Eur. J. Stat. 4 (2024) 4
doi: 10.28924/ada/stat.4.4

ANOVA F Test of Non-Null Hypothesis

Guolong Zhao*, Junxia Yang, Liufeng Zhang, Huiyu Yang

Henan Institute of Medical Sciences, Henan Academy of Medical and Pharmaceutical Sciences, 40 University Road, Zhengzhou, Henan, 450052, China

**Correspondence: zhaogzu@hotmail.com*

Abstract. ANOVA, a test of a null hypothesis, is limited in assessing the statistical significance of differences. This paper considers an ANOVA F test of the non-null hypothesis for comparing k group means. A margin is chosen for the difference of means between each group and the kth group. A non-null hypothesis is defined to be the difference equal to the margin instead of zero. Data are thus prepared under the non-null hypothesis. Then follows the derivation of the one-way ANOVA non-null F test and its power. It reduces to the classical F test on setting the margin equal to zero. The observed size of it is identical to that of the F test and is near the nominal level of significance. The observed power is close to the power in balanced designs. With the non-null F test, it enables inferences to extend to the equivalence of group means or the clinical significance of differences. An example is taken to analyze both non-inferiority trials and k-sample equivalence trials.

1. Introduction

To test a null hypothesis, the result is either statistically significant or non-significant. The non-significant result means failing to reject the null hypothesis of the equality between a new treatment and a control. Nevertheless, it does not mean that there exists a real equality [1, 2]. Regarding the significant result, it is often not enough to define success because it does not yield information about magnitude of effect, practical significance, nor clinical significance [3, 4, 5, 6, 7]. Clearly, the test of null hypothesis is unable to evaluate the equivalence of group means or the clinical significance of differences. Thus the topic shifts to a test of the non-null hypothesis.

The term non-null hypothesis was introduced by Egon Pearson (1939)[8] and used by Fisher (Good 1992)[9]. Some other terms, as we have seen, are often used synonymously: the shifted null-hypothesis, non-zero null-hypothesis, and so on [6]. Of these, the term non-null hypothesis is found in the book, A Dictionary of Statistical Terms [10]. Some tests of the non-null hypothesis are available [11, 12, 13].

Received: 28 Dec 2023.

Key words and phrases. ANOVA; clinical significance; equivalence; F test; non-null hypothesis; t test.

Up to date, the existing tests of the non-null hypothesis address categorical data only. Concerning numerical data, the most frequently used procedure is analysis of variance (or ANOVA for short) [14]. However, it refers to the null hypothesis, waiting for a non-null generalization.

This paper considers an ANOVA F test of the non-null hypothesis (or the non-null F test for short) for comparing k group means. As we show in Section 2, a margin is chosen for the difference of means between each group and the kth group. A non-null hypothesis is defined to be the difference equal to the margin instead of zero. Data are thus prepared under the non-null hypothesis. Then follows the derivation of the one-way ANOVA non-null F test in Section 3. It overlaps the two-sample non-null t test, which is then inverted to the confidence interval for analyzing the data of non-inferiority or equivalence trials. Section 4 provides an extension to the two-way ANOVA non-null F test. It overlaps the paired non-null t test. Section 5 formulates the power and sample sizes in balanced and unbalanced designs using the R Language [15]. The observed test size and power are demonstrated using Monte Carlo techniques in Section 6. Section 7 contains three examples addressing the relation between non-inferiority and k-sample equivalence trials, the clinical significance of differences, and the randomized block design, respectively. Section 8 covers various aspects of these tests such as history, contributions, properties, clinical applications, and perspectives.

2. Preparing data under non-null hypothesis

2.1. **ANOVA data organization.** ANOVA data organization is usually based on the traditional standard: Let Y_{ij}, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$, be independent numerical observations, each from an underlying normal distribution: $N(\nu_j, \sigma_j^2)$, where n_j is the sample size, ν_j the mean, and σ_j^2 the variance in the jth group. The total sample size is $n = \sum_j n_j$ and the sample fraction is $g_j = n_j/n$ with $\sum_j g_j = 1$. When carrying out an ANOVA, one assumes equal variances for the k populations: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. The hypotheses to be tested in ANOVA is $H_0 : \nu_1 = \nu_2 = \cdots = \nu_k$ *versus* H_1 : *not all the* ν_j'*s are equal*.

For a given data set, the mean ν_j is estimated by $\bar{Y}_{\cdot j} = \frac{1}{n_j}\sum_i Y_{ij}$. It follows the grand mean $\nu = \sum_j g_j \nu_j$ and its estimator $\bar{Y}_{\cdot\cdot} = \sum_j g_j \bar{Y}_{\cdot j}$. We define the difference of means between the jth and kth group and its estimator to be

$$\mu_j = \nu_j - \nu_k \text{ and } \hat{\mu}_j = \bar{Y}_{\cdot j} - \bar{Y}_{\cdot k} \tag{2.1}$$

with the averages

$$\mu = \sum_j g_j \mu_j = \nu - \nu_k \text{ and } \bar{\mu} = \sum_j g_j \hat{\mu}_j = \bar{Y}_{\cdot\cdot} - \bar{Y}_{\cdot k}.$$

Definition (2.1) does three things at once: 1. It shows the last difference $\mu_k = \nu_k - \nu_k = 0$. 2. It allows us to rewrite the hypotheses in an equivalent form $H_0 : \mu = 0$ *versus* $H_1 : \mu \neq 0$, which

provides a room for a non-null generalization. 3. We can reexpress the mean and its estimator as

$$\nu_j = \nu + (\mu_j - \mu) \text{ and } \bar{Y}_{\cdot j} = \bar{Y}_{\cdot\cdot} + (\hat{\mu}_j - \bar{\mu}), \tag{2.2}$$

which will play a key role in generalizing the F test to its non-null version.

2.2. **Process of data preparation.** The non-null hypothesis is defined by a margin of the difference. It refers to the equivalence margin in this text. In the two-sample case, this is sometimes known as the non-inferiority margin. Let Δ_j be the margin of μ_j with the average $\Delta = \sum_j g_j \Delta_j$. How to choose the values of margin Δ_j, $j = 1, 2, \ldots, k-1$, will be given later in Section 8. For the last item $j = k$, however, we set $\Delta_k = 0$ because $\mu_k = 0$ (see Definition (2.1)). As Killoy (2002)[16] mentioned, the parameter μ_j defines the range in which the margin Δ_j has its being so that the sign of Δ_j is always kept identical to that of μ_j. In clinical practice, Δ_j is the minimal detectable difference. The difference $|\mu_j|$ less than $|\Delta_j|$ implies the equivalence of group means. Conversely, the difference $|\mu_j|$ exceeding $|\Delta_j|$ is thought to be clinically important and would lead to a preference for one treatment over the other. Taking the average margin in place of zero gives the non-null hypotheses for equivalence trials

$$H_0 : |\mu| \geq |\Delta| \text{ versus } H_1 : |\mu| < |\Delta| \ (|\mu| \in (0, |\Delta|)). \tag{2.3}$$

For superiority trials, the non-null hypotheses are $H_0 : |\mu| \leq |\Delta|$ *versus* $H_1 : |\mu| > |\Delta|$ $(|\mu| \in (|\Delta|, \infty))$. Clearly, they are asymmetric to equivalence trials. As it turns out, while the null hypothesis refers to the difference of zero, the non-null hypothesis generalizes to arbitrary differences.

Under the non-null hypothesis, the mean and its estimator in (2.2) are generalized to

$$\nu_j^* = \nu + [(\mu_j - \mu) - (\Delta_j - \Delta)] \text{ and } \bar{Y}_{\cdot j}^* = \bar{Y}_{\cdot\cdot} + [(\hat{\mu}_j - \bar{\mu}) - (\Delta_j - \Delta)]. \tag{2.4}$$

If subtracting (2.2) from (2.4), we obtain

$$\nu_j^* = \nu_j - (\Delta_j - \Delta) \text{ and } \bar{Y}_{\cdot j}^* = \bar{Y}_{\cdot j} - (\Delta_j - \Delta). \tag{2.5}$$

This provides the non-null values of the grand mean $\nu^* = \sum_j g_j \nu_j^*$ and its estimator $\bar{Y}_{\cdot\cdot}^* = \sum_j g_j \bar{Y}_{\cdot j}^*$. In view of the fact $\sum_j g_j (\Delta_j - \Delta) = 0$, it must be true that

$$\nu^* = \sum_j g_j [\nu_j - (\Delta_j - \Delta)] = \nu \text{ and } \bar{Y}_{\cdot\cdot}^* = \sum_j g_j [\bar{Y}_{\cdot j} - (\Delta_j - \Delta)] = \bar{Y}_{\cdot\cdot}. \tag{2.6}$$

Remark 2.1. *The grand mean is the same regardless of whether the null or non-null hypothesis holds.*

Let Y_{ij}^* be the observations under the non-null hypothesis; then the non-null values of the sample means should be $\bar{Y}_{\cdot j}^* = \frac{1}{n_j} \sum_i Y_{ij}^*$. Referring to (2.5), we have $\frac{1}{n_j} \sum_i Y_{ij}^* = \frac{1}{n_j} \sum_i Y_{ij} - (\Delta_j - \Delta)$. That is, $\frac{1}{n_j} \sum_i Y_{ij}^* = \frac{1}{n_j} \sum_i [Y_{ij} - (\Delta_j - \Delta)]$, which implies

$$Y_{ij}^* = Y_{ij} - (\Delta_j - \Delta), \tag{2.7}$$

where $(\Delta_j - \Delta)$ is a constant.

Remark 2.2. *If the observations are normal, they are still normal under the non-null hypothesis since normal random variables plus or minus a constant are themselves normal.*

The variance σ^2 is estimated by $S_j^2 = \frac{1}{n_j-1}\sum_i (Y_{ij} - \bar{Y}_{\cdot j})^2$ with the non-null values $S_j^{2*} = \frac{1}{n_j-1}\sum_i (Y_{ij}^* - \bar{Y}_{\cdot j}^*)^2$. Since subtracting (2.5) from (2.7) gives

$$Y_{ij}^* - \bar{Y}_{\cdot j}^* = Y_{ij} - \bar{Y}_{\cdot j}, \tag{2.8}$$

it follows that

$$S_j^{2*} = \frac{1}{n_j - 1}\sum_i (Y_{ij} - \bar{Y}_{\cdot j})^2 = S_j^2. \tag{2.9}$$

Remark 2.3. *The sample variance is left unchanged in the non-null generalization of the ANOVA F test.*

3. Building one-way ANOVA of non-null hypothesis

3.1. **One-way ANOVA non-null F test.** The total sum of squares is known to be $Q = \sum_j \sum_i (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$ in the one-way ANOVA. Under the non-null hypothesis, it should be $Q^* = \sum_j \sum_i (Y_{ij}^* - \bar{Y}_{\cdot\cdot}^*)^2$. Since $\bar{Y}_{\cdot\cdot}^* = \bar{Y}_{\cdot\cdot}$ (see Remark 2.1), it is written formally as

$$Q^* = \sum_j \sum_i (Y_{ij}^* - \bar{Y}_{\cdot\cdot})^2.$$

For the treatment sum of squares, the non-null value can be treated in the same way as

$$Q_1^* = \sum_j \sum_i (\bar{Y}_{\cdot j}^* - \bar{Y}_{\cdot\cdot})^2 = \sum_j n_j (\bar{Y}_{\cdot j}^* - \bar{Y}_{\cdot\cdot})^2. \tag{3.1}$$

Concerning the error sum of squares, it is $Q_2^* = \sum_j (n_j - 1) S_j^{2*}$. From (2.9) comes the not-surprising result that

$$Q_2^* = \sum_j (n_j - 1) S_j^2 = \sum_j \sum_i (Y_{ij} - \bar{Y}_{\cdot j})^2 = Q_2, \tag{3.2}$$

which expresses the fact that the error sum of squares is invariable in the non-null generalization of the ANOVA F test based on Remark 2.3.

The one-way ANOVA non-null F statistic has the form

$$F^* = \frac{Q_1^*/(k-1)}{Q_2/(n-k)} \sim F_{1-\alpha,\, k-1,\, n-k}, \tag{3.3}$$

where the only change is using Q_1^* in place of Q_1. It accounts for the differences among group means.

Setting $\Delta_j = 0$ gives $\bar{Y}_{\cdot j}^* = \bar{Y}_{\cdot j}$ and $Q_1^* = Q_1$. So, we have $F^* = F$, meaning that the statistic reduces to the classical F statistic on setting the margin equal to zero. This is known to be the reducibility, an essential property of such tests (Zhao 2008). This property is shared by all the test statistics presented in later sections.

The F test will provide an answer by deciding whether or not the null hypothesis $H_0 : \mu = 0$, i.e., $H_0 : \nu_1 = \nu_2 = \cdots = \nu_k$ should be rejected. However, the alternative, $H_1 : \mu \neq 0$, i.e., $H_1 : \textit{not all the } \nu_j'\textit{s are equal}$, does not specify any pair of means. This problem continues to the non-null F test. Fortunately, there is a solution available from breaking down an overall null hypothesis into smaller more relevant sub-hypotheses.

3.2. **Testing non-null sub-hypotheses.** To do this, there are two general ways: the Tukey method and the contrast [17]. The Tukey method is expressed in a confidence interval and the contrast, in an F test. Therefore, the contrast is taken here since it is the F test that conforms to the topic of this text. In every set of k-sample data, there are $\binom{k}{2}$ contrasts. It is given the symbol C_i, $i = 1, 2, \ldots, \binom{k}{2}$, with $C_i = \sum_j c_j \nu_j$, where c_j is coefficients of ν_j with $\sum_j c_j = 0$. For example, in the contrast C_1, we have $C_1 = \nu_1 - \nu_2 = \mu_1$. The linear combination of the contrast is given by $C_1 = (1)\nu_1 + (-1)\nu_2 + (0)\nu_3 + \cdots + (0)\nu_k = \nu_1 - \nu_2$, where $c_j = (1, -1, 0, \ldots, 0)$. For all $s \neq t$, we have the contrast $C_i = \nu_s - \nu_t = \mu_i$ with the estimator $\hat{C}_i = \bar{Y}_{\cdot s} - \bar{Y}_{\cdot t} = \hat{\mu}_i$. The corresponding sub-hypotheses are $H_0' : \mu_i = 0$ *versus* $H_1' : \mu_i \neq 0$. The sum of squares associated with C_i is estimated by

$$Q_i = \frac{\hat{C}_i^{\,2}}{\sum_j c_j^2/n_j} = \frac{(\bar{Y}_{\cdot s} - \bar{Y}_{\cdot t})^2}{\frac{1}{n_s} + \frac{1}{n_t}} = \frac{\hat{\mu}_i^{\,2}}{\frac{1}{n_s} + \frac{1}{n_t}}.$$

The non-null sub-hypothesis with the margin Δ can be still put in the form of (2.3): $H_0' : |\mu_i| \geq |\Delta|$ *versus* $H_1' : |\mu_i| < |\Delta|$ for equivalence trials and $H_0' : |\mu_i| \leq |\Delta|$ *versus* $H_1' : |\mu_i| > |\Delta|$ for superiority trials. Then the non-null value of of Q_i should be

$$Q_i^* = \frac{(|\hat{\mu}_i| - |\Delta|)^2}{\frac{1}{n_s} + \frac{1}{n_t}}.$$

For each different pairwise sub-hypothesis test, the non-null F statistic can be constructed as

$$F_i^* = \frac{Q_i^*/1}{Q_2/(n-k)} \sim F_{1-\alpha,\, 1,\, n-k}, \tag{3.4}$$

which enables us to deal with the difference between group means ν_s and ν_t.

Testing the sub-hypothesis $H_0' : \mu_i = 0$, i.e., $H_0' : \nu_s = \nu_t$ *versus* $H_1' : \mu_i \neq 0$, i.e., $H_1' : \nu_s \neq \nu_t$ tells which two are equal and which two are not, depending upon the pairwise comparisons. This has come to the solution of the problem given above. Also, the same solution holds for testing the non-null sub-hypotheses along the lines of the pairwise comparisons.

3.3. **Two-sample data.** The one-way ANOVA non-null F test (3.3) overlaps the two-sample non-null t test. The latter is then inverted to the confidence interval for analyzing the data of non-inferiority or equivalence trials. In the two-sample case, we only need to specify Δ_1. Definition (2.1) states $\mu_2 = 0$ so that we set $\Delta_2 = 0$. The non-null F statistic (3.3) becomes

$$F^* = \frac{Q_1^*/1}{Q_2/(n-2)} \sim F_{1-\alpha,\, 1,\, n-2}. \tag{3.5}$$

The treatment sum of squares in (3.5) is $Q_1^* = n_1(\bar{Y}_{\cdot 1}^* - \bar{Y}_{\cdot\cdot})^2 + n_2(\bar{Y}_{\cdot 2}^* - \bar{Y}_{\cdot\cdot})^2$, where $\bar{Y}_{\cdot 1}^* - \bar{Y}_{\cdot\cdot} = g_2[(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}) - \Delta_1]$ and $\bar{Y}_{\cdot 2}^* - \bar{Y}_{\cdot\cdot} = g_1[-(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}) + \Delta_1]$. Writing $\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}$ as $\hat{\mu}$ and Δ_1 as Δ gives $Q_1^* = g_1 n g_2^2(\hat{\mu} - \Delta)^2 + g_2 n g_1^2(-\hat{\mu} + \Delta)^2$, which simplifies to $Q_1^* = n k_1 g_2(\hat{\mu} - \Delta)^2$ or

$$Q_1^* = \frac{(\hat{\mu} - \Delta)^2}{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The error sum of squares in (3.5) is $Q_2 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2$. This is just the numerator of the pooled variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

so that $Q_2 = S_p^2(n_1 + n_2 - 2)$. When Q_1^* and Q_2 are substituted into the square root of (3.5), one obtains

$$\sqrt{F^*} = \frac{\hat{\mu} - \Delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = t^* \sim T_{n_1+n_2-2}.$$

This is the two-sample non-null t statistic. Notice that the denominator is the same as that of its classical counterpart.

Food and Drug Administration (FDA, 2016) recommends the use of confidence intervals on the data of two-sample non-inferiority trials. Inverting the equation gives a $100(1-\alpha)\%$ confidence interval for Δ

$$\hat{\mu} - t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \Delta \leq \hat{\mu} + t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

which has several forms in this context. For instance, it takes the form

$$0 \leq |\Delta| \leq |\hat{\mu}| + t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{3.6}$$

in equivalence trials with the hypotheses $H_0 : |\mu| \geq |\Delta|$ *versus* $H_1 : |\mu| < |\Delta|$. The equivalence holds when $H_0 : |\mu| \geq |\Delta|$ is rejected with $|\hat{\mu}| < |\Delta|$. Clearly, this definition of equivalence is the same as that in International Conference on Harmonisation (1998). The equivalence has two symmetric profiles: the non-inferiority and the non-superiority. When $\mu \leq 0$ and $\Delta \leq 0$, we have the non-inferiority with the hypothesis $H_0 : \mu \leq \Delta$ *versus* $H_1 : \mu > \Delta$ and when $\mu \geq 0$ and $\Delta \geq 0$, we have the non-superiority with the hypothesis $H_0 : \mu \geq \Delta$ *versus* $H_1 : \mu < \Delta$, where the confidence intervals are expressed in the forms

$$\hat{\mu} - t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \Delta \leq 0 \text{ and } 0 \leq \Delta \leq \hat{\mu} + t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \tag{3.7}$$

respectively. Non-inferiority and non-superiority trials are converted to each other as long as treatment and control groups reverse roles. A comparison of (3.6) and (3.7) provides insights into the relationship between equivalence trials and non-inferiority trials, which will be shown numerically in an example (see Subsection 7.1).

To evaluate the clinical significance of differences, the hypothesis to be tested is $H_0 : |\mu| \leq |\Delta|$ *versus* $H_1 : |\mu| > |\Delta|$ and the confidence interval has the form

$$|\hat{\mu}| - t_{\alpha/2,\, n_1+n_2-2} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq |\Delta| \leq \infty.$$

3.4. **One-sample data.** The one-sample case has only a single mean ν. Letting ν_0 be a specified value of the mean, the difference is $\mu = \nu - \nu_0$ so that $\nu = \nu_0 + \mu$. With the margin Δ, we have $\nu^* = \nu_0 + (\mu - \Delta)$ or $\nu^* = \nu - \Delta$. The non-null hypothesis is still written in the form of (2.3).

For a sample of $Y_i \sim N(\nu, \sigma^2)$, $i = 1, 2, \ldots, n$, the mean ν is estimated by $\bar{Y} = \frac{1}{n}\sum_i Y_i$. It follows that $\bar{Y}^* = \bar{Y} - \Delta$, which implies $\frac{1}{n}\sum_i Y_i^* = \frac{1}{n}\sum_i Y_i - \Delta$ so that $Y_i^* = Y_i - \Delta$. These variables generate an identity: $Y_i^* = \nu_0 + (\bar{Y}^* - \nu_0) + (Y_i^* - \bar{Y}^*)$ or, equivalently, $(Y_i^* - \nu_0) = (\bar{Y}^* - \nu_0) + (Y_i^* - \bar{Y}^*)$. It must be true that $\sum_i (Y_i^* - \nu_0)^2 = \sum_i [(\bar{Y}^* - \nu_0) + (Y_i^* - \bar{Y}^*)]^2$. Since the cross-product term vanishes: $\sum_i (\bar{Y}^* - \nu_0)(Y_i^* - \bar{Y}^*) = 0$, it follows that $\sum_i (Y_i^* - \nu_0)^2 = \sum_i (\bar{Y}^* - \nu_0)^2 + \sum_i (Y_i^* - \bar{Y}^*)^2$, or more conveniently, $Q^* = Q_1^* + Q_2^*$. Since $Y_i^* - \bar{Y}^* = (Y_i - \Delta) - (\bar{Y} - \Delta) = Y_i - \bar{Y}$, it turns out that $Q_2^* = \sum_i (Y_i - \bar{Y})^2 = Q_2$. The statistic for testing H *versus* H_1 is

$$F^* = \frac{Q_1^*/1}{Q_2/(n-1)} \sim F_{1-\alpha,\, 1,\, n-1}, \tag{3.8}$$

where $Q_1^* = \sum_i (\bar{Y}^* - \nu_0)^2$ or $Q_1^* = \sum_i ((\bar{Y} - \nu_0) - \Delta)^2$. The square root of (3.8) is just the one-sample non-null t statistic

$$\sqrt{F^*} = \frac{(\bar{Y} - \nu_0) - \Delta}{\sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n(n-1)}}} = t^* \sim T_{n-1}.$$

Inverting the statistic gives the confidence interval, which is omitted since its form is similar to that in the two-sample case.

4. Randomized block designs under non-null hypothesis

4.1. **Non-null F test for randomized block design.** The procedure stated above can easily be extended to the two-way ANOVA, say, the randomized block design. It operates a horizontal ANOVA and a vertical ANOVA for the treatment and block effect, respectively. Hence two F ratios are calculated, one for the treatment effect and one for the block effect. By analogy, further extension results in the effects of multiple factors, which lies beyond the scope of this text.

The data structure for a randomized block design is a matrix with n rows and k columns representing the n blocks and the k levels of treatment, respectively. Here, Y_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, denote the observations associated with the application of treatment j to block i.

Consider first the treatment effect. Sample fractions in this context are constant: $g_j \equiv \frac{1}{k}$. It follows that $\bar{Y}_{\cdot j} = \frac{1}{n}\sum_i Y_{ij}$, $\nu = \frac{1}{k}\sum_j \nu_j$, and $\mu = \frac{1}{k}\sum_j \mu_j$. Similarly, the average margin is $\Delta = \frac{1}{k}\sum_j \Delta_j$. The hypotheses are the same as (2.3) in style of writing.

The non-null value for the treatment sum of squares is given by

$$Q_t^* = \sum_j \sum_i (\bar{Y}_{\cdot j}^* - \bar{Y}_{\cdot\cdot})^2 = \sum_j n(\bar{Y}_{\cdot j}^* - \bar{Y}_{\cdot\cdot})^2, \tag{4.1}$$

where $\bar{Y}_{\cdot j}^* = \bar{Y}_{\cdot j} - (\Delta_j - \Delta)$ and $\bar{Y}_{\cdot\cdot} = \frac{1}{k}\sum_j \bar{Y}_{\cdot j}$. The error sum of squares is $Q_e^* = \sum_j \sum_i (Y_{ij}^* - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j}^* + \bar{Y}_{\cdot\cdot}^*)^2$, where $Y_{ij}^* = Y_{ij} - (\Delta_j - \Delta)$, $\bar{Y}_{i\cdot} = \frac{1}{k}\sum_j Y_{ij}$, and $\bar{Y}_{\cdot\cdot}^* = \frac{1}{k}\sum_j \bar{Y}_{\cdot j}^*$. An inspection of (2.6) and (2.8) discloses that

$$Q_e^* = \sum_j \sum_i (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2 = Q_e. \tag{4.2}$$

The non-null F statistic for treatment effect is given by

$$F_t^* = \frac{Q_t^*/(k-1)}{Q_e/(n-1)/(k-1)} \sim F_{1-\alpha,\, k-1,\, (n-1)(k-1)}. \tag{4.3}$$

We are then led to consider the block effect. It is analyzed with a process in analogy to that for the treatment effect but some minor changes in subscripts. The difference between block means is defined to be $\mu_i = \nu_i - \nu_n$, $i = 1, 2, \ldots, n$, with the average $\mu^b = \frac{1}{n}\sum_i \mu_i$. Prescribing the margin Δ_i corresponding to μ_i yields the average $\Delta^b = \frac{1}{n}\sum_i \Delta_i$. The hypotheses are in accord with (2.3) as long as using μ^b in place of μ and Δ^b in place of Δ.

The block sum of squares under the non-null hypothesis is

$$Q_b^* = \sum_j \sum_i (\bar{Y}_{i\cdot}^* - \bar{Y}_{\cdot\cdot})^2 = k\sum_i (\bar{Y}_{i\cdot}^* - \bar{Y}_{\cdot\cdot})^2,$$

where $\bar{Y}_{i\cdot}^* = \bar{Y}_{i\cdot} - (\Delta_i - \Delta^b)$. The error sum of squares is still Q_e. Thus we have

$$F_b^* = \frac{Q_b^*/(n-1)}{Q_e/(n-1)/(k-1)} \sim F_{1-\alpha,\, n-1,\, (n-1)(k-1)}. \tag{4.4}$$

This is the non-null F statistic for block effect.

4.2. **Paired data.** In some clinical situations, the randomized block design may have only 2 levels of treatment, where (4.3) is written

$$F_t^* = \frac{Q_t^*/1}{Q_e/(n-1)} \sim F_{1-\alpha,\, 1,\, n-1}.$$

In this case, the sample block mean is $\bar{Y}_{i\cdot} = \frac{1}{2}Y_{i1} + \frac{1}{2}Y_{i2}$ and the grand mean, $\bar{Y}_{\cdot\cdot} = \frac{1}{2}\bar{Y}_{\cdot 1} + \frac{1}{2}\bar{Y}_{\cdot 2}$. Substituting them into (4.2) gives the error sum of squares $Q_e = \sum_j \sum_i (Y_{ij} - \frac{1}{2}Y_{i1} - \frac{1}{2}Y_{i2} - \bar{Y}_{\cdot j} + \frac{1}{2}\bar{Y}_{\cdot 1} + \frac{1}{2}\bar{Y}_{\cdot 2})^2$. Let D_i be the within-block difference with $D_i = Y_{i1} - Y_{i2}$ and $\bar{D}$ be the average with $\bar{D} = \bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}$. Then (4.2) further simplifies to

$$Q_e = \frac{1}{2}\sum_i (D_i - \bar{D})^2. \tag{4.5}$$

Here (4.1) is written $Q_t^* = n\{(\bar{Y}_{\cdot 1}^* - \bar{Y}_{\cdot\cdot})^2 + (\bar{Y}_{\cdot 2}^* - \bar{Y}_{\cdot\cdot})^2\}$, where $\bar{Y}_{\cdot 1}^* - \bar{Y}_{\cdot\cdot} = \frac{1}{2}((\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}) - \Delta_1)$ and $\bar{Y}_{\cdot 2}^* - \bar{Y}_{\cdot\cdot} = \frac{1}{2}(-(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}) + \Delta_1)$. Writing $\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}$ as $\bar{D}$ and Δ_1 as Δ gives the treatment sum of squares

$$Q_t^* = \frac{1}{2} n(\bar{D} - \Delta)^2. \tag{4.6}$$

Putting (4.5) and (4.6) into the statistic and taking the square root of the result give

$$\sqrt{F_t^*} = \frac{\bar{D} - \Delta}{\sqrt{\frac{\sum_i (D_i - \bar{D})^2}{n(n-1)}}} = t^* \sim T_{n-1}.$$

This is just the paired non-null t statistic. Inverting the equation will then yield the confidence interval, which is omitted here.

5. Power and sample size determination

5.1. **Power of non-null F test.** The power of the F statistic depends upon the noncentral F distribution $F_{1-\alpha,\, k-1,\, n-k,\, \lambda}$ under the alternative, where λ is the non-centrality parameter with $\lambda = \sum_j n_j(\nu_j - \nu)^2/\sigma^2$ and n_j is the sample size per group with $n_j \equiv n/k$ [17].

The non-null value of the non-centrality parameter is expressed as $\lambda^* = \sum_j n_j(\nu_j^* - \nu)^2/\sigma^2$, where the only change is using ν_j^* (2.5) in place of ν_j with ν and σ^2 left unchanged based on Remark 2.1. and 2.3. It follows the non-null value of Cohen effect size for ANOVA $f^* = \sqrt{\lambda^*/n}$ or

$$f^* = \sqrt{\frac{\sum_j (\nu_j^* - \nu)^2/k}{\sigma^2}}. \tag{5.1}$$

This is computationally simple after specifying k, n_j, ν_j, σ^2, and Δ_j. But it is tedious to specify ν_j one by one. Here we arrange it into an arithmetic series with a given minimal mean ν_{min} and maximal difference μ_{max}. Then we have a descending series

$$\nu_j = \nu_{min} + \mu_{max} - (j-1)\, \mu_{max}/(k-1), \tag{5.2}$$

an ascending series $\nu_j = \nu_{min} + (j-1)\, \mu_{max}/(k-1)$, or an equal series $\nu_j = \nu_{min} + \mu_{max}/2$.

The power of the non-null F test at α-level is

$$1 - \beta^* = P(F^* \geq F_{1-\alpha,\, k-1,\, n-k,\, \lambda^*} \mid H_1 \textit{ is true}), \tag{5.3}$$

which is easily calculated using the function

pwr.anova.test ($k = k$, $n = n_j$, $f = f^*$, sig.level=α)

in the R Language. When H is true, we have $\nu_j^* = \nu$ by (2.4) and $f^* = 0$ by (5.1). Then, (5.3) becomes

$$1 - \beta^* = \alpha, \tag{5.4}$$

the nominal test size, which will be exhibited numerically later in Figure 3.

When $\Delta_j = 0$, we have $\nu_j^* = \nu_j$, $f^* = f$, and $F^* = F$ so that $1 - \beta^* = 1 - \beta$, meaning that the power function reduces to its classical counterpart on setting the margin equal to zero.

Figure 1 shows the power of the non-null F test at $\alpha = 0.05$ with the total sample size fixed at $n = 160$.

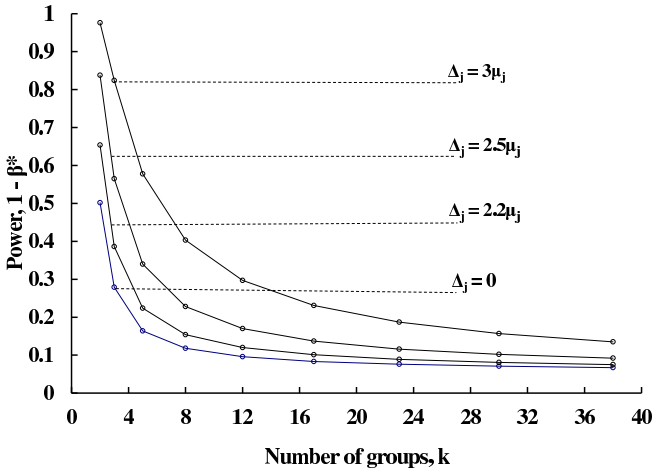

Figure 1. Power of non-null F test by number of groups with total sample size fixed

The significance level is α = 0.05. The total sample size is n = 160. The minimum mean is ν_{min} = 0.3. The maximum difference is μ_{max} = 0.125. The standard deviation is σ = 0.4. And the power of the non-null F test is given by (5.3).

In this figure, the margin is specified as $\Delta_j = 0, 2.2\mu_j, 2.5\mu_j, 3\mu_j$, which will be reused later in the simulation of equivalence trials. The power is computed from (5.3) with the other parameter values $k = 2, 3, \ldots, 38$, $\nu_{min} = 0.3$, $\mu_{max} = 0.125$, and $\sigma = 0.4$. All the curves have certain basic similarities in form. The power of the non-null F test is higher than that of the F test. Not surprisingly, as the number of groups gets larger, the power gets lower when the total sample size is fixed.

Another possibility is to take the sample size per group fixed at $n_j = 16$ instead. Then we get the power as indicated in Figure 2.

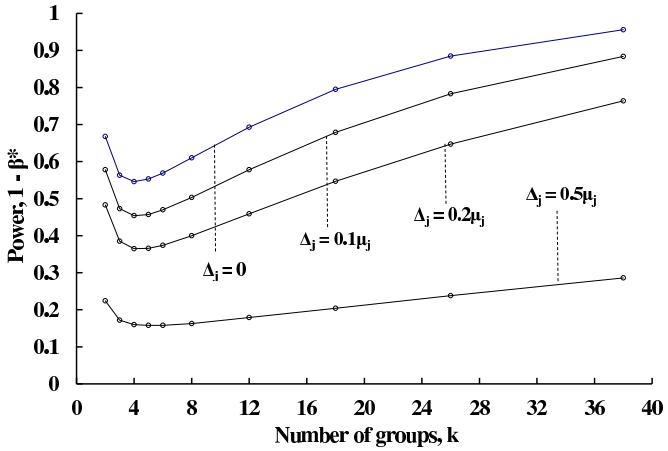

Figure 2. Similar to Figure 1, except for taking sample size per group fixed instead

The significance level is $\alpha = 0.05$. The sample size per group is $n_j = 16$. The minimum mean is $\nu_{min} = 0.3$. The maximum difference is $\mu_{max} = 0.35$. The standard deviation is $\sigma = 0.4$. And the power of the non-null F test is given by (5.3).

Here the margin is taken as $\Delta_j = 0, 0.1\mu_j, 0.2\mu_j, 0.5\mu_j$, which will be reused later in the simulation of superiority trials. The other parameter values are the same as those in Figure 1 but $\mu_{max} = 0.35$. The power of the non-null F test is lower than that of the F test. It is worth noting that over the range of k, the power decreases first and increases later with the minimum at $k = 4$.

Both Figure 1 and 2 have the margin Δ_j dependent on the difference μ_j. One might be of interest for taking Δ_j independent of μ_j. Now, the maximum margins are chosen as $\Delta_{max} = 0, 0.08, 0.16, 0.24$ for each of the maximum differences $\mu_{max} = 0, 0.04, \ldots, 0.4$, respectively. In applying the same approach with $n_j = 25$, $k = 8$, $\nu_{min} = 0.3$, and $\sigma = 0.4$, we get the power at the $\alpha = 0.05$ level as pictured in Figure 3.

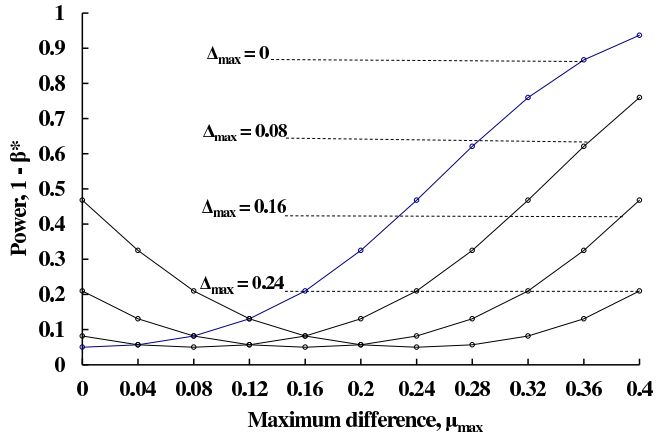

Figure 3. Power of non-null F test for values of margins independent of values of differences

The significance level is $\alpha = 0.05$. The sampe size per group is $n_j = 25$. The number of groups is $k = 8$. The minimum mean is $\nu_{min} = 0.3$. The standard deviation is $\sigma = 0.4$. And the power of the non-null F test is given by (5.3).

The curve with index $\Delta_{max} = 0$ gives the power of the F test, which increases with the increasing μ_{max}. Those labeled with $\Delta_{max} = 0.08, 0.16, 0.24$ refer to the power of the non-null F test. When $\mu_{max} < \Delta_{max}$, the power decreases as μ_{max} increases; when $\mu_{max} = \Delta_{max}$, the power equals 0.05, i.e., the nominal test size, as predicted by (5.4); and when $\mu_{max} > \Delta_{max}$, the power increases as μ_{max} increases.

5.2. **Sample size determination.** The sample size determination also relies on the power of the non-null F test (5.3) with the function pwr.anova.test. But this function provides balanced designs only. For unbalanced designs, we take a process with three steps: 1. Start by finding the sample size n_j per group using the function

$$\text{pwr.anova.test } (k = k,\ f = f^*,\ \text{sig.level}=\alpha,\ \text{power} = 1 - \beta).$$

2. Determine the total sample size $n = kn_j$. 3. Choose the sample fractions g_j deriving the unequal group sample sizes $n_j = g_j n$.

The last step involves the sample fraction g_j, which is specified as an arithmetic series here likewise. Let $\bar{g}$ be the average sample fraction with $\bar{g} = 1/k$. The minimum sample fraction is defined as $g_{min} = \eta\bar{g}$, where η is a real number with the value of $\eta \in [0, 1]$. It follows the maximum sample fraction $g_{max} = (2 - \eta)\bar{g}$. Accordingly, we have a descending series

$$g_j = g_{max} - 2(j - 1)(1 - \eta)\bar{g}/(k - 1) \tag{5.5}$$

or an ascending series

$$g_j = g_{min} + 2(j - 1)(1 - \eta)\bar{g}/(k - 1). \tag{5.6}$$

Setting $\eta = 1$ gives balanced designs and $\eta < 1$, unbalanced. Here we define the mildly, moderately, and highly unbalanced designs by $\eta = 0.9, 0.5, 0.1$, respectively. When $\eta = 0$, both (5.5) and (5.6) produce extremely unbalanced designs. It is noted that the last step often gives decimals and rounding is required.

6. Simulation studies

The use of unbalanced designs broadens the scope of simulation studies and provides more choices in describing the behavior of the non-null F test.

6.1. **Observed size of non-null F test.** Experiment 1 was conducted to measure the observed size of the non-null F test in balanced and highly unbalanced designs. The margin was taken to be $\Delta_j = 0$ throughout this experiment, depending upon this reason. Inspection of (2.2) and (2.4) discloses that $\nu_j = \nu$ when H_0 holds and $\nu_j^* = \nu$ as well when H holds. So, the function rnorm(n_j,ν_j,σ) was written here as (n_j,ν,σ) for generating normal observations in the R Language. We now give a heuristic explanation for this.

Remark 6.1. *The observed size of the non-null F test is identical to that of the F test because the sampling scheme does not distinguish between the null and non-null hypothesis.*

We fixed the total sample size at $n = 48, 192, 768$, representing small, moderate, and large sample sizes. The number of groups was taken to be $k = 3, 6$. Recall the procedure derived in Subsection 5.2. Balanced and highly unbalanced designs were implemented by (5.5) and (5.6) with $\eta = 1, 0.1$. There were 18 possible combinations of the quantities n, k, and η. We set the significance level as $\alpha = 0, 0.0125, \ldots, 0.1$. The other parameters were specified as $\nu_{min} = 0.3$, $\mu_{max} = 0$, and $\sigma = 0.4$. For each case, 1000 data sets were generated. This number yields a 95% confidence interval for the nominal size $\alpha \pm 1.96(\alpha(1-\alpha)/1000)^{1/2}$. The observed test size is given by

$$\hat{\alpha} = \sum_{1}^{1000} I\{F \geq F_{1-\alpha,\, k-1,\, n-k,\, \lambda} \,|H \; is \; true\}/1000,$$

which is corresponding to the nominal size (5.4).

The values of the observed test size for $n = 48$ appear in the upper part of Table 6.1.

Table 6.1. Observed size of non-null F test in balanced and highly unbalanced designs

n_j	Significance level α								
	0	0.0125	0.025	0.0375	0.05	0.0625	0.075	0.0875	0.1
$n = 48$									
(16, 16, 16)	0	0.017	0.027	0.040	0.051	0.064	0.079	0.094	0.109
(30, 16, 2)	0	0.012	0.022	0.036	0.056	0.067	0.073	0.081	0.095
(2, 16, 30)	0	0.016	0.030	0.043	0.057	0.068	0.081	0.087	0.097
(8, 8, 8, 8, 8, 8)	0	0.013	0.027	0.035	0.046	0.057	0.071	0.090	0.103
(15, 12, 9, 7, 4, 1)	0	0.019	0.033	0.044	0.056	0.067	0.080	0.092	0.102
(1, 4, 7, 9, 12, 15)	0	0.018	0.023	0.042	0.057	0.068	0.078	0.088	0.105
$n = 192$									
(64, 64, 64)	0	0.015	0.029	0.043	0.055	0.069	0.084	0.097	0.110
(122, 64, 6)	0	0.018	0.032	0.045	0.060	0.074	0.077	0.089	0.097
(6, 64, 122)	0	0.012	0.032	0.044	0.059	0.071	0.082	0.097	0.116
(32, 32, 32, 32, 32, 32)	0	0.015	0.028	0.036	0.052	0.065	0.076	0.091	0.100
(61, 49, 38, 26, 15, 3)	0	0.013	0.022	0.032	0.048	0.061	0.072	0.085	0.099
(3, 15, 26, 38, 49, 61)	0	0.018	0.026	0.040	0.057	0.076	0.087	0.098	0.112
$n = 768$									
(256, 256, 256)	0	0.009	0.016	0.024	0.041	0.050	0.059	0.075	0.084
(486, 256, 26)	0	0.013	0.023	0.042	0.059	0.072	0.081	0.100	0.113
(26, 256, 486)	0	0.015	0.029	0.043	0.055	0.067	0.071	0.080	0.095
(128, 128, 128, 128, 128, 128)	0	0.013	0.024	0.037	0.049	0.056	0.068	0.077	0.088
(243, 197, 151, 105, 59, 13)	0	0.007	0.021	0.038	0.049	0.067	0.085	0.103	0.116
(13, 59, 105, 151, 197, 243)	0	0.010	0.022	0.032	0.041	0.059	0.069	0.081	0.097

The balanced designs are defined with $\eta = 1$ and the highly unbalanced designs, with $\eta = 0.1$ by (5.5) and (5.6). The minimum mean is $\nu_{min} = 0.3$. The maximum difference is $\mu_{max} = 0$. And the standard deviation is $\sigma = 0.4$.

The first three rows show the entries for $k = 3$. It is not surprising that the observed test size is near the nominal level in the balanced design $n_j = (16, 16, 16)$. Even in the highly unbalanced

design $n_j = (30, 16, 2)$ or $n_j = (2, 16, 30)$, however, we have not seen any appreciable variety: the observed size still lies within the most 95% confidence intervals for the nominal size. For example, the use of the 95% confidence interval for the nominal size 0.1 results in (0.081, 0.119), which contains the observed test size 0.095 and 0.097. A similar result holds for $k = 6$ as shown in the last three rows. The remaining parts show the results for $n = 192$ and $n = 768$, which are similar to those in the upper part. Notice that the group sample size n_j may not add up the total sample size n due to rounding.

Remark 6.2. *The observed size of the non-null F test is near the nominal level in balanced designs as well as in unbalanced designs.*

Further insights about Remark 6.2 will be given in Experiment 2 and 3. We did not attempt to mention mildly and moderately unbalanced designs because the variation of the observed size is not likely to exceed that in highly unbalanced designs.

6.2. **Observed power of non-null F test.** Experiment 2 was planned to assess the power estimates of the non-null F test for a given sample size in equivalence trials. The values of margin in Figure 1 were reused here. Notice that they are larger than or equal to the differences. We fixed the total sample size at $n = 200$ and the number of groups at $k = 8$. Using (5.5) or (5.6) with $\eta = 1$ and 0.1 yielded balanced and highly unbalanced designs, respectively. Concerning k-sample equivalence trials, we took $\alpha = 0.05$ in this experiment. In two-sample non-inferiority trials, however, one may choose $\alpha = 0.025$, as recommended by FDA (2016)[18] and International Guideline ICH E9 Hirotsu (2007)[19].

The first part of Table 6.2 reports the power resulting from (5.3) with $\nu_{min} = 0.3$, $\mu_{max} = 0, 0.04, \ldots, 0.4$, and $\sigma = 0.8$.

The column "$\mu_{max} = 0$" refers to the minimum value of the power, i.e., the nominal test size 0.05 as previously mentioned in (5.4). This experiment gives that the power of the non-null F test is higher than that of the F test in equivalence trials and as the margin gets larger, the power gets higher.

We now turn to the power estimates. For each set of experimental conditions, 1000 random samples were drawn from the function rnorm(n_j,ν_j,σ). This yields a 95% confidence interval for the power $1 - \beta^* \pm 1.96(\beta^*(1 - \beta^*)/1000)^{1/2}$. The power estimates are given by

$$1 - \hat{\beta}^* = \sum_{1}^{1000} I\{F^* \geq F_{1-\alpha,\, k-1,\, n-k,\, \lambda^*} \mid H_1 \; is \; true\}/1000,$$

where the statistic F^* is calculated from (3.3).

The second part of Table 6.2 gives the power estimates from balanced designs. The column "$\mu_{max} = 0$" shows the minimum value of power estimates, i.e., the observed test size 0.05, which is equal to the nominal level. Figure 4 displays this result.

Table 6.2. Power estimates of non-null F test in equivalence trials ($\beta \in (0,1)$)

Δ_j	Maximum difference μ_{max}										
	0	0.04	0.08	0.12	0.16	0.2	0.24	0.28	0.32	0.36	0.4
	$1-\beta^*$: $n_j = (25,25,25,25,25,25,25,25)$										
0	0.050	0.052	0.057	0.067	0.082	0.103	0.131	0.166	0.210	0.264	0.325
$2.2\mu_j$	0.050	0.053	0.061	0.076	0.098	0.131	0.174	0.231	0.299	0.380	0.468
$2.5\mu_j$	0.050	0.054	0.067	0.092	0.131	0.187	0.264	0.357	0.468	0.583	0.694
$3\mu_j$	0.050	0.057	0.082	0.131	0.210	0.325	0.468	0.621	0.760	0.867	0.937
	$1-\hat{\beta}^*$: $n_j = (25,25,25,25,25,25,25,25)$										
0	0.050	0.053	0.059	0.068	0.077	0.102	0.137	0.175	0.222	0.263	0.323
$2.2\mu_j$	0.050	0.055	0.061	0.076	0.109	0.142	0.184	0.237	0.303	0.369	0.441
$2.5\mu_j$	0.050	0.054	0.066	0.099	0.142	0.195	0.269	0.354	0.441	0.568	0.689
$3\mu_j$	0.050	0.059	0.088	0.142	0.215	0.326	0.441	0.602	0.750	0.854	0.934
	$1-\hat{\beta}^*$: $n_j = (48,41,35,28,22,15,9,2)$										
0	0.059	0.061	0.063	0.065	0.072	0.084	0.108	0.132	0.165	0.195	0.232
$2.2\mu_j$	0.059	0.062	0.069	0.080	0.094	0.115	0.139	0.177	0.205	0.242	0.290
$2.5\mu_j$	0.059	0.062	0.076	0.094	0.115	0.148	0.190	0.228	0.290	0.373	0.468
$3\mu_j$	0.059	0.067	0.088	0.115	0.163	0.214	0.290	0.407	0.525	0.657	0.753
	$1-\hat{\beta}^*$: $n_j = (2,9,15,22,28,35,41,48)$										
0	0.051	0.057	0.056	0.067	0.072	0.084	0.100	0.119	0.140	0.180	0.216
$2.2\mu_j$	0.051	0.049	0.049	0.064	0.075	0.100	0.123	0.154	0.187	0.239	0.278
$2.5\mu_j$	0.051	0.048	0.057	0.073	0.100	0.129	0.167	0.228	0.278	0.359	0.457
$3\mu_j$	0.051	0.047	0.070	0.100	0.145	0.208	0.278	0.389	0.517	0.641	0.750

The significance level is $\alpha = 0.05$. The minimum mean is $\nu_{min} = 0.3$. The standard deviation is $\sigma = 0.8$. The power $1-\beta^*$ is given by (5.3). And the power estimates $1-\hat{\beta}^*$ are the fractions of F^*-values greater than or equal to the critical F-values under H_1 in 1000 sets of samples.

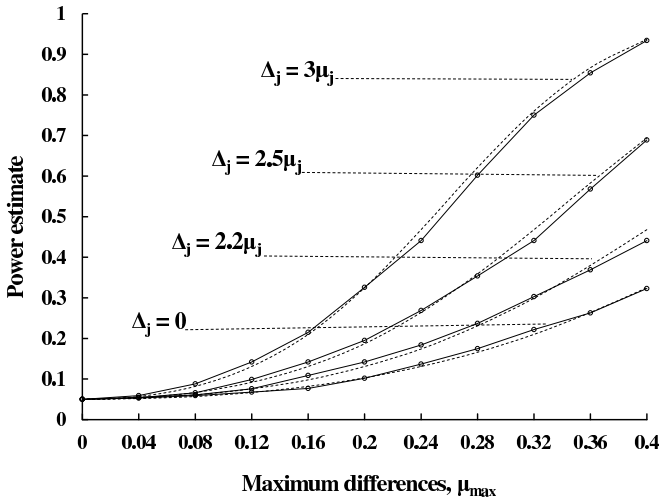

Figure 4. Power estimates versus power of non-null F test in balanced designs

The significance level is α = 0.05. The number of groups is k = 8. The minimum mean is ν_{min} = 0.3. The standard deviation is σ = 0.8. Solid curves represent the power estimates from 1000 times of simulations in the balanced designs: n_j = (25, 25, 25, 25, 25, 25, 25, 25). And dash curves indicate the power of the non-null F test given by (5.3).

On comparing the curves of the observed power and the power, we see that they almost overlap. It tells that the agreement between the observed power and the power is quite well for the non-null F tests in balanced designs.

The third part of Table 6.2 lists the power estimates from the highly unbalanced designs with descending n_j. Figure 5 portrays this result.

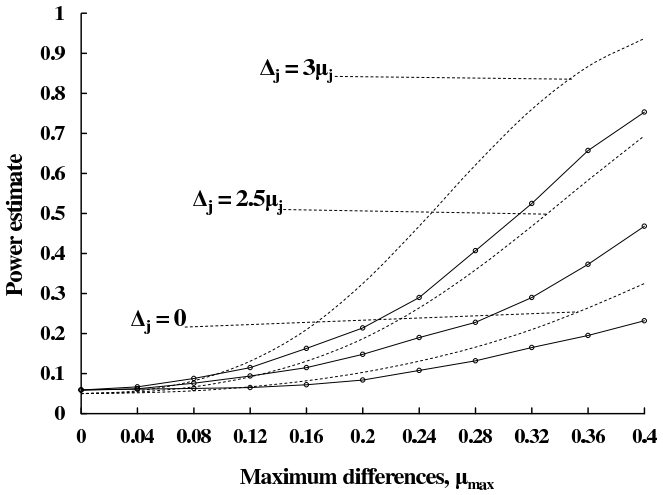

Figure 5. Power estimates versus power of non-null F test in highly unbalanced designs

The significance level is $\alpha = 0.05$. The number of groups is $k = 8$. The minimum mean is $v_{min} = 0.3$. The standard deviation is $\sigma = 0.8$. Solid curves represent the power estimates from 1000 times of simulations in the highly unbalanced designs: $n_j = (48, 41, 35, 28, 22, 15, 9, 2)$. And dash curves indicate the power of the non-null F test given by (5.3).

Clearly, the power estimates are lower than the power for both the null and non-null F tests. That is, the violation of balance may decrease the power. By contrast, we have not seen any apparent influence of unbalanced designs on the observed test size: The observed test size 0.059 lies between 0.036 and 0.064, the 95% confidence interval for the nominal size 0.05 as Remark 6.2 predicts.

A similar appearance holds for the results from the highly unbalanced designs with ascending n_j as shown in the fourth part of Table 6.2.

Experiment 3 considered the power estimates of the non-null F test for a given sample size in superiority trials. The values of margin in Figure 2 were reused here. Notice that they are smaller than or equal to the differences. The other parameter values were the same as those in Experiment 2 but $\sigma = 0.4$. Taking the same method as in Experiment 2 yielded the power and the observed power, which are laid out in Table 6.3.

In superiority trials, the power of the non-null F test is lower than that of the F test.

Looking at the column $\mu_{max} = 0$ of Table 6.3, the observed test size 0.05 in the balanced designs equals the nominal level 0.05 and the observed test size 0.059 and 0.051 in the highly unbalanced

Table 6.3. Power estimates of non-null F test in superiority trials ($\beta \in (0,1)$)

Δ_j	Maximum difference μ_{max}										
	0	0.04	0.08	0.12	0.16	0.2	0.24	0.28	0.32	0.36	0.4
	$1-\beta^*$: $n_j = (25,25,25,25,25,25,25,25)$										
0	0.05	0.057	0.082	0.131	0.210	0.325	0.468	0.621	0.760	0.867	0.937
$0.1\mu_j$	0.05	0.056	0.076	0.113	0.174	0.264	0.380	0.514	0.650	0.772	0.867
$0.2\mu_j$	0.05	0.055	0.070	0.098	0.144	0.210	0.299	0.408	0.529	0.650	0.760
$0.5\mu_j$	0.05	0.052	0.057	0.067	0.082	0.103	0.131	0.166	0.210	0.264	0.325
	$1-\hat{\beta}^*$: $n_j = (25,25,25,25,25,25,25,25)$										
0	0.05	0.059	0.077	0.137	0.222	0.323	0.471	0.607	0.747	0.859	0.943
$0.1\mu_j$	0.05	0.058	0.070	0.118	0.183	0.263	0.375	0.510	0.642	0.760	0.859
$0.2\mu_j$	0.05	0.057	0.068	0.093	0.148	0.222	0.299	0.398	0.521	0.642	0.747
$0.5\mu_j$	0.05	0.053	0.059	0.068	0.077	0.102	0.137	0.175	0.222	0.263	0.323
	$1-\hat{\beta}^*$: $n_j = (48,41,35,28,22,15,9,2)$										
0	0.059	0.063	0.072	0.108	0.165	0.232	0.324	0.435	0.538	0.644	0.742
$0.1\mu_j$	0.059	0.063	0.069	0.097	0.138	0.195	0.268	0.354	0.455	0.549	0.644
$0.2\mu_j$	0.059	0.063	0.063	0.080	0.119	0.165	0.214	0.280	0.366	0.455	0.538
$0.5\mu_j$	0.059	0.061	0.063	0.065	0.072	0.084	0.108	0.132	0.165	0.195	0.232
	$1-\hat{\beta}^*$: $n_j = (2,9,15,22,28,35,41,48)$										
0	0.051	0.056	0.072	0.100	0.140	0.216	0.294	0.408	0.531	0.645	0.755
$0.1\mu_j$	0.051	0.058	0.067	0.090	0.124	0.180	0.244	0.326	0.440	0.539	0.645
$0.2\mu_j$	0.051	0.059	0.067	0.083	0.111	0.140	0.197	0.257	0.340	0.440	0.531
$0.5\mu_j$	0.051	0.057	0.056	0.067	0.072	0.084	0.100	0.119	0.140	0.180	0.216

The significance level is $\alpha = 0.05$. The standard deviation is $\sigma = 0.4$. And the other parameter values are taken to be the same as those in Table 6.2.

designs lie within (0.036, 0.064) a 95% confidence interval for the nominal size 0.05 as predicted by Remark 6.2.

Experiment 2 and 3 were run in the range of $\beta \in (0,1)$. But our attention is often focused on $\beta = 0.1$, which will be adopted in Experiment A1 and A2 as shown in Appendix.

7. Worked examples

7.1. **Example for evaluating equivalence of group means.** Larsen and Marx (2018, p598)[17] introduced a study that tells whether the use of special walking exercises may help infants walk alone. The study was fulfilled by 23 infants being randomly assigned to one of four groups. Group 1 received special walking and placing exercises. Group 2 also had daily 12-minute exercises but were not given the special walking and placing exercises. Group 3 and 4 received no special instruction. Listed in Table 7.1 is the age (in months) at which each of the children first walk alone after seven-week training.

Table 7.1. Age when infants first walked alone (months)

Group 1	Group 2	Group 3	Group 4
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	-

Source: Larsen and Marx (2018), p598 [17].

Knowing that $\bar{Y}_{\cdot j} = (10.12, 11.38, 11.71, 12.35)$ with $\bar{Y}_{\cdot\cdot} = 11.35$ provides $\hat{\mu}_1 =$ -2.22, $\hat{\mu}_2 =$ -0.97, $\hat{\mu}_3 =$ -0.64, and $\hat{\mu}_4 = 0$ with $\bar{\mu} =$ -1.00. The ordinary analysis is performed with $\Delta_j = 0$ resulted in $Q_1 = 14.78$ and $Q_2 = 43.69$. The statistic $F = 2.14$ (P = 0.13) gives a non-significant result at the $\alpha = 0.05$ level.

The contrasts in this case $C_i = (\nu_1 - \nu_2, \nu_1 - \nu_3, \nu_1 - \nu_4, \nu_2 - \nu_3, \nu_2 - \nu_4, \nu_3 - \nu_4)$, $i = 1, 2, \ldots, 6$, are estimated as $\hat{C}_i =$ (-1.25, -1.58, -2.22, -0.33, -0.97, -0.64). Testing the sub-hypotheses from (3.4) gives the statistics $F_i =$ 2.04, 3.27, 5.87, 0.14, 1.13, 0.49 (P = 0.17, 0.09, 0.03, 0.71, 0,30, 0.49). The contrast C_3 with the maximum $|\hat{C}_3| = 2.22$ gives a significant result at the $\alpha = 0.05$ level: $F_3 = 5.87$ (P = 0.03). The others show non-significant results.

Now consider the equivalence of group means. If, for example, the margin is chosen as $\Delta_j = 2.3\hat{\mu}_j$, we have $\Delta_1 =$ -5.12, $\Delta_2 =$ -2.24, $\Delta_3 =$ -1.48, and $\Delta_4 = 0$ with the average $\Delta =$ -2.30. The treatment sum of squares under the non-null hypothesis is $Q_1^* = 24.97$. The error sum of squares is $Q_2^* = 43.69$, which equals Q_2 as (3.2) predicts. Using (3.3) gives $F^* = 3.62$ (P = 0.03) and then we reject H at $\alpha = 0.05$.

In testing the non-null sub-hypotheses H', we have $F_i^* =$ 1.45, 0.68, 0.01, 5.07, 2.10, 3.28 (P = 0.24, 0.42, 0.93, 0.04, 0.16, 0.09). Of these, the contrast C_4 with the minimum $|\hat{C}_4| = 0.33$ shows $F_4^* = 5.07$ (P = 0.04). We should reject H' at $\alpha = 0.05$ and conclude that the equivalence holds for the contrast C_4. The other contrasts are non-significant at $\alpha = 0.05$ and the results are indeterminate.

A closer look reveals that both Group 1 and 2 received the exercises and both Group 3 and 4 received no special instruction. So, we are justified to merge the former two into one treatment group and the latter two into one control group and make a simpler analysis using the two-sample non-null t test instead. The sample sizes are $n_j = (12, 11)$ with $g_j = (0.52, 0.48)$ and the sample means are $\bar{Y}_{\cdot j} = (10.75, 12.00)$ with $\hat{\mu} = -1.25$. The conventional analysis gives $Q_1 = 8.97$, $Q_2 = 49.50$, $F =$ 3,80 (P = 0.06), and $t = -1.95$ (P = 0.06) and thus the null hypothesis H_0 can not be rejected at $\alpha = 0.05$.

In considering the equivalence of the two means, if an appropriate margin is used as $\Delta = 2.1\hat{\mu}$, we have Δ = -2.62. The analysis under the non-null hypothesis gives $Q_1^* = 10.85$, $F^* = 4.60$ (P = 0.04), and $t^* = 2.15$ (the two-sided P = 0.04 and the one-sided P = 0.02). The non-null hypothesis is rejected at $\alpha = 0.025$. Using (3.6), the one-sided 97.5% confidence interval for $|\Delta|$ is computed as (0, 2.58), which fails to contain $|\Delta| = 2.62$ and so $H_0 : |\mu| = |\Delta|$ can be rejected at the $\alpha = 0.025$ level. It is in favor of such an interpretation that the two means are equivalent. Alternatively, calculating (3.7) gives (-2.58,0), the one-sided 97.5% confidence interval for Δ. Since the interval does not contain Δ = -2.62, $H_0 : \mu = \Delta$ can be rejected at the $\alpha = 0.025$ level. The interpretation turns to the non-inferiority, a profile of equivalence, implying that the mean of the treatment group is not less than the mean of control group.

7.2. **Example for evaluating clinical significance of differences.** Table 7.2 contains the transformation rates of lymphocytes from 24 healthy men in three age groups [20]. We wish to know whether there is any true difference of means among the three groups but the most interest is addressed in the differences between any pair of means, especially the clinical significance of differences.

Table 7.2. Transformation rates (%) of lymphocytes

Age (years)	Number of observations									
	1	2	3	4	5	6	7	8	9	10
11-20	58	61	61	62	63	68	70	70	74	78
41-50	54	57	57	58	60	60	63	64	66	-
61-75	43	52	55	56	60	-	-	-	-	-

Source: Sichuan Medical College (1981), p30 [20].

The sample sizes are n_j = (10, 9, 5) with g_j = (0.42, 0.38, 0.21) and the sample means are $\bar{Y}_{\cdot j}$ = (66.50, 59.89, 53.20) with $\bar{Y}_{\cdot\cdot} = 61.25$. Using (3.1) and (3.2) with $\Delta_j = 0$ yields $Q_1 = 616.31$ and $Q_2 = 662.19$. The F statistic is $F = 9.77$ (P = 0) by (3.3). One would reject H_0 at $\alpha = 0.05$. There are three contrasts in this case: $C_i = (\nu_1 - \nu_2, \nu_1 - \nu_3, \nu_2 - \nu_3)$, which are estimated as $\hat{C}_i$ = (6.61,13.30,6.69). Testing the sub-hypotheses H_0' gives F_i = (6.57, 18.70, 4.56) (P = 0.02, 0, 0.04). All the p-values imply that the differences between any pair of means are statistically significant at the level $\alpha = 0.05$, concluding that there is strong evidence that the transformation rates of lymphocytes in the three groups differ.

Now, look at the clinical significance of differences. Suppose $\Delta_j = 0.3\hat{\mu}_j$ is appropriate for the margin. Knowing that $\hat{\mu}_1 = 13.30$, $\hat{\mu}_2 = 6.69$, and $\hat{\mu}_3 = 0$ with $\bar{\mu} = 8.05$, we find $\Delta_1 = 3.99$, $\Delta_2 = 2.01$, and $\Delta_3 = 0$ with $\Delta = 2.41$. It follows the treatment sum of squares under the non-null hypothesis $Q_1^* = 301.99$ and the statistic $F^* = 4.79$ (P = 0.02). In testing the non-null sub-hypotheses H', the statistics are F_i^* = (2.64, 12.52, 1.86) (P = 0.12, 0, 0.19). We reject H' at

$\alpha = 0.05$ with respect to the contrast $C_2 = \nu_1 - \nu_3$ but not for the other two. In fact, the contrast C_2 has the maximum $|\hat{C_2}| = 13.30$. The proper interpretation is that the contrast C_2 has clinical significance but the other two are indeterminate.

7.3. **Example of randomized block design.** In a phase I trial [20], seven schistosomiasis patients received a test drug for three days. The level of serum alanine aminotransferase (ALT), a sensitive index of reflecting liver damage, was measured before and after treatment with a randomized block design. Table 7.3 lists the ALT levels.

Table 7.3. Levels of serum alanine aminotransferase (ALT) before and after treatment

Patients	Pre-treatment levels of ALT	Post-treatment levels of ALT				
		3 days	1 week	2 weeks	3 weeks	4 weeks
1	63	36	188	138	63	54
2	90	200	238	220	188	144
3	54	36	300	83	100	92
4	45	72	140	213	144	100
5	54	54	175	150	100	36
6	72	63	300	163	144	90
7	64	77	207	185	122	87

Source: Sichuan Medical College (1981), p31 [20].

With respect to the treatment effect, knowing that $\bar{Y}_{\cdot j} =$ (63.14, 76.86, 221.14, 164.57, 123.00, 86.14) with $\bar{Y}_{\cdot\cdot} = 122.48$, we find $\hat{\mu}_j =$ (-23.00, -9.29, 135.00, 78.43, 36.86, 0) with $\bar{\mu} = 36.33$. Concerning the block effect, we have $\bar{Y}_{i\cdot} =$ (90.33, 180.00, 110.83, 119.00, 94.83, 138.67, 123.67) with $\bar{Y}_{\cdot\cdot} = 122.48$ and $\hat{\mu}_i =$ (-33.33, 56.33, -12.83, -4.67, -28.83, 15.00,0) with $\bar{\mu}^b =$ -1.19.

Usual analysis with $\Delta_j = 0$ gives $Q = 202586.48$, $Q_t = 129003.33$, $Q_b = 33104.81$, and $Q_e = 40478.33$. By the use of (4.3), the F ratio for the treatment effect is $F_t = 19.12$ (P = 0). For the block effect, it is $F_b = 4.09$ (P = 0) from (4.4). Both of them are statistically significant.

With the highly statistical significance, one may further look for the clinical significance. For the treatment effect, if the appropriate margins are used as $\Delta_j = 0.3\hat{\mu}_j$, we have $\Delta_j =$ (-6.90, -2.79, 40.50, 23.53, 11.06,0) with $\Delta = 10.90$. Knowing that $Q^* = 136794.78$, $Q_t^* = 63211.63$, $Q_b = 33104.81$, and $Q_e = 40478.33$ provides $F_t^* = 9.37$ (P = 0). Turning to the block effect under the non-null hypothesis, letting $\Delta_i = 0.3\hat{\mu}_i$ gives $\Delta_i =$ (-10.00, 16.90, -3.85, -1.40, -8.65, 4.50,0) with $\Delta^b =$ -0.36. Then we have $Q^* = 185703.02$, $Q_t = 129003.33$, $Q_b^* = 16221.36$, and $Q_e = 40478.33$. By applying (4.4), we have $F_b^* = 2.00$ (P = 0.10). The small P-value for the treatment effect, yet the lack of significance for the block effect, states that the former is clinically significant at the 0.05 level but the latter not.

8. Discussion

The non-null hypothesis has a long history. A difference with clinical importance may be statistically non-significant and a statistically significant difference may be no real interest. Addressing this issue, Kirk (2001)[21] claimed that there are three questions concerning the estimated difference. First, is an observed result real or should it be attributed to chance (i.e., statistical significance)? Second, if the result is real, how large is it (i.e., effect size)? Third, is the result large enough to be meaningful and useful (i.e., clinical or practical significance)? The clinical significance is described in detail, for example, in Victor (1987)[6], Laupacis et al (1988)[22], or Kieser et al (2013)[7]. Such an idea even dates back to Kendal and Stuart (1979)[23]. Their concern concentrated on the two questions. The first is whether there is any true difference and the second is about its magnitude. The first question relates to the null hypothesis and the second, to the non-null hypothesis. The non-null hypothesis test applies to assessing the equivalence of group means or the clinical significance of differences. This has come to broaden the scope in analyzing clinical data.

It is believed that any test of the null hypothesis has a corresponding non-null version. Here are the non-null versions of the one- and two-way ANOVA F tests as well as the two-, one-sample t tests, and the paired t test. In the non-null generalization of the ANOVA F test, the treatment sum of squares changes with the margin but the grand mean, the sample variance, and the error sum of squares not (see Remark 2.1, 2.3). All the proposed tests enjoy the property of reducibility, meaning that they reduce to their classical counterparts on setting the margin equal to zero. Therefore, they are valid under both the null and non-null hypothesis.

The power of the non-null F test is higher than that of the F test in equivalence trials as described in Table 6.2. The larger the margin, the higher is the power. The required sample size of this test is smaller than that of the F test in equivalence trials, for which Appendix gives details. The results are just the contrary in superiority trials (see Table 6.3).

The observed size of the non-null F test is identical to that of the F test (see Remark 6.1) and is near the nominal level regardless of the design balanced or unbalanced as in Table 6.1, 6.2, and 6.3 (see also Remark 6.2).

The observed power of the non-null F test is close to the power as long as the design is balanced as shown in Figure 4. The power may suffer from severe violations of balance. When there are mild violations of balance, however, the observed power is still near the power as shown in the Appendix. It seems that mildly unbalanced designs would be somewhat tolerable in planning clinical trials though balanced designs are the best.

Specifying the margin is the single greatest challenge in the designing, implementation, and analysis of equivalence or non-inferiority trials, which is explained in detail in FDA (2016)[18]. This is clearly subjective, and generally relies on professional knowledge, references, academic

conferences, prior experiences, past trials, or pilot studies, etc. An appropriate valuation of the margin comes from practices and experiences. For convenience to describe the performance of the non-null F test, the margin is chosen larger than or equal to the difference for equivalence trials and smaller than or equal to the difference for superiority trials. Note that we specify $|\Delta_j| \geq 0$, $j = 1, 2, \ldots, k-1$, except the last item $j = k$, in which we set $\Delta_k = 0$ because $\mu_k = 0$. Note further that the sign of Δ_j is kept identical to that of μ_j. Situations sometimes arise where it is difficult to specify an appropriate margin. If necessary, one may even negotiate with the FDA (Rockville, Maryland, United States) about an acceptable boundary value.

To an increasing extent, active control trials are selected rather than placebo trials, especially, since the fifth edition of Helsinki Declaration published [24, 25, 1]. On the other hand, relative to the development of medical treatments, it is becoming difficult to develop more powerful drugs, hence one would be looking for new treatments that have the same efficacy but demonstrate better quality in other aspects. As a consequence, the tests of the non-null hypothesis will be increasingly useful.

Most equivalence trials or non-inferiority trials are planned in the two-sample format to compare a test drug with an active control to show that the test drug is either equivalent to or not worse than the active control. Hirotsu (2007)[26] illuminated the underlying association between these trials and superiority trials and derived a unifying approach to the three kinds of trials. Lesaffre (2008)[27] among others introduced the basic definitions of superiority, equivalence, and non-inferiority trials in a broader sense. Snapinn (2000) and Pater (2004) [28,29] among others laid the foundation for the important theory and methods of equivalence trials and non-inferiority trials. In practice, the most frequently used are non-inferiority trials, which are discussed in detail by D'Agostino Sr et al (2003)[30]. The guidance for non-inferiority trials [18] recommends the use of confidence intervals in data analysis, which is currently accepted [31, 32, 33, 34].

Clinical practice goes beyond the two-sample format when we need to compare several drugs, several doses, or several routes of administration. Then the topic becomes the k-sample equivalence trials and the inference procedures are hypothesis tests rather than confidence intervals.

The ANOVA F test of the non-null hypothesis is presented just for this context. In analyzing k-sample data, we need both the non-null hypothesis test (3.3) for overall means testing and the non-null sub-hypothesis tests (3.4) for each different pairwise comparison. In the two-sample case, the one-sided confidence interval (3.6) can be used for equivalence trials and (3.7) for non-inferiority trials. Subsection 3.3 provides insights into the relationship between non-inferiority trials and k-sample equivalence trials and Subsection 7.1 gives a numerical calculation for it. Of the two-way ANOVA non-null F tests, (4.3) is used to test the treatment effect and (4.4), the block effect. If the original observations are claimed to be normal after checking using Shapiro-Wilk test [35], so are the observations under the non-null hypothesis based on (2.7) (see also Remark 2.2). The non-null

F test is used when the data are numerical. In the case of categorical data, one may prefer the $r \times 2$ chi-square test of non-null hypothesis (Zhao 2015) instead.

With the non-null F test, it enables inferences to extend to the equivalence of group means in k-sample equivalence trials or the clinical significance of differences in clinical superiority trials.

References

[1] T.R. Fleming, Evaluation of active control trials in AIDS, J. Acquired Immune Deficiency Syndrome. 2 (1990), S82-S87.
[2] R. Temple, Problems in interpreting active control equivalence trials, Account. Res. 4 (1996), 267-275. https://doi.org/10.1080/08989629608573887.
[3] R.F. Haase, M.V. Ellis, N. Ladany, Multiple criteria for evaluating the magnitude of experimental effects, J. Counsel. Psychol. 36 (1989), 511-516. https://doi.org/10.1037/0022-0167.36.4.511.
[4] G. Greenstein, Clinical versus statistical significance as they relate to the efficacy of periodontal therapy, J. Amer. Dental Assoc. 134 (2003), 583-591. https://doi.org/10.14219/jada.archive.2003.0225.
[5] H.C. Kraemer, G.A. Morgan, N.L. Leech, J.A. Gliner, J.J. Vaske, R.J. Harmon, Measures of clinical significance, J. Amer. Acad. Child Adolesc. Psych. 42 (2003), 1524-1529.
[6] N. Victor, On clinically relevant differences and shifted nullhypotheses, Methods Inf. Med. 26 (1987), 109-116. https://doi.org/10.1055/s-0038-1635499.
[7] M. Kieser, T. Friede, M. Gondan, Assessment of statistical significance and clinical relevance, Stat. Med. 32 (2012), 1707-1719. https://doi.org/10.1002/sim.5634.
[8] E.S. Pearson, "Student" as statistician, Biometrika. 30 (1939), 210-250. https://doi.org/10.2307/2332648.
[9] I.J. Good, The Bayes/Non-Bayes compromise: A brief review, J. Amer. Stat. Assoc. 87 (1992), 597-606. https://doi.org/10.1080/01621459.1992.10475256.
[10] F.H.C. Marriott, A dictionary of statistical terms, 5th ed., Longman Scientific and Technical, Harlow (1990).
[11] C.W. Dunnett, M. Gent, Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables, Biometrics. 33 (1977), 593. https://doi.org/10.2307/2529457.
[12] G. Zhao, Tests of non-null hypothesis on proportions for stratified data, Stat. Med. 27 (2007), 1429-1446. https://doi.org/10.1002/sim.3023.
[13] G. Zhao, A test of non null hypothesis for linear trends in proportions, Comm. Stat. - Theory Meth. 44 (2013), 1621-1639. https://doi.org/10.1080/03610926.2013.776687.
[14] R.A. Fisher, Statistical methods for research workers, Oliver and Boyd, Edinburgh, 1925.
[15] W.N. Venables, D.M. Smith, the R Core Team, An introduction to R notes on R: A programming environment for data analysis and graphics, Version 3.6.1., The R Foundation for Statistical Computing (R-core@R-project.org) (2019).
[16] W.J. Killoy, The clinical significance of local chemotherapies, J. Clin. Periodontol. Supplement 2 (2002), 22-29.
[17] R.J. Larsen, M.L. Marx, An introduction to mathematical statistics and its applications, Sixth edition, Pearson, Boston, 2018.
[18] U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry. 3-6 (2016).
[19] International Conference on Harmonisation. Guidance E9: statistical principles for clinical trials. Fed Register. 63 (179), (1998).
[20] Sichuan Medical College, Health statistics, First edition, Beijing: People's Health Publishing House, 30-31, (1981).
[21] R.E. Kirk, Promoting good statistical practices: some suggestions, Educ. Psychol. Measure. 61 (2001), 213-218. https://doi.org/10.1177/00131640121971185.
[22] A. Laupacis, D.L. Sackett, R.S. Roberts, An assessment of clinically useful measures of the consequences of treatment, N. Engl. J. Med. 318 (1988), 1728-1733.
[23] S.W. Kendal, A. Stuart, The advanced theory of statistics, Volume 2, Charles Griffin and Company Limited, London, p175, (1979).
[24] R. Temple, Difficulties in evaluating positive control trials, In: Proceedings of the Biopharmaceutical Section of the American Statistical Association, 1-7 (1983).
[25] T.R. Fleming, Treatment evaluation in active control studies, Cancer Treat. Rep. 71 (1987), 1061-1065.
[26] C. Hirotsu, A unifying approach to non-inferiority, equivalence and superiority tests via multiple decision processes, Pharm. Stat. 6 (2007), 193-203. https://doi.org/10.1002/pst.305.
[27] E. Lesaffre, Superiority, equivalence, and non-inferiority trials, Bull. NYU Hosp. Joint Dis. 66 (2008), 150-154.
[28] S.M. Snapinn , Noninferiority trials, Curr. Controlled Trials Cardiov. Med. 1 (2000), 19-21.

[29] C. Pater, Equivalence and noninferiority trials - are they viable alternatives for registration of new drugs? (III), Curr. Controlled Trials Cardiov. Med. 5 (2004), 1-7.
[30] R.B. D'Agostino Sr., J.M. Massaro, L.M. Sullivan, Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics, Stat. Med. 22 (2002), 169-186. https://doi.org/10.1002/sim.1425.
[31] N. Le Saux, A randomized, double-blind, placebo-controlled noninferiority trial of amoxicillin for clinically diagnosed acute otitis media in children 6 months to 5 years of age, Can. Med. Assoc. J. 172 (2005), 335-341. https://doi.org/10.1503/cmaj.1040771.
[32] Piaggio G., Elbourne D.R., Altman D.G., Pocock S.J., Evans S.J.W., Reporting of noninferiority and equivalence randomized trials an extension of the consort statement, *JAMA*, **295**(10): 1152-1160 (2006) DOI:10.1001/jama.2012.87802
[33] S. Harbarth, E. von Dach, L. Pagani, M. Macedo-Vinas, B. Huttner, F. Olearo, S. Emonet, I. Uckay, Randomized non-inferiority trial to compare trimethoprim/sulfamethoxazole plus rifampicin versus linezolid for the treatment of MRSA infection, J. Antimicr. Chemoth. 70 (2014), 264-272. https://doi.org/10.1093/jac/dku352.
[34] G. Buzançais, C. Roger, S. Bastide, P. Jeannes, J.Y. Lefrant, L. Muller, Comparison of two ultrasound guided approaches for axillary vein catheterization: a randomized controlled non-inferiority trial, Br. J. Anaesthesia. 116 (2016), 215-222. https://doi.org/10.1093/bja/aev458.
[35] J.P. Royston, An extension of shapiro and Wilk's W test for normality to large samples, Appl. Stat. 31 (1982), 115. https://doi.org/10.2307/2347973.

Appendix: Observed power of non-null F test for given type II error

Experiment A1 investigated the power estimates of the non-null F test for a given type II error in equivalence trials. The values of margin in Figure 1 were taken again. The experiment was carried out at $\alpha = 0.05$ and $\beta = 0.1$. The sample size n_j was determined from the process described in Subsection 5.2 with $k = 4$, $\nu_{min} = 0.3$, $\mu_{max} = 0.16$, $\sigma = 0.4$, and $\eta = 1, 0.9, 0.5, 0.1$. Then we obtained the balanced, mildly, moderately, and highly unbalanced designs in descending or ascending series, respectively.

The experiment consisted of generating the observations Y_{ij} using the function rnorm(n_j,ν_j,σ). In applying the method in Subsection 3.1, we obtained the statistic F^*. Each simulation is based on 1000 replications and yields a 95% confidence interval for the power $1-\beta = 0.9$: $1-\beta \pm 1.96(\beta(1-\beta)/1000)^{1/2} = (0.881, 0.919)$. Following the same procedure that is used for Experiment 2, we obtained the observed power.

The left half of Table A.1 covers the power estimates for the descending group sample sizes.

Table A.1. Power estimates of the non-null F test in equivalence trials ($\beta = 0.1$)

n	Descending n_j				$1-\hat{\beta}^*$	Ascending n_j				$1-\hat{\beta}^*$
	n_1	n_2	n_3	n_4		n_1	n_2	n_3	n_4	
					$\Delta_j = 0$					
642	160	160	160	160	0.891	160	160	160	160	0.891
	176	166	155	144	0.903	144	155	166	176	0.899
	241	187	134	80	0.857	80	134	187	241	0.842
	305	209	112	16	0.649	16	112	209	305	0.646
					$\Delta_j = 2.2\mu_j$					
447	112	112	112	112	0.910	112	112	112	112	0.910
	123	115	108	101	0.890	101	108	115	123	0.894
	168	130	93	56	0.834	56	93	130	168	0.839
	212	145	78	11	0.634	11	78	145	212	0.646
					$\Delta_j = 2.5\mu_j$					
287	72	72	72	72	0.900	72	72	72	72	0.900
	79	74	69	65	0.913	65	69	74	79	0.904
	108	84	60	36	0.852	36	60	84	108	0.842
	137	93	50	7	0.645	7	50	93	137	0.645
					$\Delta_j = 3\mu_j$					
163	41	41	41	41	0.909	41	41	41	41	0.909
	45	42	39	37	0.903	37	39	42	45	0.897
	61	48	34	20	0.845	20	34	48	61	0.836
	78	53	29	4	0.645	4	29	53	78	0.648

The balanced designs are defined by (5.5) with $\eta = 1$. The descending mildly, moderately, and highly unbalanced designs are defined by (5.5) with $\eta = 0.9, 0.5, 0.1$ and the ascending designs by (5.6). The significance level is $\alpha = 0.05$. The P(type II error) is $\beta = 0.1$. The minimum mean is $\nu_{min} = 0.3$. The maximum difference is $\mu_{max} = 0.16$. And the standard deviation is $\sigma = 0.4$.

The first rows in each part refer to the balanced designs. The observed power is close to 0.9. The same is true for the mildly unbalanced designs as shown in the second rows in each part. With the moderately or highly unbalanced designs, however, the observed power degenerates as seen in the third and fourth rows. The right half of the table refers to ascending n_j. The results are similar

to those in the left. Furthermore, Table A1 states that the required sample size of the non-null F test is smaller than that of the F test in equivalence trials.

Experiment A2 studied the power estimates of the non-null F test for a given type II error in superiority trials. The margin was the same as that seen in Figure 2. Following the same steps that were taken in Experiment A1, we obtained the power estimates for $\beta = 0.1$ at $\alpha = 0.05$ with $\mu_{max} = 0.33$. The resulting power estimates are summarized in Table A.2.

Table A.2. Power estimates of the non-null F test in superiority trials ($\beta = 0.1$)

n	Descending n_j				$1-\hat{\beta}^*$	Ascending n_j				$1-\hat{\beta}^*$
	n_1	n_2	n_3	n_4		n_1	n_2	n_3	n_4	
					$\Delta_j = 0$					
154	38	38	38	38	0.901	38	38	38	38	0.901
	42	40	37	35	0.887	35	37	40	42	0.899
	58	45	32	19	0.843	19	32	45	58	0.837
	73	50	27	4	0.636	4	27	50	73	0.644
					$\Delta_j = 0.1\mu_j$					
189	47	47	47	47	0.907	47	47	47	47	0.907
	52	49	46	43	0.896	43	46	49	52	0.893
	71	55	39	24	0.857	24	39	55	71	0.864
	90	61	33	5	0.645	5	33	61	90	0.661
					$\Delta_j = 0.2\mu_j$					
238	60	60	60	60	0.904	60	60	60	60	0.904
	66	62	58	54	0.899	54	58	62	66	0.900
	89	69	50	30	0.846	30	50	69	89	0.855
	113	77	42	6	0.640	6	42	77	113	0.681
					$\Delta_j = 0.5\mu_j$					
604	151	151	151	151	0.912	151	151	151	151	0.912
	166	156	146	136	0.900	136	146	156	166	0.900
	226	176	126	75	0.848	75	126	176	226	0.840
	287	196	106	15	0.658	15	106	196	287	0.658

See Table 6.2 for definitions of designs. The significance level is $\alpha = 0.05$. The P(type II error) is $\beta = 0.1$. The number of groups is $k = 4$. The minimum mean is $\nu_{min} = 0.3$. The maximum difference is $\mu_{max} = 0.33$. The margins are $\Delta_j = 0, 0.1\mu_j, 0.2\mu_j, 0.5\mu_j$. And the standard deviation is $\sigma = 0.4$.

We have seen that the manner of variation in the observed power is just the same as that in Experiment A1. For another, the required sample size of the non-null F test is larger than that of the F test in superiority trials.