Simple (one-way) Analysis of Variance

Advanced statistical methods and models in experimental design

Bartosz Maćkiewicz

Example: Giancola and Corman (2007)

We will begin with a study by Giancola and Corman (2007). They were interested in studying the effects of a distracting task on aggressive behavior of subjects who had consumed a significant amount of alcohol. It is well known that alcohol often leads to aggressive behavior, but why?

Giancola and Corman began by assuming that alcohol facilitated aggression by focusing attention on more salient provocative cues rather than on less salient inhibitory ones. They reasoned that if they presented their subjects with a distracting task, attention would be focused on the task rather than on provocative cues, thus limiting aggression. However, they also reasoned that if the task became too complex, its distracting effects would disappear and aggression would take over.

Example: Giancola and Corman (2007)

Giancola and Corman asked their subjects to consume alcohol in an amount that raised their average blood alcohol level to about .10%. Subjects then participated in a task that required them to remember the order in which squares in a 3 x 3 matrix were illuminated.

The attentional demands of the task were varied by manipulating the number of squares that subjects had to keep in memory. Subjects played against a fictitious opponent who either delivered mild shocks to the subject or received mild shocks from the subject, dependent on supposed task performance. The dependent variable (aggression) was based on the severity and duration of shocks that subjects delivered to opponents when they had the opportunity.

There were five groups in this study, varying in task difficulty. Subjects had to remember the pattern of either 0, 2, 4, 6, or 8 squares.

A few facts about ANOVA...

- ANOVA, or Analysis of Variance, is a statistical method used to compare multiple group means.
- It was developed by Ronald Fisher in the 1920s.
- ANOVA facilitates the examination of the impacts of one or more experimental factors simultaneously (e.g., one-way, two-way, or three-way ANOVA).
- Examples of such factors include treatment types, gender, and types of fertilizer (a nod to Fisher).
- Each factor can have several levels (e.g., types of fertilizer such as A, B, and C; or varying degrees of humidity).

Theoretical model

Here is a structural model that underlies ANOVA:

$$X_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- X_{ij} - individual observation (i observation from j group)
- μ - grand mean in the population
- τ_j - specific treatment effect of being in Group j
- ϵij - error associated with a specific individual

Assumptions: homogeneity of variance

$$\sigma_1^2 = \sigma_2^2 = ... = \sigma_3^2 = \sigma_\epsilon^2$$

The subscript "ϵ" stands for error, and this variance is the error variance – the variance unrelated to any treatment differences, which is the variability of scores within the same condition. Homogeneity of variance would be expected to occur if the effect of a treatment is to add a constant to everyone's score – if, for example, everyone who was trying to recall the position of eight stimuli scored an extra point above the others on the aggression scale.

Assumptions: normality

A second assumption of the analysis of variance is that the observations for each condition are normally distributed around their group mean.

Because σ_e^2 represents the variability of observations around the mean of that condition, a more correct way to write the assumption is to write that error is normally distributed within conditions.

Thus you will often see the assumption stated in terms of "the normal distribution of error." or "normally distributed residuals," which means the same thing.

Moderate departures from normality are not usually fatal. We said much the same thing when looking at the t test for two independent samples, which is really just a special case of the analysis of variance.

Assumptions: independence

The third important assumption is that the observations are independent of one another.

Thus for any two observations within an experimental treatment, we assume that knowing how one of these observations stands relative to the treatment (or population) mean tells us nothing about the other observation.

This is one of the important reasons why subjects are randomly assigned to groups. Violation of the independence assumption can have serious consequences for an analysis.

Hypotheses

Here are a null hypothesis and an alternative hypothesis for ANOVA:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n$$

$$H_A : \neg H_0$$

The null hypothesis could be false in a number of ways (e.g., all means could be different from each other, the first two could be equal to each other but different from the last three, and so on), but for now we are going to be concerned only with whether the null hypothesis is completely true or is false. This is frequently referred to as the **omnibus null hypothesis**.

The Logic of the Analysis of Variance

```r
giancola <- read.csv("giancola2007.csv")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.
```

```r
head(giancola)
```

```
##   Group   dv
## 1     0 1.28
## 2     0 1.35
```

The Logic of the Analysis of Variance: two estimators of variance

Consider for a moment the effect of our three major assumptions - **normality**, **homogeneity of variance**, and the **independence of observations** (or residuals).

By making the first two of these assumptions we have said that the five distributions have the same shape and dispersion. As a result, the only way left for them to differ is in terms of their means.

The Logic of the Analysis of Variance: mean square error

We will begin by making no assumption concerning H_0 – it may be true or false. For any one treatment, the variance of the 12 scores in that group would be an estimate of the variance of the population from which the scores were drawn. Because we have assumed that all populations have the same variance, it is also one estimate of the common population variance. We can thus construct a following estimator of population variance, which is equivalent to calculating the variance for each level of the factor separately and taking a mean:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$MSE = \frac{SSE}{n - k}$$

The Logic of the Analysis of Variance: mean square error

It is important to note that this estimate does not depend on the truth or falsity of H_0, because s_j^2 is calculated on each sample separately.

```
giancola %>%
  group_by(Group) %>%
  summarize(vars = var(dv)) %>%
  summarize(MSE = mean(vars)) %>%
  pull(MSE)
```

```
## [1] 2.262875
```

The Logic of the Analysis of Variance: mean square of treatments

Now let us assume that H_0 is true. If this is the case, then our five samples of $n = 12$ cases can be thought of as five independent samples from the same population (or, equivalently, from five identical populations), and we can produce another possible estimate of the population variance.

Recall that the Central Limit Theorem states that:

$$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Therefore:

$$\sigma^2 = n s_{\bar{X}}^2$$

```
giancola %>%
  group_by(Group) %>%
  summarize(means = mean(dv), n = n()) %>%
  summarize(MST = var(means) * 12) %>%
  pull(MST)
```

The Logic of the Analysis of Variance: two estimates

We now have two estimates of the population variance (σ_ϵ^2). One of these estimates (MS_{error}) is independent of the truth or falsity of H_0. The other ($MS_{treatment}$) is an estimate of σ_ϵ^2 only as long as H_0 is true (only as long as the conditions of the central limit theorem are met; namely, that the means are drawn from one population or several identical populations).

Thus, if the two estimates agree, we will have support for the truth of H_0, and if they disagree, we will have support for the falsity of H_0.

The Logic of the Analysis of Variance: degrees of freedom

With 59 *df* overall (i.e., $N - 1$), four of these are associated with differences among treatment means and the remaining 55 are associated with variability within the treatment groups.

The total degrees of freedom is always $N - 1$, where N is the total number of observations. The number of degrees of freedom between treatments is always $k - 1$, where k is the number of treatments. The number of degrees of freedom for error is most easily thought of as what is left over. However, it can be calculated more directly as the sum of the degrees of freedom within each treatment.

The Logic of the Analysis of Variance: F ratio and F distribution

F is obtained by dividing MS_{treat} by MS_{error}.

As noted earlier, MS_{error} is an estimate of the population variance (σ^2_ϵ). Moreover MS_{treat} is an estimate of the population variance if H_0 is true, but not if it is false. If H_0 is true, then MS_{error} and MS_{treat} are both estimating the same thing, and as such they should be approximately equal. If this is the case, the ratio of one to the other will be approximately 1, give or take a certain amount for sampling error.

Thus, all we have to do is to compute the ratio and determine whether it is close enough to 1 to indicate support for the null hypothesis.

```
f_stat <- 15.6 / 2.2629
f_stat
```

```
## [1] 6.893809
```

```
pf(f_stat, 5-1, # k - 1 df
   nrow(giancola) - 5, # n - 1 df
   lower.tail = F)
```

```
## [1] 0.0001427235
```

The Logic of the Analysis of Variance: using R

We can of course do it using R:

```
fit <- aov(dv ~ factor(Group), data = giancola)
summary(fit)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Group)   4  62.46  15.615   6.901 0.000142 ***
## Residuals      55 124.46   2.263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc tests

When comparing the means for the levels of a factor in an analysis of variance, a simple comparison using t-tests will inflate the probability of declaring a significant difference when it is not in fact present.

One way to deal with this problem is to use one of the specialized *post-hoc* test, that take it into account.

Post-hoc tests: Tukey Hones Significant Differences (HSD) test

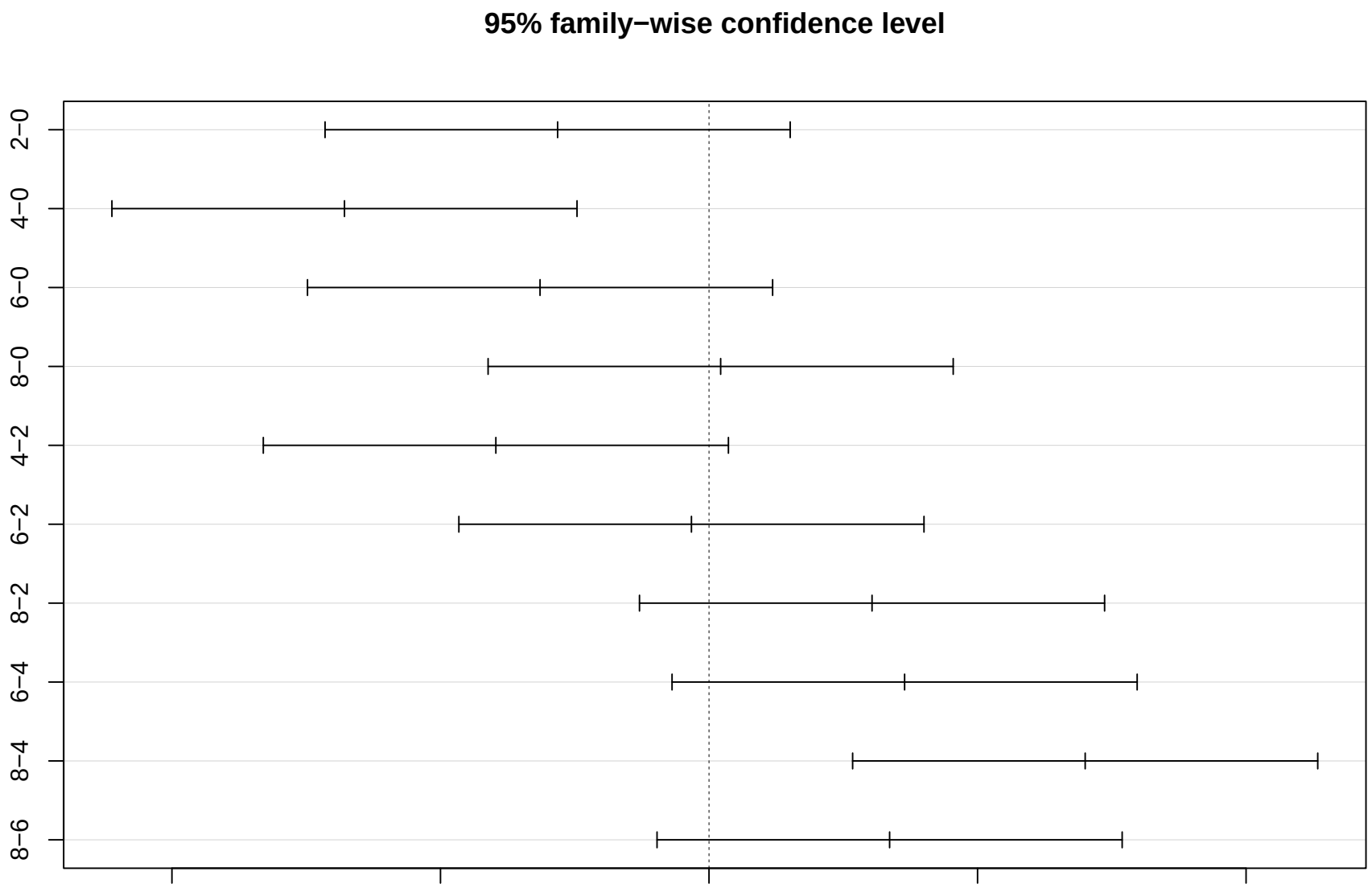
```
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dv ~ factor(Group), data = giancola)
##
## $`factor(Group)`
##             diff        lwr        upr     p adj
## 2-0 -1.12750000 -2.8595252  0.6045252 0.3639370
## 4-0 -2.71500000 -4.4470252 -0.9829748 0.0004352
## 6-0 -1.25833333 -2.9903586  0.4736919 0.2570413
## 8-0  0.08666667 -1.6453586  1.8186919 0.9999078
## 4-2 -1.58750000 -3.3195252  0.1445252 0.0872492
## 6-2 -0.13083333 -1.8628586  1.6011919 0.9995272
## 8-2  1.21416667 -0.5178586  2.9461919 0.2906677
## 6-4  1.45666667 -0.2753586  3.1886919 0.1385559
## 8-4  2.80166667  1.0696414  4.5336919 0.0002702
## 8-6  1.34500000 -0.3870252  3.0770252 0.1987677
```

Post-hoc tests: a visualization

Visualizing results from the Tukey HSD test can help in understanding the differences between group means more intuitively.

```
plot(TukeyHSD(fit))
```



Effect size

For ANOVA the most commonly used measure of effect size is η^2.

Field (2013) gives the following rules for interpreting η^2

- $\eta^2 < 0.01$ - Very small
- $0.01 \leq \eta^2 < 0.06$ - Small
- $0.16 \leq \eta^2 < 0.14$ - Medium
- $\eta^2 \geq 0.14$ - Large

```
effectsize::eta_squared(fit)
```

```
## For one-way between subjects designs, partial eta squared is
##   to eta squared. Returning eta squared.
```

```
## # Effect Size for ANOVA
##
## Parameter     | Eta2 |       95% CI
## ------------------------------------
## factor(Group) | 0.33 | [0.14, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```