Worksheet 6 - Multifactor ANOVA models

Multifactor ANOVA

- Quinn & Keough (2002) - Chpt 9

Question 1 - Nested ANOVA - one between factor

In an unusually detailed preparation for an Environmental Effects Statement for a proposed discharge of dairy wastes into the Curdies River, in western Victoria, a team of stream ecologists wanted to describe the basic patterns of variation in a stream invertebrate thought to be sensitive to nutrient enrichment. As an indicator species, they focused on a small flatworm, Dugesia, and started by sampling populations of this worm at a range of scales. They sampled in two seasons, representing different flow regimes of the river - Winter and Summer. Within each season, they sampled three randomly chosen (well, haphazardly, because sites are nearly always chosen to be close to road access) sites. A total of six sites in all were visited, 3 in each season. At each site, they sampled six stones, and counted the number of flatworms on each stone.

Format of curdies.csv data files

SEASON	SITE	DUGESIA	S4DUGESIA
WINTER	1	0.648	0.897
..	..	..	..
WINTER	2	1.016	1.004
..	..	..	..
WINTER	3	0.689	0.991
..	..	..	..
SUMMER	4	0	0
..	..	..	..

Each row represents a different stone

SEASON	Season in which flatworms were counted - fixed factor
SITE	Site from where flatworms were counted - nested within SEASON (random factor)
DUGESIA	Number of flatworms counted on a particular stone
S4DUGESIA	4th root transformation of DUGESIA variable

Open the curdies data file.

The SITE variable is supposed to represent a random factorial variable (which site). However, because the contents of this variable are numbers, R initially treats them as numbers, and therefore considers the variable to be numeric rather than categorical. In order to force R to treat this variable as a factor (categorical) it is necessary to first **convert this numeric variable into a factor** (HINT).

Notice the data set - each of the nested factors is labelled differently - there can be no replicate for the random

(nesting) factor.

Q1-1. What are the main hypotheses being tested?

a. H_0 Effect 1:

b. H_0 Effect 2:

Q1-2. In the table below, list the assumptions of nested ANOVA along with how violations of each assumption are diagnosed and/or the risks of violations are minimized.

Assumption	Diagnostic/Risk Minimization
I.	
II.	
III.	

Q1-3. Check these assumptions (HINT). Note that for the effects of SEASON (Factor A in a nested model) there are only three values for each of the two season types. Therefore, boxplots are of limited value! Is there however, any evidence of violations of the assumptions (HINT)? (Y or N) ☐
If so, assess whether a transformation will address the violations (HINT) and then make the appropriate corrections (HINT).

Q1-4. For each of the tests, state which error (or residual) term (state as Mean Square) will be used as the nominator and denominator to calculate the F ratio. Also include degrees of freedom associated with each term.

Effect	Nominator (Mean Sq, df)	Denominator (Mean Sq, df)
SEASON		
SITE		

Q1-5. If there is no evidence of violations, test the model;
S4DUGES = SEASON + SITE + CONSTANT
using a nested ANOVA (HINT). Fill (HINT) out the table below, make sure that you have treated SITE as a random factor when compiling the overall results.

Q1-6. For each of the tests, state which error (or residual) term (state as Mean Square) will be used as the nominator and denominator to calculate the F ratio. Also include degrees of freedom associated with each term.

Source of variation	df	Mean Sq	F-ratio	P-value
SEASON				
SITE				
Residuals				

Normally, we are not interested in formally testing the effect of the nested factor to get the correct F test for the nested factor (SITE), examine a representation of the anova table of the fitted linear model that assumes all factors are fixed (HINT)

Q1-7. What are your conclusions (statistical and biological)?

Q1-8. Where is the major variation in numbers of flatworms? Between (seasons, sites or stones)?

Q1-9. How might this information influence the design of future experiments on Dugesia in terms of:

a. What influences the abundance of Dugesia

b. Where best to focus sampling effort to maximize statistical power?

Q1-10. Finally, construct an appropriate **graph** to accompany the above analyses. Note that this should use the correct replicates for depicting error (HINT). **Save the graph as a jpeg image** and **import the graph into word**.

Question 2 - Two factor ANOVA

A biologist studying starlings wanted to know whether the mean mass of starlings differed according to different roosting situations. She was also interested in whether the mean mass of starlings altered over winter (Northern hemisphere) and whether the patterns amongst roosting situations were consistent throughout winter, therefore starlings were captured at the start (November) and end of winter (January). Ten starlings were captured from each roosting situation in each season, so in total, 80 birds were captured and weighed.

C(DIST,poly,2)

Format of starling.csv data files

SITUATION	MONTH	MASS	GROUP
S1	November	78	S1Nov

..	..	..	..
S2	November	78	S2Nov
..	..	..	..
S3	November	79	S3Nov
..	..	..	..
S4	November	77	S4Nov
..	..	..	..
S1	January	85	S1Jan
..	..	..	..

SITUATION Categorical listing of roosting situations
MONTH Categorical listing of the month of sampling.
MASS Mass (g) of starlings.
GROUP Categorical listing of situation/month combinations - used for checking ANOVA assumptions

Open the starling data file.

Q2-1. List the 3 null hypothesis being tested

I.

II.

III.

Q2-2. **Test the assumptions** by producing **boxplots** (HINT) and **mean vs variance plot**.

a. Is there any evidence that one or more of the assumptions are likely to be violated? (Y or N) _____

b. Is the proposed model **balanced?** (Yor N) _____

Q2-3. Now **fit a two-factor ANOVA model** (HINT) and **examine the residuals** (HINT). Any evidence of skewness or unequal variances? Any outliers? Any evidence of violations? ('Y' or 'N') _____.
Examine the ANOVA table and fill in the following table:

Source of Variation	SS	df	MS	F-ratio	Pvalue
SITUATION					

MONTH					
SITUATION : MONTH					
Residual (within groups)					

Q2-4. An **interaction plot (plot of means)** is useful for summarizing multi-way ANOVA models. Summarize the trends using an **interaction plot** (HINT).

Q2-5. In the absence of an interaction, we can examine the effects of each of the main effects in isolation. It is not necessary to examine the effect of MONTH any further, as there were only two groups. However, if we wished to know which roosting situations were significantly different to one another, we need to perform additional **multiple comparisons**. Since we don't know anything about the roosting situations, no one comparison is any more or less meaningful than any other comparisons. Therefore, a Tukey's test is most appropriate. Perform a **Tukey's test** (HINT) and summarize indicate which of the following comparisons were significant (put * in the box to indicate P< 0.05, ** to indicate P< 0.001, and NS to indicate not-significant).

Situation 1 vs Situation 2

Situation 1 vs Situation 3

Situation 1 vs Situation 4

Situation 2 vs Situation 3

Situation 2 vs Situation 4

Situation 3 vs Situation 4

Q2-6. Generate a bargraph to summarize the findings of the above Tukey's test.

Q2-7. Summarize your conclusions from the analysis.

Question 3 - Two factor ANOVA - Type II SS

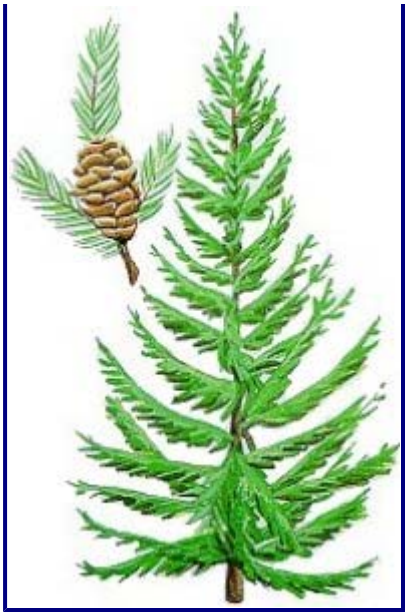
Here is a modified example from Quinn and Keough (2002). Stehman and Meredith (1995) present data from an experiment that was set up to test the hypothesis that healthy spruce seedlings break bud sooner than diseased spruce seedlings. There were 2 factors: pH (3 levels: 3, 5.5, 7) and HEALTH (2 levels: healthy, diseased). The dependent variable was the average (from 5 buds) bud emergence rating (BRATING) on each seedling. The sample size varied for each combination of pH and health, ranging from 7 to 23 seedlings. With two factors, this experiment should be analyzed with a 2 factor (2 x 3) ANOVA.

Format of stehman.csv data files

PH	HEALTH	GROUP	BRATING
3	D	D3	0.0
..	..	..	..
3	H	H3	0.8

..	..	..	..
5.5	D	D5.5	0.0
..	..	..	..
5.5	H	H5.5	0.0
..	..	..	..
7	D	D7	0.2
..	..	..	..

PH	Categorical listing of pH (not however that the levels are numbers and thus by default the variable is treated as a numeric variable rather than a factor - we need to correct for this)
HEALTH	Categorical listing of the health status of the seedlings, D = diseased, H = healthy
GROUP	Categorical listing of pH/health combinations - used for checking ANOVA assumptions
BRATING	Average bud emergence rating per seedling



Open the stehman data file.

The variable PH contains a list of pH values and is supposed to represent a factorial variable. However, because the contents of this variable are numbers, R initially treats them as numbers, and therefore considers the variable to be numeric rather than categorical. In order to force R to treat this variable as a factor (categorical) it is necessary to first **convert this numeric variable into a factor** (HINT).

Q3-1. **Test the assumptions** by producing **boxplots** and **mean vs variance plot**.

a. Is there any evidence that one or more of the assumptions are likely to be violated? (Y or N) _____

b. Is the proposed model **balanced?** (Yor N) _____

Q3-2. Now **fit a two-factor ANOVA model** and **examine the residuals**. Any evidence of skewness or unequal variances? Any outliers? Any evidence of violations? ('Y' or 'N') _____. As the model is not balanced, we will base hypothesis tests on Type II sums of squares. Produce an ANOVA table (HINT) and fill in the following table:

Source of Variation	SS	df	MS	F-ratio	Pvalue
PH					
HEALTH					
PH : HEALTH					
Residual (within groups)					

Q3-3. Summarize these trends using a **interaction plot**.

Q3-4. In the absence of an interaction, we can examine the effects of each of the main effects in isolation.

It is not necessary to examine the effect of HEALTH any further, as there were only two groups. However, if we wished to know which pH levels were significantly different to one another, we need to perform additional **multiple comparisons**. Since no one comparison is any more or less meaningful than any other comparisons, a Tukey's test is most appropriate. Perform a **Tukey's test** and summarize indicate which of the following comparisons were significant (put * in the box to indicate P< 0.05, ** to indicate P< 0.001, and NS to indicate not-significant).

pH 3 vs pH 5.5 ☐

pH 3 vs pH 7 ☐

pH 5.5 vs pH 7 ☐

Q3-5. Generate a bargraph to summarize the findings of the above Tukey's test.

Q3-6. Summarize your biological conclusions from the analysis.

Q3-7. Why aren't the 5 buds from each tree true replicates? Given this, why bother observing 5 buds, why not just use one?

Question 4 - Two factor ANOVA

An ecologist studying a rocky shore at Phillip Island, in southeastern Australia, was interested in how clumps of intertidal mussels are maintained. In particular, he wanted to know how densities of adult mussels affected recruitment of young individuals from the plankton. As with most marine invertebrates, recruitment is highly patchy in time, so he expected to find seasonal variation, and the interaction between season and density - whether effects of adult mussel density vary across seasons - was the aspect of most interest.

The data were collected from four seasons, and with two densities of adult mussels. The experiment consisted of clumps of adult mussels attached to the rocks. These clumps were then brought back to the laboratory, and the number of baby mussels recorded. There were 3-6 replicate clumps for each density and season combination.

Format of quinn.csv data files

SEASON	DENSITY	RECRUITS	SQRTRECRUITS	GROUP
Spring	Low	15	3.87	SpringLow
..	..	..	..	..
Spring	High	11	3.32	SpringHigh
..	..	..	..	..
Summer	Low	21	4.58	SummerLow
..	..	..	..	..
Summer	High	34	5.83	SummerHigh
..	..	..	..	..
Autumn	Low	14	3.74	AutumnLow
..	..	..	..	..

SEASON Categorical listing of Season in which mussel clumps were collected independent variable

DENSITY	Categorical listing of the density of mussels within mussel clump independent variable
RECRUITS	The number of mussel recruits response variable
SQRTRECRUITS	Square root transformation of RECRUITS - needed to meet the test assumptions
GROUPS	Categorical listing of Season/Density combinations - used for checking ANOVA assumptions

Open the quinn data file.

Confirm the need for a square root transformation, by examining **boxplots** and **mean vs variance plots** for both raw and transformed data. Note that square root transformation was selected because the data were counts (count data often includes values of zero - cannot compute log of zero).

Also confirm that the design (model) is **unbalanced** and thus warrants the use of Type II sums of squares. (HINT)

Q4-1. Now **fit a two-factor ANOVA model** (using the square-root transformed data and **examine the residuals**. Any evidence of skewness or unequal variances? Any outliers? Any evidence of violations? ('Y' or 'N') _____

☐. Produce an anova table based on Type II SS (HINT) and fill in the following table:

Source of Variation	SS	df	MS	F-ratio	Pvalue
SEASON	☐	☐	☐	☐	☐
DENSITY	☐	☐	☐	☐	☐
SEASON : DENSITY	☐	☐	☐	☐	☐
Residual (within groups)	☐	☐	☐		

Q4-2. Summarize these trends using a **interaction plot**. Note that graphs do not place the restrictive assumptions on data sets that formal analyses do (since graphs are not statistical analyses). Therefore, it data transformations were used for the purpose of meeting test assumptions, it is usually better to display raw data (non transformed) in graphical presentations. This way readers can easily interpret actual values in a scale that they are more familiar with.

Q4-3. The presence of a significant interaction means that we cannot make general statements about the effect of one factor (such as density) in isolation of the other factor (e.g. season). Whether there is an effect of density depends on which season you are considering (and vice versa). One way to clarify an interaction is to analyze subsets of the data. For example, you could examine the effect of density separately at each season (using four, single factor ANOVA's), or analyze the effect of season separately (using two, single factor ANOVA's) at each mussel density.

For the current data set, the effect of density is of greatest interest, and thus the former option is the most interesting. Perform the **simple main effects anovas**.

a. Was the effect of DENSITY on recruitment consistent across all levels of SEASON? (Y or N) _____

b. How would you interpret these results?

Welcome to the end of Worksheet 6!