Lab 9: Analysis of variance

STAT218

The goal of this lab is to learn how to implement analysis of variance (ANOVA) in R. In its most basic form, this comprises three steps: a visual check of the data, fitting a model and generating the ANOVA table, and interpreting results.

Remember throughout that the first goal of an analysis of variance is to test the hypothesis of no difference in means (interpreted as no effect in the case of an experiment with random treatment allocation). In notation:

This is essentially an extension of the two-sample test to arbitrarily many means. We’ll use the chicks dataset to illustrate these steps and you’ll reproduce using the anorexia dataset.

library(tidyverse)  
load('data/chicks-20d.RData')  
anorexia <- read\_csv('data/anorexia.csv')

### Preliminaries

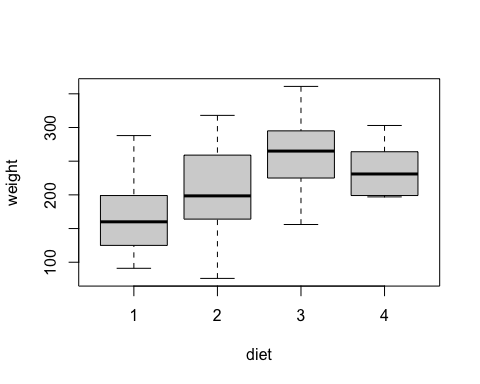
Analysis of variance should start with a brief descriptive analysis: calculation of groupwise summary statistics and visualization of the data by group. The following choices for summary statistics and visualization are standard:

* means, standard deviations, and sample sizes, since inference in ANOVA is based on these statistics
* boxplots, since in an ANOVA there will be more than two groups

# groupwise summaries: mean, sd, and n  
chicks |>  
 group\_by(diet) |>  
 summarize(mean = mean(weight),  
 sd = sd(weight),  
 n = n())

# A tibble: 4 × 4  
 diet mean sd n  
 <fct> <dbl> <dbl> <int>  
1 1 170. 55.4 17  
2 2 206. 70.3 10  
3 3 259. 65.2 10  
4 4 234. 37.6 9

# boxplots by group  
boxplot(weight ~ diet, data = chicks)



While descriptive analysis is useful in and of itself for summarizing the data – here you can see that chicks on diet 3 grew most, and the average weight in this group was 258.9 grams – it also provides a means for assessing ANOVA model assumptions:

* symmetry and unimodality within each group
* similar variation (standard deviation or variance) across groups

Greater departures are tolerable for larger sample sizes. In this instance, sample sizes are modest (9-17), so we should expect both that (a) plots and summary statistics will be tricky to assess and (b) some departure from assumptions is okay.

Most of the time, you’ll only be looking for major departures, so try not to be too sensitive to unequal standard deviations or skewness in boxplots.

|  |
| --- |
| Your turn 1 |
| Produce descriptive summaries – boxplots and summary statistics – for the change variable in the anorexia dataset. (This is the difference between pre-treatment and post-treatment weight, computed as post - pre.) Discuss whether assumptions for ANOVA are plausible.  # groupwise summaries: mean, sd, and n  # boxplots by group |

### Fitting ANOVA models with aov(...)

Fitting an ANOVA model is fairly straightforward in R. The call is:

aov(formula = <RESPONSE> ~ <GROUP>, data = <DATAFRAME>)

Where <RESPONSE> is the name of the variable of interest in <DATAFRAME> and <GROUP> is the name of the grouping variable. These should match exactly the syntax you use to make your boxplot.

When you run the example lines below, make sure to inspect the output at each stage. Look specifically to identify the parts of the analysis of variance we discussed in class:

* the fitted ANOVA model displays only sums of squares, degrees of freedom, and a pooled estimate of the standard deviation of data values about the group means (“residual standard error”)
* the output of summary(...) is required to obtain mean squares, the statistic, and the -value for the test

# fit anova model  
fit <- aov(formula = weight ~ diet, data = chicks)  
fit

Call:  
 aov(formula = weight ~ diet, data = chicks)  
  
Terms:  
 diet Residuals  
Sum of Squares 55881.02 143190.31  
Deg. of Freedom 3 42  
  
Residual standard error: 58.38915  
Estimated effects may be unbalanced

# construct anova table  
summary.aov(fit)

Df Sum Sq Mean Sq F value Pr(>F)   
diet 3 55881 18627 5.464 0.00291 \*\*  
Residuals 42 143190 3409   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

This is all that is needed to report the result of the test. At the 1% level, the test is significant (null hypothesis rejected) since , so the result is:

The data provide evidence that diet has an effect on mean weight among chicks (*F* = 5.46 on 3 and 42 degrees of freedom, *p* = 0.00291).

|  |
| --- |
| Your turn 2 |
| Test for an effect of therapeutic treatment on mean weight change among anorexia patients. Fit the ANOVA model using aov(...) and generate the ANOVA table using summary.aov(...). Interpret the result in context.  # fit anova model  # construct anova table |

### Computing the -value directly

Occasionally you’ll have sums of squares provided to you and will need to complete an ANOVA test “by hand”. In this case you’ll need to refer to the layout of the table to remember the following:

The calculations would look something like what is shown below. You’ll have to imagine that in place of extracting sums of squares and degrees of freedom from a fitted ANOVA model, you’d plug them in directly.

# view ss and df  
fit

Call:  
 aov(formula = weight ~ diet, data = chicks)  
  
Terms:  
 diet Residuals  
Sum of Squares 55881.02 143190.31  
Deg. of Freedom 3 42  
  
Residual standard error: 58.38915  
Estimated effects may be unbalanced

# run but ignore -- this extracts sums of squares and degrees of freedom  
fit.ss <- summary.aov(fit)[[1]]$`Sum Sq`  
fit.df <- summary.aov(fit)[[1]]$Df  
names(fit.df) <- names(fit.ss) <- c('treatment', 'error')  
  
# mean squares  
fit.ms <- fit.ss/fit.df  
fit.ms

treatment error   
18627.007 3409.293

# f statistic  
fstat <- fit.ms['treatment']/fit.ms['error']  
fstat

treatment   
 5.463598

# p value  
pf(fstat, df1 = fit.df['treatment'], df2 = fit.df['error'], lower.tail = F)

treatment   
0.002909054

|  |
| --- |
| Your turn 3 |
| Compute the statistic and -value manually for the test you performed above.  # plug in the sums of squares and degrees of freedom from your fitted model above fit.ss <- c(614.644, 3910.742) fit.df <- c(1, 1) # REPLACE names(fit.df) <- names(fit.ss) <- c('treatment', 'error')  # mean squares  # f statistic  # p value |

### Practice problems

In a study[[1]](#footnote-33), female mice were randomly assigned to six treatment groups to investigate whether restricting dietary intake increases life expectancy:

* [NP] mice ate unlimited amount of nonpurified, standard diet
* [N/N85] normal diet before weaning and normal diet after weaning (85 kcal/wk)
* [N/R50] normal diet before weaning and reduced calorie diet after weaning (50 kcal/wk)
* [N/R40] normal diet before weaning and reduced diet after weaning (40 Kcal/wk)

1. [L3, L4, L9] The longevity dataset contains observations of lifetime in months for 237 mice for the study above. In this problem you’ll test whether diet restriction has an effect on longevity.
   1. [L3] Make a boxplot of the data, orienting the boxes vertically, and compute summary statistics (mean, standard deviation, and sample size) for each treatement group. Do assumptions for inference using ANOVA seem met?
   2. [L4] Compute point estimates and standard errors for the mean lifetime in each group.
   3. [L9] Test for an effect of diet restriction. Provide the statistic, numerator and denominator degrees of freedom for the model, and -value.
   4. [L9] Interpret the result of the test in context following the narrative format from class.
2. [L9] A study investigating physiological differences among species of predatory crabs recorded measurements of propodus height of the claws of crabs from three predatory species. The output below shows the result of an ANOVA model fitted to this data.
   1. Construct the ANOVA table to test the hypothesis that mean weight loss does not differ among the three diets. (*Hint:* the -value from an F model can be computed as pf(fstat, num.df, denom.df, lower.tail = F).)
   2. Report the result of the test in context following the narrative style from class.

Call:  
 aov(formula = height ~ species, data = crab)  
  
Terms:  
 species Residuals  
Sum of Squares 46.22818 137.15524  
Deg. of Freedom 2 35  
  
Residual standard error: 1.979576  
Estimated effects may be unbalanced

1. Weindruch, R., Walford, R.L., Fligiel, S. and Guthrie D. (1986). The Retardation of Aging in Mice by Dietary Restriction: Longevity, Cancer, Immunity and Lifetime Energy Intake, Journal of Nutrition 116(4):641–54. [↑](#footnote-ref-33)