PROJET DE FIN DE SEMESTRE

Ce projet d'une durée de trois semaines va vous permettre de mettre en application vos connaissances étudiées dans le cours Statistiques et R durant ce semestre. Il se décompose en deux parties :

- **Exercices théoriques** : mise en pratique des notions d'inférence statistique - Bayes, test d'hypothèses. Introduction et question sur la régression linéaire simple.
- **Programmation** : étude statistique d'un ensemble de données dans le but d'établir des corrélations.

1 Exercices théoriques

1.1 Exercice 1 - Inférence Bayesienne

.

Une boîte contient deux dés à quatre faces, un dé à huit faces, et un dé à douze faces. Vous tirez un dé au hasard parmi ceux dans la boite, le lancez et obtenez un cinq.

(a) Faire une table de Bayes affichant les probabilités antérieures, vraisemblances, numérateurs de Bayes et probabilités postérieures pour les hypothèses sachant le résultat obtenu.

(b) Sachant le résultat du premier lancer, quelle est la probabilité que le résultat du prochain lancer soit sept?

1.2 Exercice 2 - Inférence fréquentiste - test d'hypothèses

.

Les données sont tirées d'une distribution binomial(5,θ) où θ est inconnue. Voici le tableau des probabilités $p(x|\theta)$ pour trois valeurs de θ.

x	0	1	2	3	4	5
$\theta = 0.5$	0.031	0.156	0.313	0.313	0.156	0.031
$\theta = 0.6$	0.010	0.077	0.230	0.346	0.259	0.078
$\theta = 0.8$	0.000	0.006	0.051	0.205	0.410	0.328

Vous voulez lancer un test de signification sur la valeur de θ. Vous avez les informations suivantes :

- Hypothèse nulle : $\theta = 0.5$
- Hypothèses alternatives : $\theta > 0.5$
- Niveau de signification : $\alpha = 0.1$

(a) Trouver la région de rejet.

(b) Calculer la puissance du test pour les deux hypothèses alternatives $\theta = 0.6$ et $\theta = 0.8$.

(c) Supposons que vous lancez l'expérience et la donnée est $x = 4$. Calculer la p-valeur de cette donnée.

1.3 Exercice 3 - Régression linéaire

Supposons des données bivariées (couple de données) $(x_1, y_1), ..., (x_n, y_n)$, $i = 1, ..., n$. Le but de la régression linéaire est de modéliser la relation entre x et y en trouvant une fonction $y = f(x)$ qui approche au mieux les données. La régression linéaire simple pour deux variables aléatoires X, Y cherche la ligne

$$y = ax + b$$

qui s'adapte au mieux aux données. Notre modèle dit que chaque y_i est prédit par x_i avec une certaine erreur ϵ_i :

$$y_i = ax_i + b + \epsilon_i$$

Donc

$$\epsilon_i = y_i - ax_i - b$$

La méthode des moindres carrés calcule les valeurs $\hat{a}$ et $\hat{b}$ de a et b qui minimise la somme des carrés des erreurs :

$$S(a, b) = \sum_i \epsilon_i^2 = \sum_i (y_i - ax_i - b)^2$$

Question : Prouver par le calcul que

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

où

$$\bar{x} = \frac{1}{n}\sum_i x_i, \;\; \bar{y} = \frac{1}{n}\sum_i y_i, \;\; s_{xx} = \frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2, \;\; s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Ici $\bar{x}$ est la moyenne échantillonnale de x, $\bar{y}$ la moyenne échantillonnale de y, s_{xx} est la variance échantillonnale de x et s_{xy} est la covariance échantillonnale de x et y.

2 Programmation - étude statistique

Nous allons appliquer ici les notions de régression linéaire à un ensemble de données réel. La base de données utilisée ici, The World FactBook de 2015, est produite par la CIA annuellement et contient des informations sur chaque pays du monde sur des aspects géographique, démographique, politique, économique, communication ou militaire. L'ensemble de données fourni comprend 279 pays (et entités) et 75 variables. Cependant cette base n'est pas pleinement construite.

2.1 Manipulation préalable

Plusieurs documents sont donnés pour que vous puissiez créer la base de données "World_FactBook.csv".

- "codes.csv" : la base de données contenant les noms des pays concernés (les autres codes doivent être négligés).
- "categories.csv" : est la base de données contenant les références des différentes variables, leur catégorie, le nom de la variable, ainsi que leurs unités (et une traduction).
- le dossier "data" : contient les données (dans des fichiers sous la forme "cxxxx.csv") associées aux références des variables contenues dans "categories.csv".

La première partie du projet de programmation est d'écrire un code R, que vous nommerez "Manipulation.r", permettant de re-créer la base "World_FactBook.csv" qui vous est donnée regroupant toutes ces informations, grâce aux documents expliqués au-dessus.

Remarque

- Attention, les données contenues dans le dossier "data" (les "cxxxx.csv") ne sont parfois pas définies pour certains pays. Dans ce cas, renseignez `NA` (pour Not Available ou missing value) dans le champ correspondant.
- Les variables ayant comme unité le dollar ont le symbole "$" dans leur table de valeurs. Il ne faut pas que ce symbole apparaisse dans la base "World_FactBook.csv".
- De plus, certaines valeurs dont l'unité est le $ sont négatives et contiennent des virgules (correspondant à mille, un milion, etc en anglais : 1,000 et 1,000,000) comme dans la variable "Current account balance" data n°2187. Nous voulons par exemple que "-$144,500,000 " devienne -144500000.

2.2 Etude

Ici, le but est d'effectuer une étude observationnelle et établir des corrélations via la régression linéaire sur l'ensemble de données "World_FactBook.csv". Vous devrez coder par vous même la construction de paramètres d'une droite de régression (grâce aux formules de la partie théorique) et savoir interpréter les résultats pour un jeu de données choisi, extrait de la base The World FactBook. Ces résultats pourront ensuite être validés à l'aide des commandes existantes dans les packages R de type `statsr`.

L'étude devra être produite sur R en utilisant un document **R Markdown** comme vu en TD.

Bonus (facultatif) : effectuer un test d'hypothèse de votre choix (pour une statistique bien réfléchie) sur la base de données The World FactBook. Voir le document "Inference3.rmd" du TD8.

3 Documents à rendre

Dans trois semaines, vous devrez rendre un dossier compressé (nommé "nom1_nom2_nom3", avec nom$i$ vos noms respectifs) contenant les documents suivants :

- "**Exercices_theoriques.pdf**" : le document contenant vos réponses aux exercices théoriques

- "**Manipulation.r**" : le fichier R créant "Wolrd_FactBook.csv" à partir des documents cités dans la sous-section 2.1
- "**Projet.rmd**" : votre étude statistique au format .rmd
- "**Projet.html**" : votre étude statistique exporté en html