08_homework_factorial_anova

Bill Perry

```r
library(skimr)
library(patchwork)
library(janitor)
```

```


Attaching package: 'janitor'
```

```
The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
library(readxl)
library(car)
```

```
Loading required package: carData
```

```r
library(broom)
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.2     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.1.0
```

```
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
✖ dplyr::recode() masks car::recode()
✖ purrr::some()   masks car::some()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```r
theme_set(theme_light())
```

Assignment Overview

This homework assignment analyzes crayfish growth data from Sargent and Lodge (2014) to examine differences in growth rates between native and invasive populations of rusty crayfish (*Orconectes rusticus*) across different lake environments using two-way ANOVA.

Learning Objectives

By completing this assignment, you will be able to:

1. **Understand two-way ANOVA concepts and applications**
2. **Perform exploratory data analysis for factorial designs**
3. **Conduct two-way ANOVA analysis**
4. **Test statistical assumptions for ANOVA**
5. **Interpret main effects and interactions**
6. **Create publication-quality figures**
7. **Write scientific methods and results sections**

Data Description

The dataset contains growth measurements from a common garden experiment where young-of-year (YOY) rusty crayfish from native (Ohio) and invasive (Wisconsin) populations were grown in enclosures in three northern Wisconsin lakes during summer 2011.

Key variables:

- - `range`: Population origin (Native vs Invasive)
- - `lake`: Lake location (Big, High, Papoose)
- - `growth_per_day`: Daily growth rate (mm/day) - response variable
- - `initial_length`: Starting length (mm)
- - `final_length`: Ending length (mm)
- - `days`: Duration of experiment

Part 1: Data Loading and Preparation

1.1 Load and Clean the Data

Part 2: Statistical Analysis Setup

2.1 Analysis Type and Model

Type of Analysis: Two-way factorial ANOVA

Model Equation: $\text{Growth Rate} = \mu + \text{Range} + \text{Lake} + (\text{Range} \times \text{Lake}) + \varepsilon$

Where:

- - Growth Rate = daily growth rate (mm/day)
- - Range = population origin (Native vs Invasive)
- - Lake = lake environment (Big vs High vs Papoose)
- - Range × Lake = interaction between range and lake
- - ε = random error

Hypotheses:

- *Main Effect*
  - *- Range:*
    - - H_0: μ_Native = μ_Invasive (no difference in growth between ranges)
    - - H_1: μ_Native ≠ μ_Invasive (difference exists between ranges)
  - *- Lake:*
    - - H_0: μ_Big = μ_High = μ_Papoose (no difference among lakes)
    - - H_1: At least one lake mean differs from others
- *Interaction Effect:*
  - - H_0: No interaction between Range and Lake
  - - H_1: Interaction exists between Range and Lake

Variables:

- - Response: growth_per_day (continuous, mm/day)
- - Factor 1: range (categorical, 2 levels: Native, Invasive)
- - Factor 2: lake (categorical, 3 levels: Big, High, Papoose)

Part 3: Exploratory Data Analysis

3.1 Summary Statistics

```
# cray_df %>% group_by(range) %>%
# skim()
```

3.2 Exploratory Visualizations

Part 4: Two-Way ANOVA Analysis

4.1 Fit the ANOVA Model

Car ANOVA for unbalanced

```
# Anova(????, type = 3)
```

4.2 Effect Sizes

```
# # note the code here is provided as it can be a mess....
# # name of datframe and such may need modificaiton....
#
# eta_squared_df <- cray_df %>%
#   group_by(range, lake) %>%
#   summarise(mean_growth = mean(growth_per_day), .groups = "drop") %>%
#   ungroup()
#
# total_ss <- sum((cray_df$growth_per_day - mean(cray_df$growth_per_day))^2)
#
# anova_summary <- summary(crayfish_anova_model)
# range_ss <- anova_summary[[1]]["range ", "Sum Sq"]
# lake_ss <- anova_summary[[1]]["lake", "Sum Sq"]
# interaction_ss <- anova_summary[[1]]["range:lake", "Sum Sq"]
#
# eta_squared_range <- range_ss / total_ss
# eta_squared_lake <- lake_ss / total_ss
# eta_squared_interaction <- interaction_ss / total_ss
#
# cat("Eta-squared (effect sizes):\n")
# cat("Range:", round(eta_squared_range, 3), "\n")
# cat("Lake:", round(eta_squared_lake, 3), "\n")
# cat("Range × Lake:", round(eta_squared_interaction, 3), "\n")
```

Eta-squared represents the **proportion of total variance explained** by the factor (range).

- Formula: η^2 = SS_between / SS_total
- Range: 0 to 1
- Interpretation: If $\eta^2 = 0.21$, then 21% of the variance in growth rate is explained by population range

Omega-squared is a **less biased estimate** of effect size than eta-squared.

- Formula: ω^2 = (SS_between - df_between × MS_within) / (SS_total + MS_within)
- Range: 0 to 1 (but can be slightly negative)
- More conservative than η^2 because it adjusts for bias in small samples

Why Calculate Both?

- **Eta-squared (η^2)**: Easier to calculate and interpret, but slightly **overestimates** effect size
- **Omega-squared (ω^2)**: More accurate, **unbiased estimate** of population effect size

Effect Size Interpretation Guidelines:

```
Effect Size     η² / ω²       Interpretation
Small           0.01          1% of variance explained
Medium          0.06          6% of variance explained
Large           0.14          14% of variance explained
```

Example Output Interpretation:

If your results show:

```
  eta_squared omega_squared
1        0.21          0.20
```

This means:

- **21%** of the variance in crayfish growth rate is explained by population range (η^2)
- **20%** is the unbiased estimate of variance explained (ω^2)
- This represents a **large effect size** (much larger than 0.14)
- Population range is a strong predictor of growth rate

Bottom line: Both metrics tell you how much of the differences in crayfish growth can be attributed to whether they're from native vs. invasive populations, with omega-squared being the more conservative (and accurate) estimate.

4.3 Post-hoc Tests

```
# Estimated marginal means for main effects
```

```
# Pairwise comparisons with Sidak correction
# Estimated marginal means for interaction effect
```

```
# Compact letter display for interaction effect
```

```
# Custom interaction plot using emmeans results
# emmeans_interaction_df <- as.data.frame(emmeans_interaction)
#
# emmeans_interaction_plot <- emmeans_interaction_df %>%
#   ggplot(aes(x = lake, y = emmean, color = range, group = range)) +
#   geom_point(size = 3, position = position_dodge(width=0.2)) +
#   geom_line(size = 1, position = position_dodge(width=0.2)) +
#   geom_errorbar(aes(ymin = lower.CL, ymax = upper.CL),
#                 width = 0.1, size = 1,
#                 , position = position_dodge(width=0.2)) +
#   labs(x = "Lake",
#        y = "Estimated Marginal Mean Growth Rate (mm/day)",
#        color = "Range",
#        title = "Estimated Marginal Means with 95% Confidence Intervals") +
#   scale_color_manual(values = c("Native" = "coral", "Invasive" = "steelblue")) +
#   theme_classic()
#
# emmeans_interaction_plot
```

Part 5: Assumption Testing

5.1 Check ANOVA Assumptions

```
# # Create diagnostic plots
# par(mfrow = c(2, 2))
# plot(crayfish_anova_model)
# par(mfrow = c(1, 1))
```

5.2 Formal Assumption Tests

Part 6: Publication Figure

6.1 Create Publication-Quality Figure

Submission Guidelines

What to turn in -

1. a quarto markdown file and dataframe if you modified the original. All of the code should be able to run with what you turn in. **(2 points)**

2. a self-contained html file showing the code and output **(2 points)**

3. annotations in the quarto file that shows or tells what is being done in the r code chunks describing what you are trying to do - credit will be given even if it does not work as long as you detail what you are doing. As we start to move into more statistics you will be expected to interpret the results. **(2 points)**

Points

- summary stats - 2 point
- assumptions and hypotheses - 3 points
- exploratory graphs - 2 point
- interpretation - 4 points
- Final figure - 1 point
- Results - 2 points