Analysis of variance

Assume we have 3 treatment groups each with an n of 4. These data could be any measurements taken under 3 different conditions (e.g. an experimental treatment, a control, a procedural control). In this exercise, we will step through the ANOVA calculations.

```r
library(tidyverse)
control <- c(120, 141, 160, 175)
proc_control <- c(138, 158, 173, 194)
exp_treatment <- c(147, 161, 180, 202)
dat <- data.frame(exp_treatment, control, proc_control)
dat_long <- dat %>% gather(key = treatment, value = y, exp_treatment:proc_control)
```

We begin by calculating the *sums of squares* (SS). This first step partitions the total SS into two sources of variation: among groups (SS_{groups}) and within groups (SS_{error}):

$$SS_{total} = SS_{groups} + SS_{error}$$
$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i n_i(\bar{x}_i - \bar{x})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

where $\bar{x}$ is the grand mean across all observations, i refers to the group to which an individual belongs, and j refers to the jth individual within a group. The SS_{error} can be rewritten as:

$$SS_{error} = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \sum_i s_i^2(n_i - 1)$$

In the code below, we calculate SS_{groups} and SS_{error} using the above equations.

```r
grand_mean <- mean(dat_long$y)
dat_summary <- dat_long %>% group_by(treatment) %>%
  summarise(mean = mean(y), s2 = var(y), n = n(),
            ss_g = n * (mean - grand_mean)^2, # among group variation (SS), for group i
            ss_e = s2 * (n - 1))              # within group variation (SS), for group i
ss_groups <- sum(dat_summary$ss_g)
ss_error <- sum(dat_summary$ss_e)
```

Next, we calculate the *mean squares* (MS), by dividing the SS by the relevant degrees of freedom, df:

$$MS_{groups} = \frac{SS_{groups}}{df_{groups}};\ df_{groups} = k - 1$$
$$MS_{error} = \frac{SS_{error}}{df_{error}};\ df_{error} = N - k$$

where N is the total number of observations and k is the number of groups.

```r
N <- length(dat_long$treatment); k <- length(unique(dat_long$treatment))
df_groups <- k - 1
df_error <- N - k
ms_groups <- ss_groups / df_groups
ms_error <- ss_error / df_error
```

Under the null hypothesis that the population means of all groups are the same, the variation among individuals belonging to different groups (represented by MS_{groups}) will be similar to the variation among individuals belonging to the same group (estimated by MS_{error}). In ANOVA, we test for a difference by calculating the variance ratio, F:

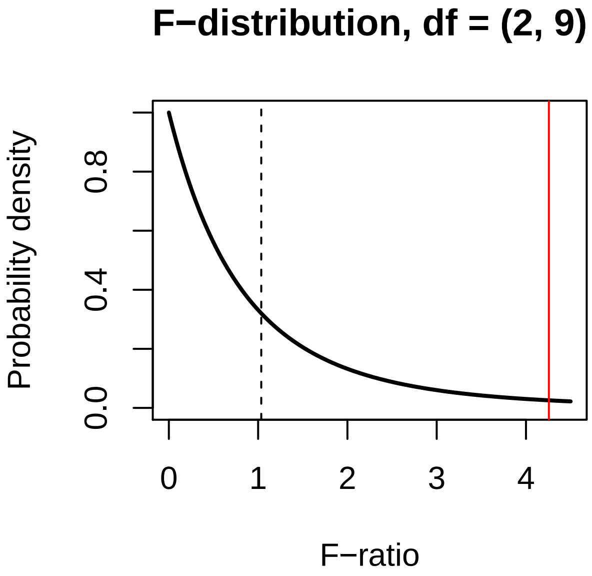
$$F = \frac{MS_{groups}}{MS_{error}}$$

F is the test statistic in ANOVA. Under H_0, it should be ~1. If H_0 is false, we expect F to be greater than 1. To calculate the P-value, we use an F-distribution that requires F, and the df for the numerator in the F-ratio (i.e., MS_{groups}), and the df for the denominator in the F-ratio (i.e., MS_{error}).

```
f <- ms_groups / ms_error
pf(q = f, df1 = df_groups, df2 = df_error, lower.tail = FALSE)
```

Here we visualize the relevant F-distribution, with the calculated F represented as a dashed vertical line. The solid red line represents the critical F-ratio for $\alpha = 0.05$.

```
x <- seq(from = 0, to = 4.5, length = 200)
y_1 <- df(x, df_groups, df_error)
plot(x, y_1, lwd = 2, type = "l", xlim = c(0, 4.5), ylim = c(0, 1),
     main = "F-distribution, df = (2, 9)", xlab = "F-ratio", ylab = "Probability density")
abline(v = f, lty = 2) # calculated F
abline(v = qf(p = 0.05, 2, 9, lower.tail = FALSE),
       lty = 1, col = "red") # Critical F-value at P = 0.05
```



Let's check our work, using the function aov. Inspect the ANOVA table, and be sure you understand all of its components.

```
mod1 <- aov(y ~ treatment, data = dat_long); summary(mod1)
```

On your own

Add 15 to each value in the experimental treatment.

- How do the mean and variance change for the experimental group?
- How do the SS and MS for groups and error change?
- How does F change, and its associated P-value?

References

Whitlock, Schluter. 2015. Analysis of Biological Data.

Prepared by R Elahi