Interrater Reliability

- Each subject assessed by multiple raters

- To what extent are the ratings within a subject homogeneous?

- Ideally, want raters to be interchangeable
## Decayed, Missing, Filled Teeth

<table>
<thead>
<tr>
<th>Patient</th>
<th>Examiner 1</th>
<th>Examiner 2</th>
<th>Examiner 3</th>
<th>Examiner 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>7</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>13</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>23</td>
<td>27</td>
<td>18</td>
</tr>
</tbody>
</table>

J.L. Fleiss *The Design and Analysis of Clinical Experiments*
Intraclass Correlation Coefficient

For continuous data, ICC often used to assess interrater reliability.

ICC is the correlation between two measurements made on the same subject.

\[
ICC = \text{Corr}(Y_{ij}, Y_{ik})
\]
ICC Properties

- Like any correlation, $-1 \leq ICC \leq 1$
- Ideally, ICC will be $\geq 0$
- ICC Values close to 1 are desirable and indicate good interrater reliability
There are numerous versions of the intraclass coefficient (ICC)... Each form is appropriate for specific situations defined by the experimental design...

Randomized Block Design

Dental example: 4 different raters randomly selected to rate each patient

Patients represent random sample of all possible patients

\[ Y_{ij} = \mu + \beta_i + \varepsilon_{ij} \]

\( \beta_i \sim N(0, \sigma_\beta^2) \) random subject effect

\( \varepsilon_{ij} \sim N(0, \sigma^2) \) experimental error
ICC for Randomized Blocks

\[ \text{Cov}(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) \]
\[ = \mu^2 + E(\beta_i^2) - \mu^2 \]
\[ = \sigma_{\beta}^2 \]

\[ \text{Var}(Y_{ij}) = \text{Var}(Y_{ij}) = \sigma_{\beta}^2 + \sigma^2 \]

\[ \therefore \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_{\beta}^2}{\sigma_{\beta}^2 + \sigma^2} = \text{ICC} \]
Estimating the ICC: Method 1

PROC GLM

1. Output mean squares to dataset

2. Use mean squares to calculate estimates of $\sigma^2_\beta$ and $\sigma^2$

3. Use the estimates of $\sigma^2_\beta$ and $\sigma^2$ to calculate the ICC
Estimating the ICC: More Methods

PROC MIXED
1. Output estimates of variance components (part of standard output) to a dataset
2. Use the estimates to calculate ICC

PROC NLMIXED
1. Calculate ICC within the procedure in a single step

%INTRACC macro
1. No programming to do!
Decayed, Missing, Filled Teeth

Start with PROC MIXED + PROC SQL Approach

ods output CovParms = cov1;
proc mixed data=dental method=ml;
  class patient;
  model score = ;
  random patient;
run;

REML not available in NLMIXED. For comparison purposes, I will use ML, and then repeat with REML
Decayed, Missing, Filled Teeth

The Mixed Procedure

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate (ML)</th>
<th>Estimate (REML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient</td>
<td>41.8611</td>
<td>50.5528</td>
</tr>
<tr>
<td>Residual</td>
<td>6.3889</td>
<td>6.3889</td>
</tr>
</tbody>
</table>
Decayed, Missing, Filled Teeth

```
proc sql;
  create table icc as
  select sum(estimate*(covparm='patient')) / sum(estimate) as icc
  from covl;
quit;
proc print data = icc ; run ;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>icc (ML)</th>
<th>icc (REML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.86759</td>
<td>0.88780</td>
</tr>
</tbody>
</table>

Good reliability in both cases
PROC NLMIXED Approach

PROC NLMIXED data=dental method=firo ;
parms mu=10  s_subj = 50  s_err = 6 ;
pred = mu + beta ;
model score ~ normal(pred, s_err) ;
random beta ~ normal(0, s_subj) 
subject = patient ;
estimate 'icc'
  s_subj/(s_subj+s_err);
run ;

Same estimate as PROC MIXED using ML

Label Estimate
icc  0.8676
Don’t want to do the programming?

- There is a user-written macro that can be downloaded from the SAS website
  - [http://support.sas.com/kb/25/031.html](http://support.sas.com/kb/25/031.html)

- The macro is called `%INTRACC`

- Computes 6 different versions of the ICC
%INTRACC and the Dental Data

Macro invocation:

```plaintext
%intracc(depvar=score, target=patient, rater=rater, nrater=4)
```

Documentation states that macro computes 6 different versions of ICC

Output includes 9 versions
### %INTRACC Output

<table>
<thead>
<tr>
<th></th>
<th>Winer reliability: mean of single score</th>
<th>Winer reliability: mean of k scores</th>
<th>Shrout-Fleiss reliability: mean of 4 scores</th>
<th>Shrout-Fleiss reliability: single score rand set</th>
<th>Shrout-Fleiss reliability: single score fxd set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.88780</td>
<td>0.96937</td>
<td>0.96937</td>
<td>0.88780</td>
<td>0.88958</td>
</tr>
<tr>
<td>Shrout-Fleiss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reliability:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fixed set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rel: rand set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean k scores</td>
<td>0.94996</td>
<td>0.96937</td>
<td>0.96990</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean k scrs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shrout-Fleiss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rel: fxd set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean k scrs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Same ICC value as PROC MIXED using REML
Dental Data Using PROC GLM

First, output the mean squares to a dataset.

Next, examine expected mean squares to determine relationship between MS’s and variance components, $\sigma^2_\beta$ and $\sigma^2$. 
ods output ModelANOVAm=ms1
   (where=(hypothesistype=3)) ;
ods output OverallANOVAm=ms0
   (where=(source='Error')) ;
proc glm data = dental ;
   class patient ;
   model score = patient ;
   random patient ;
run ;
The GLM Procedure

Source | Type III Expected Mean Square
patient | \( \text{Var(Error)} + 4 \text{ Var(patient)} \)

So, use MS(error) to estimate \( \sigma^2 \)

Use \( [\text{MS(patients)} - \text{MS(error)}]/4 \) to estimate \( \sigma^2_\beta \)
GLM Continued Again

```sql
proc sql;
  create table icc as
  select a.ms as s_err, (b.ms-a.ms)/4 as s_sub,
    ((b.ms-a.ms)/4)/(a.ms+(b.ms-a.ms)/4)
  as icc
  from ms0 a, ms1 b;
quit;
proc print data = icc ; run ;
```

Same estimates as PROC MIXED with REML

<table>
<thead>
<tr>
<th>Obs</th>
<th>s_err</th>
<th>s_subj</th>
<th>icc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.388889</td>
<td>50.5528</td>
<td>0.88780</td>
</tr>
</tbody>
</table>
Comparison for RCB Design

- GLM is cumbersome!
- %INTRACC is too much of a black box for me
- NLMIXED is simple, which I like, but no option to do REML
- MIXED seems to me like best compromise
- What about more complicated designs?
Random Rater Effect

Suppose same 4 raters assess each subject, but that the 4 raters randomly selected from larger pop’n of raters

\[ Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \]

\( \beta_i \sim N(0, \sigma^2_{\beta}) \) random subject effect

\( \tau_j \sim N(0, \sigma^2_{\tau}) \) random rater effect

\( \varepsilon_{ij} \sim N(0, \sigma^2) \) experimental error
ICC for Random Rater Effect

\[ \text{Cov}(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) \]
\[ = \mu^2 + E(\beta_i^2) - \mu^2 \]
\[ = \sigma^2_\beta \]

\[ \text{Var}(Y_{ij}) = \text{Var}(Y_{ij}) = \sigma^2_\beta + \sigma^2_\tau + \sigma^2 \]

\[ \therefore \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_\tau + \sigma^2} = ICC \]
NL MIXED and GLM

Oops!! NLMIXED can’t handle “crossed” random effects. Maybe in an upcoming version?

Here are the expected mean squares from PROC GLM:

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient</td>
<td>$\text{Var(Error)} + 4 \text{Var(patient)}$</td>
</tr>
<tr>
<td>rater</td>
<td>$\text{Var(Error)} + 6 \text{Var(rater)}$</td>
</tr>
</tbody>
</table>

I’m too lazy to do the calculations so I’ll skip ahead to PROC MIXED
ods output CovParms = cov2;
proc mixed data = dental method = REML;
  class patient rater;
  model score = ;
  random patient rater;
run;

proc sql;
create table icc2 as
  select sum(estimate*(covparm='patient'))/
       sum(estimate) as icc
  from cov2;
quit;

Obs    icc
  1  0.88958
%INTRACC Output

<table>
<thead>
<tr>
<th>Winer</th>
<th>Winer</th>
<th>Shrout-Fleiss</th>
<th>Shrout-Fleiss</th>
</tr>
</thead>
<tbody>
<tr>
<td>reliability</td>
<td>reliability: mean of</td>
<td>reliability: mean of</td>
<td>reliability: mean of</td>
</tr>
<tr>
<td>reliability:</td>
<td>single score k scores</td>
<td>4 scores</td>
<td>single score random set</td>
</tr>
<tr>
<td>0.88780</td>
<td>0.96937</td>
<td>0.96937</td>
<td>0.88780</td>
</tr>
</tbody>
</table>

Shrout-Fleiss reliability: rel: rand set rel: fxd set
fixed set mean k scores mean k scr mean k scr
0.94996 0.96937 0.96990 0.98700

- Same macro call as earlier & same output
- Same value as MIXED with REML
Raters from 2 Different Specialties

Suppose that 1\textsuperscript{st} two raters are from one dental specialty \& 2\textsuperscript{nd} two raters are from another

Only two specialities of interest

\therefore speciality is a fixed effect

2 raters picked at random from each specialty

Rater effect expected to differ by specialty
Model for Different Specialties

\[ Y_{ijk} = \mu + \beta_i + \tau_{j(k)} + \gamma_k + \epsilon_{ij(k)} \]

\[ \beta_i \sim N(0, \sigma^2_\beta) \] random subject effect

\[ \tau_{j(1)} \sim N(0, \sigma^2_{\tau_1}) \] random rater effect for 1\textsuperscript{st} specialty

\[ \tau_{j(2)} \sim N(0, \sigma^2_{\tau_2}) \] random rater effect for 2\textsuperscript{nd} specialty

\[ \gamma_k \] fixed specialty effect

\[ \epsilon_{ij} \sim N(0, \sigma^2) \] experimental error
ICC for this Model

Same subject, same specialty:

\[ ICC = \frac{\sigma^2_{\beta}}{\sigma^2_{\beta} + \sigma^2_{\tau k} + \sigma^2} \]

Same subject, different specialties:

\[ ICC = \frac{\sigma^2_{\beta}}{\sqrt{\sigma^2_{\beta} + \sigma^2_{\tau 1} + \sigma^2} \times \sqrt{\sigma^2_{\beta} + \sigma^2_{\tau 2} + \sigma^2}} \]
Challenge

Expressions for Model and ICC are easy to write out, but how to fit the model and estimate ICC?

“group=specialty” option in MIXED varies *all* parameters by specialty instead of just the rater variance.

NLMIXED would allow this if there was only one random effect, but there are 2, which it can’t handle.