More on reliability

Telemetrics lab

Department of Psychology
Northwestern University
Evanston, Illinois USA

October, 2010
Outline

1. Types of reliability
   - Alpha and its alternatives

2. Calculating reliabilities
   - Congeneric measures
   - Hierarchical structures

3. \(2 \neq 1\)
   - Multiple dimensions - falsely labeled as one
   - Using score.items to find reliabilities of multiple scales

4. Intraclass correlations
   - ICC of judges

5. Kappa
   - Cohen’s kappa
   - Weighted kappa
Types of reliability

- **Internal consistency**
  - $\alpha$
  - $\omega_{\text{hierarchical}}$
  - $\omega_{\text{total}}$
  - $\beta$

- **Intraclass Agreement**

- **Test-retest, alternate form**

- **Generalizability**

- **Internal consistency**
  - alpha, score.items
  - omega
  - iclust

- **icc**

- **wkappa, cohen.kappa**

- **cor**

- **aov**
Alpha and its alternatives

- Reliability: \( \frac{\sigma^2_t}{\sigma^2_x} = 1 - \frac{\sigma^2_e}{\sigma^2_x} \)
- If there is another test, then \( \sigma_t = \sigma_{t_1}t_2 \) (covariance of test \( X_1 \) with test \( X_2 = C_{xx} \))
- But, if there is only one test, we can estimate \( \sigma_t^2 \) based upon the observed covariances within test 1
- How do we find \( \sigma_e^2 \)?
- The worst case, (Guttman case 1) all of an item’s variance is error and thus the error variance of a test \( X \) with variance-covariance \( C_x \)
  - \( C_x = \sigma^2_e = \text{diag}(C_x) \)
  - \( \lambda_1 = \frac{C_x - \text{diag}(C_x)}{C_x} \)
- A better case (Guttman case 3, \( \alpha \)) is that that the average covariance between the items on the test is the same as the average true score variance for each item.
  - \( C_x = \sigma^2_e = \text{diag}(C_x) \)
  - \( \lambda_3 = \alpha = \lambda_1 \cdot \frac{n}{n-1} = \frac{(C_x - \text{diag}(C_x)) * n}{(n-1)C_x} \)
Guttman 6: estimating using the Squared Multiple Correlation

- Reliability \( = \frac{\sigma^2_t}{\sigma^2_x} = 1 - \frac{\sigma^2_e}{\sigma^2_x} \)
- Estimate true item variance as squared multiple correlation with other items
  \( \lambda_6 = \frac{(C_x - \text{diag}(C_x) + \Sigma(smci)}{C_x} \)
  - This takes observed covariance, subtracts the diagonal, and replaces with the squared multiple correlation
  - Similar to \( \alpha \) which replaces with average inter-item covariance
- Squared Multiple Correlation is found by \( smc \) and is just
  \( smci = 1 - 1/R_{ii}^{-1} \)
First, create some simulated data with a known structure

```r
> set.seed(42)
> v4 <- sim.congeneric(N=200,short=FALSE)
> str(v4) #show the structure of the resulting object
List of 6
$ model : num [1:4, 1:4] 1 0.56 0.48 0.4 0.56 1 0.42 0.35 0.48 0.42 ...  
  ..- attr(*, "dimnames")=List of 2  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
$ pattern : num [1:4, 1:5] 0.8 0.7 0.6 0.5 0.6 ...  
  ..- attr(*, "dimnames")=List of 2  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
  .. ..$: chr [1:5] "theta" "e1" "e2" "e3" ...  
$ r : num [1:4, 1:4] 1 0.546 0.466 0.341 0.546 ...  
  ..- attr(*, "dimnames")=List of 2  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
$ latent : num [1:200, 1:5] 1.371 -0.565 0.363 0.633 0.404 ...  
  ..- attr(*, "dimnames")=List of 2  
  .. ..$: NULL  
  .. ..$: chr [1:5] "theta" "e1" "e2" "e3" ...  
$ observed: num [1:200, 1:4] -0.104 -0.251 0.993 1.742 -0.503 ...  
  ..- attr(*, "dimnames")=List of 2  
  .. ..$: NULL  
  .. ..$: chr [1:4] "V1" "V2" "V3" "V4"  
$ N : num 200  
  - attr(*, "class")= chr [1:2] "psych" "sim"
```
Types of reliability

Calculating reliabilities

Intraclass correlations

Kappa

Congeneric measures

A congeneric model

> f1 <- fa(v4$model)
> fa.diagram(f1)

Factor Analysis

V1 V2 V3 V4
MR1 0.8
V1 1.00 0.56 0.48 0.40
V2 0.56 1.00 0.42 0.35
V3 0.48 0.42 1.00 0.30
V4 0.40 0.35 0.30 1.00

> v4$model

V1 V2 V3 V4
V1 1.00 0.55 0.47 0.34
V2 0.55 1.00 0.38 0.30
V3 0.47 0.38 1.00 0.31
V4 0.34 0.30 0.31 1.00

> round(cor(v4$observed),2)

V1 V2 V3 V4
V1 1.00 0.55 0.47 0.34
V2 0.55 1.00 0.38 0.30
V3 0.47 0.38 1.00 0.31
V4 0.34 0.30 0.31 1.00
Types of reliability

Calculating reliabilities

2 ≠ 1

Intraclass correlations

Kappa

Congeneric measures

Find $\alpha$ and related stats for the simulated data

```r
> alpha(v4$observed)

Reliability analysis
Call: alpha(x = v4$observed)

       raw_alpha std.alpha   G6(smc) average_r    mean     sd
V1         0.71     0.72     0.67       0.39  -0.036   0.72
V2         0.63     0.64     0.55       0.37  -0.048   0.70
V3         0.65     0.66     0.59       0.40  -0.119   0.91
V4         0.72     0.72     0.64       0.46  -0.034   0.72

Reliability if an item is dropped:

       raw_alpha std.alpha   G6(smc) average_r
V1         0.59     0.60     0.50       0.33
V2         0.63     0.64     0.55       0.37
V3         0.65     0.66     0.59       0.40
V4         0.72     0.72     0.64       0.46

Item statistics

       n   r   r.cor r.drop  mean    sd
V1    200 0.80 0.72  -0.015 0.93
V2    200 0.76 0.64  -0.060 0.98
V3    200 0.73 0.59  -0.119 0.92
V4    200 0.66 0.46   0.049 1.09
```
Hierarchical structures

**A hierarchical structure**

cor.plot(r9)

```
> set.seed(42)
> r9 <- sim.hierarchical()
> round(r9, 2)

V1   V2   V3   V4   V5   V6   V7   V8   V9
V1 1.00 0.56 0.48 0.40 0.35 0.29 0.30 0.25 0.20
V2 0.56 1.00 0.42 0.35 0.30 0.26 0.25 0.20 0.18
V3 0.48 0.42 1.00 0.30 0.26 0.22 0.23 0.19 0.15
V4 0.40 0.35 0.30 1.00 0.42 0.35 0.24 0.20 0.16
V5 0.35 0.30 0.26 0.42 1.00 0.30 0.20 0.17 0.13
V6 0.29 0.25 0.22 0.35 0.30 1.00 0.17 0.14 0.11
V7 0.30 0.26 0.23 0.24 0.20 0.17 1.00 0.30 0.24
V8 0.25 0.22 0.19 0.20 0.17 0.14 0.30 1.00 0.20
V9 0.20 0.18 0.15 0.16 0.13 0.11 0.24 0.20 1.00
```
Types of reliability

Hierarchical structures

α of the 9 hierarchical variables

> alpha(r9)

Reliability analysis

Call: alpha(x = r9)

raw_alpha std.alpha G6(smc) average_r
0.76 0.76 0.76 0.26

Reliability if an item is dropped:

raw_alpha std.alpha G6(smc) average_r
V1 0.71 0.71 0.70 0.24
V2 0.72 0.72 0.71 0.25
V3 0.74 0.74 0.73 0.26
V4 0.73 0.73 0.72 0.25
V5 0.74 0.74 0.73 0.26
V6 0.75 0.75 0.74 0.27
V7 0.75 0.75 0.74 0.27
V8 0.76 0.76 0.75 0.28
V9 0.77 0.77 0.76 0.29

Item statistics

r r.cor
V1 0.72 0.71
V2 0.67 0.63
V3 0.61 0.55
V4 0.65 0.59
V5 0.59 0.52
V6 0.53 0.43
V7 0.56 0.46
V8 0.50 0.39
V9 0.45 0.32
Types of reliability

Calculating reliabilities

2 ≠ 1

Intraclass correlations

Kappa

Multiple dimensions - falsely labeled as one

An example of two different scales confused as one

Correlation plot

> set.seed(17)
> two.f <- sim.item(8)
> round(cor(two.f),2)

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1.00</td>
<td>0.29</td>
<td>0.05</td>
<td>0.03</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.06</td>
<td>-0.08</td>
</tr>
<tr>
<td>V2</td>
<td>0.29</td>
<td>1.00</td>
<td>0.03</td>
<td>-0.02</td>
<td>-0.35</td>
<td>-0.33</td>
<td>0.02</td>
<td>-0.04</td>
</tr>
<tr>
<td>V3</td>
<td>0.05</td>
<td>0.03</td>
<td>1.00</td>
<td>0.34</td>
<td>-0.02</td>
<td>-0.10</td>
<td>-0.40</td>
<td>-0.39</td>
</tr>
<tr>
<td>V4</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.34</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.06</td>
<td>-0.36</td>
<td>-0.37</td>
</tr>
<tr>
<td>V5</td>
<td>-0.38</td>
<td>-0.35</td>
<td>-0.02</td>
<td>-0.01</td>
<td>1.00</td>
<td>0.33</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>V6</td>
<td>-0.38</td>
<td>-0.33</td>
<td>-0.10</td>
<td>0.06</td>
<td>0.33</td>
<td>1.00</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>V7</td>
<td>-0.06</td>
<td>0.02</td>
<td>-0.40</td>
<td>-0.36</td>
<td>0.03</td>
<td>0.04</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>V8</td>
<td>-0.08</td>
<td>-0.04</td>
<td>-0.39</td>
<td>-0.37</td>
<td>0.05</td>
<td>0.03</td>
<td>0.37</td>
<td>1.00</td>
</tr>
</tbody>
</table>

> cor.plot(cor(two.f),zlim=c(-1,1),colors=TRUE)
Note the use of the keys parameter to specify how some items should be reversed.

> alpha(two.f, keys=c(rep(1,4), rep(-1,4)))

Reliability analysis
Call: alpha(x = two.f, keys = c(rep(1, 4), rep(-1, 4)))

raw_alpha  std.alpha  G6(smc)  average_r  mean   sd
0.62       0.62       0.65    0.17    -0.0051 0.27

Reliability if an item is dropped:
raw_alpha  std.alpha  G6(smc)  average_r
V1 0.59  0.58  0.61  0.17
V2 0.61  0.60  0.63  0.18
V3 0.58  0.58  0.60  0.16
V4 0.60  0.60  0.62  0.18
V5 0.59  0.59  0.61  0.17
V6 0.59  0.59  0.61  0.17
V7 0.58  0.58  0.61  0.17
V8 0.58  0.58  0.60  0.16

Item statistics
n  r  r.cor  r.drop  mean   sd
V1 500 0.54  0.44   0.33  0.063 1.01
V2 500 0.48  0.35   0.26  0.070 0.95
V3 500 0.56  0.47   0.36 -0.030 1.01
V4 500 0.48  0.37   0.28 -0.130 0.97
V5 500 0.52  0.42   0.31 -0.073 0.97
V6 500 0.52  0.41   0.31 -0.071 0.95
V7 500 0.53  0.44   0.34  0.035 1.00
V8 500 0.56  0.47   0.36  0.097 1.02
First, make up a keys matrix to specify which items should be scored, and in which way

```r
> keys <- make.keys(nvars=8,keys.list=list(one=c(1,2,-5,-6),two=c(3,4,-7,-8)))
> keys
    one two
[1,]  1  0
[2,]  1  0
[3,]  0  1
[4,]  0  1
[5,] -1  0
[6,] -1  0
[7,]  0 -1
[8,]  0 -1
```
Now score the two scales and find $\alpha$ and other reliability estimates

```r
> score.items(keys,two.f)
Call: score.items(keys = keys, items = two.f)
(Unstandardized) Alpha:
          one  two
alpha  0.68  0.7
Average item correlation:
          one  two
average.r 0.34  0.37
Guttman 6* reliability:
          one  two
Lambda.6  0.62  0.64
Scale intercorrelations corrected for attenuation
  raw correlations below the diagonal, alpha on the diagonal
  corrected correlations above the diagonal:
          one  two
one 0.68  0.08
two 0.06  0.70
Item by scale correlations:
  corrected for item overlap and scale reliability
          one  two
V1 0.57  0.09
V2 0.52  0.01
V3 0.09  0.59
V4 -0.02  0.56
V5 -0.58 -0.05
V6 -0.57 -0.05
V7 -0.05 -0.58
V8 -0.09 -0.59
```
Reliability of judges

- When raters (judges) rate targets, there are multiple sources of variance
  - Between targets
  - Between judges
  - Interaction of judges and targets

- The intraclass correlation is an analysis of variance decomposition of these components

- Different ICC’s depending upon what is important to consider
  - Absolute scores: each target gets just one judge, and judges differ
  - Relative scores: each judge rates multiple targets, and the mean for the judge is removed
  - Each judge rates multiple targets, judge and target effects removed
What is the reliability of ratings of different judges across ratees? It depends. Depends upon the pairing of judges, depends upon the targets. ICC does an Anova decomposition.
Sources of variances and the Intraclass Correlation Coefficient

Table: Sources of variances and the Intraclass Correlation Coefficient.

<table>
<thead>
<tr>
<th></th>
<th>(J1, J2)</th>
<th>(J3, J4)</th>
<th>(J5, J6)</th>
<th>(J1, J3)</th>
<th>(J1, J5)</th>
<th>(J1 ... J3)</th>
<th>(J1 ... J4)</th>
<th>(J1 ... J6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variance estimates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$MS_b$</td>
<td>7</td>
<td>15.75</td>
<td>15.75</td>
<td>7.0</td>
<td>5.2</td>
<td>10.50</td>
<td>21.88</td>
<td>22.86</td>
</tr>
<tr>
<td>$MS_w$</td>
<td>0</td>
<td>2.58</td>
<td>7.58</td>
<td>12.5</td>
<td>1.5</td>
<td>8.33</td>
<td>7.12</td>
<td>6.80</td>
</tr>
<tr>
<td>$MS_j$</td>
<td>0</td>
<td>6.75</td>
<td>36.75</td>
<td>75.0</td>
<td>0.0</td>
<td>50.00</td>
<td>38.38</td>
<td>24.36</td>
</tr>
<tr>
<td>$MS_e$</td>
<td>0</td>
<td>1.75</td>
<td>1.75</td>
<td>0.0</td>
<td>1.8</td>
<td>0.00</td>
<td>.88</td>
<td>.86</td>
</tr>
<tr>
<td><strong>Intraclass correlations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC(1,1)</td>
<td>1.00</td>
<td>.72</td>
<td>.35</td>
<td>-.28</td>
<td>.55</td>
<td>.08</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>ICC(2,1)</td>
<td>1.00</td>
<td>.73</td>
<td>.48</td>
<td>.22</td>
<td>.53</td>
<td>.30</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>ICC(3,1)</td>
<td>1.00</td>
<td>.80</td>
<td>.80</td>
<td>1.00</td>
<td>.49</td>
<td>1.00</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>ICC(1,k)</td>
<td>1.00</td>
<td>.84</td>
<td>.52</td>
<td>-.79</td>
<td>.71</td>
<td>.21</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>ICC(2,k)</td>
<td>1.00</td>
<td>.85</td>
<td>.65</td>
<td>.36</td>
<td>.69</td>
<td>.56</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>ICC(3,k)</td>
<td>1.00</td>
<td>.89</td>
<td>.89</td>
<td>1.00</td>
<td>.65</td>
<td>1.00</td>
<td>.96</td>
<td></td>
</tr>
</tbody>
</table>
ICC of judges

**ICC is done by calling anova**

```r
aov.x <- aov(values ~ subs + ind, data = x.df)
s.aov <- summary(aov.x)
stats <- matrix(unlist(s.aov), ncol = 3, byrow = TRUE)
MSB <- stats[3, 1]
MSW <- (stats[2, 2] + stats[2, 3])/(stats[1, 2] + stats[1, 3])
MSJ <- stats[3, 2]
MSE <- stats[3, 3]
ICC1 <- (MSB - MSW)/(MSB + (nj - 1) * MSW)
ICC2 <- (MSB - MSE)/(MSB + (nj - 1) * MSE + nj * (MSJ - MSE)/n.obs)
ICC3 <- (MSB - MSE)/(MSB + (nj - 1) * MSE)
ICC12 <- (MSB - MSW)/(MSB)
ICC22 <- (MSB - MSE)/(MSB + (MSJ - MSE)/n.obs)
ICC32 <- (MSB - MSE)/MSB
```
Intraclass Correlations using the ICC function

```r
> print(ICC(Ratings), all=TRUE) # get more output than normal

$results

<table>
<thead>
<tr>
<th>type</th>
<th>type</th>
<th>ICC</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>p</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single_raters_absolute</td>
<td>ICC1</td>
<td>0.32</td>
<td>3.84</td>
<td>5</td>
<td>30</td>
<td>0.04</td>
<td>0.04</td>
<td>0.79</td>
</tr>
<tr>
<td>Single_random_raters</td>
<td>ICC2</td>
<td>0.37</td>
<td>10.37</td>
<td>5</td>
<td>25</td>
<td>0.09</td>
<td>0.09</td>
<td>0.80</td>
</tr>
<tr>
<td>Single_fixed_raters</td>
<td>ICC3</td>
<td>0.61</td>
<td>10.37</td>
<td>5</td>
<td>25</td>
<td>0.28</td>
<td>0.28</td>
<td>0.91</td>
</tr>
<tr>
<td>Average_raters_absolute</td>
<td>ICC1k</td>
<td>0.74</td>
<td>3.84</td>
<td>5</td>
<td>30</td>
<td>0.21</td>
<td>0.21</td>
<td>0.96</td>
</tr>
<tr>
<td>Average_random_raters</td>
<td>ICC2k</td>
<td>0.78</td>
<td>10.37</td>
<td>5</td>
<td>25</td>
<td>0.38</td>
<td>0.38</td>
<td>0.96</td>
</tr>
<tr>
<td>Average_fixed_raters</td>
<td>ICC3k</td>
<td>0.90</td>
<td>10.37</td>
<td>5</td>
<td>25</td>
<td>0.70</td>
<td>0.70</td>
<td>0.98</td>
</tr>
</tbody>
</table>

$summary

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>subs</td>
<td>5</td>
<td>141.667</td>
<td>28.3333</td>
<td>10.366</td>
</tr>
<tr>
<td>ind</td>
<td>5</td>
<td>153.000</td>
<td>30.6000</td>
<td>11.195</td>
</tr>
<tr>
<td>Residuals</td>
<td>25</td>
<td>68.333</td>
<td>2.7333</td>
<td>NA</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ^O***~O 0.001 ^O**~O 0.01 ^O*~O 0.05 ^O.~O 0.1 ^O ~O 1

$stats

[,1]       [,2]       [,3]
[1,] 5.000000e+00 5.000000e+00 25.000000
[2,] 1.416667e+02 1.530000e+02 68.333333
[3,] 2.833333e+01 3.060000e+01 2.733333
[4,] 1.036585e+01 1.119512e+01 NA
[5,] 1.800581e+05 9.644359e-06 NA

$MSW
[1] 7.377778

$Call
ICC(x = Ratings)
Cohen’s kappa and weighted kappa

- When considering agreement in diagnostic categories, without numerical values, it is useful to consider the kappa coefficient.
  - Emphasizes matches of ratings
  - Doesn’t consider how far off disagreements are.

- Weighted kappa weights the off diagonal distance.

- Diagnostic categories: normal, neurotic, psychotic
Cohen kappa and weighted kappa

> cohen
  [,1] [,2] [,3]
[1,]  0.44  0.07  0.09
[2,]  0.05  0.20  0.05
[3,]  0.01  0.03  0.06
> cohen.weights
  [,1] [,2] [,3]
[1,]  0  1  3
[2,]  1  0  6
[3,]  3  6  0
> cohen.kappa(cohen, cohen.weights)
Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries

<table>
<thead>
<tr>
<th></th>
<th>lower estimate</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>unweighted kappa</td>
<td>-0.92</td>
<td>0.49</td>
</tr>
<tr>
<td>weighted kappa</td>
<td>-10.04</td>
<td>10.7</td>
</tr>
</tbody>
</table>

see the other examples in ?cohen.kappa