Parametric & Nonparametric Models for Within-Groups Comparisons

- overview
- WG Designs & WG RH:/RQ:
- McNemar's test
- Cochran's tests
- WG t-test & ANOVA
- Wilcoxin's test
- Friedman's F-test

Statistics We Will Consider

DV →	Categorical	Parametric Interval/ND	Nonparametric Ordinal/~ND
univariate stats	mode, #cats	mean, std	median, IQR
univariate tests	gof X^2	1-grp t-test	1-grp Mdn test
association	X^2	Pearson's r	Spearman's r
2 bg	X^2	t- / F-test	M-W K-W Mdn
k bg	X^2	F-test	K-W Mdn
2wg	McNem Crn's	t- / F-test	Wil's Fried's
kwg	Crn's	F-test	Fried's

M-W -- Mann-Whitney U-Test Wil's -- Wilcoxin's Test Fried's -- Friedman's F-test
K-W -- Kruskal-Wallis Test
Mdn -- Median Test McNem -- McNemar's X^2 Crn's – Cochran's Test

Repeated measures designs…

There are two major kinds of these designs:

1) same cases measured on the same variable at different times or under different conditions

   - pre-test vs. post-test scores of clients receiving therapy
   - performance scores under feedback vs. no feedback conds
   - % who "pass" before versus after remedial training

2) same cases measured at one time under one condition, using different (yet comparable) measures

   - comparing math and reading scores (both T-scores, with mean=50 and std=10)
   - number of "omissions" (words left out) and "intrusions" (words that shouldn't have been included) in a word recall task
   - % who "pass" using two different tests

Repeated measures designs…

There is really a third related kind of design:

3) non-independent groups of cases measured on the same variable at different times or under different conditions

- matched-groups designs
- snow-ball sampling over time

Statistically speaking, groups-comparisons analyses divide into 2 kinds”

- independent groups designs → Between Groups designs
- dependent groups designs → within-groups & Matched-groups designs

For all dependent groups designs, the non-independence of the groups allows the separation of variance due to “differences among people” from variance due to “unknown causes” (error or residual variance)

For repeated measures designs (especially of the first 2 kinds), there are two different types of research hypotheses or questions that might be posed…

1) Do the measures have different means (dif resp dist for qual DVs)
   - are post-test scores higher than pre-test scores?
   - is performance better with feedback than without it?
   - are reading scores higher than math scores?
   - are there more omissions than intrusions?

2) Are the measures associated?
   - are the folks with the highest pre-test scores also the ones with the highest post-test scores?
   - is performance with feedback predictable based on performance without feedback?
   - are math scores and reading scores correlated?
   - do participants who make more omissions also tend to make more intrusions?

So, taken together there are four “kinds of” repeated measures analyses. Each is jointly determined by the type of design and the type of research hypothesis/question. Like this…

	Type of Hypothesis/Question	
Type of Design	mean difference	association
Different times or situations	pre-test < post-test	pre-test & post-test
Different measures	math < spelling	math & spelling

But… All the examples so far have used quantitative variables.

Qualitative variables could be used with each type of repeated measures design (dif times vs. dif measures)

Consider the difference between the following examples of repeated measures designs using a qualitative (binary) response or outcome variable

- The same % of students will be identified as needing remedial instruction at the beginning and end of the semester (dif times).
- The same students will be identified as needing remedial instruction at the end of the semester as at the beginning (dif times)
- The same % of folks will be identified as needing remedial instruction based on teacher evaluations as based on a standardized test (dif measures)
- The same folks will be identified as needing remedial instruction based on teacher evaluations as based on a standardized test (dif measures)

So, we have to expand our thinking to include 8 situations...

So, for repeated measures designs, here are the analytic "situations" and the statistic to use for each

	Type of Question/Hypothesis			
	Quant Vars		Qual Vars	
Type of Design	mean dif	assoc	% dif^	pattern^*
Different times or situations	wg t/F-test	Pearson's r	Cochrans	McNemar's X^2
Different measures	wg t/F-test	Pearson's r	Cochran's	McNemar's X^2

^ Cochran's and McNemar's are for use only with binary variables

* McNemar's looks at patterns of classification disagreements

Statistical Tests for WG Designs w/ qualitative variables

McNemar's test

Of all these tests, McNemar's has the most specific application…

- are two qualitative variable related -- Pearson's X^2
- do groups have differences on a qual variable -- Pearson's X^2
- does a group change % on a binary variable -- Cochran's
- is the the relationship between the variables revealed by an asymmetrical pattern of "disagreements" –McNemar's

e.g., more folks are classified as "pass" by the computer test but "fail" by the paper test than are classified as "fail" by the computer test but "pass" by the paper test.

	computer test	
paper test	pass	fail
pass	40	4
fail	12	32

	Cond #1	
Cond #2	value 1	value 2
value 1	a	b
value 2	c	d

$$X^2 = \frac{(b - c)^2}{(b + c)}$$

McNemar’s always has df=1

	computer test	
paper test	pass	fail
pass	40	4
fail	12	32

$$X^2 = \frac{(4 - 12)^2}{(4 + 12)}$$

$$= 4$$

Compare the obtained X^2 with $X^2_{1, .05} = 3.84$. We would reject H0: and conclude that there is a relationship between what performance on the paper test and performance on the computer test & that more uniquely fail the paper test than uniquely fail the computer test.

Cochran’s Q-test – can be applied to 2 or k-groups

The simplest “qualitative variable” situation is when the variable is binary. Then “changes in response distribution” becomes the much simple “changes in %”.

Begin the computation of Q by arranging the data with each case on a separate row. 1 = pass 0 = fail

	pretest	posttest	retention	L	L^2
S1	0	1	1	2	4
S2	0	1	0	1	1
S2	0	0	1	1	1
S2	1	1	1	3	9
S5	0	0	1	1	1
G	1	3	4		
G^2	1	9	16		

Compute the sum for each column (G) and it’s square (G^2)

Compute the sum for each row (L) and its square (L^2)

	pretest	posttest	retention	L	L^2
S1	0	1	1	2	4
S2	0	1	0	1	1
S2	0	0	1	1	1
S2	1	1	1	3	9
S5	0	0	1	1	1
G	1	3	4		
G^2	1	9	16		

k = # conditions

$$\mathbf{Q} = \frac{(k-1)*[\,(k * \Sigma G^2) - (\Sigma G)^2\,]}{(k * \Sigma L) - \Sigma L^2} = \frac{(3-1)*[(3*(1+9+16)) - (1+3+4)^2]}{(3 * (2+1+1+3+1)) - (4+1+1+9+1)} = 3.0$$

Q is compared to X^2 critical based on df = k-1 $X^2_{2, .05} = 7.81$

So we would retain H0: of no % difference across the design conditions.

Parametric tests for WG Designs using ND/Int variables

t-tests

- H0: Populations represented by the IV conditions have the same mean DV.
- degrees of freedom df = N - 1
- Range of values $-\infty$ to ∞
- Reject Ho: If $| t_{obtained} | > t_{critical}$
- Assumptions
  - data are measured on an interval scale
  - DV values from both groups come from ND & have equal STDs

ANOVA

- H0: Populations represented by the IV conditions have the same mean DV.
- degrees of freedom df numerator = k-1, denominator = N - k
- Range of values 0 to ∞
- Reject Ho: If $F_{obtained} > F_{critical}$
- Assumptions
  - data are measured on an interval scale
  - DV values from both groups come from ND with equal STD
  - for k > 2 – data from any pair of conditions are equally correlated

Nonparametric tests for WG Designs using ~ND/~Int variables

- within-subjects design - same subjects giving data under each of two or more conditions
- comparison of two or more “comparable” variables -- same subjects giving data on two variables (same/dif time)
- matched-groups design -- matched groups of two or more members, each in one of the conditions

The nonparametric RM models we will examine and their closest parametric RM counterparts…

2-WG Comparisons

Wilcoxin’s Test dependent t-test

2- or k-WG Comparisons

Friedman’s ANOVA dependent ANOVA

Let’s start with a review of applying a within groups t-test

Here are the data from such a design :

IV is Before vs. After the child “discovers” Barney (and watches it incessantly, exposing you to it as well) so..

1st Quant variable is 1-10 rating “before” discovery

2nd Quant variable is 1-10 rating “after discovery”

Before		After		Difference
s1	2	s1	6	-4
s2	4	s2	8	-4
s3	6	s3	9	-3
s4	7	s4	10	-3
M =	4.75	M =	8.25	M_d = -3.5

A WG t-test can be computed as a single-sample t-test using the differences between an individual’s scores from the 2 design conditions.

- Rejecting the H0: M_d=0, is rejecting the H0: $M_{before} = M_{after}$
- other formulas exist

When using a WG t-test (no matter what computational form_ the assumption of interval measurement properties is even "more assuming" than for the BG design. We assume …

- that each person's ratings are equally spaced -- that the difference between ratings given by S1 of "3" and "5" mean the same thing as the difference between their ratings of "8" and "10" ???
- that different person's rating are equally spaced -- that the difference between ratings given by S1 of "3" and "5" mean the same thing as the difference between ratings of "8" and "10" given by S2 ???

Nonparametric tests for WG Designs using ~ND/~Int variables

Wilcoxin's Test

If we want to avoid some assumptions, we can apply a nonparametric test. To do that we …

- Compute the differences between each person's scores
- Determine the "signed ranks" of the differences
- Compute the summary statistic W from the signed ranks

Before		After		Difference	Signed Ranks
s1	2	s1	5	3	2.5
s2	4	s2	8	4	4
s3	6	s3	9	3	2.5
s4	9	s4	7	-2	-1

The "W" statistic is computed from the signed ranks. W=0 when the signed ranks for the two groups are the same (H0:)

There are two different "versions" of the H0: for the Wilcoxin's test, depending upon which text you read.

The "older" version reads:

H0: The two sets of scores represent a population with the same distribution of scores under the two conditions.

Under this H0:, we might find a significant U because the samples from the two situations differ in terms of their:

- centers (medians - with rank data)
- variability or spread
- shape or skewness

This is a very "general" H0: and rejecting it provides little info.

Also, this H0: is not strongly parallel to that of the t-test (that is specifically about mean differences)

Over time, “another” H0: has emerged, and is more commonly seen in textbooks today:

H0: The two sets of scores represent a population with the same median under the two conditions (assuming these populations have distributions with identical variability and shape).

You can see that this H0:

- increases the specificity of the H0: by making assumptions (That’s how it works - another one of those “trade-offs”)
- is more parallel to the H0: of the t-test (both are about “centers”)
- has essentially the same distribution assumptions as the t-test (equal variability and shape)

Finally, there are also “forms” of the Wilcoxin’s Test:

With smaller samples (N < 10-50 depending upon the source ??)

- Compare the $W_{obtained}$ with a $W_{critical}$ that is determined based on the sample size

With larger samples (N > 10-50)

- with these larger samples the distribution of Uobtained values approximates a normal distribution
- a Z-test is used to compare the Uobtained with the Ucritical
- the $Z_{obtained}$ is compared to a critical value of 1.96 (p = .05)

You should notice considerable similarity between the Mann-Whitney U-test and the Wilcoxin -- in fact, there are BG and RM versions of each -- so be sure to ask the “version” whenever you hear about one of these tests.

Nonparametric tests for WG Designs using ~ND/~Int variables

Friedman’s test applies this same basic idea (comparing ranks), but can be used to compare any number of groups.

- Each subject’s DV values are converted to rankings (across IV conditions)
- Score ranks are summed within each IV Condition and used to compute a summary statistic “F”, which is compared to a critical value to test H0:
- E.g., -- more of Barney . . . (from different “stages of exposure”)

	Before		After 6 months		After 12 months	
	DV	rank	DV	rank	DV	rank
S1	3	1	7	3	5	2
S2	5	1	9	3	6	2
S3	4	2	6	3	2	1
S4	3	1	6	2	9	3

- H0: has same two “versions” as the other nonparametric tests
  - DVs from populations with same score distributions
  - DVs from populations with same median (assuming …)
- Rejecting H0: requires pairwise follow-up analyses
  - Bonferroni correction -- $p_{critical}$ = (.05 / # pairwise comps)
- Finally, there are also “forms” of Friedman’s Test:
  - With smaller samples (k < 6 & N < 14)
    - Compare the $F_{obtained}$ with a $F_{critical}$ that is determined based on the sample size & number of conditions
  - With larger samples (k > 6 or N > 14)
    - the $F_{obtained}$ is compared to a $X^2_{critical}$ value