Devoir 1 : Statistiques descriptives, modèles et estimation

Le staff de MATH0487-2

Novembre 2024

Nom(s), Prénom(s) et matricule(s) :

VOTRE TEXTE ICI

Instructions générales

Objectifs Les objectifs de ce devoir sont les suivants :

- explorer un jeu de données et en extraire différentes statistiques descriptives,
- manipuler les paramètres d'un modèle statistique et comprendre le concept de vraisemblance (*likelihood*, en anglais),
- estimer les paramètres d'un modèle statistique selon différentes méthodes et apporter un regard critique sur la modélisation,
- étudier la convergence d'un estimateur.

Délivrables Ce devoir doit être réalisé par groupe de 2 étudiants maximum. Chaque groupe doit rendre ce notebook complété, et **une version .pdf de ce notebook** qui sera utilisée pour la correction.

La date limite de soumission est fixée au **29 novembre 2024 à 20h00**. Jusqu'à cette date, vous avez la possibilité de (re)soumettre votre rapport ou votre code autant de fois que vous le souhaitez. Au-delà de cette date, il ne sera plus possible de soumettre le devoir. N'attendez pas la dernière minute pour soumettre une première version de votre travail !

La soumission doit se faire sur la plateforme Gradescope directement.

- Chaque étudiant doit s'inscrire sur Gradescope en utilisant son adresse `@student.uliege.be`. Si vous ne voyez pas le cours MATH0487 dans votre tableau de bord, contactez-nous sur Ed au plus vite (n'attendez pas la veille de la date de soumission pour vérifier que vous avez accès au cours sur Gradescope ;-).
- Chaque groupe doit soumettre un seul fichier `.ipynb` et un seul fichier `.pdf` sur Gradescope. Toutes les cellules doivent être exécutées et leurs sorties ne doivent pas être effacées avant la soumission. Assurez-vous que tous les membres du groupe sont correctement ajoutés à la soumission !
- N'oubliez pas d'assigner les pages de votre pdf aux questions sur Gradescope !

Si vous n'êtes pas familiers avec Gradescope, vous trouverez des explications sur chaque étape de la soumission ci-dessous :

- Soumission de pdf,
- Soumission de code,
- Ajout de membres de groupe.

Attention: Pour convertir votre notebook en pdf, nous conseillons l'utilisation de nbconverter. Cet outil fait partie de l'écosystème Jupyter, et peut être directement utilisé pour l'exportation de notebooks. Il peut également être installé via `pip`ou `conda`. nbconverter requiert également l'installation de Pandoc et TeX (en particulier, XeLaTeX). Les instructions pour installer ces packages sont disponibles sur la page d'installation renseignée ci-avant.

N'attendez pas la veille de la date limite de soumission pour tester que cette conversion se fait correctement avec votre installation ! Le staff peut vous aider si vous en avez le besoin, mais - nous ne règlerons pas de problème de conversion de "dernière minute", - nous n'accepterons pas de pdf illisibles, - nous ne corrigerons pas de fichier .ipynb.

Vous êtes responsables de la qualité du document soumis. Si, par exemple, une figure ne s'affiche pas correctement après conversion, il vous incombe de régler ce problème (en demandant de l'aide au staff si nécessaire).

Remarques importantes sur l'utilisation de ce notebook :

- Ne modifiez et ne supprimez pas de cellules (Markdown) contenant des consignes/questions.
- Les cellules demandant une réponse sous forme de texte ou sous forme de code sont colorées en vert,
  comme ceci.
- Remplissez uniquement les cellules prévues à cet effet :
  - soit dans une une portion réservée à votre code comme ci-dessous
    `"""VOTRE CODE ICI"""`
  - soit dans une portion réservée à une réponse écrite comme ci-dessous
    VOTRE RÉPONSE ÉCRITE ICI

  Ne créez pas de nouvelles cellules.
- Respectez le type de cellule prévu pour une question donnée : certaines questions demandent d'implémenter du code (cellules "Code", en Python) et de présenter des résultats (valeurs numériques, tables, graphes, ...), et d'autres vous demandent de fournir une réponse utilisant du texte (cellules "Markdown", incluant certaines commandes LaTeX et acceptant la syntaxe HTML).
- Lorsque vous présentez un graphique, n'oubliez pas d'indiquer un titre et / ou des noms pour les axes. Lorsque cela est nécessaire, affichez également la légende.
- Un exemple de cellule Markdown comprenant des commandes LaTeX est donné ci-dessous :

`début de l'exemple`

Ceci est un *exemple* de cellule de texte **Markdown**. Double-cliquez dessus pour voir le texte brut.

Vous pouvez utilisez certaines commandes LaTeX comme $\sin(x)$ ou α. Il est possible d'écrire des équations comme

$$\beta \dot{y} = 3x$$

ou encore

$$\begin{aligned} \beta\dot{y} &= 3x \\ \gamma\dot{x} &= 4y. \end{aligned} \tag{1}$$

Des listes sont également disponibles :

- élément 1
- élément 2
- …

ou

- élément 1
- élément 2
- …

N'hésitez pas à vous renseigner sur la syntaxe Markdown si vous avez besoin d'autres éléments (*e.g.* construire des tables).

`fin de l'exemple`

- Enfin, veillez à toujours présenter vos résultats en sortie de cellule quand cela est nécessaire (figures, valeurs numériques, etc.). **Une cellule non exécutée ou dont les valeurs calculées et demandées ne sont pas affichées sera considérée comme non implémentée.**

Si vous rencontrez des problèmes ou avez des questions concernant ces remarques, merci de contacter l'équipe pédagogique *via* le forum de Ed Discussion.

Questions Toutes vos questions sur le devoir (y compris sur l'utilisation de Python ou de Jupyter) doivent être postées dans le forum de Ed Discussion du cours sous la catégorie *Assignments/Homework* (une question par fil de discussion).

Politique de collaboration Vous pouvez discuter du devoir avec d'autres groupes, mais *vous devez écrire vous-même vos propres solutions, et écrire et exécuter vous-même votre propre code.* Copier la solution de quelqu'un d'autre, ou simplement apporter des modifications triviales pour ne pas copier textuellement, n'est pas acceptable.

Présentation du problème Le staff du cours MATH0487-2 est, comme vous (sans aucun doute), fasciné par les tremblements de terre ! Ce devoir vous propose d'explorer les différents concepts du cours théorique et des TP en les appliquant à un jeu de données réel, et ce, à l'aide d'outils numériques.

Plus précisément, les données qui vous sont fournies recueillent des observations réalisées pour 1000 tremblements de terre, pour lesquels ont été reportées plusieurs caractéristiques telles que :

- le lieu
- la date
- la magnitude
- …

C'est cette dernière grandeur, la magnitude du tremblement, que nous allons étudier à l'aide de notre échantillon. L'ensemble des données peut être chargé depuis le fichier `data_math0487.csv` de l'archive fournie.

[ ]:
```
# ces librairies devraient suffir à réaliser ce premier devoir, vous pouvez
↪évidemment en utiliser d'autres (à importer dans cette cellule).
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import seaborn as sns
from sklearn.neighbors import KernelDensity
```

Importation des données

[ ]:
```
# chargement du dataset
data = pd.read_csv("data_math0487.csv")
var = "magnitude"
data = data[var]
```

---

1 Analyse descriptive

1.1 Statistiques d'échantillon, données aberrantes, ECDF et histogramme

Dans cette première partie du devoir, vous allez explorer l'échantillon qui vous est fourni. Vous allez analyser comment sont distribuées ces données et en extraire certaines statistiques.

Pour étudier la distribution de la variable étudiée au sein de l'échantillon à votre disposition, vous allez générer 3 graphiques sur la même figure :

- la **fonction de répartition empirique** (*empirical cumulative density fucntion*);
- l'**histogramme** normalisé de manière à ce que les hauteurs des rectangles soient déterminées de sorte que les surfaces des rectangles correspondent aux fréquences relatives, appelées normalisation en densité (*density*) - considérez un nombre de classes (*bins*) égal à 15;
- la **boîte à moustaches**.

Sur ces trois graphiques, représentez également la moyenne, l'écart-type, et les quartiles ($\hat{F}^{-1}(0.25)$, $\hat{F}^{-1}(0.5)$ et $\hat{F}^{-1}(0.75)$). Les **valeurs numériques** de ces statistiques doivent également apparaître.

Conseil : Utilisez `subplots` pour créer des sous-graphes dans une seule figure.

[ ]:
```
"""VOTRE CODE ICI"""
```

Y a-t-il des données aberrantes ?

VOTRE TEXTE ICI

1.2 Estimation de densité

L'histogramme que vous avez représenté peut vous aider à postuler un modèle pour la distribution de la variable étudiée. Cependant, il est nécessaire de fixer le nombre de bins, et le graphique obtenu peut être difficile à interpréter, la visualisation dépendant du choix du nombre de bins.

La construction d'un tel histogramme peut être vue comme un empilement de blocs, chaque bloc correspondant à un point (*i.e.* une observation). Pour chaque observation se trouvant dans la plage de valeurs d'une bin, un bloc est ajouté pour cette bin. Une autre manière de construire une estimation de la densité consiste à ajouter un bloc centré en la valeur de l'observation à laquelle il correspond. Pour chaque observation, on place un bloc, non pas dans une bin arbitraire, mais à la position de l'observation sur l'axe des observations. Ensuite, on somme la hauteur des blocs en chaque point de cet axe.

Derrière cette intuition se cache la méthode de l'estimation de la densité basée sur un noyau (*Kernel Density Estimation*, KDE, en anglais). Le noyau représente le bloc que l'on place pour chaque observation. Il peut prendre beaucoup de formes, et ne se limite donc pas à un "bloc" rectangulaire.

Cette méthode est par exemple implémentée dans la librairie scikit-learn.

À l'aide de cette librairie, générez une figure superposant l'histogramme généré précédemment, et les densités estimées avec un noyau gaussien, pour les valeurs du paramètre "bandwidth" (bw) suivantes: $bw \in \{0.1, 0.3, 0.8, 1.2\}$. Faites également apparaître les données d'échantillon comme proposé dans l'exemple de l'url ci-dessus. Vous pouvez laisser les autres paramètres par défaut de la méthode de la librairie.

[ ]:
```
"""VOTRE CODE ICI"""
```

Comparez la représentation obtenue avec une KDE et un histogramme.

Que représente le paramètre "bandwidth" de la fonction utilisée ? Comment change-t-il la représentation générée ci-dessus ?

VOTRE TEXTE ICI

Afin d'étudier plus en détails les propriétés du paramètre "bandwidth", générez une nouvelle figure en suivant la méthode suivante.

Pour chaque bandwidth $bw \in \{0.1, 0.3, 0.5, 1.2\}$:

- générez 25 sous-échantillons de taille $n = 25$ de l'échantillon de départ;
- estimez, à l'aide d'une KDE et d'un kernel gaussien, pour la bw considérée, la densité pour chaque sous-échantillon;
- tracez la moyenne des densités estimées, et mettez en évidence la zone comprise entre cette moyenne et $\pm$ la variance de ces densités (au moyen par exemple de la fonction `fill_between` de la librairie `matplotlib`).

[ ]:
```
"""VOTRE CODE ICI"""
```

Comparez les résultats obtenus pour les différentes bandwidth. Quels critères sont modifiés par le choix de ce paramètre, qui pourraient entrer en compte dans le choix de la valeur à lui donner ? Le choix de cette valeur est-il évident ou résulte-t-il d'un compromis ? Justifiez.

VOTRE TEXTE ICI

---

2 Modèle et estimation des paramètres

Dans cette deuxième partie du devoir, vous allez devoir construire différents estimateurs ponctuels pour le paramètre d'un modèle statistique de vos données. Pour cela, vous utiliserez comme modèle statistique la distribution Exponentielle de paramètre λ, décalée de 6.5 unités vers la droite, dont la PDF est donnée par

$$f_Y(y;\lambda) = \begin{cases} \lambda e^{-\lambda(y-6.5)}, & \text{si } y \geq 6.5, \\ 0, & \text{si } y < 6.5, \end{cases}$$

avec $\lambda > 0$.

Intéressons nous à plusieurs méthodes pour estimer la valeur du paramètre de la distribution.

Remarque importante:

Lorsque vous calculez des statistiques en utilisant une fonction implémentée dans une librairie Python (ou autre), vérifiez dans la documentation que cette fonction calcule bien la statistique souhaitée, pour la distribution considérée, telles que définies dans le cours théorique ou dans l'énoncé.

2.1 Estimation par la méthode des moments

Une première manière pour estimer les paramètres de la distribution Exponentielle est d'utiliser la méthode des moments. Pour rappel, cette méthode consiste à faire coincider le moment théorique et le moment d'échantillon d'ordre k :

$$E(Y^k) = \frac{1}{n}\sum_{j=1}^{n} Y_j^k.$$

2.1.1 Calcul de l'estimateur

Calculez la valeur de l'estimateur $\hat{\lambda}_{\text{MoM}}$ pour votre échantillon.

[ ]:
```
"""VOTRE CODE ICI"""
```

2.2 Estimation par la méthode du maximum de vraisemblance

Une seconde méthode d'estimation des paramètres est la méthode du maximum de vraisemblance. Pour rappel, cette méthode consiste à trouver la valeur des paramètres maximisant la fonction de vraisemblance (ou, de manière équivalente, maximisant la log-vraisemblance) :

$$\hat{\lambda}_{\text{MLE}} = \arg\max_{\lambda} \ \log L(\lambda; \mathbf{y}),$$

où $\mathbf{y}$ représente les données observées.

Note : la notation log dénote dans la convention du cours le logarithme naturel.

2.2.1 Calcul des estimateurs

Calculez la valeur de l'estimateur $\hat{\lambda}_{\text{MLE}}$ pour votre échantillon.

```
[ ]: """VOTRE CODE ICI"""
```

2.3 Comparaison

Maintenant que les valeurs des estimateurs ont été déterminées, superposez les modèles estimés $f_Y(y; \hat{\lambda})$ par la méthode des moments et du maximum de vraisemblance à l'histogramme de vos données d'échantillon $\mathbf{y} = (y_1, \dots, y_n)$.

```
[ ]: """VOTRE CODE ICI"""
```

Que pouvez vous dire des modèles estimés par rapport à vos données ? Ces modèles semblent-ils coller aux données ?

Les deux méthodes donnent-elles des résultats similaires ? Justifiez.

VOTRE TEXTE ICI

---

3 Convergence des estimateurs

Cette dernière partie du devoir a pour but de vous faire explorer la convergence des estimateurs, c'est-à-dire leur évolution pour des tailles d'échantillon de plus en plus grandes. En particulier, il vous est demandé d'étudier les propriétés asymptotiques du MLE.

Afin d'étudier le biais de vos estimateurs, il est nécessaire d'utiliser une valeur "théorique" des paramètres de votre distribution. Nous allons donc, dans cette dernière partie du devoir, considérer une distribution théorique pour la variable étudiée, et générer des données simulées de cette distribution.

On fait ici l'hypothèse que la variable étudiée suit une distribution Exponentielle de paramètres $\lambda = 2.272$ (même si ce n'est (peut-être) pas le cas dans la réalité).

3.1 Biais et variance de l'estimateur

Afin d'étudier le concept de biais et d'estimer le biais de l'estimateur MLE, vous allez premièrement comparer les effets de certains paramètres sur la précision de cet estimateur.

Pour ce faire, vous allez générer les courbes suivantes, en tirant $m = 500$ échantillons différents :

- dans une première figure, l'évolution du **biais** (en valeur absolue) de $\hat{\lambda}_{\text{MLE}}$ en fonction de la taille d'échantillon n, pour $n \in \{2, 3, ..., 100\}$
- dans une seconde figure, l'évolution de la **variance** de $\hat{\lambda}_{\text{MLE}}$ en fonction de la taille d'échantillon n, pour $n \in \{2, 3, ..., 100\}$.

Au total, deux figures sont attendues, chacune contenant une courbe. Pensez à ajouter une légende ou un titre sur vos deux figures, afin d'éviter une confusion entre les courbes (en plus des labels d'axes, évidemment).

[ ]:
```
"""VOTRE CODE ICI"""
```

Qu'observez-vous par rapport à l'évolution

- du biais, et
- de la variance

de l'estimateur, lorsque la taille n des échantillons augmente. Pouvez-vous en donner une explication théorique ? En particulier, quelle propriété de l'estimateur du maximum de vraisemblance est ici observée ?

VOTRE TEXTE ICI

Quelle modification pouvez-vous apporter à l'estimateur $\hat{\lambda}_{\text{MLE}}$ afin d'obtenir un estimateur non-biaisé de λ ?

VOTRE TEXTE ICI

Cet estimateur est-il meilleur que $\hat{\lambda}_{\text{MLE}}$? Pour argumenter votre réponse, tracez tout d'abord l'évolution du **biais** et de la **variance** de ce nouvel estimateur en suivant la même marche à suivre qu'à la sous-question précédente. Superposez les courbes obtenues pour l'estimateur biaisé (générées pour les mêmes échantillons que l'estimateur non-biaisé) sur ces nouvelles figures. Veillez à choisir les dimensions de vos axes pour que la comparaison avec les graphes de la sous-question précédente soit pertinente.

[ ]:
```
"""VOTRE CODE ICI"""
```

Au vu des graphes obtenus, quel choix d'estimateur vous semble le plus pertinent? Justifiez.

VOTRE TEXTE ICI

Vous avez désormais étudié le biais de $\hat{\lambda}_{\text{MLE}}$ pour estimer λ.

La moyenne d'une distribution exponentielle étant égale à $1/\lambda$, une manière intuitive d'estimer cette moyenne est de calculer $1/\hat{\lambda}_{\text{MLE}}$.

Tracez à présent l'évolution du **biais** et de la **variance** de ce nouvel estimateur en suivant la même marche à suivre qu'aux sous-question précédentes.

[ ]:
```
"""VOTRE CODE ICI"""
```

Qu'observez-vous pour l'évolution du biais et de la variance pour cet estimateur de la moyenne ? Est-ce surprenant au vu de l'évolution du biais et de la variance de l'estimateur $\hat{\lambda}_{\text{MLE}}$ pour λ ?

VOTRE TEXTE ICI

3.2 Représentation graphique

Une autre manière de visualiser comment les estimateurs évoluent en fonction de la taille d'échantillon est de générer l'histogramme de ces estimateurs, autrement dit, de représenter leur distribution d'échantillon. On peut, dans ce cas, observer de manière plus graphique l'évolution du biais et de la variance des échantillons. Pour ce faire, pour chacun des trois estimateurs, il vous est demandé de superposer les histogrammes des estimateurs obtenus pour différentes tailles

d'échantillons $n \in \{5, 25, 100, 200\}$. Pour chaque histogramme, représentez la moyenne afin de la comparer avec la valeur à estimer (à représenter également).

Au total, ce sont trois graphiques qui sont attendus pour plus de lisibilité.

[ ]:
```
"""VOTRE CODE ICI"""
```

4 Que faut-il en retenir ?

Rassurez-vous, cette section ne contient aucun travail que vous devrez nous rendre.

Le but de celle-ci est de constituer une liste de questions que vous devriez être en mesure d'aborder à la fin de ce devoir. Ces questions ont pour but de vous aider à identifier si vous avez réellement compris ce que vous avez fait pendant ce devoir et pourquoi vous l'avez fait. Rien à rendre donc, simplement quelque chose que vous pourrez relire en guise d'introduction à votre étude future des statistiques (par exemple, avant l'examen).

- Comment peut-on représenter la distribution des données dans un échantillon ?
- Quelle est la différence entre la construction d'un histogramme et d'une KDE ?
- Qu'est ce qu'une donnée aberrante ?
- Pourquoi utilise-t-on des estimateurs ? Quelles sont les méthodes pour ce faire ?
- Pourquoi est-il possible d'étudier la variance et l'esperance d'un estimateur ?
- Qu'est ce que la convergence d'un estimateur et pourquoi est-elle souhaitable ?
- Quelles sont les propriétés asymptotiques du MLE ?
- Existe-t-il une seule manière d'estimer (de manière ponctuelle) une caractéristique de la population ? Quels sont les critères qui permettent de comparer des estimateurs ?

Bon travail !