Name: _____ LA Initials: [ ]

Lab 6: Describing Data with Summary Statistics

Learning Objectives

By the end of this lab, students will be able to:

- Explain how measures of center (mean, median) and spread (SD, SE) summarize different aspects of a distribution
- Explain how mean and median differ in skewed distributions and describe how each summarizes the data
- Create and interpret summary tables of grouped data using descriptive statistics
- Explain the difference between SD and SE in words, and describe what each does and does *not* represent
- Use the 2SE rule of thumb to describe uncertainty in group means, without making inferential claims
- Interpret different graph types used to address distributional versus mean-based questions
- Describe patterns in data clearly using both figures and concise Results-style text

Getting started

Before you begin the lab activities, complete the following steps to make sure you are fully set up.

1. **Get the Lab Worksheet.**

   Pick up a physical copy of the lab worksheet, or print one if you are working outside of class.
   Download Lab Worksheet (PDF, if needed)

2. **Open Posit Cloud and Start Lab 6.**

   Log in to **Posit Cloud**, navigate to **Your Workspace**, and create a new project for Lab 6 Describing Data

3. **Install required packages**

   Install the required packages for this lab, including **tidyverse** and lterdatasampler using the Packages tab or by running the following code ***in the Console***:

   ```
   install.packages(c("tidyverse", "lterdatasampler"))
   ```

4. **Create an R Script**

   Create a new R script and save it as **lab-6-script.R**

5. **Load the packages**

   Copy and paste this code to your R script and run it to load the packages. Remember to run both library commands (Ctrl+Enter on one line, then the next).

   ```
   # load packages -----------------------------------------------------------

   library(tidyverse)
   library(lterdatasampler)
   ```

i Checkpoint

At this point, you should have:

- These instructions open in a web browser.
- Your Lab 6 project open in Posit Cloud in another browser window.
- The required packages installed and loaded without errors
- The Lab 6 worksheet in front of you.

Do not continue until all of the above steps are working correctly.

Overview

In this lab, you will practice describing data, not testing hypotheses. The goal is to learn how to summarize biological data in ways that are meaningful, interpretable, and appropriate for the question being asked.

You will work with numerical variables and use descriptive statistics—such as the mean, median, standard deviation (SD), and standard error (SE)—to describe patterns in the data. You will also use the 2SE rule of thumb as a way to describe uncertainty in estimated means, without making inferential or causal claims.

A key focus of this lab is judgment. Different summaries and graphs answer different questions. Some plots are best for visualizing distributions and identifying skew or outliers, while others are best for comparing group means and their uncertainty. You will see examples of both, and you will practice choosing and interpreting them appropriately.

We will begin with a worked example using a large ecological dataset to walk through the workflow step by step. You will then apply the same ideas to a smaller dataset, creating summary statistics, visualizations, and a short Results-style description of the data. The skills you practice here will be used again in later labs and in your exploratory data analysis (EDA) project.

Worked Example: Describing Data Step by Step

In this section, we will walk through a complete example of how to describe biological data using summary statistics and graphs. The goal is not to learn new plot types, but to see how different graphs emphasize different aspects of the same data.

While group means can be shown on many kinds of plots, some visualizations emphasize distribution shape, skew, and outliers, whereas others emphasize central tendency and uncertainty.

You will also see how plots based on summary statistics often use a much narrower y-axis scale than plots showing raw data, which can make differences in means appear large even when they are small relative to the full range of the data.

Dataset overview: `and_vertebrates`

The `and_vertebrates` dataset contains measurements of vertebrates sampled at multiple stream sections in the Andrews Experimental Forest LTER. In this lab, we will focus on **body length** as the numerical variable of interest, and use **species** and **stream section** as categorical grouping variables. These variables allow us to summarize and visualize the data across groups using descriptive statistics and graphs.

For a full description of the `and_vertebrates` dataset and the other data samples included in the **lterdatasampler** package, see the package website at **https://lterdatasampler.lter.github.io/**. If you want to look up variable names or documentation, you can also use `?and_vertebrates` after loading the package.

To start exploring the data in R, first load the dataset and then use functions like `print()` and `glimpse()` to view its structure and contents. For example:

```
library(lterdatasampler)

print(and_vertebrates)
```

```
# A tibble: 32,209 × 16
    year sitecode section reach  pass unitnum unittype vert_index pitnumber
   <dbl> <chr>    <chr>   <chr> <dbl>   <dbl> <chr>         <dbl>     <dbl>
 1  1987 MACKCC-L CC      L         1       1 R                 1        NA
 2  1987 MACKCC-L CC      L         1       1 R                 2        NA
 3  1987 MACKCC-L CC      L         1       1 R                 3        NA
 4  1987 MACKCC-L CC      L         1       1 R                 4        NA
 5  1987 MACKCC-L CC      L         1       1 R                 5        NA
 6  1987 MACKCC-L CC      L         1       1 R                 6        NA
 7  1987 MACKCC-L CC      L         1       1 R                 7        NA
 8  1987 MACKCC-L CC      L         1       1 R                 8        NA
 9  1987 MACKCC-L CC      L         1       1 R                 9        NA
10  1987 MACKCC-L CC      L         1       1 R                10        NA
# ℹ 32,199 more rows
# ℹ 7 more variables: species <chr>, length_1_mm <dbl>, length_2_mm <dbl>,
#   weight_g <dbl>, clip <chr>, sampledate <date>, notes <chr>
```

```
glimpse(and_vertebrates)
```

```
Rows: 32,209
Columns: 16
$ year        <dbl> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987…
$ sitecode    <chr> "MACKCC-L", "MACKCC-L", "MACKCC-L", "MACKCC-L", "MACKCC-L"…
$ section     <chr> "CC", "CC", "CC", "CC", "CC", "CC", "CC", "CC", "CC", "CC"…
$ reach       <chr> "L", "L", "L", "L", "L", "L", "L", "L", "L", "L", "L", "L"…
$ pass        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ unitnum     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2…
$ unittype    <chr> "R", "R", "R", "R", "R", "R", "R", "R", "R", "R", "R", "R"…
$ vert_index  <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1, …
$ pitnumber   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ species     <chr> "Cutthroat trout", "Cutthroat trout", "Cutthroat trout", "…
$ length_1_mm <dbl> 58, 61, 89, 58, 93, 86, 107, 131, 103, 117, 100, 127, 99, …
$ length_2_mm <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ weight_g    <dbl> 1.75, 1.95, 5.60, 2.15, 6.90, 5.90, 10.50, 20.60, 9.55, 13…
$ clip        <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "N…
$ sampledate  <date> 1987-10-07, 1987-10-07, 1987-10-07, 1987-10-07, 1987-10-0…
$ notes       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
```

Question 1

List the numerical variable and the two categorical grouping variables used in this analysis.

Creating a summary table with `summarize()`

Before computing summary statistics, it is important to check whether the variables you are using contain missing values (`NA`). Summary statistics like the mean and standard deviation cannot be computed correctly if missing values are present. You can check for missing values in specific variables and remove those rows if needed. For example:

```
# Check for missing values in the variables of interest
and_vertebrates |>
  summarize(
    n_missing_length = sum(is.na(length_1_mm)),
    n_missing_species = sum(is.na(species)),
    n_missing_section = sum(is.na(section))
  )
```

```
# A tibble: 1 × 3
  n_missing_length n_missing_species n_missing_section
             <int>             <int>             <int>
1               17                 3                 0
```

```
# Remove rows with missing values in these variables
and_vertebrates_clean <-
  and_vertebrates |>
  filter(
    !is.na(length_1_mm),
    !is.na(species),
  )
```

Question 2

Why is it important to check for and remove missing values *before* computing summary statistics like the mean or standard deviation?

From this point on, we use the cleaned dataset `and_vertebrates_clean`.

A summary table is a new table created from the original data in which each row represents a group and each column represents a descriptive statistic, such as a mean, standard deviation, or sample size. Instead of listing individual observations, a summary table condenses the data into a small number of values that describe key features of each group.

Summary tables exist because many analyses and graphs require statistics that are not present in the raw data. Values such as means, standard errors, and uncertainty intervals must be computed before they can be plotted or interpreted. A summary table provides these computed values in an organized form that is easy to use for comparisons.

Raw data contain one row per observation and preserve information about distributions, skew, and outliers. Summary tables, in contrast, contain one row per group and no longer show individual values. As a result, summary tables are useful for comparing group-level patterns, but they cannot show distribution shape. This is why raw-data plots and summary-based plots often answer different questions and are used together.

Question 3

In your own words, how does a summary table differ from the raw data table?

We will now create a summary table by grouping the data by species and stream section and computing descriptive statistics for each group using `summarize()`.

```
and_vertebrates_summary <-
  and_vertebrates_clean |>
```

```
  summarize(
    median = median(length_1_mm),
    mean = mean(length_1_mm),
    sd = sd(length_1_mm),
    n = n(),
    se = sd / sqrt(n),
    upper_cl = mean + 2 * se,
    lower_cl = mean - 2 * se,
    .by = c(section, species)
  )
```

This code creates a new object called `and_vertebrates_summary`. It starts with the cleaned dataset (`and_vertebrates_clean`) and then uses `summarize()` to build a **summary table**.

Inside `summarize()`, each line creates a **new column** in the summary table:

- `median` and `mean` are two measures of **central tendency** for `length_1_mm`.
- `sd` is the **standard deviation**, which describes how spread out individual lengths are within a group.
- `n` is the **sample size** in that group (the number of rows contributing to the summaries).
- `se` is the **standard error** of the mean, computed as `sd / sqrt(n)`. This describes how precisely the group mean is estimated.
- `lower_cl` and `upper_cl` define a **2SE interval** around the mean: `mean ± 2 * se`. This is a rule-of-thumb interval used to describe uncertainty in the mean estimate.

The most important line for understanding how the table is built is:

- `.by = c(section, species)`

This tells R to compute all of these statistics **separately for each combination of `section` and `species`**. In the resulting table, each **row** corresponds to one `(section, species)` group, and each **column** corresponds to one statistic for that group.

Question 4

What does the `.by = c(section, species)` argument do in this `summarize()` call?

Finally, printing `and_vertebrates_summary` displays the completed summary table so you can inspect the group-level results.

```
and_vertebrates_summary
```

```
# A tibble: 6 × 9
  section species               median  mean    sd     n    se upper_cl lower_cl
  <chr>   <chr>                  <dbl> <dbl> <dbl> <int> <dbl>    <dbl>    <dbl>
1 CC      Cutthroat trout         88    85.3 36.0  11072 0.342     86.0     84.6
2 OG      Cutthroat trout         84    81.4 34.5   9356 0.356     82.1     80.7
3 CC      Coastal giant salama…   58    60.5 20.8   5397 0.283     61.1     60.0
4 OG      Coastal giant salama…   52    54.0 20.5   6352 0.257     54.6     53.5
5 CC      Cascade torrent sala…   39    37    3.54     9 1.18      39.4     34.6
6 OG      Cascade torrent sala…   34.5  34.3  7.12     6 2.91      40.1     28.5
```

Each row of the summary table represents one species–section combination, and the columns describe the distribution of body lengths within that group.

For Cutthroat trout, the median length is about 2.5–3 mm greater than the mean in both stream sections. This indicates left-skewed distributions, where a small number of shorter individuals pull the mean downward. Although the difference between the mean and median is modest relative to the overall variability (SD ≈ 34–36 mm), the pattern is consistent across sections.

For Coastal giant salamanders, the pattern is reversed: the mean length is about 2–2.5 mm greater than the median in both sections. This indicates right-skewed distributions, where a small number of larger individuals pull the mean upward. Notably, the *magnitude* of the mean–median difference is similar to that seen in trout, even though the direction of skew is opposite.

For Cascade torrent salamanders, sample sizes are very small (n = 6–9), and the mean and median differ by several millimeters. In this case, both measures of center are strongly influenced by individual observations, and the summaries should be interpreted cautiously.

Question 5

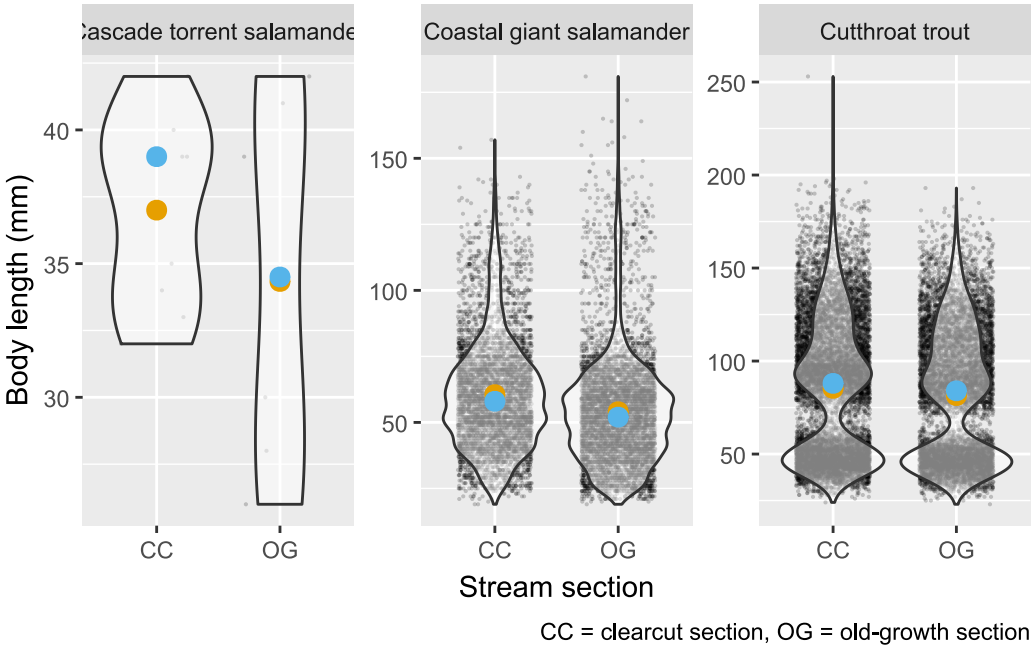
In this dataset, is the standard error (SE) larger or smaller than the standard deviation (SD)? Why?

Importantly, this summary table alone cannot show how the data are skewed, where outliers occur, or whether the distributions differ in shape. The similar mean–median differences across species could arise from very different underlying distributions. To understand those differences—and to see why the direction of skew matters—we next turn to plots of the raw data, which reveal distribution shape, spread, and outliers directly.

Question 1: How do distributions differ?

In this section, we focus on the question: How do distributions differ across groups? To answer distribution questions, we need to look at the raw data, because summary statistics alone cannot show skew, outliers, or differences in shape. Below, we use a strip (jitter) plot and a violin plot to visualize the distribution of individual body lengths in each stream section, within each species. We also overlay the group mean and median to show how measures of center can differ depending on distribution shape.

```
ggplot(
  data = and_vertebrates_clean,
  mapping = aes(x = section, y = length_1_mm)
) +
  geom_jitter(alpha = .2, shape = 16, width = 0.3, height = 0, size = .4) +
  geom_violin(alpha = .5) +
  geom_point(aes(y = mean),
             data = and_vertebrates_summary,
             color = "#E69F00", size  = 3) +
  geom_point(aes(y = median),
             data = and_vertebrates_summary,
             color = "#56B4E9", size = 3) +
  facet_wrap(~ species, scales = "free_y") +
  labs(
    x = "Stream section",
    y = "Body length (mm)",
    caption = "CC = clearcut section, OG = old-growth section"
  )
```



Question 6

Name one distributional feature you can see in this plot that is *not* visible in the summary table.

How to interpret this code

- The base `ggplot()` call uses the **raw data** (`and_vertebrates_clean`) and maps:
  - `section` to the x-axis
  - `length_1_mm` to the y-axis
    These mappings apply to any layer that does not explicitly override them.
- `geom_jitter()` and `geom_violin()` both use the **raw data**:
  - Each point represents an individual observation
  - The violin shows the overall distribution shape
  - These layers emphasize **spread, skew, and outliers**, not just averages
- The orange and blue points use a **different dataset**, `and_vertebrates_summary`:
  - This table contains one row per species–section group
  - It does *not* contain individual measurements
- In `geom_point()`, the y-aesthetic is **overridden**:
  - `aes(y = mean)` plots the group mean
  - `aes(y = median)` plots the group median
    Overriding the y-aesthetic is necessary because summary tables do not include raw values like `length_1_mm`.
- Using different datasets in different layers allows raw data and summary statistics to appear together in the same figure.
- `facet_wrap(~ species)` creates a separate panel for each species:
  - Each panel shows the same comparison across stream sections

- `scales = "free_y"` allows each species to use its own y-axis scale
  This makes distribution shape easier to see within species but prevents direct comparison of absolute values across species.

- `labs()` adds axis labels and a caption so the figure can be interpreted without additional explanation.

Question 7

Why do some `geom_*()` layers use the raw dataset, while others use the summary table?

Question 8

What does it mean to "override" an aesthetic inside a `geom_*()` call?

Interpreting the distributions

Sample size

- *Cutthroat trout* and *Coastal giant salamanders* have very large sample sizes in both stream sections.
- Large samples produce smooth violins and dense point clouds, allowing distribution shape to be assessed reliably.
- Differences in visual smoothness across panels primarily reflect **sample size**, not biological variability.

Distribution shape and modality

- Distributions differ in shape and show evidence of **skew** and, in some cases, **multiple modes**.
- For trout, the distributions appear multi-modal, consistent with the presence of distinct size classes (e.g., juveniles and adults).
- For coastal giant salamanders, distributions are more unimodal but still skewed.
- Skew and modality influence how measures of center should be interpreted and whether a single "typical" value is meaningful.

Central tendency and spread

- For trout, medians tend to be slightly higher than means, consistent with left-skewed distributions.
- For coastal giant salamanders, means tend to be slightly higher than medians, consistent with right-skewed distributions.
- The amount of variability within sections is large compared to the differences in means and medians between sections, making those differences seem small by comparison. However, even differences of only a few millimeters in body length can have important biological consequences.
  - Body size is often linked to **survival, growth rates, competitive ability, and reproductive success**.
  - As a result, small shifts in average size between stream sections could translate into meaningful differences in fitness or population dynamics.
  - Visualizing the full distribution helps place these mean differences in context, but it does not determine whether they are biologically important.

Question 9

For this dataset, does the mean or the median better represent a "typical" individual? Briefly explain your reasoning.

Implications for the next step in the analysis

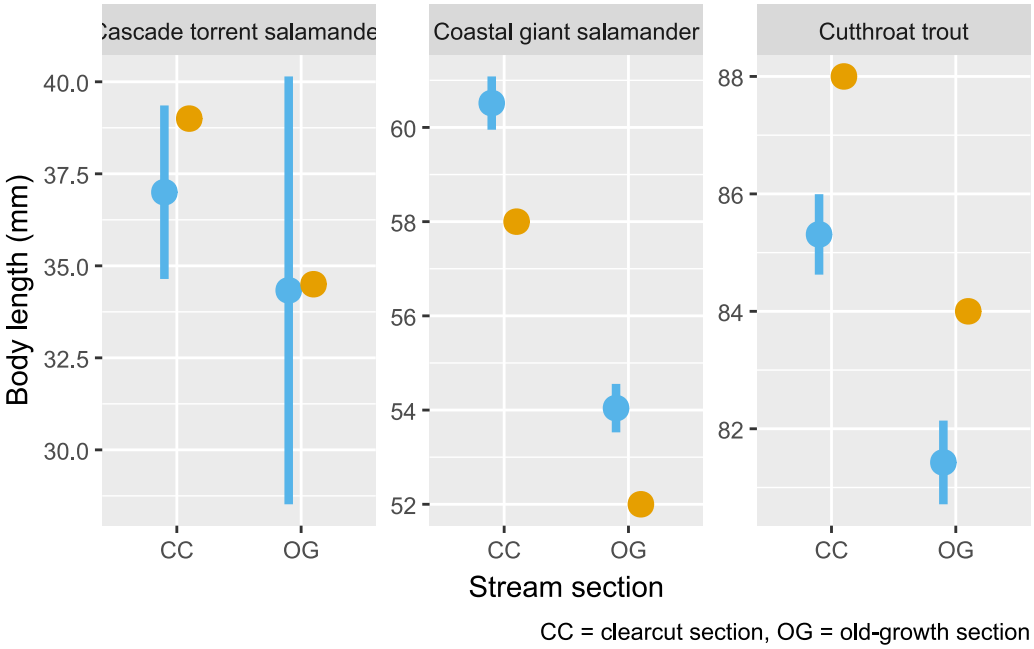
- Because raw-data plots emphasize variability, skew, and modality, they are not ideal for clearly comparing group means when those means are close together.

- To focus on differences in means and their uncertainty, we need a different type of graph that does not display individual observations.
- In the next step, we will use a plot that emphasizes means and confidence intervals, allowing small differences between groups to be seen more clearly.
- The distributional patterns observed here also inform future choices about:
  - which summary statistics are appropriate,
  - whether assumptions of parametric tests are reasonable,
  - and which visualization best matches the specific question being asked.

Question 2: How do group means compare?

In this section, we shift focus from distribution shape to group-level averages. When the goal is to compare mean values across groups—especially when differences are small relative to overall variability—plots showing raw data can obscure those differences. Instead, we use a graph that emphasizes means and their uncertainty, allowing group means to be compared more directly while still acknowledging variation in the data.

```
ggplot(
  data = and_vertebrates_summary,
  mapping = aes(x = section)
) +
  geom_point(aes(y = mean),
             color = "#56B4E9", size = 4,
             position = position_nudge(x = -0.1)) +
  geom_linerange(aes(ymin = lower_cl, ymax = upper_cl),
                 color = "#56B4E9", linewidth = 1.5,
                 position = position_nudge(x = -0.1)) +
  geom_point(aes(y = median),
             color = "#E69F00", size  = 4,
             position = position_nudge(x = 0.1)) +
  facet_wrap(~ species, scales = "free_y") +
  labs(
    x = "Stream section",
    y = "Body length (mm)",
    caption = "CC = clearcut section, OG = old-growth section"
  )
```



Question 10

Why is this graph more effective than the distribution plot for comparing group means?

How to interpret this code

- This plot uses the **summary table** (`and_vertebrates_summary`) rather than the raw data.
  - Each row in this dataset represents one **species–section group**
  - The values being plotted are **precomputed summary statistics**, not individual observations
- The base `ggplot()` call maps `section` to the x-axis.
  - No y variable is mapped globally, because different layers plot **different summary values**
- `geom_point(aes(y = mean))` plots the **group mean** for each stream section.
- `geom_linerange(aes(ymin = lower_cl, ymax = upper_cl))` draws a vertical line showing the **2SE interval** around each mean.
  - This layer emphasizes **uncertainty in the mean estimate**
  - Unlike the previous plot, raw data are not shown
- `geom_point(aes(y = median))` plots the **group median**.
  - Including the median allows direct comparison between two measures of central tendency
- The `position = position_nudge(x = …)` argument shifts layers **slightly left or right**:
  - Mean points and their 2SE intervals are nudged left
  - Median points are nudged right
  - Nudging is purely visual and prevents overlap; it does not change the data
- `facet_wrap(~ species)` creates a separate panel for each species.
  - `scales = "free_y"` allows each species to use its own y-axis scale
  - This makes mean differences within species easier to see but prevents comparison of absolute sizes across species
- `labs()` adds axis labels and a caption so the figure can be interpreted without additional context

Important points:

- This plot emphasizes **group means and uncertainty**, rather than distribution shape
- Removing raw data makes small differences in means easier to see

- Using summary-based plots is appropriate when the question is about **comparing averages**, not individual variability, and when the differences between summary statistics are small compared to the variability of the data.

Interpreting group means and uncertainty
Sample size and uncertainty

- The vertical line segments show **±2 standard errors** around each mean.
- For species with large sample sizes, these intervals are very narrow, indicating that the means are estimated with high precision.
- Where sample sizes are smaller, the 2SE intervals are wider, reflecting greater uncertainty in the mean estimates.

Means versus medians

- The median (orange point) does not always coincide with the mean (blue point).
- These differences reflect skew in the underlying distributions observed in the raw-data plots.
- The direction of the mean–median difference differs among species, reinforcing that distribution shape affects how measures of central tendency should be interpreted.

Apparent size of differences

- Differences between CC and OG means appear relatively large in this figure because the y-axis scale is narrow and does not show the full range of raw values.
- This is intentional: summary-based plots remove raw variability to make **small differences in means** easier to see.
- Even differences of only a few millimeters can be biologically meaningful, affecting traits such as survival, growth, or reproduction.

Why this graph type is appropriate here

- Raw-data plots are best for understanding distribution shape and variability.
- Mean ± uncertainty plots are better for **comparing average values across groups** when variability would otherwise obscure those differences.
- Together, these plots provide complementary views of the same data and inform how group differences should be interpreted and analyzed next.

Example Results paragraph

Body length distributions differed in shape and spread among species and overlapped substantially between stream sections (Figure 1). For both Cutthroat trout and Coastal giant salamanders, body lengths showed wide distributions with clear skew, and mean and median values did not coincide, indicating asymmetric distributions.

Despite this within-group variability, mean body length was slightly higher in clearcut (CC) sections than in old-growth (OG) sections for both species (Figure 2). For Cutthroat trout, mean body length was 85.3 mm in CC sections (approximate 95% CI: 84.6–86.0 mm) compared to 81.4 mm in OG sections (approximate 95% CI: 80.7–82.1 mm). Coastal giant salamanders showed a similar pattern, with mean body length of 60.5 mm in CC sections (approximate 95% CI: 60.0–61.1 mm) and 54.0 mm in OG sections (approximate 95% CI: 53.5–54.6 mm). Although differences in means were small relative to the full range of observed body sizes, these shifts in average body length are visible when uncertainty is summarized and raw variability is removed (Figure 2). In contrast, Cascade torrent salamanders had much smaller sample sizes, resulting in wide confidence intervals around mean body length estimates and greater uncertainty in comparisons between stream sections.

- **In-text references to figures** (Figure 1 and Figure 2)
  Results should point readers to the figures that show the patterns being described, rather than restating everything shown in the plots.

- **Description of distributions** (Figure 1)
  The paragraph describes spread, skew, and overlap in the raw data before discussing averages, because distribution shape affects how summary statistics should be interpreted.
- **Numerical reporting of group means and uncertainty** (Figure 2)
  Means and approximate 95% confidence intervals are reported directly in the text so readers do not have to extract values from the figure.
- **Clear group comparisons**
  Differences are described in terms of direction and approximate magnitude (CC vs. OG), without using hypothesis-testing language.
- **Minimal interpretation**
  The paragraph reports observable patterns and places mean differences in context, but avoids causal explanations or claims about statistical significance.

Student task: Applying the workflow to bison data

In this task, you will apply the same workflow used in the worked example to a different dataset. Using the bison data, you will create summary tables, visualize distributions, and compare group means using appropriate graphs. The goal is not to explore the data freely, but to practice making deliberate choices about how to summarize, visualize, and describe data based on the question being asked.

The `knz_bison` dataset

The `knz_bison` dataset contains records of American bison sampled at Konza Prairie, including individual **body weight** measurements and **sex**. In this task, you will focus on `animal_weight` as the numerical variable of interest and `animal_sex` as the grouping variable. This dataset is smaller and simpler than the vertebrate dataset used in the worked example, making it well suited for practicing the full workflow of summarizing data, visualizing distributions, and comparing group means.

Assignment: Bison descriptive statistics and visualization

Using the `knz_bison` dataset, you will answer the **same two questions** addressed in the worked example. Follow the same workflow and make the same kinds of analytical choices, adjusting only what is necessary for the structure of this dataset.

Step 1: Create a summary table

Before making your plots, create a summary table of `animal_weight` grouped by `animal_sex`. Your summary table should include:

- `mean`
- `median`
- `sd`
- `n`
- `se`
- `lower_cl` and `upper_cl` for an approximate 95% confidence interval using the 2SE rule of thumb

You will use these summary values in your plots.

Question 1: How do distributions differ?

Using the raw bison data, create a plot that compares the distributions of body weight between sexes. Your plot should:

- Show individual observations (e.g., jitter/strip plot)
- Show distribution shape (e.g., a violin)
- Overlay the group mean and median from your summary table

Use this plot to describe differences in shape, spread, and skew between groups.

Question 2: How do group means compare?
Create a second plot that emphasizes group means and uncertainty for body weight by sex. This plot should:

- Use your summary table (not the raw data)
- Show the group mean and an approximate 95% confidence interval (mean ± 2SE)
- Optionally include the median for comparison, as in the worked example
- Avoid plotting individual observations

Use this plot to compare group means directly, without raw-data variability obscuring differences.

Results paragraph
Complete this part of the assignment outside of class and submit it on D2L by the due date.

Create a new Microsoft Word document containing one Results-style paragraph that summarizes and compares bison body weight by sex. Your paragraph should:

- Refer explicitly to both figures you created (the distribution plot and the mean comparison plot)
- Describe relevant distributional features (e.g., shape, spread, skew) where they inform interpretation
- Report numerical values for medians.
- Report numerical values for means and approximate 95% confidence intervals.
- Briefly note whether the mean or median provides a more informative summary for bison body weight, and why.

The paragraph should read like a Results section, not a methods explanation or interpretation-heavy discussion.

Wrap-up and submission

Before leaving lab:

1. Ensure your R script is saved in your Posit Cloud project.
2. Keep your worksheet/handout for your own notes.
3. Submit **both graphs** on D2L under the *Lab 6 Graphs* assignment.

After lab:

4. Write your Results paragraph and submit it on D2L by the posted deadline.