# Probabilités et Statistiques
# Rapport TP 1 et 2

Eldis YMERAJ
Tolgan SUNER

17 Décembre 2023

# Table des matières

# Introduction

Dans la première partie des travaux pratiques, nous allons manipuler différentes lois de probabilités dans le but de les appliquer à des situations concrètes. Il s'agira également de démontrer la fiabilité de nos choix en calculant les intervalles de confiance pour nos expériences.
La seconde partie sera consacrée à la régression linéaire, une technique cruciale dans l'analyse des relations entre variables. Nous examinerons l'ajustement affine pour modéliser la dépendance entre la quantité de glaces vendues et la température de midi sur un jour donné. Les méthodes des moindres carrés et d'optimisation seront détaillées, avec une mise en œuvre pratique. L'objectif sera de déterminer un modèle de régression linéaire capable de prédire les ventes de glaces en fonction de la température.

# 1 Partie I : Les lois de probabilités

## 1.1 Lois discrètes

Une loi discrète est associée à une variable aléatoire discrète. Une variable aléatoire (v.a.) discrète est une v.a. qui prend des valeurs distinctes.

### 1.1.1 Loi de Poisson $P(\lambda)$

La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance connue et ce de manière indépendante du temps écoulé depuis l'événement précédent.
Dans le graphe ci-dessous, on va comparer la densité simulée et théorique :

- Simulée : la fonction de la librairie numpy `np.random.poisson`
- Répartition simulée : calculer grâce à la fréquence de chaque valeur issue de la fonction simulée
- Théorique : $P(X = i) = \frac{\lambda^i}{k!} e^{-\lambda}$
- Répartition théorique : $F(x) = \sum_{k=0}^{i} \frac{\lambda^k}{k!} e^{-\lambda}$

**Jeu d'essais :**

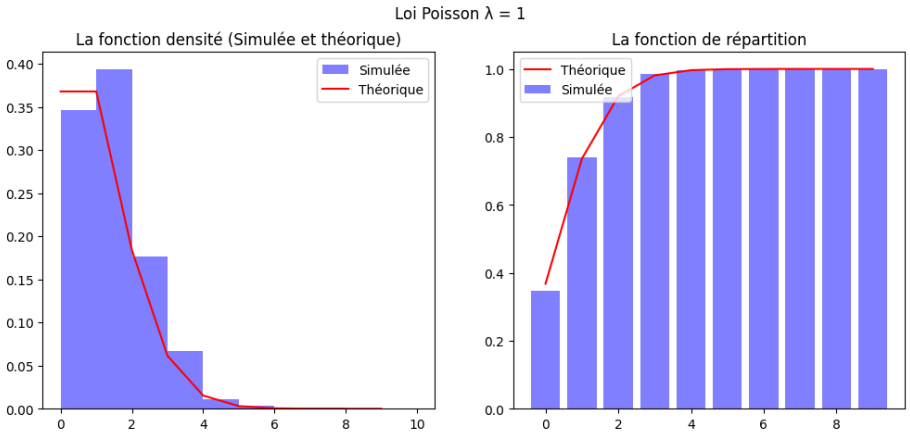


FIGURE 1 – Loi de Poisson simulée et théorique pour $\lambda = 1$ et Repartition.

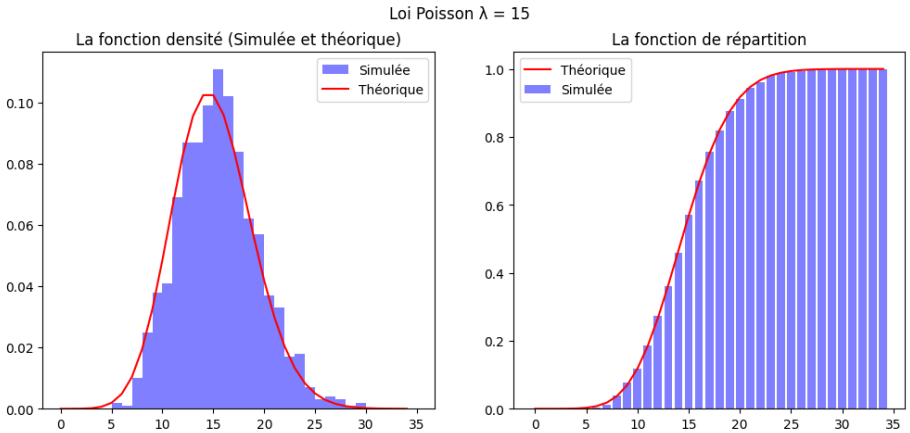


FIGURE 2 – Loi de Poisson simulée et théorique pour $\lambda = 15$ et Repartition.

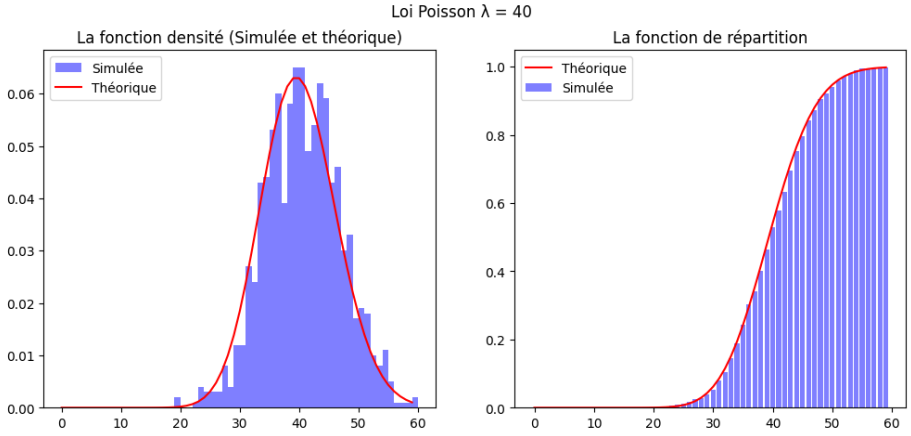


FIGURE 3 – Loi de Poisson simulée et théorique pour $\lambda = 40$ et Repartition.

### 1.1.2 Loi binomial $B(n, p)$

La loi binomiale modélise la fréquence du nombre de succès obtenus lors de la répétition de plusieurs expériences aléatoires identiques et indépendantes. Plus mathématiquement, la loi binomiale est une loi de probabilité discrète décrite par deux paramètres : `n`, le nombre d'expériences réalisées et `p`, la probabilité de succès. La variable aléatoire, somme de toutes ces variables aléatoires, compte le nombre de succès et suit une loi binomiale.
Dans le graphe si dessous on va comparer la densite simulée et théorique :

- Simulée : la fonction de la librairie numpy `np.random.binomial`
- Répartition simulée : calculer grâce à la fréquence de chaque valeur issue de la fonction simulée
- Théorique : $P(X = i) = \binom{n}{i} p^i q^{n-i}$, avec $p \in [0, 1]$ et $q = 1 - p$
- Répartition théorique : $F(x) = P(X \leq x) = \begin{cases} 1 & \text{si } x \geq n, \\ \sum_{k=0}^{x} \binom{n}{k} p^k q^{n-k} & \text{si } 0 \leq x < n, \\ 0 & \text{si } x < 0. \end{cases}$

**Jeu d'essais :**

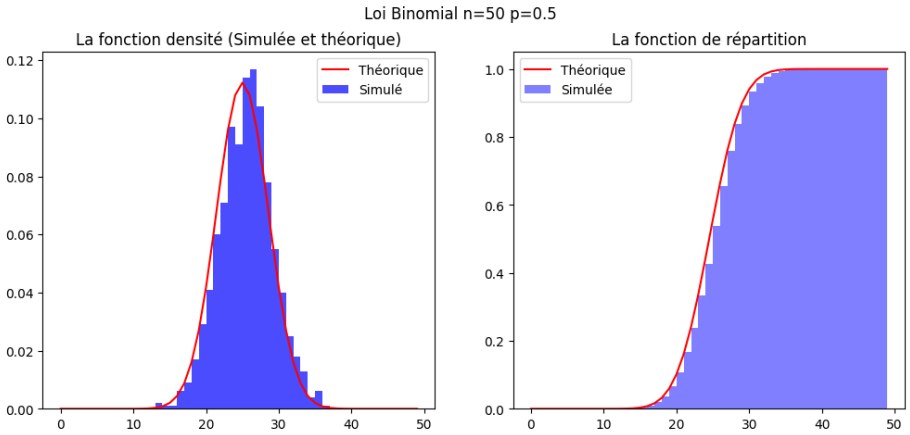


FIGURE 4 – Loi Binomial simulée et théorique pour n=50, p=0.5 et Repartition.

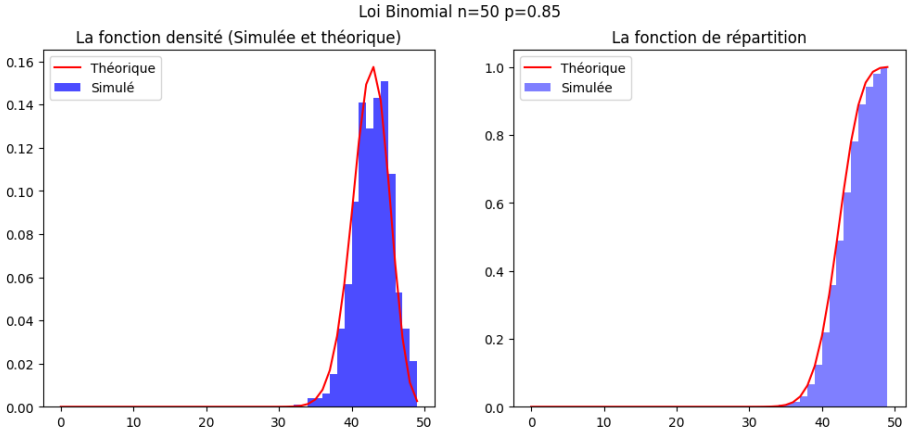


FIGURE 5 – Loi Binomial simulée et théorique pour n=50, p=0.85 et Repartition.

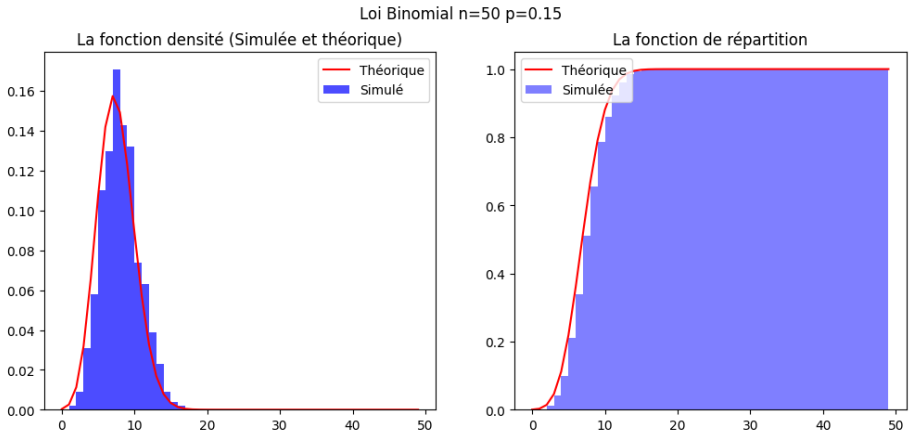


FIGURE 6 – Loi Binomial simulée et théorique pour n=50, p=0.15 et Repartition.

### 1.1.3 Loi de Poisson et Binomial

La loi binomiale peut être approximée par la loi de Poisson si le nombre d'essais $n$ est grand (généralement $n \geq 50$) et la probabilité de succès $p$ est petite, de sorte que le produit $\lambda = np$ reste modéré. Dans notre cas, nous n'avons pas pu dépasser $n = 150$ car les calculs impliquant la fonction factorielle, telle qu'elle apparaît dans la loi de Poisson, dépassent nos capacités de calcul.

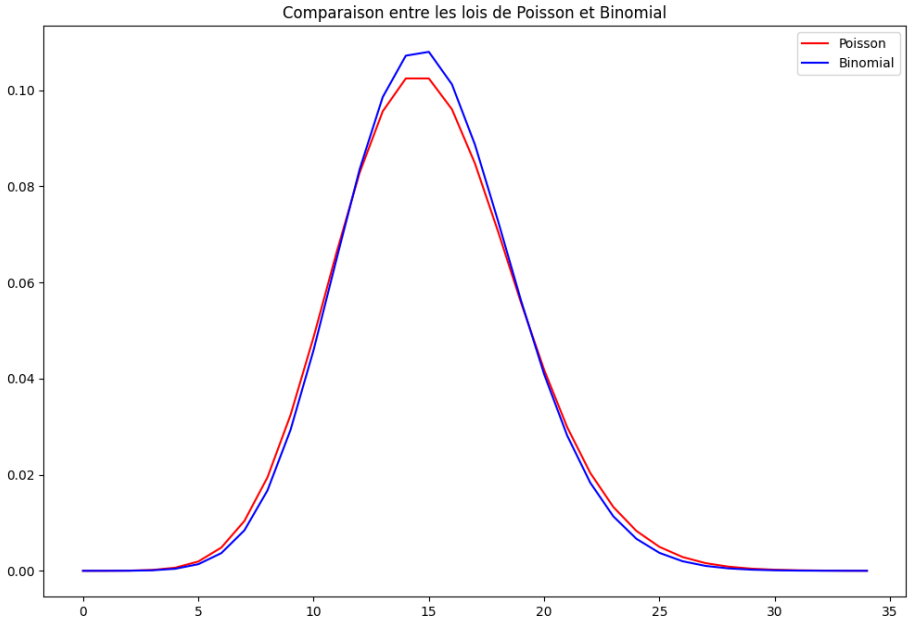

FIGURE 7 – Loi de Poisson et binomial pour $\lambda = np$, n=150 et p=0.1.

## 1.2 Lois continues

Une loi de probabilité continue est associée à une variable aléatoire continue, cette dernière étant une variable qui peut prendre n'importe quelle valeur dans un intervalle donné plutôt que des valeurs distinctes. La distribution de probabilité continue est caractérisée par une fonction de densité de probabilité, où la variable aléatoire se situe dans un intervalle particulier.

### 1.2.1 Loi Normal $N(\mu, \sigma)$

Une loi normale, est une loi de probabilité absolument continue qui dépend de deux paramètres : son espérance, un nombre réel noté $\mu$, et son écart-type, un nombre réel positif noté $\sigma$.
Dans le graphe ci-dessous, on va comparer la densité simulée et théorique :

- Simulée : la fonction de la librairie numpy `np.random.normal`
- Répartition simulée : calculer grâce à la fréquence de chaque valeur issue de la fonction simulée
- Théorique : $f(x) = \frac{e^{\frac{-1}{2}\frac{x-\mu}{\sigma}}}{\sigma\sqrt{2\pi}}$, avec $\mu =$, $\sigma =$
- Répartition théorique : $F(x) = \frac{1}{2}(1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})$, ou $erf$ est une fonction d'erreur $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}$

**Jeu d'essais :**

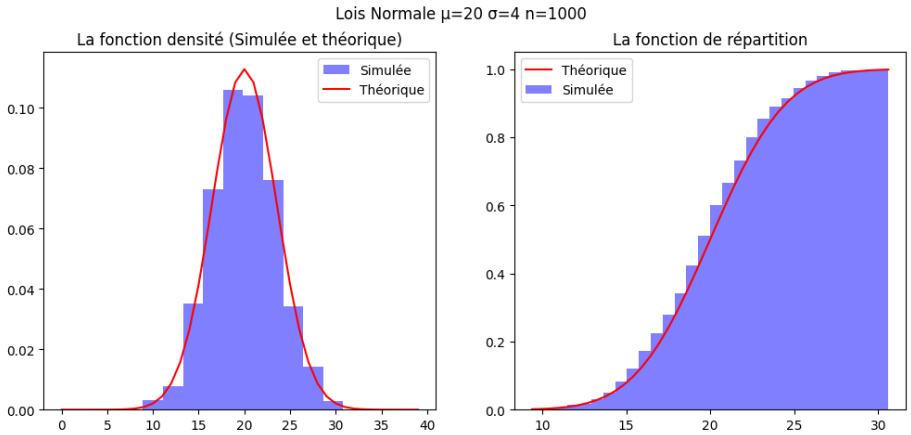


FIGURE 8 – Loi normale simulée et théorique pour n=50, p=0.15 et Repartition.

### 1.2.2 Loi Exponentielle $E(\lambda)$

La loi exponentielle est une loi de probabilité continue qui modélise le temps entre des événements indépendants et de même nature, se produisant à un

rythme constant. Plus simplement, elle décrit la distribution du temps écoulé jusqu'à ce qu'un événement spécifique se produise. Cette loi est caractérisée par un seul paramètre, le taux de la distribution, noté $\lambda$, qui représente la fréquence moyenne d'occurrence des événements.
Dans le graphe si dessous on va comparer la densite simulée et théorique :

- Simulée : la fonction de la librairie numpy `np.random.exponential`
- Répartition Simulée : calculer grâce à la fréquence de chaque valeur issue de la fonction simulée
- Théorique : $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$
- Répartition théorique : $F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$

**Jeu d'essais :**

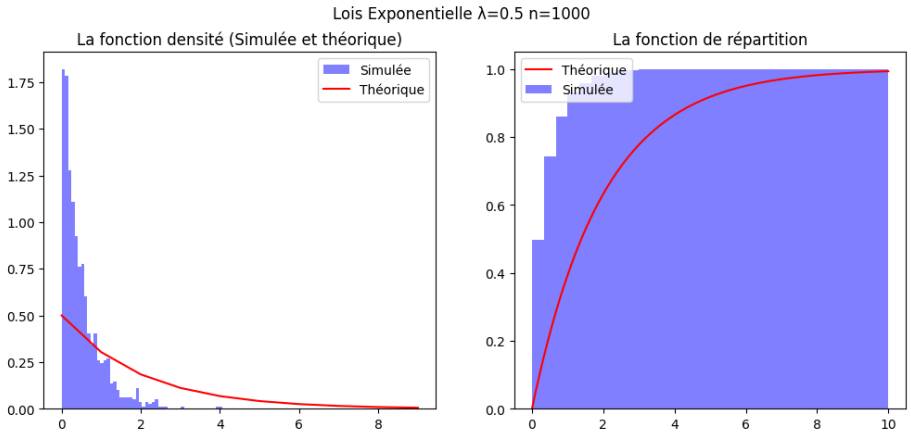


FIGURE 9 – Loi exponentielle simulée et théorique pour $\lambda = 0.5$ et Repartition.

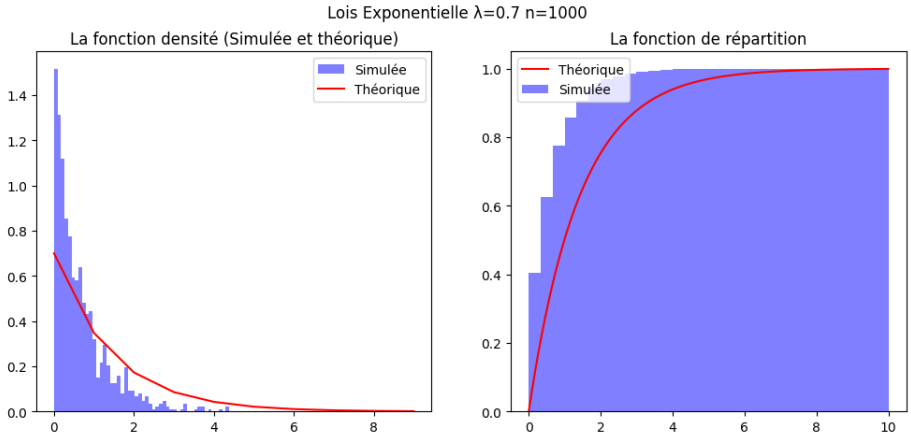

FIGURE 10 – Loi exponentielle simulée et théorique pour $\lambda = 0.7$ et Repartition.

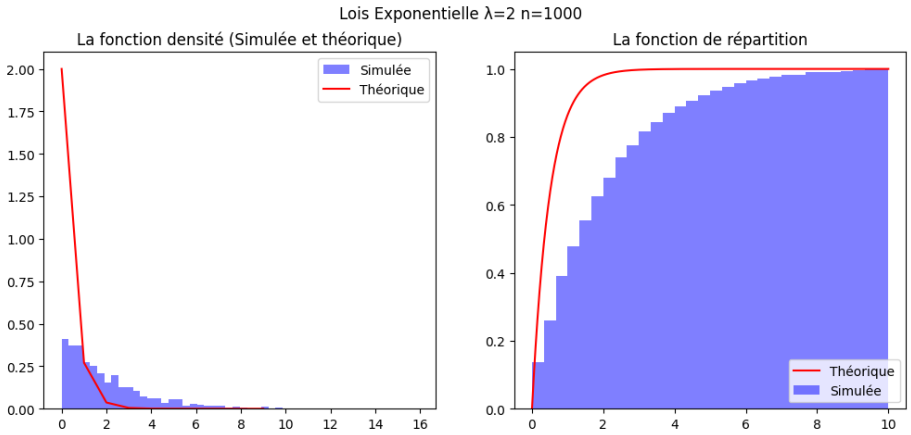


FIGURE 11 – Loi exponentielle simulée et théorique pour $\lambda = 2$ et Repartition.

# 2 Intervalle de confiance

Un intervalle de confiance est une plage de valeurs associée à un paramètre statistique, calculée à partir d'un échantillon de données. Il fournit une estimation de l'intervalle probable dans lequel se situe le vrai paramètre de la population avec une certaine probabilité. En d'autres termes, c'est une mesure de l'incertitude entourant notre estimation.

## 2.1 Temps de réaction

Nous traiterons un jeu de données des temps de réaction en supposant qu'ils suivent une loi normale et calculerons un intervalle de confiance pour chacun des deux points demandés en fonction des informations données.
Liste des temps de réaction : V = {0.98, 1.4, 0.84, 0.86, 0.54, 0.68, 1.35, 0.76, 0.79, 0.99, 0.88, 0.75, 0.45, 1.09, 0.68, 0.60, 1.13, 1.30, 1.20, 0.91, 0.74, 1.03, 0.61, 0.98, 0.91}

Pour le premier point, nous supposerons que l'on a une variance $\sigma^2 = 0.25$.
Calcul de la moyenne empirique :
Soient $X_1, ..., X_n$ des variables aléatoires (deux à deux) indépendantes, de même loi, d'espérance m et de variance $\sigma^2$. La moyenne empirique de $X_1, ..., X_n$ est la variable aléatoire $\bar{X}_n = \frac{X_1+...+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

On a alors n=25 et $\bar{X}_{25} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx 0.898$, pour $X_i \in V$

Nous connaissons la variance, nous pouvons donc lire la valeur dans la table des fractiles. Pour ce faire, on calcule $\alpha$ pour 1-$\alpha = 0.95 \Leftrightarrow \alpha = 0.05$ on lit la valeur 1.6449 et 1-$\alpha = 0.99 \Leftrightarrow \alpha = 0.01$ on lit la valeur 2.3263.

**Jeu d'essais :**

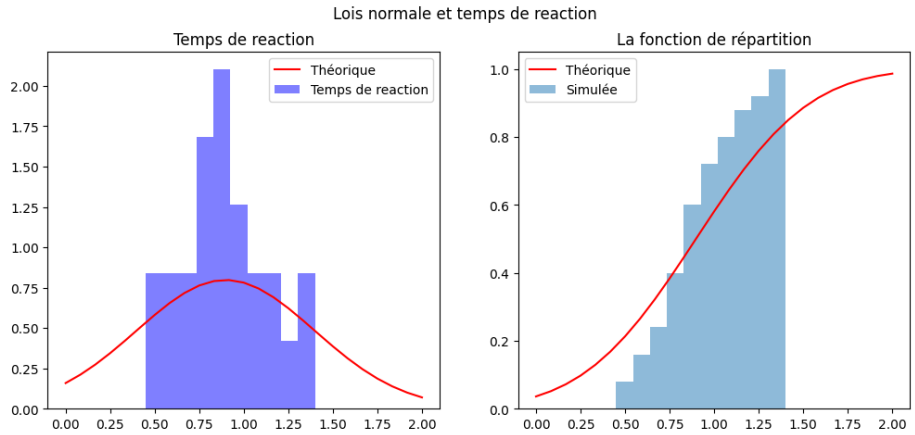


FIGURE 12 – Temps de réaction simulé et théorique par une loi normale pour $\lambda = 2$ et repartition.

Calcule de l'intervale de confiance :
IC= $\left[\bar{X}_n - t_\alpha \frac{s}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{s}{\sqrt{n}}\right]$
On note $s = \sqrt{\sigma^2}$ ici $\sigma^2 = 0.25$ alors $s = \sqrt{0.25} = 0.5$
L'intervalle de confiance à 1-$\alpha = 95\%$ :
$\left[0.898 - 1.6449\frac{0.5}{\sqrt{25}}, 0.898 + 1.6449\frac{0.5}{\sqrt{25}}\right] \approx [0.734, 1.062]$
L'intervalle de confiance à 1-$\alpha = 99\%$ :
$\left[0.898 - 2.3263\frac{0.5}{\sqrt{25}}, 0.898 + 2.3263\frac{0.5}{\sqrt{25}}\right] \approx [0.665, 1.131]$

Pour le seconde points la variance ne nous est pas donnée, nous calculerons alors la variance empirique, a besoin du coeffficient de student alors on lit dans la table de student ligne 24 (taille de l'echantillion - 1) pour 1-$\alpha = 0.95 \Leftrightarrow \alpha = 0.05$ on lit la valeur 2.0639 et 1-$\alpha = 0.99 \Leftrightarrow \alpha = 0.01$ on lit la valeur 2.797.

On note s la variance empirique $S^2_{n-1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$, alors $s = \frac{\frac{1}{25}\sum_{i=1}^{25}(X_i - \bar{X}_{25})^2}{24} \approx 0.251$
L'intervalle de confiance à 1-$\alpha = 95\%$ :
$\left[0.898 - 2.0639\frac{0.5}{\sqrt{25}}, 0.898 + 2.0639\frac{0.5}{\sqrt{25}}\right] \approx [0.794, 1.002]$
L'intervalle de confiance à 1-$\alpha = 99\%$ :
$\left[0.898 - 2.797\frac{0.251}{\sqrt{25}}, 0.898 + 2.797\frac{0.251}{\sqrt{25}}\right] \approx [0.757, 1.039]$

## 2.2 Estimation d'une proportion

Les données du probléme nous donne $n = 1000$ ètudiant dont $S_n = 673$ d'entre eux on choisis de suivre le cours d'algorithme avec un interval de confiance de $1 - \alpha = 80\% \Leftrightarrow \alpha = 0.2$
On peut calcule la moyenne empirique grâce à la formule $\bar{X}_n = \frac{S_n}{n}$
$\bar{X}_n = \frac{673}{1000} = 0.673$

$Z_\alpha = \frac{\alpha}{2} = \frac{0.2}{2} = 0.1$ dans la table des fractiles de la loi normale pour $\alpha = 0.1$ ce qui vos 1.3408.

La formule pour calcule l'intervale de confiance(IC) d'une proportion :
IC= $\left[\bar{X}_n - Z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}}, \bar{X}_n + Z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}}\right]$

L'intervale de confiance a 80% de cette population vos :
$\left[0.673 - 1.3408\sqrt{\frac{0.673(1-0.673)}{\sqrt{1000}}}, 0.673 + 1.3408\sqrt{\frac{0.673(1-0.673)}{\sqrt{1000}}}\right] \approx [0.653, 0.693]$

# 3 Partie II : Régression linéaire

## 3.1 Approximation

Nous souhaitons approximer les ventes de glaces, G = {215, 325, 185, 332, 406, 522, 412, 614, 544, 421, 445, 408} et la température, T = {14.2, 16.4, 11.9, 15.2, 18.5, 22.1, 19.4, 25.1, 23.4, 18.1, 22.6, 17.2} avec une fonction affine $f(x) = ax+b$ grâce à deux méthodes d'approximation..

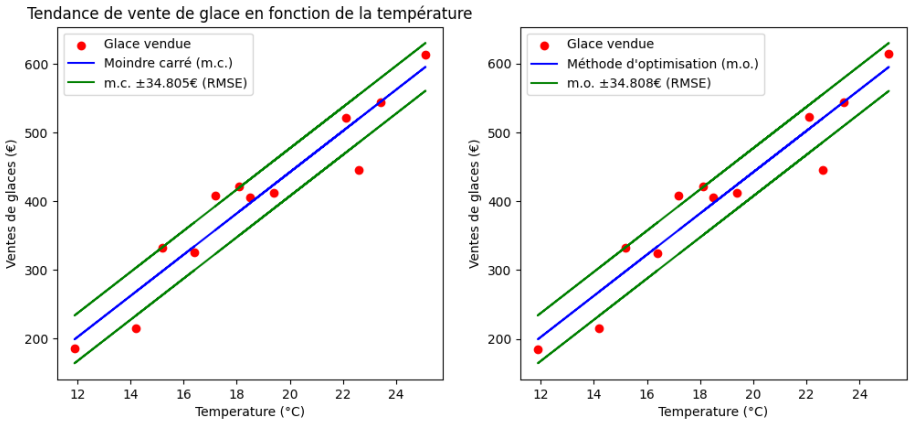


FIGURE 13 – Aprroximation par méthode des moindres carrés et d'optimisation.

## 3.2 Méthodes

### 3.2.1 Méthode des moindres carrés

La méthode des moindres carrés est une technique d'estimation des paramètres d'un modèle mathématique en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle.
Le modèle de régression linéaire est défini par l'équation $y_i = ax_i + b + \epsilon_i$, où $y_i$ est la variable à expliquer, $x_i$ est la variable explicative et $\epsilon_i$ est le terme d'erreur. Les estimateurs des moindres carrés, $\hat{a}$ et $\hat{b}$, sont obtenus en minimisant la somme totale des erreurs quadratiques définie par la fonction

$$\min_{a,b} S(a,b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Les formules des estimateurs sont les suivantes :

$$\hat{a} = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$\hat{b} = \frac{\sum_i y_i - \hat{a} \sum_i x_i}{n}$$

Une fois les paramètres estimés avec les données d'entraînement, le modèle peut être utilisé pour faire des prédictions :

$$\hat{y} = f(x) = \hat{a}x + \hat{b}$$

### 3.2.2 Méthode d'optimisation

La méthode d'optimisation est utilisée en apprentissage automatique et en optimisation mathématique. Elle permet de trouver les meilleurs paramètres pour un modèle donné. $\min J(a,b)$ permet de minimiser l'erreur. Comme J est une fonction convexe, il existe un seul minimum global [1]. Pour trouver ce minimum, on calcule le gradient de J. La méthode du gradient étant une méthode itérative, on doit trouver de bons paramètres $\gamma$ car, sans cela, la méthode pourrait ne jamais converger. La méthode s'arrête quand l'écart entre les itérations est suffisamment petit.

### 3.2.3 Quel est le modèle le plus précis ?

En appliquant les formule de la méthode du RMSE et du coefficient de détermination au fonction calcule precedement et au jeu de donnée on obtient les valeur suivant :
La valeur RMSE par la méthode des moindres carrés : $\approx 34.805$
La valeur RMSE par la méthode d'optimisation : $\approx 34.808$

1. `https://fr.wikipedia.org/wiki/Fonction_convexe`

Le coefficient de détermination de la méthode des moindres carrés vos : $\approx 0.917$
Le coefficient de détermination de la méthode d'optimisation vos : $\approx 0.91$

La méthode des moindres carrées est une meilleur approximation si l'on regarde avec la méthode RMSE mais la méthode d'optimisation est une meilleur approximation si l'on regarde le coefficient de détermination.

### 3.2.4 Vente de glace estimées pour 13, 20 et 27°C

Afin de déterminer le prix estimé a chacune de ces températures on utilise la fonction déterminé par chacune des méthode precedent.

Notons $f(x) = ax + b$ la fonction obtenue par la méthode des moindres carrés sur le jeux de donnée on obtient une valeur de $a \approx 30.088$ et $b \approx -159.474$.
La vente de glace estimé par la méthode des moindres carrés pour les températures x sont :

| Températures (°C) | Estimation des ventes (€) |
|---|---|
| 13 | 231.668 |
| 20 | 442.283 |
| 27 | 652.898 |

Table 1 – Approximation par la méthode des moindres carrés

Notons $g(x) = a'x + b'$ la fonction obtenue par la méthode d'optimisation sur le jeux de donnée on obtient une valeur de $a' \approx 29.972$ et $b' \approx -157.222$
La vente de glace estimé par la méthode d'optimisation pour les température x sont :

| Températures (°C) | Estimation des ventes (€) |
|---|---|
| 13 | 232.416 |
| 20 | 442.221 |
| 27 | 652.026 |

Table 2 – Approximation par la méthode d'optimisation

### 3.2.5 Vente de glace estimées pour 13, 20 et 27°C avec vente de 450€à 21°C

Afin de déterminé ces nouvelle valeur on ajoute 21°C et 450€a notre jeux de donnée ce qui va influencer les droites obtenus par les deux méthodes ont a alors, un nouveau (a,b) pour la méthode des moindres carrés et (a',b') par la méthode d'optimisation.
Les nouvelles valeurs de a et b pour la fonction f(x) sont : $a \approx 29.824$ et

$b \approx -156.221$

Les ventes estimées par la méthode des moindres carrés sont :

| Températures (°C) | Estimation des ventes (€) |
|---|---|
| 13 | 231.492 |
| 20 | 440.26 |
| 27 | 649.028 |

TABLE 3 – Approximation par la méthode des moindres carrés

Les nouvelles valeurs de a et b pour la fonction g(x) sont : $a' \approx 29.706$ et $b' \approx -153.903$
Les ventes estimées par la méthode d'optimisation sont :

| Températures (°C) | Estimation des ventes (€) |
|---|---|
| 13 | 232.272 |
| 20 | 440.212 |
| 27 | 649.028 |

TABLE 4 – Approximation par la méthode d'optimisation

# Conclusion

On conclut que les méthodes étudiées dans ces deux travaux pratiques ont chacune leur utilité spécifique. Les lois de probabilités discrètes, telles que Poisson et binomiale, conviennent aux données discrètes, tandis que les lois continues, comme la normale ou l'exponentielle, sont adaptées aux données continues. Les intervalles de confiance offrent une mesure précise de la fiabilité des estimations, et la régression linéaire est essentielle pour l'analyse des relations entre variables. Aucune méthode n'est supérieure en soi ; leur efficacité dépend de la nature des données et de l'objectif de l'étude.

# Table des figures