E-411-PRMA

Lecture 6

Christopher David Desjardins

3 September 2015

Standard Error Measurement

$$\sigma_{SEM} = \sigma\sqrt{1 - r_{xx}}$$

- standard error of measurement = standard deviation of test scores * square root of 1 - reliability coefficient of the test

Standard Error Measurement

$$\sigma_{SEM} = \sigma\sqrt{1 - r_{xx}}$$

- standard error of measurement = standard deviation of test scores * square root of 1 - reliability coefficient of the test
- Can use this to create confidence intervals by using normality assumption of an individual's score on a large number of tests centered at the mean
- Determines the range of plausible values for a person's true score

SEM example

A math test is administered. The test scores have a reliability of 0.80 and a standard deviation of 0.5

What is the standard error of measurement?

If Anna scored a 7.5, what range of values can we be 95% confident that her true score lies between? 99% confident?

Standard Error of the difference between two scores

$$\sigma_D = \sqrt{\sigma_{SEM_1} + \sigma_{SEM_2}}$$

$$\sigma_D = \sigma\sqrt{2 - r_1 - r_2}$$

- Can be used to compare two individuals on the same test or a different test
- Can be used to compare performance of an individual on two tests

SED example

Sigrun takes the same test as Anna and scores a 6.5. Did Anna perform significantly better on the test?

If Anna took a second test and got a score of 8 and the reliability coefficient for the second test was 0.6, did Anna do significantly better on the second test?

Validity

Validity

- What is validity?
  - An indicator of how well the test measures the latent construct(s) it claims to.
  - A determination of the appropriateness of the test scores for specific uses/users
  - Validity of the test for a given purpose, at a given time, for a given population
  - You are a lawyer presenting evidence to a judge to make the case for the validity of your instrument - validation
  - Users can conduct a validation study to assess the validity of the instrument for their purposes

Overview of Validity

- Content - Evaluation of the subjects, topic, or content covered by the items in the test
- Criterion-Related - Evaluating the relationship of scores obtained on the test to scores on other tests or measures
- Construct - Evaluation of how scores obtained on the test relate to scores on other instruments AND understanding how the test scores fit within the theoretical framework of the latent construct that the test purports to measure

Content Validity

- How adequately the test represents the latent construct of interest
- Do the items throughly and completely tap into the latent construct?
- Content valid test would have percentage of items on each topics to be proportional to the amount of time spent on these topics
- How can we be sure I am teaching the entire domain of psychological testing?
- Create a test blueprint
  - What could be conceivably measured and in what proportion
  - Number of questions, types of questions, areas covered, organization, etc

Assessing Content Validity

- Assume you are giving an instrument to measure aggressive behavior in children
- How can we assume this is measuring the construct of aggression quantitatively?
  - Experts assess whether each item is essential, useful, or not necessary to the definition of aggression
  - $CVR = \frac{n_e - (N/2)}{N/2}$
  - Where n_e is number say "essential" and N is number of experts
  - Want this larger than chance (Table 6-1)

CVR in toy example

- "Does your child bite other children?"
- 20 experts, 17 say "essential"

Letting R do the work

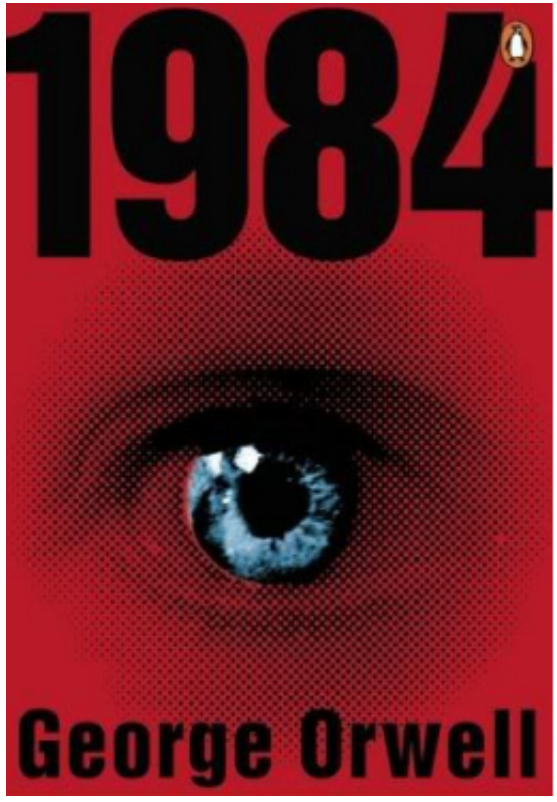
```
CVR <- function(n, essential){
  (essential - n/2)/(n/2)
}
CVR(n = 20, essential = 17)

## [1] 0.7
```

Assessing Content Validity

BUT ... expert judgement!!!

"Who controls the past controls the future; who controls the present controls the past."



source

Criterion-Related Validity

- What the test score tells you about where a person falls on the underlying construct being measured w.r.t a criterion
- A criterion is a benchmark or standard used for comparison
- Scores on a new IQ instrument, *but do you really know that high scores mean high IQ*?
  - Should be relevant
    - People that are known to have high IQs (maybe MENSA membership) should score highly on this instrument
  - Should be valid for measuring IQ
    - Who created this instrument?
    - Does it correlate with established IQ instruments (e.g. WAIS or Stanford-Binet)?

Criterion Problems

- Predict whether someone is receiving counseling services based on Beck Depression Inventory
  - Find out BDI was used to determine whether someone should receive services

Criterion Problems

- Predict whether someone is receiving counseling services based on Beck Depression Inventory
  - Find out BDI was used to determine whether someone should receive services
- In addition, to self-report and parent report, you ask teachers to rate students on externalizing behaviors
  - After all the students' scores have been calculated, ask teachers to comment on them

Criterion Problems

- Predict whether someone is receiving counseling services based on Beck Depression Inventory
  - Find out BDI was used to determine whether someone should receive services
- In addition, to self-report and parent report, you ask teachers to rate students on externalizing behaviors
  - After all the students' scores have been calculated, ask teachers to comment on them
- What is wrong with this?

Concurrent Validity

- Concurrent Validity
  - Test scores are obtained at the *same time* as the criterion measures are obtained
  - Measures of the relationship between the test and the criterion are concurrent validity evidence
  - Example?

Concurrent Validity

- Concurrent Validity
  - Test scores are obtained at the *same time* as the criterion measures are obtained
  - Measures of the relationship between the test and the criterion are concurrent validity evidence
  - Example?
- If test scores (test new) correlate with a test (test old) that has already been validated to measure the criterion, then test old can be used as a validating criterion
- When might you do this?

Predictive Validity

- Predictive Validity
  - Test scores are obtained *before* the criterion measures are obtained
  - How accurately does the test scores predict the criterion measures
    - SAT measures "college readiness"
    - What could be our future criterion?
    - What relationship would we expect between the scores and this criterion?
    - Could we use dropout (i.e. student attrition)?

Validity Coefficient

- Correlation between test scores and scores on the criterion-measure
  - Correlation between sdcores on the SAT and GPA at the end of Freshman year (criterion-measure)

Validity Coefficient

- Correlation between test scores and scores on the criterion-measure
  - Correlation between sdcores on the SAT and GPA at the end of Freshman year (criterion-measure)
- Validity coefficient affected everything a correlation is
- Range restriction from attrition in a study or self-selection
- Testtakers need to be relevant in the validation study and cover the scope of the test
- Read the test manual and make sure test is appropriate for your testtakers
  - Does their validity study map well to your target population and purpose?
- Coefficient should be high enough to matter

Incremental Validity

- Refers to the degree to which an additional predictor explains the criterion measure above and beyond that already explained by those predictors already included
- Requirement: each predictor (obviously?) must have predictive validity
  - Let predict final grade in students in a statistics course
  - We have several variables to choose from:

```
##  [1] "SECTION"  "GENDER"   "ETHDESCR" "CUM_GPA"  "CUMCREDS" "ACT_TOTL"
##  [7] "ACT_ENGL" "ACT_MATH" "ACT_READ" "ACT_SCIR" "HSPR"     "LTRGRADE"
## [13] "STATGRAD" "DEVSTDNT"
```

- What should we do?

- For simplicity, let's just look only at the continuous variables
- Correlations ...

```
##      SECTION      CUM_GPA     CUMCREDS     ACT_TOTL     ACT_ENGL
## -0.009021592  0.491283854  0.250867602  0.252158752  0.188233466
##     ACT_MATH     ACT_READ     ACT_SCIR         HSPR     STATGRAD
##  0.293462052  0.167622795  0.167987726  0.239023226  1.000000000
```

- For simplicity, let's just look only at the continuous variables
- Correlations ...

```
##      SECTION      CUM_GPA     CUMCREDS     ACT_TOTL     ACT_ENGL
## -0.009021592  0.491283854  0.250867602  0.252158752  0.188233466
##     ACT_MATH     ACT_READ     ACT_SCIR         HSPR     STATGRAD
##  0.293462052  0.167622795  0.167987726  0.239023226  1.000000000
```

- Which variable would you think is the strongest predictor of statistics grade?

- For simplicity, let's just look only at the continuous variables
- Correlations ...

```
##      SECTION      CUM_GPA     CUMCREDS     ACT_TOTL     ACT_ENGL
## -0.009021592  0.491283854  0.250867602  0.252158752  0.188233466
##     ACT_MATH     ACT_READ     ACT_SCIR         HSPR     STATGRAD
##  0.293462052  0.167622795  0.167987726  0.239023226  1.000000000
```

- Which variable would you think is the strongest predictor of statistics grade?
- Which variables might have incremental validity?

- For simplicity, let's just look only at the continuous variables
- Correlations ...

```
##      SECTION      CUM_GPA     CUMCREDS     ACT_TOTL     ACT_ENGL
## -0.009021592  0.491283854  0.250867602  0.252158752  0.188233466
##     ACT_MATH     ACT_READ     ACT_SCIR         HSPR     STATGRAD
##  0.293462052  0.167622795  0.167987726  0.239023226  1.000000000
```

- Which variable would you think is the strongest predictor of statistics grade?
- Which variables might have incremental validity?
- LET'S DO THIS TOGETHER!

Expectancy tables

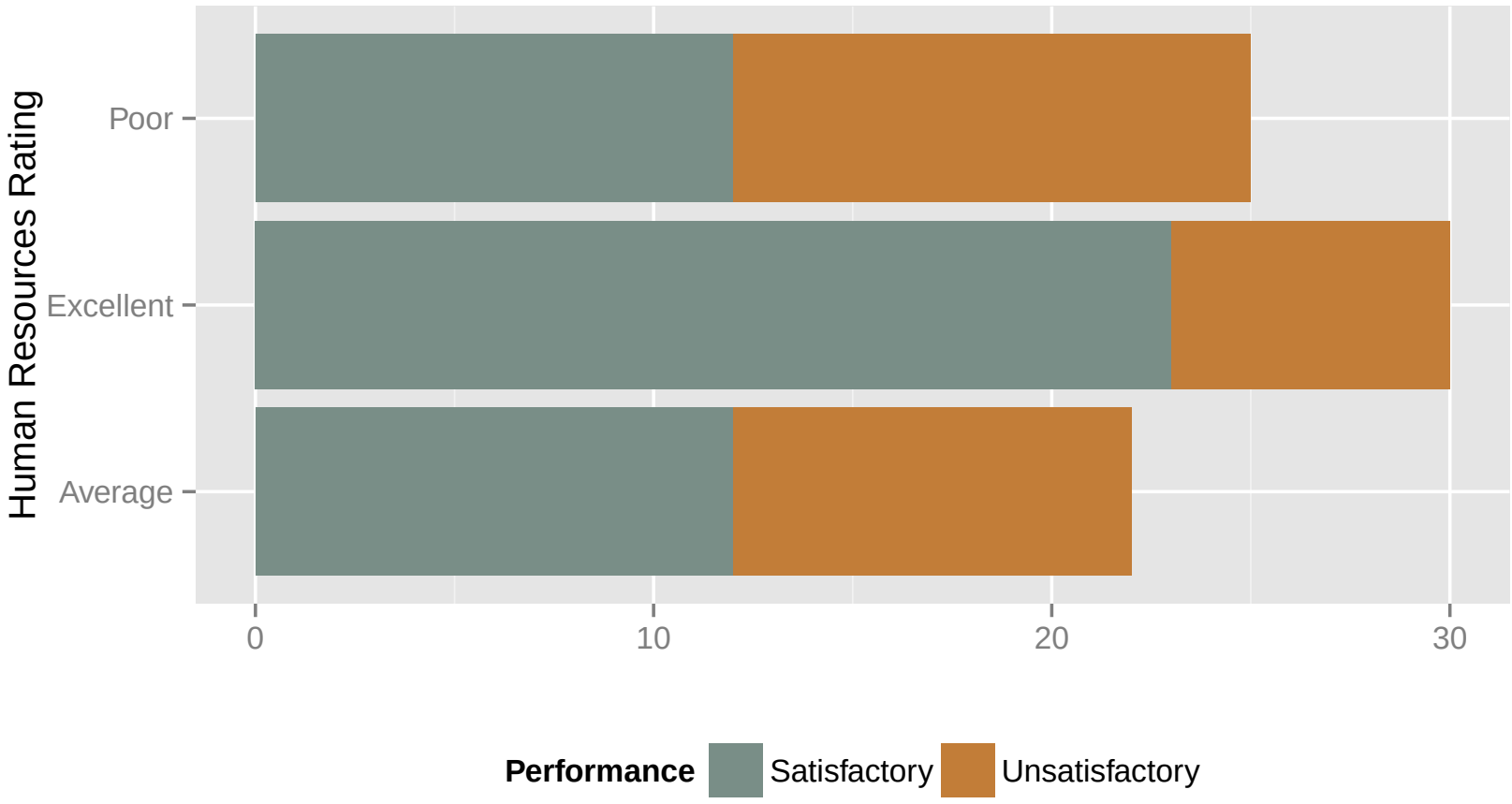
- Visualization tool
- Test scores (or applicant/client ratings) are obtained and placed into some interval (e.g. "excellent", "good", "ok", "bad", "miserable")
- You are discretizing your measure only for visualization
- Criterion measures obtained later (e.g. proficent in math or job performance)
- Create a chart that shows relationship betwen test scores and criterion measure
  - Essentially a contigency table

Expectancy tables

- Visualization tool
- Test scores (or applicant/client ratings) are obtained and placed into some interval (e.g. "excellent", "good", "ok", "bad", "miserable")
- You are discretizing your measure only for visualization
- Criterion measures obtained later (e.g. proficent in math or job performance)
- Create a chart that shows relationship betwen test scores and criterion measure
  - Essentially a contigency table
- A major omission from your book - we need to check and see if this is larger than chance alone!

HR ratings and Job Performance

	Satisfactory	Unsatisfactory
Excellent	23	7
Average	12	10
Poor	12	13



```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 5.2582, df = 2, p-value = 0.07214
```

- H_0: There is no association between HR rating and job performance

```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 5.2582, df = 2, p-value = 0.07214
```

- H_0: There is no association between HR rating and job performance
- Probably need to intervene with HR!

Construct Validity

- Evidence supporting that the test *measures* the underlying construct and that it is capable of *placing* test takers along that latent construct
- A test maker MUST have theories about the construct, it's definition, structure, and relationship to other constructs and has theories about how their test relates to other tests
- If the test fails to discern test takers, need to know why
  - Recall all the various potential sources of error in testing
- All forms of validity could be considered subsets of construct validity

Construct Validity Evidence

- Homogeniety
  - Structure of a test should be homogeneous if it is measuring a single construct
  - Responses to test items should be positively correlated with total score on the test
    - What kind of correlation is this?
    - Items that are not ... need to be removed or rewritten
    - What to do with items that have low correlations?
    - What does it mean to throw away items and rewrite them?
  - Homogeneity implies inter item agreement ... how can we measure this?

Construct Validity Evidence

- Homogeniety
  - Structure of a test should be homogeneous if it is measuring a single construct
  - Responses to test items should be positively correlated with total score on the test
    - What kind of correlation is this?
    - Items that are not ... need to be removed or rewritten
    - What to do with items that have low correlations?
    - What does it mean to throw away items and rewrite them?
  - Homogeneity implies inter item agreement ... how can we measure this?
- Change with age and pre/post
  - Testtakers taking a test in reading *should* score higher on comprehension if they are older
  - Students getting tutored in reading between a pre and post test should score higher on the post test
  - Should we be able to predict how anxiety will change as we get older?

Construct Validity Evidence - contd

- Groups higher on the measured construct should have higher scores (method of contrasted groups)
  - Administer a test measuring tendency toward violent behavior
  - Who should have higher scores: The general public or prison inamtes for assault and battery?

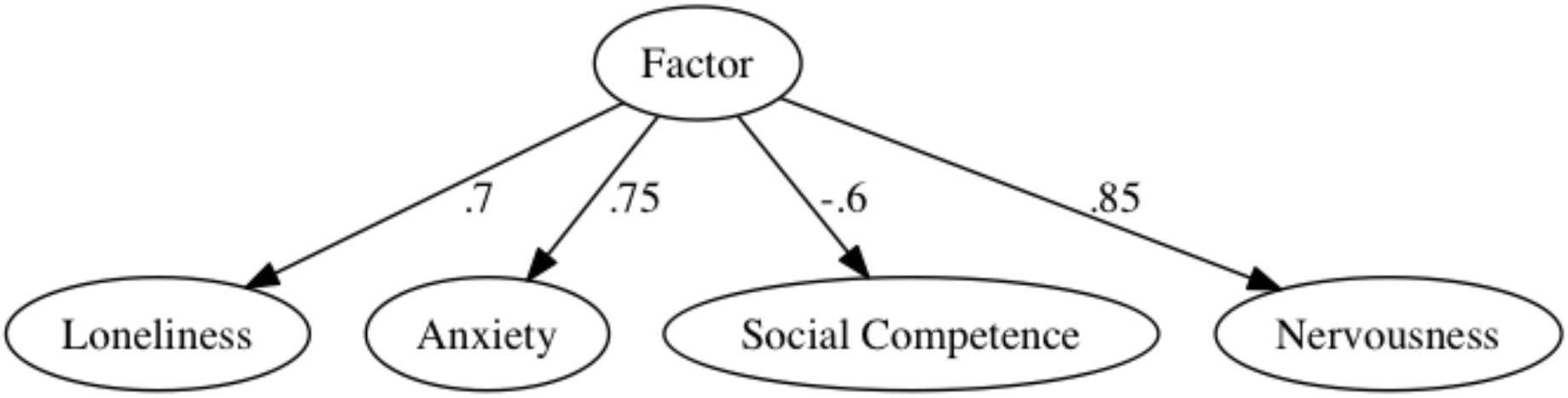
Construct Validity Evidence - contd

- Groups higher on the measured construct should have higher scores (method of contrasted groups)
  - Administer a test measuring tendency toward violent behavior
  - Who should have higher scores: The general public or prison inamtes for assault and battery?
- Convergent - Test takers IQ scores on a new test should be correlated with their IQ score from an established and validated IQ tests (or a related construct)

Construct Validity Evidence - contd

- Groups higher on the measured construct should have higher scores (method of contrasted groups)
  - Administer a test measuring tendency toward violent behavior
  - Who should have higher scores: The general public or prison inamtes for assault and battery?
- Convergent - Test takers IQ scores on a new test should be correlated with their IQ score from an established and validated IQ tests (or a related construct)
- Discriminant - Test scores should be unrelated to scores from another instrument
  - Ask students to score each other on leadership
  - Ask students to score each other on popularity
  - What does it mean if these two are uncorrelated?

Factor Analysis



- What should we call this factor?
- If Nervousness is our new instrument to measure the factor, how well does it do?
- What does it mean that social competence is negatively correlated with our factor?

Test Bias and Fairness

- Test bias - degree to which a test systematically favors one group or another
  - Can test for this statistically using logistic regression model
  - Known as differential item functioning
  - Errors by raters - too lightly, too severely, to the middle, too perfectly
- Test fairness - the degree to which a test is fair and used in an equitable way
  - What if we administer a test to a group not involved in the validation sample
  - Maybe some groups of people are just different?
- Why do we care about bias and fairness?