Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud Mammeri de Tizi-Ouzou

Faculté des Sciences

Département de Mathématiques

Mémoire de Fin d'Études

Présenté en vue de l'obtention du diplôme de

Master Académique en Mathématiques

Analyse des Tables de Contingence

Spécialité : Statistiques et Probabilités

Présenté par : M^{lle} HOUACINE Lisa

Membres du jury :

Président :	GRAICHE Farid, (UMMTO) MCB
Examinateur :	AIT MOHAMMED Noura, (UMMTO) MCB
Encadrant :	MEHIRI Mohamed, (UMMTO) MAA

Soutenu le : 02/07/2025 **Année Universitaire : 2024–2025**

Table des matières

Remerciements

Au terme de ce travail, je tiens à exprimer ma profonde gratitude à toutes les personnes qui, de près ou de loin, ont contribué à sa réalisation.

Je remercie tout d'abord Dieu, pour m'avoir accordé la force, la patience et la persévérance nécessaires tout au long de ce parcours.

Je tiens à adresser mes remerciements les plus sincères à mon encadrant, pour sa disponibilité, ses conseils précieux, son accompagnement rigoureux et sa bienveillance tout au long de ce mémoire. Son expertise et son exigence m'ont permis de progresser tant sur le plan scientifique que personnel.

Je remercie également l'ensemble des enseignants du Master "Probabilités et Statistiques" pour la qualité des enseignements dispensés, leur écoute et leur passion pour la transmission du savoir.

À mes parents, je dois tout. Leur amour, leur soutien inconditionnel et leurs sacrifices silencieux ont été le socle sur lequel j'ai construit ce parcours. Merci du fond du cœur.

Un remerciement particulier à mes frères, Mohamed El-Amine et Yanis, pour leur présence, leur affection et leur soutien constants.

À mon fiancé Titoum Amine, merci pour ton amour, ta patience et ta foi en moi, même dans les moments de doute.

Je n'oublie pas mes amies chères, Amel et Amira, pour leur amitié sincère, leur soutien moral et les innombrables moments de réconfort partagés.

Enfin, je remercie chaleureusement mes collègues de promotion pour les moments de partage, de solidarité et d'entraide. Ce mémoire est aussi le fruit de cette aventure humaine que nous avons vécue ensemble.

Avec toute ma reconnaissance,
Houacine Lisa

Introduction Générale

L'analyse des données constitue un pilier fondamental dans de nombreux domaines scientifiques tels que la médecine, les sciences sociales, l'économie, l'ingénierie, et bien d'autres.

Parmi les différents types de données rencontrées, celle des **variables catégorielles** occupe une place prépondérante. En effet, les données catégorielles décrivent des attributs qualitatifs ou des modalités, comme le genre, la couleur, la classe sociale, ou encore le mode de transport utilisé. Ces données ne se prêtent pas aux méthodes classiques de modélisation (de variables quantitatives), ce qui rend nécessaire le recours à des outils spécifiques pour les analyser de manière rigoureuse.

L'un des outils les plus couramment utilisés dans l'étude de telles données est la **table de contingence**, support de leur recueil, aussi appelée tableau croisé. Elle permet de représenter la distribution conjointe de deux ou plusieurs variables qualitatives sous forme d'**effectifs**.
À travers ces tableaux, on peut visualiser des relations potentielles entre variables et poser des hypothèses statistiques telles que l'indépendance, l'homogénéité, la symétrie -quand il s'agit de tables carrées, ou l'association entre variables/modalités.

L'objectif principal de cette étude est de présenter et d'approfondir les méthodes d'analyse des tables de contingence. En particulier, nous nous intéressons à la modélisation *linéaire*, ainsi qu'à la modélisation **log-linéaire** qui constitue une approche puissante et flexible pour étudier la structure des dépendances entre des variables catégorielles. Nous explorerons également les (statistiques des) **tests du Khi-deux**, outil fondamental pour appréhender les hypothèses précitées.

Ce mémoire est structuré en trois chapitres. Dans le **premier** chapitre, nous introduisons les notions de base relatives aux données catégorielles, aux tables de contingence, et aux lois statistiques associées telles que la loi multinomiale.
Le **deuxième** chapitre est consacré à la formalisation mathématique des modèles log-linéaires et aux formes quadratiques associées aux tests et leurs statistiques sous-jacentes.
Dans le **troisième** chapitre, on illustrera l'application pratique de ces concepts à travers des exemples concrets, des tests statistiques, et l'interprétation des résultats.

À travers cette étude, nous espérons non seulement maîtriser les fondements théoriques de l'**analyse des tables de contingence**, mais également développer une capacité critique d'interprétation des résultats, en tenant compte des limites et des conditions d'utilisation des méthodes statistiques employées.

Chapitre 1

Données Catégorielles

1.1 Données Catégorielles :

Les données catégorielles sont un ensemble de données classées en différentes catégories. Elles peuvent être obtenues, par exemple, lorsqu'une organisation collecte des informations individuelles sur son personnel.

Lorsqu'un chercheur étudie un concept, il doit recueillir des données à son sujet. Celles-ci peuvent se présenter sous différentes formes, telles que la profession, le niveau d'éducation, la couleur des yeux ou encore des préférences alimentaires. Il est essentiel de bien identifier la nature des données collectées afin de les analyser correctement.

Types de données : Les données utilisées en recherche se répartissent en deux grandes catégories :

— **Données Catégorielles :** Les données catégorielles désignent des informations stockées et identifiées selon des étiquettes ou des noms. Il s'agit d'un type de données qualitatives qui peuvent être classées en groupes, appelés **catégories**, non quantifiables numériquement.

Elles incluent des variables telles que le type de véhicule possédé, la marque de téléphone utilisé(e) ou le domaine/niveau d'étude(s) d'un individu. Bien que certaines données puissent être représentées par des nombres (comme un numéro de département ou un code client), ces chiffres n'ont pas de signification mathématique.

Exemples de données catégorielles :
— Type de musique préférée,
— Niveau d'études,
— Couleur des yeux,
— Type de transport.

Un moyen simple de différencier les données catégorielles des données numériques (communément dites quantitatives en statistique) est d'évaluer si leur moyenne peut être calculée. Si c'est possible, elles sont numériques ; sinon, elles sont catégorielles.

— **Données numériques :** Ce sont des données de variables se rapportant au numérique. Il en est ainsi de toute donnée concernant les mesures de poids, de distance, de volume -par exemple, ou toute autre observation dont le résultat est quantifié numériquement ; on en trouve de deux types : **discrètes :** (dont les modalités sont isolées -tel le nombre d'enfants à charge d'un employé) et **continues :** (dont on peut faire une répartition en classes -telles que le salaire d'un employé).

1.1.1 Types de données catégorielles

Les données catégorielles regroupent des observations classables en catégories. Elles sont souvent représentées sous forme de diagrammes à barres ou circulaires. Il existe deux principaux types de données catégorielles :

Données Nominales

Les données nominales sont constituées de catégories qui ne suivent aucun ordre spécifique. Elles sont aussi appelées **échelle nominale** et ne peuvent pas être classées

Exemples :
— Type de véhicules (voiture, moto, vélo)
— Marque de téléphone (Apple, Samsung, Huawei)
— Type de cuisine préférée (italienne, asiatique, méditerranéenne)

Données Ordinales

Les données ordinales sont caractérisées par un ordre naturel sur les diverses catégories, mais sans différence mesurable entre les valeurs. Elles interviennent et sont souvent utilisées dans des sondages (-d'opinion), des questionnaires (-à choix multiples) ou encore dans des études économiques (qualité d'une prestation de service) ou financières.

Exemple :
— Niveau de satisfaction (faible, moyen, élevé)
— Classement dans un concours (1er, 2ème, 3ème)
— Niveau d'expérience (débutant, intermédiaire, avancé)

1.1.2 Caractéristiques des données catégorielles

Les données catégorielles possèdent plusieurs caractéristiques spécifiques :
— **Catégories :** Divisées en *nominales* et *ordinales*.
— **Non numériques :** Décrites sous forme de mots ou de catégories, et non de valeurs numériques
— **Nature :** Binaires (*oui/non*) ou non binaires (plus de deux options).
— **Valeurs numériques :** Parfois représentées par des chiffres sans signification statistique.
— **Représentation Graphique :** Diagrammes en barres et circulaires pour illustrer les fréquences et/ou pourcentages.
— **Analyse :** Utilisation du mode (données nominales) et de la médiane (données ordinales).

1.1.3 Exemples de Données Catégorielles

Supposons qu'un restaurant souhaite connaître les préférences de ses clients en matière de desserts. Un sondage est réalisé auprès d'un échantillon de clients, et les résultats suivants sont obtenus :

Dessert préféré	**Fréquence absolue**
Tarte aux pommes	08
Chocolat fondant	12
Crème brûlée	05
Tiramisu	10

Table 1.1 – Exemple de données catégorielles sur les desserts préférés

Les données présentées ici sont catégorielles, car elles sont classées en différentes catégories, selon les préférences alimentaires des clients.

Remarque : Les variables catégorielles ont une importance particulière dans l'analyse statistique des données, notamment dans les domaines où les variables ne peuvent ni

être mesurées ni ordonnées de manière précise. Elles sont largement utilisées dans les enquêtes, les études de marché, les sciences sociales, médicales et bien d'autres disciplines où l'information collectée est souvent qualitative par nature. Leur traitement requiert, donc, des méthodes spécifiques qui diffèrent sensiblement de celles appliquées aux données numériques, notamment en raison de leur caractère non métrique.

Il est fondamental de distinguer les types de variables catégorielles, telles que les variables *nominales*, qui ne possèdent aucun ordre naturel entre les modalités, et les variables *ordinales*, qui présentent une hiérarchie entre les catégories. Cette distinction conditionne le choix des méthodes statistiques à employer ainsi que l'interprétation des résultats (Saporta, 2006).

Dans la suite de ce travail, nous allons nous intéresser plus spécifiquement aux **tables de contingence**, un outil fondamental pour explorer les liens entre deux ou plusieurs variables catégorielles. Ces tableaux permettent de résumer efficacement les données, de visualiser les co-occurrences des modalités, et de poser les bases de tests statistiques comme le test du Chi-deux, très utilisé pour évaluer l'indépendance entre variables, par exemple. La maîtrise de ces concepts est indispensable pour toute personne amenée à manipuler des données qualitatives dans un cadre statistique.

1.2 Table de Contingence :

En statistique, une **table de contingence** est un outil essentiel pour analyser la relation entre deux variables qualitatives. Elle permet de visualiser les effectifs communs à différentes catégories et d'identifier des éventuelles corrélations entre elles.

1.2.1 Définition d'une Table de Contingence :

Une table de contingence est un tableau à double entrée où :
- Les **lignes** correspondent aux catégories d'une première variable.
- Les **colonnes** correspondent aux catégories d'une seconde variable.
- Les **cellules** contiennent les effectifs des observations appartenant à chaque combinaison de deux catégories (l'une de la variable ligne et l'autre de la variable colonne).

1.2.2 Exemple de Table de Contingence :

Supposons que, pour un groupe de personnes, nous voudrions étudier la relation entre le genre et la boisson préférée, et que les résultats observés soient les suivants :

Boisson préférée \ Sexe	**Homme**	**Femme**	**Total**
Café	40	35	75
Thé	20	45	65
Jus	25	35	60
Total	85	115	200

Table 1.2 – Table de contingence des préférences de boisson selon le genre

n_{ij} = Nombre d'Observations où $\{X = x_i\}$ et $\{Y = y_j\}$.

Types de Tables de Contingence :

On distingue principalement deux types :

1. **Table de contingence *simple*** : compare deux variables qualitatives.
2. **Table de contingence *multiple*** : intègre une troisième variable ou plus pour une analyse plus détaillée.

1.2.3 Analyse des tables de contingence :

Effectifs Marginaux

Les **totaux de lignes et les totaux colonnes** permettent d'analyser les distributions marginales des variables.

Fréquences Relatives :

On calcule les fréquences relatives par :

$$f_{ij} = \frac{n_{ij}}{n_{Total}}$$

où :

— f_{ij} est la fréquence relative de la cellule (i, j),
— n_{ij} est l'effectif observé,
— n_{Total} est le nombre total d'observations (taille de l'échantillon)

1.2.4 Représentation Graphique :

Pour mieux visualiser les relations, on utilise :

— **Diagrammes en barres** pour comparer les fréquences (absolues ou relatives).
— **Diagrammes circulaires** pour figurer (généralement) les proportions (en termes de degrés de l'angle du secteur correspondant).
— **Mosaic Plots** pour une représentation plus détaillée.

Remarque : Les tables de contingence sont un outil puissant en analyse des données. Elles permettent d'identifier et d'interpréter les relations entre variables qualitatives de manière claire et structurée.

1.3 La Loi Multinomiale :

La **loi Multinomiale** est une généralisation de la loi Binomiale, permettant de modéliser une expérience aléatoire où chaque essai peut aboutir à l'une des k catégories possibles avec des probabilités respectives $p_1, p_2, ..., p_k$, vérifiant :

$$\sum_{i=1}^{k} p_i = 1$$

Si l'on répète cette expérience n fois de manière indépendante, et que l'on note X_i le nombre de fois où l'issue i est observée, alors le vecteur aléatoire $(X_1, X_2, ..., X_k)$ suit une **loi multinomiale** de paramètres $(n, p_1, p_2, ..., p_k)$, on note :

$$(X_1, X_2, ..., X_k) \sim \mathcal{M}(n; p_1, p_2, ..., p_k),$$

ou $N \sim \mathcal{M}(n, \mathbf{p})$, en posant $\mathbf{p} = (p_1, p_2, ..., p_k)$ et $N = (X_1, X_2, ..., X_k)$.

Cela signifie que la probabilité d'observer une répartition donnée des effectifs est donnée par :

$$P(X_1 = n_1, ..., X_k = n_k) = \frac{n!}{n_1!n_2!...n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

sous la contrainte :

$$\sum_{i=1}^{k} n_i = n$$

Ce modèle est largement utilisé en statistique, notamment dans l'analyse des **tables de contingence**, les sondages, et les expériences dont les observations sont réparties en plusieurs catégories.

1.3.1 Exemple : Le schéma de l'urne à catégories :

Considérons une partition d'un ensemble en k événements de probabilités respectives $p_1, p_2, ..., p_k$.

On répète alors indépendamment n fois l'expérience aléatoire et on compte le nombre d'occurrences des événements C_i, ce qui donne les valeurs des variables X_i.

Le vecteur aléatoire discret $(X_1, X_2, ..., X_k)$ suit alors par définition une **loi multinomiale** de paramètres $(n, p_1, p_2, ..., p_k)$.

Ce schéma se retrouve notamment dans les problèmes de sondages : une population est divisée en k catégories, et on tire **avec remise** n individus. On compte ensuite le nombre d'individus appartenant à chaque catégorie.

Ce même processus est observé dans le comptage des réalisations d'une variable aléatoire X lorsque ses valeurs sont réparties en k classes avec des probabilités p_i. On observe alors, parmi n individus, la distribution des effectifs dans ces classes. C'est cette approche qui est utilisée pour construire un **histogramme**.

Le tableau suivant illustre la répartition des catégories :

Catégorie	**Probabilité** p_i	**Nombre d'occurrences** n_i
C_1	p_1	n_1
C_2	p_2	n_2
$\vdots$	$\vdots$	$\vdots$
C_k	p_k	n_k
Total	$\sum p_i = 1$	$\sum n_i = n$

1.3.2 Propriétés de la Loi Multinomiale

Les composantes X_i du vecteur multinomial sont linéairement dépendantes, car elles sont liées par la relation :

$$\sum_{i=1}^{k} X_i = n(= \sum_{i=1}^{k} n_i).$$

On a évidemment

$$\sum_{i=1}^{k} p_i = 1$$

L'espérance et la variance sont alors :

$$E(X_i) = np_i$$

$$V(X_i) = np_i(1 - p_i)$$

La loi conditionnelle de X_i sachant que $X_j = n_j$ suit également une loi binomiale :

$$X_i|X_j \sim \mathcal{B}\left(n - n_j, \frac{p_i}{1 - p_j}\right)$$

1.3.3 Espérance et Matrice de Variance covariance :

Comme chaque X_i suit une loi $\mathcal{B}(n, p_i)$, on a :

$$E(X_i) = np_i$$

La covariance entre X_i et X_j est donnée par :

$$\text{Cov}(X_i, X_j) = -np_ip_j \quad \text{si } i \neq j$$

La **matrice de variance-covariance** Σ est donc :

$$\Sigma = n \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{bmatrix}$$

Méthode de Calcul de la Matrice de Variance-Covariance :

1. **Calcul des Variances** : Pour chaque variable X_i, on applique la formule :

$$V(X_i) = np_i(1 - p_i)$$

2. **Calcul des Covariances :** Pour un couple de variables (X_i, X_j) avec $i \neq j$, on a :

$Cov(X_i, X_j) = E(X_i.X_j) - E(X_i)E(X_j)$

et, comme $E(X_i.X_j) = n(n-1)p_ip_j$ et $E(X_i)E(X_j) = n^2p_ip_j$,

il s'ensuit que :

$Cov(X_i, X_j) = n(n-1)p_ip_j - n^2p_ip_j$

soit :

$$\mathbf{Cov}(X_i, X_j) = -np_ip_j$$

3. **Construction de la Matrice** :
   — Les termes diagonaux correspondent aux variances $V(X_i)$.
   — Les termes hors diagonale sont les covariances $\text{Cov}(X_i, X_j)$.

Ainsi, la matrice Σ représente la dispersion des valeurs prises par les variables C_i autour de leurs moyennes respectives.

1.3.4 Loi Limit

Considérons un vecteur aléatoire $(X_1, X_2, \ldots, X_k)$ suivant une loi multinomiale de paramètres n et $(p_1, p_2, \ldots, p_k)$, c'est-à-dire :

$$(X_1, \ldots, X_k) \sim \mathcal{M}(n, p_1, \ldots, p_k), \quad \text{avec } \sum_{i=1}^{k} p_i = 1$$

On a alors :
— $\mathbb{E}[X_i] = np_i$
— $\text{Var}(X_i) = np_i(1 - p_i)$
— $\text{Cov}(X_i, X_j) = -np_ip_j$ pour $i \neq j$

D'après le théorème central limite multivarié, lorsque $n \to \infty$, on a la convergence en loi :

$$\frac{1}{\sqrt{n}}(X_1 - np_1, X_2 - np_2, \ldots, X_k - np_k) \xrightarrow{d} \mathcal{N}_k(0, \Sigma)$$

où Σ est la matrice de variance-covariance du vecteur multinomial.

Remarque : La loi multinomiale joue un rôle central en statistique, notamment dans l'analyse des **tables de contingence** et dans l'application du **test du khi-deux** (χ^2).

Elle permet de modéliser des expériences où plusieurs catégories sont possibles et son approximation normale est particulièrement utile en statistique inférentielle lorsque n est grand.

1.4 Test du Khi-deux :

Le **test du khi-deux ($\chi 2$)** est un test statistique utilisé pour analyser des données catégorielles et vérifier si une différence entre des fréquences observées et attendues est significative ou due au hasard. Il existe principalement deux types de tests du khi-deux :
— **Test du khi-deux d'ajustement**
— **Test du khi-deux d'indépendance**
Le résultat est comparé à une valeur critique issue de la loi du khi-deux avec un certain nombre de degrés de liberté.

Conditions d'application :

Les conditions d'application du test du khi-deux sont les suivantes :
— Les effectifs attendus dans chaque cellule doivent être supérieurs ou egale à 5 (parfois>=10) pour garantir une approximation correcte de la loi khi-deux.
— Les expériences doivent être indépendantes.
— La taille n de L'échantillon doit être suffisamment grande.

Interprétation des résultats :

L'interprétation des résultats du test du khi-deux repose sur la p-valeur :
— Si la *p-valeur* associée au test est inférieure au seuil α (souvent 5% parfois 1%), on rejette l'hypothèse nulle. Cela signifie qu'il existe une association entre les variables ou que la distribution observée diffère de celle attendue.
— Si la *p-valeur* est supérieure à α, on ne rejette pas l'hypothèse nulle.

Lien avec la loi multinomiale et le TCL :

Le test du khi-deux repose sur l'approximation de la statistique de test par une loi du khi-deux lorsque les effectifs sont suffisamment grands. Cette approximation provient du **théorème central limite** (TCL), qui assure que la somme de variables indépendantes tend vers une loi normale.

1.4.1 Test du Khi-Deux d'*Ajustement* :

Ce test permet de vérifier si une variable catégorielle suit une distribution théorique donnée.
Hypothèses :
— H_0 : Les données suivent la distribution théorique.
— H_1 : Les données ne suivent pas la distribution théorique.
Procédure :

1. Collecter les données et déterminer les **fréquences observées** (O_i).
2. Calculer les **fréquences attendues** (E_i) selon la distribution théorique(sous H_O).
3. Calculer la statistique de test :

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{1.1}$$

où :
— O_i représente le i ème effectif observé.
— E_i représente le i ème effectif attendu (sous l'hypothèse nulle).

4. Comparer cette statistique à la valeur critique de la loi du khi-deux a $(k-1)-l$ degrés de liberté,) (l) étant le nombre de relations liant les deux distributions)
5. Si la statistique est supérieure à la valeur critique, on rejette H_0.

Exemple : Lancer de dé :

Pour vérifier si un dé est équilibré, on l'a lancé 120 fois, ce qui donne la distribution suivante :

Face	**Observé** O_i	**Attendu** E_i
1	25	20
2	20	20
3	18	20
4	22	20
5	17	20
6	18	20

Hypothèses du test :

— $H_0 : p_1 = p_2 = \cdots = p_6 = \dfrac{1}{6}$ (le dé est équilibré)

— $H_1 : \exists\, i \in \{1, \ldots, 6\}$ tel que $p_i \neq \dfrac{1}{6}$ (le dé n'est pas équilibré)

Calcul de la statistique du khi-deux :

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(25-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20}$$

$$\chi^2 = \frac{25}{20} + \frac{0}{20} + \frac{4}{20} + \frac{4}{20} + \frac{9}{20} + \frac{4}{20} = \frac{46}{20} = 2.3$$

Décision :

— **Degrés de liberté :** $k - 1 = 6 - 1 = 5$

— **Valeur critique au seuil de 5% :** $\chi^2_{0.05,5} \approx 11.07$

$$\chi^2_{\text{calculée}} = 2.3 < 11.07 = \chi^2_{\text{critique}}$$

Conclusion :

On ne rejette pas l'hypothèse nulle H_0. Il n'y a pas de différence significative entre les fréquences observées et les fréquences théoriques. Le dé semble donc **équilibré**. (En vrai, les données observées ne permettent pas de remettre en cause l'hypothèse selon laquelle le dé n'est pas truqué ; on admet donc, au seuil $\alpha = 5\%$, que celui-ci est homogène).

1.4.2 Test du Khi-Deux d'*Indépendance* :

Ce test permet de tester si deux variables catégorielles sont indépendantes.

Hypothèses :

— H_0 : Les deux variables sont indépendantes.

— H_1 : Les deux variables ne sont pas indépendantes.

Procédure :

1. Le **tableau de contingence** contiant les fréquences observées (O_{ij}).
2. Calculer les **fréquences attendues** sous l'hypothèse d'indépendance :

$$E_{ij} = \frac{n_{i.} n_{.j}}{n} \tag{1.2}$$

   avec $n_{i.} = \sum_{j=1}^{c} O_{ij}$ et $n_{.j} = \sum_{i=1}^{r} O_{ij}$
3. Calculer la statistique de test :

$$\chi^2 = \sum_{i}^{r} \sum_{j}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{1.3}$$

tq : (r) est le nombre de lignes et (c) est le nombre de colonnes

4. Comparer cette statistique à la valeur critique avec $(r-1)(c-1)$ degrés de liberté.
5. Si la statistique est supérieure à la valeur critique, on rejette H_0.

1.4.3 Exemple : Sexe et Choix de Filière :

On étudie si le choix de filière dépend du sexe. Les données observées sont :

Sexe \ Filière	**Mathématiques**	**Informatique**	**Biologie**	**Total**
Homme	30	25	20	75
Femme	20	35	10	65
Total	50	60	30	140

Hypothèses du test :
— H_0 : Le choix de filière est indépendant du sexe.
— H_1 : Le choix de filière dépend du sexe.

Fréquences attendues (sous H_0) :
On calcule les fréquences attendues selon la formule :

$$E_{ij} = \frac{n_{i.}.n_{.j}}{n}$$

Par exemple :

$$E_{11} = \frac{75 \times 50}{140} = 26.79$$

$$E_{12} = \frac{75 \times 60}{140} = 32.14$$

$$E_{13} = \frac{75 \times 30}{140} = 16.07$$

$$E_{21} = \frac{65 \times 50}{140} = 23.21$$

$$E_{22} = \frac{65 \times 60}{140} = 27.86$$

$$E_{23} = \frac{65 \times 30}{140} = 13.93$$

Calcul de la statistique du khi-deux :

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

$$\chi^2 = \frac{(30-26.79)^2}{26.79} + \frac{(25-32.14)^2}{32.14} + \frac{(20-16.07)^2}{16.07} + \frac{(20-23.21)^2}{23.21} + \frac{(35-27.86)^2}{27.86} + \frac{(10-13.93)^2}{13.93}$$

$$\chi^2 \approx \frac{10.34}{26.79} + \frac{51.00}{32.14} + \frac{15.44}{16.07} + \frac{10.30}{23.21} + \frac{50.84}{27.86} + \frac{15.44}{13.93}$$

$$\chi^2 \approx 0.39 + 1.59 + 0.96 + 0.44 + 1.82 + 1.11 = 6.31$$

Décision :
— **Degrés de liberté** : (nb lignes $-$ 1) $\times$ (nb colonnes $-$ 1) $= (2-1)(3-1) = 2$
— **Valeur critique** (à $\alpha = 5\%$) : $\chi^2_{0.05,2} = 5.99$

$$\chi^2_{\text{calculée}} = 6.31 > 5.99 = \chi^2_{\text{critique}}$$

Conclusion :
On **rejette** l'hypothèse nulle H_0. Il y a donc une dépendance significative entre le sexe et le choix de la filière. Le choix de filière semble donc **dépendre du sexe**.

1.4.4 Conclusion :

Le test du Khi-deux occupe une place importante dans l'analyse statistique des variables catégorielles. Le test d'ajustement permet de comparer une distribution observée à une distribution théorique afin de vérifier la validité d'un modèle. Cette approche est utile dans les situations où l'on souhaite évaluer si les fréquences d'un échantillon suivent une loi donnée.

Le test d'indépendance, quant à lui, sert à étudier l'existence d'une éventuelle dépendance entre deux variables qualitatives à travers l'analyse d'un tableau de contingence. Il permet de déterminer si les modalités de l'une des variables influencent celles de l'autre.

Ces deux tests reposent sur des conditions d'application précises, notamment en ce qui concerne les effectifs attendus. Lorsqu'elles sont respectées, les conclusions obtenues sont fiables et permettent d'appuyer des décisions basées sur les données. De ce fait, ces tests sont largement utilisés dans divers domaines d'application, tant en recherche qu'en pratique.

Chapitre 2

Modèle Linéaire

2.1 Forme quadratique (Rappel) :

Soit E un espace vectoriel sur un corps $\mathbb{K}$ (en général $\mathbb{R}$ ou $\mathbb{C}$), et soit $b : E \times E \to \mathbb{K}$ une forme bilinéaire.

L'application $q : E \to \mathbb{K}$ définie par $q(x) = b(x, x)$ est appelée **forme quadratique associée à la forme bilinéaire** b.

Remarques :

- L'ensemble des formes quadratiques sur E constitue un espace vectoriel.
- Si b est alternée, alors q est identiquement nulle, car $b(x, x) = 0$ pour tout x.
- L'application $b \mapsto q$ est linéaire, et son noyau est l'ensemble des formes bilinéaires alternées.

Proposition :

Toute forme quadratique q sur E est associée à une unique forme bilinéaire symétrique φ appelée *forme polaire* de q.

$$\forall x, y \in E, \quad \varphi(x, y) = \frac{1}{2}\left[q(x + y) - q(x) - q(y)\right] \tag{2.1}$$

2.1.1 Exemples classiques :

1. Si $f, g : E \to \mathbb{K}$ sont des formes linéaires, alors $q(x) = f(x)g(x)$ est une forme quadratique.
2. Si $b : E \times E \to \mathbb{K}$ est une forme bilinéaire symétrique, alors $q(x) = b(x, x)$ est une forme quadratique.
3. Sur $\mathbb{K}^n$, la forme $q(x) = \sum_{i=1}^{n} x_i^2$ est une forme quadratique.

2.1.2 Représentation matricielle d'une forme quadratique :

Soit q une forme quadratique sur un espace vectoriel E de dimension n, muni d'une base $\mathcal{B} = (e_1, \ldots, e_n)$. La forme polaire φ est représentée par une matrice symétrique $A = (a_{ij})$ telle que :

$$\forall x, x = \sum x_i e_i, \quad q(x) = X^T A X, \text{ avec } X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{2.2}$$

Changement de base : Si A est la matrice de q dans la base $\mathcal{B}$ et A' celle dans une autre base $\mathcal{B}'$, alors :

$$A' = P^T A P \tag{2.3}$$

avec P la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$.

2.1.3 Expression polynomiale d'une forme quadratique :

Toute forme quadratique q peut s'écrire sous la forme :

$$q(x) = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{1 \leq i < j \leq n} a_{ij} x_i x_j \tag{2.4}$$

Cette écriture correspond à un polynôme homogène de degré 2, que l'on peut associer à une matrice symétrique.

2.1.4 Équivalence entre formes quadratiques :

Définition :

Deux formes quadratiques q et q' sont dites *équivalentes* s'il existe un isomorphisme $u : E \to F$ tel que :

$$q'(u(x)) = q(x), \quad \forall x \in E$$

Proposition :

Deux formes quadratiques q et q' sont équivalentes **ssi** leurs matrices associées sont congruentes, c'est-à-dire qu'il existe $P \in GL_n(\mathbb{K})$ tel que :

$$A' = P^T A P$$

2.1.5 Notions de Noyau, Rang, Dimension :

— Le noyau de q est : $\ker(q) = \{x \in E \mid q(x) = 0 \text{ et } \varphi(x, y) = 0, \ \forall y \in E\}$
— Le rang de q est le rang de sa matrice associée A
— La dimension de q est la dimension de E

2.1.6 Formes régulières et formes dégénérées :

Soit E un espace vectoriel de dimension finie sur un corps $\mathbb{K}$, et soit $q : E \to \mathbb{K}$ une forme quadratique. On lui associe une forme bilinéaire symétrique φ définie par :

$$\varphi(x, y) = \frac{1}{2}(q(x + y) - q(x) - q(y)), \quad \forall x, y \in E.$$

La forme quadratique q est dite **régulière** si la forme bilinéaire associée φ est non dégénérée, c'est-à-dire si :

$$\ker(q) = \{x \in E \mid \varphi(x, y) = 0, \ \forall y \in E\} = \{0\}.$$

Dans ce cas, la matrice associée à q dans une base donnée est **inversible** (de rang maximal), et son déterminant est non nul.

À l'inverse, si $\ker(q) \neq \{0\}$, on dit que q est une forme **dégénérée**. Cela signifie que la matrice associée est **singulière**, c'est-à-dire que son rang est strictement inférieur à la dimension de l'espace.

Exemples :

— **Forme régulière** :
Soit la forme quadratique définie sur $\mathbb{R}^2$ par :

$$q(x, y) = x^2 + y^2.$$

La matrice associée est :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{avec } \det(A) = 1 \neq 0.$$

Le noyau est trivial ($\ker(q) = \{0\}$), donc q est régulière.

— **Forme dégénérée (non régulière)** :
Soit la forme quadratique suivante :

$$q(x, y) = x^2.$$

Sa matrice associée est :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{avec } \det(A) = 0.$$

Ici, $\ker(q) = \{(0, y) \mid y \in \mathbb{R}\} \neq \{0\}$, donc q est dégénérée (non régulière).

2.1.7 Vecteurs isotropes et cône isotrope :

Un vecteur $x \in E$ est *isotrope* si $q(x) = 0$. Le *cône isotrope* est l'ensemble $\mathrm{Co}(q) = \{x \in E \mid q(x) = 0\}$.

2.1.8 Conique projective :

La conique projective associée à q est l'ensemble des droites vectorielles incluses dans $\mathrm{Co}(q)$, i.e. les sous-espaces vectoriels de dimension 1 contenus dans le cône isotrope.

2.1.9 Déterminant et groupe orthogonal :

Le *déterminant* d'une forme quadratique est défini comme le déterminant de sa matrice associée : $\det(q) = \det(A)$.

Le *groupe orthogonal* $O(q)$ est l'ensemble des automorphismes u de E tels que $q(u(x)) = q(x)$ pour tout $x \in E$.

2.1.10 Diagonalisation des formes quadratiques :

Théorème de Sylvester :

Toute forme quadratique réelle peut être réduite par un changement de base à une somme de carrés (avec des signes + et -) :

$$q(x) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2$$

La *signature* d'une forme quadratique réelle est le couple (p, q) où p est le nombre de valeurs propres positives et q le nombre de valeurs propres négatives dans la forme diagonale.

2.1.11 Formes définies

— q est **positive** si $q(x) \geq 0$ pour tout $x \in E$.
— q est **négative** si $q(x) \leq 0$.
— q est **définie positive** si $q(x) > 0$ pour tout $x \neq 0$.
— q est **définie négative** si $q(x) < 0$ pour tout $x \neq 0$.

2.1.12 Formes quadratiques en statistique :

Une forme quadratique en statistique est une expression de la forme :

$$Q = X^T AX$$

où :
— X est un vecteur aléatoire de dimension n,
— A est une matrice symétrique de taille $n \times n$,
— Q est un scalaire résultant de l'expression quadratique.

Lois associées à une forme quadratique :

La loi de $Q = X^T AX$ dépend de la loi de X et des propriétés de la matrice A. Voici les cas les plus courants :

1. Cas Gaussien centré :

Si :

$$X \sim \mathcal{N}_n(0, I_n)$$

et si A est une matrice symétrique semi-définie positive, alors :

$$Q = X^T AX \sim \sum_{i=1}^{n} \lambda_i Z_i^2$$

où :
— λ_i sont les valeurs propres de A,
— $Z_i \sim \mathcal{N}(0, 1)$ indépendantes.

Si A est une **matrice de projection orthogonale**, c'est-à-dire $A = A^2 = A^T$, de rang r, alors :

$$Q = X^T AX \sim \chi_r^2$$

2. Cas non centré :

Si :

$$X \sim \mathcal{N}_n(\mu, I_n)$$

alors :

$$Q = X^T AX \sim \sum_{i=1}^{n} \lambda_i Z_i^2$$

avec :

$$Z_i \sim \mathcal{N}(m_i, 1)$$

où les m_i dépendent de μ et des vecteurs propres de A.

Dans le cas où A est une matrice de projection, Q suit une **loi du chi-deux non centrée**.

3. Cas général :

Si :

$$X \sim \mathcal{N}_n(\mu, \Sigma)$$

alors la loi de :

$$Q = X^T A X$$

est plus complexe. On fait appel à des résultats avancés sur la distribution des formes quadratiques gaussiennes, ou bien à des méthodes numériques ou asymptotiques.

2.1.13 Application : le test du Khi-deux :

Dans les tests d'indépendance (comme ceux utilisés dans les tables de contingence), la statistique de test peut s'écrire comme une forme quadratique :

$$Q = (O - E)^T V^{-1} (O - E)$$

où :

— O est le vecteur des fréquences *observées*,
— E est le vecteur des fréquences *théoriques*,
— V est la matrice de variance-covariance (parfois diagonale).

Alors :

$$Q \sim \chi^2_k$$

où k est le nombre de degrés de liberté.

2.2 Modèle Linéaire :

Le modèle linéaire est un outil fondamental en statistique. Il permet d'expliquer un phénomène aléatoire à travers une ou plusieurs variables explicatives. Grâce à sa simplicité mathématique et son efficacité pratique, il est largement utilisé dans plusieurs domaines comme l'économie, la biologie, ou encore les sciences sociales.

Dans ce chapitre, nous allons présenter les bases du modèle linéaire, à commencer par le modèle linéaire simple, puis nous aborderons le cas du modèle linéaire multiple. Nous nous appuierons principalement sur des éléments issus de la littérature statistique , en intégrant également des éléments issus de la littérature statistique, notamment du livre de Saporta (2006).

2.2.1 Le modèle linéaire simple :

Le modèle linéaire simple cherche à modéliser la relation entre une variable Y , dite dépendante, et une variable X dite explicative. Il est défini par l'équation :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, ..., n$$

où :

— Y_i est la i-ème valeur observée de la variable dépendante ;
— X_i est la i-ème valeur de la variable explicative ;
— β_0 est l'ordonnée à l'origine ;
— β_1 est le coefficient de régression ;
— ε_i est une erreur aléatoire suivant une loi normale $\mathcal{N}(0, \sigma^2)$.

2.2.2 Hypothèses du modèle linéaire simple :

Afin de garantir la validité des résultats, le modèle repose sur les hypothèses suivantes :

1. **Linéarité** : la relation entre Y et X est linéaire ;
2. **Indépendance des erreurs** : les erreurs ε_i sont indépendantes ;
3. **Homoscédasticité** : la variance des erreurs est constante ($\mathrm{Var}(\varepsilon_i) = \sigma^2$) ;
4. **Normalité des erreurs** : les erreurs suivent une loi normale.

2.2.3 Estimation par la méthode des moindres carrés :

La méthode des moindres carrés ordinaires (MCO) consiste à minimiser la somme des carrés des résidus :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Les solutions de ce problème d'optimisation donnent les estimateurs suivants :

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

2.2.4 Propriétés des Estimateurs :

Sous les hypothèses précédentes, les estimateurs obtenus présentent les propriétés suivantes :

— Ils sont **sans biais** : $\mathbb{E}[\hat{\beta}_j] = \beta_j$;
— Ils sont **efficaces**, au sens du théorème de Gauss-Markov ;
— Leur variance est estimée par :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

2.2.5 Analyse de la variance (ANOVA) :

L'analyse de la variance, couramment désignée par **ANOVA** (Analysis of Variance), est une méthode statistique permettant de comparer les **moyennes** de plusieurs groupes indépendants. Elle repose sur la décomposition de la variance totale d'une variable réponse en différentes composantes, attribuables aux **facteurs explicatifs** et à l'**erreur aléatoire**.

Soit Y_{ij} l'observation de la variable réponse pour le i-ème individu du groupe j, avec $i = 1, \ldots, n_j$ et $j = 1, \ldots, k$. On modélise Y_{ij} par le modèle suivant :

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

où :

— μ est la moyenne générale ;
— α_j est l'effet du j-ème traitement (ou groupe) ;
— $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ est une erreur aléatoire supposée indépendante et identiquement distribuée.

Décomposition de la somme des carrés

L'ANOVA repose sur la décomposition suivante de la somme des carrés totale (SCT) :

$$\text{SCT} = \text{SCE} + \text{SCR}$$

où :
— **SCT** : somme des carrés totale $= \sum_{i,j}(Y_{ij} - \bar{Y})^2$;
— **SCE** : somme des carrés expliquée $= \sum_j n_j(\bar{Y}_{.j} - \bar{Y})^2$;
— **SCR** : somme des carrés résiduelle $= \sum_{i,j}(Y_{ij} - \bar{Y}_{.j})^2$.

Test d'hypothèse

On teste l'hypothèse nulle :

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

contre l'hypothèse alternative :

$$H_1 : \exists j \neq l,\ \alpha_j \neq \alpha_l$$

Le **statistique de Fisher** associé est :

$$F = \frac{\text{SCE}/(k-1)}{\text{SCR}/(N-k)} \sim \mathcal{F}(k-1, N-k)$$

où $N = \sum_{j=1}^{k} n_j$ est la taille totale de l'échantillon.

Ce développement repose sur les fondements théoriques issus de la littérature statistique, notamment l'ouvrage de **Saporta (2006)**.

2.2.6 Le coefficient de Détermination R^2 :

Le coefficient R^2 permet de mesurer la qualité d'ajustement du modèle :

$$R^2 = 1 - \frac{\text{SCE}}{\text{SCT}}$$

Plus R^2 est proche de 1, plus le modèle explique bien la variabilité de Y.

2.2.7 Tests de signification

— **Test de Student** : il permet de tester la significativité individuelle de chaque coefficient ;
— **Test de Fisher** : il évalue la significativité globale du modèle.

2.2.8 Exemple d'Application :

On suppose que le modèle linéaire simple sur un jeu de données simulées portant sur la relation entre le *temps d'étude (en heures)* et la *note obtenue (sur 20)*. Les résultats ont montré une relation positive *significative*, avec un R^2 de 0,78. Ce qui suggère que le temps d'étude explique environ 78 % de la variabilité des notes.

Étudiant	X (heures)	Y (note)
1	2	8
2	4	10
3	6	12
4	8	15
5	10	17
6	12	18
7	14	19
8	16	20

1. Le modèle à ajuster est : $Y = \beta_0 + \beta_1 X + \varepsilon$
2. Calcul des estimateurs par la méthode des moindres carrés :
   On utilise les formules suivantes :
   $$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
   On calcule :
   $$\bar{X} = \frac{2+4+6+8+10+12+14+16}{8} = 9,$$
   $$\bar{Y} = \frac{8+10+12+15+17+18+19+20}{8} \approx 14.875$$
   Après calculs, on trouve :
   $$\hat{\beta}_1 = \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \approx 0{,}92, \quad \hat{\beta}_0 \approx 14{,}875 - 0{,}92 \times 9 = 6{,}63$$
3. Le coefficient de Détermination :
   $$R^2 = 1 - \frac{\text{SCR}}{\text{SCT}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \approx 0{,}975$$
4. **Interprétation** : Le modèle explique environ 97,5% de la variabilité des notes. Le nombre d'heures d'étude a donc une **forte influence** sur la note obtenue.

2.2.9 Le modèle linéaire multiple :

Lorsque plusieurs variables explicatives sont impliquées, on utilise le modèle :

$$Y = X\beta + \varepsilon$$

où :
— Y : vecteur $n \times 1$ des observations ;
— X : matrice $n \times p$ des variables explicatives ;
— β : vecteur $p \times 1$ des paramètres ;
— ε : vecteur des erreurs.
L'estimation des moindres carrés s'écrit :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Diagnostic du modèle :

Il est essentiel de valider les hypothèses à travers des tests comme :
— Test de normalité (Shapiro-Wilk) ;
— Test d'homoscedasticité (Breusch-Pagan) ;
— Test d'autocorrélation (Durbin-Watson) ;
— Analyse des points influents (Cook's distance).

2.2.10 Le modèle linéaire classique :

Soit Y une variable aléatoire, et $(y_1, y_2, \ldots, y_n)$ une suite de n observations indépendantes (n réalisations de Y). On suppose que Y est de moyenne $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$.

Le modèle linéaire classique pose une relation de la forme :

$$\mu_i = \sum_{j=1}^{t} \beta_j X_{ij}$$

où : X^t est le nombre de paramètres.

On suppose qu'il existe une relation linéaire entre la moyenne de Y et un ensemble de t variables explicatives, à travers des paramètres β_j.

Si l'on pose $\mu_i = \mathbb{E}(Y_i)$, alors l'équation précédente se récrit, en désignant par X_{ij} la valeur de la j-ème covariable pour l'observation i :

$$\mu_i = \sum_{j=1}^{t} X_{ij}\beta_j, \quad \text{pour } i = 1, \ldots, n$$

ou, en écriture matricielle :

$$\mu = X\beta$$

Et l'on résume le **modèle linéaire classique (MLC)** en ceci :

Les composantes de Y sont des variables normales indépendantes de variance constante σ^2 et $\mathbb{E}(Y) = \mu = X\beta$.

Généralisation d'un modèle linéaire classique :

Elle consiste à poser un concept unificateur utile quant à l'identification de la structure des données pour permettre une approche flexible à l'ajustement de modèles. Pour cela, on a recours à la définition et l'agencement de **trois (3)** composantes.

Le Modèle Linéaire Généralisé (MLG) :

1) La partie Aléatoire :

Elle consiste en des observations indépendantes $Y_1, \ldots, Y_N$ d'une distribution dans la famille exponentielle naturelle, chaque observation ayant une distribution de la forme :

$$f(y; \theta, \phi) = \exp\left\{\frac{y \cdot \theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \qquad (*)$$

a, b et c étant des fonctions connues, a étant généralement de la forme ϕ/ω_i, où ω_i est un poids connu, souvent n_i, quand Y_i est une moyenne de n_i prélèvements indépendants.

Remarque :

— Cette forme plus générale de distributions est très utile pour les familles à deux paramètres telles que les lois normale, gamma (où ϕ est un paramètre de nuisance).
— Elle l'est aussi pour les familles à un paramètre telles que la binomiale ou la loi de Poisson.
— Le cas $f(y;\theta) = a(\theta)b(y)\exp[y \cdot Q(\theta)]$ est un cas particulier, et l'on appelle $Q(\theta)$ le paramètre naturel.

2) La partie Systématique :
Elle relie des paramètres η_i à des covariables en utilisant un prédicteur linéaire :

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N,$$

ou (matriciellement) :

$$\eta = X \cdot \beta$$

où l'on a posé $\eta = (\eta_1, \ldots, \eta_N)'$ et $\beta = (\beta_1, \ldots, \beta_t)'$ est le vecteur des paramètres du modèle. X est la matrice $N \times t$ du modèle, consistant en des valeurs de variables explicatives (ou facteurs) pour les n observations.

Définition :

Le vecteur η est appelé **prédicteur linéaire**.

3) La fonction lien :
C'est toute fonction g monotone différentiable. Elle relie les deux composantes précédentes : la moyenne μ de Y au prédicteur linéaire par : $\eta_i = g(\mu_i)$.

Ainsi, le MLG est une relation entre les valeurs espérées des observations et les variables explicatives via la formule :

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N$$

Définition :

Un **modèle linéaire généralisé** est un modèle linéaire pour la transformation d'une variable ayant une distribution dans la famille exponentielle.

Remarque :

— Des expressions générales pour les deux premiers moments de Y sont tirées de la forme distributionnelle pour donner :

$$\mathbb{E}(Y) = b'(\theta), \quad \text{et} \quad \mathrm{Var}(Y) = b''(\theta) \cdot a(\phi)$$

comme moyenne et variance de Y respectivement.

— La fonction g pour laquelle $g(\mu) = \theta_i$ dans la fonction de distribution (*) est appelée **fonction lien canonique**. et pour une telle fonction , il y a la relation linéaire directe :

$$\theta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N$$

entre le paramètre naturel et le prédicteur linéaire. et comme μ_i vaut $b'(\theta_i)$, le lien canonique est l'inverse de la fonction b' .

Exemples de fonction lien :

1. **Le Logit (ou Logistique)** : Elle est donnée par

$$g(x) = \text{Logit}(x) = \log\left(\frac{x}{1-x}\right) = \log x - \log(1-x)$$

Elle est utile pour la modélisation des données binaires. Elle a l'avantage d'être très appropriée quand les données proviennent d'études rétrospectives.

2. **La Probit** : Elle est définie par

$$g(x) = \Phi(x)$$

où Φ est la fonction de répartition de la loi normale standard.

3. **Le Log-log complémentaire** : Elle est spécifiée par

$$g(x) = \log(-\log(1-x))$$

Elle est la contrepartie naturelle de la fonction lien Log-log qui est rarement utilisée car non appropriée pour $x < 0.5$, qui est plus fréquemment la région d'intérêt.

2.2.11 Modèles Linéaires Généralisés (MLG) :

Motivation : Limites du modèle linéaire classique :

Le modèle linéaire classique repose sur des hypothèses fortes, notamment la normalité de la variable réponse, la relation linéaire entre la moyenne de la variable réponse et la (les) variable(s) explicative(s), ainsi que l'homoscédasticité des erreurs. Cependant, dans de nombreux cas pratiques, ces hypothèses ne sont pas satisfaites. Par exemple, lorsque la variable réponse est binaire, comptage ou proportion, le modèle linéaire ne s'applique plus directement.

C'est dans ce contexte que les Modèles Linéaires Généralisés (GLM) offrent une extension flexible permettant de modélise une grande variété de données tout en conservant une structure linéaire au niveau du prédicteur.

Définition générale d'un MLG :

Un MLG est caractérisé par trois principales composantes :

— La **distribution** de la variable réponse Y_i, appartenant à la famille exponentielle. Cette famille inclut les distributions classiques telles que la normale, la binomiale, ou la Poisson.

— La **fonction de lien** $g(\cdot)$ qui relie l'espérance de Y_i, notée $\mu_i = \mathbb{E}[Y_i]$, à la combinaison linéaire des variables explicatives :

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

— La **structure linéaire** $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, où $\boldsymbol{\beta}$ est le vecteur des paramètres à estimer.

Cette structure permet de modéliser différentes formes de données en adaptant la distribution et la fonction de lien (adéquates).

Famille exponentielle :

Une variable aléatoire Y appartient à la famille exponentielle si sa densité ou fonction de masse peut s'écrire sous la forme :

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

où θ est le paramètre canonique, ϕ un paramètre de dispersion, et $a(\cdot), b(\cdot), c(\cdot)$ des fonctions spécifiques à la distribution.

Exemples :

— **Cas continu : Loi normale $\mathcal{N}(\mu, \sigma^2)$**
La densité peut s'écrire sous la forme exponentielle avec :

$$\theta = \mu, \quad b(\theta) = \frac{\theta^2}{2}, \quad a(\phi) = \sigma^2, \quad c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

La loi normale est donc un cas particulier de la famille exponentielle.

— **Cas discret : Loi de Poisson de paramètre λ**
La fonction de masse est :

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Elle peut s'écrire sous forme exponentielle avec :

$$\theta = \log(\lambda), \quad b(\theta) = e^\theta, \quad a(\phi) = 1, \quad c(y, \phi) = -\log(y!)$$

Ce qui confirme que la loi de Poisson appartient également à la famille exponentielle.

Soit Y une variable aléatoire à valeurs dans $\mathcal{Y}$, on dit que sa loi appartient à la **famille exponentielle** si sa fonction de densité (ou de masse) s'écrit sous la forme :

$$p(y; \theta, \phi) = \exp\left[\phi\left(y^\top \theta - b(\theta) - c(y)\right) - \frac{1}{2}s(y, \phi)\right]$$

avec :
— θ : le *paramètre naturel*,
— ϕ : le *paramètre de dispersion*,
— $b(\theta)$: la *fonction de normalisation*,
— $c(y)$: une *fonction mesurable de y*,
— $s(y, \phi)$: une *fonction correctrice* (souvent nulle).

Soit $Y = (Y_1, Y_2, \ldots, Y_n)$ un échantillon de variables aléatoires indépendantes suivant une loi de la famille exponentielle. La densité conjointe s'écrit comme un produit :

$$p(y_1, \ldots, y_n; \theta, \phi) = \prod_{i=1}^{n} \exp\left[\phi\left(y_i - b(\theta) - c(y_i)\right) - \frac{1}{2}s(y_i, \phi)\right]$$

Par propriétés de l'exponentielle, ce produit peut s'écrire :

$$p(y; \theta, \phi) = \exp\left[\phi \sum_{i=1}^{n}\left(y_i - b(\theta) - c(y_i)\right) - \frac{1}{2}\sum_{i=1}^{n} s(y_i, \phi)\right]$$

avec :
— θ : le *paramètre naturel*,
— ϕ : le *paramètre de dispersion*,
— $b(\theta)$: la *fonction de normalisation*,
— $c(y_i)$: une *fonction mesurable de* y_i,
— $s(y_i, \phi)$: une *fonction correctrice* (souvent nulle),
— $y = (y_1, \ldots, y_n)$: le vecteur des observations.

Composantes des loi susuelles de la famille exponentielle :

Composant	**Binomiale** (n, p)	**Poisson** (λ)	**Normale** (μ, σ^2)	**Gamma** (α, β)
θ	$\log\left(\frac{p}{1-p}\right)$	$\log(\lambda)$	μ	$-\frac{1}{\beta}$
$b(\theta)$	$n\log(1+e^{\theta})$	e^{θ}	$\frac{1}{2}\theta^2$	$-\log(-\theta)$
$c(y)$	$\log\binom{n}{y}$	$-\log(y!)$	$\frac{y^2}{2}$	$(\alpha-1)\log y - \log\Gamma(\alpha)$
ϕ	1	1	$\frac{1}{\sigma^2}$	α
$s(y, \phi)$	0	0	$\log(2\pi\sigma^2)$	0

Quelques lois usuelles de la famille exponentielle :
— **Normale** : $Y \sim \mathcal{N}(\mu, \sigma^2)$, telles les erreurs de mesure,
— **Binomiale** : adaptée aux données comme des fréquences ou des proportions,
— **Poisson** : adaptée aux données de comptage.

Exemple :

Considérons une étude médicale où l'on observe la réussite (succès = 1, échec = 0) d'un traitement selon la dose administrée. La variable réponse Y_i est binaire.

Un modèle logistique est adapté :

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i$$

avec $\mu_i = \mathbb{E}[Y_i]$ la probabilité de succès.

Interprétation des coefficients :

- β_0 représente le log-odds (logarithme du rapport des chances) lorsque $x = 0$. - β_1 indique l'effet d'une unité supplémentaire de x sur le log-odds.

Remarque :

Les GLM constituent une extension essentielle du modèle linéaire classique, offrant une modélisation adaptée à des données non normales via la famille exponentielle et une fonction de lien. Leur structure unifiée permet de traiter des variables binaires, de comptage, et autres avec rigueur statistique.

2.2.12 Fonctions lien et modèles binaires :

Fonction lien : définition :

La fonction lien $g(\cdot)$ transforme la moyenne $\mu = \mathbb{E}[Y]$ afin d'assurer une relation linéaire avec le prédicteur :

$$g(\mu) = \eta = \mathbf{x}^T \boldsymbol{\beta}$$

Elle doit être monotone et dérivable pour garantir la stabilité et la convergence des estimations.

Modèle logit (logistique) :

Le modèle logit utilise la fonction de lien log-odds :

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Ce modèle est largement utilisé en raison de son interprétation intuitive et de son efficacité dans les classifications binaires.

Modèle probit :

Le modèle probit utilise la fonction inverse de la fonction de répartition de la loi normale standard Φ :

$$g(\mu) = \Phi^{-1}(\mu)$$

Ce modèle repose sur l'hypothèse d'une variable latente normale sous-jacente.

Comparaison logit vs probit :

Les deux modèles fournissent des résultats similaires sur le plan prédictif, mais diffèrent légèrement dans leur formulation et interprétation. Le logit est préféré pour sa simplicité interprétative (log-odds), tandis que le probit est utilisé lorsque les hypothèses de normalité sont justifiées.

Exemple : données de proportions Considérons une campagne de sensibilisation dans différentes régions, où la variable réponse correspond à la **proportion de succès** (par exemple, le taux de personnes convaincues). Le modèle approprié est le **modèle logistique**, qui s'écrit :

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i$$

avec :
— μ_i : la proportion attendue de succès pour l'observation i,
— x_i : la valeur de la variable explicative associée à l'observation i,
— i = 1, ..., n : l'indice des observations (par exemple, les régions),
— β_0, β_1 : les coefficients du modèle à estimer.

Pour chaque observation i, on peut exprimer la proportion attendue μ_i comme :

$$\mu_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Ce modèle permet donc de relier une variable explicative x_i à une proportion de succès μ_i comprise entre 0 et 1.

Chapitre 3

Le Modèle Log-linéaire

3.1 Le Modèle Log-linéaire :

L'analyse des tables de contingence multidimensionnelles par les modèles log-linéaires se propose de faire des inférences sur un ensemble de paramètres décrivant les relations structurelle entre les variables sous-jacentes.

3.1.1 Le Modèle Log-linéaire en *Deux* Dimensions :

Soit un échantillon de taille n d'une distribution multinomiale sur les $N = sr$ cellules d'une $s \times r$-table de contingence (s lignes $\times$r$colonnes$).

Les deux variables sont indépendantes si :

$$\pi_{ij} = \pi_{i.} \cdot \pi_{.j}, \quad i = 1, \ldots, s, \quad j = 1, \ldots, r$$

Désignons par m_{ij} les fréquences (absolues) espérées sous quelque modèle.
Posons $l_{ij} = \log(m_{ij})$.
Le modèle log-linéaire pour ces fréquences s'écrit :

$$l_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad (*)$$

$$o\text{ù} : \mu = \frac{1}{sr} l_{..} = \left\{ \sum_i \sum_j \log(m_{ij}) \right\} / sr = l_{..},$$

$$\lambda_i^X = \frac{1}{r} l_{i.} - \mu = \left\{ \sum_j \log(m_{ij}) \right\} / r = l_{i.} - \mu,$$

$$\lambda_j^Y = \frac{1}{s} l_{.j} - \mu = \left\{ \sum_i \log(m_{ij}) \right\} / s = l_{.j} - \mu,$$

et :

$$\lambda_{ij}^{XY} = l_{ij} - \left(\frac{1}{s} l_{.j} + \frac{1}{r} l_{i.} \right) + \frac{1}{sr} l_{..}) = l_{ij} - l_{i.} - l_{.j} + l_{..},$$

avec les contraintes :

$$\sum_i \lambda_i^X = 0 = \quad \sum_j \lambda_j^Y = \quad \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY}.$$

En faisant l'analogie avec l'ANOVA (Analyse de la Variance), les coefficients $\mu, \lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}$ représentent respectivement :

— une moyenne globale,
— les effets principaux des deux variables (ligne et colonne),
— et l'interaction (effet des deux variables).

Le modèle $(*)$ décrit *parfaitement* tout ensemble de fréquences espérées : c'est le modèle *saturé* en deux dimensions.

Les paramètres λ_{ij}^{XY}, mesurant l'association entre X et Y, reflètent l'éloignement de X et Y de l'hypothèse d'**indépendance**, et satisfont à :

$$\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0.$$

Remarque : Si les données sont des observations d'une loi de Poisson, $Y_i \sim \mathcal{P}(\mu_i)$, alors on a un MLG avec la fonction lien g telle que

$$g(\mu_i) = \theta_i = \log(\mu_i) : \eta_i = \log(\mu_i),$$

ceci donne alors le modèle :
$log(\mu) = X\beta$
Donc, pour l'échantillonnage Poissonien, un MLL est un MLG avec fonction lien canonique, $log(m_{ij})$ étant le paramètre naturel d'une variable de (loi de) Poisson

3.1.2 Le Modèle Log-linéaire en *Trois* Dimensions :

Considérons trois variables X, Y et Z à s, r et q catégories respectivement, avec la distribution des probabilités conjointes π_{ijk} telle que $\sum \pi_{ijk} = 1$.

Soit $\pi_{ijk} = P[X = i, Y = j, Z = k]$ dans la $s \times r \times q$-table de contingence sous-jacente. Il est intéressant d'étudier la relation entre X et Y à un niveau fixé de Z (ou les deux autres cas, à savoir : la relation entre X et Z à un niveau fixé de Y, ou bien encore celle entre Y et Z à un niveau fixé de Y).

Le modèle log-linéaire saturé s'écrit :

$$log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad (**)$$

avec les contraintes : $\sum_i \lambda_i^X = ... = \sum_j \lambda_{ijk}^{XYZ} = 0$.

On considère les modèles log-linéaires (dits hiérarchiques) suivants :

— (X, Y, Z) : Indépendance des 3 variables.

— (XY, Z) (ou (XZ, Y) ou (YZ, X) -avec l'écriture adaptée du MLL) : *dépendance conditionnelle* de deux variables par rapport à la troisième (ici (XY, Z)), soit le modèle : $log(m_{ij}) = \mu + \lambda_i^X + \lambda_{kj}^{YZ} + \lambda_k^Z + \lambda_{ij}^{XY}$
le termeλ_{ij}^{XY} représentant l'association entre X et Y sachant Z, c'est-à-dire que X et Y sont conditionnellement dépendantes sachant Z.

— (XY, YZ) ou (XZ, YZ) ou (XY, XZ) : *indépendance conditionnelle* d'une paire de variables relativement à la troisième, soit le modèle :

$$log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Les deux derniers termes représentant les associations partielles, λ_{ij}^{XY} entre X et Y sachant Z et λ_{jk}^{YZ} (celle) entre Y et Z sachant X. Dans ce cas, X et Z sont conditionnellement indépendantes sachant Z (X et Z ne figurent pas ensemble).

— (XY, XZ, YZ) : Aucune paire n'est conditionnellement indépendante sachant l'autre variable, ce qui s'écrit :
$log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
c'est un cas particulier du modèle saturé où les termes à trois facteurs sont nuls pour tous i, j et k (variant de 1 à s, r et q respectivement) : On dit qu'*il n'y a pas d'interaction d'ordre* 3 entre les trois variables.

— (XYZ) : c'est le modèle le plus **général** pour 3 variables, qui tient compte de l'interaction entre celles-ci (simultanément) :
$log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$

3.1.3 Loi de Poisson :

La variable Y_i suit une loi de Poisson avec paramètre μ_i :

$$Y_i \sim \mathcal{P}(\mu_i)$$

Le lien logarithmique est utilisé :

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Exemple : Table de Contingence 2x2 :

Soit une table présentant la distribution entre exposition et maladie :

	Cas	Non-cas
Exposé	a	b
Non exposé	c	d

où on a noté (comme c'est **généralement** le cas pour une table (2×2)) :
$n_{11} = \text{a}$
$n_{12} = \text{b}$
$n_{21} = \text{c}$
$n_{22} = \text{d}$
Le modèle log-linéaire s'écrit :

$$\log(\mu_{ij}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)} + \lambda_{ij}^{(AB)}$$

3.1.4 Indépendance et Interaction :

- L'**indépendance** entre les deux variables correspond à $\lambda_{ij}^{(AB)} = 0$.
- Une **interaction** est présente si ce terme est non nul, indiquant une ***dépendance*** entre les deux variables.

3.1.5 Estimation :

Estimation par moindres carrés : cas linéaire

Dans le modèle linéaire classique, les coefficients sont estimés par minimisation de la somme des carrés des résidus ; il s'agit de résoudre les équations -dites- normales qui donnent comme solution, le vecteur :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Algorithme IRLS (Iteratively Reweighted Least Squares) :

C'est la méthode des Moindres Carrés pondérés, mais avec des itérations. Elle est de mise dans certaines applications où on a à traiter de grandes tables de contingence croisant des variables avec plusieurs -beaucoup de- modalités, engendrant ainsi des effectifs nuls, et donc des cellules vides, car certains profils réponses -croisement de catégories des variables, peuvent ne pas être observés. Et des méthodes de résolution du système d'équations normales, on privilégie le recours aux méthodes itératives. L'estimation est ainsi effectuée via un algorithme itératif (IRLS), où à chaque étape un problème de moindres carrés pondérés est résolu pour ajuster les coefficients.

Estimation par maximum de Vraisemblance (MV) :

Pour les MLG, les paramètres sont estimés en maximisant la log-vraisemblance associée à la famille exponentielle :

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \text{constante}$$

3.2 Évaluation et qualité de l'ajustement :

3.2.1 Déviance :

La déviance mesure la qualité de l'ajustement du modèle par rapport au modèle saturé :

$$D = 2\left(\ell_{sat} - \ell_{mod}\right)$$

où :

ℓ_{sat} :log-vraisemblance du modèle saturé

ℓ_{mod} :log-vraisemlance du modèle étudié,

Une valeur de **D faible** indique un **bon** ajustement.

Remarque :

Un **modèle saturé** est un modèle qui contient suffisamment de paramètres pour reproduire exactement les données observées. Dans le cadre des modèles log-linéaires, il inclut tous les effets et toutes les interactions possibles entre les variables.

3.2.2 AIC (Akaike Information Criterion)

Critère d'information qui équilibre qualité d'ajustement et complexité du modèle :

$$\text{AIC} = -2\ell + 2k$$

où k est le nombre de paramètres. Plus AIC est **faible**, **meilleur** est le modèle.

3.2.3 Test de Wald et Test du rapport de vraisemblance :

Ces tests statistiques évaluent la significativité, dans le modèle, de certains coefficients ou de l'ensemble.

3.3 Modèles non linéaires et paramètres supplémentaires :

3.3.1 Motivation :

Certains phénomènes ne sont pas bien modélisés par une relation linéaire au niveau du lien, nécessitant des extensions non linéaires.

3.3.2 Modèle d'Indépendance :

Le modèle s'écrit :

$$\log(\mu_{ij}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)}$$

où μ_{ij} est la moyenne (paramètre du modèle) associée à la cellule (i, j).

3.3.3 Hypothèses d'indépendance :

Considérons deux variables catégorielles A et B avec I et J modalités respectivement. La table de contingence donne les effectifs observés n_{ij} pour chaque paire de modalités (i, j).

L'hypothèse d'indépendance signifie que la distribution conjointe se factorise :

$$P(A = i, B = j) = P(A = i)P(B = j)$$

En termes de modèles log-linéaires, cela se traduit par l'absence du terme d'interaction :

$$\lambda_{ij}^{(AB)} = 0, \quad \forall i, j$$

3.3.4 Test d'Indépendance :

On compare la déviance du modèle d'indépendance au modèle saturé (qui ajuste exactement chaque cellule) :

$$D = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}$$

où $\hat{\mu}_{ij}$ sont les effectifs estimés sous le modèle d'indépendance.

La statistique D suit asymptotiquement une loi du chi-deux avec

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

3.3.5 Extension à plus de deux variables :

Pour K variables catégorielles, les modèles log-linéaires incluent des termes d'interaction d'ordre variable. Par exemple, pour trois variables A, B, C :

$$\log(\mu_{ijk}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)} + \lambda_k^{(C)} + \lambda_{ij}^{(AB)} + \lambda_{ik}^{(AC)} + \lambda_{jk}^{(BC)} + \lambda_{ijk}^{(ABC)}$$

3.3.6 Choix du modèle :

Le choix du modèle adéquat repose sur :
— Tests d'hypothèses sur les termes d'interaction.
— Critères d'information (AIC, BIC).
— Interprétabilité.

3.3.7 Estimation des paramètres et interprétation :

L'estimation des paramètres du modèle log-linéaire se fait par maximum de vraisemblance ou par IRLS. Les paramètres λ sont interprétés en termes de log-fréquences relatives.

Les interactions positives indiquent une association forte entre modalités.

3.4 Modèles Multinomiaux :

Définition :

Une variable catégorielle à K modalités peut être modélisée par une loi multinomiale avec paramètres $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, où $\sum_k \pi_k = 1$.

La vraisemblance d'un échantillon $(y_1, \ldots, y_K)$ est :

$$L(\boldsymbol{\pi}) = \frac{n!}{y_1! \cdots y_K!} \prod_{k=1}^{K} \pi_k^{y_k}$$

3.4.1 Modèle linéaire généralisé multinomial :

Le modèle GLM pour une variable multinomiale utilise la fonction logit généralisée :

$$\log \frac{\pi_k}{\pi_K} = \mathbf{x}^T \boldsymbol{\beta}_k, \quad k = 1, \ldots, K-1$$

où π_K est la catégorie de référence.

3.4.2 Estimation :

L'estimation des $\boldsymbol{\beta}_k$ se fait par maximum de vraisemblance, souvent via Newton-Raphson ou autres méthodes numériques.

3.4.3 Applications :

Ce modèle est très utilisé en classification, en analyse des choix, et dans l'étude des variables catégorielles dépendantes.

3.4.4 Remarque :

Les modèles linéaires généralisés constituent un cadre puissant et flexible pour l'analyse statistique des données, en particulier dans le contexte des variables catégorielles et des tables de contingence. Ils permettent de modéliser différentes structures de dépendance et facilitent l'interprétation des interactions entre variables. Leur estimation par maximum de vraisemblance et l'utilisation d'algorithmes efficaces rendent leur application pratique et fiable.

3.5 Modèle log-linéaire

Les modèles log-linéaires sont une extension des modèles linéaires, utilisés principalement pour l'analyse des données qualitatives, notamment dans les tableaux de contingence. Ils permettent d'étudier les relations entre plusieurs variables qualitatives en modélisant le logarithme des effectifs attendus dans les différentes cellules du tableau.

3.5.1 Principe du modèle :

Considérons un tableau de contingence à deux dimensions avec des variables X et Y. Le modèle log-linéaire suppose que le logarithme des effectifs attendus m_{ij} peut s'écrire comme :

$$\log(m_{ij}) = \mu + \alpha_i^{(X)} + \beta_j^{(Y)} + \gamma_{ij}^{(XY)}$$

où :
— μ est une constante (effet moyen global) ;
— $\alpha_i^{(X)}$ est l'effet de la modalité i de la variable X ;
— $\beta_j^{(Y)}$ est l'effet de la modalité j de la variable Y ;
— $\gamma_{ij}^{(XY)}$ représente l'effet d'interaction entre X et Y.
Selon les termes inclus dans le modèle, on peut tester l'indépendance ou l'interdépendance entre les variables.

3.5.2 Interprétation :

— Si $\gamma_{ij}^{(XY)} = 0$ pour tout i, j, le modèle représente l'hypothèse d'indépendance entre X et Y.
— Si les termes d'interaction sont significatifs, cela signifie qu'il existe une dépendance entre les variables.

3.5.3 Utilité du modèle :

Le modèle log-linéaire est particulièrement utile dans :
— L'analyse des données catégorielles multidimensionnelles ;
— La modélisation des interactions complexes dans des tableaux à plus de deux dimensions ;
— Les études de comportement, de santé publique, de marketing, etc.

3.5.4 Exemple :

Considérons un tableau de contingence croisant deux variables qualitatives : le genre (*Homme* ou *Femme*) et la préférence de transport (*Voiture*, *Bus* ou *Vélo*). On a observé :

transport / Genre	**Voiture**	**Bus**	**Vélo**	**Total**
Homme	20	15	15	**50**
Femme	10	25	15	**50**
Total	**30**	**40**	**30**	**100**

3.5.5 Modèle Log-linéaire sans interaction (hypothèse d'*Indépendance*) :

On suppose que les deux variables sont indépendantes. Le modèle log-linéaire s'écrit :

$$\log(m_{ij}) = \mu + \alpha_i^{(\text{Genre})} + \beta_j^{(\text{Transport})}$$

Les effectifs attendus sous cette hypothèse sont calculés par la formule :

$$\hat{m}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

où :
— $n_{i\cdot}$ est le total de la ligne i,
— $n_{\cdot j}$ est le total de la colonne j,
— n est le total général du tableau.

Transport / Sexe	**Voiture**	**Bus**	**Vélo**
Homme	15	20	15
Femme	15	20	15

3.5.6 Le Modèle log-linéaire avec interaction :

Dans ce modèle, on introduit un terme d'interaction entre les deux variables :

$$\log(m_{ij}) = \mu + \alpha_i^{(\text{Genre})} + \beta_j^{(\text{Transport})} + \gamma_{ij}^{(\text{Genre,Transport})}$$

Ce modèle permet d'expliquer parfaitement les effectifs observés.

3.5.7 Test du Khi-deux d'indépendance :

On teste l'hypothèse d'indépendance des deux variables à l'aide du test du χ^2 :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Sexe	**Transport**	n_{ij}	$\hat{m}_{ij}$	$\frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}}$
Homme	Voiture	20	15	1,67
Homme	Bus	15	20	1,25
Homme	Vélo	15	15	0
Femme	Voiture	10	15	1,67
Femme	Bus	25	20	1,25
Femme	Vélo	15	15	0
Total				**5,84**

Le nombre de degrés de liberté est :

$$ddl = (r-1)(c-1) = (2-1)(3-1) = 2$$

À un seuil de significativité $\alpha = 0{,}05$, la valeur critique du χ^2 pour 2 degrés de liberté est environ 5,99.

Conclusion : $\chi^2 = 5{,}84 < 5{,}99$, on ne rejette pas l'hypothèse d'indépendance. Il n'y a pas de preuve statistique d'une association significative entre le genre et le mode de transport.

Ce type de modèle est couramment utilisé en analyse des données catégorielles pour mieux comprendre les relations entre les variables qualitatives.

3.5.8 Remarque :

Le modèle log-linéaire complète efficacement les modèles linéaires classiques en permettant l'analyse rigoureuse des données qualitatives. Il est essentiel dans le cadre de l'analyse des tableaux de contingence, thème central de notre mémoire.

3.6 Modèle log-linéaire pour un tableau 2×2 et Odds-Ratio :

Dans le cas de données catégorielles à deux variables qualitatives binaires, il est courant de résumer l'information par un tableau de contingence 2×2. L'analyse de ce type de tableau peut se faire via un **modèle log-linéaire**, qui permet de modéliser les fréquences attendues et de

tester l'indépendance entre les deux variables. Une mesure importante dans ce contexte est le **rapport des cotes (Odds-Ratio)**, qui fournit une interprétation directe de l'association entre les modalités.

3.6.1 Tableau 2×2 et formulation du modèle :

Considérons deux variables qualitatives A et B, chacune ayant deux modalités. On note le tableau de contingence suivant :

	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

Le modèle log-linéaire saturé pour les fréquences n_{ij} s'écrit sous la forme :

$$\log(n_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

où :

— μ est l'effet global (moyenne),
— λ_i^A est l'effet de la modalité i de la variable A,
— λ_j^B est l'effet de la modalité j de la variable B,
— λ_{ij}^{AB} est l'effet d'interaction entre A et B.

Ce modèle est dit *saturé* car il contient tous les effets principaux ainsi que leur interaction. En l'absence d'interaction (i.e., $\lambda_{ij}^{AB} = 0$), le modèle réduit teste l'**indépendance** entre A et B.

3.6.2 Rapport des cotes (Odds-Ratio) :

Une mesure clé de l'association entre les variables A et B dans un tableau 2×2 est l'**odds-ratio (OR)**, défini par :

$$\text{OR} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Cette quantité permet d'interpréter la direction et l'intensité de la relation entre A et B :

— Si $\text{OR} = 1$, il n'y a pas d'association entre A et B,
— Si $\text{OR} > 1$, il existe une association positive,
— Si $\text{OR} < 1$, l'association est négative.

L'odds ratio est lié directement aux paramètres d'interaction du modèle log-linéaire, puisque :

$$\log(\text{OR}) = \lambda_{11}^{AB} - \lambda_{12}^{AB} - \lambda_{21}^{AB} + \lambda_{22}^{AB}$$

Grâce aux contraintes d'identifiabilité imposées aux effets d'interaction (somme nulle sur lignes et colonnes), cette relation peut souvent être simplifiée, notamment :

$$\log(\text{OR}) = 4 \cdot \lambda_{11}^{AB}$$

Ce lien permet d'interpréter le paramètre d'interaction comme un indicateur d'association logarithmique entre les modalités.

3.6.3 Exemple numérique :

Considérons le tableau suivant :

	B_1	B_2
A_1	40	10
A_2	20	30

L'odds ratio est :

$$\text{OR} = \frac{40 \cdot 30}{10 \cdot 20} = \frac{1200}{200} = 6$$

et donc :

$$\log(\text{OR}) = \log(6) \approx 1{,}792$$

Sous les contraintes usuelles, cela donne :

$$\lambda_{11}^{AB} \approx \frac{1}{4}\log(6) \approx 0{,}448$$

Ce paramètre d'interaction positif confirme que la co-occurrence des modalités A_1 et B_1 est plus fréquente que ce que prévoirait un modèle d'indépendance.

3.7 Modèle loglinéaire - Test de non interaction :

Cet exemple fournit un test de non interaction pour les données de la table suivante. La variable réponse, ici ordinale,

Taille de Portée	**Traitement**	**0**	**1**	**2+**	**Total**
7	A	58	11	5	74
7	B	75	19	7	101
8	A	49	14	10	73
8	B	58	17	8	83
9	A	33	18	15	66
9	B	45	22	10	77
10	A	15	13	15	43
10	B	39	22	18	79
11	A	4	12	17	33
11	B	5	15	8	28

TABLE 3.1 – Données de KASTENBAUM et LAMPHEAR (1959)

C'est le nombre de déplétions dont les trois catégories sont 0, 1 et 2+ (2 ou plus), étudiées en fonction de la Taille de Portée et du Traitement, variables explicatives, ordinale — à cinq catégories — et binaire respectivement ; c'est une table "deux facteurs, une réponse" avec $s = 10$ et $r = 3$. Les dix sous-populations sont induites par le croisement des cinq catégories de la Taille de Portée et les deux types de Traitement.

Définissons π_{i0}, π_{i1} et π_{i2} comme étant les probabilités espérées d'observer 0, 1 ou 2+ déplétions respectivement, et posons, pour $i = 1, \ldots, 10$:

$$l_{i0} = \ln\left(\frac{\pi_{i0}}{\pi_{i2}}\right), \quad l_{i1} = \ln\left(\frac{\pi_{i1}}{\pi_{i2}}\right)$$

Considérons l_{i0} et l_{i1} comme fonctions additives des effets de la Moyenne, de la Taille de Portée et du Traitement. Pour générer les statistiques-test sur les effets des facteurs, posons :

$$\begin{bmatrix} l_0 \\ l_1 \end{bmatrix}_{(20\times 1)} = \begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix}_{(20\times 12)} \times \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}_{(12\times 1)}$$

où pour $j = 0, 1$:

$$l_j = (l_{1j}, l_{2j}, \ldots, l_{10j})' \quad ; \quad \alpha_j = (\mu_j, \alpha_{1j}, \alpha_{2j}, \alpha_{3j}, \alpha_{4j}, \alpha_{5j})'$$

et la matrice suivante est obtenue d'une reparamétrisation du modèle d'analyse de variance usuel pour qu'elle soit de rang complet :

$$L = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}_{(10\times 6)}$$

En termes du modèle linéaire (transformation logarithmique)

$$\mathrm{F}(\pi) = \mathrm{K} \cdot \log(A\pi),$$

avec $A =$ Identité, et sous l'ossature de la méthode MCP, K sera la 20×30-matrice définie par :

$$K = \left[\begin{array}{cccccccccc} 1 & 0 & -1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \hline 0 & 1 & -1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{array}\right]$$

dont les dix premières lignes s'appliquent aux composantes de l_{i0} et les dix dernières à celles de l_{i1}.

Les paramètres estimés sont :

$$\hat{\mu}_0 = 0{,}945 \quad ; \quad \hat{\alpha}_{10} = -0{,}278 \quad ; \quad \hat{\alpha}_{20} = 1{,}415 \quad ; \quad \hat{\alpha}_{30} = 0{,}846$$

$$\hat{\alpha}_{40} = 0{,}195, \quad \text{et} \quad \hat{\alpha}_{50} = -\sum_{1}^{4} \hat{\alpha}_{j0} = -0{,}514;$$

De même :

$$\mu_1 = 0{,}400, \quad \hat{\alpha}_{11} = -0{,}278, \quad \hat{\alpha}_{21} = 0{,}474, \quad \hat{\alpha}_{31} = 0{,}153,$$

$$\hat{\alpha}_{41} = 0{,}072, \quad \hat{\alpha}_{51} = -\sum_{1}^{4} \hat{\alpha}_{j1} = -0{,}401.$$

En considérant l'analyse de $\hat{l}_{i0}$, le paramètre $\hat{\alpha}_{i0}$ estime l'effet du traitement A et son opposé, $\hat{\alpha}_{i1}$, celui de B : les effets des quatre premières tailles de portées sont estimés par $\hat{\alpha}_{20} \ldots \hat{\alpha}_{40}$ et l'effet de la cinquième par $\hat{\alpha}_{50} = -\sum \hat{\alpha}_{i0}$.

La valeur de la somme des carrés de la déviation du modèle est la statistique du test de non interaction qui vaut 3,1269, et elle s'accorde avec la valeur 3,128 trouvée par Berkson (1968 [16]) qui avait réanalysé ces données, la différence étant due à des erreurs d'arrondi.

Cette somme de carrés résiduelle peut être interprétée comme test de non interaction ou comme un test d'ajustement du modèle additif en les logarithmes des probabilités cellulaires, et ayant pour paramètres les effets des *Tailles de Portées* et des *Traitements*.

En adoptant la seconde interprétation, il est raisonnable de faire des tests d'hypothèses concernant ces effets.

Le test d'absence d'effet *Traitement* sur l_{i0} et l_{i1} simultanément se fait par le choix de la matrice 2×12 :

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

qui conduit à :

$$\chi^2 = 6{,}41 \text{ à } 2 \text{ ddl}, \quad \text{et est donc significative à } \alpha = 0{,}01 :$$

il y a un effet traitement significatif.

Le test d'absence d'effet de *Taille de Portée* sur $\hat{l}_{i0}$ et $\hat{l}_{i1}$ est obtenu par le choix de la matrice 8×12.

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

qui donne pour valeur de la statistique-test :

$$\chi^2 = 75{,}32 \text{ à } 8 \text{ ddl},$$

valeur hautement significative aussi.

L'effet linéaire de la taille de portée peut être testé, sous l'hypothèse d'équi-espacement, par le test de l'hypothèse

$$H_0 : -2\alpha_2 - \alpha_3 + \alpha_5 + 2\alpha_6 = 0,$$

équivalente à

$$H_0 : -4\alpha_2 - 3\alpha_3 - 2\alpha_4 - \alpha_5 = 0,$$

qui induit la 2×12 matrice suivante :

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Elle donne pour statistique-test la valeur

$$\chi^2 = 87{,}70 \text{ à } 2 \text{ ddl},$$

qui est aussi une valeur grandement significative.

Remarque :

On peut tester la même hypothèse pour $\hat{l}_{i0}$ et $\hat{l}_{i1}$ séparément, ce qui donne

$$\chi^2_0 = 52{,}17 \quad \text{et} \quad \chi^2_1 = 4{,}674 \quad \text{respectivement, à 1 ddl.}$$

La conclusion de cette analyse est qu'une transformation des proportions observées p_{ij} en les logits empiriques :

$$l^0_{i0} = \log\left(\frac{p_{i0}}{p_{i2}}\right), \quad l^0_{i1} = \log\left(\frac{p_{i1}}{p_{i2}}\right)$$

donne une échelle sur laquelle un modèle additif à effets des *Traitements* et des *Tailles de Portées*, ajuste les données de la table 4 ; ceci est confirmé par l'absence de déviation significative du modèle — la non-interaction.

Les traitements A et B sont significativement différents, ainsi que les tailles de portées ; et le nombre de déplétions varie linéairement avec les tailles de portées.

Conclusion Générale

L'analyse des ***tableaux de contingence*** constitue un outil fondamental dans l'étude des données ***catégorielles*** et demeure un thème qui sucite beaucoup d'approches. L'intérêt, toujours grandissant, du traitement de telles tables réside en ce que c'est le support du recueil de données qualitatives avec *beaucoup* de catégories, induisant un *grand nombre* de profils réponses engendrés par le croisement des diverses modalités/catégories des variables sous étude.

Au cours de ce travail, nous avons exploré différentes méthodes statistiques permettant de modéliser, tester et interpréter les relations entre variables qualitatives à travers la structure des tableaux croisés.

Nous avons commencé par une présentation générale des tableaux de contingence, en mettant l'accent sur les notions d'indépendance, de dépendance ou d'interaction. Ensuite, nous avons abordé les modèles log-linéaires, qui offrent un cadre puissant pour analyser ces tableaux, notamment grâce à leur capacité à décomposer les effets marginaux et les interactions entre les modalités.

Des tests statistiques comme le test du Chi-deux ont été utilisés pour évaluer la qualité d'ajustement des modèles aux données observées. Nous avons également illustré l'importance des hypothèses comme l'indépendance simple, l'indépendance conditionnelle et l'équidistribution dans l'interprétation des résultats.

L'ensemble de cette étude démontre que l'analyse des tableaux de contingence ne se limite pas à une simple lecture de fréquences croisées, mais constitue une véritable démarche de modélisation statistique, utile aussi bien en sciences sociales qu'en biostatistique, marketing ou toute discipline manipulant des variables qualitatives.

Au terme de ce travail, nous espérons avoir contribué à mettre en exergue la richesse des données catégorielles, et avoir permis de mieux comprendre la pertinence des outils statistiques dédiés à cet effet.

Des perspectives intéressantes peuvent être envisagées, notamment l'extension vers des modèles multinomiaux, l'analyse de correspondances ou encore l'intégration de données mixtes (quantitatives et qualitatives).

Bibliographie

[1] Saporta, G. *Probabilités, Analyse des données et Statistique*. Éditions Technip, 2006.

[2] Agresti, A. *Categorical Data Analysis*. Wiley, 2002.

[3] McCullagh, P., & Nelder, J.A. *Generalized Linear Models*. Chapman and Hall, 1989.

[4] Mehiri Mohamed *Analyse De Mesures Répétées de Données Catégorielles*, Thèse de Magister, USTHB, 1996.

[5] Yvonne M. M Bishop, S. E. Fienberg and P. W. Holland, *Discrete Multivariate Analysis*. MIT Press 1975.

[6] J. E. Grizzle, C. F. Starmer and G. G. Koch, *Analysis of Categorical Data by Linear Models*, Biometrics, 25, (489-504).

[7] Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. John Wiley and Sons, New York.

[8] Koehler, K. J., & Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communications in Statistics* .

[9] 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.

[10] 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.