Analyses of K-Group Designs : Omnibus F & Follow-up Analyses

- ANOVA for multiple condition designs
- Pairwise comparisons, alpha inflation & correction
- Alpha estimation reconsidered…
- Analytic Comparisons: Simple, Complex & Trend Analyses
- Effect sizes for k-group designs

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 → [k * (k-1)] / 2
   - But remember → if the design has only 2 conditions the Omnibus-F is sufficient; no pariwise comparsons 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

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

"The Problem" with making multiple pairwise comparisons -- "Alpha Inflation"

- As you know, whenever we reject H0:, 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 H0: 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 α = .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 α = .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 / #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

Dunnett's test

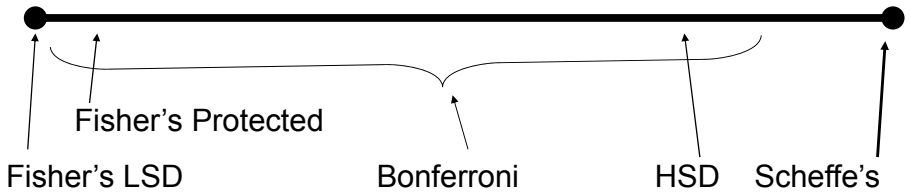
- used to compare one IV condition to all the others
- alpha correction considers non-independence of comparisons

The "tradeoff" or "continuum" among pairwise comparisons

Type II errors ↓ ↓ Type I errors

↑ Type I errors Type II errors ↑

more "sensitive" more "conservative"



Bonferroni has a "range" on the continuum, depending upon the number of comparisons being "corrected for"

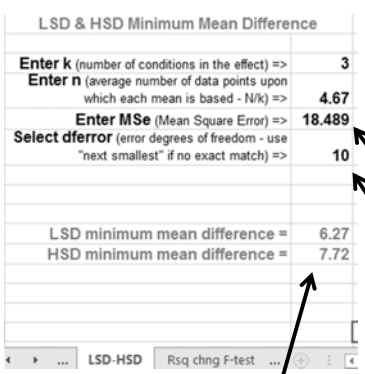
Bonferroni is slightly more conservative than HSD when correcting for all possible comparisons

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 & MSerror)
  - 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 & MSerror)
  - 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")

Using the LSD- HSD tab of xls Computator to find the mmd for BG designs



Descriptives

number of fish at store

	N	Mean	Std. Deviation
chain store	7	17.43	4.117
privately owned	3	19.33	4.041
coop	4	35.50	4.796
Total	14	23.00	9.140

k = # conditions

n = N / k = 14 / 3 = 4.67

Note: always use decimal part of n

Use the drop-down menu to set dferror. Round down!

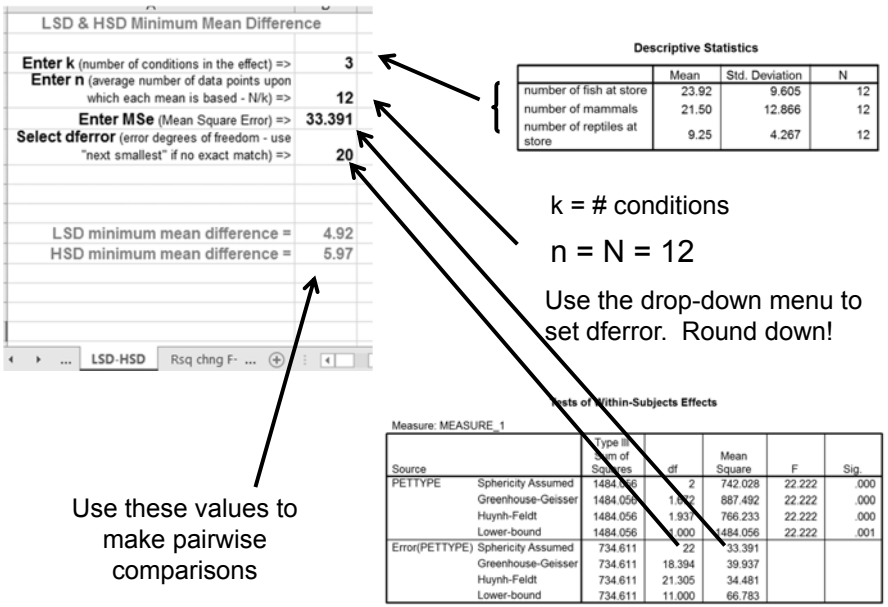
Use these values to make pairwise comparisons

ANOVA

number of fish at store

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	882.619	2	441.310	23.869	.000
Within Groups	203.381	11	18.489		
Total	1086.000	13			

Using the LSD- HSD tab of xls Computator to find the mmd for WG designs



Some common questions about applying the lsd/hsd formulas…

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

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

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

- "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)

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 <u>retain H0:</u> 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 <u>reject H0:</u> 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 <u>reject H0: using LSD</u> & <u>retain H0: using HSD</u>
  - if this happens we are confident about neither conclusion

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…

α_E / # comparisons = α_C .05 / 3 = .0167

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

With p=.05 for each of 4 coms our experiment-wise Type I error rate would be … α_E = # comparisons * α_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 Σp = .005 + .01 = .015

What is our α_E for this set of comparions? 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 αE – actual risk accumulated across rejected H0:s ???

Notice that these αE values vary dramatically !!!

Analytic Comparisons -- techniques to make specific comparisons among condition means. There are two types…

Simple Analytic Comparisons -- to compare the means of two IV conditions at a time

Rules for assigning weights:

1. Assign weight of "0" to any condition not involved in RH
2. Assign weights to reflect comparison of interest
3. Weights must add up to zero

			Tx2	Tx1	C
			40	10	40
E.g. #1	RH:	Tx1 < C (is 10 < 40 ?)	0	-1	1
E.g. #2	RH:	Tx2 < Tx1 (is 40 < 10?)	-1	1	0

How do Simple Analytic Comparisons & Pairwise Comparisons differ?

- Usually there are only k-1 analytic comparisons (1 for each df)

So, what happens with these weights?

The formula $SS_{comp} = \frac{n(\Sigma w*mean)^2}{\Sigma w^2}$ & $F = SS_{comp}/MS_{errror}$

The important part is the $\Sigma w*mean$ → multiply each mean by its weight and add the weighted means together

- if a group is weighted 0, that group is "left out" of the SS_{comp}
- if the groups in the analysis have the same means $SS_{comp} = 0$
- the more different the means of the groups in the analysis the larger SS_{comp} will be

Tx2	Tx1	C	
40	10	40	
-1	0	1	$\Sigma w*mean = (-1*40) + (0 * 10) + (1 * 40) = 0$
-1	1	0	$\Sigma w*mean = (-1*40) + (1 * 10) + (0 * 40) = -30$

Complex Analytic Comparisons -- To compare two "groups" of IV conditions, where a "group" is sometimes one condition and sometimes 2 or more conditions that are "combined" and represented as their average mean.

Rules for assigning weights:

1. Assign weight of "0" to any condition not involved in RH
2. Assign weights to reflect group comparison of interest
3. Weights must add up to zero

		Tx2	Tx1	C
		40	10	40
RH:	Control higher than average of Tx conditions (40 > 25?)	1	1	-2

Careful !!! Notice the difference between the proper interpretation of this complex comparison and of the set of simple comparisons below.

RH:	Control is poorer than	(is 40 < 40)	1	0	-1
	both of Tx conditions	(is 10 < 40)	0	1	-1

Notice the complex & set of simple comparisons have different interpretations!

Criticism of Complex Analytical Comparisons

- Complex comparisons are seldom useful for testing research hypotheses !! (Most RH are addressed by the proper set of simple comparisons!)
- Complex comparisons require assumptions about the comparability of IV conditions (i.e., those combined into a "group") that should be treated as research hypotheses !!
- Why would you run two (or more) separate IV conditions, being careful to following their different operational definitions, only to "collapse" them together in a complex comparison
- Complex comparisons are often misinterpreted as if it were a set of simple comparisons

Orthogonal and nonorthogonal sets of analytics

Orthogonal means independent or unrelated -- the idea of a set of orthogonal analytic comparisons is that each would provide statistically independent information.

The way to determine if a pair of comparisons is orthogonal is to sum the products of the corresponding weights. If that sum is zero, then the pair of comparisons is orthogonal.

Non-orthogonal Pair				Orthogonal Pair		
Tx1	Tx2	C		Tx1	Tx2	C
1	0	-1		1	1	-2
0	1	-1		1	-1	0
0	0	1	< products >	1	-1	0
Sum = 1				Sum = 0		

For a "set" of comparisons to be orthogonal, each pair must be !

Advantages and Disadvantages of Orthogonal comparison sets

Advantages

- each comparison gives statistically independent information, so the orthogonal set gives the most information possible for that number of comparisons
- it is a mathematically elegant way of expressing the variation among the IV conditions -- SS_{IV} is partitioned among the comps

Disadvantages

- "separate research questions" often doesn't translate into "statistically orthogonal comparisons" (e.g., 1 -1 0 & 1 0 -1)
- can only have # orthogonal comparisons = df_{IV}
- the comparisons included in an orthogonal set rarely address the set of research hypotheses one has (e.g., sets of orthogonal analyses usually include one or more complex comparisons)

Trend Analyses -- the ***shape*** of the IV-DV relationship

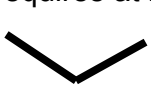
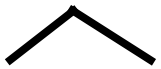
Trend analyses can be applied whenever the IV is quantitative.

• There are three basic types of trend (w/ two versions of each)

Linear Trends

positive  negative 

Quadratic Trends (requires at least 3 IV conditions)

U-shaped  inverted-U-shaped 

Cubic Trends (requires at least 4 IV conditions)

Note: Trend analyses are computed same as analytics -- using weights (coefficients) from "table" (only for =n & =spacing)

Note: As set of trend analyses are orthogonal – separate info @

Not only is it important to distinguish between the two different types of each basic trend, but it is important to identify shapes that are combinations of trends (and the different kinds)

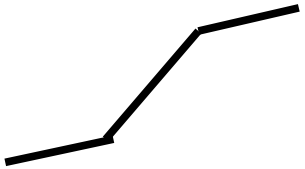
Here are two different kinds of "linear + quadratic" that would have very different interpretations

+ linear & U-shaped quadratic 

("accelerating returns" curve)

+ linear & inverted U-shape quad 

("diminishing returns" curve)

Here is a common combination of + linear & cubic 

("learning curve")

"How to mess-up interpreting analytic comparisons"

Simple Comparisons:

-- ignore the direction of the simple difference (remember you must have a difference in the ***correct direction***)

Complex Comparisons:

-- ignore direction of the difference (remember you must have a difference in the ***correct direction***)

-- misinterpret complex comparison as if it were a set of simple comparisons

Trend Analyses:

-- ignore specific pattern of the trend (remember you must have a shape in the ***correct direction or pattern***)

-- misinterpret trend as if it were a set of simple comps

-- ignore combinations of trend (e.g., the RH of a linear trend "really means" that there is a significant linear trend, and no significant quadratic or cubic trend)

-- perform trend analyses on non-quantitative IV conditions

Effect Sizes for the k-BG or k-WG → Omnibus F

The effect size formula must take into account both the size of the sample (represented by dferror) and the size of the design (represented by the dfeffect).

$$r = \sqrt{(df_{effect} * F) / (F + df_{error})}$$

The effect size estimate for a k-group design can only be compared to effect sizes from other studies with designs having exactly the same set of conditions.

There is no "d" for k-group designs – you can't reasonably take the "difference" among more than 2 groups.

Effect Sizes for k-BG → Pairwise Comparisons

You won't have F-values for the pairwise comparisons, so we will use a 2-step computation

First: $d = (M1 - M2) / \sqrt{MSerror}$

Second: $r = \sqrt{\frac{d^2}{d^2 + 4}}$

This is an "approximation formula"

Pairwise effect size estimates can be compared with effect sizes from other studies with designs having these 2 conditions (no matter what other differing conditions are in the two designs)

Effect Sizes for k-WG → Pairwise Comparisons

You won't have F-values for the pairwise comparisons, so we will use a 2-step computation

First: $d = (M1 - M2) / \sqrt{(MSerror * 2)}$

Second: $d_w = d * 2$

Third: $r = \sqrt{\frac{d_w^2}{d_w^2 + 4}}$

This is an "approximation formula"

Pairwise effect size estimates can be compared with effect sizes from other studies with designs having these 2 conditions (no matter what other differing conditions are in the two designs).

Effect Sizes for the k-BG or k-WG → Analytic Comps (Simple, Complex & Trend Analyses)

Since all three kinds of analytic comparisons always have $df_{effect} = 1$, we can use the same effect size formula for them all (the same one we used for 2-group designs).

$$r = \sqrt{F / (F + df_{error})} \qquad \text{or} \qquad r = \sqrt{t^2 / (t^2 + df)}$$

Effects size estimates from simple & complex comparisons can be compared with effect sizes from other studies with designs having the same set of conditions (no matter what other differing conditions are in the two designs).

Effect size estimates from trend analyses can only be compared with effect sizes from other studies with designs having the same set of conditions.