



Understanding one-way ANOVA using conceptual figures

Tae Kyun Kim

Department of Anesthesia and Pain Medicine, Pusan National University Yangsan Hospital and School of Medicine, Yangsan, Korea

Analysis of variance (ANOVA) is one of the most frequently used statistical methods in medical research. The need for ANOVA arises from the error of alpha level inflation, which increases Type 1 error probability (false positive) and is caused by multiple comparisons. ANOVA uses the statistic F, which is the ratio of between and within group variances. The main interest of analysis is focused on the differences of group means; however, ANOVA focuses on the difference of variances. The illustrated figures would serve as a suitable guide to understand how ANOVA determines the mean difference problems by using between and within group variance differences.

Key Words: Analysis of variance, False positive reactions, Statistics.

Introduction

The differences in the means of two groups that are mutually independent and satisfy both the normality and equal variance assumptions can be obtained by comparing them using a Student's t-test. However, we may have to determine whether differences exist in the means of 3 or more groups. Most readers are already aware of the fact that the most common analytical method for this is the one-way analysis of variance (ANOVA). The present article aims to examine the necessity of using a one-way ANOVA instead of simply repeating the comparisons using Student's t-test. ANOVA literally means analysis of variance, and the present article aims to use a conceptual illustration to explain

how the difference in means can be explained by comparing the variances rather by the means themselves.

Significance Level Inflation

In the comparison of the means of three groups that are mutually independent and satisfy the normality and equal variance assumptions, when each group is paired with another to attempt three paired comparisons¹⁾, the increase in Type I error becomes a common occurrence. In other words, even though the null hypothesis is true, the probability of rejecting it increases, whereby the probability of concluding that the alternative hypothesis (research hypothesis) has significance increases, despite the fact that it has no significance.

Let us assume that the distribution of differences in the means of two groups is as shown in Fig. 1. The maximum allowable error range that can claim "differences in means exist" can be defined as the significance level (α). This is the maximum probability of Type I error that can reject the null hypothesis of "differences in means do not exist" in the comparison between two mutually independent groups obtained from one experiment. When the null hypothesis is true, the probability of accepting it becomes $1-\alpha$.

Now, let us compare the means of three groups. Often, the

Corresponding author: Tae Kyun Kim, M.D., Ph.D.
Department of Anesthesia and Pain Medicine, Pusan National University Yangsan Hospital and School of Medicine, 20, Geumo-ro, Mulgeum-eup, Yangsan 50612, Korea
Tel: 82-55-360-2129, Fax: 82-55-360-2149
Email: anesktk@pusan.ac.kr
ORCID: <http://orcid.org/0000-0002-4790-896X>

Received: November 14, 2016.
Revised: December 5, 2016 (1st); December 8, 2016 (2nd).
Accepted: December 9, 2016.

Korean J Anesthesiol 2017 February 70(1): 22-26
<https://doi.org/10.4097/kjae.2017.70.1.22>

¹⁾ A, B, C three paired comparisons: A vs B, A vs C and B vs C.

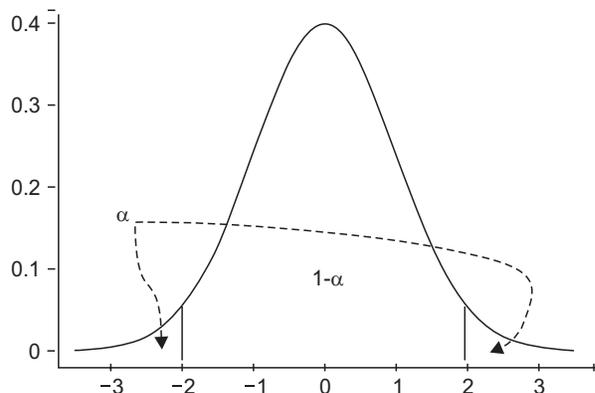


Fig. 1. Significant level alpha to reject null hypothesis.

null hypothesis in the comparison of three groups would be “the population means of three groups are all the same,” however, the alternative hypothesis is not “the population means of three groups are all different,” but rather, it is “at least one of the population means of three groups is different.” In other words, the null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_1 \neq \mu_3 \text{ or } \mu_2 \neq \mu_3$$

Therefore, among the three groups, if the means of any two groups are different from each other, the null hypothesis can be rejected.

In that case, let us examine whether the probability of rejecting the entire null hypothesis remains consistent, when two continuous comparisons are made on hypotheses that are not mutually independent. When the null hypothesis is true, if the null hypothesis is rejected from a single comparison, then the entire null hypothesis can be rejected. Accordingly, the probability of rejecting the entire null hypothesis from two comparisons can be derived by firstly calculating the probability of accepting the null hypothesis from two comparisons, and then subtracting that value from 1. Therefore, the probability of rejecting the entire null hypothesis from two comparisons is as follows:

$$1 - (1 - \alpha)(1 - \alpha)$$

If the comparisons are made n times, the probability of rejecting the entire null hypothesis can be expressed as follows:

$$1 - (1 - \alpha)^n$$

It can be seen that as the number of comparisons increases, the probability of rejecting the entire null hypothesis also in-

Table 1. Inflation of Significance Level

Number of comparisons	Significance level
1	0.05
2	0.098
3	0.143
4	0.185
5	0.226
6	0.265

Table 2. Example of One-way ANOVA

	Class A (n = 30)	Class B (n = 30)	Class C (n = 30)
156	171.2	156.6	169.3
160.4	171.3	160.1	169.4
161.7	171.5	161	169.5
163.6	171.9	161.2	170.7
163.8	172	161.4	170.7
164.8	172	162.5	172.2
165.8	172.9	162.6	172.9
165.8	173.5	162.9	173.9
166.2	173.8	163.1	173.9
168	173.9	164.4	174.1
168.1	174	165.9	174.3
168.4	175.7	166	174.9
168.7	175.8	166.3	175.4
169.4	176.7	167.3	176.7
170	187	168.9	178.7
			173.3
			183.7

Raw data of students’ heights in three different classes. Each class consists of thirty students.

creases. Assuming the significance level for a single comparison to be 0.05, the increases in the probability of rejecting the entire null hypothesis according to the number of comparisons are shown in Table 1.

ANOVA Table

Although various methods have been used to avoid the hypothesis testing error due to significance level inflation, such as adjusting the significance level by the number of comparisons, the ideal method for resolving this problem as a single statistic is the use of ANOVA. ANOVA is an acronym for analysis of variance, and as the name itself implies, it is variance analysis. Let us examine the reason why the differences in means can be explained by analyzing the variances, despite the fact that the core of the problem that we want to figure out lies with the comparisons of means.

For example, let us examine whether there are differences in the height of students according to their grades (Table 2). First, let us examine the ANOVA table (Table 3) that is commonly obtained as a product of ANOVA. In Table 3, the significance

Table 3. ANOVA Table Resulted from the Example

	Sum of squares	Freedom	Mean sum of squares	F	Significance probability
Intergroup	273.875	2	136.937	3.629	0.031
	$\sum_{i=1}^K n_i(\bar{Y}_i - \bar{Y})^2$	$(K - 1)$	$\sum_{i=1}^K n_i(\bar{Y}_i - \bar{Y})^2 / (K - 1)$		
Intragroup	3282.843	87	37.734		
	$\sum_{ij=1}^n (Y_{ij} - \bar{Y}_i)^2$	$(N - K)$	$\sum_{ij=1}^n (Y_{ij} - \bar{Y}_i)^2 / (N - K)$		
Overall	3556.718	89			

\bar{Y}_i is the mean of the group i ; n_i is the number of observations of the group i ; \bar{Y} is the overall mean; K is the number of groups; Y_{ij} is the j^{th} observational value of group i ; and N is the number of all observational values. The F statistic is the ratio of intergroup mean sum of squares to intragroup mean sum of squares.

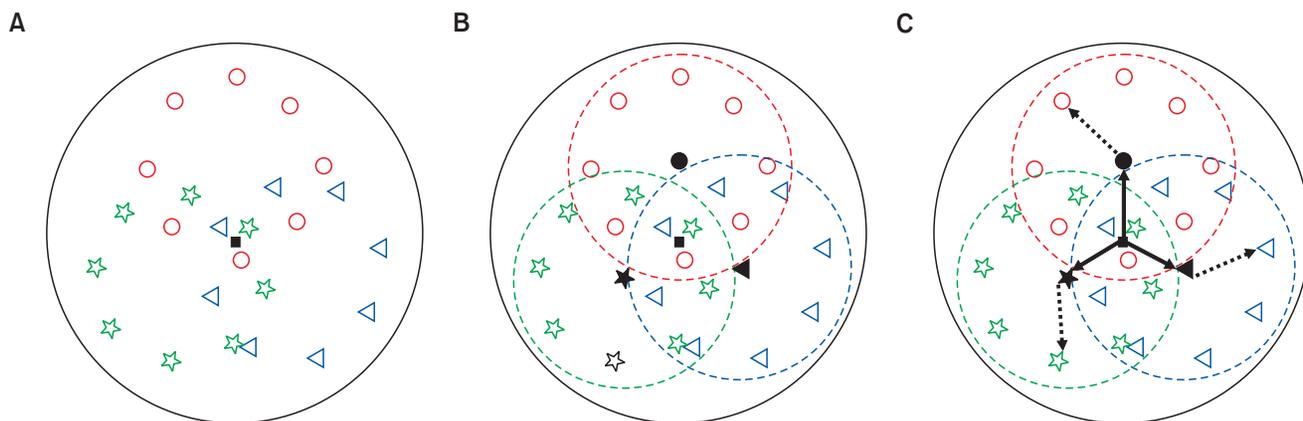


Fig. 2. Solid square is suggested as a general representative value such as mean of overall data (A). It looks reasonable to divide the data into three groups and explain the data with three different means of groups (B). To evaluate the efficiency or validity of dividing three groups, the distances from group means to overall mean and the distances from group means to each data are compared. Distance between group means and overall mean (solid arrows) stands for the inter-group variance and distance between group means and each group data (dotted arrows) stands for the intra-group variances (C).

is ultimately determined using a significance probability value (P value), and in order to obtain this value, the statistic and its position in the distribution to which it belongs, must be known. In other words, there has to be a distribution that serves as the reference and that distribution is called F distribution. This F comes from the name of the statistician *Ronald Fisher*. The ANOVA test is also referred to as the F test, and F distribution is a distribution formed by the variance ratios. Accordingly, F statistic is expressed as a variance ratio, as shown below.

$$F = \frac{\text{Intergroup variance}}{\text{Intragroup variance}} = \frac{\sum_{i=1}^K n_i(\bar{Y}_i - \bar{Y})^2 / (K - 1)}{\sum_{ij=1}^n (Y_{ij} - \bar{Y}_i)^2 / (N - K)}$$

Here, \bar{Y}_i is the mean of the group i ; n_i is the number of observations of the group i ; \bar{Y} is the overall mean; K is the number of groups; Y_{ij} is the j^{th} observational value of group i ; and N is the number of all observational values.

It is not easy to look at this complex equation and understand

ANOVA at a single glance. The meaning of this equation will be explained as an illustration for easier understanding. Statistics can be regarded as a study field that attempts to express data which are difficult to understand with an easy and simple ways so that they can be represented in a brief and simple forms. What that means is, instead of independently observing the groups of scattered points, as shown in Fig. 2A, the explanation could be given with the points lumped together as a single representative value. Values that are commonly referred to as the mean, median, and mode can be used as the representative value. Here, let us assume that the black rectangle in the middle represents the overall mean. However, a closer look shows that the points inside the circle have different shapes and the points with the same shape appear to be gathered together. Therefore, explaining all the points with just the overall mean would be inappropriate, and the points would be divided into groups in such a way that the same shapes belong to the same group. Although it is more cumbersome than explaining the entire population with just the overall mean, it is more reasonable to first form

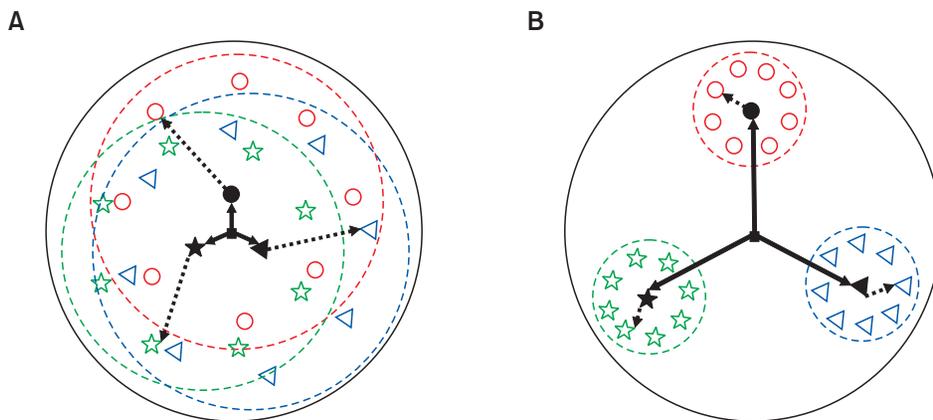


Fig. 3. Compared to the intra-group variances (dotted arrow), the inter-group variance (solid arrow) is so small that dividing three groups does not look valid (A), however, grouping looks valid when the solid arrow is much larger than the dotted arrow (B).

groups of points with the same shape and establish the mean for each group, and then explain the population with the three means. Therefore, as shown in Fig. 2B, the groups were divided into three and the mean was established in the center of each group in an effort to explain the entire population with these three points. Now the question arises as to how can one evaluate whether there are differences in explaining with the representative value of the three groups (e.g.; mean) versus explaining with lumping them together as a single overall mean.

First, let us measure the distance between the overall mean and the mean of each group, and the distance from the mean of each group to each data within that group. The distance between the overall mean and the mean of each group was expressed as a solid arrow line (Fig. 2C). This distance is expressed as $(\bar{Y}_i - \bar{Y})^2$, which appears in the denominator of the equation for calculating the F statistic. Here, the number of data in each group are multiplied, $n_i(\bar{Y}_i - \bar{Y})^2$. This is because explaining with the representative value of a single group is the same as considering that all the data in that group are accumulated at the representative value. Therefore, the amount of variance which is induced by explaining with the points divided into groups can be seen, as compared to explaining with the overall mean, and this explains inter-group variance.

Let us return to the equation for deriving the F statistic. The meaning of $(Y_{ij} - \bar{Y}_i)^2$ in the numerator is represented as an illustration in Fig. 2C, and the distance from the mean of each group to each data is shown by the dotted line arrows. In the figure, this distance represents the distance from the mean within the group to the data within that group, which explains the intragroup variance.

By looking at the equation for F statistic, it can be seen that this inter- or intragroup variance was divided into inter- and intragroup freedom. Let us assume that when all the fingers are stretched out, the mean value of the finger length is represented by the index finger. If the differences in finger lengths are compared to find the variance, then it can be seen that although

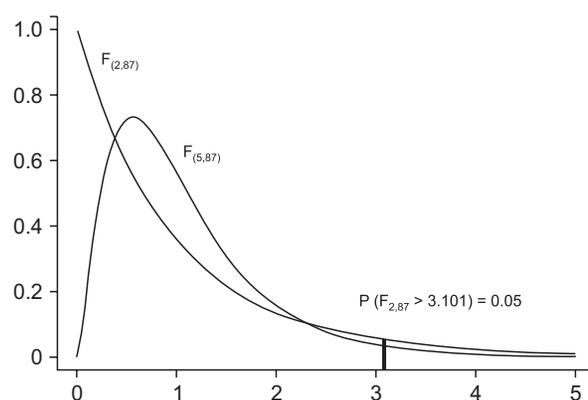


Fig. 4. F distributions and significant level. F distributions have different forms according to its degree of freedom combinations.

there are 5 fingers, the number of gaps between the fingers is 4. To derive the mean variance, the intergroup variance was divided by freedom of 2, while the intragroup variance was divided by the freedom of 87, which was the overall number obtained by subtracting 1 from each group.

What can be understood by deriving the variance can be described in this manner. In Figs. 3A and 3B, the explanations are given with two different examples. Although the data were divided into three groups, there may be cases in which the intra-group variance is too big (Fig. 3A), so it appears that nothing is gained by dividing into three groups, since the boundaries become ambiguous and the group mean is not far from the overall mean. It seems that it would have been more efficient to explain the entire population with the overall mean. Alternatively, when the intergroup variance is relatively larger than the intragroup variance, in other word, when the distance from the overall mean to the mean of each group is far (Fig. 3B), the boundaries between the groups become more clear, and explaining by dividing into three group appears more logical than lumping together as the overall mean.

Ultimately, the positions of statistic derived in this manner

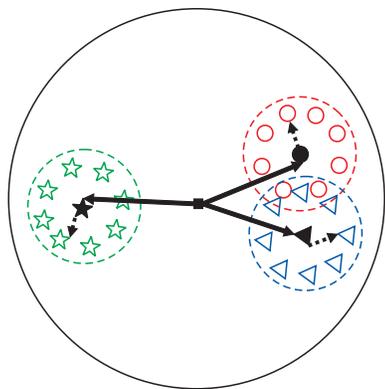


Fig. 5. It shows the schematic drawing for the necessity of post-hoc test. Post-hoc test is needed to find out which groups are different from each other.

from the inter- and intragroup variance ratios can be identified from the F distribution (Fig. 4). When the statistic 3.629 in the ANOVA table is positioned more to the right than 3.101, which is a value corresponding to the significance level of 0.05 in the F distribution with freedoms of 2 and 87, meaning bigger than 3.101, the null hypothesis can be rejected.

Post-hoc Test

Anyone who has performed ANOVA has heard of the term post-hoc test. It refers to “the analysis after the fact” and it is derived from the Latin word for “after that.” The reason for performing a post-hoc test is that the conclusions that can be derived from the ANOVA test have limitations. In other words, when the null hypothesis that says the population means of three mutually independent groups are the same is rejected, the information that can be obtained is not that the three groups are different from each other. It only provides information that the means of the three groups may differ and at least one group may show a difference. This means that it does not provide information on which group differs from which other group (Fig. 5). As a result, the comparisons are made with different pairings of groups, undergoing an additional process of verifying which group differs from which other group. This process is referred to as the post-hoc test.

The significance level is adjusted by various methods [1],

such as dividing the significance level by the number of comparisons made. Depending on the adjustment method, various post-hoc tests can be conducted. Whichever method is used, there would be no major problems, as long as that method is clearly described. One of the most well-known methods is the Bonferroni’s correction. To explain this briefly, the significance level is divided by the number of comparisons and applied to the comparisons of each group. For example, when comparing the population means of three mutually independent groups A, B, and C, if the significance level is 0.05, then the significance level used for comparisons of groups A and B, groups A and C, and groups B and C would be $0.05/3 = 0.017$. Other methods include Turkey, Schéffe, and Holm methods, all of which are applicable only when the equal variance assumption is satisfied; however, when this assumption is not satisfied, then Games Howell method can be applied. These post-hoc tests could produce different results, and therefore, it would be good to prepare at least 3 post-hoc tests prior to carrying out the actual study. Among the different types of post-hoc tests it is recommended that results which appear the most frequent should be used to interpret the differences in the population means.

Conclusions

It is believed that a wide variety of approaches and explanatory methods are available for explaining ANOVA. However, illustrations in this manuscript were presented as a tool for providing an understanding to those who are dealing with statistics for the first time. As the author who reproduced ANOVA is a non-statistician, there may be some errors in the illustrations. However, it should be sufficient for understanding ANOVA at a single glance and grasping its basic concept.

ANOVA also falls under the category of parametric analysis methods which perform the analysis after defining the distribution of the recruitment population in advance. Therefore, normality, independence, and equal variance of the samples must be satisfied for ANOVA. The processes of verification on whether the samples were extracted independently from each other, Levene’s test for determining whether homogeneity of variance was satisfied, and Shapiro-Wilk or Kolmogorov test for determining whether normality was satisfied must be conducted prior to deriving the results [2-4].

References

1. Ludbrook J. Multiple comparison procedures updated. *Clin Exp Pharmacol Physiol* 1998; 25: 1032-7.
2. Kim TK. T test as a parametric statistic. *Korean J Anesthesiol* 2015; 68: 540-6.
3. Lee Y. What repeated measures analysis of variances really tells us. *Korean J Anesthesiol* 2015; 68: 340-5.
4. Lee S. Avoiding negative reviewer comments: common statistical errors in anesthesia journals. *Korean J Anesthesiol* 2016; 69: 219-26.