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 the amount of phosphorus and genotype 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 2×3 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.
- 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:

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

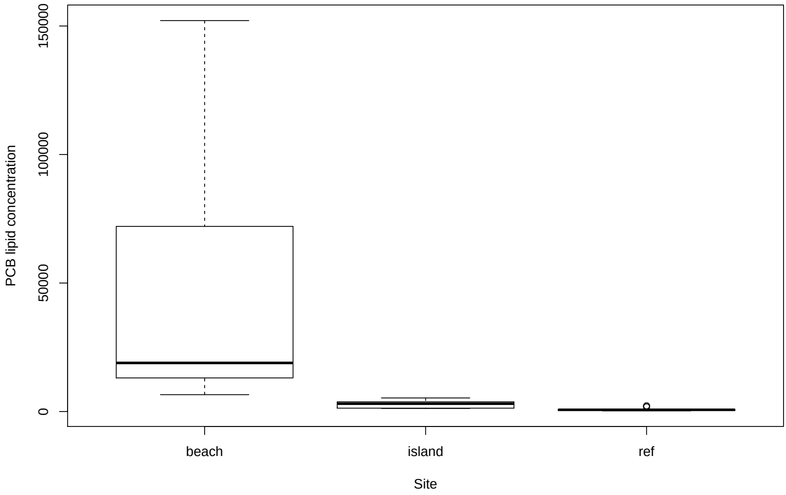
- Because we are typically dealing with factors with three or more levels, the two-sample *t*-test will not suffice.

Example: PCB lipid concentration

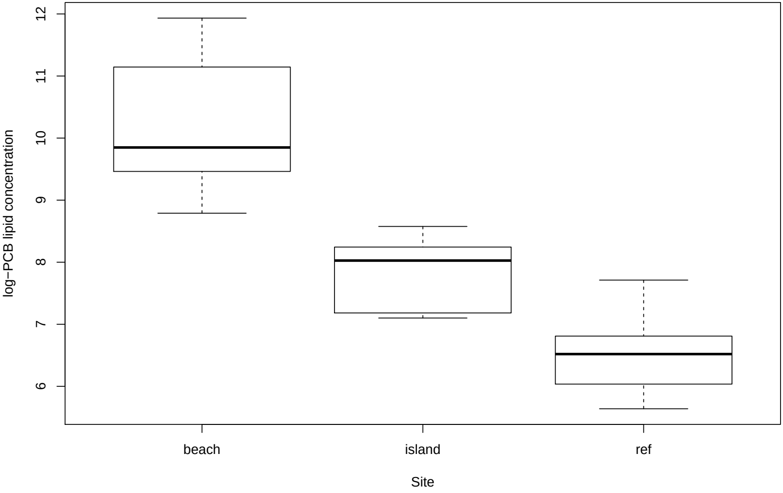
A study was done on the liver PCB concentration (`lipid`) in Black Guillemot birds in Canada at three different sites: a reference site (`ref`), a nearby island (`island`), and the beach `beach`. A boxplot of the PCB lipid concentration for the birds at the three sites shows strong positive skewness (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).

Example: PCB lipid concentration

- Let y_{ij} be the log-PCB concentration of the j-th bird at the i-th site. A simple linear model for these data has the form

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right)$$

- Testing for equality of mean in the one-way ANOVA model is equivalent to testing the null hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \ldots \alpha_k = 0$$

- The alternative hypothesis is that at least one of the α_j's is different from zero.

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:

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

- Let $\bar{y}_{i\cdot}$ be the sample mean of the response at level i.
- **Total sum of squares**: $SS\,(Total) = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - \bar{y})^2$.
- **Between sum of squares**: $SS\,(Between) = n_i\sum_{i=1}^{k}(y_{i\cdot} - \bar{y})^2$.
- **Within sum of squares**: $SS\,(Within) = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_{i\cdot})^2$.

F-test for the one-way ANOVA model

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

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

- 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 $1-\alpha$ percentile from an F distribution with $k-1$ and $n-k$ degrees of freedom.

Example: PCB lipid concentration

To test the null hypothesis

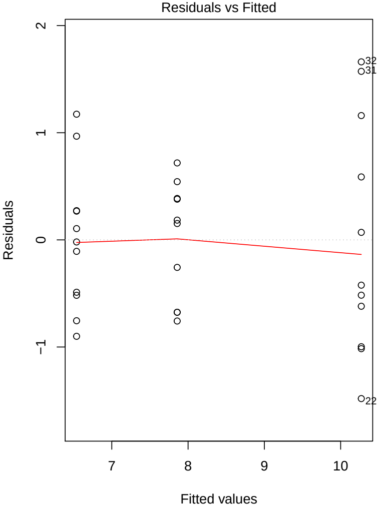
$$H_0 : \mu_{beach} = \mu_{island} = \mu_{ref}$$

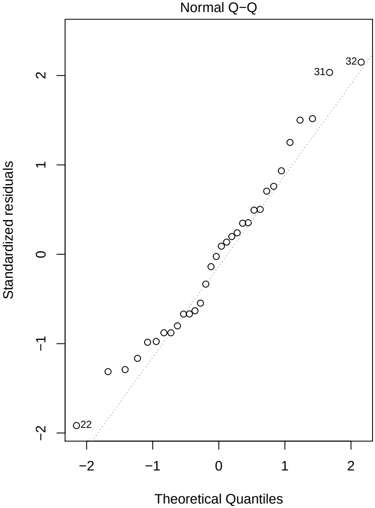
in R, we could use the aov function:

```
## 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

Just like in regression, we can use the residuals to construct model diagnostic plots:


Residuals vs Fitted
Residuals
Fitted values


Normal Q−Q
Standardized residuals
Theoretical Quantiles