Chapter 580

Pair-Wise Multiple Comparisons (Simulation)

Introduction

This procedure uses simulation analyze the power and significance level of three pair-wise multiple-comparison procedures: Tukey-Kramer, Kruskal-Wallis, and Games-Howell. For each scenario, two simulations are run: one estimates the significance level and the other estimates the power.

The term *multiple comparisons* refers to a set of two or more statistical hypothesis tests. The term *pair-wise multiple comparisons* refers to the set of all pairs of means that can be generated among the means of *k* groups. For example, suppose the levels of a factor with five groups are labeled A, B, C, D, and E. The ten possible paired-comparisons that could be made among the five groups are A-B, A-C, A-D, A-E, B-C, B-D, B-E, C-D, C-E, and D-E.

As the number of groups increases, the number of comparisons (pairs) increases dramatically. For example, a 5 group design has 10 pairs, a 10 group design has 45 pairs, and a 20 group design has 190 pairs. When several comparisons are made among the group means, the determination of the significance level of each individual comparison is much 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 H0) 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 $\mu_1, \mu_2, \cdots, \mu_k$ and unknown variance σ^2. Let $n_1, n_2, \cdots, n_k$ denote the number of subjects in each group.

The analysis of these responses is based on the sample means

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

and the pooled sample variance

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

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 Tukey-Kramer procedure, the Kruskal-Wallis procedure, and the Games-Howell procedure are the pair-wise multiple-comparison procedures that have been developed for this situation. The calculation of each of these tests is given next.

Tukey-Kramer

This test is referenced in Kirk (1982). It uses the critical values from the studentized-range distribution. For each pair of groups, the significance test between any two groups *i* and *j* is calculated by rejecting the null hypothesis of mean equality if

$$\frac{\left|\overline{Y}_i - \overline{Y}_j\right|}{\sqrt{\dfrac{\hat{\sigma}^2}{2}\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}} \geq q_{\alpha_f, k, v}$$

where

$$v = \sum_{i=1}^{k} n_i - k$$
$$= N - k$$

Kruskal-Wallis

This test is attributed to Dunn (1964) and is referenced in Gibbons (1976). It is a nonparametric, or distribution-free, test for which the assumption of normality is not necessary. It tests whether pairs of medians are equal using a rank test. Sample sizes of at least five (but preferably larger) for each treatment are recommended for 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 $\alpha = \alpha_f / \left(k(k-1) \right)$. For each pair of groups, i and j, the null hypothesis of equality is rejected if

$$\frac{|\overline{R}_i - \overline{R}_j|}{\sqrt{\dfrac{n(n+1)}{12}\left(\dfrac{1}{n_i} + \dfrac{1}{n_j} \right)}} \geq z_\alpha$$

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

Games-Howell

This test is referenced in Kirk (1982) page 120. It was developed for the case when the individual group variances cannot be assumed to be equal. It also uses critical values from the studentized-range distribution. For each pair of groups, i and j, the null hypothesis of equality is rejected if

$$\frac{\left| \overline{Y}_i - \overline{Y}_j \right|}{\sqrt{\dfrac{1}{2}\left(\dfrac{\hat{\sigma}_i^2}{n_i} + \dfrac{\hat{\sigma}_j^2}{n_j} \right)}} \geq q_{\alpha_f, k, v'}$$

where

$$v' = \frac{\left(\dfrac{\hat{\sigma}_i^2}{n_i} + \dfrac{\hat{\sigma}_j^2}{n_j} \right)^2}{\dfrac{\hat{\sigma}_i^4}{n_i^2 (n_i - 1)} + \dfrac{\hat{\sigma}_j^4}{n_j^2 (n_j - 1)}}$$

If any of the following conditions hold, then $v' = n_i + n_2 - 2$:

1. $9/10 \leq n_i / n_j \leq 10/9$

2. $9/10 \leq \left(\dfrac{\hat{\sigma}_i^2}{n_i} \right) / \left(\dfrac{\hat{\sigma}_j^2}{n_j} \right) \leq 10/9$

3. $4/5 \leq n_i / n_j \leq 5/4$ and $1/2 \leq \left(\dfrac{\hat{\sigma}_i^2}{n_i} \right) / \left(\dfrac{\hat{\sigma}_j^2}{n_j} \right) \leq 2$

4. $2/3 \leq n_i / n_j \leq 3/2$ and $3/4 \leq \left(\dfrac{\hat{\sigma}_i^2}{n_i} \right) / \left(\dfrac{\hat{\sigma}_j^2}{n_j} \right) \leq 4/3$

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 three groups labeled, A, B, and C. There are three paired comparisons in this experiment: A-B, A-C, and B-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 <u>alternative</u> 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 <u>null</u> 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 n 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

- **Tukey-Kramer**

  This is the most popular and the most often recommended.

- **Kruskal-Wallis**

  This is recommended when a nonparametric procedure is wanted.

- **Games-Howell**

  This is recommended when the variances of the groups are unequal.

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.

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 samples 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 *m1, m2, m3, …, mk* and this value is represented by *n*, the group sample sizes *N1, N2, N3, …, Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

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 ith 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.

- **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 $n1 = n2 = … = ng = n$. 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.

Grps [A – C] (Grps D – I 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)

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 Normal(4, 5) + 2 Normal (5, 6) is 4*4 + 2*5 = 26, but the mean of 4 Normal (4, 5) * 2 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 *M1* 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.

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 <u>much</u> 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

Show Various Reports & Plots

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 investigate the variety of response when an experiment is replicated under five different conditions. Previous studies have shown that the standard deviation within a group is 3.0. Researchers want to detect a shift in the mean of 3.0 or more. To accomplish this, they set the means of the first four groups to zero and the mean of the fifth group to 3.0. They want to investigate sample sizes of 5, 10, 15, and 20 subjects per group.

Although they will conduct an F-test on the data, their primary analysis will be a set of Tukey-Kramer 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 **Pair-Wise Multiple Comparisons (Simulation)** procedure window by expanding **Means**, then then clicking on **Multiple Comparisons**, and then clicking on **Pair-Wise Multiple Comparisons (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**.

Option	**Value**
Design Tab	
Solve For	**Power**
MC Procedure..	**Tukey-Kramer**
Simulations	**2000**
FWER (Alpha)..	**0.05**
n (Sample Size Multiplier)......................	**5 10 15 20**
Group Sample Size Pattern	**Equal**
Grps 1	**4**
Group 1 Distribution(s)	H0	**Normal(M0 S)**
Group 1 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)**
Equivalence Margin	**0.5**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**3**
S..	**3**

Annotated Output

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

Simulation Summary Report

Summary of Simulations of 5 Groups
MC Procedure: Tukey-Kramer 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	0.251	5.0	25	0.014	1.2	3.0	0.043	0.050	0.0	3.0	3.0
	(0.019)	[0.232	0.269]	(0.005)	[0.008	0.019]	(0.009)	[0.034 0.051]			
2	0.562	10.0	50	0.077	1.2	3.0	0.052	0.050	0.0	3.0	3.0
	(0.022)	[0.540	0.583]	(0.012)	[0.065	0.088]	(0.010)	[0.042 0.062]			
3	0.770	15.0	75	0.200	1.2	3.0	0.057	0.050	0.0	3.0	3.0
	(0.018)	[0.751	0.788]	(0.018)	[0.182	0.217]	(0.010)	[0.046 0.067]			
4	0.898	20.0	100	0.356	1.2	3.0	0.060	0.050	0.0	3.0	3.0
	(0.013)	[0.884	0.911]	(0.021)	[0.335	0.377]	(0.010)	[0.049 0.070]			

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

Summary of Simulations Report Definitions
H0: the null hypothesis that each pair of group means are equal.
H1: the alternative hypothesis that at least one pair of group means are not equal.
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 FWER.
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 5 groups has an average group sample size of 5.0 for a total sample size
of 25. This design achieved an any-pair power of 0.2505 and an all-pair power of 0.0135 using
the Tukey-Kramer M.C. Test with a target family-wise error rate of 0.050 and an actual target
family-wise error rate 0.043. 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); N(M0 S); and N(M0 S) and the alternative
distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M1 S). Other parameters used in the
simulation were: M0 = 0.0, M1 = 3.0, 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 <u>any</u> 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.

All-Pairs Power

This is the probability of detecting <u>all</u> 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 Sm|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 Sm|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

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

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 5 Groups
MC Procedure: Tukey-Kramer 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	10	0.057	0.006	0.043	0.050	0.006	0.005	0.007
2	10	0.072	0.007	0.052	0.050	0.007	0.005	0.011
3	10	0.075	0.008	0.057	0.050	0.008	0.006	0.011
4	10	0.075	0.008	0.060	0.050	0.008	0.004	0.010

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 pair-wise multiple comparisons.

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 5 Groups
MC Procedure: Tukey-Kramer M.C. Test

Sim. No.	No. of Equal/ Uneq. Pairs	Mean No. of False Pos.	Mean No. of False Neg.	Prop. Errors	Prop. Equal that were Detect.	Prop. Uneq. that were Undet.	(FDR) Prop. Detect. that were Equal	Prop. Undet. that were Uneq.	All Uneq. Pairs Power	Any Uneq. Pairs Power	Mean Pairs Power	Min Pairs Power	Max Pairs Power
1	6/4	0.04	3.58	0.362	0.007	0.896	0.091	0.375	0.014	0.251	0.046	0.004	0.109
2	6/4	0.04	2.85	0.289	0.007	0.711	0.035	0.323	0.077	0.562	0.120	0.006	0.302
3	6/4	0.03	2.11	0.214	0.006	0.527	0.017	0.261	0.200	0.770	0.192	0.002	0.478
4	6/4	0.03	1.41	0.144	0.005	0.352	0.012	0.191	0.356	0.898	0.262	0.003	0.658

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 means 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.

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 all 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 <u>all</u> of the pairs that were different in the H1 simulation.

Any Uneq. Pairs Power

This is the probability of detecting <u>any</u> 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 = 3.0, S = 3.0
MC Procedure: Tukey-Kramer M.C. Test

Hypo. Type	Groups	Group Labels	n/N	Group Mean	Ave. S.D.	Simulation Model
H0	1-4	A1-A4	5/25	0.0	2.9	N(M0 S)
H0	5	B1	5/25	0.0	3.0	N(M0 S)
H0	All			Sm=0.0	3.0	
H1	1-4	A1-A4	5/25	0.0	3.0	N(M0 S)
H1	5	B1	5/25	3.0	3.0	N(M1 S)
H1	All			Sm=1.2	3.0	

Detailed Model Report for Simulation No. 2

Hypo. Type	Groups	Group Labels	n/N	Group Mean	Ave. S.D.	Simulation Model
H0	1-4	A1-A4	10/50	0.0	3.0	N(M0 S)
H0	5	B1	10/50	0.0	3.0	N(M0 S)
H0	All			Sm=0.0	3.0	
H1	1-4	A1-A4	10/50	0.0	3.0	N(M0 S)
H1	5	B1	10/50	3.0	3.0	N(M1 S)
H1	All			Sm=1.2	3.0	

Detailed Model Report for Simulation No. 3

Hypo. Type	Groups	Group Labels	n/N	Group Mean	Ave. S.D.	Simulation Model
H0	1-4	A1-A4	15/75	0.0	3.0	N(M0 S)
H0	5	B1	15/75	0.0	3.0	N(M0 S)
H0	All			Sm=0.0	3.0	
H1	1-4	A1-A4	15/75	0.0	3.0	N(M0 S)
H1	5	B1	15/75	3.0	3.0	N(M1 S)
H1	All			Sm=1.2	3.0	

Detailed Model Report for Simulation No. 4

Hypo. Type	Groups	Group Labels	n/N	Group Mean	Ave. S.D.	Simulation Model
H0	1-4	A1-A4	20/100	0.0	3.0	N(M0 S)
H0	5	B1	20/100	0.0	3.0	N(M0 S)
H0	All			Sm=0.0	3.0	
H1	1-4	A1-A4	20/100	0.0	3.0	N(M0 S)
H1	5	B1	20/100	3.0	3.0	N(M1 S)
H1	All			Sm=1.2	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.

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	A1	A2	A3	A4	B1
A1	0.0		0.006	0.006	0.011	0.109*
A2	0.0	0.006		0.006	0.004	0.106*
A3	0.0	0.005	0.006		0.008	0.097*
A4	0.0	0.005	0.007	0.007		0.107*
B1	3.0	0.005	0.005	0.007	0.006	

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

Group	Means	A1	A2	A3	A4	B1
A1	0.0		0.007	0.007	0.008	0.294*
A2	0.0	0.006		0.007	0.006	0.280*
A3	0.0	0.011	0.007		0.006	0.302*
A4	0.0	0.005	0.007	0.005		0.281*
B1	3.0	0.008	0.009	0.008	0.007	

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

Group	Means	A1	A2	A3	A4	B1
A1	0.0		0.005	0.002	0.007	0.478*
A2	0.0	0.006		0.006	0.007	0.477*
A3	0.0	0.007	0.008		0.005	0.473*
A4	0.0	0.009	0.008	0.007		0.465*
B1	3.0	0.006	0.011	0.007	0.008	

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

Group	Means	A1	A2	A3	A4	B1
A1	0.0		0.003	0.005	0.004	0.645*
A2	0.0	0.008		0.008	0.007	0.650*
A3	0.0	0.006	0.008		0.004	0.640*
A4	0.0	0.010	0.007	0.009		0.658*
B1	3.0	0.010	0.007	0.004	0.007	

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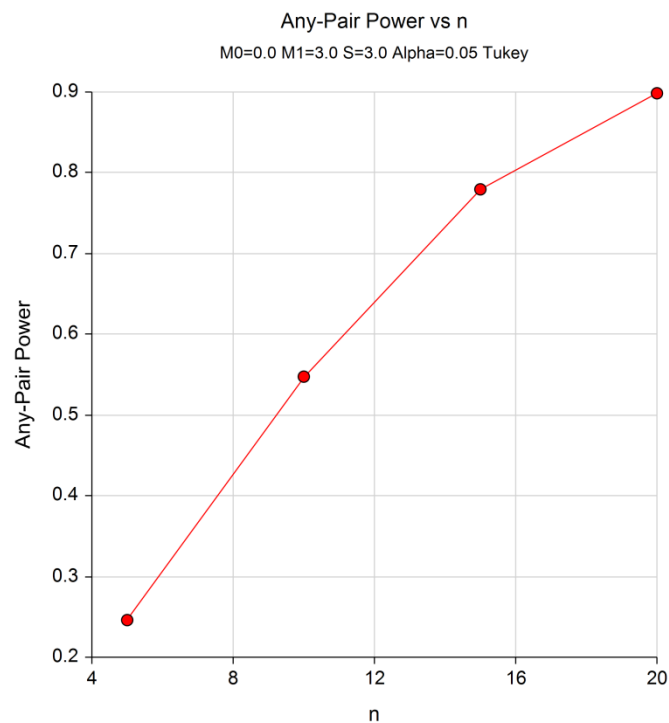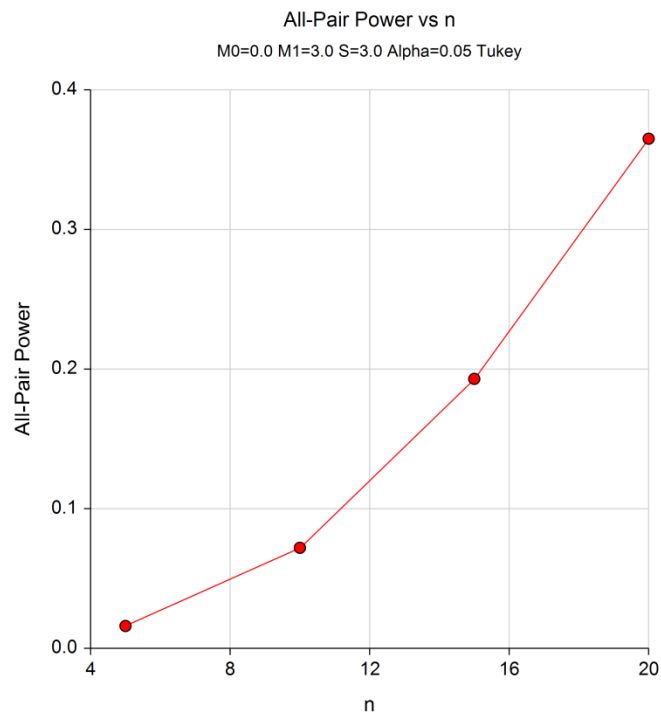
These plots give a visual presentation of the all-pairs power values and the any-pair power values.

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 **Pair-Wise Multiple Comparisons (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Pair-Wise Multiple Comparisons (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**.

Option	Value
Design Tab	
Solve For	**Power**
MC Procedure...	**Tukey-Kramer**
Simulations	**2000**
FWER (Alpha)...	**0.05**
n (Sample Size Multiplier).......................	**5 10 15 20**
Group Sample Size Pattern.....................	**Equal**
Grps 1	**4**
Group 1 Distribution(s)	H0	**Normal(M0 S)**
Group 1 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)**
Equivalence Margin	**0.5**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**3**
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 5 Group Means

Sim. No.	Total Sample Size	Target Alpha	Tukey Kramer All-Pair Power	Kruskal Wallis All-Pair Power	Games Howell All-Pair Power	Tukey Kramer Any-Pair Power	Kruskal Wallis Any-Pair Power	Games Howell Any-Pair Power
1	25	0.050	0.013	0.000	0.003	0.238	0.132	0.209
2	50	0.050	0.076	0.010	0.055	0.555	0.456	0.523
3	75	0.050	0.201	0.076	0.163	0.779	0.697	0.755
4	100	0.050	0.350	0.197	0.318	0.890	0.842	0.883

Pool Size: 10000. Simulations: 2000. Run Time: 33.82 seconds.

Family-Wise FWER Comparison for Testing the 5 Group Means

Sim. No.	Total Sample Size	Target FWER	Tukey Kramer FWER	Kruskal Wallis FWER	Games Howell FWER
1	25	0.050	0.039	0.023	0.064
2	50	0.050	0.050	0.039	0.059
3	75	0.050	0.051	0.033	0.055
4	100	0.050	0.051	0.040	0.058
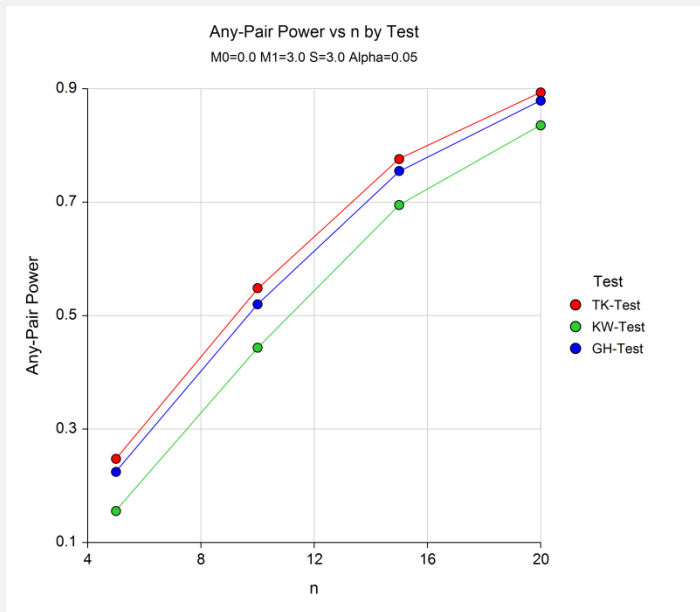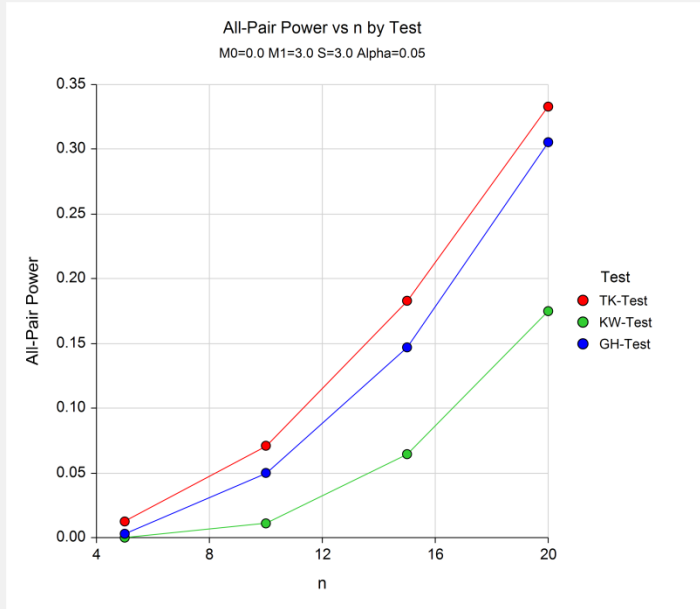
These reports show the power and FWER of each of the three multiple comparison procedures. In these simulations of groups from the normal distributions with equal variances, we see that the Tukey-Kramer procedure is the champion.

Example 3 – Validation using Ramsey

Ramsey (1978) presents the results of a simulation study that compared the all-pair power of several different multiple comparison procedures. On page 483 of this article, he presents the results of a simulation in which there were four groups: two with means of -0.7 and two with means of 0.7. The standard deviation was 1.0 and the FWER was 0.05. Tukey's multiple comparison procedure was used in the simulation. The sample size was 16 per group. Using a simulation of 1000 iterations, the all-pairs power was calculated as 0.723. Note that a confidence interval for this estimated all-pairs power is (0.703 to 0.759).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Pair-Wise Multiple Comparisons (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**.

Option	Value
Design Tab	
Solve For	**Power**
MC Procedure...	**Tukey-Kramer**
Simulations	**2000**
FWER (Alpha)...	**0.05**
n (Sample Size Multiplier).......................	**16**
Group Sample Size Pattern	**Equal**
Grps 1	**2**
Group 1 Distribution(s) | H0	**Normal(M0 S)**
Group 1 Distribution(s) | H1	**Normal(M0 S)**
Grps 2	**2**
Group 2 Distribution(s) | H0	**Normal(M0 S)**
Group 2 Distribution(s) | H1	**Normal(M1 S)**
Equivalence Margin	**0.1**
M0 (Mean|H0)	**-0.7**
M1 (Mean|H1)	**0.7**
S...	**1**

Output

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

Numeric Results

Summary of Simulations of 5 Groups
MC Procedure: Tukey-Kramer 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	0.996	16.0	64	0.729	0.7	1.0	0.052	0.050	-0.7	0.7	1.0
	(0.004)	[0.992	1.000]	(0.028)	[0.701	0.757]	(0.014)	[0.038	0.066]		

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

Note that the value found by PASS of 0.729 is very close to the value of 0.723 found by Ramsey (1978). More importantly, the value found by PASS is inside the confidence limits of Ramsey's study.

We ran the simulation five more times and obtained 0.727, 0.723, 0.714, 0.740, and 0.732. We also ran the simulation with 10,000 iterations and obtained a power of 0.736 with a confidence interval of (0.727 to 0.745).

Example 4 – Selecting a Multiple Comparison Procedure when the Data Contain Outliers

This example will investigate the impact of outliers on the power and precision of the various multiple comparison procedures when there are five groups.

A mixture of two normal distributions will be used to randomly generate outliers. The mixture will draw 95% of the data from a normal distribution with mean zero and variance one. The other 5% of the data will come from a normal distribution with mean zero and variance that ranges from one to ten. In the alternative distributions, two will have means of zero and the other three will have means of one.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by expanding **Means**, then clicking on **Multiple Comparisons**, and then clicking on **Pair-Wise Multiple Comparisons (Simulation)**. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Option   **Value**

Design Tab
Solve For .... **Power**
Multiple Comparison Procedure .... **Tukey Kramer**
Simulations **2000**
FWER (Alpha)....................................... **0.05**
n (Sample Size Multiplier)...................... **16**
Group Sample Size Pattern.................... **Equal**
Grps 1 .... **2**
Group 1 Distribution(s) | H0 .... **Normal(M0 S)[95];Normal(M0 A)[5]**
Group 1 Distribution(s) | H1 .... **Normal(M0 S)[95];Normal(M0 A)[5]**
Grps 2 .... **3**
Group 2 Distribution(s) | H0 .... **Normal(M0 S)[95];Normal(M0 A)[5]**
Group 2 Distribution(s) | H1 .... **Normal(M1 S)[95];Normal(M1 A)[5]**
Equivalence Margin **0.5**
M0 (Mean|H0)....................................... **0**
M1 (Mean|H1)....................................... **1**
S.. **1**
A.. **1 5 10**

Reports Tab
Show Comparative Reports.................... **Checked**
Show Comparative Plots .... **Checked**

Output

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

## Numeric Results

Power Comparison for Testing the 5 Group Means

Sim. No.	Total Sample Size	Target Alpha	Tukey Kramer All-Pair Power	Kruskal Wallis All-Pair Power	Games Howell All-Pair Power	Tukey Kramer Any-Pair Power	Kruskal Wallis Any-Pair Power	Games Howell Any-Pair Power
1	80	0.050	0.131	0.043	0.098	0.892	0.837	0.890
2	80	0.050	0.019	0.017	0.016	0.592	0.751	0.701
3	80	0.050	0.006	0.022	0.005	0.341	0.759	0.577

Pool Size: 10000. Simulations: 2000. Run Time: 33.82 seconds.

Family-Wise Error-Rate Comparison for Testing the 5 Group Means

Sim. No.	Total Sample Size	Target FWER	Tukey Kramer FWER	Kruskal Wallis FWER	Games Howell FWER
1	80	0.050	0.046	0.036	0.049
2	80	0.050	0.044	0.044	0.040
3	80	0.050	0.033	0.038	0.025
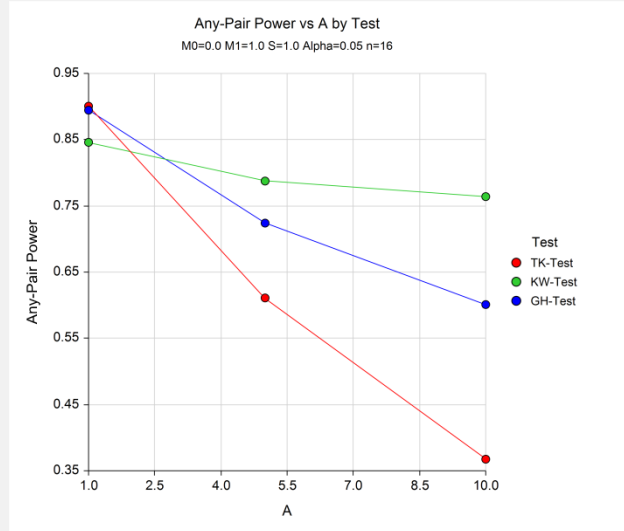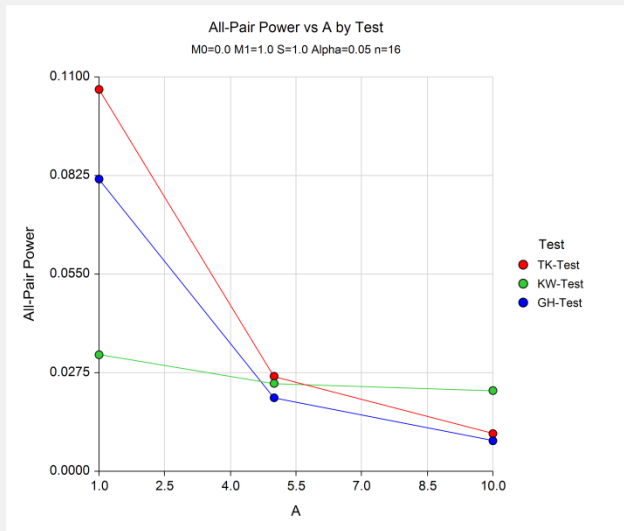
These reports show the power and FWER of each of the three multiple comparison procedures. We note that when the variances are equal (A = 1), the Tukey-Kramer procedure performs only slightly better than the others. However, as the number of outliers is increased, the Kruskal-Wallis procedure emerges as the better choice. Also note that in the case with many outliers (Simulation 3), the FWER of all procedures is below the target value.