## IV : Hétérogénéité et Segmentation Tarifaire

### I. Le Problème de l'Hétérogénéité et ses Conséquences

### A. Définition et Illustration de l'Hétérogénéité

1. Définition :

L'Hétérogénéité (Heterogeneity) désigne le fait que le niveau de risque varie structurellement entre les assurés, même s'ils appartiennent au même portefeuille. Les individus ne sont pas des répliques exactes du risque moyen $E[\bar{S}]$calculé en C3.

- **Exemple :** Le risque de sinistre pour un jeune conducteur citadin est intrinsèquement plus élevé que celui d'un conducteur expérimenté vivant à la campagne.

**2. Conséquence Mathématique :** L'hétérogénéité est la cause principale de la **surdispersion (Overdispersion)** observée dans les données de fréquence : $\mathrm{Var}[N] > \mathbb{E}[N]$. Une partie de la variance n'est pas aléatoire (stochastique) mais **systématique** (due aux facteurs de risque non modélisés).

### B. Conséquences Économiques et Éthiques

1. **Inéquité Tarifaire :** Si tout le monde paie la même Prime Pure (basée sur le $\mathbb{E}[S]$ global), les **bons risques** (risque faible) subventionnent les **mauvais risques** (risque élevé).
2. **Anti-Sélection (Adverse Selection) :** Face à cette inéquité, les bons risques ont tendance à quitter la compagnie pour trouver un assureur qui leur offre une prime plus basse, laissant l'assureur initial avec un portefeuille de plus mauvais risques. Ceci fait augmenter la prime moyenne globale et mène à une spirale de pertes.

- **Objectif de la Segmentation :** Rendre la tarification **équitable** (chaque assuré paie selon son risque propre) et **économiquement viable** (éviter l'anti-sélection).

Une grande compagnie couvre le risque RC automobile. Deux facteurs influencent la charge des sinistres: la puissance du véhicule (faible-élevée) et l'expérience du conducteur (débutant-experimenté). On suppose que la population assurée est repartie uniformément entre ces catégories (250,000 assures dans chaque catégorie). Les charges moyennes des sinistres en fonction des profils de risque sont données au tableau ci-dessous:

| | experimenté | débutant |
|---|---|---|
| Faible | 100 | 1500 |
| Puissant | 900 | 2500 |

1- Supposons que seules deux compagnies, $C_1$ et $C_2$ disons, opèrent sur le marché et que l'assurance est obligatoire. $C_1$ décide de ne pas différencier le montant des primes (et réclame 1250e a tous les assures).
Quelle est la prime pure réclamée par la compagnie C1

$$E[s] = \frac{100(250.000) + 1500(250.000) + 900(250.000) + 2500(250.000)}{250.000\ (4)} = 1250$$

La seconde compagnie $C_2$ différencie les primes sur base de la puissance du véhicule.
Quelle est les primes pures réclamées par la compagnie C2
La prime réclamée pour les véhicules a faible puissance

$E[S_F] = \frac{100(250.000)+1500(250.000))}{250.000\ (2)} = 800$
La prime réclamée pour les véhicules puissants
$E[S_P] = \frac{900(250.000)+2500(250.000))}{250.000\ (2)} = 1700$

Si l'information est parfaite et que les assurés optent systématiquement pour la compagnie dont le tarif est le plus avantageux, donnez les résultats moyens de $C_1$ et de $C_2$. Comment devrait réagir $C_1$?
la compagnie C2 n'aurait en portefeuille que des assurés dont la puissance de leurs véhicules est faibles.et la compagnie C1 n'aurait en portefeuille que des véhicules est puissants.

L'encaissement global de la compagnie C1= 500.000 (1250) = 625.000.000
le montant de sinistres dédommagés C1 = 500.000 (1700) = 850.000.000
Résultat attendu C1 =625.000.000 -850.000.000 = - 175.000.000
L'encaissement global de la compagnie C2= 500.000 (800) = 400.000.000
le montant de sinistres dédommagés C2 = 500.000 (800) = 400.000.000
Résultat attendu C1 = 400.000.0000 – 400.000.000 = 0
$C_1$ devrait différencier le montant des primes reclamées en utilisant la puissance de vehicule (800 pour les véhicules faible et 1700 pour les vehicules puissants)

2- Supposons désormais que $C_1$ et $C_2$ appliquent un tarif segmenté selon la puissance du véhicule. Si une nouvelle compagnie $C_3$ fait son entrée sur le marché en utilisant l'expérience du conducteur pour différencier les assurés (sans tenir compte de la puissance du véhicule). Quels seront les résultats des trois compagnies? Que se passera-t-il à terme sur le marché?
Les primes réclamées par C1 et C2

La prime réclamée pour les véhicules a faible puissance

$E[S_F] = \frac{100(250.000)+1500(250.000))}{250.000\ (2)} = 800$
La prime réclamée pour les véhicules puissants
$E[S_P] = \frac{900(250.000)+2500(250.000))}{250.000\ (2)} = 1700$
Les primes réclamées par C3
La prime réclamée pour les conducteurs expérimentés
$E[S_E] = \frac{100(250.000)+900(250.000))}{250.000\ (2)} = 500$
La prime réclamée pour les conducteurs débutants
$E[S_D] = \frac{1500(250.000)+2500(250.000))}{250.000\ (2)} = 2000$

la compagnie C3 aurait en portefeuille tous les assurés expérimentés. Et C1 et C2 partagent les conducteurs débutants (vehicules faibles et puissants)
L'encaissement global de la compagnie C3= 500.000 (500) = 250.000.000
le montant de sinistres dédommagés C3 = 500.000 (500) = 250.000.000
Résultat attendu C3 = 250.000.000 - 250.000.000 = 0
L'encaissement global C1 + C2 = 250.000 (800) + 250.000 (1700) = 625.000.000
le montant de sinistres dédommagés C1+ C2 = 250.000 (1500) + 250.000 (2500) = 1.000.000.000
Résultat attendu C1 +C2 = 625.000.000 – 1.000.000.000 = -375.000.000

À terme les 3 compagnies appliquent un tarif segmenté selon : la puissance du véhicule (faible-élevée) et l'expérience du conducteur (débutant-experimenté)

| | experimenté | débutant | |
|---|---|---|---|
| Faible | 100 | 1500 | 800 |
| Puissant | 900 | 2500 | 1700 |
| | 500 | 2000 | 1250 |

La prime réclamée pour les conducteurs expérimentés avec véhicules a faible puissance

$E[S_F] = \frac{100(250.000)+1500(250.000))}{250.000\ (2)} = 100$

La prime réclamée pour les conducteurs expérimentés avec véhicules puissants

$E[S_P] = \frac{900(250.000)+2500(250.000))}{250.000\ (2)} = 900$

les conducteurs débutants avec véhicules a faible puissance

$E[S_E] = \frac{100(250.000)+900(250.000))}{250.000\ (2)} = 1500$

les conducteurs débutants avec véhicules puissants

$$E[S_D] = \frac{1500(250.000) + 2500(250.000))}{250.000\ (2)} = 2500$$

**II. Les Modèles Linéaires Généralisés (GLMs)**

Les GLMs sont l'outil standard en actuariat pour réaliser la segmentation tarifaire, car ils permettent de modéliser la dépendance de la fréquence et de la sévérité vis-à-vis des facteurs de risque (âge, zone, type de véhicule) tout en respectant la nature non-Normale des données (comptage, montants positifs).

**A. Structure du GLM**

Le GLM est une extension du Modèle Linéaire Classique et comprend trois composantes :

1. **La Composante Aléatoire (Random Component) :** Spécifie la distribution de la variable de réponse ($Y$).
   - *Exemples :* Loi de **Poisson** pour la Fréquence ; Loi **Gamma** pour la Sévérité. (Doit appartenir à la famille exponentielle).

2. **Le Prédicteur Linéaire (Linear Predictor) :** Une combinaison linéaire des facteurs de risque ($X_i$).

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

3. **La Fonction de Lien (Link Function) :** Relie l'espérance de la variable de réponse $\mathbb{E}[Y]$ au prédicteur linéaire $\eta$.

$$g(\mathbb{E}[Y]) = \eta$$

- La fonction $g()$ doit être inversible. Elle garantit, par exemple, que les prédictions de fréquence ou de sévérité restent toujours positives.

**B. Le Choix de la Fonction de Lien**

| Distribution $Y$ | Fonction de Lien $g(\mu)$ | Relation $\mu = \mathbb{E}[Y]$ | Utilisé pour... |
|---|---|---|---|
| **Poisson** | Lien Logarithmique : $g(\mu) = \ln(\mu)$ | $\mu = e^\eta$ | **Modéliser la Fréquence** (garantit $\mu > 0$). |
| **Gamma** | Lien Logarithmique : $g(\mu) = \ln(\mu)$ | $\mu = e^\eta$ | **Modéliser la Sévérité** (garantit $\mu > 0$). |

## III. Application des GLMs à la Tarification

La segmentation est réalisée en appliquant deux modèles GLM distincts, un pour la fréquence et un pour la sévérité, pour obtenir une Prime Pure segmentée.

$$\mathbf{P}_{\text{Pure}}(\text{segment}) = \mathbb{E}[\mathbf{N}|\text{facteurs}] \cdot \mathbb{E}[\mathbf{X}|\text{facteurs}]$$

**A. Modélisation de la Fréquence : Régression de Poisson**

Pour la fréquence $N$, on utilise un GLM avec une distribution **Poisson** et un lien **logarithmique**.

**1. Le Modèle :**

$$\ln(\mathbb{E}[N]) = \beta_0 + \beta_1(\text{Âge du Conducteur}) + \beta_2(\text{Zone Urbaine}) + \dots$$

**2. Interprétation des Coefficients :** Le coefficient de risque $\beta_i$ mesure l'impact multiplicatif du facteur $X_i$ sur la fréquence.

- Si $\beta_2 = 0.5$ pour la variable "Zone Urbaine" : $\mathbb{E}[N]_{\text{Urbain}} = e^{0.5} \times \mathbb{E}[N]_{\text{Rural}} \approx 1.65 \times \mathbb{E}[N]_{\text{Rural}}$.
- $\rightarrow$ Un conducteur urbain a $65\%$ de sinistres en plus que le conducteur de référence (rural).

### B. Modélisation de la Sévérité : Régression Log-Gamma

Pour la sévérité $X$, on utilise un GLM avec une distribution **Gamma** (ou Tweedie) et un lien **logarithmique**.

**1. Le Modèle :**

$$\ln(\mathbb{E}[X]) = \gamma_0 + \gamma_1(\text{Type de Véhicule}) + \gamma_2(\text{Ancienneté du Contrat}) + \dots$$

**2. L'Indicateur d'Exposition (Offset) :** Dans la pratique, la régression de Poisson doit tenir compte de la période d'exposition (durée du contrat) $t_i$. On ajoute $\ln(t_i)$ comme **Offset** au prédicteur linéaire.

$$\ln(\mathbb{E}[N/t]) = \eta \iff \ln(\mathbb{E}[N]) = \eta + \ln(t)$$

## Exemple 1 : Modélisation de la Fréquence par Régression de Poisson (GLM)

L'objectif est d'estimer la **Fréquence (nombre de sinistres $N$)** en fonction des caractéristiques du conducteur, qui introduisent l'hétérogénéité. Nous utilisons un GLM avec une fonction de lien logarithmique et une distribution de Poisson (Régression de Poisson).

#### Contexte du Portefeuille (Assurance Auto)

- **Variable à Expliquer (Réponse) :** Nombre de sinistres $N$.
- **Variables Explicatives (Covariables $\mathbf{Z}$) :**
  - **Catégorie d'Âge :** Jeune (J: Moins de 25 ans) ou Senior (S: 25 ans et plus).
  - **Zone de Conduite :** Urbaine (U) ou Rurale (R).

#### Le Modèle GLM

Nous estimons le taux de sinistralité ($\lambda$) par la formule linéaire :

$$\ln(\lambda) = \beta_0 + \beta_{\text{J}} \cdot \mathbb{I}_{\text{Jeune}} + \beta_{\text{U}} \cdot \mathbb{I}_{\text{Urbaine}}$$

Où $\mathbb{I}_{\dots}$ est la fonction indicatrice (vaut 1 si la condition est remplie, 0 sinon).

#### Coefficients Estimés

Les coefficients estimés par l'actuaire sont les suivants : | Coefficient | Description | Valeur Estimée ($\hat{\beta}$) | | :--- | :--- | :--- | | $\hat{\beta}_0$ | Interception (référence : Senior en zone Rurale) | **-2.5** | | $\hat{\beta}_{\text{J}}$ | Effet "Jeune" (par rapport à Senior) | **+0.4** | | $\hat{\beta}_{\text{U}}$ | Effet "Urbain" (par rapport à Rural) | **+0.2** |

#### Calcul de la Fréquence Estimée pour Deux Segments

**Segment A : Senior en Zone Rurale (Catégorie de Référence)**

1. **Application du Modèle :**

$$\ln(\hat{\lambda}_A) = -2.5 + (0.4 \cdot 0) + (0.2 \cdot 0) = -2.5$$

2. **Fréquence Estimée ($\hat{\lambda}_A$) :**

$$\hat{\lambda}_A = e^{-2.5} \approx \mathbf{0.082} \text{ sinistres par an}$$

**Segment B : Jeune en Zone Urbaine (Segment à Haut Risque)**

1. **Application du Modèle :**

$$\ln(\hat{\lambda}_B) = -2.5 + (0.4 \cdot 1) + (0.2 \cdot 1) = -2.5 + 0.4 + 0.2 = -1.9$$

2. **Fréquence Estimée ($\hat{\lambda}_B$) :**

$$\hat{\lambda}_B = e^{-1.9} \approx \mathbf{0.150} \text{ sinistres par an}$$

##### Analyse de l'Hétérogénéité

Le modèle GLM a permis de quantifier l'hétérogénéité :

- Le segment à haut risque (B) a une fréquence attendue ($\hat{\lambda}_B = 0.150$) qui est **1.83 fois supérieure** à la fréquence du segment de référence (A) ($\hat{\lambda}_A = 0.082$).
- C'est cette différence de $\lambda$ qui justifie la segmentation et la différence de prime.

### 2. Exemple 2 : Segmentation Tarifaire Intégrée et Calcul de la Prime Nette (Chargée)

L'objectif est maintenant de combiner la fréquence estimée ($\hat{\lambda}$) avec la sévérité estimée ($\hat{\mu}$) pour calculer la Prime Nette finale.

### Hypothèses de Sévérité et Chargement

1. **Sévérité Estimée ($\hat{\mu}$) :** La modélisation de la sévérité (par une Loi Gamma ou Log-normale) a donné une espérance constante pour tous les segments : $\mathbb{E}[X] = \hat{\mu} = \mathbf{4,000}\ \mathrm{U.M.}$
2. **Volatilité :** L'écart-type de la Perte Agrégée pour le segment B est $\sigma[S_B] = \mathbf{1,200}\ \mathrm{U.M.}$
3. **Chargement de Sécurité :** La compagnie utilise un coefficient de risque $\beta = 2$.

### Calcul pour les Deux Segments

| Étape | Description | Segment A (Senior / Rural) | Segment B (Jeune / Urbain) |
|---|---|---|---|
| **1. Prime Pure ($\mathbb{E}[S]$)** | $\hat{\lambda} \times \hat{\mu}$ | $0.082 \times 4000 = \mathbf{328}\ \mathrm{U.M.}$ | $0.150 \times 4000 = \mathbf{600}\ \mathrm{U.M.}$ |
| **2. Écart-Type ($\sigma[S]$)** | Hypothèse de lissage (pour simplifier) | $\sigma[S_A] = 800\ \mathrm{U.M.}$ | $\sigma[S_B] = 1200\ \mathrm{U.M.}$ |
| **3. Chargement de Sécurité ($C$)** | $\beta \cdot \sigma[S] \quad (\beta = 2)$ | $2 \times 800 = \mathbf{1,600}\ \mathrm{U.M.}$ | $2 \times 1200 = \mathbf{2,400}\ \mathrm{U.M.}$ |
| **4. Prime Nette Chargée** | $\mathbb{E}[S] + C$ | $328 + 1,600 = \mathbf{1,928}\ \mathrm{U.M.}$ | $600 + 2,400 = \mathbf{3,000}\ \mathrm{U.M.}$ |

Ce processus montre comment les GLMs quantifient l'hétérogénéité pour créer une **segmentation tarifaire équitable et efficace** :

- Le segment à haut risque (B) paie une **Prime Pure plus élevée** (600 vs 328) car sa fréquence attendue est plus forte.
- De plus, le segment B paie un **Chargement de Sécurité plus élevé** (2,400 vs 1,600) car sa **volatilité intrinsèque** ($\sigma[S]$) est plus grande.

Le tarif final reflète à la fois le **risque moyen (Prime Pure)** et le **risque extrême (Chargement de Sécurité)**, tous deux segmentés par les variables explicatives.

## IV. Synthèse

**Synthèse :** Les GLMs permettent de passer d'une Prime Pure globale, inéquitable et sujette à l'anti-sélection, à une **structure tarifaire personnalisée**. Cette technique est le cœur de la **tarification *a priori***, où la prime est fixée avant l'événement en fonction des caractéristiques observables du risque.