

Lecture 03

Bill Perry

Lecture 2: Review of data and graphing

- We covered
- How to design a well-organized project
- How to implement good naming conventions
 - Controlled vocabulary
 - Including units in names
- Create and use metadata effectively
- Build tidy, well-structured spreadsheets
- Understand data repositories
- Create effective visualizations with ggplot2

These are variables - do you know what they mean?

TGW - yep its a thing

ODO - what do you think it is?

NO3 - what is it? Are you sure? Why might you get in legal trouble if you used this?

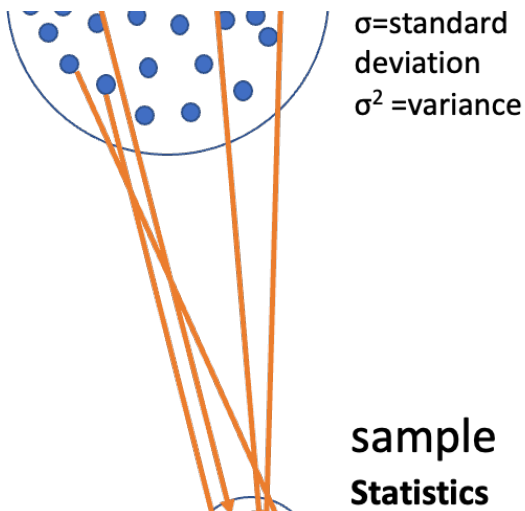


Lecture 3: Descriptive Statistics and Uncertainty in R and Tidyverse

The objectives:

- Understand why statistics is vital in biology
- Distinguish between different types of biological variables
- Learn about accuracy, precision, and bias in measurements
- Calculate and interpret measures of central tendency (mean, median, geometric mean)
- Calculate and interpret measures of spread (standard deviation, variance, IQR)
- Understand data transformations for skewed distributions
- Visualize descriptive statistics for our data
- Learn how to handle uncertainty in our data

We'll use a dataset on grayling fish from two different lakes to explore these concepts..



Lecture 3: Why Statistics is Vital in Biology

Biology is fundamentally different from fields like physics in that:

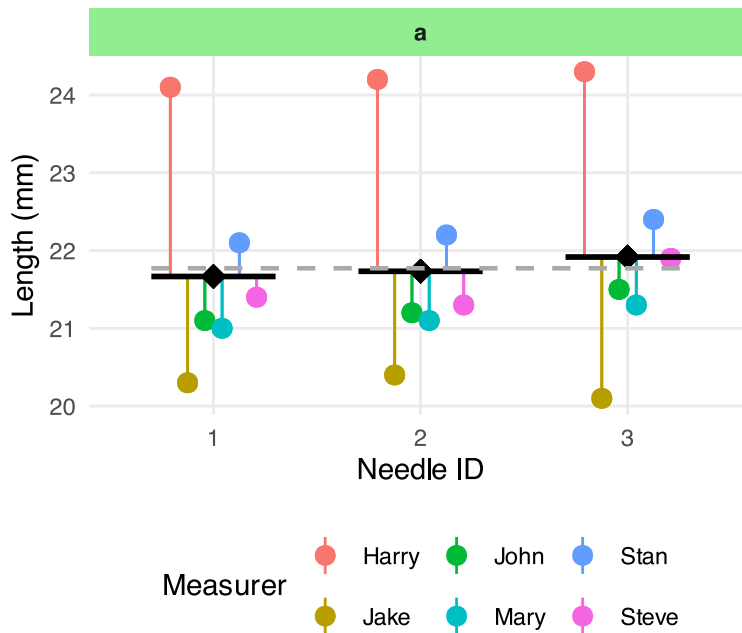
- Most biological phenomena are **probabilistic** rather than **deterministic**
 - Responses occur with some characteristic probability, not with certainty
- All biological material varies, which is essential for evolution (recall Darwin's postulates):
 - Variation exists within populations
 - Some variation is heritable
 - Some heritable variation affects survival/reproduction
- Environmental conditions (in nature, lab, or greenhouse) always vary
- Measurements include error
- Multiple unmeasured causal factors influence nearly all biological systems

Statistics helps us understand biological processes in this variable world by:

1. Condensing variation into summary form (Descriptive statistics)
2. Testing whether observations are consistent with predictions (Inferential statistics)

Pine Needle Length by Species and Measur

Points show individual measurements, lines connect



Practice Exercise 1: Pine Data Analysis

💡 Practice Exercise 1: Can you do this for the pine data we have collected?

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!
```

Lecture 3: Populations and Samples

Before we dive into descriptive statistics, let's clarify some fundamental concepts:

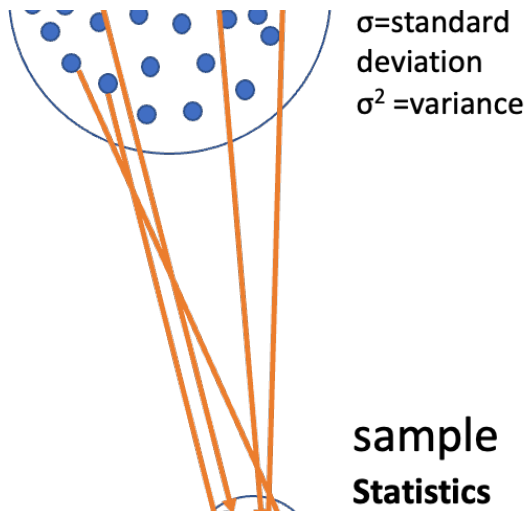
- **Population:** The entire group of things under consideration; the group for which answers obtained from measurements and statistical analysis are pertinent.
- **Sample:** A subset of the population that is actually measured.
- **Sample unit:** The individual thing drawn from the population.

Types of populations:

- **Observational population:** Usually finite but may be very large (e.g., head width of all corn earworms in a field)
- **Experimental population:** Often conceptually infinite (e.g., all possible goldenrod plants that could receive a specific fertilizer treatment)

Sampling involves

- **inference** - generalizing from what is observed in the sample to what is present in the population.
- Valid inference requires **random sampling**.



Lecture 3: Parameters vs. Statistics

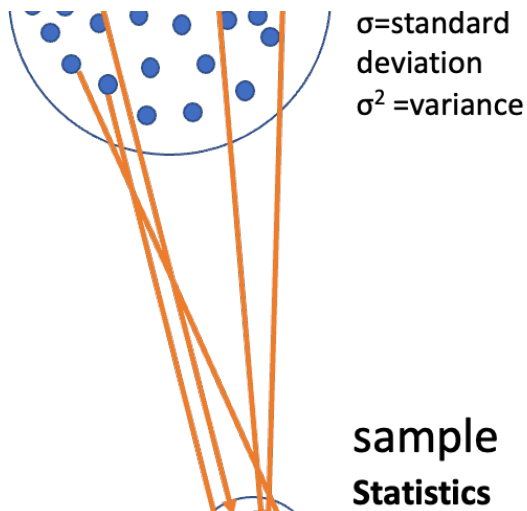
It's important to distinguish between:

- **Parameters:** True numerical values for a population (usually denoted by Greek letters)
- **Statistics:** Estimates of parameters based on samples (usually denoted by Roman letters)

For example:

- Population mean (μ) is estimated by sample mean (\bar{Y})
- Population standard deviation (σ) is estimated by sample standard deviation (s)

The standard deviation formula above includes $n-1$ in the denominator (rather than n) to provide an unbiased estimate of the population parameter.



Lecture 3: Kinds of Biological Variables

Understanding the type of variable you're working with is essential for selecting appropriate statistics:

Measurement or Quantitative Variables

- **Continuous:** Any value between extremes of scale is possible (e.g., mass, length)
- **Discrete (meristic):** Only fixed values (usually integers) between extremes are possible (e.g., bristle number, egg count)

Rank Variables (Ordinal)

- Assign only order, not quantity

- Nothing implied about relative distance between values

Categorical Variables (Qualitative)

- No quantitative information (e.g., male/female, living/dead)
- Some are simplifications of quantitative variables (e.g., color instead of wavelength)


Lecture 3: Derived Variables

Derived Variables

- **Percentages, Proportions:** Ratio of some component to total
- **Ratios:** Relation of two variables
- **Rates:** Quantity per unit (time, mass, etc.)
- **Indices:** More complex derived variables (e.g., condition index)

Let's explore our grayling fish dataset and identify the types of variables it contains.

Practice Exercise 2: Examining Grayling Data

 Practice Exercise 2: Can you do this for the pine data we have collected?

Let's examine the different data and determine what they are?

```
# 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")

# Take a look at the first few rows
head(grayling_df)
```

```
# A tibble: 6 × 5
  site lake species      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
```

Lecture 3: Accuracy, Precision, and Bias

When taking biological measurements, understanding measurement quality is essential:

- **Accuracy:** Closeness of measured value to true value
- **Precision:** Closeness of repeated measurements to each other (repeatability)
- **Bias:** Systematic departure from the true value

Accuracy is a function of both precision and bias. For statisticians, bias is usually a more serious problem than low precision because:

- It's harder to detect (true value usually unknown)

- Low precision can be compensated for by increased sample size



Not Accurate
Precise



Not Accurate
Not Precise



Practice Exercise: Sources of Error

💡 Practice Exercise 1: What are potential sources of error in pine needles or fish?

For our grayling data, potential sources of measurement error might include:

- Precision issues:
 - Variations in how fish are measured (e.g., slightly bent fish)
- Bias issues:
 - Systematic underestimation of length if measurements aren't taken from the true tip of the snout to the end of the tail
- Accuracy issues? what could they be?

Lecture 3: Measures of Central Tendency - Mean

The two most common measures of central tendency are the **mean** and the **median**.

The Arithmetic Mean The arithmetic mean is the average of a set of measurements:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

Where:

- Y_i represents each individual measurement
- n is the total number of observations

Mean length of all fish: 324.5 mm

lake	mean_length
I3	265.6061
I8	362.5980

Lecture 3: Measures of Central Tendency - Median

The Median

- The median is the middle value of a sorted dataset.
- If there is an even number of observations, it's the average of the two middle values.

Median length of all fish: 324.5 mm

lake	median_length
i3	266
i8	373

Lecture 3: Measures of Spread - Variance and Standard Deviation

The spread of a distribution tells us how variable the measurements are.

Variance and Standard Deviation

The variance is

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

The standard deviation is the square root of variance

- measures how far observations typically are from the mean and are in the units of the mean:

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

Variance of length: 4225.9 mm²

Standard deviation of length: 65 mm

lake	var_length	sd_length
i3	801.104	28.30378
i8	2,739.371	52.33901

Lecture 3: Understanding Standard Deviation

The area under the curve of a bell shaped curve within + and - 2 Standard deviations on each side includes about 95% of the data

i3 Lake Fish Length Summary:

Number of fish: 66

Mean length: 265.61 mm

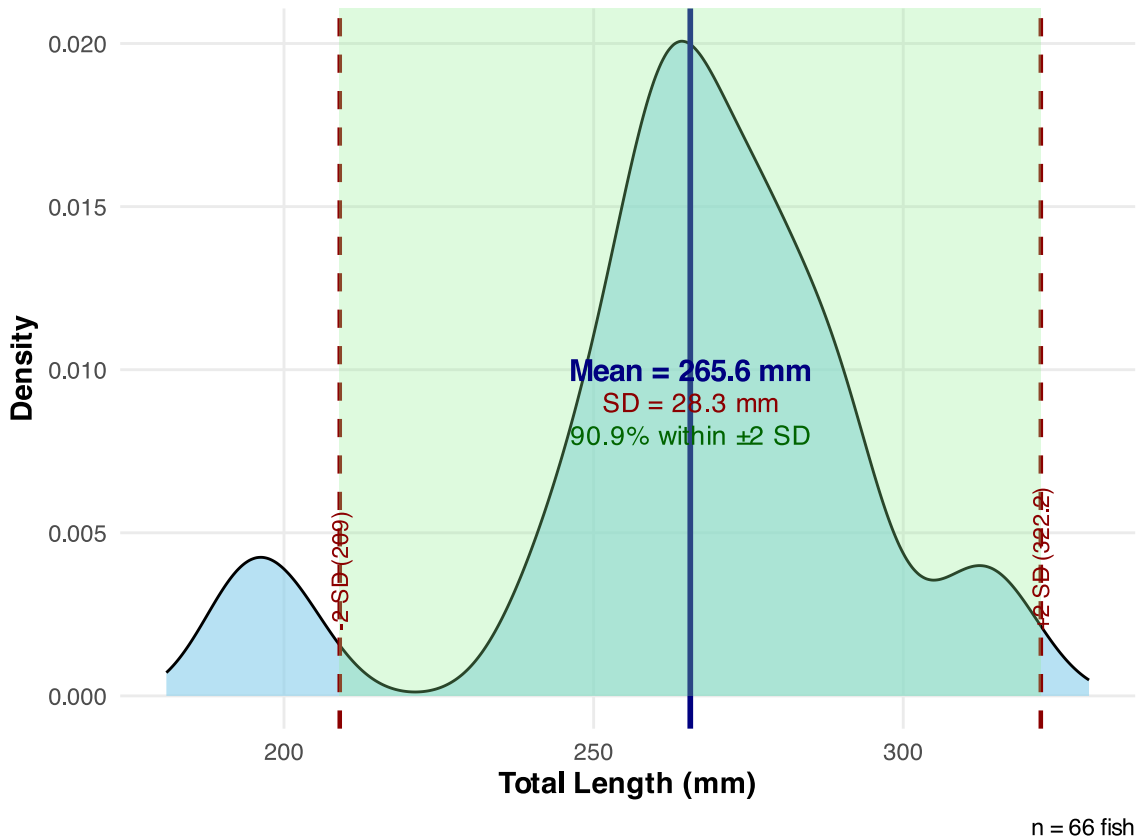
Standard Deviation: 28.3 mm

Range for ± 2 SD: 209 to 322.21 mm

Percentage within ± 2 SD: 90.91 %

Distribution of Fish Lengths in i3 Lake

Area between dashed lines represents ± 2 standard deviations from the mean



Lecture 3: Coefficient of Variation

The coefficient of variation (CV) expresses the standard deviation as a percentage of the mean:

$$CV = \frac{s}{\bar{Y}} \times 100\%$$

This is useful for comparing the variability of measurements with different units or vastly different scales.

Coefficient of variation: 10.7 %

lake	cv_length
I3	10.65630
I8	14.43444

Lecture 3: Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the data:

$$IQR = Q_3 - Q_1$$

Where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile).

First quartile: 270.75 mm

Third quartile: 377 mm

Interquartile range: 106.25 mm

lake	q1	q3	iqr
l3	256	280	24
l8	340	401	61

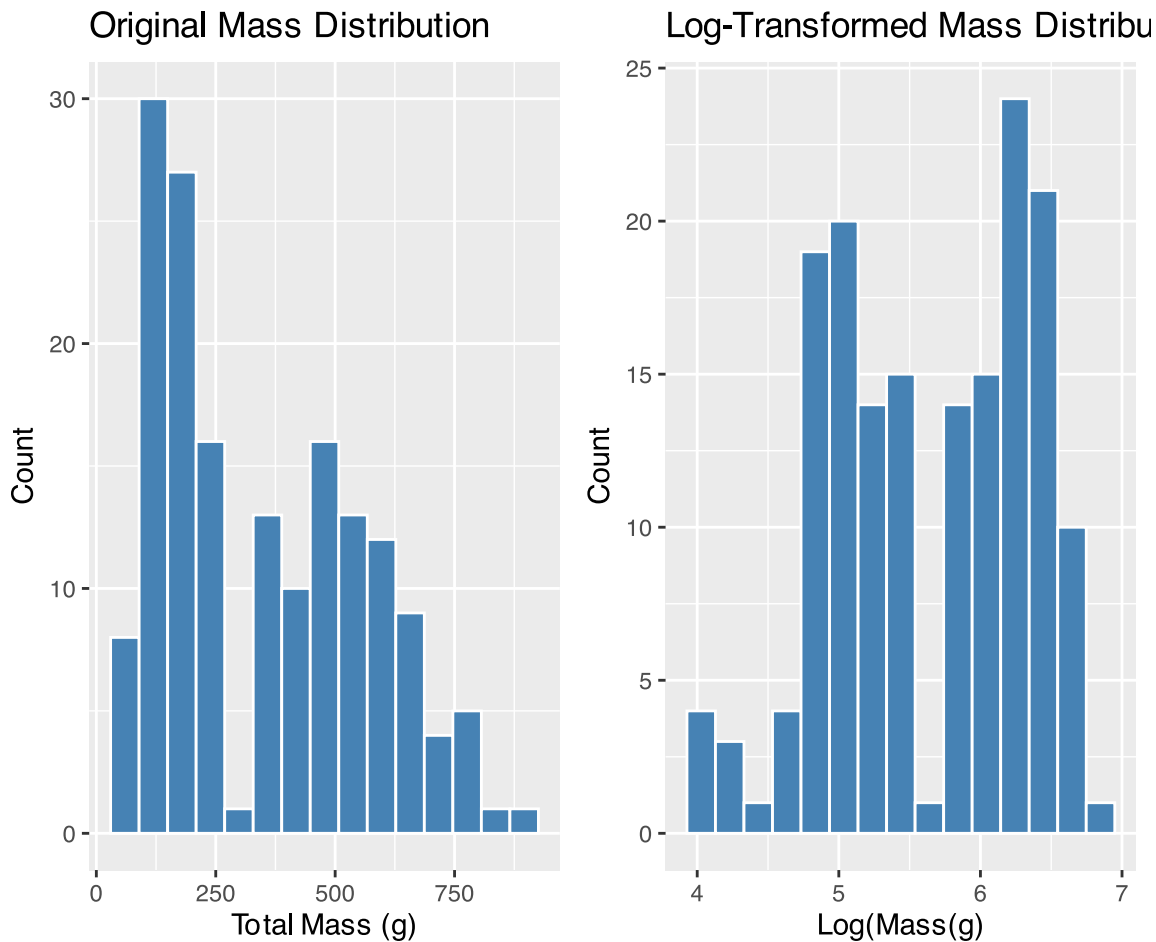
Lecture 3: Data Transformations for Skewed Distributions

Biological data are often skewed (asymmetrical), which can make the arithmetic mean less representative of central tendency. Data transformations can help address this issue.

Logarithmic Transformation

The logarithmic transformation is one of the most common for right-skewed biological data:

When data are log-normally distributed, the geometric mean often provides a better measure of central tendency than the arithmetic mean.

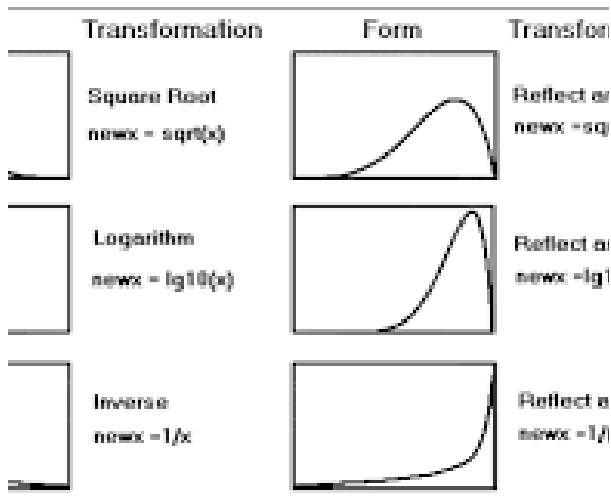


Arithmetic mean of original data: 265.6 mm

Geometric mean (back-transformed mean of logs): NA mm

Lecture 3: When to Use Transformations

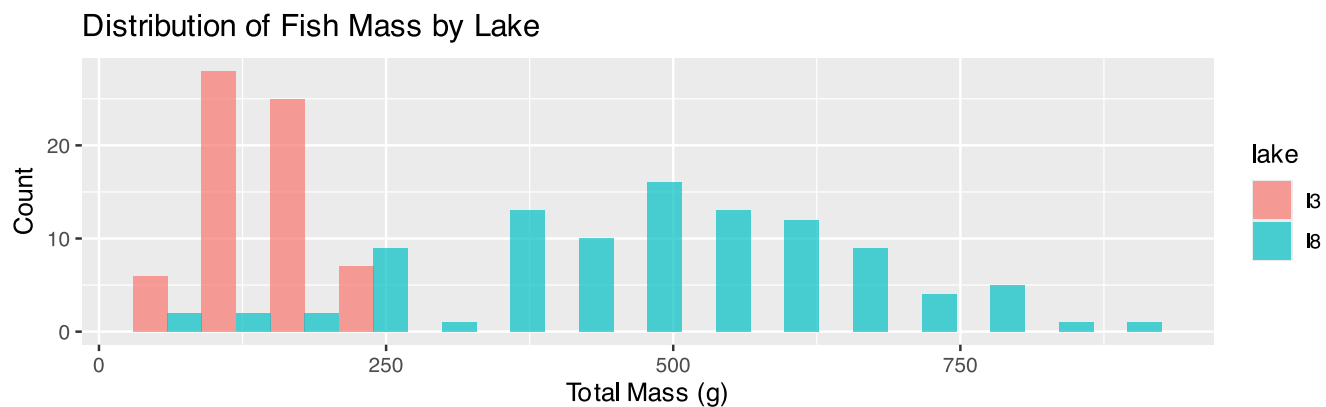
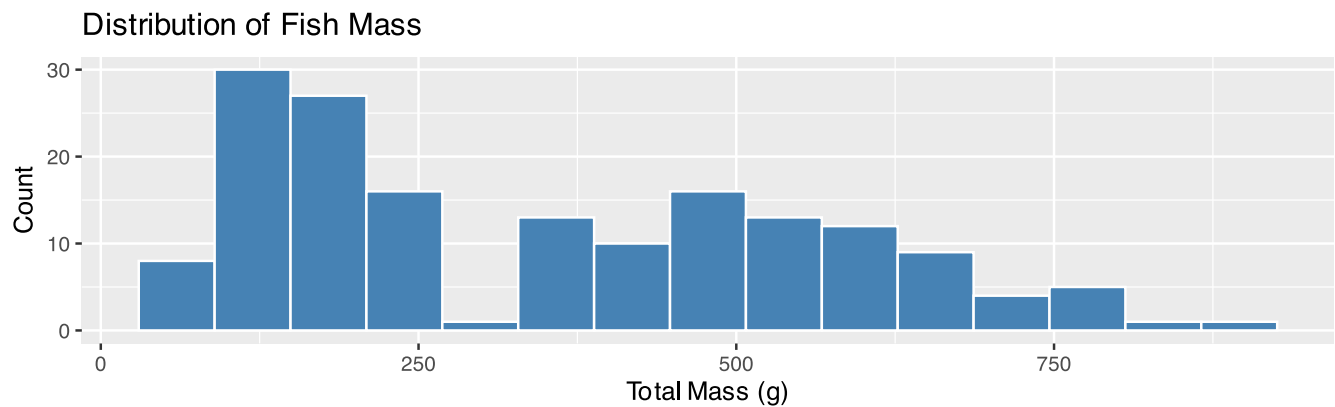
- **Log transformation:** When data are right-skewed or follow multiplicative rather than additive processes
- **Square root transformation:** For count data or data where variance increases with the mean
- **Inverse transformation:** For strongly right-skewed data
- **Arcsine square root transformation:** For proportions or percentages (though logistic regression is often preferred now)



Lecture 3: Visualizing Distributions - Histograms

Histograms

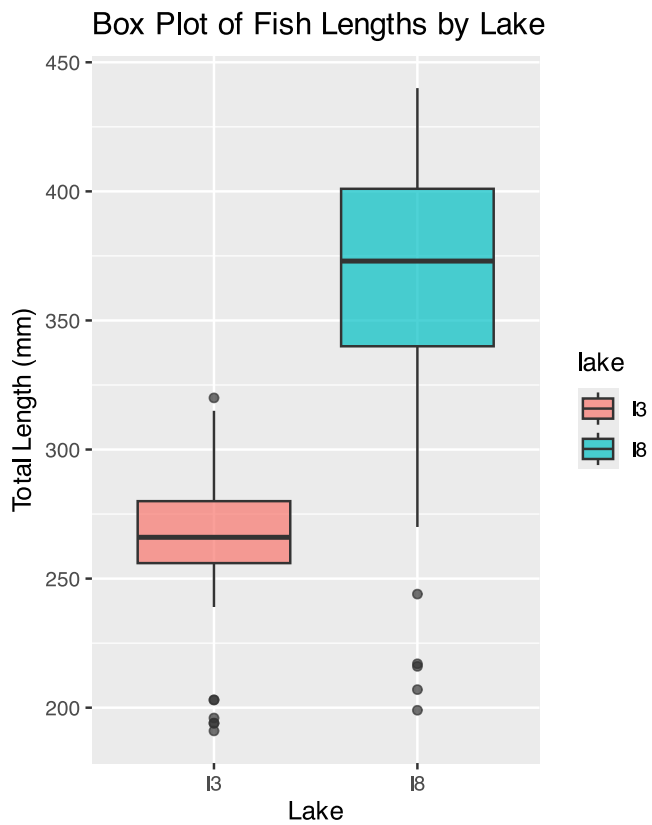
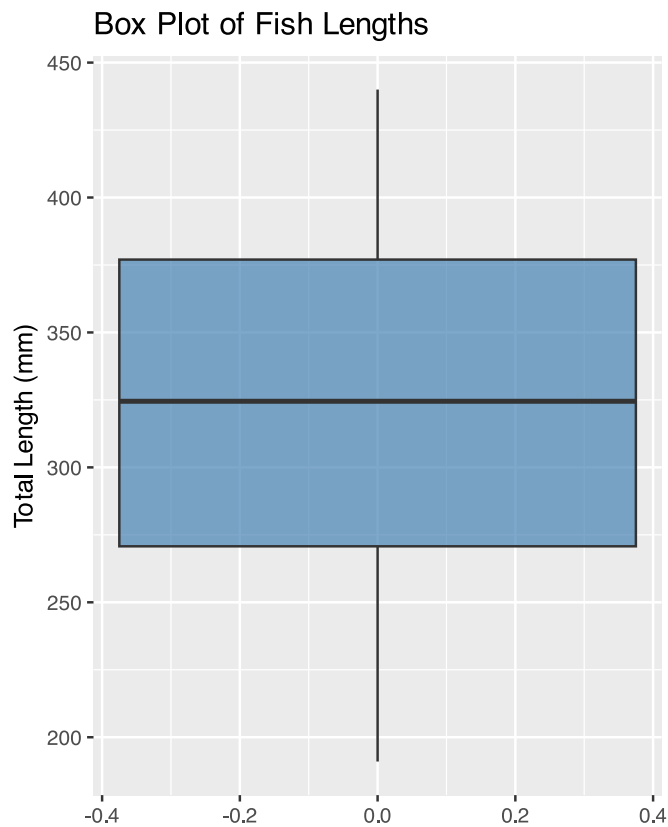
Histograms show the frequency distribution of our data.



Lecture 3: Visualizing Distributions - Box Plots

Box Plots

Box plots show the median, quartiles, and potential outliers.



Lecture 3: Comparing Mean vs. Median

The mean and median measure different aspects of a distribution:

Mean: Center of gravity of the distribution

Median: Middle value of the data

When a distribution is symmetric, the mean and median are similar. When it's skewed or has outliers, they can differ significantly.

lake	mean	median	sd	iqr	skewness
l3	265.6061	266	28.30378	24	-0.8826195
l8	362.5980	373	52.33901	61	-1.0909961

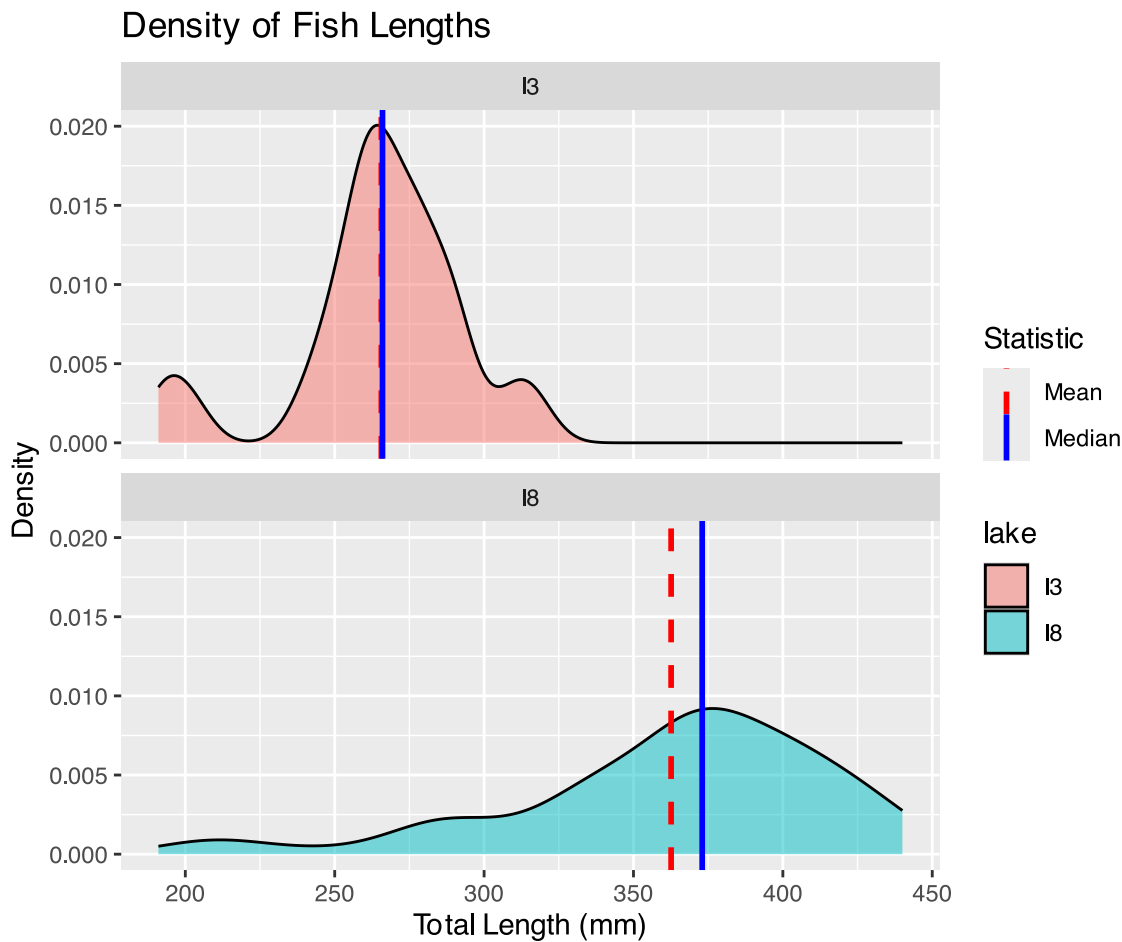
Lecture 3: Density Plot - Mean vs. Median

The mean and median measure different aspects of a distribution:

Mean: Center of gravity of the distribution

Median: Middle value of the data

When a distribution is symmetric, the mean and median are similar. When it's skewed or has outliers, they can differ significantly.



Lecture 3: Standard Deviation vs. Interquartile Range

The standard deviation and interquartile range both measure spread, but:

Standard deviation: Sensitive to outliers

Interquartile range: Robust against outliers

When the data is approximately normal, the $IQR \approx 1.35 \times \text{standard deviation}$.

lake	sd	iqr	ratio_iqr_sd
I3	28.30	24.00	0.85
I8	52.34	61.00	1.17

Lecture 3: Understanding Percentiles

Percentiles are values that divide a dataset into 100 equal parts.

The 25th percentile is the first quartile (Q1)

The 50th percentile is the median

The 75th percentile is the third quartile (Q3)

The IQR is the difference between Q3 and Q1.

Percentile	Value
10th	251.1
25th (Q1)	270.8
50th (Median)	324.5
75th (Q3)	377.0
90th	408.6

Lecture 3: Handling Missing Values

Let's examine how missing values affect our descriptive statistics by looking at the mass variable, which has some missing data.

```
[1] 2
```

```
Mean mass without handling NAs: NA g
```

```
Mean mass with na.rm=TRUE: 351.2289 g
```

lake	mean_mass	median_mass	sd_mass	n_missing
I3	150.5	147.0	42.2	0
I8	483.7	490.0	176.5	2

Lecture 3: Best Practices for Missing Values

1. Always check for missing values in your data before calculating statistics.
2. Use `na.rm = TRUE` when calculating summary statistics to handle missing values.
3. Report the number of missing values along with your statistics.
4. Consider whether the missing values are random or might introduce bias.

Sampling from a Population

Now that we have estimates of the sample we need to relate that to the population

In reality, we rarely know the true population parameters. When studying fish in lakes I3 and I8:

- The **population** includes all grayling fish in each lake
- The true population mean (μ) and standard deviation (σ) are unknown
- Our dataset is a **sample** from this population
- We use the sample mean (\bar{x}) to estimate μ
- Sampling introduces random variation in our estimates

Let's demonstrate how different samples from the same population can give different estimates.

If we could sample all fish in the lake, we would know the true mean length. But that's usually impossible in ecology!

Demonstrating Sampling Variation

Let's take several random samples from Lake I3 and see how the sample means vary:

```

# Filter for Lake I3
i3_data <- grayling_df %>% filter(lake == "I3")

# Function to take a random sample and calculate the mean
sample_mean <- function(data, sample_size) {
  sample_data <- sample_n(data, sample_size)
  return(mean(sample_data$length_mm))
}

# Take 10 different samples of size 15 from Lake I3
set.seed(123) # For reproducibility
sample_size <- 15
sample_means <- replicate(10, sample_mean(i3_data, sample_size))

# Create a data frame with sample numbers and means
samples_df <- data.frame(
  sample_number = 1:10,
  sample_mean = sample_means
)

```

Plotting Sample Variation

```

# Display the sample means
samples_df

```

	sample_number	sample_mean
1	1	269.9333
2	2	260.6000
3	3	255.2000
4	4	263.4000
5	5	275.3333
6	6	279.2667
7	7	263.7333
8	8	273.6000
9	9	264.8000
10	10	269.8667

```

# Calculate the mean and standard deviation of the sample means
mean(sample_means)

```

```
[1] 267.5733
```

```
sd(sample_means)
```

```
[1] 7.346063
```

```

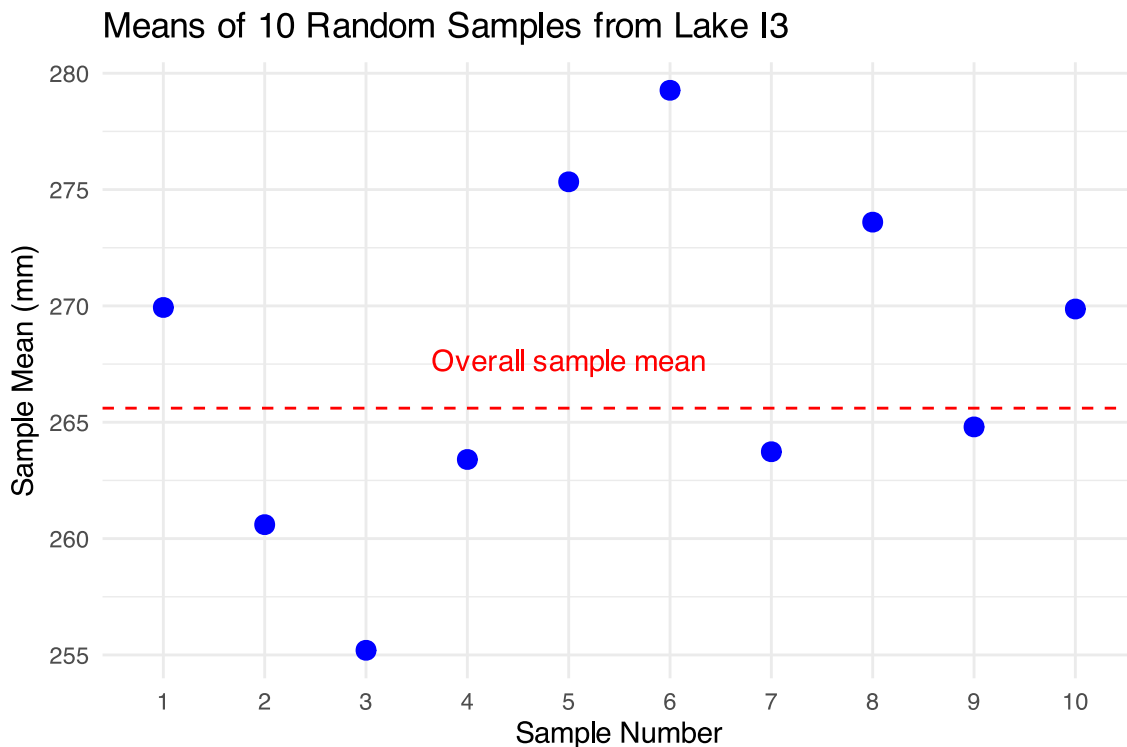
# Plot the different sample means
ggplot(samples_df, aes(x = factor(sample_number), y = sample_mean)) +
  geom_point(size = 3, color = "blue") +
  geom_hline(yintercept = mean(i3_data$length_mm),

```

```

linetype = "dashed", color = "red") +
annotate("text", x = 5, y = mean(i3_data$length_mm) + 2,
        label = "Overall sample mean", color = "red") +
labs(title = "Means of 10 Random Samples from Lake I3",
     x = "Sample Number",
     y = "Sample Mean (mm)") +
theme_minimal()

```



Notice how each sample's mean differs from the overall mean. This demonstrates sampling variation.

Standard Error: Quantifying Uncertainty

The **standard error of the mean (SEM)** measures the precision of a sample mean as an estimate of the population mean.

Formula: $SE_x = \frac{s}{\sqrt{n}}$

Where: - s is the sample standard deviation - n is the sample size

The standard error tells us: - How much uncertainty is in our estimate - How much sample means are expected to vary - How close our sample mean is likely to be to the true population mean

Remember: - Standard deviation (s) describes the variability in the individual data points - Standard error (SE) describes the variability in the sample mean itself - As sample size increases, SE decreases (more precise estimate)

Standard Error for Our Grayling Data

Let's calculate and visualize the standard error for both lakes:

```

# Calculate mean, SD, and SE for each lake
grayling_stats <- grayling_df %>%
  group_by(lake) %>%
  summarize(

```



```

mean_length = mean(length_mm),
sd_length = sd(length_mm),
n = n(),
se_length = sd_length / sqrt(n)
)

```

```

# Display the statistics
grayling_stats

```

```

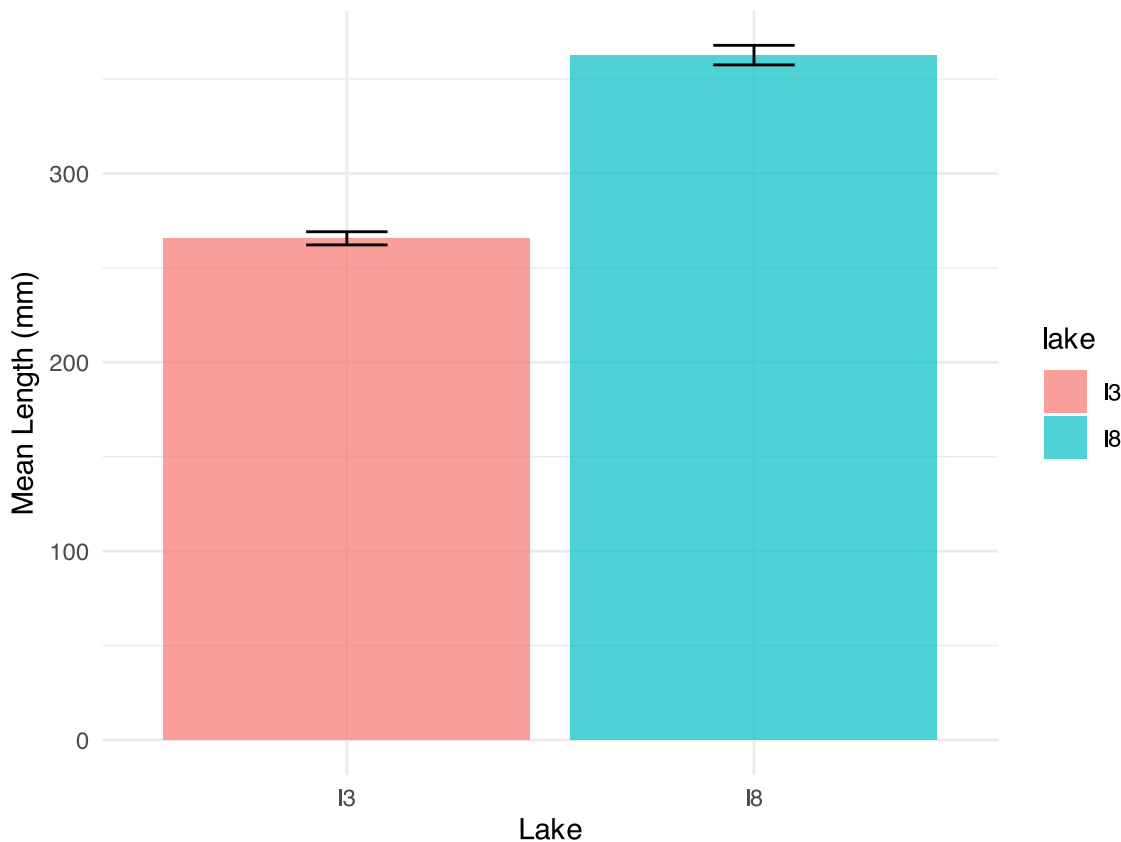
# A tibble: 2 × 5
  lake mean_length sd_length    n se_length
<chr>    <dbl>    <dbl> <int>    <dbl>
1 I3      266.     28.3    66     3.48
2 I8      363.     52.3   102     5.18

```

Visualizing Standard Error

Mean Fish Length by Lake with Standard Error

Error bars represent ± 1 standard error



Sampling Distribution of the Mean

The **sampling distribution of the mean** is the theoretical distribution of all possible sample means of a given sample size from a population.

Important properties: 1. It is centered at the population mean (μ) 2. Its standard deviation is the standard error (σ/\sqrt{n}) 3. For large sample sizes, it approaches a normal distribution (Central Limit Theorem)

The larger the sample size: - The narrower the sampling distribution - The smaller the standard error - The more precise our estimate of the population mean

Let's simulate the sampling distribution for Lake I3 fish data.

Simulating the Sampling Distribution

Let's simulate taking many samples from Lake I3 to visualize the sampling distribution:

```
# Filter for Lake I3
i3_data <- grayling_df %>% filter(lake == "I3")

# Number of samples to simulate
num_simulations <- 1000
sample_size <- 20

# Simulate many samples and calculate means
set.seed(456) # For reproducibility
simulated_means <- replicate(num_simulations, sample_mean(i3_data, sample_size))

# Calculate the mean and standard deviation of the simulated means
mean_of_means <- mean(simulated_means)
sd_of_means <- sd(simulated_means)

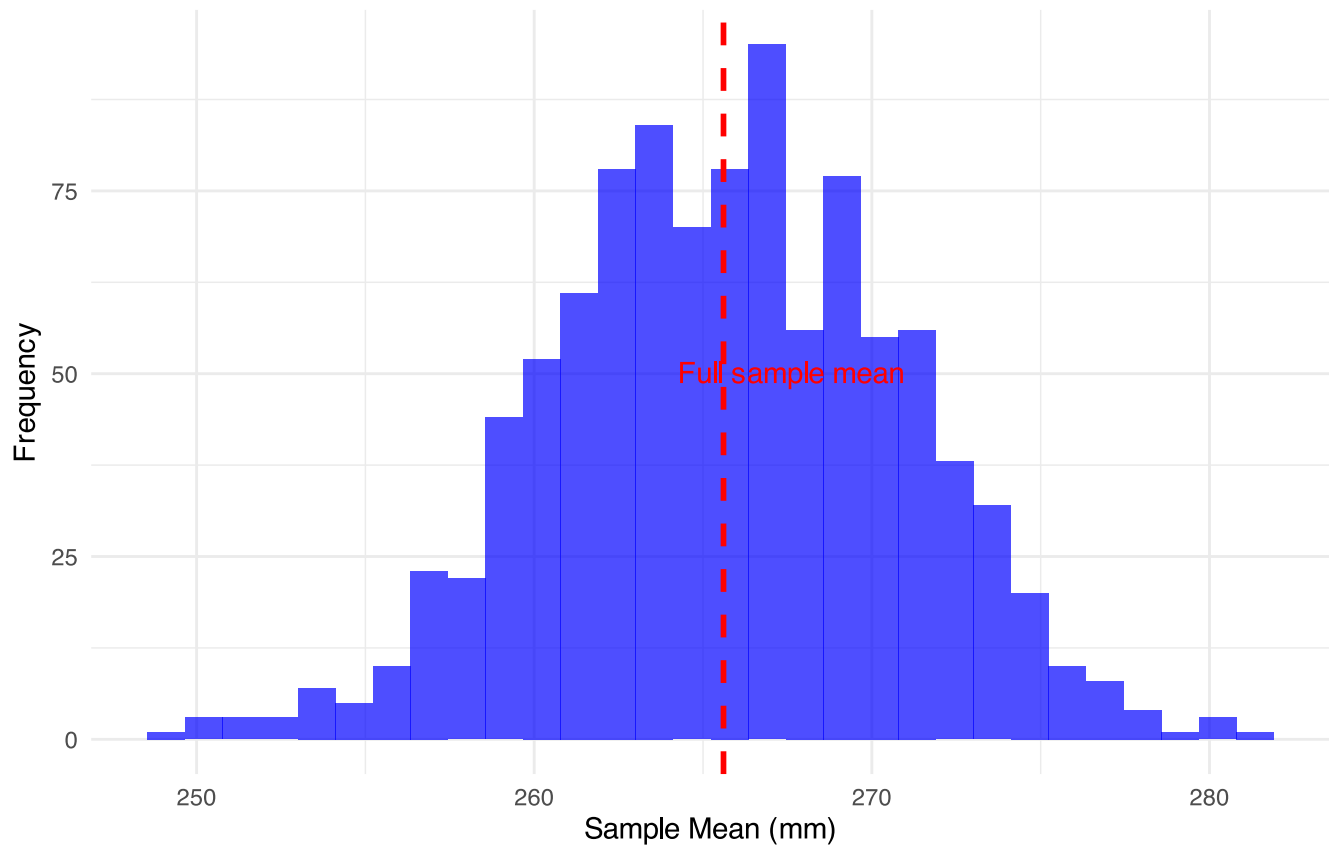
# Create a data frame with the simulated means
simulated_df <- data.frame(sample_mean = simulated_means)
```

Plotting Sampling Distribution

```
# Plot the sampling distribution
ggplot(simulated_df, aes(x = sample_mean)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  geom_vline(xintercept = mean(i3_data$length_mm),
             linetype = "dashed", color = "red", size = 1) +
  annotate("text", x = mean(i3_data$length_mm) + 2, y = 50,
          label = "Full sample mean", color = "red") +
  labs(title = "Simulated Sampling Distribution of the Mean",
       subtitle = paste("Based on", num_simulations, "samples of size", sample_size),
       x = "Sample Mean (mm)",
       y = "Frequency") +
  theme_minimal()
```

Simulated Sampling Distribution of the Mean

Based on 1000 samples of size 20



Notice that the simulated sampling distribution:

1. Is approximately normally distributed
2. Is centered around the overall sample mean
3. Has a spread that is related to the standard error

Standard Error and Sample Size

Let's see how the standard error changes with different sample sizes:

Sample Size vs. Standard Error

```
# Display the results
results
```

	sample_size	empirical_se	theoretical_se
1	5	12.349407	12.657835
2	10	8.178270	8.950441
3	20	5.558957	6.328918
4	30	3.792177	5.167540
5	50	2.099744	4.002759

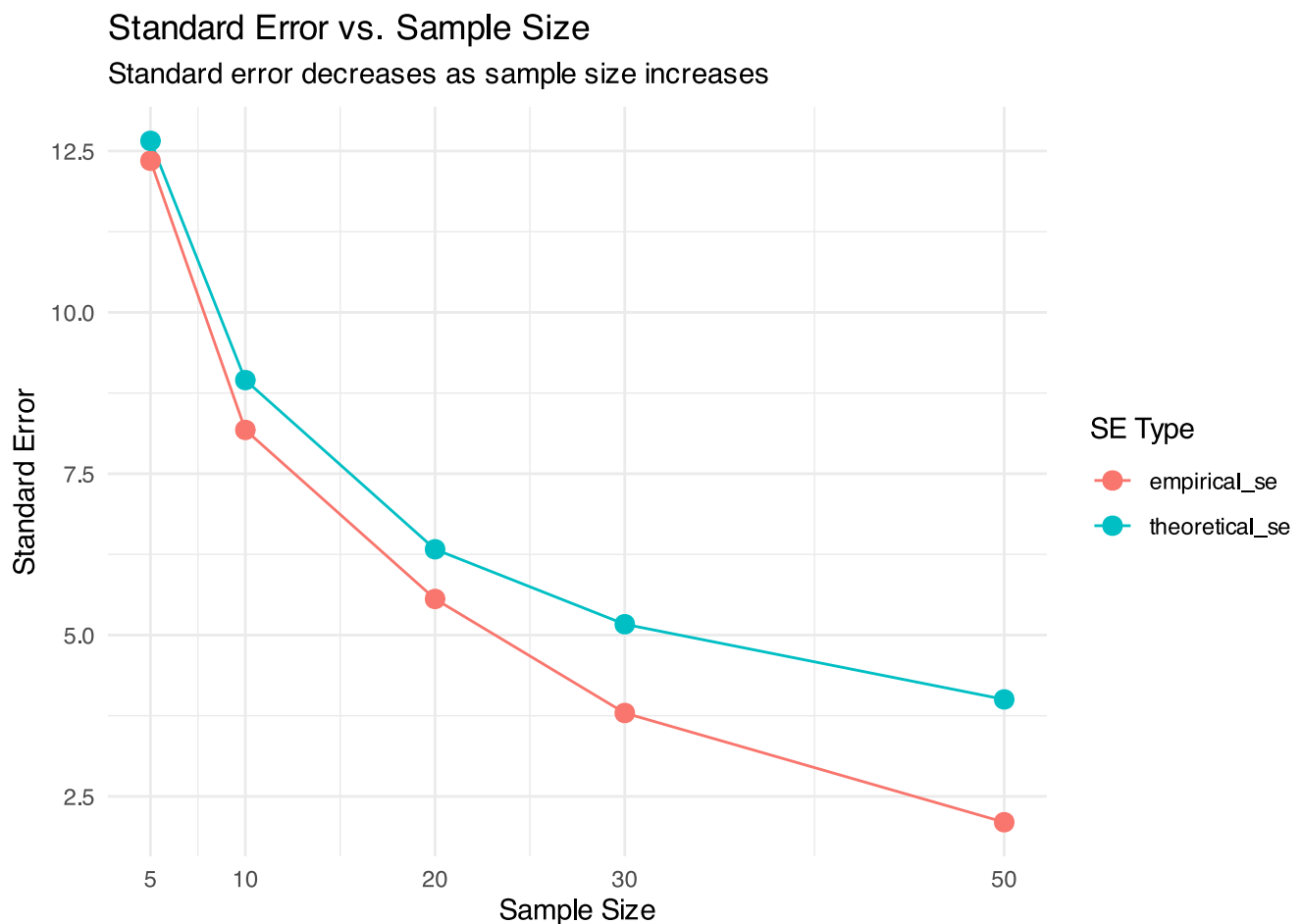
```
# Plot how SE changes with sample size
results_long <- pivot_longer(results,
                             cols = c(empirical_se, theoretical_se),
```

```

names_to = "se_type",
values_to = "standard_error")

ggplot(results_long, aes(x = sample_size, y = standard_error, color = se_type)) +
  geom_line() +
  geom_point(size = 3) +
  scale_x_continuous(breaks = sample_sizes) +
  labs(title = "Standard Error vs. Sample Size",
       subtitle = "Standard error decreases as sample size increases",
       x = "Sample Size",
       y = "Standard Error",
       color = "SE Type") +
  theme_minimal()

```



Confidence Intervals

A **confidence interval** is a range of values that is likely to contain the true population parameter.

The 95% confidence interval for the mean is approximately:

$$\bar{x} \pm 2 \times SE_{\bar{x}}$$

This “2 SE rule of thumb” means: - The interval extends 2 standard errors below and above the sample mean - About 95% of such intervals constructed from different samples would contain the true population mean

Confidence intervals provide a way to express the precision of our estimates.

Calculating Confidence Intervals for Grayling Data

Let's calculate and visualize the 95% confidence intervals for the mean fish length in each lake:

```
# Calculate 95% confidence intervals
grayling_ci <- grayling_df %>%
  group_by(lake) %>%
  summarize(
    mean_length = mean(length_mm),
    sd_length = sd(length_mm),
    n = n(),
    se_length = sd_length / sqrt(n),
    ci_lower = mean_length - 2 * se_length,
    ci_upper = mean_length + 2 * se_length
  )

# Display the confidence intervals
grayling_ci
```

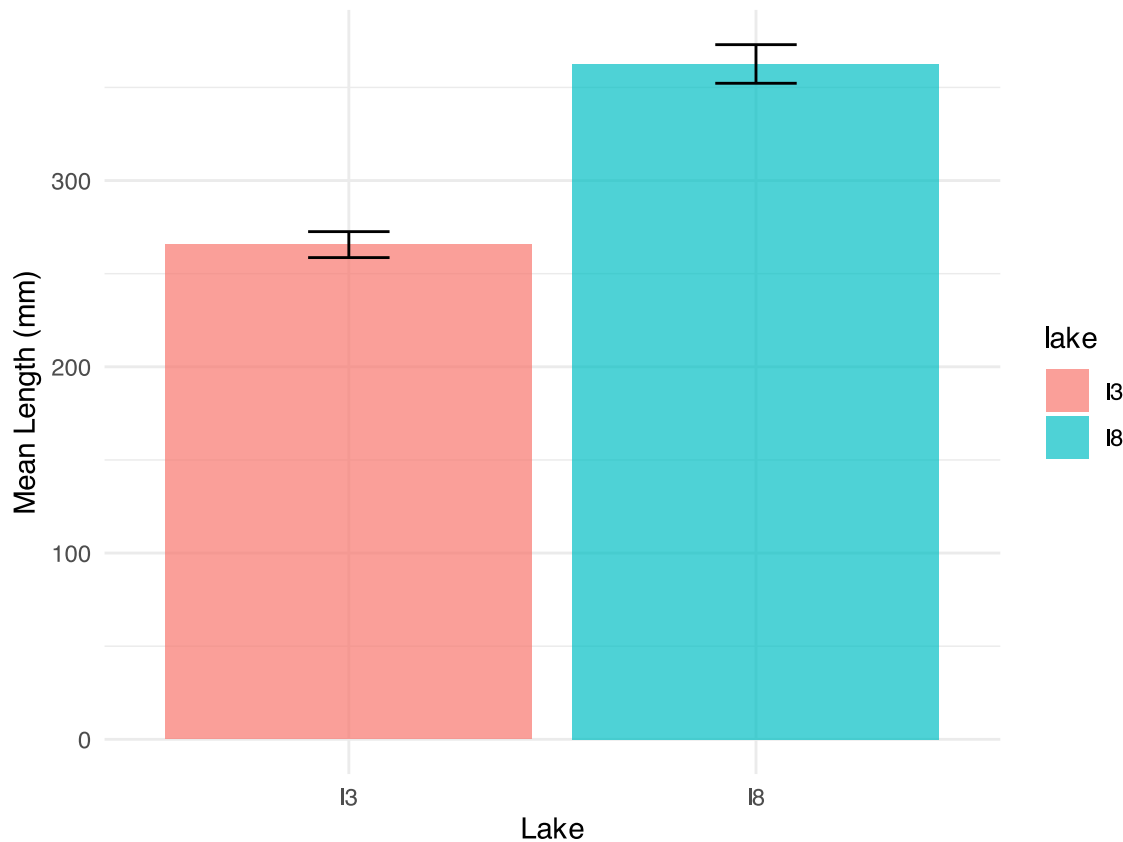
```
# A tibble: 2 × 7
  lake mean_length sd_length    n se_length ci_lower ci_upper
<chr>    <dbl>    <dbl> <int>    <dbl>    <dbl>    <dbl>
1 I3      266.      28.3    66     3.48     259.     273.
2 I8      363.      52.3   102     5.18     352.     373.
```

Visualizing Confidence Intervals

```
# Plot with confidence intervals
ggplot(grayling_ci, aes(x = lake, y = mean_length, fill = lake)) +
  geom_bar(stat = "identity", alpha = 0.7) +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper),
    width = 0.2) +
  labs(title = "Mean Fish Length by Lake with 95% Confidence Intervals",
    subtitle = "Error bars represent 95% confidence intervals",
    x = "Lake",
    y = "Mean Length (mm)") +
  theme_minimal()
```

Mean Fish Length by Lake with 95% Confidence Intervals

Error bars represent 95% confidence intervals



Different Types of Error Bars

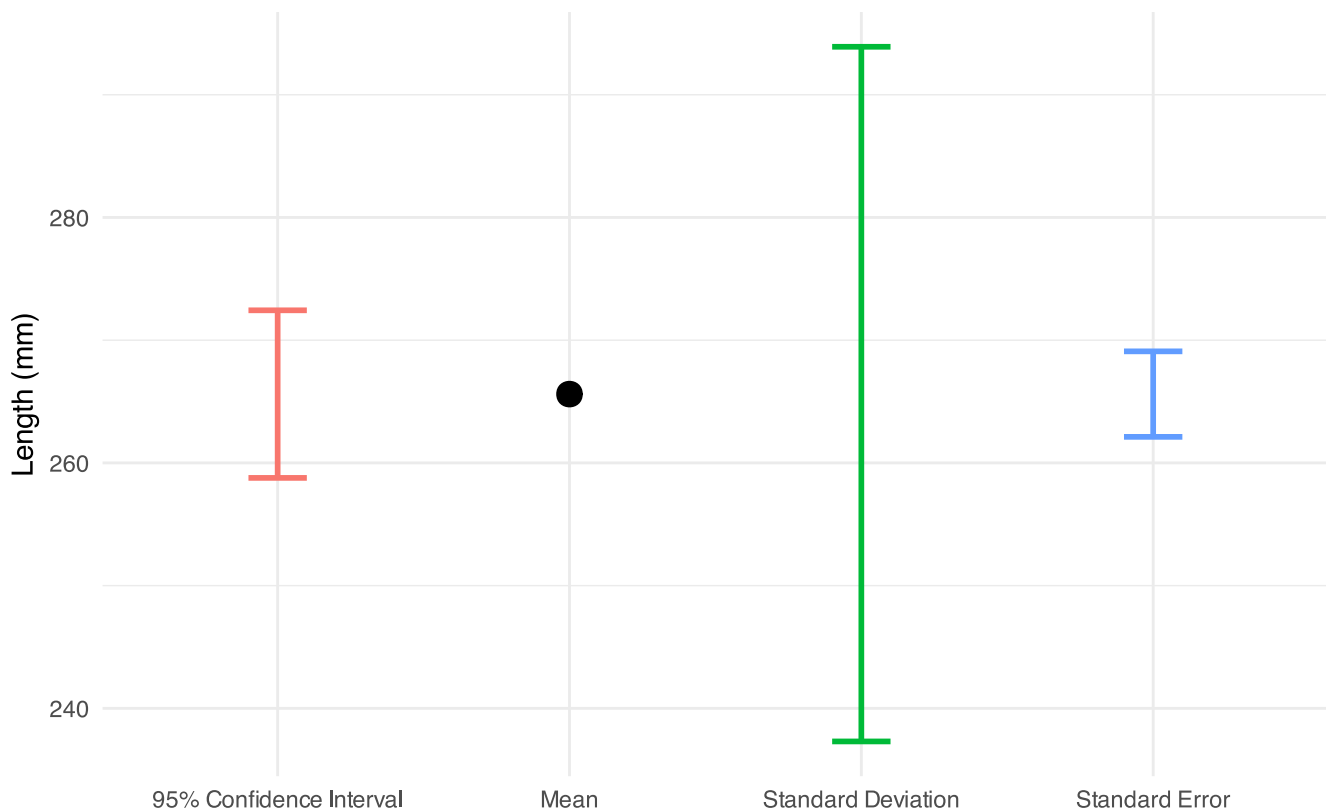
Let's compare different ways of displaying uncertainty in our estimates:

```
# Calculate statistics for different types of error bars
grayling_error_bars <- grayling_df %>% group_by(lake) %>%
  summarize(mean_length = mean(length_mm),
            sd_length = sd(length_mm), n = n(),
            se_length = sd_length / sqrt(n),
            ci_lower = mean_length - 1.96 * se_length,
            ci_upper = mean_length + 1.96 * se_length,
            one_sd_lower = mean_length - sd_length,
            one_sd_upper = mean_length + sd_length)
# Create a data frame for plotting different error types
lake_i3 <- grayling_error_bars %>% filter(lake == "I3")
error_types <- data.frame(
  error_type = c("Standard Deviation", "Standard Error", "95% Confidence Interval"),
  lower = c(lake_i3$one_sd_lower,
            lake_i3$mean_length - lake_i3$se_length,
            lake_i3$ci_lower),
  upper = c(lake_i3$one_sd_upper,
            lake_i3$mean_length + lake_i3$se_length,
            lake_i3$ci_upper))
```

Comparing Error Bar Types

```
# Plot the comparison
ggplot() +
  geom_point(data = lake_i3, aes(x = "Mean",
    y = mean_length), size = 4) +
  geom_errorbar(data = error_types,
    aes(x = error_type, ymin = lower,
      ymax = upper, color = error_type),
    width = 0.2, linewidth = 1) +
  labs(title = "Different Types of Error Bars for Lake I3",
    subtitle = "Comparing SD, SE, and 95% CI",
    x = "",
    y = "Length (mm)",
    color = "Error Bar Type") +
  theme_minimal() +
  theme(legend.position = "none")
```

Different Types of Error Bars for Lake I3
Comparing SD, SE, and 95% CI



Key Takeaways

- The **standard error** measures the precision of a sample statistic as an estimate of a population parameter
- The standard error of the mean decreases as sample size increases: $SE_x = \frac{s}{\sqrt{n}}$
- The **sampling distribution** shows the variation in sample statistics that would be expected due to random sampling
- **Confidence intervals** provide a range of plausible values for the population parameter
- Larger sample sizes provide more precise estimates (narrower confidence intervals)
- When reporting results, always include a measure of precision (SE or

CI)

For Further Practice

- Try calculating the standard error and confidence intervals for other variables in the dataset
- Experiment with different sample sizes to see how they affect the precision of estimates
- Compare the means of the two lakes using confidence intervals - do they overlap?
- Consider how these concepts extend to other statistics beyond the mean

Lecture 3: Conclusion

In this lecture, we've explored:

- Why statistics is essential in biology
- Types of biological variables and their properties
- Accuracy, precision, and bias in measurements
- Measures of central tendency (mean, median, geometric mean)
- Measures of spread (standard deviation, variance, and interquartile range)
- Data transformations for skewed distributions
- Visualization techniques for understanding distributions
- Handling missing values

These tools form the foundation of statistical analysis and will be essential as we move forward to more complex statistical methods.