Scénario: Dénombrements de colonies d'*E. coli*

Résumé

*Pratique des outils statistiques élémentaires (description et tests) sur des données de microbiologie. L'objectif est de déterminer les facteurs pouvant influencer le dénombrement de cultures d'*E.coli *: dilution et mode d'étalement dans les boîtes de Petri.*

1 Introduction

1.1 Rappel du protocole expérimental

L'expérimentation se résume ainsi : chaque étudiant doit réaliser trois boîtes (réplicats techniques) pour deux modes d'étalement de trois dilutions différentes 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 inexploitables, ils ne sont pas considérés.

Les résultats des groupes sont comparés entre eux mais il est clair qu'ils partent chacun de cultures initiales différentes.

Ce qui revient à considérer :

- Trois boîtes identifiées par les étiquettes `B1`, `B2` et `B3`
- Deux modes d'étalement par `Eb` billes ou par `Er` râteau,
- Deux dilutions : `D5` de 10^{-5} et `D6` de $^{10-6}$,
- Trois groupes `Ga`, `Gb`, `Gc` de travaux pratiques.

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 $12\times$(nb d'étudiants) mesures de dénombrement des cultures.

1.2 Précautions

Le travail décrit dans ce scénario a une vocation pédagogique et se limite nécessairement à des outils et modèles statistiques de niveau L3 (tests et analyse de variance à un facteur). En toute rigueur, il faudrait directement mettre en œuvre un modèle plus réaliste (en annexe) mais aussi plus complexe reflétant plus exactement la réalité expérimentale. Ceci reviendrait à expliquer (modéliser) la variable de dénombrement, ou sa transformation, par une combinaison de facteurs dits fixes (groupe, étalement, dilution) car contrôlés dans une planification expérimentale et éventuellement leurs interactions. Ce serait modéliser au "mieux" les différentes sources de variabilité et donc rendre plus fiable une prise de décision.

2 Exploration statistique élémentaire

2.1 Lecture des données

Les données[1] extraites sont disponibles dans le répertoire `http://wikistat.fr/data`; elles ont été regroupées dans le fichier `microbio.dat`. 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 ou C), suivi de 12 valeurs : les trois boîtes de `Eb x DO6`, les trois `Er x D06`, les trois boîtes `Eb x DO5`, les trois `Er x DO5`. Les données manquantes sont codées `"NA"` (*not available*).

Compte tenu de la spécificité de la première dilution D04 (10^{-4}), celle-ci a été supprimée du fichier.

Télécharger ce fichier dans le répertoire courant de R avant d'exécuter les commandes :

```
# Data frame à partir d'un fichier .csv
dbrutes=read.csv("microbio-2016.csv",
     sep=";",header=TRUE)
# vérification de la bonne lecture
summary(dbrutes)
```

1. Les données initiales sont dans le fichier `microbio-init-2016-2017.xls` sur *moodle*.

Remarquer la présence de nombreuses valeurs manquantes codées "NA".

2.2 Mise en forme des données

Un pré-traitement est nécessaire pour rendre accessibles les données à toute méthode statistique. Celui-ci est tellement courant qu'il est bien de savoir le faire avec le logiciel utilisé. Le travail consiste à construire la matrice dite de *design expérimental* ou de plan d'expérience et de lui faire correspondre toutes les valeurs observées sous la forme d'un seul vecteur : une valeur par combinaison des niveaux des facteurs.

```
# nombre d'étudiants par groupe
nbetud=table(dbrutes$Groupe)
na=nbetud["A"];nb=nbetud["B"];nc=nbetud["C"]
# construction du "design expérimental"
# - facteur numéro de boîte à 3 niveaux
# - facteur étalement à 2 niveaux
# - facteur dilution à 2 niveaux
# - facteur groupe à 3 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)),
   labels=c("Ga","Gb","Gc")))
# 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)
```

2.3 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.

Commenter chacun des graphiques et résultats obtenus.

```
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)
```

Les diagrammes boîtes révèlent des dénombrements atypiques mais sans que l'erreur de manipulation soit flagrante, elles sont conservées.

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)
```

Commenter les différents graphiques.

La distribution présente des caractéristiques peu "gaussiennes" qu'il est utile de corriger. Comparer les trois distributions :

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

Justifier le choix de remplacer "Nombre" par sa racine carrée [2].

```
ecoli$Nombre=sqrt(ecoli$Nombre)
# Dernières vérifications
boxplot(Nombre~Dilution,data=ecoli)
```

2. Remarque : la distribution d'une variable de comptage suit généralement une loi de Poisson ; elle est alors rendue plus "symétrique" par une transformation "racine carrée".

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

Quelles intuitions sur les résultats ces graphiques suggèrent-ils ?

3 Inférence statistique

3.1 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 ; ils sont donnés à titre indicatif.

```
# 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)
```

Commenter les résultats obtenus en justifiant précisément les réponses aux questions :

- 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 ?

3.2 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",]
```

Pour analyser les résultats d'un groupe, 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 dont les risques ne sont pas exactement contrôlés.

```
# 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)
```

Exécuter les mêmes commandes sur les deux autres groupes. Résumer dans un tableau les possibles effets observés ou non par groupe.

3.3 Intervalles de confiance

Intervalle pour les groupes

Les distributions ne sont pas gaussienne 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)
tapply(ecoliB$Nombre,ecoliB$ModEtal,t.test)
tapply(ecoliC$Nombre,ecoliC$ModEtal,t.test)
# Pour chaque dilution
tapply(ecoliA$Nombre,ecoliA$Dilution,t.test)
tapply(ecoliB$Nombre,ecoliB$Dilution,t.test)
tapply(ecoliC$Nombre,ecoliC$Dilution,t.test)
```

Élever les résultats au carré pour revenir aux valeurs initiales puis les représenter graphiquement sur un axe (un axe par groupe)afin de visualiser ces intervalles, notamment leurs tailles.

Comparer ces résultats avec le graphe précédemment obtenu.

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

Intervalles individuels

Chaque étudiant a en principe obtenu des mesures (6) pour chacune des 4 situations (2 étalements, 2 dilutions). Calculer pour chacune de ces situations les moyennes et les intervalles de confiances sur les données que vous avez observées ; les situer sur la même échelle par rapport aux intervalles de confiance de votre groupe. Commenter la précision de vos résultats par rapport à ceux du groupe.

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

3.4 Analyse en composantes principales

Cette méthode permet de détecter rapidement certains problèmes expérimentaux mais ne "supporte" pas les valeurs manquantes.

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

Il serait facile d'identifier les expérimentateurs les plus "atypiques" parmi ceux qui ont fourni toutes les valeurs.

3.5 Modèle cohérent avec la procédure expérimentale

Comme indiqué en introduction, il est préférable de poser directement le modèle correspondant à la situation expérimentale : 3 facteurs avec de possibles interactions. Il s'agit alors d'une ANOVA à trois facteurs. Ce modèle considérant l'ensemble des nombreuses observations, l'hypothèse de normalité n'a pas à être strictement vérifiée ; néanmoins, considérer les racines carrées des dénombrements assure des résultats plus pertinents.

```
# Modèle de toutes interactions d'ordre 2
lm.ecoli=glm(Nombre~(ModEtal+Dilution+Groupe)^2,
   data=ecoli)
qqnorm(lm.ecoli$residuals)
anova(lm.ecoli,test="F")
```

Commenter.