MS Module 11: Single-Factor ANOVA – practice problems

(The attached PDF file has better formatting.)

Exercise 11.1: ANOVA (one-way)

A experiment has three groups and four observations in each group. The groups have subscripts *i* = 1,2,3 and the observations have subscripts *j* = 1,2,3,4 (following the notation in the textbook).

	1	2	3	4
group 1	11	12	13	14
group 2	5	6	10	15
group 3	7	8	21	26

The columns in this table do not affect the solution. The observations are shown in increasing order for each group, but they could be shown in any order.

The groups are normally distributed with the same variance.

The null hypothesis is that the means of the groups are the same: H_0: $\mu_1 = \mu_2 = \mu_3$

We test (i) whether the null hypothesis should be rejected, (ii) which group means differ significantly, and (iii) whether the variances by group differ, using the computational formulas in the textbook.

Final exam problems ask for the SSTr, the SSE, the MSTr, the MSE, the *F* value, the *p* value, Tukey's honestly statistical difference, Levene's test, and various qualitative items and expected values.

A. What is the square of the sum of all the observations, or $x_{..}^2$?
B. What is the sum of the squares of all the observations, or $\Sigma_i \Sigma_j x_{ij}^2$?
C. What is SST, the total sum of squares?
D. What is SSTr, the treatment sum of squares?
E. What is SSE, the error sum of squares?
F. What are the total degrees of freedom?
G. What are the degrees of freedom for the group?
H. What are the degrees of freedom for the error sum of squares (SSE)?
I. What is MSTr, the mean squared deviation for the groups?
J. What is MSE, the mean squared error?
K. What is the *F* value for testing the null hypothesis?
L. If the null hypothesis is true, what is the expected *F* value?
M. What is the *p* value for this test of the null hypothesis?

Part A: The sums of the observations by group and in total are

	1	*2*	*3*	*4*	*Total*
Group 1	11	12	13	14	50
Group 2	5	6	10	15	36
Group 3	7	8	21	26	62
Total					148

The sum of all the observations is 11 + 12 + 13 + 14 + 5 + 6 + 10 + 15 + 7 + 8 + 21 + 26 = 148

The square of this sum is $148^2 = 21{,}904$

Part B: The sum of the squares of all the observations is

$$11^2 + 12^2 + 13^2 + 14^2 + 5^2 + 6^2 + 10^2 + 15^2 + 7^2 + 8^2 + 21^2 + 26^2 = 2{,}246$$

Part C: One can compute the total sum of squares two ways:

- Using the definition of the total sum of squares: subtract the mean value from each observation and compute the sum of the squares of the differences (not shown here).

- Using the computational formula: compute the sum of the squares of the observations and subtract the square of the sum of the observations divided by the number of observations:

$$SST = \Sigma_i \Sigma_j x_{ij}^2 - x_{..2} / N = 2{,}246 - 21{,}904 / 12 = 420.67$$

Question: What is the intuition for this formula?

Answer: The textbook proves this formula by expanding the definition of the total sum of squares. The intuition for the formula may be grasped by starting with no differences among the groups and random errors, so all observations are the same, and then adding random errors.

If all N observations are the same value Z, the square of the sum is $(N \times Z)^2$ and the sum of the N squares is $N \times Z^2$, so the formula $SST = \Sigma_i \Sigma_j x_{ij}^2 - x_{..2} / N$ gives zero. If the observations differ from each other, the sum of squares increases even if the square of the sum does not change.

The total sum of squares does not distinguish between (i) differences in average values by group and (ii) the variance of the error term. These two parts are reflected in the treatment sum of squares (SSTr) and the error sum of squares (SSE), not in the total sum of squares (SST).

Illustration: Keep the total $N \times Z$ but increase half the observations by a constant *k* and decrease half the observations by a constant *k*. The sum of the N squares is $\frac{1}{2} N \times (Z+k)^2 + \frac{1}{2} N \times (Z-k)^2 = N \times Z^2 + N \times k^2$.

Part D: We derive first the sum of the squares of the group totals, or $\Sigma_i x_{i.2}$

- Group 1: 11 + 12 + 13 + 14 = 50
- Group 2: 5 + 6 + 10 + 15 = 36
- Group 3: 7 + 8 + 21 + 26 = 62

The sum of squares of these group totals is $\Sigma_i x_i^2 = 50^2 + 36^2 + 62^2 = 7{,}640$

The treatment sum of squares = $SSTr = \Sigma_i x_{i.2} / J - x_{..2} / N = 7{,}640 / 4 - 21{,}904 / 12 = 84.6667$, where J is the number of observations in each group and N is the total number of observations.

Question: What is the intuition for this formula?

Answer: The formula is the same as the formula for the total sum of squares except that we use only the totals of the observations for each group.

Each observation for the group totals is the sum of four individual observations. The sum of squares of group totals looks at the variability among groups, ignoring any variability among observations in a group.

Part E: We can compute the error sum of squares SSE two ways:

The computation formula (expanding the definition of the SSE and simplifying): $SSE = \Sigma_i \Sigma_j x_{ij}^2 - \Sigma_i x_{i.2} / J$

= the sum of squares of all the observations
– the sum of squares of the group totals ÷ the number of observations in each group

$$= 2{,}246 - 7{,}640 / 4 = 336.00$$

Alternatively, we derive SSE by subtraction.

$$SST = SSTr + SSE \Rightarrow SSE = SST - SSTr = 420.6667 - 84.6667 = 336.00$$

Question: Do we infer that group differences (84.67) are not material compared to random fluctuations (336)?

Answer: The ratio of the SSTr to the SSE is affected by the number of observations in each group. The *F* test compares the MSTr to the MSE, where each sum of squares is divided by its degrees of freedom. The SSE is much smaller than the SSTr, the MSE is about the same size as the MSTr.

Part F: The degrees of freedom for the total sum of squares = the number of observations – 1 = 12 – 1 = 11.

Part G: The degrees of freedom for the groups (SSTr) = the number of groups – 1 = 3 – 1 = 2.

Part H: The degrees of freedom for the error sum of squares = the number of groups × (the observations in each group – 1) = 3 × (4 – 1) = 9.

The degrees of freedom for SST = degrees of freedom for SSTr + degrees of freedom for SSE: 11 = 2 + 9.

Part I: The treatment mean squared = SSTr / df = 84.66667 / 2 = 42.3333

Part J: The mean squared error = SSE / df = 336 / 9 = 37.3333

The degrees of freedom (df) = 2 for the groups and 9 for the random error term. The mean values show that the observed group differences imply a slightly greater variance than the random error term does.

Part K: The *F* value is 42.3333 / 37.3333 = 1.1339.

Part L: If the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ is true, the expected *F* value is close to one (if the number of observations is large). An accurate statement is that if the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ is true, the expected value of the numerator of the *F* value equals the expected value of the denominator of the *F* value. The expected value of a ratio is not equal to the ratio of the expected values, but the expected *F* value is close to one, especially if the ANOVA has many observations.

Question: Does a value greater than one imply that the *F* test is significant?

Answer: The observed group differences are caused by the combination of the true group differences and the random errors.

If the groups have the same expected means, the expected MSTr is the same as the expected MSE. In this practice problem, the MSTr is slightly greater than the MSE. This difference might be caused by chance. The *F* test shows the probability that group differences of this size or greater might arise from random fluctuations even if the expected groups means are the same. The *p* value for this *F* test gives the probability.

Part M: The *p* value depends on three items: the *F*-value, the degrees of freedom in the numerator, and the degrees of freedom in the denominator.

The cumulative distribution function for *F*=1.133929 with 2 degrees of freedom in the numerator and 9 degree of freedom in the denominator is 0.636248. The *p* value is the complement of the cumulative distribution function (for a one-tailed *F* test) = 1 – 0.636248 = 0.3638. If the null hypothesis is true, the likelihood of getting an *F* test this high is 36.38%.

Question: How does the F test vary with the F value and the degrees of freedom?

Answer: If the F value is close to one, the group expected differences are not contributing to the observed differences and the p value should be close to 50%. With 2 and 9 degrees of freedom, an F value of one gives a p value of 40.5%. If the F value is large, the p value is close to zero.

- As the degrees of freedom in the numerator increase, the p value increases
    - As the number of groups increase, testing hypotheses is harder, as apparent differences reflect the greater number of parameters. The degrees of freedom in the numerator increase, and the p value increases (becomes less significant).
- As the degrees of freedom in the denominator increase, the p value decreases.
    - As the number of observations increase, testing hypotheses is easier. The degrees of freedom in the denominator increase, and the p value decreases (becomes more significant).

Question: What does the sum of squares of each observation tell us? We are testing whether the group means differ, not whether the observations in the groups have particular variances. The analysis of variance assumes the groups have normal distributions with the same variance, and we do not test this assumption. If Group 1 has a mean observation of 20, what difference does it make if the observations are 18, 19, 21, and 22 vs 12. 18, 22, and 28?

Answer: The significance of the ANOVA depends on the variance of the random error term.

- If the observations in a group are similar, the MSE is small, and differences in the group means cause a high F value.
- If the observations in a group are dis-similar, the MSE is large, and differences in the group means cause a lower F value.

Analysis of variance compares two estimates of σ^2: the MSE and the MSTr. These two estimates are equally influential: the F value and p value depend on the ratio of the MSTr to the MSE.