Stats with Sparrows - 18

Julia Schroeder

18 Observer repeatability

Housekeeping!

```
a<-read.table("ObserverRepeatability.txt", header=T)
```

We now want to find out how much the measurement of tarsus and bill width of the same bird depends on different observers. We will use the ANOVA method to do that. Calculate the between-observer repeatability of tarsus of the porn-star female!

We know that repeatability is

$$r = \frac{s_A^2}{(s_W^2 + s_A^2)}$$

with

$$s_W^2 = MS_W$$

$$s_A^2 = \frac{MS_A - MS_W}{n_0}$$

and

$$n_0 = [\frac{1}{a-1}[\sum_{i=1}^{a} n_i - (\frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i})]$$

We have before calculated this using dplyr. If you're adventurous, try doing these sums with tapply, or write a function yourself!

a is the number of groups, that means, how many different observers have measured tarsus and bill width. n_i is the sample size in each group.

```
require(dplyr)

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
a %>%
  group_by(StudentID) %>%
  summarise (count=length(StudentID))
```

```
## # A tibble: 80 × 2
##    StudentID count
##       <fctr> <int>
## 1   AF151093     3
## 2     AH2912     2
## 3    Ak70593     1
## 4    AK70593     1
## 5     AL0109     1
## 6     AR0408     2
## 7     AR1310     1
## 8    ASR0312     2
## 9     BC1603     1
## 10    cd1302     2
## # ... with 70 more rows
```

Here we can see the repeats by student. One row here is one n_i, with the first row being n_1, second n_2 and so on until the last row, which is n_{80}. This number was hard to get to because the output said ...with 70 more rows. So more elegant is if we'd simply count how many observers there are:

```
a %>%
  group_by(StudentID) %>%
  summarise (count=length(StudentID))  %>%
    summarise (length(StudentID))
```

```
## # A tibble: 1 × 1
##   `length(StudentID)`
##                 <int>
## 1                  80
```

And this tells us that 80 students measured tarsus and bill. Thus, in the equation, a, the number of groups, is 80. Don't confuse this a with the a in R, where we named our data a!

We get the sums of the n_1 's from the previous r code to get the denominator in the equation calculating n_0:

```
a %>%
  group_by(StudentID) %>%
  summarise (count=length(StudentID)) %>%
    summarise (sum(count))
```

```
## # A tibble: 1 × 1
##   `sum(count)`
##           <int>
## 1           151
```

Good! The total sum of observations is 151. Hmm. We could have also gotten this with this easier code:

```
length(a$StudentID)

## [1] 151
```

Do you understand why? I hope so!

Next, we need to We sum the square values of each group's n. We can't use length for this, we have to use a sub-grouping function:

```
a %>%
  group_by(StudentID) %>%
  summarise (count=length(StudentID)) %>%
  summarise (sum(count^2))

## # A tibble: 1 × 1
##   `sum(count^2)`
##             <dbl>
## 1             333
```

Now we can solve the equation! Remember, a was 80, the sum of n_i was 151, the sum of squares was 333.

$$n_0 = [\frac{1}{a-1}[\sum_{i=1}^{a} n_i - (\frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i})]$$

1/79*(151-333/151) = 1.88

Yay, that's our n_0. Now that we have that, we can look back at what we mean to calculate, the repeatability:

$$r = \frac{s_A^2}{(s_W^2 + s_A^2)}$$

with

$$s_W^2 = MS_W$$

$$s_A^2 = \frac{MS_A - MS_W}{n_0}$$

Let's check the mean squres, we run an ANOVA for this:

```
mod<-lm(Tarsus~StudentID,data=a)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: Tarsus
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## StudentID 79 561.95  7.1133  4.2004 2.095e-09 ***
## Residuals 71 120.24  1.6935                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, $s_{W}^{2} = 1.69 $

and

$$s_A^2 = \frac{7.11 - 1.69}{1.88} = 2.88$$

That means, the repeatability R is

$$r = \frac{2.88}{1.69 + 2.88)} = 0.63$$

or, 63 percent of the variability of tarsus length is explained by differences in how observers measure it. That's a bit of a poor result, and we (for science) usually aim at a between-observer repeatability of 80 percent or more. If this would be a project, we would conclude that it would be better if fewer observers would measure tarsus more often. So we'd reduce the number of observers and increase n_0, and then, with fewer observers and more repeats, make sure that the repeatability is higher.

Cool! Now do it with bill width!

Now we can also test whether handedness, and which leg was measured made a difference:

```
mod<-lm(Tarsus~Leg+Handedness+StudentID,data=a)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: Tarsus
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## Leg         1   1.64  1.6393  0.9426    0.3350    
## Handedness  2   5.49  2.7473  1.5798    0.2134    
## StudentID  78 555.06  7.1162  4.0919 5.756e-09 ***
## Residuals  69 120.00  1.7391                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This really interesting. We can see handedness and leg doesn't really do anything. but, for a second, have a look at the degrees of freedom. Can you see how StudentID really drags down the df's? If we run a linear model, we really don't want to have factors in it that have many levels. Two levels are fine, but the studentIDs are a few too much. We really estimate ONE mean for each student! This model is thus a bit overparametrised. There is a solution for this, and it's called Mixed Linear Models. It actually does a bit of a mix of a linear model and an ANOVA. In real life, that's not true, but it serves as a good visualisation for now. You really want to spend more than just half an hour on this. But it's good to know it exists:

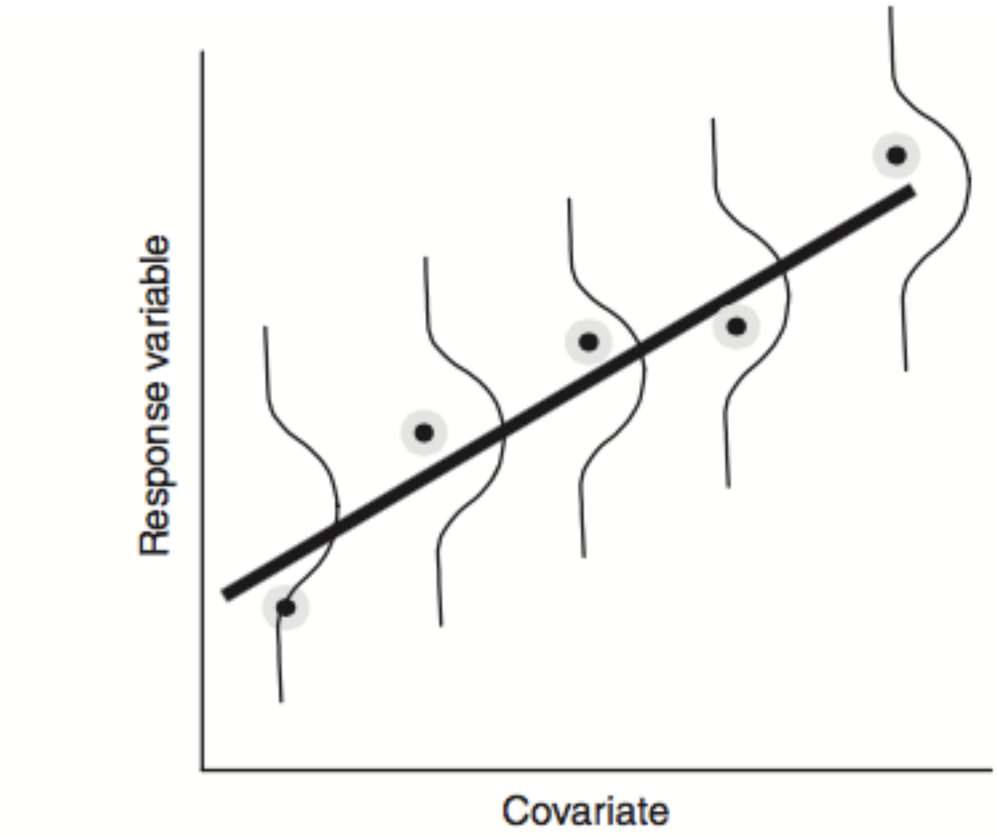
```
require(lme4)

## Loading required package: lme4

## Loading required package: Matrix

lmm<-lmer(Tarsus~Leg+Handedness+(1|StudentID),data=a)
```

In this model, we specify Leg and Handedness as *fixed* factors, as we do in a normal linear mdoel. We then do something that looks odd: (1|StudentID) is code for modelling StudentID as random effect on the intercept. That means, we model for each student, a normal distribution around each student's measurements. We then actually calculate the variance for each student! That's a *random* effect.



Let's have a look at the outcome:

```
summary(lmm)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Tarsus ~ Leg + Handedness + (1 | StudentID)
##    Data: a
##
## REML criterion at convergence: 617.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0964 -0.3545  0.0135  0.4402  3.3502
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  StudentID (Intercept) 3.028    1.740
##  Residual              1.713    1.309
## Number of obs: 151, groups:  StudentID, 80
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 18.33925    1.12055  16.366
## Legright    -0.06686    0.33815  -0.198
## HandednessL  0.25166    1.20036   0.210
## HandednessR  0.43444    1.14267   0.380
##
## Correlation of Fixed Effects:
##             (Intr) Lgrght HnddnL
## Legright    -0.093
## HandednessL -0.832 -0.069
## HandednessR -0.968 -0.039  0.825
```

We first get presented with the model that we've run. We are also informed that this model is not evaluated with ordinary least squares (OLS) method, but rather with a REML method. That's a type of maximum likelihood method, that uses residual, or restricted ML instead of OLS. You can read up on that if you are interested.

We get an overview of the scaled residuals, their median (mean should be zero if they are scale, you should knoe that!) and their spread. Then we see the parameter estimates for the random effect:

We get a variance and a standard deviation for Students. This is the *among student* variance. It is comparable (but not equal) to the MS_A 's of an ANOVA. Just this time, we get a standard deviation with the measurement, that means, we can get some grips on how much the error might be! Cool! We get that because it's a measurement of *many variances, each for each student! Cool. So, we can conclude that 3.03 of the variance is explained by differences between students! We also get the number of observations, and number of studentIDs. The residual variance is similar to the MS_W or the residual mean squares - the variance that's less within-students. Here, we can actually much easier calcuate the repeatability: it's simple the among-variance divided by the sum of among- and residual variance:

3.03/(3.03+1.71) = 0.64

That's a very close result to the repeatability that we've calculated with the ANOVA method before! Cool. Also, check this out:

```
var(a$Tarsus)

## [1] 4.547938

3.03+1.71

## [1] 4.74
```

So we actually get (within error) the variance of the tarsus with this method! Now let's go on inspecting our results:

```
summary(lmm)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Tarsus ~ Leg + Handedness + (1 | StudentID)
##    Data: a
##
## REML criterion at convergence: 617.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0964 -0.3545  0.0135  0.4402  3.3502
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  StudentID (Intercept) 3.028    1.740
##  Residual              1.713    1.309
## Number of obs: 151, groups:  StudentID, 80
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 18.33925    1.12055  16.366
## Legright    -0.06686    0.33815  -0.198
## HandednessL  0.25166    1.20036   0.210
## HandednessR  0.43444    1.14267   0.380
##
## Correlation of Fixed Effects:
##             (Intr) Lgrght HnddnL
## Legright    -0.093
## HandednessL -0.832 -0.069
## HandednessR -0.968 -0.039  0.825
```

Now we can see that indeed, which leg was measured, and with which hand, didn't make much of a difference. The left leg is the reference category here, as is ambidextrous (because both words start with letters earlier in alphanumerical order than right and left). But we don't get any p-values, that's a shame! That is because the author of this function does not feel confident with any of the suggested methods to calculate the degrees of freedom for this method. None of the methods are really satisfactory, so he decided to go without, to err on the side of caution.

What can we do? We can look at the t-values. They are all really, really small. None of it is larger than 1.7. So we can conclude that none of the fixed effects has a significant effect.

With mixed-effect models, you should do the same check-list as you've done on linear models. You should also do some checking on the violation or not of the assumptions.

When should you use random and when fixed effects? First off, you should only use *factors* for random effects. Not covariates - continuous variables should always be fixed (unless you know what you're doing). How to decide whether to put a factorial variable should be a random effect or not? There is no good answer to it. First, if you are interested in the effect size (How much longer are female tarsi than male tarsi? How much more mm do I measure compared to everybody else?), you put them on the fixed side of things. However, only if you have sufficient power - you do not want to spend all df's on fixed factors - aim at df>30, better 100 if you run linear models with fixed factors that have many, many levels. So another rule of thumb is if it has more than five levels, it goes into the random bit.

You can also put things into the random bit when you want to know how much variance that factor explains.

This is not a sufficient introduction to mixed effects modelling, and should only explain to you that these exist. If you want to use them, find books and more information to read up on before you use them!