# MATH 530/630

Integrative Lab 4 - ANOVA

## Contents



## Overview

The goal of this lab is to carefully, thoroughly, and thoughtfully conduct an analysis of variance (ANOVA). You are also asked to communicate clearly about the steps in your analysis process with others, by sharing your R code, output, and narrative. As such, your code cannot "stand alone"- it is meant to complement / enhance / support your narrative. As with our previous iLab, this lab will be due in two stages:

1. A complete knitted `html` file submitted on Sakai.
2. At that time, you'll be provided with a code key. You are asked to review your initial submission, and reflect on your own code/narrative after reviewing the key thoroughly.

Using the key, your self-assessment should include even **more** narrative; where you made mistakes, you must discuss and analyze where you went wrong, and correct them without copying/pasting directly from the key (this typically means that you need to include more narrative than we provide in the key). A good self-assessment will include:

- Assessment of the accuracy and completeness of your "initial solutions"
- Correct worked solutions with some discussion and analysis of why your initial solution was incorrect, and reflection on the source of your confusion (if you got an answer correct, this is not necessary)
- Attributions as appropriate to other students who helped you, or other sources such as lecture notes, readings, online resources, etc. that helped you

## Headnotes

- Also this slide deck on the general linear model may be helpful.
- Also this slide deck on comparing multiple sample means.
- This slide deck on post-hoc tests and p-value adjustment.

## Logistics

You will use R Markdown to construct your analysis report. You'll submit your work as an html file knit from your `.Rmd` file (please leave the default code chunk options for `eval = TRUE` and `echo = TRUE`). Your lab should serve as your own personal cheatsheet in the future for ANOVAs. Give yourself the cheatsheet you deserve!

For all things, code and narrative, if you're dissatisfied with a result, discuss the problem, what you've tried and move on (remember my 30-minute rule). You'll need this loaded at the top:

```
library(tidyverse)
library(moderndive)
library(broom)
library(infer)
library(multcomp) # for post-hoc tests
library(ggbeeswarm)
library(car) # for Anova
library(phia) # for post-hoc tests
select <- dplyr::select # deals with the namespace conflict
```

## Data

This data is associated with this publication:

Here is the paper abstract:

> "A higher incidence of osteopenia is observed among children with inherited metabolic disorders (inborn errors of metabolism, or IEMs) who consume medical food–based diets that restrict natural vitamin D–containing food sources. We evaluated the vitamin D status of children with IEMs who live in the Pacific Northwest with limited sun exposure and determined whether bone mineral density (BMD) in children with phenylketonuria (PKU), the most common IEM, correlated with diet or biochemical markers of bone metabolism. We hypothesized that children with IEMs would have lower serum vitamin D concentrations than controls and that some children with PKU would have reduced bone mineralization. A retrospective record review of 88 patients with IEMs, and 445 children on unrestricted diets (controls) found the 25-hydroxyvitamin D concentrations were normal and not significantly different between groups (IEM patients, 27.1 $\pm$ 10.9; controls, 27.6 $\pm$ 11.2). Normal BMD at the hip or spine (-2 $<$ z-score $<$ 2) was measured in 20 patients with PKU. There was a correlation between lumbar spine BMD and dietary calcium intake. We saw no evidence of low serum vitamin D in our population of children with IEMs compared with control children. We also found no evidence for reduced BMD in children with PKU on specialized diets, but BMD was associated with calcium intake. Dietary intake of essential nutrients in medical food–based diets supports normal 25-hydroxyvitamin D levels and BMD in children with IEMs, including PKU. The risk of vitamin D deficiency among patients consuming a medical food–based diet is similar to the general population."

Use this code to read in the data:

```
or_vitd <- read_csv("http://bit.ly/conj620-orvitdcsv",
                    col_types = cols(
                      season = col_factor(levels = NULL),
                      patient_type = col_factor(levels = NULL),
                      region = col_factor(levels = NULL)))
```

# EDA

Explore the `patient_type`, `season`, and `vitamin_d` variables. Recall that a new exploratory data analysis involves three things:

- Looking at the raw values.
- Computing summary statistics of the variables of interest.
  - In this case, you should be able to verify this sentence: "A retrospective record review of 88 patients with IEMs, and 445 children on unrestricted diets (controls) found the 25-hydroxyvitamin D concentrations were normal and not significantly different between groups (IEM patients, 27.1 ± 10.9; controls, 27.6 ± 11.2)."
- Creating informative visualizations.

General functions that we have used at this stage:

- `dplyr::glimpse()`
- `skimr::skim()`
  - Given the grouped design, you may wish to do a `dplyr::group_by()` first here
- `ggplot2::ggplot()`
  - `geom_histogram()` or `geom_density()` or `geom_boxplot()` for numeric continuous variables
    - You may wish to combine these with `facet_wrap()`
  - `geom_bar()` or `geom_col()` for categorical variables

You may also find your want to use `filter`, `mutate`, `arrange`, `select`, or `count`. Let your questions lead you!

# Comparing many means

- Create a plot of the mean vitamin d levels (use `stat_summary` with `fun.data = mean_cl_boot`- this gives you the mean standard error from 1000 bootstrapped replicates) across season, colored by patient type.
- State the null and alternative hypotheses for the omnibus ANOVA.
- Hazard a guess as to (a) whether there are any main effects of season or patient type, and (b) whether you think there is a significant interaction between the two variables (that is, does the effect of one variable seem to depend on the level of the other variable)? Some examples are if you think patient type matters but only in the summer, or if season matters more for cases than controls.

# Two-way ANCOVA

Let's start with two predictors in our ANOVA model, using addition first (i.e., `+`). Variables as covariates are typically added (`+`), and their effects are assumed to be additive, so this is called an analysis of covariance (ANCOVA). No interaction term is estimated, meaning we are not allowing for estimated non-parallel lines.

Our two predictors will be:

- `patient_type`: 2 levels (case vs. control)
- `season`: 2 levels (winter vs. summer)

When we include another variable in our model, like `season`, then our estimate of the effect of `patient_type` is interpreted holding the value of the covariate fixed, just like in multiple regression.

- First, predict `vitamin_d` with `patient_type`; `season` should be the second predictor in your `lm`. Run `anova` on the `lm` object.
- Next, try switching the order of the two predictors (so, `season` should now be the first predictor in your `lm` model), then run `anova` on the new `lm` object. What do you notice?
- At this point, you may wish to examine each `lm` object using `moderndive::get_regression_table()`. Do the regression coefficients change depending on the order of the predictors?
- Remember, the `anova()` command as we have used it before was used to compare two nested models. The null hypothesis was that the more complicated model was not better than the less complicated model. Use the code below to compare each of your above models to a model with only the first predictor. Write a few sentences describing the output of this code, and how it helps you understand the null/alternative hypotheses being tested when `vitamin_d ~ patient_type + season` versus `vitamin_d ~ season + patient_type` (and how they are different).

```
# patient only
vitd_patient <- lm(vitamin_d ~ patient_type, data = or_vitd)
anova(vitd_patient, vitd_plusseason)

# season only
vitd_season <- lm(vitamin_d ~ season, data = or_vitd)
anova(vitd_season, vitd_pluspatient)
```

## Types of sums of squares

The majority of the time, if you are interested in doing an ANOVA, it is unlikely that you want output where the order of the predictors in your model formula matters. This type of output is called sequential sums of squares, also known as a Type I ANOVA. Here is more information on the 3 types of sums of squares in the context of an ANOVA:

- Type I: sequential (order matters)
  - This is the default in R when you use `anova`.
  - This is rarely what you will be interested in if you are not doing a nested models comparison intentionally.
- Type II:
  - This type tests for each main effect after the other main effect. Note that no significant interaction is assumed (in other words, you should test for interaction first) and only if AB is not significant, continue with the analysis for main effects).
- Type III:
  - This type tests for the presence of a main effect after the other main effect and interaction. However, it is often not interesting to interpret a main effect if interactions are present (generally speaking, if a significant interaction is present, the main effects should not be further analysed). If the interactions are not significant, type II gives a more powerful test.

In this ANOVA predicting `vitamind_d`, the separate partial effects, or main effects, of `patient_type` and `season` would be marginal to the `patient_type:season` interaction. In general, we neither test nor interpret main effects of explanatory variables that interact. If we can rule out interaction either on theoretical or empirical grounds, then we can proceed to test, estimate, and interpret main effects. It does not generally make sense to specify and fit models that include interaction regressors but that delete main effects that are

marginal to them. Such models — which violate the principle of marginality — are interpretable, but they are not broadly applicable.

The bottom line is that for anything beyond a one-way ANOVA (so anything with more than 1 predictor or independent variable), I recommend using `car::Anova` (capital "A" Anova, not lower case "a" anova) setting the `type` argument explicitly; from `?Anova`:

> "Type-II tests are calculated according to the principle of marginality, testing each term after all others, except ignoring the term's higher-order relatives; so-called type-III tests violate marginality, testing each term in the model after all of the others."

More reading on this topic here.

## Two-way ANCOVA with Type II SS

Now, we'll perform an ANOVA using the Type II SS. The way to do this is with the `car` package, using the `Anova` function with the `type = 2` argument. Try using this code:

```
vitd_add <- lm(vitamin_d ~ season + patient_type, data = or_vitd)
Anova(vitd_add, type = 2)
```

- Try switching the order of the predictors again to confirm for yourself that they are in fact the same!
- Interpret the output here. Keep in mind that any F value where the factor has only two levels can actually be interpreted without doing post-hoc testing.
- Often with ANOVA, you want to plot the *adjusted means*, which are the fitted/predicted values from the ANOVA model (just as were created with linear regression models we have worked with before). Use this code to use the `phia` package to calculate the fitted group means and plot them. Read the vignette. Write a few sentences here about the model output and this figure, making sure to note *why* the plots that contain two colors show lines that are parallel (hint).

```
vitd_add_means <- interactionMeans(vitd_add)
plot(vitd_add_means)
```

## Two-way ANOVA with Type II SS

Now run the same ANOVA analysis, this time allowing your two predictors to interact.

- Is the interaction significant? If it is not, then the Type II sums of squares analysis is valid. What do you say?
- Use the `phia` package to plot the fitted means from this model. What do you see? What is different now? (hint).

The following two questions are demos- in this actual context you would not need post-hoc testing.

- Attempt to conduct post-hoc comparisons between seasons using Tukey's Honestly Significant Difference method (through the `multcomp` package). What does the error tell you?
- Now try to use the `phia` package to do post-hoc tests; use the Bonferroni method for adjusting the p-values.

```
testInteractions(lm_model, pairwise = "variable_name", adjustment = "p.adjust.method")
```

## Final two-way ANOVA

Try a two-way ANOVA with Type II SS to predict vitamin D from season and region, allowing the two predictors to interact.

- Interpret the output here. Is the interaction significant? If it is not, then the Type II sums of squares analysis is valid. What do you say?
- Use the `phia` package to plot the fitted means from this model. What do you see?

The following two questions are demos- in this actual context you would certainly be p-hacking your way to glory if you ran all these comparisons, even with p-value adjustment!

- Imagine we had gotten a significant effect of region. Use `phia::testInteractions` to test all pairwise differences between regions using Bonferroni p-value adjustment. Any differences stand out?
- Use the same function, adding the `fixed = "season"` argument to compare all pairwise differences between regions separately for each season, again using Bonferroni p-value adjustment. Do you feel confident there is no effect of region, and no interaction between region and season?