One-way ANOVA

Kristian Wichmann

April 3, 2017

ANOVA is short for ANalysis Of VAriance. Analysis may be done for a number of *factors*, i.e. groupings of potentially explanatory categories. This document deals with the case for one factor, hence the term "one-way".

1 The situation

Assume we have a samples, each from a different subpopulation. Each sample is of size n, and the total number of samples is called $N = na$. The question we could ask ourselves is whether there's a difference between the a subpopulation means? If $a = 2$ we can do a t-test comparison, but for $a \geq 3$ other methods are needed.

2 Notation

The stochastic variables corresponding to sample j from subpopulation i is denotes X_{ij}. The estimator of the mean of the i'th subpopulation is denoted:

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij} \tag{1}$$

The overall sample mean, also sometimes known as the *grand mean*, is denoted by $\bar{\bar{X}}$:

$$\bar{\bar{X}} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{n} X_{ij} = \frac{1}{n} \sum_{j=1}^{n} \bar{X}_j \tag{2}$$

Any actualized values of these are denoted by using lower case x'es. The true subpopulation means are denoted $\mu_1, \mu_2, \cdots, \mu_a$.

2.1 Example: Use of fertilizer

12 plots of land are divided randomly into three groups: 1, 2 and 3. Group 1 and 2 are fertilized while 3 (the control) is not. The table below shows the crop yield[1]:

Group	1	2	3
	75	74	60
	70	78	64
	66	72	65
	69	68	55
Average	$\bar{x}_1 = 70$	$\bar{x}_2 = 73$	$\bar{x}_3 = 61$

So here $a = 3, n = 4$ and $N = 12$. The means for each group is shown in the table. Finally, the grand mean is:

$$\bar{\bar{x}} = \frac{1}{3}(70 + 73 + 61) = 68 \tag{3}$$

3 Null hypothesis

We wish to be able to detect differences between the subpopulations. For this, a null hypothesis is needed. We will use the following:

$$H_0: \text{ All subpopulations are distributed as: } X_{ij} \sim N(\mu, \sigma^2) \tag{4}$$

So the subpopulations really form one, big population. Specifically $\mu = \mu_1 = \mu_2 = \cdots = \mu_a$. The variances are also the same.

4 Splitting the sum of squares

We're interested in variance, so we need to compute the total sum of squares for all the observations:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(X_{ij} - \bar{\bar{X}} \right)^2 \tag{5}$$

Now use the usual trick of inserting zero:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(X_{ij} - \bar{X}_i + \bar{X}_i - \bar{\bar{X}} \right)^2 \tag{6}$$

[1]This is from problem 10-1 from Wonnacot & Wonnacot: Introductory statistics, fifth edition.

Expand $\left(X_{ij} - \bar{X}_i + \bar{X}_i - \bar{\bar{X}}\right)^2$:

$$\left(X_{ij} - \bar{X}_i\right)^2 + \left(\bar{X}_i - \bar{\bar{X}}\right)^2 + 2\left(X_{ij} - \bar{X}_i\right)\left(\bar{X}_i - \bar{\bar{X}}\right) \tag{7}$$

When summing over the last term we get:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)\left(\bar{X}_i - \bar{\bar{X}}\right) = \sum_{i=1}^{a}\left[\left(\bar{X}_i - \bar{\bar{X}}\right)\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)\right] \tag{8}$$

But:

$$\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right) = n\bar{X}_i - n\bar{X}_i = 0 \tag{9}$$

So the sum over the cross-term vanishes and we're left with:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2 + n\sum_{i=1}^{a}\left(\bar{X}_i - \bar{\bar{X}}\right)^2 \tag{10}$$

Or, setting $SS_A = n\sum_{i=1}^{a}\left(\bar{X}_i - \bar{\bar{X}}\right)^2$ and $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$ this may be compactly written:

$$SS_T = SS_A + SS_E \tag{11}$$

The E is short for 'error', as it cannot be attributed to any explanatory factor.

4.1 Example: Use of fertilizer (cont.)

For this example, the sum of squares for the group factor (A) is:

$$SS_A = \underbrace{4}_{n}\left[(70-68)^2 + (73-68)^2 + (61-68)^2\right] = 312 \tag{12}$$

The error contribution is:

$$\begin{aligned} SS_E = &\underbrace{(75-70)^2 + (70-70)^2 + (66-70)^2 + (69-70)^2}_{\text{Group 1}} + \\ &\underbrace{(74-73)^2 + (78-73)^2 + (72-73)^2 + (68-73)^2}_{\text{Group 2}} + \\ &\underbrace{(60-61)^2 + (64-61)^2 + (65-61)^2 + (55-61)^2}_{\text{Group 3}} = \\ &156 \end{aligned}$$

Let's calculate the total sum of squares too to see if things add up:

$$\begin{aligned} SS_T = &\underbrace{(75-68)^2 + (70-68)^2 + (66-68)^2 + (69-68)^2}_{\text{Group 1}} + \\ &\underbrace{(74-68)^2 + (78-68)^2 + (72-68)^2 + (68-68)^2}_{\text{Group 2}} + \\ &\underbrace{(60-68)^2 + (64-68)^2 + (65-68)^2 + (55-68)^2}_{\text{Group 3}} = \\ &468 \end{aligned}$$

Since $468 = 312 + 156$ everything seems to be in order.

5 Terms as quadratic forms

We'd like to express the sum of squares terms as quadratic forms. This will allow us to apply Cochran's theorem to the problem.

The relevant matrices should be $N \times N$. We need to normalize the X_{ij} variables to get standard normals, and we need to decide on an ordering. Here we'll use the following:

$$\begin{array}{ccccc} U_1 = \frac{X_{11}-\mu}{\sigma}, & U_2 = \frac{X_{12}-\mu}{\sigma}, & \cdots & U_n = \frac{X_{1n}-\mu}{\sigma} \\ U_{n+1} = \frac{X_{21}-\mu}{\sigma}, & U_{n+2} = \frac{X_{22}-\mu}{\sigma}, & \cdots & U_{2n} = \frac{X_{2n}-\mu}{\sigma} \\ \vdots & \vdots & \ddots & \vdots \\ U_{(a-1)n+1} = \frac{X_{a1}-\mu}{\sigma}, & U_{(a-1)n+2} = \frac{X_{a2}-\mu}{\sigma}, & \cdots & U_N = \frac{X_{an}-\mu}{\sigma} \end{array}$$

For SS_T this is pretty straightforward; it's the same as always for a basic mean estimator:

$$\frac{SS_T}{\sigma^2} = U^T \underbrace{\left(I_N - \frac{1}{N} J_N\right)}_{B^{(T)}} U \tag{13}$$

As usual $B^{(T)}$ is idempotent with rank $N-1$. For the other terms it's useful

to introduce the following $a \times N$ matrix:

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{14}$$

This should be read as n rows of the first unit vector, n rows of the second, and so on till the a'th unit vector. Such an encoding is known as *one-hot* encoding, since a one indicates the group. It is also sometimes known as a *dummy variable*. A is called the *design matrix* for one-way ANOVA[2]. This is tied to the general linear model. Here, to solve the normal equations, we want to invert the matrix A^tA. But this is simply a diagonal matrix with n in all the diagonal entries. Hence the inverse is $\frac{1}{n}I_a$. This means that the OLS estimator is:

$$(A^tA)^{-1}A^tX = \frac{1}{n}A^tX \tag{15}$$

The entries of the $a \times 1$ matrix are simply the sums of each X corresponding to the relevant group. Hence - unsurprisingly - the BLUE estimators of the true group means are the sample group averages.

6 Distribution of sum of square terms, and mean squares

Now, let's look at the distribution of the different terms. We will assume $X_{ij} \sim N(\mu, \sigma^2)$ pr. the null hypothesis. For SS_T this means:

$$\bar{\bar{X}} \sim N\left(\mu, \frac{\sigma^2}{N}\right), \quad SS_T \sim \sigma^2\chi^2_{N-1} \tag{16}$$

[2]It's also common to denote this matrix as X, but with X_{ij} representing stochastic variables, A is hopefully less confusing

Similarly for the means of each group:

$$\bar{X}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \sum_{j=1}^{n} \left(X_{ij} - \bar{X}\right)^2 \sim \sigma^2 \chi^2_{n-1} \tag{17}$$

This implies the following distribution for SS_E:

$$SS_E \sim \sigma^2 \chi^2_{a(n-1)} \tag{18}$$

Since $\frac{SS_T}{\sigma^2}$ is χ^2 distributed, (11) along with Cochran's theorem implies, that $\frac{SS_A}{\sigma^2}$ must also be χ^2 distributed. The number of degrees of freedom should add up, so it must be $N - 1 - a(n-1) = a - 1$. So:

$$SS_A \sim \sigma^2 \chi^2_{a-1} \tag{19}$$

Now that we know the degrees of freedom for all the χ^2-distributed sum of square terms, we can also calculate the corresponding mean square (MS) expressions (sometimes informally known as variances):

$$MS_T = \frac{SS_T}{N-1}, \quad MS_A = \frac{SS_A}{a-1}, \quad MS_E = \frac{SS_E}{a(n-1)} \tag{20}$$

6.1 Example: Use of fertilizer (cont.)

We use equation (20) to get:

$$MS_T = \frac{468}{12-1} = 52.55, \quad MS_A = \frac{312}{3-1} = 156, \quad MS_E = \frac{156}{3(4-1)} = 17.33 \tag{21}$$

7 The F test statistic

We're looking for a statistic that measures how well the null hypothesis 4 is satisfied. If the null is true, we should get no contribution from the variations due to factor, i.e. SS_A. On the other hand, if SS_E is small, it means that the variance is due to random errors. In other words, the smaller the following number is, the more likely the null seems:

$$F = \frac{MS_A}{MS_E} \tag{22}$$

Because of equation (20) this quantity is F-distributed with $a-1$ numerator degrees of freedom, and $a(n-1)$ denominator degrees of freedom under H_0. Therefore it can be used to perform tests. Critical values for a given confidence level and/or p-values can be found using a table or a computer program. The null is accepted/rejected accordingly.

7.1 Example: Use of fertilizer (cont.)

Let's test at a 5% significance level. The F test statistic is:

$$F = \frac{156}{17.33} = 9 \tag{23}$$

The numerator degrees of freedom is 2, and the denominator degrees of freedom is 9. The critical value for the corresponding F-distribution is found in a table to be 4.26. Since the test statistic is larger than the critical value, the null is rejected: There really does seem to be a difference in crop yield between the groups. In fact the p-value can be calculated to be 0.0071, making the claim rather strong.

8 Unequal groupings

Sometimes the individual groups don't have the same number of elements in them.