Séance 1 : Traitement de données et ajustement de lois

Assistants : Blaise DHONT, Gauthier ROUSSEAU, Tomás TREWHELA, Daniel Vito PAPA ZANG, Zhenzhu MENG

1 Traitement d'une série de données

1. Utilisez les données de pluie à Davos.
   - Regardez le fichier de données.
   - Chargez les données.

On peut utiliser différentes fonctions pour lire un fichier, la plus universelle pour un fichier de données quelconque est [`fopen`] *suivi de* [`textscan`] *qui permet, entre autres, de sélectionner les données voulues et de sauter l'en-tête (ce qui est le cas ici), on peut aussi rapidement supprimer la première ligne et utiliser.* [`load`] *mais c'est un cas particulier*

   - La 7^e colonne vous donne les précipitations $pluie_i$ journalières.
   - Visualisez ces données en utilisant des points.
   - Faites un histogramme et déterminez la fonction de répartition. [`hist,ecdf`]
   - Calculez la densité de probabilité.
     - Fractionnez le domaine des valeurs en 30 classes de même taille $\mathrm{d}x$. Calculez le milieu de ces classes. [`max`] [`length`]
     - Comptez le nombre N_i d'événements par classe et calculez-en la probabilité P. [`hist`]
     - La densité de probabilité se calcule maintenant par

$$\rho = \frac{\mathrm{d}P}{\mathrm{d}x} = \frac{N_i}{\mathrm{d}x \times \sum N_i}$$

   - Déterminez la fréquence et le temps de retour pour une précipitation de plus de 30 mm, 50 mm et de 80 mm.
   - (si vous avez du temps) Calculez les moyennes par année. [`mean,find,==`] On fera en sorte de ne prendre en compte que les années complètes.

En hydrologie on devrait parler d'année hydrologique c'est à dire du 1er Octobre au 30 Septembre, mais par souci de temps de code nous utiliserons, ici, l'année calendaire (les plus à l'aise peuvent trouver un code pour sélectionner les jours de l'année hydrologique)

2. En utilisant les données de chutes de neige à Chamonix, déterminez :
   - les chutes de neige annuelles ; [`sum`]
   - la fréquence et le temps de retour pour une chute de plus de 30 cm. Attention à la définition donnée au temps de retour, il n'y a pas 365 jours de données par an.
   - faites un diagramme avec le temps de retour T fonction de la chute de neige ; [`ecdf`]
3. Pour les chutes journalières de pluie et de neige ainsi que pour leurs cumuls annuels, déterminez le caractère stationnaire des séries de données.
   - Fixez un seuil S et comptez le nombre $N(i)$ d'événements dépassant ce seuil parmi les i premières valeurs de votre série temporelle. [`find,cumsum`]

- Tracez cette fonction $N(i)$ et comparez-la avec une ligne droite.
- Qu'en concluez-vous sur le caractère stationnaire de ces données ?

2 Lois de probabilités et ajustements

2.1 Ajustement d'une loi

L'ajustement le plus simple est de chercher la loi ayant les mêmes moments que les données. Les moments sont:

$$
\begin{array}{lrl}
\text{Moyenne (1}^{\text{er}}\text{ moment)} & \mu = \int X dP & = \frac{1}{N}\sum x_i \\
\text{Variance (2}^{\text{e}}\text{ moment)} & \sigma^2 = \int (X-\mu)^2 dP & = \frac{1}{N}\sum (x_i - \mu)^2 \\
\vdots & \vdots & \vdots
\end{array}
$$

Le nombre de paramètres de la loi donne le nombre de moments à égaliser. Trois séries de données (*serie01.txt, serie02.txt et serie03.txt*) ont été obtenues lors d'expériences et on cherche à les caractériser à l'aide d'une loi de probabilités (loi uniforme, loi exponentielle, loi gaussienne, loi de Cauchy, loi gamma, etc.). (Voir Annexe 1 du cours pour une description de ces lois)

- Tracez premièrement l'histogramme normalisé de chacun de ces tirages. Leurs allures vous paraissent-elles familières ?
- Calculez leurs premiers moments.
- Comparez les densités de probabilité théoriques aux histogrammes de données.
- Pour les données *serie03.txt*, tracez un diagramme de quantile avec la loi trouvée précédemment et comparez le avec la bissectrice. La loi trouvée vous parait-elle adaptée? `[qqplot]`

2.2 Loi binômiale et loi de Poisson

La loi binômiale (notée **B**) vient de la somme de m variables aléatoires de type Bernoulli de paramètre p. Pour un nombre de succès petit, celle-ci peut être approchée par une loi de Poisson (notée **P**), plus simple dans la pratique. Mais sous quelles conditions ?

1. p fixé.
   - Tracez en fonction de k la probabilité $P(X = k)$ avec $X \simeq \mathbf{B}(100\ ;\ 0{,}1)$ pour $k = 1 \cdots 30$
   - Tracez en fonction de k la probabilité $P(X = k)$ avec $X \simeq \mathbf{P}(100 \cdot 0{,}1)$ pour $k = 1 \cdots 30$
2. k fixé.
   - Tracez en fonction de p la probabilité $P(X = 5)$ avec $X \simeq \mathbf{B}(100\ ;\ p)$ pour $p = 0{,}01 \cdots 0{,}2$
   - Tracez (sur la même figure) en fonction de p la probabilité $P(X = 5)$ avec $X \simeq \mathbf{P}(100 \cdot p)$ pour $p = 0{,}01 \cdots 0{,}2$
3. Condition sur p pour $k \in [0\ ;\ 20]$.
   - Calculez le maximum de l'erreur relative sur $P(X = k)$ entre la loi binômiale et la loi de Poisson pour $p = 0{,}01 \cdots 0{,}2$. On tracera le graphique de l'erreur en fonction de p.
   - Trouvez la condition sur p pour que l'erreur absolue entre les deux lois soit de moins de 10 % quel que soit k?
4. *Question facultative :* les paramètres k et p ont tous les deux une influence sur la valeur de l'erreur commise entre ces deux lois. Tracez un graphique 3D du logarithme de l'erreur en fonction de p et de k pour $p \in [0{,}001\ ;\ 0{,}5]$ et $k \in [1\ ;\ 50]$. Concluez. `[meshgrid,mesh]`

3 Le paradoxe de Petersbourg (facultatif)

Un joueur va au casino pour faire fortune. Grace à son cours de probabilité, il a développé une stratégie qu'il pense infaillible : il commence par miser la mise minimale sur une couleur (rouge ou noire). Quand il perd il rejoue en doublant son enjeu. S'il perd à nouveau il double une nouvelle fois son enjeu, etc. Une fois qu'il a gagné, il recommence avec la mise minimale.
Cette stratégie est-elle payante ?
Plus prudent, vous lui proposez de simuler cette stratégie avant de l'appliquer :

- La fortune au début est de 100 fois la mise minimale, l'enjeu de départ est alors 1. Tant que sa fortune est supérieure à l'enjeu que lui impose sa stratégie, il joue.
  - Il perd avec une probabilité de $P = 0{,}5$; la mise est doublée, sa fortune diminue de la mise perdue;
  - Il gagne avec une probabilité de $P = 0{,}5$: la mise retombe à 1 et sa fortune augmente de la mise ;
- Comme le joueur n'est pas trop téméraire, il s'arrête de jouer lorsqu'il a doublé sa fortune initiale.

1. Simulez cette stratégie. Faites un diagramme de sa fortune.
2. Enregistrez sa fortune finale et calculez en la moyenne.

Explication du paradoxe :
Pour mettre en évidence l'aspect paradoxal de ce problème, il faut considérer que, quelle que soit la mise initiale, l'espérance mathématique de gain est positive, et même infinie, pour le joueur. Pourtant, tout quidam sain d'esprit refusera de jouer à un tel jeu si la mise initiale est trop élevée. Ce comportement d'apparence irrationnelle s'appelle l'aversion au risque. Il a été formalisé par la notion de fonction d'utilité et a donné naissance à la théorie de la décision.
Une solution simple à ce paradoxe consiste à faire l'hypothèse réaliste que la banque n'est pas infiniment riche, et qu'elle va donc cesser de payer au-delà d'une certaine somme. Par exemple, si on suppose qu'elle ne dispose « que » de 4 000 000 Fr soit 2^{22} Fr, le jeu va cesser au 22e coup et la mise équitable sera alors de 22 Fr.