UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

UFR SEN
Sciences Exactes et Naturelles

Compte-rendu de TP Statistique Inférentielle.

Master 1 Mathématiques et Applications

UFR Sciences Exactes et Naturelles
Parcours Statistique pour l'Évaluation et la Prévision

PAR
GNAGO YANNICK.

ENSEIGNANT: M. DJAMAL LOUANI

.

2021/2022

Résumé

L'objet de ce TP est de voir des aspects pratiques de l'inférence statistique. La compréhension des lois de probabilités usuelles, les estimations et les tests sur des données simulées (ou réelles) seront détaillés tout au long de ce rapport.
Nos programmes informatiques seront faits avec le logiciel R.

Table des matières

Table des figures

Chapitre 1

Les Tirages d'échantillons et les diagrammes

Nous présenterons quelques méthodes de simulation de variables aléatoires de certaines lois classiques.
L'idée de représenter ces lois en passant par la fonction de répartition [1].

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

avec f la fonction de densité de la loi.[1]

Dans la suite, nous simulerons des échantillons des lois Binomiale, Poisson, exponentielle et normale et nous présenterons les diagrammes pour chacune de ces lois .

1.1 Loi de Binomiale

Cette loi est d'abord basée sur la loi de Bernoulli qui est une loi discréte [2], elle prend la valeur 1 (succès) avec la probabilité p et 0 (échec) avec la probabilité $q = 1 - p$ (i.e.$\mathbf{P}(X = x) = p^x(1-p)^{1-x}$ avec $x \in \{0, 1\}$).
Lorsque l'on répète cette loi de Bernoulli un certain nombre de fois de façon indépendante (n fois), on obtient alors une loi binomiale de paramètres n et p.
Pour maintenant l'a simulé, on va donc créer un échantillon de taille n d'une variable de loi de Bernoulli de paramètre p. Pour cela, on utilise les fonctions de la distribution binomiale de rstudio : **rbinom(n, size, prob), dbinom(x, size, prob), pbinom(q, size, prob)** où n correspond au nombre d'observations de distribution binomiale de paramètres size(nombre d'essais) et prob(probabilité de succès à chaque essai.).

1. La fonction de répartition d'une variable aléatoire X indique pour chaque valeur réelle x la probabilité que X prenne une valeur au plus égale à x. C'est la somme des probabilités des valeurs de X jusqu'à x.

2. Une loi de probabilité est dite discrète quand l'expérience aléatoire associée à cette loi ne peut prendre qu'un nombre limité de valeurs distinctes (Qualitatives ou quantitatives).

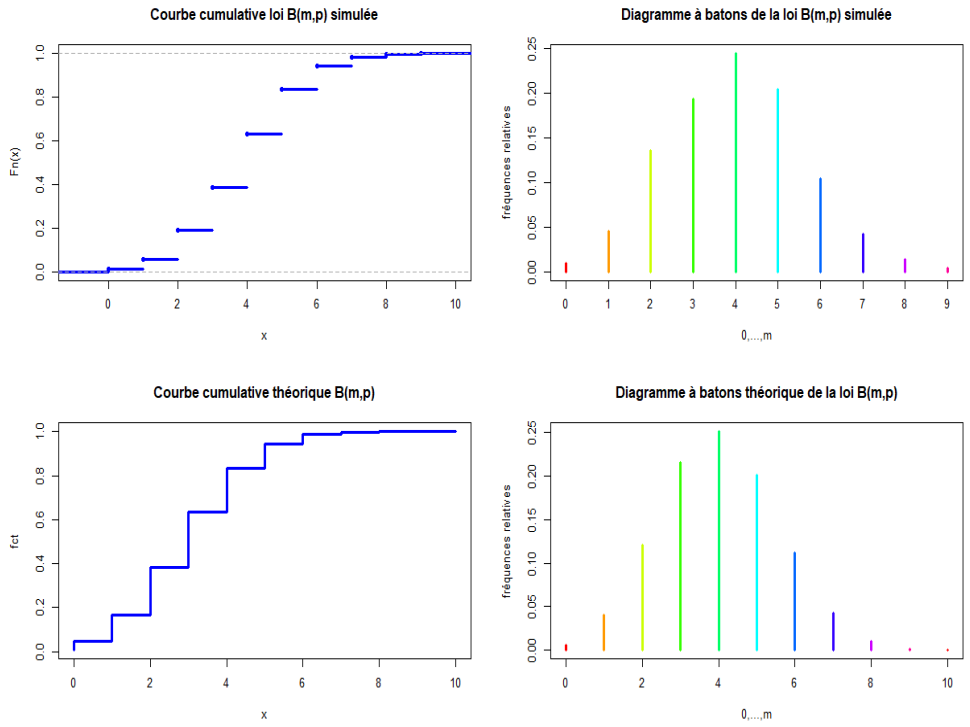


FIGURE 1.1 – Avec n = 1000, size = 10, prob = 0.4

On remarque que plus n augmente la loi simulée se rapproche de la courbe théorique et que les diagrammes à bâtons sont pareils dans les deux situations.

1.2 Loi de Poisson

La loi de Poisson de paramètre λ, ou loi des événements rares, correspond au modèle suivant :
Sur une période T, un événement arrive en moyenne λ fois. On appelle X, la variable aléatoire déterminant le nombre de fois où l'événement se produit dans la période T. X prend des valeurs entières : 0, 1, 2, ...(i.e. une loi discrète).
Cette variable aléatoire suit une loi de probabilité définie par :

$$p(k) = P(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \text{pour tout entier naturel} \quad k \in \mathbf{N}$$

Pour avoir un échantillon de taille n d'une variable de loi de Poisson de paramètre λ, on utilise les fonctions de la distribution de Poisson de rstudio : **rpois(n,lambda),ppois, dpois** où n correspond au nombre d'observations de loi Poisson de paramètre lambda.

On remarque que plus n augmente la loi simulée se rapproche de la courbe théorique et que les diagrammes à bâtons sont presque pareils dans les deux situations.

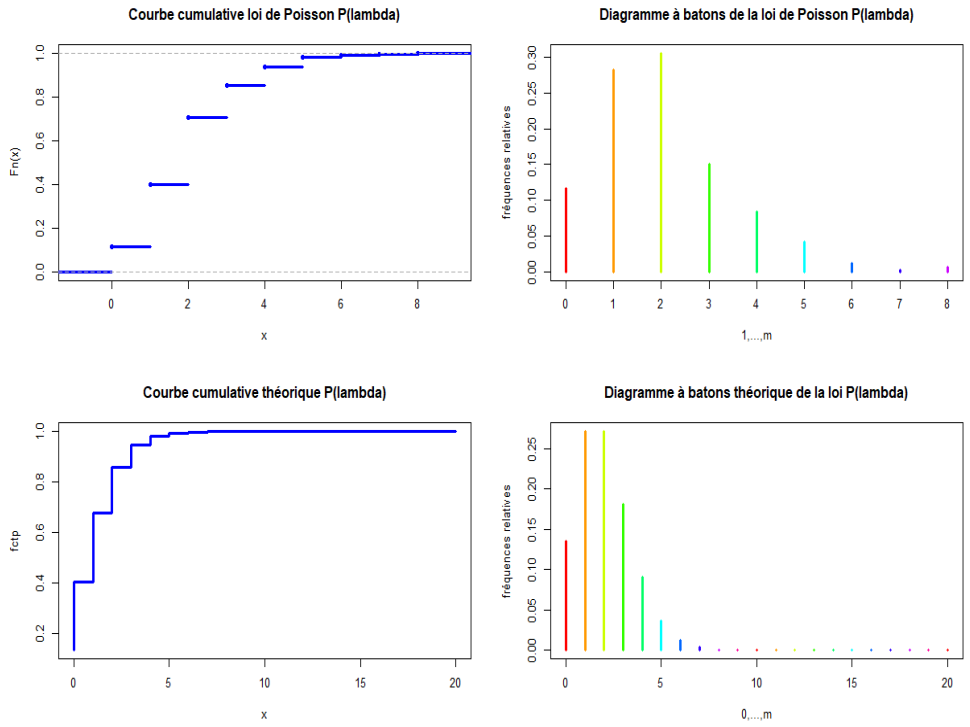


FIGURE 1.2 – Avec n = 500, size = 10, lambda=2

1.3 Loi exponentielle

Une loi exponentielle correspond au modèle suivant :
Soit X, une variable aléatoire définissant la durée de vie d'un phénomène (i.e. une loi continue [3]). Si l'espérance de vie du phénomène est $E(X)$ et si la durée de vie est sans vieillissement, i.e. si la durée de vie au-delà de l'instant T est indépendante de l'instant T, alors X a pour fonction de répartition :

$$F(x) = p(X \leq x) = \begin{cases} 0 \text{ si } x < 0 \\ 1 - e^{-\lambda x} \text{ si } x \geq 0 \end{cases}$$

On utilise les fonctions de la distribution exponentielle de rstudio : **rexp(n, rate), pexp, dexp** où n correspond au nombre d'observations de la loi exponentielle et rate correspond au paramètre λ.

3. Une loi de probabilité est dite continue quand l'expérience aléatoire associée à cette loi peut prendre n'importe quelle valeur dans un intervalle défini, ouvert ou non.

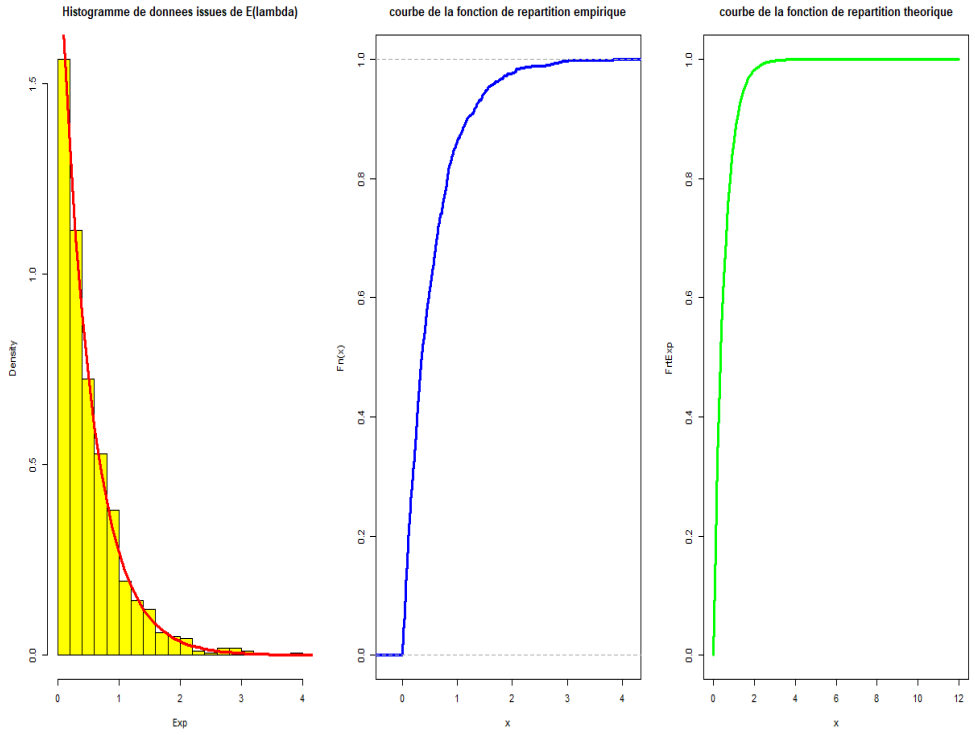


FIGURE 1.3 – Avec n = 1000,rate=2.

On remarque que plus n augmente, la courbe de la fonction de répartition empirique se rapproche de celle théorique et l'histogramme des données est suit bien la loi.

1.4 Loi Normale

La loi normale, ou distribution normale, définit une représentation de données selon laquelle la plupart des valeurs sont regroupées autour de la moyenne et les autres s'en écartent symétriquement des deux côtés. Ces paramètres sont alors : son espérance (μ) et sa variance (σ).
Pour avoir un échantillon de taille n d'une variable de loi normale de paramètre (μ, σ), on utilise les fonctions de la distrubition normale sous rstudio : **rnorm(n, mean, sd),pnorm,dnorm** où n correspond au nombre d'observations, mean : μ, sd :σ.

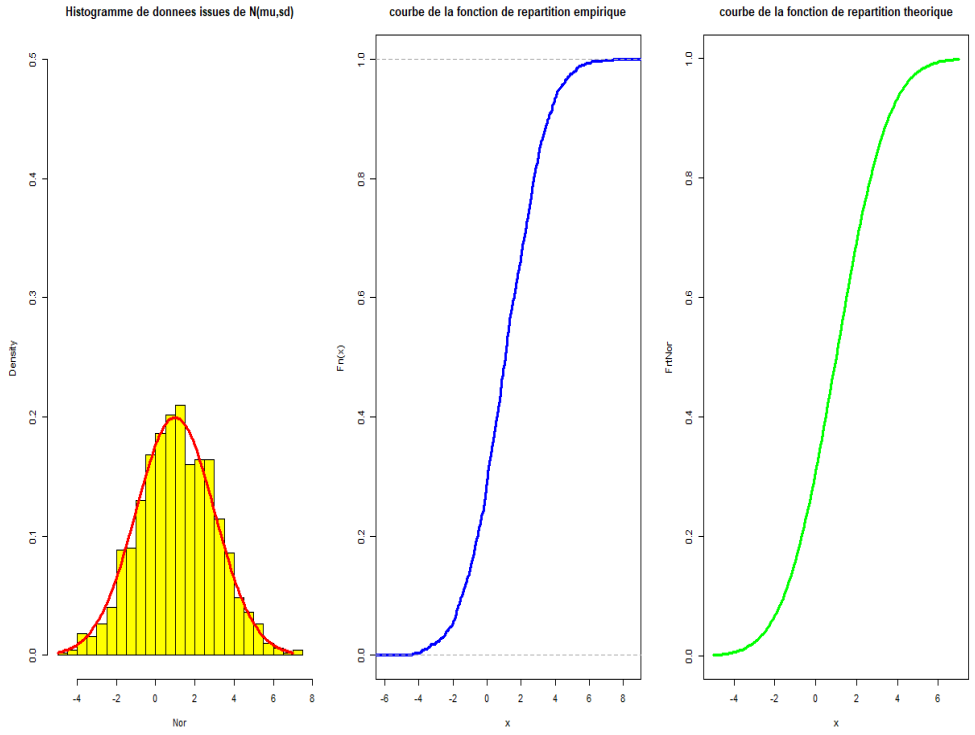


FIGURE 1.4 – Avec n = 1000, mu = 1, et sd=2 et nclass=20(Le nombre de bâtons).

Au regard de la figure 1.4, on constate que l'histogramme de l'échantillon est ajusté par la cloche d'une loi normale et que la fonction de répartition empirique est pareille à celle théorique.
Après avoir étudié ces lois classiques, nous allons maintenant poursuivre pour connaître certaines caractéristiques des lois.

Chapitre 2

Lois des grands nombres

La loi des grands nombres est l'une des propriétés les plus importantes en probabilité.
Elle a été formalisée au XVIIe siècle lors de la découverte de nouveaux langages mathématiques.

Essentiellement, la loi des grands nombres indique que lorsque l'on fait un tirage aléatoire dans un échantillon de grande taille, plus on augmente la taille de l'échantillon, plus les caractéristiques statistiques du tirage (l'échantillon) se rapprochent des caractéristiques statistiques de la population.[4]

Theorème 1 (Loi forte des grands nombres) *Si* $(X_n)_{n>0}$ *est une suite de variables aléatoires indépendantes identiquement distribuées, on a équivalence entre :*
(i) $E(|X_1|) < +\infty$,
(ii) la suite $\frac{X_1+\cdots+X_n}{n}$ *converge presque sûrement.*
De plus, si l'une de ces deux conditions équivalentes est remplie, alors la suite $\frac{X_1+\cdots+X_n}{n}$ *converge presque sûrement vers la constante* $E(X_1)$.

Theorème 2 (Loi faible des grands nombres) *Soit* X *une variable aléatoire admettant une variance. Soit* $(X_n)_{n\in N}$ *une suite de variables aléatoires indépendantes de même loi que* X. *Alors :* $\forall \varepsilon > 0\ P\left[\left|\frac{X_1+\cdots+X_n}{n} - E[X]\right| > \varepsilon\right] \longrightarrow 0$ *quand* $n \longrightarrow \infty$.

Par exemple, lorsqu'on lance un très grand nombre de fois un dé non pipé, la fréquence d'apparition du 5 tend vers 1/6. En d'autres termes, elle consiste à ne montrer que la moyenne empirique (S_n)(respectivement la variance empirique) d'un échantillon suivant une certaine loi d'espérance μ (respectivement de variance σ) convergent vers l'espérance μ(respectivement la même variance σ).
Nous allons illustrer de façon pratique la loi des grands nombres pour les lois simulées précédemment.

2.1 Loi binomiale

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{B}(10, 0.4)$. Nous l'avons simulé pour différentes valeurs de N et nous observons l'aspect de convergence

émit auparavant de la loi des grands nombres, i.e. la moyenne empirique converge vers l'espérance(ici $E(X) = 4.0$) pour une taille de l'échantillon suffisamment grande (cf figure 2.1, figure 2.2).

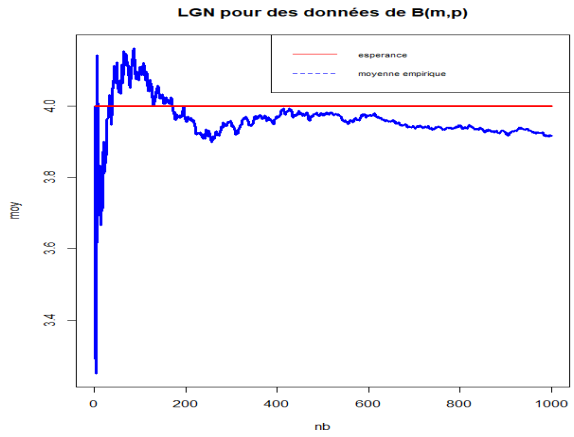


FIGURE 2.1

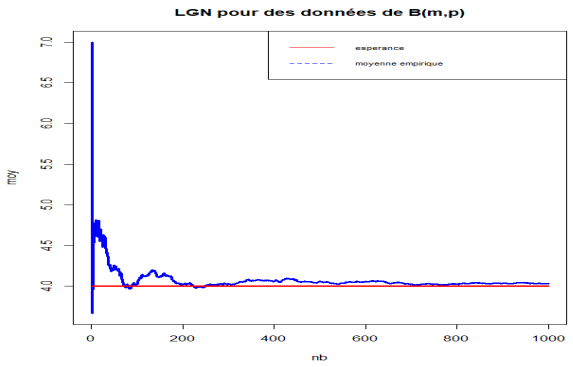


FIGURE 2.2

2.2 Loi de Poisson

Considérons un échantillon de taille $N = 500$ qui suit une loi $\mathcal{P}(2)$.

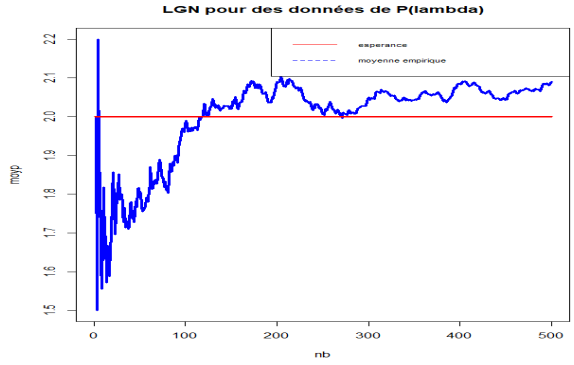


FIGURE 2.3

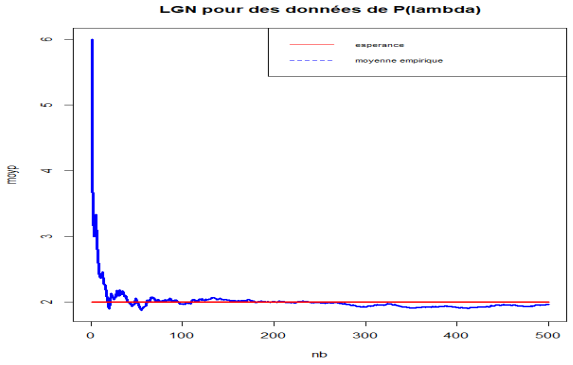


FIGURE 2.4

Nous l'avons simulé pour différentes valeurs de N et nous observons l'aspect de convergence émit auparavant de la loi des grands nombres, i.e. la moyenne empirique converge vers l'espérance(ici $E(X) = 2.0$) pour une taille de l'échantillon suffisamment grande (cf figure 2.3, figure 2.4).

2.3 Loi exponentielle

Considérons un échantillon de taille $N = 500$ qui suit une loi $\mathcal{E}(1.5)$.
Nous l'avons simulé pour différentes valeurs de N et nous observons l'aspect de

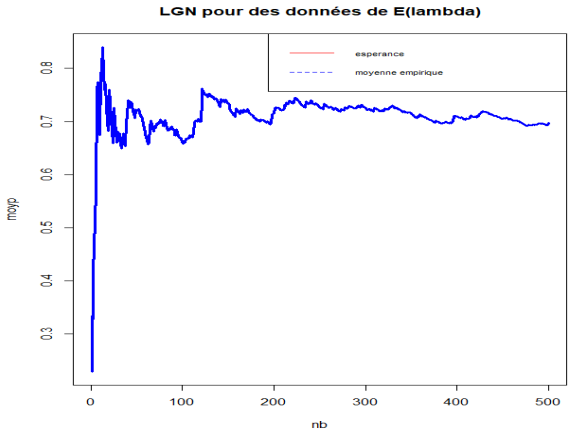


FIGURE 2.5

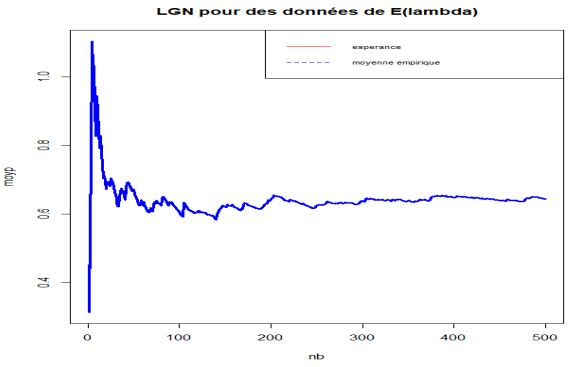


FIGURE 2.6

convergence émit auparavant de la loi des grands nombres, i.e. la moyenne empirique converge vers l'espérance(ici $E(X) = 0.66$) pour une taille de l'échantillon suffisamment grande (cf figure 2.5, figure 2.6).

2.4 Loi Normale

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{N}(1, 2)$.

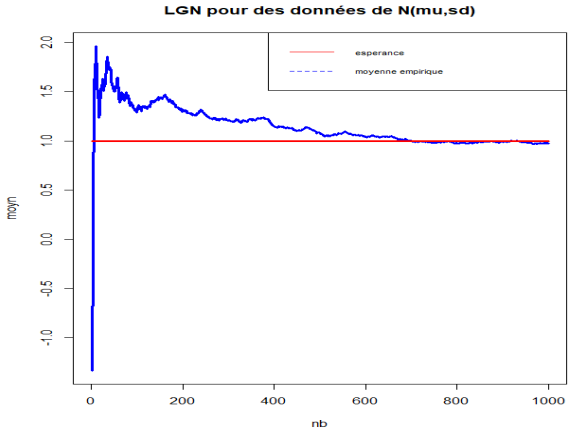


FIGURE 2.7

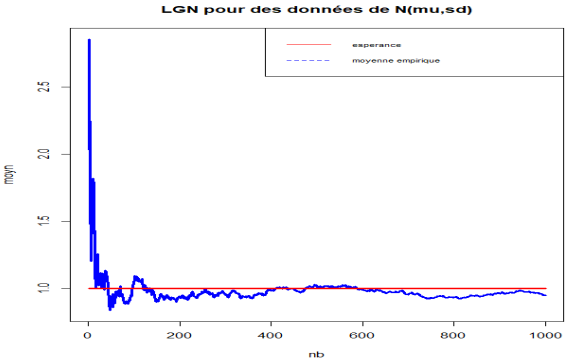


FIGURE 2.8

Nous l'avons simulé pour différentes valeurs de N et nous observons l'aspect de convergence émit auparavant de la loi des grands nombres, i.e. la moyenne empirique converge vers l'espérance(ici $E(X) = 1$) pour une taille de l'échantillon suffisamment grande (cf figure 2.7, figure 2.8).

Chapitre 3

Le théorème central limite.

Le théorème central limite (parfois appelé théorème de la limite centrale) établit la convergence en loi d'une suite de variables aléatoires vers la loi normale. Intuitivement, ce résultat affirme que toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire gaussienne [1].
En d'autres termes, cela signifie que l'histogramme de l'échantillon doit être ajusté par la cloche d'une loi normale.

Theorème 3 (Central Limite) *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. et notons $\overline{X_n}$ sa moyenne empirique.*
Si $\sigma^2 = Var(X_1) < +\infty$
Alors

$$U_n = \sqrt{n}\left(\frac{\overline{X_n} - E(X_1)}{\sigma}\right) \overset{\mathcal{L}oi}{\to} \mathcal{N}(0,1)$$

Pour représenter sur rstudio ce théorème, l'idée est qu'on va générer mm échantillons pour avoir mm points de U_n, car chaque échantillon donne un point de U_n, ensuite on trace l'histogramme pour les mm points de U_n.

3.1 Loi Binomiale

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{B}(10, 0.4)$ avec $mm = 2000$.
Nous remarquons bien la convergence de la loi binomiale vers la loi normale (cf figure 3.1).

1. la variable gaussienne est une variable aléatoire dont la densité est entièrement déterminée par la donnée de ses deux premiers moments, dit moyenne et variance.

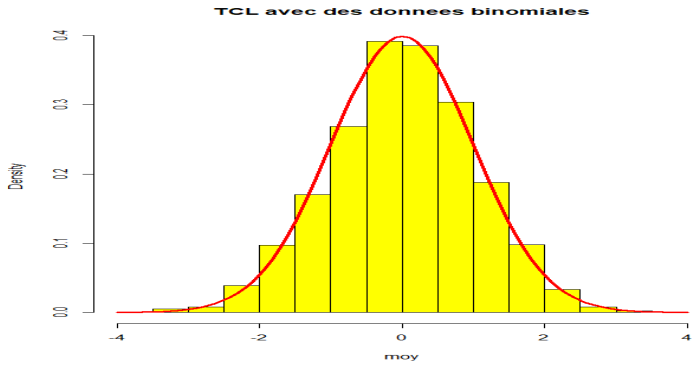


FIGURE 3.1 – TCL avec N = 1000, size = 10, prob = 0.4.

3.2 Loi de Poisson

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{P}(1.5)$ avec $mm = 2000$.
Nous remarquons bien la convergence de la loi de poisson vers la loi normale (cf figure 3.1).

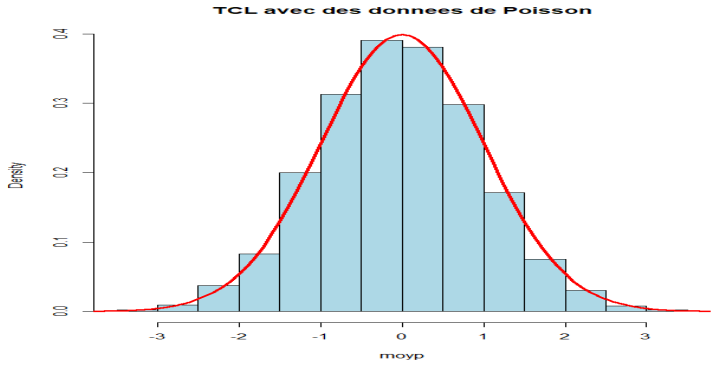


FIGURE 3.2 – TCL avec N = 1000,lambda=1.5.

3.3 Loi exponentielle

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{E}(1.5)$ avec $mm = 2000$.
Nous remarquons bien la convergence de la loi d'exponentielle vers la loi normale (cf figure 3.1).

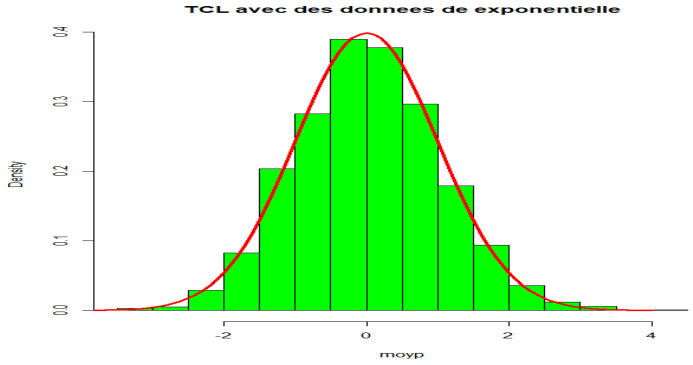


FIGURE 3.3 – TCL avec N = 1000,lambda=1.5.

3.4 Loi Normale

Considérons un échantillon de taille $N = 1000$ qui suit une loi $\mathcal{N}(1, 1)$ avec $mm = 2000$.
Nous remarquons bien la convergence de la loi d'exponentielle vers la loi normale (cf figure 3.1).

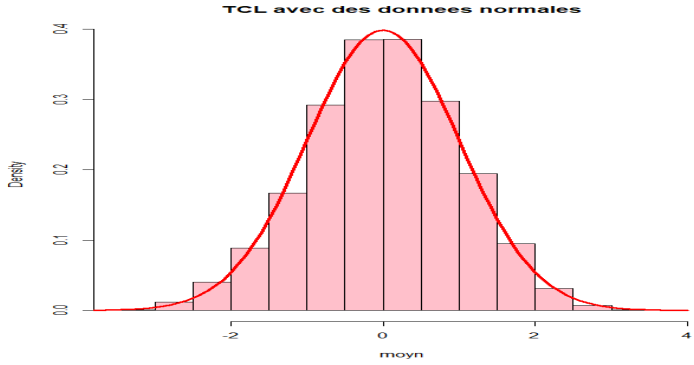


FIGURE 3.4 – TCL avec N = 1000,lambda=1.5.

Après avoir appliqué la loi des grands nombres et le théorème centrale limite aux lois classiques simulées, nous allons maintenant faire des estimations des paramètres ces lois.

Chapitre 4

Estimation ponctuelles des paramètres

Dans cette partie, nous estimerons certains paramètres des lois classiques simulées. On se donnera à chaque fois des ensembles de points (mm) où chaque point correspond à la valeur de l'estimateur pour l'échantillon.

Définition 4.0.1 *Un ESTIMATEUR de θ sera une statistique $T = f(X_1, ..., X_n)$ et sa réalisation sera notée $t = f(x_1, ..., x_n)$.*
Cet ESTIMATEUR T de θ est dit sans biais si $E(T) = \theta$.[3]

4.1 Loi binomiale

Sur Rstudio, pour estimer la taille(n) ou la probabilité de succes(p) d'une loi binomiale($\mathcal{B}(n, p)$), nous allons simuler mm échantillons de taille N de cette loi. Pour chacun de ces échantillons, on calculera la taille et la probabilité.
On remarque que les estimations de l'échantillon sont bien concentrées autour de l'esperance($E(X) = 4$) et la variance($V(X) = 2.4$).
Cela montre que les estimateurs des parametres sont sans biais (cf. figure 4.1)

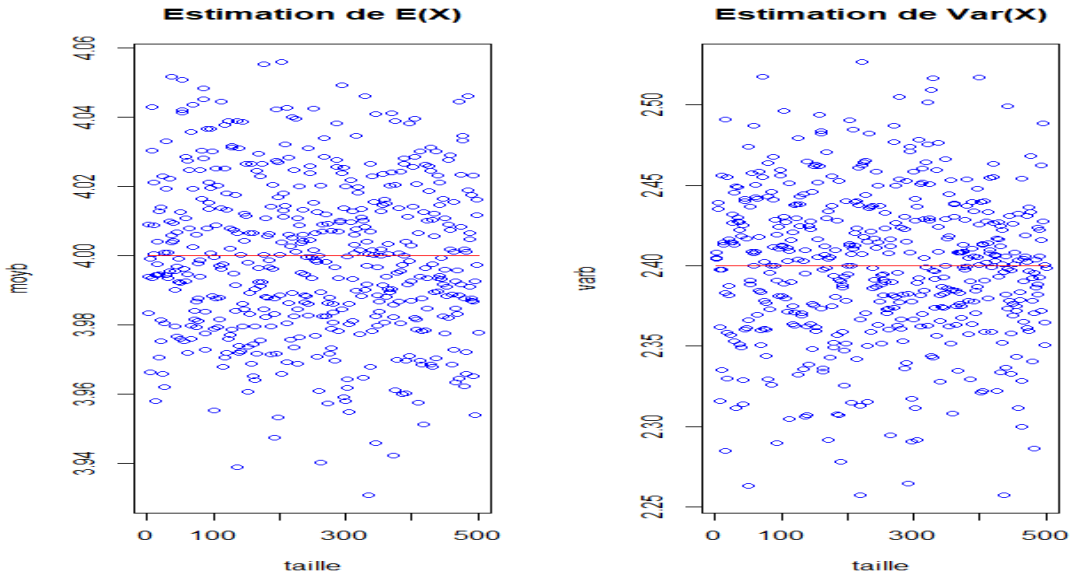


FIGURE 4.1 – Estimation des parametes d'une loi binomiale avec N = 5000, mm = 500, n = 10 et p= 0.4.

4.2 Loi normales

Sur Rstudio, pour estimer la moyenne (μ) ou l'ecart-type(σ) d'une loi normale($\mathcal{N}(\mu, \sigma^2)$), nous allons simuler mm échantillons de taille N de cette loi. Pour chacun de ces échantillons, on calculera la moyenne et l'ecart-type de l'échantillon .

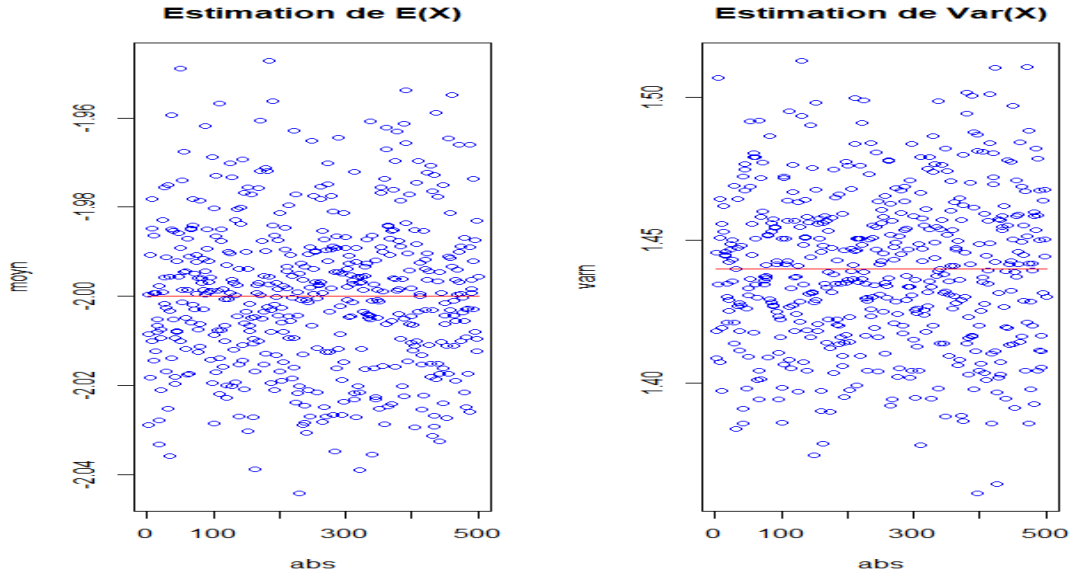


FIGURE 4.2 – Estimation des parametes d'une loi binomiale avec N = 5000, mm = 500, mu =-2 et sd= 1.2.

On remarque que l' estimation de l'échantillon sont bien concentrées autour de l'esperance($E(X) = -2$) et ce qui n'est pas le cas pour la variance($V(X) = 1.2$).

Alors, ici l'estimation de la variance est biaisé contrairement à celle de la moyenne(cf. figure 4.2).

Chapitre 5

Tests

En statistiques, un test, est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle (H_0), en fonction d'un échantillon de données.
Il faut noter qu'on ne pourra jamais conclure avec certitude dans un test statistique. Il y aura toujours des erreurs de décision qui sont :
- L'erreur de première espèce, notée α, la probabilité de rejeter (H_0) alors qu'elle est vraie.
- L'erreur de deuxième espèce, notée β, la probabilité d'accepter (H_0) alors qu'elle est fausse.
- La puissance du test pour (H_1) la probabilité de retenir (H_1) alors qu'elle est vraie ($= 1 - \beta$).

Nous effectuerons plusieurs tests sur des lois classiques dont les paramètres seront définis.

5.1 Test sur le paramètre μ d'une loi normale de variance connue(Bilatéral)

Nous aurons donc comme hypothèses : H_0 **:"**$\mu = \mu_0$**"** **contre** H_1 **:"**$\mu \neq \mu_0$**"**.
Ici, l'on considéra la loi $\mathcal{N}(\mu, 1.2)$ avec $\mu_0 = 1$.
L'idée est qu'on a simulé mm échantillons de taille N d'une variable X suivant la loi Normale et on crée un compteur alphatest que l'on initialise à 0. Pour chaque échantillon, on ajoute 1 à ce compteur si l'hypothèse était fausse. On retourne ce compteur divisé par mm, et ce résultat correspondra au risque de seconde espèce.

On remarque qu'ici, le test est sans biais i.e. la puissance du test $1 - \beta$ est supérieure au niveau α sur H_1 (β : le risque de seconde espèce).

5.2 Test sur le paramètre μ d'une loi normale de variance inconnue

Nous aurons donc comme hypothèses : H_0 **:"**$\mu = \mu_0$**"** **contre** H_1 **:"**$\mu \neq \mu_0$**"**.
Ici, l'on considéra la loi $\mathcal{N}(\mu, \sigma)$ avec $\mu_0 = 1$.

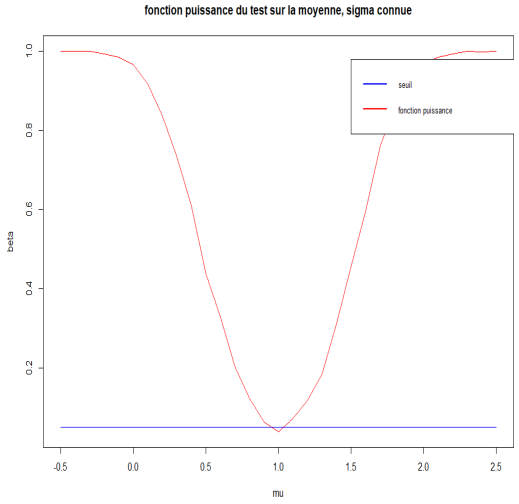


FIGURE 5.1 – Fonction puissance du test sur la moyenne d'une loi normale de variance connu.

L'idée sur rstudio est pareille que la précédente.

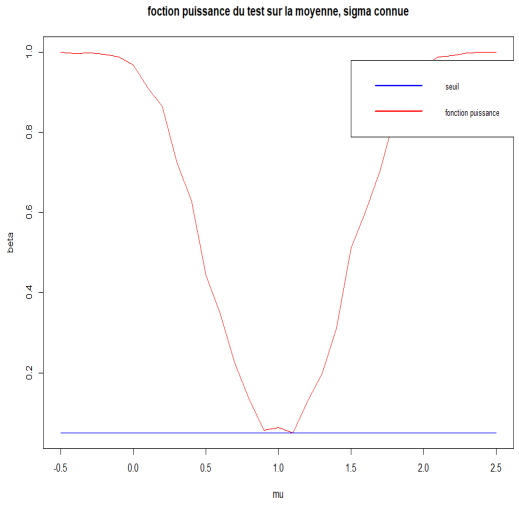


FIGURE 5.2 – Fonction puissance du test sur la moyenne d'une loi normale de variance inconnu.

On remarque qu'ici aussi, le test est sans biais i.e. la puissance du test est supérieure au niveau α sur H_1.

5.3 Test d'égalité des moyennes (d'une loi de Poisson et d'une loi exponentielle)

Nous aurons donc comme hypothèses : H_0 **:"**$\mu = \mu_0$**"** **contre** H_1 **:"**$\mu \neq \mu_0$**"**. Ici l'on considère les lois $\mathcal{P}(\lambda)$ et $\mathcal{E}(\lambda)$.

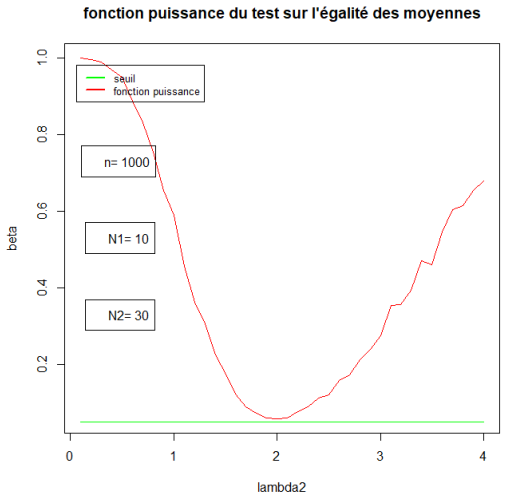


FIGURE 5.3 – fonction puissance du test sur l'égalité des moyennes.

On remarque qu'ici, le test est sans biais i.e. la puissance du test est supérieure au niveau α sur H_1.

5.4 Test sur le paramétre μ d'une loi normale de variance inconnue(unilatéral)

Nous aurons donc comme hypothèses : H_0 **:"**$\mu \leq \mu_0$**"** **contre** H_1 **:"**$\mu > \mu_0$**"**

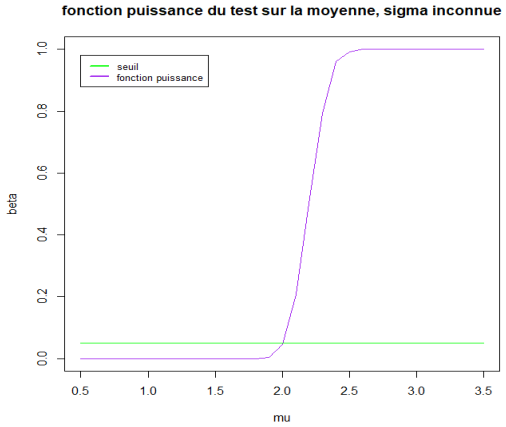


FIGURE 5.4 – fonction puissance du test unilatéral d'une loi normale($\mu_0 = 2$) avec une variance inconnue.

On remarque qu'ici, le test est sans biais i.e. le risque de première espèce est inférieure au niveau α sur H_0.

Chapitre 6

Intervalles de confiance

L'intervalle de confiance est un indicateur mathématique qui permet de chiffrer la zone d'incertitude, lors d'une enquête ou d'un sondage portant sur un échantillon de population.

Définition 6.0.1 *Soit X une v.a. dont la loi dépend d'un paramètre inconnu θ ; on appelle INTERVALLE DE CONFIANCE pour θ de niveau $1-\alpha$ (ou de seuil α), un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur de θ.*
On dit que $[t_1, t_2]$ est un intervalle de confiance de niveau $1-\alpha$ pour θ si

$$P(t_1 < \theta < t_2) = 1 - \alpha$$

(plus le niveau de confiance est élevé, plus la certitude est grande que la méthode d'estimation produira une estimation contenant la vraie valeur de θ).
Plus géneralement, le niveaux de confiance utilisée est 99%.

Sur rstudio, on simulera mm échantillons de taille N d'une variable X suivant une loi normale. Pour chaque échantillon, on calcule un intervalle de confiance pour $\mu = E(X)$, si d'une part σ est inconnu et d'autre part si σ connu, avec une erreur α. On crée ensuite un compteur *alphatest* et nous vérifions que la véritable valeur de μ est dans l'intervalle de confiance, si c'est le cas on rajoute 1 au compteur.
Finalement on divise le compteur par mm, ce qui retourne la fréquence d'apparition de la vraie valeur de μ dans les intervalles.

6.1 Intervalle de confiance pour $\mu = E(X)$ d'une loi normale(variance est inconnue)

la figure 6.1 montre qu'au fur à mesure, les variations de μ sur l'échantillon sont centrées autour de l'erreur α(le trait rouge).

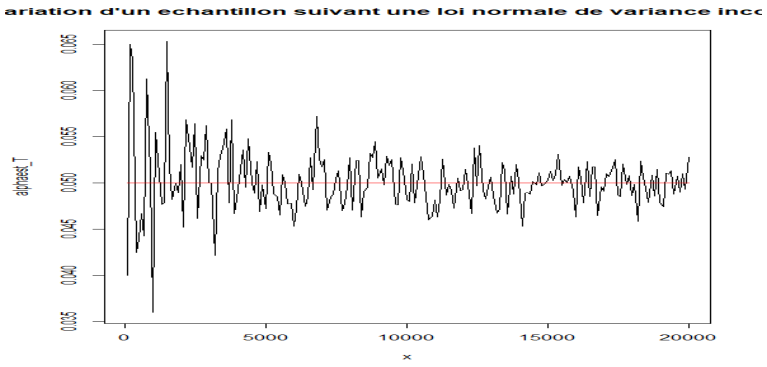


FIGURE 6.1 – IC pour une loi normale avec $N = 20$,$\mu = 1.5$, $\alpha = 0.05$.

6.2 Intervalle de confiance pour $\mu = E(X)$ d'une loi normale(variance est connue)

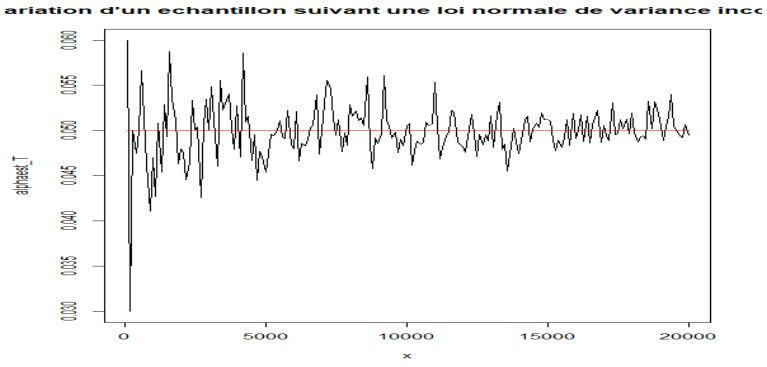


FIGURE 6.2 – IC pour une loi normale avec $N = 20$,$\mu = 1.5$,$\sigma = 1.2$, $\alpha = 0.05$.

la figure 6.2 montre qu'au fur à mesure, les variations de μ sur l'échantillon défini sont centrées autour de l'erreur α(le trait rouge).

6.3 Région de confiance pour le vecteur (μ, σ^2) de la loi normale

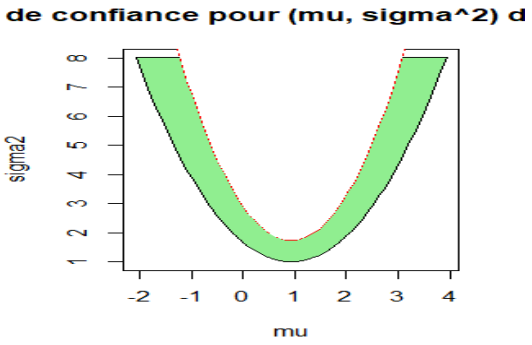


FIGURE 6.3 – IC pour (μ, σ^2) d'une loi normale avec $N = 20$,$\mu = 1.5$,$\sigma = 1.2$, $\alpha = 0.05$.

La bande verte sur la figure 6.3 représente donc la région de confiance pour (μ, σ^2) .
Jusqu'ici, nous avons donné des explications sur les différents aspects(le tirage d'échantillon, la loi des grands nombres, le TCL,...) de la statistique inférentielle ; maintenant il sera idéal de pourvoir appliquer tous ces aspects sur des données réelles.
Au regard de cela, la suite de notre travail consistera à effectuer une étude globale avec des données réelles.

Chapitre 7

Etude globale

Dans cette partie, nous allons travailler une base de données fourni par notre soin, afin de pourvoir trouver un modèle de régression linéaire multiple relatif à notre base.
Plusieurs travaux portés sur les axes développés auparavant seront effectués.
Nous avons pris comme dataset, la base de données "Auto" de la librairie "ISLR" sur la consommation d'essence, puissance et autres informations pour 392 véhicules.
Nous chercherons le meilleur modèle avec une variable qui ajuste au mieux la variable mpg(kilomètres par litre) dans notre data. On aura alors :

$$y = f(x) + \epsilon, \quad f(x) = \sum \alpha.x$$

Où $\alpha \in R$, le vecteur $x \in R^p$: les variables exogènes, le vecteur $y \in R$: la variable mpg et le bruit d'observation(l'erreur) : $\epsilon \sim \mathcal{N}(0, \sigma^2)$(l'hypothèse d'homoscédasticité).

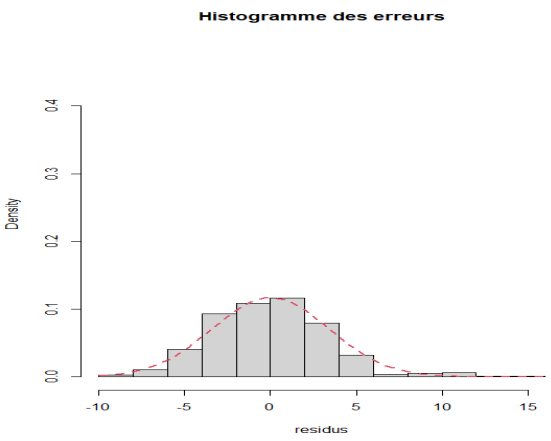
- Après l'ajustement du modèle sur la variable mpg avec toutes les autres variables, nous trouvons que les variables les plus significatives sont : horsepower, weight, year avec $AdjustedR\text{-}squared = 0.8068$ (R^2 : le coefficient de détermination [1])
  On a alors : $mpg \sim weight + year + horsepower \qquad (A)$

1. R^2 : le coefficient de détermination est une mesure de la qualité de la prédiction d'une régression linéaire.

```
call:
lm(formula = mpg ~ weight + year + horsepower, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7911 -2.3220 -0.1753  2.0595 14.3527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.372e+01  4.182e+00  -3.281  0.00113 **
weight      -6.448e-03  4.089e-04 -15.768  < 2e-16 ***
year         7.487e-01  5.212e-02  14.365  < 2e-16 ***
horsepower  -5.000e-03  9.439e-03  -0.530  0.59663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.43 on 388 degrees of freedom
Multiple R-squared:  0.8083,    Adjusted R-squared:  0.8068
F-statistic: 545.4 on 3 and 388 DF,  p-value: < 2.2e-16
```

FIGURE 7.1 – L'ajustement du modele avec une variable.

- la normalité des erreurs



Les erreurs suivent bien une loi normale.
En faisant de test de Shapiro-Wilk, on a :$p - valeur = 4.653e - 07$, donc le test est significative, alors l'erreurs ne semble suivre pas une loi Normale.

FIGURE 7.2

- l'autocorrélation des erreurs En faisant le test de Durbin-Watson, on a : $p - value = 7.205e - 15$, on conclut alors il y'a une forte corrélation des erreurs.
- Le test d'hétéroscedasticité En faisant le test de Breusch-Pagan, on a : $p - value = 1.052e - 06$, on conclut alors l'hétéroscédasticité des erreurs.

De plus, pour évaluer les hypotheses de normalité et d'homoscedascité, nous avons aussi utilisé la fonction sur rstudio "check_model" de la librairie perfomance(cf. figure 7.3) :

-Le schéma de la normalité des résidus montre une linéarité presque parfaite à part les valeurs extrêmes qui sont loin de la droite verte.

-Le schéma de l'homogénéité de la variance, nous montre que les variables exogènes ne soient pas parfaitement homogènes, mails elle est acceptable.
Les conditions d'application (La linéarité, la normalité ; l'homoscédasticité des résidus et l'absence de multi colinéarité) étant vérifiées, nous pouvons alors considérer le modèle (A) comme l'un des meilleurs modèles d'ajustement de la variable mpg.

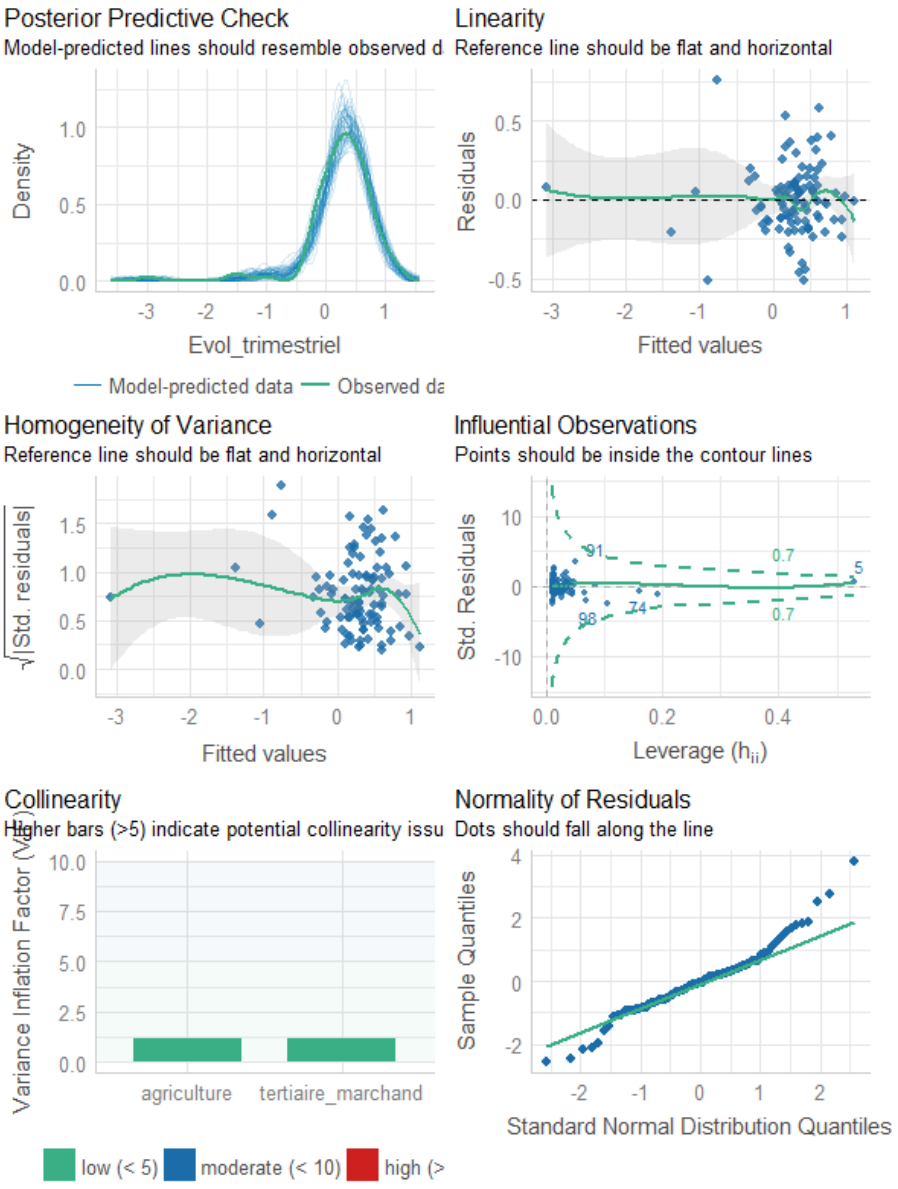


FIGURE 7.3 – L'évaluation des hypothèses de normalité et d'homoscédasticité des erreurs.

Bibliographie

[1] Jimmy Bourque and Salah-Eddine El Adlouni. *Manuel d'introduction à la statistique appliquée aux science*. Presses de l'Université Laval, 2016.

[2] François Cottet-Emard. *Probabilités et tests d'hypothèses*. De Boeck Superieur, 2014.

[3] Laurence GRAMMONT. Cours de statistiques inferentielles licence d'économie et de gestion, 2003.

[4] Jérôme Pagès. *Statistique générale pour utilisateurs. Méthodologie*. Presses universitaires de Rennes, 2010.

[5] Murray R Spiegel and LJ Stephens. *Statistique*. Ediscience, 2002.