Overview

- Review of Classical Test Theory
- Standard Error of Measurement (SEM)
- Comparing SEM to Reliability
- Reliability vs. Agreement
- Agreement

Review of Classical Test Theory

Classical Test Theory – A refresher!

If an immortal person took test an infinite number of times, what would be the mean of the observed scores?

\[ X = T + E, \]

T is the same every time you took it
E is random (sometimes +, sometimes -)
and would average out to zero, so…
average \[ X = T \]
Standard Error of Measurement (SEM)

- ...but what would the standard deviation be?...

The Normal Curve of Error

$T = 10$

Review of Classical Test Theory

The Mysteries of SEM …Revealed!

$\text{standard error of measurement} = \text{SEM}$

standard deviation of the error around any individual's true score

±2 SEM captures 95% of the error

...you’d like SEM/error to be small

Why?

How?
Large SEM

The CAT (Creative Analogies Test) has 50 items. Most people score anywhere from 20 to 100.

Amy scores 75 on the CAT. 
*What if SEM = 10?*
Then…
• 95% sure her true score is 55 – 95 (±2 SEM)

Small SEM

The CAT (Creative Analogies Test) has 50 items. Most people score anywhere from 20 to 100.

Amy scores 75 on the CAT. 
*What if SEM = 2?*
Then…
• 95% sure her true score is 71 – 79 (±2 SEM)

It’s SEMtacular!

\[ SEM = s_x \sqrt{1 - r_{xx}} \]

- \( s_x \) = SD of test scores
- \( r_{xx} \) = test reliability

- good reliability \( \rightarrow \) low SEM \( SEM = 10 \sqrt{1 - .84} = 4 \)
- poor reliability \( \rightarrow \) high SEM \( SEM = 10 \sqrt{1 - .19} = 9 \)
SEM Example

Say you have a test where 
mean = 50, \( s_x = 10 \) and \( r_{xx} = .84 \)

\[
SEM = 10 \sqrt{1-.84} = 4
\]

Fred scores 35 on this test, which means

• 95% sure his true score is 27 – 43 (±2 SEM)
• 68% sure his true score is 36 – 44 (±1 SEM)

SEM Assumptions

\[
SEM = s_x \sqrt{1-r_{xx}}
\]

• a reliability coefficient based on an appropriate measure
• the sample appropriately represents the population

Use of Cut Scores

• Cut scores are set values on a test that determines who passes and who fails the test.
  – Commonly used for licensure or certification (bar exam, medical licensure, civil service)

• What is the impact of the SEM on interpreting cut scores?
Some general rules about reliability

• Bigger is better!
  – The more items, the more occasions, and the more observers the better the reliability
• The Spearman-Brown Prophesy formula gives us the reliability of a test if the number of (similar) items is increased
• Need to compute the standard error of measurement to understand individual test scores
• A measure can not be valid if it isn’t reliable

Reliability vs. Agreement

Example 1 - Interrater Agreement

*Example*
Brutus and Maureen are clinical psychologists. They independently conduct intake interviews for 200 clients presenting mood problems.
(Ex. 1, cont.) Assigned Categories

Clients assigned to 1 of 3 categories
- Cyclothymic
- Bipolar
- Depressed

Why would you think/hope the therapists agree?

(Ex. 1, cont.) Assignments

<table>
<thead>
<tr>
<th></th>
<th>Maureen</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cyclo</td>
<td>bipo</td>
</tr>
<tr>
<td>Brutus</td>
<td>cyclo</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>bipo</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>depr</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>130</td>
</tr>
</tbody>
</table>

(Ex. 1, cont.) Convert # to Proportions

<table>
<thead>
<tr>
<th></th>
<th>Maureen</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cyclo</td>
<td>bipo</td>
</tr>
<tr>
<td>Brutus</td>
<td>cyclo</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>bipo</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>depr</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>.65</td>
</tr>
</tbody>
</table>

.53+.14+.03 = 70% interrater agreement
Example 2: 
*Brutus and Maureen: Bogus Clinicians*

What if Brutus and Maureen were actors, just pretending to be clinical psychologists as a sick little drama class project?

In other words, what if Brutus and Maureen had NO IDEA WHAT THEY WERE DOING, and all their "diagnoses" were COMPLETELY DUE TO CHANCE?

---

**(Ex. 2, cont.)** Frequencies

<table>
<thead>
<tr>
<th></th>
<th>Maureen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cyclo</td>
</tr>
<tr>
<td>Brutus</td>
<td>22</td>
</tr>
<tr>
<td>cyclo</td>
<td>23</td>
</tr>
<tr>
<td>bipo</td>
<td>22</td>
</tr>
<tr>
<td>depr</td>
<td>67</td>
</tr>
</tbody>
</table>

---

**(Ex. 2, cont.)** Proportions

<table>
<thead>
<tr>
<th>Proportions (=cell freq/200)</th>
<th>Maureen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cyclo</td>
</tr>
<tr>
<td>Brutus</td>
<td>.11</td>
</tr>
<tr>
<td>cyclo</td>
<td>.11</td>
</tr>
<tr>
<td>bipo</td>
<td>.11</td>
</tr>
<tr>
<td>depr</td>
<td>.33</td>
</tr>
</tbody>
</table>

.11+.11+.11 = 33 % interrater agreement just by chance
Wait a sec…

Two people could agree just by chance, right? (e.g., 2 people flippin’ coins)

(1) By how much do ratings beat chance agreement?

(2) What’s the MOST the ratings could beat chance by?

---

Kappa Coefficient

How Much Above Chance Agreement?

\[ \kappa = \frac{P_{\text{obs}} - P_{\text{chance}}}{1 - P_{\text{chance}}} \]

The amount above chance you do have agreement on.

\[ \text{The maximum amount beyond chance you could have agreement on.} \]

---

**Proportions (\(=\text{cell freq}/200\))**

<table>
<thead>
<tr>
<th></th>
<th>Maureen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cyclo</td>
</tr>
<tr>
<td>Brutus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>.33</td>
</tr>
</tbody>
</table>

.11+.11+.11 = 33 % interrater agreement *just by chance*
Example 3: What if a pregnancy test was claimed to be "98% effective"?

Say you had the following data on 1000 women...

<table>
<thead>
<tr>
<th>Pregnancy Test</th>
<th>preg</th>
<th>no preg</th>
</tr>
</thead>
<tbody>
<tr>
<td>preg</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>no preg</td>
<td>20</td>
<td>960</td>
</tr>
</tbody>
</table>

98% effectiveness

\[ \kappa = \left( \frac{P_{\text{obs}} - P_{\text{chance}}}{1 - P_{\text{chance}}} \right) \]

\[ \kappa = \left( \frac{.98 - .94}{1 - .94} \right) = \frac{.04}{.06} = .67 \]
What if a pregnancy test was claimed to be “98% effective”?

In these data, most (96%) were not pregnant anyway, so the test would have been correct 96% of the time even if the test always said the individual was not pregnant!

The pregnancy test does help above chance levels, but – given the sample data – not as much as the claim might lead one to believe.