Anthropological Review • Vol. 61, pp. 103–115, Poznań 1998
https://doi.org/10.18778/1898-6773.61.08

Comparisons of means of many populations

Krzysztof Kościński

Abstract
Testing hypotheses on the equality of means is very common. Methods of comparing simultaneously many populations with regard to their means are less popular than the methods of comparing two populations. For this reason, the paper deals only with the latter case. Several tests used for the verification of hypotheses on the equality of many means are described. If such a hypothesis is rejected with these tests, we can find the so-called homogenous groups. An example is given showing how to use the tests and how to interpret the results obtained.
Krzysztof Kościński 1998; *Anthropological Review*, vol. 61, Poznań 1998, pp. 103–115, tables 2. ISBN 83-86969-15-0, ISSN 0033-2003

Introduction

Methods of comparing two populations with regard to their means are very popular and easily accessible in the literature. Tests for comparing many populations are less frequent, hence this paper will deal only with cases when simultaneous comparison of more than two population means is made. Further, I will discuss only methods of the verification of hypotheses on the equality of particular mean values. MILLER [1966], OKTABA [1971] and SITEK [1973] discuss hypotheses relating to more complex interrelations between means. CIEŚLIK, SITEK [1987] describe methods of determining the populations with the highest or the lowest mean value.

The purpose of this article is to briefly describe the methods of testing hypotheses on the equality of many means with the use of a few mathematical formulae and to make readers better understand the function of certain buttons in the popular program Statistica. The work also provides anthropologists with an analysis of an example based on anthropological data.

The layout of this paper is as follows. First, I will describe some statistical tests that make it possible to verify hypotheses on the equality of means of a number of variables. Next, I will show how to apply these tests to exemplary empirical data (body height of 9-year-old boys in various years). In the next step, I will make some remarks on how to interpret the results obtained.

Let us assume that we are interested in k populations Π_1, ..., Π_k with regard to a specific metrical trait being a random variable in each population – X_1, ..., X_k. We

Institute of Anthropology
Adam Mickiewicz University
Fredry 10, 61-701 Poznań, Poland

assume that these variables have normal distributions with means $m_1, ..., m_k$ and variances $\sigma_1^2, ..., \sigma_k^2$. From each population Π_i ($i = 1,..., k$) we draw a random sample of n_i size and composed of elements $x_{i,1}, ..., x_{i,ni}$. For each sample we calculate its mean (equation 1) and variance (equation 2):

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

(2) $$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$$

If these variables have normal distribution or sample sizes are large, then sample means $\bar{x}_i$ have normal distribution with means m_i and variances σ_i^2/n_i. What we want to find out is whether the means of these populations are equal. For this purpose, we have to test hypothesis H: $m_1 = ... = m_i = ... = m_k$, where m_i is the mean of i-th population. According to this hypothesis mean values of all populations are equal.

If the hypothesis is rejected, we can search for the so-called homogenous groups. A population group is homogenous if it includes populations whose equality we do not reject (using a given test and at definite α). Among k populations we can find a few homogenous groups that usually are not disjoint sets.

Let us assume that variables $X_1, ..., X_k$ are not correlated, that they have equal variances $\sigma_1^2 = ... = \sigma_k^2$, and that sample sizes are equal $n_1 = ... = n_k$. Then, sample means $\bar{x}_i$ will also have equal variances. In this situation, if $\bar{x}_i$ is lower than $\bar{x}_{i'}$ ($i, i' = 1,..., k$) and both means belong to a homogenous group, each sample mean $\bar{x}_j$ satisfying inequality $\bar{x}_i < \bar{x}_j < \bar{x}_{i'}$ will also belong to this homogenous group. Owing to that rule, homogenous groups can be conveniently represented in a graphical form, providing that the group means have been ordered into an ascending or descending series. For instance:

$$\underline{\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_3 \quad \bar{x}_4} \quad \bar{x}_5 \quad \underline{\bar{x}_6 \quad \bar{x}_7}$$

Each underlining marks one homogenous group and comprises according populations. If sample sizes are not equal then $\bar{x}_j$ may not belong to the homogenous group including $\bar{x}_i$ and $\bar{x}_{i'}$, even if $\bar{x}_i < \bar{x}_j < \bar{x}_{i'}$. Such a situation will occur further in the article (cf. Examples).

Description of the tests

There are many tests designed for checking the equality of means of many populations at the same time. In this paper I will discuss the following ones: Scheffe test, t-Bonferroni test, Tukey test, T' test (Spjøtvoll-Stoline test), Newman-Keuls test and the least significant difference test (LSD test). If variables $X_1, ..., X_k$ are correlated, one should use Scheffe or t-Bonferroni test. If these variables are not correlated then any test may be used (special versions of Scheffe and t-Bonferroni tests are needed).

Scheffe test

Let us assume that variables X_1, ..., X_k are correlated. In order to estimate the covariance between particular variables (this is necessary when verifying hypotheses with Scheffe test), sample sizes must be equal ($n_1 = ... = n_k$) and every element from any sample must have one „equivalent" in each of the other samples. Two possible examples of such a situation are as follows: 1) we have a group of n individuals, of which each individual was studied in k experimental conditions; 2) each of k populations contains n classes (types) of elements and there are the same classes of elements in each population. In Scheffe test sample variances do not need to be equal.

The hypothesis that all population means m_1, ..., m_k are equal may be tested by making k–1 comparisons of a marginal, with regard to the value, sample mean (let us mark it as $\bar{x}_k$) with all other sample means. Thus, the hypothesis may be expressed in the following way:

$$H: m_k - m_{k-1} = m_k - m_{k-2} = ... = m_k - m_2 = m_k - m_1 = 0$$

The hypothesis will be falsified if at least one comparison is falsified. On the other hand, the equality $m_k - m_i = 0$ will be rejected when the following inequality is true:

$$(3) \qquad |\bar{x}_k - \bar{x}_i| > \sqrt{\frac{(n-1)(k-1)\left[S_k^2 + S_i^2 - 2\,\mathrm{cov}(i,k)\right]F_{k-1,n-k+1}(\alpha)}{n(n-k+1)}}$$

where $\bar{x}_i$, $\bar{x}_k$ are sample means, and S_k^2, S_i^2 are sample variances of variables X_k and X_i, computed with equations (1) and (2); $F_{k-1,\,n-k+1}(\alpha)$ is a tabular value of F-Snedecor distribution for k–1 and n–k+1 degrees of freedom and chosen α; cov(i,k) is sample covariance between these variables computed with the equation:

$$(4) \qquad \mathrm{cov}(i,k) = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{i,j} - \bar{x}_i\right)\left(x_{k,j} - \bar{x}_k\right).$$

Value α refers to all k–1 comparisons, rather than to a single comparison, i.e., if all population means are equal, there is probability α that sample results will make us reject the hypothesis on this equality, because at least one equality of pairs of means will be rejected. However, the probability of the rejection of true hypothesis with Scheffe test is not exactly equal α. It is in fact lower than α. This means that Scheffe test gives too long confidence intervals.

When variables are not correlated, we can use the following version of Scheffe test. Making k–1 comparisons between the marginal sample mean and the other sample means, the hypothesis H: $m_k = m_i$ will be rejected when the following inequality is satisfied:

$$(5) \qquad |\bar{x}_k - \bar{x}_i| > \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)S^2(k-1)F_{k-1,\nu}(\alpha)}$$

where ν stands for degrees of freedom:

(6) $$v = \left(\sum_{i=1}^{k} n_i \right) - k$$

and S^2 is the general variance calculated with the formula:

(7) $$S^2 = \frac{1}{v} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2 .$$

The total probability of error in such series of comparisons is, like for non-correlated variables, less then α.

t-Bonferroni test

This is another test employed for the verification of hypotheses on correlated variables. In t-Bonferroni test, like in Scheffe test, variances of variables can be different. In the case of the hypothesis consisting of k–1 successive comparisons of marginal sample mean with the other means, the i-th comparison (and, consequently, the hypothesis on the equality of all population means) will be rejected when the following inequality is satisfied:

(8) $$\left| \bar{x}_k - \bar{x}_i \right| > t_{n-1}(\alpha') \sqrt{\frac{S_k^2 + S_i^2 - 2\,\mathrm{cov}(i,k)}{n}}$$

where $\alpha' = \alpha/(k-1)$. t-Bonferroni test has the same shortcoming as Scheffe test has, namely too long confidence intervals. As a result, the probability of error of the first kind is lower than intended α. This feature can be useful when choosing a test for the verification of a concrete hypothesis. We shall choose the test that gives shorter confidence interval. It appears that the values of the following expressions should be compared:

$$t_{n-1}(\alpha') \quad \text{and} \quad \sqrt{\frac{(n-1)(k-1)F_{k-1,n-k+1}(\alpha)}{n-k+1}}$$

t-Bonferroni test should be chosen when the value of the expression on the left side is lower than the value of the right side expression. In the opposite case we should choose Scheffe test.

There is also another version of t-Bonferroni test, designed for non-correlated variables. Making k–1 comparisons between the marginal sample mean and the others sample means, we will deny the equality of all population means when for an i (i = 1, ..., k) the following inequality is satisfied:

(9) $$\left| \bar{x}_k - \bar{x}_i \right| > t_v(\alpha') \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j} \right) S^2}$$

where ν and S^2 are computed according to equations (6) and (7). The total probability of error in this series of comparisons is lower than α.

All the tests discussed further are applicable only for non-correlated variables.

Tukey test

Except the assumption on the lack of correlation between the variables this test is based also on an important assumption of the equality of variances ($\sigma_1^2 = \ldots = \sigma_k^2$) which can be tested with Bartlett test [GREŃ 1987].

Let us assume that we want to check the hypothesis H: $m_1 = \ldots = m_k$. Having drawn k samples of $n_1, \ldots, n_k$ sizes and having obtained means $\bar{x}_1, \ldots, \bar{x}_k$ we will reject the hypothesis if for a pair i and j ($i,j = 1, \ldots, k$) the following inequality is satisfied:

$$(10) \qquad \left|\bar{x}_i - \bar{x}_j\right| \sqrt{\frac{2}{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) S^2}} > q_{k,\nu}(\alpha)$$

where ν and S^2 are determined by equations (6) and (7); while $q_{k,\nu}(\alpha)$ is a critical value of studentized range (tables can be found for instance in OKTABA [1971]). In other words, the hypothesis will be falsified when the highest, with regard to indexes i and j, value of the left side of equation (10) is higher than the appropriate value $q_{k,\nu}(\alpha)$. In the case of equal sample sizes ($n_1 = \ldots = n_k = n$), the hypothesis will be rejected if the following inequality is satisfied:

$$(11) \qquad \left(\bar{x}_{\max} - \bar{x}_{\min}\right) \sqrt{\frac{n}{S^2}} > q_{k,\nu}(\alpha)$$

where $\bar{x}_{max}$ and $\bar{x}_{min}$ are the highest and the lowest sample mean, respectively.

If the hypothesis that all population means are equal has been rejected, we can look for homogenous groups of means. A given group of means (being part of all k means) will be referred to as homogenous if:

1. for each pair i, j equation (10) is not true (i and j refer only to those means that belong to the group) – unequal sample sizes case;
2. equation (11) is not true ($\bar{x}_{max}$ and $\bar{x}_{min}$ are chosen from among the means belonging to the group) – equal sample sizes case.

If we single out a few homogenous groups then the probability of error α refers to all groups jointly.

Spjøtvoll-Stoline test (T' test)

The test is recommended when sample sizes differ strongly from one another. Variables $X_1, \ldots, X_k$ cannot be correlated and their variances must be equal. The hypothesis about the equality of k means will be rejected if for a pair i, j ($i, j = 1, \ldots, k$) the following inequality is true:

$$(12) \qquad \left|\bar{x}_i - \bar{x}_j\right| > q'_{k,\nu}(\alpha) \frac{S}{\sqrt{\min(n_i, n_j)}}$$

where $\min(n_i,n_j)$ is the lower value of n_i and n_j. Tables with critical values $q'_{k,\nu}(\alpha)$ can be found in STOLINE [1978].

Due to the fact that, where Tukey test inserts harmonic mean of two values, T' test chooses the lower one of them and that $q_{k,\nu}(\alpha)$ values differ slightly from $q'_{k,\nu}(\alpha)$ val-

ues (if $k > 8$ and $\alpha \leq 0.2$, they can be considered equal), T' test gives longer confidence intervals than Tukey test. Since Tukey test checking the hypothesis that k means are equal exposes us to the risk of error equal α, then in the case of T' test the risk is lower than α. When sample sizes are equal, $k > 8$ and $\alpha \leq 0.2$, T' test and Tukey test are identical.

If we reject the equality of all means, we can single out homogenous groups. The method is the same as in Tukey test, that is i and j in formula (12) refer only to those means that belong to the group under study. The total probability of error is here also lower than intended α.

Newman-Keuls test

In the case of this test the hypothesis that k means are equal is checked according to exactly the same rules as these used in Tukey test (equations 10 and 11). Differences appear only when identifying homogenous groups. A group will be regarded as homogenous if for no i, j (i and j refer only to those means that belong to a given group) the following formula is true:

$$\left|\bar{x}_i - \bar{x}_j\right| \sqrt{\frac{2}{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) S^2}} > q_{l,\nu}(\alpha) \tag{13}$$

and in the case of equal sample sizes:

$$\left(\bar{x}_{\max} - \bar{x}_{\min}\right) \sqrt{\frac{n}{S^2}} > q_{l,\nu}(\alpha) \tag{14}$$

Parameter l in the above formulas is a number of means in the group under study. Because $l < k$, thus $q_{l,\nu}(\alpha) < q_{k,\nu}(\alpha)$. Therefore Newman-Keuls test gives shorter confidence intervals than Tukey test and homogenous groups found with Newman-Keuls test contain fewer means than homogenous groups found with Tukey test.

Newman-Keuls test is constructed in such a way that the probability of error α refers to each homogenous group separately. This means that α is a risk of not including a population into a homogenous group in spite of the fact that this population's mean is equal to the means of populations belonging to this group. The total probability of error is in this test higher than α.

Least significant difference test (LSD test)

In this test the hypothesis on the equality of means of all k populations (H: $m_1 = \ldots = m_k$) will be rejected when the following inequality is true:

$$F = \frac{\nu}{k-1} \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}\right)^2 n_i}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2} > F_{k-1,\nu}(\alpha) \tag{15}$$

where $\bar{x}$ is the arithmetic mean of all the measurements in all samples; $F_{k-1,\nu}(\alpha)$ is a tabular value of F-Snedecor distribution and ν is determined with formula (6). The probability of erroneous rejection of the hypothesis is equal α. If the hypothesis is rejected, we may compare pairs of sample means in order to check which of them are equal. The equality of two means will be denied if the following inequality is satisfied:

$$|\bar{x}_i - \bar{x}_j| > t_\nu(\alpha)\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S^2} \tag{16}$$

where ν and S^2 are determined with equations (6) and (7); and $t_\nu(\alpha)$ is a tabular value of t-Student distribution. We may accomplish $k(k-1)/2$ comparisons of pairs of means. A set containing populations that are equal in pairs makes a homogenous group. Because in the second step of the LSD test, α refers to a single comparison, in $k(k-1)/2$ comparisons the total probability of error is much higher than α.

What is interesting, it may happen that in spite of the rejection of the equality of all population means with F-Snedecor test (step one), t-Student test will not be able to find any pair of means that are not equal (step two).

Summary

I have discussed the following tests designed for the verification of hypotheses on the equality of all or part of k populations: Scheffe test, t-Bonferroni test, Tukey test, Spjøtvoll-Stoline test (T' test), Newman-Keuls test and the least significant difference test (LSD test). Each of these tests allows for the verification of the hypothesis on the equality of the marginal sample mean with the other ones. Whether the equality of a pair of population means will be denied or not depends on whether the absolute value of the difference between these sample means $|\bar{x}_i - \bar{x}_k|$ is higher or lower than a given critical value, specific to each test. Due to the different critical values, hypothesis H: $m_1 = ... = m_i = ... = m_k$ may be rejected with one test but not with another. Also homogenous groups may differ depending on the test used. Different critical values are responsible for different probabilities of error characteristic of the tests. The longer the confidence intervals a test gives (i. e. the higher the critical value), the lower the probability of error α.

Table 1 below presents for each test:

- critical value, which we compare with $|\bar{x}_i - \bar{x}_k|$ to state whether given means can be found as equal or not;
- probability of error for the hypothesis that all population means are equal: lower than, higher than or equal α;
- specific factor for Scheffe test, t-Bonferroni test, Tukey test, Newman-Keuls test and LSD test, because critical values of the tests contain a common factor $[(1/n_i + 1/n_j)S^2]^{1/2}$.

Table 1 refers to the comparisons of non-correlated variables, hence relevant versions of Scheffe test and t-Bonferroni test have been taken into consideration.

Table 1

test	critical value	risk of error	specific factor
Scheffe	$\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S^2(k-1)F_{k-1,\nu}(\alpha)}$	$<\alpha$	$\sqrt{(k-1)F_{k-1,\nu}(\alpha)}$
t-Bonferroni	$t_\nu(\alpha')\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S^2}$	$<\alpha$	$t_\nu(\alpha')$
LSD	$t_\nu(\alpha)\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S^2}$	$<\alpha$	$t_\nu(\alpha)$
Tukey and N.-K.	$q_{k,\nu}(\alpha)\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\frac{S^2}{2}}$	$=\alpha$	$q_{k,\nu}(\alpha)\frac{1}{\sqrt{2}}$
T'	$q'_{k,\nu}(\alpha)\sqrt{\frac{S^2}{\min(n_i,n_j)}}$	$<\alpha$	--

Examples

Every year from 1980 to 1990 body height of nine-year-old boys was studied at the Department of Human Growth Biology at Adam Mickiewicz University in Poznań. Sample sizes, means and variances are shown in Table 2.

Table 2

i	year	$\bar{x}_i$	S_i^2	n_i
1	1980	136.21	35.32	272
2	1981	134.09	37.38	480
3	1982	134.16	34.41	297
4	1983	134.78	32.38	237
5	1984	132.51	32.98	354
6	1985	133.91	35.88	426
7	1986	133.10	30.99	224
8	1987	133.20	36.86	516
9	1988	134.23	33.72	330
10	1989	133.90	34.64	766
11	1990	133.30	33.48	186
		$\bar{x} = 133.90$	$S^2 = 35.41$	$n = 4088$

Let us suppose that we want to test the hypothesis that body height of boys is the same every year, and if the hypothesis is rejected we want to determine which years differ significantly with one another and which do not.

The total variance for all samples together is determined with equation (7). Bartlett test has not rejected the hypothesis that all population variances are equal. Therefore, we can assume that the samples are taken from normal populations with means $m_1, ..., m_{11}$ and a common variance S^2, and that they are independent. These assumptions allow us for the application of all the above-discussed tests. In each case the level of significance was fixed at $\alpha = 0.05$.

Tukey test

Putting $S^2 = 35.41$ and $q_{k,\nu}(\alpha) = q_{11,4077}(0.05) = 4.55$ into formula (10), we find out that a pair of means should be considered unequal when:

$$|\bar{x}_i - \bar{x}_j| / \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} > 19.145$$

We can easily ascertain that null hypothesis (that all population means are equal) must be rejected. In order to do this it is enough to compare, for instance, sample 5 and 1, where 3.70/0.081 = 45.902 > 19.145. Still, we can distinguish homogenous groups. Let us arrange sample means in ascending order: 5, 7, 8, 11, 10, 6, 2, 3, 9, 4, 1 (sample numbers are taken from Table 2). Tukey test distinguishes the following homogenous groups: group I – 5, 7, 8, 11; group II – 7, 8, 11, 10, 6, 2, 3, 9, 4; group III – 4, 1.

T' test (Spjøtvoll-Stoline)

Because for $k > 8$ and $\alpha \leq 0.2$ $q'_{k,\nu}(\alpha) = q_{k,\nu}(\alpha)$, then $q'_{11,4077}(0.05) = 4.55$. The equality of two means will be rejected if $|\bar{x}_i - \bar{x}_k|[\min(n_i,n_j)]^{1/2} > 27.075$ (equation 12). The test singles out the following homogenous groups: group I – 5, 7, 8, 11, 10, 6; group II – 7, 8, 11, 10, 6, 2, 3, 9, 4; group III – 4, 1.

Newman-Keuls test

In this test a group of l populations is homogenous if the following inequality is satisfied for no pair of means from this group (see equation 13):

$$|\bar{x}_i - \bar{x}_j| / \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} > q_{l,4077}(\alpha)/0.238$$

In this way, we can find two homogenous groups: group I – 5, 7, 8, 11; group II – 7, 8, 11, 10, 6, 2, 3, 9, 4.

The least significant difference test (LSD test)

The hypothesis on the equality of all means is rejected, because inequality (15) is satisfied:

$$F = \frac{4077}{10}\frac{2854.527}{141895.957} = 8.203 > 1.83 = F_{10,4077}(0.05)$$

Inserting $S^2 = 35.41$ and $t_{4077}(0.05) = 1.960$ into equation (16) we find out that the equality of two means must be rejected if the following inequality is true:

$$|\bar{x}_k - \bar{x}_i| / \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} > 11.664$$

We obtain the following homogenous groups: group I – 5, 7, 8, 11; group II – 7, 8, 11, 10, 6; group III – 11, 10, 6, 2, 3, 9; group IV – 6, 2, 3, 9, 4.

Both Scheffe test and t-Bonferroni test have been designed for correlated variables, however we can use a special version of **Scheffe test** intended for non-correlated variables. Let us see the results obtained with this version of the test.

Inserting $k-1 = 10$, $S^2 = 35.41$ and $F_{10,4077}(0.05) = 1.83$ into equation (5) we obtain an inequality we will use to single out homogenous groups:

$$|\bar{x}_k - \bar{x}_i| / \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} > 25.456$$

The results are as follows: group I – 5, 7, 8, 11, 10, 6, 2, 3, 9; group II – 7, 8, 11, 10, 6, 2, 3, 9, 4; group III – 3, 9, 4, 1.

Homogenous groups obtained with the above-described tests can be shown in graphical form:

– Tukey test:	5	7	8	11	10	6	2	3	9	4	1
– T' test:	5	7	8	11	10	6	2	3	9	4	1
– Newman-Keuls test:	5	7	8	11	10	6	2	3	9	4	1
– LSD test:	5	7	8	11	10	6	2	3	9	4	1
– Scheffe test:	5	7	8	11	10	6	2	3	9	4	1

In the above-given examples five methods were used to distinguish homogenous groups among eleven definite populations and each method produced a different result. This requires an explanation.

In Tukey test $\alpha = 0.05$ refers to all groups simultaneously, that is these groups are distinguished at the same critical value at which the equality of all means is tested ($q_{k,\nu}(\alpha)$).

In T' test probability of error, like in Tukey test, refers to all groups simultaneously, and is lower than α. Longer confidence intervals are due to the fact that where Tukey test computes the harmonic mean of two samples' sizes, T' test uses the lower value of these sample sizes.

In Newman-Keuls test $\alpha = 0.05$ refers to each group separately, that is α is a risk of a population not being included into a homogenous group in spite of the fact that this population's mean is equal to the means of populations belonging to the group. Therefore, the total probability of error exceeds α (in Tukey test it is exactly equal α). For this reason, the Newman-Keuls test gives shorter confidence intervals than Tukey

test and, in consequence, rejects the equality of means 1 and 4 found out to be equal with Tukey test.

In LSD test the probability of error α refers in each case to a pair of means under concern, which entails short confidence intervals and a high total risk of error. That is why, this test is not recommended for determining homogenous groups. Still, it may be used for testing the hypothesis about the equality of all means.

Scheffe test resulted in the longest confidence intervals. This is so because the real probability of error it entails is lower than chosen $\alpha = 0.05$ (the same is true of t-Bonferroni test). As a result, Scheffe test and t-Bonferroni test are recommended for correlated variables in the case of which the other tests cannot be applied.

In the introduction, I said that if variables are independent, their variances are equal and sample sizes are equal, then, if for two sample means $\bar{x}_i$, $\bar{x}_{i'}$ from one homogenous group $\bar{x}_i < \bar{x}_{i'}$, any $\bar{x}_j$ satisfying the expression: $\bar{x}_i < \bar{x}_j < \bar{x}_{i'}$ will also belong to this homogenous group (since the variances of the arithmetic means will be equal). However, when sample sizes are not equal, the rule is not valid, which was observed on a few occasions in the above-presented examples. For instance:

- LSD test rejects the equality of means 8 and 10, in spite of the fact it rejects neither the equality of means 7 and 10 nor the equality of means 8 and 6 (see group II in the test).
- Tukey test and Newman-Keuls test reject the equality of means 8 and 4, but do not reject the equality of means 7 and 4 (see group II in these tests).

These cases prove that some homogenous groups are not in fact totally homogenous.

Interpretation of results

If any of the tests had failed to reject hypothesis H: $m_1 = ... = m_{11}$ (the hypothesis that body height means of nine-year-old boys in particular years are equal) then the interpretation of the case would have been simple. We would have stated that there were no reasons to believe that any variable (i.e. mean body height in any year) differed from any other variable.

In the above example, however, the hypothesis was rejected with each test; only homogenous groups were distinguished. The situation interpretation of which would be the easiest is the situation when particular homogenous groups are disjoint groups, for example: group I – 5, 7, 8, 11; group II – 10, 6, 2, 3; group III – 9, 4, 1. Then we would say that boys from the years 1984, 1986, 1987, 1990 on average do not differ with their body height; boys from the years 1989, 1985, 1981, 1982 are on average of the same height; similarly to boys from the years 1988, 1983, 1980, who too are on average of the same height. In turn, each two variables from different groups would have different means. From the logical point of view, this is nothing else but a division of a set (containing 11 elements) into three subsets based on the criterion of the same body height [KMITA 1973].

The homogenous groups we actually obtained (in each of the tests) are not disjoint groups. As a result, we face the following interpretation problem. Let us take a look at

the results of Tukey test. The equality of means m_5 and m_8 as well as of m_8 and m_6 was not rejected with this test, but the equality of means m_5 and m_6 was. We could interpret this fact in the following way: Nine-year-old boys measured in 1984 (Π_5) had on average the same body height as nine-year-old boys from 1987 (Π_8). The latter, in turn, had on average the same body height as nine-year-old boys from the year 1985 (Π_6). However, boys measured in 1984 (Π_5) were shorter than boys from 1985 (Π_6). This means that $m_5 = m_8$ and $m_8 = m_6$, but $m_5 < m_6$, which is in conflict with the fact that the relation of equality is transitional [KMITA 1973]. Thus, the results obtained with the tests performed fail to enable a classic division of the entire set of populations under concern. Such a division involves forming subsets where each two populations from the same subset will have equal means and each two populations from different subsets will have unequal means. The results obtained allow only for a limited recognition of the variation of the value of the studied trait between populations.

To conclude, I would like to point out to the fact that all the tests discussed in this paper, except t-Bonferroni test, are accessible in the program Statistica[1]. It is activated with the „Post-hoc comparison" button in the „ANOVA/MANOVA" module or with the same button in the „Basic statistics" module after choosing the „One-way ANOVA" option. The program gives the level of significance of the difference of the means of any pair of variables. If the level is lower than 0.05, the equality of these means is rejected (at the probability of error of 0.05). The scrollsheet displayed is useful in finding homogenous groups.

References

CIEŚLIK J., SITEK M., 1987, *Wybór najlepszych, ze względu na średnią, populacji normalnych*. Prz. Antr., 53, 35-50

GREŃ J., 1987, *Statystyka matematyczna. Podręcznik programowany*. Warszawa

KMITA J., 1973, *Wstęp do logiki i metodologii nauk*. Warszawa

MILLER R. G., 1966, *Simultaneous statistical inference*. McGraw-Hill, New York

OKTABA W., 1966, *Elementy statystyki matematycznej i metodyka doświadczalnictwa*. Warszawa

OKTABA W., 1971, *Metody statystyki matematycznej w doświadczalnictwie*. Warszawa

SITEK M.,1973, *Testy porównań wielokrotnych*. Listy biometryczne, nr 39-41

STOLINE M. R., 1978, *Tables of the studentized augmented range and applications to problems of multiple comparison*. JASA, 73

ZIELIŃSKI R., 1972, *Tablice statystyczne*. Warszawa

[1] StatSoft, Inc. (1996). STATISTICA for Windows [Computer program manual]. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104, phone: (918) 749-1119, fax: (918) 749-2217, email: info@statsoftinc.com, WEB: http://www.statsoft.com.

Streszczenie

Tematem pracy było zagadnienie równości wartości średnich wybranej cechy w wielu (>2) populacjach. W pierwszej kolejności testuje się hipotezę, że wszystkie średnie populacji są sobie równe. Jeżeli hipoteza ta zostanie odrzucona to można wyróżniać tak zwane grupy jednorodne, a więc grupy zawierające jedynie te spośród wszystkich populacji, które mają równe średnie. Opisano następujące testy służące do weryfikowania hipotez o równości wielu średnich oraz do wyróżniania grup jednorodnych: test Scheffego, test t-Bonferroniego, test Tukeya, test Newmana-Keulsa, test T' (Spjøtvolla-Stoline'a) oraz test najmniejszej istotnej różnicy. Test Scheffego i t-Bonferroniego są wskazane dla zmiennych skorelowanych (tzn. gdy istnieje korelacja między wartościami cechy w różnych populacjach), natomiast pozostałe testy mogą być stosowane tylko w przypadku zmiennych nieskorelowanych.

Część teoretyczna pracy została uzupełniona przykładem. Postawiono hipotezę, że średni wzrost dziesięcioletnich chłopców był taki sam w każdym roku z przedziału 1980–1990. Obliczenia oparto na danych zebranych przez Zakład Biologii Rozwoju Człowieka (UAM, Poznań). Każdy test odrzucił tą hipotezę, w związku z czym postanowiono znaleźć grupy jednorodne. Chociaż każdy test wyróżnił kilka takich grup to w każdym przypadku były to inne grupy, a powody tych rozbieżności zostały wyjaśnione.

Podano również w jaki sposób powyższe testy mogą być użyte w programie statystycznym Statistica firmy StatSoft.