Analyse de la variance

10 octobre 2018

Ce laboratoire doit être remis le **24 octobre à 17h sur Moodle**. Dans votre réponse pour chaque question, veuillez inclure une copie du code R utilisé (s'il y a lieu) et des résultats obtenus.

1. Prises de charbonnier en Alaska

Le fichier `sablefish.csv` contient des données de Kimura (1988) sur le nombre de prises de charbonnier par unité d'effort (*catch*) dans quatre localités d'Alaska (*location*) pour chacune des six années entre 1978 et 1983.

```
sable <- read.csv("sablefish.csv")
str(sable)
```

```
## 'data.frame':    24 obs. of  3 variables:
##  $ year    : int  1978 1978 1978 1978 1979 1979 1979 1979 1980 1980 ...
##  $ location: Factor w/ 4 levels "Chirikof","Kodiak",..: 3 1 2 4 3 1 2 4 3 1 ...
##  $ catch   : num  0.236 0.204 0.241 0.232 0.14 0.202 0.228 0.268 0.286 0.275 ...
```

Supposons que nous nous intéressons à déterminer si l'abondance de ce poisson varie d'une année à l'autre. Dans ce cas, les localités sont des blocs où le nombre de prises, une mesure indirecte de l'abondance, est mesuré à chaque année.

a) Effectuez une ANOVA pour déterminer si l'abondance varie significativement d'une année à l'autre ($\alpha = 0.05$). À partir des graphiques de diagnostic, vérifiez que les suppositions du modèle d'ANOVA soient respectées. Assurez-vous que l'année soit considérée comme une variable catégorielle (facteur).

b) Quelle est la fraction de la variance totale des prises expliquée par le modèle en (a)? Était-ce important de tenir compte des blocs définis par les localités (*location*) pour cette analyse? Justifiez votre réponse.

c) Ré-analysez le modèle en (a) avec la fonction de régression linéaire `lm`. Utilisez les contrastes appropriés pour déterminer la moyenne générale des prises (*catch*) et les déviations de cette moyenne pour chaque année.

2. Résistance à l'eau du bois

Le fichier `woodstain.csv` contient des données de Potcner et Kowalski (2004) sur la résistance à l'eau (*resistance*) d'échantillons de bois traités avec deux pré-traitements (*pretreat*) et quatre teintures (*stain*). Il y a trois réplicats pour chaque combinaison de pré-traitement et de teinture.

```
stain <- read.csv("woodstain.csv")
str(stain)
```

```
## 'data.frame':    24 obs. of  3 variables:
##  $ resistance: num  53.5 32.5 46.6 35.4 44.6 52.2 45.9 48.3 40.8 43 ...
##  $ pretreat  : int  2 2 2 2 2 2 2 2 1 1 ...
##  $ stain     : int  2 4 1 3 4 1 3 2 3 1 ...
```

Analysez les résultats de cette expérience avec une ANOVA à deux facteurs, puis répondez aux questions suivantes.

a) Y a-t-il une différence statistiquement significative ($\alpha = 0.05$) entre les différents pré-traitements et les différentes teintures? Les effets des deux facteurs sont-ils additifs, ou y a-t-il une interaction?

b) Si l'un des deux facteurs ou leur interaction a un effet significatif, comment pourriez-vous estimer la fraction de la variance totale due à cet effet?

c) Si l'un des deux facteurs ou leur interaction a un effet significatif, quel est l'estimé de la différence moyenne de résistance à l'eau entre les traitements? Quel est son intervalle de confiance à 95%?

3. Caractéristiques des variétés de choux

Le jeu de données `cabbages` inclus dans le package `MASS` présente le poids en kg (*HeadWt*) et la teneur en vitamine C (*VitC*) de choux selon la variété (cultivar *Cult*) et la date de plantage. Il y a 10 réplicats pour chacune des six combinaisons de variété et de date.

```
library(MASS)
str(cabbages)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ Cult  : Factor w/ 2 levels "c39","c52": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date  : Factor w/ 3 levels "d16","d20","d21": 1 1 1 1 1 1 1 1 1 1 ...
##  $ HeadWt: num  2.5 2.2 3.1 4.3 2.5 4.3 3.8 4.3 1.7 3.1 ...
##  $ VitC  : int  51 55 45 42 53 50 50 52 56 49 ...
```

a) Pour chacune des deux variables numériques (*HeadWt* et *VitC*), produisez un graphique de boîtes à moustaches montrant la distribution de cette variable pour chaque combinaison de *Cult* et *Date*. Dans chaque cas, semble-t-il y avoir une interaction entre les deux facteurs? Avant même de réaliser l'ANOVA, croyez-vous que les suppositions du modèle (en particulier l'égalité des variances) seront respectées?

b) Choisissez l'une des deux variables (*HeadWt* ou *VitC*) qui correspond le mieux au modèle d'ANOVA d'après votre résultat en (a). Réalisez une ANOVA à deux facteurs et déterminez si l'interaction a un effet significatif.

c) Effectuez une nouvelle ANOVA à deux facteurs pour le même modèle qu'en (b), mais sans interaction (modèle additif). Enregistrez le résultat dans une variable `an3_add`. Affichez le sommaire du modèle linéaire équivalent à cette ANOVA avec le code: `summary(lm(an3_add))`. Comment interprétez-vous chacun des coefficients du modèle linéaire?

d) (*Bonus*) En quoi le test t rapporté pour chaque coefficient dans le tableau obtenu en (c) diffère-t-il du test des étendues de Tukey, obtenu avec `TukeyHSD(an3_add)`?