Liste (non exhaustive) d'outils mobilisables pour les projets

1. Estimation paramétrique d'une série d'observations

Supposons que l'on possède une série de nombres $x_1, \dots, x_n$ dont il est raisonnable de présumer qu'il s'agit d'observations i.i.d. d'un certain phénomène, par exemple :

- le taux d'hémoglobine mesuré chez chaque individu d'une population de n patients ;
- les émissions de CO_2 de n exemplaires d'un certain modèle de voiture ;
- la température moyenne au mois d'août à Paris sur les n dernières années ;
- le taux de croissance de n actifs boursiers sur une certaine durée.

On peut vouloir représenter $x_1, \dots, x_n$ comme les réalisations indépendantes d'une variable aléatoire X de loi inconnue dans un ensemble paramétrique $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. On motivera alors le choix du *modèle* $\mathcal{P}$, puis l'on présentera un *estimateur* de θ, que l'on pourra utiliser pour construire des *intervalles de confiance* ou, lorsque cela se justifie, des *tests* : par exemple, la température moyenne au mois d'août à Paris sur les 15 dernières années est-elle distribuée de la même manière que sur le 20ème siècle ?

Une fois un estimateur $\hat{\theta}$ calculé, on peut vérifier la qualité du modèle en représentant l'histogramme des observations auquel on superpose la densité de $P_{\hat{\theta}}$. De manière plus quantitative, l'adéquation des données au modèle peut être vérifiée au moyen d'un test dans l'esprit de celui de *Kolmogorov*. On notera que la loi de la statistique de test sous l'hypothèse nulle n'est pas nécessairement connue (penser au test d'adéquation au modèle exponentiel vu en cours), et qu'il peut être utile d'estimer ses quantiles par simulation numérique.

Si le test d'adéquation des données au modèle ne se révèle pas concluant, on peut évidemment réessayer avec un autre modèle (la version française de Wikipédia donne une liste très complète de familles paramétriques de mesures de probabilité, en expliquant souvent leur intérêt dans la modélisation) : une telle étude comparative de différents modèles ne peut que valoriser un projet !

2. Étude de l'influence d'un facteur sur une variable

On suppose dans cette partie que l'on observe une série $(x_1, y_1), \dots, (x_n, y_n)$ et que l'on cherche à expliquer la variable y en fonction du facteur x.

2.1. Influence d'un facteur quantitatif sur une variable quantitative.

On s'intéresse d'abord au cas où x et y sont des variables quantitatives, par exemple :

- y_i est la consommation d'électricité sur une journée dans la ville i, et x_i est la température du jour dans cette ville ;
- y_i est le logarithme du prix de l'actif i aujourd'hui, et x_i est le logarithme du prix de cet actif hier.

La *régression linéaire* (possiblement *multiple* si les x_i sont des vecteurs) permet de vérifier si x et y sont liés par une relation linéaire. L'influence de x sur y est déterminée par le *test d'utilité des régresseurs*, et la qualité du modèle linéaire est mesurée par le *coefficient de détermination*. L'*analyse des résidus* permet de se prémunir contre la présence d'observations atypiques ou suspectes.

2.2. Influence d'un facteur quantitatif sur une variable booléenne.

On suppose maintenant que les variables y_i sont booléennes, c'est-à-dire qu'elles ne peuvent prendre que deux valeurs : par exemple, y_i encode le fait que le patient numéro i est atteint d'une maladie donnée, tandis que x_i représente le taux d'anticorps dans le sang de ce patient. La *régression logistique* permet alors d'estimer la loi de y en fonction de la valeur prise par x. L'influence du facteur x sur la variable y est déterminée par le *test du rapport de vraisemblance*.

2.3. **Influence d'un facteur qualitatif sur une variable quantitative.** On traite ici l'exemple suivant : sur un échantillon de n étudiants à l'École des Ponts, x_i est la variable booléenne encodant le fait que l'élève i soit inscrit (et assidu) au cours de Statistiques et Analyse des Données, et y_i représente le salaire en sortie d'école de cet élève. On peut alors employer les méthodes suivantes :

- l'*analyse de la variance à un facteur* permet de tester en particulier s'il y a un effet « cours de statistiques » dans le salaire en sortie d'école — on prendra néanmoins garde au fait que cette méthode se place dans le cadre gaussien ;
- les tests de *Kolmogorov-Smirnov* et *Mann-Whitney* permettent de vérifier si les deux séries $\{y_i : x_i = 0\}$ et $\{y_i : x_i = 1\}$ sont distribuées sous la même loi ;
- le *test d'indépendance du* χ_2 permet de vérifier si x et y sont indépendants — comme la variable y est a priori continue, il est nécessaire de partitionner l'ensemble des salaires possibles en un nombre raisonnable de tranches.

2.4. **Influence d'un facteur qualitatif sur une variable booléenne.** On considère enfin le cas où x_i est une variable qualitative — par exemple, le genre d'un individu, et y_i une variable booléenne : cet individu est-il au chômage ? Le *test de comparaison des proportions* pour les échantillons *non appariés* permet alors de déterminer si la proportion de chômeurs parmi les hommes est égale à la proportion de chômeurs parmi les femmes.

3. Représentation de données en grande dimension

Les méthodes d'analyse exploratoire permettent de représenter graphiquement des données comportant beaucoup de variables. L'*analyse en composante principale* permet par exemple de projeter un nuage de points en grande dimension sur le *premier plan factoriel*, qui est celui qui « déforme le moins » le nuage. Le *cercle des corrélations* permet alors de visualiser les composantes principales en fonction des variables originales. Enfin, les méthodes de *classification automatique* permettent de discerner des groupes homogènes au sein d'un nuage de points. Remarquons que l'on peut combiner le résultat d'une classification aux méthodes décrites ci-dessus pour estimer la distribution d'une certaine variable à l'intérieur de chacun des groupes obtenus par la classification, ou pour tester si l'appartenance d'un individu à un de ces groupes a une influence sur la distribution d'une variable.