Random effects in R

Read-in “calcium.csv” and load the library **lme4**.

```
caldata <- read.csv("calcium.csv")
library(lme4)
```

```
## Loading required package: Matrix
```

Suppose a turnip farmer wants to know whether the type of fertiliser he uses affects how much calcium are in the leaves of his plants. However, there are over 10,000 leaves in his crop. It is unfeasible to measure the calcium concentration of each leaf. The farmer can only measure the calcium concentration of a smaller, random sample of leaves. From those results he can infer the calcium concentration of the whole crop.

To minimise the variance in his calcium concentration calculation, should he be measuring leaves from as many different plants as possible? Or is it better to gather fewer plants and instead measure more leaves from each plant?

If there is more variance in calcium concentration between plants, then the farmer should maximise the number of different plants in his sample.

On the other hand, if there is more variance in calcium concentration between the leaves of each plant, then the farmer should maximise the number of leaves selected from each plant (and thus gather fewer plants, comparatively)

So, to see where the most variance lies - between plants or within plants - the farmer randomly selects four plants that have each been treated with a different fertiliser (A, B, C or D). He then randomly selects four leaves from each plant to be measured. Table 1 shows the calcium concentration results for each measured leaf.

Plant	Leaf	Calcium Concentration
A	1	3.28
A	2	3.09
A	3	3.03
A	4	3.03
B	1	3.52
B	2	3.48
B	3	3.38
B	4	3.38
C	1	2.88
C	2	2.8
C	3	2.81
C	4	2.76
D	1	3.34

Plant	Leaf	Calcium Concentration
D	2	3.38
D	3	3.23
D	4	3.23

Table 1: Calcium concentration in turnip leaves from selected sample

To measure the effect of fertiliser type on calcium concentration, we might normally use a simple linear model.

```
model1 <- lm(calconc~plant, data=caldata)
```

But this model doesn't work because we took multiple measurements per plant! We took four measurements from each plant, which violates the independence assumption; i.e., multiple responses from the same subject cannot be regarded as independent from each other.

The solution is that we have to add random effects terms to our model. We need to tell our model that there are actually two sources of random variation: one source due to pure random error and another source due to the random variation seen between our plants. To do this we use the *lmer* function (from the package *lme4*):

```
model2 <- lmer(calconc~plant+(1|leaf), data=caldata)
```

The term **(1|leaf)** is saying to "assume a different intercept for each leaf" (in other words, *leaf* is a random variable). We can think of this formula as telling the model that there are going to be multiple measurements per plant and these measurements will depend on each plant's baseline calcium concentration.

However, model 2 is not quite right. Model 2 is still treating the *plant* variable as a fixed variable when it should be a random variable. *Plant* is a random variable because the plants were randomly selected from a larger population of plants just like the leaves were randomly selected from larger population of leaves. Model 3 shows how we can also treat the *plant* variable as a random variable:

```
model3 <- lmer(calconc~(1|plant)+(1|leaf), data=caldata)
summary(model3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: calconc ~ (1 | plant) + (1 | leaf)
##    Data: caldata
##
## REML criterion at convergence: -24.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.8273 -0.5039 -0.2038  0.1377  1.9168
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  plant    (Intercept) 0.072801 0.26982
##  leaf     (Intercept) 0.004635 0.06808
##  Residual             0.002236 0.04729
## Number of obs: 16, groups:  plant, 4; leaf, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   3.1637     0.1396   22.66
```

See that *plant* has much more variability than *leaf* (0.073 vs. 0.005). It means there is more variation between plants than there is variation between leaves from any one plant. Therefore, the farmer should be measuring the calcium concentration from as many different plants as possible (as opposed to maximinsing the number of leaves measured from any one plant).