Comparisons and Type 1 Errors

1. Errors
   1. Type 1 - Rejecting the null when you shouldn’t
   2. Type 2 – Not rejecting the null when you should.
2. Why do you get these errors (mostly Type 1)
   1. Simultaneous testing/comparisons – when you run a bunch of tests after the main analysis, you have the problem of fishing – the probability of rejecting a real null increases
   2. Type 1 and Type 2 are a seesaw – so when you do more tests, you increase type 1 and decrease type 2 – it’s easier to control and examine type 1 with post hoc tests, while type 2 is more about power (chapter 8)
   3. Problem! So many names!
3. Research questions and type 1
   1. Family of tests – set of related research questions
      1. Your hypothesis may be that the groups are different, but then you have sub questions (is group 1 > group 2)
      2. Familywise type 1 error rate – alpha fw – probability of making at least one type 1 error in a family of tests when all nulls are true.
      3. Experimentwise error rate – alpha ew – all possible test you can run probability
   2. Alpha fw = 1 – (1-alpha)c where c is the number of comparisons
      1. FW is always larger than alpha (unless one test)
      2. Per comparison error rate = alpha
      3. SHOW EXAMPLES OF DIFFERENT ALPHAS HERE
      4. Basically the heart of all error corrections is to fix alpha by fixing the cut off score, your p (alpha) used, or your p-found
   3. The typical experiment
      1. Testing primary questions – main test you examine – the main hypotheses – generally you want the most power here, so you do not adjust alpha (so leave per comparison error rate)
      2. Special families of hypotheses – all paired comparisons, all experimental groups to control groups, specific sets of contrasts
         1. You are limited to the number of df in doing these
         2. So you need to correct for the number of tests you run
         3. Sometimes this is broken into separate sets of hypotheses (families: like main effects, and interactions) and you adjust each family separately
      3. Fishing! Exploring data for unexpected relationships – definitely need to control for random data testing
   4. What to do?
      1. Generally, you have some hypothesis (otherwise why did you run this stupid thing?)
      2. Control for the tests you thought you were going to run
      3. Control for the tests that you end up running (if more)
   5. Come up with a good story. – you have to be able to sell why you did what you did.
4. Planned comparisons
   1. There is an argument about orthogonality – most people argue that they need to be orthogonal…but rarely are real hypotheses orthogonal.
   2. Helps you from using the same variance over and over again
   3. Technically you are limited by the number of degrees of freedom, but that won’t work if you are comparing every mean to every other mean.
   4. Must you have a significant overall F value first? Most people say yes.
      1. Doesn’t mean that everyone does it this way – sometimes I just do the planned comparison, but you have to have a good reason.
5. Restricted sets of contrasts – these tests are better with smaller families of hypotheses (i.e. less comparisons over they get overly strict).
   1. RUN ALL THE TESTS ON THE SAME DATA SO YOU CAN SEE THE DIFFERNCES
   2. Bonferroni – through some fancy math, we’ve shown that the family wise error rate is < number of comparisons times alpha. Afw < c(alpha)
      1. Therefore it’s easy to correct for the problem by dividing afw by c to control where alpha = afw / c
      2. EXAMPLE – so if you have 5 tests and an alpha of p<.05 = new alpha = .01, so you can look at the SIG column on SPSS and see if they are less than .01
      3. OR you can use the Bonferroni correction option in SPSS (show example here).
      4. However, Bonferroni will over correct for large number of tests
   3. Sidak Bonferroni – less conservative version of Bonferroni
      1. Alpha = 1 – (1 – afw) 1/c
      2. There are tables in the back of the book for when you need a cut off value, but you can use basically just use the sig column OR SPSS to get these scores.
   4. Dunnett’s Test – usually used when there is a control group compared to all other groups (so one to many, but not many to many)
      1. You do not correct the p-values or alpha, but instead use a stricter cut off score.
      2. You need the total number of groups, the DF for error term, and alpha
      3. Critical difference Dunnett = tdunnett times square root (2 MSs/a / N)
         1. Need appendix here.
      4. RUN IN SPSS.
6. Pairwise comparisons – basically when you want to run everything.
   1. Tukey’s HSD – honestly significant difference (please note that Tukey B is not recommended)
      1. Compare mean difference to Dtukey = q times sqrt (MSs/a / n)
      2. Get q from the studentized range statistic table – you’ll need the number of groups, df error, and alpha fw
      3. Useful to know how to look things up because of ANCOVA.
      4. Unequal sample sizes, you should use the harmonic mean (2 / 1/n + 1/n)
      5. Unequal variance = qa sqrt (1/2 (SE2 + SE2))
   2. Fisher – Hayter – Tukey’s is very popular because it is a good way to hold the type 1 error rate in control. However this decreases power.
      1. FH is a sequential testing approach – we start with one hypothesis and then continue onto the next sets of hypotheses as long as the previous test is significance
      2. First you run the overall F test to see if it significant
      3. Then test all pairwise comparisons using a Dfh getting studentized range table, but groups – 1
      4. Dfh = q sqrt (mss/a / n)
   3. Newman-Keuls – SNK – a walk down procedure
      1. You first compare the groups that have the largest mean difference (rank them based on means)
         1. Dnk = q sqrt(MSs/a / N)
         2. However q is based on the number of groups difference between them (so if there are 5 groups, this first comparison is 5)
      2. Then walk down – so each q value is getting smaller and smaller, so easier to compare means to each other when one apart.
      3. However, this also increases type 1 error rate for those smaller differences in means.
   4. Ryan – Einot – Gabriel – Welsch (REGW) – basically modifications on the SNK to keep the type one error rate down.

|  |  |  |  |
| --- | --- | --- | --- |
| Test | When to Use | Cut off table | Formula |
| Bonferroni | When number of comparisons are small | n/a | a/c |
| Sidak Bonferroni | When number of comparisons are small, but want to be less restrictive – get more power | n/a | 1 – (1 – afw) 1/c |
| Dunnett’s | Control group compared to everyone else | Dunnett’s | Dunnett = tdunnett times square root (2 MSs/a / N) |
| Tukey’s HSD | When you want to compare everything to everything | Studentized range statistic | Dtukey = q times sqrt (MSs/a / n)  Unequal samples  Unequal variances |
| Fisher-Hayter | Step down hypothesis – if first significant, then compare this, then that  Less stringent than the tukey | Studentized range statistic | Dfh = q sqrt (mss/a / n) |
| SNK | Similar to FH, but allows you to have more power with smaller mean differences | Studentized range statistic | Dnk = q sqrt(MSs/a / N) |
| REGW | Similar to FH, but allows you to have more power with smaller mean differences, keeps type 1 error rate down | n/a | Computer |
| Scheffe | Post hoc – allows for all types of comparisons | F table | Fscheffe = (groups – 1) F (dfa, dfs/a)  Take square root for t-value |

1. Recommendations
   1. People really love the tukey test.
   2. FH is good business as well because it increases power of tukey test.
   3. Don’t use SNK, use REGW.
2. Post – Hoc error correction – these tests are for more exploratory data analyses.
   1. These types of tests normally control alpha experiment wise, which makes them more stringent.
   2. Controls for all possible combinations (including complex contrasts) – over what we discussed before.
   3. Scheffe – corrects the cut off score for the F test when doing post hoc comparisons
      1. So you can do this pairwise or drop one group, etc.
      2. Fscheffe = (groups – 1) F (dfa, dfs/a)
      3. Take square root for t-value
      4. Very conservative – loss of power
      5. Not good for just pairwise tests – it’s controlling for more than pairwise

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (I) group | (J) group | Mean Difference (I-J) | Std. Error | Sig. |
|  |
| Tukey HSD | right | left | -1.5000 | .72060 | .113 |
| equal | -3.1000\* | .72060 | .001 |
| left | right | 1.5000 | .72060 | .113 |
| equal | -1.6000 | .72060 | .086 |
| equal | right | 3.1000\* | .72060 | .001 |
| left | 1.6000 | .72060 | .086 |
| Scheffe | right | left | -1.5000 | .72060 | .134 |
| equal | -3.1000\* | .72060 | .001 |
| left | right | 1.5000 | .72060 | .134 |
| equal | -1.6000 | .72060 | .104 |
| equal | right | 3.1000\* | .72060 | .001 |
| left | 1.6000 | .72060 | .104 |
| Bonferroni | right | left | -1.5000 | .72060 | .141 |
| equal | -3.1000\* | .72060 | .001 |
| left | right | 1.5000 | .72060 | .141 |
| equal | -1.6000 | .72060 | .105 |
| equal | right | 3.1000\* | .72060 | .001 |
| left | 1.6000 | .72060 | .105 |
| Sidak | right | left | -1.5000 | .72060 | .134 |
| equal | -3.1000\* | .72060 | .001 |
| left | right | 1.5000 | .72060 | .134 |
| equal | -1.6000 | .72060 | .101 |
| equal | right | 3.1000\* | .72060 | .001 |
| left | 1.6000 | .72060 | .101 |
| Dunnett t (2-sided)b | right | equal | -3.1000\* | .72060 | .000 |
| left | equal | -1.6000 | .72060 | .064 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **pleasant** | | | | | |
|  | group | N | Subset | | |
|  | 1 | 2 | 3 |
| Student-Newman-Keulsa,b | right | 10 | 4.3000 |  |  |
| left | 10 |  | 5.8000 |  |
| equal | 10 |  |  | 7.4000 |
| Sig. |  | 1.000 | 1.000 | 1.000 |
| Tukey HSDa,b | right | 10 | 4.3000 |  |  |
| left | 10 | 5.8000 | 5.8000 |  |
| equal | 10 |  | 7.4000 |  |
| Sig. |  | .113 | .086 |  |
| Scheffea,b | right | 10 | 4.3000 |  |  |
| left | 10 | 5.8000 | 5.8000 |  |
| equal | 10 |  | 7.4000 |  |
| Sig. |  | .134 | .104 |  |
| Ryan-Einot-Gabriel-Welsch Fb | right | 10 | 4.3000 |  |  |
| left | 10 |  | 5.8000 |  |
| equal | 10 |  |  | 7.4000 |
| Sig. |  | 1.000 | 1.000 | 1.000 |
| Means for groups in homogeneous subsets are displayed.  Based on observed means.  The error term is Mean Square(Error) = 2.596. | | | | | |
| a. Uses Harmonic Mean Sample Size = 10.000. | | | | | |
| b. Alpha = | | | | | |