MB590-012
Microbiome Analysis
ANOVA & Permutation Tests

Dr. Christine Hawkes

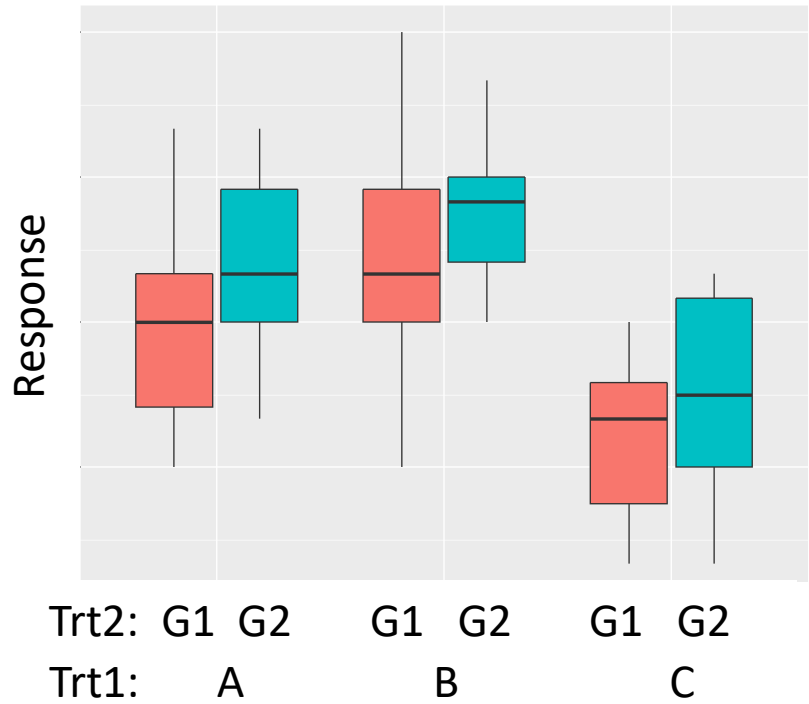
NC STATE UNIVERSITY

Today’s outline

- Review of factorial designs and ANOVA
- Microbiome data and permANOVA
- Overview of today’s datasets
- How to build ANOVA/permANOVA tables

Factorial designs

- Factors or categorical variables
- Levels selected to test hypotheses
- Can be natural or controlled



General linear models

- OLS regression from last week
- Analysis of variance (ANOVA)

- Use ordinary least squares approach to estimate model parameters
- Assume residuals are normally distributed, no correlation between samples, constant variance

- How to apply to high dimensional data?
  - Permutation ANOVA (permANOVA) with GLS – today

Analysis of Variance (ANOVA) & hypothesis testing

- Linear model of the form:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \ldots + \varepsilon_{ijk}$$

  - Y_{ijk} = individual outcome
  - μ = grand mean; average over all individuals
  - α_i, β_j = treatment effects; average over all individuals in groups i, j
  - $\alpha\beta_{ij}$ = interaction effect; average over all individuals in groups ij
  - ε_{ijk} = random errors associated with individual k (residuals)

- Testing hypothesis (H1) against null (H0)
  - H1 = All groups do not have the same mean value ($\mu_{trt} \neq \mu_{control}$)
  - H0 = All groups have the same mean value ($\mu_{trt} = \mu_{control}$)

ANOVA Parameters

- Degrees of Freedom (df)
  - number of independent observations in the data that are free to vary as parameters are estimated
- Sums of Squares (SS)
  - sum of squared differences from the mean
  - total SS = treatment SS + error SS
  - treatment SS compares group mean to grand mean for treatment ("between")
  - error SS compares individual responses to group mean ("within")
- Mean Squares (MS)
  - ratio of SS to df ("average SS")
  - describes the variability within treatments
  - MS-error estimates variation in residual errors around group means
- F statistic
  - ratio of MS treatment to MS error
  - tests whether variability between group means is larger than the variability of the observations within the groups

SS Types

- Type I = Sequential
  - Fits models according to the order of terms entered
  - Not appropriate for factorial designs
  - Sometimes used to remove effects of specific terms first (e.g., nested factors)
- Type II = Hierarchical
  - Tests each model term after all other model terms
  - Assumes no interactions
  - Not appropriate for factorial designs (although sometimes used for unbalanced designs)
- Type III = Partial or Orthogonal
  - Model terms are tested in light of every other term
  - Includes interactions
  - Appropriate for balanced factorial designs

Fixed and Random Effects

- Fixed Effects
  - Factors whose levels are experimentally determined or where interest lies in the specific effects of each level
  - If experiment was repeated, levels would be the same

- Random Effects
  - Factors whose levels are sampled from a larger population, or where interest lies in the variation among them rather than at specific levels
  - If experiment was repeated, specific levels would vary
  - Accounting for random effects allow us to better account for variation <u>within</u> groups in order to test for differences <u>among</u> treatments

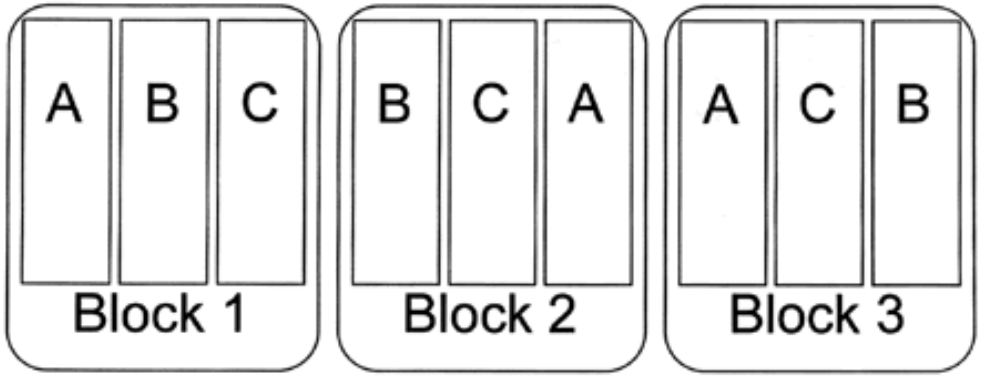
Fixed Effects ANOVA

Hypothesis Testing!

	Fixed or Random	df	SS	MS	F-ratio
A	Fixed	(a-1)	SS_A	SS_A/df_A	MS_A/MS_e
B	Fixed	(b-1)	SS_B	SS_B/df_B	MS_B/MS_e
A x B		(a-1)(b-1)	SS_{AB}	SS_{AB}/df_{AB}	MS_{AB}/MS_e
Residual (error)	Random	ab(n-1)	SS_e	SS_e/df_e	
TOTAL		abn-1			

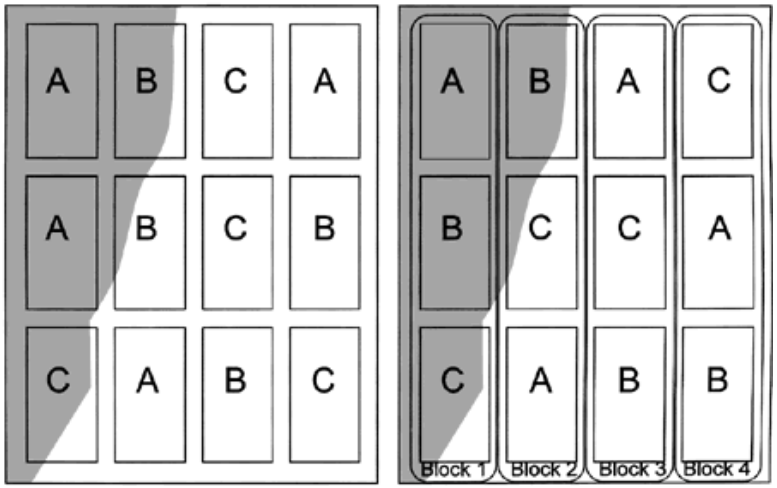
Random Effects - Blocks

- Block effects are those that apply equally to all individuals within a group, leading to a single level of correlation within groups
- Blocks are typically created to account for random variation when a fully randomized design cannot be applied
- Randomized complete block design (RCBD) is common in agriculture, where complete set of treatments is randomized in every block and there are no within-block replicates



Random Effects – Block examples

- Spatial
  - Resource gradients
  - Animals in a pen/cage
- Temporal
  - Planting, harvesting
  - Feeding, sampling
- Experimental artifacts
  - Individuals collecting data
  - Sample runs on equipment



Mixed Effects ANOVA

	Fixed or Random	df	SS	MS	F-ratio
A	Fixed	(a-1)	SS_A	SS_A/df_A	MS_A/MS_{AB}
B	Random	(b-1)	SS_B	SS_B/df_B	MS_B/MS_e
A x B		(a-1)(b-1)	SS_{AB}	SS_{AB}/df_{AB}	MS_{AB}/MS_e
Residual	Random	ab(n-1)	SS_e	SS_e/df_e	
TOTAL		abn-1			

Other Common Random Effects

- Nested
  - Random effects that are hierarchically structured
  - Hierarchical structure means certain groups of lower-level factor only found in certain groups of higher-level factor
- Repeated measures
  - Multiple measures of the same subject made over time
- Both end up further partitioning the within-groups SS

	Ecotypes			
	A	B	C	D
Species1	1	1	0	0
Species2	0	0	1	1

	Date			
	J	**F**	**M**	**A**
Subject1	5	5	4	3
Subject2	2	2	2	2
Subject3	4	5	6	7

Mixed Effects ANOVA with single Nested Random Effect

	Fixed or Random	df	SS	MS	F-ratio
A	Fixed	(a-1)	SS_A	SS_A/df_A	$MS_A/MS_{B(A)}$
B(A)	Random	(b-1)a	SS_{AB}	SS_{AB}/df_{AB}	$MS_{B(A)}/MS_e$
Residual	Random	ab(n-1)	SS_e	SS_e/df_e	
TOTAL		abn-1			

Microbiome Data

- High *p* (OTUs) relative to *n* (samples)
- Typically violates assumptions of: linearity, normality, heteroscedasticity, error independence
- Invalidates parametric tests and the *F*-distribution for hypothesis testing
- Requires nonparametric tests (& transformation) via
  - Permutation tests – resampling approach
  - Generalized linear models – likelihood approach

permANOVA with RRPP::lm.rrpp

- Non-parametric, handles cases where #ASVs>>#samples
- Dependent var can be raw data or distance matrix (uses inter-observation distances)
- Type I, II, or III SS
- OLS or GLS models
  - OLS assumes no correlation between samples and constant variance
  - GLS (or weighted least squares) modifies OLS by accounting for inequality of variance across groups
  - GLS in RRPP requires you to provide a covariance matrix (i.e., matrix giving covariance between each pair of elements – see R::cov)
  - We will focus on OLS today to save on computation time
- For mixed models, specify MS error terms for each parameter (otherwise assumes residuals)

Collyer & Adams 2018 Methods Eco Evo doi: 10.1111/2041-210X.13029

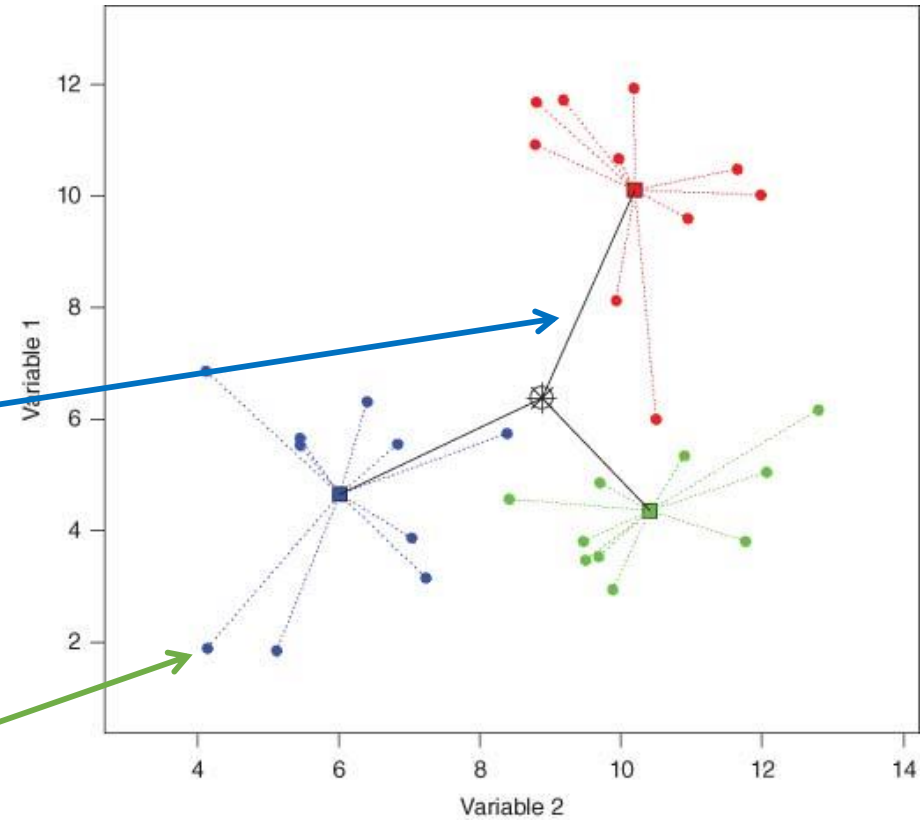
RRPP Resampling Procedures

- Calculate SS, MS, F from inter-observation distances in multivariate space

$SS_T = SS_A + SS_R$

SS_A = among-group (treatments)
= sum of the squared distances from individual group centroids to the overall centroid (solid black lines)

SS_R = residual (within-group)
= sum of the squared distances to centroids from individual sampling units (replicates) to their own group centroid (colored dotted lines)



Anderson 2017 Stat Ref https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841

RRPP Resampling Procedures

- Repeatedly randomizes residuals from null models to create a null distribution for comparison to full models (null+effect)
- Resample by shuffling data
  - Raw data: Resample rows
  - Distance matrix: Resample rows and columns jointly
- Can control # iterations with rrpp::iter
- Compare F stat to distribution of resampled F statistics

	ASV2	ASV3	ASV4	ASV5	ASV6	ASV7	ASV8	ASV9	ASV10
M1551P81	22	65	13679	92	1980	5123	7079	82	79
M1551P29	2	4	2496	183	781	1278	2699	449	0
M1551P90	2	70	2428	292	2273	401	2870	90	14
M1551P48	59	41	1323	305	2570	498	5123	40	76
M1551P52	3	2	6919	2	13	29029	33	29	14
M1551P31	5	472	1597	439	1158	525	3039	75	706
M1551P77	7	33	1175	297	6852	259	3353	1068	301
M1551P37	3	90	2750	509	3796	950	5668	35	80

	P81	P29	P90	P48	P52	P31	P77	P37
P81	1							
P29	0.6	1						
P90	0.0	0.9	1					
P48	0.6	0.6	0.2	1				
P52	0.6	0.6	0.8	0.6	1			
P31	0.7	0.5	0.2	0.8	0.7	1		
P77	0.0	0.8	0.7	0.2	0.3	0.9	1	
P37	0.0	0.8	0.6	0.8	1.0	0.4	0.4	1

Anderson 2017 Stat Ref https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841

permANOVA with RRPP::lm.rrpp

```
otu.rrpp <- RRPP::lm.rrpp(OTU ~ A*B,
                data = SAM,
                SS.type="III",
                seed="random",
                iter=1000)

anova(otu.rrpp, effect.type = "F",
     error = c("A:B", "Residuals", "Residuals"))
```

permANOVA Alternatives in R

Distance-based methods

- ANOSIM – sensitive to heterogeneity of variance
- adonis2 – limited model specification options

Model-based methods

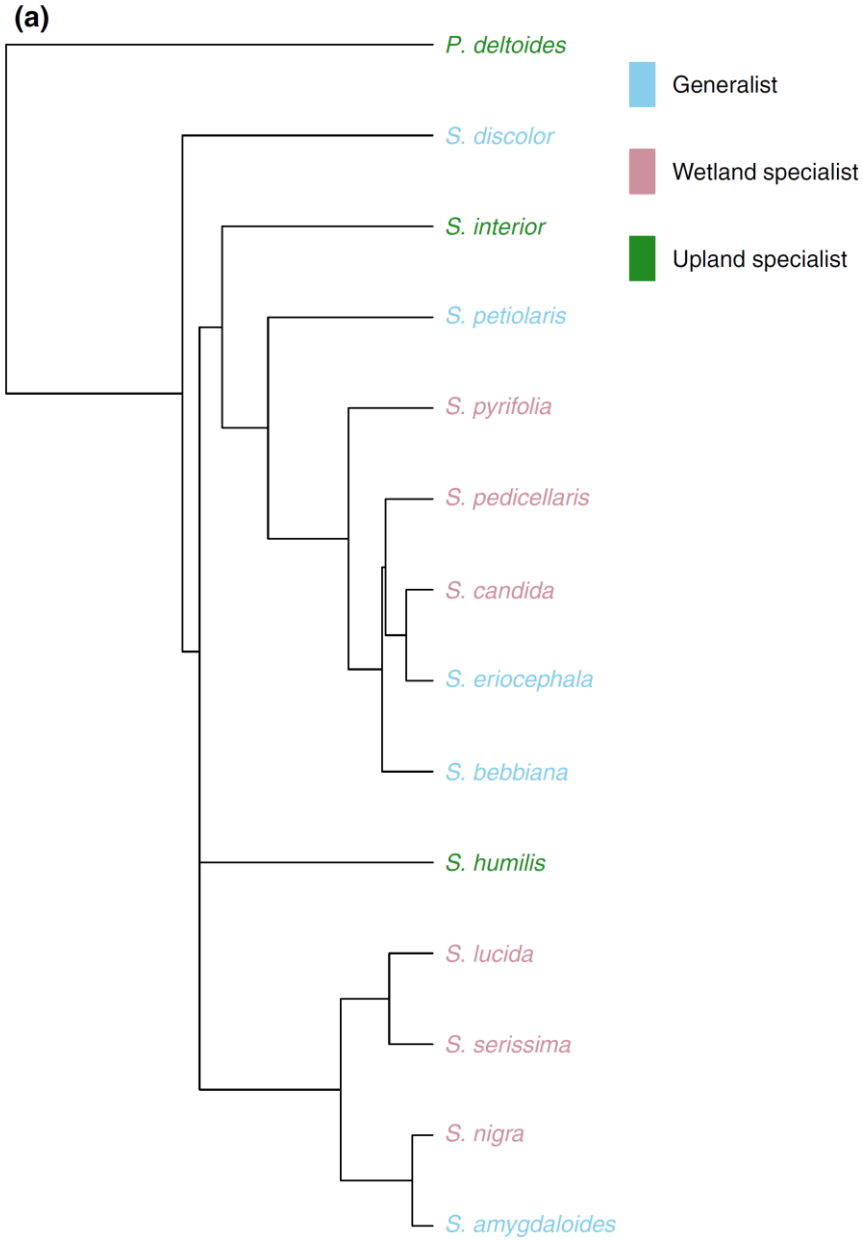
- mvabund –fits separate GLMs to each species; useful for unbalanced designs; LASSO penalty; no mixed model except via multi-model comparisons; computationally intensive
- gllvm – uses latent variables

See Collyer & Adams 2018 Appendix 3 Table S1 for more details

Today's data – Erland et al. 2017

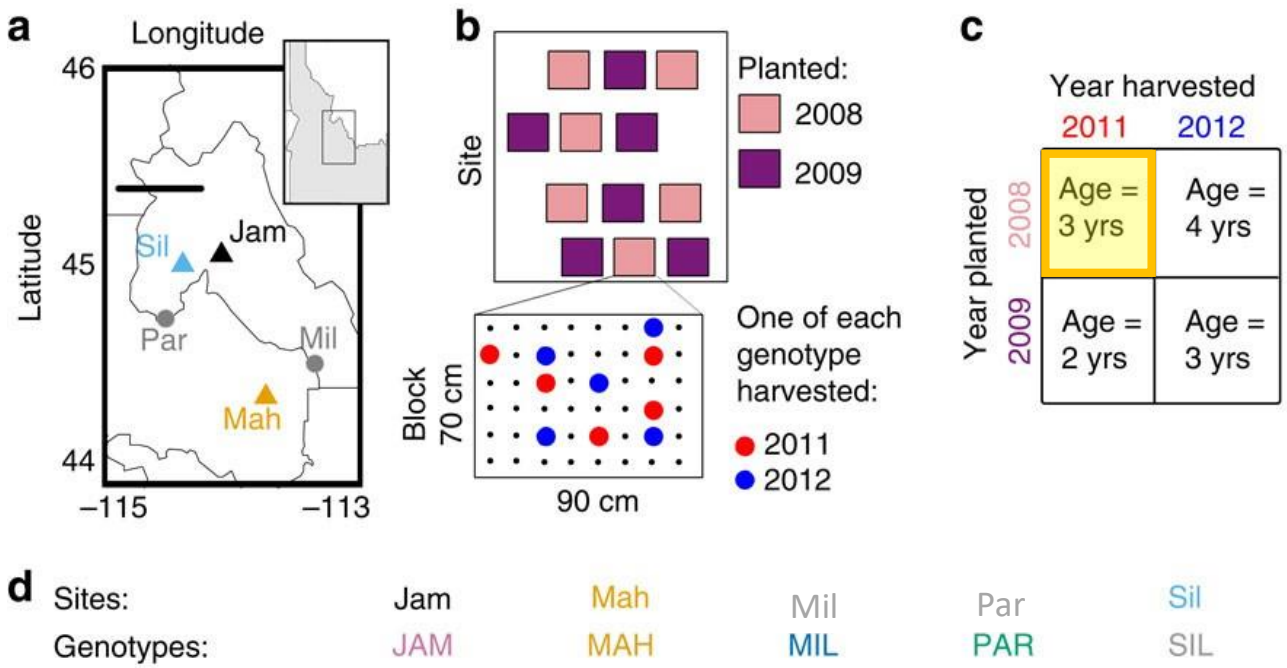
DOI:10.1111/mec.14576

- **Treatment** – upland vs. lowland common gardens (fixed)
- **Spp** – 14 willow species (random)
- **Plot** – a spatial block for treatment (random); (note: unbalanced, might actually be nested in Garden.Location)



Today's coding exercise data

- Wagner et al. 2016 Nature Communications (also wk3) https://www.nature.com/articles/ncomms12151
- Root and leaf samples (Type, n=2)
- Genotypes (n=5)
- Sites (n=5)
- Block nested in Site (n=6 per site)
- Limited to one cohort and one experiment



https://www.nature.com/articles/ncomms12151

How to build the ANOVA table?

	Source →	**T**	**S**	**P**	**e**
	Fix or Rand →	F	R	R	R
	Levels →	a	b	c	n
	Subscript→	i	j	k	l
Source					
$\mathbf{T_i}$	(a-1)				
$\mathbf{S_j}$	(b-1)				
$\mathbf{P_k}$	(c-1)				
$\mathbf{TS_{ij}}$	(a-1)(b-1)				
$\mathbf{e_{l(ijk)}}$	abc(n-1)				

- First, set up table with each factor in the model and residual error
  - Treatment (T), Spp (S), and Plot (P), Treatment*Spp (TS), Error (e)
- For each factor indicate:
  - Fixed or random
  - Levels
  - Subscripts for replication
- Calculate df
  - num levels for subscripts inside () * num levels -1 for subscripts outside ()

Underwood (1997) Experiments in Ecology. Cambridge University Press
Montgomery (1976) Design and Analysis of Ecological Experiments. John Wiley & Sons

Rule 1: If term in row has column's subscript and

(a) column subscript is not in brackets (not nested)

(i) enter 0 if column subscript represents a *fixed* factor

(ii) enter 1 if column subscript represents a *random* factor

(b) column subscript is in brackets (nested) enter 1

	Source →	**T**	**S**	**P**	**e**
	Fix or Rand →	F	R	R	R
	Levels →	a	b	c	n
	Subscript→	i	j	k	l
Source					
$\mathbf{T_i}$	(a-1)	0			
$\mathbf{S_j}$	(b-1)		1		
$\mathbf{P_k}$	(c-1)			1	
$\mathbf{TS_{ij}}$	(a-1)(b-1)	0	1		
$\mathbf{e_{l(ijk)}}$	abc(n-1)	1	1	1	1

Rule 1: If term in row has column's subscript and

(a) column subscript is not in brackets (not nested)

(i) enter 0 if column subscript represents a *fixed* factor

(ii) enter 1 if column subscript represents a *random* factor

(b) column subscript is in brackets (nested) enter 1

Rule 2: if term in row does not have column's subscript, enter the # levels of the factor

	Source →	**T**	**S**	**P**	**e**
	Fix or Rand →	F	R	R	R
	Levels →	a	b	c	n
	Subscript→	i	j	k	l
Source					
$\mathbf{T_i}$	(a-1)	0	b	c	n
$\mathbf{S_j}$	(b-1)	a	1	c	n
$\mathbf{P_k}$	(c-1)	a	b	1	n
$\mathbf{TS_{ij}}$	(a-1)(b-1)	0	1	c	n
$\mathbf{e_{l(ijk)}}$	abc(n-1)	1	1	1	1

Rule 3: for each row, identify components that belong in the MS (rows that share the subscript)
Rule 4: multiply each component by the product of all row entries that represent that component, omitting columns with that subscript
Rule 5: in the residual row, the multiplier is always 1

	Source →	T	S	P	e	MS Estimate
	Fix or Rand →	F	R	R	R	
	Levels →	a	b	c	n	
	Subscript→	i	j	k	l	
Source						
T_i	(a-1)	0	b	c	n	$\sigma^2_e + cn\sigma^2_{TS} + bcn\sigma^2_T$
S_j	(b-1)	a	1	c	n	
P_k	(c-1)	a	b	1	n	
TS_{ij}	(a-1)(b-1)	0	1	c	n	
$e_{l(ijk)}$	abc(n-1)	1	1	1	1	

Rule 3: for each row, identify components that belong in the MS (rows that share the subscript)
Rule 4: multiply each component by the product of all row entries that represent that component, omitting columns with that subscript
Rule 5: in the residual row, the multiplier is always 1

	Source →	**T**	**S**	**P**	**e**	**MS Estimate**
	Fix or Rand →	F	R	R	R	
	Levels →	a	b	c	n	
	Subscript→	i	j	k	l	
Source						
$\mathbf{T_i}$	(a-1)	0	b	c	n	$\sigma^2_e + cn\sigma^2_{TS} + bcn\sigma^2_T$
$\mathbf{S_j}$	(b-1)	a	1	c	n	$\sigma^2_e + acn\sigma^2_S$
$\mathbf{P_k}$	(c-1)	a	b	1	n	$\sigma^2_e + abn\sigma^2_P$
$\mathbf{TS_{ij}}$	(a-1)(b-1)	0	1	c	n	$\sigma^2_e + cn\sigma^2_{TS}$
$\mathbf{e_{l(ijk)}}$	abc(n-1)	1	1	1	1	

Finally identify the MS error term for F-ratio denominator as the term that *contains all the other components except the target factor itself (and the residual error)*

	Source →	**T**	**S**	**P**	**e**	**MS Estimate**	**F-ratio denom**
	Fix or Rand →	F	R	R	R		
	Levels →	a	b	c	n		
	Subscript→	i	j	k	l		
Source							
$\mathbf{T_i}$	(a-1)	0	b	c	n	$\sigma^2_e + cn\sigma^2_{TS} + bcn\sigma^2_T$	TS
$\mathbf{S_j}$	(b-1)	a	1	c	n	$\sigma^2_e + acn\sigma^2_S$	Residual
$\mathbf{P_k}$	(c-1)	a	b	1	n	$\sigma^2_e + abn\sigma^2_P$	Residual
$\mathbf{TS_{ij}}$	(a-1)(b-1)	0	1	c	n	$\sigma^2_e + cn\sigma^2_{TS}$	Residual
$\mathbf{e_{l(ijk)}}$	abc(n-1)	1	1	1	1		

Let’s practice! Switch to html