Chapter 9: Analysis of Variance (ANOVA)

Brandon M. Greenwell

November 12, 2017

Recall that the two-sample t-test is used to compare the means of two populations or two experimental groups. In many experiments there are more than two conditions or treatments to compare in which case the two sample *t*-test will not suffice. This chapter introduces *analysis of variance* (ANOVA) and the associated test of equality of more than two means.

Terminology

Example: An experiment is carried out to see how genotype and the amount of phosphorus affect plant growth. The biomass of the plant is the measured response. Three genotypes and two phosphorus levels (high and low) were used in the experiment. There were 10 plants grown at each experimental combination.

- This is an example of a 3×2 factorial experiment.

Terminology

- **Response variable:** The response (or dependent) variable in this example is the biomass of the plant.
- **Factors:** The independent variables which *may* influence the response variable (genotype and phosphorus in this example).
- **Levels:** The individual levels of each factor. For example, genotype has three levels and phosphorus has two.
- **Treatment:** A combination of factor levels. In this example, there are six possible treatments:
  - genotype 1 and low phosphorus;
  - genotype 2 and low phosphorus;
  - genotype 3 and low phosphorus;
  - genotype 1 and high phosphorus;
  - genotype 2 and high phosphorus;
  - genotype 3 and high phosphorus.

Terminology

- **Replicates:** The number of experimental units (plants in this example) per treatment. In this example, there are a total of ten replicates because there are ten plants grown at each treatment combination.
- **Balanced design:** This occurs when each treatment combination uses the same number of replicates.

Single factor ANOVA

- An experiment with a single factor is called a single factor ANOVA or a one-way ANOVA.
- The two-sample t-test was used to compare the means of two populations (think of this as a one-way ANOVA with only two factor levels).
- In a one-way ANOVA, the classical statistical inference is to test if the means across the different levels of the factor are equal or not:

$$H_0 : \mu_1 = \mu_2 = \ldots \mu_k$$

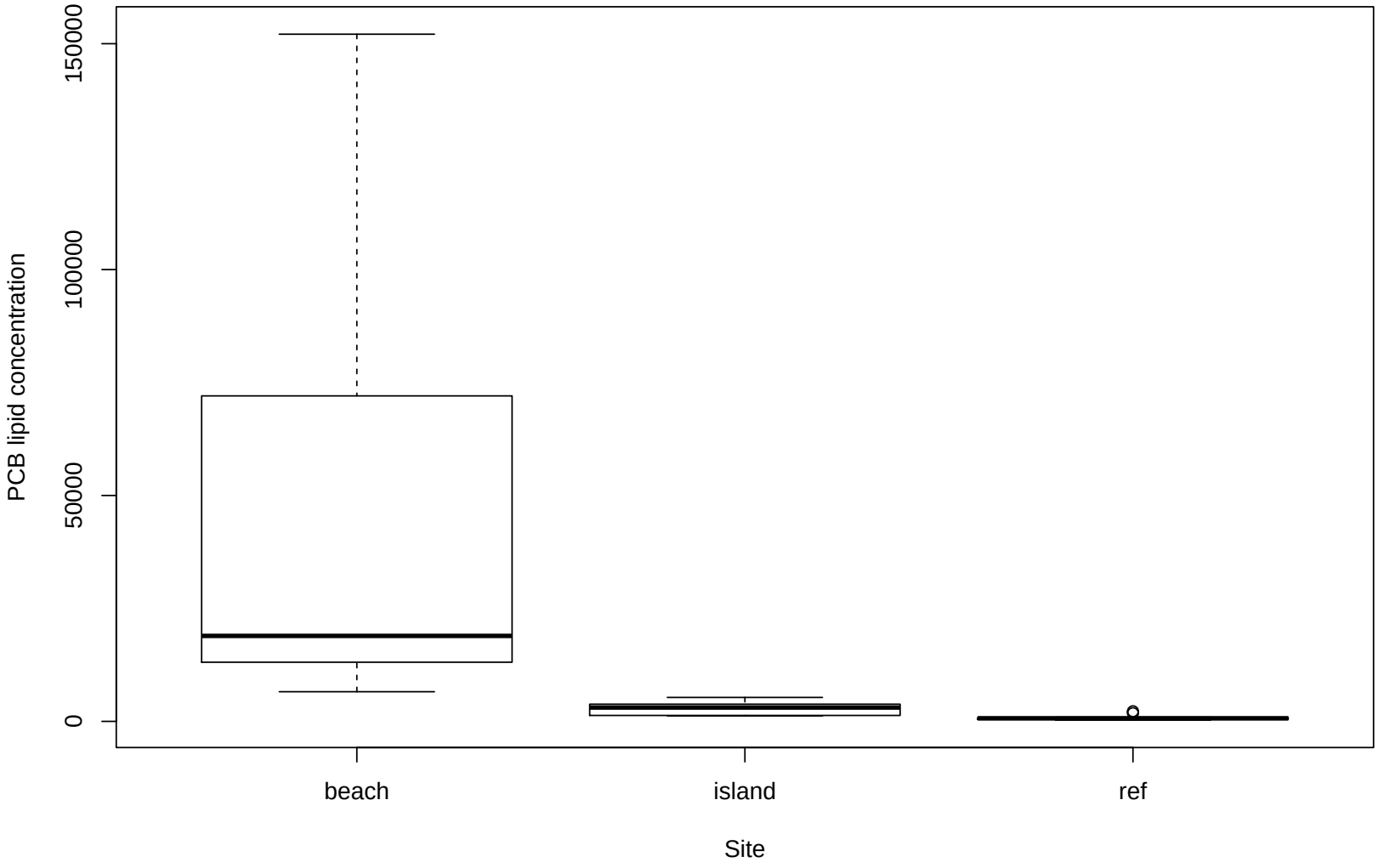
- Because we are typically dealing with factors with three or more levels, the two-sample t-test will not suffice.

Example: PCB lipid concentration

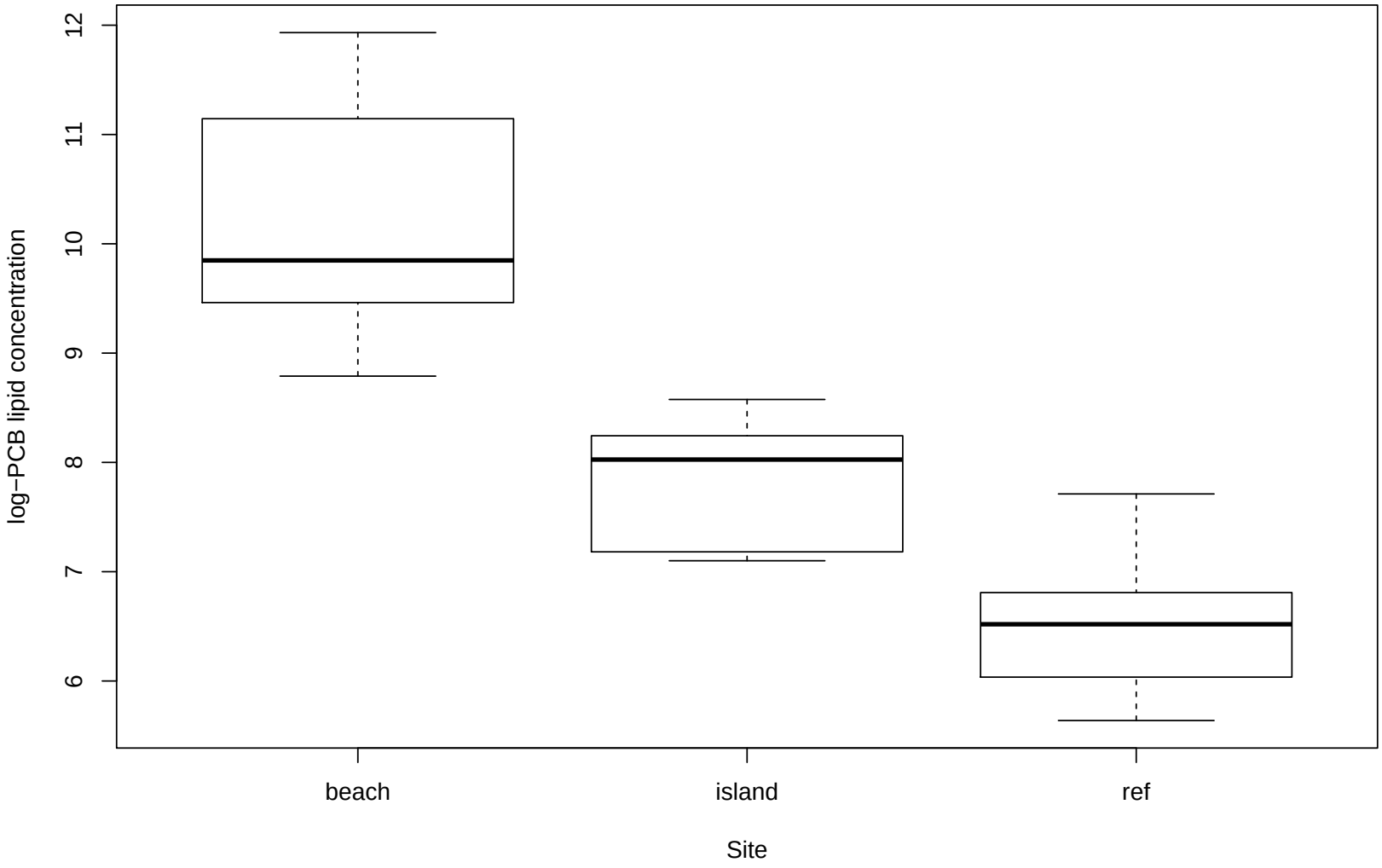
A study was done on the liver PCB concentration (`lipid`) in Black Guillemot birds in Canada at three different sites: a reference site (`ref`), a nearby island (`island`), and the beach (`beach`). A boxplot of the PCB lipid concentration for the birds at the three sites shows strong positive skewness (i.e., right skewed data) which is quite common for data on concentrations of toxins. Because the data is strongly skewed, a log transformation was applied. The goal of the study is to compare the mean log-PCB lipid concentration in the birds at the three sites.

Note that these data are observational, not experimental. That is, the data were collected by observing the birds at the three sites. In an experiment, the experimenter would assign units at random to the different treatments.

Example: PCB lipid concentration



Example: PCB lipid concentration



Example: PCB lipid concentration

Let μ_{beach}, μ_{island}, and μ_{ref} be the mean log-PCB lipid concentration of Black Guillemot birds at the three site. We are interested in testing

$$H_0 : \mu_{beach} = \mu_{island} = \mu_{ref}$$

The alternative (or research) hypothesis is that the means are not all equal (but two of them could be).

F-test for the one-way ANOVA model

To set up the one-way ANOVA hypothesis test procedure, we need to partition the overall variability in the response variable into two components: a component due to the variability **within** each factor level and the variability **between** the factor level means:

$$SS\,(Total) = SS\,(Between) + SS\,(Within)$$

F-test for the one-way ANOVA model

Let y_{ij} denote the j-th observation at level i and $\bar{y}_{i\cdot}$ be the sample mean of the response at level i. Then, we define the following:

- **Total sum of squares**: $SS\left(Total\right) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2$.
- **Between sum of squares**:
  $SS\left(Between\right) = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}\right)^2$.
- **Within sum of squares**:
  $SS\left(Within\right) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$.

F-test for the one-way ANOVA model

- If the factor level means are all equal, then the within group variance and between group variance should be equal.
- The F-test of the one-way ANOVA model is based on a comparison of these two variances:

$$F = \frac{MS\left(Between\right)}{MS\left(Within\right)} = \frac{SS\left(Between\right)/\left(k-1\right)}{SS\left(Within\right)/\left(n-k\right)}$$

- Where $n = n_1 + n_2 + \cdots + n_k$ is the total sample size.

F-test for the one-way ANOVA model

- If the null hypothesis (i.e., equality of means) is true, then the F statistic

$$F = \frac{MS\,(Between)}{MS\,(Within)}$$

  has an F distribution with $k-1$ and $n-k$ degrees of freedom, denoted $F_{k-1,n-k}$.
- Correspondingly, we will reject the null hypothesis whenever F is greater than the $1-\alpha$ percentile from an F distribution with $k-1$ and $n-k$ degrees of freedom.
- The p-value for the test is given by the area to the right of F under an $F_{k-1,n-k}$ distribution: $Pr\left\{F_{k-1,n-k} > F\right\}$.

Example: PCB lipid concentration

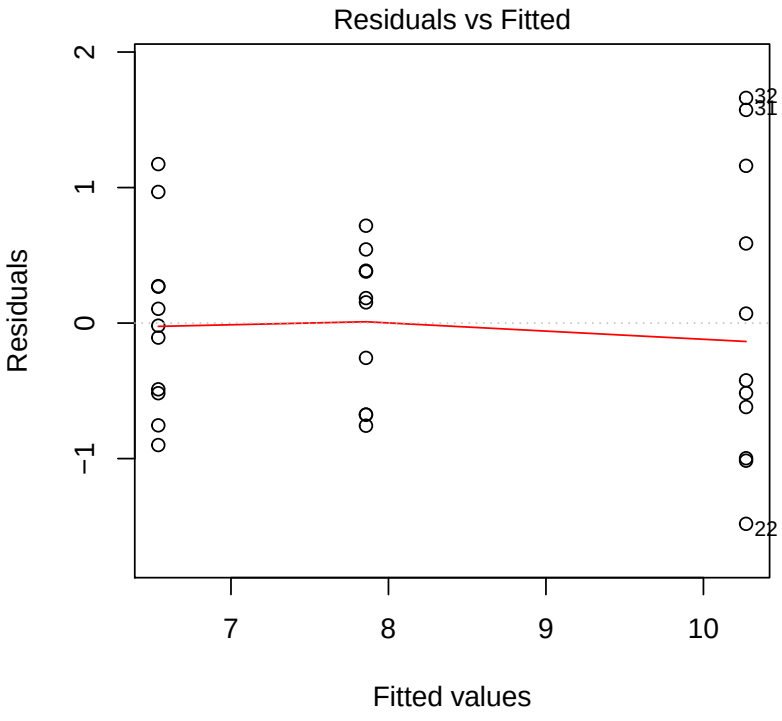
To test the null hypothesis

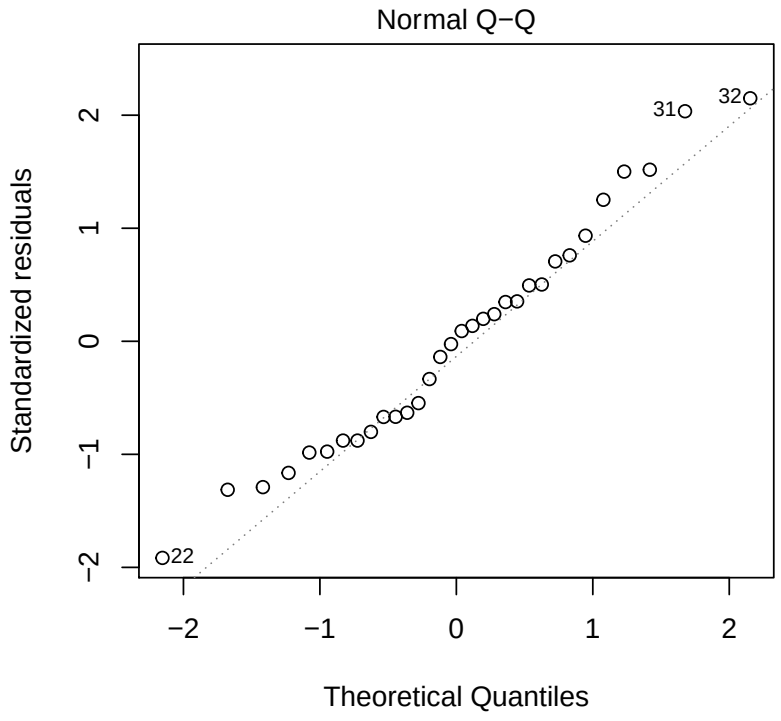
$$H_0 : \mu_{beach} = \mu_{island} = \mu_{ref}$$

in R, we could use the aov function:

```
## Call:
##    aov(formula = log(lipid) ~ site)
##
## Terms:
##                       site Residuals
## Sum of Squares  78.67164  19.06337
## Deg. of Freedom         2        29
##
## Residual standard error: 0.8107758
## Estimated effects may be unbalanced
```

Example: PCB lipid concentration





Multiple comparisons

- In the previous example, the F-test allowed us to conclude that the average $\log(PCB)$ level in the livers of the black Guillemot differed depending among the three sites.
- However, the F-test in an ANOVA does not tell us where the differences, if any occurred, occurred!
- Typically, the next step in an ANOVA is do determine where the differences are.
- One approach is to compute a confidence interval for the difference in the means for all possible pairs of means using the methods of **Chapter 7**.

Multiple comparisons

- Recall that the ANOVA model assumes a common variance between all k groups.
- In an ANOVA setting, we can the MSE as a pooled (across all factor levels) estimate of the error variance: $\widehat{\sigma}^2 = MSE$.
- If our goal is to compare all pairs of means in a one-way ANOVA, then how many comparisons are possible?
- The problem is that by doing multiple tests, we are inflating the type I error rate α—this is known as the multiple testing problem.
    - Suppose there is a 0.01 chance you get a speeding ticket on any given day that you drive to work and you drive to work every day for a year. Is the probability of getting a speeding ticket at least once during the year equal to 0.01?

The Bonferroni method

- The Bonferroni method of multiple comparisons is very simple!
- Specify a confidence level $1-\alpha$ for all pairwise comparisons
- For each pair of means μ_i and μ_j, for $i, j = 1, 2, \ldots, k$ $(i \neq j)$, a confidence interval for the difference $\mu_i - \mu_j$ is

$$\bar{y}_i - \bar{y}_j \pm t_{1-\frac{\alpha}{2g}, n-k} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- where $k = k\left(k-1\right)/2$.

Tukey honest significant differences (HSD)

- Tukey's HSD procedure is similar to the Bonferroni procedure, but less conservative.
- For the Tukey HSD procedure, we simply replace the pairwise confidence intervals used in the Bonferroni procedure with

$$\bar{y}_i - \bar{y}_j \pm q_{1-\alpha,n-k,k}/2\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

- where $q_{1-\alpha,n-k,k}$ is the $1-\alpha$ percentile from the *studentized range distribution*.

Power and sample size

As with any statistical test, it is important to determine the required sample size to achieve adequate power.

Recall the following principals regarding power and sample size:

- Higher power requires larger sample size (all else held constant).
- The larger the sample the more powerful the test (all else held constant).
- A larger effects size increases power (for a fixed sample size) or requires a fewer sample to achieve a particular power.
- As the error variance decreases, the power increases for a fixed sample size.
- As the significance level α decreases, the power decreases. In other words, to maintain a particular power, the sample size increases with α.

Power and sample size

- The biggest challenge in power analyses (or sample size calculations) is specifying an effect size and obtaining an estimate of the variance!!!!
- This is relatively easy when comparing two means, but what about comparing more than two means?
- One common effects size to use is to specify the smallest difference δ that we would like to detect between the two most different means and let

$$\phi = \sqrt{\frac{n\delta^2}{2k\sigma^2}}$$

The Kruskal-Wallis test

- The Kruskal-Wallis test is a nonparametric alternative to the ANOVA F-test.
- If the data strongly deviate from normality (as did the lipid concentration data before taking the logarithm).
- The Kruskal-Wallis test, like many nonparametric tests, is based on ranks, rather than the raw data.
- In R, the function `kruskal.test` performs a Kruskal-Wallis rank sum test of the null that the location parameters of the distribution of y are the same in each group (sample). The alternative is that they differ in at least one.

Two-way ANOVA

- The one-way ANOVA is a GLM with a single categorical predictor (i.e., factor).
- In practice, many experiments contain more than one factor.
- It is usually more efficient to run a single experiment with multiple factors than it is to run individual experiments for each factor alone.
- By running an experiment with all factors at once gives us the possibility to study **interaction effects**.

Two-way ANOVA: heart weight example

A study was conducted where the weights of hearts of patients with (previously) normal mitral valves prior to infective endocarditis (i.e., an infection of the endocardial surface of the heart) were measured. The patients were classified based on race (black and white) and on sex (male and female). Thus, this is a two-factor study with race as one factor and sex as the other factor. In other words, this is a 2×2 factorial experiment. The response variable y is heart weight.

Two-way ANOVA: heart weight example

We can model the data from this type of experiment using a linear model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where

- y_{ijk} is the response for the k-th subject at the i-th level of the first factor (race) and the j-th level for the second factor (sex).
- μ is the overall mean heart weight for the entire population.
- α_i represents the effect for the i-th level of the first factor (race); hence, α_1 represents the mean effect for blacks and α_2 represents the mean effect for whites.
- β_j represents the effect for the j-th level of the second factor (sex).
- γ_{ij} represents the interaction effect between the two factors race and sex.
- ϵ_{ijk} represents random error.

Two-way ANOVA: heart weight example

In order to estimate α_i, β_j, and γ_{ij}, we need to place some constraints on them. The standard convention is

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = 0,$$

$$\sum_{i=1} \gamma_{ij} = 0, \quad \forall j,$$

and

$$\sum_{j=1} \gamma_{ij} = 0, \quad \forall i$$

Two-way ANOVA: heart weight example

The analysis of two-factor experiments is typically carried out using F-tests (similar to the one-way ANOVA). A standard approach two two-way ANOVAs is as follows:

- **Step 0:** (Optional) Perform an overall F-test of the null hypothesis that all three parameters in the model are zero (i.e., no differences in the mean response among all the treatments).
- **Step 1:** Perform an F-test for the interaction term. If we reject, then we would skip steps 2 and 3 and proceed to use a multiple comparison procedure.
- **Step 2:** Perform an F-test for the first factor. If we reject the null hypothesis, then we conclude that the mean response differs across the different levels of this factor.
- **Step 3:** Perform an F-test for the second factor. If we reject the null hypothesis, then we conclude that the mean response differs across the different levels of this factor.