Outline

Some ways to quantify Reliability:

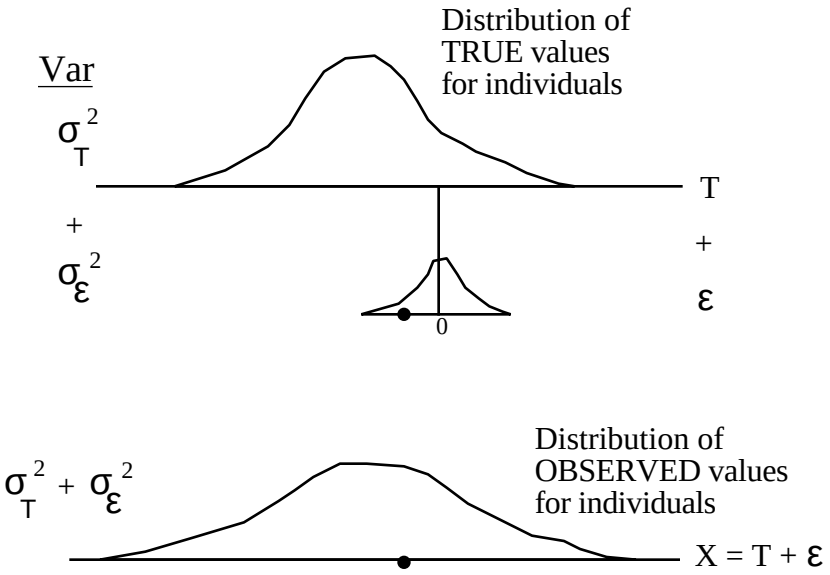
- Reliability Coefficient
- Internal Consistency (Cronbach's α)

Implications:

-

.....

Model for Reliability



"True" scores / values not knowable;

Variance calculation assumes that the distribution of errors is independent of T

Reliability Coefficient

$r_{XX} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_\varepsilon}$ i.e. the fraction of observed variation that is 'real'

Note that one can 'manipulate' r by choosing a large or small σ^2_T

Effect of # of Items on Reliability Coefficient
(if all items have same variance and same intercorrelations)

SCALE 2 N Times more items than SCALE 1

$r_{SCALE\ 2} = \frac{N \times r_{SCALE\ 1}}{1 + [N-1] \times r_{SCALE\ 1}}$

e.g.

Scale	# Items	r
1	10	0.4
2	20 (× 2)	0.57
3	30 (× 3)	0.67

Cronbach's α

k items

$$\alpha = \frac{k \times \bar{r}}{1 + [k-1] \times \bar{r}}$$, where $\bar{r}$ = average of inter-item correlations

α is an estimate of the expected correlation of one test with an alternative form with the same number of items.

α is a lower bound for r_{XX} i.e $r_{XX} \geq \alpha$

$r_{XX} = \alpha$ if items are parallel.

parallel

Average [item 1] = Average [item 2] =Average [item 3] = ...

Variance[item 1] = Variance[item 2] =Variance [item 3] = ...

Correlation[item 1, item 2] = Correlation[item 1, item 3] = ...

=Correlation[item 2, item 3] = ...

INTRACLASS CORRELATIONS (ICC's)

- **Various versions**

  TEST-RETEST

  INTRA-RATER

  INTER-RATER...

- **Formed as Ratios of various Variances**

  e.g. $$\frac{\sigma^2_{TRUE}}{\sigma^2_{TRUE} + \sigma^2_{ERROR}}$$

  with estimates of various σ^2 's substituted for the σ^2 's .

  Estimates of various components typically derived from ANOVA.

- **Note the distinction between DEFINITIONAL FORM (involving PARAMETERS) and COMPUTATIONAL FORM (involving STATISTICS)**

  Fleiss Chapter 1 good here; Norman & Streiner not so good!!)

ICC's (Portnoy and Wilkins)

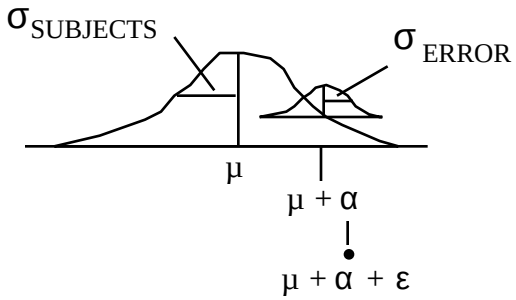
(1) multiple (unlabeled) measurements of each subject

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

(3) as in (2), but these raters studied are the only raters of interest

.....

(1) multiple (unlabeled) measurements of each subject



$$\text{ICC} = \frac{\sigma^2_{\text{SUBJECTS}}}{\sigma^2_{\text{SUBJECTS}} + \sigma^2_{\text{ERROR}}}$$

Model for observed data:

y[subject i, measurement j] = $\mu + \alpha_i + \varepsilon_{ij}$

EXAMPLE 1

This example is in the spirit of the way the ICC was first used, as a measure of the greater similarity within families than between families: Study by Bouchard (NEJM) on weight gains of 2 members from each of 12 families: It is thought that there will be more variation between members of different families than between members of the same family: family (genes) is though to be a large source of variation; the two twins per family are thought of as 'replicates' from the family and closer to each other (than to others) in their responses. Here the "between" factor is family i.e. families are the subjects and the two twins in the family are just replicates and they don't need to be labeled (if we did label them 1 and 2, the labels would be arbitrary, since the two twins are thought to be 'interchangeable'. (weight gain in Kg over a summer)

model: weight gain for person j in family i = $\mu + \mu + \alpha_i + e_{ij}$

1-way Anova and Expected Mean Square (EMS)

Source	Sum of Sq	d.f	Mean Square	Expected Mean Square
Between (families)	99	11	9.0	$\sigma^2_{\text{"error"}} + k_0\,\sigma^2_{\text{between}}$
Error(Within families)	30	12	2.5	$\sigma^2_{\text{"error"}}$
Total	129	23		

In our example, we measure k=2 members from each family, so k_0 is simply 2

[if the k's are unequal, k_0 is somewhat less than the average k... k_0 = average k – (variance of k's) / (n times average k) ...see Fleiss page 10]

Estimation of parameters that go to make up ICC

2.5 is an estimate of $\sigma^2_{\text{"error"}}$

9.0 is an estimate of $\sigma^2_{\text{"error"}} + 2\,\sigma^2_{\text{between}}$

6.5 is an estimate of $2\,\sigma^2_{\text{between}}$

$\frac{6.5}{2}$ is an estimate of $\sigma^2_{\text{between}}$

$$\frac{\frac{6.5}{2}}{\frac{6.5}{2} + 2.5} = \frac{3.25}{3.25 + 2.5} = 0.57$$

is an estimate of ICC $= \dfrac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{error}}}$

COMPUTATIONAL Formula for "1-way" ICC

$$\frac{\frac{MSbetween - MSwithin}{k_0}}{\frac{MSbetween - MSwithin}{k_0} + MSwithin}$$

$$= \frac{MSbetween - MSwithin}{MSbetween + (k_0 - 1)MSwithin} \quad \text{[shortcut]}$$

is an estimate of the ICC

Notes:
• Streiner and Norman start on page 109 with the 2-way anova for inter-observer variation. There are mistakes in their depiction of the SSerror on p 110 [it should be $(6\text{-}6)^2+(4\text{-}4)^2+(2\text{-}1)^2 +...(8\text{-})^2 =10$. If one were to do the calculations by hand, one usually calculates the SStotal and then obtains the SSerror by subtraction]
• They then mention the 1-way case, which we have discussed above, as "the observer nested within subject" on page 112
• Fleiss gives methods for calculating CI's for ICC's.

EXAMPLE 2: INTRA-OBSERVER VARIATION FOR 1 OBSERVER

Computations performed on earlier handout...

Var(SUBJECT) = 23.67 Var(ERROR) = 1.38

$\hat{ICC}$ = 23.67 / (23.67 + 1.38) = 0.94

An estimated 94% of observed variation in earsize measurements by this observer is 'real' .. i.e. reflects true between-subject variability.

Note that I say 'an estimated 94% ...". I do this because the 94% is a statistic that is subject to sampling variability (94% is just a point estimate or a 0% Confidence Interval). An interval estimate is given by say a 95% confidence interval for the true ICC (lower bound of a 1-sided CI is 82% ... see previous handout)

Increasing Reliability by averaging several measurements

In 1-way model: $y_{i,j} = \mu + \alpha_i + e_{ij}$

where $var[\alpha_i] = \sigma^2_{\text{between subjects}}$; $var[e_{ij}] = \sigma^2_{\text{"error"}}$

Then if we average k measurements, i.e.,

$$ybar_i = \mu + \alpha_i + ebar_i$$

then

$$\underset{i}{Var}[ybar_i] = \sigma^2_{between} + \frac{\sigma^2_{\text{"error"}}}{k}$$

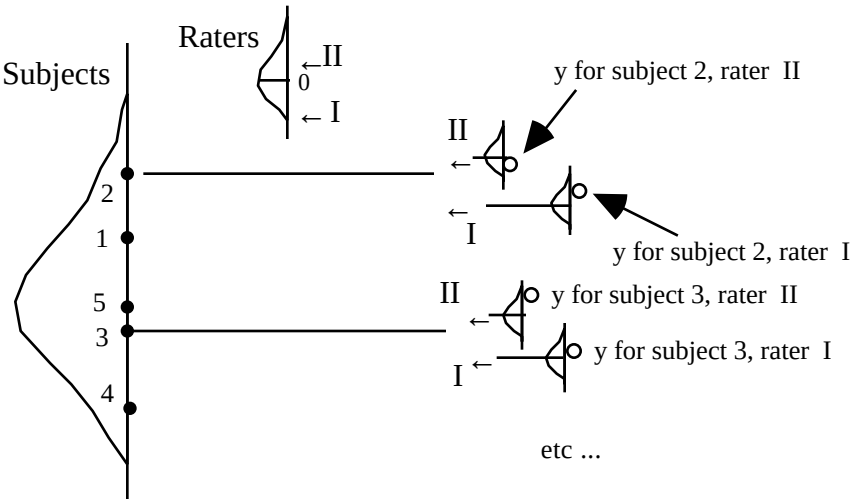
$$\text{So } ICC[k] = \frac{\sigma^2_{between}}{\sigma^2_{between} + \frac{\sigma^2_{\text{"error"}}}{k}}$$

This is called **"Stepped-Up" Reliability**.

ICC's (Portnoy and Wilkins).

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

• **Model**



$$\mu + \alpha_{[subject]} + \beta_{[rater]} + \varepsilon$$

$$\sigma^2_{subjects} \quad \sigma^2_{raters} \quad \sigma^2_{error}$$

• **From 2- way data layout (subjects x Raters)**

estimate $\sigma^2_{"subjects"}$, $\sigma^2_{"raters"}$ and $\sigma^2_{"error"}$ by 2-way ANOVA

• **Substitute variance estimates in appropriate ICC form**

e.g. 2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64

subject	5				6				7				6			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

ESTIMATING INTER-OBSERVER VARIATION from occasion=1;

PROC GLM in SAS ==> estimating components 'by hand'

```
INPUT subject rater occasion earsize; if occasion=1; (32 obsns)

proc glm; class subject rater; model earsize=subject rater / ss3;
  random subject rater;

 General Linear Models Procedure: Class Level Information

Class     Levels    Values
SUBJECT        8    1 2 3 4 5 6 7 8
RATER          4    1 2 3 4 Number of observations in data set = 32

                          Sum of      Mean
Source                DF  Squares    Square   F Value   Pr > F
Model                 10  764.500    76.45    78.80     0.0001
Error                 21   20.375     0.97
Corrected Total       31  784.875

R-Square       C.V.      Root MSE     EARSIZE Mean
0.974040   1.534577      0.98501        64.1875

Source    DF  Type III SS   Mean Square   F Value   Pr > F
SUBJECT    7   734.875000     104.98      108.20   0.0001
RATER      3    29.625000       9.87       10.18   0.0002

Source       Type III Expected Mean Square
SUBJECT      Var(Error) + 4 Var(SUBJECT)
RATER        Var(Error) + 8 Var(RATER)
```

So... solving 'by hand' for the 3 components...

```
Var(Error) + 4 Var(SUBJECT) = 104.98
Var(Error)                  =   0.97
        ==>  4 Var(SUBJECT) = 104.01
        ==>    Var(SUBJECT) = 104.01 / 4 = 26.00

Var(Error) + 8 Var(RATER)   =   9.87
Var(Error)                  =   0.97
        ==>  8 Var(RATER)   =   8.90
        ==>    Var(RATER)   =   8.90 / 8 =  1.11

               Var(Error)                =  0.97
```

Estimating Variance components using PROC VARCOMP in SAS

```
proc varcomp; class subject rater; model earsize = subject rater;

                          Estimate
Variance Component         EARSIZE
Var(SUBJECT)                26.00
Var(RATER)                   1.11
Var(Error)                   0.97
```

• **ICC: "Raters Random"** (Fleiss § 1.5.2)

$$\underline{ICC} = \frac{Var(SUBJECT)}{Var(SUBJECT) + Var(RATER) + Var(Error)} = \frac{26.00}{26.00+1.11+0.97} = 0.93$$

1-sided 95% Confidence Interval (see Fleiss p 27)

df for F in CI: (8-1)= 7 and v* , where

$$v^* = \frac{(8-1)(4-1)(4 \cdot 0.93 \cdot 10.18 + 8[1+(4-1) \cdot 0.93]-4 \cdot 0.93)^2}{(8-1) \cdot 4^2 \cdot 0.93^2 \cdot 10.18^2 + (8[1+(4-1) \cdot 0.93]-4 \cdot 0.93)^2} = 8.12$$

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{8(104.98 - 3.5 \cdot 0.97)}{8 \cdot 104.98 + 3.5 \cdot [4 \cdot 9.87 + (8 \cdot 4 - 8 - 4) \cdot 0.97]} = \underline{0.78}$$

• **ICC: if use one "fixed" observer** (see Fleiss p 23, strategy 3)

$$ICC = \frac{Var(SUBJECT)}{Var(SUBJECT) + Var(Error)} = \frac{26.00}{26.00 + 0.97} = 0.96$$

lower limit of 95% 1-sided CI (eqn 1.49: F = 2.5 ; 7 & 7x3=21 df)

$$ICC = \frac{104.98 - 2.5}{104.98 + (4-1) \cdot 2.5} = 0.91$$

USING ALL THE DATA SIMULTANEOUSLY

(can now estimate subject x Rater interaction .. i.e extent to which raters 'reverse themselves' with different subjects)

Components of variance when use **both** measurements (all 64 obsns)

```
proc varcomp;
 class subject rater;
 model earsize = subject rater;
```

```
proc varcomp;
class subject rater;
model earsize = subject rater
                subject*rater;
```

Variance Component	Estimate EARSIZE
Var(SUBJECT)	25.52
Var(RATER)	0.70
Var(Error)	1.37

Variance Component	EARSIZE
Var(SUBJECT)	25.47
Var(RATER)	0.67
Var(SUBJECT*RATER)	0.31
Var(Error)	1.13

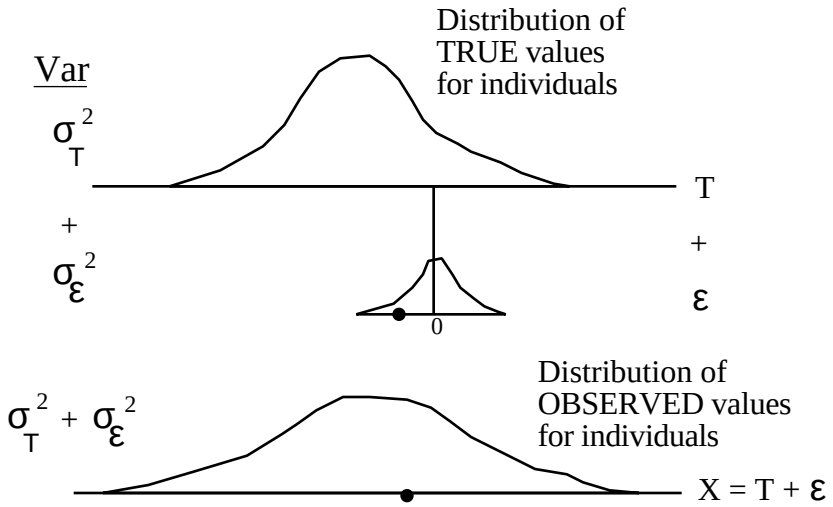
LINK between STANDARD ERROR OF MEASUREMENT and RELIABILITY COEFFICIENT

Example: GRE Tests (cf blurb from Educational Testing Service)

Standard Error of Measurement = 23 points

Reliability Coefficient: R = 0.93

recall...



$\sigma^2_e = 23$ ==> $\sigma^2_e = 529$;

$$R = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_e} = 0.93 \Rightarrow \sigma^2_T = \frac{R \times \sigma^2_e}{1 - R} = \frac{0.93 \times 529}{1 - 0.93} = 7028$$

$$\sigma^2_T + \sigma^2_e = 7028 + 529 = 7557 \Rightarrow \sqrt{\sigma^2_T + \sigma^2_e} = \sqrt{7557} = 87$$

So if 3 SD's on either side of the mean of 500 covers most of the observed scores, this would give a range of observed scores of 500 – 261 = 239 to 500 + 261 = 761.

Another way to say it (see Streiner and Norman, bottom of page 119) :-

$$\sigma_e = \sqrt{\sigma^2_T + \sigma^2_e} \times \sqrt{1 - R} = \text{SD[observed scores]} \times \sqrt{1 - R}$$

Confidence Intervals / Sample Sizes for ICC's

see Fleiss...

CI's based on F distribution tables;

CI's not symmetric;

More interested in 1-sided CI's i.e. (lower bound, 1) i.e. ICC ≥ 0.xx;

See also Donner and Eliasziw.

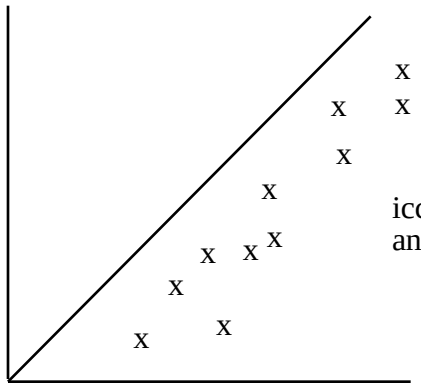
NOTE: If interested in ICC that incorporates random raters, then sample size must involve both # of raters and # of raters

CI will be very wide if use only 2 or 3 raters

Approach sample size as "n's or raters and subjects needed for a sufficiently narrow CI.

Why Pearson's r is not always a good [or practical] measure of reproducibility

1. It does not pick up "shifts"

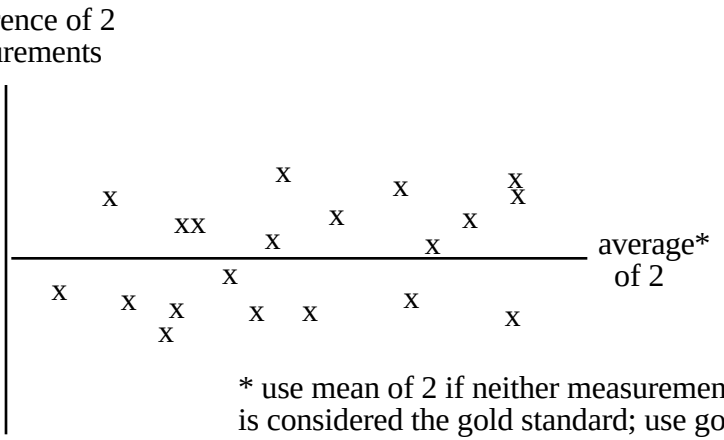


icc includes "shifts" and is lower than r

2. not practical if > 2 measurements or variable # of measurements per subject

   ICC 'made for' such situations

Method of Bland & Altman [Lancet]



* use mean of 2 if neither measurements is considered the gold standard; use gold standard otherwise

\+++ see biases quickly

can explain to your in-laws
(can you explain ICC to them?)

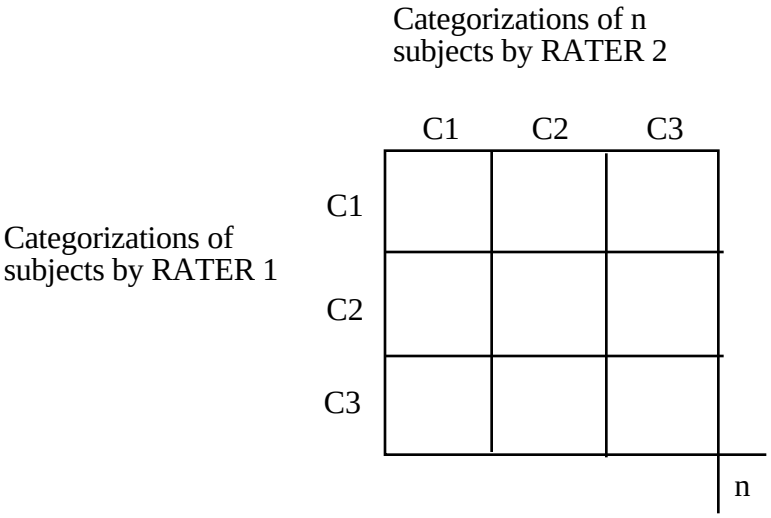
emphasises errors in measurements scale itself
(like ±23 in GRE score)

– – – if don't know real range, magnitudes of standard error of measurement not helpful (see Norman & Streiner)

cannot use with > 2 measurements

doesn't generalize to <u>raters</u>

Assessing reproducibility of measurements made on a CATEGORICAL scale



See chapter 13 in Fleiss's book on Rates and Proportions or pp 516-523 of Chapter 26 of Portnoy and Wilkins

• Simple Measure

$$\% \text{ agreement} = \frac{\#\text{ in diagonal cells}}{n} \times 100$$

• Chance-Corrected Measure

$$\kappa = \frac{\% \text{ agreement} - \% \text{ agreement expected by chance}^*}{100\% \text{ agreement} - \% \text{ agreement expected by chance}}$$

* expected proportion = $\sum$ p[row]*p[col] --- $\sum$ over the diagonals

(see Aickin's arguments against 'logic' of chance-correction: Biometrics 199)

can give weights for 'partial' agreement

if > 2 raters, use range or average of pairwise kappas

with quadratic weights, weighted kappa = icc