# Analysis of variance

*October 2, 2019*

Answers for this lab must be submitted before **October 9th at 5pm on Moodle**. In your answer for each question, please include a copy of the R code used (if applicable) and the results obtained.

## 1. Sablefish catches in Alaska

The file `sablefish.csv` contains data from Kimura (1988) on catches of sablefish per unit effort in four locations in Alaska for each of the six years between 1978 and 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 ...
```

Suppose we are interested in determining whether the abundance of this fish varies from year to year. In this case, the locations are blocks where the catch per unit effort, an indirect measure of abundance, is measured each year.

a) Perform an ANOVA to determine if abundance varies significantly from year to year ($\alpha = 0.05$). From the diagnostic graphs, verify that the assumptions of the ANOVA model are met. Make sure the year is a categorical variable (factor).

b) What is the fraction of the total variance in catch explained by the model in (a)? Was it important to consider the locations as blocks for this analysis? Justify your answer.

c) Re-analyze the model in (a) with the linear regression function `lm`. Use the appropriate contrasts to determine the mean catch and the deviation from this mean for each year.

## 2. Water resistance of wood

The `woodstain.csv` file contains data from Potcner and Kowalski (2004) on the water resistance (*resistance*) of wood samples treated with two pre-treatments (*pretreat*) and four stains (*stain*). There are three replicates for each combination of pre-treatment and stain.

```
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 ...
```

Analyze the results of this experiment with a two-way ANOVA, then answer the questions below.

a) Is there a statistically significant difference ($\alpha = 0.05$) between different pre-treatments and different stains? Are the effects of the two factors additive, or is there an interaction?

b) If one of the two factors or their interaction has a significant effect, how could you estimate the fraction of the total variance due to that effect?

c) If one of the two factors or their interaction has a significant effect, what is the estimate of the average difference in water resistance between treatments? What is its 95% confidence interval?

## 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) For each of the two numeric variables (*HeadWt* and *VitC*), produce a boxplots graph showing the distribution of that variable for each combination of *Cult* and *Date*. In each case, does there seem to be an interaction between the two factors? Before even performing the ANOVA, do you think that the assumptions of the model (especially the equality of variances) will be respected?

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

c) Perform a new two-way ANOVA for the same model as in (b), but without interaction (additive model). Save the result in an `an3_add` variable. Display the summary of the linear model equivalent to this ANOVA with the code: `summary (lm (an3_add))`. How do you interpret each of the coefficients of the linear model?

d) How is the *t*-test reported for each coefficient in the table in (c) different from the results of Tukey's range test, obtained with `TukeyHSD(an3_add)`?