Stat 139: Statistical Sleuthing through Linear Models

Section 6: ANOVA

TF: Reagan Rose (rrose@g.harvard.edu)

October 12, 2015

Basic Concepts

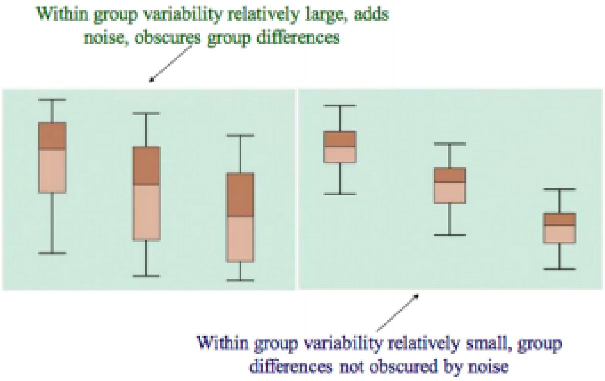
Analysis of variance (ANOVA) is used to analyze data of multiple groups. We collect data about subjects in each group, and want to know whether the groups are "significantly different".

Of course, we need a statistical measure by which to quantify "statistically different" groups, and this is where the ANOVA comes in. ANOVA essentially boils down to the following comparison: Is the variability in the dataset mainly caused by

(a) Variability within groups? (the group means are similar, but there is a lot of variance of observations around each group mean)

(b) Variability between groups? (the group means are substantially spread out around the overall mean)

Question: How can we tell whether the group means are "substantially" spread out?
Answer: When the variance of the group means (from the overall mean) is large compared to the within-group variance around each group mean.



Theory

For an ANOVA comparing k groups, our null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu.$$

More specifically, we assume that the data have been generated in the following way:

$$Y_{11}, \ldots, Y_{1n_1} \sim^{iid} N(\mu_1, \sigma^2)$$
$$\vdots$$
$$Y_{k1}, \ldots, Y_{kn_k} \sim^{iid} N(\mu_k, \sigma^2).$$

This is equivalent to saying

$$Y_{ij} = \mu_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$$

where Y_{ij} is the observation for the jth person in group i. Therefore, under the **null hypothesis**, we are assuming that **all data was generated from the same distribution**, and therefore group membership does not matter:

$$Y_{ij} \sim^{iid} N(\mu, \sigma^2),$$

Now since (in real life), we never know μ or σ, we must use our best estimates of μ and σ to test whether our null hypothesis is true. To do so, we compute the sample means $\bar{Y}_1, \ldots, \bar{Y}_k$ and sample variances $S_1^2, \ldots, S_k^2$ for each of the groups. Under H_0, all of the sample variances are measuring variability of the observations from the same mean μ, so we can pool the variances to obtain

$$S_W^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) S_i^2 = \frac{1}{n-k} SSW,$$

which is called the **within-groups mean square** or **mean square error**. Also under H_0, we can consider the sample means $\bar{Y}_i$ as k draws from a normal distribution centered at μ, so we can take the overall mean $\bar{Y}$ and consider

$$S_B^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2 = \frac{1}{k-1} SSB,$$

which is called the **between-groups mean square**.

Key point of ANOVA: If H_0 is true, then S_W^2 and S_B^2 are estimating the **same quantity**, σ^2. If H_0 is not true, then we should have greater between-groups mean square than within-groups mean square. To test this, we define

$$F = \frac{S_B^2}{S_W^2} = \frac{SSB/(k-1)}{SSW/(n-k)} \sim F_{k-1,n-k}.$$

Contrasts

Recall that our null hypothesis for an ANOVA is $H_0 : \mu_1 = \ldots = \mu_k$ and our alternative hypothesis is that **at least one μ is different**. Therefore, the ANOVA only tells if H_0 can be accepted (and so we have evidence that at least one μ is different). **It does not tell us which μ is significantly different.** To find out, we use **contrast tests**, which are closely related to the two-sample t-test.
In contrast testing, we have the hypotheses:

$$H_0 : a_1\mu_1 + a_2\mu_2 + \ldots + a_k\mu_k = 0$$
$$H_a : a_1\mu_1 + a_2\mu_2 + \ldots + a_k\mu_k \neq 0,$$

where $\sum_i a_i = 1$. Then our test statistic is

$$T_{obs} = \frac{\sum_i a_i \bar{Y}_i}{S_p\sqrt{\sum_i \frac{a_i^2}{n_i}}}.$$

You should recognize this as a standard normal divided by a chi-squared (under H_0), and in fact

$$T_{obs} \sim t_{df=n-k},$$

where n is the total sample size ($n = \sum_i n_i$) and k is the number of groups we're comparing. So we can find a (2-sided) p-value for this hypothesis by calculating

$$p - value = P(|t| > T_{obs}) = P(t < -|T_{obs}|) + (1 - P(t < |T_{obs}|)).$$

Formula Cheat Sheet

ANOVA is (in my opinion) where the number of formulas and the specific notation gets to be a bit too much to handle. Here's some formuals you will need to have on hand when using ANOVA. Here I'm assuming we're doing ANOVA on k groups, each with n_i observations, $i = 1, \ldots, k$ for a total of $n = \sum_i n_i$ observations.

$$SSB = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$

$$SSW = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

$$T = \frac{\sum_i a_i \bar{Y}_i}{S_p\sqrt{\sum_i \frac{a_i^2}{n_i}}} \quad \text{(for contrasts)}$$

Exercises

Question 1. ANOVA by hand

The data below show a summary of highway gas mileage for three types of vehicles: midsize cars, SUV's and pickup trucks.

Group	n	Mean	Std. Dev
Midsize	31	25.8	2.56
SUV	31	22.68	3.67
Pickup	14	21.29	2.76
Overall	76	69.77	8.99

(a) Use ANOVA to test whether there are any significant differences in gas mileage between the groups. Comment on whether the assumptions of the ANOVA were met.

(b) Does driving a larger car (SUV or pickup) decrease mileage? Use a contrast test to investigate.

(c) Now compare SUV versus pickups on gas mileage using contrasts. Is the difference significant?

(d) Pretend instead that we had not observed the Midsize data at all, and were interested in calculating the differences in means between SUVs and pickups. In this case, the appopriate test would be an independent samples t-test. Run this test and compare your results to those in part (c). Which result do you trust more?

Question 2. ANOVA in R

Download the dataset INSURANCE_CLAIMS.CSV from Canvas and import it into R. This is a typical dataset you might see working as a data analyst in industry. We are interested in preventing fraudulent insurance claims, so we collect some data and examine it to see if there are any notable variables associated with fraudulent claims.

(a) Run some exploratory analyses. Do your data meet the assumptions for ANOVA? If not, transform the data to alleviate the problem.

(b) Run a t-test for for the differences in mean claim amount based on fraudulence, then run an ANOVA on the same. Prove to yourself that the t-test and ANOVA are equivalent when testing two groups.

(c) Conduct a one-way ANOVA to evaluate whether claim amount is associated with claim type. Comment on your results.

(d) Conduct a two-way ANOVA to evaluate whether claim amount is associated with claim type and fraudulence. Comment on your results.