03\_Class\_Activity

Bill Perry

# In class activity 3:

## What did we do last time in activity 2?

* Setting up a project and variable names and code names
* How to use the pipe command %>%
* How to create descriptive statistics of a sample

p\_df %>%   
 summarize(  
 mean\_length = mean(len\_mm, na.rm = TRUE),  
 sd\_length = sd(len\_mm, na.rm = TRUE),  
 n\_length = sum(!is.na(len\_mm)))

* More graphs…
* ggplot(data = p\_df, aes(x=len\_mm, fill = wind)) +  
   geom\_histogram( binwidth = 2,   
  # sets the width in units of the bins - try different nubmers  
   position = position\_dodge2(width = 0.5))
* 
* What questions do you have and what is unclear - what did not work so far when you started the homework?

# Introduction

In this active learning module, we’ll explore real data from fish populations in Alaska. We’ll focus on understanding:

* How to create and interpret frequency distributions
* How sample size affects our view of a population
* How distributions differ among lakes

We’ll use the tidyverse package for data manipulation and visualization.

## Setup

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

# # Install the patchwork package if needed  
  
# install.packages("patchwork")  
library(patchwork)  
  
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

# Read in the data file  
sculpin\_df <- read\_csv("data/sculpin.csv")

Rows: 1052 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (2): lake, species  
dbl (3): site, total\_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(sculpin\_df)

# A tibble: 6 × 5  
 site lake species total\_length\_mm mass\_g  
 <dbl> <chr> <chr> <dbl> <dbl>  
1 146 E 01 slimy sculpin 53 1.25  
2 146 E 01 slimy sculpin 61 1.9   
3 146 E 01 slimy sculpin 53 1.75  
4 146 E 01 slimy sculpin 77 4.25  
5 146 E 01 slimy sculpin 45 0.9   
6 146 E 01 slimy sculpin 48 0.9

## Basic Data Summary

Let’s first check what lakes are in our dataset:

# Get a list of unique lakes  
unique(sculpin\_df$lake)

[1] "E 01" "E 05" "NE 12" "NE 14" "S 06" "S 07" "Toolik"

How many fish do we have from each lake?

# Count observations by lake  
sculpin\_df %>%  
 group\_by(lake) %>%   
 summarize(sculpin\_n = sum(!is.na(total\_length\_mm)))

# A tibble: 7 × 2  
 lake sculpin\_n  
 <chr> <int>  
1 E 01 79  
2 E 05 14  
3 NE 12 180  
4 NE 14 37  
5 S 06 132  
6 S 07 73  
7 Toolik 208

# Part 1: Creating Frequency Distributions

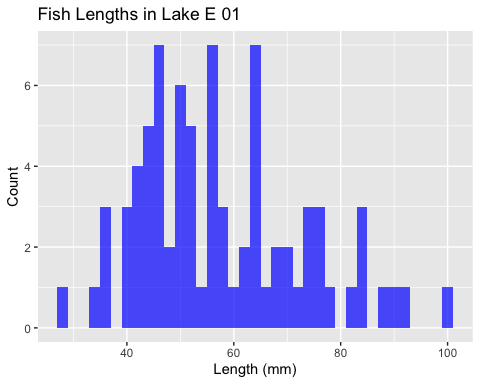
## Basic Histograms

A histogram shows how many observations fall into certain ranges (or “bins”).

Let’s create a simple histogram of fish lengths from Lake E 01 :

# Filter for Toolik Lake and create a histogram  
sculpin\_df %>%  
 filter(lake == "E 01") %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_histogram(binwidth = 2, fill = "blue", alpha = 0.7) +  
 labs(title = "Fish Lengths in Lake E 01",  
 x = "Length (mm)",  
 y = "Count")

Warning: Removed 189 rows containing non-finite outside the scale range  
(`stat\_bin()`).



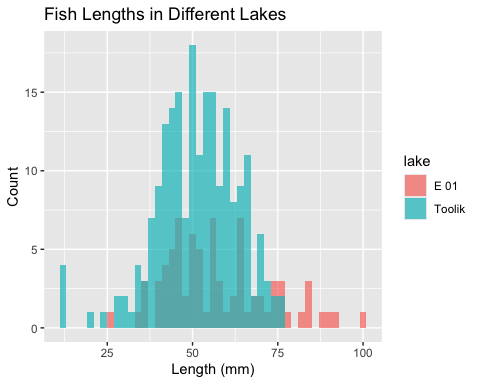
|  |
| --- |
| Activity 1 |
| Try changing the binwidth parameter to 5 and then to 1. How does the appearance of the histogram change? |

## Comparing Lakes

Now let’s compare two lakes

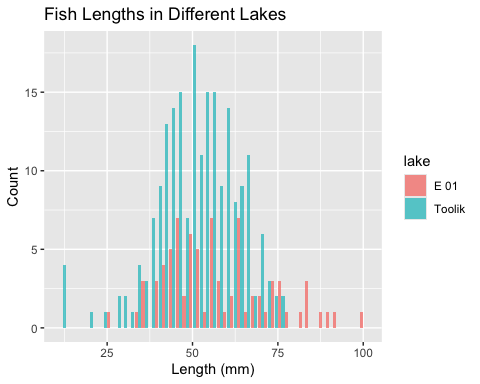
# Compare histograms from Toolik and E 01 lakes  
sculpin\_df %>%  
 filter(lake %in% c("Toolik", "E 01")) %>%  
 ggplot(aes(x = total\_length\_mm, fill = lake)) +  
 geom\_histogram(binwidth = 2, alpha = 0.7,   
 position = "identity") +  
 labs(title = "Fish Lengths in Different Lakes",  
 x = "Length (mm)",  
 y = "Count")

Warning: Removed 268 rows containing non-finite outside the scale range  
(`stat\_bin()`).



# Compare histograms from Toolik and E 01 lakes  
sculpin\_df %>%  
 filter(lake %in% c("Toolik", "E 01")) %>%  
 ggplot(aes(x = total\_length\_mm, fill = lake)) +  
 geom\_histogram(binwidth = 2, alpha = 0.7,   
 position = position\_dodge2(width=1)) +  
 labs(title = "Fish Lengths in Different Lakes",  
 x = "Length (mm)",  
 y = "Count")

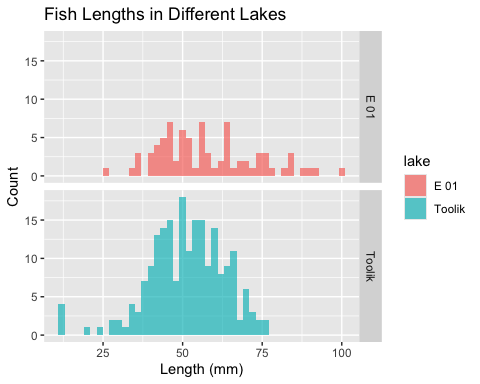
Warning: Removed 268 rows containing non-finite outside the scale range  
(`stat\_bin()`).



Now let’s compare two lakes side by side:

# Compare histograms from Toolik and E 01 lakes  
sculpin\_df %>%  
 filter(lake %in% c("Toolik", "E 01")) %>%  
 ggplot(aes(x = total\_length\_mm, fill = lake)) +  
 geom\_histogram(binwidth = 2, alpha = 0.7, position = "identity") +  
 labs(title = "Fish Lengths in Different Lakes",  
 x = "Length (mm)",  
 y = "Count") +  
 # facet\_wrap(~lake, ncol = 1) +  
 facet\_grid(lake~.)

Warning: Removed 268 rows containing non-finite outside the scale range  
(`stat\_bin()`).



|  |
| --- |
| Activity 2 |
| Choose two new lakes to compare. What differences do you notice in their distributions? |

Add code here

# enter code here

# Part 2: Sample Size Effects

Let’s explore how the sample size affects what we see.

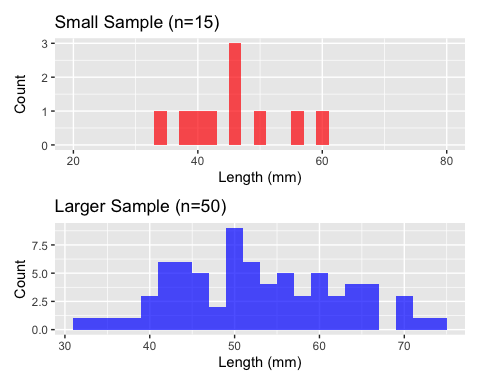
## Small vs. Large Samples

We’ll randomly select different sample sizes from Toolik Lake:

# Set a seed for reproducibility  
set.seed(123)  
  
# Create small sample (15 fish)  
small\_sample <- sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 sample\_n(15)  
  
# Create larger sample (50 fish)  
larger\_sample <- sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 sample\_n(100)  
  
# Plot both samples  
p1 <- small\_sample %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_histogram(binwidth = 2, fill = "red", alpha = 0.7) +  
 # coord\_cartesian(xlim = c(20,80)) +  
 labs(title = "Small Sample (n=15)",  
 x = "Length (mm)",  
 y = "Count") +  
 coord\_cartesian(xlim = c(20,80))  
  
p2 <- larger\_sample %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_histogram(binwidth = 2, fill = "blue", alpha = 0.7) +  
 # coord\_cartesian(xlim = c(20,80)) +  
 labs(title = "Larger Sample (n=50)",  
 x = "Length (mm)",  
 y = "Count")  
  
  
# Display the plots side by side  
p1 + p2 +  
 plot\_layout(ncol = 1)

Warning: Removed 5 rows containing non-finite outside the scale range  
(`stat\_bin()`).

Warning: Removed 26 rows containing non-finite outside the scale range  
(`stat\_bin()`).



|  |
| --- |
| Activity 3 |
| Try changing the sample sizes. What happens when you use very small samples (n=5)? What about larger samples (n=150)? |

add code here

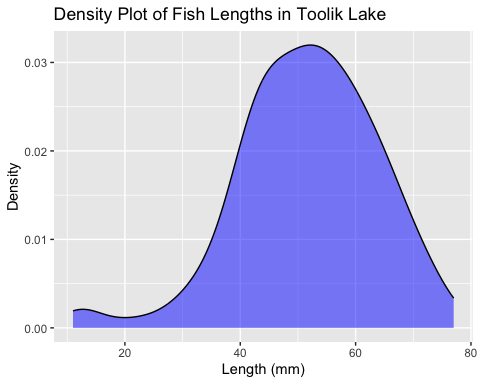
# enter code here

# Part 3: From Histograms to Density Plots

Density plots give us a smoothed version of the histogram:

# Create a density plot  
sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_density(fill = "blue", alpha = 0.5) +  
 labs(title = "Density Plot of Fish Lengths in Toolik Lake",  
 x = "Length (mm)",  
 y = "Density")

Warning: Removed 79 rows containing non-finite outside the scale range  
(`stat\_density()`).

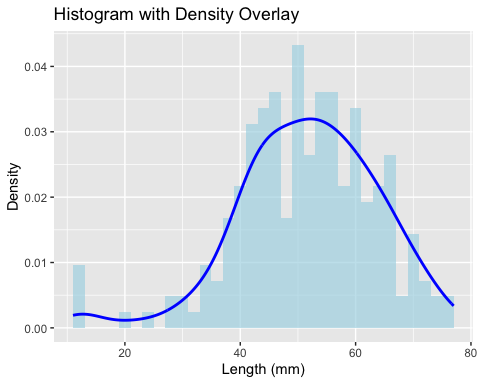


We can overlay the histogram and the density plot:

# Combine histogram and density plot  
sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_histogram(aes(y = after\_stat(density)), binwidth = 2,   
 fill = "lightblue", alpha = 0.7) +  
 geom\_density(color = "blue", linewidth = 1) +  
 labs(title = "Histogram with Density Overlay",  
 x = "Length (mm)",  
 y = "Density")

Warning: Removed 79 rows containing non-finite outside the scale range  
(`stat\_bin()`).

Warning: Removed 79 rows containing non-finite outside the scale range  
(`stat\_density()`).



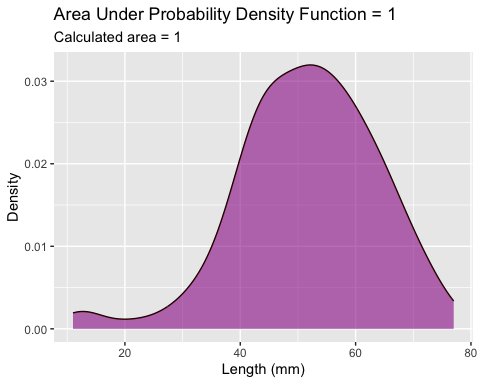
|  |
| --- |
| Activity 4 |
| Create a density plot comparing multiple lakes. Which lakes have similar distributions? Which ones are different? |

Try code here using patchwork or facet\_grid

#Enter code here#

# Function to calculate area under density curve  
calculate\_density\_area <- function(data\_vector) {  
 # Remove NA values  
 data\_vector <- data\_vector[!is.na(data\_vector)]  
   
 # Calculate density  
 dens <- density(data\_vector)  
   
 # Calculate area using numeric integration (trapezoidal rule)  
 # Area should be approximately 1  
 dx <- diff(dens$x)  
 y\_avg <- (dens$y[-1] + dens$y[-length(dens$y)]) / 2  
 area <- sum(dx \* y\_avg)  
 return(area)  
}  
  
# Apply to Toolik lake data  
toolik\_data <- sculpin\_df %>%   
 filter(lake == "Toolik") %>%   
 pull(total\_length\_mm)  
  
area\_value <- calculate\_density\_area(toolik\_data)  
  
# Create plot with calculated area  
sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 ggplot(aes(x = total\_length\_mm)) +  
 geom\_density(fill = "blue", alpha = 0.4) +  
 geom\_area(stat = "density", fill = "red", alpha = 0.3) +  
 labs(title = "Area Under Probability Density Function = 1",  
 subtitle = paste("Calculated area =", round(area\_value, 4)),  
 x = "Length (mm)",  
 y = "Density")

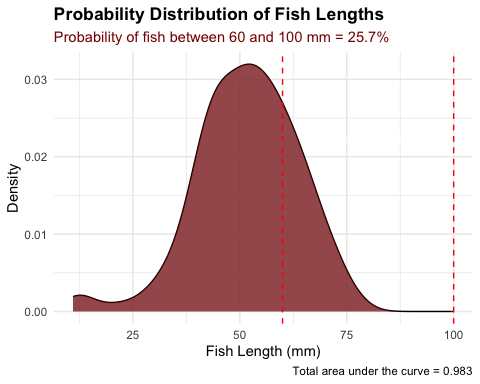
Warning: Removed 79 rows containing non-finite outside the scale range  
(`stat\_density()`).  
Removed 79 rows containing non-finite outside the scale range  
(`stat\_density()`).



This can be adapted to calculate the area of a subset of the plot

# ------- PART 3: SET STUDENT INPUT VALUES -------  
# Students can change these values to calculate different probabilities  
  
# For this example, let's calculate the probability of fish between 40mm and 60mm  
lower\_bound <- 60 # Student would change this value  
upper\_bound <- 100 # Student would change this value  
  
  
  
# ------- PART 1: PREPARE THE DATA -------  
  
# Filter data for just one lake to keep it simple for students  
toolik\_fish <- sculpin\_df %>%  
 filter(lake == "Toolik") %>%  
 filter(!is.na(total\_length\_mm)) # Remove any missing values  
  
# ------- PART 2: CREATE A FUNCTION TO CALCULATE PROBABILITY -------  
  
# This function calculates the probability of finding a fish with length between  
# lower\_bound and upper\_bound using the empirical distribution of our data  
calculate\_probability <- function(data\_vector, lower\_bound, upper\_bound) {  
 # First, we create a density object from our data  
 dens <- density(data\_vector)  
   
 # Find indices of x-values that fall within our bounds  
 indices <- which(dens$x >= lower\_bound & dens$x <= upper\_bound)  
   
 # If we have no points in the range, return 0  
 if(length(indices) <= 1) {  
 return(0)  
 }  
   
 # Get x and y values within our bounds  
 x\_values <- dens$x[indices]  
 y\_values <- dens$y[indices]  
   
 # Calculate the area using the trapezoidal rule  
 # (average height × width) for each segment, then sum all segments  
 widths <- diff(x\_values)  
 avg\_heights <- (y\_values[-1] + y\_values[-length(y\_values)]) / 2  
 area\_in\_range <- sum(widths \* avg\_heights)  
   
 # Return the calculated probability  
 return(area\_in\_range)  
}  
  
  
# ------- PART 4: CALCULATE THE PROBABILITY -------  
  
# Calculate the probability for the specified range  
probability <- calculate\_probability(toolik\_fish$total\_length\_mm, lower\_bound, upper\_bound)  
  
# Calculate the total area to show that the complete distribution sums to approximately 1  
total\_area <- calculate\_probability(toolik\_fish$total\_length\_mm,   
 min(toolik\_fish$total\_length\_mm),  
 max(toolik\_fish$total\_length\_mm))  
  
# ------- PART 5: CREATE THE VISUALIZATION -------  
  
# Create a plot showing the full distribution and the selected range  
ggplot(toolik\_fish, aes(x = total\_length\_mm)) +  
 # First, plot the overall density curve  
 geom\_density(fill = "lightblue", alpha = 0.5) +  
   
 # Then highlight our region of interest in a different color  
 geom\_area(stat = "density",   
 fill = "darkred",   
 alpha = 0.7,  
 xlim = c(lower\_bound, upper\_bound)) +  
   
 # Add vertical lines to clearly mark the boundaries  
 geom\_vline(xintercept = lower\_bound, linetype = "dashed", color = "red") +  
 geom\_vline(xintercept = upper\_bound, linetype = "dashed", color = "red") +  
   
 # Add informative labels  
 labs(  
 title = "Probability Distribution of Fish Lengths",  
 subtitle = paste0("Probability of fish between ", lower\_bound,   
 " and ", upper\_bound, " mm = ",   
 round(probability \* 100, 1), "%"),  
 caption = paste("Total area under the curve =", round(total\_area, 3)),  
 x = "Fish Length (mm)",  
 y = "Density"  
 ) +  
   
 # Add text annotations to explain the areas  
 annotate("text", x = (lower\_bound + upper\_bound)/2,   
 y = max(density(toolik\_fish$total\_length\_mm)$y) \* 0.7,  
 label = paste0("Area = ", round(probability, 3)),  
 color = "white", size = 4) +  
   
 # Make the plot look nicer  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(face = "bold"),  
 plot.subtitle = element\_text(color = "darkred")  
 )

Warning in geom\_area(stat = "density", fill = "darkred", alpha = 0.7, xlim =  
c(lower\_bound, : Ignoring unknown parameters: `xlim`



# Part 4: Summary Statistics

Let’s calculate basic summary statistics for each lake:

# Calculate mean, standard deviation, and sample size by lake  
sculpin\_df %>%  
 group\_by(lake) %>%  
 summarize(  
 mean\_length = mean(total\_length\_mm),  
 sd\_length = sd(total\_length\_mm),  
 count = n(),  
 .groups = "drop"  
 ) %>%  
 arrange(desc(count))

# A tibble: 7 × 4  
 lake mean\_length sd\_length count  
 <chr> <dbl> <dbl> <int>  
1 Toolik NA NA 287  
2 E 01 NA NA 268  
3 NE 12 49.8 15.2 180  
4 S 06 54.0 10.9 132  
5 E 05 NA NA 75  
6 S 07 55.6 12.7 73  
7 NE 14 47.3 10.5 37

## WOAH - what happened there - there are NA values in the data

you need to either remove missing values or you can do that in the formulas

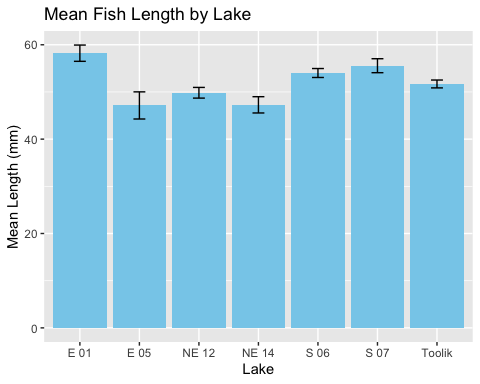
*What is the advantage to manually removing or doing it in formulas?*

# Calculate mean, standard deviation, and sample size by lake  
sculpin\_stats\_df <- sculpin\_df %>%  
 group\_by(lake) %>%  
 summarize(  
 mean\_length = mean(total\_length\_mm, na.rm = TRUE),  
 sd\_length = sd(total\_length\_mm, na.rm = TRUE),  
 se\_length = sd(total\_length\_mm, na.rm = TRUE)/ sum(!is.na(total\_length\_mm))^.5,  
 count = sum(!is.na(total\_length\_mm)),  
 .groups = "drop"  
 ) %>%  
 arrange(desc(count))  
sculpin\_stats\_df

# A tibble: 7 × 5  
 lake mean\_length sd\_length se\_length count  
 <chr> <dbl> <dbl> <dbl> <int>  
1 Toolik 51.7 12.0 0.834 208  
2 NE 12 49.8 15.2 1.13 180  
3 S 06 54.0 10.9 0.949 132  
4 E 01 58.2 15.3 1.72 79  
5 S 07 55.6 12.7 1.48 73  
6 NE 14 47.3 10.5 1.72 37  
7 E 05 47.1 10.8 2.88 14

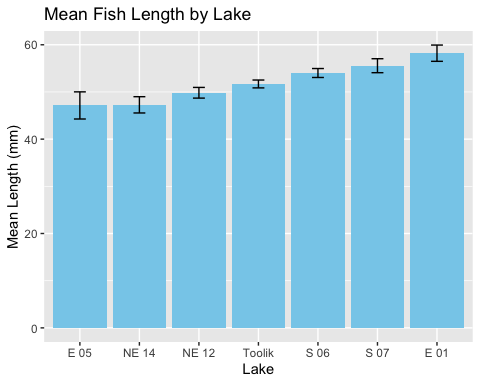
Now let’s visualize these statistics:

# Create a bar plot of mean lengths with error bars  
sculpin\_df %>%   
 ggplot(aes(lake, total\_length\_mm)) +  
 stat\_summary(  
 fun = mean, na.rm = TRUE,   
 geom = "bar",  
 fill = "skyblue"  
 ) +  
 stat\_summary(  
 fun.data = mean\_se, na.rm = TRUE,   
 geom = "errorbar",   
 width = 0.2) +  
 labs(title = "Mean Fish Length by Lake",  
 x = "Lake",  
 y = "Mean Length (mm)")



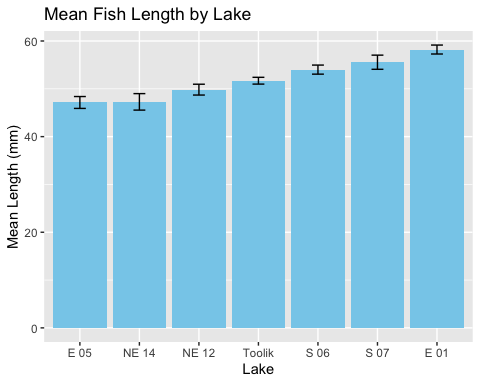
We could also do this from the dataframe we just made

# Create a bar plot of mean lengths with error bars  
sculpin\_stats\_df %>%   
 ggplot(aes(x = reorder(lake, mean\_length), y = mean\_length)) +  
 geom\_bar(stat = "identity",   
 fill = "skyblue") +  
 geom\_errorbar(aes(  
 ymin = mean\_length - se\_length,   
 ymax = mean\_length + se\_length),  
 width = 0.2  
 ) +  
 labs(  
 title = "Mean Fish Length by Lake",  
 x = "Lake",  
 y = "Mean Length (mm)")



The power of the pipe command is you can do this without hving to make a new dataframe

# Create a bar plot of mean lengths with error bars  
sculpin\_df %>%  
 group\_by(lake) %>%  
 summarize(  
 mean\_length = mean(total\_length\_mm, na.rm = TRUE),  
 sd\_length = sd(total\_length\_mm, na.rm = TRUE),  
 se\_length = sd\_length / sqrt(n()),  
 count = n(),  
 .groups = "drop"  
 ) %>%  
 filter(count >= 10) %>% # Only include lakes with sufficient sample size  
 ggplot(aes(x = reorder(lake, mean\_length), y = mean\_length)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 geom\_errorbar(aes(ymin = mean\_length - se\_length,   
 ymax = mean\_length + se\_length),  
 width = 0.2) +  
 labs(title = "Mean Fish Length by Lake",  
 x = "Lake",  
 y = "Mean Length (mm)")



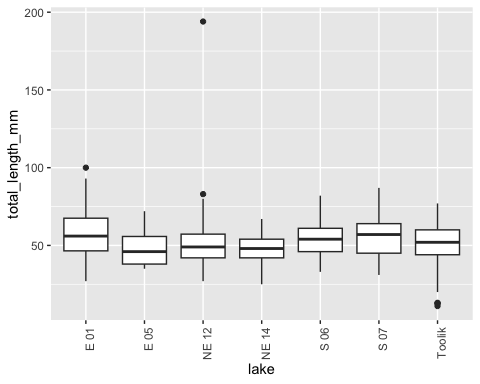
|  |
| --- |
| Activity 5 |
| Based on the mean plot and what you’ve seen in the distributions, what can you say about fish sizes in different lakes? Are there lakes with particularly large or small fish?  We will start to ask how different are they and is it by chance?  This is the inductive phase of doing reserach. |

# Part 5: Guided Challenges

Now it’s your turn to explore the data! Work with your partner to complete these challenges:

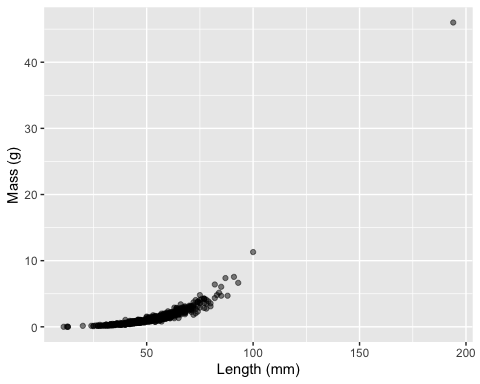
1. Find the lake with the widest range of fish lengths (hint: use the range() function)
2. Create box and whisker plots to compare fish lengths across lakes:

# Example boxplot code to get you started  
sculpin\_df %>%  
 filter(!is.na(total\_length\_mm)) %>%  
 ggplot(aes(x = lake, y = total\_length\_mm)) +  
 geom\_boxplot() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



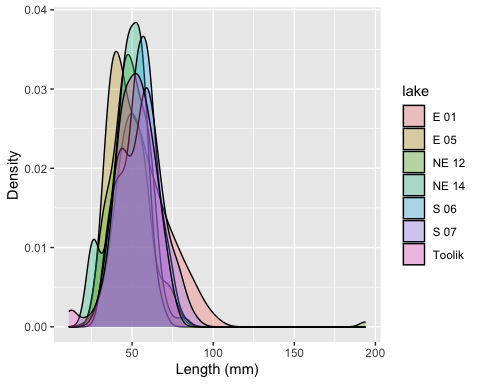
1. Explore if there’s a relationship between fish length and mass:

# Starting code for length-mass relationship  
sculpin\_df %>%  
 filter(!is.na(total\_length\_mm), !is.na(mass\_g)) %>%  
 ggplot(aes(x = total\_length\_mm, y = mass\_g)) +  
 geom\_point(alpha = 0.5) +  
 labs(x = "Length (mm)", y = "Mass (g)")



1. Try creating a density plot that shows all lakes in different colors:

# Starting code for multi-lake density plot  
sculpin\_df %>%  
 filter(!is.na(total\_length\_mm)) %>%  
 ggplot(aes(x = total\_length\_mm, fill = lake)) +  
 geom\_density(alpha = 0.3) +  
 labs(x = "Length (mm)", y = "Density")



# Reflection Questions

After completing the activities, discuss these questions with your group:

1. How does sample size affect our view of a population’s characteristics?
2. Why might fish lengths be different in different lakes?
3. What are the advantages and disadvantages of histograms versus density plots?
4. What additional data would help you better understand these fish populations?