# Ajustement-Hr-Bordeaux-Sujet

February 13, 2022

# 1 Analyse de l'humidité relative à Bordeaux en 2018

## 1.1 Description

Dans ce document nous analysons un échantillon d'humidité relative mesurée à Bordeaux en 2018.

L'humidité relative (HR) de l'air est le "rapport de la pression partielle de la vapeur d'eau contenue dans l'air sur la pression de vapeur saturante (ou tension de vapeur) à la même température". C'est une mesure que l'on peut considérer entre 0 et 100 % dans la plupart des situations à la surface de la Terre (mais la formation des nuages et la précipitation qui peut s'ensuivre nécessite une HR légèrement supérieure à 100%).

L'objectif de cet exercice est de trouver une loi de distribution appropriée pour cet échantillon.

## 1.2 Données

Il s'agit d'un échantillon de taille 2779 associée à une mesure toutes les trois heures du 1er janvier au 31 décembre 2018 :

`Humidite-relative-Bordeaux-2018.csv`

Les données sont issues du site www.infoclimat.fr. Infoclimat est une association à but non lucratif régie par la loi de 1901 créée le 15 Octobre 2003 par les co-auteurs du site internet www.infoclimat.fr (dont la création remonte au 7 Octobre 2001).

## 1.3 Loi Beta

La variable $X$ suit la distribution Beta si sa densité de probabilité est :

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}$$

pour tout $x \in ]a, b[$ où

- $a$ est le minimum,
- $b$ est le maximum,
- $\alpha > 0$ est le premier paramètre de forme,
- $\beta > 0$ est le second paramètre de forme,

et $B$ est la fonction Beta définie par :

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

La log-densité de la loi beta est :

$$\log(f(x)) = (\alpha - 1)\log(x-a) + (\beta - 1)\log(b-x) - (b-a)\log(\alpha + \beta - 1) - \log(B(\alpha, \beta))$$

pour tout $x \in ]a, b[$.

### 1.4 Moments et méthode des moments

Les moments de la loi Beta sont

$$\mathbb{E}(X) = a + (b-a)\frac{\alpha}{\alpha + \beta}$$

et

$$\mathrm{Var}(X) = (b-a)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

On peut inverser les équations précédentes et calculer $\alpha$ et $\beta$ en fonction de la moyenne $\mu$ et de la variance $\sigma^2$ :

$$\alpha = \frac{\mu - a}{b - a}\left(\frac{(b-\mu)(\mu - a)}{\sigma^2} - 1\right)$$

et

$$\beta = \frac{b - \mu}{\mu - a}\alpha.$$

On suppose que l'on dispose d'un échantillon $x_1, \ldots, x_n$ de réalisations indépendantes issues de la loi Beta.

La méthode des moments consiste à utiliser les équations précédentes pour estimer les paramètres de la loi Beta. La méthode des moments utilise les estimateurs suivants. Soient $x_{min}$ et $x_{max}$ le minimum et le maximum de l'échantillon :

$$\hat{x}_{min} = \min(x_1, ..., x_n), \quad \hat{x}_{max} = \max(x_1, ..., x_n).$$

On estime l'étendue par :

$$\hat{\delta} = \hat{x}_{max} - \hat{x}_{min}.$$

Les paramètres de borne sont estimés par l'équation :

$$\hat{a} = \hat{x}_{min} - \frac{\hat{\delta}}{n+2}, \quad \hat{b} = \hat{x}_{max} + \frac{\hat{\delta}}{n+2}.$$

Soient $\bar{x}$ la moyenne empirique et $\hat{\sigma}^2$ la variance empirique non biaisée. Alors les paramètres de forme sont estimés par les équations :

$$\hat{\alpha} = \frac{\bar{x} - a}{b - a} \left( \frac{(b - \bar{x})(\bar{x} - a)}{\hat{\sigma}^2} - 1 \right)$$

et

$$\hat{\beta} = \frac{b - \bar{x}}{\bar{x} - a} \alpha.$$

### 1.5 Estimation par maximum de vraisemblance

On suppose que l'on dispose d'un échantillon $x_1, \ldots, x_n$ de réalisations indépendantes issues de la loi Beta. La log-vraisemblance est :

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(x_i, \theta))$$

ce qui implique:

$$\ell(\theta) = (\alpha - 1) \sum_{i=1}^{n} \log(x_i - a) + (\beta - 1) \sum_{i=1}^{n} \log(b - x_i) - n(\alpha + \beta - 1) \log(b - a) - n \log(B(\alpha, \beta))$$

où $\theta = (\alpha, \beta, a, b)$ est le vecteur des paramètres.

Pour les paramètres de borne, l'estimateur du maximum de vraisemblance est :

$$\hat{a} = \min(x_1, ..., x_n), \quad \hat{b} = \max(x_1, ..., x_n).$$

Pour estimer les paramètres de forme $\alpha$ et $\beta$, on peut alors utiliser une méthode d'optimisation et maximiser la log-vraisemblance. Toutefois, la log-vraisemblance est singulière lorsque $x = \min(x_1, ..., x_n)$ et $x = \max(x_1, ..., x_n)$. Pour résoudre ce problème, on peut élargir un peu les estimateurs des paramètres de borne et utiliser plutôt :

$$\hat{a} = \min(x_1, ..., x_n)(1 - \epsilon), \quad \hat{b} = \max(x_1, ..., x_n)(1 + \epsilon),$$

où $\epsilon > 0$ est un réel proche de zéro. En pratique, on peut choisir $\epsilon = 10^{-16}$, par exemple.

## 1.6 Références

- https://www.infoclimat.fr
- Humidité relative. (2019, janvier 31). Wikipédia, l'encyclopédie libre. Page consultée le 10:03, janvier 31, 2019 à partir de http://fr.wikipedia.org/w/index.php?title=Humidit%C3%A9_relative&oldid=156332196.

## 1.7 Questions

## 1.8 Question 1 : description des données

*Questions* :

- Utiliser la classe `Sample_ImportFromTextFile` pour lire le fichier suivant `"Humidite-relative-Bordeaux-2018.csv"`. Dans ce but, on utilisera le séparateur de colonnes `;`.
- Afficher la taille de l'échantillon.
- Utiliser la classe `HistogramFactory` pour créer l'histogramme de cette échantillon et la méthode `drawPDF()` pour dessiner l'échantillon.
- Observer la distribution : que remarquez-vous ?

## 1.9 Question 2 : ajustement de la loi bêta

La classe `BetaFactory` utilise la méthode des moments pour estimer les paramètres de la loi Beta.

*Questions* :

- Utiliser la classe `BetaFactory` pour identifier les paramètres de la loi bêta qui s'ajustent aux données et créer la distribution `HRBeta` correspondante. Qu'observez-vous ?
- Dessiner la PDF de la loi ajustée et superposer l'histogramme des données pour valider graphiquement l'ajustement. L'ajustement est-il satisfaisant ?
- Utiliser la classe `VisualTest_DrawQQplot` pour valider graphiquement l'ajustement. L'ajustement est-il satisfaisant ?
- Utiliser la classe `Lilliefors` pour réaliser le test de Kolmogorov-Smirnov dans le cas où les paramètres sont estimés. Quel est le résultat du test ?

## 1.10 Question 3 : distribution de l'estimation des paramètres

On souhaite estimer la distribution des paramètres estimés pour la loi bêta.

*Questions* :

- Utiliser la méthode `buildEstimator` de la classe `BetaFactory` pour créer l'estimateur de la distribution des paramètres. Quelle est la méthode utilisée dans ce cas particulier ?
- Pour chaque marginale de cette distribution, dessiner la densité de probabilité.
- Pour chaque marginale de cette distribution, calculer un intervalle de confiance bilatéral de niveau 95%.
- Quels sont les paramètres qui semblent précisément estimés et quels sont ceux qui ne le sont pas ?

### 1.11 Question 4 : recherche de la loi s'ajustement le mieux au sens du BIC

On souhaite déterminer la loi qui s'ajuste le mieux au sens du critère Bayesian Information Criterion. Ce critère permet de pénaliser les distributions ayant un nombre plus élevé de paramètres.

*Questions* :

- Utiliser la classe `DistributionFactory_GetContinuousUniVariateFactories` pour créer la variable `distcoll` contenant la liste des distributions continues. Combien y-a-t-il de distributions dans cette liste ?
- Utiliser la classe `FittingTest_BestModelBIC` pour identifier la loi qui s'ajuste le mieux au sens du critère BIC.

### 1.12 Question 5 : avec une méthode d'estimation à noyau

On souhaite utiliser une méthode d'estimation non paramétrique pour approcher la distribution de l'échantillon.

*Questions* :

- Utiliser la classe `KernelSmoothing` pour créer un distribution fondée sur une méthode à noyau. Dessiner la distribution. Qu'observez-vous ?
- Utiliser la méthode `setBoundingOption` pour configurer la méthode de correction du bord droit de la distribution. Dessiner la distribution. Qu'observez-vous ?
- Utiliser la classe `Kolmogorov` pour réaliser le test de Kolmogorov-Smirnov. Quel est le résultat du test ?

### 1.13 Question 6 : synthèse

*Questions* :

- Dans le même graphe, dessiner l'histogramme empirique de l'échantillon, la loi de `HRBeta` et l'estimation à noyau.
- Que peut-on conclure ?

### 1.14 Question 7 : estimation d'une loi bêta par la méthode du maximum de vraisemblance

On souhaite estimer les paramètres d'une distribution bêta par la méthode du maximum de vraisemblance.

*Questions* :

- Utiliser les méthodes `getMin` et `getMax`pour obtenir les minimums et maximum de l'échantillon. Puis utiliser les instuctions présentées dans l'introduction pour estimer les paramètres de borne $a$ et $b$.
- Utiliser la classe `MaximumLikelihoodFactory` pour créer l'objet `factory` contenant un estimateur fondé sur la méthode du maximum de vraisemblance.
- Utiliser la méthode `setKnownParameter` pour fixer les paramètres $a$ et $b$ aux valeurs calculées précédemment.
- Les paramètres $\alpha$ et $\beta$ sont strictement positifs. Utiliser les instructions suivantes pour configurer les bornes inférieures de la recherche :

```
bounds_lower = [ot.SpecFunc.MinScalar] * 2
bounds_upper = [1.0] * 2  # Disable anyway (ignored)
interval = ot.Interval(bounds_lower, bounds_upper, [True] * 2, [False] * 2)
factory.setOptimizationBounds(interval)
```

- Estimer les paramètres en utilisant la méthode `build`. Quels sont les valeurs des paramètres paramètres estimés par la méthode ?
- Créer un graphique qui superpose l'histogramme empirique de l'échantillon et la distribution Bêta dont les paramètres ont étés estimés par la méthode du maximum de vraisemblance. Que peut-on dire de la qualité de l'ajustement ?
- Dessiner le QQ-Plot pour cette distribution.
- Conclure.

### 1.15 Question 8 : dessin de la log-vraisemblance

On souhaite visualiser la forme de la log-vraisemblance en fonction de $\alpha$ et $\beta$. Puisque c'est une fonction bidimensionnelle, il est plus facile d'observer les iso-valeurs de la log-vraisemblance.

*Questions* :

- Définir la fonction `log_likelihood` qui prend en entrée un point de dimension 2 représentant les paramètres $\alpha$ et $\beta$ et retourne un point de dimension 1 représentant la log-vraisemblance.
- Créer l'objet `log_likelihood_Py` comme une instance de la classe `PythonFunction` pour créer une fonction OpenTURNS correspondant à la fonction de log-vraisemblance `log_likelihood` définie précédemment.
- Utiliser la méthode `draw` de la fonction `log_likelihood_Py` pour dessiner les iso-valeurs de la log-vraisemblance.
- Superposer dans le même graphique le point correspondant à l'estimateur du maximum de vraisemblance.
- Commenter le résultat.

### 1.16 Question 9 : estimation de la distribution des paramètres

On souhaite estimer la distribution de l'estimateur de $\theta$ par la méthode du maximum de vraisemblance.

*Questions* :

- Utiliser la méthode `buildEstimator` de l'objet `factory` associé à la classe `MaximumLikelihoodFactory` pour obtenir une distribution de $\widehat{\theta}$.
- Puisque c'est une distribution en dimension 4, on ne peut pas la visualiser facilement. Utiliser la méthode `getMarginal` pour stocker dans la variable `alpha_beta_distribution` la loi jointe du vecteur $\left(\widehat{\alpha}, \widehat{\beta}\right)$ dont les indices sont 0 et 1.
- Utiliser la méthode `draw` de la distribution pour représenter les iso-valeurs de la densité de probabilité de la loi de l'estimateur.
- Qu'observez-vous ? Qu'en concluez-vous ?
- On souhaite déterminer les valeurs les plus probables du vecteur $\left(\widehat{\alpha}, \widehat{\beta}\right)$ pour évaluer un domaine de confiance bidimensionnel. Utiliser la méthode `computeMinimumVolumeLevelSetWithThreshold` pour déterminer le levelset de volume minimal contenant 95% de la distribution du vecteur aléatoire $\left(\widehat{\alpha}, \widehat{\beta}\right)$.
- Qu'observez-vous ? Pouvez-vous expliquer la forme particulière du domaine de confiance ?

- Que pouvez-vous conclure sur les valeurs vraisemblables des paramètres $\hat{\alpha}$ et $\hat{\beta}$ ?

## 1.17 Question 10 : Synthèse

L'objectif de cette question est d'avoir une vue d'ensemble sur l'estimation des paramètres par les différentes méthodes.

*Questions* :

- Dans un tableau, rassemblez les paramètres estimés par les différentes méthodes.
- Comparer les résultats et analysez les différences.
- Comment pourrait-on aller plus loin dans l'analyse ou la mise en oeuvre des différentes méthodes d'estimation ?

[ ]:

```

```