ENSEEIHT - Toulouse INP

Rapport BE Statistiques

Modélisation de la vitesse du vent par une loi de Weibull

Sam Alonso-Virissel
Promo 2023

Encadrant : Corentin Lubeigt

17 mars 2021

Table des matières

Introduction

Ce BE porte sur l'étude de valeurs variables de vitesse du vent suivant une loi de Weibull $\mathcal{W}(\theta, p)$. Dans un premier temps, on générera ces données. Puis, on étudiera un estimateur efficace et non-biaisé de θ^p. Ensuite, on s'intéressera à un test statistique afin de déterminer si l'on se trouve en période de vent calme ou pas. Finalement, on cherchera à caractériser la loi d'un ensemble de valeurs données.

1 Génération d'un signal test

K signaux $\mathbf{y} = (y_1, ..., y_N)^T$ de loi de Weibull $\mathcal{W}(\theta, p)$ sont générés en appliquant la fonction de répartition inverse de cette loi de Weibull à $X \sim \mathcal{U}(]0, 1[)$. La fonction de répartition est donnée dans le sujet [1] (p 5) :

$$F(x; \theta, p) = 1 - \exp\left[-\left(\frac{x}{\theta}\right)^p\right]$$

On en déduit :

$$F^{-1}(X; \theta, p) = \theta(-\ln(1 - X))^{\frac{1}{p}} \tag{1.1}$$

X est générée comme cela : `X = rand(N,K)`. On applique la fonction de répartition inverse issue de l'équation 1.1. On obtient alors $Y \sim \mathcal{W}(\theta, p)$ comme demandé.

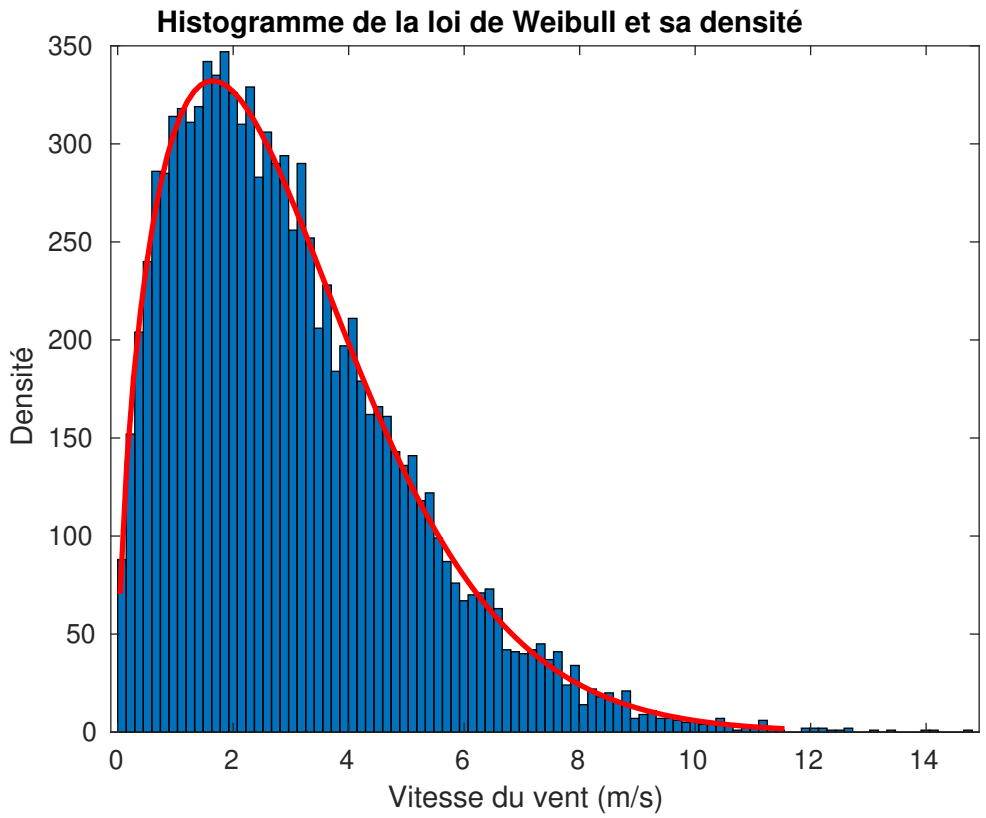

FIGURE 1.1 – Histogramme de données générées par la méthode précédente avec $N = 10\,000$, $K = 1$, $\theta = 3,3$ et $p = 1,5$.

La modélisation en rouge correspond avec l'histogramme des valeurs générées. Les données suivent donc vraisemblablement une loi de Weibull.

Le tableau suivant compare les moyennes et variances calculées avec les fonctions `mean` et `var` avec leurs valeurs théoriques rappelées dans [1] [1] :

$$\textit{Moyenne} : \mu = \theta\Gamma\left(1+\frac{1}{p}\right) \qquad \textit{Variance} : \sigma^2 = \theta^2\Gamma\left(1+\frac{2}{p}\right) - \mu^2 \tag{1.2}$$

	Moyenne	Variance
Numérique	`2.969003`	`3.984124`
Théorique	`2.979059`	`4.091267`

TABLE 1 – Moyenne et variance numérique et théorique pour $N = 10\,000$, $K = 1$, $\theta = 3,3$ et $p = 1,5$.

Il y a une erreur de l'ordre de $0,3\%$ pour la moyenne et 2% pour la variance. Encore une fois, les valeurs numériques sont proches de celles attendues théoriquement.

Afin de vérifier que ce n'est pas un coup de chance, on effectue $K = 500$ réalisations de signaux de $N = 1000$ éléments avec la même technique. θ et p sont inchangés.

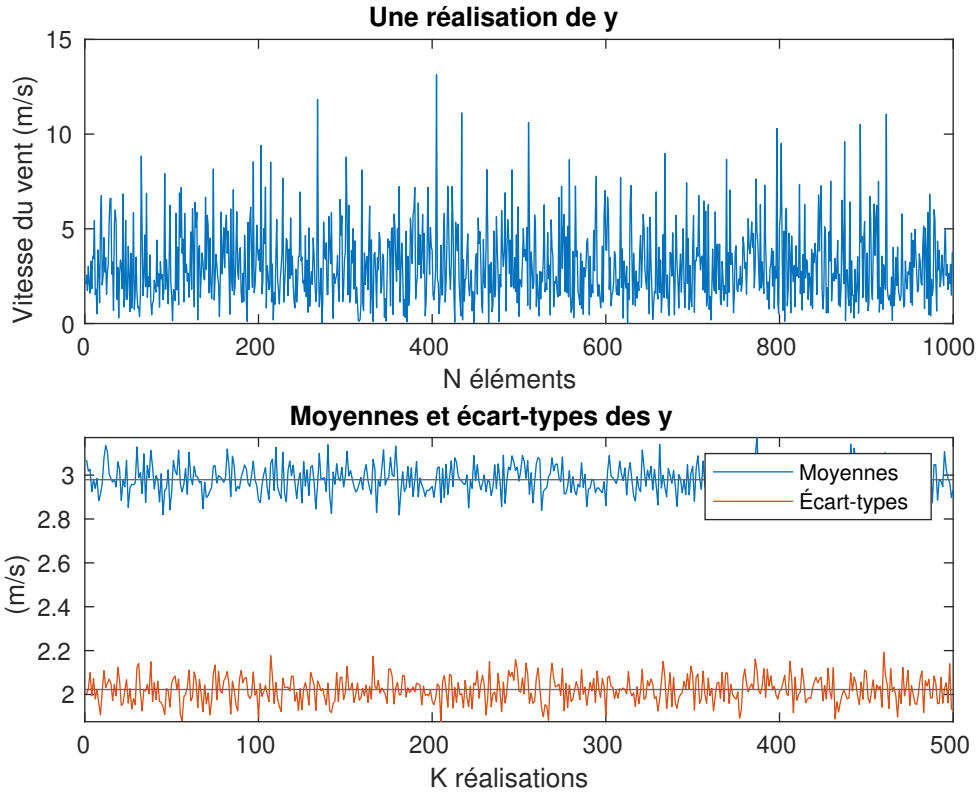

FIGURE 1.2 – Une réalisation de $N = 1000$ éléments avec $\theta = 3,3$ et $p = 1,5$. Moyennes et écart-types de $K = 500$ réalisations avec les mêmes paramètres. Les lignes sont les valeurs théoriques.

La variance n'étant pas de même dimension que la moyenne, l'écart-type est privilégié. Les moyennes et écart-types fluctuent bien autour des valeurs exactes calculées dans le tableau 1. Les réalisations ont donc les moyennes et variances attendues.

1. Dans notre version du sujet, la moyenne et la variance données sont fausses mais elles ont été corrigé à l'oral par l'encadrant.

2 Estimation statistique

On définit un estimateur simplifié issu de l'estimateur du maximum de vraisemblance en posant $a = \theta^p$:

$$\hat{a}_{MV} = \frac{1}{N}\sum_{i=1}^{N} Y_i^p \tag{2.1}$$

D'après le sujet [1], $\hat{a}_{MV}$ est non-biaisé et l'estimateur efficace de a. La borne de Cramér-Rao `BRC` renvoyée par la fonction `estimateur_mv` est donc la variance théorique de $\hat{a}_{MV}$:

$$\text{var}[\hat{a}_{MV}] = \frac{a^2}{N}$$

`alpha_est` est calculé en vectorisant l'équation 2.1 et en y injectant une matrice `Y` générée section 1.

Le tableau suivant récapitule les valeurs calculées.

	Moyenne	Variance	B.R.C
Numérique	`5.9938`	`0.036206`	-
Théorique	`5.9947`	`0.035937`	`0.035937`

TABLE 2 – Moyennes et variances numériques et théoriques de l'estimateur $\hat{a}_{MV}$. Borne de Cramér-Rao de $\hat{a}_{MV}$. $K = 500$.

L'erreur sur la moyenne est de l'ordre de 10^{-2} % et celle sur la variance de $0,7$ %. Les valeurs numériques sont très proches de celles attendues théoriquement. Le programme est vraisemblablement bon.

3 Détection

3.1 Courbes C.O.R théoriques

On cherche maintenant à établir un test statistique nous permettant de détecter si les données correspondent à un vent calme ou pas. Les hypothèses associées sont les suivantes :

$$H_0 : a = a_0 \text{ avec } \frac{2}{a_0}Y_i^p \sim \chi_2^2, i = 1, ..., N \tag{3.1}$$

$$H_1 : a = a_1 > a_0 \text{ avec } \frac{2}{a_1}Y_i^p \sim \chi_2^2, i = 1, ..., N \tag{3.2}$$

Pour calculer numériquement la puissance du test théorique : $\pi = 1 - \beta = 1 - G_{2N}\left(\frac{2\lambda_\alpha}{a_1}\right)$, on calcule tout d'abord le seuil de décision :

$$\lambda_\alpha = \frac{a_0}{2}G_{2N}^{-1}(1-\alpha) \tag{3.3}$$

où G_{2N} est la fonction `chi2cdf(X, 2N)`, G_{2N}^{-1} la fonction `chi2inv(X, 2N)` et α, `alpha = (0.01:0.01:0.99)`.

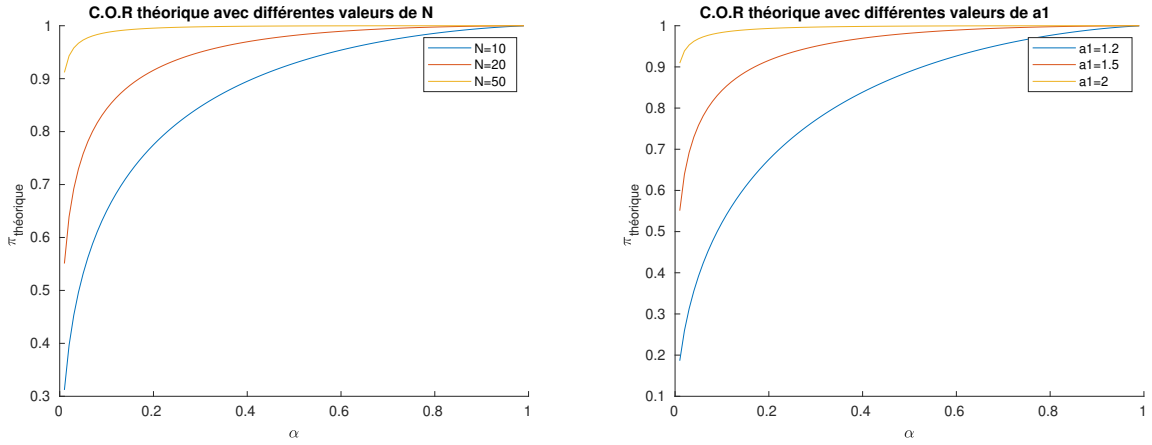

FIGURE 3.1 – Quand ils ne varient pas, les paramètres sont $a_0 = 0.9$, $a_1 = 1.5$ et $N = 20$.

Au vu du premier graphique de la Figure 3.1, plus N est grand et meilleur est le test statistique. En effet, il y a plus de données à tester donc plus qui vérifient [Rejeter $H_0|H_1$ vraie].

De même sur le deuxième graphique, plus a_1 est grand et meilleur est le test. En effet, plus a_1 et a_0 sont distants, plus il est simple de différencier leurs réalisations associées.

3.2 Courbes C.O.R numériques

Pour tracer une courbe C.O.R numériquement, il faut tout d'abord générer des signaux associés à l'hypothèse H_1 (3.2). On remarque que :

$$Y_i^p \sim \frac{a_1}{2}\chi_2^2$$

Donc on génère K réalisations de N éléments avec cette formule vectorisée. Une matrice de nombres aléatoires suivant une loi du χ_2^2 est générée par la fonction `chi2rnd(2, N, K)`.

Le seuil de décision est le même que pour la partie théorique 3.3.

On calcule la statistique de test de Neyman-Pearson (3.4) avec la fonction `sum` appliquée aux données précédemment générées.

$$T(\mathbf{Y}) = \sum_{i=1}^{N} Y_i^p \tag{3.4}$$

Ensuite, on compte les valeurs qui appartiennent à la région critique du test (3.5) dans une matrice `R` de taille $K \times 99$. K réalisation avec 99 valeurs de α possible.

$$R_\alpha = \{\mathbf{y} \in \mathbb{R}^N | T(\mathbf{y}) > \lambda_\alpha\} \tag{3.5}$$

Finalement, la puissance du test calculée numériquement est, à α fixé, la moyenne des K valeurs de `R` correspondantes. i.e `p = mean(R)`.

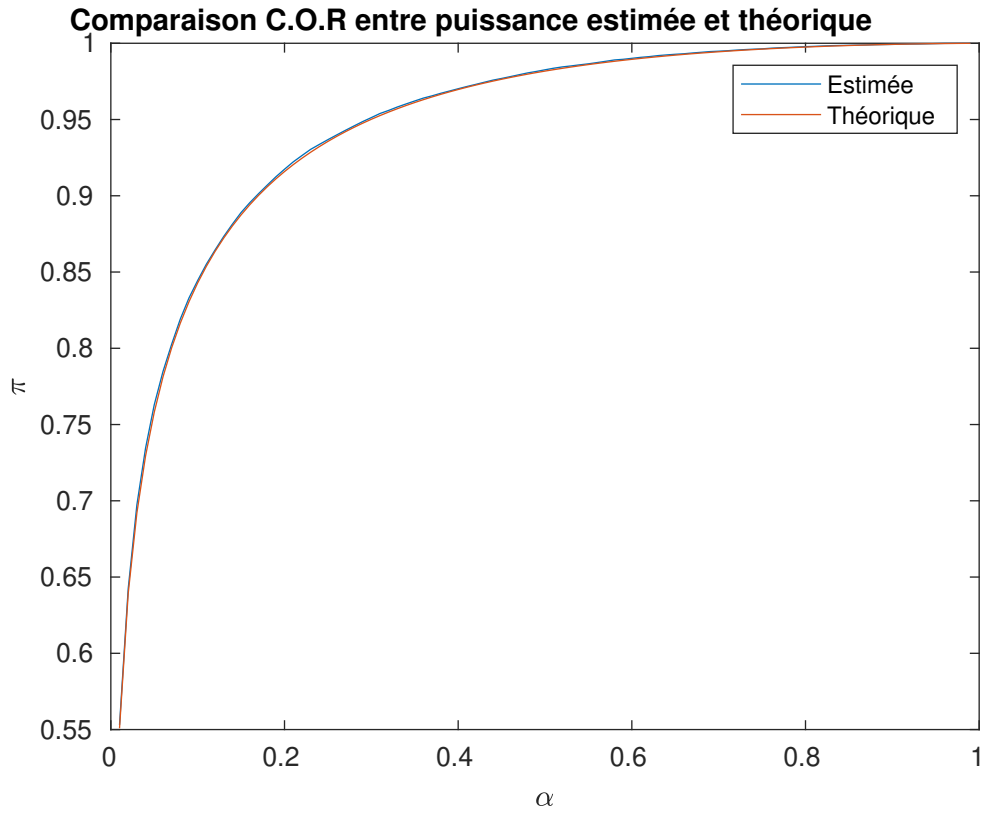

FIGURE 3.2 – Comparaison entre une courbe C.O.R calculée à l'instant et une théorique telle que vu Figure 3.1. $a_0 = 0.9$, $a_1 = 1.5$, $N = 20$ et $K = 50\,000$.

Les 2 courbes se superposent quasi parfaitement. L'estimation numérique est donc bonne pour ces paramètres de calcul.

4 Analyse d'un fichier de données

4.1 Fonctions de répartition

Dans cette section, on s'intéresse à un fichier de données fourni `wind.mat`.

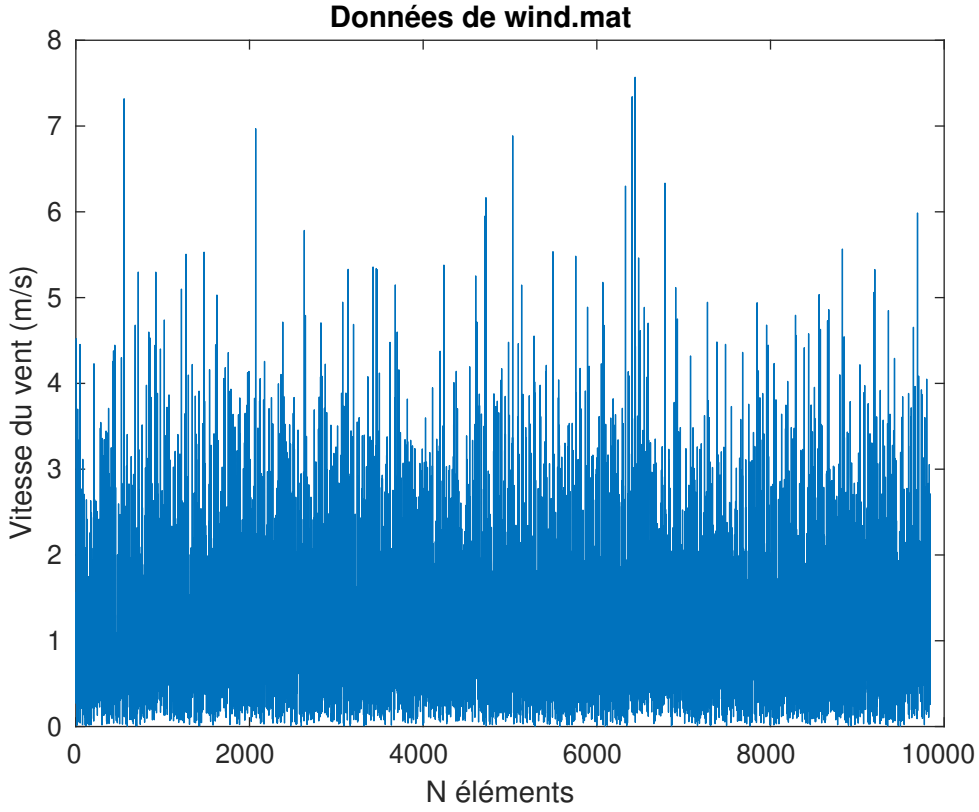

FIGURE 4.1 – Représentation des mesures de vitesse de vent contenues dans le vecteur `test`.

Les estimés des paramètres θ et p (notés $\hat{\theta}$ et $\hat{p}$) associés aux données contenues dans ce vecteur `test` sont obtenues à l'aide de la méthode du maximum de vraisemblance par la fonction `wblfit`. La fonction de répartition théorique est alors générée grâce à la fonction `wblcdf(x, theta, p)` où `x = sort(test)` et `theta` et `p` les estimés de θ et p.

La fonction de répartition empirique F est définie comme suit :

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\texttt{test(i)} \leqslant x} \tag{4.1}$$

Avec N le nombre d'éléments de `test`.

On peut ainsi tracer $F(X)$ avec X les valeurs de `test` triées dans l'ordre croissant par la fonction `sort`.

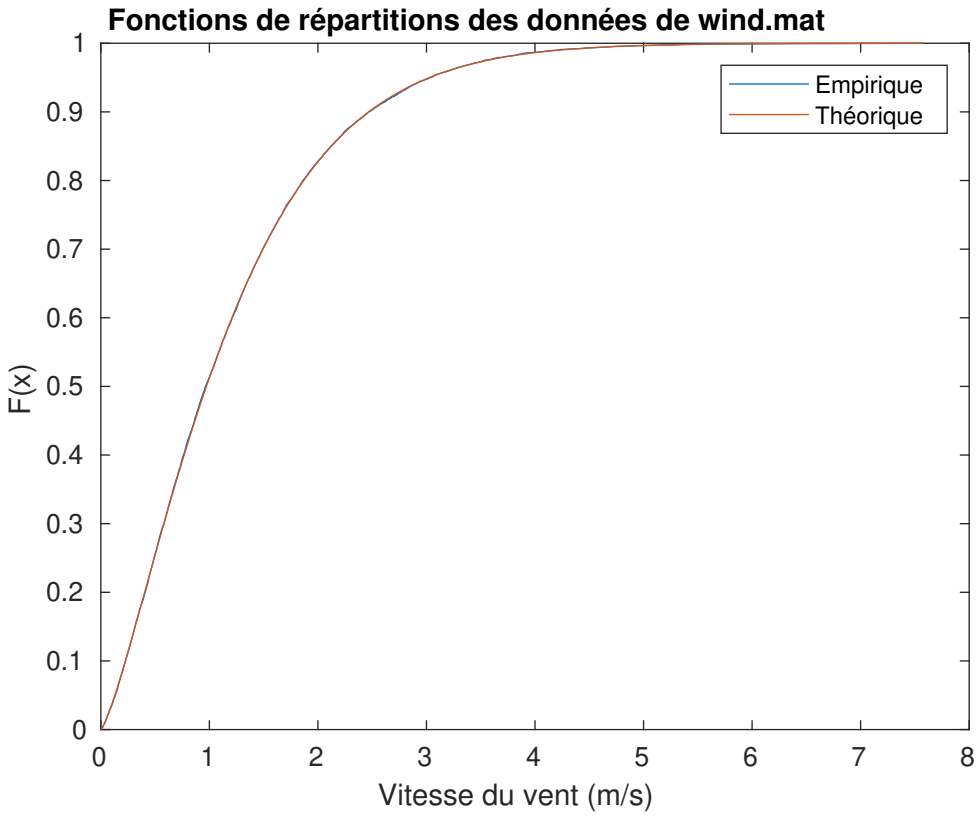

FIGURE 4.2 – Comparaison entre la fonction de répartition théorique d'une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ et la fonction de répartition empirique des données de `test`.

Les 2 fonctions de répartitions se superposent, on peut donc vraisemblablement penser que les données de `test` suivent une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$.

4.2 Test de Kolmogorov

Pour confirmer ce résultat, on applique à `test` un test de Kolmogorov.

Pour cela, on crée une fonction qui prend en argument un vecteur $\mathbf{y}$, son nombre d'éléments N_y et les paramètres estimés $\hat{\theta}$, $\hat{p}$ et renvoie les vecteur $\mathbf{E}^+$ et $\mathbf{E}^-$ définit ci-dessous.

$$E_i^+ = \left| \frac{i}{N} - F_W(y_i; \hat{\theta}, \hat{p}) \right|$$
$$E_i^- = \left| \frac{i-1}{N} - F_W(y_i; \hat{\theta}, \hat{p}) \right|$$

$F_W(y_i; \hat{\theta}, \hat{p})$ est la fonction de répartition d'une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ calculée par la fonction `wblcdf(y, theta, p)`.

On obtient alors comme statistique de test de Kolmogorov par cette méthode :

$$D_n = \texttt{5.039850e-03}$$

Cette valeur étant très faible (i.e le maximum des écarts entre théorie et numérique est très faible), il est raisonnable de penser que les données suivent une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$.

La fonction `kstest` renvoie exactement la même statistique de test : `5.039850e-03` donc le calcul numérique effectué précédemment est juste dans ces conditions.

De plus, la fonction `kstest` renvoie un seuil S_α pour $\alpha = 0,05$ de `1.367460e-02` $> D_n$ donc `test` suit une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ avec un risque $\alpha = 0,05$.

Conclusion

Au cours de ce BE, beaucoup de notions de statistiques ont été abordé et cela a permis d'avoir une visualisation concrète de ce qu'on manipulait en cours. La dernière partie traitant un fichier de données, elle montrait une application des statistiques sur des données inconnues que l'on a finalement identifiées. Ce BE était donc très formateur, en statistiques mais aussi en Matlab car c'est la première fois qu'on l'utilisait sur un devoir noté.

Références

[1] Équipe pédagogique de statistiques. Sujet du BE - Probas/Stats - 1MFEE. 2021.