03\_Homework

Bill Perry

# Homework Week 2

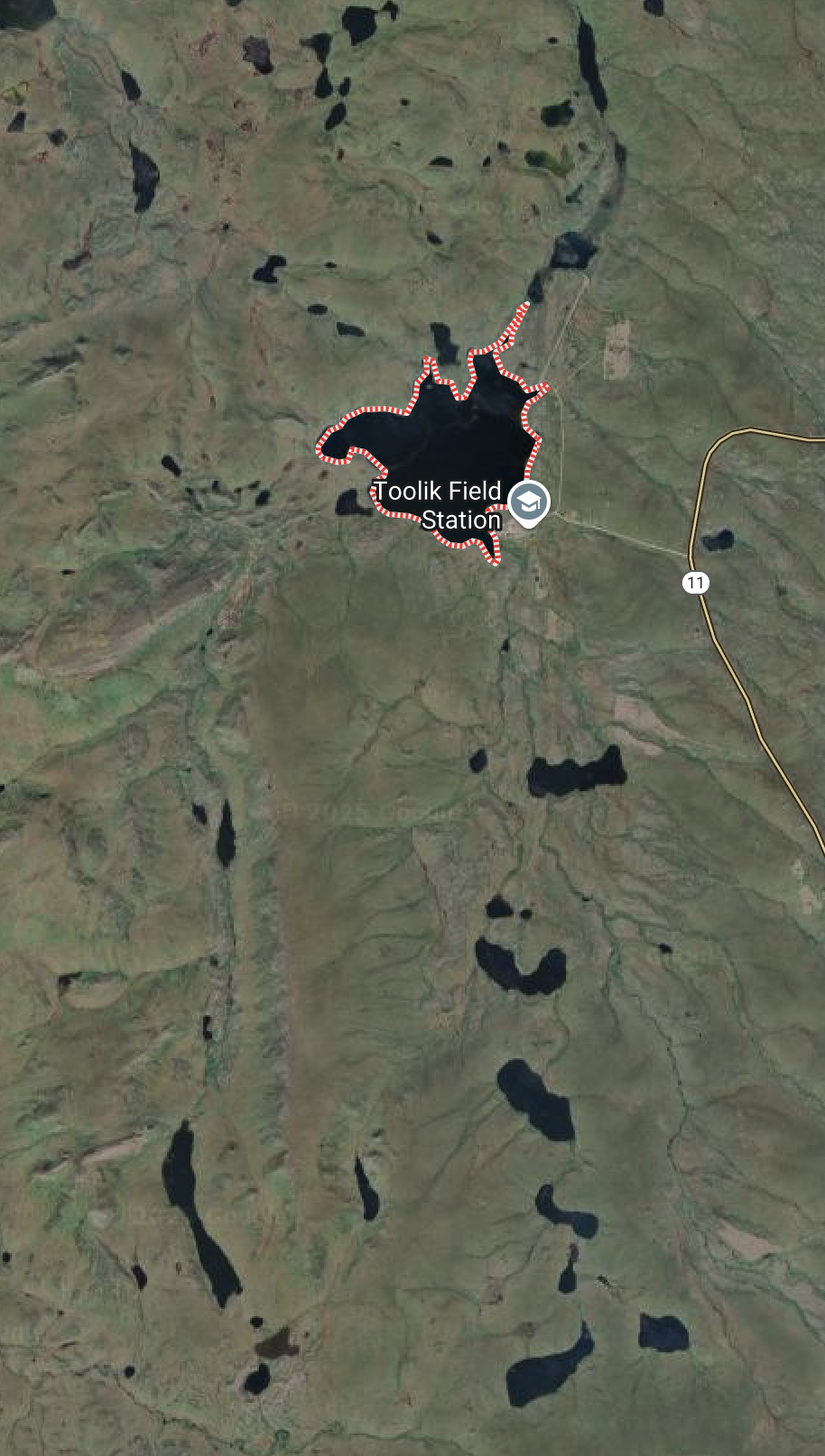
This is an assignment for you to practice coding and redo the work we do in class with a few twists on a new dataframe practicing to create new projects and writing new code. I realize you could copy the code from lecture and although you will get the code right, I urge you to retype it form scratch as it will be learned so much faster. This is a new language for you and if you dont “type” == “speak” the language you would remember it…. really and try breaking things. Dont be afraid you can download a new version or fix it… that is how we learn.

# Background

These data were collected by Mike McDonald and his laboratory at Toolik lake Alaska - below from google. At this site they were interested in fishes like Lake Trout. In this system there are some trout that live in lakes with forage fish (fish food) and other lakes where they do not get access to forage fish and eat snails. We will look at two lakes, one just to the south of this image called Island Lake and another, NE12 which is in the upper left.



lake trout



# Objectives and goals

In this homework we will explore the data on lake trout in these two lakes. As this is a lot of practice we will start on length and we will repeat the tests on mass.

Today we will test this on lake trout from NE 12 and Island Lake, the lake with forage fish but sample numbers are low

* We need to explore the lenght and weight data graphically
  + as a whole
    - box plots
    - histograms
  + By mainland and island
    - box plots
    - histograms
  + describe the output
    - do data look normally distributed
    - are there outliers
    - does the variation look about the same in each plot
* Generate summary statistics
  + mean, mode, min, max, variance, standard deviation, standard error, N
* If you were to go fish and catch at random what is the mean mass of fish you would catch at random if you sampled 10, 20, or 40 fish - please report mean and SE
* # # Note you can use   
  # set.seed(456) # makes it repeatable so we can check ; )  
    
  # Create samples of different sizes  
  # small\_sample <- df %>% sample\_n(times\_sampled)  
    
  # # Calculate mean and standard error for each sample  
  # small\_result\_df <- small\_sample %>%   
  # summarize(  
  # mean = mean(length\_mm, na.rm = TRUE),  
  # se = sum(!is.na(length\_mm)/sqrt(n))  
  # )
* Now if you wanted to fish at these lakes
  + what is the chance you will catch a fish larger than 450 mm and how does it differ by lake
  + what is the chance you will catch a fish larger than 1500 g

We’ll use the tidyverse package for data manipulation and visualization, along with patchwork for combining plots.

## Setup

First, let’s load the packages we need and the dataset:

# Load required packages  
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 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
✔ purrr 1.0.4   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(patchwork)  
  
# Read in the data file  
w3\_df <- read\_csv("data/lake\_trout.csv") %>% filter(lake %in% c("Island Lake","NE 12"))

Rows: 1502 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (3): sampling\_site, species, lake  
dbl (2): length\_mm, mass\_g  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Look at the first few rows  
head(w3\_df)

# A tibble: 6 × 5  
 sampling\_site species length\_mm mass\_g lake   
 <chr> <chr> <dbl> <dbl> <chr>   
1 Island Lake lake trout 640 2600 Island Lake  
2 Island Lake lake trout 650 2350 Island Lake  
3 Island Lake lake trout 585 2200 Island Lake  
4 Island Lake lake trout 720 3950 Island Lake  
5 Island Lake lake trout 880 6800 Island Lake  
6 Island Lake lake trout 830 3200 Island Lake

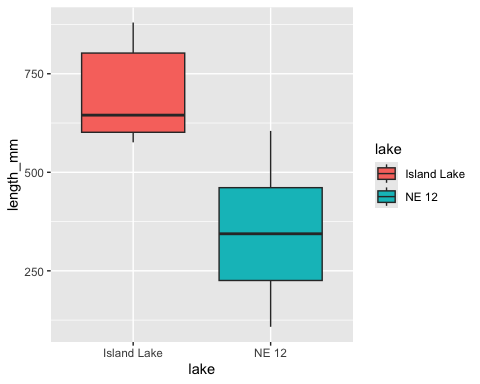
Let’s calculate some basic statistics for mice

# Calculate basic statistics   
w3\_stats <- w3\_df %>%   
 group\_by(lake) %>%   
 summarize(  
 mean\_length = mean(length\_mm, na.rm = TRUE),  
 sd\_length = sd(length\_mm, na.rm = TRUE),  
 n = sum(!is.na(length\_mm)),  
 se\_length = sd\_length / sqrt(n)  
 )  
  
# Display the statistics  
w3\_stats

# A tibble: 2 × 5  
 lake mean\_length sd\_length n se\_length  
 <chr> <dbl> <dbl> <int> <dbl>  
1 Island Lake 698. 121. 10 38.2   
2 NE 12 348. 127. 323 7.05

w3\_df %>% ggplot(aes(x=lake, y = length\_mm, fill=lake )) +geom\_boxplot()

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_boxplot()`).



w3\_df %>% ggplot(aes(x=mass\_g, fill=lake )) +geom\_histogram()+facet\_wrap(~lake)

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 2 rows containing non-finite outside the scale range  
(`stat\_bin()`).

