

# Statistique : devoir, décembre 2020

Le sujet est composé de **5 exercices indépendants**. Vous devrez répondre aux questions sur un document Markdown avec une sortie au format **html ou pdf**. Ce document devra afficher les codes R ainsi que les sorties qui permettent de répondre aux questions. A la fin de l'épreuve vous enverrez par email le **fichier de sortie compilé correctement au format html ou pdf** ainsi que le **fichier source au format Rmd** par email à laurent.rouviere@univ-rennes2.fr. La qualité du document markdown sera prise en compte dans le barème tout comme la structure et l'élégance des codes **R**.

On utilisera les packages suivants :

```
library(tidyverse)
theme_set(theme_classic(base_size=10))
library(lubridate)
```

## Exercice 1 (quelques calculs de probabilités)

1. On considère  $X$  une variable de loi binomiale  $B(10, 0.4)$ .
  - a) Calculer les probabilités (on donnera les résultats sans utiliser de fonctions **R** mais en justifiant brièvement).

$$P(X = -1), P(X \leq -1) \quad \text{et} \quad P(X \geq -1).$$

- b) Calculer les probabilités (on peut utiliser des fonctions **R** à partir de maintenant).

$$P(X = 1), P(X = 4) \quad \text{et} \quad P(X = 10).$$

- c) Calculer les probabilités

$$P(X \leq 3), P(X > 4), P(X > 3.5) \quad \text{et} \quad P(2 \leq X \leq 8).$$

2. On considère ici  $Y$  une variable de loi normale d'espérance 3 et de variance 1 (notée  $N(3, 1)$ ).
  - a) Calculer les probabilités

$$P(Y = 3) \quad \text{et} \quad P(Y = 0).$$

- b) Calculer les probabilités

$$P(Y \leq 2), P(Y < 2) \quad \text{et} \quad P(Y > 2).$$

- c) Calculer les probabilités

$$P(2 \leq Y \leq 4) \quad \text{et} \quad P(Y \leq 2 \text{ ou } Y \geq 3.5).$$

## Exercice 2 (intervalle de confiance)

On considère les données sur les iris de Fisher. Calculer un intervalle de confiance de niveau 95% pour les paramètres suivants :

- La longueur de Pétales moyenne
- La largeur de Sépales moyenne de l'espèce Setosa
- La longueur de Sépales moyenne pour les iris des espèces Setosa ou Virginica

### Exercice 3 (IC pour un sondage)

Au cours d'une élection avec deux candidats  $A$  et  $B$  on réalise un sondage pour estimer la proportion  $p$  inconnue d'électeurs qui vont voter pour  $A$ . On interroge 1004 personnes, sur ces 1004 personnes 478 déclarent qu'elles vont voter pour  $A$ . Donner un intervalle de confiance à 90% pour le paramètre  $p$ .

### Exercice 4 (Graphes ggplot)

On utilisera les fonctions du package **ggplot** pour faire les graphes demandés.

1. Tracer les densités gaussiennes des lois  $N(0, 1)$  et  $N(1, 1)$  sur un même graphe avec deux couleurs différentes (et une légende qui permet d'identifier les densités).
2. Même question mais sur deux graphes séparés (côte à côte).

### Exercice 5 (Données sur le covid)

Le jeu de données `data_covid_2020.csv` contient des informations sur les nombres de cas confirmés et de décès entre le 1er mars 2020 et le 21 novembre 2020 (quelques journées d'observations peuvent être manquantes pour certains pays). Il contient 7 colonnes :

- `date` : identifiant de la journée de l'observation
- `country` : pays
- `population` : la population totale du pays
- `confirmed` : nombre de cas positifs dans le pays `country` confirmés depuis le début de l'épidémie jusqu'à la date `date`
- `deaths` : nombre de décès dans le pays `country` depuis le début de l'épidémie jusqu'à la date `date`
- `deaths.day` : nombre de décès observé le jour `date` dans le pays `country`
- `confirmed.day` : nombre de cas positifs observés le jour `date` dans le pays `country`.

On souhaite calculer différents indicateurs et visualiser ces données. On utilisera les verbes **dplyr** et **ggplot** pour répondre aux questions suivantes.

1. Importer les données et afficher un résumé (fonction **summary**).
2. Convertir la première colonne du jeu de données (colonne **date**) en objet **date** à l'aide de la fonction **as\_date** du package **lubridate**.
3. Combien de pays sont représentés dans le tableau ? On pourra utiliser le verbe **n\_distinct**.
4. Afficher les pays présents dans l'étude.
5. Quel est le nombre de jours considéré dans ce tableau (nombre de jours entre la première observation et la dernière ?)
6. Comparer la distribution de la variable **confirmed** de chaque pays à l'aide d'un boxplot.
7. Tracer, pour chaque pays, les courbes représentant le nombre de morts par jours avec une couleur différente. Il s'agit de représenter la variable **deaths.day** sur l'axe des  $y$  en fonction de la **date** sur l'axe des  $x$ .
8. Proposer une version plus lisse de ces courbes à l'aide de la fonction **geom\_smooth**. On pourra utiliser l'option **span** de **geom\_smooth** avec des petites valeurs si les courbes sont trop lissées.
9. On s'intéresse uniquement à la France. Calculer le nombre de cas confirmés moyen par jour (moyenne de la variable **confirmed.day**) ainsi que le nombre moyen de morts par jour pour toute la période observée. Les résultats devront être présentés sur un tableau à 1 ligne et deux colonnes.
10. Même question pour chaque pays. Les résultats devront cette fois être présentés sur un tableau à **p** lignes et 3 colonnes (**p** représentant le nombre de pays).

11. Ordonner les pays en fonction du nombre total de morts au 21 novembre 2020. On affichera sur une colonne le pays et sur l'autre le nombre total de morts. On pourra créer la date du 21 novembre à l'aide de

```
t1 <- ymd("2020-11-21")
```

12. Ajouter aux données une variable `deaths.pop` égale au rapport du nombre de morts cumulé (`deaths`) divisé par la population (`population`).
13. Représenter, pour chaque pays, les courbes représentant la variable `deaths.pop` avec une couleur différente.
14. Ordonner les pays en fonction de la variable `deaths.pop` au 21 novembre 2020. On affichera sur une colonne le pays et sur l'autre la variable `deaths.pop`.
15. On s'intéresse uniquement à la France. Calculer le nombre de morts par mois. On pourra créer une variable qui permet d'identifier le mois d'une observation à l'aide du verbe **separate**
16. Même question pour tous les pays. On pourra visualiser les résultats dans un tableau à double entrée avec en ligne le pays et en colonne le mois de l'étude.
17. **Question ouverte** : proposer des indicateurs numériques ou graphiques qui permettent de visualiser et/ou comparer les deux vagues dans les pays européens.