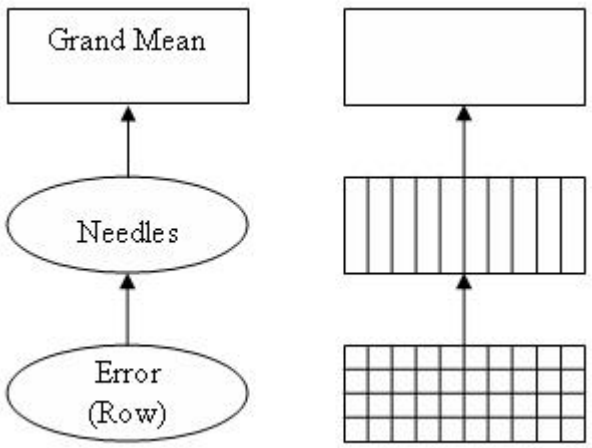
1. Ten needles were randomly selected from a large branch of a loblolly pine tree. The stomata (microscopic breathing holes) on loblolly pine needles are arranged in rows. On each needle, 4 rows were randomly chosen and the number of stomata per centimeter were determined for each row. The resulting data are shown below and are in the file needle.txt.

Needle #	1	2	3	4	5	6	7	8	9	10
	149	136	143	121	148	129	127	134	117	129
	143	139	142	133	121	134	130	137	128	132
	138	129	124	126	124	127	123	119	117	131
	131	143	134	130	128	113	125	130	118	137

(a) Write down the random effects model appropriate to this problem identifying all terms used. State the distributional assumptions. Why is a random effects model more appropriate than a fixed effects model?

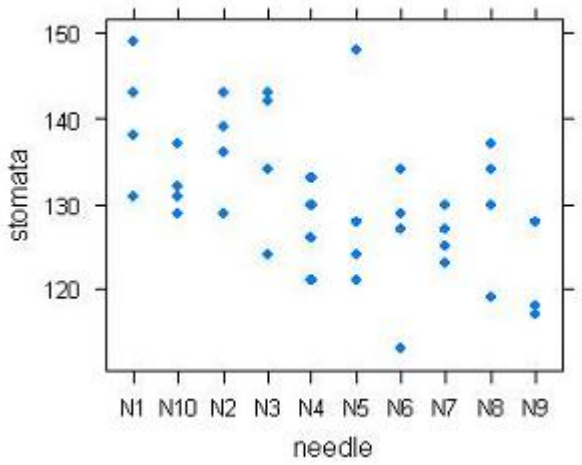
The model is Number of Stomata (per centimeter) = Grand Mean + (Needle, Random) + error(Row). Symbolically, $y_{ij} = \mu + a_i + e_{ij}$, where $a_i \sim Normal(0, \sigma_a^2)$ and $e_{ij} \sim Normal(0, \sigma_e^2)$. A random effects model is appropriate because the needles are drawn at random from the tree. We want to generalize our results to the population of needles on the tree.

(b) Draw a nesting diagram for the model variables as in the notes.



(c) Examine a plot of the stomata counts versus needle. Are the random effects model assumptions reasonable?

The variation within each needle is smaller than the overall variation. There is also a fairly symmetric and even distribution within needles.



(d) Estimate all relevant variance components defined in (a) using both lmer and from computations with sums of squares using an ANOVA table from an analysis using lm (or aov) based on the expected mean square error (EMS) expressions of $\sigma_e^2 + n\sigma_a^2$ a for "treatment" and σ_e^2 for error. (See end of notes for *Random Effects in R.*) Are the estimates similar?

```
> summary(needle.lmer)

Random effects:
 Groups   Name        Variance Std.Dev.
 needle   (Intercept) 22.651   4.7593
 Residual             53.808   7.3354

> summary(needle.aov)

Error: needle
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  9 1299.73  144.41

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 30 1614.25   53.81
```

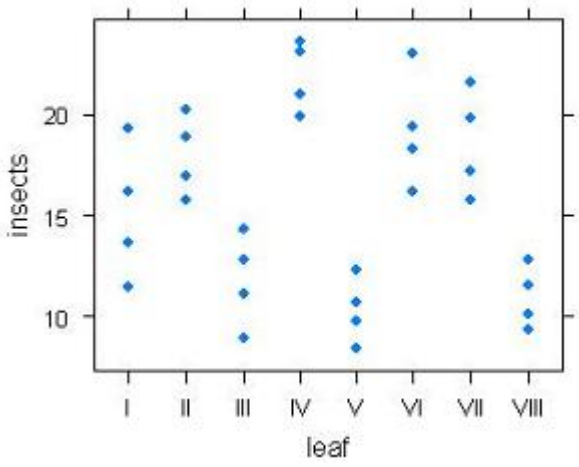
From the above output, $\sigma_e^2 = MSError = 53.81$ and $\sigma_a^2 = \frac{MSTr - MSError}{4} = 22.65$. These estimates are exactly the same as those found using the lmer function.

2. Suppose it is of interest to estimate the mean number (μ) of parasitic insects per unit leaf weight for a particular tree. Eight leaves were randomly selected from the tree. From each leaf, four small disks were cut. For each disk the number of insects per unit leaf weight was determined. The data presented below are also in the file leaf.txt.

Leaf #	I	II	III	IV	V	VI	VII	VIII
	11.4	20.2	14.3	23.6	8.4	18.3	21.6	12.8
	19.3	17.0	11.1	23.1	10.7	16.2	15.8	9.3
	16.2	15.8	12.8	19.9	12.3	23.0	17.2	11.5
	13.6	18.9	8.9	21.0	9.8	19.4	19.8	10.1

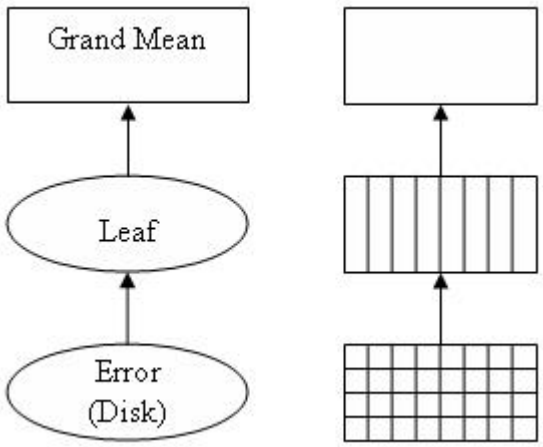
(a) Examine a dotplot of the data that shows the insect data plotted against each leaf. Do the mean counts look similar for each leaf? Is the spread of counts similar for each leaf? (You do not need to include the plot in your solution, but you may if it makes you happy.)

The spread of counts looks relatively similar from leaf to leaf. However, the mean counts do not look similar. Leaves III, V, and VIII have consistently low densities of insects, leaves IV, VI, and VII have higher densities, while leaves I and II are in the middle.



(b) Give a suitable model for describing these data, identifying all terms in the model and identifying any distribution assumptions. Draw a nesting diagram for the model variables as in the notes, indicating which variables should be modeled as random effects and which as fixed.

The model is Number of Insects (per unit leaf weight) = Grand Mean + (Leaf, Random) + error(Disk). Symbolically, $y_{ij} = \mu + a_i + e_{ij}$, where $a_i \sim Normal(0, \sigma_a^2)$ and $e_{ij} \sim Normal(0, \sigma_e^2)$.



(c) Find a 95% CI for μ assuming a t distribution with 7 degrees of freedom.

```
> 15.728-qt(0.975,7)*1.536
[1] 12.09594
> 15.728+qt(0.975,7)*1.536
[1] 19.36006
```

(d) Use the mcmcsamp function with a sample size of 10,000 to find a 95% credible region for μ. How does this region differ from that found in (c)? Is its width much larger, much smaller, or about the same?

```
> leaves.lmer = lmer(insects ~ 1 + (1|leaf))
> leaves.samp = mcmcsamp(leaves.lmer, n=10000)
> CI95 = apply(leaves.samp,2,function(x)quantile(x,prob = c(0.025, 0.975)))
> CI95[,1]
    2.5%    97.5%
12.19793 19.30082
```

This region is slightly smaller than the region found in part (b).

(e) Fit a model for insects using leaf as a fixed effect using lm. Based on this model find a 95% CI for μ. How many degrees of freedom are used here? How does this interval compare with the intervals from parts (c) and (d)? Which interval or intervals are most appropriate?

```
> 15.728-qt(0.975,24)*sqrt(2.339)
[1] 12.57152
> 15.728+qt(0.975,24)*sqrt(2.339)
[1] 18.88448
```

This interval is smaller than the previous two. If we fix the effect, we have less variability introduced. Therefore the standard error decreases as compared to a random effects model. The fixed effect interval is inappropriate since leaves were sampled from the tree, not disks. The leaf is the "experimental unit". The disks are a way to measure each leaf without having to measure the entire leaf, but we expect (and see) disks to be more similar within leaves than across leaves.

(f) In a balanced experiment with k leaves and s disks sampled from each leaf, the expression for the variance of the intercept treating leaf as a random effect is as follows.

$$V(\hat{\mu}) = \frac{\sigma_a^2}{k} + \frac{\sigma_e^2}{ks}$$

Verify that this expression is consistent with the summary in R using lmer.

```
##Result using given expression
> 17.5099/8 + 5.4699/(8*4)
[1] 2.359672

##Result using se from lmer function in R
> 1.536^2
[1] 2.359296
```

The resulting estimates are nearly the same.

(g) Suppose that 16 leaves had been selected and two disks per leaf had been taken. Assume that the same estimates for σ_e^2 and σ_a^2 are obtained as in the actual experiment (using notation from lecture). What would the estimate of the variance of $\hat{\mu}$ be in this case? Compare this with the estimate for the variance of $\hat{\mu}$ in the actual experiment. Interpret the comparison. Is it better to sample more leaves or more disks per leaf given a fixed total sample size?

```
> 17.5099/16 + 5.4699/(16*2)
[1] 1.265303
```

This estimate is smaller than the estimate based on 8 leaves with 4 disks each. Note that the denominator of the second term ($16 \times 2 = 32$) is the same for both scenarios ($8 \times 4 = 32$). Therefore the only difference comes from the first term. By having more leaves, we are minimizing this term. Therefore it is better to sample more leaves given a fixed total sample size. This is true in this case and in general when there is more leaf-to-leaf variability than within leaf variability.

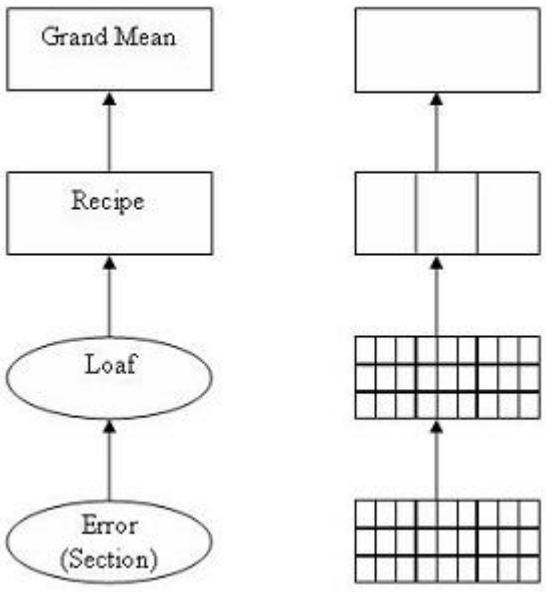
3. An experiment was conducted to compare the amount of calcium in loaves of bread made with three different recipes. Two loaves were made from each of the three recipes. Measurements of calcium were made on small sections obtained from the loaf interior; three small sections

were obtained for each loaf and measured for calcium concentration. The data were recorded as follows. Note that the "loaf number" was used to identify the two loaves for a given recipe. The three observations in each group are the calcium measurements on the small sections; the units are mg of calcium per gm of bread. The data is also in the file bread.txt.

	Recipe		
	A	B	C
Loaf 1	0.18	0.19	0.07
	0.15	0.16	0.10
	0.16	0.18	0.08
Loaf 2	0.14	0.23	0.09
	0.12	0.20	0.12
	0.14	0.20	0.10

(a) Write down a random effects model appropriate for this experiment. Identify the terms of the model. Draw a nesting diagram for the model variables as in the notes.

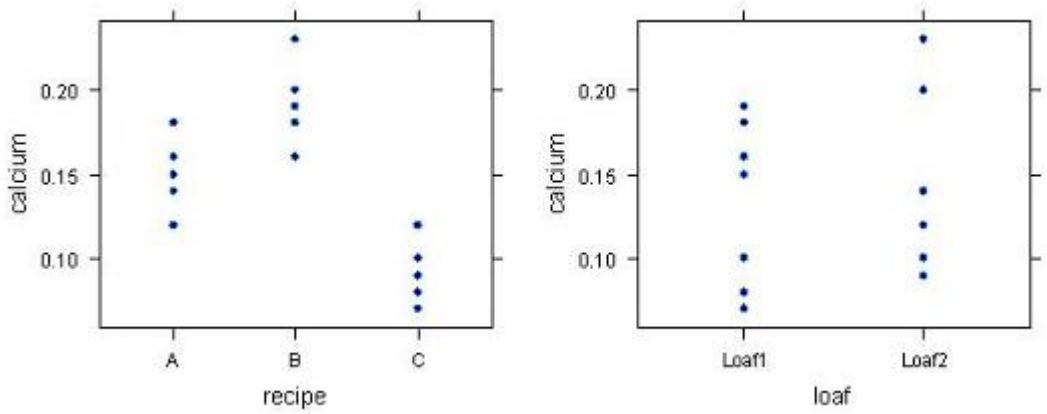
The model we want to fit is Calcium (mg) = Grand Mean + Treatment(Recipe, fixed) + Block(Loaf, random) + error(Section). Recipe is treated as fixed because we are interested in the amount of calcium found based on these three specific options. Loaf is treated as random because we do not wish to test the difference between loaves - any differences should be do to random chance. Mathematically, $y_{ijk} = \mu + \alpha_i + b_{ij} + e_{ijk}$, where $b_{ij} \sim Normal(0, \sigma_b^2)$ and $e_{ijk} \sim Normal(0, \sigma_e^2)$ where $i = 1, 2, 3$; $j = 1, 2$; $k = 1, 2, 3$.



(b) Plot the data versus loaf and versus recipe. Summarize your observations. Is the model reasonable?

According to the plot of calcium versus loaf, there does not appear to be large variation between the two loaves. The three recipes do, however, appear to differ. Recipe B

seems to have the most calcium and recipe C has the least. A statistical test should be performed to verify if these differences are significant.



(c) Is there much difference in the recipes effect on calcium content in bread? Summarize your findings.

```
> summary(bread.lmer)
Linear mixed-effects model fit by REML
Formula: calcium ~ 1 + recipe + (1 | recipe:loaf)
   Data: bread
    AIC   BIC logLik MLdeviance REMLdeviance
 -64.86 -61.3  36.43     -93.17       -72.86
Random effects:
 Groups      Name        Variance   Std.Dev.
 recipe:loaf (Intercept) 0.00032593 0.018053
 Residual                0.00022778 0.015092
number of obs: 18, groups: recipe:loaf, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.14833    0.01417  10.465
recipeB      0.04500    0.02005   2.245
recipeC     -0.05500    0.02005  -2.744

Correlation of Fixed Effects:
        (Intr) recipB
recipeB -0.707
recipeC -0.707  0.500
```

The loaf effect does not appear to be significant. The variance associated with loaf, nested in recipe, is only 0.0003. The loaf differences are significantly different, supporting our findings in part (b). According to the summary statistics, recipe B has the highest calcium content (0.045 mg above recipe A) while recipe C has the lowest (0.055 mg below recipe A).