Test Scores, Norms, and Score Interpretation

What We Will Cover in This Section.

- Test scores
- Norms
- Types of interpretation

Raw Score

Count or sum of the number of items endorsed or number of items correctly answered.
Norms

Norm
Behavior that is usual or typical for members of a group.

Norms
Reference scores against which an individual’s scores are compared.

Norming
Process of establishing norms for a test.

Norman
A fat guy who drinks beer in Cheers.

Establishing Norms

1. Target population.
2. Standardization sample.
   - Size.
   - Geographic representation.
   - Socioeconomic level.
   - Ethnic composition.
Types of Norms

Reference Groups

- National norms.
- Regional norms.
- Local norms.
- Sub-group norms.
### Table 2. Word, raw score: Total Sample

<table>
<thead>
<tr>
<th>Valid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>5.5</td>
<td>5.9</td>
<td>6.5</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>14.6</td>
<td>15.7</td>
<td>22.2</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>18.9</td>
<td>20.3</td>
<td>42.5</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>24.4</td>
<td>26.1</td>
<td>68.6</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>7.3</td>
<td>7.8</td>
<td>76.5</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>11.6</td>
<td>12.4</td>
<td>88.9</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>9.1</td>
<td>9.8</td>
<td>98.7</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>.6</td>
<td>.7</td>
<td>99.3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>.6</td>
<td>.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>153</td>
<td>93.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Missing System</td>
<td>11</td>
<td>6.7</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>164</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6. Word, raw score: Undergraduate Students

<table>
<thead>
<tr>
<th>Valid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>8.1</td>
<td>8.1</td>
<td>8.1</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>37.8</td>
<td>37.8</td>
<td>45.9</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>29.7</td>
<td>29.7</td>
<td>75.7</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>18.9</td>
<td>18.9</td>
<td>94.6</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2.7</td>
<td>2.7</td>
<td>97.3</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2.7</td>
<td>2.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

### Table 9. Word, raw score: Graduate Students

<table>
<thead>
<tr>
<th>Valid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>.9</td>
<td>.9</td>
<td>.9</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5.2</td>
<td>5.2</td>
<td>6.0</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>8.6</td>
<td>8.6</td>
<td>14.7</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>17.2</td>
<td>17.2</td>
<td>31.9</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>28.4</td>
<td>28.4</td>
<td>60.3</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>10.3</td>
<td>10.3</td>
<td>70.7</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>15.5</td>
<td>15.5</td>
<td>86.2</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>12.1</td>
<td>12.1</td>
<td>98.3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>.9</td>
<td>.9</td>
<td>99.1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>.9</td>
<td>.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Developmental: Age Norms

• Uses.
  – Initially for establishing ‘mental age’.
  – BASAL AGE highest age at and below which all tests are passed.

• Issues.
  – Mental age units are not constant with chronological age.
  – Rapid growth spurts.
  – Not an interval scale.
  – Cannot be interpreted as overall performance.

Developmental: Grade Norms

• Uses.
  – Educational achievement tests.
  – May be expressed in years and months.

• Issues.
  – Instructional content varies from school to school.
  – Not really relevant at high school level.
  – Interpretation.

Developmental: Ordinal Scales

• Uses.
  – Behavioral and cognitive development.
  – Piaget.
  – Gesell.

• Issues.
  – Assume one level is prerequisite to the next.
Major Issue

1. How do you compare performance of different people across different forms of a test?
2. How to compare performance of people on the same test across different administration techniques?

Equating

Process of establishing the relationship between scores on alternate forms of a test or the same test across time.

Equating: Same Sample Technique

1. Get a large representative sample of examinees.
2. Give half the sample form A and the other half form B.
3. Compute the mean and standard deviation for each form.
4. Compute the z-scores for the raw scores for each test.
5. For ease of interpretation transform the z-scores to some common transformed score.
Equating: Different Sample Technique

1. Identify a sample of key questions from the earlier test version.
2. Insert these items in the current version of the test.
3. Compare the performance on the key questions across the two groups.

National Anchor Norms

Uses
- Compare performance on different tests.

Issues.
- Does not imply that the tests are interchangeable.
- Different psychometric properties.

Within Group Comparisons
Percentile

Percent of people in the standardization sample who fall at or below a given score.

Percentiles: Benefits & Issues

- **Benefits.**
  - Easy to compute.
  - Easy to understand.
  - Suitable to all tests.
  - Widely used.
  - Good if the distribution is normal.

- **Issue.**
  - Inequality of units at extremes of the normal curve.
  - Problem if the distribution is not normal.
Quartiles

Divides the distribution into four groups each containing 25% of the sample.

Standard Scores

A score that has been converted from one scale into another scale.
Concept

*z* scores

\[ z = \frac{X - M}{S} \]

Example

Mean = 50  SD = 10

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Z for 65</th>
<th>Z for 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>( \frac{65 - 50}{10} ) = 1.5</td>
<td>( \frac{65 - 50}{10} ) = 1.5</td>
</tr>
<tr>
<td>50</td>
<td>( \frac{50 - 50}{10} ) = 0</td>
<td>( \frac{50 - 50}{10} ) = 0</td>
</tr>
<tr>
<td>45</td>
<td>( \frac{45 - 50}{10} ) = -0.5</td>
<td>( \frac{45 - 50}{10} ) = -0.5</td>
</tr>
</tbody>
</table>
Where they are

Standard Scores: Benefits

- Statistically meaningful.
  - Communicates in terms of the normal curve.
  - Direct reference to the mean.
- Translate directly to percentiles.
- Comparison within same test.
- Comparison across different tests.

Standard Scores: Issues

- May appear to be artificial.
- Assumes a normal distribution.
- Hard to communicate to lay people.
Transformed Scores (T scores)

A score that has been converted to a scale with a mean of 50 and a standard deviation of 10.

Z Score to t-Score

\[ X_{\text{New}} = (Z_{\text{Old}} \times SD_{\text{New}}) + \overline{X}_{\text{New}} \]

Mean = 50    SD = 10

<table>
<thead>
<tr>
<th>What is the T-score for a z-score of 1.5?</th>
<th>What is the T-score for a z-score of -1?</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ = (1.5 x 10) + 50</td>
<td>IQ = (-1 x 10) + 50</td>
</tr>
<tr>
<td>T = 15 + 50</td>
<td>T = -10 + 50</td>
</tr>
<tr>
<td>T = 65</td>
<td>T = 40</td>
</tr>
</tbody>
</table>

Other Standard Scores.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>SAT and GRE</td>
<td>500</td>
<td>100</td>
</tr>
</tbody>
</table>
Stanines

*Divides the distribution into nine groups. The width of the interval between adjacent groups is ½ a standard deviation.*

Criterion Referenced Testing

*Has the person mastered a pre-established level of competence on some key content domain.*
Criterion Referenced Testing: Uses & Issues

**Uses.**
- Focus on what a person knows.
- Assess minimal standards.

**Issues.**
- Establishing the content domain.
- How to establish a score to indicate mastery?
- Useful for basic skills and knowledge.

Actuarial vs. Clinical Interpretation

**Actuarial**
Interpreting test scores in terms of their statistical properties (testing).

**Clinical**
Interpreting test scores in terms of their meaning (assessment).

THE END
AVA Standardization Sample

**Assertiveness (V-1) Raw Scores**

- **Std. Dev.** = 4.93
- **Mean** = 8.0
- **N** = 3629.00

**AVA Standardization Sample**

**Conformity (V-4) Raw Scores**

- **Std. Dev.** = 4.22
- **Mean** = 4.7
- **N** = 3629.00