symmetry

MDPI

Article

Logical Contradictions in the One-Way ANOVA and Tukey–Kramer Multiple Comparisons Tests with More Than Two Groups of Observations

Vladimir Gurvich [1] and Mariya Naumova [2,*]

1 Higher School of Economics, National Research University, 101978 Moscow, Russia; vgurvich@hse.ru or vladimir.gurvich@gmail.com
2 Rutgers Business School, Rutgers University, Piscataway, NJ 08854, USA
* Correspondence: mnaumova@business.rutgers.edu

Abstract: We show that the one-way ANOVA and Tukey–Kramer (*TK*) tests agree on any sample with two groups. This result is based on a simple identity connecting the Fisher–Snedecor and studentized probabilistic distributions and is proven without any additional assumptions; in particular, the standard ANOVA assumptions (independence, normality, and homoscedasticity (*INAH*)) are not needed. In contrast, it is known that for a sample with $k > 2$ groups of observations, even under the *INAH* assumptions, with the same significance level α, the above two tests may give opposite results: (i) ANOVA rejects its null hypothesis $H_0^A : \mu_1 = \ldots = \mu_k$, while the *TK* one, $H_0^{TK}(i,j) : \mu_i = \mu_j$, is not rejected for any pair $i, j \in \{1, \ldots, k\}$; (ii) the *TK* test rejects $H_0^{TK}(i,j)$ for a pair (i,j) (with $i \neq j$), while ANOVA does not reject H_0^A. We construct two large infinite pseudo-random families of samples of both types satisfying *INAH*: in case (i) for any $k \geq 3$ and in case (ii) for some larger k. Furthermore, case (ii) ANOVA, being restricted to the pair of groups (i,j), may reject equality $\mu_i = \mu_j$ with the same α. This is an obvious contradiction, since $\mu_1 = \ldots = \mu_k$ implies $\mu_i = \mu_j$ for all $i, j \in \{1, \ldots, k\}$. Such contradictions appear already in the symmetric case for $k = 3$, or in other words, for three groups of d, d, and c observations with sample means $+1, -1$, and 0, respectively. We outline conditions necessary and sufficient for this phenomenon. Similar contradictory examples are constructed for the multivariable linear regression (MLR). However, for these constructions, it seems difficult to verify the Gauss–Markov assumptions, which are standardly required for MLR. Mathematics Subject Classification: 62 Statistics.

Keywords: one-way ANOVA; Tukey–Kramer multiple comparisons test; ANOVA assumptions; independence; normality; homoscedasticity; Fisher–Snedecor distribution; studentized range

check for updates

Citation: Gurvich, V.; Naumova, M. Logical Contradictions in the One-Way ANOVA and Tukey–Kramer Multiple Comparisons Tests with More Than Two Groups of Observations. *Symmetry* **2021**, *13*, 1387. https://doi.org/10.3390/sym13081387

Academic Editor: Serge Lawrencenko

Received: 14 June 2021
Accepted: 11 July 2021
Published: 30 July 2021

1. One-Way ANOVA and Tukey–Kramer Multiple Comparisons Tests

We use standard statistical definitions and notation; the reader can find more details in [1] or [2].

1.1. One-Way ANOVA

Consider an arbitrary sample that consists of k groups of randomly chosen real values. A group $j \in \{1, \ldots, k\}$ contains n_j values $x_{\ell j}$ with $\ell = 1, \ldots, n_j$. Then, $n = n_1 + \ldots + n_k$ is the total number of values in the sample.

Standardly, $\bar{x}_j$ and μ_j denote the *sample* and *population means* for $j = 1, \ldots, k$.

We test

$$H_0^A : \mu_1 = \ldots = \mu_k$$
$$H_1^A : \textit{not all } \mu_i \textit{ are the same}, i = 1, \ldots, k.$$

The *one-way ANOVA* test rejects the null hypothesis H_0^A with significance α, that is, with confidence $100(1-\alpha)\%$, if and only if

$$F_{stat} > F_{crit}(\alpha, k-1, n-k),$$

or equivalently, if the p-value corresponding to F_{stat} is less than α.

Here, $F_{crit}(\alpha, k-1, n-k)$ is the critical value of the Fisher–Snedecor distribution corresponding to the significance level α, with degrees of freedom of the numerator $df_1 = k-1$ and of the denominator $df_2 = n-k$.

The value F_{stat} is given by the ratio

$$F_{stat} = \frac{MS(Tr)}{MSE},$$

where

$$MS(Tr) = \frac{SS(Tr)}{k-1}, \quad SS(Tr) = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=j+1}^{k} n_j n_i (\bar{x}_j - \bar{x}_i)^2,$$

$$\bar{\bar{x}} = \frac{1}{n}\sum_{j=1}^{k}\sum_{\ell=1}^{n_j} x_{\ell j} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j. \tag{1}$$

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

Thus, ANOVA rejects H_0^A if and only if

$$MSE < (n(k-1)F_{crit}(\alpha, k-1, n-k))^{-1}\sum_{j=1}^{k}\sum_{i=j+1}^{k} n_j n_i (\bar{x}_j - \bar{x}_i)^2. \tag{2}$$

1.2. Tukey–Kramer's Test

For each pair $i, j \in \{1, \ldots k\}$ we test the null hypothesis:

$$H_0^{TK}(i,j) : \mu_i = \mu_j$$
$$H_1^{TK}(i,j) : \mu_i \neq \mu_j$$

for all $i \neq j$.

Tukey [3] proposed a procedure for testing these hypotheses in case of equal group sizes $n_1 = \ldots = n_k$. Then, it was extended in [4,5] to arbitrary group sizes. This test, called the Tukey–Kramer (TK) test, uses the studentized range statistic

$$Q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{\frac{MSE}{n}}},$$

where $\bar{y}_{max}$ and $\bar{y}_{min}$ are the largest and the smallest sample means, out of a collection of k sample means.

TK rejects H_0^{TK} if and only if

$$|\bar{x}_i - \bar{x}_j| > CR(\alpha, k, n, i, j), \tag{3}$$

where the *critical range* (CR) is defined by the formula

$$CR(\alpha, k, n; i, j) = Q(\alpha, k, n-k)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \tag{4}$$

and $Q(\alpha; k, n-k)$ is the critical value of the studentized range Q corresponding to the significance level α, with degrees of freedom of numerator $df_1 = k$ and of denominator $df_2 = n - k$.

Equivalently, (3) can be stated as

$$MSE < 2Q^{-2}(\alpha, k, n-k)(\bar{x}_i - \bar{x}_j)^2 \frac{n_i n_j}{n_i + n_j}. \tag{5}$$

1.3. Comparing ANOVA and Tukey–Kramer Tests

Both ANOVA and *TK* tests are based on the following standard assumptions: independence, normality, and homoscedasticity (*INAH*). The last one means that all groups have equal standard deviations, $\sigma_1 = \ldots = \sigma_k$. Both rejection criteria (2) and (5) are based on these assumptions; see [1,2,6] for more details.

In this note, we concentrate on the agreement between the above two tests rather than on their validity. Both inequalities (2) and (5) have the same left-hand side, *MSE*, which can be any number and is irrelevant for the sake of comparison of two tests.

By definition, H_0^A holds if and only if H_0^{TK} holds for all pairs (i, j) with $i \neq j$. When $k = 2$, there is only one such pair, and hence, the ANOVA and *TK* tests should agree, and indeed they are. In Section 2, we prove it for an arbitrary sample. In particular, even if the *INAH* assumptions are not met, still both tests either reject their null hypothesis or both do not, for any fixed significance level α.

However, when $k > 2$, even under *INAH* assumptions, the ANOVA and TK tests may disagree and both cases (i) and (ii) defined in the Abstract may occur. Case (i) is not a paradox. Indeed, if $H_0^{TK}(i, j) : \mu_i = \mu_j$ holds with significance slightly larger than α, then it is not rejected by the *TK* test. This may hold for all pairs $i, j \in \{1, \ldots, k\}$ with $i \neq j$. Yet, the number of these pairs $\frac{k(k-1)}{2}$ is more than 1 when $k > 2$. Thus, ANOVA may reject $H_0^A : \mu_1 = \ldots = \mu_k$ with significance α.

Somewhat surprisingly, the inverse also happens: H_0^A may hold with a fixed α, while H_0^{TK} may be rejected for some pair (i, j) with the same α. Such examples are known. Hsu [7], on page 177, remarks: "An unfortunate common practice is to pursue multiple comparisons only when the null hypothesis of homogeneity is rejected."

We construct two large families of samples of both types considered above. In Section 3.1, we provide two randomly generated samples with three groups in each, $k = 3$, and in Section 3.2, two infinite families of pseudo-random samples with $k \geq 3$ for type (i) and with some larger k for type (ii). It is important to note that these constructions are realized under *INAH* assumptions.

When $k > 2$, Formula (4) appears to be somewhat strange: the critical range is defined for a given pair (i, j) via the value of *MSE* that depends on all observations, in all k groups. These observations are independent random variables; hence, their values in a group ℓ cannot affect the equality $H_0^{TK}(i, j) : \mu_i = \mu_j$ whenever i, j, ℓ are pairwise distinct. Moreover, group ℓ may not relate to groups i and j at all.

1.4. Modified Tukey–Kramer Test

Given significance level α, a sample with $k > 2$ groups, and a pair (i, j) with $i \neq j$, let us modify the *TK* test for $H_0^{TK}(i, j) : \mu_i = \mu_j$ by eliminating all groups but i and j from the sample. Thus, we obtain a new sample with $k' = 2$, $n' = n_i + n_j$, and

$$MSE' = \frac{SSE'}{n-2}, \quad SSE' = \sum_{i'=1}^{n_i} (x_{i'i} - \bar{x}_i)^2 + \sum_{j'=1}^{n_j} (x_{j'j} - \bar{x}_j)^2. \tag{6}$$

Then, we define

$$CR'(\alpha, k, n; i, j) = Q(\alpha, 2, n' - 2)\sqrt{\frac{MSE'}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}. \tag{7}$$

In Section 3.3, assuming homoscedasticity ($\sigma_1 = \ldots = \sigma_k$) and also that $n_1 = \ldots = n_k$ and $\frac{n}{k}$ is large enough, we will show that $CR' \leq CR$. Hence, the modified TK test rejects $H_0^{TK}(i,j) : \mu_i = \mu_j$ whenever the standard TK test does.

Remark 1. *Note that in general, inequality $CR' \leq CR$ may fail since MSE may be much smaller than MSE'. Indeed, if $s_\ell = 0$ (resp., small) for all $\ell \in \{1,\ldots,k\} \setminus \{i,j\}$, while $s_i > 0$ and $s_j > 0$ (resp., large), then $SSE' = SSE > 0$ (resp., $SSE - SSE'$ may be an arbitrarily small non-negative number). Notice however that the homoscedasticity assumption might not hold when s_i and s_j differ significantly. Furthermore, $n-k$ may be much larger than $n'-2$ resulting in $Q(k,n-k)\sqrt{MSE} < Q(2,n'-2)\sqrt{MSE'}$.*

1.5. Counterintuitive Examples with Symmetric Samples of Two and Three Groups

Another infinite set of pseudo-random examples will be constructed in Section 4. Given two groups of observations with $n_1 = n_2 = d$, $\bar{x}_1 = -1, \bar{x}_2 = 1$, $\sigma_1 = \sigma_2 = \sigma$, and a significance level α, we find d, and α such that ANOVA rejects $H_0 : \mu_1 = \mu_2$, for some σ, with confidence $1-\alpha$. Then, we add a third group of observations with $n_3 = c$, $\bar{x}_3 = 0$, $\sigma_3 = \sigma$ and show that $H_0^A : \mu_1 = \mu_2 = \mu_3$, for some σ, is not rejected by ANOVA with the same confidence when $0 < c < d$.

As we already mentioned, this is a logical contradiction. Let us add that condition $c < d$ seems counterintuitive. Indeed, $\bar{x}_3 = 0$, hence group 3, contains values that are typically between $\bar{x}_1 = -1$ and $\bar{x}_2 = 1$, which could be viewed as "an argument" in support of $\mu_1 = \mu_2 = \mu_3$. Furthermore, the larger c is, the stronger is this argument.

1.6. ANOVA Is Not Inclusion Monotone on the Subsets of Its k Groups of Observations $\{1,\ldots,k\}$

Given a significance level α, a sample with $k > 2$ groups, and a pair $i,j \in \{1,\ldots,k\}$ with $i \neq j$, recall case (ii) (the TK test rejects $H_0^{TK}(i,j) : \mu_i = \mu_j$, while ANOVA does not reject $H_0^A : \mu_1 = \ldots = \mu_k$).

Reduce the sample to only two groups i and j, eliminating $k-2$ remaining groups, and apply the ANOVA and modified TK tests to the obtained sample. According to the previous subsection, the latter still rejects the equality $\mu_i = \mu_j$ and, based on Theorem 1, ANOVA also rejects it, while $\mu_1 = \ldots = \mu_k$ is not rejected. This is a contradiction.

1.7. Logical Contradictions in F- and t-Tests of Multivariable Linear Regression (MLR)

The general multivariable linear regression model with k predictors $X^1,\ldots,X^k$ and response Y can be written as

$$Y = \beta_0 + \beta_1 X^1 + \ldots + \beta_k X^k + \epsilon.$$

The properties of the estimators of the coefficients β_i are derived under the Gauss–Markov assumptions (GMA); see, for example, [8].

Commonly used tests in regressions analysis are the F-test

$$H_0^F : \beta_1 = \ldots = \beta_k = 0$$
$$H_1^F : \text{at least one } \beta_i \text{ is not 0, for } i = 1,\ldots,k,$$

and the t-test for individual coefficients β_i for $i \in \{1,\ldots,k\}$ as follows:

$$H_0^{t_i} : \beta_i = 0$$
$$H_1^{t_i} : \beta_i \neq 0.$$

It is well-known that the F- and t-tests are equivalent in the case of simple linear regression (SLR), that is, when $k = 1$. In this case, the p-values of the tests are equal due to identity $F(1,\nu) = t^2(\nu)$ for all natural ν, where $F(1,\nu)$ is a Fisher–Snedecor random variable with $df_1 = 1$ and $df_2 = \nu$, and $t(\nu)$ is a random variable having Student's distribution with the degrees of freedom ν.

Yet, for MLR, $k > 1$, logical contradictions similar to ones outlined in Sections 1.3–1.5 appear. With the same α, the F- and t-tests for MLR may give opposite results as follows:

(j) F-test rejects $H_0^F : \beta_1 = \ldots = \beta_k = 0$, while $H_0^{t_i} : \beta_i = 0$ is rejected for no $i \in \{1, \ldots, k\}$;
(jj) F-test does not reject H_0^F, while t-test rejects $H_0^{t_i}$ for some (or even for all) $i \in \{1, \ldots, k\}$.

Similarly to case (i) of ANOVA, case (j) is not a paradox: $H_0^{t_i}$ cannot be rejected with significance α for each particular i, but it can be rejected with this significance for at least one i. In contrast, case (jj) is an obvious contradiction, since $H_0^F : \beta_1 = \ldots = \beta_k = 0$ implies $H_0^{t_i} : \beta_i = 0$ for every $i \in \{1, \ldots, k\}$.

The corresponding examples are shown in Section 5 for $k = 2$ with the following inequalities for the p-values:

- $p_{12} > p_1$ and $p_{12} > p_2$ in case (j) in Section 5;
- $p_{12} < p_1$ and $p_{12} < p_2$ in case (jj) in Section 5,

where p_{12} is the p-value of the F-test, while p_1 and p_2 are the p-values of the t-tests for β_1 and β_2, respectively.

Similarly to Section 1.5 for ANOVA, we will show that MLR may not be a monotonic inclusion on the set $\{1, \ldots, k\}$ of its predictors. More precisely, consider MLR F-test with k predictors and eliminate $k-1$ of them, all but i, obtaining k SLR problems, one for each predictor X^i and the same response Y, for $i \in \{1, \ldots, k\}$. Denote p'_i the p-value of the SLR test i. (Recall that the F- and t-tests for SLR are equivalent and have equal p-values.)

The example also has the following property: after removing predictor X^1, we obtain $p_{12} > 0.05 > p'_2$. Hence, for significance level $\alpha = 0.05$, the null hypothesis $H_0^F : \beta_1 = \beta_2 = 0$ is not rejected, while p_2 and p'_2 are both less than 0.05. Thus, Case (jj) holds and, furthermore, both predictors X^1 and X^2 are not significant, while X^2 alone is significant.

Let us remark finally that our constructions of Sections 1.3–1.5 satisfy *INAH* assumptions for ANOVA; however, it seems difficult to verify the GMA, which are standardly required for MLR.

2. Two Groups, $k = 2$

In this case, a unique pair $(i, j) = (1, 2)$ of means and multiple comparisons turn into a single one. ANOVA and TK tests' null hypotheses H_0 coincide, stating that $\mu_1 = \mu_2$.

Theorem 1. *In the case of two groups, k = 2, ANOVA and TK tests are equivalent.*

Proof. It is enough to show that inequalities (2) and (5) are equivalent when $k = 2$. In this case, Formulas (2) and (5) can be rewritten as follows:

$$MSE < \frac{n_1 n_2}{n} F_{crit}^{-1}(\alpha, 1, n-2)(\bar{x}_1 - \bar{x}_2)^2,$$

and

$$MSE < \frac{2 n_1 n_2}{n} Q^{-2}(\alpha, 2, n-2)(\bar{x}_1 - \bar{x}_2)^2,$$

where

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

Thus, it is sufficient to prove the identity

$$Q^2(\alpha, 2, n-2) = 2 F_{crit}(\alpha, 1, n-2),$$

which is implied by the following lemma.

Let $F(1,\nu)$ be a Fisher–Snedecor random variable with $df_1 = 1$ and $df_2 = \nu$, and $Q(2,\nu)$ be a random variable having studentized range distribution with the number of groups $k = 2$ and the degrees of freedom ν. □

Lemma 1. *Equation* $2F(1,\nu) = Q^2(2,\nu)$ *holds.*

Proof. The probability density function of studentized range Q in case $k = 2$ is given by

$$\begin{aligned} f_Q(q;2,\nu) &= \frac{4\sqrt{2\pi}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty s^\nu \phi(\sqrt{\nu}s)\left[\int_\infty^\infty \phi(z+qs)\phi(z)dz\right]ds \\ &= \frac{4\sqrt{2\pi}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty s^\nu \frac{1}{\sqrt{2\pi}}e^{-\frac{\nu s^2}{2}}\left[\int_\infty^\infty \frac{1}{2\pi}e^{-\frac{(z+qs)^2+z^2}{2}}dz\right]ds; \end{aligned}$$

see [3]. We transform this formula as follows: substitute $u = \sqrt{2}z$ to obtain

$$\begin{aligned} f_Q(q;2,\nu) &= \frac{2\sqrt{2}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty s^\nu \frac{1}{\sqrt{2\pi}}e^{-\frac{\nu s^2}{2}}\left[\int_\infty^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{\left(u+\frac{\sqrt{2}qs}{2}\right)^2+\frac{q^2s^2}{2}}{2}}du\right]ds \\ &= \frac{2\sqrt{2}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty s^\nu e^{-\frac{\left(\nu+\frac{q^2}{2}\right)s^2}{2}}ds. \end{aligned}$$

Then, by substituting $t = s^2$,

$$f_Q(q;2,\nu) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty t^{\frac{\nu-1}{2}}e^{-\frac{\left(\nu+\frac{q^2}{2}\right)t}{2}}dt,$$

and by substituting $y = \frac{\left(\nu+\frac{q^2}{2}\right)t}{2}$,

$$\begin{aligned} f_Q(q;2,\nu) &= \frac{\sqrt{2}\nu^{\frac{\nu}{2}}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\left(\nu+\frac{q^2}{2}\right)^{\frac{\nu+1}{2}}}\int_0^\infty y^{\frac{\nu+1}{2}-1}e^{-y}dy \\ &= \frac{\sqrt{2}\nu^{\frac{\nu}{2}}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\left(\nu+\frac{q^2}{2}\right)^{\frac{\nu+1}{2}}}\Gamma\left(\frac{\nu+1}{2}\right). \end{aligned} \tag{8}$$

For $X = \frac{Q^2}{2}$, we obtain

$$P[X \le x] = P\left[\frac{Q^2}{2} \le x\right] = P\left[Q \le \sqrt{2x}\right],$$

and then

$$f_X(x) = \frac{d}{dx}P\left[Q \le \sqrt{2x}\right] = \frac{1}{\sqrt{2x}}f_Q(\sqrt{2x};2,\nu),$$

which, based on (8), implies that

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2x}}\frac{\sqrt{2}\nu^{\frac{\nu}{2}}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)(\nu+x)^{\frac{\nu+1}{2}}}\Gamma\left(\frac{\nu+1}{2}\right) \\ &= \frac{1}{B\left(\frac{1}{2},\frac{\nu}{2}\right)}(\nu x)^{-\frac{1}{2}}\left(1+\frac{x}{\nu}\right)^{-\frac{\nu+1}{2}}, \end{aligned} \tag{9}$$

where $B(a,b)$ is the beta function.

It is well known (see, for example, [9]) that (9) defines the probability density function of the Fisher–Snedecor distribution with degrees of freedom of the numerator $df_1 = 1$ and of the denominator $df_2 = \nu$. $\square$

This proves the Theorem.

Note that Theorem 1 holds for an arbitrary sample. In particular, the p-values for ANOVA and *TK* tests are equal regardless of the validity of assumptions *INAH*.

3. Some Cases When ANOVA and *TK* Tests Disagree

In this section, we provide several examples where the considered two tests disagree: (i) H_0^A is rejected, while H_0^{TK} is not or (ii) vice versa. In Section 3.1, we provide two randomly generated samples illustrating (i) and (ii) with three groups in each, $k = 3$; and in Section 3.2, we construct two infinite families of pseudo-random samples with $k \geq 3$ for (i) and with some larger k for (ii).

3.1. Two Examples with Three Groups

Using R, we generated two random samples with $k = 3$ groups, $n_1 = n_2 = n_3 = 10$, from normal distributions with parameters $\mu_1 = 10$, $\mu_2 = 25$, $\mu_3 = 40$, and $\sigma_1 = \sigma_2 = \sigma_3 = 25$. (see Table 1)

3.1.1. Case 1: ANOVA Rejects H_0^A, While *TK* Does Not Reject H_0^{TK}

Let us fix the significance level $\alpha = 0.05$; then, *TK* does not distinguish any pair μ_i and μ_j (see Table 2), while Table 3 shows that the ANOVA test rejects the null hypothesis H_0^A.

Table 1. Generated random sample for Case 1.

Group 1	Group 2	Group 3
33.73617429	41.34327861	1.949654854
6.532109599	−2.29596015	64.73534452
−15.87068125	37.80911436	17.47791461
24.41853292	−38.5488504	43.91077426
32.52469512	28.81447508	15.70006485
−36.67775074	91.99464773	54.51355702
4.946144821	54.96895462	31.54941908
−11.48789077	77.16877	0.819316511
32.04750431	95.1318948	73.84330239
−24.02530782	6.722405014	45.92357859

Table 2. The results of *TK* test for the example in Case 1.

Group	Diff	lwr	upr	p adj
Group2-Group1	34.696519918	−1.294542536	70.68758237	0.06045
Group3-Group1	30.427939620	−5.563122834	66.41900207	0.10952
Group3-Group2	−4.268580298	−40.259642752	31.72248216	0.95353

Table 3. ANOVA table for the example in Case 1.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
group	2	7159.763	3579.881	3.39789	0.04828
Residuals	27	28,446.164	1053.562		

3.1.2. Case 2: ANOVA Does Not Reject H_0^A, While *TK* Rejects H_0^{TK}

Let us fix again the significance level $\alpha = 0.05$. The generated random sample for Case 2 is shown in Table 4. Then, Table 5 shows that *TK* distinguishes μ_1 and μ_3 at significance level $\alpha = 0.05$. In contrast, Table 6 shows that the ANOVA test does not reject H_0^A for the same α. In this case, we can apply the approach suggested in Sections 1.4 and 1.5. Let us reduce the sample by eliminating group 2 and apply the ANOVA and (modified) *TK* test.

Table 4. Generated random sample for Case 2.

Group 1	Group 2	Group 3
19.65656273	30.47282693	97.66594506
31.63471018	2.359493274	37.29706457
5.474716521	25.94822801	37.28238885
7.325738946	−6.706730014	−9.215515132
47.16633	56.00337827	44.75306142
−28.99487682	22.37945513	72.60365833
14.99564807	73.81358543	21.39501942
48.12035772	5.44699726	71.5651277
25.54178184	−3.745973145	63.33149261
−16.61305101	48.61987107	26.01262136

Table 5. The results of *TK* test for the example in Case 2.

Group	Diff	lwr	upr	p adj
Group2-Group1	−10.0283214	−40.8231285393	20.76648573	0.70172
Group3-Group1	30.8382946	0.0434874678	61.63310174	0.04962
Group3-Group2	20.8099732	−9.9848339369	51.60478033	0.23279

Table 6. ANOVA table for the example in Case 2.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
group	2	4948.742	2474.371	3.20804	0.05624
Residuals	27	20,825.208	771.304		

According to Theorem 1, these two tests are equivalent: p-value is 0.02432 (see Tables 7 and 8) for the equality $\mu_1 = \mu_2$ in both cases. Yet, for the original sample of three groups, the p-value was 0.05624 for the equality $\mu_1 = \mu_2 = \mu_3$. This is an obvious contradiction: ANOVA rejects $\mu_1 = \mu_2$ with confidence 97.5% but cannot reject the stronger statement $\mu_1 = \mu_2 = \mu_3$ (which is easier to do) even with confidence 95%.

Recall that this example was generated by R under *INAH* assumptions. This did not take too many trials: with given parameters $k = 3$, $n_1 = n_2 = n_3 = 10$, $\mu_1 = 10$, $\mu_2 = 25, \mu_3 = 40$, and $\sigma_1 = \sigma_2 = \sigma_3 = 25$, about each 20 trials provide an example with such properties.

Table 7. ANOVA table for groups 1 and 3 in Case 2.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
group	1	4755.002	4755.002	6.044	0.02432
Residuals	18	14,161.164	786.731		

Table 8. The results of modified *TK* test for groups 1 and 3 in Case 2.

Group	Diff	lwr	upr	p adj
Group3-Group1	30.8382946	4.484804399	57.1917848	0.02431

3.2. Two Large Families of Examples with K Groups

In this subsection, we consider samples with k groups such that n is divisible by k, and

$$n_1 = \ldots = n_k = \frac{n}{k}, \tag{10}$$

$$s_1 = \ldots = s_k = s, \tag{11}$$

where s_i is the standard deviation of the ith group. In this case, we have

$$SSE = k\left(\frac{n}{k} - 1\right)s^2, \quad MSE = \frac{SSE}{n-k} = s^2. \tag{12}$$

3.2.1. Case 1: ANOVA Rejects H_0^A, While TK Does Not Reject H_0^{TK}

Based on (2) and (5), this happens if and only if

$$\begin{aligned} &(n(k-1)F_{crit}(\alpha, k-1, n-k))^{-1} \sum_{j=1}^{k} \sum_{i=j+1}^{k} n_j n_i (\bar{x}_j - \bar{x}_i)^2 \\ &> MSE > 2Q^{-2}(\alpha, k, n-k)(\bar{x}_i - \bar{x}_j)^2 \frac{n_i n_j}{n_i + n_j}, \end{aligned} \tag{13}$$

which implies

$$\begin{aligned} &Q^2(\alpha, k, n-k) \sum_{j=1}^{k} \sum_{i=j+1}^{k} n_j n_i (\bar{x}_j - \bar{x}_i)^2 \\ &> 2n(k-1)F_{crit}(\alpha, k-1, n-k)(\bar{x}_i - \bar{x}_j)^2 \frac{n_i n_j}{n_i + n_j}. \end{aligned} \tag{14}$$

Consider any sample with k groups, k is even, satisfyng (10) and (11), and

$$\bar{x}_1 = \ldots = \bar{x}_{\frac{k}{2}} = 1, \bar{x}_{\frac{k}{2}+1} = \ldots = \bar{x}_k = 0. \tag{15}$$

Based on (12), $MSE = s^2$, and (14) can be rewritten as

$$\left(\frac{n}{k}\right)^2 \frac{k^2}{4} Q^2(\alpha, k, n-k) > 2\frac{n^3}{k^2(2\frac{n}{k})} F_{crit}(\alpha, k-1, n-k),$$

which can be simplified to

$$G(\alpha, k, n-k) = \frac{Q^2(\alpha, k, n-k)}{F_{crit}(\alpha, k-1, n-k)} - 4\left(1 - \frac{1}{k}\right) > 0. \tag{16}$$

Function $G(\alpha, k, n-k)$ has the following properties:

1. $G(\alpha, k, n-k) \equiv 0$ if $k = 2$;
2. $G(\alpha, k, n-k)$ is monotonically increasing with respect to $n-k$ and converging as $n \to \infty$ for each k;
3. $G(\alpha, k, n-k) > 0$ for $k \geq 3$ and all $n-k \geq 0$ for $\alpha = 0.005, 0.01, 0.025, 0.05, 0.1, 0.25$, and 0.5.

It is not our goal to study function $G(\alpha, k, n-k)$ in detail; we are primarily interested only in its positivity, required by condition (16). The required inequality (16) holds for any $k \geq 3$.

Given an even k and n divisible by k, we generate a desired pseudo-random sample as follows: It satisfies (10), (11), (15), and in addition, whenever (16) holds, we still can choose $s^2 = MSE$ satisfying (13).

3.2.2. Case 2: ANOVA Does Not Reject H_0^A, While TK Rejects H_0^{TK}

Based on (2) and (5), this happens if and only if

$$\begin{aligned}&(n(k-1)F_{crit}(\alpha,k-1,n-k))^{-1}\sum_{j=1}^{k}\sum_{i=j+1}^{k}n_jn_i(\bar{x}_j-\bar{x}_i)^2\\&< MSE < 2Q^{-2}(\alpha,k,n-k)(\bar{x}_i-\bar{x}_j)^2\frac{n_in_j}{n_i+n_j},\end{aligned} \tag{17}$$

which implies

$$\begin{aligned}&Q^2(\alpha,k,n-k)\sum_{j=1}^{k}\sum_{i=j+1}^{k}n_jn_i(\bar{x}_j-\bar{x}_i)^2\\&< 2n(k-1)F_{crit}(\alpha,k-1,n-k)(\bar{x}_i-\bar{x}_j)^2\frac{n_in_j}{n_i+n_j}.\end{aligned} \tag{18}$$

Note that if $k=2$, we obtain (8).

Consider any sample with k groups satisfying (10) and (11), and

$$\bar{x}_1 = 1, \bar{x}_2 = \ldots = \bar{x}_k = 0. \tag{19}$$

Then, (18) turns into

$$\left(\frac{n}{k}\right)^2(k-1)Q^2(\alpha,k,n-k) < \frac{2n(k-1)\left(\frac{n}{k}\right)^2}{2\frac{n}{k}}F_{crit}(\alpha,k-1,n-k),$$

which simplifies to

$$H(\alpha,k,n-k) = \frac{Q^2(\alpha,k,n-k)}{F_{crit}(\alpha,k-1,n-k)} - k < 0. \tag{20}$$

Since s can be chosen arbitrarily, we can always find MSE satisfying (17) whenever (18) holds.

Function $H(\alpha,k,n-k)$ shares properties (1) and (2) of $G(\alpha,k,n-k)$; furthermore, $H(\alpha,k,n-k) > 0$ for sufficiently small k, and $H(\alpha,k,n-k) < 0$ for sufficiently large k. Again, it is not our goal to study $H(\alpha,k,n-k)$ in detail since we are primarily interested only in its negativity required by condition (20).

The signs of $H(\alpha,k,n-k)$ depending on k are shown in Table 9 for some values of α. The second (resp., third) column contains the values of k such that $H(\alpha,k,n-k) > 0$ (resp., $H(\alpha,k,n-k) < 0$) for all n. Missing values of k correspond to the cases when the sign of $H(\alpha,k,n-k)$ depends on n.

One can see that the required inequality (20) holds when the number of groups k is large enough.

Given k and n divisible by k, we generate the desired pseudo-random sample as follows: It satisfies (10), (11), (19), and in addition, whenever (20) holds, we still can choose $s^2 = MSE$ satisfying (17).

Remark 2. *We variate the choice of sample means in* (15) *and* (19) *to increase the feasible area for* (16) *and* (20)*, respectively. Obviously,* $k \geq 4\left(1-\frac{1}{k}\right)$ *and equality hold if and only if* $k=2$.

Remark 3. *We can extend considerably the family of the constructed examples by relaxing equalities* (11), (15)*, and* (19)*, and replacing them with approximate equalities.*

Table 9. The signs of $H(\alpha, k, n-k)$ for selected α depending on k.

$\boldsymbol{\alpha}$	$\boldsymbol{H(\alpha, k, n-k) > 0}$	$\boldsymbol{H(\alpha, k, n-k) < 0}$
0.005	$3 \leq k \leq 10$	$k \geq 14$
0.01	$3 \leq k \leq 10$	$k \geq 14$
0.025	$3 \leq k \leq 10$	$k \geq 13$
0.05	$3 \leq k \leq 10$	$k \geq 12$
0.1	$3 \leq k \leq 10$	$k \geq 12$
0.25	$3 \leq k \leq 10$	$k \geq 11$
0.5	$3 \leq k \leq 9$	$k \geq 10$

3.3. Critical Range in Modified TK Test

In Section 1.4, we modified the standard *TK* multiple comparisons test, replacing it with the pairwise comparison version as follows: Given significance level α, a sample with $k > 2$ groups, and a pair $(i, j) \in \{1, \ldots, k\}$ with $i \neq j$, consider the null hypothesis for the corresponding two groups, $H_0^{TK}(i, j) : \mu_i = \mu_j$ and eliminate all groups but i and j from the sample, obtaining a new one with $k' = 2,\ n' = n_i + n_j$. For the standard and modified *TK* tests, the critical ranges $CR = CR(\alpha, k, n-k; i, j)$ and $CR' = CR'(\alpha, k, n-k; i, j)$ and the corresponding values of MSE and MSE' are given by Formulas (1), (4), (6) and (7).

We are looking for conditions implying the inequality $CR' \leq CR$, in which case the modified *TK* test rejects $H_0^{TK}(i, j)$ whenever the standard one does. In general, this inequality may fail; see Remark 1.

Let us assume *INAH*, and in addition (10) and (11). As we know, in this case, $MSE = MSE' = s^2$ and formulas for CR and CR' are simplified as follows:

$$\begin{aligned} CR &= Q(\alpha, k, n-k)s\sqrt{\frac{k}{n}}, \\ CR' &= Q(\alpha, 2, n'-k')s\sqrt{\frac{k'}{n'}} \\ &= Q\left(\alpha, 2, 2\frac{n}{k} - 2\right)s\sqrt{\frac{2}{2n/k}} = Q\left(\alpha, \frac{k}{k/2}, \frac{n-k}{k/2}\right)s\sqrt{\frac{k}{n}}. \end{aligned}$$

Thus, in the considered case,

$$\frac{CR'}{CR} = \frac{Q(\alpha, 2, \nu)}{Q(\alpha, 2\ell, \nu\ell)},$$

where $\ell = \frac{k}{2} \geq 1$ and $\nu = n' - k' = 2\left(\frac{n}{k} - 1\right)$.

The critical value of the studentized range $Q(\alpha, 2\ell, \nu\ell)$ monotonically increases with ℓ when $\nu = 2\left(\frac{n}{k} - 1\right)$ is large enough; see Table 10.

Table 10. Conditions for monotonic increasing of the studentized range $Q(\alpha, 2\ell, \nu\ell)$.

$\boldsymbol{\alpha}$	**0.005**	**0.01**	**0.025**	**0.05**	**0.1**	**0.25**	**0.5**
	$\nu \geq 7$	$\nu \geq 5$	$\nu \geq 4$	$\nu \geq 3$	$\nu \geq 2$	$\nu \geq 1$	$\nu \geq 1$

In these cases, $CR' \leq CR$, and hence, conclusions of Section 1.5 are applicable. Recall the construction of Section 3.2 Case 2, in which ANOVA does not reject $H_0^A : \mu_1 = \ldots = \mu_k$, while the standard *TK* test rejects $H_0^{TK}(i, j) : \mu_i = \mu_j$. This pseudo-random construction satisfies *INAH*. Let us apply ANOVA and *TK* tests to the reduced sample that consists of only two groups i and j, with the remaining $k-2$ groups eliminated. Based on the above arguments, the modified TK test still rejects its hypothesis $H_0^{TK}(i,j) : \mu_i = \mu_j$ and, based on Theorem 1, ANOVA rejects it as well. However, ANOVA does not reject a stronger hypothesis $H_0^A : \mu_1 = \ldots = \mu_k$, with the same significance level α, which is an obvious contradiction.

4. Symmetric Samples with Two and Three Groups

4.1. Two Groups

Consider two groups 1 and 2 with d observations in each, that is, $k = 2,\ n_1 = n_2 = d,\ n = n_1 + n_2 = 2d$, with means $\bar{x}_1 = -1,\ \bar{x}_2 = -1$, and standard deviations $\sigma_1 = \sigma_2 = \sigma$. We can assume that $INAH$ holds.

Obviously, $SS(Tr) = MSR = 2d$; furthermore, according to (1),

$$SSE = \sigma^2,\ MSE = SSE/(n-k) = \sigma^2/(2d-2),$$

$$F_{stat} = MSR/MSE = 4d(d-1)\sigma^{-2}.$$

According to (2), ANOVA rejects its null hypothesis $H_0^A : \mu_1 = \mu_2$ if and only if

$$4d(d-1)\sigma^{-2} > F_{crit}(\alpha, k-1, n-k) = F_{crit}(\alpha, 1, 2d-2) \tag{21}$$

As for the TK (in this case just Tukey test), we have $|\bar{x}_1 - \bar{x}_2| = 2$, and based on (4), the critical range is given by formula

$$CR(1,2) = Q(\alpha, 2, 2d-2)\sqrt{MSE/d} = Q(\alpha, 2, 2d-2)\sigma/\sqrt{2d(d-1)}.$$

Thus, according to (3), the Tukey test rejects $\mu_1 = \mu_2$ if and only if

$$8d(d-1)\sigma^{-2} > Q^2(\alpha, 2, 2d-2). \tag{22}$$

Criteria (21) and (22) are equivalent, based on Lemma 1.

4.2. Three Groups

Let us add to groups 1 and 2 one more group 3 of c observations, obtaining $k = 3$ and $n = n_1 + n_2 + n_3 = 2d + c$. Furthermore, set $\bar{x}_3 = 0$ and $\sigma_3 = \sigma$ and assume that $INAH$ holds. Then, based on (1) and (2),

$$SS(Tr) = 2d,\ MSR = SSR/(k-1) = d;$$

$$SSE = \sigma^2,\ MSE = SSE/(n-k) = \sigma^2/(2d+c-3),$$

$$F_{stat} = MSR/MSE = d(2d+c-3)\sigma^{-2}.$$

Thus, ANOVA rejects its null hypothesis $H_0^A : \mu_1 = \mu_2 = \mu_3$ if and only if

$$d(2d+c-3)\sigma^{-2} > F_{crit}(\alpha, k-1, n-k) = F_{crit}(\alpha, 2, 2d+c-3). \tag{23}$$

As for the TK test, we have $|\bar{x}_1 - \bar{x}_2| = 2$, while based on (4), the critical range

$$CR(1,2) = Q(\alpha, k, n-k)\sqrt{MSE/d} = \sigma Q(\alpha, 3, 2d+c-3)/\sqrt{d(2d+c-3)}.$$

Thus, based on (3), the TK test rejects $\mu_1 = \mu_2$ if and only if

$$4d(2d+c-3)\sigma^{-2} > Q^2(\alpha, 3, 2d+c-3). \tag{24}$$

4.3. ANOVA for $k = 2$ and $k = 3$

ANOVA rejects $\mu_1 = \mu_2$ and does not reject $\mu_1 = \mu_2 = \mu_3$ if and only if (21) holds while (23) fails. It happens, for some σ, if and only if

$$\frac{F_{crit}(\alpha, 1, 2a)}{F_{crit}(\alpha, 2, 2a + b)} < 2\frac{2a}{2a + b}, \tag{25}$$

where $a = d - 1,\ b = c - 1$.

Remark 4. *Obviously, the set of feasible σ is an interval. In fact, this holds for all our examples showing logical contradictions. Such examples are relatively rare because the interval for σ is small.*

Consider $\alpha = 0.05$. Inequality (25) holds whenever $b < a$, or equivalently, $c < d$. It seems that (25) can be solved, with respect to a and b, explicitly. Consider the following sequence of positive integers $S = (6, A^2, B^4, A, B^5, (A, B^6)^\infty)$, where $A = (8, 9, 8, 8, 9)$, $B = (8, 9, 8, 8, 9, 8, 8, 9)$; a power denotes the number of repetitions. Thus, S is a quasi-periodic sequence with the period (A, B^6) of length $5 + 8 \cdot 6 = 53$. To each a we assign a nonnegative integer $a(s)$ uniquely defined by the inequalities

$$\sum_{i=1}^{a(s)} s_i \le a < \sum_{i=1}^{a(s)+1} s_i.$$

Then, (25) holds if and only if $b < a + a(s)$. This criterion is confirmed by computations for $a \le 500$. We conjecture that it holds for all a and that similar criteria hold for all α.

4.4. TK Test for $k = 2$ and $k = 3$

TK rejects equality $\mu_1 = \mu_2$ for $k = 2$ and does not reject it for $k = 3$ if and only if (22) holds while (24) fails. It happens, for some σ, if and only if

$$\frac{Q^2(\alpha, 3, 2a + b)}{Q^2(\alpha, 2, 2a)} < \frac{2a + b}{2a}, \tag{26}$$

Consider $\alpha = 0.05$. Inequality (26) holds whenever $b \ge a$, or equivalently, $c \ge d$. Again, it seems that (26) can be solved, with respect to a and b, explicitly. Consider sequence of positive integers $S = (3, 7, (8, 7^7, 8, 7^6)^\infty)$. It is quasi-periodic with the period $(8, 7^7, 8, 7^6)$ of length 15.

Then, (26) holds if and only if $b \ge a - a(s)$. This criterion is confirmed by computations for $a \le 500$. We conjecture that it holds for all a and that similar criteria hold for all α.

5. Logical Contradictions in Multivariable Linear Regression

Here, we provide examples presented in Section 1.7, see Tables 11–18.

5.1. Construction for Case (j): F-Test Rejects $H_0^F : \beta_1 = \beta_2 = 0$, While $H_0^{t_1} : \beta_1 = 0$, $H_0^{t_2} : \beta_2 = 0$ Are Not Rejected by t-Tests with the Same Significance $\alpha = 0.05$

Note that $p_1 = 0.05598 > 0.05$, $p_2 = 0.28837 > 0.05$, $p_{12} = 0.04865 < 0.05$. Hence, Case (*j*) holds.

Table 11. Generated random sample for Case (j).

X1	X2	Y
1.713673333	0.891652019	1.718488057
0.932830925	0.353231823	1.311861467
−0.053673724	1.132586717	1.903344806
1.055482137	0.248411619	1.582305067
−0.248355435	−0.174256727	2.607296494
−0.004449867	0.115550588	2.352411276
0.086988258	−0.833496007	2.558602277
0.687284914	−0.417171685	1.721811264
−0.253474712	0.045371123	1.982673543
0.135747949	−0.145817805	2.309533234

Table 12. Regression output for the sample in Table 11.

Coefficients	Estimate	Std. Error	t Value	$Pr(>\|t\|)$
(Intercept)	2.193398826	0.12112005	18.10929596	3.87188×10^{-7}
X1	−0.397578417	0.173782994	−2.28778667	0.05599
X2	−0.225853392	0.196597153	−1.148813135	0.28837
Residual standard error: 0.321371159 on 7 degrees of freedom				
Multiple R-squared: 0.578426902, Adjusted R-squared: 0.457977446				
F-statistic: 4.802237548 on 2 and 7 DF, p-value: 0.04865				

5.2. Constructions for Case (jj): F-Test Does Not Reject H_0^F, While t-Tests Reject $H_0^{t_1}$ or $H_0^{t_2}$ with the Same Significance

Here $p_1 = 0.08480$, $p_2 = 0.04959$, and $p_{12} = 0.12690$. Hence, F-test does not reject H_0^F, while t-tests reject $H_0^{t_1}$ and $H_0^{t_2}$ with significance $\alpha = 0.1$.

The next example also illustrates case (jj) and, in addition, shows that F-test can be as not monotone inclusion on the set of predictors.

Table 13. Generated random sample for Case (jj).

X1	X2	Y
1.173699045	1.507797593	1.693611518
1.527866588	1.204880159	1.719565524
−0.237756887	0.321525784	2.313343543
0.424876707	0.372472796	2.215619921
0.155008273	−0.382097849	1.752313506
0.078297635	0.202406996	1.985018225
−0.739378749	−1.77490523	1.280511608
−0.325947264	−0.170739193	1.751709441
0.057639294	0.025498039	2.285726127
0.317517151	0.439566564	1.809984615

Table 14. Regression output for the sample in Table 13.

Coefficients	Estimate	Std. Error	t Value	$Pr(>\|t\|)$
(Intercept)	1.927174047	0.094242869	20.44901714	1.67739×10^{-7}
X1	−0.534309107	0.266285323	−2.006528569	0.08480
X2	0.478129512	0.20172263	2.37023239	0.04959
Residual standard error: 0.271926659 on 7 degrees of freedom				
Multiple R-squared: 0.445574672, Adjusted R-squared: 0.287167435560639				
F-statistic: 2.812842909 on 2 and 7 DF, p-value: 0.12690				

Table 15. Another generated random sample for Case (jj).

X1	X2	Y
1.568562319	0.927834903	1.612462698
1.48286001	1.09773946	2.466033052
−0.573115658	0.981537183	2.518881417
−0.050008016	−0.49329821	1.301806858
0.165268254	−0.397500853	1.436310825
−0.306203404	−0.193130393	2.072432714
−0.399941489	−0.096035236	1.573998771
0.21069356	0.603984432	1.827021014
0.431810105	−0.383312909	1.933014077
0.080628207	0.231611299	2.002964703

Table 16. Regression output for the sample in Table 15.

Coefficients	Estimate	Std. Error	t Value	$Pr(>\|t\|)$
(Intercept)	1.8040579	0.1149434	15.69519	1.0317×10^{-6}
X1	−0.1814855	0.1720305	−1.05496	0.32649
X2	0.5168505	0.2013693	2.56668	0.03719
Residual standard error: 0.3324189 on 7 degrees of freedom				
Multiple R-squared: 0.486189, Adjusted R-squared: 0.3393858				
F-statistic: 3.311843 on 2 and 7 DF, p-value: 0.09723				

Note that $p_2 = 0.03719 < 0.05 < 0.09723 = p_{12}$. Hence, Case (jj) holds. Furthermore, eliminating predictor $X1$ yields the following SLR table:

Table 17. Regression output for the generated random sample for Case (jj) with independent variable $X1$ only.

Coefficients	Estimate	Std. Error	t Value	$Pr(>\|t\|)$
(Intercept)	1.86701480	0.14635023	12.75717	1.3432×10^{-6}
X1	0.02864453	0.19718444	0.14527	0.88809
Residual standard error: 0.4332275 on 8 degrees of freedom				
Multiple R-squared: 0.002631, Adjusted R-squared: −0.122040				
F-statistic: 0.0211027 on 1 and 8 DF, p-value: 0.8880929				

Table 18. Regression output for the sample in Table 15 with independent variable $X2$ only.

Coefficients	Estimate	Std. Error	t Value	$Pr(>\|t\|)$
(Intercept)	1.7797247	0.1133974	15.69458	2.7117×10^{-7}
X2	0.4157529	0.1783505	2.33110	0.04808
Residual standard error: 0.3347572 on 8 degrees of freedom				
Multiple R-squared: 0.404497, Adjusted R-squared: 0.330059				
F-statistic: 5.434026 on 1 and 8 DF, p-value: 0.04808				

We observe again that $p'_2 = 0.04808 < 0.05 < 0.09723 = p_{12}$. Thus, F test states that both predictors $X1$ and $X2$ are insignificant, while $X2$ alone is significant at $\alpha = 0.05$.

6. Concluding Remarks

Both ANOVA and TK multiple comparisons tests with k groups may result in logical contradictions when $k > 2$, even if $INAH$ assumptions hold. Therefore, the good old approach of using pairwise comparisons instead of multiple ones is a bit slower but more reliable. Furthermore, all contradictions disappear if we replace the ANOVA and TK tests by their pairwise versions, applying them for any pair of groups $i, j \in \{1, \cdots, k\}$ with $i \neq j$. Then, according to Theorem 1, these two tests become equivalent.

Similar contradictions appear for the linear regression with the number of predictors $k > 1$ (MLR). Already for $k = 2$, with the same level of significance α, it may happen that t-test rejects $H_0^{t_1} : \beta_1 = 0$, while F-test fails to reject the stronger null hypothesis $H_0^F : \beta_1 = \beta_2 = 0$.

In general, estimating the quality of a prediction made by ANOVA or MLR seems much more doubtful than the prediction itself.

Author Contributions: Data curation, M.N.; Supervision, V.G. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This paper was prepared within the framework of the HSE University Basic Research Program.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller, I.; Miller, M.; John, E. *Freund's Mathematical Statistics*, 6th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2004.
2. Montgomery, D.C. *Design and Analysis of Experiments*, 9th ed.; John Wiley: New York, NY, USA, 2017.
3. Tukey, J. Comparing Individual Means in the Analysis of Variance. *Biometrics* **1949**, *5*, 99–114. [CrossRef] [PubMed]
4. Kramer, C.Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **1956**, *12*, 307–310. [CrossRef]
5. Kramer, C.Y. Extension of multiple range tests to group correlated adjusted means. *Biometrics* **1957**, *13*, 13–18. [CrossRef]
6. Hayter, A.J. A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. *Ann. Stat.* **1984**, *12*, 61–75. [CrossRef]
7. Hsu, J.C. *Multiple Comparisons: Theory and Methods*; Chapman & Hall: London, UK, 1996.
8. Wooldridge, J.M. *Introductory Econometrics: A Modern Approach*, 5th ed.; Thomson/South-Western: Mason, OH, USA, 2012.
9. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics*; Springer Series in Statistics (Perspectives in Statistics); Kotz, S., Johnson, N.L., Eds.; Springer: New York, NY, USA, 1992.