Analysis of variance (ANOVA)

Definition: ANOVA is a statistical procedure for determining whether or not the means of more than two populations are equal (hypothesis testing again ☺).

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots\mu_i = \ldots\mu_{\mathrm{m}} \;\; = \mu$$

$$H_1 \; : \mu_i \neq \mu$$

Note: In analysis of variance, it must be assumed that:

- **all populations being studied have the same variance**, regardless of whether or not their means are equal
- the response variable (quantitative statistical attribute) **is normally distributed** for each population
- the observations **must be independent**

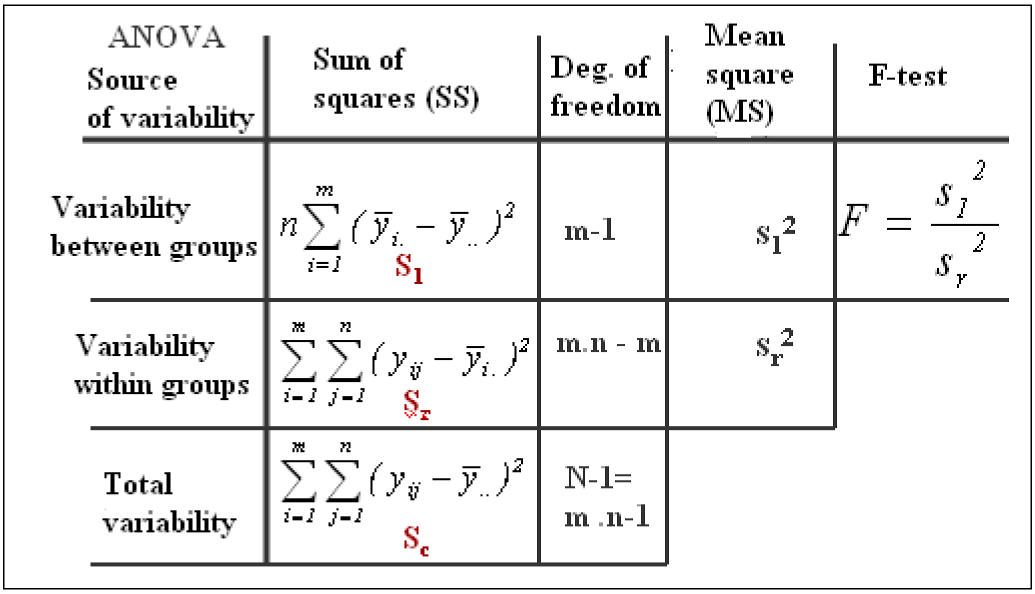
The main idea of analysis of variance is the decomposition of total variability (Sc) into:

- the **variability between groups** and (S_1)
- the **variability within groups (residual variability) –** Sr

As for ANOVA, we analyze the influence of some factors (usually qualitative) on the observed (quantitative) statistical attribute. On the basis of this assumption, analysis of variance can take on these forms:

- one-way ANOVA (single-factor ANOVA) – influence of one single factor
- two-way ANOVA (two-way ANOVA with or without replication) – influence of two factors

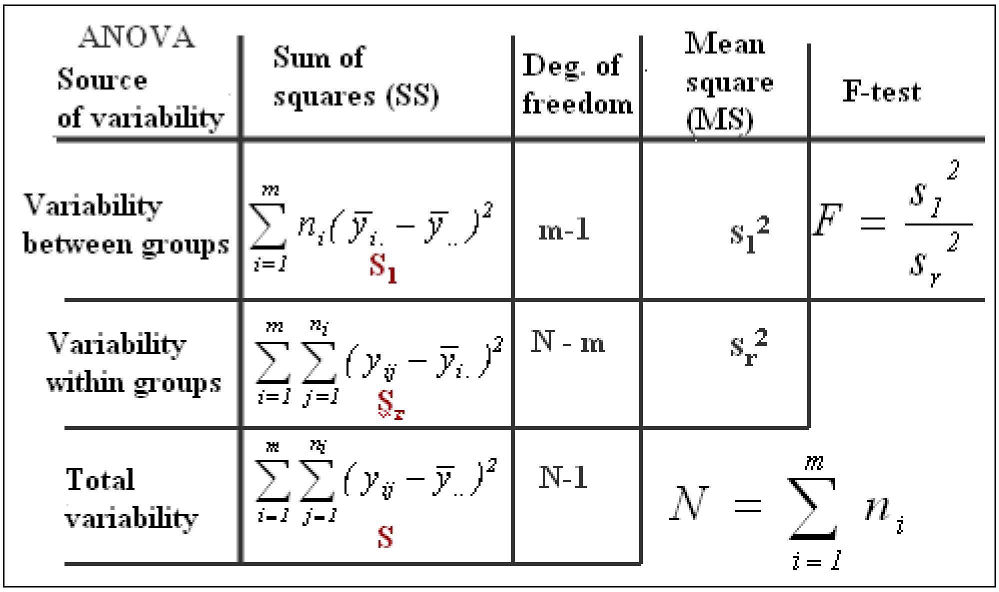
ANOVA-single factor (number of observations is equal)



ANOVA Source of variability	Sum of squares (SS)	Deg. of freedom	Mean square (MS)	F-test
Variability between groups	$n\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2$ S_1	m-1	s_1^2	$F = \dfrac{s_1^2}{s_r^2}$
Variability within groups	$\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$ S_r	m.n - m	s_r^2	
Total variability	$\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$ S_c	N-1= m .n-1		

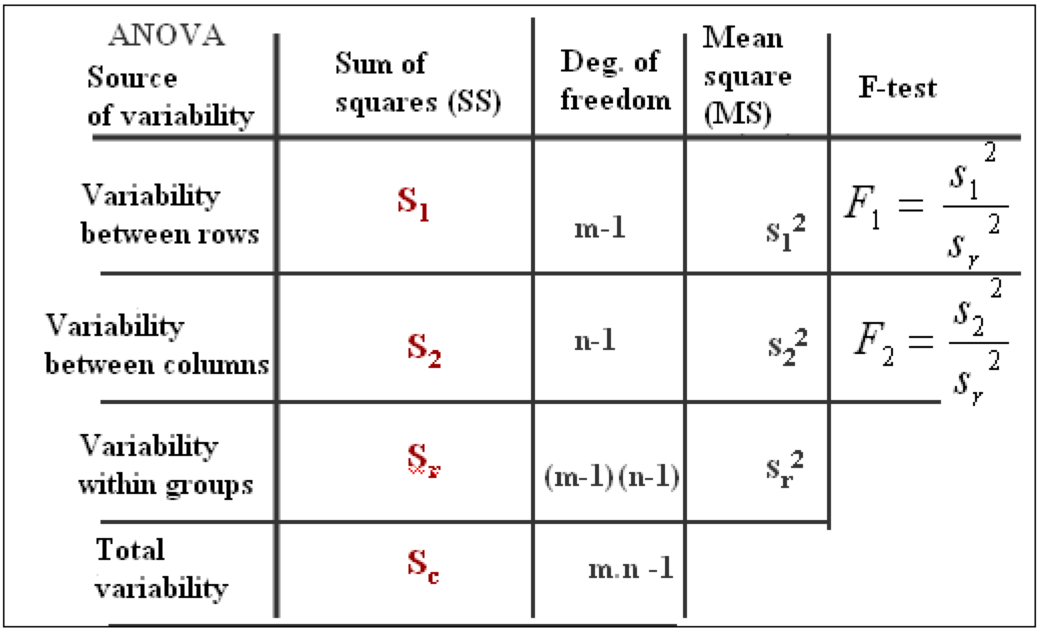
In this case, we use the Fisher F distribution, i.e. to compute the critical value of F-test, we shall use the function FINV (alpha; (m-1), (N-m)) where m is the number of rows, n is the number of columns and N is the product of m and n.

ANOVA-single factor (number of observations is not equal)



ANOVA Source of variability	Sum of squares (SS)	Deg. of freedom	Mean square (MS)	F-test
Variability between groups	$\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$ S_1	m-1	s_1^2	$F=\frac{s_1^2}{s_r^2}$
Variability within groups	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$ S_r	N - m	s_r^2	
Total variability	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$ S	N-1	$N=\sum_{i=1}^{m} n_i$	

Two-way ANOVA



ANOVA Source of variability	Sum of squares (SS)	Deg. of freedom	Mean square (MS)	F-test
Variability between rows	S_1	m-1	s_1^2	$F_1=\frac{s_1^2}{s_r^2}$
Variability between columns	S_2	n-1	s_2^2	$F_2=\frac{s_2^2}{s_r^2}$
Variability within groups	S_r	(m-1)(n-1)	s_r^2	
Total variability	S_c	m.n -1		

$$S_1 = n\sum_{i=1}^{m}(\bar{y}_{i}.-\bar{y}..)^2$$

variability between rows

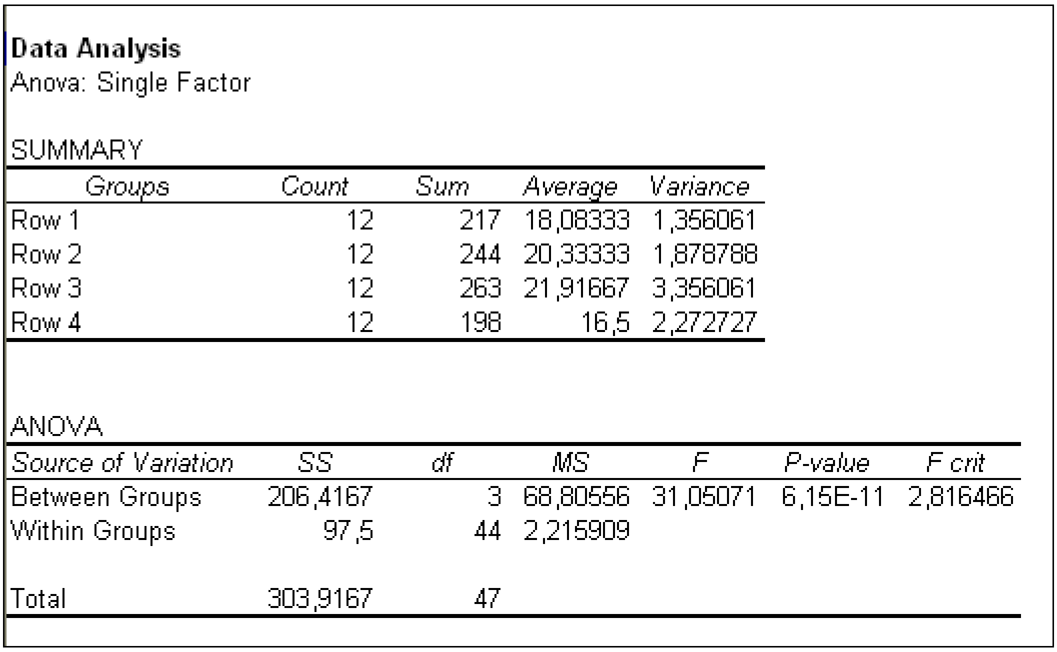
$$S_2 = m\sum_{j=1}^{n}(\bar{y}._{j}-\bar{y}..)^2$$

variability between columns

$$S_r = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i}. - \bar{y}._{j} + \bar{y}..)^2$$ variability within groups

$$S_c = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}..)^2$$ total variability

Using Data Analysis in MS Excel for the calculation procedure of ANOVA



Data Analysis
Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Row 1	12	217	18,08333	1,356061
Row 2	12	244	20,33333	1,878788
Row 3	12	263	21,91667	3,356061
Row 4	12	198	16,5	2,272727

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	206,4167	3	68,80556	31,05071	6,15E-11	2,816466
Within Groups	97,5	44	2,215909			
Total	303,9167	47				

Explanation:
In the first part of the table, simple descriptive characteristics are present (count, sum, average and variance).

The second part of the table is the necessary output for ANOVA:

- Source of variation – Between Groups, Within Groups, Total
- Sum of squares (SS)
- Degrees of freedom (df) – Total (n-1); Within Groups (n-k); Between Groups (k-1) where n is the number of observations and k is the number of groups
- Mean square (MS) – SS/df
- Test statistics for the F-test (F) – MS of Between Groups / MS of Within Groups
- Probability (p-value)
- Critical value for the F-test (F crit)

There are two ways how to determine if we accept or reject the null hypothesis (H0):

- by comparing the test statistics and the critical value (in this case test statistics – 31,05 is greater than the critical value – 2,81, i.e. we reject the null hypothesis (H0) and accept the alternate hypothesis (Ha, H1))
- by comparing the p-value with the level of significance (alpha) – in this case p-value (6,15E-11) is smaller than the level of significance (0,05 or 0,01 as default), i.e. we reject the null hypothesis (H0) and accept the alternate hypothesis (Ha, H1)

As you can see, the final result (i.e. whether we accept or reject the null hypothesis) should always be the same regardless the way you use.

Note: if we reject the null hypothesis and accept the alternate hypothesis, pairwise multiple comparison tests could follow to see between which pairs of countries were detected significant differences in the mean test scores (this is only an example). MS Excel cannot calculate these pairwise multiple comparison tests, you're advised to use another professional statistical software packages, i.e. SAS, Statistica etc.

The most frequently used pairwise multiple comparison tests are as follows:

- Scheffe‘s test
- Tukey‘s HSD test
- Fisher‘s LSD test (LSD)
- SNK test (Studentov-Newmanov-Keulsov test)
- Duncan’s test
- Bonferroni‘s test

Summary: the logic behind ANOVA is based on the development of two independent estimates of the common population varianceσ^2. One estimate of σ^2 is based on the variability among the sample means themselves and the other estimate of σ^2 is based on the variability of the data within each sample. By comparing these two estimates ofσ^2, we will be able to determine if the population means are equal. **Since the methodology uses a comparison of variances, it is referred to as analysis of variance.**