Multiple-Group Research Designs & ANOVA

- Limitations of 2-group designs
- "Kinds" of Treatment & Control conditions
- Kinds of Causal Hypotheses
- k-group ANOVA & Pairwise Comparisons
- Alpha Inflation & Alpha Correction

Limitations of 2-cond Designs

- 2-cond designs work well to conduct basic treatment evaluations
  - they allow us to investigate whether or not a specific treatment has "an effect"
  - usually by comparing it to a "no treatment" control
  - e.g., does a new treatment program work to help socially anxious clients (compared to no treatment)?
- However as research questions/hypotheses become more sophisticated and specific, we often require designs that have multiple IV conditions

"Kinds" of Conditions to Include in Research Designs
Tx Conditions

- Ways treatment conditions differ
  - amount of treatment
    - receiving therapy once vs. twice each week
    - getting 0, 1, 5 or 10 practice trials before testing
  - kind of treatment
    - receiving Cognitive vs. Gestalt clinical therapy
    - whether or not there is feedback on practice trials
  - combinations of treatment components
    - receiving both "talk" therapy vs. "combined drug & talk" therapy
    - receiving "10 practices without feedback" vs. "2 practices with feedback"

The "Secret" is to be sure the selection of conditions matches the research hypotheses you started with !!!

Different Kinds of “Control” Conditions

- “No Treatment” control
  - – Asks if the Tx works “better than nothing”
- “Standard Tx” control
  - – Asks if the Tx works “better than usual”
- “Best Practice” Control
  - – Asks if the Tx works “better than the best known”
- “Pseudo Tx” Control
  - – Asks if TX works “without a specific component”

The “Secret” is to be sure the selection of conditions matches the research hypotheses you started with !!!

An important point to remember...

Not every design needs a “no treatment control” group !!!!

Remember, a design needs to provide “an comparison of appropriate conditions” to provide a test of the research hypothesis !!!

What would be the appropriate “control group” to answer each of the following ?

My new Tx works better than the currently used behavioral therapy technique	Group receiving the behavioral therapy.
My new Tx works better than “no treatment”	Group receiving no treatment.
My new Tx works because of the combo of the usual and new behavioral components	Pseudo-Tx group
My new TX works better when given by a Ph.D. than by a Masters-level clinician	Groups receiving the Tx from the two types of clinicians.

The “Secret” is to be sure the selection of conditions matches the research hypotheses you started with !!!

Causal Hypotheses for Multiple Condition Designs

Sometimes there is more than one component to a “treatment,” and so, there are multiple differences between the IV conditions. When this happens, you must distinguish..

Causal Hypotheses about “treatment comparisons”

-- hypothesis that the difference between the DV means of the IV conditions is caused by the ***combination*** of treatment component differences

Causal Hypotheses about “identification of causal elements”

-- hypothesis that the difference between the DV means of the IV conditions is caused by a specific (out of two or more) treatment component difference (good use of pseudo-Tx controls)

The “Secret” is to be sure the condition comparison matches the specific type of causal research hypotheses !!!!

For example… I created a new treatment for social anxiety that uses a combination of group therapy (requiring clients to get used to talking with other folks) and cognitive self-appraisal (getting clients to notice when they are and are not socially anxious). Volunteer participants were randomly assigned to the treatment condition or a no-treatment control. I personally conducted all the treatment conditions to assure treatment integrity. Here are my results using a DV that measures "social context tolerance" (larger scores are better).

F(1,38) = 9.28, p = .001, Mse = 17.3

Group therapy & self-appraisal	Cx
52	25

Which of the following statements will these results support?

"Here is evidence that the combination of group therapy & cognitive self-appraisal increases social context tolerance." ???

Yep -- treatment comparison causal statement

" You can see that the treatment works because of the cognitive self-appraisal; the group therapy doesn't really contribute anything."

Nope -- identification of causal element statement & we can't separate the role of group therapy & self-appraisal

Same story... I created a new treatment for social anxiety that uses a combination of group therapy (requiring clients to get used to talking with other folks) and cognitive self-appraisal (getting clients to notice when they are and are not socially anxious). Volunteer participants were randomly assigned to the treatment condition or a no-treatment control. I personally conducted all the treatment conditions to assure treatment integrity.

What conditions would we need to add to the design to directly test the second of these causal hypotheses...

The treatment works because of the cognitive self-appraisal; the group therapy doesn't really contribute anything."

Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control

Let's keep going …

Here's the design we decided upon. Assuming the results from the earlier study replicate, we'd expect to get the means shown below.

Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control
52	25	52	25

What means for the other two conditions would provide support for the RH:

The treatment works because of the cognitive self-appraisal; the group therapy doesn't really contribute anything."

Another example… The new on-line homework I've been using provides immediate feedback for a set of 20 problems. To assess this new homework I compared it with the online homework I used last semester which 10 problems but no feedback. I randomly assigned who received which homework and made sure each did the correct type. The DV was the % score on a quiz given the day the homework was due. Here are the results ...

F(1,42) = 6.54, p = .001, Mse = 11.12

Old Hw	New Hw
72	91

Which of the following statements will these results support?

"Here is evidence that the new homework is more effective because it provides immediate feedback!"

Nope -- identification of causal element statement -- with this design we can't separate the role of feedback and number of problems

"The new homework seems to produce better learning!"

Yep -- treatment comparison causal statement

Same story... The new on-line homework I've been using provides immediate feedback for a set of 20 problems. To assess this new homework I compared it with the online homework I used last semester which 10 problems but no feedback. I randomly assigned who received which homework and made sure each did the correct type.

What conditions would we need to add to the design to directly test the second of these causal hypotheses...

"Here is evidence that the new homework is more effective because it provides immediate feedback!"

Hint: Start by asking what are the "differences" between the "new" and "old" homeworks -- what are the "components" of each treatment???

"New Hw" 20 problems w/ feedback	20 problems w/o feedback	10 problems w/ feedback	"Old Hw" 10 problems w/o feedback

Let's keep going …

Here's the design we decided upon. Assuming the results from the earlier study replicate, we'd expect to get the means shown below.

"New Hw" 20 problems w/ feedback	20 problems w/o feedback	10 problems w/ feedback	"Old Hw" 10 problems w/o feedback
91	75	89	72

What means for the other two conditions would provide support for the RH:

"Here is evidence that the new homework is more effective because it provides immediate feedback!"

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

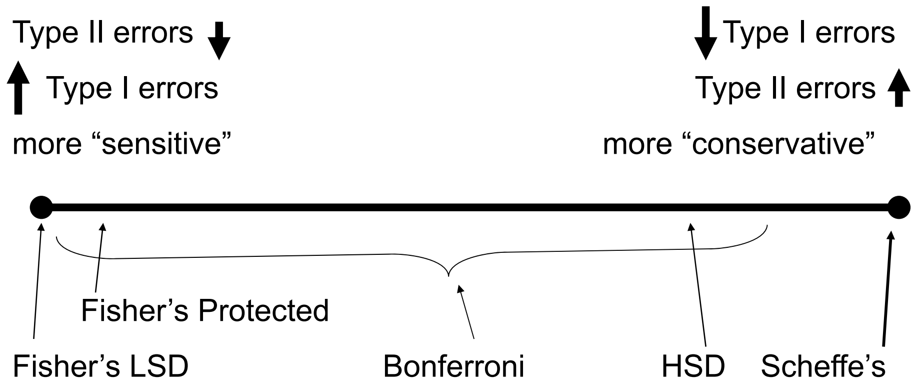
Tukey’s HSD (honestly significant difference)

- 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

more “sensitive”

↓ Type I errors

Type II errors ↑

more “conservative”

Fisher’s Protected

Fisher’s LSD Bonferroni HSD Scheffe’s

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

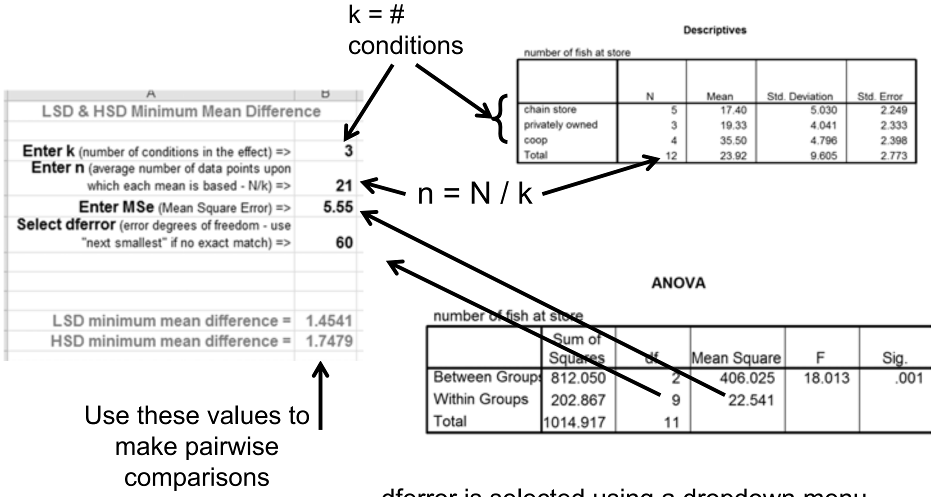
Bonferroni is slightly more conservative than HSD when correcting for all possible comparisons

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 XLS Computator to find the mmd for BG designs



A	B
LSD & HSD Minimum Mean Difference	
Enter k (number of conditions in the effect) =>	**3**
Enter n (average number of data points upon which each mean is based - N/k) =>	**21**
Enter MSe (Mean Square Error) =>	**5.55**
Select dferror (error degrees of freedom - use "next smallest" if no exact match) =>	**60**
LSD minimum mean difference =	1.4541
HSD minimum mean difference =	1.7479

Descriptives

number of fish at store

	N	Mean	Std. Deviation	Std. Error
chain store	5	17.40	5.030	2.249
privately owned	3	19.33	4.041	2.333
coop	4	35.50	4.796	2.398
Total	12	23.92	9.605	2.773

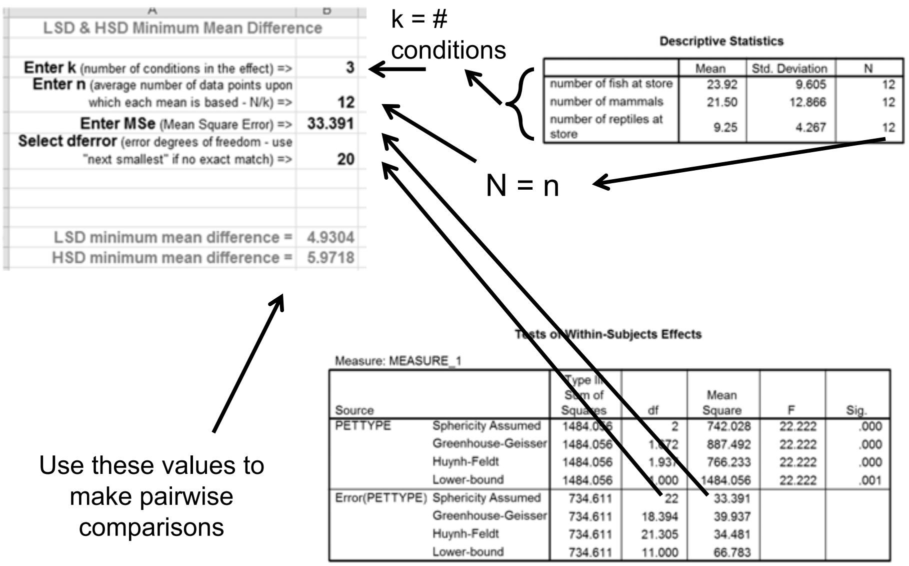
ANOVA

number of fish at store

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	812.050	2	406.025	18.013	.001
Within Groups	202.867	9	22.541		
Total	1014.917	11			

dferror is selected using a dropdown menu – use smaller value to be conservative

Using the xls Computator to find mmd for WG designs



A	B
LSD & HSD Minimum Mean Difference	
Enter k (number of conditions in the effect) =>	**3**
Enter n (average number of data points upon which each mean is based - N/k) =>	**12**
Enter MSe (Mean Square Error) =>	**33.391**
Select dferror (error degrees of freedom - use "next smallest" if no exact match) =>	**20**
LSD minimum mean difference =	4.9304
HSD minimum mean difference =	5.9718

Descriptive Statistics

	Mean	Std. Deviation	N
number of fish at store	23.92	9.605	12
number of mammals	21.50	12.866	12
number of reptiles at store	9.25	4.267	12

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
PETTYPE	Sphericity Assumed	1484.056	2	742.028	22.222	.000
	Greenhouse-Geisser	1484.056	1.672	887.492	22.222	.000
	Huynh-Feldt	1484.056	1.937	766.233	22.222	.000
	Lower-bound	1484.056	1.000	1484.056	22.222	.001
Error(PETTYPE)	Sphericity Assumed	734.611	22	33.391		
	Greenhouse-Geisser	734.611	18.394	39.937		
	Huynh-Feldt	734.611	21.305	34.481		
	Lower-bound	734.611	11.000	66.783		

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 $\Sigma p = .005 + .01 = .015$

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 !!!