Introduction

La vitesse du vent peut être modélisée par une variable aléatoire Y de loi de Weibull $\mathcal{W}(\theta, p)$ comme illustré sur la figure ci-dessous issue de [3][1]. L'objectif de ce BE est tout d'abord d'étudier certaines propriétés d'un estimateur du paramètre θ (pour p connu) construit à partir de n données $x_1, ..., x_n$ de loi de Weibull $\mathcal{W}(\theta, p)$. Dans un second temps, on s'intéressera à un test statistique permettant de décider si on est dans une période de vent calme correspondant à $\theta < 1\text{m s}^{-1}$ ou à une période de vent fort correspondant à $\theta > 1\text{m s}^{-1}$.

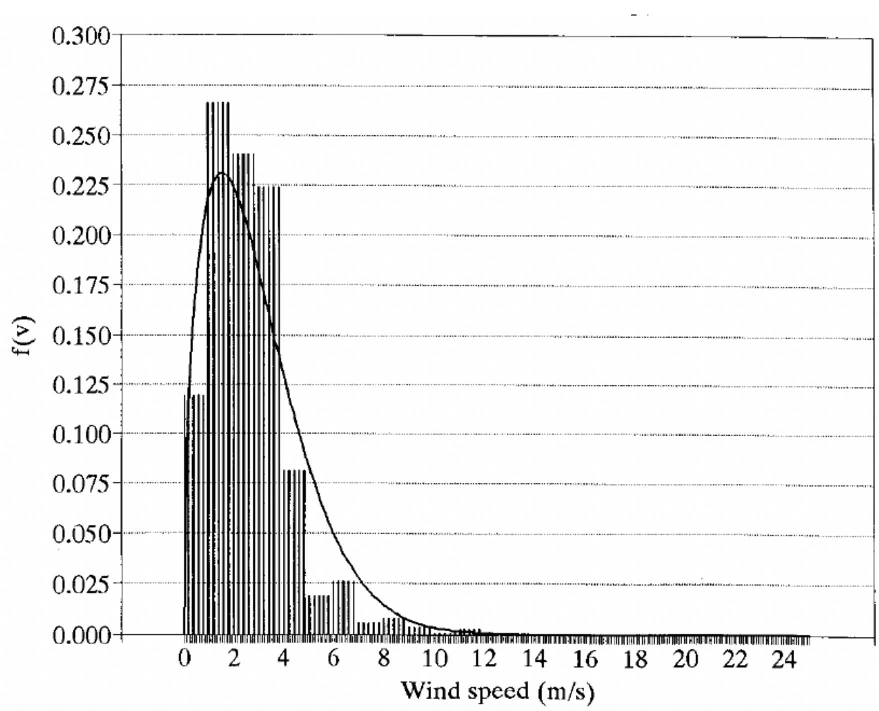

Figure 1: Histogramme de la vitesse du vent comparé à la densité de Weibull (figure extraite de [3]).

[1] Certaines propriétés de la loi de Weibull sont données à la fin de cet énoncé.

Travail à effectuer

1. **Génération d'un signal test**

   (a) Écrire une fonction `Y = generer(`θ`,`p`,`N`,`K`)` qui renvoie une matrice `Y` de taille $N \times K$, dont chaque colonne contient une réalisation du signal $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ de loi de Weibull $\mathcal{W}(\theta, p)$:

   - Pour effectuer cette génération, on utilisera le fait que si $F(x; \theta, p)$ est la fonction de répartition de cette loi de Weibull et que X est une variable aléatoire de loi uniforme sur l'intervalle $]0, 1[$, alors $Y = F^{-1}(X; \theta, p)$ suit la loi de Weibull $\mathcal{W}(\theta, p)$;
   - Les paramètres d'entrée sont θ et p : paramètres de la loi de Weibull, N : nombre de points d'un signal observé, K : nombre de signaux observés ;
   - Pour générer les réalisations du signal, on utilisera la fonction `rand(M,N)` de Matlab qui génère une matrice X de taille $M \times N$ constituée de réalisations indépendantes d'une loi uniforme sur l'intervalle $]0, 1[$ et on appliquera la fonction $Y = F^{-1}(X; \theta, p)$.

   (b) Tester cette fonction avec $N = 10000$, $K = 1$, $\theta_0 = 3.3$ et $p = 1.5$. Vérifier que l'histogramme des données générées est en accord avec la densité de la loi de Weibull à l'aide de la fonction `histfit`. Déterminer la moyenne et la variance des données générées à l'aide des fonctions `mean` et `var` et comparer avec les valeurs théoriques.

   (c) Générer une matrice de données avec $N = 1000$, $K = 500$, $\theta_0 = 3.3$ et $p = 1.5$. Afficher une réalisation du signal $\boldsymbol{y}$ (c'est-à-dire une colonne de la matrice `Y`) et tracer ensuite la moyenne et la variance des colonnes de `Y` à l'aide des fonctions `mean` et `var`. Comparer avec les résultats obtenus à la question précédente et commenter.

2. **Estimation statistique**

   (a) *Etude théorique* : montrer que l'estimateur du maximum de vraisemblance de θ construit à partir des observations $y_1, ..., y_N$ est défini par

   $$\widehat{\theta}_{\text{MV}} = \left(\frac{1}{N} \sum_{i=1}^{N} Y_i^p \right)^{1/p}.$$

   Comme cet estimateur n'est pas simple à étudier, on s'intéresse plutôt à $a = \theta^p$. En appliquant le principe d'invariance fonctionnelle, on obtient l'estimateur du maximum de vraisemblance de a

   $$\widehat{a}_{\text{MV}} = \frac{1}{N} \sum_{i=1}^{N} Y_i^p.$$

   (b) *Etude théorique* : montrez que

   - $\widehat{a}_{\text{MV}}$ est un estimateur non-biaisé de $a = \theta^p$, c'est-à-dire

   $$E\left[\widehat{a}_{\text{MV}}\right] = a$$

   - la variance de $\widehat{a}_{\text{MV}}$ est définie par

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

   - $\widehat{a}_{\text{MV}}$ est l'estimateur efficace de a, c'est-à-dire que sa variance est égale à la borne de Cramér-Rao des estimateurs non-biaisés de a.

   (c) Ecrire une fonction `alpha_est= estimateur_mv(Y,`p`,`N`,`K`)`, qui renvoie l'estimateur $\widehat{a}_{\text{MV}}$ pour chacune des K réalisations de $y = (y_1, \ldots, y_N)^T$, à partir de la matrice `Y` construite à la question 1. On obtient alors K valeurs de $\widehat{a}_{\text{MV}}$, notées $\left(\widehat{a}_{\text{MV}}(k)\right)_{k=1,\ldots,K}$.

(d) Représenter graphiquement les valeurs $(\widehat{a}_{\text{MV}}(k))_{k=1,\ldots,K}$ ainsi que leur moyenne et leur variance et comparer avec les valeurs théoriques.

3. **Détection** On cherche à étudier les performances d'un test statistique qui permet de détecter si les données y_i correspondent à un vent calme ($a = a_0 < 1ms^{-1}$) ou pas ($a = a_1 > 1ms^{-1}$). Pour simplifier, on supposera dans ce BE que les valeurs de a_0 et a_1 sont connues.

   (a) *Etude théorique* : montrez que si Y_i suit une loi de Weibull $\mathcal{W}(\theta, p)$, alors $Z_i = \frac{2}{a} Y_i^p$ suit une loi du chi2 à 2 degrés de liberté, i.e., $Z_i \sim \chi_2^2$.
   Les deux hypothèses associées à la détection d'un vent calme sont alors définies par

   $$H_0 : \quad a = a_0 \qquad H_1 : \quad a = a_1 > a_0 \tag{1}$$

   (b) *Etude théorique* : montrez que

   - la statistique de test issue du théorème de Neyman-Pearson associée à ces deux hypothèses s'écrit
   $$T(\boldsymbol{Y}) = \sum_{i=1}^{N} Y_i^p.$$
   - pour une probabilité de fausse alarme α, la région critique du test (zone de rejet de H_0) est définie par
   $$R_\alpha = \left\{ \boldsymbol{y} \in \mathbb{R}^N | T(\boldsymbol{y}) > \lambda_\alpha \right\}. \tag{2}$$
   - le seuil de décision s'écrit
   $$\lambda_\alpha = \frac{a_0}{2} G_{2N}^{-1}(1-\alpha) \tag{3}$$
   où G_{2N}^{-1} est l'inverse de la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté (une loi χ_{2N}^2). On montre également que la probabilité de non-détection (ou risque de 2ème espèce) du test s'exprime sous la forme suivante
   $$\beta = G_{2N}\left(\frac{2\lambda_\alpha}{a_1}\right)$$
   où G_{2N} est la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté.

   On souhaite tracer les courbes théoriques de la puissance du test $\pi = 1 - \beta$ en fonction de la probabilité de fausse alarme α, puis retrouver ces courbes par simulations.

   (c) En utilisant les fonctions `chi2inv` et `chi2cdf`, écrire une fonction

   π `= pi_theorique(`a_0`,`a_1`,`L`)`

   qui renvoie la puissance théorique π du test pour $\alpha \in \{0.01, 0.02, \ldots, 0.98, 0.99\}$ en fonction des paramètres a_0, a_1 et $L = 2N$. Tracer les courbes obtenues pour $a_0 = 0.9$, $a_1 = 1.5$ et différentes valeurs de N ($N \in \{10, 20, 50\}$). Tracer les courbes obtenues pour $a_0 = 0.9$, $N = 20$ et différentes valeurs de a_1 ($a_1 \in \{1.2, 1.5, 2\}$). Commenter les résultats obtenus.

   (d) On cherche maintenant à retrouver ces résultats par simulation. Puisque a_0 est un paramètre connu, on peut déterminer le seuil du test λ_α pour toute valeur de α en utilisant (3). Pour estimer la puissance du test, il suffit donc d'estimer la probabilité $P[\text{Rejeter } H_0 \, | H_1 \text{ vraie}]$. Pour cela

   - Ecrire une fonction qui génère K réalisations de signaux de longueur N associés à l'hypothèse H_1 du test (1).
   - Écrire une fonction

   $\hat{\pi}$ `= pi_estimee(`a_0`,`a_1`,`L`,`K`)`

qui renvoie la puissance estimée $\hat{\pi}$ du test pour $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, en fonction de a_0, a_1, L, et du nombre de simulations K. La puissance sera estimée à l'aide des signaux associés à l'hypothèse H_1 générés à la question précédente. Superposer la courbe COR théorique obtenue avec la fonction `pi_theorique` et la courbe COR estimée $\hat{\pi}$ avec $a_0 = 0.9$, $a_1 = 1.5$, $N = 20$ et $K = 50000$ ou $K = 1000$. Commenter.

4. **Analyse d'un fichier de données** On désire dans cette partie analyser un fichier de données contenant des mesures de vitesse de vent.

   (a) Charger le fichier `wind.mat` et représenter graphiquement les mesures de vitesse de vent contenues dans le vecteur `test`.

   (b) À l'aide de la fonction `wblfit`, déterminer des estimées des paramètres θ et p associées à ce vecteur de données (notées $\widehat{\theta}$ et $\widehat{p}$) obtenues à l'aide de la méthode du maximum de vraisemblance.

   On désire vérifier qu'il est raisonnable de penser que les données du vecteur `test` sont distribuées suivant une loi de Weibull $\mathcal{W}\left(\widehat{\theta}, \widehat{p}\right)$ à l'aide d'un test de Kolmogorov.

   (c) Représenter sur la même figure la fonction de répartition de la loi de Weibull $\mathcal{W}\left(\widehat{\theta}, \widehat{p}\right)$ évaluée aux données du fichier `wind.mat` (rangées par ordre croissant avec la fonction `sort` et notées $y_1 < y_2 < ... < y_N$) et la fonction de répartition empirique de ces données (tout d'abord en ne considérant que $N = 100$ données puis avec la totalité du fichier).

   (d) Écrire une fonction qui permet de calculer les écarts E_i^+ et E_i^- définis par

   $$E_i^+ = \left| \frac{i}{N} - F_{\mathrm{W}}\left(y_i; \widehat{\theta}, \widehat{p}\right) \right|, \quad E_i^- = \left| \frac{i-1}{N} - F_{\mathrm{W}}\left(y_i; \widehat{\theta}, \widehat{p}\right) \right|$$

   pour $i = 1, ..., N$, où N est le nombre de données x_i du vecteur `test` et $F_{\mathrm{W}}\left(x_i; \widehat{\theta}, \widehat{p}\right)$ est la fonction de répartition d'une loi de Weibull $\mathcal{W}\left(\widehat{\theta}, \widehat{p}\right)$ (fonction `wblcdf`). En déduire la valeur de la statistique du test de Kolmogorov et conclure.

   (e) Vérifier le résultat de la question précédente en utilisant la fonction `kstest` de Matlab.

References

[1] A. C. Cohen. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. *Technometrics*, 7(4):579–588, 1965.

[2] A. Joarder, H. Krishna, and D. Kundu. Inferences on Weibull parameters with conventional type-I censoring. *Comput. Stat. Data Anal.*, 55(1):1–11, 2011.

[3] M. Martin, L. V. Cremades, and J. M. Santabárbara. Analysis and modelling of time series of surface wind speed and direction. *Journal of Climatology*, 19(2):197–209, 1999.

- *Loi de Weibull* $\mathcal{W}(\theta, p)$ $\qquad p > 0, \theta > 0,\ x \in \mathbb{R}^+$
  - *densité :* $f(x;\theta,p) = \frac{p}{\theta}\left(\frac{x}{\theta}\right)^{p-1} \exp\left[-\left(\frac{x}{\theta}\right)^p\right] \mathcal{I}_{\mathbb{R}^+}(x)$
  - *Fonction de répartition :* $F(x;\theta,p) = 1 - \exp\left[-\left(\frac{x}{\theta}\right)^p\right] \mathcal{I}_{\mathbb{R}^+}(x)$
  - *Moyenne :* $\mu = \theta\Gamma\left(1 + \frac{1}{p}\right)$
  - *variance :* $\theta^2\Gamma\left(1 + \frac{2}{p}\right) - \mu^2$
- *Loi du chi2 à* L *degrés de liberté* χ^2_L $\qquad L > 0,\ x \in \mathbb{R}^+$
  - *définition : si* $X_1, ..., X_L$ *sont* L *variables aléatoires indépendantes de même loi normale* $\mathcal{N}(0,1)$ *alors la loi de* $\sum_{l=1}^L X_l^2$ *est une loi du* χ^2_L
  - *densité :* $f(x;L) = \frac{1}{2^{\frac{L}{2}}\Gamma\left(\frac{L}{2}\right)} x^{\frac{L}{2}-1} \exp\left(-\frac{x}{2}\right) \mathcal{I}_{\mathbb{R}^+}(x)$
  - *Moyenne :* L
  - *variance :* $2L$

Estimation des paramètres d'une de loi de Weibull $\mathcal{W}(\theta, p)$ [1, 2]

Pour estimer les deux paramètres d'une loi de Weibull par la méthode du maximum de vraisemblance, il est préférable de re-paramétrer la densité en posant $a = \theta^p$. La log-vraisemblance s'écrit alors

$$\ln L(y_1, ..., y_N; a, p) = N \ln\left(\frac{p}{a}\right) + (p-1)\sum_{n=1}^N \ln y_n - \frac{1}{a}\sum_{n=1}^N (y_n^p \ln y_n)$$

qui admet pour dérivées partielles

$$\begin{aligned} \frac{\partial \ln L(y_1, ..., y_N; a, p)}{\partial a} &= -\frac{N}{a} + \frac{1}{a^2}\sum_{n=1}^N y_n^p \\ \frac{\partial \ln L(y_1, ..., y_N; a, p)}{\partial p} &= \frac{N}{p} + \sum_{n=1}^N \ln y_n - \frac{1}{a}\sum_{n=1}^N (y_n^p \ln y_n)\,. \end{aligned} \tag{4}$$

En annulant ces deux dérivées partielle, on montre facilement que p est solution de l'équation

$$\frac{1}{p} = h(p) = \frac{\sum_{n=1}^N (y_n^p \ln y_n)}{\sum_{n=1}^N y_n^p} - \frac{1}{N}\sum_{n=1}^N \ln y_n$$

qui est une équation de point fixe $p = \frac{1}{h(p)}$ qu'on peut résoudre 1) en trouvant une solution initiale suffisamment proche de la solution recherchée ou 2) à l'aide de l'algorithme de Newton-Raphson basé sur la récursion

$$p_{k+1} = p_k - \frac{g(p_k)}{g'(p_k)}$$

où $g(p) = p - \frac{1}{h(p)}$.