ANOVA Models

Advanced Biostatistics

Dean Adams

Lecture 3

EEOB 590C

ANOVA (Analysis of Variance): Y~X

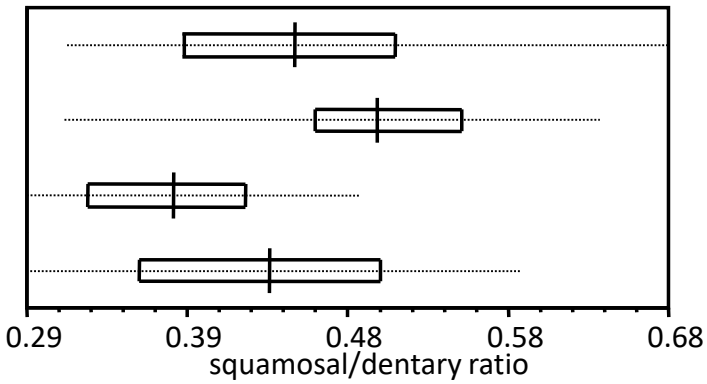
-X is categorical (groups), Y is continuous

H_0: no relationship between X & Y (i.e., no difference among groups)

-Compare variation between groups to variation within groups (e.g., are males & female different in height?)

Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (μ is grand mean, α_i is i^{th} group mean, and ε_{ij} is error)

-Standard test statistic: F-ratio (ratio of variances)



$$F \approx \frac{\sigma^2_{btwngroups}}{\sigma^2_{withingroups}}$$

1: Independence: ε_{ij} of variates must be independent

-SOME reasons for violation: spatial or temporal autocorrelation (variates close in space or time are similar).
-Autocorrelation test: Serial Independence (if data not independent, model non-independence in error variance)

2: Normality: requires normally distributed ε_{ij}

-Test distribution of deviates from mean for skewness, kurtosis (also other methods).
-If not normal, transform, use nonparametric methods, or randomization

3: Homoscedasticity: equal variance

-Test: F-tests, Bartlett's test, F-max test.
-If heteroscedastic, transform or use nonparametric (NOTE: ANOVA models fairly robust to homoscedasticity violations)

ANOVA based on partitioning Total Sums-of-Squares SST

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - \overline{\overline{Y}}\right)^2$$

SST=SSM+SSE (M = model; E = error)

Variance components and F-ratio obtained

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{df}SS$$

$$F \approx \frac{\sigma^2_{btwngroups}}{\sigma^2_{withingroups}}$$

If F-ratio = 1.0, no difference in variance ($\sigma^2_b=\sigma^2_w$)

SST can be partitioned into various components (for more complex designs)

Model contains single grouping variable (e.g., species or sex)

Procedure

Y ~ A

-Obtain overall and group means

-Estimate SSM (and σ^2=MSM) of group means vs. overall mean

-Estimate SSE (and σ^2=MSE) of individuals vs. their group mean

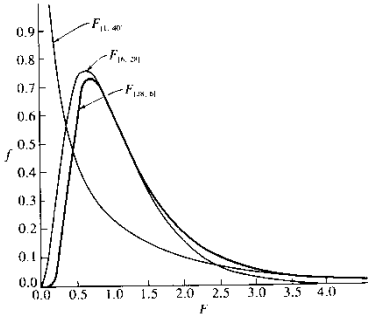
-Calculate F and assess significance (df1,df2)

Source	df	SS	MS=σ	F	P
Group	a-1	$SSM = \sum_{i=1}^{a} n_i \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$	SSM/df	MSM/MSE	
Error	$\sum n_i - a$	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_i\right)^2$	SSE/df		
Total	n-1	$SST = \sum_{i=1}^{a}\sum_{j=1}^{n} \left(Y_{ij} - \bar{\bar{Y}}\right)^2$			

*Note: SSE for can be obtained as SSE = SST – SSB

Standard approach:

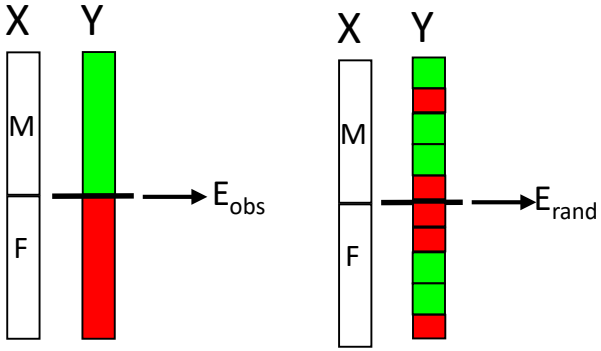
-Compare F-ratio to F-distribution with appropriate df

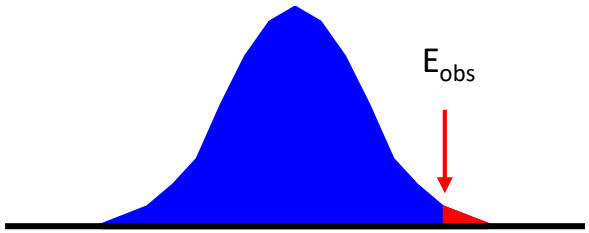


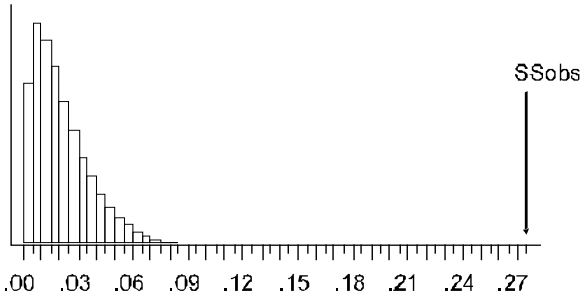
Resampling Alternative:

-Shuffle individuals among groups and generate distribution of possible

F-values (note: for single-factor ANOVA, shuffling individuals IS a residual randomization)







For 2 groups ($n_1=n_2$), ANOVA will yield identical results to t-test

$$\mathbf{X} = [m \quad m \quad m \quad m \quad m \quad f \quad f \quad f \quad f \quad f]$$
$$\mathbf{Y} = [5 \quad 4 \quad 4 \quad 4 \quad 3 \quad 7 \quad 5 \quad 6 \quad 6 \quad 6]$$

```
> t.test(y~x)

t = 4.4721, df = 8, p-value = 0.002077

> anova(lm(y~x))

             Df Sum Sq Mean Sq F value   Pr(>F)
x             1     10    10.0      20 0.002077 **
```

NOTE: For this case, one can derive F-ratio from the t-statistic

$$t^2 = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right)^2}{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \frac{n\left(\overline{Y}_1 - \overline{Y}_2\right)^2}{2s_p^2}$$

b/c $n_1=n_2=n$ rewrite as:

Numerator:

$$n\left(\overline{Y}_1 - \overline{Y}_2\right)^2 = \left(n\left(\overline{Y}_1 - \overline{\overline{Y}}\right)^2 + n\left(\overline{Y}_2 - \overline{\overline{Y}}\right)^2\right) = \sum n\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2$$

ANOVA identifies group differences, but not WHICH ONES

Pairwise multiple comparisons required (1 vs 2, 1 vs 3, etc)

MANY approaches (most derived from t-statistic)

-Example: Fishers' LSD (Least Significant Difference)

-Calculate $t_{obs} = \left|\bar{Y}_i - \bar{Y}_j\right|$

-Compare to smallest significant standardized difference:

$$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Other approaches: T', Tukey-Kramer, GT2, Duncan's multiple range, Scheffe's F-test

(choice depends on = or ≠ sample sizes, etc.)

Resampling method for pairwise comparisons

-Estimate group means

-Calculate matrix of D_{euclid}

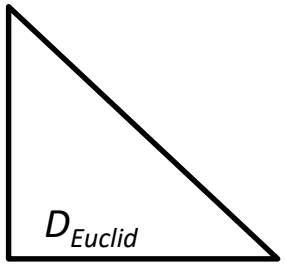
-Shuffle specimens into groups

-Estimate means and D_{rand}

-Assess D_{obs} vs. D_{rand}

-Repeat

$$\mathbf{D}_{obs} = \sqrt{\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j\right)^t \left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j\right)}$$



e.g., Adams 2004. *Ecology*

Multiple testing of data increases chance of 'finding' significance, so adjust critical-α for # comparisons (k)

Various approaches

Dunn-Sidak: $\alpha' = 1-(1-\alpha)^{1/k}$

Bonferroni: $\alpha' = \frac{\alpha}{k}$

Sequential Bonferroni: , $\alpha_k^1 = \frac{\alpha}{k}$ but recalculated each time (k=1,2,3. . .) for ORDERED comparisons

Bonferroni considered quite conservative: sequential Bonferroni much less so

After a bad winter storm (Feb. 1, 1898), Bumpus retrieved 136 sparrows in Rhode Island (about ½ died)

Collected the following measurements on each:

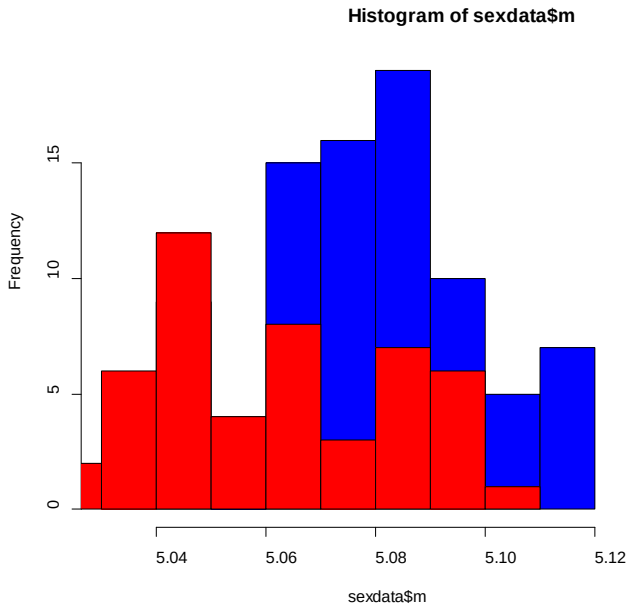
1) Alive/dead
2) Weight
3) Total length
4) Wing extent
5) Beak-head length
6) Humerus
7) Femur
8) Tibiotarsal
9) Skull
10) Keel-sternum
11) male/female

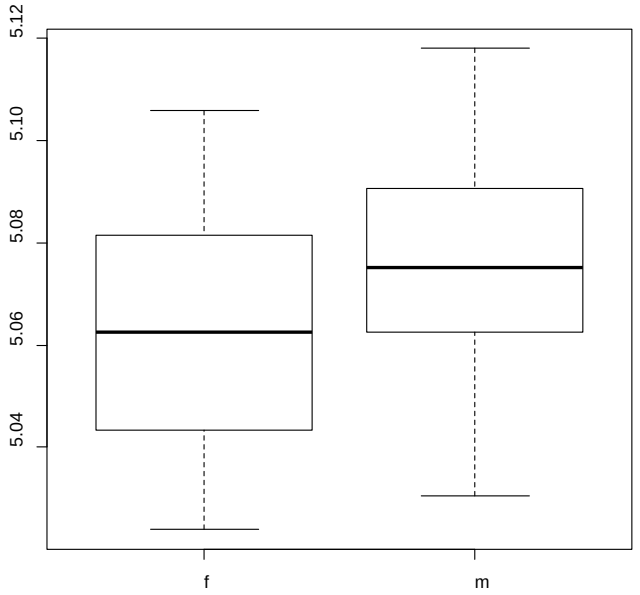
To investigate natural selection, examined whether there was a difference in alive vs. dead birds

Bumpus, H. C. 1898. *Woods Hole Mar. Biol. Sta.*

Single-factor ANOVA

```
> anova(lm(TotalLength~sex))
               Df   Sum Sq   Mean Sq F value    Pr(>F)
sex             1 0.007460 0.0074597  16.667 7.617e-05 ***
Residuals     134 0.059975 0.0004476
```





Males larger than females

ANOVA with 4 groups (male-alive, male-dead, female-alive,female-dead)

```
> gp<-as.factor(paste(sex,surv))
> anova(lm(TotalLength~gp))
             Df   Sum Sq   Mean Sq F value    Pr(>F)
gp            3 0.014515 0.0048382  12.068 4.957e-07 ***
```

Pairwise comparisons*

```
> pairwise.t.test(TotalLength, gp, p.adj = "none")

        f FALSE f TRUE  m FALSE
f TRUE  0.259   -       -
m FALSE 1.2e-05 3.5e-07 -
m TRUE  0.260   0.024   9.1e-05
```

Via Randomization (D_{euclid} below, P above diagonal)

	Fem Dead	Fem Surv	Male Dead	Male Surv
Fem Dead	0	0.330 NS	**0.002**	0.328 NS
Fem Surv	0.0066	0	**0.001**	**0.042**
Male Dead	0.0229	0.0295	0	**0.002**
Male Surv	0.0053	0.0118	0.0176	0

*Shows P-values only

ANOVA multiple X variables (e.g, species AND sex)

Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

Y ~ A + B

In simple case, SSA & SSB are independent of one another, and are thus calculated as before: e.g. $SSA = \sum_{i=1}^{a} n_i \left(\bar{Y}_i - \bar{\bar{Y}} \right)^2$

Source	df	SS	MS	F	P
Factor A	a-1	SSA	SSA/df	MSA/MSE	
Factor B	b-1	SSB	SSB/df	MSB/MSE	
Error	df_{tot}-(df_A+df_B)	SSE	SSE/df		
Total	n-1	SSA+SSB+SSE			

*Note: SSE IS affected by factors (SSE = SST –(SSA+SSB)

Think of data for Factorial ANOVA as table of reactions

	Factor B (gp 1)	Factor B (gp 2)
Factor A (gp1)	Responses AB_{11}	Responses AB_{12}
Factor A (gp 2)	Responses AB_{21}	Responses AB_{22}

Find means of rows, columns, and individual cells, and comparison of these yields tests for Factors A & B

NOTE: This ignores the interaction (A:B)

Interactions measure the joint effect of main effects A & B
Identifies whether response to A dependent on level of B

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

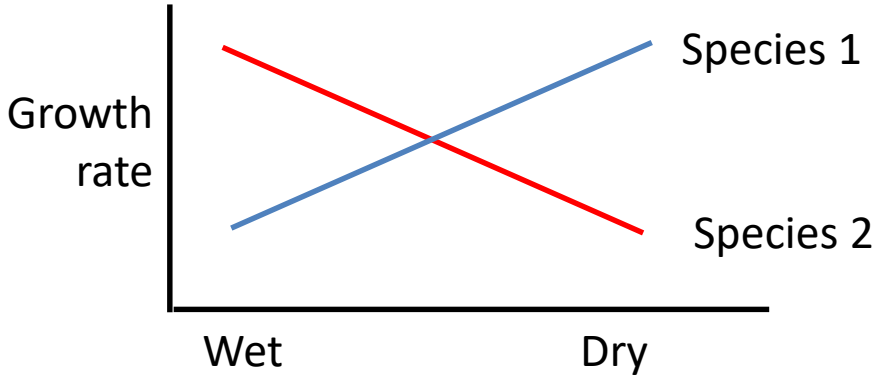
Y ~ A + B+A:B

Estimated as:

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\left(\bar{Y}_{AB} - \bar{Y}_i - \bar{Y}_j + \bar{\bar{Y}}\right)^2 = n_{AB}\sum_{AB=1}^{rc}\left(\bar{Y}_{AB} - \bar{\bar{Y}}\right)^2 - SSA - SSB$$

Significant interaction: main effects not interpretable without clarification (e.g., species 1 larger than species 2 ONLY in wet environments…)

VERY common in biology



Note: The study of trade-offs (reaction norms) in evolutionary ecology is based on the study of interactions

Other factors can be added for general factorial designs (e.g., 3-way factorial with interaction)

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

Y ~ A + B+C+A:B+A:C+B:C+A:B:C

Adding factors also means adding many interactions

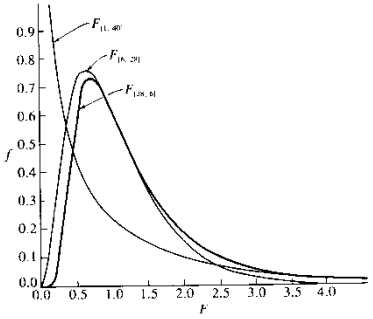
In general, assess significance of interactions, pool non-significant interaction Mean Square with MSE**, then interpret main effects whenever possible

Inclusion of more factors is 'neater', and allows assessment of **interactions**, but requires larger N

**Rules for pooling MSA*B with MSE depend upon significance, the type of effect (fixed or random), and several other factors (See Box 10.3 in *Biometry* for details)

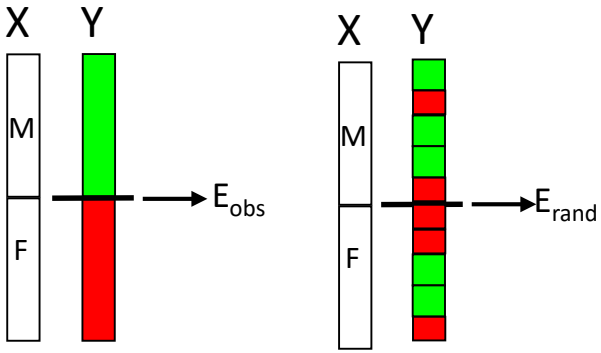
Standard approach:

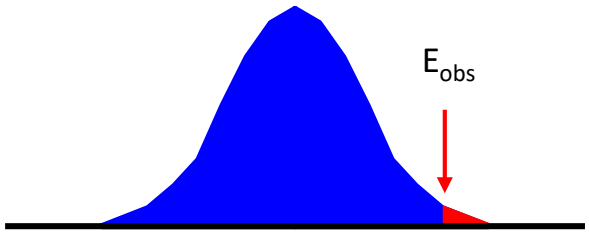
-Compare F-ratio to F-distribution with appropriate df

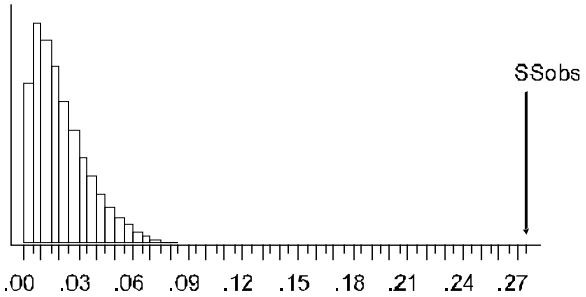


Resampling Alternative:

-Residual randomization: shuffle residuals from reduced model to assess that factor (e.g., remove A×B and use randomization to test A×B factor)







Bumpus data with two factors (sex, surv):

```
> anova(lm(TotalLength~sex*surv))
              Df   Sum Sq   Mean Sq F value    Pr(>F)
sex            1 0.007460 0.0074597 18.6069 3.122e-05 ***
surv           1 0.006121 0.0061212 15.2683 0.0001483 ***
sex:surv       1 0.000934 0.0009337  2.3289 0.1293803
```

*NOTE: no significant interaction

Previous:

	Fem Dead	Fem Surv	Male Dead	Male Surv
Fem Dead	0	0.330 NS	**0.002**	0.328 NS
Fem Surv	0.0066	0	**0.001**	**0.042**
Male Dead	0.0229	0.0295	0	**0.002**
Male Surv	0.0053	0.0118	0.0176	0

Pairwise comparisons

Via Randomization (D_{euclid} below, P above diagonal)

	Fem Dead	Fem Surv	Male Dead	Male Surv
Fem Dead	0	0.883 NS	0.154 NS	0.427 NS
Fem Surv	0.0066	0	0.625 NS	0.875 NS
Male Dead	0.0229	0.0295	0	0.179 NS
Male Surv	0.0053	0.0118	0.0176	0

***NOTE: significance of pairwise tests depends *heavily on null model used for permutation!!*

TRADITIONAL STATISTICAL VIEW

For standard approaches, evaluating factors (via F-tests) requires knowledge of *expected mean squares (EMS)*

-EMS describe the sources of variation for the effects in the model, and the sources against which effects are evaluated

FIXED EFFECTS (model I ANOVA): assumes differences among groups are due to treatment effects determined (FIXED) by the investigator (e.g., compare 3 drugs)

RANDOM EFFECTS (model II ANOVA): differences among groups are due to treatment effects and a random component NOT fixed by the investigator. Groups are chosen as a *random* sample from a larger population (e.g., compare variation among families: the variation is of interest, not the chosen families)

TRADITIONAL STATISTICAL VIEW

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

FIXED EFFECTS: Var of Y_{ij} = variance of μ + variance of α_i + variance of ε_{ij}

But since α are fixed, expected $k\sigma^2_\alpha = 0$

- Thus, EMS_{trt}= variance of ε_{ij}

= σ^2_ε

RANDOM EFFECTS: Var of Y_{ij} = variance of μ + variance of α_i + variance of ε_{ij}

EMS_{trt}= variance of α_i + variance of ε_{ij}

= $k\sigma^2_\alpha + \sigma^2_\varepsilon$ (EMS contains variation among treatments)

TRADITIONAL STATISTICAL VIEW

Determining whether a factor is fixed or random is tricky but *can be* important

This must be determined prior to final assembly and analysis of ANOVA table, because EMS is not the same (thus MS comparisons in F-tests will be different)

-Consider whether you have 'set' the levels of a factor in experiment, or have chosen 'representatives'

-Then consider what other levels may or may not exist

-Determine EMS for various factors

-When all else fails…

CONSULT STATISTICIAN AND STATISTIC BOOKS!!

Some factors are not independent, but are hierarchical
Levels of B 'nested' within levels of A:

Example: Compare CO_2 production among habitats (A) measured from multiple trees (B)

Trees logically nested within habitats

Note: B1 in each group is NOT the same! It is a place-holder (e.g., 1st tree sampled)

Nested factors ALWAYS RANDOM effects!

*NOTE: no interaction of A:B obtained when nested terms are used!!!
-because SS(A/B) = SSB + SSA:B

Model:

SSA: $Y_{ijk} = \mu + \alpha_i + \beta_j(\alpha_i) + \varepsilon_{ijk}$ (reflects summation across levels of B)

SS(A/B): $SSA = nb\sum_{i=1}^{a}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2$ ('subgroups' of A)

NOTE: SS for Nested term [SS(A/B)] is sum of SSB + SSA:B

$$SS(A/B) = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_j - \overline{Y}_i\right)^2$$

Source	df	SS	MS		F	P
Factor A	a-1	SSA	SSA/df		**MSA/MS(A/B)**	
Factor A/B	a(b-1)	SS(A/B)	SS(A/B)/df		**MS(A/B) /MSE**	
Error	ab(n-1)	SSE	SSE/df			
Total	n-1	SSA+SS(A/B)+SSE				

NOTE: because of nested relationship and because B is a random effect, MSA tested vs. MS(A/B)

Determining independent vs. nested can be tricky. Consider:

1: Is there correspondence of levels across factors, or do they represent sub-divisions?

2: Is second factor fixed/random?

3: Is interaction term meaningful?

Example 1: Is sex nested within species, or is it an independent factor?

SEX IS AN INDEPENDENT FACTOR:

Levels of sex correspond across species

Sex is fixed effect

Sex:species interaction term meaningful

Example 2: Is tree nested within habitat (for CO_2 experiment), or is it an independent factor?

TREE COULD BE NESTED FACTOR:

Levels of tree not common across habitats

Trees are random effects (trees selected as representatives of the sample)

Tree:habitat interaction not biologically meaningful

Because SS can be partitioned, much more complicated ANOVA models are possible (e.g., 5 way factorial with 2 nested effects, some fixed some random…)

Often, researchers are prohibited by $$$ and cannot measure replicates for all levels of various treatments

Various experimental designs exist to extract SS for factors in incomplete data (e.g., latin-squares, randomized complete block, randomized incomplete block, etc.)

Maximize the power of analysis vs. sampling effort, **PROVIDED THEY ARE SET UP CORRECTLY** (i.e. are considered a priori)

SEE STATISTICS Dept. for HOW to generate such designs

For balanced data sets, everything above holds

For unbalanced data, things are more complicated

Different factor SS can be obtained with different ORDER of factors in the model

```
> anova(lm(TotalLength~sex+surv))
             Df   Sum Sq   Mean Sq F value    Pr(>F)
sex           1 0.007460 0.0074597  18.423 3.383e-05 ***
surv          1 0.006121 0.0061212  15.117 0.0001588 ***

> anova(lm(TotalLength~surv+sex))
             Df   Sum Sq   Mean Sq F value    Pr(>F)
surv          1 0.004127 0.0041266  10.191  0.001761 **
sex           1 0.009454 0.0094544  23.349 3.666e-06 ***
```

Why is this the case?

Difference due to calculations of type I vs. type III SS

Type I (sequential) SS: calculated in order found in model

Estimate SS for effect given model up to that point

e.g., $SSA = A\left(\beta_A \mid \overline{\overline{Y}}\right)$ then find $SSB = B\left(\beta_B \mid \beta_A, \overline{\overline{Y}}\right)$

Type III (marginal) SS: calculated relative to all other model effects

All terms in model and estimate effect SS by removing it

e.g., $SSA = A\left(\beta_A \mid \beta_B, \overline{\overline{Y}}\right)$ $SSB = B\left(\beta_B \mid \beta_A, \overline{\overline{Y}}\right)$

Type I SS can change depending on order, while Type III do not

Type I SS remain additive (they sum to SSModel) while Type III do not

For balanced designs, Type I = Type III

Type III: Issues

1: H_0 and H_a illogical for type III:

-Obtain SS_A by doing $SS_{Full} - SS_{Full\ w.out\ A}$

-Obtain SS_B by doing $SS_{Full} - SS_{Full\ w.out\ B}$

-Obtain $SS_{A:B}$ by doing $SS_{Full} - SS_{Full\ w.out\ A:B}$

BUT what sense is this model: Y~A + A:B (ie, how can one have an interaction without the main effect???)

2: SS from Type III VIOLATE original property described by Fisher

SS partitioning: SST= SSA+SSB, etc. ***NOT*** the case with type III

```
> anova(lm(TotalLength~paste(sex,surv))) # total SSModel: A, B, A:B
Analysis of Variance Table

Response: TotalLength
                Df   Sum Sq   Mean Sq F value    Pr(>F)
paste(sex, surv)  3 0.014515 0.0048382  12.068 4.957e-07 ***
Residuals       132 0.052920 0.0004009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> sum(anova(lm(TotalLength~sex*surv))[[2]][1:3])  #TYPE I
[1] 0.01451466
> sum(drop1(aov(TotalLength~sex*surv),.~.)[[2]][2:4]) #TYPE III
[1] 0.00973504
```

DCA perspective: Type I preferred (though it forces user to think about H_0 H_a!).

Some software (SAS, JMP, Minitab) use Type III
Other packages (R, S-plus, Mathematica) use Type I
Philosophical debate amongst statisticians

Venables argues logic behind type III SS nonsensical:

> "The objection to Type III sums of squares is that they encourage naive users to do silly things such as test main effects in the presence of interactions, without really asking whether the test makes sense or not, that is, whether it really addresses a question of any interest."
>
> "There is, by any sensible reckoning, only ONE type of sum of squares, and it always represents an improvement **sum of squares of the outer (or alternative) model over the inner (or null hypothesis) model. What the SAS highly dubious classification of sums of squares does is to encourage users to concentrate on the null hypothesis model and to forget about the alternative."**

Translation: H_a matters, so think CAREFULLY about your model before pushing buttons!

$Y \sim A + B + A:B$

This *IS* a set of sequential hypotheses tests (Type I SS)

Model	Test	H_0
Y ~ A	Does A explain more variation than the grand mean?	**Y ~ 1**
Y ~ A + B	Does B|A explain more variation than A alone?	**Y ~ A**
Y ~ A + B + A:B	Does the interaction help in explaining more variation than A and B separately?	**Y ~ A + B**

Requires appropriate resampling procedure with exchangeable units for each H_0

Residual randomization is the most appropriate for factorial models

Anderson and ter Braak 2003. *J. Stat. Comput. Simul.*
Collyer and Adams 2007. *Ecology.*
Collyer et al. 2015. *Heredity.*

Sequentially permute Y_{resid} from reduced model ($H_{0.r}$) with fewer terms
Tests SS for H_1, while holding constant SS in corresponding H_0

Procedure

1: Estimate parameters and observed test statistic (E_{obs}) from full model
2: Remove one term from the model; calculate predicted values ($\hat{\mathbf{Y}}$) and residuals (ε) for the reduced model
3: Permute residuals (ε) and add them to the predicted values to obtain randomized values
4: Calculate test statistic for random data (E_{rand})
5: Repeat *i* times to obtain empirical distribution of statistic

Higher statistical power for factorial designs (Anderson and terBraak, 2003)
Tests appropriate H_0 for each term in model (Collyer, Sekora, and Adams 2015)
Uses correct exchangeable units for hypothesis testing (Collyer, Sekora, and Adams 2015)

NOTE: Using permutation, the concepts of EMS are more clear. One is explicitly describing what hypothesis is tested against which.

(DCA: using permutation, traditional concepts such as EMS, type of SS, etc. come clearly into focus. Type I SS requires clear thinking of H_0 & H_a, so hypothesis testing is rigorously defined).

ANOVA an incredibly flexible tool for testing hypotheses

Can add factors, based on theory of partitioning SS

Type I retains that theory, type III does not

Permutations are incredibly useful

Think critically about H_0 & H_a: BOTH matter!