03\_Class\_Activity

Bill Perry

# In class activity 3:

# 

# What did we do last time?

* Implement data pipeline best practices
* Apply controlled vocabulary and naming conventions
* Create effective visualizations
* Customize plots for publication quality
* Combine multiple plots into composite figures
* ggplot(name\_df, aes(x\_variable, y\_variable, color = categorical\_variable)) +  
  # dataframe, aesthetics(x and y variables, mapping of color or fill or shape) +   
   geom\_point() +  
  # this it the geometry you want and can add more layers like  
   geom\_line()
* What questions do you have and what is unclear
* What did not work so far when you started the homework?

# Objectives and goals for today

## Today’s Objectives

1. Implement descriptive statistics in R
2. Calculate measures of central tendency and spread
3. Compare distributions of data from different groups
4. Create effective visualizations of descriptive statistics
5. Interpret the meaning of these statistics in a biological context



# Part 1: Setting Up Your Environment

First, let’s load the necessary packages and import our data:

# Load required packages  
library(tidyverse) # For data wrangling and visualization  
library(knitr) # For creating tables  
library(moments) # For calculating skewness and kurtosis  
  
# Set a consistent theme for our plots  
theme\_set(theme\_minimal(base\_size = 12))

## Getting the data

|  |
| --- |
| Practice Exercise 1: Loading and Examining the Grayling Data |
| We’ll be working with data on arctic grayling fish from two different lakes (I3 and I8).  # Write your code here to read in the file # How do you examine the data - what are the ways you think and lets try it!  # Load the grayling data grayling\_df <- read\_csv("data/gray\_I3\_I8.csv")  Rows: 168 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.  # View the first few rows head(grayling\_df)  # A tibble: 6 × 5  site lake species total\_length\_mm mass\_g  <dbl> <chr> <chr> <dbl> <dbl> 1 113 I3 arctic grayling 266 135 2 113 I3 arctic grayling 290 185 3 113 I3 arctic grayling 262 145 4 113 I3 arctic grayling 275 160 5 113 I3 arctic grayling 240 105 6 113 I3 arctic grayling 265 145 |

# Examine the data structure  
glimpse(grayling\_df)

Rows: 168  
Columns: 5  
$ site <dbl> 113, 113, 113, 113, 113, 113, 113, 113, 113, 113, 113,…  
$ lake <chr> "I3", "I3", "I3", "I3", "I3", "I3", "I3", "I3", "I3", …  
$ species <chr> "arctic grayling", "arctic grayling", "arctic grayling…  
$ total\_length\_mm <dbl> 266, 290, 262, 275, 240, 265, 265, 253, 246, 203, 289,…  
$ mass\_g <dbl> 135, 185, 145, 160, 105, 145, 150, 130, 130, 71, 179, …

# Get a statistical summary  
summary(grayling\_df)

site lake species total\_length\_mm  
 Min. :113 Length:168 Length:168 Min. :191.0   
 1st Qu.:113 Class :character Class :character 1st Qu.:270.8   
 Median :118 Mode :character Mode :character Median :324.5   
 Mean :116 Mean :324.5   
 3rd Qu.:118 3rd Qu.:377.0   
 Max. :118 Max. :440.0   
   
 mass\_g   
 Min. : 53.0   
 1st Qu.:151.2   
 Median :340.0   
 Mean :351.2   
 3rd Qu.:519.5   
 Max. :889.0   
 NA's :2

# How many fish do we have from each lake?  
kable(grayling\_df %>%  
 count(lake))

| lake | n |
| --- | --- |
| I3 | 66 |
| I8 | 102 |

# Questions to Consider:

1. What variables are in our dataset?
2. What are their data types?
3. Are there any missing values?
4. What is the range of fish lengths in our dataset?
5. How many fish were sampled from each lake?

# Part 2: Calculating Descriptive Statistics

## Let’s calculate various descriptive statistics for our data:

|  |
| --- |
| Practice Exercise 2: Measures of Central Tendency |
| Let’s recreate the basic histogram of fish lengths from our last class. Use the sculpin\_df data frame that’s already loaded.  # Write your code here to read in the file # How do you examine the data - what are the ways you think and lets try it! # Calculate the mean and median fish length mean\_length <- mean(grayling\_df$total\_length\_mm) median\_length <- median(grayling\_df$total\_length\_mm)  cat("Mean fish length:", round(mean\_length, 1), "mm\n")  Mean fish length: 324.5 mm  cat("Median fish length:", median\_length, "mm\n")  Median fish length: 324.5 mm |

# Calculate mean and median by lake  
grayling\_df %>%  
 group\_by(lake) %>%  
 summarise(  
 mean\_length = mean(total\_length\_mm),  
 median\_length = median(total\_length\_mm)  
 ) %>%  
 kable(caption = "Mean and median fish length by lake", digits = 1)

Mean and median fish length by lake

| lake | mean\_length | median\_length |
| --- | --- | --- |
| I3 | 265.6 | 266 |
| I8 | 362.6 | 373 |

## Summarizing data - two ways

lets say we want to summarize the data and need to get n, means, standard deviation, standard error

We could do the following - if we had missing cells the code below would give an error

mean(grayling\_df$total\_length\_mm)

[1] 324.494

mean(grayling\_df$total\_length\_mm, na.rm = TRUE) # removes missing values

[1] 324.494

length(grayling\_df$total\_length\_mm)

[1] 168

* the length counts missing and non-missing data
* however this would get old if we had to do this for everything and then to do it for the different groupings - lee and windward…

## We need to learn to pipe things

* the dataframe –> pipe command that feed the dataframe into –> next command

grayling\_df %>% summarize(mean\_length = mean(total\_length\_mm, na.rm = TRUE))

# A tibble: 1 × 1  
 mean\_length  
 <dbl>  
1 324.

## What is cool is we can do a lot of different things now

grayling\_df %>%   
 summarize(  
 mean\_length = mean(total\_length\_mm, na.rm = TRUE),  
 sd\_length = sd(total\_length\_mm, na.rm = TRUE),  
 n\_length = n())

# A tibble: 1 × 3  
 mean\_length sd\_length n\_length  
 <dbl> <dbl> <int>  
1 324. 65.0 168

## Super cool code in case there are missing values

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

# A tibble: 1 × 3  
 mean\_length sd\_length n\_length  
 <dbl> <dbl> <int>  
1 324. 65.0 168

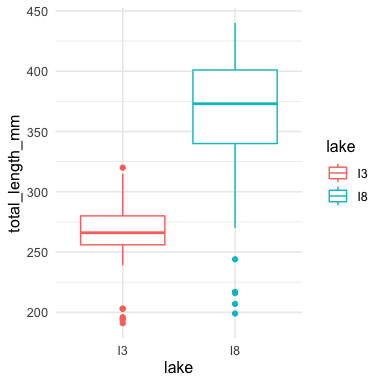
# Now for Spread…

|  |
| --- |
| Practice Exercise 3: Measures of Spread |
| # Write your code here to read in the file # Calculate standard deviation, variance, and IQR sd\_length <- sd(grayling\_df$total\_length\_mm) var\_length <- var(grayling\_df$total\_length\_mm) iqr\_length <- IQR(grayling\_df$total\_length\_mm)  cat("Standard deviation:", round(sd\_length, 1), "mm\n")  Standard deviation: 65 mm  cat("Variance:", round(var\_length, 1), "mm²\n")  Variance: 4225.9 mm²  cat("Interquartile range:", iqr\_length, "mm\n")  Interquartile range: 106.25 mm |

|  |
| --- |
| Exercise 4: Calculate Quartiles and Percentiles |
| # Calculate quartiles for overall data quartiles <- quantile(grayling\_df$total\_length\_mm, probs = c(0.25, 0.5, 0.75)) cat("First quartile (Q1):", quartiles[1], "mm\n")  First quartile (Q1): 270.75 mm  cat("Second quartile (Median):", quartiles[2], "mm\n")  Second quartile (Median): 324.5 mm  cat("Third quartile (Q3):", quartiles[3], "mm\n")  Third quartile (Q3): 377 mm  # Calculate a more comprehensive set of percentiles percentiles <- quantile(grayling\_df$total\_length\_mm,   probs = c(0.1, 0.25, 0.5, 0.75, 0.9))  # Display the percentiles kable(  data.frame(  Percentile = c("10th", "25th (Q1)", "50th (Median)", "75th (Q3)", "90th"),  Value = percentiles  ),  caption = "Key Percentiles of Fish Length (mm)",  digits = 1 )  Key Percentiles of Fish Length (mm)   |  | Percentile | Value | | --- | --- | --- | | 10% | 10th | 251.1 | | 25% | 25th (Q1) | 270.8 | | 50% | 50th (Median) | 324.5 | | 75% | 75th (Q3) | 377.0 | | 90% | 90th | 408.6 | |

### Note you could add a box plot by lake to see this if you wanted

grayling\_df %>%   
 ggplot(aes(lake, total\_length\_mm, color= lake))+  
 geom\_boxplot()



|  |
| --- |
| Exercise 5: Calculate the Coefficient of Variation |
| The coefficient of variation (CV) is the standard deviation expressed as a percentage of the mean:  # Calculate coefficient of variation cv\_length <- sd\_length / mean\_length \* 100 cat("Coefficient of variation:", round(cv\_length, 1), "%\n")  Coefficient of variation: 20 %  # Calculate by lake grayling\_df %>%  group\_by(lake) %>%  summarise(  mean\_length = mean(total\_length\_mm),  sd\_length = sd(total\_length\_mm),  cv\_length = sd\_length / mean\_length \* 100  ) %>%  kable(caption = "Coefficient of variation by lake", digits = 1)  Coefficient of variation by lake   | lake | mean\_length | sd\_length | cv\_length | | --- | --- | --- | --- | | I3 | 265.6 | 28.3 | 10.7 | | I8 | 362.6 | 52.3 | 14.4 | |

### Questions to Consider:

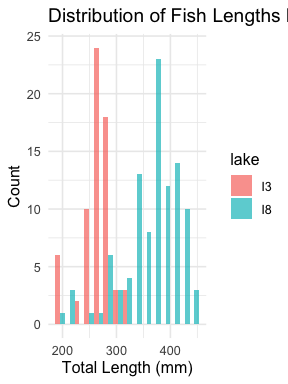
1. How do the means and medians compare within each lake? What might this tell you about the distribution?
2. Which lake has more variable fish lengths? How can you tell?
3. Why might the coefficient of variation be useful when comparing variability between different measurements (e.g., length vs. mass)?

## Part 3: Visualizing Distributions

Visualizations can help us better understand the descriptive statistics we’ve calculated.

|  |
| --- |
| Exercise 6: Creating Histograms |
| One of the best ways to look at data is a histogram - and we will do it again  # Create a histogram of all fish lengths ggplot(grayling\_df, aes(x = total\_length\_mm)) +  geom\_histogram(bins = 15, fill = "steelblue", color = "white") +  labs(  title = "Distribution of Fish Lengths",  x = "Total Length (mm)",  y = "Count"  ) |

# Create histograms by lake  
ggplot(grayling\_df, aes(x = total\_length\_mm, fill = lake)) +  
 geom\_histogram(bins = 15, position = "dodge", alpha = 0.7) +  
 labs(  
 title = "Distribution of Fish Lengths by Lake",  
 x = "Total Length (mm)",  
 y = "Count"  
 )



|  |
| --- |
| Exercise 7: Creating Box Plots |
| Personally I like box plots  # Create a box plot comparing fish lengths by lake # Create a box plot comparing fish lengths by lake ggplot(grayling\_df, aes(x = lake, y = total\_length\_mm, fill = lake)) +  geom\_boxplot() +  labs(  title = "Box Plot of Fish Lengths by Lake",  x = "Lake",  y = "Total Length (mm)"  ) |

|  |
| --- |
| Exercise 9: Creating Density Plots |
| Now these will be really important later on  ## Create density plots ggplot(grayling\_df, aes(x = total\_length\_mm, fill = lake)) +  geom\_density(alpha = 0.5) +  labs(  title = "Density Plot of Fish Lengths by Lake",  x = "Total Length (mm)",  y = "Density"  ) |

### Questions to Consider:

1. Which visualization best shows the differences in fish lengths between lakes?
2. What can you learn from the violin plots that might not be apparent from the box plots?
3. How would you interpret the cumulative frequency distribution?
4. What patterns or insights can you identify from these visualizations?

## Part 4: Interpreting the Results

Based on our analysis, we can make the following observations:

1. **Lake Differences**: Fish from Lake I8 are generally larger than those from Lake I3, both in length and mass.
2. **Variability**: Lake I8 shows greater variability in fish lengths and masses than Lake I3, as indicated by higher standard deviations and IQRs.
3. **Distribution Shape**:
   * Lake I3 fish lengths are more symmetrically distributed.
   * Lake I8 fish lengths show a slight negative skew, suggesting a few smaller fish pulling the distribution to the left.
4. **Length-Mass Relationship**: Both lakes show a strong positive correlation between fish length and mass, following an approximately cubic relationship (mass increases with the cube of length).

## Guided Questions for Deeper Understanding of descriptive statistics

1. **Biological Interpretation**: What ecological factors might explain the differences in fish size between the two lakes?
2. **Statistical Reasoning**: Why might we prefer to use the median and IQR instead of the mean and standard deviation in some cases?
3. **Data Visualization**: Which visualization method was most effective for comparing the two lakes? Why?
4. **Scientific Communication**: How would you concisely summarize these findings in a scientific paper?
5. **Further Analysis**: What additional analyses might be useful to better understand this dataset?