Tableaux de contingence et ANOVA

29 septembre 2021

1. Sélection florale d'une espèce de bourdon

En suivant l'activité de butinage de bourdons fébriles (*Bombus impatiens*) sur un site, vous notez le nombre de visites des bourdons sur quatre genres de plantes, ainsi que la proportion des fleurs de chaque genre sur le site.

Genre	Nombre de visites	Proportions des fleurs du site
Rubus	8	0.12
Solidago	8	0.24
Trifolium	18	0.33
Vaccinium	11	0.31

L'hypothèse nulle pour cette étude est que *B. impatiens* visite chaque genre proportionnellement à sa prévalence sur le site.

a) Selon l'hypothèse nulle, quelles sont les fréquences attendues pour les visites à chaque genre de plante?

Solution

```
# Les fréquences attendues sont les proportions multipliées par le nombre total de visites
f_obs <- c(8, 8, 18, 11)
prop <- c(0.12, 0.24, 0.33, 0.31)
f_att <- prop * sum(f_obs)
f_att
```

```
## [1]  5.40 10.80 14.85 13.95
```

b) Testez l'hypothèse nulle avec la fonction `chisq.test` dans R, avec un seuil de signification de 5%. Si l'hypothèse nulle est rejetée, quel(s) genre(s) sont plus ou moins visités que prévus?

Solution

```
chisq.test(f_obs, p = prop)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  f_obs
## X-squared = 3.2698, df = 3, p-value = 0.3519
```

La valeur p (0.35) est supérieure à 0.05, donc nous ne pouvons pas rejeter l'hypothèse nulle selon laquelle chaque genre est visité proportionnellement à sa prévalence.

2. Butinage de trois espèces de bourdon

Sur le même site que l'exercice précédent, vous observez l'activité de butinage de deux autres espèces de bourdons (*B. affinis* et *B. ternarius*). Voici le tableau de contingence montrant le nombre de visites par

espèce de bourdon et par genre de plante.

	Rubus	*Solidago*	*Trifolium*	*Vaccinium*
B. affinis	10	9	15	8
B. impatiens	8	8	18	11
B. ternarius	20	4	6	5

a) Quelle hypothèse nulle pouvez-vous tester à partir de ce tableau? Quelle est l'hypothèse alternative?

Solution

L'hypothèse nulle est que la distribution des visites entre genres de plantes ne varie pas selon l'espèce de bourdon. L'hypothèse alternative est que certains bourdons sont plus associés à certaines plantes.

b) Créez une matrice représentant ce tableau dans R, puis testez l'hypothèse nulle mentionnée en (a) avec la fonction `chisq.test`, avec un seuil de signification de 5%.

Solution

```
# Tableau de contingence
tab <- rbind(
    c(10, 9, 15, 8),
    c(8, 8, 18, 11),
    c(20, 4, 6, 5)
)
rownames(tab) <- c("affinis", "impatiens", "ternarius")
colnames(tab) <- c("Rubus", "Solidago", "Trifolium", "Vaccinium")
tab
```

```
##           Rubus Solidago Trifolium Vaccinium
## affinis      10        9        15         8
## impatiens     8        8        18        11
## ternarius    20        4         6         5
```

```
khi2 <- chisq.test(tab)
khi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 16.569, df = 6, p-value = 0.01101
```

L'hypothèse nulle est rejetée car la valeur p est inférieure à 0.05.

c) D'après les résultats du test en (b), quel est le nombre de degrés de liberté du χ^2? Comment cette valeur est-elle calculée?

Solution

6 degrés de liberté: (# de rangées - 1) x (# de colonnes - 1)

d) Comment pouvez-vous consulter les fréquences attendues selon l'hypothèse nulle, ainsi que les résidus?

Solution

```
khi2$expected
```

```
##              Rubus Solidago Trifolium Vaccinium
## affinis   13.08197 7.229508  13.42623  8.262295
## impatiens 14.01639 7.745902  14.38525  8.852459
```

```
## ternarius 10.90164 6.024590  11.18852  6.885246
```

```
khi2$residuals
```

```
##                Rubus    Solidago  Trifolium   Vaccinium
## affinis   -0.8521018  0.65847538  0.4295012 -0.09125145
## impatiens -1.6070082  0.09129897  0.9530597  0.72178774
## ternarius  2.7556068 -0.82484694 -1.5511631 -0.71846940
```

e) Si l'hypothèse nulle est rejetée, quelle paire bourdon-plante a le résidu le plus positif, et laquelle a le résidu le plus négatif? Comment interprétez-vous ces résidus?

Solution

Le résidu le plus positif est pour la paire *ternarius-Rubus*, le résidu le plus négatif est pour la paire *impatiens-Rubus*. Donc, il semble que *B. ternarius* a une préférence pour *Rubus* tandis que *B. impatiens* évite cette plante.

3. Caractéristiques de choux plantés à différentes dates

Le jeu de données `cabbages` inclus dans le package `MASS` présente le poids en kg (*HeadWt*) et le nombre d'unités de 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.

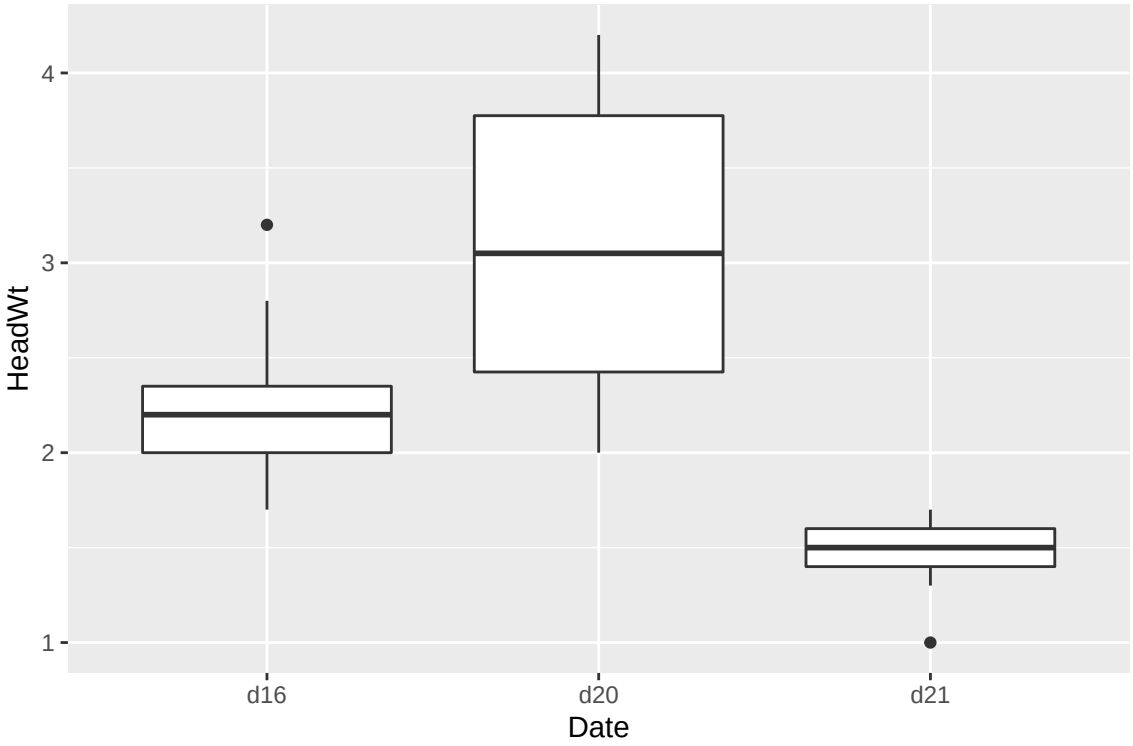
```
library(MASS)
str(cabbages)
```

```
## '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 ...
```

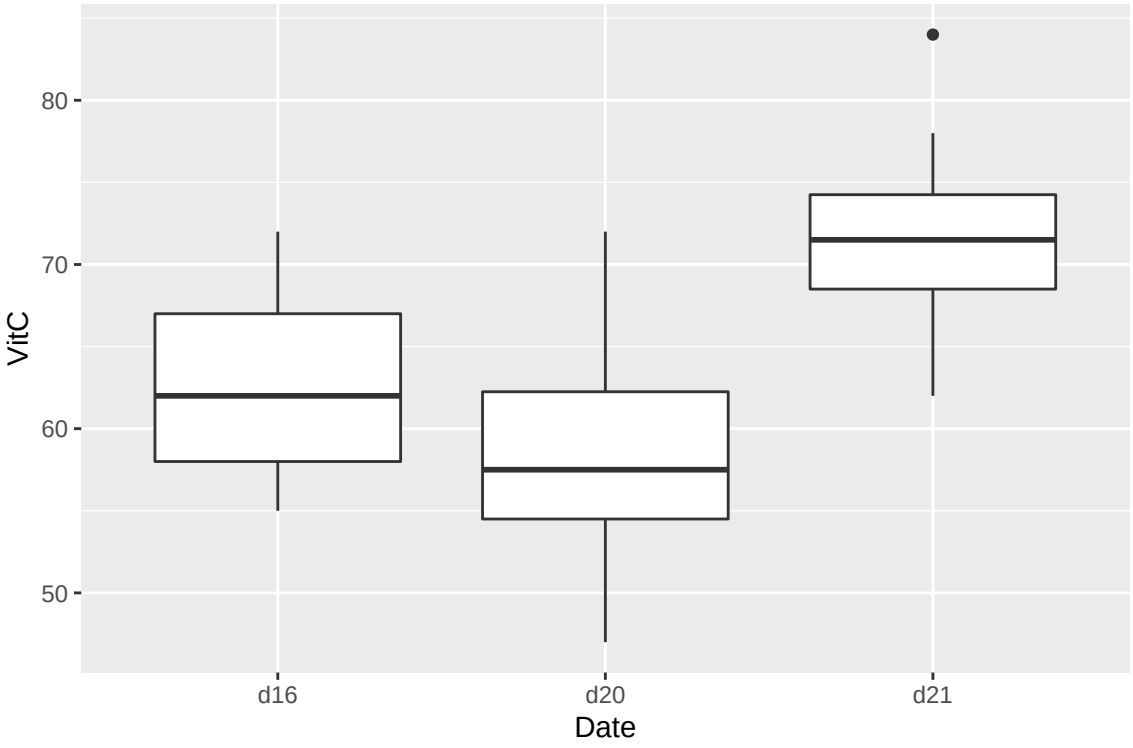
a) Choisissez le sous-ensemble des données correspondant à la variété c52. Pour chacune des deux variables numériques (*HeadWt* et *VitC*), produisez un graphique de boîtes à moustaches montrant la distribution de cette variable selon la date de plantage. Avant même de réaliser l'ANOVA, croyez-vous que les suppositions du modèle (en particulier l'égalité des variances) seront respectées dans chaque cas?

Solution

```
library(dplyr)
library(ggplot2)
c52 <- filter(cabbages, Cult == "c52")
ggplot(c52, aes(x = Date, y = HeadWt)) +
    geom_boxplot()
```




```
ggplot(c52, aes(x = Date, y = VitC)) +
    geom_boxplot()
```



En comparant l'étendue des boîtes à moustache, la variance de *HeadWt* est très différente selon la date; pour *VitC*, la variance est plus uniforme d'un cas à l'autre.

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 un facteur et déterminez si la moyenne de cette variable varie significativement ($\alpha = 0.05$) selon la date de plantage.

Solution

```
an_vitc <- aov(VitC ~ Date, data = c52)
summary(an_vitc)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Date         2  886.2   443.1   10.08 0.000538 ***
## Residuals   27 1187.0    44.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'effet de la date est significatif.

c) Selon un test des étendues de Tukey, entre quelles dates retrouve-t-on une différence significative ($\alpha = 0.05$)? Quel est l'estimé de chacune des différences significatives?

Solution

```
TukeyHSD(an_vitc)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = VitC ~ Date, data = c52)
##
## $Date
##         diff        lwr       upr     p adj
## d20-d16 -3.6 -10.952045  3.752045 0.4553220
## d21-d16  9.3   1.947955 16.652045 0.0110721
## d21-d20 12.9   5.547955 20.252045 0.0004962
```

Les différences estimées par paires (`d20-d16`, `d21-d16` et `d21-20`) sont indiquées dans la colonne `diff`, leur intervalle de confiance à 95% par les colonnes `lwr` et `upr` et la valeur p ajustée dans `p adj`. La différence est significative si l'intervalle `(lwr, upr)` n'inclut pas zéro et la valeur `p adj` est inférieure à 0.05.

Il y a donc une différence significative entre d21 et d16 (9.3 unités de plus pour d21) et entre d21 et d20 (12.9 unités de plus pour d21).