ANOVA & Pairwise Comparisons

- ANOVA for multiple condition designs
- Pairwise comparisons and RH Testing
- Alpha inflation
- LSD and HSD procedures
- Effect sizes for k-group ANOVA
- Power analysis for k-group ANOVA

H0: Tested by 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 pairwise 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

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 (next page)
- Compute the minimum pairwise mean diff
- 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 !!!

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

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) =>	4.67
Enter MSe (Mean Square Error) =>	18.489
Select dferror (error degrees of freedom - use "next smallest" if no exact match) =>	10
LSD minimum mean difference =	6.27
HSD minimum mean difference =	7.72

LSD-HSD | Rsq chng F-test

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!

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			

Use these values to make pairwise comparisons

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

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 =	4.92
HSD minimum mean difference =	5.97

LSD-HSD | Rsq chng F-

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

k = # conditions

n = N = 12

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

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		

Use these values to make pairwise comparisons

Example analysis of a multiple IV conditions design

Tx1	Tx2	Cx
50	40	35

For this design, $F(2,27)=6.54$, $p < .05$ 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 < .05 was obtained.

Compute the pairwise mean differences.

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

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 !

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 < .05 was obtained. The minimum mean difference is 3

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

Tx1 vs. Tx2 7 Tx1 vs. C 4 Tx2 vs. C 2

Your Conclusions ?

Complete support for the RH: !!

“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 the 3 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 the more pairwise comparisons that 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

LSD vs. HSD Pairwise Comparisons

- Least Significant Difference (LSD)
  - Sensitive -- no correction for alpha inflation
    - smaller minimum mean difference than for HSD
    - More likely to find pairwise mean differences
      - Less likely to make Type II errors (Miss)
      - More likely to make Type I errors (False Alarm)
- Honest Significant Difference (HSD)
  - Conservative -- alpha corrected
    - larger minimum mean difference than for LSD
    - Less likely to find pairwise mean differences
    - More likely to make Type II errors
    - Less likely to make Type I errors
- Golden Rule: Perform both!!!
  - If they agree, there is less chance of committing either a Type I or Type II error !!!

LSD vs. HSD -- 3 Possible Outcomes for a Specific Pairwise Comparison

1 Both LSD & HSD show a significant difference

- having rejected H0: with the more conservative test (HSD) helps ensure that this is not a Type I error

2 Neither LSD nor HSD show a signif difference

- having found H0: with the more sensitive test (LSD) helps ensure this isn't a Type II error

- Both of these are "good" results, in that there is agreement between the statistical conclusions drawn from the two pairwise comparison methods

LSD vs. HSD -- 3 Possible Outcomes for a Specific Pairwise Comparison

3 Significant difference from LSD, but no significant difference from HSD

- This is a problem !!!
- Is HSD right & the sigdif from LSD a Type I error (FA)?
- Is the LSD is right & H0: from HSD a Type II error (miss) ?
- There is a bias toward "statistical conservatism" in Psychology -- using more conservative HSD & avoiding Type I errors (False alarms)

- A larger study may solve the problem -- LSD & HSD may both lead to rejecting H0: with a more powerful study
- Replication is the best way to decide which is "correct"

Here's an example…

A study was run to compare 2 treatments to each other and to a no-treatment control. The resulting means and mean differences were ...

	M	Tx1	Tx2
Tx1	12.3		
Tx2	14.6	2.3	
Cx	19.9	**7.6	* 5.3

Based on LSD mmd = 3.9

Based on HSD mmd = 6.7

Conclusions:

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

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

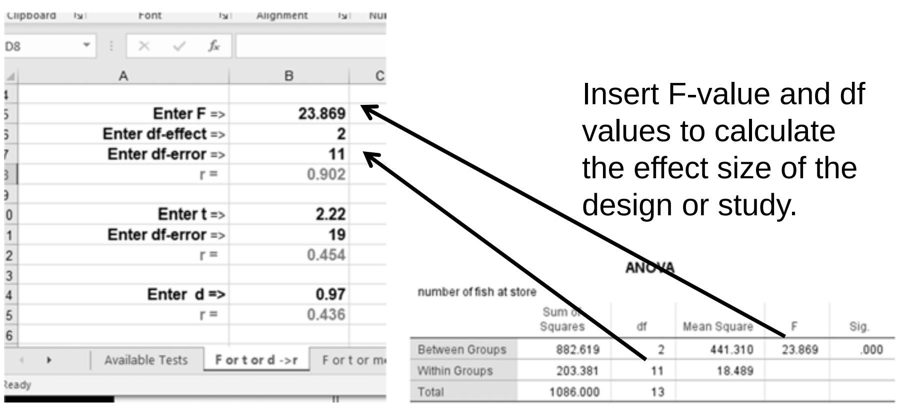
The effect size formula must take into account both the size of the sample (represented by dferror) and the size of the design (represented by the dfeffect).

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

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 the k-BG or k-WG → Omnibus F



Insert F-value and df values to calculate the effect size of the design or study.

The effect size estimate for a k-group design can only be compared to effect sizes from other studies with designs having 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: $r = \sqrt{\dfrac{d^2}{d^2 + 4}}$

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-BG → Pairwise Comparisons

Effect Size (r & d) for Pairwise Mean Comparison	
Select the type of ANOVA design =>	Between Groups (options: Between Groups / Within-Groups)
Enter mean #1 =>	17.43
Enter mean #2 =>	19.33
Enter MSe (Mean Square Error) =>	18.49
r =	0.216
d =	0.442

Use the drop-down menu to choose BG or WG design.

Copy in the means of the two groups being compared.

Copy in the MSerror

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

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			

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 k-WG → Pairwise Comparisons

Effect Size (r & d) for Pairwise Mean Comparison	
Select the type of ANOVA design =>	Within-Groups (options: Between Groups / Within-Groups)
Enter mean #1 =>	23.92
Enter mean #2 =>	21.5
Enter MSe (Mean Square Error) =>	33.39
r =	0.284
d =	0.419

Use the drop-down menu to choose BG or WG design.

Copy in the means of the two groups being compared.

Copy in the MSerror

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	[illegible].672	887.492	22.222	.000
	Huynh-Feldt	1484.056	1.9[illegible]7	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		

Power Analyses for k-BG designs

Important Symbols

S is the total # of participants in that pairwise comp

n = S / 2 is the # of participants in each condition of that pairwise comparison

N = n * k is the total number or participants in the study

Example

- the smallest pairwise effect size for a 3-BG study was .25
- with r = .25 and 80% power S = 120
- for each of the 2 conditions n = S / 2 = 120 / 2 = 60
- for the whole study N = n * k = 60 * 3 = 180

Power Analyses for k-WG designs

Important Symbols

S is the total # of participants in that pairwise comp

For WG designs, every participant is in every condition, so…

S is also the number of participants in each condition

Example

- the smallest pairwise effect size for a 3-WG study was .20
- with r = .20 and 80% power S = 191
- for each condition of a WG design n = S = 191
- for the whole study N = S = 191