Managing Complex Data Structures

Hsueh-Sheng Wu

CFDR Workshop Series

Summer 2010

BGSU
Center for Family and Demographic Research

Outline

- What are complex data structures?
- Signs of data having a complex structure
- Why is there a need to learn about complex data structure?
- How to manage complex data structure?
  - Merge data
  - Reshape the data
  - Generate variables
- Conclusions

What Are Complex Data Structures?

Data with simple structure:

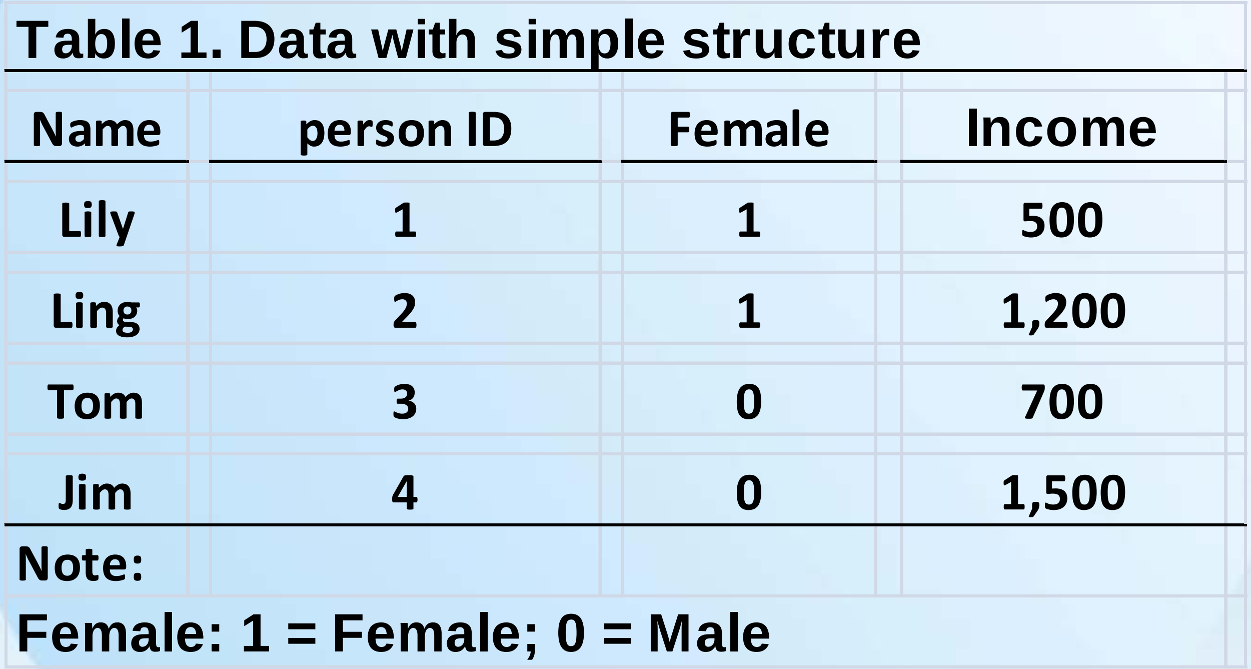

Table 1. Data with simple structure

Name	person ID	Female	Income
Lily	1	1	500
Ling	2	1	1,200
Tom	3	0	700
Jim	4	0	1,500

Note:

Female: 1 = Female; 0 = Male

What Are Complex Data Structures? (Continued)

Example 1: Constructs nested within individuals over time

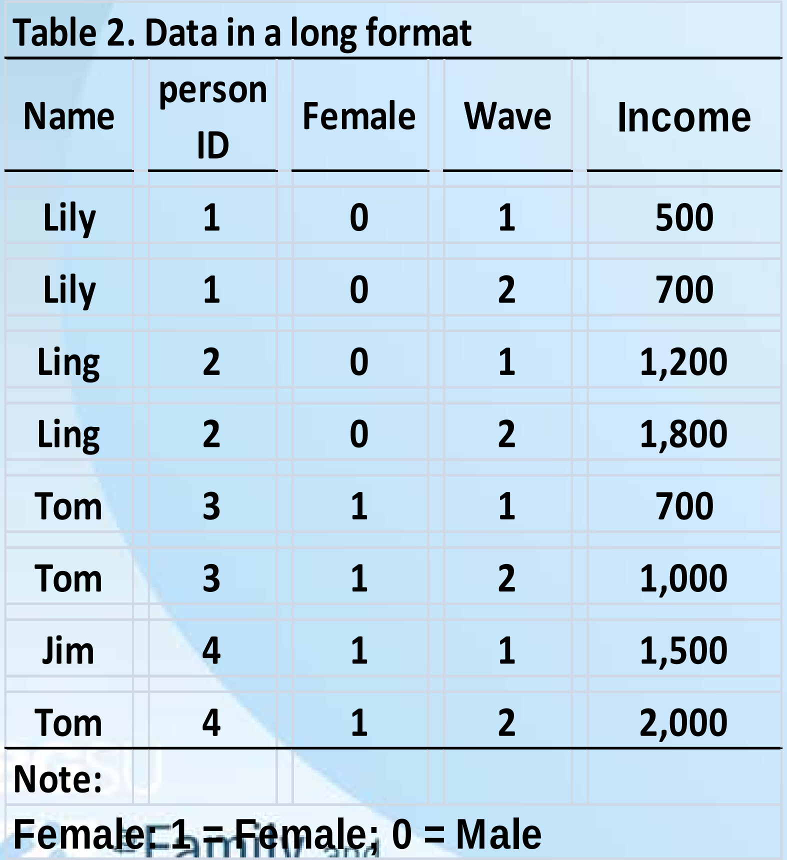

Table 2. Data in a long format				
Name	person ID	Female	Wave	Income
Lily	1	0	1	500
Lily	1	0	2	700
Ling	2	0	1	1,200
Ling	2	0	2	1,800
Tom	3	1	1	700
Tom	3	1	2	1,000
Jim	4	1	1	1,500
Tom	4	1	2	2,000
Note:				
Female: 1 = Female; 0 = Male				

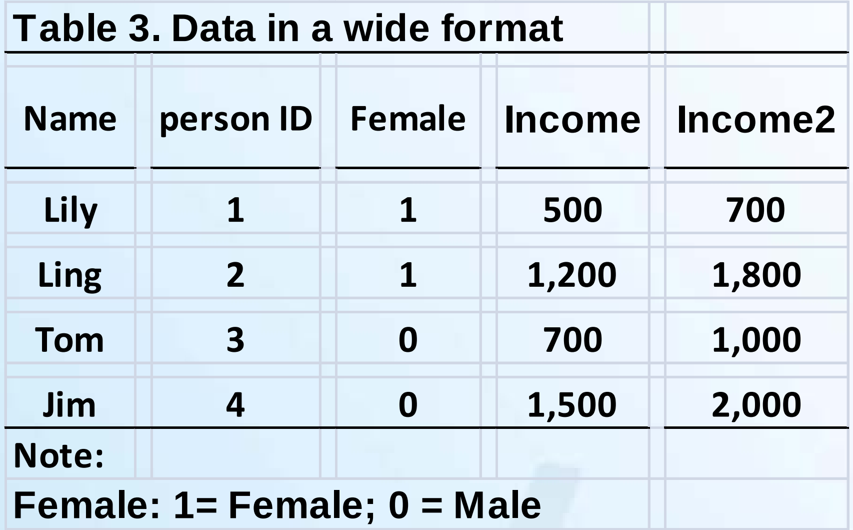

Table 3. Data in a wide format				
Name	person ID	Female	Income	Income2
Lily	1	1	500	700
Ling	2	1	1,200	1,800
Tom	3	0	700	1,000
Jim	4	0	1,500	2,000
Note:				
Female: 1= Female; 0 = Male				

What Are Complex Data Structures? (Continued)

Example 2: Individuals nested within a larger unit (e.g., a couple)

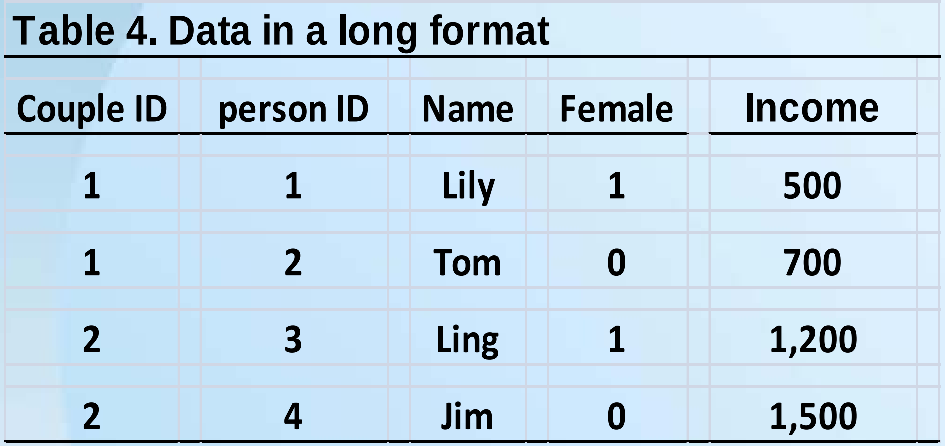

Table 4. Data in a long format

Couple ID	person ID	Name	Female	Income
1	1	Lily	1	500
1	2	Tom	0	700
2	3	Ling	1	1,200
2	4	Jim	0	1,500

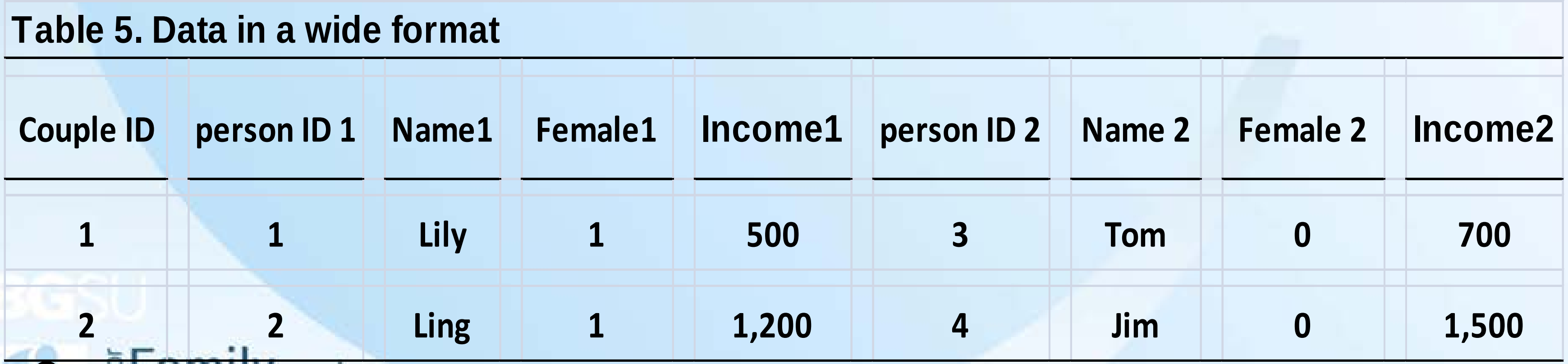

Table 5. Data in a wide format

Couple ID	person ID 1	Name1	Female1	Income1	person ID 2	Name 2	Female 2	Income2
1	1	Lily	1	500	3	Tom	0	700
2	2	Ling	1	1,200	4	Jim	0	1,500

2.

Signs of Data Having Complex Data Structure

- Data have duplicate IDs
- Data have multiple ID variables
- Data have no duplicate IDs nor multiple ID variables, but have variables with similar names

BGSU

Center for Family and Demographic Research

Why Do We Need Complex Data Structure?

Reasons of having such a structure:

- Conceptually necessity: if you want to examine change on individuals over time or understand how higher-level variables influence lower-level variables for individuals
- Analytic requirements: some analytic methods use the wide form of data and some others use the long form of data

Consequences: Complex data structure indicates multiple layers (or unit of observations) in the data. As a result, difficulties exist for the following three tasks:

- Merging data
- Reshaping data
- Generating new variables

Merge Data

- Given the nested structure of data, you often need to combine a lower level of data into a higher level data. You can choose to create a data set into a wide form or a long form.
- Original data

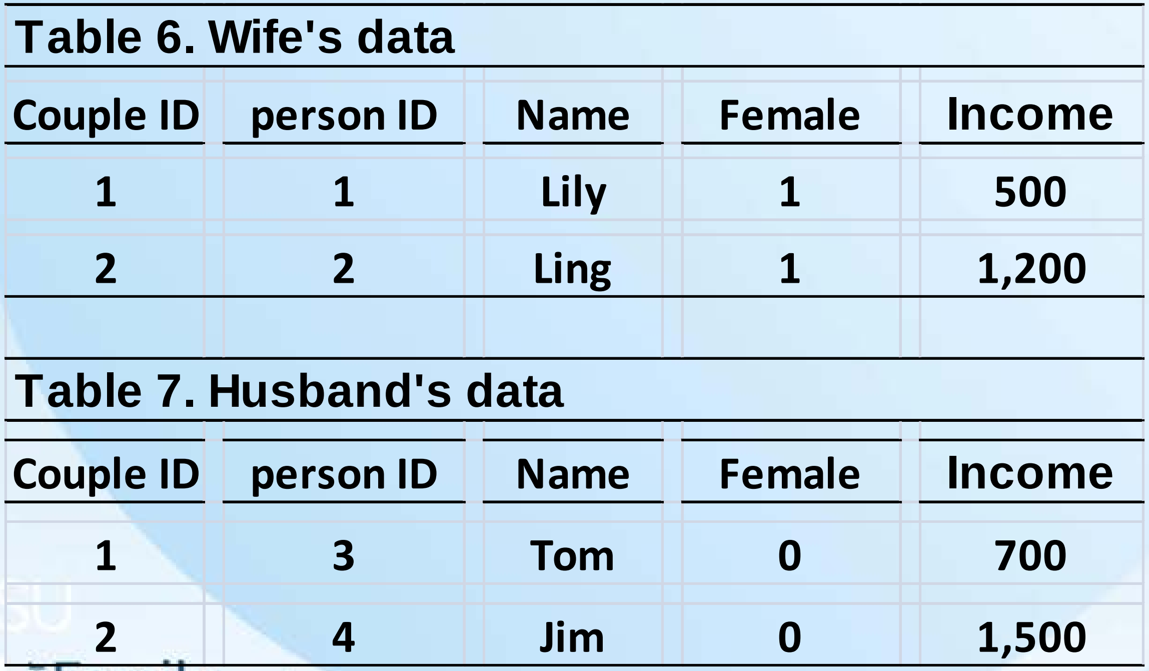

Table 6. Wife's data

Couple ID	person ID	Name	Female	Income
1	1	Lily	1	500
2	2	Ling	1	1,200

Table 7. Husband's data

Couple ID	person ID	Name	Female	Income
1	3	Tom	0	700
2	4	Jim	0	1,500

Merge Data (Continued)

- Merged data in a long format

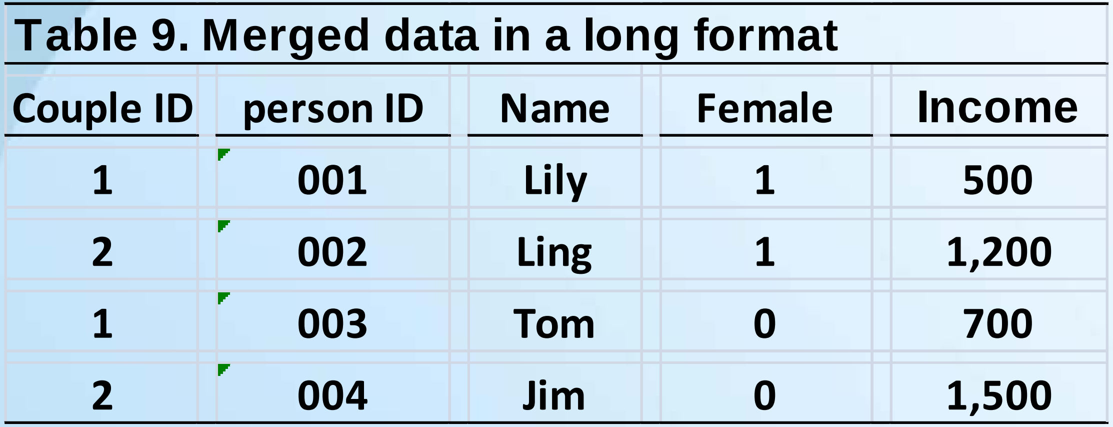

Table 9. Merged data in a long format

Couple ID	person ID	Name	Female	Income
1	001	Lily	1	500
2	002	Ling	1	1,200
1	003	Tom	0	700
2	004	Jim	0	1,500

- Stata commands

```
use c:\temp\wife_data, clear
append using c:\temp\husband_data
save c:\temp\couple_long.dta, replace
```

Merge Data (Continued)

- Merged data in a wide format

Table 8. Merged data in a wide format

Couple ID	person ID 1	Name 1	Female 1	Income1	person ID 2	Name 2	Female 2	Income2
1	1	Lily	1	500	3	Tom	0	700
2	2	Ling	1	1,200	4	Jim	0	1,500

- Stata commands

```
use c:\temp\wife-data, clear
sort couple_id
save, replace
use c:\temp\husband-data, clear
rename  personID personID2
rename  name  name2
rename  female female2
rename  income  income 2
sort couple_id
save, replace
use c:\temp\wife_data, clear
merge couple_id using c:\temp\husband_data
save c:\temp\couple_wide.dta, replace
```

Reshape Data

- Shape data from a long format to a wide format

```
use c:\temp\couple_long.dta
sort coupleID
by coupleID: gen n=_n
reshape wide personID name female income, i(coupleID) j(n)
```

- Reshape data from a wide format to a long format

```
use c:\temp\couple_wide.dta
reshape long name female income, i(coupleID) j(newvar)
```

Generate Variables

Complications: You have data at the lower level, but you want to generate variables at the higher level. You need to use different methods for the wide data format than for the long data format.

- How many households are in the data?
- How many people in each of the household?
- What are the total income of the household?
- What are the average income of the household
- The maximum income of the household?
- Which person in the household has the highest income?

Generate Variables (Continued)

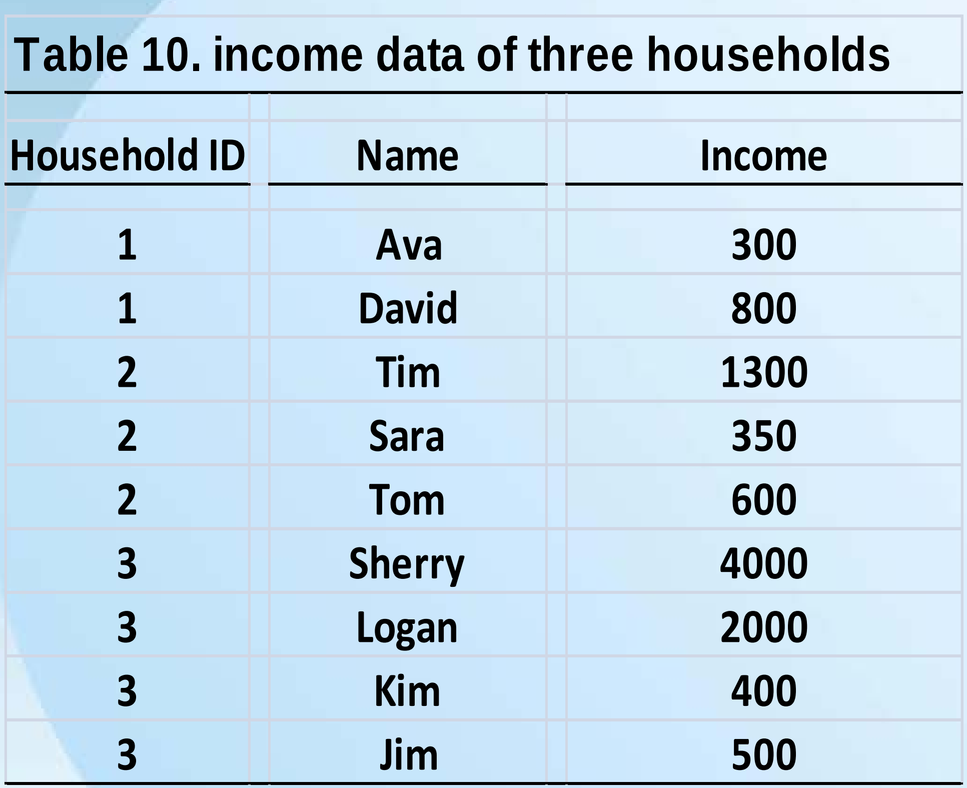

Table 10. income data of three households

Household ID	Name	Income
1	Ava	300
1	David	800
2	Tim	1300
2	Sara	350
2	Tom	600
3	Sherry	4000
3	Logan	2000
3	Kim	400
3	Jim	500

Generate Variables (Continued)

- For data in a long format

```
sort  household_id
by  household_id: gen n=_n
by  household_id: gen N=_N
by  household_id: egen t_income = sum(income)
by  household_id: egen m_income = mean(income)
by  household_id: egen max_income = max(income)
list name if income == max_income
```

Generate Variables (Continued)

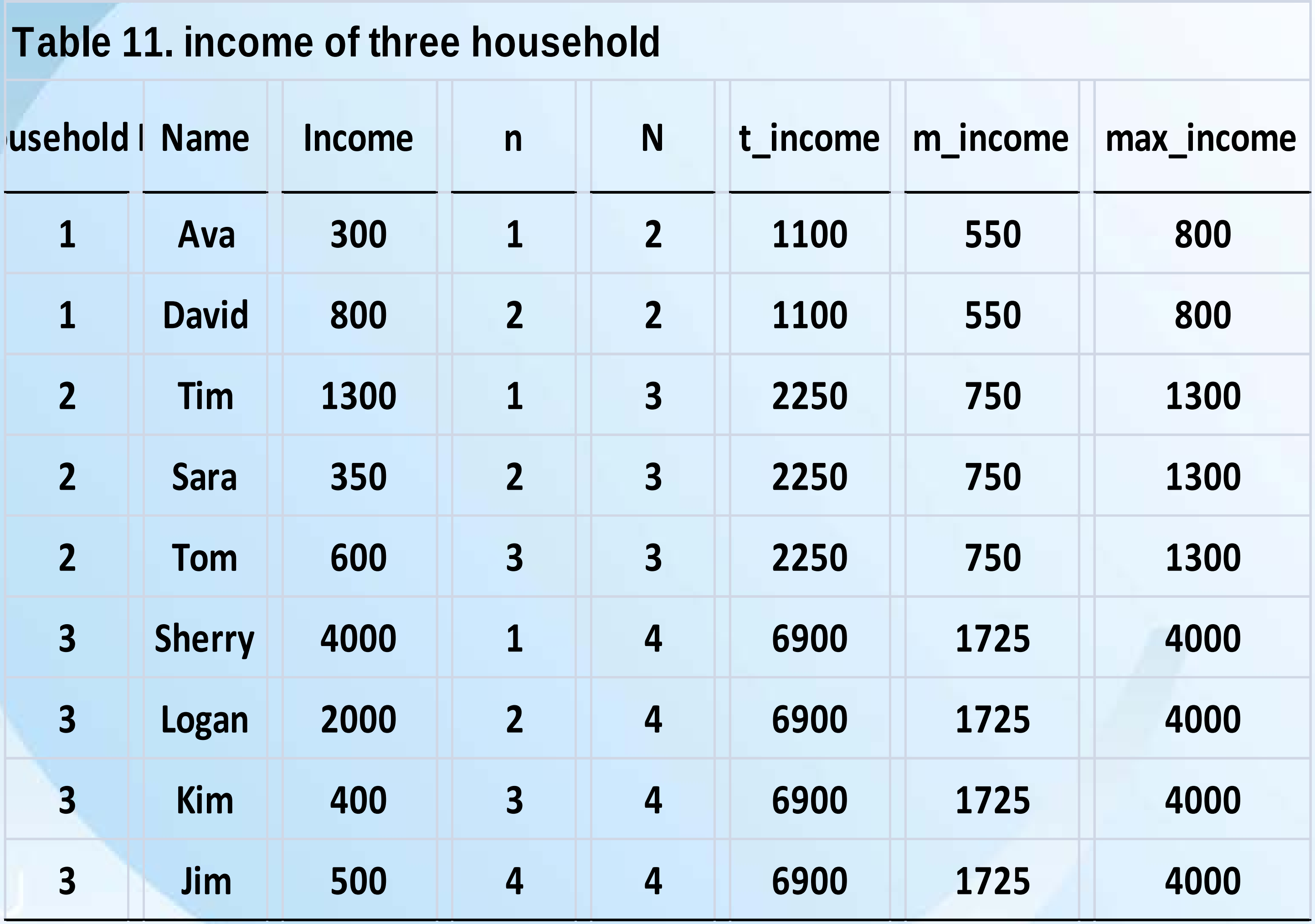

Table 11. income of three household

usehold I	Name	Income	n	N	t_income	m_income	max_income
1	Ava	300	1	2	1100	550	800
1	David	800	2	2	1100	550	800
2	Tim	1300	1	3	2250	750	1300
2	Sara	350	2	3	2250	750	1300
2	Tom	600	3	3	2250	750	1300
3	Sherry	4000	1	4	6900	1725	4000
3	Logan	2000	2	4	6900	1725	4000
3	Kim	400	3	4	6900	1725	4000
3	Jim	500	4	4	6900	1725	4000

Generate Variables (Continued)

For data in a wide format

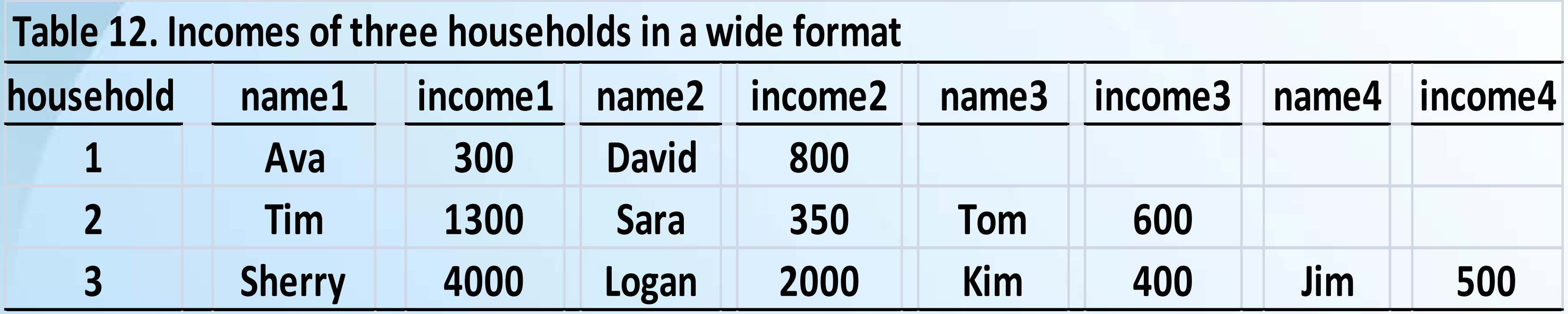

Table 12. Incomes of three households in a wide format

household	name1	income1	name2	income2	name3	income3	name4	income4
1	Ava	300	David	800				
2	Tim	1300	Sara	350	Tom	600		
3	Sherry	4000	Logan	2000	Kim	400	Jim	500

Generate Variables (Continued)

For data in a wide format

```
tab1 household_id
egen N=rownonmiss(income1 income2 income3 income4)
egen t_income = rowtotal( income1 income2 income3 income4)
egen m_income = rowmean( income1 income2 income3 income4)
egen max_income = rowmax( income1 income2 income3 income4)
```

Generate Variables (Continued)

```
gen income=income1
gen name=name1

program define loop
local i = 2
while `i' <= 4 {
replace name=name`i' if income`i'>income & income`i'~=.
replace income=income`i' if income`i'>income & income`i'~=.
local i = `i'+1
}
end
quietly loop
program drop loop
```

Generate Variables (Continued)

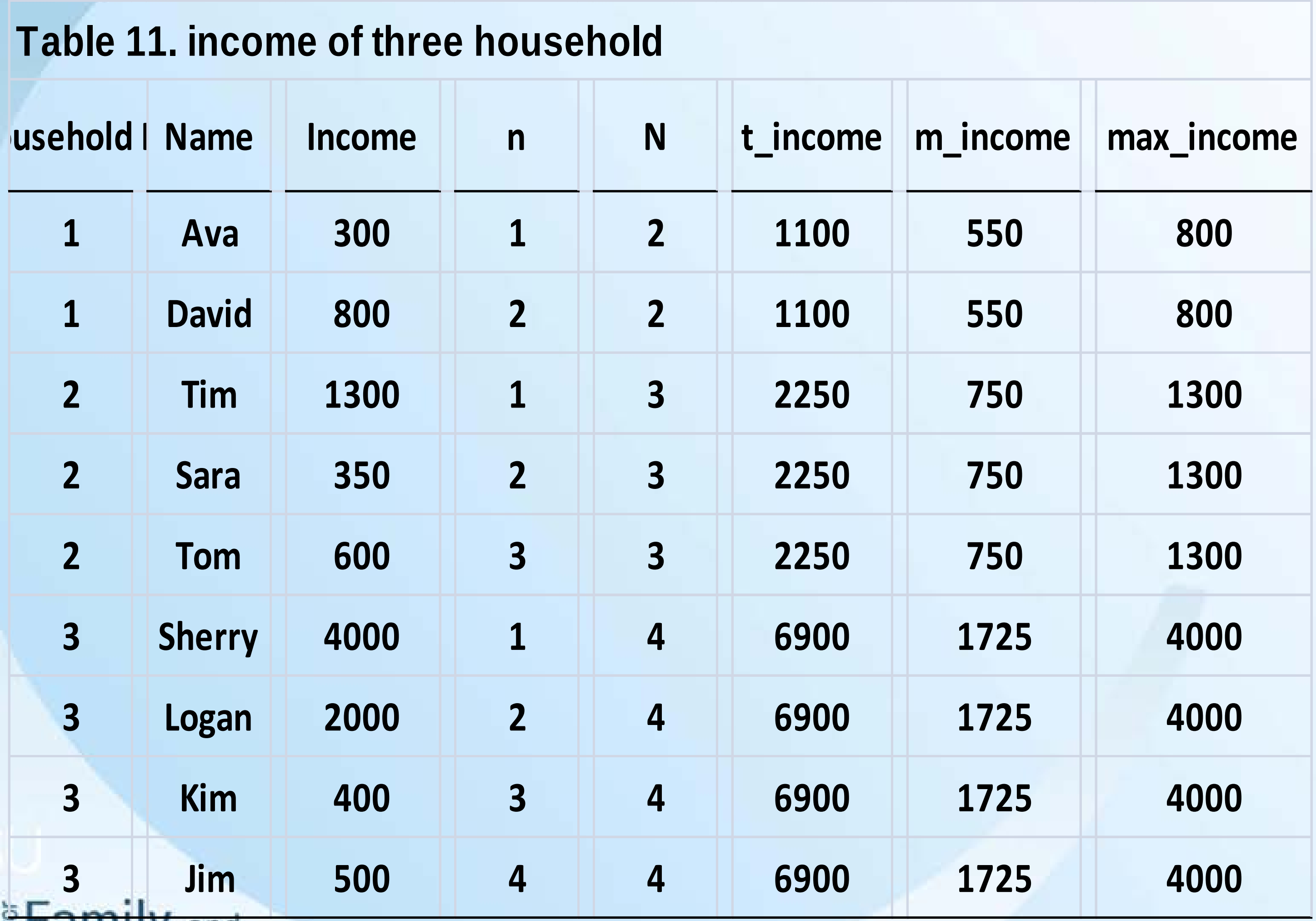

Table 11. income of three household

usehold I	Name	Income	n	N	t_income	m_income	max_income
1	Ava	300	1	2	1100	550	800
1	David	800	2	2	1100	550	800
2	Tim	1300	1	3	2250	750	1300
2	Sara	350	2	3	2250	750	1300
2	Tom	600	3	3	2250	750	1300
3	Sherry	4000	1	4	6900	1725	4000
3	Logan	2000	2	4	6900	1725	4000
3	Kim	400	3	4	6900	1725	4000
3	Jim	500	4	4	6900	1725	4000

Conclusions

- Complex data structure is necessary both conceptually and analytically.
- Complex data structure implies multiple layers of data, which creates complexities in merging data, reshaping data, and generating variables.
- Data in different formats requires different programming syntax.
- If you have questions in managing data in complex structures, contact Hsueh-Sheng wu at wuh@bgsu.edu or 372-3119