**One-Way Multiple Comparisons Tests[[1]](#footnote-1)©**

**Error Rates**

The **error rate per comparison**, ***αpc*** , is the probability of making a Type I error on a single comparison, assuming the null hypothesis is true.

 The **error rate per experiment**, ***αPE*** , is the expected number of Type I errors made when making *c* comparisons, assuming that each of the null hypotheses is true. It is equal to the sum of the per comparison alphas. If the per comparison alphas are constant, then ***αPE*** = ***cαpc***,

 The **familywise error rate**, ***αfw*** , is the probability of making one or more Type I errors in a family of *c* comparisons, assuming that each of the *c* null hypotheses is true. If the comparisons are independent of one another (orthogonal), then . For our example problem, evaluating four different teaching methods, if we were to compare each treatment mean with each other treatment mean, *c* would equal 6. If we were to assume those 6 comparisons to be independent of each other (they are not), then .

**Multiple *t* tests**

 One could just use multiple *t*-tests to make each comparison desired, but one runs the risk of greatly inflating the familywise error rate (the probability of making one or more Type I errors in a family of *c* comparisons) when doing so. One may use a series of **protected t‑tests** in this situation. This procedure requires that one first do an omnibus ANOVA involving all *k* groups. If the omnibus ANOVA is not significant, one stops and no additional comparisons are done. If that ANOVA is significant, one makes all the comparisons e wishes using *t*-tests. If you have equal sample sizes and homogeneity of variance, you can use , which pools the error variance across all *k* groups, giving you *N* - *k* degrees of freedom. If you have homogeneity of variance but unequal *n*’s use: . *MSE* is the error mean square from the omnibus ANOVA. If you had heterogeneous variances, you would need to compute separate variances *t*-tests, with adjusted *df.*

 The procedure just discussed (protected *t*-tests) is commonly referred to as **Fisher’s LSD test**. LSD stands for “Least Significant Difference.” If you were making comparisons for several pairs of means, and *n* was the same in each sample, and you were doing all your work by hand (as opposed to using a computer), you could save yourself by solving substituting the critical value of *t* in the formula above, entering the *n* and the *MSE*, and then solving for the (smallest) value of the difference between means which would be significant (the least significant difference). Then you would not have to compute a *t* for each comparison, you would just find the difference between the means and compare that to the value of the least significant difference.

 While this procedure is not recommended in general (it does not adequately control familywise alpha), there is one special case when it is the best available procedure, and that case is when *k* = 3. In that case Fisher’s procedure does hold *αfw* at or below the stated rate and has more power than other commonly employed procedures. The interested student is referred to the article “A Controlled, Powerful Multiple-Comparison Strategy for Several Situations,” by Levin, Serlin, and Seaman (*Psychological Bulletin*, 1994, *115*: 153-159) for details and discussion of how Fisher’s procedure can be generalized to other 2 *df* situations.

**Linear Contrasts**

 One may make simple or complex comparisons involving some or all of a set of *k* treatment means by using linear contrasts. Suppose I have five means, which I shall label A, B, C, D, and E. I can choose contrast coefficients to compare any one mean or subset of these means with any other mean or subset of these means. The sum of the contrast coefficients must be zero. All of the coefficients applied to the one set of means must be positive, all those applied to the other set must be negative. Means left out of the contrast get zero coefficients. There are some advantages of using a “**standard set of weights,**” so I shall do so here: The coefficients for the one set must equal +1 divided by the number of conditions in that set while those for the other set must equal -1 divided by the number of conditions in that other set. The sum of the absolute values of the coefficients must be 2.

 Suppose I want to contrast combined groups A and B with combined groups C, D, and E. A nonstandard set of coefficients is −3, −3, 2, 2, 2, and a standard set of coefficients would be −.5, −.5, 1/3, 1/3, 1/3. If I wanted to contrast C with combined D and E , a nonstandard set of coefficients would be 0, 0, -2, 1, 1 and a standard set of coefficients would be 0, 0, −1, .5, .5.

 A **standard contrast** is computed as .

 To test the significance of a contrast, compute a contrast sum of squares this way:

. When the sample sizes are equal, this simplifies to . Each contrast will have only one treatment *df*, so the contrast *MS* is the same as the contrast *SS*. To get an *F* for the contrast just divide it by an appropriate *MSE* (usually that which would be obtained were one to do an omnibus ANOVA on all *k* treatment groups).

 For our example problem, suppose we want to compare combined groups C and D with combined groups A and B. The A, B, C, D means are 2, 3, 7, 8, and the coefficients are −.5, −,5, +.5, +.5. . Note that the value of the contrast is quite simply the difference between the mean of combined groups C and D (7.5) and the mean of combined groups A and B (2.5). , and *F*(1, 16) = 125/.5 = 250, *p* << .01.

 To construct **a confidence interval about **, simply go out in each direction ., where . With equal sample sizes, this simplifies to . When one is constructing multiple confidence intervals, one can use Bonferroni to adjust the per contrast alpha. Such intervals have been called simultaneous or joint confidence intervals. For the contrast above, . With no adjustment of the per-comparison alpha, and *df* = 16, a 95% confidence interval is 5 ± 2.12(.3162), which extends from 4.33 to 5.67.

 A population **standardized contrast**, , can be estimated by , where *s* is the standard deviation of just one of the groups being compared (Glass’ Δ), the pooled standard deviation of the two groups being compared (Hedges’ *g*), or the pooled standard deviation of all of the groups (the square root of the *MSE*). For the contrast above, , a whopper effect.

 SAS and other statistical software can be used to obtain the *F* for a specified contrast. Having obtained a contrast *F* from your computer program, you can compute . For our contrast, .

 An **approximate confidence interva**l for a standardized contrast *d* can be computed simply by taking the confidence interval for the contrast and dividing its endpoints by the pooled standard deviation (square root of *MSE*). In this case the confidence interval amounts to , where . For our contrast,  and a 95% confidence interval is 7.07 ±2.12(.447), running from 6.12 to 8.02. More simply, we take the unstandardized confidence interval, which runs from 4.33 to 5.67, and divide each end by the standard deviation (.707) and obtain 6.12 to 8.02.

 At <http://www.psy.unsw.edu.au/research/PSY.htm> one can obtain **PSY: A Program for Contrast Analysis**, by Kevin Bird, Dusan Hadzi-Pavlovic, and Andrew Isaac. This program computes unstandardized and approximate standardized confidence intervals for contrasts with between-subjects and/or within/subjects factors. It will also compute simultaneous confidence intervals. Contrast coefficients are provided as integers, and the program converts them to standard weights. For an example of the use of the PSY program, see my document [PSY: A Program for Contrast Analysis](http://core.ecu.edu/psyc/wuenschk/docs30/PSY-ANOVA1.doc).

 An **exact confidence interval for a standardized contrast** involving independent samples can be computed with my SAS program [Conf\_Interval-Contrast.sas](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm). Enter the contrast *t* (the square root of the contrast *F*, 15.81 for our contrast), the *df* (16), the sample sizes (5, 5, 5, 5), and the standard contrast coefficients (−.5, −.5, .5, .5) and run the program. You obtain a confidence interval that extends from 4.48 to 9.64. Notice that this confidence interval is considerably wider than that obtained by the earlier approximation.

 One can also use ***η2* or partial *η2*** as a measure of the strength of a contrast, and use my program [Conf-Interval-R2-Regr.sas](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm) to construct a CI. For ***η2*** simply take the *SScontrast* and divide by the *SSTotal*. For our contrast, that yields *η2* = 125/138 = .9058. To get the confidence interval for *η2* we need to compute a modified contrast *F*, adding to the error term all variance not included in the contrast and all degrees of freedom not included in the contrast.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | *SS* | *df* | *MS* | *F* |
| ~~Teaching Method~~ | ~~130~~ | ~~3~~ | ~~43.33~~ | ~~86.66~~ |
|  AB vs. CD | 125 | 1 | 125 | 250 |
| Error | 8+5=13 | 16+2=18 | 13/18=.722 | 173.077 |
| Total | 138 | 19 |  |  |

. Feed that *F* and *df* to my SAS program and you obtain an *η2* of .9058 with a confidence interval that extends from .78 to .94.

 Alternatively, one can compute a **partial *η2*** as . Notice that this excludes from the denominator all variance that is explained by differences among the groups that are not captured by the tested contrast.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | *SS* | *df* | *MS* | *F* |
| Teaching Method | 130 | 3 | 43.33 | 86.66 |
|  AB vs. CD | 125 | 1 | 125 | 250 |
| Error | 8 | 16 | 0.50 |  |
| Total | 138 | 19 |  |  |

 For partial *η2* enter the contrast *F*(1, 16) = 250 into my program and you obtain *η2* = .93985 with a confidence interval extending from .85 to .96.

**Orthogonal Contrasts**

 One can construct *k* - 1 orthogonal (independent) contrasts involving *k* means. If I consider *ai* to represent the contrast coefficients applied for one contrast and *bj* those for another, for the contrasts to be orthogonal it must be true that . If you have equal sample sizes, this simplifies to . Consider the following set of contrast coefficients involving groups A, B, C, D, and E and equal sample sizes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | B | C | D | E |
| +.5 | +.5 | −1/3 | −1/3 | −1/3 |
| +1 | −1 | 0 | 0 | 0 |
| 0 | 0 | 1 | −.5 | −.5 |
| 0 | 0 | 0 | +1 | −1 |

 If we computed a *SS* for each of these contrasts and summed those *SS*, the sum would equal the treatment *SS* which would be obtained in an omnibus ANOVA on the *k* groups. This is beautiful, but not necessarily practical. The comparisons you make should be meaningful, whether or not they form an orthogonal set.

**Studentized Range Procedures**

 There is a number of procedures available to make “a posteriori,” “posthoc,” “unplanned” multiple comparisons. When one will compare each group mean with each other group mean, *k*(*k ‑* 1)/2 comparisons, one widely used procedure is the Student-Newman-Keuls procedure. As is generally the case, this procedure adjusts downwards the “per comparison” alpha to keep the alpha familywise at a specified value. It is a “layer” technique, adjusting alpha downwards more when comparing extremely different means than when comparing closer means, thus correcting for the tendency to capitalize on chance by comparing extreme means, yet making it somewhat easier (compared to non-layer techniques) to get significance when comparing less extreme means.

 To conduct a **Student-Newman-Keuls (SNK)** analysis:

 a. Put the means in ascending order of magnitude.

 b. “*r*” is the number of means spanned by a given comparison.

 c. Start with the most extreme means (the lowest vs. the highest), where *r* = *k*.

 d. Compute *q* with this formula: , assuming equal sample sizes and homogeneity of variance. *MSE* is the error mean square from an overall ANOVA on the *k* groups. Do note that the SNK, and multiple comparison tests in general, were developed as an alternative to the omnibus ANOVA. One is not required to do the ANOVA first, and if one does do the ANOVA first it does not need to be significant for one to do the SNK or most other multiple comparison procedures.

 e. If the computed *q* equals or exceeds the tabled critical value for the studentized range statistic, *qr,df*, the two means compared are significantly different, you move to the step g. The *df* is the *df* for *MSE*.

 f. If *q* was not significant, stop and, if you have done an omnibus ANOVA and it was significant, conclude that only the extreme means differ from one another.

 g. If the *q* on the outermost layer is significant, next test the two pairs of means spanning (*k ‑*1) means. Note that *r*, and, thus the critical value of *q*, has decreased. From this point on, underline all pairs not significantly different from one another, and do not test any other pairs whose means are both underlined by the same line.

 h. If there remain any pairs to test, move down to the next layer, etc. etc.

 i. Any means not underlined by the same line are significantly different from one another.

 For our sample problem. which was presented in the previous handout, “One-Way Independent Samples Analysis of Variance,” with alpha at .01:

 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | A | B | C | D |
| Mean | 2 | 3 | 7 | 8 |

 2. *n* = 5

 

 3. A vs D: *r* = 4, *df* = 16, , critical value .01 = 5.19, *p* < .01

 4. *r* = 3, *q.01* = 4.79

 a. A vs C: , *p* < .01

 b. B vs D: , *p* < .01

 5. *r* = 2, *q.01* = 4.13

 a. A vs B: , *p* > .01

 b. B vs C: , *p* < .01

 c. C vs D: , *p* > .01

 6.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Group | A | B |  | C | D |
| Mean | 2 | 3 |  | 7 | 8 |

 What if there are **unequal sample sizes**? One solution is to use the harmonic mean sample size computed across all *k* groups. That is, . Another solution is to compute for each comparison made the harmonic mean sample size of the two groups involved in that comparison, that is, . With the first solution the effect of *n* (bigger *n*, more power) is spread out across groups. With the latter solution comparisons involving groups with larger sample sizes will have more power than those with smaller sample sizes.

 If you have disparate variances, you should compute a *q* that is very similar to the separate variances *t*-test earlier studied. The formula is: , where *t′* is thefamiliar separate variances *t*-test. This procedure is known as the **Games and Howell procedure**.

 When using this unpooled variances *q* one should also adjust the degrees of freedom downwards exactly as done with Satterthwaite’s solution previously discussed.

 Consult our text book for details on how the SNK can have a familywise alpha that is greatly inflated if the omnibus null hypothesis is only party true.

**Relationship Between *q* And Other Test Statistics**

 The studentized range statistic is closely related to *t* and to *F*. If one computes the pooled-across-*k* -groups *t*, as done with Fisher’s LSD, then . If one computes an *F* from a “planned comparison,” then . For example, for the A vs C comparison with our sample problem: 

.

 The contrast coefficients to compare A with C would be −.5, 0, .5, 0.

 The contrast .

, .

**Tukey’s (a) Honestly Significant Difference Test**

 This test is applied in exactly the same way that the Student-Newman-Keuls is, with the exception that *r* is set at *k* for all comparisons. This test is more conservative (less powerful) than the Student-Newman-Keuls.

**Tukey’s (b) Wholly Significant Difference Test**

 This test is a compromise between the Tukey (a) and the Newman-Keuls. For each comparison, the critical *q* is set at the mean of the critical *q* were a Tukey (a) being done and the critical *q* were a Newman-Keuls being done.

**Ryan’s Procedure (REGWQ)**

 This procedure, the Ryan / Einot and Gabriel / Welsch procedure, is based on the *q* statistic, but adjusts the per comparison alpha in such a way (Howell provides details in our text book) that the familywise error rate is maintained at the specified value (unlike with the SNK) but power will be greater than with the Tukey(a). I recommend its use with four or more groups. With three groups the REGWQ is identical to the SNK, and, as you know, I recommend Fisher’s procedure when you have three groups. With four or more groups I recommend the REGWQ, but you can’t do it by hand, you need a computer (SAS and SPSS will do it).

**Other Procedures**

**Dunn’s Test (The Bonferroni *t* )**

 Since the familywise error rate is always less than or equal to the error rate per experiment, , an inequality known as **the Bonferroni inequality**, one can be sure that alpha familywise does not exceed some desired maximum value by using an adjusted alpha per comparison that equals the desired maximum alpha familywise divided by *c*, that is, . In other words, you compute a *t* or an *F* for each desired comparison, usually using an error term pooled across all *k* groups, obtain an exact *p* from the test statistic, and then compare that *p* to the adjusted alpha per comparison. Alternatively, you could multiply the *p* by *c* and compare the resulting “ensemble-adjusted *p*” to the maximum acceptable familywise error rate. Dunn has provided special tables which give critical values of *t* for adjusted per comparison alphas, in case you cannot obtain the exact *p* for a comparison. For example, *t120* = 2.60, alpha familywise = .05, *c* = 6. Is this *t* significant? You might have trouble finding a critical value for *t* at per comparison alpha = .05/6 in a standard *t* table.

**Dunn-Sidak Test**

 Sidak has demonstrated that the familywise error rate for *c* nonorthogonal (nonindependent) comparisons is less than or equal to the familywise error rate for *c* orthogonal comparisons, that is,

. Thus, one can adjust the per comparison alpha by using an adjusted criterion for rejecting the null: Reject the null only if . This procedure, the Dunn-Sidak test, is more powerful than the Dunn test, especially when *c* is large.

**Holm-Bonferroni Test**

 This is a modification of the Bonferroni that is less conservative (more powerful). [See this example of its calculation](https://www.statisticshowto.com/holm-bonferroni-method/). There is also a slightly more powerful Holm-Sidak procedure.

**Scheffé Test**

 To conduct this test one first obtains *F*-ratios testing the desired comparisons. The critical value of the *F* is then adjusted (upwards). The adjusted critical *F* equals (the critical value for the treatment effect from the omnibus ANOVA) times (the treatment degrees of freedom from the omnibus ANOVA). This test is extremely conservative (low power) and is not generally recommended for making pairwise comparisons. It assumes that you will make all possible linear contrasts, not just simple pairwise comparisons. It is considered appropriate for making complex comparisons (such as groups A and B versus C, D, & E; A, B, & C vs D & E; etc., etc.).

**Dunnett’s Test**

 The Dunnett *t* is computed exactly as is the Dunn *t*, but a different table of critical values is employed. It is employed when the only comparisons being made are each treatment group with one control group. It is somewhat more powerful than the Dunn test for such an application.

**Benjamini-Hochberg Test**

 This procedure controls the False Discovery Rate (FDR) rather than familywise error. The FDR is the percentage of significant results that are Type I errors. See [Benjamini-Hochberg](https://www.jstor.org/stable/pdf/2346101.pdf?casa_token=WY9_NFElO80AAAAA:qL_7yMpUOsmYX4NAppracRhnOBg53RyGWU92xMC4iKDsluvbpbeue2E7gt8U7uLofnDp24sVI9CgDD2LZui-3CV-mpy0Ccg4UYLr5Xn9Vs980t3jm-0) and the example in Chapter 12 of Howell’s Statistical Methods for Psychology (2013). [Also see my example](http://core.ecu.edu/psyc/wuenschk/docs30/Benjamin-Hochberg.docx).

**Presenting the Results of Pairwise Contrasts**

 I recommend using a table like that below. You should also give a brief description of the pattern of significant differences between means, but do not mention each contrast that was made in a large set of pairwise contrasts.

 “Teaching method significantly affected test scores, *F*(3, 16) = 86.66, *MSE* = 0.50, *p* < .001, η2 = .94, 95% CI [.82, .94].” Pairwise comparisons were made with Tukey’s HSD procedure, holding familywise error at a maximum of .01. As shown in Table 1, the computer intensive and discussion centered methods were associated with significantly better student performance than that shown by students taught with the actuarial and book only methods. All other comparisons fell short of statistical significance.

*Table 1*

*Mean Quiz Performance By Students Taught With Different Methods*

|  |  |
| --- | --- |
| Method of Instruction | Mean |
| Actuarial | 2.00A |
| Book Only | 3.00A |
| Computer Intensive | 7.00B |
| Discussion Centered | 8.00B |

Note. Means sharing a letter in their superscript are not significantly different at the .01 level according to a Tukey HSD test.

**Familywise Error, Alpha-Adjustment, and the Boogey Men**

 One should keep in mind that the procedures discussed above can very greatly lower power. Unless one can justify taking a much greater risk of a Type II error as the price to pay for keeping the conditional probability of a Type I error unreasonably low, I think it not good practice to employ these techniques. So why do I teach them? Well, because others will expect you (and me) to use these techniques even if we think them unwise. Please read my rant about this: [Familywise Alpha and the Boogey Men](http://core.ecu.edu/psyc/wuenschk/docs30/FamilywiseAlpha.htm).

**Can I Make These Comparisons Even If The ANOVA Is Not Significant?**

 Yes, with the exception of Fisher’s procedure. The other procedures were developed to be used instead of ANOVA, not following a significant ANOVA. You don’t even need to do the ANOVA, and if you do, it does not need to be significant to be permitted to make multiple comparisons. There is much misunderstanding about this. Please read [Pairwise Comparisons](http://core.ecu.edu/psyc/wuenschk/StatHelp/Pairwise.htm).

### [fMRI Gets Slap in the Face with a Dead Fish](http://neuroskeptic.blogspot.com/2009/09/fmri-gets-slap-in-face-with-dead-fish.html) – an example of research where so many comparisons are made that spurious effects are very likely to be found.

 Yes, it is possible for the ANOVA to be significant but none of the pairwise comparisons be significant. See [Tukey\_LSD.docx](http://core.ecu.edu/psyc/wuenschk/docs30/Tukey_LSD.docx).

[Return to Karl’s Stats Lessons Page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm)

Copyright 2022, Karl L. Wuensch - All rights reserved.

1. © Copyright 2010, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)