Dénombrements de colonies d'*E. coli*

- Deux dilutions : D5 de 10^{-5} et D6 de 10^{-6} ,
- Quatre groupes Ga, Gb, Gc et Gd de travaux pratiques.

Préambule

Objectif : utiliser les outils vus sous R Studio pour l'étude de données récoltées en TP de Microbiologie dans le cadre du dénombrement de colonies d'Escherichia coli.

1 Travail attendu

Un compte-rendu, au format.pdf, de 10 pages maximum est à envoyer, par binôme, à votre enseignant-e de TD par mail, au plus tard le 06 janvier 2023.

Les questions et le code ci-dessous ne sont que des suggestions sur lesquelles baser votre étude statistique, qui doit pouvoir se lire sans avoir connaissance de ce document. Il s'agit ici d'exploiter et d'interpréter les données obtenues en TP de Microbiologie de 2020-2021.

2 Protocole expérimental

L'expérimentation se résume ainsi : chaque étudiant doit réaliser trois boîtes pour chacun des deux modes d'étalement (avec un râteau ou avec des billes) de trois dilutions différentes (facteurs 10^{-4} , 10^{-5} , et 10^{-6}) d'une culture spécifique à son groupe de travaux pratiques de Microbiologie. Un problème de comptage (colonies trop denses) pour la dilution la plus concentrée rend les résultats inexploitable ; ils ne sont donc pas considérés.

Les données à analyser se présentent comme suit :

- Trois boîtes identifiées par les étiquettes B1, B2 et B3
- Deux modes d'étalement par Eb billes ou par Er râteau,

Chaque étudiant a ainsi préparé en principe $2 \times 3 \times 2 = 12$ boîtes sur lesquelles il a réalisé un comptage des unités formant colonie (UFC / mL) d'*E. coli*. Ce qui représente en tout $12 \times (\text{nombre d'étudiants})$ mesures de dénombrement des cultures.

3 Exploration statistique élémentaire

3.1 Lecture des données

Les données extraites ont été regroupées, uniformisées et anonymées. Elles sont disponibles sur la page Moodle du cours dans le fichier *microbio_2022.csv*. Ce fichier est en format ".csv" avec des "," comme séparateurs. Les données sont rangées en colonnes avec une ligne par étudiant contenant le nom du groupe (A, B, C ou D), suivi de 12 valeurs : les trois boîtes de Eb x D06, les trois Er x D06, les trois boîtes Eb x D05, les trois Er x D05.

```
# Importation des données via un data frame
# à partir d'un fichier .csv
dbrutes=read.csv("microbio_2022.csv",
                 sep=";", header=TRUE)
# vérification de la bonne lecture
summary(dbrutes)
```

On notera la présence de valeurs manquantes codées "NA" : elles correspondent aux boîtes contaminées ou aux données non relevées.

3.2 Mise en forme des données

Un pré-traitement est nécessaire pour rendre accessibles les données à toute méthode statistique; on l'effectue à l'aide du code suivant.

```
# nombre d'étudiants par groupe
nbtud=table(dbrutes$Groupe)
na=nbtud["A"];nb=nbtud["B"];nc=nbtud["C"];
nd=nbtud["D"]
# construction du "design expérimental"
# - facteur numéro de boîte à 3 niveaux
# - facteur étalement à 2 niveaux
# - facteur dilution à 2 niveaux
# - facteur groupe à 4 niveaux
design=expand.grid(
  NumBoite=factor(1:3, labels=c("B1", "B2", "B3")),
  ModEtal=factor(1:2, labels=c("Eb", "Er")),
  Dilution=factor(1:2, labels=c("D5", "D6")),
  Groupe=factor(c(rep(1,na), rep(2,nb), rep(3,nc),
    rep(4,nd)), labels=c("Ga", "Gb", "Gc", "Gd"))
# vectorisation des données de dénombrement
Nombre=as.vector(t(dbrutes[,2:13]))
# constitution du data frame
ecoli=data.frame(design,Nombre)
# vérifier la bonne cohérence des données
ecoli[1:10,];summary(ecoli)
```

Ce qui suit est une trame d'étude. Vous êtes libres de l'adapter comme vous l'entendez, l'objectif étant d'étudier les données statistiques obtenues, notamment l'impact du facteur de dilution, du mode d'étalement et les écarts entre les différents groupes de TD.

3.3 Exploration

La première approche descriptive des données permet de se faire une idée sur la qualité des mesures (distributions, valeurs atypiques, variance, valeurs manquantes...) et de décider de l'opportunité ou non d'une transformation de la variable Nombre.

C'est aussi l'opportunité de repérer, et de supprimer, d'éventuelles valeurs aberrantes. Pour supprimer une ligne (par exemple ici la 23) :

```
dbrutes=dbrutes[-23,]
```

Attention au décalage de ligne qui peut s'en suivre.

On commence par des diagrammes en boîte :

```
hist(ecoli$Nombre)
boxplot(ecoli$Nombre)
shapiro.test(ecoli$Nombre)
boxplot(Nombre~Dilution, data=ecoli)
boxplot(Nombre~Groupe, data=ecoli)
boxplot(Nombre~ModEtal, data=ecoli)
```

Si les diagrammes boîtes révèlent des dénombrements atypiques mais sans que l'erreur de manipulation soit flagrante, on les conserve.

Les données présentent un **problème évident d'échelle de mesure dû à la dilution**. Ceci est corrigé avant de refaire les graphiques :

```
ecoli$Nombre[ecoli$Dilution=="D6"]=
  ecoli$Nombre[ecoli$Dilution=="D6"]*10
hist(ecoli$Nombre)
boxplot(ecoli$Nombre)
shapiro.test(ecoli$Nombre)
boxplot(Nombre~Dilution, data=ecoli)
boxplot(Nombre~Groupe, data=ecoli)
boxplot(Nombre~ModEtal, data=ecoli)
```

La distribution présente des caractéristiques peu "gaussiennes"; on essaie de trouver une transformation appropriée Comparer les trois distributions :

```
hist(ecoli$Nombre)
hist(sqrt(ecoli$Nombre))
hist(log(ecoli$Nombre))
```

Si vous trouvez une transformation appropriée, vous pouvez l'effectuer :

```
#remplacer XXX par la transformation choisie
ecoli$Nombre=XXX (ecoli$Nombre)
# Dernières vérifications
boxplot (Nombre~Dilution, data=ecoli)
boxplot (Nombre~Groupe, data=ecoli)
boxplot (Nombre~ModEtal, data=ecoli)
boxplot (Nombre~ModEtal+Dilution+Groupe,
        data=ecoli)
```

4 Inférence statistique

4.1 Effet global d'un facteur

Attention, cette section utilise une approche élémentaire mais enchaîne de nombreux tests sans contrôle possible (tests multiples) du niveau de chacun (résultats globaux); les résultats sont à examiner avec prudence.

```
# Moyennes par groupe
aggregate (Nombre~Groupe, FUN=mean, data=ecoli)
# Variance par groupe
aggregate (Nombre~Groupe, FUN=var, data=ecoli)
# Comparaison des groupes
kruskal.test (Nombre~Groupe, data=ecoli)

# Moyennes par dilution
aggregate (Nombre~Dilution, FUN=mean, data=ecoli)
# Variances par dilution
aggregate (Nombre~Dilution, FUN=var, data=ecoli)
# Comparaison des dilutions
wilcox.test (Nombre~Dilution, data=ecoli)

# Moyennes par mode d'étalement
aggregate (Nombre~ModEtal, FUN=mean, data=ecoli)
# Variances par mode d'étalement
aggregate (Nombre~ModEtal, FUN=var, data=ecoli)
# Comparaison des modes d'étalement
wilcox.test (Nombre~ModEtal, data=ecoli)
```

On pourra se poser les questions suivantes :

- Faut-il faire plus (ou moins) confiance à un groupe plutôt qu'un autre ?
- Quelle dilution est-il préférable d'utiliser ?
- Quel mode d'étalement ?

4.2 Effets par groupe

Les cultures dénombrées par chaque groupe n'étant pas identiques, l'effet "groupe" est évident sur ces données. L'étude des effets des autres facteurs peut en être impactée. Il est proposé de faire une étude par groupe toujours dans l'idée d'utiliser des outils élémentaires.

```
ecoliA=ecoli[ecoli$Groupe=="Ga", ]
ecoliB=ecoli[ecoli$Groupe=="Gb", ]
ecoliC=ecoli[ecoli$Groupe=="Gc", ]
ecoliD=ecoli[ecoli$Groupe=="Gd", ]
```

Pour analyser les résultats d'un groupe, on peut utiliser les commandes suivantes, en justifiant les choix d'un test pour considérer l'effet ou non d'un des facteurs (mode d'étalement, dilution) sur le dénombrement. il s'agit encore d'une situation de tests multiples à examiner avec prudence.

```
# Normalité ?
shapiro.test (ecoliA[, "Nombre"])
# Quel est l'effectif ?
# Mode d'étalement
# test de Fisher d'égalité des variances
var.test (Nombre~ModEtal, data=ecoliA)
# test de Student ou de Welsh
t.test (Nombre~ModEtal, var.equal=T, data=ecoliA)
# test non paramétrique
wilcox.test (Nombre~ModEtal, data=ecoliA)
# Dilution
# test de Fisher d'égalité des variances
var.test (Nombre~Dilution, data=ecoliA)
# test de Student ou de Welsh
t.test (Nombre~Dilution, var.equal=F, data=ecoliA)
# test non paramétrique
wilcox.test (Nombre~Dilution, data=ecoliA)
```

Il est possible d'exécuter les mêmes commandes sur les trois autres groupes et de résumer dans un tableau les possibles effets observés ou non par groupe.

4.3 Intervalles de confiance

Intervalle pour les groupes

Les distributions ne sont pas gaussiennes mais, à titre indicatif, il est intéressant de calculer des intervalles de confiance des moyennes pour chaque groupe.

```
# Pour chaque mode d'étalement
tapply(ecoliA$Nombre,ecoliA$ModEtal,t.test)
# Pour chaque dilution
tapply(ecoliA$Nombre,ecoliA$Dilution,t.test)
```

On pourra transformer les données pour revenir aux valeurs initiales puis les représenter graphiquement sur un axe (un axe par groupe) afin de visualiser ces intervalles, notamment leurs tailles.

```
boxplot(Nombre~ModEtal+Dilution+Groupe,data=ecoli)
```

Intervalles individuels

Chacun a en principe obtenu des mesures (6) pour chacune des 4 situations (2 étalements, 2 dilutions). Vous pouvez calculer pour chacune de ces situations les moyennes et les intervalles de confiances sur les données que vous avez observées puis les situer sur la même échelle par rapport aux intervalles de confiance de votre groupe.

```
# Estimation d'un intervalle de confiance
# à partir de 6 valeurs :
t.test(XXX(c(260,252,288,280,276,292)))
...
```

4.4 Analyse en composantes principales

On peut faire une ACP, pour repérer les valeurs expérimentales les plus atypiques. Cette méthode permet de détecter rapidement certains problèmes expérimentaux mais ne "supporte" pas les valeurs manquantes.

```
groupe1=na.exclude(dbrutes)[,1]
acp=prcomp(na.exclude(dbrutes[,-1]),scale=T)
# on exclut les lignes contenant des données NA
biplot(acp)
plot(acp$x,type="p",col=as.integer(groupe1))
# un groupe se démarque-t-il ?
```

D'autres axes d'étude sont également envisageables libre à vous de les explorer.