

Exercises for statistical inference and stuff

Timothée Bonnet

January 16, 2019

Contents

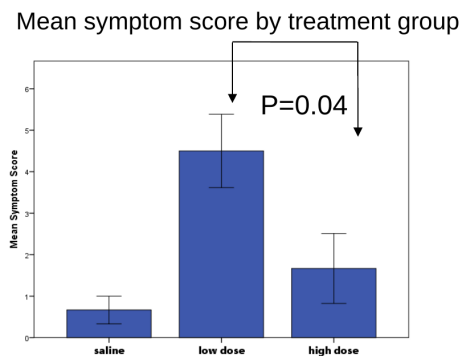
1	Understanding error structure	2
1.1	Broken experiments	2
	Exercise 1 <i>A significant result?</i>	2
	Exercise 2 <i>A better experimental design</i>	2
1.2	Broken models?	3
	Exercise 3 <i>lm vs. lmer vs. lm+</i>	3
2	Interpreting simple mixed models	3
2.1	Summary	3
	Exercise 4 <i>Reading a summary in lme4</i>	3
	Exercise 5 <i>Extracting the right number</i>	3
	Exercise 6 <i>Prediction</i>	3
2.2	Tests	4
	Exercise 7 <i>Testing variance components in lme4</i>	4
	Exercise 8 <i>CI for variance components in lme4</i>	4
	Exercise 9 <i>Testing variance components in MCMCglmm</i>	4
3	Flexible variance structures	4
3.1	Crossed or nested?	4
	Exercise 10 <i>Plants and shelters</i>	4
	Exercise 11 <i>PCR plate</i>	4
3.2	Beyond random intercepts	5
	Exercise 12 <i>Beetles: build a model</i>	5
	Exercise 13 <i>Beetles: look at the model</i>	5
	Exercise 14 <i>Beetles: interpret</i>	5
3.3	Correlated random effects	5

1 Understanding error structure

1.1 Broken experiments

* Exercise 1 A significant result?


We carry out a vaccine challenge experiment. There are three experimental groups (saline control / low vaccine dose / high vaccine dose), with 6 mice in each group. For each group all mice are housed together in the same cage (so there are three cages). All mice are challenged with *Shigella*, their symptom intensity is scored on day 8. A one-way ANOVA identifies an effect of the treatment, and pair-wise Bonferroni tests show the low-dose to be statistically significant different from the two other groups:



What is suspicious in the result? What part of the design may explain the result? How to improve the design (there are at least two different ways)?

* Exercise 2 A better experimental design

We study how a membrane protein intakes external molecules in frog eggs. The target molecules are radioactively labelled so we can measure intake. We created five mutant lines to test what part of the protein controls intake. We propose to measure six tubes of ten eggs every week. Each week we will test a different genotype (i.e., the control or one of the five mutants):

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
<u>Wild-type</u>	Mutant 1	Mutant 2	Mutant 3	Mutant 4	Mutant 5
					
10 eggs/tube	10 eggs/tube	10 eggs/tube	10 eggs/tube	10 eggs/tube	10 eggs/tube
					
x 6	x 6	x 6	x 6	x 6	x 6

How much information about mutants can we extract from this experiment? How to improve the design?

1.2 Broken models?

Let's look at models with data from the an experiment similar to the mice example.

* Exercise 3 **lm vs. lmer vs. lm+**

Load the dataset “challenge2.csv”, and test the effect of the vaccine treatment (“Group”) on the immune response (“Percent.loss”). First fit a fixed-effect only linear model and/or an anova. Is there a clear effect of treatment?

Now add a random effect for Cage (e.g., using the package lme4), what happens to the effect of treatment?

Finally, fit Cage and Group as fixed effects in a simple linear model and/or anova. What happens, why?

Which of the three models should you trust or not?

2 Interpreting simple mixed models

2.1 Summary

* Exercise 4 **Reading a summary in lme4**

Load the dataset “smm.csv” to fit a linear mixed model of y as a function of x, with years as a random intercept. Does x have a significant effect on y? How much variation is explained by differences among years? Is there a measure of uncertainty for this estimate?

* Exercise 5 **Extracting the right number**

After you fit a linear mixed model in lme4, look at the structure of the model (`str(model)`). It is very complicated, and notoriously dangerous to try and extract numbers from there. Fortunately, `summary()`, `VarCorr()`, `fixef()`, `sigma()` and `ranef()` should give you most of what you need. Try to understand what numbers these functions return.

** Exercise 6 **Prediction**

Use the function `predict` to plot model predictions. Because you predict from a mixed model, you have an option called `re.form` to control what random effects are included in the prediction. What is the meaning of the predictions from `re.form=~ 0`, `~ (1|location)` and `~ (1|location) + (1|year)` ?

2.2 Tests

Sometimes random effects are part of the experimental design and are in the models only to control for confounding effects. But sometimes we care about their value or their statistical significance.

**** Exercise 7 Testing variance components in lme4**

Use the function `anova` to test the statistical significance of the random effect “years” and “location”. What are the p-values? How are the p-values calculated here and should you trust this calculation?

*** Exercise 8 CI for variance components in lme4**

Use the function `confint` to estimate confidence intervals for the variance components for “year” and “location”.

**** Exercise 9 Testing variance components in MCMCglmm**

`lme4` can have difficulties estimating random effect parameters when models get a bit complex. A very powerful alternative, I recommend `MCMCglmm`. Try and fit a model of `y` with `x` as a fixed effect, and year and location as random effects. Use `summary` and `plot` to discuss the importance of the random effects.

3 Flexible variance structures

3.1 Crossed or nested?

**** Exercise 10 Plants and shelters**

Load the dataset “`respi.csv`”. We are interested in the interactive effect of genotype and temperature on dark respiration (`rrarea1`). Plants were measured several times in different temperatures, and we need to control for that. Fit a mixed model to test the interaction temperature-genotype. Does `plantID` matter? What if you include shelter as an additional random effect?

Plants with the same name in different shelters were not the same plants. How to account for that?

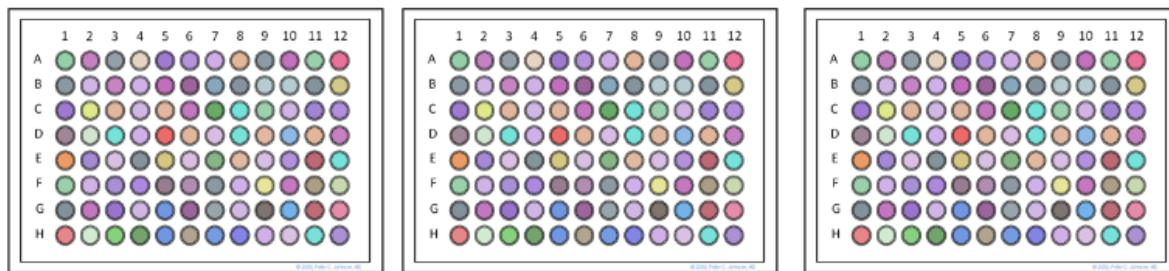
How do the variance component for `plantID` change when you account for it? Why?

*** Exercise 11 PCR plate**

What can affect results on a PCR plate? What is the difference between accounting for such effects with the two following random structures?

Crossed: `(1|plate) + (1|row) + (1|column)`

Nested: `(1|plate) + (1|plate:row) + (1|plate:column)`



3.2 Beyond random intercepts

So far we have considered random effects around intercepts only (that is the meaning of the “1” in the lme4 syntax (1| re)). But random effects can be around fixed effects. You may have heard of “random interactions”, “random slopes”, “random regressions”...

** Exercise 12 Beetles: build a model

Load the dataset “beetles.csv”. It contains (fake) data from an (real) experiment on gene-by-environment interactions. The variable of interest is the mass of beetles born in two different environments, from different parents, and in different cages. Assuming that we can measure genetic variation with parent random effects, we wonder if different genomes respond differently to different environments. **Build the model corresponding to this question in lme4.**

(hints: you could start from a `lm()` of mass modeled by environment, then add random intercepts, and finally a little something more).

** Exercise 13 Beetles: look at the model

What are the variances related to genetic differences? How are they correlated? Does genetic variation explain a lot of the total variation we observe? Try and draw a representation of genetic variation in the two environments.

*** Exercise 14 Beetles: interpret

Interpret model outputs (use raw numbers and / or graphs) to answer the following: Is there evidence for genetic variation? Do the two environment differ in their effects on beetles?

Is there evidence for genetic variation in the response to the environment?

Does that mean that genomes good at environment 1 are bad at environment 2?

3.3 Correlated random effects

If we have time, ask me to talk or give examples of models with genetic or phylogenetic effects, or spatial auto-correlation.