Statistics in Psychosocial Research
Lecture 4
Reliability II

Lecturer: Jeannie-Marie Leoutsakos
Outline

- Review of ANOVA
- Intra-Class Correlations
- Reliability Examples
- Other Research Designs
Are the true means for each group different from each other?

Compare amounts of variance within & between groups
\( i=1 \ldots, I \) indexes groups, \( j=1, \ldots n_i \) indexes members of group

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DF</th>
<th>Sum of Squares (SS)</th>
<th>Mean Square (MS)</th>
<th>F-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Group</td>
<td></td>
<td>( \sum n_i (\bar{Y}_i - \bar{Y})^2 )</td>
<td>( MSB = \frac{SSB}{DF} )</td>
<td>( \frac{MSB}{MSW} )</td>
</tr>
<tr>
<td>Within Group</td>
<td></td>
<td>( \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 )</td>
<td>( MSW = \frac{SSW}{DF} )</td>
<td></td>
</tr>
</tbody>
</table>
```plaintext
. oneway score1 group

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>4054.76741</td>
<td>2</td>
<td>2027.38371</td>
<td>2030.92</td>
<td>0.0000</td>
</tr>
<tr>
<td>Within groups</td>
<td>1494.39042</td>
<td>1497</td>
<td>.998256793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5549.15783</td>
<td>1499</td>
<td>3.70190649</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
. oneway score13 group

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>4145.64545</td>
<td>2</td>
<td>2072.82273</td>
<td>2.42</td>
<td>0.0891</td>
</tr>
<tr>
<td>Within groups</td>
<td>1281245.47</td>
<td>1497</td>
<td>855.8754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1285391.12</td>
<td>1499</td>
<td>857.499079</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Intraclass correlation: Assessing inter-rater reliability

• As before, reliability defined as:
  \[
  \frac{\text{variance in true scores}}{\text{variance in observed scores}}
  \]

• For the intra-class correlation the specific form of this equation can take on at least six different forms

• The correct form to use depends on the study design and the researcher’s assumptions about the patients and subjects (or items)

• I will discuss three designs, each with two ICCs
Overview

- **Unique Design:** Each of the $I$ subjects rated by a unique set of $m$ raters ($m>1$), such that the total number of raters, $R$, is $m*I$

- **Fixed Design:** Each subject is rated by each of the same $m$ raters, such that the total number of raters, $R$ is $m$. These raters are the only raters of interest.

- **Random Design:** $m$ raters are drawn from a larger pool of raters. Each of the $I$ subjects is rated by each of the $m$ raters. Again, the total number of raters, $R$ is $m$.

**NOTE:** raters might be people or questionnaire items
Unique Design

• No Overlap of Raters

\[ rater_1 \quad rater_2 \quad rater_3 \quad rater_4 \quad rater_5 \quad rater_6 \]

\[ s_1 \quad s_2 \quad s_3 \]

• \( m=2, \ I=3 \)  \# of raters=\( m*I=6 \)
Fixed Design

- Total Overlap of Raters

\[ \text{rater}_1 \quad \text{rater}_2 \quad \text{rater}_3 \]

\[ \text{s}_1 \quad \text{s}_2 \quad \text{s}_3 \]

- \( m=3, \ n=3 \quad \# \text{ of raters}=m=3 \)
Random Design

- Total Overlap of Raters, but raters drawn from a pool.
Types of Reliability

• There are two (at least) types of reliability associated with each of these designs.
  – Reliability of mean ratings: reliability of average of all ratings per subject
  – Reliability of one individual rating: reliability of a single rating of one subject

• Which will be higher?
• Why?
Unique Rater Design ICC

Equation to estimate reliability of rating means

\[
\frac{\text{Between Mean Square Variance} - \text{Within Mean Square Variance}}{\text{Between Mean Square Variance}}
\]

\[
\frac{\text{MSB} - \text{MSW}}{\text{MSB}}
\]
Between Mean Score Variance (Each TA is a group): Observed mean variance
Between Mean Score Variance

Degree to which mean score of rated subjects differ from grand mean

\[ S_b^2 = \frac{1}{I - 1} \left( m \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 \right) \approx \sigma_b^2 \]

- \( I \) = number of people being rated (# of TAs)
- \( \bar{Y}_i \) = mean score for each TA rated
- \( \bar{Y} \) = overall mean of scores for whole sample
- \( m \) = number of raters for each mean
Unique Rater Design

1) Between Mean Score Variance, steps in Stata
   a) calculate mean scores for each individual
      *e.g. egen meanta = rmean(score1 score2 score3)
   b) calculate overall mean
      *e.g. egen grandmean = mean(meanta)
   c) calculate deviation of individual mean from group mean
      *e.g. gen bsquarei = 3*(meanta-grandmean)^2
   d) add up all deviations in (c)
      *e.g. egen bsquare = sum(bsquarei)
   e) divide sum of squares by degrees of freedom
      *e.g. display bsquare/(10-1) =
Unique Rater Design (cont’d)

2) Within Mean Score Variance: Degree to which individual scores differ from a subject’s mean score

\[ S_w^2 = \frac{1}{I (m - 1)} \sum_{i=1}^{I} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_i)^2 \approx \sigma_w^2 \]

Where:
- \( I = \) number of individuals being rated (# of TAs)
- \( R = \) number of raters
- \( Y_{ij} = \) score of each individual rater
- \( \bar{Y}_i = \) mean score of each person rated
- \( m = \) number of raters for each mean

Note: \( R = m \)
Unique Rater Design

3) Within Mean Score Variance, steps in Stata
   a) calculate mean scores for each individual
      *e.g. egen meanid = rmean(score1 score2 score3)
   b) calculate deviation of rater from individual mean
      *e.g. gen wsquarei = (score1-meanid)^2 + (score2-meanid)^2 + (score3-meanid)^2
   c) add up deviations in (b) across all individuals
      *e.g. egen wsquare =sum(wsquarei)
   d) divide sum of squares by degrees of freedom
      *e.g. display wsquare/I*(m-1) =
Unique Rater Design

Shortcut: Use procedure ‘oneway’ in Stata

First, must “reshape” data.

```
. reshape long score, i(ta) j(rater)
```
Using ANOVA in STATA to Calculate Variance

Example:

```
.oneway score ta
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>114.00</td>
<td>9</td>
<td>12.67</td>
</tr>
<tr>
<td>Within groups</td>
<td>30.00</td>
<td>20</td>
<td>1.50</td>
</tr>
<tr>
<td>Total</td>
<td>144.00</td>
<td>29</td>
<td>4.97</td>
</tr>
</tbody>
</table>

\[ ICC = \frac{MSB - MSW}{MSB} \]

\[ ICC = \frac{12.67 - 1.50}{12.67} = .8816 \]
Important Note

1. Reliability is a group-specific statistic.

2. The greater the variance in the true scores of a population, the higher the reliability of the measure (if observed variance is constant)

\[
\text{Reliability} = \frac{\text{Variance in true scores}}{\text{Variance in observed scores}}
\]
Reliability for Individual Ratings

So far we’ve calculated reliability of the mean score for each TA.

What is the average reliability of each individual rating of the TA?
Reliability of Individual Scores in Unique Rater Design

\[
\frac{MSB - MSW}{MSB + (m - 1) MSW}
\]

Where \( m \) = number of raters per TA

Continuing with our example:

\[
\text{Reliability} = \frac{(12.67 - 1.50)}{12.67 + (3 - 1) \times 1.50} = .7128
\]
Fixed Rater Design

1) Each subject rated by each of the same m raters, who are the only raters of interest

2) examples:

3) Computation involves two-way analysis of variance

4) Before: two sources of error, (differences across individuals, and error inherent to the measurement) Error now only has one source: error due to individuals is ‘controlled.’
Fixed Rater Design

Recall that the equation for Unique Rater Design was:

\[
\frac{MSB - MSW}{MSB}
\]

Which can also be expressed as:

\[
\frac{MSB - (MSRater + MSE)}{MSB}
\]

The equation for the fixed rater design is very similar:

\[
\frac{MSB - (MSE)}{MSB}
\]
Rater Mean Variance

![Graph showing the means of different raters]

- **Rater Mean Variance**
- **Graph** showing the mean values for different raters.
Fixed Rater Design

Rater Mean Score Variance: Degree to which raters’ mean scores differ from those of the overall mean

\[
s_{r}^2 = \frac{1}{(m - 1)} (I) \sum_{j=1}^{m} (\overline{Y}_j - \overline{Y})^2 \approx \sigma_{r}^2
\]

Where:

- \( m \) = number of raters (in fixed design, R=m)
- \( I \) = number of subjects evaluated (# of TAs)
- \( \overline{Y}_j \) = mean score of rater
- \( \overline{Y} \) = overall mean mean score for sample
Fixed Rater Design

Steps in Stata

1. Calculate overall mean

2. Calculate mean for each rater
   
   *e.g.* egen r1mean=mean(rater1)
   
   egen r2mean=mean(rater2)...

3. Calculate deviation of rater mean from overall mean
   
   *e.g.* display N*(r1mean-grandmean)^2 + N*(r2mean-grandmean)^2...

4. Calculate error square variance
   
   *error square variance = (within square variance – rater square variance)*

   *divide by difference in degrees of freedom to get error variance*
Using ANOVA in STATA to Calculate Variance

Example:

```
. anova score ta Rater
```

<table>
<thead>
<tr>
<th>Source</th>
<th>Partial SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>134.00</td>
<td>11</td>
<td>12.1818182</td>
</tr>
<tr>
<td>ta</td>
<td>114.00</td>
<td>9</td>
<td>12.6666667</td>
</tr>
<tr>
<td>Rater</td>
<td>20.00</td>
<td>2</td>
<td>10.00</td>
</tr>
<tr>
<td>Residual</td>
<td>10.00</td>
<td>18</td>
<td>.555555556</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>----</td>
<td>-----------</td>
</tr>
<tr>
<td>Total</td>
<td>144.00</td>
<td>29</td>
<td>4.96551724</td>
</tr>
</tbody>
</table>

ICC for Fixed Rater Design, Group mean =

\[
\frac{MSB - MSE}{MSB} = \frac{12.67 - .56}{12.67} = .96
\]
Fixed Rater Design

Equation to estimate reliability for individual rater’s scores:

\[
\frac{MSB - MSE}{MSB + (m-1)*MSE}
\]

Where \( R = m = \) number of raters

Final Estimate:

\[
\frac{12.67 - .56}{12.67 + (2)(.56)} = .8782
\]
Random Rater Design

1. Randomly-selected raters evaluate each subject

2. Computation involves two-way analysis of variance

3. Error has two sources again, but error due to individual raters is reduced

4. Deciding between Random and Fixed design:

   Would you wish to generalize findings from this sample to situations with a different set of raters? If so, you would use the random rater design.
Random Rater Design: Reliability for Mean Score of Each Subject

\[
\frac{MSB - MSE}{MSB + ((MSRater - MSE) / I)}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>Partial SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>134.00</td>
<td>11</td>
<td>12.1818182</td>
</tr>
<tr>
<td>ta</td>
<td>114.00</td>
<td>9</td>
<td>12.6666667</td>
</tr>
<tr>
<td>Rater</td>
<td>20.00</td>
<td>2</td>
<td>10.00</td>
</tr>
<tr>
<td>Residual</td>
<td>10.00</td>
<td>18</td>
<td>.55555556</td>
</tr>
<tr>
<td>Total</td>
<td>144.00</td>
<td>29</td>
<td>4.96551724</td>
</tr>
</tbody>
</table>

2) Take into account error for rater bias

3) \( ICC = \frac{12.67 - 0.56}{12.67 + (10 - 0.56)/10} = .89 \)
Random Rater Design: Reliability for Individual Score

Source | Partial SS | df | MS
-------+------------+----+-----+
Model  | 134.00     | 11 | 12.1818182
Ta     | 114.00     | 9  | 12.6666667
Rater  | 20.00      | 2  | 10.00
Residual | 10.00  | 18 | .555555556
-------+------------+----+-----+
Total  | 144.00     | 29 | 4.96551724

$$\frac{MSB - MSE}{MSB + (m - 1) \cdot MSE + m \cdot (MSRater - MSE) / I}$$

$$ICC = \frac{12.67 - 0.56}{12.67 + (2 \times .56) + 3 \times (10.0 - .56)/10} = .72$$
Summary

1. **Unique Rater Design**: Each subject rated by a different set of \( m \) raters; formulas use between and within mean square variance.

2. **Fixed Rater Design**: Each target is rated by each of the same \( m \) raters, who are the only raters of interest; formulas use between and error square variance.

3. **Random Rater Design**: \( m \) raters, in (2), were drawn from a random sample of raters; formula uses between and error square variance, adjusting for rater variance.
Which ICC Is Most Appropriate?

Scenario 1: A target child’s three best friends all report on the target child’s level of drug use.

Scenario 2: You develop a screener to help identify victims of domestic abuse in emergency rooms; each patient is to be rated by three nurses at each hospital and you use the mean score in your analyses.

a) Which ICC would give you the estimated reliability for the nurses at your one pilot hospital?

b) Which ICC would give you an estimate of the reliability for the measure when used by different nurses at different hospitals?

c) Which ICC would give you an estimate for the reliability of the measure if it were to be administered by only one nurse instead of three?
Question 1

State the conditions under which the Unique Rater ICC (for mean values of an item) is identical to the value of the Fixed Rater ICC (for mean values of an item)?

Answer in terms of the variance of between, within, and rater sum of squares.
Solution 1

\[
\frac{MSB - (MSRater + MSE)}{MSB} = \frac{MSB - MSE}{MSB}
\]

\[
MSE = MSRater + MSE
\]

\[
MSRater = 0
\]
Question 2

You have developed a new survey measure of bipolar disorder on the basis of a pilot population composed of one third with severe symptoms, one third with mild symptoms, and one third without any symptoms. It turns out that your measure has a high reliability of .90. You find funding and administer your survey to a nationally representative sample, only to find that your reliability is now much lower.

What might be the explanation?
Solution 2

\[ .90 = \frac{MSB_{\text{pilot}} - MSW}{MSB_{\text{pilot}}} \]

There is inherent assumption here that the national sample will have the same makeup with regard to severity. If that’s not so, then the reliability may drop because the between-person variance in the national sample was lower than it was in the pilot sample, while the within-person variance was presumably about the same.
Question 3

Doesn’t high reliability imply that two measures of the same characteristic will yield the same answer? If so, why do I see graphs that imply higher reliability when sample variability is higher?
It is important to keep in mind that there are two types of variance: within-person variance and between-person variance. It is correct that when the within-person variance is high a measure typically will have low reliability. The within-person variance is a measure of the error variance, and the higher the error variance of a measure the lower its reliability. With high levels of within-person variance, measures of the same characteristic on multiple occasions will lead to different answers.

In contrast, high levels of between-person variance help improve the reliability of a measure. The more between-person variance in a population, the greater the proportion of variance that is due to the true underlying characteristic in proportion to the variance due to error, and the greater the overall reliability.
Imagine two graphs, Figure 1 and Figure 2, in which all respondents have the same mean score. If Figure 2 shows a wider spread in individual means than is shown in Figure 1, which of the two graphs has the higher reliability.
Solution 4

Figure 1 has the higher reliability. The between-person variance in both graphs is the same (all respondents have the same mean score). The within-person variance is higher in Figure 2 than it is in Figure 1 (indicated by a wider spread across the individual means). Therefore, Figure 1 has higher reliability.
Question 5

Using observed score as the y axis and true score as the x axis, draw a measure with a negative covariance between true score and error term.
Solution 5
Question 6

Using observed score as the y axis and true score as the x axis, draw a measure with a positive covariance between true score and error term.
Solution 6

Observed Score (Positive Correlation) vs True Score
Question 7

Scenario 1: Reported correlation between years of educational attainment and adults’ scores on an anti-social personality (ASP) disorder scale is about .30; reported reliability of the education scale is about .95; reported reliability for the ASP scale it is about .70.

Scenario 2: Reported reliability of the education scale is the same (.95); reported reliability of the ASP disorder scale is now .40.

What is the observed correlation between the two measures in Scenario 2?
Solution 7

\[ r_{TxTy} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} = \frac{.30}{\sqrt{.95 \times .70}} = .367883 \]

solve for \( r_{xy} \)

\[ r_{xy} = r_{TxTy} \times \sqrt{r_{xx}r_{yy}} = .367883 \times \sqrt{.95 \times .40} = .227 \]
Question 8

A. How are the alpha and the split-half reliability coefficient conceptually related?

B. For mean scores, how are the alpha and the Fixed Rater ICC related?
Solution 8

A. Cronbach’s alpha is the average of all possible split-half reliabilities.

B. Cronbach’s alpha is mathematically equivalent to the Fixed Rater ICC for mean scores.
Question 9

For a ten-item scale with an average inter-item correlation of .25, the reliability is .75. What about a twenty-item scale with the same average inter-item correlation? How about fifteen items? How about 5?
Solution 9

Use Spearman-Brown Prophesy Formula

Reliability New Scale = N(R)/(1+(N-1)R)
N = (number of desired items)/(number of items in observed scale)
R = reliability of observed scale

For 20 item scale: R = (2*.75)/(1+.75) = .86
For 15 item scale: R = (1.5*.75)/(1+.5*.75) = .82
For 5 item scale: R = (.5*.75)/(1-.5*.75) = .60
Question 10

Two psychiatrists disagree when rating a dichotomous child health outcome among 100 children. In ten of the cases, Dr. Green rated the outcome as present when Dr. Brown rated it as absent. In another ten cases, the reverse occurred; Dr. Brown rated the outcome as present while Dr. Green rated it as absent.

If both Dr. Green and Dr. Brown agree that fifty children have the outcome, what will be the value of the Kappa coefficient?

If they agree that 70 children have the outcome, will the Kappa be higher or lower?
Solution 10

\[ \begin{array}{|c|c|c|c|}\hline & + & - & \hline + & 50 & 10 & 60 \hline - & 10 & 30 & 40 \hline & 60 & 40 & 100 \hline \end{array} \]

Proportion of observed agreement = \( \frac{80}{100} = .8 \)

\[ \begin{array}{|c|c|c|c|}\hline & + & - & \hline + & 60*60/100=36 & 60*40/100=24 & 60 \hline - & 60*40/100=24 & 40*40/100=16 & 40 \hline & 60 & 40 & 100 \hline \end{array} \]

Proportion of expected agreement = \( \frac{36+16}{100} = .52 \)

\[ \kappa = \frac{prop_{obs} - prop_{ex}}{1 - prop_{ex}} = \frac{.8 - .52}{1 - .52} = .58 \]
Question 11

Measures of self-reported discrimination sometimes violate the assumptions of classical test theory. Please provide a substantive example of violation for each of the three assumptions.
Solution 11

\( E(x) = 0 \) could be violated if the true score is underreported as a result of social desirability bias

\( \text{Cov}(T_x,e) = 0 \) could be violated if people systematically overreported or underreported discrimination at either high or low extremes of the measure

\( \text{Cov}(e_i,e_j) = 0 \) could be violated if discrimination was clustered within certain areas of a location, and multiple locations were included in the analysis pool.
Question 12

An 10-item ASP measure with a reliability of .6 and an HIV risk-behavior measure with a reliability of .5 correlate at .30. How many additional items with similar item-level reliability must be added to the ASP measure to make the observed correlation \( \geq .35 \)?
Solution 12

1. Solve for true correlation

2. The true correlation is constant; therefore, $r_{xx}$ (and/or $r_{yy}$) must get bigger to raise the observed correlation.

$$r_{xx} = \left( \frac{r_{xy(\text{obs})}}{r_{TxTy}} \right)^2 \left( \frac{.35}{.547723} \right)^2 \div \frac{.5}{.5} = .817$$
Solution 12 (cont’d)

3. Determine how many items to add by using Spearman-Brown prophesy formula.

\[ R_{des} = \frac{NR_{obs}}{1-(N-1)R_{obs}} \]

4. Solve for \( N \)

\[ N = \frac{R_{des} + R_{obs}R_{des}}{R_{obs} + R_{des}R_{obs}} = \frac{.817 + (.817* .6)}{.6 + (.817* .6)} = 1.199 \approx 1.2 \]

\[ N = \frac{\text{\# items}_{des}}{\text{\# items}_{obs}} \]

\[ \text{\# items}_{des} = N \times \text{\# items}_{obs} = 1.2 \times 10 = 12 \]

Solution: The ASP scale must have \( \geq 12 \) items for an expected observed correlation with HIV risk-behavior of .35 or greater.
Other Research Designs

- We saw, with the fixed ICC, how we could partition the variance, and reduce MSE

```
. anova score ta
Number of obs =  30
Root MSE = 1.22474
          Adj
Source    Partial SS  df  MS    
Model     114          9 12.6666667
          114          9 12.6666667
Residual  30          20  1.5
Total     144         29 4.96551724
```

```
. anova score ta rater
Number of obs =  30
Root MSE = .745356
          Adj
Source    Partial SS  df  MS
Model     134         11 12.1818182
          114          9 12.6666667
rater     10          2  10
          10          18 .555555556
Residual  144         29 4.96551724
Total     144         29 4.96551724
```
Fixed Effects

(a) Set by experimenter (eg, treatment in an RCT)

(b) it is unreasonable to generalize beyond conditions. (eg, reading ability as a function of grade in school)

(c) when the # of possibilities is small, and all are included in the study design (eg, sex, in a study with both males and females)
Random Effects

(a) Multiple possible values (e.g., personality measures, age).

(b) Study subjects are considered a representative sample from a larger population.

(c) Experimenter wishes to generalize the results of the study beyond the study sample.
• We already saw an example of this with the fixed and random ICC’s.

• Part of a larger group of study designs under the heading of “generalizability theory” popularized by Cronbach, and others.

• Can take 140.655 (LDA) and/or 140.656 (Multilevel models)