Andrew Gard - equitable.equations@gmail.com

Understanding ANOVA

Analysis of variance, or ANOVA, is used to test whether a quantitative variable and a categorical variable are independent.

Analysis of variance, or ANOVA, is used to test whether a quantitative variable and a categorical variable are independent. More specifically, it is used to test the hypotheses,

$$
\begin{aligned}
H_0 &: \quad \mu_1 \;=\; \mu_2 \;=\; \cdots \;=\; \mu_m \\
H_a &: \quad \mu_i \neq \mu_j \;\; \text{for some} \;\; i \neq j.
\end{aligned}
$$

where μ_i is the mean of the quantitative variable in the i^{th} group.

Analysis of variance, or ANOVA, is used to test whether a quantitative variable and a categorical variable are independent. More specifically, it is used to test the hypotheses,

$$\begin{aligned} H_0 &: \quad \mu_1 \; = \; \mu_2 \; = \; \cdots \; = \; \mu_m \\ H_a &: \quad \mu_i \neq \mu_j \;\; \text{for some} \;\; i \neq j. \end{aligned}$$

where μ_i is the mean of the quantitative variable in the i^{th} group.

For instance, doctors may wonder whether people in different age groups respond equally to a new blood pressure medicine, or whether several different medications are all equally effective.

The fundamental idea of ANOVA is to compare the variability of the data *within* the groups to the variability *between* the groups.

The fundamental idea of ANOVA is to compare the variability of the data *within* the groups to the variability *between* the groups. As an illustration, consider the following two plots, each of which shows the distribution of a sample of a single quantitative variable across three groups.

The corresponding group means are the same in each plot.

The fundamental idea of ANOVA is to compare the variability of the data *within* the groups to the variability *between* the groups. As an illustration, consider the following two plots, each of which shows the distribution of a sample of a single quantitative variable across three groups.

The corresponding group means are the same in each plot. *The differences between group means is more significant in the first case because of the lower variability within groups.*

More specifically, ANOVA looks at the ratio of the between-groups variance and within-groups variance, both appropriately weighted.

More specifically, ANOVA looks at the ratio of the between-groups variance and within-groups variance, both appropriately weighted. More specifically, the test statistic is

$$F = \frac{SS_T/(m-1)}{SS_E/(n-m)}$$

where m is the number of groups, n is the total sample size, and

$$SS_T = \sum_i n_i(\overline{Y}_i - \overline{Y})^2 = \text{Sum of squared errors between groups}$$

$$SS_E = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2 = \text{Sum of squared errors within groups}$$

More specifically, ANOVA looks at the ratio of the between-groups variance and within-groups variance, both appropriately weighted. More specifically, the test statistic is

$$F = \frac{SS_T/(m-1)}{SS_E/(n-m)}$$

where m is the number of groups, n is the total sample size, and

$$SS_T = \sum_i n_i(\overline{Y}_i - \overline{Y})^2 = \text{Sum of squared errors between groups}$$

$$SS_E = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2 = \text{Sum of squared errors within groups}$$

If the null hypothesis is true, this ratio has a known distribution (an F-distribution) which we can use to compute a p-value and make a decision.

Analysis of variance has a few requirements.

- The observations must all be independent.
- The response variable must be normally distributed within each group.
- The variances within the groups must be equal.

Analysis of variance has a few requirements.

- The observations must all be independent.
- The response variable must be normally distributed within each group.
- The variances within the groups must be equal.

ANOVA is fairly robust when it comes to the second two assumption (particularly normality).

Analysis of variance has a few requirements.

- The observations must all be independent.
- The response variable must be normally distributed within each group.
- The variances within the groups must be equal.

ANOVA is fairly robust when it comes to the second two assumption (particularly normality). Independence is more serious as it's usually a consequence to the way the study was designed.

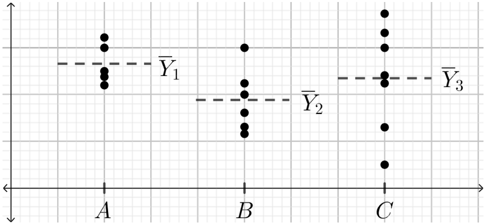
Example. Researchers measure the petal lengths (in mm) of a certain species of flower at three different latitudes. The sample data is shown below.

$$\begin{array}{lccccccc} A: & 6.4, & 6.0, & 4.8, & 5.0, & 4.4 & & \\ B: & 4.5, & 4.0, & 2.3, & 2.6, & 3.2, & 6.0 & \\ C: & 7.5, & 6.6, & 4.8, & 1.0, & 2.6, & 6.0, & 4.5 \end{array}$$

Example. Researchers measure the petal lengths (in mm) of a certain species of flower at three different latitudes. The sample data is shown below.

$$\begin{array}{llllllll} A: & 6.4, & 6.0, & 4.8, & 5.0, & 4.4 & & \\ B: & 4.5, & 4.0, & 2.3, & 2.6, & 3.2, & 6.0 & \\ C: & 7.5, & 6.6, & 4.8, & 1.0, & 2.6, & 6.0, & 4.5 \end{array}$$

A simple plot illustrates the variability within and between groups in this sample.



Analysis of variance is *always* performed using technology, for instance using the `aov` function in R.

Analysis of variance is *always* performed using technology, for instance using the `aov` function in R. Results are typically reported in a table like this one:

Source	df	SS	Mean Square	F	p-value
Treatment	2	6.83	3.42	1.18	.335
Error	15	43.55	2.90		
Total	17	50.38			

Analysis of variance is *always* performed using technology, for instance using the `aov` function in R. Results are typically reported in a table like this one:

Source	df	SS	Mean Square	F	p-value
Treatment	2	6.83	3.42	1.18	.335
Error	15	43.55	2.90		
Total	17	50.38			

In practice, you probably only care about the p-value, which represents the probability of data like that found in the sample occurring by chance when the null hypothesis is true.

Analysis of variance is *always* performed using technology, for instance using the `aov` function in R. Results are typically reported in a table like this one:

Source	df	SS	Mean Square	F	p-value
Treatment	2	6.83	3.42	1.18	.335
Error	15	43.55	2.90		
Total	17	50.38			

In practice, you probably only care about the p-value, which represents the probability of data like that found in the sample occurring by chance when the null hypothesis is true. Lower p-values indicate stronger evidence against the claim that the quantitative variable is independent from the categorical one.

Even when H_0 is rejected, the conclusion of an analysis of variation is rather weak. The alternative hypothesis

$$H_a: \quad \mu_i \neq \mu_j \quad \text{for some} \quad i \neq j$$

just says that at least two of the group means are different. It doesn't say which ones. To answer that question, a post-hoc (after the fact) test is needed.

Even when H_0 is rejected, the conclusion of an analysis of variation is rather weak. The alternative hypothesis

$$H_a: \quad \mu_i \neq \mu_j \quad \text{for some} \quad i \neq j$$

just says that at least two of the group means are different. It doesn't say which ones. To answer that question, a post-hoc (after the fact) test is needed.

The most common post-hoc test is the **Tukey honest significant differences test**. Roughly speaking, this test performs multiple t-tests between groups, limiting the total probability of a type I error by demanding smaller p-values in those pairwise tests.