TP d'introduction au logiciel R

Rappel : Les lignes en commentaires sont précédées par #. Pour commenter/décommenter plusieurs lignes, il suffit de les sélectionner, puis dans le menu Code > Commenter/Décommenter.
Dans RStudio, on exécute une ligne par CTRL+ENTRÉE

L'échantillon $x_1, \ldots, x_i, \ldots, x_n$ est noté $\mathbf{x}$. Les commandes R sont en italique.

1 Quelques exercices sur les tests

Exercice 1. Le contexte est celui de l'exercice 2 du chapitre 3.

Répondre à la question 1 en utilisant le logiciel R. Attention : la commande *var* de R correspond à la variance corrigée, et la commande *sd* à l'écart-type corrigé. Autrement dit on a : $s^{obs}_{ech} = sd(\mathbf{x})$

Exercice 2. Le contexte est celui de l'exercice 4 du chapitre 5.

1. Commencer par rappeler la condition à vérifier pour faire le t.test du cours (cf Section 2.2.2 du Chapitre 5). Cette condition est-elle satisfaite ? Pour répondre à cette question on fera un test de Shapiro-Wilk (cf Section 4 du Chapitre 5). La commande R est $shapiro.test(\mathbf{z})$. A votre avis, à quoi correspond ici le vecteur $\mathbf{z}$?

2. Pour faire le t.test du cours, la commande R est

$$t.test(\mathbf{x}, \mathbf{y}, paired = \text{TRUE}, alternative = "greater");$$

l'option *paired=TRUE* signifiant qu'il s'agit de donnée appariées. Quelle décision prenez-vous si $\alpha = 0.05$? A quoi correspond les quantités t et df affichées par R.

3. Une autre façon de procéder est de poser $\mathbf{z} = \mathbf{x} - \mathbf{y}$ et d'utiliser la commande $t.test(\mathbf{z}, alternative = "greater")$: le justifier. Si on ne spécifie pas H_0 dans cette instruction, l'option par défaut retenue par R est $\mu = 0$.

Exercice 3. Le contexte est celui de l'exercice 5 du chapitre 5.

Faire le test d'adéquation avec R. La commande R est *chisq.test(V,p=proba)* où V est le vecteur des n_i, et où *proba* est le vecteur des p_i^{ref}.

1. Quelle décision prenez-vous si $\alpha = 5\%$?

2. A quoi correspondent les quantités $X - squared$ et df affichées par R ?

3. Retrouvez la valeur de la p-value fournie par R. Indications : remarquez que la p-value est égale à $\Pr(T > T_{obs} \,|\, H_0)$ où T désigne la statistique de test, puis utilisez la commande *pchisq* de R. Le principe de cette commande est le suivant : si $Y \sim \chi^2(q)$ alors $pchisq(r, q)$ retourne la valeur de $\Pr(Y \leq r)$ où r désigne un réel positif.

Exercice 4. On souhaite tester l'hypothèse selon laquelle le pH d'une eau de source est neutre ; ce qui correspondant à un pH de 7. Pour ce faire on dispose de 15 prélèvements réalisés à des instants différents de la journée ; les mesures de pH sont les suivantes :

$$6.96, 7.13, 7.06, 6.95, 6.92, 7.05, 7.08, 6.95, 6.99, 7.03, 7.10, 6.87, 7.02, 7.05, 6.96.$$

Faire ce test en utilisant le logiciel R. Indications. On commencera par s'assurer que le test de Shapiro-Wilk ne rejette pas l'hypothèse nulle de normalité des X_i (cf la Section 2.1.1 du polycop). Quand n est petit, la commande R pour faire un test portant sur la moyenne μ est $t.test(\mathbf{x}, mu = \mu_0)$ où μ_0 désigne la valeur de mu que l'on souhaite tester (l'alternative H_1 par défaut retenue par R étant $\mu \neq \mu_0$). Donner la valeur de la p-value retournée par R et conclure ; on prendra un risque de première espèce égal à 5%.

2 Analyse des données du TP d'Endocytose 2013

Il peut-être utile d'effacer la mémoire de R pour éviter les confusions avec des variables préalablement définies. On utilise la commande suivante.

```
rm(list=ls())
```

A. Chargement du fichier de données

Si vous êtes connecté à internet utilisez la première option

```
dat = read.table("http://www.math.univ-toulouse.fr/~sgerchin/docs/DataEndocytose.csv",
      sep=";",dec=",",h=T)
```

Sinon, commencer par définir le répertoire de travail (où se trouve le code et le fichier de données.)

- Sous RStudio, naviguer dans le menu : Session -> Set working directory -> To source file location
- Sinon en utilisant la commande
  `setwd("chemin d'accès")`

Puis

```
dat = read.table("DataEndocytose.csv",sep=";",dec=",",h=T)
```

B. Prise en main

1. Sur RStudio, cliquer sur la variable "dat", dans l'espace Environnement (en haut à droite) afin de visualiser les donnees dans un tableur integré. On peut aussi utiliser `print(dat)`.
   Pour voir la forme du tableau de données on utilise

   ```
   str(dat)
   ```

   Décrivez le tableau. Quels sont le nom des colonnes, que représentent-elles ?
2. Quelques commandes utiles

   ```
   # dat$nom de colonne : permet d'afficher uniquement la colonne correspondante
   dat[3,] # affiche la ligne 3
   dat[,4] # affiche la colonne 4
   dat$temps
   # Pour sélectionner uniquement certaines valeurs on peut procéder ainsi :
   dat[dat$binome==4,]
   dat[dat$temps==30,]
   subset(dat,temps==30)
   ```

   Que donne la commande suivante : `subset(dat,concentration==40 \& temps ==60)` ?
3. Dans la suite, on s'interessera à la densité optique corrigée mesurée par les différents binômes. Comment afficher cette colonne ?

C. Représentation graphique

1. Tracer d'un cinétique pour un binome donné.

   ```
   tab=dat[dat$binome==3,] #on créé un tableau associé au binome 3
   plot(tab$temps, tab$DO.corrige, xlab='temps',ylab='Densite optique corrigée')
   ```

   Tester différents binomes, observez-vous des différences ?
2. On souhaite maintenant afficher toutes les données simultanément

   ```
   # Trace global de l'evolution de la densite optique
   # (toutes concentrations en HRP confondues)
   plot( dat$temps,dat$DO.corrige, xlab='temps',ylab='Densité optique corrigée')
   # pour mieux visualiser on peut mettre des couleurs par concentration :
   plot( dat$temps,dat$DO.corrige,col=(dat$concentration)/10,
         pch=16,xlab='temps',ylab='Densité optique corrigée')
   legend('topleft',legend=c('10','20','30','40'),col=1:4,pch=16)
   ```
3. On peut aussi tracer des boxplots en fonction du temps

   ```
   boxplot(DO.corrige ~ temps, data=dat, notch=FALSE,
   xlab="Temps (min)", ylab="Densite optique corrigée", main="Evolution densite optique corrigée")
   ```

4. Enfin, on peut résumé l'information à l'aide des densités optiques moyennes otenues aux différents temps de mesures.

```
library(gplots)
plotmeans(DO.corrige ~ temps, data=dat)
```

Que représentent les intervalles en bleu ? Comment sont-ils calculés ?

5. Comme ces mesures dépendent fortement des cocentrations initiale en HRP, on peut aussi tracer une courbe par concentration

```
par(las=2)
plotmeans(DO.corrige ~ interaction(temps,concentration,sep=','),
   data=dat, p=0.95, connect=list(1:7,8:14,15:21,22:28),
   ylab="Densité optique corrigé",xlab="Temps, Concentration",
   main="Densite optique en fonction de la concentration de la sonde")
# Avec des boxplots
par(las=2)
boxplot(DO.corrige ~ concentration, data=subset(dat,temps ==20),
   ylab="densite optique corrigée",xlab="Concentration HRP (mug/mL)",
   main="Densite optique corrigée en fonction de concentration HRP \n(à t = 20 min)")
```

Quelle est la source de la variabilité observée ?

D. Premiers tests statistiques

On presente plusieurs tests statistiques pour :
- comparer des moyennes,
- comparer des ecarts-types,
- tester la normalité (condition requise pour le t-test).

1. On veut tester s'il y a une difference de densités optiques moyennes entre deux temps d'expérience (ici 10min et 60min) pour une concentration initialle fixée (ici 40 $\mu g/mL$). Quel test va-t-on employer ? Faut-il vérifier des conditions ?

```
# On se donne les échantillons
ech1 = subset(dat,concentration==40 & temps ==10)
ech2 = subset(dat,concentration==40 & temps ==60)
```

Pour tester la normalité de la variable $X - Y$ on utilise le test de Shapiro Wilk

```
shapiro.test(ech1$DO.corrige-ech2$DO.corrige)
```

- Si on accepte l'hypothèse de normalité, on effectue un $t-$test
  `t.test(ech1$DO.corrige, ech2$DO.corrige, paired=TRUE)`
- Sinon, on choisit un test non-paramétrique de Mann-Witnney-Wilcoxon
  `wilcox.test(ech1$DO.corrige, ech2$DO.corrige, paired=TRUE)`

2. Faire de même entre les temps 40 et 60 minutes en utilisant le troisième échantillon suivant :

```
ech3 = subset(dat,concentration==40 & temps ==40)
```

3. Comparaison entre les moyennes de densité optiques corrigées pour des concentrations de HRP différentes (10 $\mu g/mL$ et 40 $\mu g/mL$) après 20 minutes.

```
ech1b = subset(dat,concentration ==10 & temps ==20)
ech2b = subset(dat,concentration ==40 & temps ==20)
```

Quel test va-t-on employer ? Faut-il vérifier des conditions ?
Pour tester la normalité des échantillons, on utilise le test de Shapiro Wilk.

```
shapiro.test(ech1b$DO.corrige)
shapiro.test(ech2b$DO.corrige)

var.test(ech1$DO.corrige, ech2$DO.corrige)
```

- Si on accepte ces hypothèses, on effectue un $t-$test
  `t.test(ech1b$DO.corrige, ech2b$DO.corrige, paired=FALSE)`
- Sinon, on choisit un test non-paramétrique de Mann-Witnney-Wilcoxon
  `wilcox.test(ech1b$DO.corrige, ech2b$DO.corrige, paired=FALSE)`

4. Faire de même à $t = 60$ minutes.

E. L'absorption est-elle spécifique ou non ?

On cherche à savoir si la vitesse d'absorption du HPR par les cellules est linéaire en fonction de la concentration en HPR ou s'il y a un phénomène de saturation. Pour cela, on trace les graphiques de l'évolution de la densité optique en fonction de la concentration pour différents temps.

```
par(mfrow=c(1,1))
plotmeans(DO.corrige ~ interaction(concentration,temps,sep=","),
          data=dat, p=0.95, connect=list(1:4,5:8,9:12,13:16,17:20,21:24,25:28),
          ylab="densite optique corrigée (moy +- et)",xlab="Concentration,Temps",
          main="Densite optique corrigée en fonction \n de la concentration de la sonde\n pour différen
```

Si on s'interesse uniquement à la densité optique corrigée après 20 minutes en fonction de la concentration initiale, on peut faire trois graphiques différents. Le premier correspond aux données brutes, le second donne les moyennes et les intervalles de confiance et le dernier des boxplots. Comment sont- construits ces graphiques ?

```
dat$DO.corrige[dat$temps==20]
plot(x=dat$concentration[dat$temps==20],y=dat$DO.corrige[dat$temps==20])

plotmeans(DO.corrige ~ concentration,
          data=subset(dat, temps ==20), p=0.95,
          ylab="densite optique (moy +- et)",xlab="Concentration sonde (mug/mL)",
          main="Densite optique en fonction de concentration HRP (à t = 20 min)")

boxplot(DO.corrige ~ concentration,
          data=subset(dat, temps ==60),
          ylab="densite optique (moy +- et)",xlab="Concentration sonde (mug/mL)",
          main="Densite optique en fonction de concentration HRP (à t = 20 min)")
```

Que pensez vous de ces graphiques ?