Multiple Group X^2 Designs & Follow-up Analyses

- X^2 for multiple condition designs
- Pairwise comparisons & RH Testing
  - Alpha inflation
  - Effect sizes for k-group X^2
  - Power Analysis for k-group X^2
- gof-X^2 & RH Testing
  - Alpha inflation
  - Power Analyses

ANOVA vs. X^2

- Same as before
  - – ANOVA – BG design and a quantitative DV
  - – X^2 -- BG design and a qualitative/categorical DV

While quantitative outcome variables have long been more common in psychology, there has been an increase in the use of qualitative variables during the last several years.

- improvement vs. no improvement
- diagnostic category
- preference, choice, selection, etc.

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 whether or not the participants was "socially comfortable" in a large-group situation

$X^2(1) = 9.882$, $p = .005$

	Group therapy & self-appraisal	Cx
Comfortable	45	25
Not comfortable	10	25

Which of the following statements will these results support?

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

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
45	25	45	25
10	25	10	25

What responses 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."

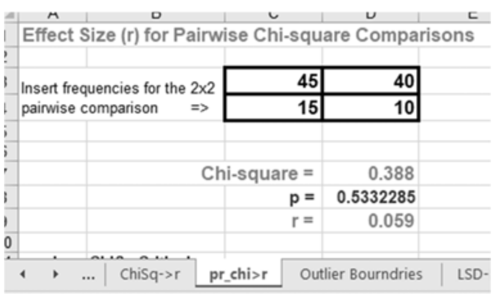
Omnibus X^2 vs. Pairwise Comparisons

- Omnibus X^2
  - overall test of whether there are any response pattern differences among the multiple IV conditions
  - Tests H0: that all the response patterns are equal
- Pairwise Comparison X^2
  - specific tests of whether or not each pair of IV conditions has a response pattern difference
- 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

Pairwise Comparisons for X^2

Using the Effect Size Computator, just plug in the cell frequencies for any 2x2 portion of the k-group design



Effect Size (r) for Pairwise Chi-square Comparisons

Insert frequencies for the 2x2 pairwise comparison =>	45	40
	15	10
Chi-square =		0.388
p =		0.5332285
r =		0.059

It calculates the p-value & effect size for each pairwise comparison.

Example of pairwise analysis of a multiple IV condition design

	Tx1	Tx2	Cx
Comfortable	45	40	25
Not comfortable	15	10	20

$X^2(2) = 7.641$, $p = .034$

	Tx1	Tx2
C	45	40
~C	15	10

$X^2(1) = .388$, $p > .05$

Retain H0: Tx1 = Tx2

	Tx1	Cx
C	45	25
~C	15	20

$X^2(1) = 4.375$, $p < .05$

Reject H0: Tx1 > Cx

	Tx2	Cx
C	40	25
~C	10	20

$X^2(1) = 6.549$, $p < .05$

Reject H0: Tx2 > Cx

What to do when you have a RH:

The RH: was, "In terms of the % who show improvement, immediate feedback (IF) is the best, with delayed feedback (DF) doing no better than the no feedback (NF) control."

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

IF > DF   IF > NF   DF = NF

Run the omnibus X^2 -- is there a relationship ?

	IF	DF	NF
Improve	78	40	65
Not improve	10	32	18

$X^2(2) = 23.917$, $p<.001$

Perform the pairwise X^2 analyses

	IF	DF
i	78	40
~i	10	32

$X^2(1)=22.384$, $p<.001$

Reject H0: IF > DF

	IF	NF
i	78	65
~i	10	18

$X^2(1)=3.324$, $p>.05$

Retain H0: IF = NF

	DF	NF
i	40	65
~i	32	18

$X^2(1)=9.137$, $p<.005$

Reject H0: DF < NF

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

RH:	IF > DF	IF > NF	DF = NF
well ?	supported	not supported	not supported

We would conclude that the RH: was partially supported !

Remember that pairwise comparisons are the same thing as simple analytic comparisons. It is also possible to perform complex comparisons with X^2

The RH: was, "In terms of the % who show improvement, those receiving feedback will do better than those receiving the no feedback (NF) control."

	IF	DF	NF
Improve	78	40	65
Not improve	10	32	18

	FB	NF
i	118	65
~i	42	18

$X^2(1)=.661$, $p>.05$

Reject H0: DF = NF

As with ANOVA, complex comparisons can be misleading if interpreted improperly → we would not want to say that "both types of feedback are equivalent to no feedback" ← that statement is false based on the pairwise comparisons.

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

There is no equivalent to HSD for X^2 follow-ups

- We can "Bonferronize" p = .05 / #comps to hold the experiment-wise Type I error rate to 5%
  - 2 comps → $X^2(1, .025) = 5.02$
  - 3 comps → $X^2(1, .0167) = 5.73$
  - 4 comps → $X^2(1, .0125) = 6.24$
  - 5 comps → $X^2(1, .01) = 6.63$
- As with ANOVA → when you use a more conservative approach you can find a significant omnibus effect but not find anything to be significant when doing the follow-ups!

Alpha Corrected pairwise comparisons for Chi-square

Effect Size (r) for Pairwise Chi-square Comparison

Insert frequencies for the 2x2 pairwise comparison =>	45	35
	15	10
Chi-square =		0.109
p =		0.7408568
r =		0.032

Bonferroni-corrected p-values for various numbers of pairwise comparisons

Number of conditions	Number of comparisons	critical p-value
2	1	0.05
	2	0.025
3	3	0.0167
	4	0.0125
	5	0.01
4	6	0.0083
5	10	0.005
6	15	0.0033

The computator also shows the critical Chi-square value for different p-values for "corrected" comparisons.

First: Compute the pariwise chi-square.

Second: Determine the p-value to use for NHST of that pairwise comparison.

If this were the comparison of two conditions from a 3-condition design…

3 conditions requires 3 comparisons, so we would use the p-value of .0167

Based on this Bonferroni-corrected p-value of .0167, we would conclude that this pariwise comparison (with a p-value of .032) is no significant!

k-group Effect Sizes

When you have more than 2 groups, it is possible to compute the effect size for "the whole study".

Effect Size (r) for Chi-Square

Enter Chi-Square =>	4.850
Enter df =>	1
Enter N =>	80
r =	0.246

Include the X^2, the total N and click the button for df > 1

However, this type of effect size is not very helpful, because:

-- you don't know which pairwise comparison(s) make up the r

-- it can only be compared to other designs with exactly the same combination of conditions

Pairwise Effect Sizes

Just as RH: for k-group designs involve comparing 2 groups at a time (pairwise comparisons)… The most useful effect sizes for k-group designs are computed as the effect size for 2 groups (effect sizes for pairwise comparisons)

Effect Size (r) for Pairwise Chi-square Comparisons

Insert frequencies for the 2x2 pairwise comparison =>	10	20
	20	10
Chi-square =		6.667
p =		0.0098233
r =		0.333

p-value	ChiSq Critical
0.05	3.84
0.025	5.02
0.0167	5.73
0.0125	6.24
0.01	6.63

The effect size computator calculates the effect size for each pairwise X^2 it computes

k-group Power Analyses

As before, there are two kinds of power analyses;;;

A priori power analyses

- conducted before the study is begun
- start with r & desired power to determine the needed N

Post hoc power analysis

- conducted after retaining H0:
- start with r & N and determine power & Type II probability

Power Analyses for k-group 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 X^2 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 * 60 = 180

As X^2 designs get larger, the required 2x2 follow-up analyses can get out of hand pretty quickly. For example …

Outcome	Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control
Improve	45	26	40	27
Stay same	10	22	12	23
Get worse	5	23	5	29

This would require 18 2x2 comparisons:

- 6 each for pairwise comparisons among the 4 IV conditions for each of improve/same, same/worse and improve/worse.
- The maximum experiment wise alpha would be 18*.05 or a 90% chance of making at least one Type I error.
- To correct for this we'd need to use a p-value of .05/18 = .003 for each of the 18 comparisons
- Which, in turn, greatly increases the chances of making Type II errors

Another approach to analyzing larger designs is to use gof-X^2 to describe response patterns of each condition or to test RH: that are phrased in terms of the response pattern predictions.

Outcome	Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control
Improve	45	26	40	27
Stay same	10	22	12	23
Get worse	5	23	5	29

For this design we would run 4 gof X^2 analyses.

As with the 2x2, there is no equivalent to HSD for X^2 follow-ups

- One approach is to use p=.01 for each pairwise comparison, reducing the alpha inflation
- Another is to "Bonferronize" p = .05 / #comps to hold the experiment-wise Type I error rate to 5%

The RH: for this study was that: The treatment works because of the cognitive self-appraisal; the group therapy doesn't really contribute anything."

Based on this we would expect that both the combined and self-appraisal conditions would have more "improve" than "stay same" or "get worse".

We would also expect a "flat" response profile for both the no-treatment and group therapy conditions.

Outcome	Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control
Improve	45	26	40	27
Stay same	10	22	12	23
Get worse	5	23	5	29

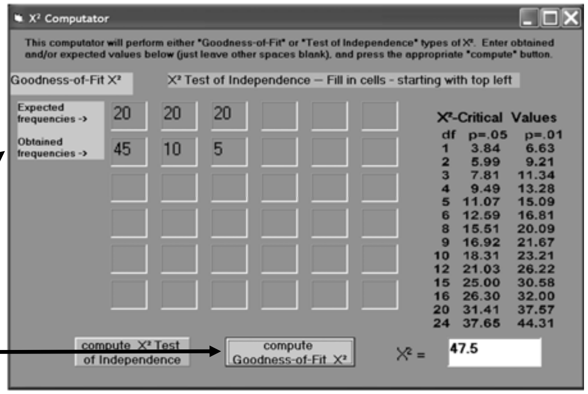
For the Group Therapy & Self-Appraisal condition…

- to perform the gof-X2 we need the expected frequency for the equiprobability H0:
- with n=60 and equiprobability, the expected frequency for each condition is 2

Enter the expected frequencies (usually representing equiprobability)

Enter the cell frequencies

Be sure to click the blue compute button



With df=2 (k-1) → $X^2(.01) = 9.21$ and so, $p < .01$

We'd conclude that this condition does not have equiprobability and that the response pattern matches the RH:

Here are the results of the follow-up analyses…

Outcome	Group therapy & self-appraisal	Group therapy	Self-appraisal	No-treatment control
Improve	45	26	40	27
Stay same	10	22	12	23
Get worse	5	23	5	29
	$X^2(2)=47.5$, $p<.001$	$X^2(2)=.37$, $p>.05$	$X^2(2)=36.11$, $p<.001$	$X^2(2)=.71$, $p>.05$

We would conclude that there is complete support for the RH: that → The treatment works because of the cognitive self-appraisal; the group therapy doesn't really contribute anything."

There are a couple of problems with X^2 follow-ups that you should consider…

- The follow-up analyses – both the 2x2 and the gof – have substantially less power than the onimibus test
  - So, it is possible to find a "significant overall effect that isn't anywhere"
  - The likelihood of this increases if you use alpha correction
- Neither the 2x2 nor the gof analyses are really "complete"
  - both analyses tell you that there is a pattern, but not what the pattern is
  - some recommend using 2-cell gof analyses to identify the specific location of the pattern – others point out the enormous alpha inflation or alpha correction involved…
    - for the example, each of the 18 2x2 follow-ups that is significant would require 2 additional 2-cell gof → as many as 18 + 36 follow-up analyses for a 3x4 design!!!