E-411-PRMA

Lecture 4

Christopher David Desjardins

27 August 2015

This week

- Classical test theory
- Reliability

CTT model

$$\underbrace{X}_{\text{observed score}} = \underbrace{T}_{\text{true score}} + \underbrace{E}_{\text{error}}$$

Sources of measurement error

- Random Error

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error
- Examples of each?

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error
- Examples of each?
- Which poses a bigger threat to a

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error
- Examples of each?
- Which poses a bigger threat to a
  - Consistent measure?

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error
- Examples of each?
- Which poses a bigger threat to a
  - Consistent measure?
  - Validity?

Sources of measurement error

- Random Error
  - Unpredictable and inconsistent sources of error
- Systematic Error
  - Constant and predictable source of error
- Examples of each?
- Which poses a bigger threat to a
  - Consistent measure?
  - Validity?
  - Reliability?

- Construction
- Administration
- Scoring
- Interpretation

Reliability def'n

$$\text{reliability} = \frac{\overbrace{\sigma_T^2}^{\text{true score variance}}}{\underbrace{\sigma_X^2}_{\text{observed score variance}}}$$

Types of Reliability

- Test-Retest
  - Coefficient of stability
  - What are sources of error here?

Types of Reliability

- Test-Retest
  - Coefficient of stability
  - What are sources of error here?
- Parallel Forms
  - Means and variances of the test scores are equivalent
  - Coefficient of equivalence
  - Parallel forms reliability
  - What are sources of error here?

Types of Reliability

- Test-Retest
  - Coefficient of stability
  - What are sources of error here?
- Parallel Forms
  - Means and variances of the test scores are equivalent
  - Coefficient of equivalence
  - Parallel forms reliability
  - What are sources of error here?
- Alternate Forms
  - Versions of a test designed to be parallel
  - Not necessarily parallel
  - Alternate forms reliability

Types of Reliability

- Test-Retest
  - Coefficient of stability
  - What are sources of error here?
- Parallel Forms
  - Means and variances of the test scores are equivalent
  - Coefficient of equivalence
  - Parallel forms reliability
  - What are sources of error here?
- Alternate Forms
  - Versions of a test designed to be parallel
  - Not necessarily parallel
  - Alternate forms reliability
- Fortunately, we can calculate measures of internal consistency

Different types of associations

- Pearson's product-moment correlation only appropriate when variables are continuous and are interval/ratio scales
- Alternatives when variables are dichotomous either naturally or artifically (assumed to have a continuous underlying scale)
  - Phi coefficient, equivalent to Pearson's correlation but for dichotomous variables
  - Polychoric coefficient, an index of association between two artifically ordinal variables
  - Tetrachoric coefficient, an index of association between two artifically dichotomized variables
  - Point-biserial coefficient, an index of association betweeen a dichotomous and a continous variable
  - Biserial coefficient, an index of association betweeen an artificially dichotomous and a continous variable
  - Spearman Rank-Order coefficient, an index of association where at least one variable is ordinal
  - Kendall's tau, alternative to Spearman

Internal consistency

- Measure level of consistency or agreement between items

Internal consistency

- Measure level of consistency or agreement between items
- A test is unidimensional if all the items measure the same latent construct

Internal consistency

- Measure level of consistency or agreement between items
- A test is unidimensional if all the items measure the same latent construct
- The more items measure just one construct, the higher the internal consistency

Internal consistency

- Measure level of consistency or agreement between items
- A test is unidimensional if all the items measure the same latent construct
- The more items measure just one construct, the higher the internal consistency
- Is it always possible or desirable to have a test that measures just one thing?

Split-Half Reliability

- Obtained by correlating (Pearson's coefficient) two pairs of scores from equivalent halves of a single test then apply a correction
- Creating two equivalent forms of a test
  - What are the steps to calculate a split-half reliability?
  - How might we consider making splits?
  - What do we need to be careful of?

Why a correction?

- Uncorrected reliability is biased downward

Why a correction?

- Uncorrected reliability is biased downward
  - Measurement error
  - Range restriction

Why a correction?

- Uncorrected reliability is biased downward
  - Measurement error
  - Range restriction
- Which should have higher reliability?
  - Test A, a math test consisting of math problems and word problems, or Test B, a math test consisting of just math problems?
  - Test C, which is 25 items long, or Test D, which is 50 items long?
  - Test E, which is a timed test, or Test F, which is the same test as E but you can take as long as you need on the test?

Spearman-Brown correction for split half

Arbitrary number of splits

$$r_{SB} = \frac{nr_{hh}}{1 + (n-1)r_{hh}}$$

Split-half

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

Spearman-Brown correction for split half

Arbitrary number of splits

$$r_{SB} = \frac{nr_{hh}}{1 + (n - 1)r_{hh}}$$

Split-half

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

- where r_{hh} is the correlation of the two halves.

Spearman-Brown correction for split half

Arbitrary number of splits

$$r_{SB} = \frac{nr_{hh}}{1 + (n-1)r_{hh}}$$

Split-half

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

- where r_{hh} is the correlation of the two halves.
- If we have a desired reliability, we can use the following formula to find out how much we have to increase the test by.

Spearman-Brown correction for split half

Arbitrary number of splits

$$r_{SB} = \frac{nr_{hh}}{1 + (n-1)r_{hh}}$$

Split-half

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

- where r_{hh} is the correlation of the two halves.
- If we have a desired reliability, we can use the following formula to find out how much we have to increase the test by.

$$N = \frac{r_{\text{desired}}(1 - r_{hh})}{r_{hh}(1 - r_{\text{desired}})}$$

Spearman-Brown correction for split half

Arbitrary number of splits

$$r_{SB} = \frac{nr_{hh}}{1 + (n-1)r_{hh}}$$

Split-half

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

- where r_{hh} is the correlation of the two halves.
- If we have a desired reliability, we can use the following formula to find out how much we have to increase the test by.

$$N = \frac{r_{\text{desired}}(1 - r_{hh})}{r_{hh}(1 - r_{\text{desired}})}$$

- What assumptions are we making about these new items?

Working with Spearman-Brown

Consider the following scenarios:

- What is the reliability of a test when the Pearson's correlation between two halves of a test is 0.6?

Working with Spearman-Brown

Consider the following scenarios:

- What is the reliability of a test when the Pearson's correlation between two halves of a test is 0.6?
- A test is 40 items long. The items have been split in half, total scores on each half have been calculated, and the correlation is 0.5. How long should the test be to have a reliability of 0.9?

```
# Create Spearman-Brown correction
sb <- function(r_hh){
  2 * r_hh / (1 + r_hh)
}

# Run the function
sb(0.6)

## [1] 0.75

# Find the new test length
new_length <- function(r_sb, r_hh, n){
  ceiling(r_sb * (1 - r_hh) / (r_hh * (1 - r_sb)) * n)
}

# Run the function
new_length(0.9, 0.5, 40)

## [1] 361
```

More on internal consistency measures

- Want high correlations among items (inter-item consistency)
- Higher the inter-item consistency, higher the homogeniety of the test (i.e. unidimensionality)
- Heterogeneity is desired when measuring a multifaceted psychological variables
  - Examples?
- Kuder-Richardson 20
  - Statistic of choice for dichotomous items reliability
  - If a test is heterogenous, K-R 20 will have lower reliability than a split-half
- Coefficient alpha
  - Mean of all possible split-half correlations
  - Appropriate for nondichotomous variables

Formulas

KR-20

$$r_{kr20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

Coefficient alpha (Cronbach's alpha)

$$r_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

- k, is the number of items
- pq and σ_i^2 are the product of the proportion answering an item correctly (p) and incorrectly (q) and the variance of a nondichotomous items, respectively.
- σ^2 is the variance of the total test scores

Formulas

KR-20

$$r_{kr20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

Coefficient alpha (Cronbach's alpha)

$$r_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

- k, is the number of items
- pq and σ_i^2 are the product of the proportion answering an item correctly (p) and incorrectly (q) and the variance of a nondichotomous items, respectively.
- σ^2 is the variance of the total test scores
- Is bigger always better?

Formulas

KR-20

$$r_{kr20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

Coefficient alpha (Cronbach's alpha)

$$r_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

- k, is the number of items
- pq and σ_i^2 are the product of the proportion answering an item correctly (p) and incorrectly (q) and the variance of a nondichotomous items, respectively.
- σ^2 is the variance of the total test scores
- Is bigger always better?
- What is too small?

Formulas

KR-20

$$r_{kr20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

Coefficient alpha (Cronbach's alpha)

$$r_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

- k, is the number of items
- pq and σ_i^2 are the product of the proportion answering an item correctly (p) and incorrectly (q) and the variance of a nondichotomous items, respectively.
- σ^2 is the variance of the total test scores
- Is bigger always better?
- What is too small?
- This is an abused statistic!
- Consider reporting 95% confidence intervals

By Hand

What is the KR-20 for this toy example?

Item 1	Item 2	Item 3
0	1	0
1	1	0
1	1	1

What is the Coefficient alpha for this toy example?

Item 1	Item 2	Item 3
4	3	4
4	3	3
5	5	5

LSAT

From the R description in the irtoys package:

> *The LSAT is a classical example in educational testing for measuring ability traits. This test was designed to measure a single latent ability scale.*

This is on 1000 subjects and 5 questions

This single latent ability should be what?

```r
lsat <- read.csv("http://cddesja.github.io/classes/e411prma2015-1/lecture3/data/lsat.csv")
kr20 <- function(data){
  p <- colMeans(data)
  q <- 1 - colMeans(data)
  num <- sum(p * q)
  denom <- var(rowSums(data))
  k <- ncol(data)
  k / (k - 1) * (1 - num / denom)
}
kr20(lsat)
```

```
## [1] 0.2959522
```

```r
coef_alpha <- function(data){
  num <- sum(apply(data, 2, var))
  denom <- var(rowSums(data))
  k <- ncol(data)
  k / (k - 1) * (1 - num / denom)
}
coef_alpha(lsat)
```

```
## [1] 0.2949972
```

```r
# 95% confidence interval
cocron::cronbach.alpha.CI(coef_alpha(lsat), n = nrow(lsat), items = 5)
```

```
## lower.bound upper.bound
##   0.2234738   0.3618025
```

Average proportional distance

- Focuses on **differences** not **similiarity** between items
- The APD method evaluates internal consistency by looking at the difference between test scores
- It works by:
  1. Calculating the absolute difference between scores for all the items
  2. Averaging the difference between scores
  3. Dividing by number of response options on the test minus one
- APD less than .2 excellent internal consistency
- Not effected by length of the test

Inter-rater reliability

- Two raters measure the same behavior
  - For example: Number of aggressive behaviors observed in a child during play time.
  - Degree to which these raters report the same incidence of aggressive behaviors is a measure of reliablity
- Correlate scores from raters (e.g. Pearson's or Spearman's rho, etc)
- Important thing to note: test scores have reliability NOT test

IRR example

Two parents are administered the CBCL (an instrument to identify problem behaviors in children) on their four children. How well do their scores for the section *Aggressive Behavior* agree (i.e. what is their inter-parent reliability)?

Child	Parent 1	Parent 2
1	5.5	6.0
2	5.2	5.2
3	4.6	4.0
4	6.6	5.6

Make sure you understand Table 5-4!

Test affects on reliability

- More homogeneous, higher reliability
- More static the characteristic, higher reliability
- Restriction range, lower reliability
- Power (difficult test with no prefect scores) vs. speed test (time limitations)
    - If speed, reliability estimates may be too high bc items are too easy
    - Everyone expected to get all of them right
    - Test-retest, alternate-forms, or split halves from two independently timed half tests
- Criterion-referenced, lower variability, lower reliability
    - If everyone has met the standard/criteria!

Calculating True Score

- Erla takes 3 tests (parallel forms) in math
- She gets an 8, 7, and 7.5
- What should we estimate as her true score/ability in math?
- Do you think that score is her true score?

Calculating True Score

- Erla takes 3 tests (parallel forms) in math
- She gets an 8, 7, and 7.5
- What should we estimate as her true score/ability in math?
- Do you think that score is her true score?
- We need a way to quantify uncertainty about Erla's score

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

$$\sigma_D = \sqrt{\sigma^2_{SEM_1} + \sigma^2_{SEM_2}}$$

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

- 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?