# Week 7: Comparing Groups – Boxplots & Statistical Tests

Ronnie Bailey-Steinitz

2025-11-03

## Week 7: Comparing Groups – Boxplots & Statistical Tests

**Goal:** Learn to compare groups statistically and visually using t-tests, Wilcoxon tests, and ANOVA; interpret assumptions; and create boxplots that clearly communicate group differences.

---

### 0. Load Packages and Data

```
library(tidyverse)
library(here)
library(janitor)
```

## Skills Learning – Lecture

This week we compare *numeric vs categorical variables* — for example, body weight across species.

These tests ask: *Is the mean (or median) of one group significantly different from another?*

### 1. Load Data

```
penguins <- read_csv(here("Week 1/Palmer Penguins Raw.csv")) %>%
  janitor::clean_names() %>%
  mutate(full_name = species,
         species = word(species, 1))
```

```
## Rows: 344 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): studyName, Species, Region, Island, Stage, Individual ID, Clutch C...
## dbl  (7): Sample Number, Bill Length (mm), Bill Depth (mm), Flipper Length (...
##
## 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.
```

## 2. Visualize Groups

Boxplots are an easy way to explore potential differences between groups before running a statistical test. Each box shows:

- the median (horizontal line inside the box)
- the interquartile range (IQR) — where the middle 50% of data fall
- potential outliers (points outside the whiskers)

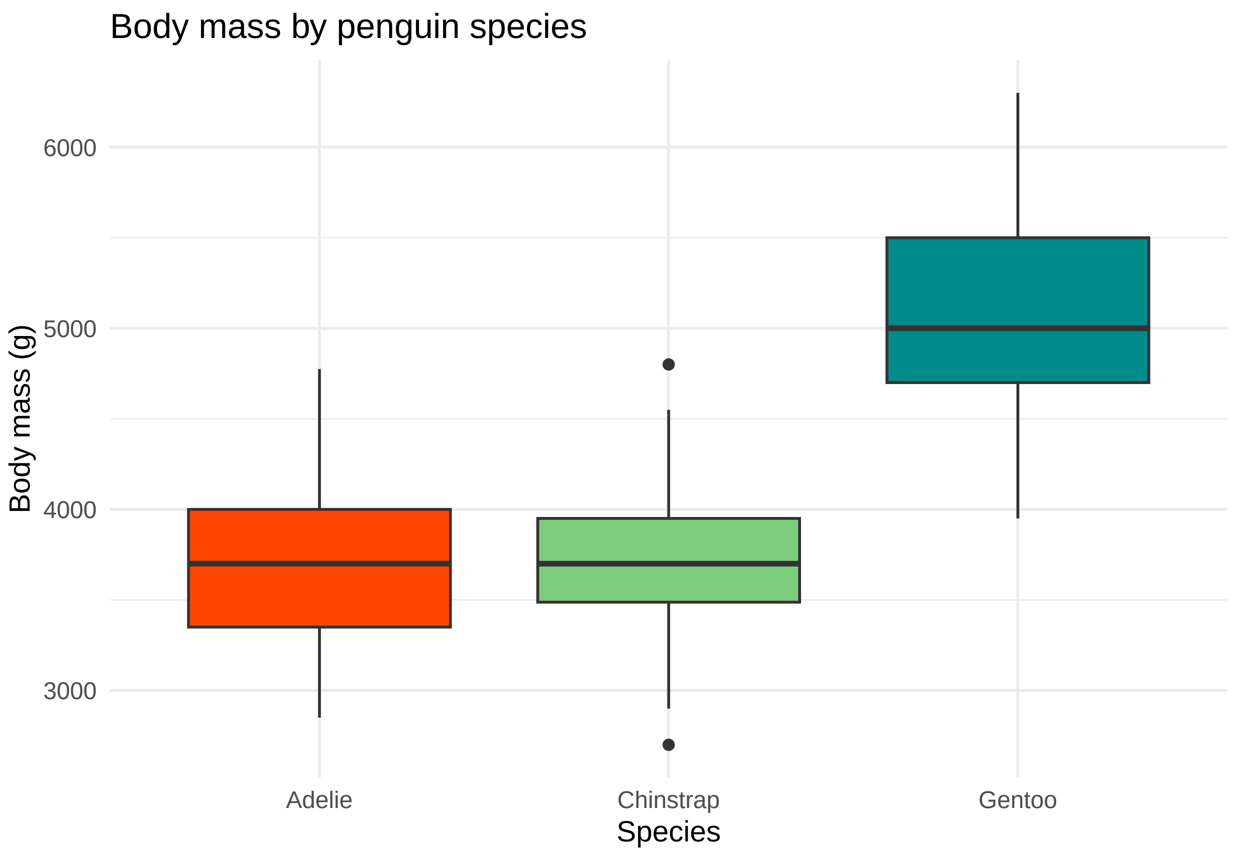
You can look for visual evidence of group differences by checking whether:

- the medians or boxes are clearly separated (suggesting different group means or medians), and
- there is little or no overlap between the boxes (suggesting a stronger difference).

However, because boxplots summarize spread and not sampling error, **they only suggest patterns**. You still need a formal statistical test (t-test or ANOVA) to determine whether the observed difference is statistically significant.

```r
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
  labs(x = "Species",
       y = "Body mass (g)",
       title = "Body mass by penguin species") +
  scale_fill_manual(values = c("orangered", "palegreen3", "cyan4")) +
  theme_minimal() +
  theme(legend.position = "none") # add this because the legend is unnecessary here
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



The boxplot suggests that Gentoo penguins are generally heavier than the other two species, but we will confirm this using statistical tests.

### 3. Compare groups with t-test

To simplify, let's compare only two groups:

```r
penguins2 <- penguins %>%
  filter(species != "Adelie")

# alternatively, instead of exclusion of "Adelie",
# I could have filtered by inclusion of the other two species:
# (spp == "Chinstrap" | spp == "Gentoo")

t.test(body_mass_g ~ species, data = penguins2)
```

```
##
##  Welch Two Sample t-test
##
## data:  body_mass_g by species
## t = -20.628, df = 170.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Chinstrap and group Gentoo is not equa
## 95 percent confidence interval:
##  -1471.440 -1214.416
## sample estimates:
## mean in group Chinstrap    mean in group Gentoo
##                3733.088                5076.016
```

> This t-test asks: Are the mean body masses of Chinstrap and Gentoo penguins statistically different?

The default type of t-test is a **Welch Two Sample T-test** which compares the means of two independent groups (in our case, Chinstrap vs Gentoo).

It adjusts the degrees of freedom if the group variances are unequal (Welch's correction).

**t-value** measures how far apart the group means are relative to the variability in the data. A large absolute t-value (far from 0) indicates the group means are very different relative to the spread within each group.

- Here, `t = -20.6` is extremely large in magnitude, showing a very strong difference.

**df** is the degrees of freedom. This reflects how much independent information is used to estimate variability. It's roughly related to your sample size. Here, around `170` effective degrees of freedom.

**p-value** is the probability of seeing a difference this large if the true means were equal (the null hypothesis). A p-value this small (`< 0.001`) indicates an extremely strong statistical difference between the groups.

For this t-test, we can say:

> There is a statistically significant difference in mean body mass between Chinstrap and Gentoo penguins ($p < 0.001$).

**confidence interval** gives the range of plausible values for the true mean difference (Chinstrap − Gentoo).

- Because the entire interval is negative, it tells you that Chinstrap penguins are, on average, lighter than Gentoo penguins. The true difference in means is estimated to be between `1214` and `1471` grams lighter.

Last, the sample estimates:

- Chinstrap mean = 3733 g
- Gentoo mean = 5076 g

Difference = $3733 - 5076 = -1343$ g (matches the CI and sign of the t-value)

So, the **negative t-value** and **negative confidence interval** simply reflect that the first group listed (Chinstrap) has a smaller mean than the second (Gentoo).

In a manuscript, you would report it like this:

> "Gentoo penguins were significantly heavier than Chinstrap penguins (mean ± SD: 5076 g vs. 3733 g; Welch's t-test, t(170.4) = –20.63, $p < 0.001$). The mean difference in body mass was approximately 1.34 kg, with a 95% confidence interval ranging from –1471 to –1214 g."

## 4. Compare more than 2 groups

Now, let's compare all species

```r
model1 <- aov(body_mass_g ~ species, data = penguins)

summary(model1)
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)    
## species       2 146864214 73432107   343.6 <2e-16 ***
## Residuals   339  72443483   213698                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

In a manuscript, you would say:

> "Body mass differed significantly among penguin species (F(2, 339) = 343.6, p < 0.001). Species explained a large proportion of the variation in body mass, with mean differences confirmed as highly significant across groups."

## 5. Post-hoc Tukey comparison

If the ANOVA is **significant**, we can use post-hoc comparisons:

```r
TukeyHSD(model1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = body_mass_g ~ species, data = penguins)
## 
## $species
##                        diff       lwr       upr     p adj
## Chinstrap-Adelie   32.42598 -126.5002  191.3522 0.8806666
## Gentoo-Adelie    1375.35401 1243.1786 1507.5294 0.0000000
## Gentoo-Chinstrap 1342.92802 1178.4810 1507.3750 0.0000000
```

The results here show something interesting: there is a significant difference in body mass between **Gentoo and Adelie** (difference of 1375 g), and between **Gentoo and Chinstrap** (difference of 1342 g), but *not* between **Chinstrap and Adelie** (only 32 g, which isn't enough to appear different, statistically).

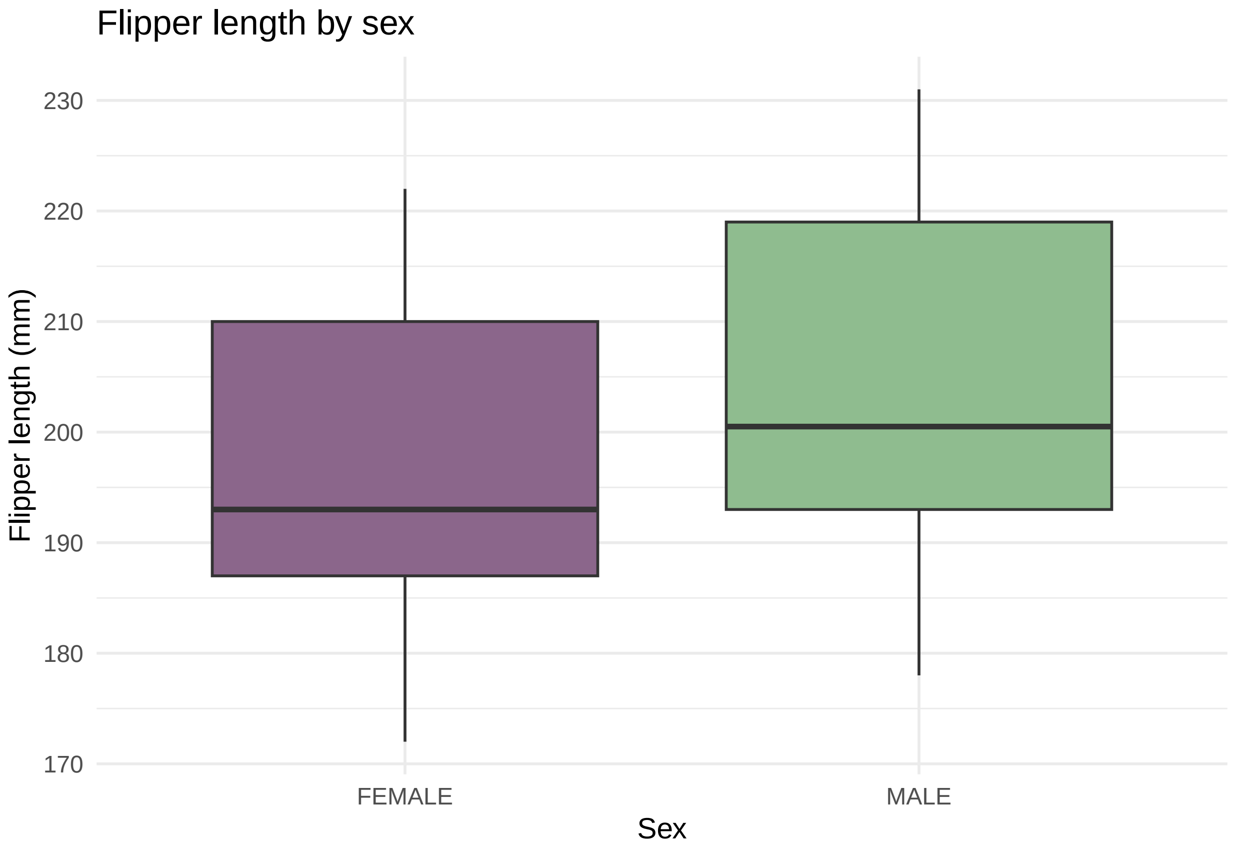
You can report a Tukey comparison like this:

> "Post-hoc Tukey tests revealed that Gentoo penguins were significantly heavier than both Adelie (p < 0.001) and Chinstrap (p < 0.001) penguins, while there was no significant difference in body mass between Chinstrap and Adelie species (p = 0.88)."

## 6. Example:

```r
sex_diff <- penguins %>%
  filter(!is.na(sex))

ggplot(sex_diff, aes(x = sex, y = flipper_length_mm, fill = sex)) +
  geom_boxplot() +
  labs(x = "Sex", y = "Flipper length (mm)", title = "Flipper length by sex") +
  scale_fill_manual(values = c("plum4", "darkseagreen")) +
  theme_minimal() +
  theme(legend.position = "none")
```



```r
t.test(flipper_length_mm ~ sex, data = penguins)
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm by sex
## t = -4.8079, df = 325.28, p-value = 2.336e-06
## alternative hypothesis: true difference in means between group FEMALE and group MALE is not equal to
## 95 percent confidence interval:
##  -10.064811  -4.219821
## sample estimates:
## mean in group FEMALE   mean in group MALE
##             197.3636             204.5060
```

The results here are: t = -4.8079, df = 325.28, p-value = 0.000002336

So, we can report it like this:

> “Male penguins had significantly longer flippers than females (mean ± SD: 204.5 mm vs. 197.4 mm; Welch’s t-test, t(325.28) = −4.81, p = 0.0000023). The difference in mean flipper length was approximately 6.9 mm, with a 95% confidence interval from −10.1 to −4.2 mm.”

## BONUS CONTENT:

### Checking Model Assumptions

**Validity:** The t-test is a *parametric* test and assumes the data meets certain criteria. Violating these assumptions can lead to inaccurate p-values and incorrect conclusions.

We’ll check normality visually and test for equal variance.

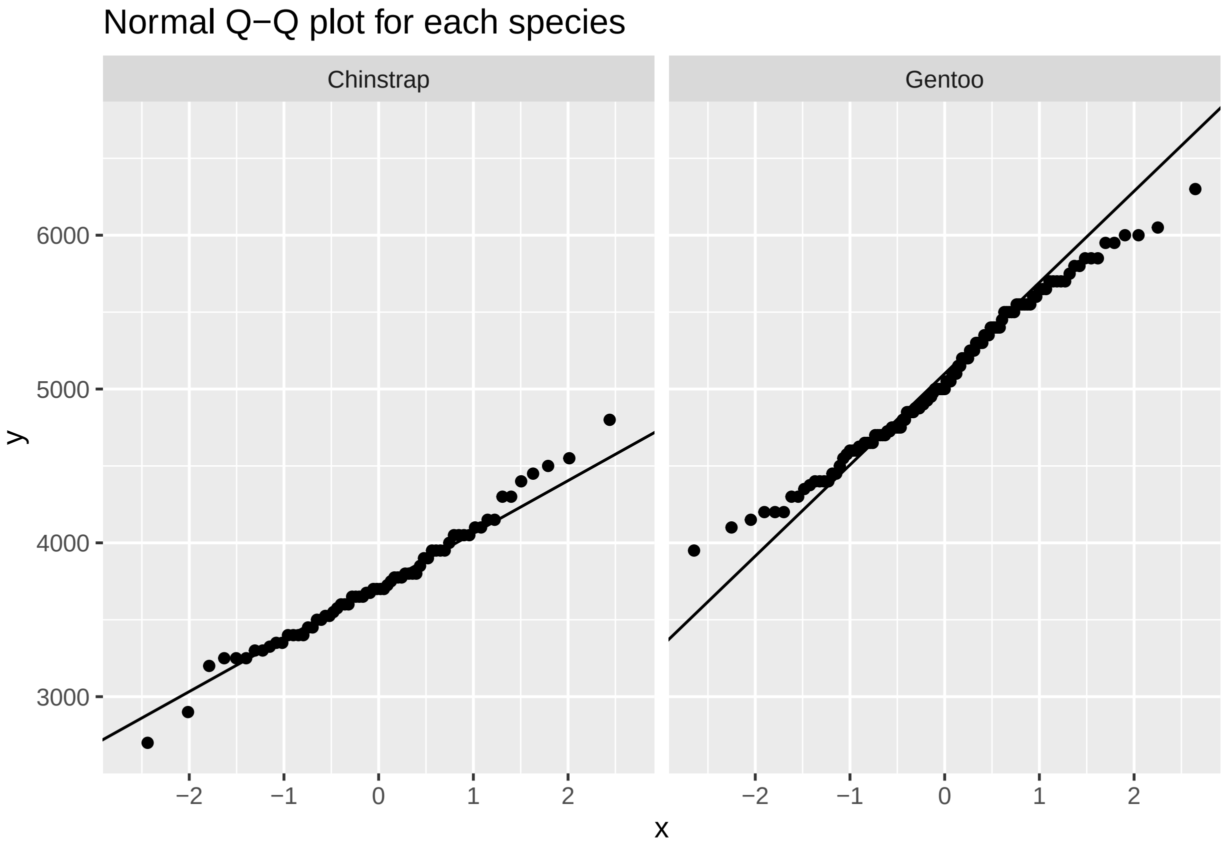
### Normality

Visual check: Use a Q-Q plot (Quantile-Quantile plot) to see if data points fall approximately along a straight diagonal line.

```r
# check for normality
ggplot(penguins2, aes(sample = body_mass_g)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ species) +
  labs(title = "Normal Q-Q plot for each species")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_qq()`).
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_qq_line()`).
```



**Variance**

**Homogeneity of variance:** For independent samples, you need to check if the variances of the two groups are **roughly equal**. Visually check for extreme outliers, which can heavily influence t-test results.

```
# test for equality of variances
var.test(body_mass_g ~ species, data = penguins2)
```

```
##
##  F test to compare two variances
##
## data:  body_mass_g by species
## F = 0.58124, num df = 67, denom df = 122, p-value = 0.01559
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3852832 0.9004223
## sample estimates:
## ratio of variances
##          0.5812443
```

**Non-parametric Wilcoxan Test**

If the normality assumption is violated, consider using a Wilcoxan non-parametric test.

Non-parametric tests are “distribution-free” and do not require the data to be normally distributed, making them a good alternative when assumptions are not met

```r
wilcox.test(body_mass_g ~ species, data = penguins2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  body_mass_g by species
## W = 131, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Wilcoxon rank sum test with continuity correction
```

## Learning Objectives Recap

By the end of Week 7, you should be able to:

- Choose the appropriate test for comparing group means/medians
- Check normality and variance assumptions
- Conduct `t.test()`, `wilcox.test()`, and `aov()` in R
- Visualize group comparisons effectively with boxplots
- Interpret and communicate results in plain language