

# Modification of Tukey's Additivity Test

Petr Šimeček, Marie Šimečková

*Institute of Animal Science, Přátelství 815, 10400 Prague, Czech Republic*

---

## Abstract

In this paper we discuss testing for an interaction in the two-way ANOVA with just one observation per cell. The known results are reviewed and a simulation study is performed to evaluate type I and type II risks of the tests. It is shown that the Tukey and Mandel additivity tests have very low power in case of more general interaction scheme. A modification of Tukey's test is developed to resolve this issue. All tests mentioned in the paper have been implemented in R package `AdditivityTests`.

### *Key words:*

two-way ANOVA, additivity tests, Tukey additivity test

---

## 1 Introduction

In many applications of statistical methods, it is assumed that the response variable is a sum of several factors and a random noise. In a real world this may not be an appropriate model. For example, some patients may react differently to the same drug treatment or the influence of fertilizer may be influenced by the type of a soil. There might exist an interaction between factors. A testing for such interaction will be referred here as **testing of additivity hypothesis**.

If there is more than one observation per cell then standard ANOVA techniques may be applied. Unfortunately, in many cases it is infeasible to get more than one observation taken under the same conditions. For instance, it is not logical to ask the same student the same question twice.

---

\* This research was supported by the grants MZE 0002701404 and NAZV QH81312. We are indebted to Dieter Rasch and an anonymous reviewer for their comments to the text.

*Email address:* `simecek@gmail.com`, `simeckova.marie@vuzv.cz` (Petr Šimeček, Marie Šimečková).

We restrict ourselves to a case of two factors, i.e. two-array model, when the response in  $i^{th}$  row and  $j^{th}$  column is modeled as

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad (1)$$

where

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

and the  $\epsilon_{ij}$  are normally distributed independent random variables with zero mean and variance  $\sigma^2$ .

To test the additivity hypothesis

$$H_0: \gamma_{ij} = 0 \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad (2)$$

a number of tests have been developed. The Section 2 recollects the known additivity tests, see also Alin and Kurt (2006) and Boik (1993a).

In Section 3 the power of the tests described in Section 2 is compared by means of simulation. While Tukey test has relatively good power when the interaction is a product of the main effects, i.e. when  $\gamma_{ij} = k\alpha_i\beta_j$  ( $k$  is a real constant), its power for more general interaction is very poor.

It should be reminded that Tukey (1949) did not originally propose his test for any particular type of interaction. Actually after a small modification derived in Section 4 the power of the test improves dramatically. There exist some issues when a sample size is not large enough that may be resolve by a permutation test or bootstrap.

## 2 Additivity Tests

This section recalls the known additivity tests of hypothesis (2) in model (1). Let  $\bar{y}_{..}$  denotes the overall mean,  $\bar{y}_{i.}$  the  $i^{th}$  row's mean and  $\bar{y}_{.j}$  the  $j^{th}$  column's mean. The matrix  $R = [r_{ij}]$  will stand for a residual matrix with respect to the main effects

$$r_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$$

The decreasingly ordered list of eigenvalues of  $RR^T$  will be denoted by  $\kappa_1 > \kappa_2 > \dots > \kappa_{\min(a,b)-1}$ , and its scaled versions equal

$$\omega_i = \frac{\kappa_i}{\sum_k \kappa_k}, \quad i = 1, 2, \dots, \min(a, b) - 1.$$

If the interaction is present we may expect that some of  $\omega_i$  coefficients will be substantially higher than others.

**Tukey test:** Introduced in Tukey (1949). Tukey test first estimates row and column effects and then tests for the interaction of a type  $\gamma_{ij} = k\alpha_i\beta_j$  ( $k = 0$  implies no interaction). Tukey test statistic  $S_T$  equals

$$S_T = MS_{int}/MS_{error},$$

where

$$MS_{int} = \frac{\left(\sum_i \sum_j y_{ij}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})\right)^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}$$

and

$$MS_{error} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 - a \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 - b \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 - MS_{int}}{(a-1)(b-1) - 1}.$$

Under the additivity hypothesis  $S_T$  is  $F$ -distributed with 1 and  $(a-1)(b-1)-1$  degrees of freedom.

**Mandel test:** Introduced in Mandel (1961). Mandel test statistic  $S_M$  equals

$$S_M = \frac{\sum_i (z_i - 1)^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}{a-1} / \frac{\sum_i \sum_j ((y_{ij} - \bar{y}_{i.}) - z_i(\bar{y}_{.j} - \bar{y}_{..}))^2}{(a-1)(b-2)},$$

where

$$z_i := \frac{\sum_j y_{ij}(\bar{y}_{.j} - \bar{y}_{..})}{\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}.$$

Under the additivity hypothesis  $S_M$  is  $F$ -distributed with  $a-1$  and  $(a-1)(b-1)$  degrees of freedom.

Definitions of the three later tests slightly differ from their original versions. For  $a, b$  fixed, a simulation may be used to get the critical values.

**Johnson – Graybill test:** Introduced in Johnson and Graybill (1972). Johnson – Graybill test statistic is just  $S_J = \omega_1$ . The additivity hypothesis is rejected if  $S_J$  is high.

**Locally best invariant (LBI) test:** See Boik (1993b). LBI test statistic equals (up to a monotonic transformation)

$$S_L = \sum_{i=1}^{\min(a,b)-1} \omega_i^2.$$

The additivity hypothesis is rejected if  $S_L$  is high.

**Tusell test:** See Tusell (1990). Tusell test statistic equals (up to a constant)

$$S_U = \prod_{i=1}^{\min(a,b)-1} \omega_i.$$

The additivity hypothesis is rejected if  $S_U$  is low.

As will be verified in the next section, Tukey and Mandel tests are appropriate if  $\gamma_{ij} = k\alpha_i\beta_j$  while Johnson – Graybill, LBI and Tusell omnibus tests are suitable in cases of more complexed interactions.

### 3 Simulation Study

In this section simulation results about power of the additivity tests are presented. According to Šimečková and Rasch (2008) the type-I-risk of the tests mentioned in Section 2 is not touched even when one of the effects in (1) is considered as random. The mixed effects model used for the simulation study is as (1) where  $\mu$ ,  $\alpha_i$  are constants, and  $\beta_j$  are independent normally distributed random variables with zero mean and variance  $\sigma_\beta^2$ .

Two possible interaction schemes were under inspection:

- A)  $\gamma_{ij} = k\alpha_i\beta_j$  where  $k$  is a real constant.
- B)  $\gamma_{ij} = k\alpha_i\delta_j$  where  $\delta_j$  are independent normally distributed random variables with zero mean and variance  $\sigma_\beta^2$ , independent of  $\beta_j$  and  $\epsilon_{ij}$ , and  $k$  a real constant.

The  $\epsilon_{ij}$  are independent normally distributed random variables with zero mean,  $\mu = 0$ , and unit variance,  $\sigma^2 = 1$ .

The other parameters are equal to  $\mu = 0$ ,  $\sigma_\beta^2 = 2$ ,  $\sigma^2 = 1$ ,  $a = 10$ ,

$$(\alpha_1, \dots, \alpha_{10}) = (-2.03, -1.92, -1.27, -0.70, 0.46, 0.61, 0.84, 0.94, 1.07, 2.00).$$

Two possibilities are considered for the  $b$ , either  $b = 10$  or  $b = 50$ , and 10 different values between 0 and 12 are considered for the interaction parameter  $k$ .

For each combination of parameters' values a dataset was generated based on the model (1), the tests of additivity were done and their results were noted down. The step was repeated 10 000 times. The estimated power of the test is the percentage of the positive results. All tests were done on  $\tilde{\alpha} = 5\%$  level.

The dependence of the power on  $k$  is visualized in Figure 1. As we can see, while Tukey and Mandel tests outperformed omnibus tests for interaction A

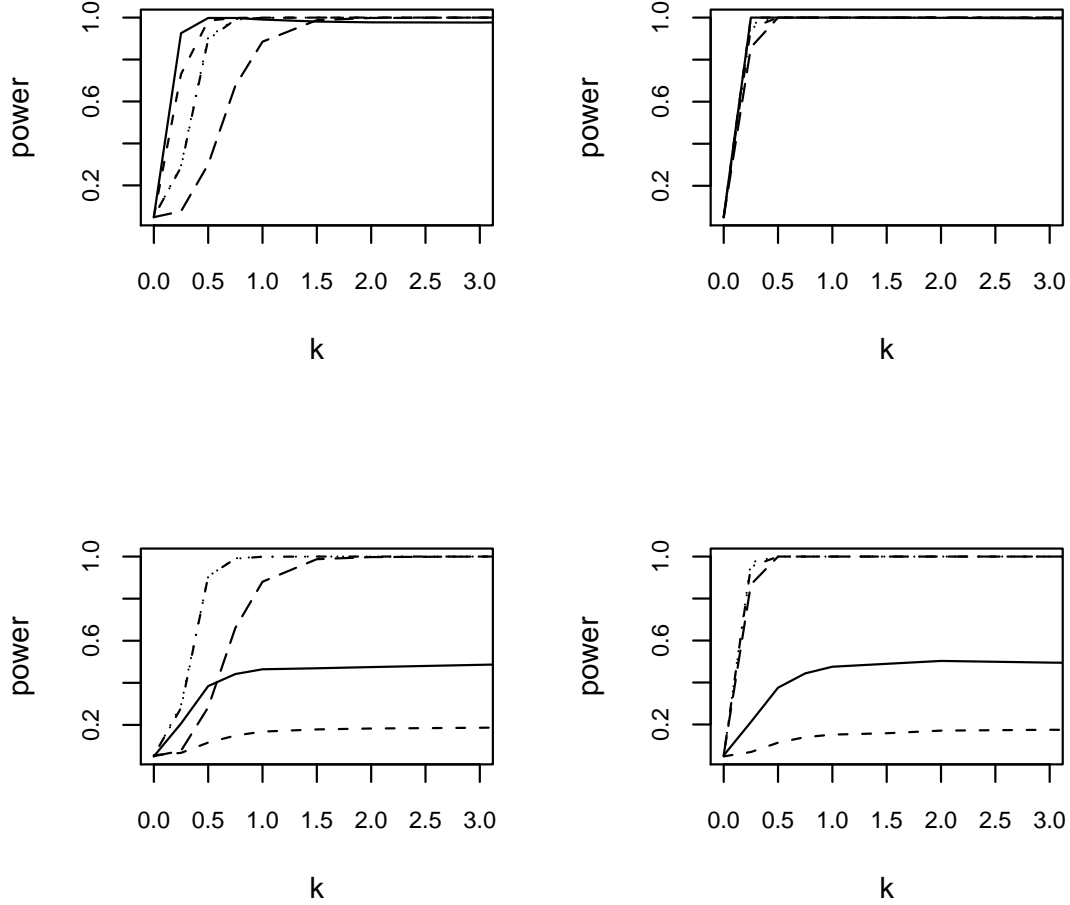


Fig. 1. Power dependence on  $k$ ,  $b$  ( $b = 10$  left,  $b = 50$  right) and interaction type ( $A$  up,  $B$  down). Tukey test solid line, Mandel test dashed line, Johnson – Graybill test dotted line, LBI test dot-dash line, Tusell test long dash line.

and low  $k$  and  $b$ , they completely fail to detect the interaction B even for a large value of  $k$  and  $b = 50$ . Therefore, it is desirable to develop a test which is able to detect a spectrum of practically relevant alternatives while still has the power comparable to the Tukey and Mandel tests for the most common interaction scheme A.

#### 4 Modification of Tukey Test

In Tukey test a model (1)

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + k\alpha_i\beta_j + \epsilon_{ij} \quad (3)$$

is tested against a submodel (2)  $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ . The estimators of row effects  $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$  and column effects  $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$  are calculated in the

same way in both models although the dependency of  $y_{ij}$  on these parameters is not linear for the full model.

The main idea behind a presented modification is that the full model (3) is fitted by a nonlinear regression and tested against a submodel  $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$  by a likelihood ratio test. The estimates of row and column effects therefore differ for each model.

#### 4.1 Non-adjusted test

Under additivity hypothesis the maximum likelihood estimators of parameters can be calculated simply as  $\hat{\mu} = \bar{y}_{..}$ ,  $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$  and  $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$ . Residual sum of squares equals

$$RSS_0 = \sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

In the full model (3) the parameters' estimates are computed iteratively. Let us first take  $\hat{\alpha}_i^{(0)} = \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ ,  $\hat{\beta}_j^{(0)} = \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$  and

$$\hat{k}^{(0)} = \frac{\sum_i \sum_j (y_{ij} - \hat{\alpha}_i^{(0)} - \hat{\beta}_j^{(0)} - \hat{\mu}) \cdot \hat{\alpha}_i^{(0)} \cdot \hat{\beta}_j^{(0)}}{\sum_i \sum_j (\hat{\alpha}_i^{(0)})^2 \cdot (\hat{\beta}_j^{(0)})^2}.$$

The  $\hat{k}^{(0)}$  is equal to the estimator of  $k$  in the classical Tukey test.

The iteration procedure continues by updating estimates one by one (while the rest of parameters are fixed):

- $\hat{\alpha}_i^{(n)} = \frac{\sum_j (y_{ij} - \hat{\mu} - \hat{\beta}_j^{(n-1)}) \cdot (1 + \hat{k}^{(n-1)} \cdot \hat{\beta}_j^{(n-1)})}{\sum_j (1 + \hat{k}^{(n-1)} \cdot \hat{\beta}_j^{(n-1)})^2}$
- $\hat{\beta}_j^{(n)} = \frac{\sum_i (y_{ij} - \hat{\mu} - \hat{\alpha}_i^{(n-1)}) \cdot (1 + \hat{k}^{(n-1)} \cdot \hat{\alpha}_i^{(n-1)})}{\sum_i (1 + \hat{k}^{(n-1)} \cdot \hat{\alpha}_i^{(n-1)})^2}$
- $\hat{k}^{(n)} = \frac{\sum_i \sum_j (y_{ij} - \hat{\alpha}_i^{(n-1)} - \hat{\beta}_j^{(n-1)} - \hat{\mu}) \cdot \hat{\alpha}_i^{(n-1)} \cdot \hat{\beta}_j^{(n-1)}}{\sum_i \sum_j (\hat{\alpha}_i^{(n-1)})^2 \cdot (\hat{\beta}_j^{(n-1)})^2}$

Surprisingly, it seems that one iteration is just enough to converge in a vast majority of cases. Therefore, for a simplicity reason let us define

$$RSS = \sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i^{(1)} - \hat{\beta}_j^{(1)} - k^{(1)} \hat{\alpha}_i^{(1)} \hat{\beta}_j^{(1)})^2.$$

The likelihood ratio statistic of the modified Tukey test, i.e. a difference of twice log-likelihoods, equals

$$\frac{RSS_0 - RSS}{\sigma^2}$$

and is asymptotically  $\chi^2$ -distributed with 1 degree of freedom.

The consistent estimate of a residual variance  $\sigma^2$  equals  $s^2 = \frac{RSS}{ab-a-b}$  and  $\frac{RSS}{\sigma^2}$  is approximately  $\chi^2$ -distributed with  $ab-a-b$  degrees of freedom. Thus, using a linear approximation of the nonlinear model (3)

$$\frac{RSS_0 - RSS}{\frac{RSS}{ab-a-b}} \tag{4}$$

is  $F$ -distributed with 1 and  $ab-a-b$  degrees of freedom. Easy manipulation of (4) gives the modified Tukey test which rejects the additivity hypothesis if and only if

$$RSS_0 > RSS \left( 1 + \frac{1}{ab-a-b} F_{1,ab-a-b}(1-\tilde{\alpha}) \right),$$

where  $F_{1,ab-a-b}(1-\tilde{\alpha})$  stands for  $1-\tilde{\alpha}$  quantile of  $F$ -distribution with 1 and  $ab-a-b$  degrees of freedom.

Now we will return to the simulation study from Section 3. For interaction A the power of the modified test is almost equal to the power of Tukey test. For interaction B the power of the tests is compared on Figure 2, the power of modified test is much higher than the power of Tukey test.

Theoretically, we may expect the modified test to be conservative because just one iteration does not find precisely the maximum of model (3) likelihood. However, as we will see in the following part a situation for a small number of rows or columns is quite opposite.

#### 4.2 Small sample adjustment

If the left part of Figure 2 would be magnified enough it will show that the modified test does not work properly (type-I-risk  $\doteq 6\%$ ). The reason is that the likelihood ratio test statistic converges to  $\chi^2$ -distribution rather slowly (see Bartlett (1937)) and a correction for small sample size is needed. We present two possibilities that are recommended if a number of rows or columns is below 20 (empirical threshold based on simulations).

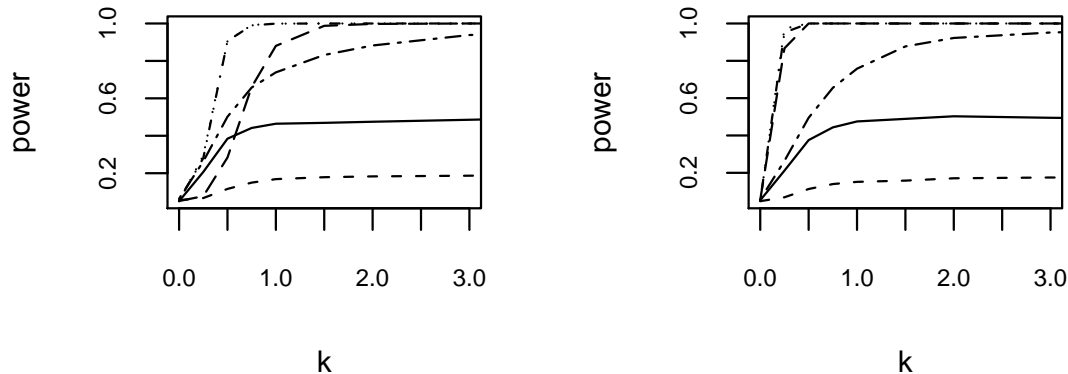


Fig. 2. Power dependence on  $k$ ,  $b$  ( $b = 10$  left,  $b = 50$  right) for interaction type  $B$ . Tukey test solid line, Mandel test dashed line, Johnson – Graybill test dotted line, LBI test dot-dash line, Tusell test long dash line, modified Tukey test two dash line. The proposed modification improved Tukey test and for large  $k$  almost reach power of omnibus tests.

One possibility to overcome this obstacle is a permutation test, i.e. generate data as follows

$$y_{ij}^{(perm)}(t) = \hat{\mu} + \hat{\alpha}_i^{(0)} + \hat{\beta}_j^{(0)} + r_{\pi_{ij}(t)}, \quad t = 1, \dots, N^{(perm)}$$

where  $\pi(t)$  is a random permutation of indexes of  $R$  matrix. For each  $t$  the statistic of interest  $S^{(perm)}(t) = RSS_0(t) - RSS(t)$  is computed. The critical value equals  $(1 - \tilde{\alpha}) \cdot 100\%$  quantile of  $S^{(perm)}(t)$ ,  $t = 1, \dots, N^{(perm)}$ .

The second possibility is to estimate the residual variance  $s^2 = \frac{RSS}{ab-a-b}$  and then generate samples of a distribution

$$y_{ij}^{(sample)}(t) = \hat{\mu} + \hat{\alpha}_i^{(0)} + \hat{\beta}_j^{(0)} + \epsilon_{ij}^{(NEW)}(t), \quad t = 1, \dots, N^{(sample)}$$

where  $(\epsilon_{ij}^{(NEW)})(t)$  are i.i.d. generated from a normal distribution with zero mean and variance  $s^2$ . This is simply parametric bootstrap on residuals.

The proposed statistic of interest is  $\text{abs}(k^{(1)})$  mirroring deviation from null hypothesis  $k = 0$ . As in the permutation test the additivity hypothesis is rejected if more than  $(1 - \tilde{\alpha}) \cdot 100\%$  of sampled statistics lie below the statistic based on real data.

## 5 Conclusion

We have proposed a modification of the Tukey additivity test. The modified test performs almost as good as Tukey test when the interaction is a product of main effects but should be recommended if we also request reasonable power



in case of more general interaction schemes. Problems with small sample size may be overcome by permutation test or parametric bootstrap on residuals.

All mentioned tests are implemented in R package `AdditivityTests` that may be downloaded on <http://github.com/rakosnicek/additivityTests>. As far as we are informed, this is the first R implementation of additivity tests with the exception of the Tukey test.

## References

- Alin, A. and Kurt, S. (2006). Testing non-additivity (interaction) in two-way anova tables with no replication. *Statistical Methods in Medical Research*, 15:63–85.
- Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests. *Royal Society of London Proceedings Series A*, 160:268–282.
- Boik, R. (1993a). A comparison of three invariant tests of additivity in two-way classifications with no replications. *Computational Statistics & Data Analysis*, 15:411–424.
- Boik, R. (1993b). Testing additivity in two-way classifications with no replications: the locally best invariant test. *Journal of Applied Statistics*, 20:41–55.
- Johnson, D. and Graybill, F. (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, 67:862–868.
- Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, 56:878–888.
- Šimečková, M. and Rasch, D. (2008). Additivity tests for the mixed model in the two-way anova with single sub-class numbers – type-I-risk. Lifestat 2008, Poster presentation.
- Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics*, 5:232–242.
- Tusell, F. (1990). Testing for interaction in two-way anova tables with no replication. *Computational Statistics & Data Analysis*, 10:29–45.