Ch. 4: Reliability

Reliability

- Constructs & Measurement
- History
- Classical Test Score Theory
- Four Kinds of Reliability
- Standard Error of Measurement
- Increasing Reliability

Constructs & Measurement

- Psychology as "soft science"
- Construct
  - exists but can't be directly measured
  - examples
- Measurement
  - "true value" - intelligence
  - measured or *observed* value (e.g. IQ test score)
  - discrepancy - "*error*"
- How to conceptualize *error*?

History 1

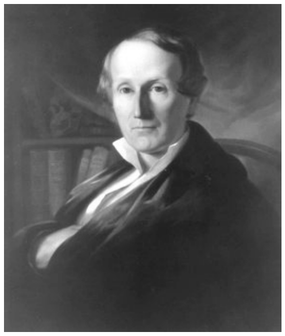
- 1896 - Karl Pearson - product-moment correlation (for continuous variables)
- 1904 - Charles Spearman - "*The proof and measurement of association between two things" - Rho* - correlation for Ordinal variables

History

- Pearson, Spearman, Thorndike (1900-1907)
  - Basic reliability theory
- Kuder, Richardson (1937), Cronbach (1989)
  - Reliability coefficients
- Bartholomew & Knott (1990s)
  - Latent variable theory
- Drasgow et al (late 1990s)
  - Item Response Theory (IRT)

Samuel George Morton

- Polygenism
  - Humans are composed of different species
- Craniometry
- Biological Determinism
- "Scientific Racism"
- d. 1851



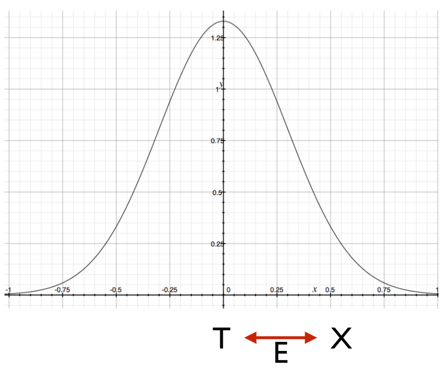
- 50 years before Spearman's work

Classical Test-Score Theory

Classical Test-Score Theory

- T= True Score
- X = Observed
- E = Error

- X = T+E
- E = X-T



Classical Test-Score Theory

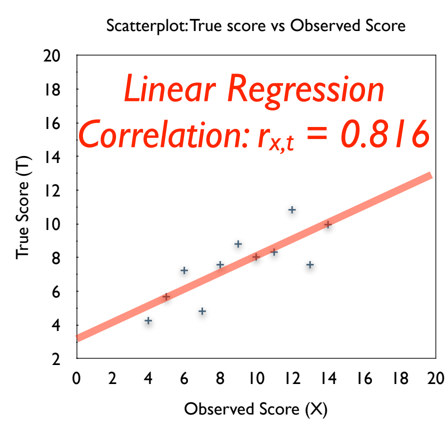
- True score (T) : the "actual" score that exists
- Observed score (X) : score as measured by a test
- Error (E) : difference between Observed and True score
- X = T + E
- E = X - T
- Assumptions: True scores have no variability. Errors are random (e.g. a normal distribution with mean of zero)
- Reliability = correlation between Observed score and True score

Classical Test-Score Theory: Reliability

- Reliability = correlation between Observed score and True score
- $R_{X,T}$

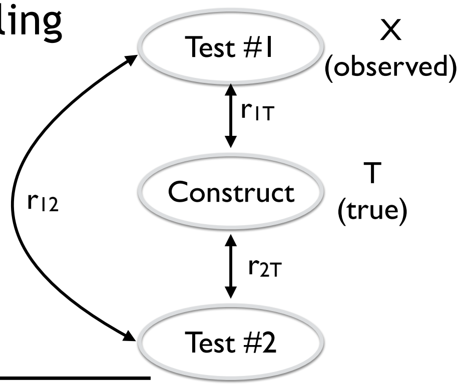
Classical Test Score Theory

X	T
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



Domain Sampling

- How to calculate r_{1T}
- Any two tests r_{12}
- r_{1j} = average of all pairs



$$r_{1,T} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} r_{i,j}}{N^2}}$$

Domain Sampling

- Problem: no way to measure True score / no possible way to measure every possible item
- Sample a limited subset of items, do this in multiple ways
- Create one or more tests
- For two given tests, correlation between the two tests will be lower than the correlation between one test and the True score
- $r_{1t} = \sqrt{r_{1j}}$

Domain Sampling Example

- Correlation of any 2 random sample tests
- $r_{1t} = \sqrt{r_{1j}}$
- $r_{1t} = \sqrt{0.64}$
- $r_{1t} = 0.80$
  - unbiased estimate of “true” reliability

Models of Reliability

- Most reliability measures are Correlation coefficients
- Alternate definition: Reliability is the ratio of the variance of True scores to the variance of the Observed scores
  - $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$
- Or, it’s the “Signal to Noise” ratio
  - $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}$
- A test with reliability of r^2=0.40 means that 40% of variation in test scores is due to variation in the “true” score, and 60% of variation is random or chance factors.

Sources of Error

- “Error” is considered the difference between True score and Observed score
- Where does Error arise?
  - Measurement errors
  - Change in True score
  - Sampling

Measuring Reliability in Practice

- Since True score is hidden, can’t use the direct formula: $R_{X,T}$
- Instead
  - think about sources of error
  - practical methods
  - *estimate* reliability

Test-Retest Reliability

- Test-Retest
  - administer same test across some time period
  - compute correlation between two administrations:
  - same subjects, same test, two administrations
  - Issue -- what is “error”?
    - actual change in true score
    - carryover or practice effects

Parallel Forms Reliability

- Also called “Alternate Forms”
  - administer two versions of the test to same subjects (often on same day)
  - compute correlation between two administrations
  - same subjects, different test forms, two administrations
  - Pros: more rigorous method of determining reliability
  - Cons: difficult to do: have to make a new test

Internal Consistency Reliability

Internal Consistency Reliability

- Give single test, calculate internal consistency of various subsets of items
- Only one test, one administration, same group of subjects
- Old: Split half method
- New: **Cronbach’s Alpha (α)**
  - estimates a lower bound for reliability
  - α of .70 to .80 is borderline
  - α of .80 is ok
  - α of .90 or higher is good

Inter-Rater Reliability

- Observational data differs from self-report data.
- Even though most behavioral rating systems attempt to be precise, errors occur (e.g. was that a “hit” or a “punch”?)
- We must consider the reliability of different observers (also called “raters”)
- **Cohen’s Kappa**
  - ranges from -1 to +1
  - “poor” < .40
  - “good” .40 to .75
  - “excellent” > .75

Reliability: errors & methods

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach’s Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

Quiz: What kind of Reliability?

Procedure	Source of error?	What kind?
Olympic judges giving consistent scores for a gymnastics performance	People	Inter-Observer
Correlation between your IQ test score taken at age 12 and again at age 13	Time	Test-Retest
Correlation between scores on 2 versions of the midterm (assuming each student takes both versions)	Item Selection	Parallel Forms
Correlation between student scores on questions 1-25 vs 26-50 of the midterm.	Item Selection	Internal Consistency

Quiz: What kind of Reliability?

Procedure	Source of error?	What kind?
Olympic judges giving consistent scores for a gymnastics performance		
Correlation between your IQ test score taken at age 12 and again at age 13		
Correlation between scores on 2 versions of the midterm (assuming each student takes both versions)		
Correlation between student scores on questions 1-25 vs 26-50 of the midterm.		

Summary

- Reliability
  - how consistent measured scores are
- Error
  - E = X - T
- What kind of Error?
  - test-retest, domain sampling, internal consistency, observer-differences

Standard Error of Measurement

- Desire to answer question "how close is this test result to the true result"
- If we know the Reliability (r) of the test, we can estimate the likely range of true values
- Given
  - S = std dev of measured scores
  - r = reliability coefficient of test

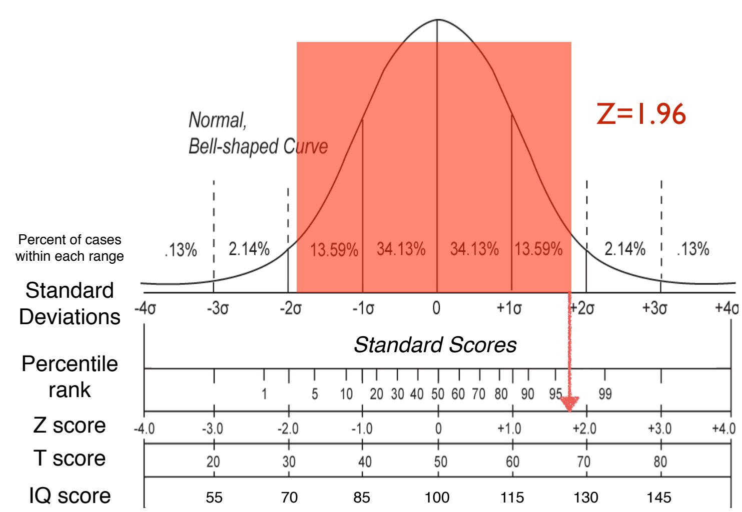
$$SEM = S\sqrt{1 - r}$$

SEM example: IQ

- Example: a person scored 106 on an IQ test, that has a reliability of 0.89. What is the 95% confidence interval of the their true score
- S = 14
  r = 0.89

$$SEM = S\sqrt{1 - r}$$
$$SEM = 14\sqrt{1 - 0.89}$$
$$SEM = 4.64$$

- Next, compute a confidence interval



Confidence Interval

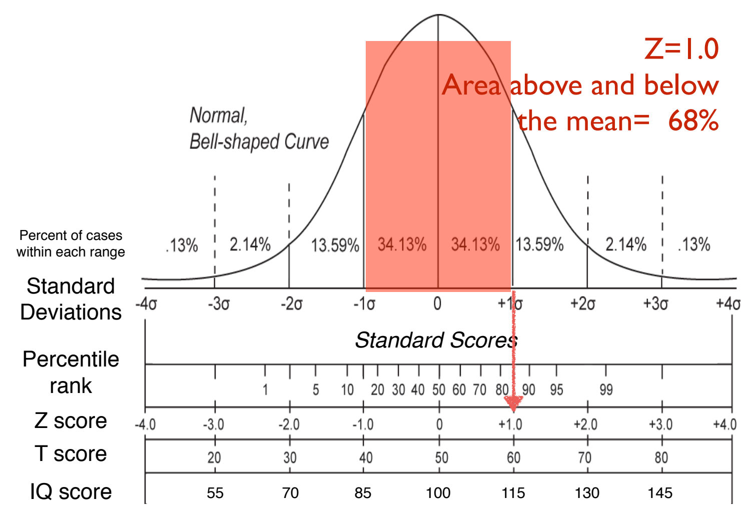
- “How likely is a true score to fall within a range”
- Z = z-score associated with % range
- Confidence interval = Z * SEM
- Example:
  - 95% confidence interval : Z = 1.96
  - SEM = 4.64
  - 1.96 * 4.64 = 9.1
  - 95% CI = ± 9.1 points
  - Range = X±CI
    - 106 ± 9.1 = range from 96.9 … 115.1

SEM Exercise

- This is for practice, not scored for points

Common Z scores & Confidence Levels

Z Score	Area above mean	Area above + below Mean	Proportion as %
0.00	0.000		0%
0.13	0.051		
0.67	0.249		
1.00	0.341	0.682	68%
1.64	0.449		
1.96	0.475		95%
2.57	0.495		



How reliable?

- r = .70 or .80 or higher is often considered “good enough” for much research
- r > .90 is very good
  - may not be worth effort to go higher
- Some real-world tests have r > 0.9
  - example: modern IQ tests

Increasing Reliability

Increasing Reliability

- **Increase N** (number of questions, items or tests)…
- **Focus** on common characteristic…
- Other methods (covered later)
  - Use **Item Analysis** ("discriminability analysis") to find items that best measure a single characteristic
  - Use **Factor Analysis** to determine sub-characteristics of a single test

Increase N

- N = number of questions or items or tests
- Formulas exist to determine how much to increase N by to reach a certain level of reliability
- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$

  N_d = new N (times old N)

  r_d = desired level of reliability

  r_o = observed level of reliability

Increase N - Examples

- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$
- Example:
  - 20-item CES-D has reliability of .87.
  - We need r = 0.95
  - N_d = 2.82
  - new N is 2.82 x 20 = 56 items

Increase N - Examples

- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$
  - Your 40-item test has reliability of .50.
  - You want .90.
  - N_d = 9.0
  - new N is 9 x 40 = 360!

Focus Test

- Reliability increases as a test focuses on a single concept or characteristic ("construct")
- Trying to capture multiple concepts in a single test reduces reliability
- Methods:
  - Informal — remove items with poor face validity (chapter 5)
  - Statistical:
    - Discriminability Analysis (chapter 6)
    - Factor Analysis (chapter 13)

Reliability Summary

- Measurement Error occurs in all fields -- Psychology focuses on it
- Kind of Reliability : *where* the error came from
- Improving Reliability: more items, focusing test, factor analysis
- Reliability is useful: calculate SEM and Confidence Intervals
- Reliability is not Validity: Reliable tests aren't automatically valid
- A reliable test *may* be valid