Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Ninth lecture, December 9, 2024

Analysis of variance

Analysis of variance

The two sample t-test generalizes into analysis of variance.

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

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

For example, we could test if there is a difference in the mean monthly salary in the 10 largest cities in Germany.

ANOVA

ANOVA can be generalized in several different ways:

- In multivariate analysis of variance, MANOVA, the tested expected values are vectors. We could test the equality of the mean monthly salary *and* weekly overtime in the 10 largest cities in Germany.
- The population could also be divided into groups based on multiple factors (multifactor ANOVA), of which some can be continuous (analysis of covariance, ANCOVA). For example, we could divide people into groups based on the city they live in and based on their gender.
- In multifactor MANOVA, the tested expected values are vectors.

  ⋮

In what follows, we only consider cases where the population is divided into groups with respect to just one factor and the expected value that is tested is univariate.

ANOVA

Let $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$ be observed values of a random variable x_j, $j \in \{1, \ldots, k\}$. Assume that the observations $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$ are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_j, \sigma^2)$, $j \in \{1, \ldots, k\}$.

That is, we now have k random samples from univariate normal distributions, and the variance of all the k normal distributions are assumed to be equal.

Assume that the k samples are independent.

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

Alternative hypothesis $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$.

ANOVA

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 the variation within the groups. If the first part is much larger than the second part, there is evidence against the null hypothesis and it can be rejected.

The test of equality of the expected values is based on comparison of between-groups variance and within-groups variance. Hence the name of the method – analysis of variance.

Calculate the group means

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j}$$

and the combined sample mean

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

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

ANOVA

Consider the total sum of squares

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x})^2,$$

the variance between groups (group sum of squares)

$$SSG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{x}_j - \overline{x})^2 = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2,$$

and the variance within groups (error sum of squares)

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_j)^2 = \sum_{j=1}^{k} (n_j - 1) s_j^2,$$

where $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_j)^2$.

Now the total sum of squares $SST = SSG + SSE$.

ANOVA

- The F-test statistic
$$F = \frac{n-k}{k-1}\frac{SSG}{SSE}.$$
- Under the null hypothesis, the test statistic follows the F-distribution with parameters $(k-1)$ and $(n-k)$.
- The expected value of the test statistic under H_0 is $\mathbb{E}[F] = \frac{n-k}{n-k-2}$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- Python:
`F_stat,p_value = scipy.stats.f_oneway(group1,...,groupk)`

F-distribution

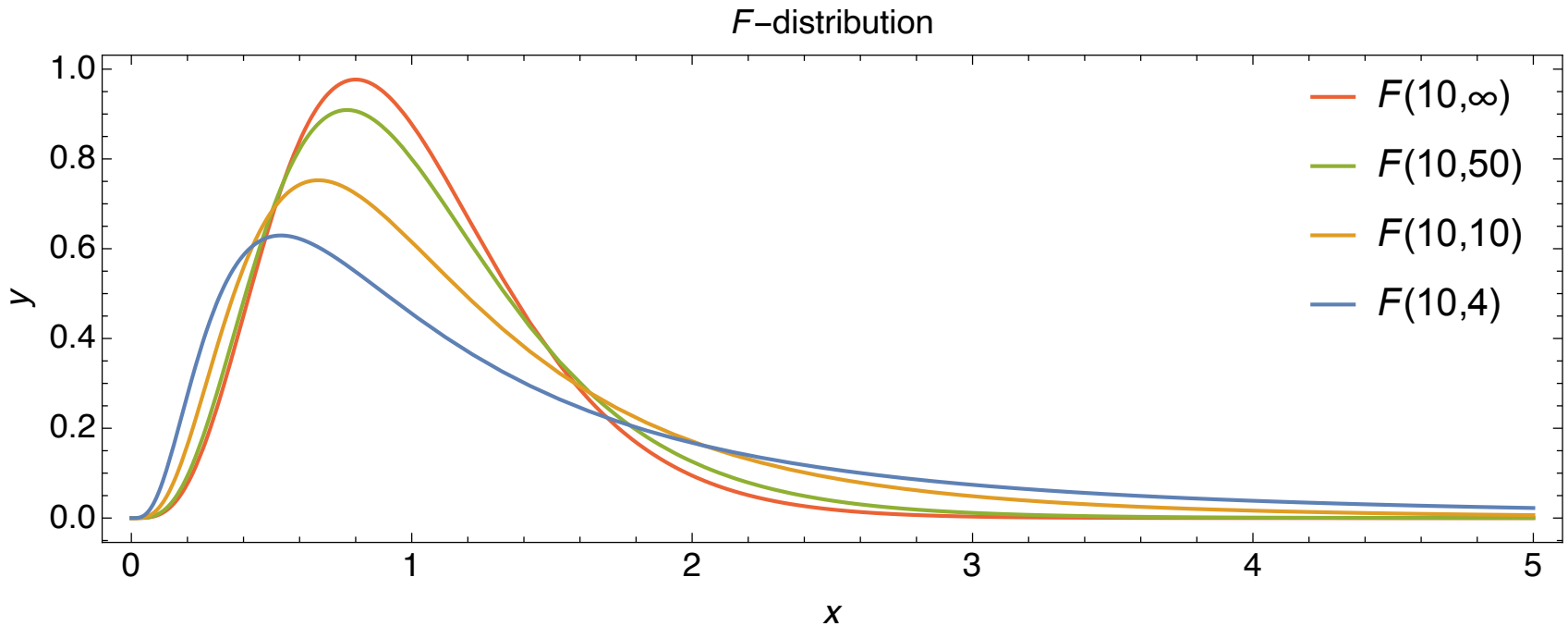

Figure: F-distributions with different parameters.

Numerical example

A research group was set to study whether the expected value of a specific laboratory test L differs between patients that are on different medications (A, B, C). 10 patients receiving medication A (group 1), 10 patients receiving medication B (group 2), and 10 receiving medication C (group 3) were picked randomly and lab test L was taken. The next table shows the accurately measured laboratory test results.

Group 1 (A)	Group 2 (B)	Group 3 (C)
0.111	0.109	0.119
0.123	0.107	0.124
0.109	0.103	0.125
0.120	0.104	0.117
0.115	0.098	0.111
0.112	0.110	0.120
0.117	0.101	0.118
0.110	0.115	0.116
0.119	0.099	0.122
0.116	0.111	0.119

Table: Laboratory test results for groups 1, 2, and 3.

The group means are $\overline{x}_1 = 0.1152$, $\overline{x}_2 = 0.1057$, and $\overline{x}_3 = 0.1191$ and the combined mean is $\overline{x} = 0.1133$. The group variances are $s_1^2 = 2.173333 \cdot 10^{-5}$, $s_2^2 = 3.134444 \cdot 10^{-5}$, and $s_3^2 = 1.654444 \cdot 10^{-5}$. The total sum of squares is

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x})^2 = \sum_{i=1}^{10} (x_{1,i} - 0.1133)^2 + \sum_{i=1}^{10} (x_{2,i} - 0.1133)^2 + \sum_{i=1}^{10} (x_{3,i} - 0.1133)^2 = 0.001576667,$$

the group sum of squares is

$$SSG = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2 = \sum_{j=1}^{3} 10 \cdot (\overline{x}_j - 0.1133)^2 = 0.0009500667,$$

and the error sum of squares is

$$\begin{aligned} SSE &= \sum_{j=1}^{k} (n_j - 1) s_j^2 \\ &= 9 \cdot (2.173333 \cdot 10^{-5} + 3.134444 \cdot 10^{-5} + 1.654444 \cdot 10^{-5}) \\ &= 0.0006265999. \end{aligned}$$

The value of the test statistic is

$$F = \frac{n-k}{k-1}\frac{SSG}{SSE} = \frac{27}{2} \cdot \frac{0.0009500667}{0.0006265999} = 20.46904.$$

Under the null hypothesis, the test statistic follows the F-distribution with parameters 2 and 27. The one-tailed critical value on 5% significance level is $3.354 < 20.46904$. The null hypothesis can be rejected.

Solution using Python:

```
import pandas as pd
from scipy.stats import f_oneway
data = pd.DataFrame({
"A": [.111,.123,.109,.120,.115,.112,.117,.110,.119,.116],
"B": [.109,.107,.103,.104,.098,.110,.101,.115,.099,.111],
"C": [.119,.124,.125,.117,.111,.120,.118,.116,.122,.119]
})
F_stat,p_value = f_oneway(*data.T.values)
```

Pairwise comparison

Pairwise comparison

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

The next step in the analysis is pairwise comparison. The goal in pairwise comparison is to identify 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 the 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 1 error is inflated greatly above its set level.

Bonferroni's method for pairwise comparison

Consider pairwise comparison of the expected values. There are $c = \frac{k(k-1)}{2}$ pairs in total to compare. Let us consider analyzing the pairs with the t-test.

Let β be the significance level of the c pairwise comparisons, i.e., the (upper bound for the) probability that H_0 is incorrectly rejected in a single comparison. Let α 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 one type 1 error during the c tests.

Probability theory shows that $\alpha \leq c\beta$. For this reason, if the significance level α is chosen for the combined comparison, the individual comparisons must be done on level $\beta = \frac{\alpha}{c}$. (For pairwise tests, instead of p-value α, one looks for p-values $\leq \frac{\alpha}{c}$.) This is known as the Bonferroni correction.

Example. We want to investigate, on significance level $\alpha = 0.05$, the differences in expected values for $k = 5$ groups. Then there are $c = 10$ pairs to compare. The t-test should be carried out at significance level $\frac{0.05}{10} = 0.005$ for each of the 10 pairs.

Bartlett's test for equality of variances

Bartlett's test for equality of variances

ANOVA makes two key assumptions:

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

As usual, the first assumption can (by CLT) be replaced with a large enough sample size n.

The second assumption is also required 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

Let $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$ be observed values of a random variable x_j, $j \in \{1, \ldots, k\}$. Assume that the observations are i.i.d. and follow a normal distribution $\mathcal{N}(\mu_j, \sigma_j^2)$, $j \in \{1, \ldots, k\}$. Assume that all the k samples are independent.

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

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

Bartlett's test for equality of variances

Let

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

and

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

Let

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

where

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

and

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

Bartlett's test for equality of variances

- Bartlett's test statistic

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

where

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

and

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

- If the sample size is large, then under the null hypothesis the test statistic approximately follows the χ^2 distribution with $(k-1)$ degrees of freedom.
- The expected value of the test statistic under H_0 is $\mathbb{E}[B] = k - 1$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false. The null hypothesis H_0 is rejected if the p-value is small enough.
- Python:

`Q_stat,p_value = scipy.stats.bartlett(group1,...,groupk)`

Kruskal–Wallis test

Analysis of variance

ANOVA tests the equality of the expected values of normally distributed samples. Next we consider a non-parametric alternative to ANOVA.

Kruskal–Wallis test

The Kruskal–Wallis test is similar to one way analysis of variance, but it does not require the normality assumption.

It is a generalization of the two sample Wilcoxon rank test.

Kruskal–Wallis test

Let $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$ be observed values of a continuous random variable x_j, $j \in \{1, \ldots, k\}$. Assume that the observations $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$ are i.i.d. Assume also that the k samples are independent and that the variables x_j, $j \in \{1, \ldots, k\}$, follow the same distribution up to location shifts (i.e., x_j follow otherwise the same distribution, but possibly with different medians) and assume that the variables x_j have population medians m_j, $j \in \{1, \ldots, k\}$.

The null hypothesis $H_0 : m_1 = m_2 = \cdots = m_k$.

The alternative hypothesis $H_1 : m_i \neq m_j$ for some $i \neq j$.

Kruskal–Wallis test

The Kruskal–Wallis test is based on examining the ranks of the observations.

Kruskal–Wallis test

Combine the groups $x_{1,j}, x_{2,j}, \ldots, x_{n_j,j}$, $j \in \{1, \ldots, k\}$, into one big sample $z_1, z_2, \ldots, z_n$, where $n = \sum_{j=1}^{k} n_j$. Order the observations z_s from the smallest value to the largest value. Let $R(z_s)$ be the rank of the observation z_s in the combined sample $z_1, z_2, \ldots, z_n$.

Calculate the group means of the ranks

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

and the mean of the combined sample

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

Kruskal–Wallis test

Consider the group sum of squares, which describes the variation of the ranks between the groups

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

and the total sum of squares, which describes the variation of the ranks in the combined sample

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

Kruskal–Wallis test

- Kruskal–Wallis test statistic

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

- Under the null hypothesis H_0, if the sample size is large, the test statistic approximately follows the χ^2 distribution with $k-1$ degrees of freedom.
- Under H_0, the (asymptotic) expected value of the test statistic is $k-1$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p-value is small enough.
- Python:
  `K_stat,p_value = scipy.stats.kruskal(group1,...,groupk)`

Kruskal–Wallis test

Statistical software often calculate exact p-values for the Kruskal–Wallis test when the sample size is small. With large sample sizes, the calculation of exact p-values requires intense computations and in these cases asymptotic p-values (based on the χ^2 distribution) are used.

Discrete distributions

We assumed above that the observations follow some continuous distribution. However, the 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 this 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.

ANOVA vs. Kruskal–Wallis

ANOVA is explicitly a test for the equality of the expected values. The Kruskal–Wallis test can, technically, be seen as a comparison of the expected ranks. Hence, the Kruskal–Wallis test is in fact more general than a test for the equality of the medians. It tests whether the probability that a random observation from each group is equally likely to be above or below a random observation from another group. The test is sensitive to differences in medians and that is why it is usually considered a test for the equality of medians.

Example

Consider three student groups (1, 2, 3) 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)

Example

Calculate the rank means within 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

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

Calculate the group sum of squares:

$$\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2 = 7 \cdot (7.714286 - 12.5)^2 + 7 \cdot (7.857143 - 12.5)^2$$
$$+ 10 \cdot (19.1 - 12.5)^2 = 746.8143$$

and the total sum of squares:

$$\sum_{s=1}^{n} (R(z_s) - \bar{r})^2$$
$$= (14 - 12.5)^2 + (4.5 - 12.5)^2 + (12 - 12.5)^2 + (7 - 12.5)^2$$
$$+ (4.5 - 12.5)^2 + (2 - 12.5)^2 + (10 - 12.5)^2 + (11 - 12.5)^2$$
$$+ (3 - 12.5)^2 + (8.5 - 12.5)^2 + (8.5 - 12.5)^2$$
$$+ (17 - 12.5)^2 + (1 - 12.5)^2 + (6 - 12.5)^2$$
$$+ (22 - 12.5)^2 + (20 - 12.5)^2 + (22 - 12.5)^2 + (24 - 12.5)^2$$
$$+ (18 - 12.5)^2 + (19 - 12.5)^2 + (22 - 12.5)^2$$
$$+ (16 - 12.5)^2 + (13 - 12.5)^2 + (15 - 12.5)^2$$
$$= 1147.$$

Example

Now

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

The p-value of the test is clearly less than 0.05 – the value 5.79 corresponds approximately to p-value 0.05. The null hypothesis can be rejected. There is a statistically significant difference in the exam success between the three groups.

Solution using Python:

```
import pandas as pd
from scipy.stats import kruskal
nan = float('nan')
data = pd.DataFrame({
"1": [18,11,17,14,11,9.5,16,nan,nan,nan],
"2": [16.5,10,15,15,20.5,8,12,nan,nan,nan],
"3": [23,22,23,24,21,21.5,23,20,17.5,19]})
K_stat,p_value = kruskal(*data.T.values,nan_policy='omit')
```

Bonferroni's method pairwise comparison

If the null hypothesis of the Kruskal–Wallis test is rejected, then the next step is pairwise comparison.

Bonferroni's method pairwise comparison

Compare the pairwise equality/difference of the medians. There are $c = \frac{k(k-1)}{2}$ pairs to compare.

The first idea would be to use the Wilcoxon two sample rank test for pairwise comparison It should be noted that if the comparison is done on significance level α, then the pairwise comparisons should be done on significance level $\beta = \frac{\alpha}{c}$. For example, if the significance level 0.05 is used for the combined comparison, then the pairwise comparisons can be used to reject the null hypothesis if the p-value is smaller than or equal to $\frac{0.05}{c}$.

Numerical example

Previously we applied ANOVA to examine whether the expected value of a specific laboratory test L differs between patients that are on different medications (A, B, C). The null hypothesis was rejected. We are now a bit worried about the normality assumption and we decide to conduct pairwise comparisons using the Wilcoxon two sample rank test. We decide to use significance level 0.05.

Group 1 (A)	Group 2 (B)	Group 3 (C)
0.111	0.109	0.119
0.123	0.107	0.124
0.109	0.103	0.125
0.120	0.104	0.117
0.115	0.098	0.111
0.112	0.110	0.120
0.117	0.101	0.118
0.110	0.115	0.116
0.119	0.099	0.122
0.116	0.111	0.119

Table: Laboratory test results for groups 1, 2, and 3.

There are $c = \frac{k(k-1)}{2} = 3$ pairs to compare. Thus, in pairwise comparison, we reject the null hypothesis (equality of the medians) if the p-value is smaller than or equal to $\frac{0.05}{3} = 0.0166\ldots$

Wilcoxon rank sum test with continuity correction
data: A and B
W=91, p-value = 0.002169
alternative hypothesis: true location shift is not equal to 0

Wilcoxon rank sum test with continuity correction
data: A and C
W=26, p-value = 0.07478
alternative hypothesis: true location shift is not equal to 0

Wilcoxon rank sum test with continuity correction
data: B and C
W=1.5, p-value = 0.0002821
alternative hypothesis: true location shift is not equal to 0

Two (A vs. B and B vs. C) of the p-values are smaller than $\frac{0.05}{3} = 0.0166\ldots$ We conclude that the median of the laboratory test L of the patients that are on medication B differs from the median of the laboratory test L of the patients that are on medication A or on medication C.