



Faculté

des **sciences économiques** et de **gestion**

Université de Strasbourg

Année universitaire

2017/2018

Année universitaire 2021/2022

## Master 1

*Data Science pour l'économie et l'entreprise du futur*

*Statistique et Économétrie*

Semestre 2 / Contrôle continu

Mars 2022

Matière : Techniques de programmation

Enseignant : Pierre Pelletier

Durée : 2H00

Tout document autorisé

Calculatrice autorisée

---

### Sujet :

#### Exercice 1 (2pts)

```
names = c()
for(i in 1:1000){
  nb_letters = 7
  vowels = letters[c(1,5,9,15,21,25)]
  consonants = letters[-c(1,5,9,15,21,25)]
  is_vowel = sample(c(T,F),1)
  name = c()
  while(length(name) <= nb_letters){
    if(is_vowel){
      name = c(name,sample(vowels,1))
      is_vowel = F
    } else {
      nb_c = sample(c(1,2),1)
      name = c(name,sample(consonants,nb_c))
      is_vowel = T
    }
  }
  name = paste(name,collapse = '')
  names = c(names,name)
}
names
```

1) Expliquez ligne par ligne le code ci-dessus.

## **Exercice 2 (4pts)**

- 1) Créez un vecteur 'X' allant de 1 à 1000 inclus.**
- 2) Ecrivez une boucle pour ajouter un nombre aléatoire (tiré d'une loi normale de moyenne 0 et de variance 1) à chacun des éléments du vecteur 'X' .**
- 3) Réalisez la même opération sans réaliser de boucle. Est-il toujours préférable de vectoriser une opération ? Argumentez.**
- 4) Créez à la main le vecteur suivant (2,3,4,5,0,2,3,4,5). Ecrivez une fonction permettant de trouver le seul élément unique (ici le 0).**

## **Exercice 3 (3pts)**

- 1) Ouvrez le fichier 'phone\_numbers.csv', dans la variable 'phone\_number'. Combien de numéros contiennent la suite de chiffres '27' ?**
- 2) Créez une nouvelle variable binaire nommée 'first\_number', notez 1 si le numéro commence par '337', 0 sinon. Trouvez deux manières de répondre à la question.**
- 3) Combien de numéros de téléphone ont pour 4ème chiffre un chiffre pair ?**

## **Exercice 5 (5pts)**

- 1) Créez une fonction avec deux arguments, 'date' et 'month'. Cette fonction renvoie 1 si la date de naissance est comprise entre le 21/05 et le 21/06 inclus, elle renvoie 0 sinon. (les dates sont comprises entre 1 et 31, les mois entre 1 et 12)**
- 2) En utilisant la fonction de la question 1, ouvrez le fichier 'birth\_dates.csv', combien de personnes ont une date d'anniversaire située entre le 21 mai et le 21 juin?**
- 3) Ecrivez une fonction pour générer des dates de naissances entre 1950 et 2020 aléatoirement (utilisez une loi uniforme). (Attention aux années bissextiles → Exercice 3 - Chapitre 1)**

## **Exercice 6 (5pts)**

- 1) Joignez 'birth\_dates.csv' à la base de données 'phone\_numbers.csv' via la variable 'ID'. Gardez uniquement les observations présentes dans les deux bases de données.**

**2) Groupez les individus par année de naissance, quel est le nombre de personnes présentes dans chacun des groupes ? Représentez également graphiquement cette information avec un histogramme.**

**3) Générez une adresse mail pour chaque individu, cette adresse mail est composée de la première lettre du prénom, du nom et des deux derniers chiffres de l'année de naissance.**

**Ex : Si le nom est 'Xxxxx', le prénom 'Yyy' et l'année de naissance est '1999' alors l'adresse mail est : y.xxxx\_99@etu.unistra.fr**

### **Exercice 7 (7pts)**

**1) Créez deux vecteurs de 1000 nombres aléatoires tirés du loi normale, 'X' (moyenne 0, variance 1) et 'Y' (moyenne 10, variance 1).**

**2) Faites une régression linéaire et récupérez la P-value pour le coefficient de 'X'.**

**3) Faites une fonction réalisant les instructions suivantes :**

- génération d'un nouveau vecteur X de 1000 nombre aléatoire**
- régression linéaire du nouveau X sur Y**
- renvoi de la P-value du coefficient X**

**4) Faites une boucle pour lancer 500 fois la fonction créée précédemment, stockez les résultats.**

**5) Représentez graphiquement la distribution des p-values.**

**6) Quelle est la proportion de P-value < 0.05 ?**

**7) Est-il possible d'accélérer la procédure en parallélisant l'opération ? Argumentez.**