Lab 11 - Math 58B: Data + ANOVA + LM

your name here

not due

Lab Goals

The goal of the lab is to work holistically with a dataset in order to gain information from the dataset in a variety of ways. First, describing the data graphically and numerically will allow an understanding of the variables. Running inferential techniques (here ANOVA and regression) allow for conclusions about the population at hand.

- plotting data
- summarizing data
- ANOVA output
- linear model output

Getting started

For formatting the the linear model output, use the **broom** package. Also, **tidyverse** will continue to be used for `ggplot()` and data wrangling.

For using data wrangling verbs, you might see: https://r4ds.had.co.nz/transform.html (remembering the verbs `filter()`, `arrange()`, `select()`, `mutate()`, `summarize()`, and `group_by()`.)

For `ggplot()`, look for inspiration on this cheat sheet: https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-visualization.pdf

Load packages & data

The data is on penguins near Palmer Station, Antarctica from the **palmerpenguins** package.

> Includes measurements for penguin species, island in Palmer Archipelago, size (flipper length, body mass, bill dimensions), and sex.

```
library(palmerpenguins)

data(penguins)
```

Q1. Look at the data

What are the observational units? What are the variables? Are the variables numeric or categorical? How many levels does each categorical variable have?

Q2. Plot the data

Try making many different plots.
* Use new `geoms_XXX()`s. * The usual ones: `geom_point()`, `geom_boxplot()`, `geom_jitter()`, `geom_bar()`,... * New ones: `geom_smooth()` `geom_line()`, `geom_violin()` * Make the plots more sophisticated: `geom_text()` (to add labels to points), `facet_wrap(~ var_for_grouping)`, 'scale_color_

Q3. Summarize the data

Try things like:

- tabulate the categorical variables (use `select()` and then `table()`)
- find the `mean()`, `median()`, `sd()`, `n()` broken down by (i.e., `group_by()`) species and/or year (you can `group_by()` two variables at a time!)
- if there are any missing values, you might need to include `na.rm = TRUE` to tell the function to "remove" the "NA" values

Q4. ANOVA

Run an ANOVA on the `body_mass_g` variable broken down by `species`. That is, you'll be testing the following:

$$H_0 : \mu_{Adelie} = \mu_{Chinstrap} = \mu_{Gentoo}$$

where μ represents the average `body_mass_g` for the entire species in the Palmer Station region of Antarctica.

Look at the output:

- Find MSG and MSE.
- Find the F test statistic.
- Find the p-value for the test above.
- Make a plot (or look to earlier plots) which tell a consistent story about whether or not the null hypothesis has been rejected.

```
penguins %>%
  lm(body_mass_g ~ species, data = .) %>%
  anova()
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 146864214 73432107  343.63 < 2.2e-16 ***
## Residuals 339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q5. Linear model on two numeric variables

Choose two numeric variables, and run a linear model. You may remember the code from earlier in the semester (look back to Lab 2).

- Plot the two variables (use `geom_smooth()` and remember to set `method = "lm"` and `se = FALSE`).
- Add a least squares regression line to the plot.
- Run a regression (using `lm()`) to find the intercept and the slope of the model.
- Notice that with the output we get a p-value for each of the two statistics (the intercept and the slope). What is the null hypothesis for the slope test? That is, what do you think the slope parameter is set to be in the null claim?

To Turn In

The only thing to turn in is the next installment of the project. The lab, per se, is not due. Use the tools in the lab (and the solutions to the lab which are posted) to investigate your own dataset.

Project instructions are here: https://m58-intro-stats.netlify.app/project.html

```
praise()
```

```
## [1] "You are tiptop!"
```