Compulsory Assignment STAT 210 - Design of Experiments and Analysis of Variance

August 2020

Lars Erik Gangsei

August 14, 2020

Practical info

Upload your solution in one pdf-file by Monday August,24th 23:59h. You may cooperate with other students, but it is not acceptable to copy other's solutions. It is not important to type mathematical symbols nicely. You can, if you prefer, write by hand, scan and upload as one pdf-file. If you chose to write by hand, make sure that your solution is readable for others. You may write your solution in English, Norwegian or any other Scandinavian language.

Software

You may use R or other statistical software. Limit computer output as much as possible in your paper. All computer output should be commented. Use 5% significance level and calculate 95 % confidence intervals unless stated otherwise.

Data

There is two dataset that are to be used in the exercise, 'Listeria.csv' and 'FishingExperiment.csv'. Both datasets, and their associated meatadata, are available at the ***"Datafolder"*** at Canvas.

Data used in the exercises are simulated (not real). However, they are simulated in a way so that they are plausible, based on knowledge from similar experiments.

1 Listeria

The data for this excersise is stored in the file "Listeria.csv".

Listeria is a genus of bacteria, which might cause the disease Listeriosis. Listeria might grow on different food products, like cold cuts of ham ("kokt skinkepålegg" in Norwegian). Cold cuts of ham is available in a large number of brands, differing slightly in thickness, packing method, producers, addings etc. A research group wanted to test the hypothesis if the brand of cold cuts of ham had any effect on Listeria growth. They soon realized that they could not test all brands (you might regard the number of brands as infinite). Thus they randomly picked 6 brands of cold cuts of ham and for those 6 brands they randomly picked 8 replicates. Each replicate was inoculated, in randomly order, with Listeria to a start concentration of 100 colony-forming units per gram (cfu/g). The samples were stored at constant temperature, 4 degrees celsius for 10 days. At that point, the Listeria concentration was registered (in cfu/g).

a) Write down a suitable model for this experiment. Give an interpretation of all variables and model parameters.

b) What are the model assumptions regarding the random error?

c) Fit the model. Test the model assumptions regarding the random error. ***Hint***: In R/ RStudio you might use the 'mixml' package and something like:

```
require('mixlm')
... <- lm(...~r(),data = ...)
```

c) Conduct a log-transformation (ln) of the listeria concentration, and store it in your dataset. ***Hint***: In R/ RStudio you might use something like:

```
... <- log(Listeria$ListeriaNumber)
```

d) Fit a new model using the same basic model as in **a)**, but with the log transformed concentration from **c)** as response ("y"). What is the interpretation of y_{it} now? Test the model assumptions regarding the random error now.

e) Test if there is an effect of brand (of cold cuts of ham) on Listeria concentration.

f) Estimate all model parameters.

g) What is the interpretation of the intraclass correlation $\left(\frac{\sigma_t^2}{\sigma_t^2+\sigma^2}\right)$? Find the estimate, and the 95 % confidence interval for the intraclass correlation.

h) Make a 95% CI for the overall mean.

2 Fishing experiment

The data for this exercise is stored in the file "FishingExperiment.csv".

A nature manager is responsible for sport fishing after trout in 4 different lakes. He suspects that the type of hook and the hour of day (morning or evening) may influence the yield of trout, and that those factors might interact with each other and the lake they are used in. Thus he design a 3 factor factorial design with the three factors Lake (4 levels, 'Lake1', 'Lake2', 'Lake3' and 'Lake4'), Hook (2 levels 'Lure' ("Sluk" in Norwegian) and 'Wobler') and Time (2 levels, 'Morning' or 'Evening'). He sets opp a totally randomized scheme so that he get 3 replications for every "cell" (16 "cells" a total of N = 48 observations) in the design. Each observation is the yield (in hg) of trout after 4 hours fishing.

NB! Excercise a) will be teached wednesday 19th. If you are trying to solve this exercise before that, you might go directly to question **b)** and further on.

He uses the three-factor analysis of variance model, i.e.:
$y_{ijkt} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + e_{ijkt}, e_{ijkt} \overset{i.i.d}{\sim} N(0, \sigma^2)$
$i = 1, \ldots, 4$ and $j = 1, 2$ and $k = 1, 2$ and $t = 1, 2, 3$, and all the "usual restrictions on parameter sums".

y_{ijkt} is the yield in lake i, with hook j at Time k. $\tau_1, \ldots, \tau_4$ are main effects of Lake1-4, β_1 and β_2 are effects of hook ('Lure' and 'Wobler' respectively) and γ_1 and γ_2 are effects of Time ('Morning' and 'Evening' respectively). Second and third order interactions follow 'common notation'.

a) Test if Time (morning or evening) could be excluded as factor in the analysis. ***Hint***: Fit a full model using the full three-factor analysis of variance model (above), and thereafter a reduced model (full two-factor analysis of variance model, with 'Lake' and 'Hook', omitting Time), and apply the partial F-test.

Whatever your anwer were in **a)**, now consider the two-factor factorial design, i.e.:
$y_{ijt} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + e_{ijt}, e_{ijt} \overset{i.i.d}{\sim} N(0, \sigma^2)$
$i = 1, \ldots, 4$ and $j = 1, 2$ and $t = 1, \ldots, 6$, and all the "usual restrictions".

b) Make an interaction plot between Lake and Hook. Use yield as response. Comment.

c) Test formally if the interaction terms could be excluded from the model. State the hypothesis.

d) Find the parameter estimates for main effect of Lake3 and the interactionterm between Lake2 and Wobler

e) Use Tukey's test to test which combinations of Hooks and Lakes that have different expected yields, keeping the overall significance level at 5%.