ANOVA, nested

One of the most frequent uses of a Nested ANOVA is to control statistically for extraneous, but potentially important, sources of variation. You should use a Nested ANOVA whenever there are additional sources of variation that could confound your interpretation of the main treatment effect. The simplest Nested ANOVA has two factors or treatments: a main effect A, of which there are a different levels. Levels of the second factor, B, are nested within each level of A. Factor B accounts for variation contributed by differences among sub-groups to which the same level of Factor A is applied (e.g. subgroups of individuals to which a drug is administered, duplicate aquaria with different sets of animals fed the same diet, replicate transects within study sites). There is a different randomly-selected set of b levels of Factor B within each level of Factor A. Each nested sub-group has n replicates. In most biological situations, Factor B is a random factor. Factor A may be either fixed or random, depending on the specific biological hypothesis being tested. For each individual observation:

$$x_{k(j(i))} = \mu + A_i + B(A)_{j(i)} + \epsilon_{k(j(i))}$$

where μ is the parametric mean of the population, A_i is the added effect of the ith Level of Factor A, $B(A)_{j(i)}$ is the effect due to the jth subgroup of B nested within the ith Level of Factor A, and $\epsilon_{k(j(i))}$ is the error or residual term. $B(A)_{j(i)}$ is a random variable with an Expected Value of zero and a variance $\sigma^2_{B(A)}$. Subscript notation describes the pattern of nesting.

Assumptions

1. The individual observations are independent and unbiased (i.e. the A_i's, $B(A)_{j(i)}$'s and $\epsilon_{k(j(i))}$'s are uncorrelated).
2. Within every nested group (B) within each level of Factor A, the $\epsilon_{k(j(i))}$'s are normally distributed with Expected Value = 0 and variance = σ^2_{error} (i.e. variances of all levels of B within each A have equal variances).
3. Factor B is (usually) Random (its levels are a small random sample from the entire population of all *possible* levels that could have been chosen for the experiment), E($\overline{B(A_i)}$) = 0 with variance $\sigma^2_{B(A)}$ for each level of A_i.
4. For a fixed effects model, $\sum^a A_i$; for a random effects model, E($\overline{A}$) = 0 with variance = σ^2_A.

There are two null hypotheses to test: all $B(A)_{j(i)}$'s = 0 (i.e., $\sigma^2_{B(A)} = 0$), and all A_i's = 0 or $\sigma^2_A = 0$.

Source	df	Sum of squares	Expected MS	F
Among A	a - 1	$bn\sum_{i=1}^{a}(\overline{x}_i - \overline{x})^2$	$\sigma^2_e + n\sigma^2_{B(A)} + \frac{bn}{a-1}\sum_{i=1}^{a}(A_i - \overline{A})^2$	$\frac{MS_A}{MS_{B(A)}}$
B within A	a(b-1)	$n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{x}_{j(i)} - \overline{x}_i)^2$	$\sigma^2_e + n\sigma^2_{B(A)}$	$\frac{MS_{B(A)}}{MS_{within}}$
Within	ab(n-1)	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{x}_{k(j(i))} - \overline{x}_{j(i)})^2$	σ^2_e	
Total	abn - 1	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{x}_{k(j(i))} - \overline{x})^2$		

Example: mussel recruitment on rocky intertidal shores

Mussels are not uniformly distributed in the rocky intertidal, but rather occur in the upper mid-zone. One possible explanation is that larvae settle selectively in the upper regions and avoid the lower regions. To test this (biological) hypothesis, we set up the following experiment. We chose three intertidal Heights (high, mid, & low). Within each Height zone we randomly chose 3 replicate Locations, separated horizontally by 20-30 m. The replicate Locations will tell us how much medium-scale spatial variation there is within the different Height Zones (we hope it's small, but we won't know for sure if we use just one Location at each Height). Within each Location, we attached 4 replicate plates to the rocks within ~1 m of each other. The plates were deployed in spring when mussel larvae are abundant in the water column. After 3 weeks, we collected all the plates and counted the newly settled mussels on each. The Locations are nested within Heights, since a given Location can be associated with one and only one Height. Here, Heights are a fixed factor and Locations-within-Heights are random.

The linear model is:

$$x_{k(j(i))} = \mu + Height_i + Location(Height)_{j(i)} + \epsilon_{k(j(i))}$$

where i is the subscript for Heights (i:1 to $a = 3$), j is the subscript for Locations within Heights (j:1 to $b = 3$), and k is the subscript for replicate plates in each Locations (k:1 to $n = 4$).

We create the dataset below. Although there are 3 locations (1,2,3) nested within each site, it is *always* advisable to create *unique* identifiers for your nested term.

```
library(tidyverse); library(broom)
dat <- tibble(
  shore_height = c(rep("high", 12), rep("mid", 12), rep("low", 12)),
  location = as.factor(rep(c(rep(1, 4), rep(2, 4), rep(3, 4)), 3)),
  mussel_n = c(30, 23, 30, 23, 18, 24, 29, 23, 25, 20, 19, 20,
               35, 30, 35, 25, 34, 31, 28, 27, 26, 39, 33, 29,
               28, 24, 31, 29, 42, 35, 37, 37, 33, 43, 32, 35),
  loc = paste(shore_height, location, sep = ""))
```

We run the linear model using `lm`. Note that the syntax for nested terms in `R` is sometimes expressed as "`shore_height / loc`", or "`shore + shore_height:loc`" (e.g., using `lmer` for mixed-effects models) - but in our case it does not matter because we are going to calculate the F and P 'by hand'. Why? Because by default, `lm` uses the MS_{error} as the denominator for all F-ratios. But in a nested ANOVA, the appropriate denominator for the F-ratio of a term is the MS of the term immediately below it. In the code below, we use the function `lead` to achieve this goal.

```
m1 <- lm(mussel_n ~ shore_height + loc, data = dat)
m1_df <- tidy(anova(m1)); m1_df # default output from `lm`
m1_df <- mutate(m1_df, f = meansq / lead(meansq), # calculate f, p 'by hand'
                p = pf(q = f, df1 = df, df2 = lead(df), lower.tail = FALSE)); m1_df
```

Source	df	Sum of squares	Mean square	F	P
shore_height	2	660.667	330.333	7.095	0.026
loc	6	279.333	46.556	2.775	0.031
Residuals	27	453.000	16.778	NA	NA

Compare the output of `m1_df` before and after you create the proper F and P values. How did they change? Try using the nested syntax described above - the results should not change, but prove it to yourself. What can we conclude about the 'Height' and 'Location' effects?

Note that the df, SS, MS are calculated appropriately. It is only the F, P that we cannot trust - check that these values are correct whenever using any 'black-box' ANOVA software (and there are lots of options in `R`).

We can use the function `aov` below, but note that the syntax is a bit wonky, and frankly, we might as well just calculate F, P ourselves - because `aov` does not calculate P for the nested term.

```
aov1 <- aov(mussel_n ~ shore_height + Error(loc), data = dat); summary(aov1)
```

On your own

- Check the assumptions of the nested ANOVA (plot the data, residuals).
- Run the nested model with `location` instead of `loc`. What happens when you don't use a unique nested identifier for each location? Specifically - what happens to the df? Why?
- Run a one-way ANOVA without the nested term. How do df, SS, MS, F, P change?

References

Underwood. 1997. Experiments in Ecology.

Prepared by R Elahi, J Watanabe