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 $\mu_1, \mu_1, \cdots, \mu_k$ and unknown variance σ^2. Let $n_1, n_1, \cdots, n_k$ denote the number of subjects in each group.

The analysis of these responses is based on the sample means

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

and the pooled sample variance

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

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 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

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

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 $\alpha = \alpha_f \,/\, \big(k(k-1)\big)$. For each pair of groups, *i* and *j*, the null hypothesis of equality is rejected if

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

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

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

where

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

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

1. $9/10 \leq n_i \,/\, n_j \leq 10/9$
2. $9/10 \leq \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \leq 10/9$
3. $4/5 \leq n_i \,/\, n_j \leq 5/4$ and $1/2 \leq \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \leq 2$
4. $2/3 \leq n_i \,/\, n_j \leq 3/2$ and $3/4 \leq \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \leq 4/3$

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 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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Random Seed	**3480541** (for Reproducibility)
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
Set A Grps	**4**
Set A Group Distribution(s)|H0	**Normal(M0 S)**
Set A Group Distribution(s)|H1	**Normal(M0 S)**
Set B Grps	**1**
Set B Group Distribution(s)|H0	**Normal(M0 S)**
Set B Group Distribution(s)|H1	**Normal(M1 S)**
Set C Grps	**0**
Equivalence Margin	**0.5**
M0 (Mean|H0) Parameter Value(s)	**0**
M1 (Mean|H1) Parameter Value(s)	**3**
Parameter 1 Label	**S**
Parameter 1 Value(s)	**3**

Output

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

Simulation Summary Report

Summary of Simulations

Solve For:	Power
Multiple-Comparison Procedure:	Tukey-Kramer M.C. Test
Number of Groups:	5
Number of Comparisons:	10

		Sample Size			Standard Deviation of Group Means	Within-Group Standard Deviation	Family-Wise Error Rate				
Scenario	Any-Pair Power	Average Group n	Total N	All-Pairs Power	Sm|H1	SD|H1	Actual FWER	Target FWER	M0	M1	S
1	0.262 (0.019) [0.242 0.281]	5	25	0.017 (0.006) [0.011 0.022]	1.2	3	0.042 (0.009) [0.033 0.051]	0.05	0	3	3
2	0.560 (0.022) [0.538 0.581]	10	50	0.077 (0.012) [0.065 0.089]	1.2	3	0.054 (0.01) [0.044 0.063]	0.05	0	3	3
3	0.785 (0.018) [0.766 0.803]	15	75	0.190 (0.017) [0.173 0.207]	1.2	3	0.049 (0.009) [0.04 0.058]	0.05	0	3	3
4	0.895 (0.013) [0.881 0.908]	20	100	0.338 (0.021) [0.317 0.359]	1.2	3	0.054 (0.01) [0.044 0.064]	0.05	0	3	3

Pool Size: 10000. Simulations: 2000. Run Time: 1.81 seconds.
Equivalence Margin: 0.5. User-Entered Random Seed: 3480541

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.
Pair	Each comparison of two-group means is a "pair."
Any-Pair Power	The estimated probability of detecting at least one unequal pair. The second row provides the precision and a 95% confidence interval for Any-Pair Power, (Any-Pair Power Precision) [95% LCL and UCL], based on the number of simulations.
n	The average of the group sample sizes.
N	The combined sample size of all groups.
All-Pairs Power	The estimated probability of detecting all unequal pairs. The second row provides the precision and a 95% confidence interval for All-Pairs Power, (All-Pairs Power Precision) [95% LCL and UCL], based on the number of simulations.
Sm|H1	The standard deviation of the set of group means under H1.
SD|H1	The pooled, within-group standard deviation under H1.
FWER	Family-Wise Error Rate. The probability of detecting at least one equal pair assuming H0.
Actual FWER	The FWER estimated by the alpha simulations.
Target FWER	The user-specified FWER. The second row provides the precision and a 95% confidence interval for FWER, (FWER Precision) [95% LCL and UCL], based on the number of simulations.
M0, M1, S, etc.	The value(s) of the user-specified distribution parameters.

Summary Statements

A one-way design with 5 groups will be used to test each group mean against every other group mean (pair-wise). The comparisons will be made using Tukey-Kramer pairwise tests with a target overall (family-wise) Type I error rate (α) of 0.05. Based on 2000 simulations of the null distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M0 S), and of the alternative distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M1 S) (where M0 = 0, M1 = 3, and S = 3), with an average group sample size of 5 (for a total of 25 subjects), the any-pair power (probability of detecting at least one of the pair differences) is 0.2615, and the all-pair power (probability of detecting all of the pair differences) is 0.0165. (Additional details: The standard deviation of the group means under the null hypothesis is 0, and the standard deviation of the group means under the alternative hypothesis is 1.2. The average of the within-group standard deviations, assuming the alternative hypothesis distributions, is 3. The actual family-wise Type I error rate (α), based on the null hypothesis distribution simulations, is 0.042.)

Dropout-Inflated Sample Size

Average Group Sample Size n	Group	Dropout Rate	Sample Size Ni	Dropout-Inflated Enrollment Sample Size Ni'	Expected Number of Dropouts Di
5	1 - 5	20%	5	7	2
	Total		25	35	10
10	1 - 5	20%	10	13	3
	Total		50	65	15
15	1 - 5	20%	15	19	4
	Total		75	95	20
20	1 - 5	20%	20	25	5
	Total		100	125	25

n	The average group sample size.
Group	Lists the group numbers.
Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
Ni	The evaluable sample size for each group at which power is computed (as entered by the user). If Ni subjects are evaluated out of the Ni' subjects that are enrolled in the study, the design will achieve the stated power.
Ni'	The number of subjects that should be enrolled in each group in order to obtain Ni evaluable subjects, based on the assumed dropout rate. Ni' is calculated by inflating Ni using the formula Ni' = Ni / (1 - DR), with Ni' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
Di	The expected number of dropouts in each group. Di = Ni' - Ni.

Dropout Summary Statements

Anticipating a 20% dropout rate, group sizes of 7, 7, 7, 7, and 7 subjects should be enrolled to obtain final group sample sizes of 5, 5, 5, 5, and 5 subjects.

.
.
.

References

Devroye, Luc. 1986. Non-Uniform Random Variate Generation. Springer-Verlag. New York.
Hsu, Jason. 1996. Multiple Comparisons: Theory and Methods. Chapman & Hall. London.
Kirk, Roger E. 1982. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole. Pacific Grove, California.
Matsumoto, M. and Nishimura,T. 1998. 'Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator.' ACM Trans. On Modeling and Computer Simulations.
Ramsey, Philip H. 1978. 'Power Differences Between Pairwise Multiple Comparisons', JASA, vol. 73, no. 363, pages 479-485.

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.

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.

Standard Deviation of Group 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.

Within-Group Standard Deviation 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 (Alpha) Simulations

Error Rate Summary from H0 (Alpha) Simulations

Multiple-Comparison Procedure:	Tukey-Kramer M.C. Test
Number of Groups:	5
Number of Comparisons:	10

		Type I Errors		Family-Wise Error Rate		Individual Comparison Alpha		
Scenario	Total Sample Size	Average Number	Proportion	Actual	Target	Average	Minimum	Maximum
1	25	0.067	0.007	0.042	0.05	0.007	0.003	0.008
2	50	0.070	0.007	0.054	0.05	0.007	0.005	0.010
3	75	0.066	0.007	0.049	0.05	0.007	0.004	0.009
4	100	0.070	0.007	0.054	0.05	0.007	0.005	0.009

Pair	Two groups whose means will be compared.
Type I Error	Rejecting the hypothesis that two groups have equal means when the means are actually equal.
Family	The set of all possible pairs of the groups.
Scenario	An identification number that signifies a set of options. This is used across the various reports.
Total Sample Size	The combined sample size of all groups.
Average Number of Type I Errors	The average number of Type I errors per family (set).
Proportion of Type I Errors	The proportion of pairs with equal means that were falsely concluded to be unequal.
FWER	The Family-Wise Error Rate. The probability that at least one of the pairs was falsely concluded as being unequal among the Alpha simulations.
Actual FWER	The computed FWER value from the Alpha simulations.
Target FWER	The user-specified value of the FWER.
Average Indiv. Comp. Alpha	The average of all individual comparison alphas.
Minimum Indiv. Comp. Alpha	The minimum of all individual comparison alphas.
Maximum Indiv. Comp. Alpha	The maximum of all individual comparison alphas.

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.

Number of Comparisons

This value is shown in the subtitles. 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.

Total Sample Size

The combined sample size of all groups.

Average Number of Type-1 Errors

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

Proportion Type-1 Errors

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

Actual Family-Wise Error Rate

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

Target Family-Wise Error Rate

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

Average Individual Comparison Pairs Alphas

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.

Minimum Individual Comparison Pairs Alphas

This is the minimum of all the individual comparison alphas.

Maximum Individual Comparison Pairs Alphas

This is the maximum of all the individual comparison alphas.

Error Rate Summary for H1 (Power) Simulations

Error Rate Summary from H1 (Power) Simulations

Multiple-Comparison Procedure: Tukey-Kramer M.C. Test
Number of Groups: 5
Number of Comparisons: 10

	Number of Pairs		Average Number			Proportion of Pairs Whose Means Are				Power		Individual Comparison Power		
Scenario	Equal	Not Equal	False Pos.	False Neg.	Proportion of Pairs with Errors	Equal but Detected	Unequal but Undetected	Detected but Equal	Undetected but Unequal	All-Pairs	Any-Pair	Ave	Min	Max
1	6	4	0.04	3.55	0.359	0.007	0.887	0.088	0.373	0.017	0.262	0.049	0.004	0.116
2	6	4	0.03	2.85	0.288	0.006	0.713	0.028	0.323	0.077	0.560	0.118	0.004	0.299
3	6	4	0.05	2.06	0.211	0.008	0.516	0.023	0.257	0.190	0.785	0.198	0.005	0.500
4	6	4	0.04	1.43	0.147	0.006	0.358	0.015	0.194	0.338	0.895	0.261	0.005	0.645

Pair	Two groups whose means will be compared.
Type-I Error	Rejecting the hypothesis that two groups have equal means when the means are equal.
Family	The set of all possible combinations of the groups.
Scenario	An identification number that signifies a set of options. This is used across the various reports.
Equal	Number of pairs with equal means under H1.
Not Equal	Number of pairs with unequal means under H1.
False Pos.	Average number of pairs with equal means that are falsely concluded as being different.
False Neg.	Average number of pairs with unequal means that are falsely not detected (concluded as being different).
Pairs with Errors	The proportion of pairs that exhibit either a Type I or a Type II error.
Equal but Detected	The proportion of pairs whose means are equal but are concluded as being different (detected).
Unequal but Undetected	The proportion of pairs whose means are unequal but are concluded as being equal (undetected).
Detected but Equal	The proportion of pairs whose means concluded as being different (detected) but were actually equal. The is also known as the False Discovery Rate (FDR).
Undetected but Unequal	The proportion of pairs whose means concluded as being equal (undetected) but were actually unequal.
All-Pairs Power	The power of the test that all pairs with unequal means were also concluded to be unequal.
Any-Pair Power	The power of the test that at least one of the pairs with unequal means were also concluded to be unequal.

Average Indiv. Comp. Power	The average of all individual comparison powers.
Minimum Indiv. Comp. Power	The minimum of all individual comparison powers.
Maximum Indiv. Comp. Power	The maximum of all individual comparison powers.

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

Number of Pairs That Are Equal

The number of pairs for which the means were equal under H1.

Number of Pairs That Are Not Equal

The number of pairs for which the means were different under H1.

Average Number of 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.

Average Number of 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.

Proportion of Pairs with Errors

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

Proportion of Pairs Whose Means Are Equal but Detected

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

Proportion of Pairs Whose Means Are Unequal but Undetected

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

Proportion of Pairs Whose Means Are Detected but 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*.

Proportion of Pairs Whose Means Are Undetected but Unequal

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

All-Pairs Power

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

Any-Pair Power

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

Ave, Min, and Max Individual Comparison Powers

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

Detailed Model Reports

Detailed Model Report for Scenario 1

Target FWER = 0.05, M0 = 0, M1 = 3, S = 3
Multiple-Comparison Procedure: Tukey-Kramer M.C. Test
Number of Groups: 5

Hypothesis Type	Groups	Group Labels	n	N	Group Mean	Standard Deviation of Group Means Sm	Average Group Standard Deviation Ave(SD)	Simulation Model
H0	1-4	A1-A4	5	25	0		3	N(M0 S)
H0	5	B1	5	25	0		3	N(M0 S)
H0	All					0.0	3	
H1	1-4	A1-A4	5	25	0		3	N(M0 S)
H1	5	B1	5	25	3		3	N(M1 S)
H1	All					1.2	3	

Hypothesis Type	H0 for the alpha simulations and H1 for the power simulations.
Groups	The groups reported on this line. 'All' is used for the average of all groups.
Group Labels	The labels that are used in the individual alpha-level reports.
n	n is the average group size
N	N is the combined sample size.
Group Mean	The average of the groups specified on this line of the report.
Sm	The standard deviation of the mean values. This value provides an index of how different the individual means are.
Ave(SD)	The average standard deviation of all within-group standard deviations reported on this line.
Simulation Model	The distribution used to simulate data for the groups reported on this line.

(More Reports Follow)

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

Hypothesis 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 is the average sample size of the groups.

N

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.

Sm

This is the standard deviation of the group-mean values. This value provides an index of how different the individual means are.

Ave(SD)

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 in Scenario 1

Group	Means	A1	A2	A3	A4	B1
A1	0		0.009	0.008	0.007	0.112*
A2	0	0.006		0.006	0.008	0.111*
A3	0	0.007	0.007		0.004	0.116*
A4	0	0.007	0.008	0.003		0.113*
B1	3	0.006	0.008	0.007	0.005	

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.

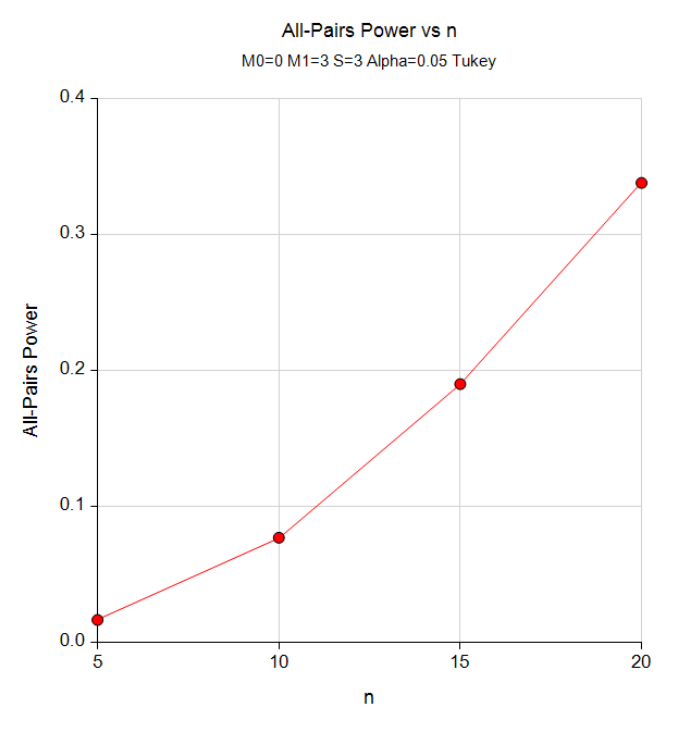
(More Reports Follow)

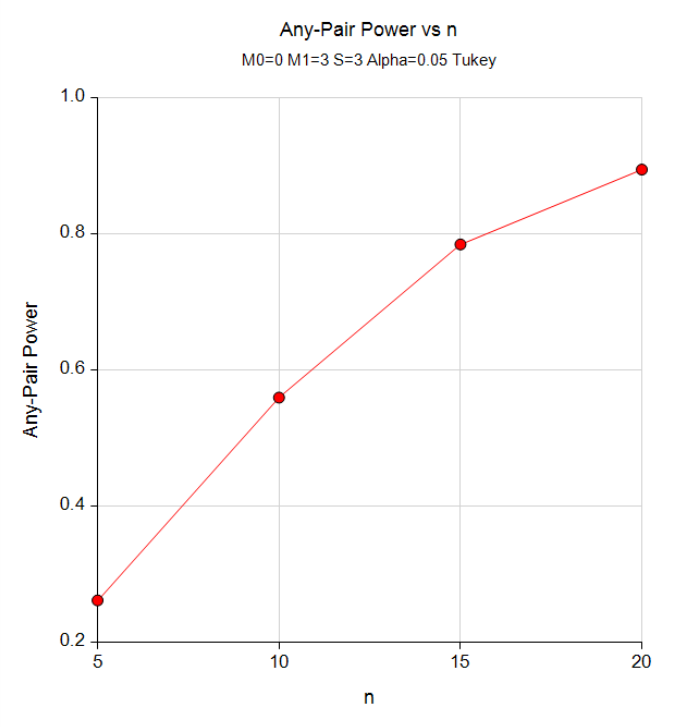
This report shows the individual probabilities of rejecting each pair. When a pair was actually different, the value is the power of that test. These power values are starred.

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

Plots





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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Random Seed	**3547196** (for Reproducibility)
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
Set A Grps	**4**
Set A Group Distribution(s)|H0	**Normal(M0 S)**
Set A Group Distribution(s)|H1	**Normal(M0 S)**
Set B Grps	**1**
Set B Group Distribution(s)|H0	**Normal(M0 S)**
Set B Group Distribution(s)|H1	**Normal(M1 S)**
Set C Grps	**0**
Equivalence Margin	**0.5**
M0 (Mean|H0) Parameter Value(s)	**0**
M1 (Mean|H1) Parameter Value(s)	**3**
Parameter 1 Label	**S**
Parameter 1 Value(s)	**3**

Reports Tab	
Show Comparative Reports	**Checked**

Comparative Plots Tab	
Comparative All-Pairs Power Plot	**Checked**
Comparative Any-Pair Power Plot	**Checked**

Output

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

Power Comparison for Testing the Pairs of Group Means

Number of Groups: 5

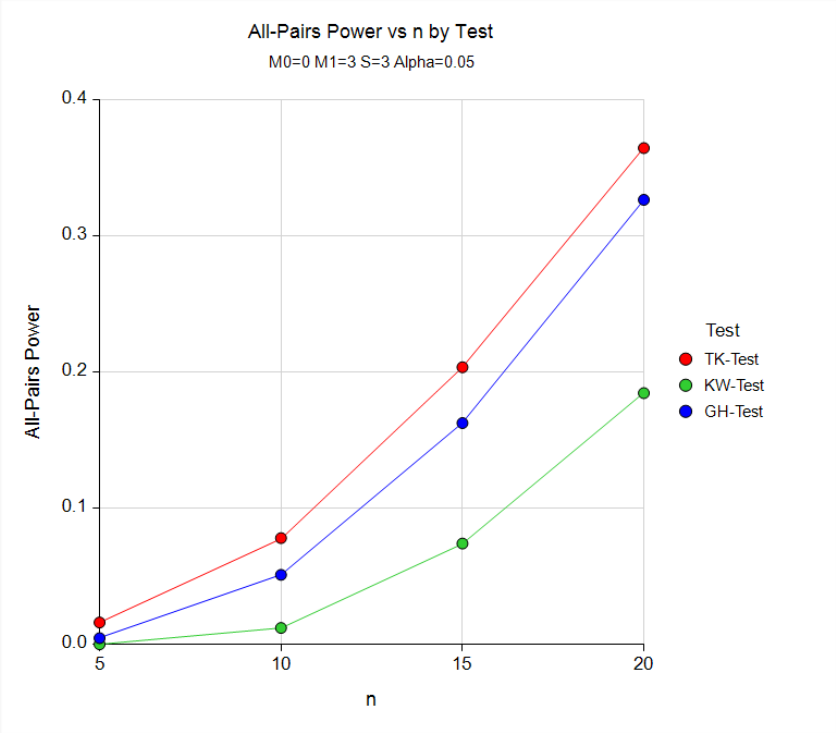
			Power					
			All-Pairs			Any-Pair		
Scenario	Total Sample Size	Target Alpha	Tukey Kramer	Kruskal Wallis	Games Howell	Tukey Kramer	Kruskal Wallis	Games Howell
1	25	0.05	0.016	0.000	0.005	0.273	0.153	0.238
2	50	0.05	0.078	0.012	0.051	0.558	0.450	0.516
3	75	0.05	0.204	0.074	0.163	0.773	0.690	0.751
4	100	0.05	0.365	0.185	0.327	0.897	0.846	0.883

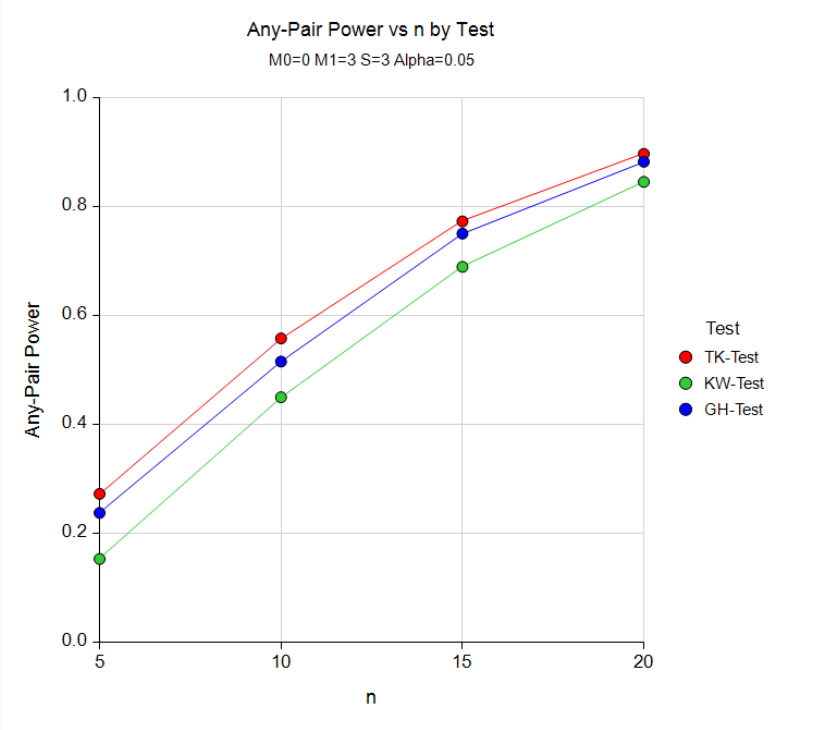
Family-Wise Error Rate Comparison for Testing the Pairs of Group Means

Number of Groups: 5

		Family-Wise Error Rate			
Scenario	Total Sample Size	Target	Tukey Kramer	Kruskal Wallis	Games Howell
1	25	0.05	0.045	0.027	0.071
2	50	0.05	0.054	0.035	0.055
3	75	0.05	0.050	0.038	0.057
4	100	0.05	0.057	0.042	0.061

Plots





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 (1978)

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).

For reproducibility, we'll use a random seed of 5123067.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Random Seed	**5123067** (for Reproducibility)
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**16**
Group Sample Size Pattern	**Equal**
Set A Grps	**2**
Set A Group Distribution(s)|H0	**Normal(M0 S)**
Set A Group Distribution(s)|H1	**Normal(M0 S)**
Set B Grps	**2**
Set B Group Distribution(s)|H0	**Normal(M0 S)**
Set B Group Distribution(s)|H1	**Normal(M1 S)**
Set C Grps	**0**
Equivalence Margin	**0.1**
M0 (Mean|H0) Parameter Value(s)	**-0.7**
M1 (Mean|H1) Parameter Value(s)	**0.7**
Parameter 1 Label	**S**
Parameter 1 Value(s)	**1**

Output

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

Summary of Simulations

Solve For:	Power
Multiple-Comparison Procedure:	Tukey-Kramer M.C. Test
Number of Groups:	4
Number of Comparisons:	6

Scenario	Any-Pair Power	Sample Size: Average Group n	Sample Size: Total N	All-Pairs Power	Standard Deviation of Group Means Sm|H1	Within-Group Standard Deviation SD|H1	Family-Wise Error Rate: Actual FWER	Family-Wise Error Rate: Target FWER	M0	M1	S
1	0.995 (0.003) [0.992 0.998]	16	64	0.719 (0.02) [0.699 0.738]	0.7	1	0.054 (0.01) [0.044 0.063]	0.05	-0.7	0.7	1

Pool Size: 10000. Simulations: 2000. Run Time: 0.49 seconds.
Equivalence Margin: 0.1. User-Entered Random Seed: 5123067

Note that the value found by **PASS** of 0.719 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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Random Seed	**3577400** (for Reproducibility)
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**16**
Group Sample Size Pattern	**Equal**
Set A Grps	**2**
Set A Group Distribution(s)|H0	**Normal(M0 S)[95];Normal(M0 A)[5]**
Set A Group Distribution(s)|H1	**Normal(M0 S)[95];Normal(M0 A)[5]**
Set B Grps	**3**
Set B Group Distribution(s)|H0	**Normal(M0 S)[95];Normal(M0 A)[5]**
Set B Group Distribution(s)|H1	**Normal(M1 S)[95];Normal(M1 A)[5]**
Set C Grps	**0**
Equivalence Margin	**0.5**
M0 (Mean|H0) Parameter Value(s)	**0**
M1 (Mean|H1) Parameter Value(s)	**1**
Parameter 1 Label	**S**
Parameter 1 Value(s)	**1**
Parameter 2 Label	**A**
Parameter 2 Value(s)	**1 5 10**

Reports Tab	
Show Comparative Reports	**Checked**

Comparative Plots Tab	
Comparative All-Pairs Power Plot	**Checked**
Comparative Any-Pair Power Plot	**Checked**

Output

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

Power Comparison for Testing the Pairs of Group Means

Number of Groups: 5

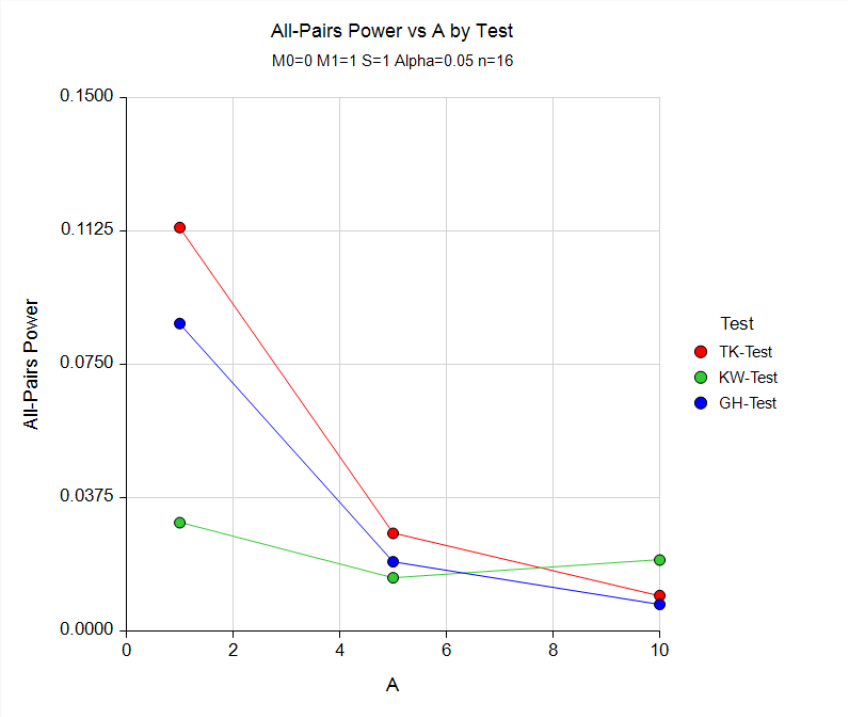
			Power					
			All-Pairs			Any-Pair		
Scenario	Total Sample Size	Target Alpha	Tukey Kramer	Kruskal Wallis	Games Howell	Tukey Kramer	Kruskal Wallis	Games Howell
1	80	0.05	0.114	0.031	0.087	0.890	0.837	0.884
2	80	0.05	0.028	0.015	0.020	0.589	0.763	0.714
3	80	0.05	0.010	0.020	0.008	0.310	0.730	0.544

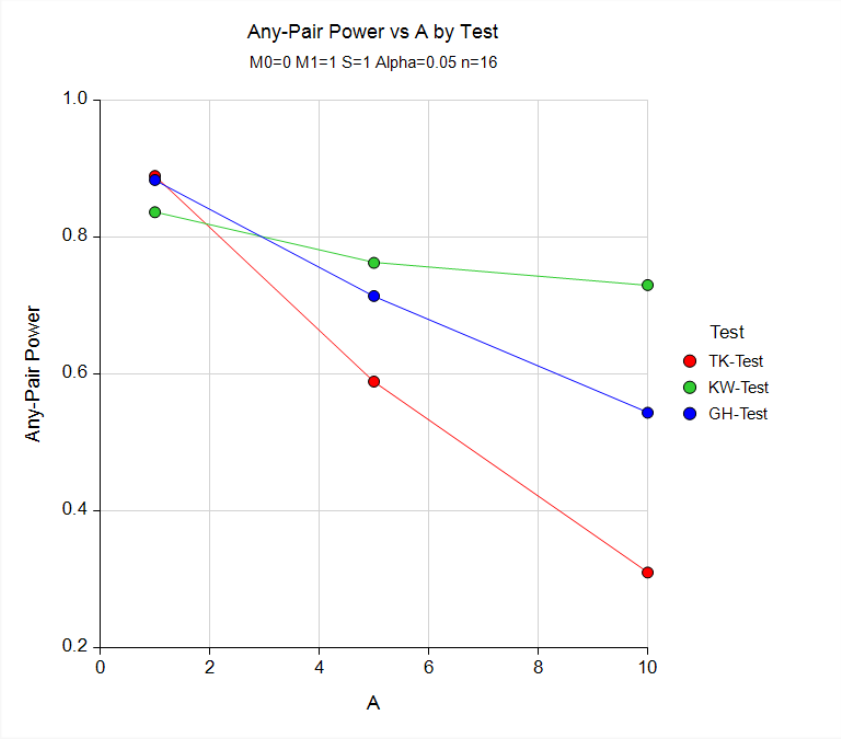
Family-Wise Error Rate Comparison for Testing the Pairs of Group Means

Number of Groups: 5

		Family-Wise Error Rate			
Scenario	Total Sample Size	Target	Tukey Kramer	Kruskal Wallis	Games Howell
1	80	0.05	0.051	0.040	0.056
2	80	0.05	0.046	0.039	0.039
3	80	0.05	0.039	0.038	0.025

Plots





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.