Basic Concepts of Classical Test Theory

Lecture 9
Feb 28, 2019

CLPS 2908
Multivariate Statistical Techniques

1. Test Components

- A "test" is
  - any variable, item, trial
  - though often a multi-item scale or measure.
- Every test has a **true-score** component and an **error** component

$$\boxed{x_i = T_i + e_i}$$

measurement = latent/true score + error

latent score ≈ average score on universe of items

DEFINE: $T = \varepsilon v(x)$ and **DEFINE**: $e = x - T$

Then $\varepsilon v(e) = \varepsilon v(x - T) = \varepsilon v(x) - \varepsilon v(T) = T - T = 0$; **hence** $\varepsilon v(e) = 0$

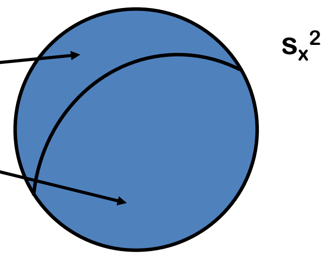
2. Variance of a Test

$x = T + e$

$$s_x^2 = s_T^2 + s_e^2 + 2cov_{T.e}$$

- By assumption, **errors** (random variable) and **true score** (constant) are uncorrelated, hence $cov_{T.e} = 0$

$$s_x^2 = s_T^2 + s_e^2$$

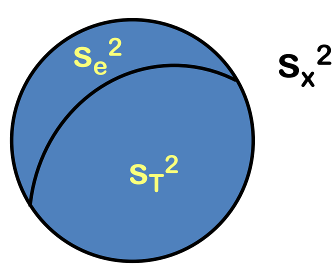


3. Reliability

= the proportion of **true score** variance out of **total** measured variance: $\dfrac{s_T^2}{s_x^2}$

$$\text{or}: \quad 1 - \frac{s_e^2}{s_x^2}$$

= the proportion of **nonerrror** variance



Look familiar? Total variance, explained variance….

Deriving Reliability (Nunally)

Aim: assess item's *representativeness* of true score, $\boldsymbol{r}_{iT}$

If $\bar{r}_{ij}$ for any item X_i with all other $X_j \approx$ constant, then

$$\bar{r}_{ij} = r^2_{i\sum i} = r^2_{iT}$$

pp. 176-177

If reliability measure (written as r_{xx}) can estimate $\bar{r}_{ij}$,

then $r_{xx} \approx r^2_{iT}$

= **variance of all X_i explained by $\boldsymbol{T}$.**

Thus, reliability is s_T^2 / s_x^2 .

3. Reliability (cont'd)

= the proportion of true-score variance out of total measured variance $\dfrac{s_T^2}{s_x^2}$

- Types of reliability estimates
  - Test re-test (canceling *context*-specific s_e^2)
  - Split-half (canceling *item*-specific s_e^2)
  - Internal consistency → Cronbach's α
  - Inter-rater agreement

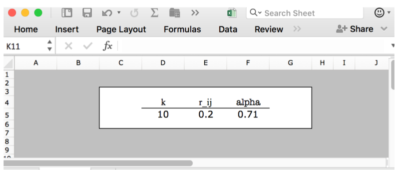
4. Formulas

N is the factor by which you lengthen the test. So if *N* = 2, the number of items doubles.

Spearman-Brown "prophecy" formula for varying test length

$$r_{new} = \frac{N\, r_{old}}{1 + (N-1) r_{old}}$$

$$\downarrow$$



Standardized Cronbach's α: $$\frac{k\, \bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}}$$

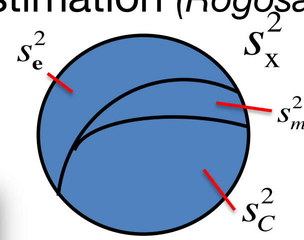
"Represents the expected correlation of one test with an alternative form containing the same number of items"

$$= \frac{k}{k-1}\left[1 - \frac{\Sigma \sigma_k^2}{\sigma_{Total}^2}\right]$$

SPSS provides both

5. Misconceptions

- "True score" does not necessarily mean you are "truly" (validly) measuring ***construct*** of interest.
- s_T^2 isn't pure construct variance s_c^2.
  - Method variance, s_m^2 is measured in s_T^2 as well
  - So are response biases, stable context, etc.
- Reliability ≠ accuracy of point estimation *(Rogosa)*
  - Shoe size distribution + variance
  - If s_x^2 is large, s_e^2 is still considerable, even if reliability is > .80



$$s_T^2 = 4.21,\ s_e^2 = 1.0, \text{ therefore } s_x^2 = 5.21$$
$$r_{xt} = 4.21/5.21 = 0.808$$

6. How to Increase Reliability

- Write items clearly
- Remove ambiguous or multi-part items
- Remove items with unfamiliar content
- Make test instructions easy to understand
- Standardize conditions for administration
- Minimize fatigue effects (shorter, interesting test)
- Make test longer (use "prophecy formula")
  - but shortening the test may focus on the best items
- If applicable, train raters (minimize guessing, subjective judgments)

Thanks to Gerard Saucier!

7. Basics of Item Analysis

- Item sampling, test construction
- Item selection → first run
  - Item-total correlations, inter-item correlations
  - Internal consistency (α)
  - How much homogeneity is desirable?
  - Multi-dimensionality ➛ PCA (or item grouping)
- Item exclusion → next run

See SPSS output examples

```
RELIAB VAR = anxious tense disorgan shy
harsh quiet lax
 /SCALE (AGREEABLENESS) = anxious tense
disorgan shy harsh
  quiet lax
 /STA = CORR
 /SUM = TOTAL CORR.
```

Warnings

The covariance matrix is calculated and used in the analysis.

Case Processing Summary

		N	%
Cases	Valid	234	80.1
	Excluded(a)	58	19.9
	Total	292	100.0

a Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.572	.565	7

Inter-Item Correlation Matrix

	anxious	tense	disorgan	shy	harsh	quiet	lax
anxious	1.000	.677	.058	.156	.154	.271	-.055
tense	.677	1.000	.026	.206	.223	.239	-.161
disorgan	.058	.026	1.000	.096	.146	.032	.350
shy	.156	.206	.096	1.000	.004	.712	.062
harsh	.154	.223	.146	.004	1.000	.003	.038
quiet	.271	.239	.032	.712	.003	1.000	.041
lax	-.055	-.161	.350	.062	.038	.041	1.000

The covariance matrix is calculated and used in the analysis.

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Inter-Item Correlations	.156	-.161	.712	.873	-4.425	.045	7

The covariance matrix is calculated and used in the analysis.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
anxious	25.44	35.646	0.401	0.482	0.491
tense	25.96	35.921	0.382	0.503	0.499
disorgan	26.50	38.448	0.199	0.152	0.574
shy	26.02	35.171	0.394	0.522	0.492
harsh	26.97	41.815	0.170	0.074	0.573
quiet	25.96	35.260	0.422	0.536	0.483
lax	26.49	43.496	0.091	0.160	0.597