes.

Posons $X = \sum_{i=1}^{n} X_i$, variable aléatoire qui compte le nombre d'étudiants ayant reussi : $X \to B(180, \frac{1}{3})$ (loi binomiale de paramètres n=180 et $p = \frac{1}{3}$)

$$P(X = k) = C_{180}^{k}(\frac{1}{3})^{k}(\frac{2}{3})^{180-k}, \quad 0 \leq k \leq 180.$$

Le nombre moyen d'étudiants ayant reussi est : $E(X) = np = 180 \times \frac{1}{3} = 60$.

1.5 Loi Multinomiale $\mathcal{M}(n, p_1, p_2, ..., p_k)$

Le modèle multinomial est une généralisation du modèle binomial permettant de modéliser une variable aléatoire discrète prenant plus de deux modalités. Il est utilisé pour représenter la probabilité d'observer un certain nombre d'occurrences dans des catégories mutuellement exclusives, lors de n répétitions indépendantes d'une même expérience aléatoire.

Définition 1.5.1 *Soit* $\mathbf{X} = (X_1, X_2, \ldots, X_K)$ *une variable aléatoire vectorielle représentant les effectifs observés dans chacune des* K *catégories. On dit que* $\mathbf{X}$ *suit une loi multinomiale de paramètres* $n \in N^*$ *et* $\mathbf{p} = (p_1, p_2, \ldots, p_K)$*, avec* $p_k \geq 0$ *pour tout* k *et* $\sum_{k=1}^{K} p_k = 1$*, si sa fonction de masse de probabilité est donnée par :*

$$P(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{x_1! x_2! \ldots x_K!} p_1^{x_1} p_2^{x_2} \ldots p_K^{x_K},$$

sous les contraintes :

$$\sum_{k=1}^{K} x_k = n \quad et \quad \sum_{k=1}^{K} p_k = 1,$$

$$x_k \in \{0, 1, ..., n\} \quad p_k \in [0, 1] \quad pour\ tout\ k \in \{1, \ldots, K\}.$$

Cela signifie que sur les n *essais réalisés,* x_k *observations appartiennent à la catégorie* k*, avec une probabilité* p_k*.*

Espérance et variance

Pour chaque catégorie k, l'espérance et la variance sont données par :

$$E[X_k] = np_k, \quad \mathrm{Var}(X_k) = np_k(1 - p_k).$$

1.5.1 Matrice des variances covariances

Une matrice de variance/covariance est une matrice carrée qui comporte les variances et les covariances associées à plusieurs variables. Les éléments de la diagonale de la matrice sont les variances des variables, tandis que les éléments hors diagonale représentent les covariances entre toutes les paires possibles de variables.

Pour rappel $Cov(X, X) = V(X)$

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Var(X_n) \end{pmatrix}.$$

Dans le cas multinomial :
Pour $i = j$: $\quad cov(X_i, X_j) = V(X_i) = np_i(1 - p_i)$;

si $i \neq j \quad Cov(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j)$.
On a donc : $E(X_iX_j) = n(n-1)p_ip_j$ et $E(X_i)E(X_j) = n^2p_ip_j$.
Alors : $Cov(X_i, X_j) = n(n-1)p_ip_j - n^2p_ip_j$.
Donc : $Cov(X_iX_j) = -np_ip_j$.

$$\Sigma = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_n \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_n \\ \vdots & \vdots & \ddots & \vdots \\ -p_np_1 & -p_np_2 & \cdots & p_n(1-p_n) \end{pmatrix}.$$

Exemple

Supposons que l'on interroge 100 personnes sur leur fruit préféré parmi trois options : *Pomme*, *Banane* et *Orange*. Le nombre de réponses pour chaque fruit suit une distribution binomiale avec $n = 100$ et p_{Pomme}(resp $p_{\text{Banane}}, p_{\text{Orange}}$) ; tandis que le vecteur$(X_p; X_B; X_O) \sim \mathcal{M}(100\,;\, \mathrm{P}_P; P_B; P_O)$. Le modèle permet d'estimer ces proportions ou de tester l'hypothèse d'une répartition uniforme des préférences.

1.6 Loi Hypergéométrique

Considérons une population finie de taille N, dont une proportion p (soit Np individus) possède une certaine caractéristique. On effectue un tirage aléatoire **sans remise** d'un échantillon de taille n dans cette population. Le tirage peut se faire en une seule fois ou progressivement, mais la composition de la population change à chaque tirage.

Soit X la variable aléatoire représentant le nombre d'individus de l'échantillon qui possèdent la propriété étudiée. Alors X suit une **loi hypergéométrique** de paramètres (N, Np, n). Les valeurs possibles de X sont données par :

$$\min(X) = \max(0,\ n - N(1-p)), \quad \max(X) = \min(n,\ Np).$$

La fonction de masse de probabilité de la loi hypergéométrique s'écrit :

$$P(X = x) = \frac{C_{Np}^{x} C_{N(1-p)}^{n-x}}{C_N^n}, \quad \text{pour } x \in [\max(0, n - N(1-p)),\ \min(n, Np)]\,.$$

Le nombre total d'échantillons possibles est C_N^n. Le numérateur de la probabilité représente :

— C_{Np}^{x} : le nombre de façons de choisir x individus parmi ceux qui possèdent la propriété sous intérêt ;

— $C_{N(1-p)}^{n-x}$: le nombre de façons de choisir les $n - x$ individus ne possédant pas la propriété.

Le rapport $\dfrac{n}{N}$ est appelé **taux de sondage**.

On peut interpréter la variable aléatoire X comme la somme de n variables indicatrices :

$$X = X_1 + X_2 + \cdots + X_n,$$

où chaque X_i est une variable de Bernoulli prenant la valeur 1 si le i^{e} individu tiré possède la propriété, et 0 sinon. Ces variables ne sont pas indépendantes (en raison du tirage sans remise), mais nous allons montrer qu'elles ont toutes la même espérance.

Considérons d'abord la première variable :

$$E[X_1] = P(X_1 = 1) = \frac{Np}{N} = p.$$

Étudions maintenant $E[X_2]$. Par la formule des probabilités totales :

$$P(X_2 = 1) = P(X_2 = 1 \mid X_1 = 1) \cdot P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 0) \cdot P(X_1 = 0).$$

On a :

$$P(X_2 = 1 \mid X_1 = 1) = \frac{Np - 1}{N - 1}, \quad \text{et} \quad P(X_2 = 1 \mid X_1 = 0) = \frac{Np}{N - 1}.$$

Ainsi :

$$\begin{aligned} P(X_2 = 1) &= \left(\frac{Np - 1}{N - 1}\right) \cdot p + \left(\frac{Np}{N - 1}\right) \cdot (1 - p) \\ &= \frac{(Np - 1)p + Np(1 - p)}{N - 1} \\ &= \frac{Np(p - 1 + 1)}{N - 1} = \frac{Np}{N - 1} \cdot \left(p - \frac{1}{Np} + 1 - p\right) \\ &= p. \end{aligned}$$

Ce calcul montre que :

$$E[X_2] = p,$$

ce qui est aussi vrai pour toutes les autres variables X_i. Ainsi, bien que les X_i soient dépendantes, elles ont toutes la même espérance p.

Espérance et variance

L'espérance de X est donnée par :

$$E[X] = np,$$

identique à celle d'une loi binomiale.

En revanche, la variance est différente en raison de la dépendance entre les tirages :

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).$$

Chaque X_i suit une loi de Bernoulli de paramètre p, donc :

$$\mathrm{Var}(X_i) = p(1-p), \quad \text{et donc } \sum \mathrm{Var}(X_i) = np(1-p).$$

La covariance entre deux variables X_i et X_j $(i \neq j)$ est :

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \frac{Np-1}{N-1}\, p - p^2 = -\frac{p(1-p)}{N-1}.$$

Il existe $\frac{n(n-1)}{2}$ couples (i, j), donc le second terme devient :

$$2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) = 2\, \frac{n(n-1)}{2} \left(-\frac{p(1-p)}{N-1} \right) = -\frac{n(n-1)}{N-1} p(1-p).$$

Finalement, la variance est :

$$\mathrm{Var}(X) = np(1-p) \left(1 - \frac{n-1}{N-1} \right) = np(1-p) \cdot \frac{N-n}{N-1}.$$

Ce facteur de correction $\frac{N-n}{N-1}$ (appelé « facteur de correction pour population finie ») traduit l'effet de la dépendance entre les tirages, qui réduit la dispersion par rapport au cas du tirage avec remise.

1.7 Théorème Central-Limite (TCL)

Cas univarié La théorème central-limite établit la convergence vers la loi de Gauss sous des hypothèses peu contraignantes.

Théorème 1.7.1 *Soit $(X_n)_n$ une suite de variables aléatoires indépendantes de même loi d'espérance μ et d'écart-type σ. Alors :*

$$\tfrac{1}{\sqrt{n}} \left(\tfrac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma} \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Cas multivarié De même que pour des lois à une dimension on peut établir le résultat suivant : Soit $X_1, X_2, ..., X_n$ une suite de vecteurs aléatoires indépendants de même loi, d'espérance $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$, alors :

Théorème 1.7.2 *On a :*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

La méthode delta Dans le cadre de l'inférence asymptotique, il est souvent nécessaire de connaître la distribution d'une *fonction d'un estimateur*, et non seulement celle de l'estimateur lui-même. La méthode delta constitue une approche classique pour obtenir cette distribution de manière approximative, à partir d'un développement de Taylor au premier ordre.

Définition 1.7.1

*Cas **univarié***

Soit $\hat{\theta}_n$ un estimateur de θ tel que :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

et soit $g : R \to R$ une fonction dérivable en θ. Alors, la méthode delta affirme que :

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \cdot \sigma^2).$$

Autrement dit, l'estimateur $g(\hat{\theta}_n)$ est lui aussi asymptotiquement normal, et sa variance peut être approximée par :

$$\mathrm{Var}(g(\hat{\theta}_n)) \approx \frac{1}{n} \cdot [g'(\theta)]^2 \cdot \sigma^2.$$

*Cas **multivarié***

Dans un cadre multivarié, la méthode delta peut être généralisée comme suit.

Soit $\hat{\boldsymbol{\theta}}_n \in R^K$ un vecteur d'estimateurs asymptotiquement normaux :

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_K(\mathbf{0}, \Sigma),$$

et soit $g : R^K \to R$ une fonction différentiable. Alors :

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\theta})' \Sigma \nabla g(\boldsymbol{\theta})\right),$$

où $\nabla g(\boldsymbol{\theta})$ désigne le gradient de g évalué en $\boldsymbol{\theta}$.

1.8 Loi Gamma

Une variable aléatoire positive X suit une loi Gamma $\Gamma(t, \lambda)$ de paramètres positifs t et λ, si sa densité de probabilité est définie par :

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{t-1}}{\Gamma(t)} & si\ x \geq 0. \\ 0 & sinon \end{cases}$$

Γ étant la fonction eulérienne telle que : $\Gamma(k) = \int_0^\infty x^{k-1}e^{-t}dt$.

1.9 Loi du chi-deux

Soit $Z_1, Z_2, ..., Z_n$ des variables aléatoires indépendantes et identiquement distribuées de loi normale centrée et réduite : $Z_i \to N(0, 1), \forall i$.

Alors, la variable aléatoire $X = \sum_1^n Z_i^2$ suit la loi du chi-deux, χ^2, à n degrés de liberté, sa fonction de densité est :

$$f_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{(\frac{n}{2}-1)}e^{(\frac{-x}{2})} & si\ x > 0. \\ 0 & sinon \end{cases}$$

Espérance et variance

l'espérance et la variance de la variable$X \sim \chi^2$ sont données par les formules suivantes :

$$E(X) = n \text{ et } Var(X) = 2n.$$

1.10 Test du chi-deux

Les tests du Chi-deux (χ^2) sont des tests statistiques non paramétriques utilisés pour analyser des données catégorielles. Il en existe deux variantes principales : le test d'adéquation (ou d'ajustement) et le test d'indépendance.

1.10.1 Test d'adéquation

Ce test permet de vérifier si la répartition observée d'une variable catégorielle est conforme à une loi théorique donnée.

Hypothèses :

— H_0 : les observations suivent la loi multinomiale avec probabilités $p_1, \ldots, p_k$.

— H_1 : la distribution observée ne suit pas la loi multinomiale (inadéquation).

Statistique de test :

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$$

Degrés de liberté :

$$\text{ddl} = K - 1$$

1.10.2 Test d'indépendance

Ce test permet de vérifier l'indépendance statistique entre deux variables qualitatives à partir d'un tableau de contingence.

Hypothèses :

— H_0 : les deux variables sont indépendantes.
— H_1 : les deux variables sont dépendantes.

Statistique de test :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où :

$$E_{ij} = \frac{O_{i\cdot} \cdot O_{\cdot j}}{n}$$

— O_{ij} : fréquence observée dans la cellule (i, j)
— $O_{i\cdot}$: total de la ligne i
— $O_{\cdot j}$: total de la colonne j
— n : taille totale de l'échantillon

Degrés de liberté :

$$\text{ddl} = (I - 1)(J - 1)$$

avec I le nombre de lignes et J le nombre de colonnes du tableau de contingence.

1.11 Formes quadratiques

Sous certaines conditions, des formes quadratiques définies sur des vecteurs gaussiens suivent des lois du χ^2. Ces résultats sont fondamentaux en statistique dans les problèmes de décomposition de variance.

Théorème 1.11.1 *Soit X une variable aléatoire suivant une loi normale en dimension p, de vecteur d'espérance $\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{\Sigma}$ supposée régulière. Alors, la statistique :*

$$D^2 = (X - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu}),$$

suit une loi du χ^2 à p degrés de liberté

Il suffit de se rappeler que si $Y_1, Y_2, \ldots, Y_p$ sont des variables aléatoires indépendantes suivant une loi normale centrée réduite, c'est-à-dire :

$$Y_i \sim \mathcal{N}(0,1), \quad \text{indépendantes pour } i = 1, \ldots, p,$$

alors :

$$D_2 = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2.$$

Considérons maintenant un vecteur Y gaussien centré-réduit, de composantes indépendantes, et intéressons-nous à la forme quadratique générale :

$$Q = Y'AY = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij} Y_i Y_j,$$

où A est une matrice symétrique réelle de taille $p \times p$.

Nous allons établir la forme de la fonction caractéristique de Q, ce qui permettra ensuite de déduire dans quels cas cette forme quadratique Q suit une loi du χ^2.

Théorème 1.11.2 *Soit Y un vecteur aléatoire gaussien centré-réduit de dimension p (i.e. $Y \sim \mathcal{N}_p(0, I_p)$) et soit A une matrice symétrique réelle de taille $p \times p$. On considère la forme quadratique :*

$$Q = Y^\top AY.$$

Alors, la fonction caractéristique de Q est donnée par :

$$\varphi_Q(t) = E[e^{itQ}] = \left(\det(I - 2itA)\right)^{-1/2}.$$

Démonstration.

Commençons par écrire :

$$\varphi_Q(t) = E\left[\exp(itQ)\right] = E\left[\exp(itY^\top AY)\right].$$

Comme A est symétrique réelle, elle est diagonalisable. Il existe donc une matrice orthogonale P telle que :

$$A = P^\top \Lambda P,$$

où $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ est la matrice diagonale des valeurs propres de A.

Posons alors :

$$Z = PY.$$

Comme P est orthogonale (i.e. $P^\top P = I$), le vecteur Z reste un vecteur gaussien centré-réduit de composantes indépendantes : $Z \sim \mathcal{N}_p(0, I_p)$.

Ainsi, on a :

$$Y^\top AY = Y^\top P^\top \Lambda PY = Z^\top \Lambda Z = \sum_{j=1}^{p} \lambda_j Z_j^2.$$

Donc :

$$\varphi_Q(t) = E\left[\exp\left(it\sum_{j=1}^{p}\lambda_j Z_j^2\right)\right] = \prod_{j=1}^{p} E\left[\exp(it\lambda_j Z_j^2)\right].$$

Or, si $Z_j \sim \mathcal{N}(0,1)$, alors :

$$E\left[\exp(it\lambda_j Z_j^2)\right] = \frac{1}{\sqrt{1-2it\lambda_j}},$$

pour tous les t tels que $\Re(1-2it\lambda_j) > 0$.

Ainsi :

$$\varphi_Q(t) = \prod_{j=1}^{p} \frac{1}{\sqrt{1-2it\lambda_j}} = \left(\prod_{j=1}^{p}(1-2it\lambda_j)\right)^{-1/2} = \left(\det(I-2itA)\right)^{-1/2}.$$

Théorème 1.11.3 *Soit Y un vecteur aléatoire suivant une loi normale centrée réduite, et A une matrice symétrique réelle. Alors :*

$$Q = Y^\top AY \sim \chi_k^2 \quad \textit{si et seulement si } A \textit{ est un projecteur orthogonal,}$$

c'est-à-dire si :

$$A^2 = A \quad et \quad A = A^\top.$$

Dans ce cas, le rang de A est égal au nombre de degrés de liberté de la loi du χ^2, soit :

$$rang(A) = k.$$

Justification. Si A est un projecteur orthogonal, alors ses valeurs propres sont égales à 0 ou 1. La fonction caractéristique de Q devient :

$$\varphi_Q(t) = \left(\det(I-2itA)\right)^{-1/2},$$

ce qui correspond à celle d'une loi du χ_k^2. La réciproque est immédiate par identification de la fonction caractéristique.

Théorème 1.11.4 ***Théorème de CRAIG***

Soient $Q_1 = Y^\top A_1 Y$ et $Q_2 = Y^\top A_2 Y$ deux formes quadratiques définies sur le même vecteur $Y \sim \mathcal{N}_p(0, I_p)$, avec A_1 et A_2 deux matrices symétriques. Alors :

$$Q_1 \textit{ et } Q_2 \textit{ sont indépendantes} \quad \Leftrightarrow \quad A_1 A_2 = 0.$$

Démonstration. La fonction caractéristique conjointe de (Q_1, Q_2) est donnée par :

$$\varphi_{Q_1+Q_2}(t_1, t_2) = E\left[\exp\left(it_1Q_1 + it_2Q_2\right)\right] = \left(\det(I - 2it_1A_1 - 2it_2A_2)\right)^{-1/2}.$$

Comparons avec le produit des fonctions caractéristiques individuelles :

$$\varphi_{Q_1}(t_1) \cdot \varphi_{Q_2}(t_2) = \left(\det(I - 2it_1A_1)\right)^{-1/2} \cdot \left(\det(I - 2it_2A_2)\right)^{-1/2}.$$

On a égalité des deux expressions pour tout t_1, t_2 si et seulement si :

$$\det(I - 2it_1A_1 - 2it_2A_2) = \det(I - 2it_1A_1) \cdot \det(I - 2it_2A_2),$$

ce qui est vrai si et seulement si $A_1A_2 = 0$, c'est-à-dire que les deux matrices sont orthogonales pour le produit matriciel.

Nous pouvons enfin énoncer le résultat le plus important concernant les fonnes quadratiques qui généralise la propriété d'additivité du X^2

Théorème 1.11.5 ***Théorème de Cochran*** *Soient $Q_1, Q_2, \ldots, Q_k$ des formes quadratiques définies pour un vecteur aléatoire $Y \sim \mathcal{N}_p(0, I_p)$, telles que :*

$$\sum_{j=1}^{k} Q_j = Y^\top Y,$$

c'est-à-dire que ces formes réalisent une décomposition du carré de la norme de Y.

Alors, les trois conditions suivantes sont équivalentes :

1. *$\sum_{j=1}^{k} \operatorname{rang}(Q_j) = p$;*
2. *chaque Q_j suit une loi du χ^2 avec un certain nombre de degrés de liberté ;*
3. *les variables Q_j sont indépendantes.*

Ce théorème a pour équivalent en algèbre linéaire le résultat suivant :
Soient $A_1, A_2, \ldots, A_k$ des matrices symétriques d'ordre p, telles que :

$$\sum_{j=1}^{k} A_j = I_p.$$

Alors, les trois conditions suivantes sont équivalentes :

1. $\sum_{j=1}^{k} \operatorname{rang}(A_j) = p$;
2. $A_j^2 = A_j$ (i.e., chaque A_j est un projecteur orthogonal) ;
3. $A_iA_j = 0$ pour $i \neq j$ (i.e., les projecteurs sont orthogonaux entre eux).

Géométriquement ce théorème est une extension du théorème de Pythagore et de sa réciproque à la décomposition d'un vecteur et donc du carré de sa norme, sur des sous-espaces deux à deux orthogonaux. L'orthogonalité est ici synonyme d'indépendance pour des vecteurs gaussiens.

1.12 Analyse de la Variance (ANOVA)

L'analyse de la variance (ANOVA) est une méthode statistique utilisée pour comparer les moyennes de plusieurs groupes indépendants afin de déterminer si au moins une moyenne diffère significativement des autres. Elle est particulièrement utile dans les expériences où l'on étudie l'effet d'un facteur qualitatif sur une variable quantitative.

Modèle statistique

Soit un facteur qualitatif A ayant k modalités (groupes), et une variable aléatoire Y mesurée dans chaque groupe. On note :

— Y_{ij} : la j-ème observation dans le groupe i ;
— n_i : le nombre d'observations dans le groupe i, avec $i = 1, \ldots, k$;
— $n = \sum_{i=1}^{k} n_i$: le nombre total d'observations.

Le modèle d'ANOVA à un facteur s'écrit :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

où :

— μ est la moyenne générale ;
— α_i est l'effet du i-ème groupe (déviation de la moyenne générale) ;
— $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ sont des erreurs indépendantes et identiquement distribuées.

Hypothèses

Les hypothèses posées sont :

— $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, c'est-à-dire $\mu_1 = \mu_2 = \cdots = \mu_k$;
— H_1 : il existe au moins un $i \neq j$ tel que $\mu_i \neq \mu_j$.

Les conditions techniques d'application sont :

1. Les observations sont indépendantes ;
2. Les variables Y_{ij} suivent une loi normale ;
3. Les variances sont égales dans les groupes (homoscédasticité).

Décomposition de la variance

La somme des carrés totale (SCT) peut être décomposée en deux composantes :

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}_{\text{SCI}},$$

où :

— $\bar{Y}_i$ est la moyenne du groupe i ;
— $\bar{Y}$ est la moyenne générale.

Statistique du test

La statistique-test est la statistique de Fisher (suit la loi de Fisher sous H_0) :

$$F = \frac{MSC_{\text{entre}}}{MSC_{\text{intra}}} = \frac{SCE/(k-1)}{SCI/(N-k)},$$

où :
— MSC_{entre} est la moyenne des carrés entre groupes ;
— MSC_{intra} est la moyenne des carrés intra-groupes.
Sous l'hypothèse nulle H_0, la statistique suit une loi de Fisher :

$$F \sim \mathcal{F}(k-1, N-k).$$

Décision

On compare la valeur de F obtenue à la valeur critique $F_\alpha(k-1, N-k)$ pour un niveau de signification α donné (souvent $\alpha = 0{,}05$). On rejette H_0 si :

$$F_{\text{calculée}} > F_{\text{critique}},$$

ou si la p-valeur associée est inférieure à α.

Interprétation

Si l'hypothèse nulle est rejetée, cela signifie qu'au moins une des moyennes des groupes est significativement différente des autres. Pour identifier précisément les groupes qui diffèrent, on peut effectuer des tests post-hoc tels que le test de Tukey.

1.13 Statistique de Wald

La statistiques de Wald est un outil d'inférence statistique utilisé pour tester des hypothèses concernant les paramètres d'un modèle statistique, notamment dans les modèles de régression. Elle permet de déterminer si un paramètre estimé est significativement différent d'une valeur hypothétique, souvent zéro. Cette statistique est particulièrement utile dans le contexte de l'estimation par maximum de vraisemblance et est fréquemment utilisée dans la régression logistique et d'autres modèles.

Le test de Wald permet de tester la véracité d'une hypothèse sur la vraie valeur d'un ou plusieurs paramètres d'un modèle, en se basant sur leurs estimateurs et leur variance asymptotique. Soit $\hat{\theta}$ un estimateur du paramètre θ, obtenu par exemple via la méthode du maximum de vraisemblance. Le test repose sur la comparaison entre $\hat{\theta}$ et une valeur hypothétique θ_0, sous l'hypothèse nulle $H_0 : \theta = \theta_0$.

Test sur un seul paramètre

Lorsque l'hypothèse nulle ne porte que sur un seul paramètre, la statistique de test de Wald s'écrit :

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})}.$$

Sous l'hypothèse nulle H_0, cette statistique suit asymptotiquement une loi du χ^2 à **un degré de liberté**. Plus la valeur de W est grande, plus on a de raisons de rejeter H_0.

Test sur plusieurs paramètres

Le test de Wald peut être utilisé pour :
— tester une seule hypothèse portant sur plusieurs paramètres,
— ou tester simultanément plusieurs hypothèses sur un ou plusieurs paramètres.

Soit $\hat{\boldsymbol{\theta}}_n$ un estimateur d'un vecteur de paramètres $\boldsymbol{\theta}$ de dimension $p \times 1$, obtenu à partir d'un échantillon de taille n. On suppose que cet estimateur suit asymptotiquement une loi Normale :

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

où V est la matrice des variances covariances asymptotique de $\hat{\boldsymbol{\theta}}_n$.

On souhaite tester l'hypothèse nulle suivante :

$$H_0 : R\boldsymbol{\theta} = r \qquad \text{contre} \qquad H_1 : R\boldsymbol{\theta} \neq r,$$

où :
— R est une matrice de dimension $q \times p$ (avec q le nombre de restrictions),
— r est un vecteur de dimension $q \times 1$.

La statistique de Wald s'écrit alors :

$$W = (R\hat{\boldsymbol{\theta}}_n - r)' \left[R \left(\frac{\hat{V}_n}{n} \right) R' \right]^{-1} (R\hat{\boldsymbol{\theta}}_n - r),$$

où $\hat{V}_n$ est un estimateur convergent de la matrice V. Sous H_0, la statistique W converge en loi vers une distribution χ^2 à q degrés de liberté :

$$W \xrightarrow{\mathcal{L}} \chi^2_q$$

Démonstration

Supposons que :

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V).$$

Alors, par le théorème de Slutsky, on a :

$$\sqrt{n}(R\hat{\boldsymbol{\theta}}_n - r) = R\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{\lceil}{\to} \mathcal{N}(0, RVR').$$

Remarque 1.13.1 *Le théorème de Slutsky : Soient deux suites de variables aléatoires* (X_n) *et* (Y_n) *définies sur un espace de probabilité commun. Si :*

$$X_n \xrightarrow{d} X \quad et \quad Y_n \xrightarrow{p} c,$$

où X *est une variable aléatoire et* c *est une constante réelle, alors :*

1. $X_n + Y_n \xrightarrow{d} X + c$,
2. $X_n Y_n \xrightarrow{d} cX$,
3. *Si* $c \neq 0$, *alors* $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$.

On sait qu'une forme quadratique d'une variable gaussienne suit une loi du χ^2. Donc :

$$\left[\sqrt{n}(R\hat{\boldsymbol{\theta}}_n - r)\right]' (RVR')^{-1} \left[\sqrt{n}(R\hat{\boldsymbol{\theta}}_n - r)\right] \xrightarrow{\mathcal{D}} \chi^2_Q.$$

En simplifiant, on retrouve la statistique de Wald :

$$W = (R\hat{\boldsymbol{\theta}}_n - r)' \left[R\left(\frac{V}{n}\right)R'\right]^{-1} (R\hat{\boldsymbol{\theta}}_n - r) \xrightarrow{\mathcal{D}} \chi^2_q.$$

1.14 Conclusion

Dans ce chapitre, nous avons rappelé les notions fondamentales relatives aux données catégorielles et au modèle multinomial, qui constitue une base incontournable pour modéliser la distribution d'effectifs parmi plusieurs catégories. Après avoir défini formellement cette loi, nous avons étudié ses propriétés statistiques essentielles, telles que l'espérance, la variance et la covariance, ainsi que les méthodes usuelles d'estimation des paramètres.

Nous avons également introduit les différentes formes de représentation des données catégorielles, ainsi que quelques outils classiques d'analyse. Ces éléments théoriques forment un socle nécessaire pour appréhender les modèles plus complexes développés dans la suite de ce mémoire.

Le chapitre suivant s'inscrit dans cette continuité en introduisant le modèle Dirichlet-Multinomial généralisé, qui permet de prendre en compte la variabilité supplémentaire souvent observée dans les données réelles, en modélisant la dépendance entre observations de manière plus souple que le modèle multinomial classique.

Chapitre 2 :

Le modèle Dirichlet-Multinomial Généralisé

2.1 Introduction

L'analyse de données catégorielles en présence d'une dépendance intra-groupe ou d'une surdispersion nécessite le recours à des modèles statistiques plus flexibles que le modèle multinomial classique. Ce chapitre est consacré à l'étude du modèle Dirichlet-Multinomial généralisé, un cadre probabiliste permettant de mieux modéliser la variabilité observée dans des contextes où les hypothèses d'indépendance et d'homogénéité ne sont pas toujours satisfaites.

Nous commencons par rappeler les fondements du modèle multinomial généralisé, en abordant l'espérance, la variance, et les méthodes classiques d'estimation des paramètres. Ensuite, nous introduisons le modèle Dirichlet-Multinomial, qui constitue une première généralisation utile dans les cas de **surdispersion**. Nous présentons ses propriétés mathématiques, notamment les expressions de l'espérance et de la matrice de variance-covariance.

Enfin, nous étudierons en détail le modèle Dirichlet-Multinomial Généralisé, qui permet d'intégrer une structure de corrélation intra-groupe explicite. Nous développerons les expressions théoriques de l'espérance et de la variance de ce modèle, les méthodes d'estimation des paramètres (en particulier la corrélation intra-groupe ρ et le facteur de variation supplémentaire C), ainsi que les tests statistiques permettant de vérifier les hypothèses structurelles du modèle. Ces outils sont essentiels pour valider l'adéquation du modèle aux données et garantir la fiabilité des inférences statistiques.

2.2 Modèle Multinomial Généralisé

Considérons un système composé de S unités observées simultanément à n instants distincts. À chaque instant t, chaque unité s est classée dans l'un des I états possibles, ces états étant mutuellement exclusifs.

On note par X_{ist} la variable aléatoire qui prend la valeur 1 si, au temps t, la s-ième unité est observée dans l'état i, et 0 sinon. La probabilité que $X_{ist} = 1$ est supposée constante, égale à π_i, indépendamment de l'unité s et du temps t.

Par ailleurs, les observations effectuées à des instants différents sont supposées indépendantes. Pour chaque unité s, on définit le vecteur

$$\mathbf{X}_{st} = (X_{1st}, X_{2st}, \ldots, X_{Ist})',$$

représentant la répartition dans les différentes catégories à l'instant t.
Ce vecteur suit une loi multinomiale de paramètre $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_I)'$ et de taille 1, $X_{st} \sim \mathcal{M}(1, \boldsymbol{\pi})$ (c'est-à-dire qu'à chaque instant, une seule catégorie est observée par unité).

En cumulant les observations sur les n instants, le total des observations dans la catégorie i pour l'unité s, défini par $X_{is} = \sum_{t=1}^{n} X_{ist}$, suit une loi binomiale de paramètres n et π_i, car il s'agit de la somme de n essais de Bernoulli indépendants.

Tallis (1962) a proposé une généralisation du modèle multinomial classique, appelée *modèle multinomial généralisé*, qui prend en compte **la dépendance** entre les unités statistiques. Cette dépendance est modélisée à l'aide d'un paramètre de corrélation commun ρ, introduit pour relier les différentes unités entre elles.

Pour formaliser ces dépendances, Tallis a spécifié la *fonction génératrice des moments conjointe* des variables X_{is}, c'est-à-dire des totaux par catégorie et par unité, ce qui permet de caractériser leur comportement conjoint sous l'effet de la corrélation ρ.

La Fonction Génératrice des Moments

Soit le vecteur aléatoire $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iS})'$ représentant des variables catégorielles prenant des valeurs dans l'ensemble $\{0, 1, \ldots, n\}$. On suppose que $\boldsymbol{X}_i$ suit un modèle de mélange à deux composantes :

— avec probabilité $\rho \in [0, 1]$, toutes les composantes de $\boldsymbol{X}_i$ dépendantes (sont égales à une même valeur k) choisie selon une loi discrète de probabilités $(p_{i0}, \ldots, p_{in})$;

— avec probabilité $1 - \rho$, les variables $X_{i1}, \ldots, X_{iS}$ sont indépendantes et identiquement distribuées selon la même loi discrète $(p_{i0}, \ldots, p_{in})$.

La fonction génératrice des moments (FGM) de $\boldsymbol{X}_i$ est définie par :

$$G_i(\boldsymbol{u}) = E\left[e^{\boldsymbol{u}'\boldsymbol{X}_i}\right] = E\left[\exp\left(\sum_{s=1}^{S} u_s X_{is}\right)\right],$$

où $\boldsymbol{u} = (u_1, \ldots, u_S)'$.

Première composante : dépendance totale Si $\boldsymbol{X}_i = (k, \ldots, k)$, alors :

$$\sum_{s=1}^{S} u_s X_{is} = k \sum_{s=1}^{S} u_s, \quad \text{et donc} \quad e^{\sum_{s=1}^{J} u_s X_{is}} = \left(\prod_{s=1}^{S} e^{u_s}\right)^k.$$

La contribution de cette composante à la FGM est :

$$\rho \sum_{k=0}^{n} p_{ik} \left(\prod_{s=1}^{S} e^{u_s}\right)^k.$$

Deuxième composante : indépendance des catégories Sous cette hypothèse, les X_{is} sont indépendantes avec :

$$P(X_{is} = k) = p_{ik}, \quad \text{pour } s = 1, \ldots, S.$$

On a donc :

$$E\left[e^{\sum_{s=1}^{S} u_s X_{is}}\right] = \prod_{s=1}^{S} E[e^{u_s X_{is}}] = \prod_{s=1}^{S} \left(\sum_{k=0}^{n} p_{ik} e^{k u_s}\right).$$

On note cette expression :

$$p(e^{u_s}) = \sum_{k=0}^{n} p_{ik} e^{k u_s},$$

et donc la contribution est :

$$(1-\rho)\prod_{s=1}^{S} p(e^{u_s}).$$

En combinant les deux cas, on obtient l'expression finale de la fonction génératrice des moments :

$$G_i(\boldsymbol{u}) = \rho \sum_{k=0}^{n} p_{ik} \left(\prod_{s=1}^{S} e^{u_s} \right)^k + (1-\rho) \prod_{s=1}^{S} p(e^{u_s}), \tag{2.2.1}$$

avec :

$$p(e^{u_s}) = \sum_{k=0}^{n} p_{ik} e^{k u_s}.$$

2.2.1 Espérance et Variance dans le Modèle Multinomial Généralisé

Soit $\boldsymbol{X}_s \in \{0,1\}^K$ le vecteur indicateur de la catégorie observée pour la s-ième variable, avec $s = 1, \ldots, S$. On note $\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s$ le vecteur global des totaux par catégorie. On suppose que chaque $\boldsymbol{X}_s \sim \mathcal{M}(1, \pi)$ de parmètre $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, avec une corrélation intra-variables modélisée par un paramètre $\rho \in]0,1[$.

Espérance Soit $\boldsymbol{X}_s \in \{0,1\}^K$ le vecteur indicateur de la catégorie observée pour la s-ième variable, avec $s = 1, \ldots, S$. Autrement dit, $\boldsymbol{X}_s = (X_{s1}, \ldots, X_{sK})'$ est un vecteur aléatoire tel que :

$$X_{sk} = \begin{cases} 1 & \text{si l'observation } s \text{ appartient à la catégorie } k, \\ 0 & \text{sinon,} \end{cases} \qquad \text{avec } \sum_{k=1}^{K} X_{sk} = 1.$$

Ce vecteur encode une unique observation parmi K catégories possibles, selon une loi multinomiale de taille 1, notée $\mathcal{M}(1, \boldsymbol{\pi})$, où $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ est un vecteur de probabilités tel que $\pi_k \geq 0$ pour tout k et $\sum_{k=1}^{K} \pi_k = 1$. On suppose en outre que les variables présentent une dépendance intra-groupe modélisée par un paramètre de corrélation $\rho \in]0,1[$.

Le vecteur total des observations est défini par la somme des vecteurs indicateurs individuels :

$$\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s.$$

Espérance. Par linéarité de l'espérance, on a :

$$E(\boldsymbol{X}) = \sum_{s=1}^{S} E(\boldsymbol{X}_s).$$

Calculons d'abord l'espérance de $\boldsymbol{X}_s$. Pour toute composante $k = 1, \ldots, K$, la variable X_{sk} est une variable indicatrice, donc :

$$E(X_{sk}) = 1 \cdot P(X_{sk} = 1) + 0 \cdot P(X_{sk} = 0) = P(X_{sk} = 1) = \pi_k.$$

Par conséquent, l'espérance du vecteur $\boldsymbol{X}_s$ s'écrit :

$$E(\boldsymbol{X}_s) = (\pi_1, \ldots, \pi_K)' = \boldsymbol{\pi}.$$

En remplaçant dans la somme, on obtient l'espérance du vecteur global :

$$E(\boldsymbol{X}) = \sum_{s=1}^{S} \boldsymbol{\pi} = S\boldsymbol{\pi}.$$

Cas multinomial généralisé : si chaque vecteur $\boldsymbol{X}_s$ représente une somme de n observations indépendantes suivant la même loi multinomile $\boldsymbol{X}_j \sim \mathcal{M}(n, \boldsymbol{\pi})$, alors :

$$E(\boldsymbol{X}_s) = n\boldsymbol{\pi}, \quad \text{et ainsi :} \quad E(\boldsymbol{X}) = \sum_{s=1}^{S} n\boldsymbol{\pi} = nS\boldsymbol{\pi}.$$

Variance On considère le vecteur total $\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s$, où chaque $\boldsymbol{X}_s \in \{0, 1\}^K$ est un vecteur indicateur décrivant l'appartenance de l'observation s à l'une des K catégories. Chaque vecteur suit une loi multinomiale de taille 1 avec paramètre $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$, tel que $\sum_{k=1}^{K} \pi_k = 1$ et $\pi_k \geq 0$.

Nous souhaitons calculer la matrice des variances-covariances de $\boldsymbol{X}$, notée $\mathrm{Cov}(\boldsymbol{X})$.

Formule de la variance d'une somme vectorielle

On applique la formule générale de la variance d'une somme de vecteurs aléatoires :

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}\left(\sum_{s=1}^{S} \boldsymbol{X}_s\right) = \sum_{s=1}^{S} \mathrm{Var}(\boldsymbol{X}_s) + \sum_{\substack{s,k=1 \\ s \neq k}}^{S} \mathrm{Cov}(\boldsymbol{X}_s, \boldsymbol{X}_k).$$

Cette formule se décompose en deux termes :
— La somme des variances individuelles,
— La somme des covariances croisées entre vecteurs différents.

Variance d'un vecteur indicateur $\boldsymbol{X}_s$

Le vecteur $\boldsymbol{X}_s$ est un vecteur indicateur de dimension K, où une seule composante vaut 1 et les autres 0. La probabilité que la composante k prenne la valeur 1 est π_k. Ainsi, la variable aléatoire X_{sk} suit une loi de Bernoulli de paramètre π_k, et on a :

$$E(X_{sk}) = \pi_k, \quad \mathrm{Var}(X_{sk}) = \pi_k(1 - \pi_k).$$

De plus, pour deux composantes distinctes $k \neq l$, on a :

$$\mathrm{Cov}(X_{sk}, X_{sl}) = E(X_{sk}X_{sl}) - \pi_k\pi_l.$$

Or, comme une seule catégorie est observée à chaque tirage, on a nécessairement $X_{sk}X_{sl} = 0$, donc :

$$E(X_{sk}X_{sl}) = 0 \quad \Rightarrow \quad \mathrm{Cov}(X_{sk}, X_{sl}) = -\pi_k\pi_l.$$

Ainsi, la matrice des variances-covariances du vecteur $\boldsymbol{X}_s$ est :

$$\mathrm{Var}(\boldsymbol{X}_s) = \boldsymbol{M}_{\boldsymbol{\pi}} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top.$$

Matriciellement, cela donne :

$$[\boldsymbol{M}_{\boldsymbol{\pi}}]_{kl} = \begin{cases} \pi_k(1-\pi_k) & \text{si } k = l \\ -\pi_k\pi_l & \text{si } k \neq l \end{cases}$$

C'est une matrice symétrique, semi-définie positive, de rang $K-1$, car la somme des composantes de chaque $\boldsymbol{X}_j$ vaut 1.

Covariances croisées(entre vecteurs $\boldsymbol{X}_s$ et $\boldsymbol{X}_k$, $s \neq k$)

Dans le cas classique (multinomial simple), les vecteurs $\boldsymbol{X}_s$ sont supposés indépendants, donc les covariances croisées sont nulles.

Cependant, dans le modèle multinomial généralisé, on introduit une dépendance entre les vecteurs $\boldsymbol{X}_s$ à l'aide du paramètre $\rho \in]0,1[$, qui modélise la corrélation intra-groupe. On suppose que :

$$\mathrm{Cov}(\boldsymbol{X}_s, \boldsymbol{X}_k) = \rho\boldsymbol{M}_{\boldsymbol{\pi}}, \quad \text{pour } s \neq k.$$

Il y a $S(S-1)$ paires distinctes (s,k) avec $s \neq k$, donc la somme des covariances croisées donne :

$$\sum_{\substack{s,k=1 \\ s\neq k}}^{S} \mathrm{Cov}(\boldsymbol{X}_s, \boldsymbol{X}_k) = \rho S(S-1)\boldsymbol{M}_{\boldsymbol{\pi}},$$

En combinant les deux parties, on obtient :

$$\mathrm{Var}(\boldsymbol{X}) = S\boldsymbol{M}_{\boldsymbol{\pi}} + \rho S(S-1)\boldsymbol{M}_{\boldsymbol{\pi}} = S\left[1 + (S-1)\rho\right]\boldsymbol{M}_{\boldsymbol{\pi}}.$$

C'est-à-dire :

$$\boxed{\mathrm{Var}(\boldsymbol{X}) = S\left[1 + (S-1)\rho\right]\left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right)}$$

Cette expression montre que la dépendance entre les observations, introduite par le paramètre ρ, amplifie la variance totale proportionnellement à $1 + (S-1)\rho$. Lorsque $\rho = 0$, on retrouve le cas d'indépendance : $\text{Var}(\boldsymbol{X}) = S\boldsymbol{M}_{\boldsymbol{\pi}}$. Alors :

$$\boxed{E(\mathbf{X}) = nS\boldsymbol{\pi}, \text{Var}(\boldsymbol{X}) = S\left[1 + (S-1)\rho\right](\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')}$$

Estimateur non biaisé Puisque $E(\boldsymbol{X}) = Sn\boldsymbol{\pi}$, un estimateur non biaisé de $\boldsymbol{\pi}$ est donné par :

$$\widehat{\boldsymbol{\pi}} = \frac{1}{Sn}\boldsymbol{X}.$$

2.2.2 Méthode d'estimation des paramètres

Tallis (1962) a proposé des estimateurs pour le paramètre de corrélation intra-groupe ρ, mais il n'a pas abordé de méthode d'inférence concernant le vecteur de probabilité $\boldsymbol{\pi}$. Nous présentons ici une méthode asymptotique, permettant de construire des tests statistiques pour des fonctions de $\boldsymbol{\pi}$.

Considérons un plan d'expérience dans lequel les observations sont organisées en S groupes, chacun contenant n individus. À chaque instant t, on observe un vecteur $\boldsymbol{X}_{st} = (X_{\mathbf{1}st}, X_{\mathbf{2}st}, \ldots, X_{\mathbf{I}st})'$ représentant les comptages dans I catégories pour l'individu t du groupe s. Chaque vecteur suit une loi multinomiale de paramètre $\boldsymbol{\pi}$, commun à tous les groupes. Les dépendances intra-groupes sont modélisées par un coefficient de corrélation $\rho \in]0, 1[$.

Pour effectuer l'inférence, on définit d'abord un vecteur de dimension IS, noté $\boldsymbol{X}_{t(S)} = \mathbf{1}_S \otimes \boldsymbol{X}_{st}$, où ($\mathbf{1}_S$ est un vecteur de'1' de dimention S : $\mathbf{1}_S = (1, 1....., 1)'_S$, et $\otimes$ désigne le produit matriciel direct).

Remarque 2.2.1 *Le produit matriciel direct , aussi appelé le produit de kronecker, est une opération entre deux matrices (ou vecteurs) qui produit une matrice de plus grandes dimensions.*

Soient deux matrices :

$$A \in R^{m\times n}, \quad B \in R^{p\times q}$$

Le produit de Kronecker de A et B, noté $A \otimes B$, est une matrice de dimension $(mp) \times (nq)$ obtenue en multipliant chaque élément a_{ij} de A par la matrice B :

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}$$

Propriétés

1. *Le produit matriciel direct n'est pas commutatif $A \otimes B \neq B \otimes A$*

2. Le produit matriciel direct est associatif $(A \otimes B) \otimes C = A \otimes (B \otimes C)$

Le vecteur agrégé s'écrit alors :

$$\boldsymbol{X}_{(S)} = \sum_{t=1}^{n} \boldsymbol{X}_{t(S)}.$$

Sous l'hypothèse que les vecteurs $\boldsymbol{X}_{t(S)}$ sont indépendants et admettent des moments finis, le théorème central limite(multivarié) s'applique, et on a :

$$\sqrt{n}\left(\boldsymbol{X}_{(S)} - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

avec $\boldsymbol{\mu} = \mathbf{1}_S \otimes \boldsymbol{\pi}$ et $\boldsymbol{\Sigma} = \mathbf{M}_\pi \otimes \mathbf{R}$.

La matrice $\mathbf{M}_\pi$ est la matrice des variances-covariances de la loi multinomiale :

$$\mathbf{M}_\pi = \operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}',$$

et $\mathbf{R}$ est la matrice de corrélation intra-groupe de taille $S \times S$, avec 1 sur la diagonale et ρ ailleurs :

$$\mathbf{R} = (1-\rho)\mathbf{I}_S + \rho \mathbf{1}_S \mathbf{1}_S'.$$

On obtient ensuite la statistique globale en sommant les composantes sur les groupes via une transformation linéaire :

$$\boldsymbol{X} = B\boldsymbol{X}_{(S)}, \quad \text{avec } B = \mathbf{1}_S' \otimes \mathbf{I}_I,$$

où $\mathbf{I}_I$ est la matrice identité de taille I. Par stabilité de la loi normale multivariée sous transformation linéaire, on a :

$$\sqrt{n}\left(\boldsymbol{X} - nS\boldsymbol{\pi}\right) \xrightarrow{d} \mathcal{N}_I\left(\mathbf{0},\ S(1+(S-1)\rho)\,\mathbf{M}_\pi\right). \tag{2.2.2}$$

Cela permet de construire une statistique de Wald (§1.16) pour tester une hypothèse nulle de la forme $H_0 : f(\boldsymbol{\pi}) = \mathbf{0}$, où f est une fonction vectorielle différentiable. La statistique s'écrit :

$$\chi^2_W = \frac{nS}{1+(S-1)\rho}\left[f(\hat{\boldsymbol{\pi}})\right]'\left[S\mathbf{M}_\pi S'\right]^{-}\left[f(\hat{\boldsymbol{\pi}})\right], \tag{2.2.3}$$

où $\hat{\boldsymbol{\pi}} = \boldsymbol{X}/(nS)$ est l'estimateur empirique de $\boldsymbol{\pi}$, et J est la matrice jacobienne de f évaluée en $\boldsymbol{\pi}$:

$$J = \left.\frac{\partial f}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}}.$$

Rappel : Une inverse généralisée d'un matrice A est toute matrice, notée A^- telle que : $A^-AA^- = A^-$et $AA^-A = A$

L'inverse généralisée $[J\mathbf{M}_\pi J']^-$ est utilisée dans le cas où la matrice n'est pas inversible. La statistique χ^2_W converge asymptotiquement vers une loi du χ^2 avec un nombre de degrés de liberté égal au rang de la matrice $J\mathbf{M}_\pi J'$.

Cette approche permet donc de tester des contrastes, linéaires ou non, sur le vecteur de probabilités $\boldsymbol{\pi}$ tout en tenant compte de la structure de dépendance introduite par ρ.

2.3 Modèle Dirichlet-Multinomial

Le modèle Dirichlet-Multinomial est une généralisation du modèle multinomial classique, permettant de prendre en compte la variabilité des proportions à travers les unités et dans le temps. Ce modèle est flexible pour des données de cette nature. Son intérêt est qu'il permet de prendre en compte l'éventuelle **surdispersion** dans les données , ce qui n'est pas capté par un modèle multinomial classique.

Le modèle Dirichlet-Multinomial est un modèle hiérarchique à deux niveaux, défini comme suit :

- Au premier niveau, on tire un vecteur aléatoire avec des probabilités $\boldsymbol{\pi}$ qui suit le loi de Dirichlet

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{D}(\boldsymbol{\alpha}) \quad \text{où } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K),\ \alpha_k > 0.$$

Remarque 2.3.1 *La loi de Dirichlet est une loi de probabilité définie sur le simplexe Δ de dimension K :*

$$\Delta_K = \left\{ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \in R^K \;\middle|\; \pi_k \geq 0,\ \sum_{k=1}^{K} \pi_k = 1 \right\}.$$

On dit que $\boldsymbol{\pi} \sim \mathrm{D}(\alpha_1, \ldots, \alpha_K)$ si sa densité de probabilité est donnée par :

$$f(\pi_1, \ldots, \pi_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \quad \textit{pour } \boldsymbol{\pi} \in \Delta_{K-1},$$

où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ avec $\alpha_k > 0$, et

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)},$$

est la fonction bêta multivariée.

- Au second niveau, conditionnellement à $\boldsymbol{\pi}$, on tire un vecteur $\boldsymbol{X}$ suivant une loi multinomiale :

$$\boldsymbol{X} = (X_1, \ldots, X_K) \mid \boldsymbol{\pi} \sim \mathcal{M}(n, \boldsymbol{\pi}) \quad \text{avec } \sum_{k=1}^{K} X_k = n.$$

L'objectif est de déterminer la fonction de masse de la loi marginale de $\boldsymbol{X}$, en "intégrant" la loi conditionnelle par rapport à la densité de $\boldsymbol{\pi}$.

Loi conditionnelle La loi conditionnelle de $\boldsymbol{X}$ sachant $\boldsymbol{\pi}$ est donnée par la loi multinomiale :

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) = \frac{n!}{x_1! \cdots x_K!} \prod_{k=1}^{K} \pi_k^{x_k} \quad \text{où } \sum_{k=1}^{K} x_k = n.$$

On effectue n essais indépendants. Lors de chaque essai i, on observe une catégorie parmi K, représentée par un vecteur indicateur $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_K^{(i)})$ tel que :

$$X_k^{(i)} = \begin{cases} 1 & \text{si la catégorie } k \text{ est observée au } i\text{-ième essai,} \\ 0 & \text{sinon,} \end{cases} \quad avec \sum_{k=1}^{K} X_k^{(i)} = 1$$

Soit $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ le vecteur de probabilités des catégories. On suppose que chaque $\boldsymbol{X}^{(i)}$ suit une loi discrète sur les vecteurs sur la base canonique, avec :

$$P(\boldsymbol{X}^{(i)} = \boldsymbol{e}_k) = \pi_k, \quad \text{pour } k = 1, \ldots, K.$$

On définit le vecteur total des comptages $\boldsymbol{X} = (X_1, \ldots, X_K)$ comme la somme des vecteurs indicateurs :

$$\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{X}^{(i)}, \quad \text{on a } X_k = \sum_{i=1}^{n} X_k^{(i)}.$$

Pour une séquence de résultats compatible avec $\boldsymbol{x} = (x_1, \ldots, x_K)$, la probabilité est :

$$P(X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_K = x_K) = \prod_{k=1}^{K} \pi_k^{x_k},$$

car il y a x_k tirages où la catégorie k a été observée.

Le nombre de façons de répartir n essais en x_1 catégories 1, x_2 catégories 2, etc., est donné par :

$$\frac{n!}{x_1! \cdots x_K!}.$$

La probabilité totale d'obtenir un vecteur de comptage $\boldsymbol{x}$ est donc :

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) = \frac{n!}{x_1! \cdots x_K!} \prod_{k=1}^{K} \pi_k^{x_k}, \quad \text{avec } \sum_{k=1}^{K} x_k = n.$$

La loi marginale de $\boldsymbol{X}$ s'obtient par intégration de la loi conjointe :

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int_{\Delta_K} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) \cdot f(\pi_1, \cdots, \pi_k)\, d\boldsymbol{\pi}$$

$$= \int_{\Delta_K} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) = \frac{n!}{x_1! \cdots x_K!} \prod_{k=1}^{K} \pi_k^{x_k} \cdot \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} d\boldsymbol{\pi}$$

où $\Delta_K = (\pi_1, ..., \pi_k) / \sum_{k=1}^{K} \pi_k = 1$

$$= \frac{n!}{x_1! \cdots x_K!} \cdot \frac{1}{B(\boldsymbol{\alpha})} \int_{\Delta_K} \prod_{k=1}^{K} \pi_k^{x_k + \alpha_k - 1} \, d\boldsymbol{\pi}.$$

L'intégrale ci-dessus est reconnue comme la fonction bêta multivariée $B(\boldsymbol{x} + \boldsymbol{\alpha})$. Ainsi, on obtient :

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!}{x_1! \cdots x_K!} \cdot \frac{B(\boldsymbol{x} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}.$$

En remplaçant les fonctions bêta multivariées par leurs expressions avec la fonction gamma, on a :

$$B(\boldsymbol{x} + \boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(x_k + \alpha_k)}{\Gamma(n + \alpha_0)}, \quad B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}.$$

D'où la fonction de masse de la loi de Dirichlet Multinomiale :

$$P(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{\prod_{k=1}^{K} x_k!} \cdot \frac{\prod_{k=1}^{K} \Gamma(x_k + \alpha_k)}{\Gamma(n + \alpha_0)} \cdot \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \quad \text{où } \alpha_0 = \sum_{k=1}^{K} \alpha_k. \tag{2.3.1}$$

Remarque 2.3.2 *La présence de la fonction gamma s'explique par la généralisation des factorielles aux réels positifs. Elle intervient naturellement dans la définition des lois de Dirichlet et Bêta multivarié, ainsi que dans les intégrales sur le simplexe* (Δ_k).

2.3.1 Cas particulier : convergence vers la loi multinomiale.

Lemme :

$$\mathbf{Lim}_{\alpha_0 \to \infty} P(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} (\pi_k)^{x_k},$$

qui est la distribution multinomiale.

Preuve En se rappelant (propriété de la fonction Gamma) que :

$$\Gamma(x + k) = \Gamma(x) \prod_{l=1}^{k} (x + l - 1),$$

pour x réel et k entier, on peut écrire le noyau h de (2.3.1) comme le rapport de R_1 à R_2 avec :

$$R_1 = \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)} \qquad et \qquad R_2 = \frac{\Gamma(x + \alpha_0)}{\Gamma(\alpha_0)}.$$

On aura, puisque $\sum x_k = n$:

$$R_1 = (\prod_{k=1}^{K} \prod_{l=1}^{x_k} ((l-1)\frac{1}{\alpha_0} + \frac{\alpha_k}{\alpha_0}))(\alpha_0^n), \quad R_2 = (\prod_{l=1}^{x_k} ((l-1)\frac{1}{\alpha_0} + 1)(\alpha_0^n);$$

Ce qui donne :

$$h = \frac{\prod_{k=1}^{K} \prod_{l=1}^{x_k} ((l-1)\frac{1}{\alpha_0} + \frac{\alpha_k}{\alpha_0})}{\prod_{l=1}^{x_k} ((l-1)\frac{1}{\alpha_0} + 1)}$$

$$= \prod_{k=1}^{K} (\prod_{l=1}^{x_k} \frac{((l-1)\frac{1}{\alpha_0} + \frac{\alpha_k}{\alpha_0})}{((l-1)\frac{1}{\alpha_0} + 1)}).$$

Et alors :

$$\mathbf{Lim}_{\alpha_0 \to \infty} h = \prod_{k=1}^{K} \prod_{l=1}^{x_k} \pi_k = \prod_{k=1}^{K} (\pi_k)^{x_k},$$

qui est le noyau de la distribution multinomiale.

2.3.2 Espérance et Variance

Considérons le vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_K)$ suivant une loi Dirichlet Multinomiale avec n essais et un paramètre de concentration $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. On note $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ la somme des paramètres, et $\pi_k = \frac{\alpha_k}{\alpha_0}$ la probabilité "moyenne" associée à la k-ième catégorie. Le modèle est hiérarchique et défini comme suit :

— $\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K)$,
— $\boldsymbol{X} \mid \boldsymbol{\theta} \sim \mathcal{M}(n, \boldsymbol{\theta})$.

1. Espérance On utilise la formule de l'espérance totale :

$$E[X_k] = E_{\boldsymbol{\theta}}\left[E[X_k \mid \boldsymbol{\theta}]\right] = E_{\boldsymbol{\theta}}[n\theta_k] = n \cdot E[\theta_k] = n \cdot \frac{\alpha_k}{\alpha_0} = n\pi_k.$$

Donc, le vecteur d'espérance est :

$$\boxed{E[\boldsymbol{X}] = n\boldsymbol{\pi}, \quad \text{où } \boldsymbol{\pi} = \left(\frac{\alpha_1}{\alpha_0}, \ldots, \frac{\alpha_K}{\alpha_0}\right)'.}$$

2. Variance On applique la formule de la variance totale :

$$\mathrm{Var}(X_k) = E[\mathrm{Var}(X_k \mid \boldsymbol{\theta})] + \mathrm{Var}(E[X_k \mid \boldsymbol{\theta}]).$$

Les deux termes sont :

$$\mathrm{Var}(X_k \mid \boldsymbol{\theta}) = n\theta_k(1-\theta_k), \quad \Rightarrow \quad E[\mathrm{Var}(X_k \mid \boldsymbol{\theta})] = n\pi_k(1-\pi_k) \cdot \frac{\alpha_0}{\alpha_0+1},$$

$$E[X_k \mid \boldsymbol{\theta}] = n\theta_k.$$

Alors :

$$Var(E[X_k \mid \boldsymbol{\theta}]) = \mathrm{Var}(n\theta_k) = n^2 \cdot \mathrm{Var}(\theta_k) = n^2 \cdot \frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0^2(\alpha_0+1)}.$$

En combinant, on obtient :

$$\boxed{\mathrm{Var}(X_k) = n\pi_k(1-\pi_k) \cdot \frac{\alpha_0+n}{\alpha_0+1}.}$$

Lorsque α est petit, la variance devient plus grande, ce qui reflète une plus grande variabilité entre les vecteurs de probabilités.

3. Covariance De même, pour $i \neq j$, la covariance est donnée par :

$$\mathrm{Cov}(X_i, X_j) = E[\mathrm{Cov}(X_i, X_j \mid \boldsymbol{\theta})] + \mathrm{Cov}(E[X_i \mid \boldsymbol{\theta}], E[X_j \mid \boldsymbol{\theta}]).$$

On a :

$$\mathrm{Cov}(X_i, X_j \mid \boldsymbol{\theta}) = -n\theta_i\theta_j \quad \Rightarrow \quad E[\mathrm{Cov}(X_i, X_j \mid \boldsymbol{\theta})] = -n \cdot \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)},$$

$$\mathrm{Cov}(n\theta_i, n\theta_s) = -n^2 \cdot \frac{\alpha_i\alpha_s}{\alpha_0^2(\alpha_0+1)}.$$

En combinant :

$$\boxed{\mathrm{Cov}(X_i, X_s) = -n\pi_i\pi_s \cdot \frac{\alpha_0+n}{\alpha_0+1}.}$$

La matrice des variances covariances

La matrice de variance-covariance de l'estimateur $\hat{\boldsymbol{\pi}}$ est donnée par :

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}) = \frac{1}{nS}\left(1 + \frac{nS-1}{\alpha_0+1}\right)\left[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right],$$

Soit :

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}) = \tfrac{1}{nS}\left(1+\tfrac{nS-1}{\alpha_0+1}\right) \times [\begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & pk \end{pmatrix} - \begin{pmatrix} \pi_1^2 & \pi_1\pi_2 & \cdots & \pi_1\pi_k \\ \pi_2\pi_1 & \pi_2^2 & \cdots & \pi_2\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ \pi_k\pi_1 & \pi_k\pi_2 & \cdots & \pi_k^2 \end{pmatrix}].$$

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}) = \tfrac{1}{nS}\left(1+\tfrac{nS-1}{\alpha_0+1}\right) \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_k \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_1\pi_k & -\pi_2\pi_k & \cdots & \pi_k(1-\pi_k) \end{pmatrix}$$

où

1. $\mathrm{diag}(\boldsymbol{\pi})$ est la matrice diagonale contenant les π_i sur la diagonale .
2. $\boldsymbol{\pi}\boldsymbol{\pi}'$ est le produit matriciel du vecteur $\boldsymbol{\pi}$ avec lui-même :(oú $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_k)$ et $\boldsymbol{\pi'} = (\pi_1, \pi_2, ..., \pi_k)'$) .
3. Cette expression décrit comment les variances et les covariances des composantes du vecteur $\hat{\boldsymbol{\pi}}$ expriment la structure de dépendance dans le modèle.

2.3.3 La relation entre les modèles MMG et DM

La relation entre DM et MMG peut être établie en comparant les expressions de leurs matrices des variances covariances.

Dans le modèle DM, la variabilité entre observations est modélisée par l'introduction d'un vecteur de probabilités aléatoire $\boldsymbol{\pi} \sim \mathrm{D}(\boldsymbol{\alpha})$, ce qui induit une surdispersion. La variance d'un vecteur de comptage $\boldsymbol{X} \sim \mathcal{DM}(n, \boldsymbol{\alpha})$ s'écrit :

$$\mathrm{Var}(X_k) = n\pi_k(1-\pi_k) \cdot \frac{\alpha_0 + n}{\alpha_0 + 1}$$

Dans le modèle MMG, la dépendance entre observations est modélisée par un paramètre de corrélation intra-groupe $\rho \in [0, 1]$, supposé constant. La variance du vecteur $\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s$, somme des S vecteurs indicateurs corrélés, est donnée par :

$$\mathrm{Var}_{MMG}(\boldsymbol{X}) = S\,(1 + (S-1)\rho)\left[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\right].$$

En identifiant les expressions des variances, on obtient :

$$1 + (n-1)\rho = \frac{\alpha_0 + n}{\alpha_0 + 1}$$

D'où la relation directe entre le paramètre de corrélation ρ du modèle MG et le paramètre α_0 du modèle DM :

$$\rho = \frac{1}{\alpha_0 + 1}$$

2.4 Modèle Dirichlet-Multinomial Généralisé

Un modèle Dirichlet-Multinomial Généralisé est développé, dans lequel les vecteurs observés de dénombrements peuvent être corrélés, comme dans le modèle multinomial généralisé. Supposons que S unités soient sélectionnées au hasard dans une population, pour laquelle les vecteurs de proportions sont distribués selon une loi de Dirichlet de paramètre α et $\boldsymbol{u} = (u_1, u_2, ..., u_T)'$

Comme dans le modèle multinomial généralisé, les vecteurs sont identiquement distribués mais non indépendants. Les observations faites à l'instant t sur les S individus sont également corrélées deux à deux, cette corrélation étant mesurée par le paramètre ρ.

2.4.1 La Distribution Dirichlet-Multinomiale généralisée

La loi Dirichlet-Multinomiale généralisée (GDM) est une distribution hiérarchique définie comme suit :

— Le vecteur des proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ suit une loi de Dirichlet généralisée :

$$\boldsymbol{\pi} \sim GD(\boldsymbol{\alpha}, \boldsymbol{u}),$$

où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ et $\boldsymbol{u} = (u_1, \ldots, u_k)$ sont des paramètres strictement positifs.

On dit que $\boldsymbol{\pi} \sim \mathrm{GD}(\boldsymbol{\alpha}, \boldsymbol{u})$ si sa densité de probabilité est donnée par :

$$f(\pi_1, \ldots, \pi_K) = \prod_{k=1}^{K} \left[\frac{\Gamma(\alpha_k + u_k)}{\Gamma(\alpha_k)\Gamma(u_k)} \right] \pi_k^{\alpha_k - 1} \left(1 - \sum_{s=1}^{k-1} \pi_s \right)^{u_k - 1}$$

— Conditionnellement à $\boldsymbol{\pi}$, le vecteur de comptages $\boldsymbol{X} = (X_1, \ldots, X_K)$ suit une loi multinomiale :

$$\boldsymbol{X} \mid \boldsymbol{\pi} \sim \mathrm{M}(n, \boldsymbol{\pi}).$$

La loi conditionnelle de $\boldsymbol{X} \mid \boldsymbol{\pi}$ est :

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) = \frac{n!}{x_1! \cdots x_K!} \prod_{k=1}^{K} \pi_k^{x_k}, \quad \text{avec } \sum_{k=1}^{K} x_k = n.$$

La loi marginale de $\boldsymbol{X}$ (distribution GDM) est obtenue par intégration :

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int_{\Delta_K} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\pi}) f(\boldsymbol{\pi})\, d\boldsymbol{\pi},$$

$$P(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} \left[\frac{\Gamma(\alpha_k + u_k)}{\Gamma(\alpha_k)\Gamma(u_k)} \cdot \frac{\Gamma(x_k + \alpha_k)\Gamma(n_k - x_k + u_k)}{\Gamma(n_k + \alpha_k + u_k)} \right],$$

où les quantités n_k sont définies récursivement par :

$$n_1 = n, \quad \text{et pour } k \geq 2 : \quad n_k = n - \sum_{s=1}^{k-1} x_s.$$

2.4.2 Espérance et variance

Soit :
— S : nombre d'unités (groupes et individus),
— n : nombre d'observations par unité,
— $N = nS$: nombre total d'observations,
— $\boldsymbol{u} = (u_1, \ldots, u_I)'$: vecteur des proportions moyennes (espérance de la loi de Dirichlet).

On suppose que chaque vecteur de comptage $\boldsymbol{X}_s$ suit (conditionnellement) une loi multinomiale :

$$\boldsymbol{X}_s \mid \boldsymbol{\pi}_s \sim \mathcal{M}(n, \boldsymbol{\pi}_s), \quad \text{avec } \boldsymbol{\pi}_s \sim \mathcal{D}(\boldsymbol{\alpha}).$$

Espérance. L'espérance conditionnelle et l'espérance marginale de $\boldsymbol{X}_s$ sont données par :

$$E[\boldsymbol{X}_s \mid \boldsymbol{\pi}_s] = n\boldsymbol{\pi}_s, \quad E[\boldsymbol{X}_s] = nE[\boldsymbol{\pi}_s] = n\boldsymbol{u}.$$

En sommant sur les S groupes :

$$\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s \quad \Rightarrow \quad E[\boldsymbol{X}] = \sum_{s=1}^{S} E[\boldsymbol{X}_s] = Sn\boldsymbol{u} = N\boldsymbol{u}.$$

Variance–Covariance de $\boldsymbol{X}_s$. La variance totale de $\boldsymbol{X}_s$ est obtenue à partir de la formule de la variance totale :

$$\mathrm{Var}(\boldsymbol{X}_s) = E[\mathrm{Var}(\boldsymbol{X}_s \mid \boldsymbol{\pi}_s)] + \mathrm{Var}(E[\boldsymbol{X}_s \mid \boldsymbol{\pi}_s]).$$

a) Variance conditionnelle :

$$\mathrm{Var}(\boldsymbol{X}_s \mid \boldsymbol{\pi}_s) = n\left[\mathrm{diag}(\boldsymbol{\pi}_s) - \boldsymbol{\pi}_s\boldsymbol{\pi}_s'\right],$$
$$\Rightarrow E[\mathrm{Var}(\boldsymbol{X}_s \mid \boldsymbol{\pi}_s)] = n\left[E[\mathrm{diag}(\boldsymbol{\pi}_s)] - E[\boldsymbol{\pi}_s\boldsymbol{\pi}_s']\right].$$

b) Variance de l'espérance conditionnelle :

$$\mathrm{Var}(E[\boldsymbol{X}_s \mid \boldsymbol{\pi}_s]) = \mathrm{Var}(n\boldsymbol{\pi}_s) = n^2\,\mathrm{Var}(\boldsymbol{\pi}_s).$$

Or, si $\boldsymbol{\pi}_s \sim \mathcal{D}(\boldsymbol{\alpha})$, alors :

$$\mathrm{Var}(\boldsymbol{\pi}_s) = \frac{1}{1+\alpha_0}\left[\mathrm{diag}(\boldsymbol{u}) - \boldsymbol{u}\boldsymbol{u}'\right].$$

En combinant les deux composantes, la variance totale de $\boldsymbol{X}_s$ devient :

$$\mathrm{Var}(\boldsymbol{X}_s) = \left(n + \frac{n^2}{1+\alpha_0}\right)\left[\mathrm{diag}(\boldsymbol{u}) - \boldsymbol{u}\boldsymbol{u}'\right].$$

On définit alors :

$$R = n + \frac{n^2}{1+\alpha_0} = \frac{n(1+n+\alpha_0)}{1+\alpha_0},$$

et donc :

$$\mathrm{Var}(\boldsymbol{X}_s) = R \cdot \left[\mathrm{diag}(\boldsymbol{u}) - \boldsymbol{u}\boldsymbol{u}'\right].$$

Variance du total $\boldsymbol{X} = \sum_{s=1}^{S} \boldsymbol{X}_s$. Les vecteurs $\boldsymbol{X}_s$ étant corrélés entre eux via une corrélation intra-groupe $\rho \in]0, 1[$, la variance totale du vecteur $\boldsymbol{X}$ s'écrit :

$$\mathrm{Var}(\boldsymbol{X}) = N \cdot R \cdot (1 + \rho(S-1)) \cdot [\mathrm{diag}(\boldsymbol{u}) - \boldsymbol{u}\boldsymbol{u}'],$$

où :

- $N = nS$: nombre total d'observations ;
- $R = \dfrac{n(1+n+\alpha_0)}{1+\alpha_0}$: facteur d'inflation de la variance ;
- ρ : coefficient de corrélation intra-groupe ;
- $\boldsymbol{u}$: vecteur des proportions moyennes .

2.4.3 L'estimation de la corrélation

L'estimation de la corrélation intra-temporelle constitue une étape essentielle dans l'analyse de données catégorielles observées à plusieurs instants dans le temps. En effet, les distributions observées à différents moments peuvent présenter une dépendance, appelée corrélation intra-temporelle, qui influence la validité des inférences statistiques.

Pour mesurer cette corrélation commune ρ entre les S instants d'observation, on commence par calculer la matrice empirique de corrélation $R = (r_{ss'})$, où chaque coefficient $r_{ss'}$ est la corrélation de Pearson entre les vecteurs de fréquences ou d'effectifs des catégories observés aux instants s et s'. Pour deux instants distincts s et s', $r_{ss'}$ est défini par :

$$r_{ss'} = \frac{\sum_{i=1}^{I}(X_{si} - \bar{X}_s)(X_{s'i} - \bar{X}_{s'})}{\sqrt{\sum_{i=1}^{I}(X_{si} - \bar{X}_s)^2}\sqrt{\sum_{i=1}^{I}(X_{s'i} - \bar{X}_{s'})^2}},$$

où $\mathbf{X}_s = (X_{s1}, \ldots, X_{sI})$ et $\mathbf{X}_{s'} = (X_{s'1}, \ldots, X_{s'I})'$ sont les vecteurs des effectifs ou fréquences aux instants s et s', et $\bar{X}_s$ (resp. $\bar{X}_{s'}$) est la moyenne des composantes du vecteur $\mathbf{X}_s$ (resp. $\mathbf{X}_{s'}$). Ensuite, l'estimateur $\hat{\rho}$ est obtenu en faisant la moyenne de toutes les corrélations hors diagonale, c'est-à-dire la moyenne des corrélations entre toutes les paires distinctes d'instants. Cette moyenne est calculée en sommant les corrélations $r_{ss'}$ pour tous les couples (s, s') tels que $s < s'$, puis en divisant par le nombre total de paires distinctes, qui est égal à $\frac{S(S-1)}{2}$:

$$\hat{\rho} = \frac{2}{S(S-1)} \sum_{1 \le s < s' \le S} r_{ss'}.$$

Cette moyenne représente la corrélation intra-temporelle moyenne et permet d'ajuster les modèles statistiques en tenant compte de la dépendance existant entre les observations temporelles. Ignorer cette corrélation peut conduire à une sous-estimation de la variance réelle et à des erreurs dans les tests d'hypothèses. Ainsi, cette estimation est cruciale pour améliorer la robustesse et la fiabilité des analyses portant sur des données temporellement corrélées.

Une fois l'estimateur $\hat{\rho}$ de la corrélation intra-temporelle obtenu, il devient possible d'ajuster les statistiques-test classiques en introduisant un facteur de correction noté F_c, permettant

de compenser l'excès de variation induit par la dépendance entre les observations. Ce facteur de variation supplémentaire F_c est crucial pour éviter la sous-estimation de la variance et rendre les inférences valides. Plusieurs méthodes existent pour estimer F_c. Une approche simple, largement utilisée et facilement implémentable dans les logiciels statistiques, consiste à s'appuyer sur la statistique de Pearson issue du test d'indépendance dans un tableau de contingence à double entrée de dimension $I \times S$.

Dans ce cadre, une estimation pratique de F_c est donnée par :

$$\hat{F}_c = \frac{T}{(I-1)(S-1)},$$

où T désigne la statistique de Pearson calculée à partir du tableau de contingence, c'est-à-dire :

$$T = \sum_{i=1}^{I} \sum_{s=1}^{S} \frac{(X_{si} - E_{si})^2}{E_{si}},$$

avec X_{si} les effectifs observés et E_{si} les effectifs théoriques attendus sous l'hypothèse d'indépendance, donnés par :

$$E_{si} = \frac{(\sum_{s'} X_{s'i})(\sum_{i'} X_{si'})}{N},$$

où N est le total général du tableau.

Cette approche, bien que simple, fournit une estimation consistante du facteur de correction F_c et permet ainsi de réajuster les variances estimées dans les tests asymptotiques. Il est aussi possible d'estimer F_c à l'aide de modèles de régression ou des méthodes de type ANOVA (§1.14), comme suggéré par Koehler et Wilson (1986), ou encore par les techniques proposées par Brier (1980) pour quantifier l'effet de regroupement. Toutefois, dans le cadre de l'analyse de données catégorielles à structure temporelle, l'usage de la statistique de Pearson comme base pour le calcul de $\hat{F}_c$ reste une méthode efficace, simple à mettre en œuvre, et suffisante dans la plupart des cas pratiques.

2.4.4 Test des hypothèses du modèle

Dans l'application du modèle multinomial généralisé, deux hypothèses fondamentales sont généralement posées :

- **(a)** H_0 : Les corrélations entre les unités X_s et $X_{s'}$ sont constantes pour tout $s \neq s'$.
- **(b)** H_0' : Les vecteurs aléatoires $\boldsymbol{X}_s$ sont identiquement distribués selon une loi multinomiale.

Dans cette section, nous présentons des tests statistiques permettant de vérifier la validité de ces hypothèses.

Test de la constance des coefficients de corrélation

Soient I observations dans chaque population et S catégories. On note X_{is} la k-ème composante du vecteur d'observations de la i-ème unité.

La corrélation empirique entre les catégories s et s' est donnée par :

$$r_{ss'} = \frac{\sum_{i=1}^{I}(X_{is} - \bar{X}_s)(X_{is'} - \bar{X}_{s'})}{\sqrt{\sum_{i=1}^{I}(X_{is} - \bar{X}_s)^2} \cdot \sqrt{\sum_{i=1}^{I}(X_{is'} - \bar{X}_{s'})^2}}, \tag{2.4.1}$$

où $\bar{X}_s = \frac{1}{I}\sum_{i=1}^{I} X_{is}$ est la moyenne empirique dans la catégorie s.

Pour chaque catégorie s, on calcule la moyenne des corrélations avec les autres catégories :

$$r_s = \frac{1}{S-1}\sum_{\substack{s'=1 \\ s' \neq s}}^{S} r_{ss'}. \tag{2.4.2}$$

Ensuite, on définit la corrélation moyenne globale entre catégories :

$$\bar{r} = \frac{2}{S(S-1)}\sum_{s<s'} r_{ss'}. \tag{2.4.3}$$

Puis, on introduit la quantité suivante :

$$w = (S-1)\cdot\left[1-(1-\bar{r})^2\right]\cdot\left[S-(S-2)(1-\bar{r})^2\right]^{-1}. \tag{2.4.4}$$

La statistique-test pour évaluer la constance des corrélations est donnée par :

$$\chi^2 = (S-1)(1-\bar{r})^{-2}\left[\sum_{s<s'}(r_{ss'}-\bar{r})^2 - w\sum_{s=1}^{S}(r_s-\bar{r})^2\right]. \tag{2.4.5}$$

Sous l'hypothèse nulle de constance des corrélations entre toutes les paires de catégories, la statistique χ^2 suit approximativement une loi du χ^2 avec $\frac{(S+1)(S-2)}{2}$ degrés de liberté.

$$H_0 : \boldsymbol{\pi}_s = \boldsymbol{\pi}_0, \quad \text{pour tout } s = 1, \ldots, S. \tag{2.4.6}$$

où $\boldsymbol{\pi}_s$ est le vecteur des probabilités de la loi multinomiale associée à la s-ème population, et $E[\boldsymbol{X}_s] = n\boldsymbol{\pi}_s$.

On définit :

— L'estimateur empirique de chaque vecteur $\boldsymbol{\pi}_s$:

$$\hat{\boldsymbol{\pi}}_s = \frac{1}{n}\boldsymbol{X}_s,$$

— L'estimateur global moyen :

$$\hat{\boldsymbol{\pi}}_0 = \frac{1}{S}\sum_{s=1}^{S}\hat{\boldsymbol{\pi}}_s.$$

La statistique du test d'ajustement est donnée par :

$$T = (\hat{\Pi}^{(S)} - \hat{\boldsymbol{\pi}}_{\mathbf{0}})' \left[I_{mS} - \frac{1}{S} \mathbf{1}_S \mathbf{1}_S' \otimes I_m \right] (\hat{\Pi}^{(S)} - \hat{\boldsymbol{\pi}}_{\mathbf{0}}), \tag{2.4.7}$$

où :
— $\hat{\Pi}^{(S)}$ est le vecteur concaténé de tous les $\hat{\boldsymbol{\pi}}_s$,
— I_{mS} est la matrice identité de dimension mS,
— $\mathbf{1}_S$ est un vecteur de dimension J contenant des 1,
— $\otimes$ désigne le produit de Kronecker.

Sous l'hypothèse nulle, cette statistique suit asymptotiquement une loi du χ^2 avec $(m-1)(S-1)$ degrés de liberté, où m est le nombre de modalités de la variable catégorielle.

Interprétation

Ces deux tests permettent de valider les deux fondements du modèle multinomial généralisé :

1. La structure de corrélation constante entre les unités .
2. L'homogénéité des loi multinomiales entre les populations .

Le rejet de l'une de ces hypothèses remettrait en cause l'adéquation du modèle aux données.

2.5 Conclusion

Ce chapitre a permis d'introduire et d'approfondir les modèles statistiques utilisés pour traiter les données catégorielles en présence de **dépendance** intra-groupe et de **surdispersion**. Après avoir présenté le modèle multinomial généralisé, nous avons examiné en détail le modèle Dirichlet-Multinomial et sa généralisation.

Le modèle Dirichlet Multinomial Généralisé constitue une avancée significative pour modéliser des données issues de plans expérimentaux complexes ou de contextes où l'indépendance des observations au sein des groupes ne peut être garantie. L'introduction de la structure de corrélation intra-groupe à travers le paramètre ρ et le facteur de correction supplémentaire F_c permet de mieux capturer la variabilité réelle des données.

Nous avons donnée les expressions explicites de l'espérance et de la variance de ce modèle, ainsi que les méthodes d'estimation des paramètres impliqués. Des tests statistiques ont également été présentés afin de vérifier les hypothèses de validité du modèle, notamment **l'homogénéité** des vecteurs de probabilité et la **constance** de la corrélation intra groupes.

Chapitre3 : Application du modèle DMG à des données réelles

3.1 Introduction

L'analyse des données médicales catégorielles longitudinales, telles que le suivi des symptômes chez des patients asthmatiques, pose des défis statistiques importants. En effet, ces données sont souvent caractérisées par une dépendance temporelle entre les observations répétées sur un même individu, ainsi que par une surdispersion par rapport au modèle multinomial classique.

Le modèle Dirichlet-Multinomial généralisé (DMG) constitue une extension pertinente du modèle multinomial en intégrant une structure de corrélation intra-temporelle, ce qui permet de mieux capturer la variabilité et la dépendance présente dans les données. Ce modèle offre ainsi une meilleure flexibilité et une plus grande précision dans les estimations des paramètres, ainsi que dans les tests d'hypothèses.

Ce chapitre est dédié à l'application du modèle DMG à un jeu de données médicales -simulées, correspondant au suivi quotidien de cent (100) patients asthmatiques pendant 5 jours, avec observation des symptômes respiratoires principaux.
Nous décrivons d'abord les données, puis présentons les méthodes d'ajustement et d'estimation des paramètres du modèle. Enfin, nous interprétons les résultats obtenus, en réalisant des tests de validation des hypothèses du modèle et en discutant de leur pertinence dans ce contexte médical.

3.2 Description des données

Les données utilisées dans cette étude proviennent d'un suivi longitudinal simulé portant sur 100 patients asthmatiques observés quotidiennement pendant une période de 5 jours consécutifs. Chaque patient est interrogé chaque jour afin de déterminer le symptôme respiratoire principal ressenti.

Les modalités observées pour le symptôme sont au nombre de quatre :

— A : absence de gêne respiratoire,
— T : toux,
— E : essoufflement,
— S : sifflements respiratoires.

Ainsi, pour chaque patient et chaque jour, la variable enregistrée est qualitative (catégorielle), prenant une des quatre modalités ci-dessus. Les données ont donc une structure longitudinale, avec des observations répétées sur chaque individu.

La table de contingence des fréquences par symptôme et par jour est présentée dans le tableau 1, illustrant la répartition des modalités observées au fil du temps.

Cette structure de données permet d'étudier l'évolution des symptômes dans le temps, ainsi que la dépendance potentielle entre les observations quotidiennes, ce qui justifie l'utilisation d'un modèle statistique prenant en compte cette corrélation intra-temporelle.

Listing 1 – Simulation et graphique des symptômes

```
set.seed(42)
n <- 100  # nombre de patients
```

TABLE 1 – Tableau de contingence des symptômes par jour

Symptôme	Jour 1	Jour 2	Jour 3	Jour 4	Jour 5
A	26	24	27	23	25
T	38	36	34	39	35
E	23	26	24	25	27
S	13	14	15	13	13

```
S <- 5     # nombre de jours
modalites <- c("A", "T", "E", "S")
probs <- c(0.25, 0.35, 0.25, 0.15)

data <- replicate(S, sample(modalites, n, replace = TRUE, prob = probs))
colnames(data) <- paste0("Jour_", 1:S)

table_sympt <- sapply(1:S, function(j) table(factor(data[, j], levels = moda
colnames(table_sympt) <- paste0("Jour_", 1:S)
rownames(table_sympt) <- modalites

print(table_sympt)

if (!require(ggplot2)) install.packages("ggplot2")
library(ggplot2)
library(reshape2)

df <- as.data.frame(table_sympt)
df$Symptome <- rownames(df)
df_long <- melt(df, id.vars = "Symptome", variable.name = "Jour", value.name =

ggplot(df_long, aes(x = Jour, y = Effectifs, group = Symptome, color = Sympto
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  labs(title = "  volution   des sympt mes respiratoires sur 5 jours",
       x = "Jour",
       y = "Nombre de patients",
       color = "Sympt me") +
  theme_minimal()
```

3.3 Ajustement du modèle Dirichlet-Multinomial Généralisé

Le modèle Dirichlet-Multinomial généralisé (DMG) est une extension du modèle multinomial classique qui permet de modéliser la surdispersion et la corrélation intra-temporelle dans des données catégorielles observées de manière répétitive (dans le temps).

Présentation du modèle

Soit S le nombre d'instants d'observation (ici, 5 jours) et m le nombre de modalités/-catégories (ici, 4 symptômes). Pour chaque instant $s = 1, \ldots, S$, on observe un vecteur de comptages $\mathbf{X}_s = (X_{s1}, X_{s2}, \ldots, X_{sm})$ représentant les effectifs de chaque modalité.

Sous le modèle DMG, ces vecteurs $\mathbf{X}_s$ sont corrélés, et la distribution conjointe tient compte d'une corrélation intra-temporelle moyenne ρ entre les différentes observations.

Estimation des paramètres

Les paramètres principaux à estimer sont :

— Les proportions marginales $\pi = (\pi_1, \ldots, \pi_m)$ représentant la probabilité d'observer chaque modalité.

— La corrélation intra-temporelle moyenne $\hat{\rho}$, estimée à partir des corrélations de Pearson entre les vecteurs de fréquences des jours.

— Le facteur de correction $\hat{F}_c$, qui ajuste la variance des tests pour tenir compte de la dépendance entre observations.

L'estimation de π est obtenue par la moyenne empirique des proportions sur les jours :

$$\hat{\pi} = \frac{1}{S} \sum_{s=1}^{S} \hat{\pi}_s, \quad \text{avec} \quad \hat{\pi}_s = \frac{\mathbf{X}_s}{n_s},$$

où n_s est le total des observations au jour s.

La corrélation intra-temporelle $\hat{\rho}$ est estimée par la moyenne des coefficients de corrélation de Pearson entre tous les couples distincts de jours :

$$\hat{\rho} = \frac{2}{S(S-1)} \sum_{1 \leq s < s' \leq S} r_{ss'},$$

où $r_{ss'}$ est la corrélation empirique entre les vecteurs $\mathbf{X}_s$ et $\mathbf{X}_{s'}$.

Le facteur de correction $\hat{F}_c$ est calculé à partir de la statistique de Pearson T issue du tableau de contingence des modalités par jours :

$$\hat{F}_c = \frac{T}{(m-1)(S-1)}, \quad \text{avec} \quad T = \sum_{i=1}^{m} \sum_{s=1}^{S} \frac{(X_{si} - E_{si})^2}{E_{si}},$$

et E_{si} les effectifs théoriques sous l'hypothèse d'indépendance, calculés par :

$$E_{si} = \frac{(\sum_{s'} X_{s'i})\,(\sum_{i'} X_{si'})}{N},$$

où N est le total général des observations.

Implémentation sous R

L'ajustement du modèle a été réalisé à l'aide du logiciel R, en suivant les étapes suivantes :

— Simulation des données de comptages par jour et par modalité ;
— Calcul des tableaux de contingence ;
— Estimation des proportions marginales et de la corrélation intra-temporelle,
— Calcul du facteur de correction $\hat{F}_c$;
— Réalisation des tests statistiques pour valider les hypothèses du modèle.

Un extrait du code R utilisé est présenté ci-dessous (voir section 3.1 pour le code complet) :

Listing 2 – Estimation des paramètres du modèle DMG

```
# Estimation des proportions marginales
pi_hats <- apply(table_sympt, 2, function(col) col / sum(col))
pi_bar <- rowMeans(pi_hats)

# Estimation de la corr lation intra-temporelle
rss <- matrix(0, S, S)
for (s in 1:(S - 1)) {
  for (s2 in (s + 1):S) {
    rss[s, s2] <- cor(table_sympt[, s], table_sympt[, s2])
    rss[s2, s] <- rss[s, s2]
  }
}
rho_hat <- mean(rss[upper.tri(rss)])

# Calcul du facteur de correction Fc
N <- sum(table_sympt)
row_totals <- rowSums(table_sympt)
col_totals <- colSums(table_sympt)
E <- outer(row_totals, col_totals) / N
T_stat <- sum((table_sympt - E)^2 / E)
Fc_hat <- T_stat / ((m - 1) * (S - 1))
color = "Sympt me") +
```

Cette procédure permet de prendre en compte la dépendance entre observations et d'obtenir des estimations robustes pour le modèle Dirichlet-Multinomial Généralisé adapté aux données longitudinales médicales.

3.4 Interprétation des résultats

L'application du modèle Dirichlet-Multinomial Généralisé aux données médicales simulées a permis d'estimer les principaux paramètres du modèle, à savoir les proportions marginales des symptômes, la corrélation intra-temporelle moyenne entre jours d'observation, et le facteur de correction lié à la surdispersion. Des tests statistiques ont également été réalisés afin de valider les hypothèses fondamentales du modèle.

1. Proportions marginales estimées

   Le vecteur des proportions marginales $\hat{\pi}$ des symptômes, obtenu comme moyenne des proportions journalières, est le suivant :

   $$\hat{\pi} = \begin{pmatrix} \hat{\pi}_A \\ \hat{\pi}_T \\ \hat{\pi}_E \\ \hat{\pi}_S \end{pmatrix} = \begin{pmatrix} 0{,}25 \\ 0{,}36 \\ 0{,}25 \\ 0{,}14 \end{pmatrix}$$

   Cela signifie que la toux est le symptôme le plus fréquent, suivi à parts égales de l'essoufflement et de l'absence de gêne respiratoire. Les sifflements sont moins fréquemment observés.

2. Corrélation intra-temporelle

   L'estimation de la corrélation moyenne entre les jours d'observation donne :

   $$\hat{\rho} = 0{,}956.$$

   Cette valeur très élevée indique une forte dépendance entre les distributions observées à différents jours. Ce résultat confirme l'importance de prendre en compte la structure de corrélation dans l'analyse, ce que permet le modèle DMG.

3. Surdispersion et Facteur de correction

   Le facteur de correction $\hat{F}_c$, estimé à partir de la statistique de Pearson issue du tableau de contingence, est :

   $$\hat{F}_c = 2{,}468.$$

   Ce facteur supérieur à 1 indique la présence d'une surdispersion significative dans les données, c'est-à-dire d'une variabilité plus importante que celle prédite par un modèle multinomial classique.

4. Résultats des tests statistiques

   Deux tests ont été réalisés pour valider les hypothèses de structure du modèle.

   — **Test de *constance* des Corrélations** : la statistique du test est $\chi^2 = 1{,}57$ avec 6 degrés de liberté. La p-valeur étant supérieure à 0,1, on ne rejette pas l'hypothèse nulle (de constance des corrélations entre jours).

   — **Test d'*Homogénéité* des lois multinomiales** : la statistique obtenue est $T = 10{,}8$, suivant approximativement une loi du χ^2 à 12 degrés de liberté. La p-valeur d'environ 0,55 ne permet pas de rejeter l'hypothèse d'homogénéité des distributions. Les lois peuvent donc être considérées homogènes à travers le temps.

Les résultats numériques montrent que le modèle DMG est bien adapté aux données étudiées. Il permet de tenir compte à la fois de la forte corrélation entre observations répétées et de la surdispersion présente dans les données. Les tests de validation confirment que les hypothèses du modèle sont raisonnablement satisfaites. Ce modèle se révèle ainsi pertinent pour l'analyse de données médicales longitudinales catégorielles.

3.5 Discussion

L'application du modèle Dirichlet-Multinomial généralisé (DMG) aux données médicales de suivi des symptômes asthmatiques a mis en évidence plusieurs points importants.

Tout d'abord, l'estimation d'une corrélation intra-temporelle significative souligne l'importance de prendre en compte la dépendance entre observations répétées. Ignorer cette corrélation, comme c'est le cas dans un modèle multinomial classique, conduirait à sous-estimer la variance des estimateurs et pourrait biaiser les conclusions statistiques.

Le facteur de correction $\hat{F}_c$, supérieur à 1, confirme la présence d'une surdispersion, phénomène fréquent dans les données médicales longitudinales où les individus peuvent présenter des profils hétérogènes ou des comportements similaires à travers le temps.

Les tests réalisés ont validé les hypothèses fondamentales du modèle : la constance des corrélations entre les différents jours et l'homogénéité des distributions marginales. Ces résultats renforcent la pertinence du modèle DMG dans ce contexte, et garantissent la validité des inférences statistiques effectuées.

Cependant, certaines limites doivent être soulignées. La taille de l'échantillon, ici fixée à 100 patients, pourrait influencer la précision des estimations et la puissance des tests. De plus, la simulation ne reflète pas nécessairement toute la complexité des données réelles, comme covariables cliniques ou des effets individuels, qui pourraient être intégrés dans des extensions du modèle.

Enfin, l'application de ce modèle offre un cadre statistique robuste pour l'analyse de données médicales longitudinales catégorielles, mais son implémentation demande une (bonne) maîtrise des outils statistiques et informatiques, notamment en R. L'amélioration des méthodes d'estimation et la prise en compte de structures temporelles plus complexes restent des pistes intéressantes pour des travaux futurs.

Ainsi, le modèle DMG constitue une avancée notable pour la modélisation des données corrélées, et son utilisation dans le domaine médical est prometteuse pour une meilleure compréhension des phénomènes cliniques observés dans le temps.

3.6 Conclusion

Dans ce chapitre, nous avons appliqué le modèle Dirichlet-Multinomial Généralisé à un jeu de données médicales simulées, représentant le suivi quotidien de cent (100) patients asthmatiques sur une période de cinq (5) jours. L'objectif était de modéliser la distribution des

symptômes respiratoires catégoriels en tenant compte à la fois de la dépendance temporelle entre les observations répétées et du phénomène de la surdispersion.

Après avoir décrit la structure des données, nous avons procédé à l'ajustement du modèle, estimé les proportions marginales des symptômes, la corrélation moyenne entre les jours, ainsi que le facteur de correction $\hat{F}_c$ permettant d'ajuster les variances lors de l'estimation et des tests d'hypothèse.

Les résultats ont mis en évidence :

— une forte corrélation intra-temporelle ($\hat{\rho} = 0{,}956$) confirmant la dépendance des observations dans le temps ;

— une surdispersion significative ($\hat{F}_c = 2{,}468$) par rapport au modèle multinomial classique ;

— la validité des hypothèses du modèle (corrélation constante et homogénéité des lois multinomiales), justifiée par des tests statistiques non significatifs.

Ces résultats confirment la pertinence du modèle DMG dans le contexte des données médicales longitudinales catégorielles. Il permet de capturer des caractéristiques importantes des données, souvent ignorées dans les modèles plus simples, et améliore la robustesse des inférences statistiques. Ce travail ouvre ainsi la voie à des analyses plus poussées, notamment l'introduction de covariables cliniques ou environnementales pour expliquer la variabilité observée.

Conclusion Générale

L'objectif de ce mémoire était de présenter et d'appliquer le modèle Dirichlet Multinomial Généralisé (DMG), une **extension** puissante du modèle multinomial pour l'analyse de données **catégorielles** comportant de la **surdispersion** et de la **dépendance** temporelle.

Dans une **première** partie, nous avons rappelé les notions fondamentales relatives à quelques lois de probabilité discrètes -utiles au thème abordé dans ce travail, ainsi que les outils d'inférence statistique nécessaires à la compréhension du modèle DMG. Ensuite, nous avons détaillé la construction du modèle Dirichlet Multinomial Généralisé, ses propriétés théoriques, ainsi que l'estimation des paramètres et les tests d'hypothèses permettant de valider son application.

L'analyse empirique (**application**) menée dans le troisième chapitre, basée sur des données médicales simulées, a permis de mettre en évidence les apports du modèle DMG et son utilité, voire sa nécessité. Les résultats ont, d'ailleurs, montré une forte corrélation entre les observations répétées et une surdispersion notable, deux caractéristiques prises en compte par ce modèle. Les tests statistiques ont confirmé la validité des hypothèses sous-jacentes, renforçant la crédibilité des inférences réalisées.

Ce travail met en lumière l'importance d'utiliser des modèles adaptés à la structure des données, en particulier dans les contextes médicaux ou les observations sont souvent dépendantes dans le temps. Le modèle DMG s'inscrit ainsi comme un outil flexible, puissant, et pertinent pour des analyses futures, notamment lorsqu'il s'agit d'introduire des covariables ou de comparer plusieurs groupes.

Des perspectives d'amélioration et d'extension du modèle sont envisageables, notamment en explorant des versions hiérarchiques, bayésiennes, ou mixtes du modèle DMG. De même, l'application à des données réelles plus complexes constituerait un prolongement naturel de cette étude.

Bibliographie

1. Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, 27, 162–167.
2. Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. John Wiley and Sons, New York.
3. Cochran, W. G. (1943). Analysis of variance for percentages based on unequal numbers. *Journal of the American Statistical Association*, **38**, 287–301.
4. Koehler, K. J., & Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communication in Statistics* .
5. Lawley, D. N. (1963). On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics*, **34**, 149–151.
6. Moore, D. S. (1977). Generalized inverses, Wald's method and construction of chi-squared tests of fit. *Journal of the American Statistical Association*, **72**, 131–137.
7. Moseman, J. E. (1962). On the compound multinomial distribution, the multivariate-t distribution, and correlation among proportions. *Biometrika*, **49**, 65–82.
8. Saporta, G. (2006). *Probabilités, analyse de données et statistique* (2^{e} édition). Paris : Éditions Technip.
9. **Tallis, G. M.** (1962). The use of a generalized multinomial distribution in the estimation of correlation in discrete data. *Journal of the Royal Statistical Society, Series B*, **24**, 530–534.
10. Tallis, G. M. (1964). Further models for estimating correlation in discrete data. *Journal of the Royal Statistical Society, Series B*, **26**, 82–85.
11. Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482.
12. Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effects in the Dirichlet-multinomial model. *Communications in Statistics, Theory and Methods*, **A15**(4).