

Probability and Statistics for Final Year Engineering Students

By Yoni Nazarathy, Last Updated: May 1, 2011.

Lecture 5s:

Comparison of Means: The Analysis of Variance: ANOVA

The Analysis of Variance (ANOVA) is one of the most widely used basic statistical techniques in experimental design and data analysis. In contrast to the name, ANOVA is **NOT** designed to study the variance of populations but rather **uses the variance** to perform statistical inference (hypothesis testing) about the **mean: ANOVA is used to compare the mean of several populations.**

To illustrate the basic set up, consider the following **example**: Assume you are comparing the behavior of 3 different pick-up algorithms for the robotic-arm: Alg 1, Alg 2 and Alg 3. You are applying the 3 algorithms on the beta-tested robotic arm for 24 hours, letting each of the algorithms run for 8 hours and obtain the following data:

$$x_{11}, \dots, x_{1n_1}$$

$$x_{21}, \dots, x_{2n_2}$$

$$x_{31}, \dots, x_{3n_3}$$

With $n_1 = n_2 = n_3 = 8$ (in our example). So we have a total of 24 data points.

Each x_{ij} is the number of items successfully picked up using Alg i during hour j.

The basic thing to do now is to calculate the three sample means:

$$\bar{x}_i = \frac{\sum_{j=1}^n x_{ij}}{n_i}$$

and compare them. Say for illustration that we got $\bar{x}_1 = 141.75$, $\bar{x}_2 = 152.25$, $\bar{x}_3 = 157.625$.

Is Alg 3 really the best algorithm? Or is this just statistical error?

ANOVA Allows us to statistically answer this question by formulating the following Hypothesis test:

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

H_1 : Otherwise (i.e. one of the means is different)

We first specify the procedure used for carrying out this Hypothesis test and then briefly discuss where it comes from.

Steps in carrying the test:

1. **Collect the data** (make sure it is randomized! – In our example this may mean randomly assigning algorithms to hours in the 24 hour schedule, as opposed to doing 8 hours Alg 1 then 8 hours Alg 2 and then 8 hours Alg 3).
2. **Calculate the test statistic (F):**

$$F = \frac{MS_{Treatments}}{MS_E}.$$

We now specify the numerator and denominator:

MS_E is the **pooled sample variance** (which we now define briefly):

In situations where we have samples taken from k different populations (e.g. in our k=3 Algorithm case) yet **we assume** the variance of each of the populations is the same, while the means may (or may not) differ, the following is a good (unbiased) estimator of the variance:

$$S_p^2 = \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{\sum_{i=1}^k (n_i - 1)}.$$
 Here S_i^2 is the sample variance of population i.

This is sometimes called the **pooled variance** (and is a good thing to know w/out relation to ANOVA). In the context of ANOVA we take $MS_E = S_p^2$.

$MS_{Treatments}$ captures the variability between populations:

$$MS_{Treatments} = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{k-1}.$$

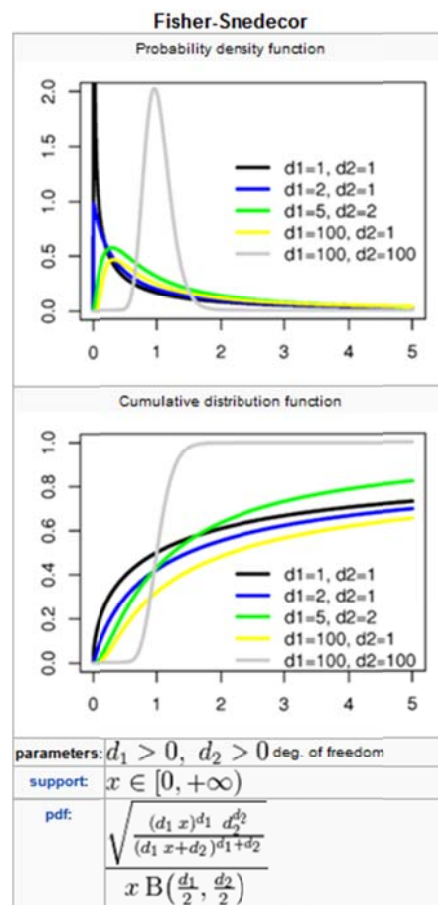
Here \bar{x} is the sample mean of all observations (a total of $n_1 + \dots + n_k$). As opposed to that, \bar{x}_i is the sample mean of only the i'th population.

3. **Find the critical value.** Under H_0 (and additional model assumptions specified below) F is distributed according to a Fisher-Snedecor F-distribution (F distribution for short) with parameters (**degrees of freedom** of numerator and denominator): k-1 and $\sum_{i=1}^k n_i - k$.

We can summarize this as:

$$F_{statistic} = \frac{MS_{Treatments} = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{k-1}}{MS_E = \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{\sum_{i=1}^k (n_i - 1)}} \sim_{H_0} \sim Fdistribution(k-1, \sum_{i=1}^k n_i - k)$$

We don't give go into the details of the F distribution, but here are some basics taken from Wikipedia:



The B() in the denominator is the beta-function which is defined using an integral (that typically has to be computed numerically).

Under H_0 , it is “not too likely” to get a “high” value for the statistic F. The F distribution quantifies this. We thus choose our critical value to be the $1 - \alpha'$ th percentile of the F distribution (denote this as $F_{1-\alpha}$).

The decision rule is: If $F > F_{1-\alpha}$, reject H_0 . Otherwise do not reject H_0 .

Worked out example:

The following can be used to simulate our example (in Mathematica):

```
alg1=Table[Round[Random[NormalDistribution[150,10]]],{8}];
alg2=Table[Round[Random[NormalDistribution[150,10]]],{8}];
alg3=Table[Round[Random[NormalDistribution[150,10]]],{8}];
x1=Mean[alg1];
x2=Mean[alg2];
x3=Mean[alg3];
totalMean=Mean[{x1,x2,x3}];
MStreat=(8(x1-totalMean)2+8(x2-totalMean)2+8(x3-totalMean)2)/(3-1);
MSE=(7 Variance[alg1] + 7 Variance[alg2]+7Variance[alg3])/(8+8+8-3);
F=MStreat/MSE
```

For example, assume the data is:

Alg1 = {167,99,152,178,140,133,142,123}

Alg2 = {106,202,173,158,155,129,147,148}

Alg3 = {135,161,179,176,140,192,136,142}

Thus, $\bar{x}_1 = 141.75$, $\bar{x}_2 = 152.25$, $\bar{x}_3 = 157.625$ (as in the start of the lecture above).

The total mean is $\bar{x} = 150.542$ and the F statistic (after some calculations is) 0.5169.

The F-distribution which we use has degrees of freedom 2 and 21. And if we choose $\alpha = 0.05$

Then the critical value is $F_{0.95}^{(2,21)} = 3.47$. Since $0.5169 < 3.47$ we **do not reject H0**.

Model Assumptions and the Underlying idea of ANOVA:

The underlying idea of ANOVA is decomposition of the **sum of squares** SS_{total} into $SS_{treatments}$ and SS_{error} :

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \text{after some algebra} = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2.$$

Under the assumption of H_0 (all means are equal) the variability in both SS_{total} and $SS_{treatments}$ is due to random error. If we assume:

1. That the probability distribution of each of the populations are normal.
2. That the population variances are equal.

This in turn implies that the F statistic is distributed according to the F distribution (under H_0).

If on the other hand H_1 is the case, and not all means are equal, then we have that the variability in SS_{error} is still due to random error, but the variability in $SS_{\text{treatments}}$ is also due to the different population means. In this case, it is most likely to get high values of the F statistic (and if high enough we reject H_0).

The above procedure is sometimes called **one-way** analysis of variance. We are checking the effect of different levels of **one treatment** on a population. In our example the **treatment** was the algorithm type. There is also **two-way** ANOVA and more complicated procedures where we are looking at the effects of several categories of treatments in parallel.