

Pairwise Comparisons

- ANOVA for multiple condition designs
- Pairwise comparisons and RH Testing
- Alpha inflation & Correction
- LSD & HSD procedures
- Alpha estimation reconsidered...

H0: Tested by k-grp ANOVA

- Regardless of the number of IV conditions, the H0: tested using ANOVA (F-test) is ...
 - “all the IV conditions represent populations that have the same mean on the DV”
- When you have only 2 IV conditions, the F-test of this H0: is sufficient
 - there are only three possible outcomes ...
 $T=C$ $T<C$ $T>C$ & only one matches the RH
- With multiple IV conditions, the H0: is still that the IV conditions have the same mean DV...
 - $T_1 = T_2 = C$ but there are many possible patterns
 - Only one pattern matches the Rh:

Omnibus F vs. Pairwise Comparisons

- Omnibus F
 - overall test of whether there are any mean DV differences among the multiple IV conditions
 - Tests H0: that all the means are equal
- Pairwise Comparisons
 - specific tests of whether or not each pair of IV conditions has a mean difference on the DV
- How many Pairwise comparisons ??
 - Formula, with $k = \#$ IV conditions
pairwise comparisons = $[k * (k-1)] / 2$
 - or just remember a few of them that are common
 - 3 groups = 3 pairwise comparisons
 - 4 groups = 6 pairwise comparisons
 - 5 groups = 10 pairwise comparisons

How many Pairwise comparisons – revisited !!

There are two questions, often with different answers...

1. How many pairwise comparisons can be computed for this research design?
 - Answer $\rightarrow [k * (k-1)] / 2$
 - But remember \rightarrow if the design has only 2 conditions the Omnibus-F is sufficient; no pairwise comparisons needed

2. How many pairwise comparisons are needed to test the RH:?
 - Must look carefully at the RH: to decide how many comparisons are needed
 - E.g., The ShortTx will outperform the control, but not do as well as the LongTx
 - This requires only 2 comparisons
ShortTx vs. control ShortTx vs. LongTx

Process of statistical analysis for multiple IV conditions designs

- Perform the Omnibus-F
 - test of H0: that all IV conds have the same mean
 - if you retain H0: -- quit
- Compute all pairwise mean differences
- Compute the minimum pairwise mean diff
 - formulas are in the Stat Manual -- ain't no biggie!
- Compare each pairwise mean diff with minimum mean diff
 - if mean diff > min mean diff then that pair of IV conditions have significantly different means
 - be sure to check if the "significant mean difference" is in the hypothesized direction !!!

Example analysis of a multiple IV conditions design

Tx1	Tx2	Cx
50	40	35

For this design, $F(2,27)=6.54$, $p = .005$ was obtained.

We would then compute the pairwise mean differences.

Tx1 vs. Tx2 10 Tx1 vs. C 15 Tx2 vs. C 5

Say for this analysis the minimum mean difference is 7

Determine which pairs have significantly different means

Tx1 vs. Tx2	Tx1 vs. C	Tx2 vs. C
Sig Diff	Sig Diff	Not Diff

What to do when you have a RH:
 The RH: was, "The treatments will be equivalent to each other, and both will lead to higher scores than the control."

Determine the pairwise comparisons, how the RH applied to each ...

Tx1 = Tx2 Tx1 > C Tx2 > C

Tx1	Tx2	Cx
85	70	55

For this design, $F(2,42)=4.54$, $p = .012$ was obtained.

Compute the pairwise mean differences.

Tx1 vs. Tx2 ____ Tx1 vs. C ____ Tx2 vs. C ____

Cont. Compute the pairwise mean differences.

Tx1 vs. Tx2 15 Tx1 vs. C 30 Tx2 vs. C 15

For this analysis the minimum mean difference is 18

Determine which pairs have significantly different means

Tx1 vs. Tx2 No Diff !	Tx1 vs. C Sig Diff !!	Tx2 vs. C No Diff !!
--------------------------	--------------------------	-------------------------

Determine what part(s) of the RH were supported by the pairwise comparisons ...

RH:	Tx1 = Tx2	Tx1 > C	Tx2 > C
results	Tx1 = Tx2	Tx1 > C	Tx2 = C
well ?	supported	supported	not supported

We would conclude that the RH: was partially supported !

Your turn !! The RH: was, "Treatment 1 leads to the best performance, but Treatment 2 doesn't help at all."

What predictions does the RH make ?

Tx1 Tx2 Tx1 C Tx2 C

Tx1	Tx2	Cx
15	9	11

For this design, $F(2,42)=5.14$, $p = .010$ was obtained. The minimum mean difference is 3

Compute the pairwise mean differences and determine which are significantly different.

Tx1 vs. Tx2 ____ Tx1 vs. C ____ Tx2 vs. C ____

Your Conclusions ?

“The Problem” with making multiple pairwise comparisons -- “Alpha Inflation”

- As you know, whenever we reject H_0 , there is a chance of committing a Type I error (thinking there is a mean difference when there really isn't one in the population)
 - The chance of a Type I error = the p-value
 - If we reject H_0 : because $p < .05$, then there's about a 5% chance we have made a Type I error
- When we make multiple pairwise comparisons, the Type I error rate for each is about 5%, but that error rate “accumulates” across each comparison -- called “alpha inflation”
 - So, if we have 3 IV conditions and make 3 the pairwise comparisons possible, we have about ...
 $3 * .05 = .15$ or about a 15% chance of making at least one Type I error

Alpha Inflation

- Increasing chance of making a Type I error as more pairwise comparisons are conducted

Alpha correction

- adjusting the set of tests of pairwise differences to “correct for” alpha inflation
- so that the overall chance of committing a Type I error is held at 5%, no matter how many pairwise comparisons are made

Here are the pairwise comparisons most commonly used -- but there are several others

Fisher's LSD (least significance difference)

- no Omnibus-F – do a separate F- or t-test for each pair of conditions
- no alpha correction -- use $\alpha = .05$ for each comparison

Fisher's “Protected tests”

- “protected” by the omnibus-F -- only perform the pairwise comparisons IF there is an overall significant difference
- no alpha correction -- uses $\alpha = .05$ for each comparison

Scheffe's test

- emphasized importance of correction for Alpha Inflation
- pointed out there are "complex comparisons" as well as "pairwise" comparisons that might be examined
- E.g., for 3 conditions you have...
 - 3 simple comparisons Tx1 v. Tx2 Tx1 v. C Tx2 v. C
 - 3 complex comparisons – by combining conditions and comparing their average mean to the mean of other condition
Tx1+Tx2 v. C Tx1+C v. Tx2 Tx2+C v. Tx1
- developed formulas to control alpha for the total number of comparisons (simple and complex) available for the number of IV conditions

Bonferroni (Dunn's) correction

- pointed out that we don't always look at all possible comparisons
- developed a formula to control alpha inflation by "correcting for" the actual number of comparisons that are conducted
- the p-value for each comparison is set = $.05 / \# \text{comparisons}$

Tukey's HSD (honestly significant difference)

- pointed out the most common analysis was to look at all the simple comparisons – most RH: are directly tested this way
- developed a formula to control alpha inflation by "correcting for" the number of pairwise comparisons available for the number of IV conditions

Dunnnett's test

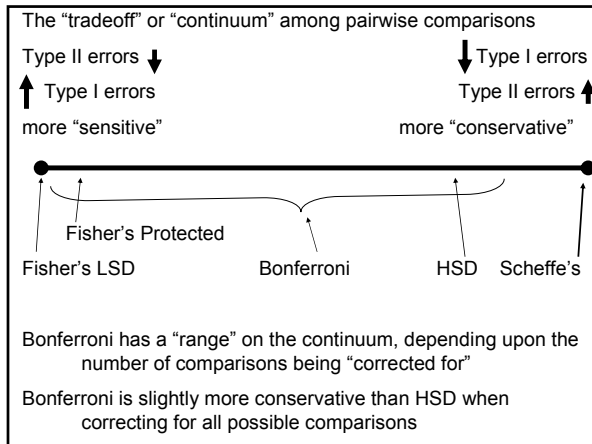
- used to compare one IV condition to all the others
- alpha correction considers non-independence of comparisons

Two other techniques that were commonly used but which have "fallen out of favor" (largely because they are more complicated than others that work better)

Newman-Keuls and Duncan's tests

- used for all possible pairwise comparisons
- called "layered tests" since they apply different criterion for a significant difference to means that are adjacent than those that are separated by a single mean, than by two mean, etc.
- Tx1-Tx3 have adjacent means, so do Tx3-Tx2 and Tx2-C. Tx1-Tx2 and Tx3-C are separated by one mean, and would require a larger difference to be significant. Tx1-C would require an even larger difference to be significant.

Tx1	Tx3	Tx2	C
10	12	15	16



So, now that we know about all these different types of pairwise comparisons, which is the "right one" ???

Consider that each test has a build-in BIAS ...

- "sensitive tests" (e.g., Fisher's Protected Test & LSD)
 - have smaller mmd values (for a given n & MS_{Error})
 - are more likely to reject H₀: (more power - less demanding)
 - are more likely to make a Type I error (false alarm)
 - are less likely to make a Type II error (miss a "real" effect)
- "conservative tests" (e.g., Scheffe' & HSD)
 - have larger mmd values (for a given n & MS_{Error})
 - are less likely reject H₀: (less power - more demanding)
 - are less likely to make a Type I error (false alarm)
 - are more likely to make a Type II error (miss a "real effect")

Computing Pairwise Comparisons by Hand

The two most commonly used techniques (LSD and HSD) provide formulas that are used to compute a "minimum mean difference" which is compared with the pairwise differences among the IV conditions to determine which are "significantly different".

$$d_{LSD} = \frac{t * \sqrt{2 * MS_{Error}}}{\sqrt{n}}$$

t is looked-up from the t-table based on $\alpha=.05$ and the $df = df_{Error}$ from the full model

$$d_{HSD} = \frac{q * \sqrt{MS_{Error}}}{\sqrt{n}}$$

q is the "Studentized Range Statistic" -- based on $\alpha=.05$, $df = df_{Error}$ from the full model, and the # of IV conditions

For a given analysis LSD will have a smaller minimum mean difference than will HSD.

Some common questions about applying the lsd/hsd formulas...

What is "n" for a within-groups design ?

Since "n" represents the number of data points that form each IV condition mean (in index of sample size/power), $n = N$ (since each participant provides data in each IV condition)

What is "n" if there is "unequal-n" ?

Use the "average n" from the different conditions. This is only likely with BG designs -- very rarely is there unequal n in WG designs, and most computations won't handle those data.

Earlier in the lecture we discussed a General procedure for pairwise comparisons..

- Compute the obtained mean difference for all pairs of IV conditions
- Compute the minimum mean difference (MMD e.g., 6.1 for LSD)
- Compare the obtained and minimum difference for each pair
 - If |obtained mean difference| > minimum mean difference then conclude those means are significantly different

	Cx	Tx1	Tx2		
Cx	20.3			no mean dif	Cx = Tx1
Tx1	24.6	4.3		mean dif	Tx2 > Cx
Tx2	32.1	11.8	7.5	mean dif	Tx2 > Tx1

Remember to check the DIRECTION of mean differences when evaluating whether RH: is supported or not !!!

But, still you ask, which post test is the "right one" ???

Rather than "decide between" the different types of bias, I will ask you to learn to "combine" the results from more conservative and more sensitive designs.

If we apply both LSD and HSD to a set of pairwise comparisons, any one of 3 outcomes is possible for each comparison

- we might retain H0: using both LSD & HSD
 - if this happens, we are "confident" about retaining H0:, because we did so based not only on the more conservative HSD, but also based on the more sensitive LSD
- we might reject H0: using both LSD & HSD
 - if this happens we are "confident" about rejecting H0: because we did so based not only on the more sensitive LSD, but also based on the more conservative HSD
- we might reject H0: using LSD & retain H0: using HSD
 - if this happens we are confident about neither conclusion

Here's an example...

A study was run to compare 3 treatments to each other and to a no-treatment control. The resulting means and mean differences were found.

	M	Tx1	Tx2	Tx3
Based on LSD mmd = 3.9	Tx1	12.3		
Based on HSD mmd = 6.7	Tx2	14.6	2.3	
	Tx3	18.8	6.5*	2.2
	Cx	22.9	10.6**	8.3** 4.1*

Conclusions:

- confident that Cx > Tx1 & Cx > Tx2 -- ~~H0~~: lsd & hsd
- confident that Tx2 = Tx1 & Tx3 = Tx2 -- H0: w/ both lsd & hsd
- not confident about Tx3 - Tx1 or Cx - Tx3 -- lsd & hsd differed
 - next study should concentrate on these comparisons

Applying Bonferroni

Unlike LSD and HSD, Bonferroni is based on computing a "regular" t/F-test, but making the "significance" decision based on a p-value that is adjusted to take into account the number of comparisons being conducted.

Imagine a 4-condition study - three Tx conditions and a Cx. The RH: is that each of the TX conditions will lead to a higher DV than the Cx. Even though there are six possible pairwise comparisons, only three are required to test the researcher's hypothesis. To maintain an experiment-wise Type I error rate of .05, each comparison will be evaluated using a comparison-wise p-value computed as

If we wanted to hold out experiment-wise Type I rate to 5%, we would perform each comparison using...

$$\alpha_E / \# \text{ comparisons} = \alpha_C \quad .05 / 3 = .0167$$

We can also calculate the experiment-wise for a set of comps...

With p=.05 for each of 4 comps our experiment-wise Type I error rate would be ... $\alpha_E = \# \text{ comparisons} * \alpha_C = 4 * .05 = 20\%$

A few moments of reflection upon "Experiment-wise error rates" the most commonly used α_E estimation formula is ...

$$\alpha_E = \alpha_C * \# \text{ comparisons}$$

e.g., $.05 * 6 = .30$, or a 30% chance of making at least 1 Type I error among the 6 pairwise comparisons

But, what if the results were as follows (LSDmmd = 7.0)

	Tx1	Tx2	Tx3	C
Tx1	12.6			
Tx2	14.4	1.8		
Tx3	16.4	3.8	2.0	
C	22.2	9.6*	7.8*	5.8

We only rejected H0: for 2 of the 6 pairwise comparisons. We can't have made a Type I error for the other 4 -- we retained the H0: !!!

At most our α_E is 10% -- 5% for each of 2 rejected H0:s

Here's another look at the same issue...
imagine we do the same 6 comparisons using t-tests, so we get exact p-values for each analysis...
Tx2-Tx1 p. = .43 Tx3-Tx1 p. = .26 Tx3-Tx2 p. = .39
C-Tx1 p. = .005* C-Tx2 p. = .01* C-Tx3 p. = .14
We would reject H0: for two of the pairwise comparisons ...
We could calculate α_E as $\Sigma p = .005 + .01 = .015$

What is our α_E for this set of comparisons? Is it ...
.05 * 6 = .30, *a priori* α_E – accept a 5% risk on each of the possible pairwise comparisons ???
.05 * 2 = .10, post hoc α_E – accept a 5% risk for each rejected H0: ???
.005 + .01 = .015, exact post hoc α_E – actual risk accumulated across rejected H0:s ???

Notice that these α_E values vary dramatically !!!
