Université de Thiès **Année Académique 2019-2020**
U.F.R S.E.S **Master 1 SDA**

Travail Pratique de SDD4142-Statistiques

Travail à faire en groupe de 3 personnes.
à rendre avant 20 avril 2020 à $23h59$
(remise sur github)

Question 1 ***Simulation de lois*** *(3 pts)*

1. Simuler un échantillon de taille 10000 suivant une loi binomiale $\mathcal{B}(30, 0.2)$. Tracer l'histogramme de l'échantillon obtenu.
2. Simuler un échantillon de taille 10000 suivant une loi normale $\mathcal{N}(3, .4)$. Tracer la fonction de densité de l'échantillon obtenu. Choisir un intervalle contenant 0 pour domaine de représentation.
3. Simuler un échantillon de taille 10000 suivant une loi gamma $\gamma(10, .5)$. Tracer la fonction de densité de l'échantillon obtenu. Choisir un intervalle contenant 0 pour domaine de représentation.

Question 2 ***Régression linéaire simple*** *(4 pts)*
Nous donnons les couples d'observations suivants :

x_i	18	7	14	31	21	5	11	16	26	29
y_i	55	17	36	85	62	18	33	41	63	87

1. La première étape est d'obtenir les données. Enregistrer-les dans un format adapté pour une lecture par la suite avec Python.
2. Représentez les y_i en fonction des x_i. A la vue de cette représentation, pouvons-nous soupçonner une liaison linéaire entre ces deux variables ?
3. Déterminer pour ces observations la droite des moindres carrés, c'est-à-dire donner les coefficients de la droite des moindres carrés.
4. Donner les ordonnées des y_i calculés par la droite des moindres carrés correspondant aux différentes valeurs des x_i.
5. Tracer ensuite la droite sur le même graphique.
6. Quelle est une estimation plausible de Y à $x_i = 21$?
7. Quel est l'écart entre la valeur observée de Y à $x_i = 21$ et la valeur estimée avec la droite des moindres carrés ? Comment appelons-nous cet écart ?
8. Est-ce que la droite des moindres carrés obtenue en 2. passe par le point $(\bar{x}, \bar{y})$? Pouvons-nous généraliser cette conclusion à n'importe laquelle droite de régression ?

Question 3 ***Données réelles*** *(10 pts)*

Les données (voir fichier smp.csv) de cet exercice proviennent d'une étude de santé mentale en prison. Cette étude a été réalisée entre 2003 et 2004 dans les prisons françaises. Pour la description des variables voir fichier `presentation_donnees.pdf`.

1. Enregistrer les données dans un format adapté pour une lecture par la suite avec Python sachant que la première ligne du fichier smp.csv correspond au noms des variables. Vérifier si vous avez une structure de 799 observations et 26 variables.
2. Changer les types des variables. Vous devez obtenir le résultat suivant :

```
age              float64
prof            category
duree           category
discip          category
n.enfant         float64
n.fratrie          int64
ecole           category
separation      category
juge.enfant     category
place           category
abus            category
grav.cons       category
dep.cons        category
ago.cons        category
ptsd.cons       category
alc.cons        category
subst.cons      category
scz.cons        category
char            category
rs              category
ed              category
dr              category
suicide.s        float64
suicide.hr      category
suicide.past    category
dur.interv       float64
```

3. Calculer la moyenne, la variance, et l'écart type pour chacune des variables suivantes : age,n.enfant, n.fratrie, dur.interv. Donner les 3 premiers quantiles pour la variable age.
4. Tracer le boxplot pour la variable age. Quelles conclusions en tirez-vous ?
5. Afficher les données pour les agriculteurs qui ont plus de 2 enfants.
6. Calculer les fréquences des modalités de la variable prof. Quelle est la catégorie modale ?
7. Tracer le diagramme circulaire de la variable profession
8. Donner les moyennes des âges par profession

9. Donner la table des effectifs pour les variables prof incluant les "NaN".
10. Donner le nombre de "Nan" pour chaque variable.
11. Supprimer toutes les lignes contenant des "Nan".
12. Tracer l'histogramme et la densité de la variable age sur la même figure.
13. Discrétisez la variable age. Pour ce faire on ajoutera une variable dans le DataFrame des données une nouvelle variable nommée `age_classe`. Cette variable aura 4 classes :

$$[\min(age), Q1],\ \ ]Q1, Q2],\ \ ]Q2, Q3],\ \ ]Q3, \max(age)].$$

ou Q1, Q2, Q3 sont respectivement les 3 premiers quantiles de la variable age, $\min(age)$ et $\max(age)$ respectivement la plus petite et la plus grande valeur de la variable age.

14. Donner les fréquences des modalités de la nouvelle variable `age_classe`.

Question 4 ***Méthode de Monte Carlo*** *(3 pts)*
Soit $I_2 = \int_0^1 \sqrt{1-x^2}\ dx$

Estimer I_2 par une méthode de Monte Carlo avec $n = 10000$

Observer par graphique l'évolution de cette estimation lorsque n varie et vérifier la cohérence avec la valeur théorique $I_2 = \frac{\pi}{4}$.

Remise

A Remettre :
Un fichier .py (pour ceux qui travaillent sur Spyder) ou .ipynb (pour ceux qui travaillent sur Jupyter Notebook) par question. Les noms des membres du groupe seront écris dans l'entête de chaque fichier
Un fichier .pdf contenant les réponses aux questions
Plagiat interdit : Ne recopiez pas les codes trouvés sur internet même si vous pouvez vous en inspirez.

BON TRAVAIL !