MS-C1620 Statistical inference

11 Analysis of variance

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2020–2021
Period III–IV

Contents

1. Analysis of Variance

2. Kruskal-Wallis test

ANOVA

Analysis of variance (ANOVA) is the generalization of the two sample *t*-test for more than two populations.

In analysis of variance the population consists of two or more independent groups. Observations are assumed to follow normal distribution. Each group is independently sampled.

ANOVA tests the equality of the expected values of the groups.

- Is there a difference in mean salary in the 10 largest cities in Finland?
- Is there a difference in the average lengths of products made in different production lines?

ANOVA

ANOVA, assumptions

- Let $x_{1j}, x_{2j}, \ldots, x_{n_j j}$ be i.i.d. observed values of a $\mathcal{N}(\mu_j, \sigma^2)$-distributed random variable x_j, $j = 1, \ldots, k$. Assume that the k samples are independent.

We thus have k independent random samples from univariate normal distributions. The variances of all k distributions are assumed to be the same.

ANOVA, hypotheses

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

$H_1 : \mu_h \neq \mu_j$ for some groups $h \neq j$.

ANOVA, the basic idea

In analysis of variance, the total variance is divided into two parts. The first part measures the variation between the group means and the second part measures variation within the groups.

If the first part is much larger than the second part, there is evidence against the null hypothesis and we reject it. Otherwise, it is plausible that the group means are equal.

Hence, the test of the equality of the expected values is based on the comparison of between-groups variance and within-groups variance (giving the name of the method).

ANOVA, components

To conduct ANOVA we calculate,

1. the group means

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij},$$

   where n_j is the group size of the jth group,

2. the combined sample mean

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij},$$

   where $n = \sum_{j=1}^{k} n_j$,

ANOVA, components

3. the variance between groups (group sum of squares)

$$SSG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \quad \text{and}$$

4. the variance within groups (error sum of squares)

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2.$$

ANOVA

ANOVA, test statistic

- The F-test statistic,

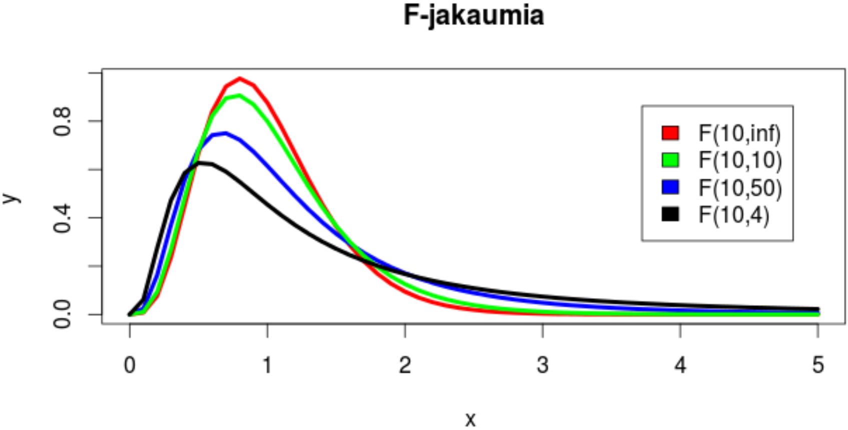
$$F = \frac{SSG \,/\, (k-1)}{SSE \,/\, (n-k)},$$

  follows under H_0 the $F-$distribution with the parameters $(k-1)$ and $(n-k)$.
- The expected value of the test statistic under H_0 is $\frac{n-k}{n-k-2}$ and **large** values of the test statistic give evidence against the null hypothesis.

F-distribution

F-distribution is a family of distributions indexed by two parameters.

It is rarely encountered outside of theoretical results.



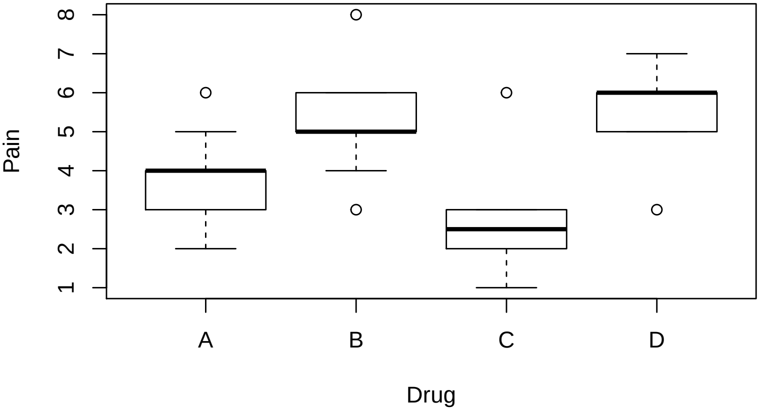
Example

- A drug company wants to test the effectiveness of 4 drugs in relieving headache. The company recruited 40 volunteers and randomized them to take one of the four drugs during their next headache. The subjects reported their pain after one hour of taking the drug on a scale 1-10 (1 = no pain).

A	4	3	2	4	6	5	4	3	4	4
B	6	5	4	5	8	5	3	5	5	6
C	2	3	1	3	2	6	3	3	2	2
D	6	7	6	5	5	5	6	3	6	7

- We are interested in studying whether the drugs differ in their effectiveness on the significance level $\alpha = 0.05$.

Example



Kuva: Boxplots of pain by drug.

- The group means are,

$$\bar{x}_A = 3.9 \quad \bar{x}_B = 5.2 \quad \bar{x}_C = 2.7 \quad \bar{x}_D = 5.6,$$

  and the combined sample mean is $\bar{x} = 4.35$.
- The variance between groups is,

$$SSG = 52.1,$$

  and the variance within groups is,

$$SSE = 55.$$

- This gives the F-statistic value,

$$F = \frac{(40-4)52.1}{(4-1)55} = 11.367,$$

  and the p-value of 0.0000216, indicating that at least one of the treatments differs in effectiveness from the others.

Multiple testing problem

If the null hypothesis is rejected based on the F-test, then we know that at least two of the groups differ (but we do not know which ones!).

The next step in the analysis is usually to find out the groups with statistically significant differences in expected values.

A simple way to do this is to analyze the groups in pairs of two with t-test.

There are $C = \frac{k(k-1)}{2}$ pairs in total to compare and conducting all possible comparisons has the side effect that the probability of type I error is inflated greatly above its set level.

This is called the *multiple testing problem*.

Bonferroni correction

Let β be the significance level of the C pair-wise comparisons, i.e, the (upper bound for the) probability that H_0 is incorrectly rejected in a single comparison, i.e., the probability of type I error in a single comparison.

Let γ be the probability that H_0 is incorrectly rejected in at least one test, when the test is repeated C times, i.e., the probability of making at least type I error during the C tests.

Probability theory shows that,

- if the tests are independent (which they most likely are not), then $\gamma = 1 - (1 - \beta)^C$.
- in the general case, we have the bound $\gamma \leq C\beta$.

Thus, to be absolutely sure that the probability of making at least one type I error during the C tests is at most some α, the individual comparisons must be done on significance level $\beta = \frac{\alpha}{C}$.

Bonferroni correction

Bonferroni correction

That is, for each pair μ_j, μ_k we conduct a *t*-test to test for their equality and reject the null hypothesis $H_0 : \mu_j = \mu_k$ if the corresponding *p*-value satisfies

- $p < \frac{\alpha}{C}$, or equivalently
- $pC < \alpha$ (each *p*-value is magnified *C*-fold).

This is known as the *Bonferroni correction.*

Example, continued

- The drug company wants to know exactly which of the drugs differ from each other in effectiveness.
- To keep the 5 % Type I error rate, the pairwise comparisons are done on a significance level $\beta = \alpha / C = 0.05/6 = 0.0083$.

Pair	H_0
AB	Not rejected
AC	Not rejected
AD	Rejected
BC	Rejected
BD	Not rejected
CD	Rejected

- Drug D is statistically significantly different from drugs A and C and furthermore drug B is statistically significantly different from drug C.

Assumptions of ANOVA

ANOVA makes two key assumptions:

1. The groups are normally distributed.
2. The groups have equal variances.

As usual, the first of the assumptions can (by central limit theorem) be replaced with a large enough sample size n.

The second one is required also for large samples. However, ANOVA is robust to moderate violations from it. As a rule of thumb, the largest group variance should be at most 4 times the smallest group variance.

The variance assumption can also be tested using *Bartlett's test.*

Bartlett's test for equality of variances

Bartlett's test, assumptions

Let $x_{1j}, x_{2j}, \ldots, x_{n_j j}$ be i.i.d. observed values of a $\mathcal{N}(\mu_j, \sigma_j^2)$-distributed random variable x_j, $j = 1, \ldots, k$. Assume that the k samples are independent.

Bartlett's test, hypotheses

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

$H_1 : \sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

Bartlett's test for equality of variances

To conduct Bartlett's test for equality of variances we calculate,

1. The individual variance estimates

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2,$$

2. The pooled variance estimate

$$s^2 = \frac{1}{n - k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2,$$

3. and the components of the test statistic,

$$Q = (n-k) \ln s^2 - \sum_{j=1}^{k} (n_j - 1) \ln s_j^2$$

$$h = 1 + \frac{1}{3(k-1)} \left(\left(\sum_{j=1}^{k} \frac{1}{n_j - 1} \right) - \frac{1}{n-k} \right).$$

Bartlett's test for equality of variances

Bartlett's test, test statistic

- Bartlett's test statistic,

$$B = \frac{Q}{h},$$

  follows, for large n, under H_0 approximately the χ^2-distribution with $k-1$ degrees of freedom.
- The expected value of the test statistic under H_0 is approximately $k-1$ and **large** values of the test statistic suggest that the null hypothesis H_0 is false.

ANOVA and linear regression

ANOVA is closely related to multiple linear regression.

In fact, it is equivalent to regressing the x-variable on the indicator variables of the groups,

$$x_{ij} = \beta_0 + \beta_1 \mathtt{I}(j = 1) + \beta_2 \mathtt{I}(j = 2) + \cdots + \beta_{k-1} \mathtt{I}(j = k - 1) + \epsilon_{ij},$$

where e.g. $\mathtt{I}(j = 1) = 1$ if $j = 1$ and $\mathtt{I}(j = 1) = 0$ otherwise, and the independent errors $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

The above model includes only $k - 1$ indicators as adding the kth would make the model parameters *unidentifiable* (no unique solution).

ANOVA and linear regression

Parameter interpretation:

- β_0 is the expected value μ_k of the kth group (*the reference group*).
- β_ℓ, $\ell = 1, \ldots, k-1$ are the differences $\mu_\ell - \mu_k$ between the expected values of the other groups and the group k.

Statistical software usually sets the first or last group as the reference group.

The ANOVA test on slide 8 is equivalent to testing the null hypothesis $\beta_1 = \ldots = \beta_{k-1} = 0$ in the regression formulation for ANOVA.

Similar tests for testing simultaneously whether multiple regression coefficient are zero exist also for standard linear regression. However, they are out of our scope.

Contents

1 Analysis of Variance

2 Kruskal-Wallis test

Kruskal-Wallis test

Kruskal-Wallis test is a non-parametric alternative to the analysis of variance. That is, it avoids the need for the normality assumption.

It is a generalization of the two-sample rank test/Wilcoxon rank sum test to more than two groups.

Kruskal-Wallis test tests the null hypothesis that k independent samples all come from the same distribution.

Kruskal-Wallis test

Kruskal-Wallis test, assumptions

- Let $x_{1j}, x_{2j}, \ldots, x_{n_j j}$ be an i.i.d random sample from the distrbution F_j, $j = 1, \ldots, k$, and let the k samples be independent.
- Assume further that the groups distributions $F_1, \ldots, F_k$ are equal up to location shifts (i.e. the distribution have the same "shape" but possibly different medians/locations) and denote the distributions' medians by m_j, $j = 1, \ldots, k$.

Kruskal-Wallis test, hypotheses

$H_0 : m_1 = m_2 = \cdots = m_k$.

$H_1 : m_j \neq m_k$ for some j, k.

Kruskal-Wallis test

To compute the Kruskal-Wallis test,

1. Combine the groups $x_{1j}, x_{2j}, \ldots, x_{n_j j}$, $j = 1, \ldots, k$, into one larger sample $z_1, z_2, \ldots, z_n$, where $n = \sum_{j=1}^{k} n_j$.
2. Order the observations z_s from the smallest to the largest and let $R(z_s)$ be the rank of the observation z_s in the combined sample $z_1, z_2, \ldots, z_n$.
3. Calculate the group means of the ranks

$$\bar{r}_j = \frac{1}{n_j} \sum_{z_s = x_{ij}, i=1}^{n_j} R(z_s)$$

and the mean rank of the combined sample

$$\bar{r} = \frac{1}{n} \sum_{s=1}^{n} R(z_s).$$

Kruskal-Wallis test

4. Compute the group sum of squares, which describes the variance of the ranks between the groups

$$\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2,$$

5. and the total sum of squares, which describes the variance of the ranks in the combined sample

$$\sum_{s=1}^{n} (R(z_s) - \bar{r})^2.$$

Kruskal-Wallis test

Kruskal-Wallis test, assumptions

- Kruskal-Wallis test statistic,

$$K = (n-1)\frac{\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2}{\sum_{s=1}^{n}(R(z_s) - \bar{r})^2},$$

  follows, for large n, under H_0 approximately the χ^2_{k-1} -distribution.
- Under H_0, the expected value of the test statistic is approximately $k-1$ and **large** values of the test statistic suggest that the null hypothesis H_0 is false.

Kruskal-Wallis test, some notes

- Statistical software can often calculate exact p-values of the Kruskal-Wallis test when the sample size is small.
- With large sample sizes, calculation of the exact p-values would require large amounts of computing and in these cases the asymptotic p-values (based on the above-mentioned χ^2-distribution) are used.
- We assumed above that the observations follow a continuous distribution. However, Kruskal-Wallis test can be used for discrete observations as well. Then it is possible that some of the observations have the same rank. In that case, all those observations are assigned to have the median of the corresponding ranks.
- For example, if two observations have the same rank corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same ranks corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Numerical example

Consider three student groups and their statistics exam scores. The table below displays the scores and the corresponding ranks (in parenthesis).

Group 1	Group 2	Group 3
18.0 (14)	16.5 (11)	23 (22)
11.0 (4.5)	10.0 (3)	22 (20)
17.0 (12)	15.0 (8.5)	23 (22)
14.0 (7)	15.0 (8.5)	24 (24)
11.0 (4.5)	20.5 (17)	21 (18)
9.5 (2)	8.0 (1)	21.5 (19)
16.0 (10)	12.0 (6)	23 (22)
		20.0 (16)
		17.5 (13)
		19.0 (15)

Numerical example

- Calculate the rank means within the groups

$$\bar{r}_1 = \frac{1}{7}(14 + 4.5 + 12 + 7 + 4.5 + 2 + 10) = \frac{54}{7} = 7.714286,$$

$$\bar{r}_2 = \frac{1}{7}(11 + 3 + 8.5 + 8.5 + 17 + 1 + 6) = \frac{55}{7} = 7.857143,$$

$$\bar{r}_3 = \frac{1}{10}(22+20+22+24+18+19+22+16+13+15) = \frac{191}{10} = 19.1,$$

- and the mean rank of the combined sample

$$\bar{r} = \frac{1}{24}(54 + 55 + 191) = \frac{300}{24} = 12.5.$$

Numerical example

- Calculate the group sum of squares

$$\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2 = 7 * (7.714286 - 12.5)^2 + 7 * (7.857143 - 12.5)^2$$

$$+10 * (19.1 - 12.5)^2 = 746.8143,$$

- and the total sum of squares

$$\sum_{s=1}^{n} (R(z_s) - \bar{r})^2$$

$$= (14 - 12.5)^2 + (4.5 - 12.5)^2 + (12 - 12.5)^2 + (7 - 12.5)^2$$
$$+(4.5 - 12.5)^2 + (2 - 12.5)^2 + (10 - 12.5)^2 + (11 - 12.5)^2$$
$$+(3 - 12.5)^2 + (8.5 - 12.5)^2 + (8.5 - 12.5)^2$$
$$+(17 - 12.5)^2 + (1 - 12.5)^2 + (6 - 12.5)^2$$
$$+(22 - 12.5)^2 + (20 - 12.5)^2 + (22 - 12.5)^2 + (24 - 12.5)^2 + (18 - 12.5)^2$$
$$+(19 - 12.5)^2 + (22 - 12.5)^2 + (16 - 12.5)^2 + (13 - 12.5)^2 + (15 - 12.5)^2$$
$$= 1147$$

Numerical example

- Now

$$K = (n-1)\frac{\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2}{\sum_{s=1}^{n}(R(z_s) - \bar{r})^2} = (24-1)\frac{746.8143}{1147} = 14.97535.$$

- $p-$value of the test is 0.00056 so the null hypothesis of equal medians is rejected.
- We conclude that there is a statistically significant difference in the exam scores between the groups.

Bonferroni's method for pairwise comparison of the medians

- If the null hypothesis of the Kruskal-Wallis test is rejected, then the analysis can be continued by finding the groups with statistically significant differences in medians.
- A simple idea is to apply the two-sample rank test for pairwise comparisons, with a total of $C = \frac{k(k-1)}{2}$ pairs to compare.
- If the combined comparison is to be done on significance level α, then the pairwise comparisons should be done on significance level $\beta = \frac{\alpha}{C}$ (as in ANOVA).
- For example, if significance level 0.05 is desired for the combined comparison, then the pairwise comparisons reject the corresponding null hypotheses if the p-value is smaller than $\frac{0.05}{C}$.