Statistics 354 (Fall 2018)
Analysis of Variance: Comparing Several Means

Remark. These notes are from an elementary statistics class and introduce the Analysis of Variance technique for comparing several population means. They are meant as both a review and a reminder of how ANOVA works. In STAT 354, we will use an Analysis of Variance technique to assess the strength of the linear regression relationship which has the advantage that it can be extended quite easily from simple linear regression to multiple linear regression.

Recall that the two sample t-test allows us to compare the means of two independent populations based on data from two independent samples.

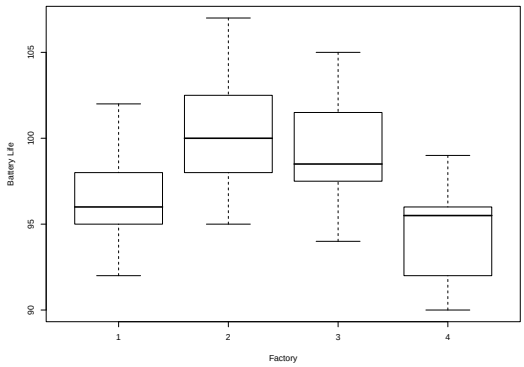
Suppose that we are interested in comparing the means from $I \geq 2$ independent populations based on I independent simple random samples from those populations, i.e., we want to compare the means of more than two populations.

The technique is called *analysis of variance*, or more compactly, *ANOVA*.

Example. A particular brand of battery is manufactured in one of four factories. Ideally, the mean life of the batteries should be the same from each factory. Battery life (in weeks) of random samples from each factory are as follows.

	Factory 1	Factory 2	Factory 3	Factory 4
	102	107	98	99
	96	102	98	97
	92	103	102	91
	95	95	105	96
	94	97	94	95
	101	100	97	90
	97	99	99	95
	96		101	96
	95			92
	98			96
mean	96.60	100.43	99.25	94.70
SD	3.06	3.99	3.37	2.83
sample size	10	7	8	10

Do the data indicate that the mean battery life is different between the factories? Boxplots show that there is variability within each factory, but there is also variability between factory means.



Our goal is to test the global hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

We will accomplish this using analysis of variance.

It is important to note that the probability of a Type I error (rejecting H_0 when H_0 is true) for performing all pairs of comparisons is much greater than for a single comparison. For instance, in this example where $I = 4$, there are 6 possibilities, namely

$$\begin{array}{ll} \text{(i) } H_0 : \mu_1 = \mu_2, & \text{(ii) } H_0 : \mu_1 = \mu_3, \\ \text{(iii) } H_0 : \mu_1 = \mu_4, & \text{(iv) } H_0 : \mu_2 = \mu_3, \\ \text{(v) } H_0 : \mu_2 = \mu_4, & \text{(vi) } H_0 : \mu_3 = \mu_4. \end{array}$$

Therefore, we have 6 opportunities to reject

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Suppose that the significance level for each of these tests is 0.05. The overall significance level α (i.e., the probability of rejecting at least once) is

$$\alpha = P(\text{rejecting at least once}) = 1 - P(\text{rejecting none}) = 1 - (0.95)^6 = 0.2649.$$

Thus, we need to develop a more useful way of performing *multiple comparisons*. In general, statistical methods for dealing with multiple comparisons usually have two steps.

(1) An *overall test* to see if there is good evidence of *any* differences among the parameters we want to compare.

(2) A detailed *follow-up analysis* to decide which of the parameters differ and to estimate how large the differences are.

ANOVA is designed for the first step. The t-test can then be used as *part* of the second step.

The ANOVA F-test

Suppose that we have an independent simple random sample from each of I populations and that the jth population has a normal $N(\mu_j, \sigma^2)$ distribution, where σ is the common standard deviation in all the populations. (It is irrelevant whether or not σ is known. We do, however, require that σ be the same for each population.)

Assume that the jth sample has size n_j, sample mean $\overline{x}_j$, and sample standard deviation s_j. In general, we can list our notation as follows.

Population	1	2	$\cdots$	I
Sample	x_{11}	x_{21}		x_{I1}
	x_{12}	x_{22}		x_{I2}
	$\vdots$	$\vdots$		$\vdots$
	x_{1n_1}	x_{2n_2}		x_{In_I}
Mean	$\overline{x}_1$	$\overline{x}_2$		$\overline{x}_I$
Standard Deviation	s_1	s_2		s_I

Our goal is to make comparisons between $\mu_1, \mu_2, \ldots, \mu_I$ based on the observed data.

Remark. If all the sample sizes are the same, we sometimes say that the design is *balanced*.

Let $N = n_1 + n_2 + \cdots + n_I$ denote the total number of observations, and let

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{I} \sum_{k=1}^{n_j} x_{jk} = \frac{1}{N} \sum_{j=1}^{I} n_j \overline{x}_j$$

denote the average of all observations taken together. (We sometimes call $\overline{x}$ the *grand mean*.)
In order to test the hypothesis that all I populations have the same mean,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{against} \quad H_a : \text{not all of the } \mu_j \text{ are equal,}$$

we use the ANOVA F statistic given by

$$F = \frac{MSG}{MSE}$$

where the *mean square for groups* is

$$MSG = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_I(\overline{x}_I - \overline{x})^2}{I - 1}$$

and the *mean square for error* is

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{N - I}.$$

When H_0 is true, the F statistic has the F distribution with $I - 1$ numerator degrees of freedom and $N - I$ denominator degrees of freedom.

That is, the degrees of freedom of the F statistic are a pair $(I - 1, N - I)$ which are the same as the denominators of MSG and MSE, respectively.

We usually call the numerators of MSG and MSE the *sums of squares* so that the *sum of squares among groups* (or *sum of squares between groups*) is

$$SS(\text{among}) = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_I(\overline{x}_I - \overline{x})^2$$

and the *sum of squares within groups* is

$$SS(\text{within}) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2.$$

An *ANOVA table* lists all of this information. Software packages will always give you an ANOVA table containing the following information.

Source	*df*	*SS*	*MS*
variation among groups	$I - 1$	$SS(\text{among})$	$MSG = SS/df$
variation within groups	$N - I$	$SS(\text{within})$	$MSE = SS/df$

Example (continued). The ANOVA table for the battery factories is as follows.

Source	*df*	*SS*	*MS*
variation among groups	$4-1=3$	170.9	$MSG = 57.0$
variation within groups	$35-4=31$	331.4	$MSE = 10.7$

Thus, the F test statistic is

$$F = \frac{MSG}{MSE} = \frac{57.0}{10.7} = 5.33$$

and the degress of freedom are $df = (3, 31)$. From a table, we find that the critical values corresponding to $df = (3, 25)$ are

$$4.68 < 5.33 < 7.45$$

so that the P-value satisfies

$$0.001 < P\text{-value} < 0.01.$$

Thus, we can reject H_0 at the 1% level and conclude that this is strong evidence that the mean battery life differs between the four factories.

Conditions for validity of ANOVA

The ANOVA procedure is valid if the following conditions are satisfied.

(i) The groups of observations can be regarded as random samples from their respective populations. That is, we have I independent SRSs, one from each of the I populations. In particular, observations within each group are independent, and observations across groups are independent.

(ii) The jth population has a normal $N(\mu_j, \sigma^2)$ distribution with unknown mean μ_j.

(iii) All I populations have the same standard deviation σ.

Like the t-test, the ANOVA procedure is robust meaning that departures from normality are permitted as long as the sample sizes are sufficiently large. In fact, even for roughly symmetric distributions with no outliers, the ANOVA procedure can be used for a sample of size 4.

Furthermore, the results of the ANOVA F test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample size.

Intuition for ANOVA

- Variability in the observed data results from variation in the population means (if they are truly different) and from variation within each population itself.

- Differences between sample means estimate differences in the populations means, but the observed differences may simply reflect the error in estimating $\mu_1, \ldots, \mu_I$ due to sampling variation within a population.
- Conclude the population means are not all equal if the differences in $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_I$ are sufficiently large that it is unlikely that the difference is due to estimation error due to within population variation.

Additional remarks

- Rejection of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ does not allow one to determine which means are different. For that, a more detailed analysis is needed (such as the Newman-Keuls procedure).
- ANOVA can be used for $I = 2$ populations in which case it gives exactly the same results as the two-sample t test.
- Failure to reject H_0 does not mean that H_0 is true. It simply means that the differences in the μ_j may be too small to detect given the available data.