Compte rendu de statistiques : Les Mass-Shooting aux Etats-Unis

Par Antoine Buffandeau, Ruben Léon et Théo Lissarrague.

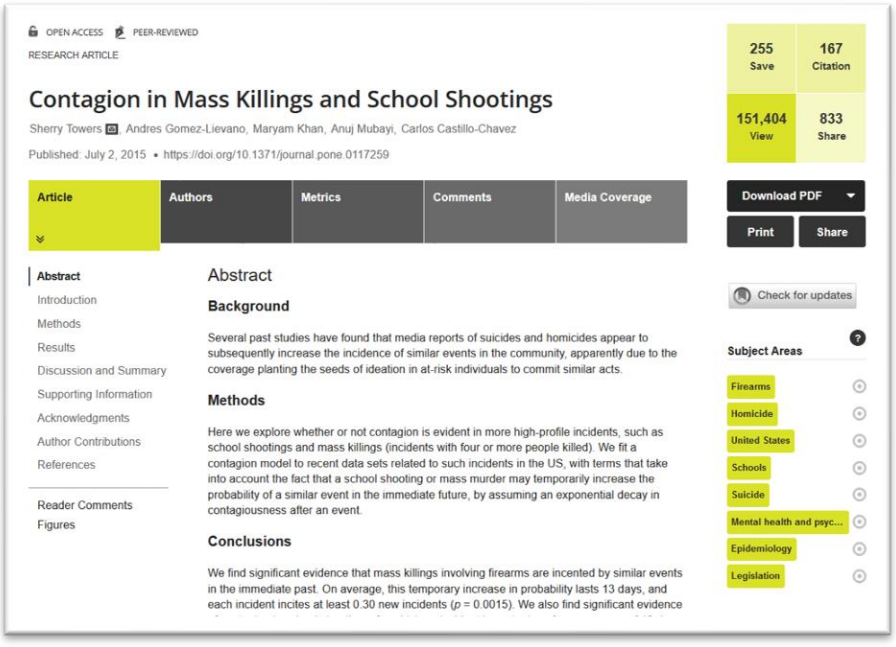
Table des matières

Préface

Nous avons réalisé ce projet durant notre cursus en classe préparatoire intégrée à l'école d'ingénieur ESILV à Paris La Défense. À la suite de notre enseignement en statistiques sous R qui impliquait la découverte du langage R, du logiciel RStudio et l'utilisation de celui-ci dans l'étude de statistiques diverses et variées. Nous avons suivi cet enseignement durant notre 2e année de classe préparatoire et nous devions réaliser une étude statistique sur un thème choisi.

Lors de la recherche d'un sujet, nous nous sommes focalisés sur des données officielles, si possible Étatique, pour avoir des données relativement objectives et permettre d'éviter un maximum de biais issue des sources elles-mêmes. Nous nous sommes alors tournés vers les chiffres de la criminalité dans le monde.

Au fils des débats nous avons évoqué l'article « Contagion in Mass Killings and School Shootings » de Sherry Towers publié en juillet 2015. Cet article de recherche évoque présente ces crimes comme une épidémie que l'on peut modéliser avec des outils statistiques développé pour lutter contre des maladies infectieuses. Nous souhaitions retrouver la même conclusion avec nos données ainsi que nos outils.


OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Contagion in Mass Killings and School Shootings

Sherry Towers, Andres Gomez-Lievano, Maryam Khan, Anuj Mubayi, Carlos Castillo-Chavez

Published: July 2, 2015 • https://doi.org/10.1371/journal.pone.0117259

255 Save | 167 Citation | 151,404 View | 833 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

Abstract
Introduction
Methods
Results
Discussion and Summary
Supporting Information
Acknowledgments
Author Contributions
References
Reader Comments
Figures

Abstract

Background

Several past studies have found that media reports of suicides and homicides appear to subsequently increase the incidence of similar events in the community, apparently due to the coverage planting the seeds of ideation in at-risk individuals to commit similar acts.

Methods

Here we explore whether or not contagion is evident in more high-profile incidents, such as school shootings and mass killings (incidents with four or more people killed). We fit a contagion model to recent data sets related to such incidents in the US, with terms that take into account the fact that a school shooting or mass murder may temporarily increase the probability of a similar event in the immediate future, by assuming an exponential decay in contagiousness after an event.

Conclusions

We find significant evidence that mass killings involving firearms are incented by similar events in the immediate past. On average, this temporary increase in probability lasts 13 days, and each incident incites at least 0.30 new incidents ($p = 0.0015$). We also find significant evidence

Check for updates

Subject Areas

Firearms | Homicide | United States | Schools | Suicide | Mental health and psyc... | Epidemiology | Legislation

Nonobstant, nous nous sommes rendu compte que nous sommes aujourd'hui trop limités par nos connaissances. Nous avons donc gardé le même set de données, et nous avons décidé de nous tourner vers l'étude comparative des préjugés sur les auteurs de ces crimes.

Introduction

La notion de tuerie de masse par un individu seul désigne l'assassinat de plusieurs personnes sur un même laps de temps. Des profileurs du FBI ont identifié cette notion comme le meurtre ou la tentative de meurtre d'au moins 4 personnes lors d'un évènement particulier. Elle se déroulent généralement dans un lieu unique et se soldent par le suicide ou l'élimination par les forces de l'ordre de l'auteur. Pour des raisons d'efficacité, ces attaques se déroulent généralement à l'arme à feu d'où le terme anglais de « Mass Shooting ». Les raisons sont généralement racistes, terroristes, sociales (harcèlement, licenciement litigieux, etc.) et/ou dues à des troubles psychologiques.

Cette notion se différencie donc de la tuerie en séries qui est un enchaînement de meurtres, généralement plus discret dans différentes unités de lieux et d'espaces et de la tuerie de masse perpétré par un État par l'intermédiaire de ces agents (génocide, assassinat de groupe politique, massacre de prisonniers ou de civils en temps de guerre, etc.).

Pour beaucoup d'entre nous, il est compliqué de comprendre les justifications de ces meurtres. Nombre d'entre nous ont alors tendance à déshumaniser les auteurs de ces crimes. Au 19ème siècle, on étudiait le corps des tueurs pour mettre en évidence le physique type qui trahissait la nature criminelle de ces individus. Encore aujourd'hui, en fonction des sociétés, on s'imagine qu'une partie de la population est à l'origine de la majorité de ces tueries.

Nous avons donc décidé de mener une étude comparative sur les auteurs des différentes tueries de masse pour voir s'il est possible de déterminer un éventuelle profil type. Nous préjugeons d'une grande influence de la société sur ces données donc nous nous sommes limités uniquement à un pays. Nous avons étudié les Etats-Unis car ils ont été les premiers à avoir défini et à utiliser la notion de tueur de masse et leurs bases de données sur les criminels sont très riches.

Problématique

En nous aidant de certaines des archives étatiques des Etats-Unis en matière de fusillade de masse, nous nous poserons la problématique suivante :

Le bilan humain d'un mass shooting est-il dépendant du type de personne qui l'a commis ?

Nous tenterons de répondre à cette problématique en deux actions distinctes :

- La préparation de notre base de données ;
- L'analyse de cette dernière en effectuant des statistiques bivariées sur l'ensemble des données statistiques choisies.

Ceci dans un objectif final d'essayer de dresser un profil type d'un tueur de masse.

Présentation de la base de données

Nous avons choisi pour support de notre étude, une base de données issu du site Kaggle (cliquez ici pour y accéder). Ce site est en effet très répandu et réputé pour tout ce qui concerne l'obtention de données utilisables gratuitement. Cette base de données est composée de plusieurs versions et nous avons choisit d'exploiter celle nommée « Mass Shootings Dataset Ver 5.csv » car c'est celle qui contenait le plus de critères différents.

Nous avons par la suite passé un temps conséquent à homogénéiser, choisir et rendre exploitables les données issues de cette base, pour ne garder qu'à la fin cinq critères dont trois qualitatifs (le genre du tireur, son origine ethnique et sa condition mentale) et deux quantitatifs (son âge et le nombre total de victimes, morts et blessés).

Nous supposons que la base de données que nous avons sélectionnée est un regroupement de différentes autres car nous avons dû réaliser tout le code suivant pour la rendre exploitable :

```
setwd(dir= "C:\\Users\\ruben\\Documents\\R\\Repos")
MS=read.csv("Mass Shooting.csv",sep=",", row.name = 1)

#selection et uniformisation des données
MS = MS[c(3,9,10,11,13,16,17,18)]  #selectionne bonnes colonnes
MS = MS[c(155:323),c(1:8)]           #selectionne bonnes lignes
row.names(MS) <- 1:nrow(MS)          #remet compteur à 1


MS$Gender = gsub("Female","F",MS$Gender)
MS$Gender = gsub("Male","M",MS$Gender)

MS$Race = gsub("White American or European American","White",MS$Race)
MS$Race = gsub("White","White",MS$Race)
MS$Race = gsub("White/Some other Race","White",MS$Race)
MS$Race = gsub("Black American or African American","Black",MS$Race)
MS$Race = gsub("Black","Black",MS$Race)
MS$Race = gsub("Black/Unknown","Black",MS$Race)
MS$Race = gsub("Native American or Alaska Native","Native",MS$Race)
MS$Race = gsub("Asian American","Asian",MS$Race)
MS$Race = gsub("Asian American/Other","Asian",MS$Race)
MS$Race = gsub("Asian/Other","Asian",MS$Race)
MS$Race = gsub("Some other race","Other",MS$Race)
MS$Race = gsub("Two or more races","Other",MS$Race)

MS$Age = gsub(",.*","",MS$Age)
MS <- MS[c(-7,-26,-37,-84,-112,-115,-136,-167,-5),]    #supprime ligne age nul
row.names(MS) <- 1:nrow(MS)                            #remet compteur à 1
```

Ainsi, nous avons commencé par importer la base de données, puis sélectionner les lignes et colonnes que nous considérions comme utiles à notre étude. Puis nous avons beaucoup la commande gsub qui permet de remplacer un mot mit en paramètre par un autre, ce qui nous à permit de poser une nomenclature fixe aux critères pour faciliter l'exploitation qui va suivre (par exemple dans cette base le sexe masculin pouvait être décrit par « Male » ou « M », et nous avons arbitrairement choisi de fixer l'appellation « M »). Enfin, nous avons supprimé les lignes incomplètes car nous nous sommes rendu compte qu'une case vide pose des problèmes pour les fonctions nécessaires au traitement.

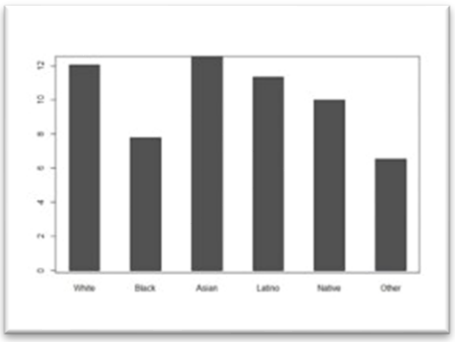
Nous avions donc finalement la base de données dont voici l'aperçu :

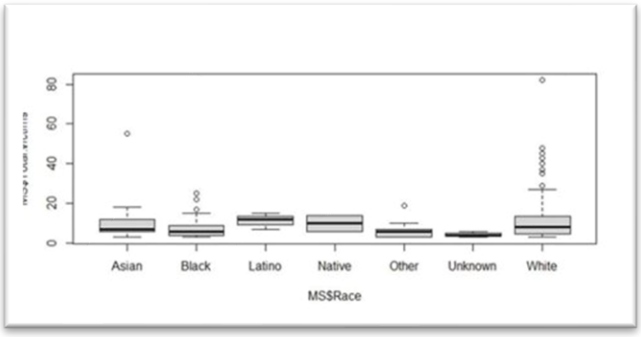
	Date	Fatalities	Injured	Total.victims	Age	Mental.Health.Issues	Race	Gender
1	5/23/2014	6	13	19	22	Yes	White	M
2	4/29/2014	1	6	6	19	No	White	M
3	4/3/2014	3	12	15	34	Unclear	Latino	M
4	4/2/2014	4	16	19	34	Yes	Other	M
5	2/20/2014	4	2	6	44	No	Native	F
6	11/1/2013	1	3	4	23	No	White	M
7	10/21/2013	2	2	3	12	Unknown	Other	M
8	9/16/2013	13	3	15	34	Yes	Black	M
9	8/14/2013	4	0	4	40	Yes	White	M
10	8/7/2013	4	4	8	44	Unknown	Black	M
11	8/5/2013	3	3	6	59	No	White	M
12	7/26/2013	7	0	7	42	Unclear	Latino	M
13	6/7/2013	6	3	8	23	Yes	White	M
14	4/24/2013	7	1	7	43	Unknown	White	M
15	4/21/2013	5	0	4	27	Unknown	Black	M
16	3/13/2013	5	2	6	64	No	White	M
17	2/19/2013	4	3	6	20	Yes	White	M
18	2/3/2013	4	2	7	33	Yes	Black	M
19	1/30/2013	3	1	3	70	Unknown	White	M
20	1/19/2013	5	0	5	15	Yes	Other	M
21	12/14/2012	28	2	29	20	Yes	White	M
22	12/11/2012	3	1	3	22	No	Other	M
23	10/21/2012	4	4	7	45	No	Black	M
24	9/27/2012	7	2	8	36	Yes	White	M
25	8/5/2012	7	4	10	40	No	White	M
26	7/20/2012	12	70	82	24	Yes	White	M

Comparaison des données qualitatives

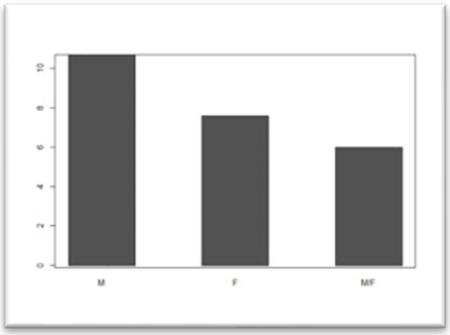
Pour commencer, intéressons-nous aux trois critères qualitatifs, soient le genre, l'ethnie et la condition mentale du tireur. Pour ce faire, nous allons faire une comparaison bivariée avec le nombre de victime pour chacun de ces critères.

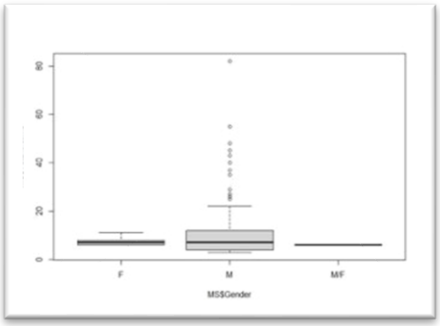
Ainsi, après l'application des fonctions de R nécessaires, nous obtenons ceci :



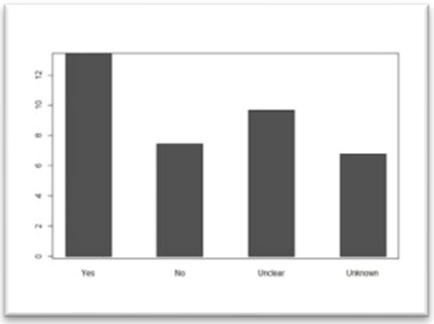


Nb victimes d'après l'ethnie du tireur





Nb victimes d'après le sexe du tireur



Nb victimes d'après la condition mentale du tireur

Pour tester ensuite la corrélation entre toutes ces données, nous avons calculé le coefficient e^2 entre ces valeurs sachant que ce coefficient est donné par la formule :

$$e^2 = \frac{variance\ inter - groupe}{variance\ totale}$$

Sachant que plus ce coefficient est proche de 1, plus les données sont corrélées.

Ainsi, nous avons calculé la variances inter groupe puis la variance totale pour chacun des critères, puis le coefficient en question avec le code suivant (appliqué ici à l'ethnie) :

```
> vartot = var(MS$Total.victims)
> R=c("White","Black","Asian","Latino","Native","Other")
> varinter = rep(0,6)
> for(i in range(6)){
+   varinter[i] = var(MS[MS$Race == R[i],4])
+ }
> ecarre = sum(varinter)/vartot
> print(ecarre)
[1] 0.1932295
```

Et nous trouvons alors les coefficients suivants :

-e^2(ethnie) = 0,1932295

-e^2(condition mentale) = 0,08733021

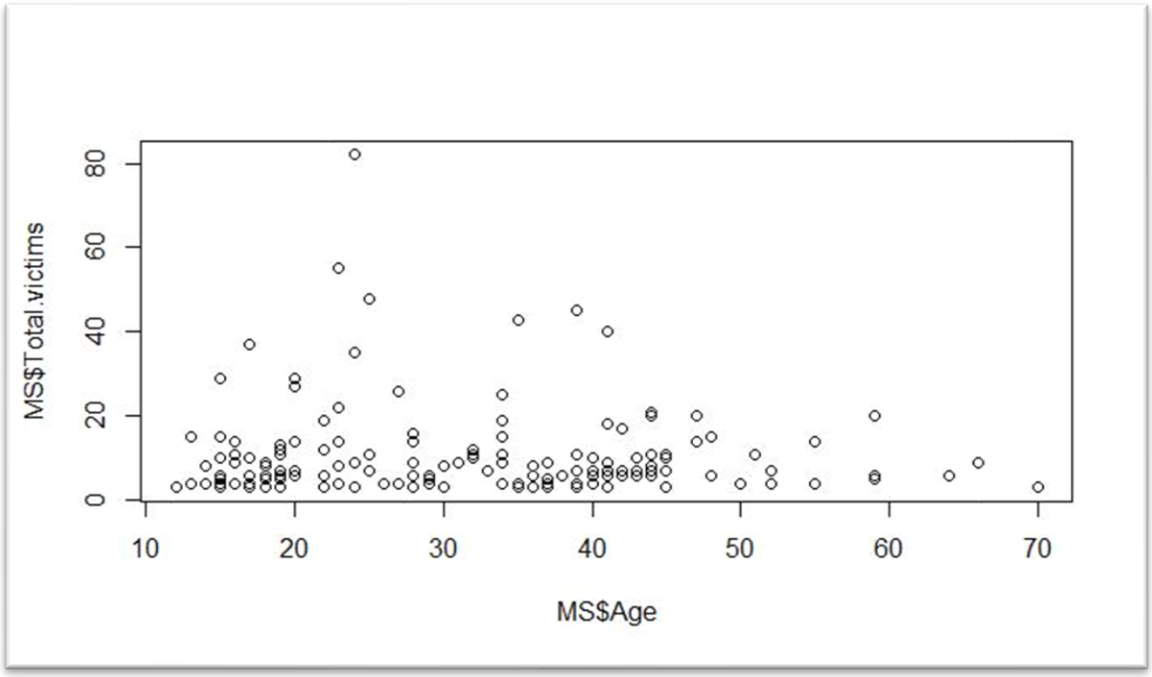
-e^2(sexe) = non défini

Ainsi, nous voyons que nos données sont peu corrélées.

Comparaison des données quantitatives

Afin de résoudre notre problématique et de savoir si un profil type serait plus à même ou non de réaliser une fusillade de masse, il s'avère également utile de déterminer s'il existe une corrélation entre le nombre de victimes et l'âge du tueur.

Pour cela, nous avons réalisé dans un premier temps un nuage de point afin de visualiser si la représentation des mesures de ces deux variables quantitatives semble démontrer l'existence d'une potentielle relation entre ces deux dernières.



Graphique n° 1 : Nuage de point représentant le nombre de victimes d'un auteur d'un mass shooting en fonction de l'âge de ce dernier (sans droite de tendance).

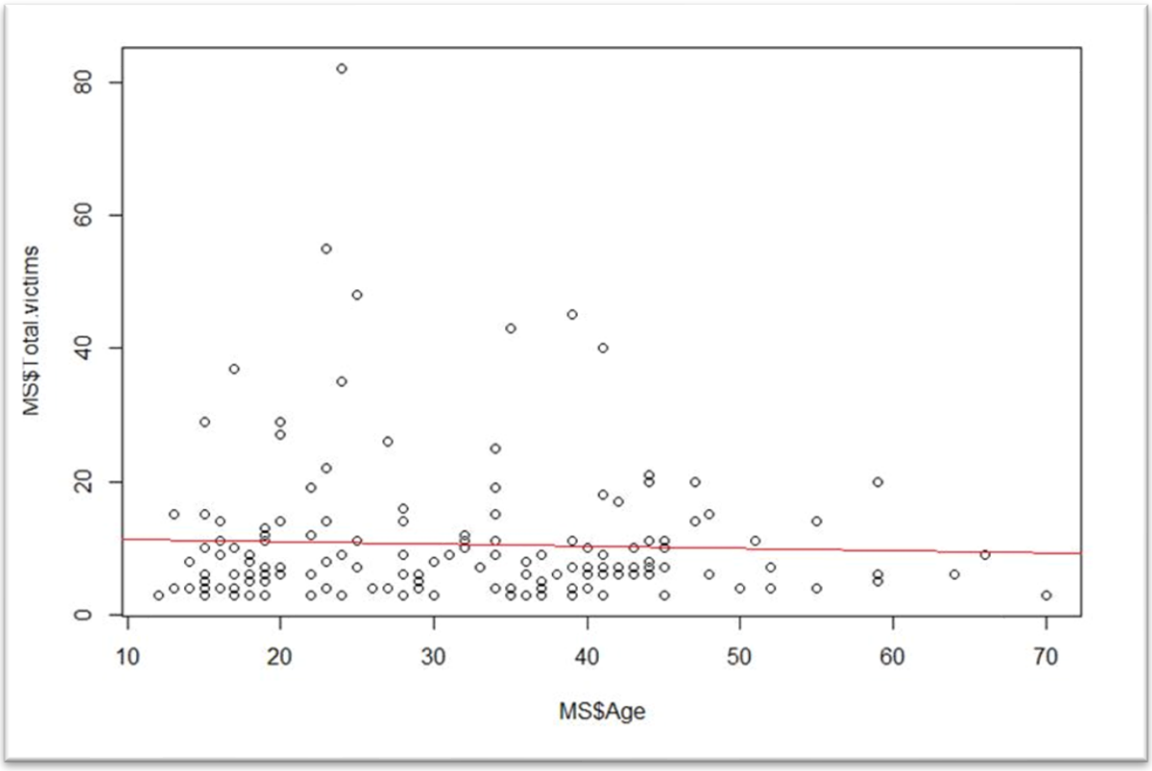
A première vue, les points sont concentrés et dispersés par endroit et ne semblent pas former de lignes distinctes. Il semblerait qu'il n'existe donc pas de relation linéaire au premier regard (1ère hypothèse).

Par ailleurs, les points semblent regroupés en deux parties distinctes : l'une de 10 à environ 26 ans dans laquelle les points vont du coin inférieur gauche au coin supérieur droit et l'autre d'environ 26 à 70 ans dans laquelle c'est l'inverse : les points vont du coin supérieur gauche au coin inférieur droit. La seconde hypothèse de l'existence d'une relation positive, puis négative respectivement, peut alors être émise entre ces deux variables.

Remarque :

L'élaboration de ce nuage de point semble également mettre en évidence l'existence de certaines valeurs extrêmes. Certaines valeurs semblent en effet assez éloignées par rapport à l'ensemble de la tendance (exemple : E_{xemple} (26 ;80)). Des valeurs qui pourraient faussées nos résultats par la suite.

Afin de vérifier nos hypothèses, nous allons faire usage des outils à notre disposition et vérifier si notre logiciel ne trouve pas de relation linéaire positive ou négative en réalisant une régression linéaire.



Graphique n° 1bis : Nuage de point représentant le nombre de victimes d'un auteur d'un mass shooting en fonction de l'âge de ce dernier (avec droite de tendance).

La droite de régression linéaire obtenue se rapproche alors d'une droite constante. Un résultat pas très pertinent et concluant vis-à-vis de notre problématique. Nous ne pouvons donc pas encore conclure directement sur une potentielle relation linéaire avec ce modèle de régression linéaire.

Afin de résoudre notre problème nous allons alors vérifier le test de corrélation de Pearson qui permet de mesurer le degré de liaison entre ces deux variables quantitatives étudiées, autrement dit dans le but de vérifier ou non l'existence d'une corrélation entre ces variables.
Formule du coefficient de corrélation de Pearson :

$$r_{x,y}{}^2 = \left(\frac{cov(x,y)}{\sigma_x \sigma_y}\right)^2$$

Avec :

- x : la variable quantitative représentant le nombre de victime d'un auteur d'une fusillade de masse ;
- y : la variable quantitative représentant l'âge d'un auteur d'une fusillade de masse ;
- $r_{x,y}{}^2$: le coefficient de détermination ;
- $\sigma_x = \sqrt{var(x)}$ et $\sigma_y = \sqrt{var(y)}$

Les résultats obtenus à l'aide du logiciel RStudio sont les suivants :

```
> cor.test(MS$Age, MS$Total.victims, method = "pearson")

        Pearson's product-moment correlation

data:  MS$Age and MS$Total.victims
t = -0.49739, df = 158, p-value = 0.6196
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1935105  0.1163335
sample estimates:
        cor
-0.03953895
```

Le coefficient de corrélation $r_{x,y} = 0.001563$ étant très proche de 0, il semble montrer que la relation entre ces deux variables est nulle.

Or, de plus, d'après ces résultats, l'hypothèse alternative nous précise que la corrélation n'est pas égale à 0 (« true correlation is not equal to 0 »). Cela suggère donc que la relation entre nos deux variables quantitatives ne peut être attribuée au hasard. Il existe donc bien une relation entre nos deux jeux de données quantitatives mais nous ne sommes toujours pas sûr de sa nature linéaire car $r_{x,y} \approx 0$ ce qui peut souligner la faiblesse de pertinence de ce modèle.

Nous nous sommes alors intéressés à un second test de corrélation : celui de Spearman afin d'essayer d'avoir une réponse précise.
Nous avons alors utilisé à nouveau RStudio pour déterminer le coefficient de corrélation de Spearman et avons obtenu les résultats suivants :

```
> cor.test(MS$Age, MS$Total.victims, method = "spearman")

        Spearman's rank correlation rho

data:  MS$Age and MS$Total.victims
S = 615410, p-value = 0.2153
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09848515
```

En comparant le résultat du coefficient de Spearman avec celui de Pearson, nous avons remarqué alors que celui de corrélation de Spearman est supérieur à celui de Pearson. Une observation qui indique par convention une éventuelle existence d'une relation non linéaire entre ces deux variables quantitatives.

Nous pouvons donc conclure à la véracité d'une relation non linéaire existante entre le nombre de victimes par auteur de fusillade de masse avec l'âge de cet auteur. Cependant, la possibilité d'un cas biaisé dans notre modèle réside toujours. Une éventualité qui contrarie nos résultats et qui n'affirme pas à 100% la véracité de cette conclusion.

Un lien semble donc existant mais sans certitude absolue !

Limite de l'étude

Notre étude se heurte à plusieurs limites. Dans un premier temps nous regrettons l'absence d'évocation d'éléments que nous aurions pu relier au niveau social des auteurs de mass shooting. En effet, l'indication du patrimoine, du revenu ou simplement de l'adresse de l'auteur relié au prix moyen des loyers dans le quartier aurait pu nous permettre de voir s'il y a une corrélation linéaire entre la précarité et les mass shooting. De la même façon, une évocation du relationnel (solitaire, sociable, célibataire, en couple, en famille), d'un casier judiciaire ou d'une addiction de la personne aurait pu permettre de dresser un profil ou au contraire montrer la non-corrélation de ces variables.

De plus, l'évocation de la préméditation ou non des mass shooting aurait été intéressante à analyser. Est-ce que le déchaînement de violence d'un individu sur un coups de tête est plus dangereux que l'organisation méthodique d'une attaque.

Enfin, nous nous sommes limités à des études de corrélations linéaires alors que certaines données, notamment l'âge, donnent l'impression qu'il y a une corrélation curviligne. De plus, les différents paramètres ont été analysés un par un et non dans leur ensemble. Il est possible que plusieurs profils bien défini (Hommes âgés avec des problème mentaux et hommes jeunes sans problèmes mentaux) aient brouillé notre analyse comparative.

Conclusion

Après la préparation de notre base de données, issue de plusieurs bases de données recensées par certains états des Etats-Unis en matière de fusillade de masse, nous avons analysé nos données en réalisant des analyses statistiques bivariées sur l'ensemble des données que nous avons sélectionnées. Ceci dans le but de déterminer si certains critères ont un impact sur le nombre de victimes réalisées lors d'une fusillade de masse ; ce qui nous permettrait de dresser un profil type d'auteur de fusillade de masse.

Nous avons donc d'abord effectué une comparaison des données qualitatives mais avons déterminé que ces données n'étaient pas corrélées. Ainsi, d'après nos analyses il s'avèrerait que le nombre de victime réalisé par un auteur d'une fusillade de masse ne dépendrait ni de l'ethnie, ni du sexe et ni de la condition mentale de cet individu.

De même, nous avons tenté de trouver une corrélation entre le nombre de victimes recensées lors d'un mass-shooting et l'âge de l'auteur. Cette analyse a abouti sur le fait que nous ne pouvions pas conclure avec certitude sur l'existence d'une telle corrélation. Ainsi, nous ne pouvons pas affirmer avec certitude que l'âge d'un auteur impliquerait l'hypothèse d'un nombre de victime plus élevé.

Ainsi, à la suite de cette étude, nous ne pouvons conclure avec certitude sur la potentielle corrélation entre nos critères sélectionnés et le nombre élevé de victime. Nous ne pouvons donc pas dresser un profil type précis sur les auteurs de fusillade de masse. Néanmoins, ce projet, même s'il aboutit à une incertitude, a été l'occasion de découvrir plus concrètement le logiciel RStudio et d'appliquer les notions de statistiques vues en cours tout en découvrant le domaine complexe que forment les fusillades de masse. Ainsi, même si nous ne pouvons conclure sur un résultat concret et précis, ce projet nous a permis d'enrichir nos compétences dans divers domaines, ce en quoi nous en sommes particulièrement fiers.

Bibliographie :

- Base de données : https://www.kaggle.com/datasets/zusmani/us-mass-shootings-last-50-years
- Logiciel de Statistiques sous R : https://www.rstudio.com/