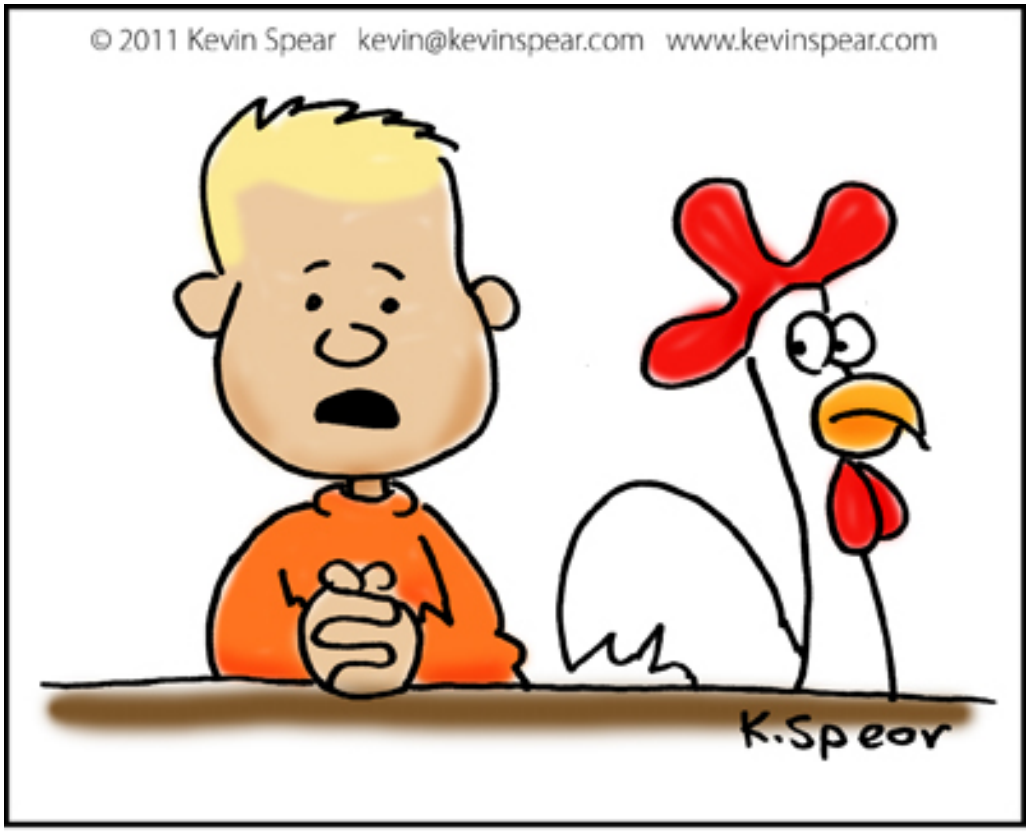
Estimating Reliability

RCMAR/EXPORT Methods Seminar Series

Drew (Cobb Room 131)/ UCLA (2nd Floor at Broxton)

December 12, 2011

Anonymous Dedication



"I want to thank you for showing me
what it truly means to be a chicken."

Reliability Minimum Standards

- 0.70 or above (for group comparisons)
- 0.90 or higher (for individual assessment)
  - SEM = SD $(1 - \text{reliability})^{1/2}$

Two Raters' Ratings of GOP Debate Performance on *Excellent* to *Poor* Scale

- Bachman Turner Overdrive (Good, Very Good)
- Ging Rich (Very Good, Excellent)
- Rue Paul (Good, Good)
- Gaylord Perry (Fair, Poor)
- Romulus Aurelius (Excellent, Very Good)
- Sanatorium (Fair, Fair)

Cross-Tab of Ratings

Rater 2 \ Rater 1	P	F	G	VG	E	Total
P	0	1				1
F		1				1
G			1			1
VG			1	0	1	2
E				1	0	1
Total	0	2	2	1	1	6

Calculating KAPPA

$$\mathbf{P_C =} \frac{(0 \times 1) + (2 \times 1) + (2 \times 1) + (1 \times 2) + (1 \times 1)}{(6 \times 6)} = 0.19$$

$$\mathbf{P_{obs.} =} \frac{2}{6} = 0.33$$

$$\mathbf{Kappa =} \frac{0.33 - 0.19}{1 - 0.19} = \mathbf{0.17} \quad (0.52, 0.77)$$

Linear and Quadratic Weighted Kappa

	P	F	G	VG	E
P	1	.75 (.937)	.50 (.750)	.25 (.437)	0
F	.75 (.937)	1	.75 (.937)	.50 (.750)	.25 (.437)
G	.50 (.750)	.75 (.937)	1	.75 (.937)	.50 (.750)
VG	.25 (.437)	.50 (.750)	.75 (.937)	1	.75 (.937)
E	0	.25 (.437)	.5 (.750)	.75 (.937)	1

$W_i = 1 - (i/ (k - 1)$ I = number of categories ratings differ by k = n of categories

$W_i = 1 - (i^2 / (k - 1)^2$

Intraclass Correlation and Reliability

Model	Reliability	Intraclass Correlation
One-way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (k-1)MS_{WMS}}$
Two-way fixed	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS}}$
Two-way random	$\frac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS} + k(MS_{JMS} - MS_{EMS})/N}$

BMS = Between Ratee Mean Square N = n of ratees
WMS = Within Mean Square k = n of replicates
JMS = Item or Rater Mean Square
EMS = Ratee x Item (Rater) Mean Square

01 34
02 45
03 33
04 21
05 54
06 22

Reliability of Performance Ratings

Source	df	SS	MS
Candidates (BMS)	5	15.67	3.13
Raters (JMS)	1	0.00	0.00
Cand. x Raters (EMS)	5	2.00	0.40
Total	11	17.67	

$$\text{2-way R} = \frac{2\,(3.13 - 0.40)}{2\,(3.13) + 0.00 - 0.40} = \boxed{0.89}$$

ICC = 0.80

GOP Presidential Candidates Responses to Two Questions about Their Health

- Bachman Turner Overdrive (Good, Very Good)
- Ging Rich (Very Good, Excellent)
- Rue Paul (Good, Good)
- Gaylord Perry (Fair, Poor)
- Romulus Aurelius (Excellent, Very Good)
- Sanatorium (Fair, Fair)

01 34
02 45
03 33
04 21
05 54
06 22

Two-Way Fixed Effects (Cronbach's Alpha)

Source	df	SS	MS
Respondents (BMS)	5	15.67	3.13
Items (JMS)	1	0.00	0.00
Resp. x Items (EMS)	5	2.00	0.40
Total	11	17.67	

$$\text{Alpha} = \frac{3.13 - 0.40}{3.13} = \frac{2.93}{3.13} = \boxed{0.87}$$

Overall Satisfaction of 12 Patients with 6 Doctors (2 patients per doctor)

- Dr. Overdrive (p1: Good, p2: Very Good)
- Dr. Rich (p3: Very Good, p4: Excellent)
- Dr. Paul (p5: Good, p6: Good)
- Dr. Perry (p7: Fair, p8: Poor)
- Dr. Aurelius (p9: Excellent, p10: Very Good)
- Dr. Sanatorium (p11: Fair, p12: Fair)

01 34
02 45
03 33
04 21
05 54
06 22

Reliability of Ratings of Doctor

Source	df	SS	MS
Respondents (BMS)	5	15.67	3.13
Within (WMS)	6	2.00	0.33
Total	11	17.67	

$$\text{1-way} = \frac{3.13 - 0.33}{3.13} = \frac{2.80}{3.13} = \boxed{0.89}$$

Candidates Perceptions of the U.S. Economy in November & December, 2011

- Bachman Turner Overdrive (Good, Very Good)
- Ging Rich (Very Good, Excellent)
- Rue Paul (Good, Good)
- Gaylord Perry (Fair, Poor)
- Romulus Aurelius (Excellent, Very Good)
- Sanatorium (Fair, Fair)

Which model would you use to estimate reliability?

Reliability and SEM

- For z-scores (mean = 0 and SD = 1):
  - Reliability = $1 - SE^2$
  - So reliability = **0.90** when SE = 0.32

- For T-scores (mean = 50 and SD = 10):
  - Reliability = $1 - (SE/10)^2$
  - So reliability = **0.90** when SE = 3.2

In the past 7 days

I was grouchy [1st question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 56.1 SE = 5.7 (rel. = 0.68)

In the past 7 days …

I felt like I was read to explode [2nd question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 51.9 SE = 4.8 (rel. = 0.77)

In the past 7 days …

I felt angry [3rd question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 50.5 SE = 3.9 (rel. = 0.85)

In the past 7 days …

I felt angrier than I thought I should [4th question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 48.8 SE = 3.6 (rel. = 0.87)

In the past 7 days …

I felt annoyed [5th question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 50.1 SE = 3.2 (rel. = 0.90)

In the past 7 days …

I made myself angry about something just by thinking about it. [6th question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 50.2 SE = 2.8 (rel = 0.92)

Theta and SEM estimates

- 56 and 6 (reliability = .68)
- 52 and 5 (reliability = .77)
- 50 and 4 (reliability = .85)
- 49 and 4 (reliability = .87)
- 50 and 3 (reliability = .90)
- 50 and <3 (reliability = .92)

Thank you.

Powerpoint file posted at URL below (freely available for you to use, copy or burn):

http://gim.med.ucla.edu/FacultyPages/Hays/

http://www.chime.ucla.edu/measurement/wip.htm

Contact information:

drhays@ucla.edu 310-794-2294

For a good time call 8675309 or go to:

http://twitter.com/RonDHays

Appendices
ANOVA Computations

- Candidate/Respondents SS

  $(7^2+9^2+6^2+3^2+9^2+4^2)/2 - 38^2/12 = \underline{15.67}$

- Rater/Item SS

  $(19^2+19^2)/6 - 38^2/12 = \underline{0.00}$

- Total SS

  $(3^2+ 4^2+4^2+5^2+3^2+3^2+2^2+1^2+5^2+4^2+2^2+2^2) - 38^2/10 = \underline{17.67}$

- Res. x Item SS= Tot. SS – (Res. SS+Item SS)

```
options ls=130 ps=52 nocenter;
options nofmterr;

data one;
 input id 1-2 rater 4 rating 5;
CARDS;
01 13
01 24
02 14
02 25
03 13
03 23
04 12
04 21
05 15
05 24
06 12
06 22
;
run;
**************;
```

```
proc freq;
tables rater rating;
run;
******************.
                   ;
proc means;
var rater rating;
run;
****************************************.
                                        ;
proc anova;
class id rater;
model rating=id rater id*rater;
run;
****************************************.
                                        ;
```

```
data one;
input id 1-2 rater 4 rating 5;
CARDS;
01 13
01 24
02 14
02 25
03 13
03 23
04 12
04 21
05 15
05 24
06 12
06 22
;
run;
*****************************************************************;
%GRIP(indata=one,targetv=id,repeatv=rater,dv=rating,
      type=1,t1=test of GRIP macro,t2=);
```

GRIP macro is available at: http://gim.med.ucla.edu/FacultyPages/Hays/util.htm

```
data one;
 input id 1-2 rater1 4 rater2 5;
 control=1;
CARDS;
01 34
02 45
03 33
04 21
05 54
06 22
;
run;
**************.
                ;
DATA DUMMY;
INPUT id 1-2 rater1 4 rater2 5;
CARDS;
01 11
02 22
03 33
04 44
05 55
RUN;
```

```
DATA NEW;
SET ONE DUMMY;
PROC FREQ;
TABLES CONTROL*RATER1*RATER2
/NOCOL NOROW NOPERCENT AGREE;
*****************************************.
                                         ;
data one;
 set one;
 *****************************************.
                                         ;
proc means;
var rater1 rater2;
run;
*****************************************.
                                         ;
proc corr alpha;
var rater1 rater2;
run;
```

Guidelines for Interpreting Kappa

Conclusion	Kappa	Conclusion	Kappa
Poor	< .40	Poor	< 0.0
Fair	.40 - .59	Slight	.00 - .20
Good	.60 - .74	Fair	.21 - .40
Excellent	> .74	Moderate	.41 - .60
		Substantial	.61 - .80
		Almost perfect	.81 - 1.00
Fleiss (1981)		Landis and Koch (1977)	