Université de Montréal

# La régression de Poisson multiniveau généralisée au sein d'un devis longitudinal: un exemple de modélisation du nombre d'arrestations de membres de gangs de rue à Montréal entre 2005 et 2007

par

Amélie Rivest

Département de Sociologie

Faculté des Arts et Sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès Sciences (M.Sc.)
en sociologie

Décembre, 2012

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé:

**La régression de Poisson multiniveau généralisée au sein d'un devis longitudinal: un exemple de modélisation du nombre d'arrestations de membres de gangs de rue à Montréal entre 2005 et 2007**

Présenté par :

Amélie Rivest

a été évalué par un jury composé des personnes suivantes :

Anne Calves, président-rapporteur

Éric Lacourse, directeur de recherche

Jean-Pierre Guay, membre du jury

# Résumé

Les données comptées (count data) possèdent des distributions ayant des caractéristiques particulières comme la non-normalité, l'hétérogénéité des variances ainsi qu'un nombre important de zéros. Il est donc nécessaire d'utiliser les modèles appropriés afin d'obtenir des résultats non biaisés. Ce mémoire compare quatre modèles d'analyse pouvant être utilisés pour les données comptées : le modèle de Poisson, le modèle binomial négatif, le modèle de Poisson avec inflation du zéro et le modèle binomial négatif avec inflation du zéro. À des fins de comparaisons, la prédiction de la proportion du zéro, la confirmation ou l'infirmation des différentes hypothèses ainsi que la prédiction des moyennes furent utilisées afin de déterminer l'adéquation des différents modèles. Pour ce faire, le nombre d'arrestations des membres de gangs de rue sur le territoire de Montréal fut utilisé pour la période de 2005 à 2007. L'échantillon est composé de 470 hommes, âgés de 18 à 59 ans. Au terme des analyses, le modèle le plus adéquat est le modèle binomial négatif puisque celui-ci produit des résultats significatifs, s'adapte bien aux données observées et produit une proportion de zéro très similaire à celle observée.



**Mots-clés** : Données comptées, Analyse multiniveaux longitudinale, gang de rue, loi de Poisson, loi binomiale négative et modèles avec inflation du zéro

# Abstract

Count data have distributions with specific characteristics such as non-normality, heterogeneity of variances and a large number of zeros. It is necessary to use appropriate models to obtain unbiased results. This memoir compares four models of analysis that can be used for count data: the Poisson model, the negative binomial model, the Poisson model with zero inflation and the negative binomial model with zero inflation. For purposes of comparison, the prediction of the proportion of zero, the confirmation or refutation of the various assumptions and the prediction of average number of arrrests were used to determine the adequacy of the different models. To do this, the number of arrests of members of street gangs in the Montreal area was used for the period 2005 to 2007. The sample consisted of 470 men, aged 18 to 59 years. After the analysis, the most suitable model is the negative binomial model since it produced significant results, adapts well to the observed data and produces a zero proportion very similar to that observed.


**Keywords** : Count data, longitudinal multilevel analysis, street gang, Poisson law, negative binomial law, zero-inflated models

# Table des matières

# Liste des tableaux

# Liste des figures

# Liste des sigles

NB : Modèle binomial négatif (*Binomial Negative*)

GML : Modèle linéaire généralisé *(Generalized Linear Model)*

GMML : Modèle mixte linéaire généralisé *(Generalized Mixed Linear Model)*

P : Modèle de Poisson (*Poisson Model*)

ZINB : Modèle Binomial Négatif avec inflation du zéro (*Zero inflated binomial negative*)

ZIP : Modèle de Poisson avec inflation du zéro (*Zero inflated Poisson*)

Le calcul des probabilités s'applique également aux choses de toutes espèces, morales ou physiques, et ne dépend aucunement de leur nature, pourvu que dans chaque cas, l'observation fournisse les données numériques nécessaires à ses applications. –*Siméon Denis Poisson*

# Remerciements

Je voudrais tout d'abord remercier mon directeur de maitrise, Éric Lacourse. Ses précieux commentaires et sa confiance m'ont permis de pondre ce mémoire en méthodologie et statistiques.

Un immense merci mon collègue Stéphane Paquin pour ses idées et ses conseils plus utiles les uns que les autres.

Un deuxième immense merci à Alain Girard, statisticien au GRIP, pour ses habiletés techniques dans les approches multiniveaux utilisant la distribution de Poisson.

Un troisième merci aux chercheurs de la section de Recherche et de Planification du SPVM soit : Mathieu Charest, Rémi Boivin et Maurizio D'élia sous la supervision de Michelle Côté. Mon stage fût le commencement de toute cette belle aventure.

Merci également à Émilie pour nos diners et nos soirées à discuter de statistiques et de la vie en général. Ces deux années furent un pur plaisir à tes côtés.

Merci à tous mes amis pour leur encouragement constant me donnant ainsi la volonté de mener à terme ce mémoire.

Merci aussi à Marc-A pour ses conseils de design et son support en fin de parcours.

Pour terminer, un remerciement n'est pas assez pour souligner l'aide de mes parents sur tous les plans de ma réussite actuelle. Sans eux, je ne serais pas où je suis aujourd'hui.

Du fond du cœur, merci à tous!

# Introduction- Les statistiques et la sociologie

Depuis les études de Durkheim, la statistique en sociologie a évolué d'une manière phénoménale. En effet, dans les dernières décennies, de plus en plus de sociologues ont recours à des méthodes diverses en statistiques afin d'analyser concrètement des phénomènes sociétaux. Certes, la statistique n'est qu'un outil de plus pour les sociologues, mais combien pertinent pour étudier des phénomènes sociaux à petite ou plus grande échelle. Les données peuvent être recueillies soit à l'aide de questionnaires, d'entretiens, de statistiques officielles entre autres, mais l'enjeu principal est, et sera toujours, la bonne utilisation des données et si la compilation est relativement simple, l'analyse statistique nécessite des compétences particulières. Il est possible de travailler sur une panoplie de sujets possibles en sociologie à condition de bien choisir sa variable dépendante et par le fait même, les bonnes analyses statistiques s'y rattachant. Chaque méthode, chaque variable, chaque type de distributions possèdent ses propres particularités et il est impossible de passer outre ces spécificités pour mener une analyse statistique sociologique poussée avec le moins de biais possible permettant ainsi de généraliser les résultats de l'échantillon à la population. Dans le cadre de ce mémoire, nous nous concentrerons sur un phénomène à petite échelle, soit les gangs de rue et plus spécialement encore, le nombre d'arrestations de ses membres à travers le temps. L'utilisation de cette variable dépendante ne permet pas d'employer les techniques statistiques plus conventionnels comme la régression linéaire par moindres carrés. Des analyses plus poussées suivant la loi des phénomènes rares doivent donc être mobilisées.

Il n'y a donc pas de meilleures méthodes pour se familiariser avec des techniques statistiques avancées que de débuter par les définir autant au niveau théorique que statistiques. Dans le premier chapitre, nous débuterons avec un volet historique menant aux études sur les données comptées. Par la suite, la définition d'une donnée comptée ainsi que les origines du premier modèle d'application sur ce type de données, soit la Loi de Poisson, sera définie dans le chapitre deux. À travers les années, plusieurs perfectionnements se sont effectués, basés sous la loi de Poisson, afin d'obtenir des résultats qui s'ajustent plus

fidèlement aux données observées. Le chapitre trois traitera des modèles linéaires généralisés, plus spécialement le modèle binomial négatif ainsi que les modèles avec inflation du zéro. Dans le chapitre quatre, une analyse des définitions et du contexte historique du sujet d'étude, soit les gangs de rues est exposée. Une fois les modèles bien définis, la problématique sera développée dans le chapitre cinq et la description des données utilisées se retrouve dans le chapitre six. Dans le chapitre sept, les variables prédictives ont été examinées en fonction de leur signification et de leur influence sur la variable dépendante pour ainsi déterminer ultimement, un modèle parcimonieux qui reflète le plus fidèlement possible les variations observées sur le nombre d'arrestations des membres de gangs de rue. Pour conclure sur la validité des modèles employés, ils seront comparés un à un en fonction de leurs forces et de leurs faiblesses par rapport à leur capacité prédictive dans un but final de sélectionner le meilleur modèle d'analyse lorsque nous sommes confrontés à un phénomène rare.

# Chapitre 1 : L'analyse des évènements

Dans les sciences sociales, plus particulièrement en sociologie, l'intérêt envers l'étude historique des évènements est grandissant depuis son arrivée au début des années 60. L'engouement trouve racine dans la capacité des analyses historiques d'évènements à étudier la dynamique et la nature du phénomène au même moment, dans un contexte social défini, ce que ne permettent pas les études transversales. Ces dernières se définissent comme étant l'étude du changement par rapport à deux points dans le temps, l'avant et l'après. Il s'agit d'établir une association entre les variables et non de déterminer s'il existe une relation causale. L'étude historique des évènements se défini quant à elle comme un schéma d'occurrence et de corrélations dans un temps donné (Yamaguchi, 1991) ou plus précisément, selon Ulrich Mayer and Brandon Tuma (1990) comme étant des méthodes pour examiner les changements à travers différents états à l'intérieur d'un intervalle de temps basé sur une séquence temporelle complète pour un échantillon sélectionné. Il s'agit de mesurer le changement entre les états pour un individu à chaque temps de mesure. Les précurseurs dans l'utilisation de l'analyse historique des évènements se concentrent principalement en bio-statistiques, dans le contrôle de qualité en ingénierie et dans la méthodologie de recherche avec l'estimation des meilleurs modèles à l'aide des paramètres statistiques. Cependant, depuis les quatre dernières décennies, ces méthodes d'analyse sont de plus en plus appliquées en sociologie et en économie. Ce genre d'étude laisse une possibilité infinie aux chercheurs. Par exemple, les démographes peuvent étudier le nombre de naissance dans la dernière décennie, les sociologues peuvent étudier le taux de chômage des pays industrialisés dans les vingt dernières années, les criminologues peuvent étudier le nombre de crimes commis par les résidents de Montréal dans les cinq dernières années etc. L'analyse historique des évènements exploite toute l'information disponible dans une base de données en analysant les évènements par rapport à des covariables qui peuvent affecter la nature du changement à travers le temps. Certaines méthodes utilisent également une analyse descriptive du changement afin de dresser un portrait global du phénomène. D'autres méthodes orientent l'analyse du changement en comparant deux ou plusieurs

groupes. Certains chercheurs se servent alors de l'analyse historique des évènements pour estimer des paramètres visant à être inférés à la population à l'aide de divers modèles statistiques. Par exemple, ils utilisent les processus aléatoires à travers le temps dans la théorie des probabilités afin de déterminer les chances d'apparition d'un phénomène.

Dans l'analyse historique des évènements, l'occurrence de l'évènement se traduit comme étant un taux par rapport à un évènement dans un temps donné. Celui-ci varie en fonction du temps et des individus étudiés. Il peut s'agir d'un phénomène plus rare ou plus courant, seulement les modèles statistiques d'analyse vont différés. Le taux se qualifie de deux façons : un taux de risque ou un taux de transition. Il n'y a pas de différences fondamentales entre les deux expressions, il s'agit seulement d'une question de terminologie. Le taux de risque est davantage utilisé dans les milieux médicaux comme un nombre de décès par rapport à une maladie alors que le taux de transition est davantage utilisé dans les sciences sociales pour compter un phénomène précis comme un taux de chômage ou de divorce. Mathématiquement, ces taux mesurent la probabilité par unité de temps que l'évènement ou le changement dans l'état surviennent dans une période de risque, soit la période d'étude en question (Mayer et Tuma, 1990). Les taux sont toujours positifs et peuvent être supérieurs à 1. Les modèles utilisant ces taux de transition ou de risque postulent que diverses covariables influent à la baisse ou à la hausse, ou un mixte des deux, l'occurrence d'un phénomène à travers le temps.

Il y a plusieurs avantages à utiliser l'analyse historique des évènements versus les études transversales. Premièrement, l'analyse historique des évènements tient compte au même moment du changement en soi ainsi que de la dynamique de celui-ci. Il n'est pas nécessaire alors, contrairement aux études transversales, de superposer plusieurs analyses pour tirer des conclusions quant à la nature et à la dynamique du changement. Deuxièmement, l'étude des évènements est plus informative que les études transversales et permet de déterminer la continuité du phénomène et non seulement, si le changement se produit ou non. Elle peut fournir des informations sur la période antécédente à l'étude et permettre ainsi une meilleure exploration des données. Également, elle maximise le

pronostic de la puissance des modèles statistiques. Troisièmement, une des principales forces de ce genre d'étude est la capacité de déterminer de quelle façon le phénomène change à travers le temps. Celui-ci est envisagé sur une période continue permettant de mettre l'emphase sur la forme du changement et de déterminer des périodes où celui-ci est plus marqué ou plus stable. En résumé, il est incontestable que la grande force de ce type d'analyse est sa capacité à représenter le changement de façon continue et observer les moindres changements à des points précis dans le temps (Blossfeld, Hamerle, & Mayer, 1989).

Par contre, le principal désavantage de l'utilisation de ce type d'analyse se rapporte aux temps nécessaire pour observer et compiler des observations. Il s'agit d'un type d'analyse couteux en temps et en argent. Les informations ne peuvent être compilées annuellement par exemple sans laisser l'année s'écouler. Le chercheur est carrément à la merci du temps, impossible d'aller plus vite. Afin de contrecarrer l'effet du temps, plusieurs chercheurs vont effectuer leurs études de façon rétrospective afin de diminuer légèrement celui-ci sur la compilation des données. Par contre, les critiques fusent rapidement dans la communauté scientifique lorsque le phénomène étudié est trop lointain. Également, lorsque celui-ci est étudié après coup et centrer sur un élément précis, il devient rapidement obsolète. Le deuxième désavantage reproché à l'étude historique de l'évènement est la grande quantité d'informations que génère une analyse longitudinale. Il est facile pour un chercheur de s'égarer dans les questions de recherche possibles ainsi que dans les théories pouvant être mobilisées afin de répondre adéquatement à une question de recherche. À notre avis, ce désavantage peut constituer également un avantage. Une panoplie d'informations peut permettre de répondre à une panoplie de questions de recherche, ou du moins, tenter d'y répondre en partie. Il suffit de bien sélectionner les théories et les questions de recherche en mettant l'accent sur les variables réellement utiles. Le troisième désavantage qui limite généralement l'utilisation de ce genre d'analyse concerne directement les chercheurs. Ceux-ci se sentent souvent abasourdis devant une méthode d'analyse dynamique à travers le temps. La structure des données peut s'avérer rapidement complexe avec les paramètres fixes et aléatoires. De plus, ce genre d'analyse est souvent méconnu ou mal utilisé dans la communauté scientifique.

### *1.1 L'étude des trajectoires de vie*

L'étude des trajectoires de vie est également de plus en plus utilisé dans les études sociologiques même si à la base, il s'agit d'évènements survenant au niveau individuel. Ce paradigme interdisciplinaire a émergé dans les dernières années suite aux avancés techniques en méthodologie et statistiques. L'objectif principal de ce genre d'analyse est la représentation des processus sociétaux analysés au niveau de l'individu (Blossfeld et al., 1989). Celui-ci est alors placé au cœur des phénomènes sociaux qui le régissent et l'approche des trajectoires de vie suggère de le conceptualiser dans ses dimensions temporelles et contextuelles. Ce grand objectif se subdivise également en deux. Premièrement, l'étude des trajectoires de vie vise à expliquer les évènements qui surviennent individuellement dans des schémas sociaux collectifs à l'intérieur d'une conceptualisation et d'un cadre empirique commun. Deuxièmement, elle vise à représenter les processus sociaux qui génèrent ces évènements et ces trajectoires. L'observation de ce type de données implique de cartographier les mouvements successifs de cohortes étudiés à travers un certain type d'évènement au niveau individuel tout en tenant compte de l'environnement, c'est-à-dire de l'incubateur dans laquelle émerge le phénomène (Blossfeld et al., 1989). Il ne s'agit pas seulement d'effectuer une meilleure description des processus qui guident ou façonne les comportements individuels, mais aussi de les relier entre eux (Mayer et Tuma, 1990). Les évènements sont donc expliqués, à travers quelques concepts, mais plus spécialement par rapport aux croyances culturelles à propos de la biographie de l'individu, de ses séquences institutionnelles, de ses rôles et de ses positions dans la société, du cadre légal régit par son âge ainsi que par sa capacité subjective.

L'étude des trajectoires de vie est sans aucun doute un produit des structures et des forces sociales existantes dans notre société, mais l'influence de la société envers l'individu n'est pas exhaustive. Les variations dans la vie «normale» d'un individu sont inévitables comme la naissance d'un enfant, le décès d'un proche, une maladie etc. Ils reflètent les processus sociaux systématiques qui sont plus ou moins prévisibles dans un niveau agrégé. De plus, les variations dans les institutions n'affectent pas uniquement la vie des individus,

mais peuvent, lorsque les variations sont présentes en grand nombre et de façon importante, générer une nouvelle structure sociale et des nouvelles institutions. Les forces sociales en causes ne sont pas uniquement des retombées des institutions sociales vers l'individu, mais également une influence des interactions individuelles vers l'individu, ce qui produit le même effet, soit des nouvelles structures sociales et institutions. L'étude des trajectoires de vie permet donc de saisir au même moment l'influence du contexte social et de l'individu à l'aide d'un phénomène agrégé au niveau individuel choisi et étudier par un chercheur.

L'étude des trajectoires de vie fut caractérisée d'innovatrice, car ce type d'analyse a créé une rupture définie des barrières imposées entre l'étude du macro et du micro dans les anciennes méthodes d'analyse du changement (Blossfeld et al., 1989). Cette méthode d'analyse a également permis de transcender les écoles théoriques ainsi que les disciplines pour former un discours relativement commun entre la microéconomie, la démographique et la sociologie.

### *1.1.1 Le noyau théorique sous-jacent*

Il existe huit assertions à l'intérieur du noyau théorique de l'étude des trajectoires de vie. Elles seront dénombrées brièvement afin de saisir l'essence derrière les analyses effectuées dans les chapitres suivants (paraphrasé de Ulrich Mayer and Brandon Tuma (1990)). Premièrement, les structures sociales sont conçues comme des éléments inter-reliés dans un temps indéterminé. Elles ne sont pas considérées comme étant adjacentes l'une à l'autre, mais bien nichées l'une dans l'autre. Deuxièmement, les éléments de la structure sociale sont un produit à la fois des actions individuelles, des processus organisationnels, de la force des institutions et du contexte historique. Troisièmement, l'histoire récente a favorisé l'émergence de ce type d'analyse grâce aux découvertes statistiques permettant de mettre en lumière les différences par rapport à l'individu lui-même à travers le temps ainsi qu'à son groupe d'appartenance. Également, les actions individuelles sont gouvernées par divers domaines institutionnels dans lesquels les individus gravitent tout au long de leurs vies passant d'un à l'autre à tout moment. Quatrièmement, les trajectoires de vie doivent

être considérées à l'intérieur du contexte propre à la cohorte de naissances des individus. La compétition pour l'accès à l'emploi ainsi que les caractéristiques propres à chaque cohorte sont des variables importantes pour la forme des trajectoires. Cinquièmement, l'étude des trajectoires de vie est plus représentative de la réalité avec l'utilisation des analyses multiniveaux longitudinales. Les différents niveaux permettent de relier les individus entre eux par rapport à leur développement individuel, aux institutions formelles, à la cohorte, au groupe ethnique, à la localité, au pays, etc. Le temps permet de les relier à leur âge de naissance, à l'âge de leurs institutions, au contexte historique et aux changements propres à leur époque ayant des conséquences importantes comme un changement dans une loi ou une catastrophe naturelle. Sixièmement, la trajectoire de vie ne doit pas être pensée en terme d'âge individuel. Ce type d'étude sous-entend que la durée dans une position ou d'une situation est plus importante que l'âge chronologique. En fait, les mécanismes psychologiques et biologiques sont indépendants des mécanismes qui régissent les trajectoires de vie. Un même évènement peut modifier la trajectoire de vie d'un individu à 17 ans ou 30 ans de façon tout à fait similaire. Septièmement, un évènement dans une sphère de la vie ne peut être analysé séparément des autres sphères de l'individu dû à l'interaction entre les différentes institutions et processus sociaux. Un évènement dans le domaine de l'emploi à nécessairement des répercussions sur la vie familiale, l'aspect monétaire, etc. L'étude des trajectoires a comme fondement la proposition qu'un évènement ne peut être analysé isolément. Les évènements et les phases constituent un système causal endogène régi par lui-même. Huitièmement et dernièrement, le timing des évènements de vie a un impact important sur la trajectoire. Un évènement précoce ou tardif n'a aucunement la même importance sur les résultats subséquents, car la durée de l'exposition au phénomène est déterminante dans certains cas.

### *1.2 La différence entre trajectoires de vie et analyse historique des évènements*

Semblables à première vue, l'analyse historique des évènements et les trajectoires de vie doivent être considérés comme une intersection entre deux lignes de recherche. Bien

que les deux types d'analyses soient un ensemble de modèles et de méthodes statistiques, l'analyse historique des évènements oriente la résolution de problème indépendamment des trajectoires de vie. En fait, la dernière découle naturellement de la première. S'appliquant à des sujets divers, l'analyse historique des évènements est particulièrement appropriée dans la dynamique de la recherche longitudinale, à l'aide des trajectoires. En effet, cette dernière semble être le résultat d'un progrès statistique issu de l'analyse historique des évènements permettant d'étudier différents temps de mesure simultanément. En résumé, l'utilisation des analyses historiques des évènements et des trajectoires de vie offre aux chercheurs en sciences sociales une base solide pour observer les schémas des variations à travers le temps et pour détecter un ordre qui ne semble pas apparent à première vue. Elle aide aussi à démêler l'impact des facteurs nichés dans différents niveaux ainsi que les dimensions propres à l'étude du temps.

Ce cadre théorique sera fort utile pour la réalisation des analyses suivantes portant sur un phénomène individuel, soit le nombre d'arrestations dans le contexte d'un réseau social criminalisé. Par contre, l'utilisation de ce type de données nécessite un cadre de mesure particulier.

# Chapitre 2 : Les données comptées et la loi de Poisson

Pourquoi utiliser des modèles spécifiques pour traiter les données comptées ? Le nombre d'évènements relatifs à un nombre de crimes est, la plupart du temps, très faible pour une majorité d'individus et à l'inverse, il peut être très élevé pour une minorité. Cela implique une distribution fortement asymétrique. Les modèles statistiques pour analyser des données ayant ce type de distribution doivent donc être adaptés en conséquence et ne peuvent suivre les modèles de régression linéaire classique par moindres carrés.

## *2.1 Les données comptées*

L'étude des évènements reliés à la criminalité peut s'observer de deux façons, soit en terme d'occurrence par unité de temps ou soit par un taux. Ce type de données est appelé en anglais, les *count data* traduit en français dans ce mémoire par données comptées. En général, l'observation des données se fait par rapport à une région spatiale ou à un intervalle de temps précis. Par définition, une donnée comptée est une donnée discrète non négative (0 à $\infty$) (Hilbe, 2007). Il s'agit par exemple du nombre d'homicides dans une année à Montréal ou encore le taux de vol par habitant dans la province de Québec. Tous les modèles basés sur des données comptées visent à déterminer les facteurs qui influencent, à la hausse ou à la baisse, le nombre d'évènements ou le taux calculé.

Au niveau statistique, les données comptées ne peuvent être traitées comme des variables continues. En effet, la plupart des tests statistiques couramment utilisés comme la régression par moindres carrés et certains modèles multiniveaux sont basés sur le postulat de la normalité (loi gaussienne) de la distribution des données observées. En guise de rappel, la normalité s'observe avec la forme de la distribution des variables à l'étude (Fox, 1998). Les données prennent généralement la forme d'une cloche symétrique. Au niveau mathématique, une distribution normale contient 68% des valeurs à plus ou moins un écart-

type et 95% des valeurs à plus ou moins deux écarts-types. Également, une donnée comptée est en soi intrinsèquement hétéroscédastique (variance inégale) et asymétrique positive (Hilbe, 2007). Par contre, comme mentionne Atkins and Gallop (2010), beaucoup de chercheurs passent outre ces postulats et compte sur un nombre de cas suffisants pour pallier aux problèmes de non-normalité dans l'étude des évènements. Cependant, l'estimation des erreurs types, sans égard à la taille de l'échantillon, peut s'avérer inexacte et biaiser les tests de signification statistique. Il est alors nécessaire de tenir compte de cette particularité en utilisant, généralement, des modèles statistiques basés sur la loi de Poisson ou également appelés la loi des évènements rares.

## *2.2 La loi de Poisson*

Depuis quelques siècles, les statisticiens ont reconnu les difficultés de travailler avec des données comptées. Ils ont développé des distributions basées sur la loi de Poisson qui est une extension de la loi de Bernoulli avec la probabilité qu'un évènement survienne égal à $p$ et ne survienne pas égal à $(1-p)$ (MacDonald & Lattimore, 2010).

### *2.2.1 Historique*

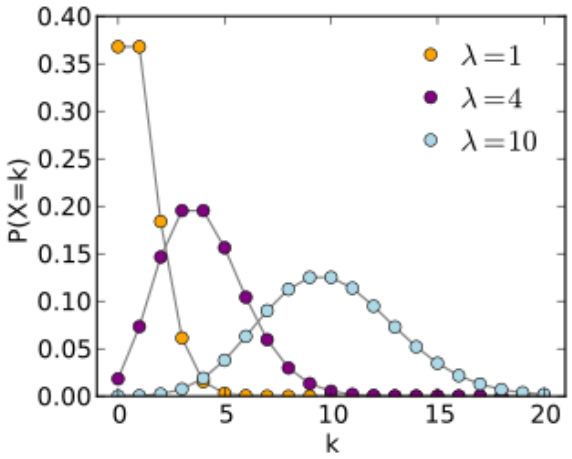
La loi de Poisson fut développée par M. Siméon Denis Poisson (1781-1840). Ses domaines de compétence se rapportent à la géométrie, à la physique, mais celui qui lui vaudra sa notoriété est sans aucun doute les mathématiques. Il a écrit entre 300 et 400 ouvrages, mais le livre le plus marquant est celui sur la distribution de Poisson intitulé *Recherches sur la probabilité des jugements en matière criminelle et matière civile*, publié en 1837. Autrefois réservée à l'étude des phénomènes rares, la loi de Poisson est de plus en plus employée dans le domaine de la télécommunication, dans le contrôle de qualité, la biologie, la météorologie et la finance.

Dans son livre classique, Siméon Poisson démontre que certaines variables aléatoires dénombrent le nombre d'occurrences qui survient dans un temps donné suivant cette équation :

$$\Pr(y|\mu) = \frac{\exp(-\mu_i)\,\mu_i{}^{y}}{y!} \tag{5}$$

où *exp* est l'exponentielle, $y!$ est la factorielle de $y$ et $\mu$ est la moyenne à estimer. Dans la loi de Poisson, il y a un paramètre inconnu soit la moyenne ($\mu$). La forme que prend la distribution est exposée dans le graphique suivant :

**Figure 1: La distribution de Poisson**




Une des particularités de la distribution de Poisson se rapporte à sa moyenne. Plus celle-ci augmente plus la distribution se rapproche de la normalité comme le démontre le graphique précédent. Plus le lambda ($\lambda$) est élevé, plus la distribution tend vers la normalité. Par contre, les évènements étudiés considérés comme des données comptées ont, la plupart du temps, une moyenne faible se caractérisant par une distribution asymétrie avec beaucoup plus de valeurs faibles qu'élevées.

Dans le cadre de ce mémoire, les analyses multiniveaux longitudinales seront utilisées, il est donc nécessaire de définir les équations de l'utilisation de la Loi de Poisson

dans un contexte de données comptées. Tout d'abord, comme les analyses multiniveaux sont à la base une régression, il est nécessaire de définir au préalable les équations de régression pour enchainer avec les équations multiniveaux. En régression, il est assumé que les évènements sont indépendants les uns des autres et leur occurrence est un taux d'incidence constant ($\lambda$) défini comme étant une probabilité instantanée d'un nouvel évènement dans un intervalle de temps. La moyenne prédite en fonction est $y$ et elle est notée ainsi où ($\lambda$) est un lambda, soit le nombre d'occurrences et t est l'intervalle de temps :

$$\mu = \lambda t \tag{6}$$

De plus, lorsque les valeurs observées sont différentes entre les unités ou les sujets ($i$) en fonction des covariables, la moyenne est modélisée de manière log-linéaire. Donc, pour une covariable ($x_i$), un modèle multiplicatif pour obtenir la moyenne prédite, est spécifié ainsi :

$$\lambda_i t = \exp(\beta_1 + \beta_2) = \exp(\beta_1) \exp(\beta_2) \tag{7}$$

Par contre, la régression ordinaire de Poisson est irréaliste à utiliser dans les faits, car le postulat d'indépendance entre les observations ne peut être respecté avec l'utilisation de données comptées. Par exemple, le nombre d'arrestations à un temps subséquent peut dépendre du nombre d'arrestations au temps précédent, mais surtout à un nombre illimité de caractéristiques non observables qui sont stables dans le temps.

Le traitement des données en niveaux permet de distinguer méthodologiquement et statistiquement l'impact des caractéristiques de l'environnement, de l'individu et du temps sur la variable à l'étude. Au niveau du nombre d'arrestations, le département de police, les effectifs, les caractéristiques du quartier, l'usage de la force et du pouvoir par rapport à l'âge des individus (ou le temps) sont tous des exemples d'influence externe ou interne à l'individu pouvant être quantifié par l'analyse multiniveau (Johnson, 2010).

Cette dernière est une méthode qui permet également de travailler avec des données qui possèdent des patrons de variabilités complexes en mettant l'emphase sur les données nichées l'une dans l'autre (Snijders & Bosker, 1999). La grande force de ce type d'analyse consiste à tenir compte simultanément de l'effet du contexte et/ou du temps et du changement individuel. Ce genre de modélisation est possible grâce aux coefficients fixes et aux paramètres aléatoires. Cela permet de tenir compte, du changement intra-individuel et interindividuel (J.D. Singer & B.Willet, 2003). Dans ce mémoire, l'analyse multiniveaux à deux niveaux est utilisée. Le niveau 1 représente le changement d'un individu à travers le temps et le niveau 2 représente le changement d'un individu à l'autre conditionnel à une variable indépendante. Voici l'équation utilisée afin de mener à terme les analyses multiniveaux :

$$\mu_{ij} = E(y_{ij}|x_{ij}, \zeta_{1j}) = \exp\left(\beta_0 + \beta_1(Temps)_{1i} + \beta_2(Violent)_{ij}\right) \quad (8)$$

où $\mu_{ij}$ est la moyenne prédite du nombre d'arrestations, $x_{ij}$ les prédicteurs inclus (le temps et la violence) et $\zeta_{1j}$ le terme d'erreur (la variance) aléatoire. Cette modélisation multiniveaux sera appliquée à tous les modèles présentés dans ce mémoire.

En bref, le principal avantage d'utiliser l'analyse multiniveau longitudinale est la grande flexibilité des modèles permettant de travailler avec des données dépendantes entre elles possédant également une distribution de variance non normale, deux caractéristiques présentes lors de l'utilisation du nombre d'arrestations comme variable dépendante.

*2.2.2 Les postulats statistiques*

Tous modèles statistiques possèdent une série de postulats à respecter, le modèle de Poisson ne fait pas exception. (Atkins & Gallop, 2010; Hilbe, 2007). Premièrement, l'absence du phénomène (soit la valeur zéro) doit être possible dans la distribution. Également, le nombre de zéro ne doit pas être excessif et dépasser la probabilité suggérée

par la Loi de Poisson. Un nombre très important de zéro justifie le choix de modèles modifié en zéro. Les valeurs doivent être des nombres absolus, entiers et positifs. La distribution peut se subdiviser en deux parties bien distinctes, la fréquence de la valeur zéro et les fréquences associées aux autres valeurs (1 à $\infty$). Au niveau des mesures de tendance centrale et de dispersion, la moyenne influence fortement la forme de la distribution de la variable dépendante, plus elle est élevée, plus la distribution tend vers une distribution normale. De plus, l'écart-type doit être égal à la moyenne, si la variance est plus élevée que la moyenne, il y a présence de surdispersion ce qui justifie le modèle d'analyse suivant la distribution binomiale négative exposée dans la partie suivante.

# Chapitre 3 : Les modèles linéaires généralisés

Les études en criminologie et sociologie de la déviance portent principalement sur des données observationnelles. Certains économétriciens ont donc étendu la loi de Poisson aux modèles linéaires généralisés (MacDonald & Lattimore, 2010). Ils seront exposés dans ce chapitre ainsi que les modèles découlant de la loi de Poisson, soit la loi binomiale négative et les distributions avec inflation du zéro notamment le modèle de Poisson avec inflation du zéro (ZIP) ainsi que le modèle binomial négatif avec inflation du zéro (ZINB).

## *3.1 Les modèles linéaires généralisés (GLM)*

Concrètement, les GLM sont l'unification des trois types de régressions (linéaire, logistique et poisson) permettant ainsi à la variable dépendante de posséder une distribution arbitraire et non normale. Les GLM permettent également aux variables d'être reliées par une fonction de lien (voir annexe III) maximisant l'amplitude de la variance à chaque temps de mesure. De plus, la variance des observations peut être fonction de la moyenne de la distribution et non considérée constante comme c'est le cas dans les modèles basés sur le postulat de la normalité.

Les analyses sont fondées sur une distribution issue de la famille des exponentielles incluant les distributions normales, binomiales ainsi que les distributions de Poisson. Les valeurs prédites peuvent être calculées selon l'équation suivante :

$$\mathrm{E}(Y) = \mu = g^{-1}(\mathrm{X}\beta) \tag{6}$$

où $\mathrm{E}(Y)$ est la valeur prédite et $g$ la fonction de lien.

Quant à la variance des GLM, elle est uniquement une fonction de la valeur prédite suivant l'équation suivante :

$$\mathrm{Var}(Y) = V(\mu) = V(g^{-1}(\mathrm{X}\beta)) \tag{7}$$

### *3.2 Les modèles linéaires multiniveaux généralisés (GLMM)*

Suivant les postulats de bases mentionnées précédemment, les modèles linéaires multiniveaux généralisés sont également utilisés pour modéliser des données comptées. La principale distinction entre ces deux types de modèles est la latitude laissée aux paramètres pour l'estimation des coefficients à l'aide du maximum de vraisemblance. En effet, les GLMM possèdent deux éléments clés soit l'indépendance conditionnelle donnant les effets aléatoires ainsi que la distribution de ces derniers. Les GLMM permettent de laisser varier des paramètres de façon aléatoire afin de capturer l'hétérogénéité non-observée.

### *3.3 La loi binomiale négative*

L'utilisation des modèles basés sur la loi de Poisson suppose une variance égale à la moyenne. En réalité, une des propriétés couramment observées dans les données comptées est une variance supérieure à la moyenne. Il existe des individus dans l'échantillon avec un nombre élevé d'occurrences alors que la grande majorité des gens se retrouve à une fréquence observée de 0 ou 1. Par exemple, la plupart des membres de gangs ne possèdent aucune arrestation alors qu'une minorité possède plus de 10 arrestations au cours de la même année. Ces résultats produisent alors une variance plus élevée conditionnellement à une moyenne faible. C'est pourquoi le modèle de Poisson est désadapté dans la plupart des cas. La loi binomiale négative devient alors un incontournable.

La loi binomiale négative est aussi une distribution de probabilité discrète. Cette loi se traduit comme étant une expérience qui consiste en une série de tirages indépendants, donnant une probabilité $p$ de «succès» et une probabilité d'«échec» de $r$ (1-$p$) (Hilbe, 2007*)*. Celle-ci se poursuit jusqu'à l'obtention d'un nombre donné $n$ de «succès». Il en résulte que la variable aléatoire représentant le nombre d'«échecs» suit une loi binomiale négative. L'équation de la loi binomiale négative est la suivante :

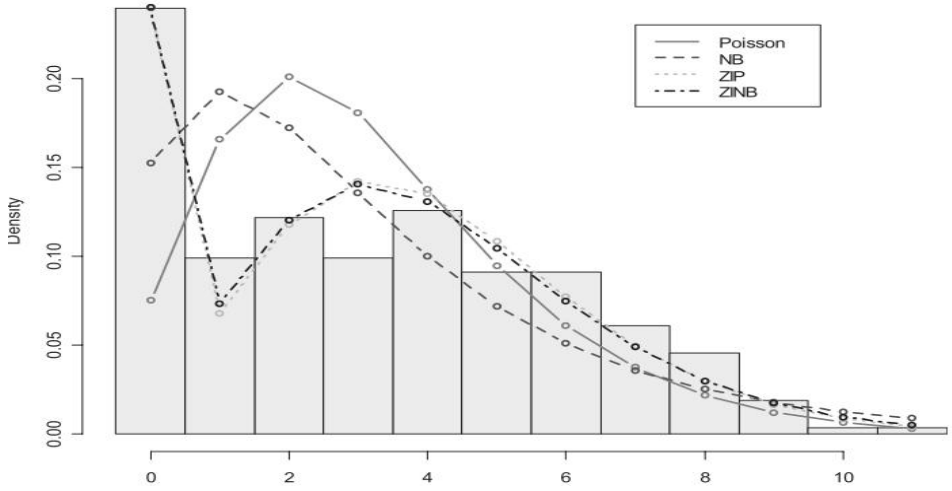
$$\Pr(\gamma_i|u_i) = \frac{\Gamma(\gamma_i + \nu_i)}{\gamma_i!\,\Gamma(\nu_i)}\left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i}\left(\frac{\mu_i}{\nu_i + \mu_i}\right)^{\gamma_i} \quad (8)$$

où $\mu_i$ est la moyenne à estimer et $\nu_i$ est le paramètre de surdispersion. À la différence de la loi de Poisson, il y a deux paramètres à estimer lors de la modélisation soit la moyenne et la surdispersion. L'ajout de ce dernier paramètre supplémentaire permet d'ajuster la variance du modèle indépendamment de la moyenne.

*3.3.1 La forme de la distribution négative binomiale*

Le graphique suivant expose la forme d'une distribution binomiale en comparaison avec la distribution de type ZIP et ZINB qui seront vues dans la section suivante.

**Figure 2 : Les formes de distributions NB, ZIP et ZINB[2]**




Il est possible d'observer que la distribution binomiale négative possède des valeurs plus élevées que la distribution de Poisson. L'asymétrie est moins prononcée avec une moyenne plus élevée, mais toujours avec une variance supérieure, phénomène appelé la surdispersion.

[2] Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43), 21*(4), p.728

*3.3.2 Qu'est-ce que la surdispersion ?*

Tout d'abord, la dispersion se définit comme une mesure quantifiable de la variabilité des données autour d'une valeur centrale. Il existe trois types de dispersion : l'équidispersion, la surdispersion et la sous-dispersion. Théoriquement, dans le modèle de Poisson, le quotient de la variance sur la moyenne est égal à 1, il y a donc équidispersion. La surdispersion s'observe lorsque le quotient est supérieur à 1 et la sous-dispersion s'observe par un quotient inférieur à 1. Avec une variable dite comptée, la surdispersion est un phénomène très courant. Statistiquement parlant, le paramètre associé à la surdispersion se nomme *k*.

*3.3.3 Distinction entre surdispersion réelle et apparente*

En pratique, un quotient supérieur à 1 n'est pas automatiquement signe de surdispersion, car celle-ci peut être seulement apparente et non réelle. Il existe un test plus précis pour déterminer s'il y a réellement présence de surdispersion et dans les deux cas, utiliser les correctifs nécessaires. Tout simplement, une présence réelle de surdispersion est détectable en divisant le Chi-Carré ($\chi^2$) de Pearson par le nombre de degrés de liberté du modèle. Si le quotient est supérieur à 1.25, l'utilisation du modèle binomial négatif doit être envisagée. Par contre, avec un grand échantillon, un quotient de 1.05 est aussi synonyme de présence réelle de surdispersion. Si le quotient se situe entre 1.01 et 1.25, pour un échantillon modéré, il existe quelques correctifs afin de conserver le modèle de Poisson et éviter de traiter la surdispersion. Premièrement, une ou des variables explicatives dans le modèle peuvent être absentes. Il faut bien réfléchir sur les facteurs qui peuvent influencer le phénomène à l'étude. Deuxièmement, il peut y avoir présence de données extrêmes devant être retirées du modèle. Pour ce faire, une analyse des résidus et de la dispersion des données (avec l'écart-type) est simple et efficace. Troisièmement, le nombre d'itérations permettant au modèle de converger est peut-être insuffisant. Il est recommandé d'augmenter la limite permise d'itérations dans le logiciel utilisé. Quatrièmement, un ou des prédicteurs inclus dans le modèle peuvent ne pas être sur une bonne échelle. Une révision des échelles des différents prédicteurs peut s'éviter le traitement de la

surdispersion. Dernièrement, la relation linéaire entre la fonction de lien, les données observées et les prédicteurs est peut-être mal spécifiée ou carrément erronée.

Si, après tous les efforts précédents, la surdispersion persiste, elle est considérée comme réelle. Celle-ci peut être causée de trois façons. Un, il existe peut-être une très grande variation entre les données observées et les données théoriques. Deux, il y a peut-être une grande corrélation positive entre les données et trois, les postulats d'utilisation des modèles ayant une structure de Poisson sont peut-être violés. Par exemple, la variable dépendante contient des décimales ou des nombres négatifs.

La surdispersion, synonyme de grande hétérogénéité de la variance, est souvent causée par un nombre excessif de zéros. Les modèles ZIP et ZINB sont les plus couramment utilisés pour traiter ce phénomène statistique.

### *3.4 Les modèles avec inflation du zéro*

Sous la loi de Poisson, le modèle d'analyse tend à sous-estimer la probabilité d'absence de l'évènement. La loi binomiale négative quant à elle permet de ne pas sous-estimer cette probabilité avec l'inclusion du paramètre de surdispersion, ce qui permet d'augmenter la variance conditionnelle sans modifier la moyenne conditionnelle. L'utilisation de modèles avec inflation du zéro permet de modifier la structure de la moyenne pour modéliser l'occurrence de zéro au sein d'une distribution en générant des processus différents pour l'occurrence et l'absence de l'évènement. Pour utiliser de nouveau l'exemple du nombre d'arrestations, un nombre plus élevé d'arrestations peut être expliqué par l'utilisation de la violence dans la commission des délits alors qu'un nombre plus faible d'arrestations peut être expliqué par l'âge plus avancé d'un individu considérant, en théorie, qu'il commet moins de délits, il est donc moins à risque d'être arrêtés. Les modèles avec inflation du zéro tiennent compte de ces différences à l'aide de l'augmentation de la variance inconditionnelle et la probabilité d'obtenir une donnée comptée égale à zéro.

Le modèle avec inflation de zéro proposé par Mullahy (1986) assume que la population est divisée en deux groupes : les personnes ayant des valeurs zéro et celles ayant des valeurs d'un et plus. Un individu qui est dans le groupe des zéros à un paramètre de probabilité égal à $\psi$ alors qu'un individu dans le groupe des uns et plus à une probabilité égale à $1 - \psi$ où $\psi$ est un paramètre inconnu à estimer. Illustrons avec un exemple portant sur le nombre d'arrestations des membres de gangs de rue. Un membre de gang qui est susceptible d'être arrêté durant l'année devrait être, en théorie, pourrait être classé dans le groupe des uns et plus. À l'inverse, un membre qui est en prison n'est pas à risque d'être arrêté, il se retrouvera impérativement dans le groupe des zéros. Par contre, il est difficile de déterminer concrètement dans lequel des groupes un individu est susceptible de toujours se retrouver considérant que quelques jours de prison sont possibles, que le membre peut être en voyage et non physiquement présent à Montréal ou bien encore il est moins actif au niveau criminel diminuant ainsi ses chances d'être arrêtés. Également, toujours par rapport aux zéros, Muthen (2010) définit deux types précis de zéro : les zéros considérés aléatoires et les zéros considérés structurels. Ces derniers s'observent lorsque l'individu à un score de 0 à tous les temps de mesures alors que les zéros aléatoires s'observent lorsqu'un individu à zéro a l'un ou l'autre des temps de mesures.

Malheureusement, il n'existe pas de critère clair concernant la proportion de zéros acceptable ou non pour justifier l'utilisation de la loi binomiale négative ou les modèles modifiés en zéro. L'article de Yau, Wang, & Lee (2003) propose alors de comparer l'indice d'ajustement, le BIC, des quatre modèles afin de sélectionner le plus adéquat. Les deux modèles les plus couramment utilisés sont exposés dans les sections suivantes.

### *3.4.1 Le modèle de Poisson avec inflation du zéro (ZIP)*

Le modèle de Poisson avec inflation du zéro fût développé par Lambert (1992) et modifié par Greene (1994). L'innovation dans ce modèle est la possibilité pour le paramètre $\psi$ de varier selon les caractéristiques individuelles. Tout comme les modèles avec inflation

du zéro, les données sont générées par deux processus. Les données positives (incluant quelques zéros) suivent l'équation de Poisson (équation 7) alors que le groupe zéro est une fonction des caractéristiques individuelles par le biais du paramètre $\psi$. Celui-ci est déterminé par une fonction *probit* ou *logit* (Gelman & Hill, 2007). Ce dernier est utilisé lorsque la valeur à chaque temps de mesure ($\gamma_i$) représente le nombre de succès dans un certain nombre ($n_i$) d'essais suivant une distribution logistique.

Concernant l'équation permettant de modéliser une distribution ZIP, il s'agit d'introduire une probabilité d'appartenance au groupe présentant des valeurs zéro. Pour un $\gamma > 0$, l'équation est la suivante :

$$\Pr(\gamma_i|u_i) = (1-\psi)\ \frac{\exp(-\mu_i)\,{\mu_i}^{\gamma_i}}{\gamma_i!} \tag{9}$$

La principale différence avec l'équation conventionnelle de la loi de Poisson est l'ajout du paramètre $\psi$ à estimer. Le modèle ZIP contient alors deux paramètres à estimer. Dans le graphique 2, il est possible d'observer que la distribution ZIP possède plus de zéros que la distribution binomiale négative. Par contre, le changement dans la fréquence pour chaque valeur est plus abrupt que la distribution binomiale négative. Il y a davantage de petites valeurs avec une plus grande fréquence.

### *3.4.2 Le modèle binomial négatif avec inflation du zéro (ZINB)*

Le modèle binomial négatif avec inflation du zéro est un modèle modifié en zéro, mais tout comme la loi binomiale négative, la variance peut excéder la moyenne. En fait, le ZINB est un modèle binomial négatif avec l'ajout de la probabilité d'observer la valeur zéro suivant cette équation :

$$\Pr(\gamma_i|u_i) = (1-\psi)\,\frac{\Gamma(\gamma_i + \nu_i)}{\gamma_i!\,\Gamma(\nu_i)}\left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i}\left(\frac{\mu_i}{\nu_i + \mu_i}\right)^{\gamma_i} \quad (10)$$

Dans le modèle ZINB, il y a 3 paramètres à estimer soit la moyenne prédite ($\mu$), le paramètre de surdispersion ($\upsilon$) et la probabilité d'observer des zéros ($\psi$) qui utilise également le *probit* ou le *logit*. Dans le graphique 2, la distribution du ZINB est la plus drastique au niveau de l'asymétrie. En effet, la fréquence de zéros est beaucoup plus élevée que toutes les autres valeurs ce qui résulte une distribution fortement influencée par sa moyenne.

En conclusion, l'implantation de la loi de Poisson est sans aucun doute un outil théorique incontournable pour traiter des données comptées, mais son application est très limitée par rapport à ses différents postulats principalement l'obligation d'égalité de la variance et de la moyenne. C'est dans cette optique que la loi binomiale négative et les modèles avec inflation du zéro furent inventés et appliqués afin de permettre aux chercheurs de traiter adéquatement toutes les particularités propres aux données comptées comme la surdispersion et l'inflation du zéro. Les modèles basés sur la loi de Poisson et la loi binomiale négative possèdent la même structure, mais ce dernier introduit de l'hétérogénéité non observée permettant de travailler avec une variance qui excède la moyenne. Quant aux modèles modifiés en zéro, ils sont souvent utilisés pour leur capacité à générer deux processus distincts, un considérant la probabilité d'observer des zéros et un autre considérant la distribution des valeurs positives. En effet, les facteurs sont souvent différents pour expliquer l'occurrence ou non d'un évènement et les variations dans le compte de cet évènement.

# Chapitre 4 : Le gang de rue

Dans ce quatrième chapitre, nous traiterons brièvement les sujets de l'étude en portant une attention particulière sur la définition d'un gang de rue. De plus, un portrait historique de l'émergence de ce phénomène au Canada et au Québec sera exposé.

## *4.1 Qu'est-ce qu'un gang de rue ?*

Tout d'abord, il n'existe pas de consensus sur ce qu'est un gang de rue ainsi que sur la manière d'identifier les délinquants qui en font partie. Depuis la définition proposée par Trasher (1927), plusieurs dizaines de définitions furent proposées aux différents acteurs qui interagissent avec les gangs de rue. Cependant, une définition semble résister à travers le temps et est la plupart du temps utilisé, aux États-Unis, dans les travaux concernant les gangs de rue. Il s'agit de la définition de Miller (1980) :

> «A youth gang is a self-formed association of peers, bound together by mutual interest, with identifiable leadership, well-developed lines of authority and other organizational features, who act in concert to achieve a specific purpose which generally includes the conduct of illegal activity and control over a particular territory, facility or type of enterprise.»

En 2005, Klein a proposé une définition qui amène un certain consensus chez les différents acteurs : «a street gang is any durable, street-oriented youth group whose own identity includes involvement in illegal activity». Celle-ci étant plus courte et plus générale, il est plus simple d'inclure, sous le phénomène des gangs de rue, une grande variété de gangs à travers les différents pays. Plus près de nous, le Service de Police de la Ville de Montréal (SPVM) s'est doté d'une définition ayant été approuvée par le Renseignement Criminel du Canada en 1991 et est depuis ce temps, la définition de référence dans le paysage québécois. Pour le SPVM, un gang de rue «est un regroupement plus ou moins structuré d'adolescents et de jeunes adultes qui privilégient la force et l'intimidation du groupe pour accomplir des actes criminels, et ce, dans le but d'obtenir pouvoir et reconnaissance ou de contrôler des sphères d'activités lucratives». Une fois la définition sélectionnée, il est primordial de dresser la liste des critères qui permet d'identifier les

membres qui font partie des gangs de rue. Pour ce faire, le SPVM propose six critères[3] permettant de considérer un individu comme étant un membre, autant pour le système de justice que pour les différents corps policiers;

1. Des renseignements provenant d'une source fiable (membre du gang, membre d'un gang rival, source de la collectivité, autorités scolaires, commerçants, citoyens);

2. Un rapport de surveillance de la police confirmant que la personne entretient des rapports avec des membres reconnus du gang;

3. L'aveu de la personne en question;

4. La participation directe ou indirecte de la personne à un crime de gang;

5. Le résultat d'un procès confirmant l'adhésion de la personne à un gang;

6. Des marques d'identification au gang, accomplissement de rituels initiatiques, possession d'articles et de symboles propres au gang : tatouages, armes, vêtements, etc.

Au final, pour qu'un individu soit officiellement considéré comme un membre à part entière d'un gang de rue, il doit impérativement remplir le critère numéro quatre ainsi que deux autres critères.

À première vue, la définition d'un gang ainsi que les différents critères d'identification semblent précis. Dans les faits, il est difficile de prouver hors de tout doute que la commission d'un délit criminel est posée pour le compte du gang. La criminalité est principalement juvénile et plus importante dans les quartiers défavorisés. Dans ceux-ci, plusieurs adolescents se rattachent à un regroupement qui n'est pas toujours criminalisé dans le sens entendu par les divers intervenants. De plus, plusieurs adolescents portent des signes distinctifs sans être réellement impliqués dans un «vrai» gang criminalisé. Bref, la transformation du phénomène des gangs de rue en un sujet d'étude comporte son lot de difficultés.

---

[3] Jean-Pierre Guay et Judith Gaumont, Le phénomène des gangs de rue au Québec, Rapport présenté au Gouvernement du Québec, 2009, p.15

*4.1.1 Un bref portait historique du phénomène des gangs de rue aux États-Unis*

À Montréal, il existe deux grandes bannières sous lesquelles s'identifient plusieurs gangs de rue : les Crips et les Bloods. Celles-ci n'ont pas été fondées à Montréal, ni même au Canada, mais bien aux États-Unis, plus précisément à Los Angeles. Dressons un bref portrait historique des Bloods et des Crips afin de mieux saisir l'influence de ces deux bannières américaines sur les gangs de rue québécois.

Conséquence de l'immigration massive de 1911 à 1970 aux États-Unis, des millions d'immigrants se sont entassés dans les grandes villes créant ainsi des quartiers très défavorisés et surtout multiethniques. La proximité et la cohabitation de diverses ethnies, dans un contexte de pauvreté, mènent inévitablement à des guerres raciales (Howell, 2012). Pour se défendre, plusieurs individus ont décidé de former des gangs afin d'être en mesure de mieux se protéger les uns des autres. Ce qui devait être à la base une source de protection se transforme rapidement en un marché lucratif concentré principalement sur la vente de drogue. Ce phénomène s'observe un peu partout aux États-Unis, mais principalement à New York, à Chicago et à Los Angeles. C'est dans cette dernière que furent créés les deux plus grands consortiums de gang de rue soit les Crips et les Bloods (Delaney, 2006).

Le gang des *Crips* fût fondé vers la fin des années 60 par deux adolescents nommés Stanley Tokkie Williams et Raymond Washington à Los Angeles. Ces derniers étaient à la tête de deux plus petits gangs appelés les *Baby Avenus West* et *East* qui furent fusionnés afin de former les *Cribs*. À travers le temps, les membres furent surnommées *cripples* ce qui mena au nom que l'on connaît actuellement, les Crips. Ce gang ne revendique aucune affiliation politique, ni cause, ni meilleures conditions de vie, il s'agit simplement d'un rassemblement d'adolescents voulant se défendre contre les attaques des blancs. Vers les années 1980, la consommation de «crack» est de plus en plus importante à L.A, les Crips décident donc de s'approprier le marché de cette drogue par le biais des gangs mexicains. Ce trafic permet au *Crips* de prendre de l'expansion et de devenir, l'un des plus grands

gangs de rue des États-Unis. La soudaine richesse des membres leur permet de quitter leur ghetto et de s'installer un peu partout aux États-Unis et au Canada pour conquérir un nouveau marché de trafic de drogue. Une fois arrivées au pays, plus précisément à Montréal et Toronto, les Crips prennent de l'ampleur, toujours avec le trafic de drogue, mais également avec la prostitution. Les membres s'installent principalement dans les quartiers défavorisés de Montréal-Nord et St-Léonard à Montréal, où se concentre une grande population d'origine africaine et haïtienne.

Le gang des *Crips* utilise la couleur bleue comme identité visuelle et sont en majorité des Afro-Américains. Les membres portent des bandanas bleus sur la tête ou au cou, des souliers bleus et des chandails bleus. Outre le style vestimentaire, les membres s'identifient entre eux à l'aide de la main gauche placée en lettre «C» et avec une danse de pied nommé le «C-Walk». De plus, les graffitis et les langages codés, surtout en prison, permettent de reconnaître les membres des Crips.

Également formé à Los Angeles en 1972, les membres des Bloods se sont affiliés pour se défendre contre les Crips qui prenaient de plus en plus d'ampleur dans la rue et dans le marché de la drogue. Le gang des Bloods est une alliance entre plusieurs petits gangs soit les DenverLanes, les Bishops et les L.A. Brim. Toujours en nombre inférieur, les membres des Bloods sont réputés comme étant très violents afin de gagner le respect de la rue et de résister aux Crips. Tout comme les Crips, les membres des Bloods favorisent le marché des drogues dures. La quantité d'informations sur les Bloods étant limitée, il est difficile d'affirmer sans aucun doute que le déplacement des Bloods des États-Unis vers le Canada s'est effectué pour les mêmes raisons que les Crips. Cependant, il est fort possible que la richesse de certains membres leur ait permis de quitter leur ghetto et d'immigrer au Canada.

Les Bloods ont la couleur rouge comme identité visuelle et sont également en majorité des Afro-Américains. Tout comme les Crips, ils portent des bandanas rouges, des chandails rouges et des souliers rouges. Pour ce qui est de la signature gestuelle, les

membres des Bloods placent leurs deux mains ensemble pour former la lettre «B» de coté. Les graffitis et les langages codés sont également utilisés en prison et dans la rue pour faciliter l'identification des membres entre eux.

Il n'existe pas de différences majeures entre les Bloods et les Crips, aucun des deux gangs ne revendique une appartenance politique ou une cause précise. L'appartenance à une ou l'autre de ces bannières dépend en grande partie du quartier de résidence des membres et de l'affiliation de ses pairs ou de sa famille.

### *4.2.1 La formation des gangs de rue au Canada*

La principale cause de l'émergence et de la formation des gangs de rue canadiens se rapporte aux politiques d'immigration du Canada. Comme l'indique le Canadian Police Survey on Youth Gangs effectué en 2002, 82% des membres de gangs de rue proviennent d'une minorité visible de première ou de deuxième génération. La signature du traité de Genève ouvrant les portes du Canada aux réfugiés politiques a permis à une panoplie d'individus de vivre une vie meilleure. Par contre, ces portes toutes grandes ouvertes ont également permis à des individus malintentionnés de venir faire la loi et l'ordre dans les rues du Québec et du Canada. Par exemple, comme le dénote Chettleburgh (2007), un nombre important de réfugiés politiques d'Haïti, sous le joug de Jean-Claude Duvalier, ont formé l'un des plus grands et des plus violents gangs de trafiquants de drogues au Canada. De plus, les politiques d'immigration de Trudeau, Mulroney et Chrétien (trois anciens premiers ministres du Canada) se concentraient sur l'aspect économique favorisant ainsi l'immigration dans les grands centres canadiens comme Montréal, Toronto et Vancouver. Même si les immigrants sont la plupart du temps plus scolarisés que les Canadiens d'origine, la non-reconnaissance de la scolarité une fois arrivée au pays, favorise la pauvreté ou l'emploi précaire des immigrants. Depuis la récession économique de 1991, la moitié des immigrants vivent dans un état de pauvreté. Celui-ci n'est pas un synonyme de criminalité, mais il en est l'une des principales causes.

Aux États-Unis, l'émergence des gangs se concentrait principalement sur le rassemblement des individus en fonction de leur ethnie afin de maximiser leur sécurité par rapport aux autres groupes. Avec le temps, les gangs américains deviennent de plus en plus mixtes et les affiliations aux deux grandes bannières tendent à diminuer. Au Canada, ce phénomène d'ethnie exclusive dans le gang n'est que très peu présent. En effet, beaucoup de gangs canadiens sont hybrides ;

> «this means that, more and more, they are multi-ethnic in composition ; they are involved in every conceivable criminal activity that produces money ; they do not display the typical characteristics of traditional street gangs (such displaying colors, which today attracts unwanted attention) ; and their members display fluid affiliations, sometimes belonging to more than one gang.»[4]

Par contre, les deux grandes bannières ne doivent pas être rejetées d'emblée lors d'études sur les gangs de rue canadiens et québécois. Selon Descormiers (2008), « sur le territoire montréalais, nous assistons à une quasi-reproduction d'une dynamique conflictuelle se résumant aux adversaires Crips versus Bloods. Les gangs montréalais imitent leurs idoles, pairs américains, pionniers West-Coast du mouvement et de la culture sous-jacente au phénomène des gangs[5]».

Pour conclure ce chapitre, le phénomène des gangs de rue à Montréal est surtout situé dans des quartiers défavorisés avec une concentration importante de familles immigrantes. De plus, le taux de criminalité de ces quartiers est aussi plus élevé que la moyenne montréalaise. Ces quartiers sont donc un lieu de prédilection pour le développement des gangs de rue. Ce petit volet historique avait pour but de mettre en lumière le contexte social dans lequel baigne ses membres.

---

[4] Michael C. Chettleburgh, *Young Thugs : Inside the Dangerous World of Canadian Street Gangs*, HarperCollinsPublishersLtd, 2007, p.21

[5] Karine Descormiers, *Le réseau social des gangs montréalais : accès à la dynamique relationnelle par l'entretion de groupe*, Mémoire de maitrise, Université de Montréal, p. 55

# Chapitre 5 : La problématique

Afin de comparer les différents modèles définis préalablement, il est nécessaire de travailler avec une variable dépendante ayant les caractéristiques postulées par les différents modèles : une variance supérieure à la moyenne ainsi qu'un nombre important de zéros. Pour ce faire, le nombre d'arrestations des membres de gangs de rue est un exemple idéal, car, comme il sera noté dans les sections suivantes, cette variable possède une variance supérieure, un nombre très important de zéros et ces deux composantes varient à travers le temps.

De prime abord, il y a trois critères à respecter pour être en mesure d'utiliser des techniques de mesure longitudinale (Judith D. Singer & Willet, 2003). Premièrement, il doit au minimum y avoir trois points de mesures dans le temps pour déterminer s'il y a présence réelle de changement. Même si la base de données comporte initialement huit points de mesures dans le temps, uniquement trois temps de mesure sont employés dans ce mémoire. Le but n'étant pas de dresser un portrait global des risques d'arrestations, mais bien de comparer des modèles statistiques entre eux. Deuxièmement, la variable dépendante doit varier systématiquement à travers le temps. Cette deuxième condition est également respectée, car le nombre d'arrestations peut varier d'une année à l'autre, et ce, pour tous les individus. Troisièmement, l'unité de mesure doit être assez sensible pour capter les variations du phénomène. Comme une arrestation est un phénomène relativement rare, même dans un milieu criminel, l'unité de mesure annuelle n'est ni trop longue, ni trop courte et elle permet de bien saisir les différences intra- et inter- individuel. L'arrestation est un évènement rare, donc une mesure mensuelle ou hebdomadaire serait trop courte.

## *5.2 Les objectifs de la recherche*

Le but de ce mémoire est de comparer quatre modèles d'analyses utiliser pour traiter les données avec une structure poissonnienne soit : le modèle de Poisson (P), le modèle

binomial négatif (NB), le modèle de Poisson avec inflation du zéro (ZIP) et le modèle binomial négatif avec inflation du zéro (ZINB). Pour ce faire, différentes hypothèses seront postulées dans la partie suivante par rapport aux changements relatifs aux risques d'arrestations. Celles-ci permettront de déterminer si le changement dans les coefficients peut différer de façon assez importante pour infirmer ou confirmer une même hypothèse de recherche issue des mêmes données par rapport à l'utilisation d'un modèle qui s'ajuste plus ou moins adéquatement aux données. Également, comme il sera analysé dans le chapitre 6, les modèles seront également comparés en fonction de leur capacité à prédire les moyennes ainsi que la proportion observée de zéros dans l'échantillon par rapport au nombre d'arrestations à travers le temps.

### *5.3 Les questions de recherche*

Dans ce projet de recherche, les démarches de modélisation attribuables à une variable dépendante ayant une structure poisonnienne à travers le temps seront étudiées. Plus précisément, le lien entre le nombre d'arrestations et les variables explicatives relatives à l'utilisation de la violence et le temps qui passe seront analysés selon les différents modèles statistiques exposés dans les deux chapitres précédents. Le mémoire tentera donc de répondre au deux questions de recherche suivantes :

1) Quelles sont les variations sur les statuts initiaux et les taux de changement, entre les différents modèles, par rapport au nombre d'arrestations des membres des gangs de rue à Montréal de 2005 à 2007 ?
2) Lequel des quatre modèles est le plus adéquat dans la prédiction des moyennes observées et dans la prédiction de la probabilité de zéros pour les individus violents et les individus non-violents ?

## *5.4 Les hypothèses postulées*

Cette section contient les hypothèses postulées pour chaque variable indépendante en lien avec les changements relatifs aux risques d'arrestations.

### *5.4.1 Les changements à travers le temps*

Selon les résultats de l'étude de Alfred Blumstein and Cohen (1979), le nombre d'arrestations est généralement stable dans la carrière criminelle d'un délinquant. Comme il s'agit d'un phénomène rare, le nombre d'arrestations est peu élevé à travers le temps, même pour les délinquants les plus actifs. Fait important, les auteurs mentionnent que le risque d'arrestations est le même pour tous les délinquants. Les membres de gangs de rue de cet échantillon possèdent alors des probabilités équivalentes d'être arrêtés à travers le temps. En accord avec les résultats présentés par Blumstein et Cohen (1979), le nombre d'arrestations devrait être stable à travers le temps. Cependant, certains facteurs peuvent modifier cette stabilité comme l'usage de la violence des membres.

### *5.4.2 L'utilisation de la violence*

Est-ce que l'utilisation de la violence permet de prédire le nombre d'arrestations à travers le temps par rapport aux individus non violents ? Une des études les plus influentes et toujours actuelle en criminologie sur le risque d'arrestations par rapport au type de crime est celle d'Albert Blumstein (1986). Ce dernier, à l'aide des calculs de probabilités à l'époque, en vient à la conclusion que les individus violents possèdent 5 % plus de chances d'être arrêtés par rapport à tous les individus qui commettent d'autres types de crimes comme les vols et les cambriolages. À la lumière des résultats de Blumstein, nous postulons que les probabilités d'être arrêtés sont plus élevées pour les individus violents, se traduisant par une ordonnée à l'origine significativement plus élevée ainsi qu'un taux de changement positif ou stable à travers le temps, mais toujours plus élevé que les individus non violents.

# Chapitre 6 : La méthodologie

Ce chapitre présente en détail la méthodologie employée afin de répondre à la question de recherche : quelles sont les variations sur les statuts initiaux et les taux de changement par rapport aux nombres d'arrestations pour les membres des gangs de rue à Montréal de 2005 à 2007 ? On y retrouve une section portant sur l'échantillon, suivit des différentes variables et leurs mesures, les statistiques descriptives ainsi que la stratégie analytique du présent mémoire.

## *6.1 L'échantillon*

La construction de la base de données a débuté au mois de janvier de l'année 2009 dans la cadre d'un projet d'envergure du SPVM portant l'évolution du phénomène des gangs de rue à Montréal pour la période de 2001 à 2008. Afin d'identifier les individus gravitant de près ou de loin dans le milieu des gangs de rue, la méthode de sélection nommée «Boule-de-neige» (ou en anglais *SnowBall*) fût utilisée. De façon stratégique, la construction de l'échantillon débute avec des individus identifiés selon un critère bien précis (Wright, Decker, Redfern, & Smith, 1992) qui dans notre cas, est la liste des membres officiels des gangs de rue les plus actifs, selon le SPVM en 2001. À la base, l'échantillon contenait 2221 membres en règles. De ces membres, les personnes ayant été en contact direct (observé par les policiers-patrouilleurs) ou étant des co-délinquants ou bien ayant été dénoncé par des sources fiables comme étant relié à des membres de gangs furent sélectionnés[6]. Après cette étape, 6443 individus furent dans le milieu des gangs de rue à Montréal. Toutes les informations ont été tirées des fiches d'observation et de co-délinquance ainsi que des rapports d'évènement provenant des contrôles d'identités du SPVM lors des enquêtes sur les membres de gangs de rue à Montréal.

[6] Seulement les individus déjà enregistrés dans les bases de données du SPVM ainsi que ceux possédant des informations primaires comme la date de naissance et le nombre d'arrestations furent conservés dans la banque de données du projet.

### *6.2 La sélection de l'échantillon final*

Premièrement, il est important de mentionner que les femmes ont été exclues de la sélection. Le justificatif derrière cette exclusion se base sur quelques études en criminologie (D'Élia, 2009; Fournier, Cousineau, & Hamel, 2011; Hébert, Hamel, & Savoie Ginette, 1997) affirmant que les femmes qui gravitent autour du milieu des gangs de rue sont, généralement, des conjointes non criminellement actives ou des prostituées sous l'emprise des proxénètes membres de gangs de rue. De ce fait, la création des trajectoires criminelles des femmes, dans le cadre de ce projet d'étude, nous apparaissait non-pertinente. Deuxièmement, en conformité avec la loi québécoise sur la protection de la jeunesse[7], tous les individus âgés de moins de 18 ans ont étés exclus de l'analyse. Troisièmement, afin de s'assurer d'analyser des individus actifs criminellement, nous avons sélectionnés uniquement les individus déjà arrêtés au moins une fois entre 1997 et 2000. Originalement, la période d'étude se situe entre 2001 à 2008. Par contre, pour les besoins de ce mémoire, uniquement trois temps de mesures furent employés (2005 à 2007). Au final, l'échantillon total est composé de 470 hommes actifs dans le milieu des gangs de rue à Montréal pour une proportion de 263 membres et de 207 individus reliés à ces derniers.

### *6.3 Les mesures*

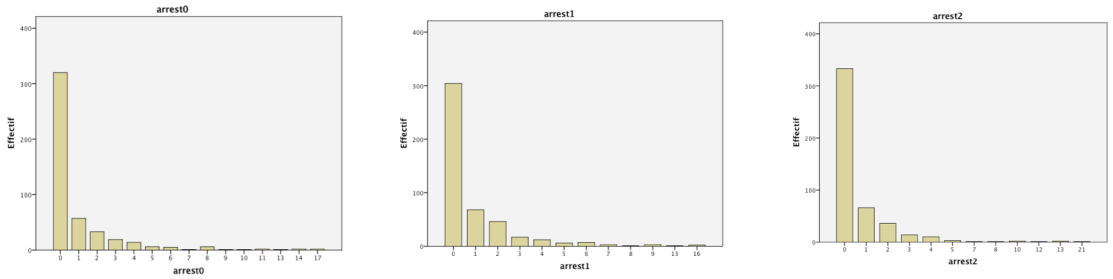
La variable dépendante est un nombre d'arrestations annuelles effectuées par les policiers du SPVM pour la période de 2005 à 2007. Une arrestation consiste à transporter une personne, selon une infraction ou un crime commis, au poste de police pour une

[7] La loi québécoise sur les jeunes contrevenants s'applique aux jeunes de 12 à 17 ans. Elle restreint l'accès aux informations de l'adolescent déviant, plus spécialement dans les dossiers de la police. Les informations de l'adolescent ne peuvent être conservées que si ce dernier est retrouvé coupable de l'infraction pour laquelle un dossier a été monté. Dans le cas contraire, si l'adolescent est acquitté, les informations doivent être détruites. Dans notre échantillon, il était impossible de déterminer avec certitude si les jeunes avaient condamnés ou acquittés. Par souci de respect de la loi, tous les adolescents furent supprimés de l'analyse.

interrogation. Il n'est pas nécessaire de porter des accusations contre un individu pour compiler une arrestation. Les valeurs de la variable varient de 0 à 21. Il n'y a aucune décimale ou valeur négative.

Le graphique suivant expose la distribution du nombre d'arrestations à tous les temps de mesures. Sans équivoque, les distributions sont asymétriques positives et non-normales avec un nombre très important de zéros à tous les temps de mesures. On comparaison avec la figure 1, la distribution s'apparente grandement à une distribution de type ZINB. Les analyses avec ce modèle pourront confirmer si ce modèle prédit le mieux les données observées.

Figure 3 **: Les distributions annuelles du nombre d'arrestations**




Au premier niveau, la variable indépendante est le temps, c'est-à-dire l'année dans laquelle le nombre d'arrestations a été comptabilisé. Pour s'assurer que l'ordonnée à l'origine représente le nombre d'arrestations au premier temps de mesure, l'échelle a été convertie de 0 à 2, où 0 représente l'année 2005 et 2 représente l'année 2007.

Au niveau deux, une variable est incluse. Afin de faciliter l'interprétation des résultats, celle-ci est traitée de façon dichotomique (Farrington & Loeber, 2000).

La variable se rapporte à l'utilisation de la violence par un individu. Il s'agit de voies de faits, d'agressions armées, de tentatives de meurtre et d'homicides. Cette variable a été transformée en dichotomique, soit 0 pour aucun crime et 1 pour un crime ou plus usant de

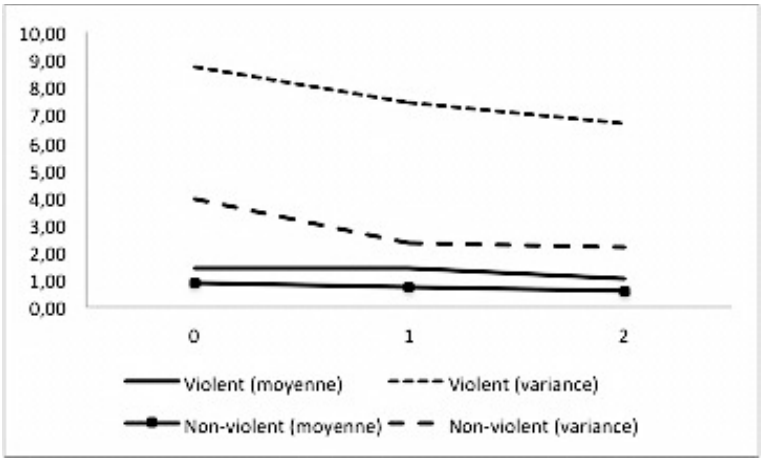
violence. Celle-ci est fixe et se rapporte au temps zéro, soit en 2005. Dans l'échantillon, 26% des individus sont considérés comme violents contre 74% de personnes non-violentes.

*6.3.1 La moyenne annuelle selon le prédicteur*

Afin de saisir la forme de la distribution, voici les moyennes et les variances annuelles du prédicteur de violence inclus dans les analyses aux trois temps de mesures. Uniquement la moyenne et la variance sont exposées dans les graphiques suivants considérant qu'il s'agit des deux statistiques descriptives les plus importantes lors de l'utilisation des modèles de données comptées.

Le graphique suivant expose les moyennes annuelles (droites pleines) ainsi que les variances (droites pointillées) aux trois temps de mesures en lien avec l'utilisation de la violence dans le milieu des gangs de rue. Au niveau de la moyenne, les personnes violentes sont légèrement plus arrêtées à travers le temps que les personnes non violentes.

Figure 4 **: La moyenne et la variance selon la variable de violence**




De plus, les deux droites sont stables à travers le temps. Cependant, la variance entre le nombre d'arrestations est beaucoup plus élevée pour les individus violents. En

d'autres mots, il existe dans cet échantillon une plus grande hétérogénéité dans le nombre d'arrestations. Les personnes non violentes possèdent une variance moins élevée. Cependant, peu importe le groupe d'appartenance, la variance est supérieure à la moyenne signe que le paramètre de surdispersion devrait être modélisé dans les analyses subséquentes.

### *6.3.2 La proportion de zéros*

L'étude des phénomènes rares implique un nombre généralement important de personnes n'ayant pas vécu l'évènement se traduisant par une proportion importante de zéros. Le tableau suivant expose la proportion de zéros observée dans notre échantillon en fonction des catégories du prédicteur à tous les temps de mesures.

**Tableau 1: La proportion de zéros conditionnelle à la violence**

| ***Temps*** | ***0*** | ***1*** | ***2*** |
|---|---|---|---|
| Pourcentage Total (n=470) | 0,68 | 0,65 | 0,71 |
| Violent (n=123) | 0,69 | 0,57 | 0,63 |
| Non-violent (n=347) | 0,69 | 0,67 | 0,74 |

Sans distinction du prédicteur, la proportion de zéros dans l'échantillon varie de 65% à 71%. Comme il est exposé dans le graphique 1, les individus violents sont plus arrêtés, la proportion de zéros est donc inférieure. Globalement, celle-ci demeure quand même très élevée à tous les temps de mesures.

## *6.4 La stratégie d'analyse*

La stratégie d'analyse du présent mémoire se base sur celle employée par J.D. Singer and B.Willet (2003) consistant à entrer une variable à la fois en débutant du niveau un vers le niveau deux. Cette méthode permet de s'assurer d'une certaine constance sur les

variations et permet également de bien saisir l'apport statistique au niveau des ordonnées à l'origine et des taux de changements des différents prédicteurs.

**Tableau 2: La logique d'introduction des variables**

| | ***Variables*** | Nom | Modèle A | Modèle B | Modèle C |
|---|---|---|---|---|---|
| | **Dépendante**<br>Nombre d'arrestations | *arrest0-2* | *X* | *X* | *X* |
| *Niveau 1* | **Indépendante**<br>Temps | *0-2* | | *X* | *X* |
| *Niveau 2* | Utilisation de violence | *Violent* | | | *X* |

Le premier modèle (A) est celui de la grande moyenne d'arrestations excluant l'effet du temps et des prédicteurs. Cette grande moyenne est calculée par rapport à toutes les variations de la moyenne de tous les individus dans l'échantillon. Deuxièmement, suit le modèle de croissance inconditionnel (B) au niveau 1. Le temps est ajouté afin de déterminer les variations dans le nombre d'arrestations à travers les trois années de l'étude sans l'effet des prédicteurs. Troisièmement, au niveau 2, le prédicteur de violence sera ajouté à l'effet du temps (modèle C).

Finalement, pour la modélisation du modèle de Poisson, du modèle binomial négatif et des modèles modifiés en zéro, le logiciel SAS (version 9.2) a été utilisé. Pour le modèle de Poisson et binomial négatif la procédure GLIMMIX a été employée. Cette procédure est utilisée généralement pour les données non normales avec une structure hiérarchique des effets aléatoires en supposant que ces derniers possèdent une distribution normale. La procédure GLIMMIX utilise la méthode d'estimation du pseudo-likelihood. Pour les modèles ZIP et ZINB, la procédure NLMIXED fut utilisée, car il s'agit de modèles plus complexes. NLMIXED permet également de travailler sur des données comptées, non linéaires, mais requiert plus de codifications manuelles. La méthode d'estimation ,des modèles ZIP et ZINB, est basée sur la quadrature de Gauss qui est une approximation de la valeur numérique d'une intégrale permettant de modéliser des probabilités comme c'est le cas avec les données comptées (Atkins & Gallop, 2010).

# Chapitre 7 : Les résultats

## *7.1 La modélisation de Poisson*

Le premier modèle utilisé dans ce mémoire est potentiellement le moins adéquat pour ce type de données. Car le principal postulat d'égalité de la variance et de la moyenne n'est pas respecté. Par le fait même, ce modèle ne tient pas compte de la surdispersion des données. Une modélisation linéaire en fonction du temps sera préconisée.

Les coefficients issus de la modélisation basée sur le modèle de Poisson sont exposés dans le tableau 3. Les coefficients inscrits sont en log, c'est-à-dire qu'ils doivent ultimement être transformés avec la fonction exponentielle pour calculer, au final, les moyennes prédites. De plus, l'erreur-type est ajoutée entre parenthèses à côté de chaque coefficient.

Considérant que nous cherchons à déterminer le meilleur modèle pour prédire les données, les coefficients seront analysés selon chaque modèle uniquement dans le modèle Poisson, ensuite nous nous concentrerons sur les modèles C.

**Tableau 3: Les coefficients du modèle de Poisson**

| | | Paramètres | Modèle A | Modèle B | Modèle C |
|---|---|---|---|---|---|
| **Effets fixes** | *Statut initial* | $\gamma_{00}$ | -0,14***(0,02) | -0,95***(0,11) | -1,07***(0,12) |
| | *Taux de changement* | $\gamma_{10}$ | | -0,36***(0,06) | -0,36***(0,06) |
| | *Violent* | $\gamma_{02}$ | | | 0,50***(0,19) |
| **Effets aléatoires** | Statut initial | $\zeta_{01}$ | 2,10*** | 2,17***(0,27) | 2.11***(0,26) |
| | Taux de changement | $\zeta_{02}$ | | 0,41***(0,08) | 0,41***(0,08) |
| | *Pearson Chi-Square/DF* | ---------- | 4,71 | 0,88 | 0,53 |
| | *BIC* | ---------- | 4858.22 | 3374.99 | 3374.30 |

**a. ***,**,*,ł, indique une différence significative au seuil de 0,001, 0,01, 0,05 et compris entre 0,05 et 0,06 respectivement.**

Le statut initial au modèle A expose la grande moyenne d'arrestations pour tous les individus à travers le temps. En étant significative, elle nous indique que la moyenne d'arrestations dans l'échantillon est différente de zéro. La variance étant égale à 2,10 montre le changement dans le temps pour un individu, une fois ajouté au niveau de la grande moyenne, le résultat expose la moyenne spécifique aux individus dans l'échantillon.

Également, le coefficient de surdispersion démontré par le Pearson Chi-Square divisé par le nombre de degrés de liberté est égal à 4,71. Comme celui-ci est supérieur à 1,25, ceci suggère la présence de surdispersion dans les données, mais cette surdispersion semble être amoindrie par l'introduction de la variable « temps » dans le modèle.

Dans le modèle B, le coefficient associé à la pente de changement est égal à -0,36 ($\gamma_{10}$) et est significatif. Autrement dit, la pente moyenne de tous les individus de l'échantillon est différente de zéro et elle est décroissante. L'hypothèse de départ mentionnant une stabilité dans le nombre d'arrestations est donc infirmée.

Le modèle C contient l'introduction de la variable de violence. Au niveau du coefficient de l'ordonnée à l'origine pour les individus violents ($\gamma_{02}$), il est égal à 0,50 et il est significatif ($p<0,001$). Autrement dit, la différence entre le statut moyen des individus non violents ($\gamma_{00}$ =-1,07, $p<0,001$) est de 0,50 supérieur pour les individus violents. Au final, le statut initial des individus violents est de -0,57. Ce modèle confirme l'hypothèse de départ qui postulait que les individus violents sont plus fréquemment arrêtés, du moins, au début de la période d'étude et ce, d'une manière constante à travers le temps. Aucune interaction n'a été détectée entre la violence et le temps. Ce même constat s'applique également au trois modèles suivants. De plus, même avec l'ajout de la violence, la trajectoire moyenne de changement est identique.

### *7.2 La modélisation binomiale négative*

Dans cette deuxième partie, le modèle avec une distribution binomiale négative a été employé afin de déterminer les variations sur le nombre d'arrestations. En théorie, ce type de modèle s'emploie couramment sur les données comptées considérant que la variance est plus importante que la moyenne au sein de la variable dépendante. Ce modèle tient compte de cette particularité dans la production des coefficients de prédiction. Par contre, la grande proportion de zéros dans la variable dépendante n'est pas prise en considération. Les coefficients issus du modèle binomial négatif sont exposés dans le tableau 4. Seulement les ordonnées à l'origine furent laissées aléatoires, car le modèle ne parvenait pas à converger avec le caractère aléatoire des pentes. Cette non-convergence peut s'expliquer soit par une variation trop faible du nombre d'arrestations à travers le temps ou soit par l'ajout du paramètre de surdispersion (k) qui tient également compte de la variance à travers le temps.

**Tableau 4: Les coefficients du modèle binomial négatif**

| | | Paramètres | Modèle A | Modèle B | Modèle C |
|---|---|---|---|---|---|
| **Effets fixes** | | | | | |
| | *Statut initial* | $\gamma_{00}$ | -0,14**(0,06) | -0,75***(0,11) | -0,87**(0,12) |
| | *Taux de changement* | $\gamma_{10}$ | | -0,15*(0,06) | -0,15*(0,06) |
| | *Violent* | $\gamma_{02}$ | | | 0,50**(0,18) |
| **Effets aléatoires** | *Statut initial* | $\zeta_{01}$ | 1.67***(0,23) | 1.67***(0,23) | 1.60***(0,23) |
| | *Échelle* | k | 1,15***(0,16) | 1.11***(0,16) | 1.11***(0,16) |
| | *Pearson Chi-Square/DF* | ---------- | 1,1 | 0,51 | 0,52 |
| | *BIC* | ---------- | 3377.60 | 3260.01 | 3258.22 |

**a. ***,**,*,ł, indique une différence significative au seuil de 0,001, 0,01, 0,05 et compris entre 0,05 et 0,06 respectivement.**

Le coefficient de pente au modèle C dénote une diminution significative du nombre d'arrestations à travers le temps ($\gamma_{10}$ = -0,15, p<0,05). L'hypothèse de départ est infirmée, le nombre d'arrestations n'est pas stable à travers le temps.

Au niveau du changement sur l'ordonnée à l'origine, le coefficient des individus violents est égal à 0,50 ($\gamma_{02}$) et il est significatif ($p<0,01$). À la lumière des résultats précédents, l'hypothèse concernant les individus violents est confirmée.

### *7.3 Le modèle ZIP*

Considérant la grande proportion de zéro à tous les temps de mesures, il est nécessaire d'utiliser les modèles avec inflation du zéro. Dans cette section, le premier modèle testé est le ZIP. Celui-ci tient compte du nombre important de zéro en créant simultanément deux types de distributions. La première étant ceux ayant été victimes de l'évènement versus les non-victimes, soit la partie logistique. La deuxième, est une distribution de Poisson avec un nombre de zéros déterminé par le logiciel permettant d'obtenir une variance égale à la moyenne avec l'ajout des valeurs un et plus. Le but des modèles avec inflation du zéro est justement de tenir compte du nombre important de zéro afin de le diminuer pour estimer la proportion de zéro en trop. Autrement dit, le modèle contrôle l'inflation du zéro par lui-même. Par contre, ce modèle basé sur la loi de Poisson ne tient pas compte de la surdispersion des données. C'est pourquoi il devrait être moins adapté que le modèle ZINB.

Comme la distribution est séparée en deux, il y a deux parties distinctes à analyser dans ce tableau, la partie dite logistique et celle de Poisson avec inflation du zéro. Également, uniquement l'ordonnée à l'origine varie aléatoirement, car le temps aléatoire dans une régression logistique n'est pas approprié.

#### *7.3.1 Les résultats de la partie logistique*

Les résultats de la partie logistique dans le modèle ZIP sont difficilement interprétables à leur état brut, car il mesure l'inflation du zéro dans le modèle (Atkins & Gallop, 2007). Il est recommandé de les transformer avec la fonction exponentielle

($e^{\beta}$) afin de les analyser comme des rapports de cotes. Par contre, les coefficients logistiques ne sont pas utilisés pour valider ou infirmer les hypothèses. Ils seront uniquement utilisés dans le calcul final des moyennes prédites.

*7.3.2 Les résultats de la partie Poisson*

Dans cette partie, le temps fût laissé aléatoire cependant afin de capter le plus de variation possible, mais en vain. L'hypothèse relative à une stabilité du nombre d'arrestations est infirmée par défaut considérant que le coefficient du taux de changement n'est pas significatif, même constat pour la variable de violence.

**Tableau 5: Les coefficients du modèle ZIP**

| **Partie logistique** | | **Paramètres** | **Modèle A** | **Modèle B** | **Modèle C** |
|---|---|---|---|---|---|
| Effets fixes | | | | | |
| | *Statut initial* | $\gamma_{00}$ | -0,68***(0,10) | 1,29*(0,59) | -0,87 (0,18) |
| | *Taux de changement* | $\gamma_{10}$ | | -0,71ł (0,36) | 0,15 (0,56) |
| | *Violent* | $\gamma_{02}$ | | | 0,50 (0,56) |
| Effets aléatoires | *Statut initial* | $\zeta_{01}$ | 2,61***(0,78) | 4,51**(1,80) | 4,35*(2,03) |
| | *Taux de changement* | $\zeta_{02}$ | ----- | ----- | ----- |
| **Partie Poisson** | | | | | |
| Effets fixes | *Statut initial* | $\gamma_{00}$ | 1,08***(0,05) | -0,44ł(0,23) | -0,50 (0,30) |
| | *Taux de changement* | $\gamma_{01}$ | | 0,22 (0,15) | 0,21 (0,21) |
| | *Violent* | $\gamma_{02}$ | | | 0,33 (0,19) |
| Effets aléatoires | | | | | |
| | *Statut initial* | $\zeta_{01}$ | 1,05 (0,00) | 1,80***(0,36) | 1,74***(0,44) |
| | *Taux de changement* | $\zeta_{02}$ | | 0,24***(0,07) | 0,24 ***(0,07) |
| **BIC** | | | 3314.7 | 3306,5 | 3304,5 |

**a. ***,**,*,ł, indique une différence significative au seuil de 0,001, 0,01, 0,05 et compris entre 0,05 et 0,06 respectivement.**

## *7.4 Le modèle ZINB*

Quant à ce quatrième modèle, il tient compte de l'inflation du zéro et de la surdispersion des données. En théorie, c'est ce modèle qui est le plus adapté à la distribution du nombre d'arrestations des membres de gang de rue considérant un nombre très important de zéro et une grande variation dans le nombre d'arrestations. Les coefficients du modèle ZINB sont condensés dans le tableau six.

**Tableau 6 : Les coefficients du modèle ZINB**

| **Partie logistique** | | **Paramètres** | **Modèle A** | **Modèle B** | **Modèle C** |
|---|---|---|---|---|---|
| Effets fixes | | | | | |
| | *Statut initial* | $\gamma_{00}$ | 0,11ł (0,16) | 1,37***(0,35) | 1,17**(0,38) |
| | *Taux de changement* | $\gamma_{10}$ | | -0,75***(0,17) | -0,72***(0,16) |
| | *Violent* | $\gamma_{02}$ | | | 0,49 (0,46) |
| Effets aléatoires | *Statut initial* | $\zeta_{01}$ | 2,38***(0,55) | 4,61***(1,70) | 4,44**(1,65) |
| | *Taux de changement* | $\zeta_{02}$ | ----- | ----- | ----- |
| **Partie NB** | | | | | |
| Effets fixes | *Statut initial* | $\gamma_{00}$ | -0,01 (0,10) | -0,47***(0,18) | -0,53**(0,18) |
| | *Taux de changement* | $\gamma_{01}$ | | 0,24*(0,10) | 0,22*(0,09) |
| | *Violent* | $\gamma_{02}$ | | | 0,34(0,19) |
| Effets aléatoires | | | | | |
| | *Statut initial* | $\zeta_{01}$ | 0,99(0,13)*** | 1,85***(0,35) | 1,77***(0,34) |
| | *Taux de changement* | $\zeta_{02}$ | | 0,23***(0,08) | 0,23***(0,08) |
| | *Échelle* | k | 0,10***(0,03) | 0,05 (0,03) | 0,05(0,03) |
| **BIC** | | | 3302,3 | 3301,6 | 3305,7 |

**a. ***,**,*,ł, indique une différence significative au seuil de 0,001, 0,01, 0,05 et compris entre 0,05 et 0,06 respectivement.**

### *7.4.1 Les résultats de la partie logistique*

Les coefficients logistiques ne sont également pas analysés dans cette partie pour les mêmes raisons mentionnées dans le modèle ZIP. Ils seront utilisés uniquement à des fins de calculs des moyennes prédites finales.

### *7.4.2 Les résultats de la partie binomiale négative*

Dans le modèle B ($\gamma_{10} = 0{,}24, p < 0{,}05$) et C ($\gamma_{10} = 0{,}22, p < 0{,}05$), il y a une donc une augmentation du nombre d'arrestations à travers le temps, car les coefficients sont positifs. L'hypothèse de la stabilité du nombre d'arrestations est infirmée dans ce modèle, mais elle est à l'opposé des modèles de P et NB. En effet, ceux-ci proposaient un taux de changement négatif à travers le temps. Pour ce qui est des individus violents, le coefficient est non-significatif.

Pour conclure ce chapitre, les quatre modèles testés amènent des résultats assez différents au niveau du taux de changement. En effet, le nombre d'arrestations est décroissant dans les modèles P et NB alors qu'il est croissant dans le modèle ZINB et non-significatif dans le modèle ZIP. Également, le modèle NB ne permet pas d'inclure le temps aléatoire en raison de la non-convergence du modèle. Au niveau de l'hypothèse de violence, celle-ci se confirme uniquement dans le modèle P et NB. Les membres violents possèdent un statut initial plus élevé que les membres non violents.

# Chapitre 8 : La comparaison des modèles

Dans ce chapitre, tous les modèles effectués dans le chapitre cinq seront comparés sous trois angles afin de déterminer leurs différences, leurs ressemblances et leur capacité prédictive dans l'optique de déterminer le plus efficient sur tous les plans. Selon MacDonald and Lattimore (2010), il n'existe pas de meilleur modèle à proprement parlé, tout dépend du type de données utilisé ainsi que de leur capacité prédictive par rapport à ces mêmes données. Le premier volet consiste à comparer les hypothèses de recherche en fonction de tous les types de modélisation. Le deuxième volet vise à comparer tous les coefficients calculés dans les modèles C issus de tous les types de modélisation à l'aide des courbes prototypiques et de la variance. Le troisième volet permettra de comparer la probabilité de zéro afin de déterminer la capacité prédictive des modèles.

## *8.1 Les hypothèses de recherche*

Dans le tableau 10, les X représentent les hypothèses infirmées, les crochets représentent celles confirmées et le sigle NS représente des résultats non significatifs. La première hypothèse postulant une stabilité du nombre d'arrestations à travers le temps peut se comparer uniquement sur les données dites comptées. Il est intéressant de constater que seulement dans le modèle de Poisson et le modèle NB, le nombre d'arrestations diminue significativement à travers le temps ce qui infirme l'hypothèse de la stabilité du nombre d'arrestations à travers le temps. Dans le ZIP, il n'existe aucune variation significative à travers le temps alors que le ZINB infirme l'hypothèse de départ avec une augmentation du nombre d'arrestations à travers le temps.

**Tableau 7: Les hypothèses de recherche confirmées ou infirmées**

| Modèles | Poisson | NB | ZIP | ZINB |
|---|---|---|---|---|
| *Hypothèses* | | | | |
| 1.Stabilité à travers le temps | X | X | NS | X |
| 2. Ordonnée (ou risque) plus élevée pour les individus violents | ✔ | ✔ | NS | NS |

Au niveau de l'ordonnée à l'origine, les individus violents débutent significativement plus haut sur la droite que les individus non violents dans le modèle P et NB. Le modèle ZIP et ZINB ne permettent pas de prendre une décision considérant que les coefficients sont non significatifs.

Le but de présenter les hypothèses confirmées ou infirmées est de démontrer que le choix d'un modèle plus ou moins adapté peut produire des résultats différents au point de tirer des conclusions contraires. Par contre, les modèles sont consistants dans le sens que les modèles avec inflation du zéro produisent les sensiblement les mêmes conclusions alors que le modèle de Poisson et le modèle NB produisent également les mêmes conclusions.

### *8.2 L'ajustement des modèles*

Le premier élément permettant de comparer l'ajustement des modèles est le BIC (*Bayesian Information Criterion*), où le plus faible est le mieux. Celui-ci est utilisé pour comparer les modèles entre eux en fonction du nombre de paramètres inclus ainsi que de la taille de l'échantillon (Singer & Willet, 2003). Par contre, il est irréalisable de comparer tous les modèles entre eux, car ceux-ci ne sont pas nichés les uns dans les autres. Selon (Atkins & Gallop, 2010), le modèle de Poisson est niché dans le modèle NB alors que le ZIP est niché dans le ZINB. Il est alors possible de comparer le P avec le NB et le ZIP avec le ZINB. Dans les deux premiers modèles, le NB possède le BIC le moins élevé (3374,30 versus 3258,22) alors que dans les deux derniers modèles, le ZIP possède un BIC très

légèrement inférieur (3304,5 versus 3305,7). De ces résultats, les modèles NB et ZIP sont les plus adaptés aux données.

Les résultats présentés dans le tableau huit donnent un aperçu numérique des résultats obtenus. Dans les analyses multiniveaux longitudinales, les trajectoires prototypiques sont un outil puissant pour communiquer les résultats et se représenter plus facilement l'impact des prédicteurs (J.D. Singer & B.Willet, 2003). D'une façon un peu différente, comme ce mémoire est une comparaison de modèles statistiques, les courbes prototypiques seront utilisées dans l'optique de déterminer quel modèle se rapproche le plus des données observées. Cependant, comme il s'agit d'analyses multiniveaux longitudinales avec des paramètres aléatoires, le calcul des moyennes prédites se fait différemment que par simple multiplication. Pour ce faire, la loi de Monte-Carlo (Hammersley & Handscomb, 1964) a été utilisée pour calculer les moyennes prédites. Celle-ci est une méthode visant à calculer une valeur numérique en utilisant des procédés aléatoires avec la loi des grands nombres ce qui s'applique bien dans les analyses ci-présentes. De plus, les modèles sont comparés à tous les temps de mesures pour les individus violents et les non-violents considérant que les moyennes sont différentes pour ces deux groupes.

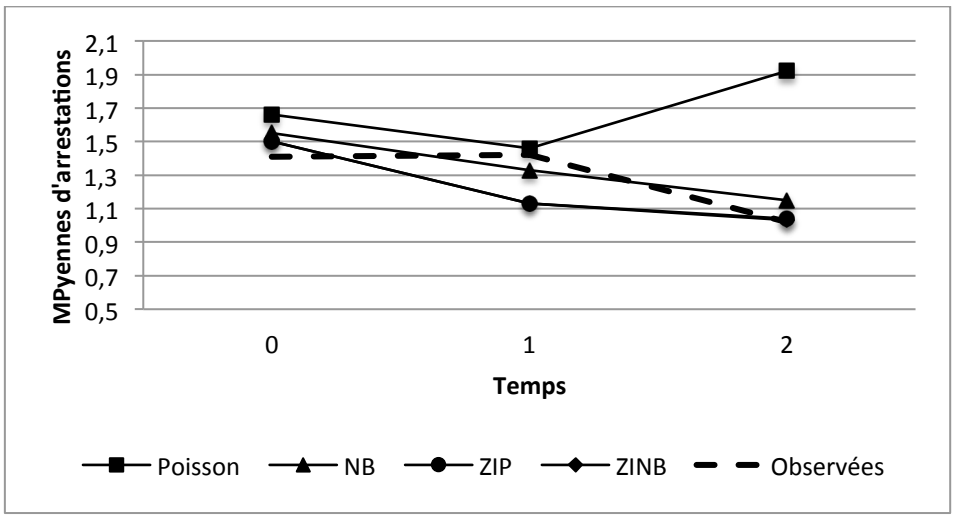
**Tableau 8: Les coefficients de tous les modèles C**

| | **Poisson** | **NB** | **ZIP** | | **ZINB** | |
|---|---|---|---|---|---|---|
| **Effets fixes** | | | *Poisson* | *Logistique* | *NB* | *Logistique* |
| Ordonnée | -1,07***(0,12) | -0,87**(0,12) | -0,50 (0,30) | -0,87 (0,18) | -0,53**(0,18) | 1,17**(0,38) |
| Temps | -0,36***(0,06) | -0,15*(0,06) | 0,21 (0,21) | 0,15 (0,56) | 0,22*(0,09) | -0,72***(0,16) |
| Violent | 0,50***(0,19) | 0,50**(0,18) | 0,33 (0,19) | 0,50 (0,56) | 0,34(0,19) | 0,49 (0,46) |
| **Effets aléatoires** | | | | | | |
| Ordonnée | 2.11***(0,26) | 1.60***(0,23) | 1,74***(0,44) | 4,35*(2,03) | 1,77***(0,34) | 4,44**(1,65) |
| Effets temps | 0,41***(0,08) | ---------- | 0,24 ***(0,07) | ---------- | 0,23***(0,08) | ---------- |
| Échelle | ---------- | 1.11***(0,16) | ---------- | ---------- | 0,05(0,03) | ---------- |
| **BIC** | 3374.30 | 3258.22 | 3304,5 | | 3305,7 | |

Débutons la comparaison avec les individus considérés comme étant violents (figure 5) dans le milieu des gangs de rue à Montréal. La courbe en pointillé représentant les moyennes observées sera comparée avec les autres courbes se rapportant chacune à un modèle spécifique. Notez que le ZIP et le ZINB sont superposés dans le graphique.

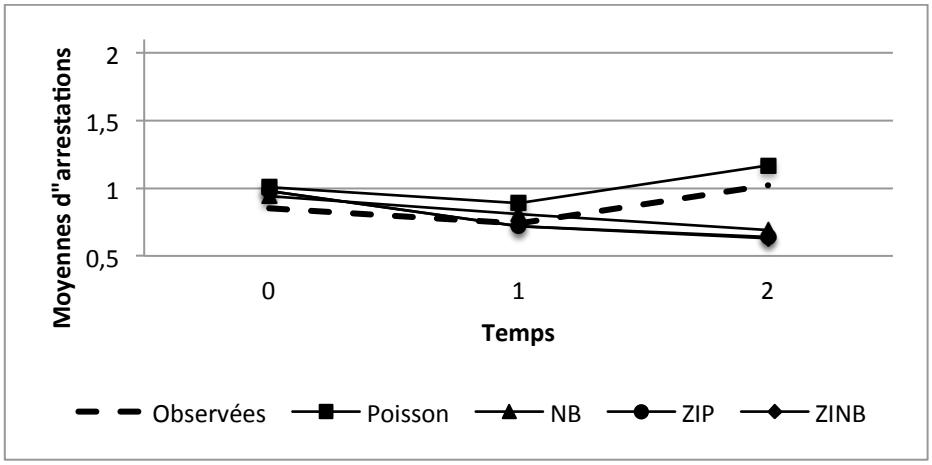
Au temps 0, toutes les courbes sont très près l'une de l'autre, les modèles prédisent bien la moyenne à ce premier temps. Au deuxième temps, les modèles ZIP et ZINB s'écartent davantage de la moyenne observée.

**Figure 5 : Les moyennes prédites et observées pour les individus violents**




Les courbes de P et de NB, quant à elle, sont relativement près de la courbe des moyennes observées, mais le P estime un peu à la hausse la moyenne alors que le NB sous-estime légèrement la moyenne. Au temps trois, la courbe de Poisson s'éloigne considérablement de la moyenne observée alors que la NB, ZIP et ZINB sont très près. Au final, nous affirmons que le modèle NB est celui qui prédit le mieux les moyennes par rapport à la courbe des données observées pour les individus violents.

**Figure 6 : Les moyennes prédites et observées pour les individus non violents**




Au temps 0 et 1, les moyennes prédites de tous les modèles sont très près des moyennes observées. Il n'y a pas de différence significative entre les modèles. Au temps 3, le modèle de Poisson surestime légèrement la moyenne alors que les modèles NB, ZIP et ZINB surestiment légèrement la moyenne. Par contre, le modèle de Poisson suit la forme de la courbe à l'inverse des trois autres. Pour ce qui est des membres de gangs de rue non-violents, le modèle de Poisson semble prédire le plus adéquatement les moyennes observées pour les individus non violents.

À la lumière des graphiques précédents, le modèle NB est le plus adapté pour les individus violents alors que le modèle de Poisson l'est davantage pour les non-violents. Cependant, dans ce dernier groupe, le modèle NB prédit également mieux les moyennes d'arrestations que le ZIP ou le ZINB. Du coup, nous considérons que le NB est celui qui prédit le plus adéquatement le nombre d'arrestations des membres de gangs de rue.

### *8.3 La prédiction du zéro*

Une autre méthode afin de poser un diagnostic sur la capacité prédictive des modèles consiste à comparer les différences entre la proportion de zéros prédits et observés (Yinglin et al., 2012). Pour ce faire, la loi de Monte-Carlo fut également utilisée afin de calculer la proportion de zéros prédits pour tous les modèles. Le calcul se fait à l'aide des

équations définies dans le chapitre 1 et 2 pour tous les modèles en tenant compte des paramètres aléatoires. Les équations réduites permettant de calculer la probabilité du zéro, pour tous les modèles, sont exposées dans le tableau 7. Une fois de plus, les violents et les non-violents furent séparés, car la proportion de zéro n'est pas identique pour les deux groupes et ce, à tous les temps de mesure.

Le modèle ZINB est le moins performant dans sa prédiction du zéro avec 28% plus de zéros qu'observé dans les données. À l'inverse, le modèle de Poisson et le NB prédisent, respectivement, 5% et 4% moins de zéros.

**Tableau 9 : Les équations de la probabilité de zéro**

| $\Pr(\gamma_i \mid 0)=$ | |
|---|---|
| $\exp(-\mu)$ | Le modèle de Poisson *(P)* |
| $\left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i}$ | Le modèle binomial négatif (NB) |
| $(1-\psi)(\exp(-\mu))$ | Le modèle de Poisson avec inflation du zéro (ZIP) |
| $(1-\psi)\left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i}$ | Le modèle binomial négatif avec inflation du zéro (ZINB) |

Le modèle ZIP est le plus adéquat dans sa prédiction du zéro avec 2% de moins. Pour les individus violents, le modèle ZIP est le plus approprié pour prédire l'absence du phénomène, c'est-à-dire la proportion de gens qui n'ont jamais été arrêtés.

**Tableau 10: Les différences de prédiction dans la proportion de zéros pour les individus violents**

| *Probabilité zéro* | ***Observé*** | **Poisson** | Différence | **NB** | Différence | **ZIP** | Différence | **ZINB** | Différence |
|---|---|---|---|---|---|---|---|---|---|
| Temps 0 | 0,69 | 0,52 | -0,17 | 0,56 | -0,13 | 0,59 | -0,1 | 0,90 | *0,21* |
| Temps 1 | 0,57 | 0,59 | 0,02 | 0,59 | 0,02 | 0,60 | 0,03 | 0,91 | *0,34* |
| Temps 2 | 0,63 | 0,63 | 0,00 | 0,62 | -0,01 | 0,63 | 0,00 | 0,92 | *0,29* |
| *Différence moyenne* | ------ | ------ | -5% | ------ | -4% | | -2% | | 28% |

Pour les individus non violents, le modèle ZIP prédit parfaitement la proportion d'individus n'ayant jamais été arrêtés. Tout comme le tableau précédent, le modèle NB arrive en deuxième avec une prédiction moyenne inférieure de 2% suivi du modèle de Poisson avec 3% de moins. Également, le ZINB produit 23% plus de zéro, il surestime la proportion d'individus qui ne seront jamais arrêtés à travers le temps.

**Tableau 11: Les différences de prédictions dans la proportion de zéros pour les individus non violents**

| *Probabilité zéro* | ***Observé*** | **Poisson** | Différence | **NB** | Différence | **ZIP** | Différence | **ZINB** | Différence |
|---|---|---|---|---|---|---|---|---|---|
| Temps 0 | 0,69 | 0,62 | -0,07 | 0,65 | -0,04 | 0,68 | -0,01 | 0,92 | *0,23* |
| Temps 1 | 0,67 | 0,68 | 0,01 | 0,68 | 0,01 | 0,70 | 0,03 | 0,93 | *0,26* |
| Temps 2 | 0,74 | 0,71 | -0,03 | 0,71 | -0,03 | 0,73 | -0,01 | 0,94 | *0,2* |
| *Différence moyenne* | ------ | ------ | -3% | ------ | -2% | | 0% | | 23% |

Par rapport à la capacité des modèles à prédire la proportion du zéro, pour les individus violents et non violents, le modèle ZIP est le plus adapté suivi de très près par le modèle NB. Tout comme les moyennes prédites, ce sont ces deux modèles qui se démarquent positivement par leur capacité d'adaptation aux données en cause.

# Chapitre 9 : Discussion

Il est difficile d'affirmer, sans aucun doute, la supériorité d'un modèle sur un autre. Dans les faits, chaque modèle utilisé possède ses forces et ses faiblesses permettant de s'adapter d'une façon différente aux données utilisées. Dans un premier temps, les modèles vus dans les chapitres précédents seront passés en revue par rapport aux variations au niveau des statuts initiaux et des taux de changement. Dans un deuxième temps, cette discussion sera conclue en répondant à la deuxième question de recherche soit : lequel des quatre modèles est le plus adéquat dans sa prédiction des moyennes observées et dans sa prédiction de la probabilité de zéros pour les individus violents et les individus non-violents ?

Le premier modèle testé dans ce chapitre est le modèle de Poisson. Celui-ci possède une longue série de postulats statistiques à respecter. En utilisant une variable dépendante catégorisée comme étant des données comptées, le modèle de Poisson est difficilement applicable par rapport à son postulat d'égalité entre la moyenne et la variance. En effet, la principale caractéristique des données comptées se rapporte à sa distribution asymétrique provoquée par une variance excédant la moyenne. Ce modèle fut quand même testé afin d'observer de quelle façon se comportent les résultats dans un contexte où plusieurs postulats statistiques sont violés. À la grande surprise, le modèle s'ajuste relativement bien sur tous les plans. Au niveau des coefficients, il existe des différences significatives à travers le temps et entre les individus violents et non violents par rapport au nombre d'arrestations infirmant ainsi l'hypothèse de la stabilité du nombre d'arrestations à travers le temps et confirmant une ordonnée à l'origine plus élevée pour les individus violents. L'ordonnée à l'origine ainsi que la pente ont été considérés comme aléatoires permettant d'identifier de la variance à ces deux niveaux. Pour ce qui est des moyennes prédites, le modèle de Poisson estime bien les moyennes observées au temps 0 et au temps 1, mais la prédiction au temps 2 est moins juste pour les individus violents. Cependant, pour les individus non violents, le modèle semble le plus adéquat dans sa prédiction des moyennes d'arrestations. Cela peut s'expliquer du fait que la variation du groupe des individus non-

violents est moins dispersée. La surdispersion étant moins présente, le modèle est plus à même de produire de meilleurs coefficients. Pourquoi ? Le modèle de Poisson utilise la loi normale pour produire l'estimation de ses coefficients alors une plus faible dispersion des données permet de s'approcher davantage d'une distribution normale. Pour ce qui est de la probabilité de zéros prédite par le modèle, la distribution de Poisson arrive en 3e place autant au niveau des individus violents que des non-violents. Étonnamment, nous nous serions attendus à un pire résultat dans sa capacité à produire la bonne proportion de zéros. Au final, le modèle de Poisson n'est pas idéal, mais il s'ajuste assez bien aux données observées.

Le deuxième modèle testé dans ce mémoire est le binomial négatif. La grande force de ce modèle est l'ajout du paramètre de surdispersion à estimer permettant au modèle d'avoir plus de latitude pour tenir compte de la variance du nombre d'arrestations. Par contre, comme le mentionnent Atkins et Gallop (2010), dans un devis longitudinal, l'ajout de ce paramètre peut impliquer une non-convergence du modèle à travers le temps. C'est exactement ce phénomène qui se produit dans les analyses avec le modèle NB. L'ajout du paramètre de surdispersion et d'une composante aléatoire pour le temps provoque la non-convergence du modèle. Au niveau des coefficients, les résultats s'avèrent significatifs engendrant des conclusions similaires au modèle de Poisson à l'exception de l'absence de variance aléatoire à travers le temps. Concernant les moyennes prédites, le NB produit de très bonnes prédictions pour le groupe des individus violents et un peu moins pour le groupe des individus non violents. À l'inverse du modèle de Poisson, moins les données sont dispersées, moins le modèle s'ajuste aux données. Dans la prédiction de la probabilité de l'absence du phénomène, le NB fait aussi très bien autant pour le groupe des individus violents que du groupe des individus non-violents, il se retrouve en deuxième position dans ces deux groupes.

Le troisième modèle utilisé dans ce mémoire est le ZIP. Celui-ci se base sur l'équation de la loi de Poisson, mais ajoute un paramètre d'inflation du zéro. La grande force de ce modèle est de tenir compte de l'inflation du zéro, c'est-à-dire d'une fréquence

plus élevée de la valeur 0 qu'attendue dans la distribution de Poisson ou binomiale négative. Le graphique 3 démontre bien que la fréquence de la valeur 0 est vraiment plus élevée que toutes les autres valeurs. Par contre, ce modèle ne tient pas compte de la surdispersion des données pouvant ainsi sous-estimer cette variation. Est-ce phénomène qui se produit si l'on observe que les coefficients s'avèrent tous non-significatifs ? Comme les estimations tiennent compte de la grande variation des résultats, le seuil de différence est plus élevé pour produire des différences significatives. Au niveau des moyennes prédites, le modèle est efficace au temps 0 et 1 et un peu moins au temps 2 et ce, autant pour le groupe des personnes violentes et des non-violentes. En gros, ce modèle s'ajuste bien aux données observées. Également, au niveau de la prédiction du zéro, ce modèle est le plus efficient dans sa prédiction avec un score tout prêt de ce que l'on observe. Au final, le modèle ZIP est très efficace dans sa prédiction du zéro et des moyennes prédites, mais la non-signification des résultats le relègue au second rang.

Le quatrième et dernier modèle testé dans ce mémoire est le ZINB. Dans les faits, ce modèle devait être le plus adéquat, car il tient compte de deux principales particularités des données, soit la surdispersion et l'inflation du zéro. Par conséquent, le modèle estime trois paramètres, la moyenne, le paramètre d'inflation et le paramètre de surdispersion. Ajouter à cela, des paramètres aléatoires au niveau du temps et de l'ordonnée à l'origine, il en résulte que le modèle semble surparamétrisé dans un contexte longitudinal. La plus grande lacune de ce modèle est son incapacité à prédire correctement la probabilité du zéro alors qu'il devrait être le plus efficace. En effet, une prédiction de près de 25% supérieure aux proportions observées est jugée inadéquate en comparaison aux trois modèles précédents où la différence moyenne varie de trois à cinq pourcents. Pour cette raison, le modèle semble beaucoup moins performant. Au niveau des résultats cependant, ceux-ci s'avèrent significatifs ce qui démontre une plus grande puissance statistique de ce modèle. De plus, les moyennes prédites sont presque identiques au modèle ZIP en prédisant convenablement les moyennes observées.

Lequel des quatre modèles est préférable dans sa prédiction des moyennes observées et dans sa prédiction de la probabilité du zéro pour le groupe des individus violents et le groupe des individus non-violents ? En se fiant aux comparaisons précédentes, le modèle binomial négatif semble être le modèle le plus adéquat pour traiter le nombre d'arrestations des membres de gangs de rue à Montréal. En effet, en plus de produire des résultats significatifs, il est assez près des valeurs observées avec une proportion de zéro très similaire à la distribution. Selon Greene (2007), la surdispersion doit être traitée en priorité ce que fait le modèle NB. À première vue, le ZINB semblait être préférable pour traiter en même temps un nombre important de zéros ainsi que la surdispersion des données. Par contre, comme il a été mentionné précédemment, le seuil d'une proportion de zéro trop importante n'a jamais été clairement défini. De plus, le paramètre aléatoire du temps ainsi que le paramètre de surdispersion semblent provoquer une surparamétrisation du modèle. Toujours selon Greene, avoir un grand nombre de zéros dans les données observées ne veut pas nécessairement dire que les modèles ZIP ou ZINB doivent être absolument employés comme le confirme ce mémoire. Cependant, le modèle ZIP fait également bien au niveau des moyennes prédites et de la prédiction de l'inflation. Par contre, les résultats ne s'avèrent pas significatifs.

# La conclusion et les limites

Il est nécessaire pour conclure ce mémoire d'exposer les différentes limites du mémoire par rapport aux analyses utilisées et également par rapport à l'échantillon sélectionnée.

Au niveau des analyses du mémoire, le fait d'avoir sélectionné uniquement trois points de mesure peut avoir joué sur la capacité prédictive de certains modèles. En effet, les modèles tenant compte de la surdispersion (soit le NB et ZINB) ont peut-être été influencé par le manque de variation du nombre d'arrestations à travers le temps. Comme il est observé dans la figure 4, la trajectoire observée est relativement stable à travers le temps, la possibilité d'y trouver beaucoup de variations peut être plus difficile. Également, une mesure avec trois temps, ne permet pas de tester un effet quadratique ou cubique du temps dans les modèles.

Deuxièmement, l'ajustement des modèles aux moyennes observées fut comparé à l'œil, il est donc difficile d'identifier les légères différences entre les modèles. Dans une analyse future, il serait intéressant de déterminer une méthode statistique permettant de comparer plus précisément l'ajustement des modèles aux données. Cependant, uniquement pour les modèles avec inflation du zéro, il existe une méthode peu connu et peu utilisé par les chercheurs, le test de Vuong (Vuong, 1989). Ce dernier semble intéressant mais fort complexe afin d'identifier le modèle qui s'ajuste le mieux entre le ZIP et ZINB.

Troisièmement, l'ajout d'une seule variable prédictive, en occurrence la violence, n'implique pas à elle seule, les variations du nombre d'arrestations. L'ajout de d'autres variables écologiques ou criminelles pourrait permettre de cibler davantage de variations non-observées.

Quatrièmement, la documentation propre aux analyses multinveaux longitudinales utilisant des modèles linéaires généralisés est peu fournie. Les modèles sont récents et en développement.

Pour terminer, les études en sociologie de la déviance et en criminologie contrôlent la propension ou l'activité criminelle du délinquant à l'aide des questionnaires auto révélés. Cette méthode permet de déterminer si l'individu est réellement actif sur la scène criminelle dans la période étudiée. Cependant, les membres de gang de rue sur lesquels se base cette étude ne peuvent être rencontrés considérant que les données proviennent des observations et des rapports de police. Il s'agit de données confidentielles ne pouvant être utilisées généralement pour des études afin de ne pas nuire aux enquêtes et aux procès en cour. De ce fait, il est impossible de déterminer si l'individu n'ayant aucune arrestation à éviter la répression du système judiciaire, est en prison ou si son activité criminelle est nulle ou très faible durant la période d'étude diminuant par le fait même ces risques d'arrestations. Deuxièmement, outre les crimes violents, les types de délits de moindre gravité comme les vols, le trafic de drogue ou même le proxénétisme ne sont pas inclus comme une variable contrôle dans les analyses. Il est possible que la police concentre leurs efforts et leurs ressources sur les individus les plus «dangereux» maximisant ainsi les risques d'arrestations. Finalement, la loi sur la protection de la jeunesse ne permet pas de travailler sur les individus mineurs. Généralement, le plus haut taux de criminalité se retrouve à l'adolescence. Les individus de l'échantillon à l'étude sont tous majeurs et réputés pour être moins actifs criminellement. Par contre, comme il a été mentionné précédent, Ouimet (2005) contredit cette théorie avec l'exemple des gangs de rue et du crime organisé.

# Bibliographie

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43), 21*(4), 726-735.

Blossfeld, H.-P., Hamerle, A., & Mayer, K. U. (1989). *Event History Analysis*. New Jersey: Lawrence Erlbaum Associates.

Blumstein, A. (1986). *Criminal Careers and Career Criminals*. Etats-Unis.

Blumstein, A., & Cohen, J. (1979). Estimation of Individual Crime Rates from Arrest Records. *The Journal of Criminal Law and Criminology, 70*(4), 561-585.

Chettleburgh, M. (2007). *Young Thugs: Inside the Dangerous World of Canadian Street Gangs*: HarperCollins Canada.

D'Élia, M. (2009). La violence chez les jeunes: Un portrait chiffré de la délinquance et de la victimisation (pp. 1-14). Section Recherche et Planification: Service de Police de la Ville de Montréal.

Delaney, T. (2006). *American Street Gangs*: Pearson Prentice Hall.

Descormiers, K. (2008). *Le réseau social des gangs montréalais : accès aux dynamiques relationnelles par l ' entrevue de groupe.* (Maitrise).

Farrington, D. P., & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health, 10*(2), 100-122.

Fournier, M., Cousineau, M.-m., & Hamel, S. (2011). La victimisation : un aspect marquant de l'expérience des jeunes filles dans les gangs *Criminologie, 37*(1), 149-166.

Fox, W. (1998). *Statisiques Sociale* (L. Imbeau, Trans.): DeBoeck University.
Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression Multilevel/Hierarchical Models*: Cambridge University Press.

Jean-Pierre Guay et Judith Gaumont, Le phénomène des gangs de rue au Québec, Rapport présenté au Gouvernement du Québec, 2009, 1-49

Greene, W. H. (1994). *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*.

Greene, W. H. (2007). *Functional Form and Heterogeneity in Models for Count Data*: Foundations and Trends in Econometrics.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo Method*. New York: Wiley.

Hébert, J., Hamel, S., & Savoie Ginette, J. (1997). «Jeunesses et gangs de rue», Phase I; Revue de la littérature.

Hilbe, J. M. (2007). *Negative Binomial Regression*: Cambridge University Press.

Howell, J. C. (2012). *Gangs in America's Communities*: SAGE Publications Inc.

Johnson, B. D. (2010). Multilevel Analysis in the Study of Crime and Justice *Handbook of Quantitative Criminology* (pp. 615-648): Springer.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics, 34*(1), 1-14.

MacDonald, J. M., & Lattimore, P. K. (2010). Count Models in Criminology *Handbook of Quantitative Criminology* (pp. 683-698): Springer.

Mullahy, J. (1986). Specification and testing of some modified count data. *Journal of Econometrics*(33), 341-365.

Singer, J. D., & B.Willet, J. (2003). Doing Data Analysis with the Multilevel Model for Change *Applied Longitudinal Data Analysis: Modeling change and event occurence* (pp. 63): Oxford University Press.

Singer, J. D., & Willet, J. B. (2003). Treating Time More Flexibly *Applied Longitudinal Data Analysis : Modeling change and event occurence* (pp. 644): Oxford University Press.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE Publications.

Ulrich Mayer, K., & Brandon Tuma, N. (1990). *Event history analysis in life course research* Wisconsin: The University of Wisconsin Press.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica, 57*(2), 307-333.

Wright, R., Decker, S. H., Redfern, A. K., & Smith, D. L. (1992). A Snowball's Chance in Hell: Doing Fieldwork with Active Residential Burglars. *Journal of Research in Crime and Delinquency, 29*(2), 148-161.

Yamaguchi, K. (1991). *Event History Analysis* (Vol. 28). London: SAGE Publications.

Yinglin, X., Morrison-Beedy, D., Ma, J., Changyong, F., Wendi, C., & Xin, T. (2012). Modeling Count Outcomes from HIV Risk Reduction Interventions: A comparaison of Competing Statistical Models for Count Responses. *AIDS Research and Treatment* 1-11.

# ANNEXES

## ANNEXE I- Les différentes fonctions de liens

**Common distributions with typical uses and canonical link functions**

| Distribution | Support of distribution | Typical uses | Link name | Link function | Mean function |
|---|---|---|---|---|---|
| Normal | real: $(-\infty, +\infty)$ | Linear-response data | Identity | $\mathbf{X}\boldsymbol{\beta} = \mu$ | $\mu = \mathbf{X}\boldsymbol{\beta}$ |
| Exponential<br>Gamma | real: $(0, +\infty)$ | Exponential-response data, scale parameters | Inverse | $\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$ | $\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$ |
| Inverse Gaussian | real: $(0, +\infty)$ | | Inverse squared | $\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$ | $\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$ |
| Poisson | integer: $[0, +\infty)$ | count of occurrences in fixed amount of time/space | Log | $\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$ | $\mu = \exp(\mathbf{X}\boldsymbol{\beta})$ |
| Bernoulli | integer: $[0, 1]$ | outcome of single yes/no occurrence | Logit | $\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$ | $\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$ |
| Binomial | integer: $[0, N]$ | count of # of "yes" occurrences out of N yes/no occurrences | | | |
| Categorical | integer: $[0, K)$ | outcome of single K-way occurrence | | | |
| | K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1 | | | | |
| Multinomial | K-vector of integer: $[0, N]$ | count of occurrences of different types (1 .. K) out of N total K-way occurrences | | | |