The intra-class correlation coefficient (ICC)

(1)

Form all possible ordered pairs of items from the same cluster (Y_{ij}, Y_{ik}), $j \neq k$

How many pairs are there? $\sum_{i=1}^{N} M_i(M_i - 1)$

The ICC is the usual correlation, computed on these ordered pairs

An ICC close to 0 is desirable.

(2)

Alternate expression for ICC:

Source	SS	df
Between	SSB	N-1
Within	SSW	K-N
Total	SST	K-1

$$ICC = 1 - \frac{K}{K-N}\frac{SSW}{SST}$$ (Note the similarity to R^2_{adj})

Worst case: $SSW = 0 \Rightarrow ICC = 1$

Best case: $SSW = SST \Rightarrow ICC = 1 - \frac{K}{K-N}$

$$= -\frac{N}{K-N} \approx 0$$

From last time: $\bar{y}_{clus} = \frac{\hat{t}}{K}$, $\hat{t} = N\bar{t}$

$$V[\bar{y}_{clus}] = \frac{1}{K^2} N^2 \frac{s_t^2}{n}\left(1 - \frac{n}{N}\right)$$

What happens if K is unknown?

This occurs if you don't know M_i for the clusters not chosen

Then $\hat{t} = N\bar{t}$ is still good

But $\bar{y}_{clus} = \frac{\hat{t}}{K}$ requires an estimator of K

How can we estimate K?

$$\hat{K} = N\bar{m}$$

$\bar{m}$: sample mean of the cluster sizes

Then $\bar{y}_{clus} = \frac{\hat{t}}{\hat{K}} = \frac{N\bar{t}}{N\bar{m}} = \frac{\bar{t}}{\bar{m}}$

This is a ratio estimator, so it is asymptotically unbiased

(5)

$$V[\bar{y}_{clus}] \approx \frac{1}{\bar{M}^2} \frac{S_e^2}{n}\left(1 - \frac{n}{N}\right),$$

where $S_e^2 = S_t^2 + B^2 S_m^2 - 2B S_{tm}$

$\bar{M}$ = average cluster size for population $= \frac{K}{N}$

S_t^2 = population variance of cluster totals

$B = \frac{\bar{T}}{\bar{M}}$, $\bar{T}$ = Average cluster total for population $= \frac{t}{N}$

(6)

S_m^2 = population variance of cluster sizes

S_{tm} = population covariance between the cluster totals & cluster sizes

$$\hat{V}[\bar{y}_{clus}] = \frac{1}{\bar{m}^2} \frac{s_e^2}{n}\left(1 - \frac{n}{N}\right),$$

$$s_e^2 = s_t^2 + \bar{y}_{clus}^2 s_m^2 - 2\bar{y}_{clus} s_{tm}$$

2-stage cluster sampling

1st stage: select n clusters from N by SRSWOR

2nd stage: select a sample of size m_i from the M_i items in each cluster, SRSWOR

Before: $\hat{t} = N\bar{t} = N \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} t_i$

But t_i is now unknown

Estimate t_i by $\hat{t}_i = M_i \bar{y}_i$

So $\hat{t} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i$

Law of Iterated Expectations

$$E\big(E[X|Y]\big) = E[X]$$

Application to variances:

$$V[X] = E[X^2] - \big(E[X]\big)^2$$

$$V[X] = E_1(E_2(X^2)) - [E_1(E_2(X))]^2 \quad (9)$$

↑ Conditional expectation, given what happened at stage 1

$$= E_1(E_2(X^2)) - E_1((E_2(X))^2) + E_1((E_2(X))^2) - [E_1(E_2(X))]^2$$

$$= E_1[E_2(X^2) - (E_2(X))^2] + E_1((E_2(X))^2) - [E_1(E_2(X))]^2$$

$$V[X] = E_1(V_2(X)) + V_1(E_2(X)) \quad (10)$$

Next time, we will apply this to $\hat{t}$.