1 Anova Models

Types of Anova models

Anova models can be divided into three basic types with respect to the categorical variables (we have already seen two, though have not really differentiated them)
They are:

- Model I (fixed effects)
- Model II (random effects)
- Mixed model (some random some fixed effects)

1.1 Model I Anova

Model I Anova

This is the anova you have encountered and we have looked at in some depth:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

In this model, specific levels of the treatment are chosen *a priori* by the researcher for a specific reason.
This model tests the hypothesis that the τ_i values are all zero (or that the means of the groups are the same). An example would be the guinea pig vitamin C/orange juice experiment.

1.2 Model II Anova

Model II Anova

You will be happy to know that fitting a model II anova is exactly like a model I. The difference comes in the subsequent interpretation. The model can be written like this:

$$Y_{ij} = \mu + A_i + \epsilon_{ij}$$

where A_i is a random effect that results from treating the categorical variable as a random draw from some distribution of possible values.

- Instead of comparing means,
- assess the amount of variation in model due to A_i.
- This is generally called estimating *variance components*

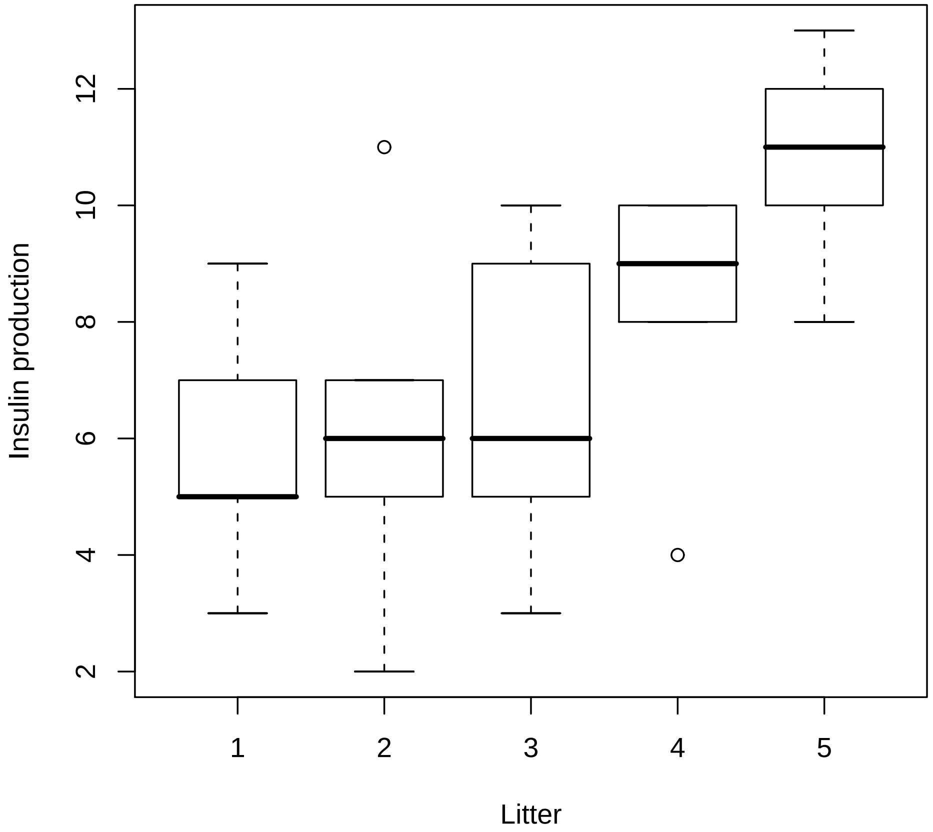
- H_0 : variation across groups is no greater than variation within groups

Mouse insulin

```
> mice

   litter insulin
1       1       9
6       1       7
11      1       5
16      1       5
21      1       3
2       2       2
7       2       6
12      2       7
17      2      11
22      2       5
3       3       3
8       3       5
13      3       9
18      3      10
23      3       6
4       4       4
9       4      10
14      4       9
19      4       8
24      4      10
5       5       8
10      5      10
15      5      12
20      5      13
25      5      11
```

Mouse data



Variance components

For a one-way anova, a sample table looks like this:

Source	df	SS	MS	Expected MS
Group	$k-1$	SS_G	$SS_G/(k-1)$	$\sigma^2 + n_0\sigma^2$
Error	$N-k$	SS_E	$SS_E/(N-k)$	σ^2

So the expected MS for groups equals the level of variance within groups plus an additional component equal to the average sample-size (n_o) times the among group variance.

Variance components II

Based on the table in the previous slide, it becomes possible to estimate the additional variance produced by groups. This can be called the addded component of variance.

$$s_A^2 = \frac{MS_G - MS_E}{n_0}$$

This can also be expressed as a percent of total variation. When dealing with an unbalanced design, n_0 is not a simple arithmetic mean. and can be given by this number.

$$n_0 = \frac{1}{k-1}\left(\Sigma^k n_i - \frac{\Sigma^k n_i^2}{\Sigma^k n_i}\right)$$

Analyzing mouse data

```
> mice.aov <- aov(insulin ~ litter, data = mice)
> mice.aov.summary <- summary(mice.aov)
> mice.aov.summary

            Df Sum Sq Mean Sq F value  Pr(>F)
litter       4  83.84   20.96  3.0733 0.03991 *
Residuals   20 136.40    6.82
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

It is possible to reject H_0: variation among litters is greater than that within litters. In other words, significant genetic variation seems to exist for insulin production.

Interpreting results

So you've set up your anova and you get a significant error term. Where do you go from there?

- Model I: you can then start to compare the means of your fixed effects (planned and unplanned post-hoc tests [next weeks topic])
- Model II: you can estimate how much variation is explained by a random factor (variance components)

What proportion of total variation?

What proportion of total variance is present among groups?

```
> mice.aov.summary

            Df Sum Sq Mean Sq F value  Pr(>F)
litter       4  83.84   20.96  3.0733 0.03991 *
Residuals   20 136.40    6.82
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

```
> var.due2.litter <- (20.96 - 6.82)/5
> var.due2.litter

[1] 2.828

> var.total <- 6.82 + var.due2.litter
> var.total

[1] 9.648

> prop.due2.litter <- var.due2.litter/var.total
> prop.due2.litter

[1] 0.2931177
```

Nested Anova

In instances where multiple observations are made within subgroups and the subgroups are located within groups, or the amount of variation at different levels of a heirarchy are desired, *Nested Anova* comes into play.

Subgroups are always random effects (model II) and groups at the highest level of hierarchy can be either random or fixed effects. If all levels of the hierarchy are random effects, then the model is called a *pure model II* anova. If the top level is model I, the model is known as a *mixed nested model* .

Nested Anova model

Source	df	SS	MS	Expected MS
Among Group	$k-1$	SS_G	$SS_G/(k-1)$	$\sigma^2 + n\sigma^2_{\text{subwithingrp}} + nb\sigma^2_{\text{grp}}$
Subgroups within Group	$k(b-1)$	SS_S	$SS_S/(k(b-1))$	$\sigma^2 + n\sigma^2_{\text{subwithingrp}}$
Error within subgroup	$kb(n-1)$	SS_E	$SS_E/kb(n-1)$	σ^2

Nested Anova Example

This example comes from S&R Box 10.1.

Wing-length was measured on mosquitos raised in three cages. Four mosquitos were chosen from each cage and two measurements were performed upon each mosquito wing.

```
> mosquito <- read.csv("nested.csv", header = T)
> names(mosquito)

[1] "cage"    "female"  "winglen"
```

Nested Anova Example cont.

```
> mosquito$cage <- as.factor(mosquito$cage)
> mosquito$female <- as.factor(mosquito$female)
> mosquito.aov <- aov(winglen ~ cage + female %in% cage, data = mosquito)
> summary(mosquito.aov)

            Df  Sum Sq Mean Sq F value    Pr(>F)
cage         2  665.68  332.84  255.70 1.452e-10 ***
cage:female  9 1720.68  191.19  146.88 6.981e-11 ***
Residuals   12   15.62    1.30
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

> f <- 332.84/191.19
> f

[1] 1.740886

> 1 - pf(f, 2, 9)

[1] 0.2295346
```

Alternative formula

It is also possible to specify the nested model with '/'

```
> summary(aov(winglen ~ cage/female, data = mosquito))

            Df  Sum Sq Mean Sq F value    Pr(>F)
cage         2  665.68  332.84  255.70 1.452e-10 ***
cage:female  9 1720.68  191.19  146.88 6.981e-11 ***
Residuals   12   15.62    1.30
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

Nested Anova Example cont.

In a pure model II anova (like this example), it is convenient to estimate variance components. Remember:

Source	df	SS	MS	Exp. MS
Among Group	$k-1$	SS_G	$SS_G/(k-1)$	$\sigma^2 + n\sigma^2_{\text{subwithingrp}} + nb\sigma^2_{\text{grp}}$
Subgroups within Group	$k(b-1)$	SS_S	$SS_S/(k(b-1))$	$\sigma^2 + n\sigma^2_{\text{subwithingrp}}$
Error within subgroup	$kb(n-1)$	SS_E	$SS_E/kb(n-1)$	σ^2

Estimating variance components II

```
> s2.grp <- (332.84 - 191.19)/(2 * 4)
> s2.subgrp <- (191.19 - 1.3)/2
> s2 <- 1.3
> pct.grp <- 100 * s2.grp/(s2.grp + s2.subgrp + s2)
> pct.subgrp <- 100 * s2.subgrp/(s2.grp + s2.subgrp + s2)
> pct.resid <- 100 * s2/(s2.grp + s2.subgrp + s2)
> print(paste("percentage of variance explained by group:", pct.grp))

[1] "percentage of variance explained by group: 15.5384429745176"

> print(paste("subgroup within group:", pct.subgrp))

[1] "subgroup within group: 83.3207182896195"

> print(paste("residual:", pct.resid))

[1] "residual: 1.14083873586292"
```

Pseudoreplication

Nested analyses are a nice way to think about pseudoreplication
You've seen how pseudoreplication could occur, this is the effect on the components of an anova table: Pseudoreplication is essentially a situation where the df used in a test are larger than they should be. This occurs when samples are non-independent, yet are treated as such.
Another way of thinking about it is that the error df are too large.

Unbalanced data

Unbalanced data complicate the calculation of variance components in nested designs. This is true both for hand-calculations and using R. We are not going down that road. If you need to calculate var components by hand look at Box 10.6 in S&R. If you want to use R, you will need to download and become familiar with the R library 'lme4' and the book by Pinerõ and Bates that documents it.