Chi-squared test and ANOVA

September 30, 2020

1. Floral selection of a bumblebee species

By following the foraging activity of bumblebees *Bombus impatiens* on a site, you note the number of bumblebee visits on four genera of plants, as well as the proportion of flowers of each genus on the site.

Genus	Number of visits	Proportions of the flowers of the site
Rubus	8	0.12
Solidago	8	0.24
Trifolium	18	0.33
Vaccinium	11	0.31

The null hypothesis for this study is that *B. impatiens* visits each genus in proportion to its prevalence on the site.

a) According to the null hypothesis, what are the expected frequencies for visits to each genus of plant?

Solution

```
# Expected frequencies are proportions x total visits
f_obs <- c(8, 8, 18, 11)
prop <- c(0.12, 0.24, 0.33, 0.31)
f_exp <- prop * sum(f_obs)
f_exp
```

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

b) Test the null hypothesis with the `chisq.test` function in R, with a significance level of 5%. If the null hypothesis is rejected, which genus or genera are more or less visited than expected?

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
```

The p-value (0.35) is greater than 0.05, so we cannot reject the null hypothesis that each genus is visited proportionally to its prevalence.

2. Foraging activity of three bumblebee species

On the same site as the previous exercise, you observe the foraging activity of two other bumble bee species (*B. affinis* and *B. ternarius*). Here is the contingency table showing the number of visits by bumblebee species and plant genus.

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

a) What null hypothesis can you test from this table? What is the alternative hypothesis?

Solution

The null hypothesis is that the distribution of visits between plant genera doesn't vary between the bumble bee species. The alternative hypothesis is that some bumble bee species are more associated with some of the genera.

b) Create a matrix representing this table in R, then test the null hypothesis mentioned in (a) with the `chisq.test` function, with a significance level of 5%.

Solution

```
# Contingency table
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
```

The null hypothesis is rejected as the p-value is smaller than 0.05.

c) Based on the test results in (b), what is the number of degrees of freedom of the χ^2 distribution? How is this value calculated?

Solution

6 degrees of freedom: (# of rows - 1) x (# of columns - 1)

d) How can you check the expected frequencies according to the null hypothesis, as well as the residuals?

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) If the null hypothesis is rejected, which bumblebee-plant pair has the most positive residual, and which has the most negative residual? How do you interpret these residuals?

Solution

The largest positive residual is for the *ternarius-Rubus* pair, the largest negative residual is for the *impatiens-Rubus* pair. Therefore, it appears that *B. ternarius* prefers *Rubus* whereas *B. impatiens* avoids it.

3. Characteristics of cabbage varieties

The `cabbages` dataset included in the`MASS` package shows the weight in kg (*HeadWt*) and the vitamin C content (*VitC*) of cabbages according to the cultivar (*Cult*) and the planting date. There are 10 replicates for each of the six combinations of cultivar and date.

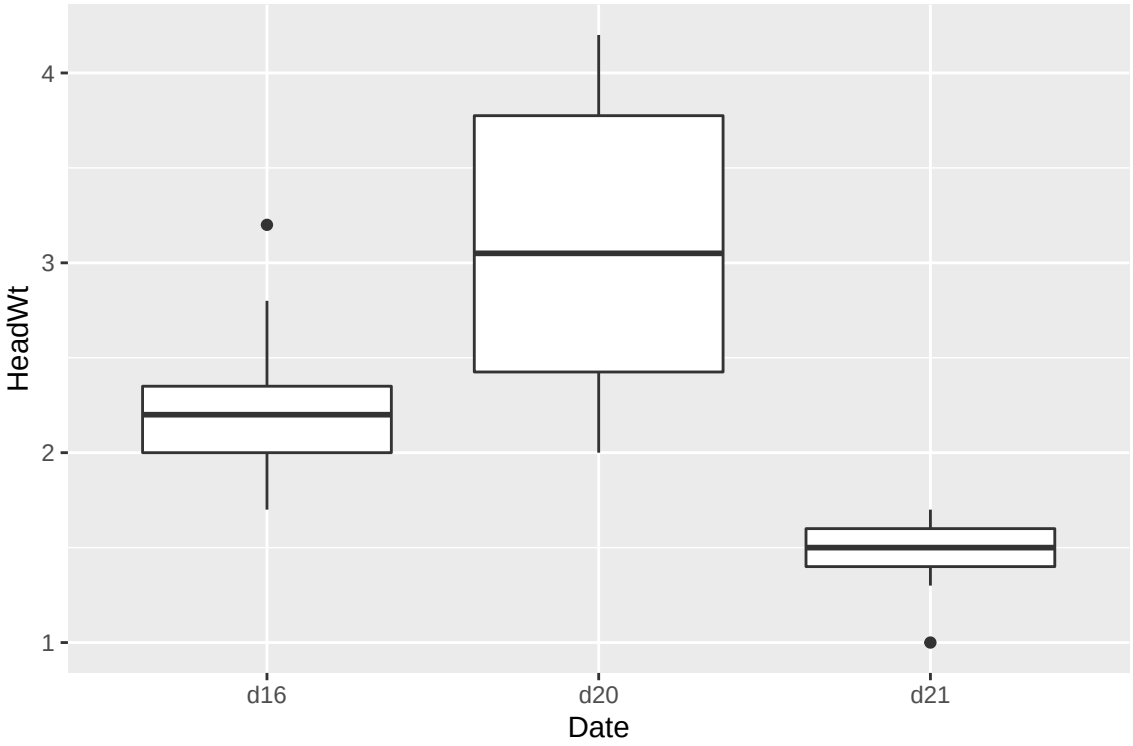
```
library(MASS)
str(cabbages)
```

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

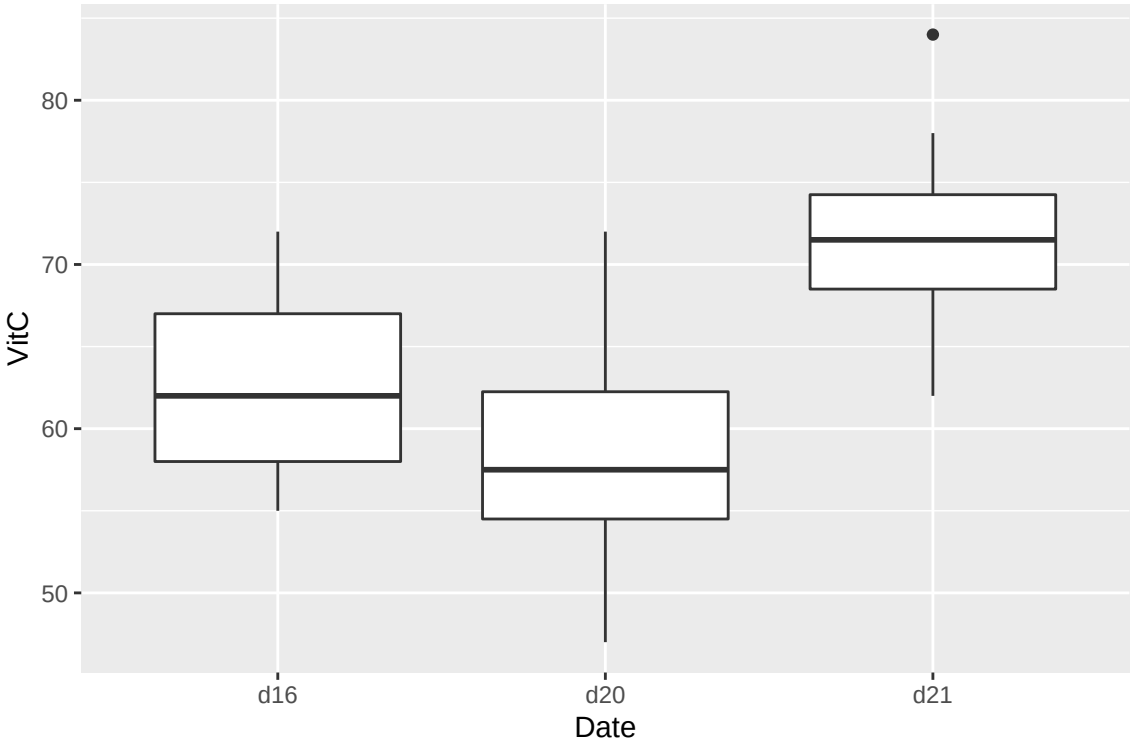
a) Select the subset of data corresponding to cultivar c52. For each of the two numeric variables (*HeadWt* and *VitC*), produce a boxplots graph showing the distribution of that variable for each planting date. Before even performing the ANOVA, do you think that the assumptions of the model (especially the equality of variances) will be respected in each case?

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()
```



By comparing the width of the boxplots, we see that the variance of *HeadWt* is very different depending on the planting date, whereas it is more uniform for *VitC*.

b) Choose one of the two variables (*HeadWt* or *VitC*) that best fits the ANOVA model based on your result in (a). Perform a one-way ANOVA and determine if the planting date has a significant effect ($\alpha = 0.05$).

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
```

Planting date has a significant effect.

c) According to Tukey's range test, between which planting dates are there significant differences ($\alpha = 0.05$)? What is the estimate of those significant differences?

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
```

The estimated pairwise differences (`d20-d16`, `d21-d16` and `d21-d20`) are found in column `diff`, their 95% confidence interval in columns `(lwr, upr)` and the adjusted p-value in column `p adj`. Differences are significant if the confidence interval doesn't include 0 and the `p adj` is smaller than 0.05.

Therefore, there is a significant difference between d21 and d16 (9.3 units greater for d21) and between d21 and d20 (12.9 units greater for d21).