Lab 6: Analysis of Variance

In this lab, we will learn a few additional helpful functions to use when tidying data. We will also learn to conduct analysis of variance tests in R. The learning goals of this lab are to:

- Learn the `group_by()`, `ungroup()`, and `summarize()` functions
- Run a one-way ANOVA using the `aov()` function
- Run post-hoc Tukey's HSD tests using the `TukeyHSD()` function
- Communicate the results of an ANOVA, graphically and in text

In the first part of today's lab, your instructor will walk through the code and relevant functions with you.

In the second part of today's lab, you will be asked to complete your assignment in a new RMarkdown file. Please be sure to answer assignment questions using full sentences, including information as you would in a final report. *Submit your RMarkdown file rendered as a pdf to the class Canvas site under the Assignments folder. Lab assignments are due at 11:59pm the day before your next lab session meets.*

Setup

Create a new project for this week's lab named `week6` in the `ENV710` folder. Open and save a new R Markdown script named `week6script.Rmd` in this project. Edit the title of the script accordingly, and add a first **Setup** header to your script. Then, create and run a new code chunk to load in the packages you will use today.

```r
# Load necessary packages.
library(here)
library(tidyverse)
library(moments)
library(palmerpenguins)

# Save the palmerpenguins dataset into your Environment.
penguin_dat <- penguins
```

Functions for grouping and summarizing data

When you'd like to quickly summarize the contents of a dataframe and impose a particular grouping on the data, the `group_by()` and `summarize()` functions allow you to do exactly that.

The `group_by()` function takes the name of the column that you want to group by, which needs to be categorical. When using `group_by()`, however, don't forget to `ungroup()` your data the close of your pipe so that grouping is not carried forward into other analyses you might want to conduct.

The `summarize()` function will perform the summary operations on the groups you've specified and created a new dataframe containing the summary outputs.

```r
# Calculate mean and standard deviation values of penguin flipper
# length according to the island they live on.
flipper_summary <- penguin_dat %>%
  group_by(island) %>%
  summarize(mean = mean(flipper_length_mm, na.rm = TRUE),
            sd = sd(flipper_length_mm, na.rm = TRUE)) %>%
  ungroup()
```

Practice Problem: Create a new dataframe that calculates the median values and variance in bill length according to penguin sex. Don't forget to ungroup your data at the end of your pipe!

Conducting Analysis of Variance (ANOVA) tests

Analysis of Variance or ANOVA tests may be used when you want to examine differences in more than two sample means. Remember the following hypotheses apply to an ANOVA:

Null Hypothesis: All sample means are equal.

Alternative Hypothesis: At least two sample means, among all samples, are not equal.

Before running an ANOVA, you must also ensure that your data meets the following requirements:

- Samples are independent of one another.
- Data within samples is normally distributed.
- Variances among samples are equal.

Create a new header in your script titled **One-Way ANOVAs**.

Examining Data Distribution

We will first examine the data we are using to ensure that sample distributions are approximately normal.

```
# Examine penguin body mass by species
penguin_fig1 <- ggplot(penguin_dat, aes(x = body_mass_g,
                        fill = species)) + # base plot
  geom_histogram() + # creates histograms
  scale_fill_manual(values = c("darkorange","purple","cyan4")) + # customize colors
  labs(x = "Body Mass (g)",
       y = "Count") + # edits labels
  facet_grid(.~species) + # facets by species to avoid overlap
  theme_bw() + # removes the grey background
  theme(legend.position = "none") # removes the legend

# View figure.
penguin_fig1
```

All but Gentoo appear reliably normally distributed, so will also calculate skew and kurtosis for this species to be more certain of moving forward with the one-way ANOVA test.

```
# Filter out only Gentoo values.
gentoo_dat <- penguin_dat %>%
  filter(species == "Gentoo")

# Calculate skew.
skewness(gentoo_dat$body_mass_g, na.rm = TRUE)

# Calculate kurtosis.
kurtosis(gentoo_dat$body_mass_g, na.rm = TRUE)
```

The data appears to be approximately symmetrical (skewness = 0.06) and moderately platykurtic (kurtosis = 2.26), so we will proceed with the assumption that these body mass data for the Gentoo species are approximately normally distributed.

Note, many data you will deal with in the future will not be normally distributed. In these cases you have two options on how to proceed:

(1) Apply a log transformation. This is standard practice to adjust raw values so that they do conform to a normal distribution and you may continue with applying a statistical test that requires normally-distributed data, like a one-way ANOVA. In R, you may apply either a log based 10 transformation (`log10()`) or a natural log transformation (`log()`) - either option is frequently used, just be sure to state clearly which you have performed.

(2) Stop and re-consider what kind of statistical test you should use. Later this week and early next, we'll cover a few alternative options when you encounter data that is not normally distributed but you would still like to compare values between groups.

Bartlett's Test for Equal Variance

The next thing we need to ensure is that variances are approximately equal across species. The Bartlett test for homogeneity of variances is based on a null hypothesis that the variance between all groups included in the data are the same. Typically, you will use a 95% confidence level (5% significant level) when applying this test.

```r
# Perform Bartlett test.
penguin_var <- bartlett.test(penguin_dat$body_mass_g, penguin_dat$species)

# Examine results.
penguin_var
```

Because p= 0.05, we do not have sufficient evidence to reject the null hypothesis, so we may assume that variance across the three groups (species) in this dataset are equal and proceed with conducting a one-way ANOVA.

There is also the **helpful rule of thumb** that states that if the *largest* sample variance is less than 4 times the *smallest* sample variance, you may still assume variances are equal across samples and conduct an ANOVA.

Analysis of Variance (ANOVA)

```r
# Perform a one-way ANOVA.
penguin_ANOVA <- aov(body_mass_g ~ species, data = penguin_dat)

# Examine the results.
summary(penguin_ANOVA)
```

If $p \geq 0.05$, you may conclude that there are no significant differences between mean values of any of the samples included in your dataset. If this is the case, you may stop here and report your results as such.

If $p < 0.05$, you may conclude that there is a significant difference in mean values between **at least two** of your samples and continue forward to conduct a Tukey's Honestly Significant Difference (HSD) test.

Post-hoc Tukey's HSD Test

```
# Perform a Tukey's HSD test.
penguin_Tukey <- TukeyHSD(penguin_ANOVA)

# Examine the results.
penguin_Tukey
```

Notice how the results of the Tukey's HSD display the p-value for each of the pair-wise comparisons between species. Based on these values, we may conclude that there is in fact a significant difference in body mass between the Gentoo and Adelie species as well as the Gentoo and Chinstrap species.

Communicating Results

When graphically presenting the results of an ANOVA, it is common to include small letters above each of the samples to notate which samples were significant from others. We'll do exactly that in the following plot.

We'll first calculate the mean and standard deviation values and then overlay those over top of the raw data points to provide better context for data distribution amongst the three species.

```
# Calculate means and standard deviation.
penguin_summary <- penguin_dat %>%
  group_by(species) %>%
  summarize(mean = mean(body_mass_g, na.rm = TRUE),
            sd = sd(body_mass_g, na.rm = TRUE)) %>%
  ungroup()

# We also need to rename the "mean" column - you
# will see why in a moment.
penguin_summary <- penguin_summary %>%
  rename(body_mass_g = mean)

# Create a plot to display results.
# Because we are using multiple datasets in the same figure,
# we will specify the aesthetics in each individual data call
penguin_fig2 <- ggplot() +
  # add raw data points
  geom_jitter(data = penguin_dat,
              aes(x = species, y = body_mass_g,
                  color = species),
              alpha = 0.5, size = 0.5) +
  # add summary statistics
  geom_point(data = penguin_summary,
             aes(x = species, y = body_mass_g,
                 color = species),
             size = 3) +
  # add error bars
  geom_errorbar(data = penguin_summary,
                aes(x = species,
                    ymin = body_mass_g-sd,
                    ymax = body_mass_g+sd,
                    color = species),
                width = 0.10, size = 1) +
  # add text annotations
  annotate("text", x = "Adelie", y = 5500,
           label = "a", size = 8) +
  annotate("text", x = "Chinstrap", y = 5500,
           label = "a", size = 8) +
  annotate("text", x = "Gentoo", y = 6100,
           label = "b", size = 8) +
  # edit colors
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  # label axes
  labs(x = "Species",
       y = "Body Mass (g)") +
  # remove grey background
  theme_bw() +
  # remove legend
  theme(legend.position = "none")

# View figure.
penguin_fig2
```

When reporting the results of an ANOVA it's best practice to include many of the measures of central tendency and data spread as we've seen in previous weeks. Furthermore, you'll want to include the degrees of freedom, the F-statistic, and p-value of the ANOVA as well as the p-values for any of the post-hoc tests that yielded significant differences. An example of how we might report our findings from this particular test can be found below:

Penguin species displayed significant differences in body mass as determined by one-way ANOVA (F(2, 339) = 343.6, $p < 0.001$). Post-hoc testing by Tukey's HSD revealed that mean body mass for Adelie (mean = 3,700 g, s.d. = 459 g) and Chinstrap (mean = 3,733 g, s.d. = 384 g) did not differ significantly, while Gentoo penguins (mean = 5,076 g, s.d. = 504 g) differed significantly from both species.

Wrap Up

Before beginning work on your assignment, be sure to *Save* and *Knit* your script to ensure it's all working properly.

Assignment

Complete the following questions by using the appropriate analyses in a new R Markdown script. Your script should include an appropriate title, load in the proper packages and data, and be structured using helpful headers. Please take care to follow the guidelines listed above when reporting the results of your statistical tests. You will submit a pdf file that displays *all* code and outputs. Be sure to submit that script, containing your rendered figures and with any requested answers as *full sentences*, via Canvas.

For this assignment, you will be using the dataset we saw on the very first day of class collected by the H.J. Andrews Long Term Ecological Research program within the Mack Creek watershed in Oregon - `and_vertebrates.csv`.

Note, the questions below ask you to proceed through the ANOVA protocol that we covered today in lab, but in some instances, **the data may not meet the necessary criteria**, and therefore you may not perform a one-way ANOVA for each question. Next week, you will learn how to address the issue of when your data may be non-normally distributed but you would still like to compare central tendency values across groups.

(1) Coastal Giant Salamander -

a. Filter the dataset only for coastal giant salamanders in cascades (C), pools (P), and side channels (SC). Create a figure that helps you to evaluate whether their snout-to-tail length

(`length_2_mm`) by habitat type (`unittype`) is normally distributed. You may also calculate skew and kurtosis values to help with your decision-making. Are these data normally distributed? Why or why not? If they are not, apply a log-transform (`log10()`) to the length data and re-evaluate if the data now meets the criteria to be considered normally-distributed.

b. If you found the data from part a were not normally distributed and a log-transform did not change this finding, you may stop here and proceed to question 2. If you found the data from part a were normally distributed, conduct a Bartlett's test for equal variance to determine if these data also satisfy the need for homogeneity of variances across groups. Do these data have approximately equal variances? Why or why not? Remember, the data may not pass the Bartlett's test, but if they adhere to the rule of thumb mentioned above, you may proceed with a one-way ANOVA.

c. If you found the data from part b did not display approximately equal variances, you may stop here. If you found the data from part b did display approximately equal variances, conduct a one-way ANOVA to see if there is a significant difference between coastal giant salamander lengths across cascades, pools, and side channels of Mack Creek. If you find evidence of significant differences, perform a post-hoc Tukey's HSD test to determine which are significant from which habitats. Communicate your findings as a figure (with an appropriate caption) and in a sentence, as it might appear in a final report.

(2) Cutthroat Trout -

a. Filter the dataset only for cutthroat trout and create a figure that helps you to evaluate whether their snout-to-fork length (`length_1_mm`) by reach (`reach`) is normally distributed. You may also calculate skew and kurtosis values to help with your decision-making. Are these data normally distributed? Why or why not? If they are not, apply a log-transform (`log10()`) to the length data and re-evaluate if the data now meets the criteria to be considered normally-distributed.

b. If you found the data from part a were not normally distributed and a log-transform did not change this finding, you may stop here. If you found the data from part a were normally distributed, conduct a Bartlett's test for equal variance to determine if these data also satisfy the need for homogeneity of variances across groups. Do these data have approximately similar variances? Why or why not? Remember, the data may not pass the Bartlett's test, but if they adhere to the rule of thumb mentioned above, you may proceed with a one-way ANOVA.

c. If you found the data from part b did not display equal variances, you may stop here. If you found the data from part b did display equal variances, conduct a one-way ANOVA to see if there is a significant difference between cutthroat trout lengths across lower, middle, and upper reaches of Mack Creek. If you find evidence of significant differences, perform a post-hoc Tukey's HSD test to determine which are significant from which reaches. Communicate your findings as a figure (with an appropriate caption) and a sentence, as it might appear in a final report.

Works Cited:

Gregory, S.V. and I. Arismendi (2020). Aquatic Vertebrate Population Study in Mack Creek, Andrews Experimental Forest, 1987 to present v. 14. *Environmental Data Initiative.* https://www.doi.org/10.6073/pasta/7c78d662e847cdbe33584add8f809165