Ch. 4: Reliability

- History
- Classical Test Score Theory
  - Domain Sampling
  - Models of reliability
  - Sources of error
- Estimating Reliability
  - Test-Retest
  - Parallel Forms
  - Internal Consistency / Cronbach's α
- Difference Scores

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

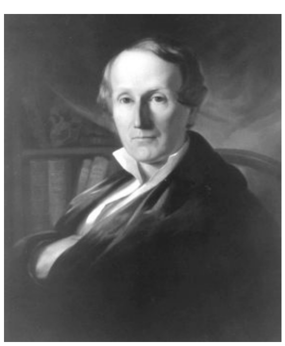
- 1896 - Karl Pearson - product-moment correlation (for continuous variables)
- 1904 - Charles Spearman - "*The proof and measurement of association between two things" - Rho* - correlation for Ordinal variables

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

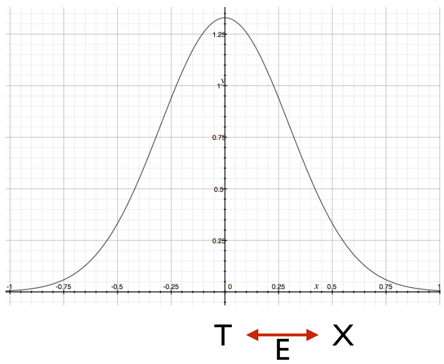


Classical Test-Score Theory

- 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

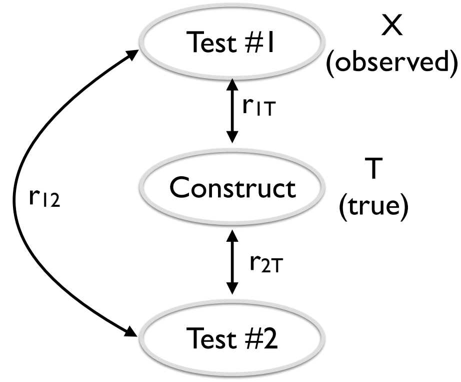
Classical Test-Score Theory

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



Domain Sampling

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



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
- $\rho^2_{XT} = \dfrac{\sigma^2_T}{\sigma^2_X}$
- A test with reliability of r=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

Test-Retest Reliability

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

Parallel Forms Reliability

- Parallel Forms
  - administer two versions of the test to same subjects (often on same day)
  - compute correlation between two administrations
  - Pros: most rigorous method of determining reliability
  - Cons: difficult to do, is not often done

Internal Consistency Reliability

- Give single test, calculate internal consistency of various subsets of items
- Split halves methods exist, but have generally been supplanted by...
- 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

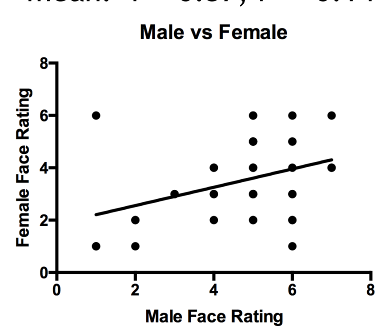
- Observational data differs from self-report data.
- Even though most behavioral rating systems attempt to be precise, errors occur (e.g. was that a "hit" or a "punch"?)
- We must consider the reliability of different observers (also called "raters")
- Cohen's Kappa
  - ranges from -1 to +1
  - "poor" < .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
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between forms
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

Estimating Reliabliity

- Exercise 02 - What did the linear regression mean? $r = 0.37$, $r^2 = 0.14$



Linear regression results:
$R^2 = 0.14$ $p = 0.013$, $N = 45$

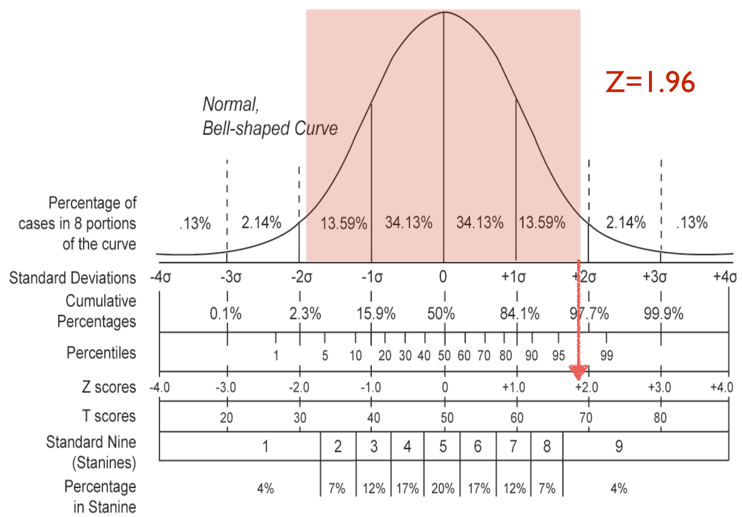
Figure 3: Linear Regression of Male vs. Female face Ratings

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
- SEM = $S\sqrt{1-r}$
- S = std dev of measured scores
- r = reliability coefficient of test

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
- SEM = $S\sqrt{1-r}$
  S = 14
  r = 0.89
- SEM = $14\sqrt{1-.89} = 4.64$
- 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

Real-world example: SAT

	Reading	Math	Writing
Mean	501	515	493
SD	112	121	112
Reliability Coefficient	0.91	0.92	0.89
SEM	31	31	34

SEM Example : SAT

- Example: a person scored 500 on the SAT Math test, that has a R=0.92 and SD=121. What is the 95% confidence interval of the their true score
- SEM = $S\sqrt{1-r}$
  S = 121
  r = 0.92
- SEM = 121 * sqrt(1- 0.92) = 121 * sqrt(.08) = 34.2
- 95% confidence interval = Z score of 1.96.
- 95% confidence interval = Z * SEM = 67.03
- 500 ± 67 gives Range of (433... 567)

Reliability of Difference Scores

- Common need is to compute the difference between two scores or two tests, with known reliability
- Unfortunately, taking the difference dramatically reduces the reliability
- E.g. for two tests with reliability .90 and .70 that are correlated to each other by .70, a difference score has a reliability of .33

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 time / effort to get higher

Increasing Reliability

- **Increase N** (number of questions, items or tests)
  - (example next slide)
- **Focus** on common characteristic
  - tests are more reliable if all items measure a <u>single</u> characteristic
- Use **Factor Analysis** to determine sub-characteristics of a single test
- Use **Item Analysis** ("discriminability analysis") to find items that best measure a single characteristic
- **Statistically correct** for attenuation

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
- Example: 20-item CES-D has reliability of .87. We need .95. N_d = 2.82, so new N is 2.82 x 20 = 56

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 .95. N_d = 2.82, so new N is 2.82 x 20 = 56
  - Your 40-item test has reliability of .50. You want .90. N_d = 9.0, so new N is 9 x 40 = 360!

(Re)Focus Test

- Reliability increases the more the test focuses on a single concept or characteristic
- Trying to capture multiple concepts in a single test reduces reliability
- Methods:
  - Ad-hoc / informal -- face validity of items and remove those that don't fit
  - Statistical:
    - Factor Analysis
    - Discriminability Analysis.

Chapter 4 Summary

- Measurement Error occurs in all fields -- Psychology has a special focus on it
- Reliability : more than one type, to measure it we need to specify *where* the measurement error comes from
- If a test is Unreliable, it is irrelevant whether or not it is Valid. Reliability is a foundation.
- Reliability can be improved through ad-hoc (informal) methods, factor analysis and discriminant analysis, and statistically
- When reliability is known, we get SEM, and from SEM we get Confidence Intervals

Reliability Summary

- Reliability: consistency of measurement
- Source of error —> how to measure reliability
- Reliability coefficients ~ correlation

- Reliability is NOT Validity
- Reliability is a foundation