

Chapter 585

Multiple Comparisons of Treatments vs. a Control (Simulation)

Introduction

This procedure uses simulation to analyze the power and significance level of two multiple-comparison procedures that perform two-sided hypothesis tests of each treatment group mean versus the control group mean using simulation. These are Dunnett's test and the Kruskal-Wallis test. For each scenario, two simulations are run: one estimates the significance level and the other estimates the power.

The term *multiple comparisons of treatments versus a control* refers to the set of comparisons of each treatment group to a control group. If there are k groups of which $k-1$ are treatment groups, there will be $k-1$ tests.

When several comparisons are made among the group means, the interpretation for each comparison becomes more complex because of the problem of *multiplicity*. *Multiplicity* here refers to the fact that the chances of making at least one incorrect decision increases as the number of statistical tests increases. The method of *multiple comparisons* has been developed to account for this multiplicity.

Error Rates

When dealing with several simultaneous statistical tests, both individual-wise and experiment wise error rates should be considered.

1. **Comparison-wise error rate.** This is the probability of a type-I error (rejecting a true H_0) for a particular test. In the case of the five-group design, there are ten possible comparison-wise error rates, one for each of the ten possible pairs. We will denote this error rate α_c .
2. **Experiment-wise (or family-wise) error rate.** This is the probability of making one or more type-I errors in the set (family) of comparisons. We will denote this error rate α_f .

The relationship between these two error rates when the tests are independent is given by

$$\alpha_f = 1 - (1 - \alpha_c)^C$$

where C is the total number of comparisons in the family. For example, if α_c is 0.05 and C is 10, α_f is 0.401.

There is about a 40% chance that at least one of the ten pairs will be concluded to be different when in fact they are all the same. When the tests are correlated, as they are among a set of pair-wise comparisons, the above formula provides an upper bound to the family-wise error rate.

The techniques described below provide control for α_f rather than α_c .

Technical Details

The One-Way Analysis of Variance Design

The discussion that follows is based on the common one-way analysis of variance design which may be summarized as follows. Suppose the responses Y_{ij} in k groups each follow a normal distribution with means $\mu_1, \mu_2, \dots, \mu_k$ and unknown variance σ^2 . Let n_1, n_2, \dots, n_k denote the number of subjects in each group. The control group is assumed to be group one.

The analysis of these responses is based on the sample means

$$\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

and the pooled sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{\sum_{i=1}^k (n_i - 1)}$$

The F test is the usual method of analysis of the data from such a design, testing whether all of the means are equal. However, a significant F test does not indicate which of the groups are different, only that at least one is different. The analyst is left with the problem of determining which of the groups are different and by how much.

The Dunnett procedure and a special version of the Kruskal-Wallis procedure have been developed for this situation. The calculation of each of these tests is given next.

Dunnett's Test

Dunnett (1955) developed a test procedure for simultaneously comparing each treatment with a control group. It uses the critical values of a special t distribution given in Dunnett (1955). For each treatment and control pair, the significance test is calculated by rejecting the null hypothesis of mean equality if

$$\frac{|\bar{Y}_i - \bar{Y}_1|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_1} \right)}} \geq t_{1-\alpha/2, k, N-k}^{Dunnett} \quad i = 2, \dots, k$$

Kruskal-Wallis Test

This test is attributed to Dunn (1964) and is referenced in Gibbons (1976). It is a nonparametric, or distribution-free, test which means that normality is not assumed. It tests whether the medians of each treatment-control pair are equal using a rank test. Sample sizes of at least five (but preferably larger) for each treatment are recommended for the use of this test. The error rate is adjusted on a comparison-wise basis to give the experiment-wise error rate, α_f . Instead of using means, it uses average ranks as the following formula indicates, with $\alpha = \alpha_f / (2(k-1))$. For each treatment and control pair, the significance test is calculated by rejecting the null hypothesis of median equality if

Multiple Comparisons of Treatments vs. a Control (Simulation)

$$\frac{|\bar{R}_i - \bar{R}_1|}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_1} \right)}} \geq z_{\alpha} \quad i = 2, \dots, k$$

Note that, when necessary, the usual adjustment for ties is made.

Definition of Power for Multiple Comparisons

The notion of the power of a test is well-defined for individual tests. Power is the probability of rejecting a false null hypothesis. However, this definition does not extend easily when there are a number of simultaneous tests.

To understand the problem, consider an experiment with four groups labeled, C, A, B, and D. Suppose C is the control group. There are three paired comparisons in this experiment: A-C, B-C, and D-C. How do we define power for these three tests? One approach would be to calculate the power of each of the three tests, ignoring the other two. However, this ignores the interdependence among the three tests. Other definitions of the power of the set of tests might be the probability of detecting at least one of the differing pairs, exactly one of the differing pairs, at least two of the differing pairs, and so on. As the number of pairs increases, the number of possible definitions of power also increases. The two definitions that we emphasize in *PASS* were recommended by Ramsey (1978). They are *any-pair power* and *all-pairs power*. Other design characteristics, such as average-comparison power and false-discovery rate, are important to consider. However, our review of the statistical literature resulted in our focus on these two definitions of power.

Any-Pair Power

Any-pair power is the probability of detecting at least one of the pairs that are actually different.

All-Pairs Power

All-pairs power is the probability of detecting all of the pairs that are actually different.

Simulation Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how each test is to be carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the power of each test.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the significance level of each test.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in *PASS* to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draws the random numbers from this pool. This second method can cut the running time of the simulation by 70%!

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design and Options tabs. For more information about the options of other tabs, go to the Procedure Window chapter.

Design 1 Tab

The Design 1 tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for: *Power*, *Sample Size (Any-Pair)*, or *Sample Size (All-Pair)*. *Any-pair power* is the probability of detecting at least one pair from among those that are actually different. *All-pairs power* is the probability of detecting all pairs that are actually different.

Note that the search for sample size may take several minutes because a separate simulation must be run for each trial value of n . You may find it quicker and more informative to solve for the Power for a range of sample sizes.

Test

MC Procedure

Specify which pair-wise multiple comparison procedure is to be reported from the simulations. The choices are

- **Dunnett Test**
This is the most popular and the most often recommended.
- **Kruskal-Wallis**
This is recommended when a nonparametric procedure is wanted.

Multiple Comparisons of Treatments vs. a Control (Simulation)

Simulations

Simulations

This option specifies the number of iterations, M , used in the simulation. As the number of iterations is increased, the running time and accuracy are increased as well.

The precision of the simulated power estimates are calculated using the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'Precision' amount given in the table.

| Simulation Size M | Precision when Power = 0.50 | Precision when Power = 0.95 |
|------------------------|--------------------------------|--------------------------------|
| 100 | 0.100 | 0.044 |
| 500 | 0.045 | 0.019 |
| 1000 | 0.032 | 0.014 |
| 2000 | 0.022 | 0.010 |
| 5000 | 0.014 | 0.006 |
| 10000 | 0.010 | 0.004 |
| 50000 | 0.004 | 0.002 |
| 100000 | 0.003 | 0.001 |

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Power and Alpha

Power

This option is only used when *Solve For* is set to *Sample Size (All-Pairs)* or *Sample Size (Any-Pair)*.

Power is defined differently with multiple comparisons. Although many definitions are possible, two are adopted here. *Any-pair power* is the probability of detecting at least one pair of the means that are different. *All-pairs power* is the probability of detecting all pairs of means that are truly different. As the number of groups is increased, these power probabilities will decrease because more tests are being conducted.

Since this is a probability, the range is between 0 and 1. Most researchers would like to have the power at least at 0.8. However, this may require extremely large sample sizes when the number of tests is large.

FWER (Alpha)

This option specifies one or more values of the *family-wise error rate* (FWER) which is the analog of alpha for multiple comparisons. FWER is the probability of falsely detecting (concluding that the means are different) at least one comparison for which the true means are the same. For independent tests, the relationship between the individual-comparison error rate (ICER) and FWER is given by the formulas

$$FWER = 1 - (1 - ICER)^C$$

or

$$ICER = 1 - (1 - FWER)^{1/C}$$

where '^' represents exponentiation (as in $4^2 = 16$) and C represents the number of comparisons. For example, if $C = 5$ and $FWER = 0.05$, then $ICER = 0.0102$. Thus, the individual comparison tests must be conducted using a Type-1 error rate of 0.0102, which is much lower than the family-wise rate of 0.05.

Multiple Comparisons of Treatments vs. a Control (Simulation)

The popular value for FWER remains at 0.05. However, if you have a large number of comparisons, you might decide that a larger value, such as 0.10, is appropriate.

Sample Size

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values separated by blanks or commas may be entered. A separate analysis is performed for each value listed here.

The group sample sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by $m_1, m_2, m_3, \dots, m_k$ and this value is represented by n , the group sample sizes $N_1, N_2, N_3, \dots, N_k$ are calculated as follows:

$$N_1 = [n(m_1)]$$

$$N_2 = [n(m_2)]$$

$$N_3 = [n(m_3)]$$

etc.

where the operator, $[X]$ means the next integer after X , e.g. $[3.1] = 4$.

For example, suppose there are three groups and the Group Sample Size Pattern is set to $1, 2, 3$. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to $2, 4, 6, 8, 10$, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

| | | |
|----|----|----|
| 2 | 4 | 6 |
| 4 | 8 | 12 |
| 6 | 12 | 18 |
| 8 | 16 | 24 |
| 10 | 20 | 30 |

As a second example, suppose there are three groups and the Group Sample Size Pattern is $0.2, 0.3, 0.5$. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

$(0.2)12 = 2.4$ which is rounded up to the next whole integer, 3.

$(0.3)12 = 3.6$ which is rounded up to the next whole integer, 4.

$(0.5)12 = 6$.

Note that in this case, $3+4+6$ does not equal n (which is 12). This can happen because of rounding.

Group Sample Size Pattern

The purpose of the group sample size pattern is to allow several groups with the same sample size to be generated without having to type each individually.

A set of positive, numeric values (one for each row of distributions) is entered here. Each item specified in this list applies to the whole row of distributions. For example, suppose the entry is $1\ 2\ 1$ and Grps 1 = 3, Grps 2 = 1, Grps 3 = 2. The sample size pattern used would be $1\ 1\ 1\ 2\ 1\ 1$.

The sample size of group i is found by multiplying the i^{th} number from this list by the value of n and rounding up to the next whole number. The number of values must match the number of groups, g . When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

Multiple Comparisons of Treatments vs. a Control (Simulation)

- **Equal**

If all sample sizes are to be equal, enter *Equal* here and the desired sample size in n . A set of g 1's will be used. This will result in $n1 = n2 = \dots = ng = n$. That is, all sample sizes are equal to n .

Effect Size

These options specify the distributions to be used in the two simulations, one set per row. The first option specifies the number of groups represented by the two distributions that follow. The second option specifies the distribution to be used in simulating the null hypothesis to determine the significance level (alpha). The third option specifies the distribution to be used in simulating the alternative hypothesis to determine the power.

Note that group number one is the control group.

Grps [Ctrl – B] (Grps C – H are found on the Data 2 tab)

This value specifies the number of groups specified by the H0 and H1 distribution statements to the right. Usually, you will enter '1' to specify a single H0 and a single H1 distribution, or you will enter '0' to indicate that the distributions specified on this line are to be ignored. This option lets you easily specify many identical distributions with a single phrase.

The total number of groups g is equal to the sum of the values for the three rows of distributions shown under the Data 1 tab and the six rows of distributions shown under the Data 2 tab.

Note that each item specified in the *Group Sample Size Pattern* option applies to the whole row of entries here. For example, suppose the *Group Sample Size Pattern* was *1 2 1* and Grps 1 = 3, Grps 2 = 1, and Grps 3 = 2. The sample size pattern would be *1 1 1 2 1 1*.

Note that since the first group is the control group, the value for Grps 1 is usually set to one.

Group Distribution(s)|H0

This entry specifies the distribution of one or more groups under the null hypothesis, H0. The magnitude of the differences of the means of these distributions, which is often summarized as the standard deviation of the means, represents the magnitude of the mean differences specified under H0. Usually, the means are assumed to be equal under H0, so their standard deviation should be zero except for rounding.

These distributions are used in the simulations that estimate the actual significance level. They also specify the value of the mean under the null hypothesis, H0. Usually, these distributions will be identical. The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Each of the parameters should be replaced with a number or parameter name.

Distributions with Common Parameters

Beta(Shape1, Shape2, Min, Max)

Binomial(P, N)

Cauchy(Mean, Scale)

Constant(Value)

Exponential(Mean)

Gamma(Shape, Scale)

Gumbel(Location, Scale)

Laplace(Location, Scale)

Logistic(Location, Scale)

Multiple Comparisons of Treatments vs. a Control (Simulation)

Lognormal(Mu, Sigma)

Multinomial(P1, P2, P3, ..., Pk)

Normal(Mean, Sigma)

Poisson(Mean)

TukeyGH(Mu, S, G, H)

Uniform(Min, Max)

Weibull(Shape, Scale)

Distributions with Mean and SD Parameters

BetaMS(Mean, SD, Min, Max)

BinomialMS(Mean, N)

GammaMS(Mean, SD)

GumbelMS(Mean, SD)

LaplaceMS(Mean, SD)

LogisticMS(Mean, SD)

LognormalMS(Mean, SD)

UniformMS(Mean, SD)

WeibullMS(Mean, SD)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on *Data Simulation* and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

Most of the distributions have been parameterized in terms of their means since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of $4 \text{ Normal}(4, 5) + 2 \text{ Normal}(5, 6)$ is $4*4 + 2*5 = 26$, but the mean of $4 \text{ Normal}(4, 5) * 2 \text{ Normal}(5, 6)$ is not exactly $4*4*2*5 = 160$ (although it is close).

Group Distribution(s)|H1

Specify the distribution of this group under the alternative hypothesis, H1. This distribution is used in the simulation that determines the power. A fundamental quantity in a power analysis is the amount of variation among the group means. In fact, classical power analysis formulas, this variation is summarized as the standard deviation of the means.

The important point to realize is that you must pay particular attention to the values you give to the means of these distributions because they are fundamental to the interpretation of the simulation.

For convenience in specifying a range of values, the parameters of the distribution can be specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *MI* is reserved for the value of the mean under the alternative hypothesis.

A list of the distributions that are available and the syntax used to specify them is given above.

Multiple Comparisons of Treatments vs. a Control (Simulation)

Equivalence Margin

Specify the largest difference for which means from different groups will be considered equal. When specifying group distributions, it is possible to end up with scenarios where some means are slightly different from each other, even though they are intended to be equivalent. This often happens when specifying distributions of different forms (e.g. normal and gamma) for different groups, where the means are intended to be the same. The parameters used to specify different distributions do not always result in means that are EXACTLY equal. This value lets you control how different means can be and still be considered equal.

This value is not used to specify the hypothesized mean differences of interest. The hypothesized differences are specified using the means (or parameters used to calculate means) for the null and alternative distributions.

This value should be much smaller than the hypothesized mean differences.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for $M0$ in the distribution specifications given above. $M0$ is intended to be the value of the mean hypothesized by the null hypothesis, $H0$.

You can enter a list of values using the syntax $0\ 1\ 2\ 3$ or $0\ to\ 3\ by\ 1$.

M1 (Mean|H1)

These values are substituted for $M1$ in the distribution specifications given above. Although it can be used wherever you want, $M1$ is intended to be the value of the mean hypothesized by the alternative hypothesis, $H1$.

You can enter a list of values using the syntax $0\ 1\ 2\ 3$ or $0\ to\ 3\ by\ 1$.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax $0\ 1\ 2\ 3$ or $0\ to\ 3\ by\ 1$.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

These options let you specify whether you want to generate the standard reports and plots.

Show Inc's & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Options Tab

The Options tab contains a random number pool size option.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Example 1 – Power at Various Sample Sizes

An experiment is being designed to compare the responses of three different treatments to a control. Previous studies have shown that the standard deviation within a group is 3.0. Researchers want to detect a shift in the mean of at least 2.0. To accomplish this, they set the mean of the control group to zero and the three treatment means to 2.0. They want to investigate sample sizes of 5, 10, 15, and 20 subjects per group.

Their primary analysis will be a set of Dunnett multiple comparison tests. They set the FWER to 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Multiple Comparisons of Treatments vs. a Control (Simulation)**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

| <u>Option</u> | <u>Value</u> |
|------------------------------------|---------------------|
| Design Tab | |
| Solve For | Power |
| MC Procedure..... | Dunnett |
| Simulations | 2000 |
| FWER (Alpha)..... | 0.05 |
| n (Sample Size Multiplier)..... | 5 10 15 20 |
| Group Sample Size Pattern | Equal |
| Grps 1 | 1 |
| Control Distribution H0..... | Normal(M0 S) |
| Control Distribution H1..... | Normal(M0 S) |
| Grps 2 | 3 |
| Group 2 Distribution(s) H0 | Normal(M0 S) |
| Group 2 Distribution(s) H1 | Normal(M1 S) |
| Equivalence Margin | 0.1 |
| M0 (Mean H0) | 0 |
| M1 (Mean H1) | 2 |
| S..... | 3 |

Report Tab

All reports except Comparative **Checked**

Multiple Comparisons of Treatments vs. a Control (Simulation)

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Simulation Summary Report

Summary of Simulations of the Control Group and the 3 Treatment Groups
MC Procedure: Dunnett's M.C. Test

| Sim. No. | Any-Pairs Power | Group Smpl. Size n | Total Smpl. Size N | All-Pairs Power | S.D. of Means Sm H1 | S.D. of Data SD H1 | Actual FWER | Target FWER | M0 | M1 | S |
|----------|------------------|--------------------|--------------------|------------------|---------------------|--------------------|------------------|-----------------|-----|-----|-----|
| 1 | 0.187 (0.017) | 5.0 [0.170 | 20 0.204] | 0.016 (0.005) | 0.9 [0.010 | 3.0 0.021] | 0.055 (0.010) | 0.050 [0.045 | 0.0 | 2.0 | 3.0 |
| 2 | 0.358 (0.021) | 10.0 [0.337 | 40 0.379] | 0.045 (0.009) | 0.9 [0.036 | 3.0 0.054] | 0.050 (0.010) | 0.050 [0.040 | 0.0 | 2.0 | 3.0 |
| 3 | 0.507 (0.022) | 15.0 [0.485 | 60 0.529] | 0.098 (0.013) | 0.9 [0.085 | 3.0 0.111] | 0.042 (0.009) | 0.050 [0.033 | 0.0 | 2.0 | 3.0 |
| 4 | 0.623 (0.021) | 20.0 [0.602 | 80 0.644] | 0.162 (0.016) | 0.9 [0.145 | 3.0 0.178] | 0.045 (0.009) | 0.050 [0.035 | 0.0 | 2.0 | 3.0 |

Pool Size: 10000. Simulations: 2000. Run Time: 26.45 seconds.

Summary of Simulations Report Definitions

H0: the null hypothesis that each treatment mean is equal to the control mean.

H1: the alternative hypothesis that at least one treatment mean differs from the control mean.

Pair: each comparison of a treatment mean and a control mean is a 'pair'.

All-Pairs Power: the estimated probability of detecting all unequal pairs.

Any-Pairs Power: the estimated probability of detecting at least one unequal pair.

n: the average of the group sample sizes.

N: the combined sample size of all groups.

Family-Wise Error Rate (FWER): the probability of detecting at least one equal pair assuming H0.

Target FWER: the user-specified FWE.

Actual FWER: the FWER estimated by the alpha simulation.

Sm|H1: the standard deviation of the group means under H1.

SD|H1: the pooled, within-group standard deviation under H1.

Second Row: provides the precision and a confidence interval based on the size of the simulation for Any-Pairs Power, All-Pairs Power, and FWER. The format is (Precision) [95% LCL and UCL Alpha].

Summary Statements

A one-way design with 3 treatment groups and one control group has an average group sample size of 5.0 for a total sample size of 20. This design achieved an any-pair power of 0.187 and an all-pair power of 0.0155 using the Dunnett's Test procedure for comparing each treatment mean with the control mean. The target family-wise error rate was 0.050 and the actual family-wise error rate was 0.055. The average within group standard deviation assuming the alternative distribution is 3.0. These results are based on 2000 Monte Carlo samples from the null distributions: N(M0 S); N(M0 S); N(M0 S); and N(M0 S) and the alternative distributions: N(M0 S); N(M1 S); N(M1 S); and N(M1 S). Other parameters used in the simulation were: M0 = 0.0, M1 = 2.0, and S = 3.0.

This report shows the estimated any-pairs power, all-pairs power, and FWER for each scenario. The second row shows three 95% confidence intervals in brackets: the first for the any-pairs power, the second for the all-pairs power, and the third for the FWER. Half the width of each confidence interval is given in parentheses as a fundamental measure of the precision of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Any-Pairs Power

This is the probability of detecting any of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

Multiple Comparisons of Treatments vs. a Control (Simulation)

All-Pairs Power

This is the probability of detecting all of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

Group Sample Size n

This is the average of the individual group sample sizes.

Total Sample Size N

This is the total sample size of the study.

S.D. of Means $S_{m|H1}$

This is the standard deviation of the hypothesized means of the alternative distributions. Under the null hypothesis, this value is zero. This value represents the magnitude of the difference among the means that is being tested. It is roughly equal to the average difference between the group means and the overall mean.

Note that the effect size is the ratio of $S_{m|H1}$ and $SD|H1$

S.D. of Data $SD|H1$

This is the within-group standard deviation calculated from samples from the alternative distributions.

Actual FWER

This is the value of FWER (family-wise error rate) estimated by the simulation using the H0 distributions. It should be compared with the Target FWER to determine if the test procedure is accurate.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the Actual FWER.

Target FWER

The target value of FWER.

M0

This is the value entered for M0, the group means under H0.

M1

This is the value entered for M1, the group means under H1.

S

This is the value entered for S, the standard deviation.

Error-Rate Summary for H0 Simulation

Error Rate Summary from H0 (Alpha) Simulation of 4 Groups
MC Procedure: Dunnett's M.C. Test

| Sim. No. | No. of Equal Pairs | Mean No. of Type-1 Errors | Prop. Type-1 Errors | Prop. (No. of Type-1 Errors > 0) FWER | Target FWER | Mean Pairs Alpha | Min Pairs Alpha | Max Pairs Alpha |
|----------|--------------------|---------------------------|---------------------|---------------------------------------|-------------|------------------|-----------------|-----------------|
| 1 | 3 | 0.067 | 0.022 | 0.055 | 0.050 | 0.022 | 0.020 | 0.025 |
| 2 | 3 | 0.060 | 0.020 | 0.050 | 0.050 | 0.020 | 0.017 | 0.023 |
| 3 | 3 | 0.048 | 0.016 | 0.042 | 0.050 | 0.016 | 0.016 | 0.017 |
| 4 | 3 | 0.055 | 0.018 | 0.045 | 0.050 | 0.018 | 0.018 | 0.019 |

Multiple Comparisons of Treatments vs. a Control (Simulation)

This report shows the results of the H0 simulation. This simulation uses the H0 settings for each group. Its main purpose is to provide an estimate of the FWER.

No. of Equal Pairs

Since under H0 all means are equal, this is the number of unique pairs of the groups. Thus, this is the number of treatment groups.

Mean No. of Type-1 Errors

This is the average number of type-1 errors (false detections) per set (family).

Prop. Type-1 Errors

This is the proportion of type-1 errors (false detections) among all tests that were conducted.

Prop. (No. of Type-1 Errors>0) FWER

This is the proportion of the H0 simulations in which at least one type-1 error occurred. This is called the family-wise error rate.

Target FWER

This is the target value of FWER that was set by the user.

Mean Pairs Alpha

Alpha is the probability of rejecting H0 when H0 is true. It is a characteristic of an individual test. This is the average alpha value over all of the tests in the family.

Min Pairs Alpha

This is the minimum of all of the individual comparison alphas.

Max Pairs Alpha

This is the maximum of all of the individual comparison alphas.

Error-Rate Summary for H1 Simulation

Error Rate Summary from H1 (Power) Simulation of 4 Groups
MC Procedure: Dunnett's M.C. Test

| Sim. No. | No. of Equal Pairs | Mean No. of False Pos. | Mean No. of False Neg. | Prop. Errors | (FDR) | | | | All Uneq. Pairs Power | Any Uneq. Pairs Power | Mean Pairs Power | Min Pairs Power | Max Pairs Power |
|----------|--------------------|------------------------|------------------------|--------------|-------------------------------|------------------------------|-------------------------------|------------------------------|-----------------------|-----------------------|------------------|-----------------|-----------------|
| | | | | | Prop. Equal that were Detect. | Prop. Uneq. that were Undet. | Prop. Detect. that were Equal | Prop. Undet. that were Uneq. | | | | | |
| 1 | 0/3 | 0.00 | 2.74 | 0.913 | 0.000 | 0.913 | 0.000 | 1.000 | 0.016 | 0.187 | 0.088 | 0.082 | 0.092 |
| 2 | 0/3 | 0.00 | 2.45 | 0.816 | 0.000 | 0.816 | 0.000 | 1.000 | 0.045 | 0.358 | 0.184 | 0.182 | 0.188 |
| 3 | 0/3 | 0.00 | 2.13 | 0.711 | 0.000 | 0.711 | 0.000 | 1.000 | 0.098 | 0.507 | 0.289 | 0.286 | 0.292 |
| 4 | 0/3 | 0.00 | 1.84 | 0.613 | 0.000 | 0.613 | 0.000 | 1.000 | 0.162 | 0.623 | 0.388 | 0.380 | 0.392 |

This report shows the results of the H1 simulation. This simulation uses the H1 settings for each group. Its main purpose is to provide an estimate of the power.

No. of Equal Pairs/Unequal Pairs

The first value is the number of pairs for which the control mean and the treatment mean were equal under H1. The second value is the number of pairs for which the means were different under H1.

Mean No. False Positives

This is the average number of equal pairs that were declared as being unequal by the testing procedure. A *false positive* is a type-1 (alpha) error.

Multiple Comparisons of Treatments vs. a Control (Simulation)

Mean No. False Negatives

This is the average number of unequal pairs that were not declared as being unequal by the testing procedure. A *false negative* is a type-2 (beta) error.

Prop. Errors

This is the proportion of type-1 and type-2 errors.

Prop. Equal that were Detect.

This is the proportion of the equal pairs in the H1 simulations that were declared as unequal.

Prop. Uneq. that were Undet.

This is the proportion of the unequal pairs in the H1 simulations that were not declared as being unequal.

Prop. Detect. that were Equal (FDR)

This is the proportion of detected pairs in the H1 simulations that were actually equal. This is often called the *false discovery rate*.

Prop. Undet. that were Uneq.

This is the proportion of undetected pairs in the H1 simulations that were actually unequal.

All Uneq. Pairs Power

This is the probability of detecting all of the pairs that were different in the H1 simulation.

Any Uneq. Pairs Power

This is the probability of detecting any of the pairs that were different in the H1 simulation.

Mean, Min, and Max Pairs Power

These items give the average, the minimum, and the maximum of the individual comparison powers from the H1 simulation.

Detail Model Report

Detailed Model Report for Simulation No. 1
 Target FWER = 0.050, M0 = 0.0, M1 = 2.0, S = 3.0
 MC Procedure: Dunnett's M.C. Test

| Hypo. Type | Groups | Group Labels | n/N | Group Mean | Ave. S.D. | Simulation Model |
|------------|--------|--------------|------|------------|-----------|------------------|
| H0 | 1 | Ctrl | 5/20 | 0.0 | 3.0 | N(M0 S) |
| H0 | 2-4 | A1-A3 | 5/20 | 0.0 | 3.0 | N(M0 S) |
| H0 | All | | | Sm=0.0 | 3.0 | |
| H1 | 1 | Ctrl | 5/20 | 0.0 | 3.0 | N(M0 S) |
| H1 | 2-4 | A1-A3 | 5/20 | 2.0 | 3.0 | N(M1 S) |
| H1 | All | | | Sm=0.9 | 3.0 | |

Detailed Model Report for Simulation No. 2

| Hypo. Type | Groups | Group Labels | n/N | Group Mean | Ave. S.D. | Simulation Model |
|------------|--------|--------------|-------|------------|-----------|------------------|
| H0 | 1 | Ctrl | 10/40 | 0.0 | 3.0 | N(M0 S) |
| H0 | 2-4 | A1-A3 | 10/40 | 0.0 | 3.0 | N(M0 S) |
| H0 | All | | | Sm=0.0 | 3.0 | |
| H1 | 1 | Ctrl | 10/40 | 0.0 | 3.0 | N(M0 S) |
| H1 | 2-4 | A1-A3 | 10/40 | 2.0 | 3.0 | N(M1 S) |
| H1 | All | | | Sm=0.9 | 3.0 | |

Multiple Comparisons of Treatments vs. a Control (Simulation)

| Detailed Model Report for Simulation No. 3 | | | | | | |
|--|--------|--------------|-------|------------|-----------|------------------|
| Hypo. Type | Groups | Group Labels | n/N | Group Mean | Ave. S.D. | Simulation Model |
| H0 | 1 | Ctrl | 15/60 | 0.0 | 3.0 | N(M0 S) |
| H0 | 2-4 | A1-A3 | 15/60 | 0.0 | 3.1 | N(M0 S) |
| H0 | All | | | Sm=0.0 | 3.0 | |
| H1 | 1 | Ctrl | 15/60 | 0.0 | 2.9 | N(M0 S) |
| H1 | 2-4 | A1-A3 | 15/60 | 2.0 | 3.0 | N(M1 S) |
| H1 | All | | | Sm=0.9 | 3.0 | |
| Detailed Model Report for Simulation No. 4 | | | | | | |
| Hypo. Type | Groups | Group Labels | n/N | Group Mean | Ave. S.D. | Simulation Model |
| H0 | 1 | Ctrl | 20/80 | 0.0 | 3.0 | N(M0 S) |
| H0 | 2-4 | A1-A3 | 20/80 | 0.0 | 3.0 | N(M0 S) |
| H0 | All | | | Sm=0.0 | 3.0 | |
| H1 | 1 | Ctrl | 20/80 | 0.0 | 3.0 | N(M0 S) |
| H1 | 2-4 | A1-A3 | 20/80 | 2.0 | 3.0 | N(M1 S) |
| H1 | All | | | Sm=0.9 | 3.0 | |

This report shows details of each row of the previous reports.

Hypo. Type

This indicates which simulation is being reported on each row. H0 represents the null simulation and H1 represents the alternative simulation.

Groups

Each group in the simulation is assigned a number. This item shows the arbitrary group number that was assigned.

Group Labels

These are the labels that were used in the individual alpha-level reports. Note that the control group is labeled 'Ctrl'.

n/N

n is the average sample size of the groups. N is the total sample size across all groups.

Group Mean

These are the means of the individual groups as specified for the H0 and H1 simulations.

Ave. S.D.

This is the average standard deviation of all groups reported on each line. Note that it is calculated from the simulated data.

Simulation Model

This is the distribution that was used to simulate data for the groups reported on each line.

Probability of Rejecting Equality

| Probability of Rejecting the Equality of Each Pair. Simulation No. 1 | | | | | |
|--|-------|-------|--------|--------|--------|
| Group | Means | Ctrl | A1 | A2 | A3 |
| Ctrl | 0.0 | | 0.082* | 0.092* | 0.090* |
| A1 | 2.0 | 0.025 | | | |
| A2 | 2.0 | 0.020 | | | |
| A3 | 2.0 | 0.022 | | | |

Multiple Comparisons of Treatments vs. a Control (Simulation)

Probability of Rejecting the Equality of Each Pair. Simulation No. 2

| Group | Means | Ctrl | A1 | A2 | A3 |
|-------|-------|-------|--------|--------|--------|
| Ctrl | 0.0 | | 0.188* | 0.184* | 0.182* |
| A1 | 2.0 | 0.023 | | | |
| A2 | 2.0 | 0.020 | | | |
| A3 | 2.0 | 0.017 | | | |

Probability of Rejecting the Equality of Each Pair. Simulation No. 3

| Group | Means | Ctrl | A1 | A2 | A3 |
|-------|-------|-------|--------|--------|--------|
| Ctrl | 0.0 | | 0.292* | 0.286* | 0.289* |
| A1 | 2.0 | 0.016 | | | |
| A2 | 2.0 | 0.016 | | | |
| A3 | 2.0 | 0.017 | | | |

Probability of Rejecting the Equality of Each Pair. Simulation No. 4

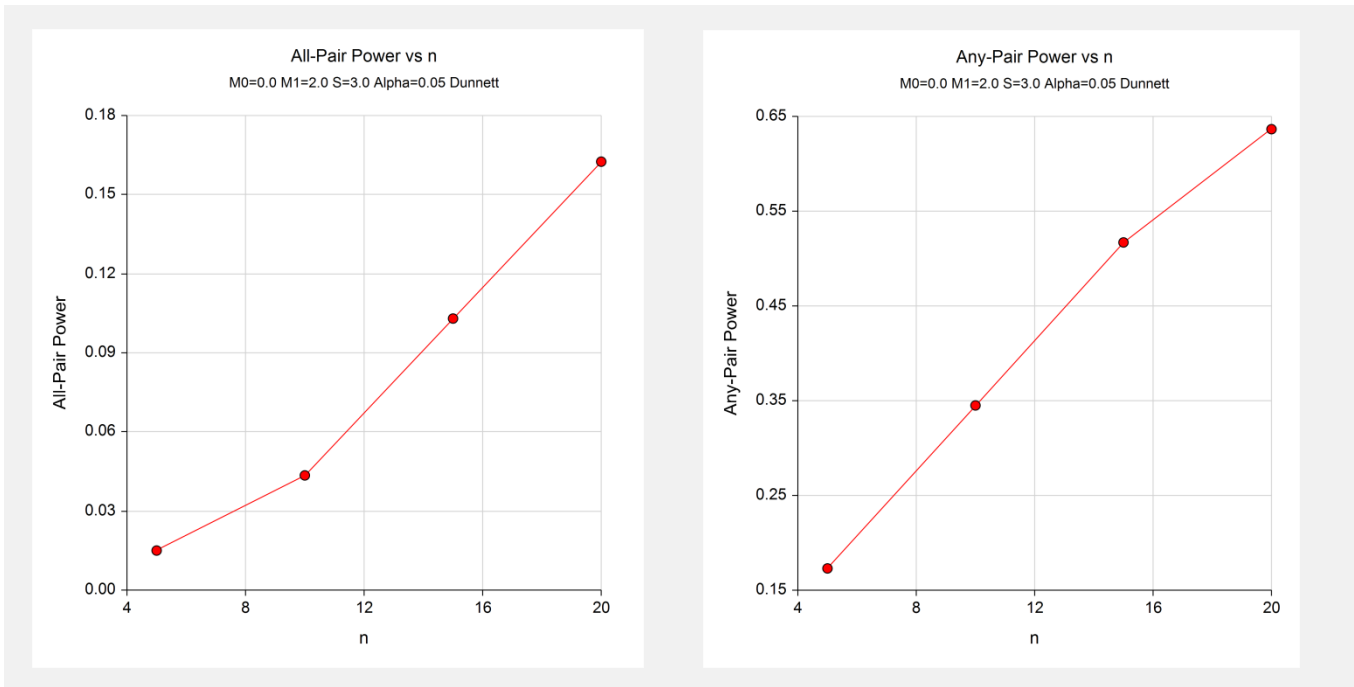
| Group | Means | Ctrl | A1 | A2 | A3 |
|-------|-------|-------|--------|--------|--------|
| Ctrl | 0.0 | | 0.392* | 0.380* | 0.392* |
| A1 | 2.0 | 0.019 | | | |
| A2 | 2.0 | 0.018 | | | |
| A3 | 2.0 | 0.018 | | | |

Individual pairwise powers from the H1 (Power) simulation are shown in the upper-right section. Individual pairwise significance levels from the H0 (Alpha) simulation are shown in the lower-left section. * Starred values are the powers of pairs that are unequal under H1.

This report shows the individual probabilities of rejecting each pair. When a pair was actually different, the value is the power of that test. These power values are starred.

The results shown on the upper-right section of each simulation report are from the H1 simulation. The results shown on the lower-left section of the report are from the H0 simulation.

Plots Section



These plots give a visual presentation of the all-pairs power values and the any-pair power values.

Example 2 – Comparative Results

Continuing with Example 1, the researchers want to study the characteristics of alternative multiple comparison procedures.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Multiple Comparisons of Treatments vs. a Control (Simulation)**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

| <u>Option</u> | <u>Value</u> |
|--------------------------------------|---------------------|
| Design Tab | |
| Solve For | Power |
| MC Procedure..... | Dunnett |
| Simulations | 2000 |
| FWER (Alpha)..... | 0.05 |
| n (Sample Size Multiplier)..... | 5 10 15 20 |
| Group Sample Size Pattern..... | Equal |
| Grps 1 | 1 |
| Control Distribution H0..... | Normal(M0 S) |
| Control Distribution H1..... | Normal(M0 S) |
| Grps 2 | 3 |
| Group 2 Distribution(s) H0 | Normal(M0 S) |
| Group 2 Distribution(s) H1 | Normal(M1 S) |
| Equivalence Margin | 0.1 |
| M0 (Mean H0) | 0 |
| M1 (Mean H1) | 2 |
| S..... | 3 |
| Reports Tab | |
| Comparative Reports..... | Checked |
| Comparative Any-Pair Power Plot..... | Checked |
| Comparative All-Pair Power Plot..... | Checked |

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Power Comparison for Testing the Control Group Versus the 3 Treatment Groups

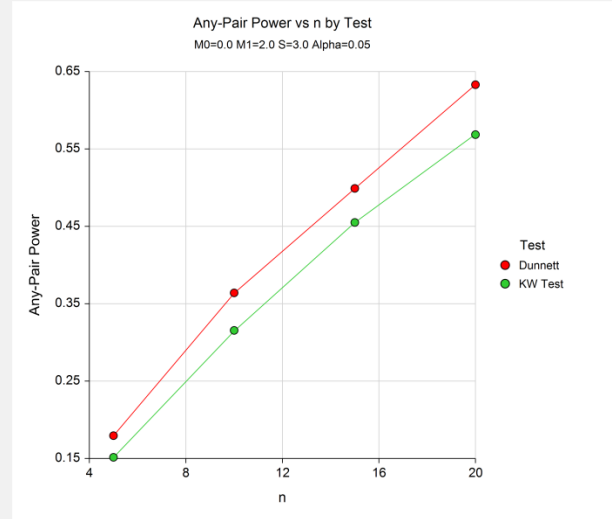
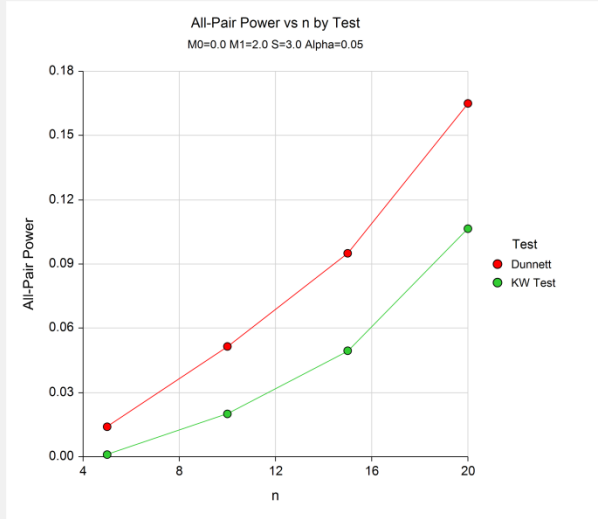
| Sim. No. | Total Sample Size | Target Alpha | Dunnett | | Kruskal Wallis | |
|----------|-------------------|--------------|----------------|----------------|----------------|----------------|
| | | | All-Pair Power | Any-Pair Power | All-Pair Power | Any-Pair Power |
| 1 | 20 | 0.050 | 0.016 | 0.187 | 0.002 | 0.155 |
| 2 | 40 | 0.050 | 0.045 | 0.358 | 0.018 | 0.309 |
| 3 | 60 | 0.050 | 0.098 | 0.507 | 0.056 | 0.474 |
| 4 | 80 | 0.050 | 0.162 | 0.623 | 0.117 | 0.581 |

Pool Size: 10000. Simulations: 2000. Run Time: 26.18 seconds.

Multiple Comparisons of Treatments vs. a Control (Simulation)

Family-Wise FWER Comparison for Testing the Control Group Versus the 3 Treatment Groups

| Sim. No. | Total Sample Size | Target FWER | Dunnett FWER | Kruskal Wallis FWER |
|----------|-------------------|-------------|--------------|---------------------|
| 1 | 20 | 0.050 | 0.055 | 0.044 |
| 2 | 40 | 0.050 | 0.050 | 0.040 |
| 3 | 60 | 0.050 | 0.042 | 0.034 |
| 4 | 80 | 0.050 | 0.045 | 0.038 |



These reports show the power and FWER of both of the multiple comparison procedures. In these simulations of groups from the normal distributions with equal variances, we see that the Dunnett's procedure is the champion.

Example 3 – Validation using Dunnett

Mukerjee, Robertson, and Wright (1987) page 909 present an example of a sample size calculation which was first discussed on page 1116 of Dunnett (1955). In this example there are five treatments and one control. The control mean and four of the treatment means are the same, while the fifth treatment mean is different.

The value of the within-group standard deviation is 1.0. Four treatment means and the control mean are -0.182574 and the fifth treatment mean is 0.912871. The FWER is 0.05 in the article, but this is for a one-sided test. Since *PASS* is finding the power for two-sided tests, we set FWER at 0.10. When the per-group sample size is 16, the all-pairs power is 0.80. When the per-group sample size is 21, the all-pairs power is 0.90. Note that since there is only one treatment that is different from the control, the all-pairs power is equal to the any-pairs power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Multiple Comparisons of Treatments vs. a Control (Simulation)**. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

| <u>Option</u> | <u>Value</u> |
|------------------------------------|---------------------|
| Design Tab | |
| Solve For | Power |
| MC Procedure..... | Dunnett |
| Simulations | 2000 |
| FWER (Alpha)..... | 0.10 |
| n (Sample Size Multiplier)..... | 16 21 |
| Group Sample Size Pattern | Equal |
| Grps 1 | 1 |
| Control Distribution(s) H0 | Normal(M0 S) |
| Control Distribution(s) H1 | Normal(M0 S) |
| Grps 2 | 1 |
| Group 2 Distribution(s) H0 | Normal(M0 S) |
| Group 2 Distribution(s) H1 | Normal(M1 S) |
| Grps 3 | 4 |
| Group 3 Distribution(s) H0 | Normal(M0 S) |
| Group 3 Distribution(s) H1 | Normal(M0 S) |
| Equivalence Margin | 0.01 |
| M0 (Mean H0) | -0.182574 |
| M1 (Mean H1) | 0.912871 |
| S..... | 1.0 |

Multiple Comparisons of Treatments vs. a Control (Simulation)

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Summary of Simulations of the Control Group and the 5 Treatment Groups MC Procedure: Dunnett's M.C. Test

| Sim. No. | Any-Pairs Power | Group Smpl. Size n | Total Smpl. Size N | All-Pairs Power | S.D. of Means Sm H1 | S.D. of Data SD H1 | Actual FWER | Target FWER | M0 | M1 | S |
|----------|------------------|--------------------|--------------------|------------------|---------------------|--------------------|------------------|-----------------|----------------|-----|-----|
| 1 | 0.795 (0.018) | 16.0 [0.777 | 96 0.812] | 0.795 (0.018) | 0.4 [0.777 | 1.0 0.812] | 0.095 (0.013) | 0.100 [0.082 | -0.2 0.107] | 0.9 | 1.0 |
| 2 | 0.894 (0.014) | 21.0 [0.880 | 126 0.907] | 0.894 (0.014) | 0.4 [0.880 | 1.0 0.907] | 0.112 (0.014) | 0.100 [0.098 | -0.2 0.125] | 0.9 | 1.0 |

Pool Size: 10000. Simulations: 1000. Run Time: 4.28 seconds.

For the first case when $n = 16$, PASS obtained a power of 0.795 which is very close to the value of 0.80 found by Muterjee et al. (1987). Indeed, 0.80 is within the confidence limits of 0.777 to 0.812. Similarly, the power for the second case when $n = 21$ is found as 0.894 which is very close to the article's value of 0.90.