Lab 11: analysis of variance

STAT218

The goal of this lab is to learn how to implement analysis of variance in R. In its most basic form, this comprises three steps:

1. Construct a graphical display of the data
2. Calculate grouped summaries
3. Fit an ANOVA model and generate the ANOVA table

Boxplots are commonly used as graphical displays, which you already know how to make. We will include that step but focus on graphical summaries and ANOVA calculations. You’ll learn a few new commands, some of which come from the widely-used tidyverse package:

* grouped summaries using group\_by() and summarize()
* fitting anova models with aov()

## Performing ANOVA in R

Here we’ll reproduce an analysis of variance using the anorexia data. Data come from a study of young anorexic women that measured weight change before and after therapeutic treatment for two therapies, cognitive behavioral therapy (CBT) and family treatment (FT), and a control. Treatment groups were randomly allocated among the study participants, so inferences can be made about the causal effects of the therapies.

library(tidyverse)  
anorexia <- read\_csv('data/anorexia.csv')  
head(anorexia)

To perform an analysis of variance, we need to identify the variable of interest and the grouping variable. These are:

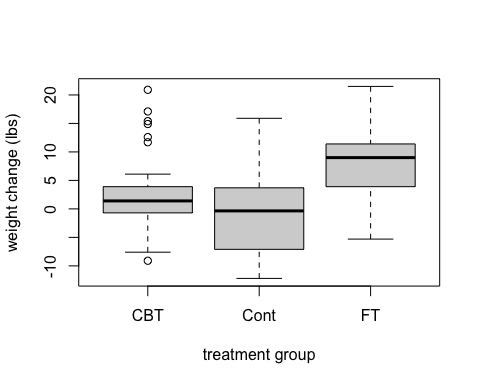
* [variable] change, which gives the change in weight after treatment in pounds
* [grouping] Treat, which indicates the treatment group

As an aside, notice that the raw data are actually before-treatment and after-treatment weight measurements. So there is pairing in the data! We won’t really account for this explicitly, and will just treat change as the sole variable of interest.

### Step 1: graphical summary

It’s always a good idea to plot data before any analysis. This helps identify any unexpected patterns. Boxplots are a fairly standard summary for analysis of variance – after all, we are comparing distributions of a numerical variable (here, weight change) across several categories (here, treatment groups), so this on some level can be thought of as a numeric/categorical comparison.

# boxplot  
boxplot(change ~ Treat, data = anorexia,  
 xlab = 'treatment group', ylab = 'weight change (lbs)')



|  |
| --- |
| Your turn |
| Sometimes you’ll prefer to see the treatment groups in a particular order. This can be accomplished by manually specifying the order, from left to right, in which you want each ‘level’ to appear. This is illustrated below  # rearrange order of treatment groups boxplot(change ~ factor(Treat, levels = c('Cont', 'FT', 'CBT')), data = anorexia,  xlab = 'treatment group', ylab = 'weight change (lbs)')  Try changing the order of treatment groups so that, from left to right, the plot shows FT, then CBT, then the control group. |

### Step 2: grouped summaries

A next step is to calculate means, standard deviations, and sample sizes for each group. This provides you with point estimates (group means), a way to check the equal-variance assumption (comparing standard deviations), and an understanding of how balanced the treatment groups are (sample size).

You already have a bit of experience computing numerical summaries like means and standard deviations. In principle, you could divide up the observations by group, and then calculate each statistic separately. However, there is an easier way:

# grouped summaries: mean, standard deviation, sample size  
anorexia |>  
 group\_by(Treat) |>  
 summarize(group.mean = mean(change),  
 group.sd = sd(change),  
 group.n = n())

# A tibble: 3 × 4  
 Treat group.mean group.sd group.n  
 <chr> <dbl> <dbl> <int>  
1 CBT 3.01 7.31 29  
2 Cont -0.450 7.99 26  
3 FT 7.26 7.16 17

This syntax will look a little unfamiliar. The symbol |> is called the ‘pipe’ operator, and it passes the result of one command as the first argument to the next. Read it as “then”: take anorexia *then* group by treatment *then* summarize.

The summarize step has a series of lines like group.mean = mean(change). These specify a name for the summary on the left, and the actual calculation, in terms of variables in the dataframe, on the right. They are of the form: summary name = calculation. You could use this to compute other statistics too, if you wish.

|  |
| --- |
| Your turn |
| Modify the example above to *also* compute a median for each group.  anorexia |>  group\_by(Treat) |>  summarize(group.mean = mean(change),  group.sd = sd(change),  group.n = n(),  ...) # add a summary here |

### Step 3: ANOVA model and table

The purpose of the graphical and group summaries is to inspect the data before conducting inference. This affords you an opportunity to consider whether model assumptions (normality and equal variance within each group) are reasonable, and to notice anything unusual that might require further scrutiny (outliers, unexpected patterns, etc.).

Analysis of variance models can be fit in a variety of ways in R, but we will use aov(). The syntax is quite similar to what you use to make a boxplot. It requires two inputs:

* a formula specifying the variable of interest and grouping factor
* a dataframe where the variable names in the formula can be found

# fit anova model  
fit <- aov(change ~ Treat, data = anorexia)  
  
# generate table  
summary(fit)

Df Sum Sq Mean Sq F value Pr(>F)   
Treat 2 615 307.32 5.422 0.0065 \*\*  
Residuals 69 3911 56.68   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

|  |
| --- |
| Your turn |
| Try inspecting the fitted model fit. It will show you the sums of squares (raw variability estimates) and corresponding degrees of freedom, along with a “residual standard error”, which gives MSE. The term “residual” is used in place of “error”.  # inspect fitted anova model fit  Call:  aov(formula = change ~ Treat, data = anorexia)  Terms:  Treat Residuals Sum of Squares 614.644 3910.742 Deg. of Freedom 2 69  Residual standard error: 7.528441 Estimated effects may be unbalanced   1. Match each number in the output to the terms in the ANOVA table. 2. What calculations are performed to obtain the MSG and F terms? (Don’t actually do the calculations, just write what they would be.) |

## Practice problem

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.

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)

Test whether diet restriction has an effect on longevity by carrying out the following:

1. Make a boxplot of the data, produce group summaries, write the hypotheses you are testing, and produce an ANOVA table.
2. Write a narrative summary of the test result.

# read in data and preview  
longevity <- read.csv('data/longevity.csv')  
head(longevity)

Lifetime Diet  
1 35.5 NP  
2 35.4 NP  
3 34.9 NP  
4 34.8 NP  
5 33.8 NP  
6 33.5 NP

# visualize the data -- make a boxplot  
  
# calculate grouped summaries  
  
# fit anova model and produce table