Chapter 5

MAXIMUM RESISTANCES OF THE FRIEDMAN TWO-WAY ANALYSIS OF VARIANCE BY RANKS

§ 5.1 Introduction

A randomized complete block design (RCBD) is a design that tries to isolate the homogeneous experimental units so that we can randomly assign the treatments to these units. Each set of these homogeneous experimental units is called a *block*. We can think of blocks as a second factor, another source of variation in the responses, although the primary analyses are carried out only on the treatments (e.g. testing of hypotheses and post hoc tests). The purpose of blocking is to reduce the experimental error or extraneous variation. Assuming that we have t treatments and b blocks, the classical way to analyze such a layout is via the analysis of variance approach. The RCBD layout has each treatment randomly assigned *once* in each block to an experimental unit, i.e. each block only has a single replicate of a treatment. The observed

responses are both a function of the treatment and a function of the block. The model is written as

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij},\ i = 1...b,\ \ j = 1...t$$

where y_{ij} is the observation corresponding to block *i* and treatment *j*, μ is an overall grand mean of the observations, β_i is the i^{th} block effect, τ_j is the j^{th} treatment effect, and ε_{ij} is a random error term. The standard hypothesis tested is

$$H_0 : \tau_j = 0 \quad \forall j$$

$$H_1 : \text{at least one nonzero } \tau_j .$$

We reject the null hypothesis of no treatment effects if we obtain an observed *F*-value greater than a critical $F_{(t-1,\,(t-1)(b-1)),\alpha}$, where α is the specified level of the test. While this is the classical way of analyzing data from this layout, there are some underlying assumptions that must be met for this test to be optimal. The assumptions that are made on the data are that the errors are i.i.d. and normal with mean zero and constant variance σ^2, and that there does not exist a block by treatment interaction. If one or more of these assumptions are violated, then this is not the proper or optimal procedure to analyze data from such an experiment.

§ 5.2 The Friedman Two-Way Analysis of Variance by Ranks

One nonparametric analogue to the classical ANOVA is Friedman's Two Way Analysis of Variance by Ranks (Friedman, 1937). The layout, model, and hypotheses are the same as described above, but instead of using and analyzing the actual observations, they are replaced by their respective ranks. Specifically, we rank the observations *within each block* from 1 to *t* for

all *b* blocks, and analyze the treatment differences in the ranks. Friedman's test statistic X_F^2 is defined as (along with the computational form):

$$X_F^2 = \frac{12}{tb(t+1)} \sum_{j=1}^{t} \left[R_{\cdot j} - \frac{b(t-1)}{2} \right]^2 = \frac{12}{tb(t+1)} \sum_{j=1}^{t} R_{\cdot j}^2 - 3b(t+1).$$

Again, *t* and *b* are defined as before and $R_{\cdot j}$ is the sum of ranks for the j^{th} treatment across all blocks. The asymptotic distribution of X_F^2 has been shown to be chi-square with *t* - 1 degrees of freedom, and we reject the null hypothesis of no treatment difference if X_F^2 is greater than a critical $\chi^2_{(t-1),\alpha}$, for an α-level test. The assumptions for this test are not as stringent as with the classical analysis of variance. For example, the measurement scale must be at least ordinal and the distribution of the errors must be continuous. This test is generally more powerful than the standard *F*-test in ANOVA when the data are nonnormal, and/or outliers are present.

§ 5.3 Maximum Resistance

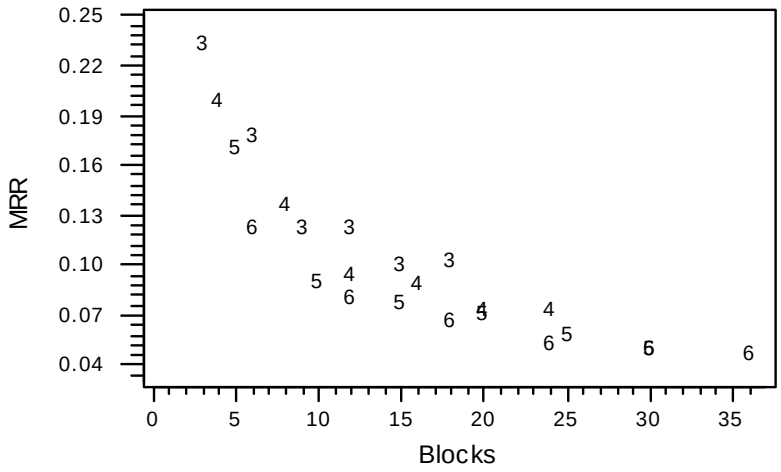
Restated again, the maximum resistance is a measure of how robust a test is when the data are in the least favorable position with regards to a desired conclusion. For a given layout, finding the maximum resistance entails starting the test statistic at the least favorable value, then one by one contaminating the sample with bad measurements until the desired conclusion is obtained. One such scheme, called the addition contamination scheme, would involve adding bad data to the already existing sample, thus increasing the sample size. The replacement contamination scheme would involve replacing existing observations with bad measurements,

thus keeping the sample size constant. As mentioned previously, the replacement contamination scheme is a more realistic scenario and used more frequently in other measures of robustness. This will be the scheme used in this research. Mathematically, the maximum resistance is the ratio of the maximum number of contaminants necessary to break down a test to the total sample size. We begin by investigating the maximum resistance to rejection for the Friedman Test.

§ 5.4 The Maximum Resistance to Rejection

As mentioned in Chapter 1, Ghassemian examined the resistance for the Friedman test with regards to voltage calibration in power systems (1997). For voltages around transformers, there may exist more than one meter value for the same voltage level. Thus, assume that there are t of these meter values. These meter values are taken over sequential time periods, and assume that there are b time periods. The Comparative Voltage Calibration method (CVC) is a method of finding the largest voltage cluster of meter values that are alike. Here we have a two way layout where the meters are the treatments, and the time periods are the blocks, and the desired conclusion is to *fail to reject* the null hypothesis of equal meters. If this hypothesis is rejected, then a multiple comparison procedure is employed to find out which meters do agree. Of the meters that do agree, the median value is taken as the calibrated value. These meter values, however, are naturally very noisy data, which would suggest using a nonparametric procedure such as the Friedman test for calibration. (One of the assumptions in using the Friedman test is that the blocks are mutually independent of each other. This is clearly not the case for the voltage calibration problem; the blocks are correlated over time. However, Jensen and Hui (1982) showed that there are certain types of exchangeable dependencies across blocks that are permitted. However, the dependency does not impact any resistance calculations.)

Ghassemian advocated the use of the Friedman test, but observed that this test is not nearly as robust as had been hoped and the test failed in certain scenarios. This prompted the question of exactly how resistant the Friedman test is. He briefly studied the maximum resistance to rejection of the Friedman test for certain cases and developed an iterative algorithm to solve for the number of contaminants necessary to break down the test. Unfortunately, he had not mathematically formulated the maximum resistance to rejection and his work stopped at the algorithm. We pick up where he left off and derive the maximum resistance to rejection for the Friedman test. As an example of some of the maximum resistance calculations, Figure 5.1 shows how the maximum resistance to rejection decreases as the number of blocks increases, for a fixed number of treatments, with the number of treatments indicated by the numeric value in the plot. We start by analyzing the case where the number of treatments is equal to the number of blocks and then expand to the case where the number of blocks is an integer multiple of the number of treatments. We then derive the resistance as a function of t, b, and $\chi^2_{t-1,\alpha}$.



Note: Values of the blocks are 1-6 times the number of treatments

Figure 5.1. Friedman Maximum Resistance to Rejection for 3-6 Treatments

§ 5.4.1 The Maximum Resistance to Rejection (*b* = *t* case)

To obtain the maximum resistance to rejection (*MRR*), we need to start the test statistic X_F^2 as small as possible, *and* have a configuration of the ranks that is most resistant to perturbations in the data. By looking at the computational form of the test statistic, to increase X_F^2, we need to increase the sum of the squared rank sums. So, we desire to find an optimal configuration of the ranks that is most resistant to perturbations in the data, and we want to find the fastest way to increase the sum of the squared rank sums. This is a minimax problem where we want to minimize over all configurations of the ranks the maximum sum of squared rank sums. If we first restrict our attention to the case where we have the same number of blocks as the number of treatments (i.e. $b = t$), then the optimal (most resistant) configuration of the ranks is where the ranks are set in a Latin Square type arrangement. For example, in a four-treatment and four-block layout, Table 5.1 shows one type of Latin Square.

Table 5.1. Most Resistant Configuration of the Ranks for Rejection (b = t = 4).

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
Block 1	1	2	3	4
Block 2	4	1	2	3
Block 3	3	4	1	2
Block 4	2	3	4	1
$\sum_{i=1}^{t} R_{ij}$	10	10	10	10

The optimal manner to increase the ranks in the range of interest is to contaminate the ranks within just one treatment, and without loss of generality, we will work with the first treatment. It is here that it should be noted that for some small designs, this might not be the optimal scheme of contamination for certain α-levels, due to the fact that these small samples do not provide much statistical power for rejection. However, these are anomalies with regard to the contamination scheme, as witnessed with all other cases studied. The range of interest is the value of X_F^2 from zero until the critical χ^2 value. To begin, the contamination scheme starts by finding the '1' in the treatment that we want to manipulate and changing it to the highest possible rank, which would be '*t*', and adjusting all other ranks within that block accordingly. So, after one contaminant in the above example (we define the number of contaminants as *m*), the configuration of ranks and rank sums would look as follows in Table 5.2. The second step would then find the '2' in the same treatment and change it to a *t*, and so on until rejection. Table 5.3 shows the example above after the second contamination.

Table 5.2. Configuration of Ranks After One Contaminant

$m = 1$	Treatment 1	Treatment 2	Treatment 3	Treatment 4
Block 1	4	1	2	3
Block 2	4	1	2	3
Block 3	3	4	1	2
Block 4	2	3	4	1
$\sum_{i=1}^{t} R_{ij}$	13	9	9	9

Table 5.3. Configuration of Ranks After Two Contaminants

$m = 2$	Treatment 1	Treatment 2	Treatment 3	Treatment 4
Block 1	4	1	2	3
Block 2	4	1	2	3
Block 3	3	4	1	2
Block 4	4	2	3	1
$\sum_{i=1}^{t} R_{ij}$	15	8	8	9

The pattern of the ranks sums can be characterized after *m* contaminations as a strictly increasing function in the first treatment, a decreasing function in (*t* - *m*) treatments, and a decreasing function up to a point in (*m* - 1) treatments. With this in mind, we succinctly formalize the sum of the squared rank sums after *m* contaminations.

Lemma 5.1: For the case where the number of blocks, *b*, is equivalent to the number of treatments, *t*, the ranks are in a Latin Square type configuration, and the optimal contamination scheme is used, then the sum of squared rank sums after *m* contaminants can be expressed as:

$$f_R(m) = \sum_{j=1}^{t} R_{\cdot j}^2 = \left[\overline{R} + \sum_{i=1}^{m}(t-i)\right]^2 + (t-m)(\overline{R}-m)^2 + \sum_{i=1}^{m-1}(\overline{R}-i)^2, \qquad (5.1)$$

where in general, $\overline{R} = \dfrac{b(t+1)}{2}$, the hypothesized rank sum for each treatment.

Proof of Lemma 5.1:

We look at each term of the equation, going from left to right. At $m = 0$, every rank sum is equal to $\overline{R}$. For the one treatment that is being contaminated, the rank sum increases by *t*-1 after one contaminant, *t*-1 + *t*-2 after 2 contaminants, and so on. Thus after *m* contaminants, the increase is *t*-1 + *t*-2 +…+ *t*-*m*, which is succinctly written as

$$\overline{R} + \sum_{i=1}^{m} (t - i).$$

Also it is easy to see that after the first contaminant, that treatment's rank sum increases by *t*-1 and the *t*-1 other ranks sums all decrease by 1. After the second contaminant is added, *t*-2 rank sums decrease by 1 and one stays constant at its previous value, and will stay constant throughout future contaminations, and so on. Thus, after *m* contaminants, *t-m* rank sums have decreased a total of *m*, which is expressed as

$$(t - m)(\overline{R} - m).$$

Of the remaining *m*-1 rank sums, the first became constant after 1 contaminant, the second after 2 contaminants, and so on, and can be stated as

$$\sum_{i=1}^{m-1} (\overline{R} - i).$$

Finally, by squaring all the terms involving $\overline{R}$, this yields the sum of squared ranks sums after *m* contaminants as

$$\sum_{j=1}^{t} R_{\cdot j}^2 = \left[\overline{R} + \sum_{i=1}^{m} (t - i) \right]^2 + (t - m)(\overline{R} - m)^2 + \sum_{i=1}^{m-1} (\overline{R} - i)^2,$$

and this completes the proof.

Our objective is to find the number of contaminants, *m*, such that the test will reject the null hypothesis. By setting the Friedman test statistic equal to the critical chi-square value, the minimum value of the test statistic in order to reject, and manipulating the terms, we can use Lemma 5.1 to obtain a function in terms of *m*, the number of contaminants necessary for rejection. Mathematically, we have

$$X_F^2 = \frac{12}{tb(t+1)} \sum_{j=1}^{t} R_{\cdot j}^2 - 3b(t+1) = \chi^2_{(t-1),\alpha},$$

and after manipulating the terms, we see that

$$\sum_{j=1}^{t} R_{\cdot j}^2 = \frac{\chi^2_{(t-1),\alpha} + 3b(t+1)}{12}\left[tb(t+1)\right].$$

Since we are assuming the number of blocks, *b*, is equal to the number of treatments, *t*, this simplifies to

$$\sum_{j=1}^{t} R_{\cdot j}^2 = \frac{\chi^2_{(t-1),\alpha} + 3t(t+1)}{12}\left[t^3 + t^2\right], \text{ or}$$

$$f_R(m) = c_\alpha \;\Leftrightarrow\; f_R(m) - c_\alpha = 0.$$

The subscript '*R*' denotes 'rejection', and 'α' denotes the level of the test, which determines the critical constant. We can substitute (5.1) for the sum of the squared ranks sums, subtract the constant determined by *t*, the chi-square critical value, and α, and solve for *m* such that $f_R(m) - c_\alpha = 0$. By expanding the summations in (5.1), we can rewrite the function as

$$f_R(m) = \frac{m^4}{4} - \frac{m^3(1+6t)}{6} + m^2\left(-\frac{1}{4} + t^2\right) + \frac{m}{6} + t\overline{R}^2,$$

which is a quartic in *m*. Thus for any number of treatments *t* and blocks *b*, such that *b* = *t*, and corresponding $\chi^2_{(t-1),\alpha}$ critical value, we can solve for the roots in the equation $f_R(m) - c_\alpha = 0$. Since this equation is a polynomial of order four, four roots exist and we are specifically interested in the *lowest positive real valued* root. This particular root will represent the lowest number of bad data necessary to switch an acceptance to a rejection. If the test is able to reject the null hypothesis for a specified α-level using the contamination scheme described earlier,

then this root *will* be real valued. A typical plot of the function appears in Figure 5.2 (here, $t = b = 5$), and the correct solution for m is the lowest positive root ($m \approx 3.46$).

Since the polynomial in m is of order four, we know a closed form expression of the root exists based on Galois theory. Using Galois theory, if we define

$$\alpha = \frac{1+6t}{6},\ \beta = 1-4t^2+\frac{1}{9}(1+6t)^2+\frac{1}{3}(-1+4t^2),\ \chi = 2^{1/3}(7+144c_\alpha^*+24t-24t^2+48t^4)$$

$$\delta = 3(162+9072c_\alpha^*-648t+15552tc_\alpha^*+1080t^2+15552t^2c_\alpha^*+2592t^3-2592t^4+3456t^6+$$
$$\left([-4(432c_\alpha^*+12(1+6t)+9(-1+4t^2)^2)^3+(324+1296c_\alpha^*(1+6t^2)-7776(-1+4t^2)+108(1+6t)(-1+4t^2)\right.$$
$$\left.+54(-1+4t^2)^3)^2]^{1/3}\right)^{1/2}$$

$$\gamma = -\frac{16}{3}+\frac{8}{27}(1+6t)^3-\frac{8}{3}(1+6t)(-1+4t^2),$$

and

$$c_\alpha^* = -\frac{3t(t+1)+\chi^2_{(t-1),\alpha}}{12}t^2(t+1)+t\overline{R}^2,$$

then formally we state the following.

Theorem 5.1: Assume the number of blocks, b, is equivalent to the number of treatments, t, and the ranks are in a Latin Square configuration. Then for the Friedman test, the minimum number of contaminants necessary to change an acceptance to a rejection is:

$$P_t^F = \alpha-\frac{1}{2}\left(\beta+\frac{\chi}{\delta}+\frac{\delta}{27(2^{1/3})}\right)^{1/2}+\frac{1}{2}\left(2\beta-\frac{\chi}{\delta}-\frac{\delta}{27(2^{1/3})}-\frac{\gamma}{4\left(\beta+\chi/\delta+\delta/27(2^{1/3})\right)^{1/2}}\right)^{1/2}$$

and the maximum resistance to rejection equals

$$\rho_t^F = \frac{\left\lfloor P_t^F+1\right\rfloor_{gif}}{t^2}.$$

The subscript '*t*' indicates the number of blocks, the superscript '*F*' refers to Friedman, and the subscript '*gif*' stands for '*greatest integer function*'.

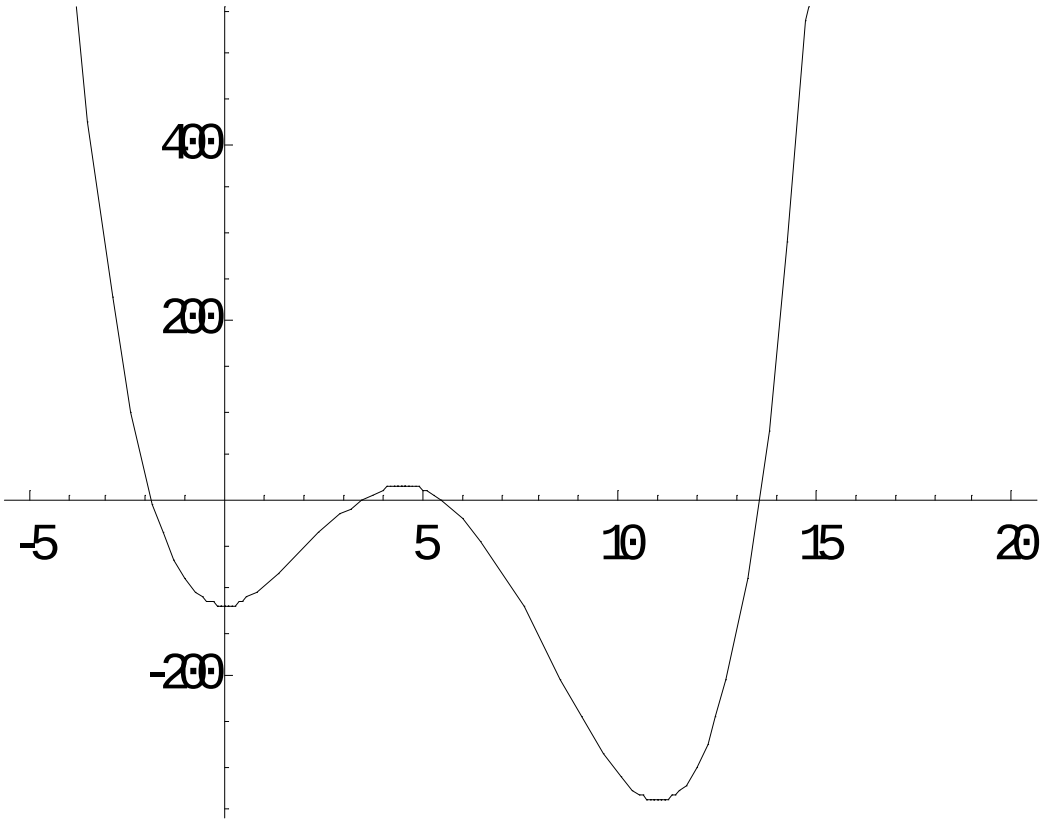


Figure 5.2. Plot of $f_R(m) - c_\alpha$ vs. m *(b = t =5)*

§ 5.4.2 The Maximum Resistance to Rejection (*b* = *nt* case)

For the case where we have an integer multiple of treatments as blocks (i.e. $b = nt$, $n \in Z^+$), the least favorable initial configuration of the ranks would consist of *n* Latin Squares. The optimal manner to contaminate the ranks is to focus on one treatment and as before, change each 1 to *t* (there are *n* of these), then change the 2's to *t*'s, etc. We can again formalize the sum of squared rank sums for this scenario after *m* contaminations as (this is without proof):

$$f_R(m) = \sum_{j=1}^{t} R_{\cdot j}^2 = \left[\overline{R} + \sum_{j=1}^{t-1} \sum_{i=(j-1)n+1}^{jn} (t-j) I(i \le m) \right]^2 + \sum_{j=2}^{t} \left[\overline{R} - \sum_{i=1}^{m} I(i \le (j-1)n) \right]^2,$$

where $I(.)$ is an indicator function and $\overline{R}$, defined as in equation (5.1), = $nt(t+1)/2$ (since $b = nt$). Again we would like to expand the summations, subtract the constant and solve for the root of the function $f_R(m) - c_\alpha = 0$. However, this is difficult to do with the form of the above equation due to the fact that the range for *i* involves *j* (*i* is dependent on *j*), as well as the fact that the formula involves indicator variables. To alleviate this problem, we rewrite the above expression in terms of just one index, and in terms of greatest integer functions. Specifically, we have

Lemma 5.2: For the case where the number of blocks, *b*, is equivalent to a constant multiple of the number of treatments, *t*, such that $b = nt$, $n \in Z^+$, the ranks are in a Latin Square type configuration, and the optimal contamination scheme is used, then the sum of squared rank sums after *m* contaminants can be expressed as

$$f_R(m) = \sum_{j=1}^{t} R_{\cdot j}^2 = \left[\overline{R} + \sum_{i=1}^{g} n(t-i) - (t-g)(ng-m) \right]^2 + (t-g)(\overline{R} - m)^2 + \sum_{i=1}^{g-1} (\overline{R} - in)^2, \tag{5.2}$$

where $g = \left[(m+n-1)/n\right]_{gif}$, and *gif* stands for the greatest integer function (proof of Lemma 5.2 in Appendix B).

One can easily verify for $n = 1$, the formula reduces exactly to (5.1) derived for the $b = t$ case. With this function in terms of only one index (i) and no indicators, we can now derive the number of contaminants necessary to reject the null hypothesis. We start by dropping the greatest integer function on g, expanding on the summations, obtaining a function in terms of m, subtracting the constant c_α, then solving for m in the same fashion as illustrated in the previous section. (It should be noted here that once the greatest integer functions are deleted, (5.2) now becomes approximate, but still extremely accurate.) After expanding, the function is again a quartic in m, and again we are interested in the *lowest positive real valued* root. A typical plot of (5.2) appears in Figure 5.3 (for the case of $t = 5$ and $n = 5$, thus $b = 25$). Using Galois theory, we obtain a closed form expression of this root. If we define

$$\alpha = \frac{n(1+6t)}{6},\ \beta = 2 - 2n + n^2 - 4n^2t^2 + \frac{n^2(1+6t)^2}{9} - \frac{2 - 2n + n^2 - 4n^2t^2}{3},$$

$$\chi = 432n^2c_\alpha^* + 12n^2(1+6t)(3 - 3n + n^2 + 6t - 6tn) + 9(-2 + 2n - n^2 + 4n^2t^2)^2,$$

$$\delta = 1296n^4(1+6t)^2c_\alpha^* + 324n^2(3 - 3n + n^2 + 6t - 6tn)^2 - 7776n^2(-2 + 2n - n^2 + 4n^2t^2)c_\alpha^* + 108n^2(1+6t)(3 - 3n + n^2 + 6t - 6tn)(-2 + 2n - n^2 + 4n^2t^2) + 54(-2 + 2n - n^2 + 4n^2t^2)^3,$$

$$\varepsilon = \frac{8}{27}n^3(1+6t) - \frac{16}{3}n(3 - 3n + n^2 + 6t - 6tn) - \frac{8}{3}n(1+6t)(-2 + 2n - n^2 + 4n^2t^2), \text{ and}$$

$$c_\alpha^* = -\frac{-3+10n-9n^2+n^3(2+t^2\chi_\alpha^2+t^3\chi_\alpha^2)}{12n^2},$$

then we can approximate the number of contaminants necessary to force rejection as

$$P_{nt}^F \approx \alpha - \frac{1}{2}\left(\beta + \frac{2^{1/3}\chi}{9\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}} + \frac{\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}}{9\cdot 2^{1/3}}\right)^{1/2} +$$

$$\frac{1}{2}\left(2\beta - \frac{2^{1/3}\chi}{9\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}} - \frac{\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}}{9\cdot 2^{1/3}} - \frac{\varepsilon}{4\left(\beta+\frac{2^{1/3}\chi}{9\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}}+\frac{\left(\delta+\sqrt{-4\chi^3+\delta^2}\right)^{1/3}}{9\cdot 2^{1/3}}\right)^{1/2}}\right)^{1/2}$$

and approximate the maximum resistance to rejection as

$$\rho_{nt}^F \approx \frac{\left\lfloor P_{nt}^F + 1\right\rfloor_{gif}}{nt^2}.$$

The subscript '*nt*' indicates the number of blocks, and the superscripts '*F*' and '*gif*' are as before. We tested this root as an approximation of the true number of contaminants necessary to force rejection. Out of 115 cases where *t* ranged from 3 to 25 and *n* ranged from 2 to 6 (thus *b* ranged from 6 – 150), and $\alpha = .05$, the root obtained by the approximation described above was correct 113 times.

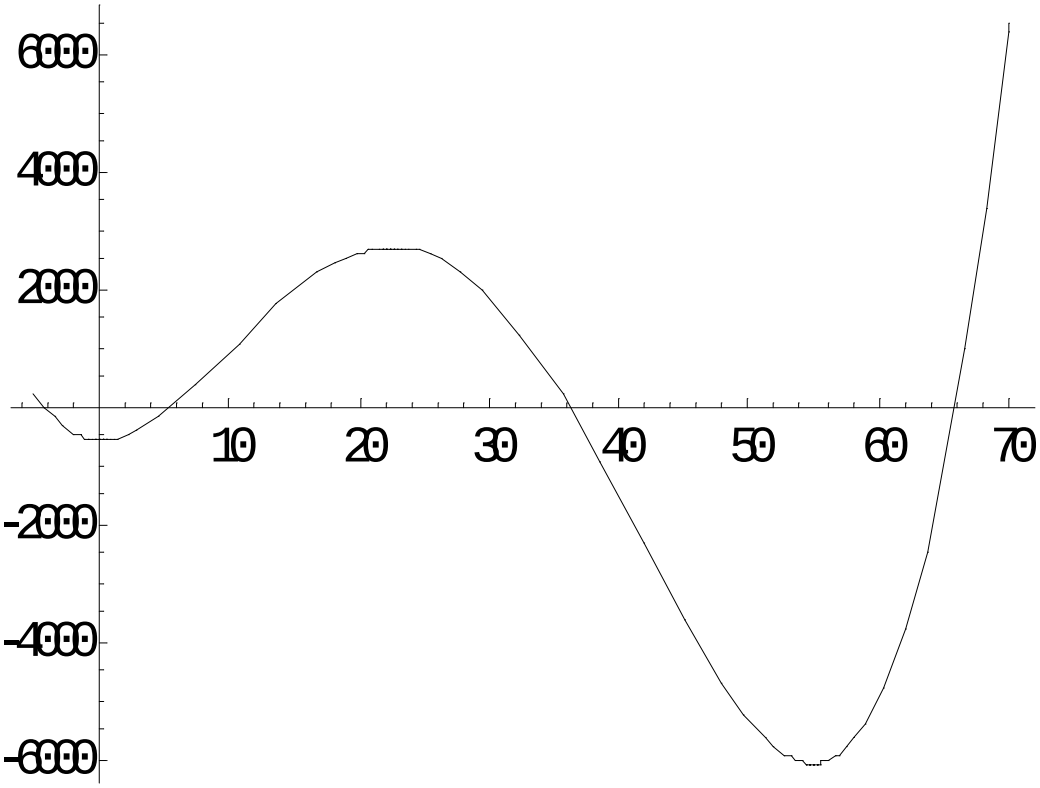


Figure 5.3. Plot of $f_R(m) - c_\alpha$ vs. m *(b = 25, t = 5)*

§ 5.5 The Maximum Resistance To Acceptance

In deriving the maximum resistance to acceptance, we need to start the test statistic at the worst possible value. This is not difficult to accomplish since the test statistic is based on the ranks of the data, which is finite, thus the test statistic has a finite maximum. For the Friedman test, the worst case scenario for acceptance is where all the highest ranks are associated with one treatment, the second highest ranks with another, and so on. This will give the highest possible observed X_F^2 value, or equivalently, the highest possible sum of the squared rank sums. For a simple example, consider the 4-treatment and 4-block design. Table 5.4 displays the most resistant configuration of the ranks for acceptance. Without loss of generality, we will assume that the first treatment contains all the highest ranks, the second treatment contains the second highest ranks, and so on. As a side note, a simple check will show that the maximum sum of the squared rank sums for any number of treatments *t*, and any number of blocks *b*, is

$$\max \sum_{j=1}^{t} R_{\cdot j}^2 = \frac{b^2}{6}(1+t)(1+2t).$$

We would like to obtain the number of contaminants that will bring this maximum value down below the critical constant. That is, we desire to find a formula as a function of *m*, such that

$$\sum_{j=1}^{t} R_{\cdot j}^2 \le \frac{\chi^2_{(t-1),\alpha} + 3b(t+1)}{12} bt(t+1)$$

or

$$f_A(m) \le c_\alpha.$$

The subscript '*A*' denotes 'acceptance', and 'α' still denotes the level of the test, which determines the critical constant.

Table 5.4. Most Resistant Configuration of the Ranks for Acceptance (b = t = 4)

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
Block 1	4	3	2	1
Block 2	4	3	2	1
Block 3	4	3	2	1
Block 4	4	3	2	1
$\sum_{i=1}^{t} R_{ij}$	16	12	8	4

§ 5.5.1 The Maximum Resistance to Acceptance (*b* = *t* case)

We begin by examining the case where the number of blocks is equivalent to the number of treatments ($b = t$). Since we know the worst configuration of the ranks that results in the highest observed X_F^2 value, we then only need an optimal pattern of contaminating the ranks. We need a scheme that brings down the test statistic, and equivalently the sum of squared rank sums, the fastest. After comparing many candidate schemes, we conclude that the optimal way of contamination is to start in the first treatment and first block and change its rank of t (the highest rank) to a 1, adjusting all other ranks within that block accordingly. The chosen block is actually arbitrary, but we start with block 1 and systematically contaminate downward through the blocks. The second step would occur in the second block, last treatment by changing the 1 to a t, adjusting all other ranks within that block accordingly. The third and fourth steps would then occur in the third block and involve changing the second treatment's entry to a 1 and the first treatment's to a 2, respectively. The fifth and sixth contaminants would then occur in the fourth block and involve changing the second to last treatment's entry from a 2 to a t and then the last treatment's entry from a 1 to a t-1. Just for emphasis, the seventh through ninth moves would occur in the fifth block and start with changing the third treatment's entry to a 1, the

second treatment's entry to a 2, and the first treatment's entry to a 3. This pattern of contamination brings the sum of squared rank sums down the fastest because each contamination minimizes the dispersion of the set of rank sums. There is a distinct pattern to the contamination scheme, one that we would like to characterize mathematically. For visualization, the first six contaminated steps for a five-treatment, five-block design are displayed in tabular form in Tables 5.5-5.7, starting with the initial configuration of the ranks. The contaminations are in bold and the order in which they are contaminated is displayed as a superscript. The rank sums are calculated after the 0^{th}, 2^{nd} and 6^{th} contaminant have been added.

Table 5.5. Configuration of the Ranks at m = 0 (b = t = 5)

$m = 0$	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
Block 1	5	4	3	2	1
Block 2	5	4	3	2	1
Block 3	5	4	3	2	1
Block 4	5	4	3	2	1
Block 5	5	4	3	2	1
$\sum_{i=1}^{t} R_{ij}$	25	20	15	10	5

Table 5.6. Optimal Placement of First and Second Contaminants

$m = 1, 2$	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
Block 1	$\mathbf{1}^{(1)}$	5	4	3	2
Block 2	4	3	2	1	$\mathbf{5}^{(2)}$
Block 3	5	4	3	2	1
Block 4	5	4	3	2	1
Block 5	5	4	3	2	1
$\sum_{i=1}^{t} R_{ij}$	20	20	15	10	10

Table 5.7. Optimal Placement of Third Through Sixth Contaminants

m = 3 - 6	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
Block 1	1	5	4	3	2
Block 2	4	3	2	1	5
Block 3	**2**$^{(4)}$	**1**$^{(3)}$	5	4	3
Block 4	3	2	1	**5**$^{(5)}$	**4**$^{(6)}$
Block 5	5	4	3	2	1
$\sum_{i=1}^{t} R_{ij}$	15	15	15	15	15

To derive the formula for the sum of squared rank sums we need to define what we call a set or 'group' of contaminants, which will be designated as G. We define the i^{th} set of contaminants, G_i, as the set of contaminants in the i^{th} ordered set of two blocks. As an example, if we consider G_1, this will be the ordered number of contaminants in the 1st two blocks. Based on the contamination scheme described above, this will be the first and second contaminant; the first block contains the first contaminant, and the second block contains the second. Thus, $G_1 = \{m: m = 1, 2\}$. For the second group, we would contaminate in order two observations in the third block (the third and fourth contaminants) and two in the fourth (the fifth and sixth). Therefore $G_2 = \{m: m = 3, 4, 5, 6\}$. Similarly, $G_3 = \{m: m = 7, 8, 9, 10, 11, 12\}$. Mathematically we have

Lemma 5.3: Define G_g as the set or 'group' of contaminants m in the i^{th} ordered set of two blocks based on the optimal contamination scheme, and g as the index of the group G_g. Then for the case where the number of blocks is equivalent to the number of treatments, and for a

given number of contaminants, *m*, the corresponding group can be found by $g = \left[\sqrt{m}+.5\right]_{gif}$, where *gif* is 'greatest integer function'.

Proof of Lemma 5.3:

It can be easily seen that the upper limit of the contaminants, m_u, for a given group G_g is $m_u = g^2 + g$, and the upper limit for the (g-1)st group, m_l, is $m_l = (g-1)^2 + (g-1) = g^2 - g$. So, we can say that

$$m_l < m \le m_u, \text{ or } g^2 - g < m \le g^2 + g .$$

By completing the square on each side of the inequality, we obtain

$$g^2 - g + .25 < m < g^2 + g + .25, \text{ or}$$

$$(g-.5)^2 < m < (g+.5)^2.$$

Note that by completing the square, the upper limit now becomes a strict inequality since $m, g \in Z^+$, and because $m, g \in Z^+$, the lower limit still stays a strict inequality. Taking the square root gives us

$$g - .5 < \sqrt{m} < g + .5, \text{ or}$$

$$g < \sqrt{m} + .5 < g + 1.$$

Therefore, we can say that

$$g = \left\lfloor \sqrt{m} + .5 \right\rfloor_{gif},$$

and this completes the proof.

We use this relationship to derive the sum of the squared rank sums. The formula derived is conditional on whether the number of contaminants is smaller or larger than g^2. This is due to the contamination scheme itself. If $m \leq g^2$, then we are contaminating higher ranks by decreasing them (since we start with the higher ranks), and if $m > g^2$, we are contaminating the lower ranks by increasing them. For each type of contamination, whether decreasing or increasing ranks, the rank sums adjust in very similar fashion. Unfortunately, the adjustment is not so similar that it can be summarized mathematically in one concise formula, but rather in two formulae; one for when the higher ranks are being decreased ($m \leq g^2$), and one for when the lower ranks are being increased ($m > g^2$). Formally,

Lemma 5.4: For the case where the number of blocks, *b*, is equivalent to the number of treatments, *t*, the ranks are such that all the highest ranks are within one treatment, then next highest ranks with another treatment, etc., and the optimal contamination scheme is used, then the sum of squared rank sums after *m* contaminants can be expressed as

$$f_A(m) = \sum_{j=1}^{t} R_{\cdot j}^2 = (g^2 - m)(t^2 - tg + t)^2 + m^*(t^2 - tg + g)^2 + \sum_{i=1}^{t-2g} \left(t(g+i) + m^*\right)^2 + g\left(tg + m^*\right)^2, \quad \text{if } m \leq g^2 \tag{5.3}$$

$$f_A(m) = \sum_{i=1}^{t} R_{\cdot j}^2 = (m - g^2)(tg + t)^2 + m^{**}(tg + g)^2 + \sum_{i=1}^{t-2g} \left(t(g+i) + m^{**}\right)^2 + g\left(t^2 - tg + m^{**}\right)^2, \quad \text{if } m > g^2 \tag{5.4}$$

where $m^* = g - g^2 + m$, $m^{**} = g + g^2 - m$, and $g = \lfloor \sqrt{m} + .5 \rfloor_{gif}$ (proof of Lemma 5.4 in Appendix B).

For the case where $m = g^2$, both formulae reduce to one equivalent formula. These formulae are still only a function of *m*, since *g*, m^* and m^{**} are a function of *m*. As with maximum resistance to rejection for $b = nt$ in Section 5.4.2, we drop the greatest integer function in *g*, expand on the summation, subtract the critical constant c_α, and try to solve for the roots of $f_A(m) - c_\alpha$. For these formulae, a correction factor was used after dropping the greatest integer function used for the value *g*. This is because when ignoring this function, the value *g* will always overestimate the true number desired. As an example, if we obtain a value of $g = \sqrt{m} + .5 \in [2.0, 3.0)$, the desired value of *g* in Lemma 5.4 is the lower bound of 2.0 (using the greatest integer function). However, if we subtract 0.5 from the original function, then using the same example, $g = \sqrt{m} \in [1.5, 2.5)$ is now centered about the desired value, thereby eliminating the constant overestimation. (Note: A correction factor was tried for the definition of *g* in Section 5.4.2, but not used due to the very high accuracy of the results without the correction.) After this modification of the value *g*, we can substitute for *g*, m^*, and m^{**} their respective functions in terms of *m*, in (5.3) and (5.4). Once this is done, the summations and polynomial terms are expanded. A little algebra will show that the sum of squared ranks sums defined in Lemma 5.4, for both (5.3) and (5.4) reduce to a single approximation formula given by

$$f_A(m) = \sum_{j=1}^{t} R_{\cdot j}^2 \approx \frac{t}{6}\left(4t\sqrt{m} + 8m^{3/2} - 6m(1+t^2) + t^2(1+3t+2t^2)\right). \qquad (5.5)$$

This is a great simplification to the formulae given in Lemma 5.4, and very accurate in approximating the sum of squared rank sums given in (5.3) and (5.4). Notice that this function is a polynomial of order 1.5 as opposed to 4 for the maximum resistance to rejection cases. A typical plot of this function is displayed in Figure 5.4, for $t = 5$, and as before, we are interested in the expression of the lowest positive real root. Using (5.5) we can express the approximate maximum number of bad data to force acceptance for the Friedman test, A_t^F, and approximate the maximum resistance to acceptance, $\alpha_t^F \approx \frac{\lfloor A_t^F + 1 \rfloor}{t^2}$. The expression of this root has been suppressed from this text, but can be found in Appendix D. This approximation was tested against the true maximum number of bad data necessary to force an acceptance and for values of t = 3-25, the root was correct for 22 of 23 cases, as shown in Appendix C, Table C.9.

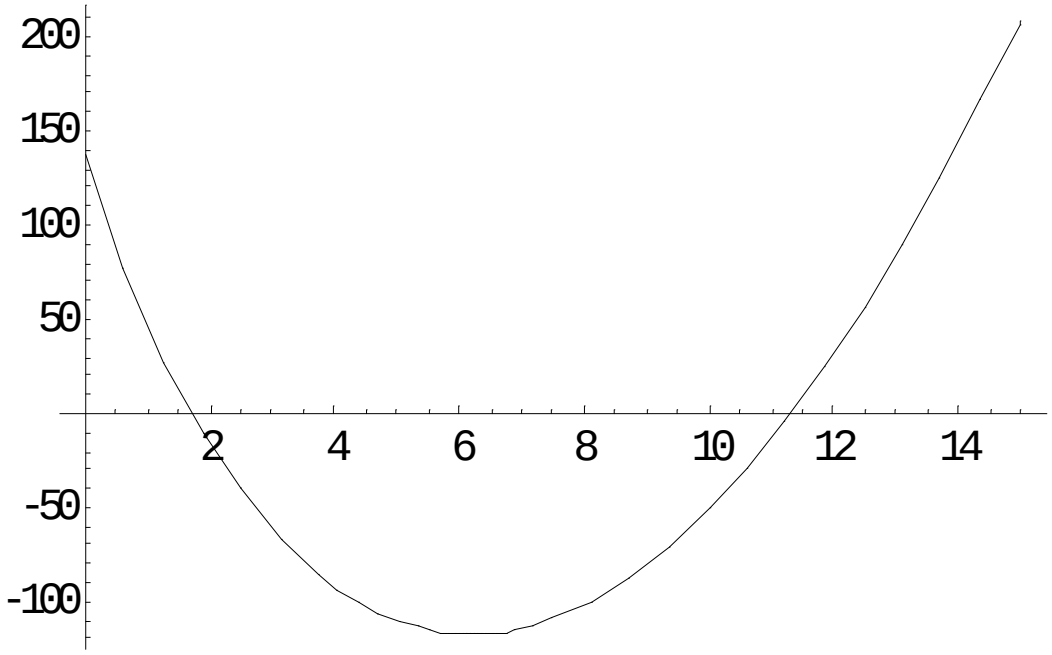


Figure 5.4. Plot of $f_A(m) - c_\alpha$ vs. m *(b = t = 5)*

§ 5.5.2 The Maximum Resistance to Acceptance (*b* = *nt* case)

For the case where the number of blocks equals the number of treatments, the optimal contamination scheme is just an extension to the previously described case. Again without loss of generality, we assume that all of the highest ranks are associated with the first treatment, the next highest possible ranks are with the second, and so on. We begin contaminating exactly as in the $b = t$ case. That is, we change in the first treatment a rank of t into a 1, and then t^{th} treatment, change a 1 into a t. However, this first sequence or group of two contaminations, defined earlier as G_1, is now repeated n times. For the next set of steps, we again mimic the $b = t$ case. We contaminate the second treatment by changing a rank of (t-1) to 1, then change a rank of t in the first treatment to 2. This is followed by changing the $(t\text{-}1)^{st}$ treatment's entry from a 2 to a t and then the t^{th}'s treatment's entry from a 1 to a t-1. Now, this sequence of four contaminations, defined earlier as G_2, is repeated n times, and so on. This would lead us to redefine what constitutes a set or group of contaminants. Formally we have

Lemma 5.5: Define G_g as the set or 'group' of contaminants m in the i^{th} ordered set of $2n$ rows based on the contamination scheme described, and g as the index of the group G_g. Then for the case where the number of blocks is an integer multiple of the number of treatments, such that $b = nt$, $n \in Z^+$, and for a given number of contaminants, m, the corresponding group can be found by $g = \left\lfloor \sqrt{m/n} + .5 \right\rfloor_{gif}$. The proof of Lemma 5.5 can be found in Appendix B.

Table 5.8 helps visualize the changing of the rank sums during the contamination scheme. Listed are the rank sums for a six-treatment twelve-block (n = 2) layout after the m^{th} contamination. The rank sums in bold indicate in which treatment the contamination took place, m is the number of contaminants in the data, and g is the associated group of the contaminants.

Table 5.8. Treatment Rank Sums After the m^{th} Contaminant (b = 12, t = 6)

m	*g*	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5	Treatment 6
0	0	72	60	48	36	24	12
1	1	**67**	62	49	37	25	13
2	1	66	61	48	36	24	**18**
3	1	**61**	61	49	37	25	19
4	1	60	60	48	36	24	**24**
5	2	60	**56**	49	37	25	25
6	2	**56**	56	50	38	26	26
7	2	55	55	49	37	**30**	26
8	2	54	54	48	36	30	**30**
9	2	54	**50**	49	37	31	31
10	2	**50**	50	50	38	32	32
11	2	49	49	49	37	**36**	32
12	2	48	48	48	36	36	**36**
13	3	48	48	**45**	37	37	37
14	3	48	**45**	45	38	38	38
15	3	**45**	45	45	39	39	39
16	3	44	44	44	**42**	39	39
17	3	43	43	43	42	**42**	39
18	3	42	42	42	42	42	**42**

At this time, a mathematical function representing the sum of squared rank sums using this contamination scheme is difficult to derive without the help of some inflexible indicator functions. We examined a couple of 'near optimal' contamination schemes, ones that we can characterize mathematically and that reasonably approximate the true sum of squared rank sums. The schemes are 'near optimal' in the sense that we either modify the optimal scheme, or use the optimal scheme but approximate the rank sums. The latter is what we used and the results were extremely accurate.

To describe how we approximate the rank sums using the optimal scheme, we used the fact that within a group of contaminants, G_i, we are contaminating the ranks associated with *2i* treatments; half of the treatments contain the higher ranks that are being decreased, and half contain the lower ranks that are being increased. For example, in G_1 we are contaminating the ranks in 2(1) = 2 treatments. Once we start in G_2, the number of treatments receiving contaminants increases to 4, and so on. We also used the fact that from the first contaminant within a group through the last, the change in rank sum for each of the treatments receiving these contaminants is equal to $n \cdot t$, with obviously higher treatment rank sums decreasing this amount and the lower rank sums increasing this amount. For the example corresponding to Table 5.8, n = 2 and t = 6, thus the change is 12. We can see in the first group that the first treatment rank sum changes from 72 to 60 and the sixth changes from 12 to 24. In the second group, the first and second treatment rank sums change from 60 to 48 while the fifth and sixth change from 24 to 36. Finally, the rank sums corresponding to the treatments not being contaminated stay fairly constant at their starting values. So, as an approximation to the sum of squared rank sums, we first estimate the non-contaminated treatment rank sums with their

starting value. For the contaminated treatment rank sums, we take the total possible change in the rank sums of $n \cdot t$, and divide this by the total number of contaminants for a given group, which is equal to $2 \cdot n \cdot g$. This yields the average change in the rank sum for each contaminant in a given group, and we uniformly add or subtract this value (depending if the rank sum is increasing or decreasing) with each contamination. As a concrete example using Table 5.8, the average change in the rank sum for the treatments contaminated in G_1 is 12/4 = 3, for G_2 it is 12/8 = 1.5, etc. Putting everything together in functional form yields an approximation to the true sum of squared rank sums as:

$$\sum_{j=1}^{t} R_{\cdot j}^2 \approx g\left(gnt + \frac{m^* t}{2g} \right)^2 + \sum_{i=1}^{t-2g} \left((g+i)nt\right)^2 + g\left((t+1-g)nt - \frac{m^* t}{2g} \right)^2,$$

where $m^* = m - ng(g-1)$, and $g = \left\lfloor \sqrt{m/n} + .5 \right\rfloor_{gif}$.

By approximating the greatest integer function on g by subtracting 0.5 as a correction factor, and substituting in for g and m^* their respective functions, we can write this approximation in polynomial form as

$$\sum_{j=1}^{t} R_{\cdot j}^2 \approx \frac{4}{3} mnt^2 \sqrt{m/n} + \frac{1}{6} n^2 t^2 \sqrt{m/n} - mnt^3 + \frac{n^2 t^3}{6} + \frac{n^2 t^4}{2} + \frac{n^2 t^5}{3}. \qquad (5.6)$$

This is a polynomial in m of order 1.5 as with the $b = t$ case described earlier. Using (5.6) we can express the approximate maximum number of bad data to force acceptance for the Friedman test, A_t^F, and approximate the maximum resistance to acceptance, $\alpha_{nt}^F \approx \left[\mathrm{A}_{nt}^F + 1\right]_{gif} / nt^2$, for

when the number of blocks is a constant multiple of the number of treatments. The expression of this root has been suppressed from this text, but can be found in Appendix D. This approximation was tested against the true maximum number of bad data necessary to force an acceptance for values of t = 3-10 and n = 1-5, the root was correct for 39 of 40 cases, as shown in Appendix C, Table C.10.

One question that remains is what exactly is the maximum *MRA*? That is, what is the limit as the sample goes to infinity? One step in answering this question is to consider (5.6). This equation represents a very accurate approximation to the sum of squared rank sums. Therefore the root of this equation accurately approximates the number of bad data to force an acceptance, and thus accurately approximates the *MRA*. By using the expression of the approximate *MRA* derived from (5.6), and taking the limit as the number of blocks, and equivalently n, goes to infinity, we obtain an expression as a function of only the number of treatments, t. That is,

$\lim_{n\to\infty} MRA(n,t) = g(t)$. This expression can be found in Appendix D. Now, if we let the number of treatments approach infinity, we can see that the limit of the Friedman *MRA* is ¼. Figure 5.5 displays this graphically, where the abscissa is t and the ordinate is $g(t)$.

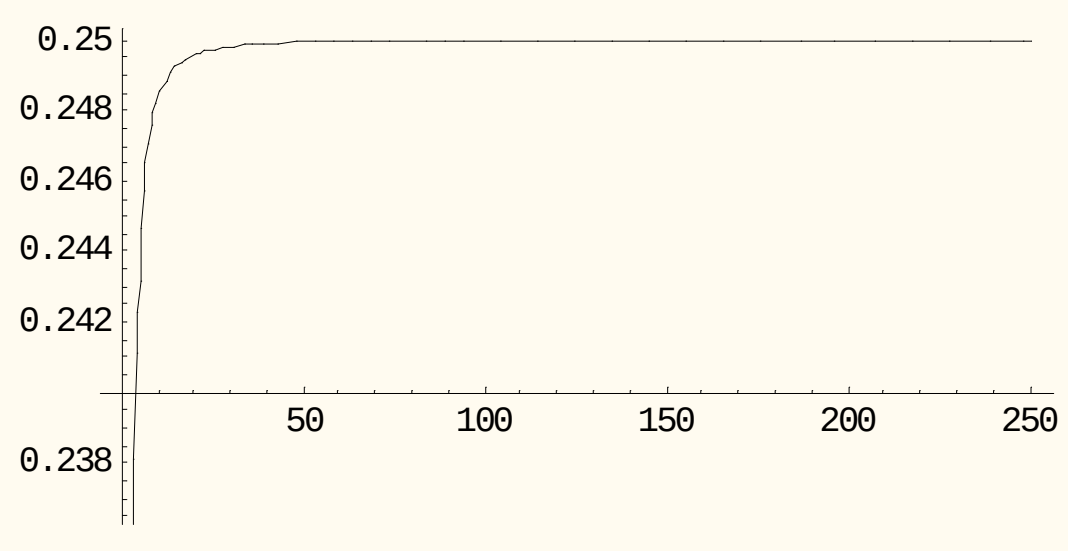


Figure 5.5. Plot of Maximum MRA of the Friedman Test

§ 5.6 Concluding Remarks

For the maximum resistance to rejection (*MRR*), the cases where the number of blocks is not a constant multiple of the number of treatments, the optimal configuration of the ranks varies on a case by case basis. So it is difficult to generalize the *MRR* to a generic *b-t* case. However, we can approximate the *MRR* by using the fact that for a fixed number of treatments, the number of contaminants necessary to force a rejection is an increasing function of the number of blocks. Therefore, for a number of blocks b^*, such that $(n-1)t < b^* < nt$, the number of contaminants necessary to force rejection, $\lfloor P^F_{b^*} + 1 \rfloor_{gif}$, is bounded by $\lfloor P^F_{(n-1)t} + 1 \rfloor_{gif}$, and $\lfloor P^F_{nt} + 1 \rfloor_{gif}$. Thus the *MRR* is bounded by

$$\frac{\lfloor P^F_{b^*} + 1 \rfloor_{gif}}{tb^*} \in \left[\frac{\lfloor P^F_{(n-1)t} + 1 \rfloor_{gif}}{tb^*}, \frac{\lfloor P^F_{nt} + 1 \rfloor_{gif}}{tb^*} \right], \text{ or}$$

$$\rho^F_{b^*} \in \lfloor \rho^F_{(n-1)t}, \rho^F_{nt} \rfloor.$$

For the maximum resistance to acceptance (*MRA*), the cases where the number of blocks is not a constant multiple of the number of treatments, the optimal contamination scheme varies on a case by case basis. Again, it is difficult to generalize the *MRA* to a generic *b-t* case. However, we can approximate the *MRA* by using the same fact that for a fixed number of treatments, the number of contaminants necessary to force an acceptance is an increasing function in the number of blocks. Therefore, for a number of blocks b^*, such that $(n-1)t < b^* < nt$, the number of contaminants necessary to force rejection, $\left\lfloor A^F_{b^*} + 1 \right\rfloor_{gif}$, is bounded by $\left\lfloor A^F_{(n-1)t} + 1 \right\rfloor_{gif}$ and $\left\lfloor A^F_{nt} + 1 \right\rfloor_{gif}$. Thus the *MRA* is bounded by

$$\frac{\left\lfloor A^F_{b^*} + 1 \right\rfloor_{gif}}{tb^*} \in \left[\frac{\left\lfloor A^F_{(n-1)t} + 1 \right\rfloor_{gif}}{tb^*}, \frac{\left\lfloor A^F_{nt} + 1 \right\rfloor_{gif}}{tb^*} \right], \text{ or}$$

$$\alpha^F_{b^*} \in \left\lfloor \alpha^F_{(n-1)t}, \alpha^F_{nt} \right\rfloor.$$