REACH Fantasy Statistics #1

How and why to calculate within-subject variance and between-subject variance in EMA/Multilevel data

Wei-Lin Wang

2020.04.03

Outline

- Overview
- The definition of WSV and BSV
- The issues of calculating BSV
- The strategies for dealing with the issues
- Coding examples
- More about WS and BS decomposition

Overview

- In this mini talk, we will discuss what kinds of issues that we may encounter when using WSV and BSV. Then, we will learn different strategies to handle the issues. The goal of the talk is to let everyone to have the knowledge and the skills to compute the WSV and BSV properly.
- We may go above and beyond the calculation of the WSV and BSV, and discuss more about when we should use WS and BS decomposition if we have enough time.

What is WSV and BSV?

- WSV means within-subject variance, and it refers to the deviance from the subject (group) mean.

- BSV means between-subject variance, and it refers to the deviance from the population (grand) mean.

Within-subject variance

- The formula:

$$WSV_{ij} = x_{ij} - \bar{x}_j$$

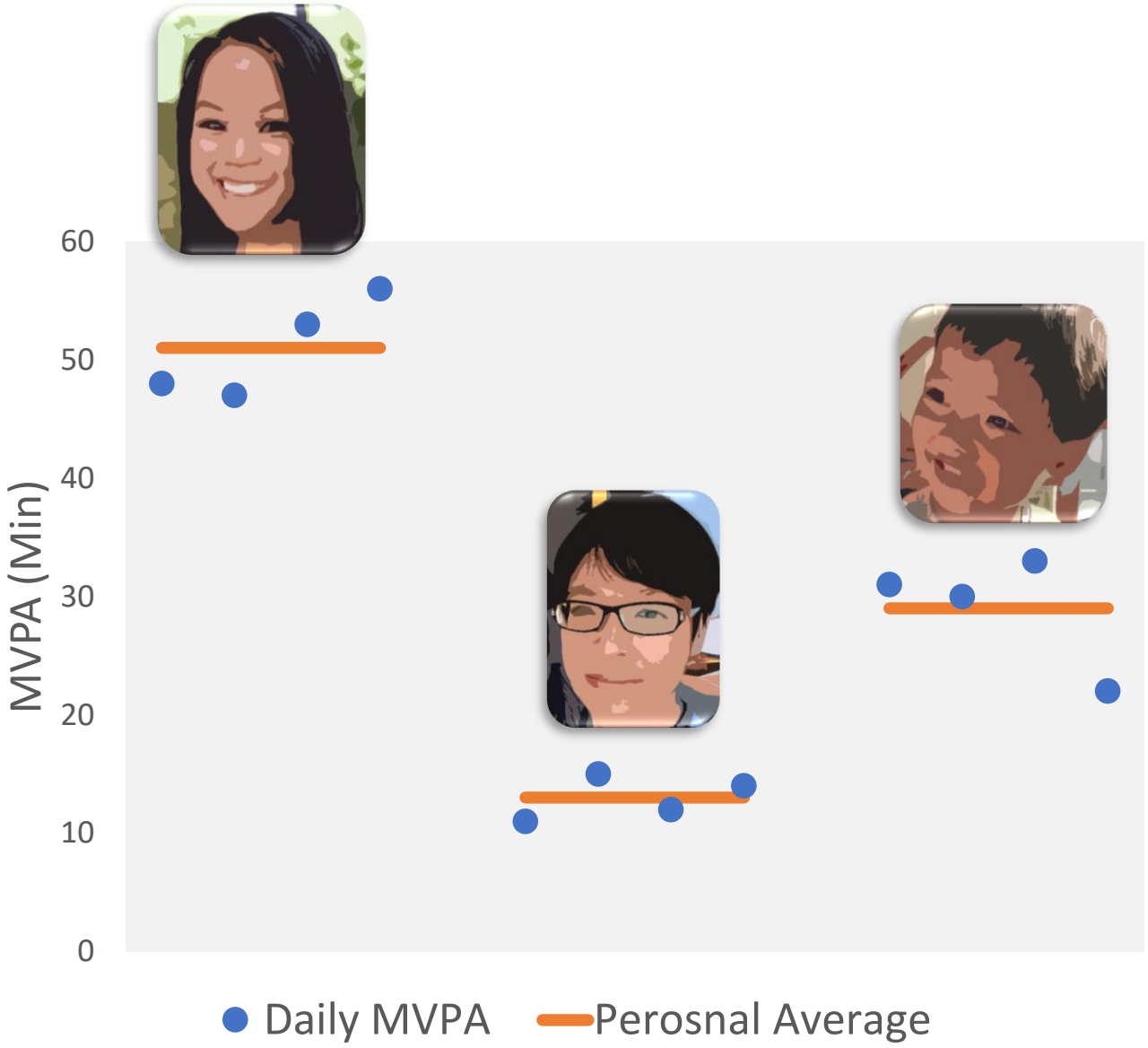
$$\bar{x}_j = \frac{\sum x_{ij}}{n_{ij}}$$

Where i refers to the index for the observation (eg., prompt)

j refers to the index for the subject (eg., person)

Example Data

Subject	Day_Num	MVPA
BD	1	48
BD	2	47
BD	3	53
BD	4	56
WLW	1	11
WLW	2	15
WLW	3	12
WLW	4	14
ST	1	31
ST	2	30
ST	3	33
ST	4	22



Within-subject variance example

Subject	Day_Num	MVPA	Group Mean	WSV
BD	1	48	51	-3
BD	2	47	51	-4
BD	3	53	51	2
BD	4	56	51	5
WLW	1	11	13	-2
WLW	2	15	13	2
WLW	3	12	13	-1
WLW	4	14	13	1
ST	1	31	29	2
ST	2	30	29	1
ST	3	33	29	4
ST	4	22	29	-7

Between-subject variance

- The formula:

$$BSV_j = \bar{x}_j - \bar{x}_{grand}$$

$$\bar{x}_{grand} = \frac{\sum \bar{x}_j}{n_j}$$

Where j refers to the index for the subject (eg., person)

Between-subject variance

- The formula:

$$BSV_j = \bar{x}_j - \bar{x}_{grand}$$

$$\bar{x}_{grand} = \frac{\sum \bar{x}_j}{n_j}$$

It's the average of subject average.

Where j refers to the index for the subject (eg., person)

Example Data

Subject	Day_Num	MVPA
BD	1	48
BD	2	47
BD	3	53
BD	4	56
WLW	1	11
WLW	2	15
WLW	3	12
WLW	4	14
ST	1	31
ST	2	30
ST	3	33
ST	4	22

However, the subjects in EMA data are at level two.

The data format is long format, which means each row is one time point per subject.

Between-subject variance

- The formula:

$$BSV_j = \bar{x}_j - \bar{x}_{grand}$$

$$\bar{x}_{grand} = \frac{\sum \bar{x}_j}{n_j} \approx \frac{\sum_{j=1}^{k} \left(\sum x_{ij} \right)}{\sum_{j=1}^{k} n_{ij}}$$

Where j refers to the index for the subject (eg., person)

k refers to the maximum number of the subject

Between-subject variance

- The formula:

$$BSV_j = \bar{x}_j - \bar{x}_{grand}$$

$$\bar{x}_{grand} = \frac{\sum \bar{x}_j}{n_j} \approx \boxed{\frac{\sum_{j=1}^{k}\left(\sum x_{ij}\right)}{\sum_{j=1}^{k} n_{ij}}}$$

Raw Grand Mean (Unweighted)

Where j refers to the index for the subject (eg., person)

k refers to the maximum number of the subject

Between-subject variance example

Subject	Day_Num	MVPA	Group Mean	Grand Mean	BSV
BD	1	48	51	31	20
BD	2	47	51	31	20
BD	3	53	51	31	20
BD	4	56	51	31	20
WLW	1	11	13	31	-18
WLW	2	15	13	31	-18
WLW	3	12	13	31	-18
WLW	4	14	13	31	-18
ST	1	31	29	31	-2
ST	2	30	29	31	-2
ST	3	33	29	31	-2
ST	4	22	29	31	-2

Between-subject variance example

Subject	Day_Num	MVPA	Group Mean	Grand Mean	BSV
BD	1	48	51	31	20
BD	2	47	51	31	20
BD	3	53	51	31	20
BD	4	56	51	31	20
WLW	1	11	13	31	-18
WLW	2	15	13	31	-18
WLW	3	12	13	31	-18
WLW	4	14	13	31	-18
ST	1	31	29	31	-2
ST	2	30	29	31	-2
ST	3	33	29	31	-2
ST	4	22	29	31	-2

$$\text{Grand Mean} = \frac{(48 + 47 + \cdots + 33 + 22)}{12} = 31$$

Between-subject variance example

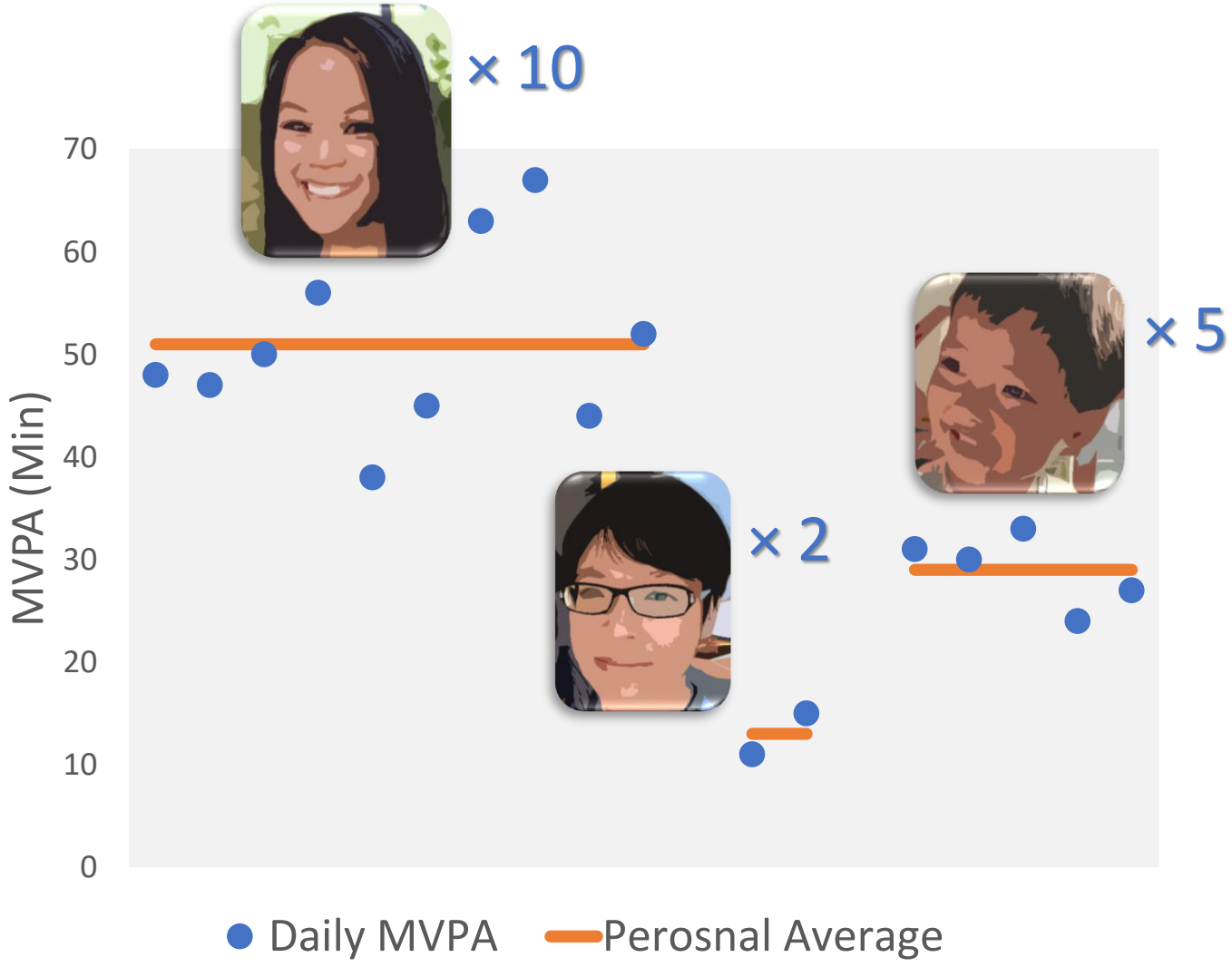
Subject	Day_Num	MVPA	Group Mean	Grand Mean	BSV
BD	1	48	51	31	20
BD	2	47	51	31	20
BD	3	53	51	31	20
BD	4	56	51	31	20
WLW	1	11	13	31	-18
WLW	2	15	13	31	-18
WLW	3	12	13	31	-18
WLW	4	14	13	31	-18
ST	1	31	29	31	-2
ST	2	30	29	31	-2
ST	3	33	29	31	-2
ST	4	22	29	31	-2

$$\text{Grand Mean} = \frac{(48 + 47 + \cdots + 33 + 22)}{12} = 31$$

Looks Good, Right?

Subject	Day_Num	MVPA
BD	1	48
BD	2	47
BD	3	50
BD	4	56
BD	5	38
BD	6	45
BD	7	63
BD	8	67
BD	9	44
BD	10	52
WLW	1	11
WLW	2	15
ST	1	31
ST	2	30
ST	3	33
ST	4	24
ST	5	27

Unbalanced data structure



Subject	Day_Num	MVPA	Group Mean	Raw Grand Mean	Raw BSV
BD	1	48	51	40.1	10.9
BD	2	47	51	40.1	10.9
BD	3	50	51	40.1	10.9
BD	4	56	51	40.1	10.9
BD	5	38	51	40.1	10.9
BD	6	45	51	40.1	10.9
BD	7	63	51	40.1	10.9
BD	8	67	51	40.1	10.9
BD	9	44	51	40.1	10.9
BD	10	52	51	40.1	10.9
WLW	1	11	13	40.1	-27.1
WLW	2	15	13	40.1	-27.1
ST	1	31	29	40.1	-11.1
ST	2	30	29	40.1	-11.1
ST	3	33	29	40.1	-11.1
ST	4	24	29	40.1	-11.1
ST	5	27	29	40.1	-11.1

Subject	Day_Num	MVPA	Group Mean	Raw Grand Mean	Raw BSV
BD	1	48	51	40.1	10.9
BD	2	47	51	40.1	10.9
BD	3	50	51	40.1	10.9
BD	4	56	51	40.1	10.9
BD	5	38	51	40.1	10.9
BD	6	45	51	40.1	10.9
BD	7	63	51	40.1	10.9
BD	8	67	51	40.1	10.9
BD	9	44	51	40.1	10.9
BD	10	52	51	40.1	10.9
WLW	1	11	13	40.1	-27.1
WLW	2	15	13	40.1	-27.1
ST	1	31	29	40.1	-11.1
ST	2	30	29	40.1	-11.1
ST	3	33	29	40.1	-11.1
ST	4	24	29	40.1	-11.1
ST	5	27	29	40.1	-11.1

$$\text{Grand Mean} = \frac{(48 + 47 + \cdots + 24 + 27)}{17} = 40.1$$

Subject	Day_Num	MVPA	Group Mean	Raw Grand Mean	Raw BSV
BD	1	48	51	40.1	10.9
BD	2	47	51	40.1	10.9
BD	3	50	51	40.1	10.9
BD	4	56	51	40.1	10.9
BD	5	38	51	40.1	10.9
BD	6	45	51	40.1	10.9
BD	7	63	51	40.1	10.9
BD	8	67	51	40.1	10.9
BD	9	44	51	40.1	10.9
BD	10	52	51	40.1	10.9
WLW	1	11	13	40.1	-27.1
WLW	2	15	13	40.1	-27.1
ST	1	31	29	40.1	-11.1
ST	2	30	29	40.1	-11.1
ST	3	32	29	40.1	-11.1
ST	4	24	29	40.1	-11.1
ST	5	27	29	40.1	-11.1

$$\text{Grand Mean} = \frac{(48 + 47 + \cdots + 24 + 27)}{17} = 40.1$$

Something Wrong!?

2

Raw Grand Mean

10

Subject	Day_Num	MVPA	Group Mean	Raw Grand Mean	Grand Mean	BSV
BD	1	48	51	40.1	31	20
BD	2	47	51	40.1	31	20
BD	3	50	51	40.1	31	20
BD	4	56	51	40.1	31	20
BD	5	38	51	40.1	31	20
BD	6	45	51	40.1	31	20
BD	7	63	51	40.1	31	20
BD	8	67	51	40.1	31	20
BD	9	44	51	40.1	31	20
BD	10	52	51	40.1	31	20
WLW	1	11	13	40.1	31	-18
WLW	2	15	13	40.1	31	-18
ST	1	31	29	40.1	31	-2
ST	2	30	29	40.1	31	-2
ST	3	33	29	40.1	31	-2
ST	4	24	29	40.1	31	-2
ST	5	27	29	40.1	31	-2

Subject	Day_Num	MVPA	Group Mean	Raw Grand Mean	Grand Mean	BSV
BD	1	48	51	40.1	31	20
BD	2	47	51	40.1	31	20
BD	3	50	51	40.1	31	20
BD	4	56	51	40.1	31	20
BD	5	38	51	40.1	31	20
BD	6	45	51	40.1	31	20
BD	7	63	51	40.1	31	20
BD	8	67	51	40.1	31	20
BD	9	44	51	40.1	31	20
BD	10	52	51	40.1	31	20
WLW	1	11	13	40.1	31	-18
WLW	2	15	13	40.1	31	-18
ST	1	31	29	40.1	31	-2
ST	2	30	29	40.1	31	-2
ST	3	29	29	40.1	31	-2
ST	4	24	29	40.1	31	-2
ST	5	27	29	40.1	31	-2

$$\text{Grand Mean} = \frac{(51 + 13 + 29)}{3} = 31$$

Unbiased Estimate

Issues of computing raw grand mean

- The estimate of raw grand mean is problematic when there is an unbalanced structure. We all know the structure of EMA data are most likely to be unbalanced.
- The estimation could be even more biased when there is an association between data structure and factors which we are interested in.

MATCH – Mother data

		Positive affect
Prompts by wave	<= 20	2.48
	21 - 30	2.56
	> 30	2.64

		Window aggregated MVPA minutes [-120m, +120m]
Prompts by wave	<= 20	5.08
	21 - 30	6.01
	> 30	5.95

Strategies for dealing with unbalanced data

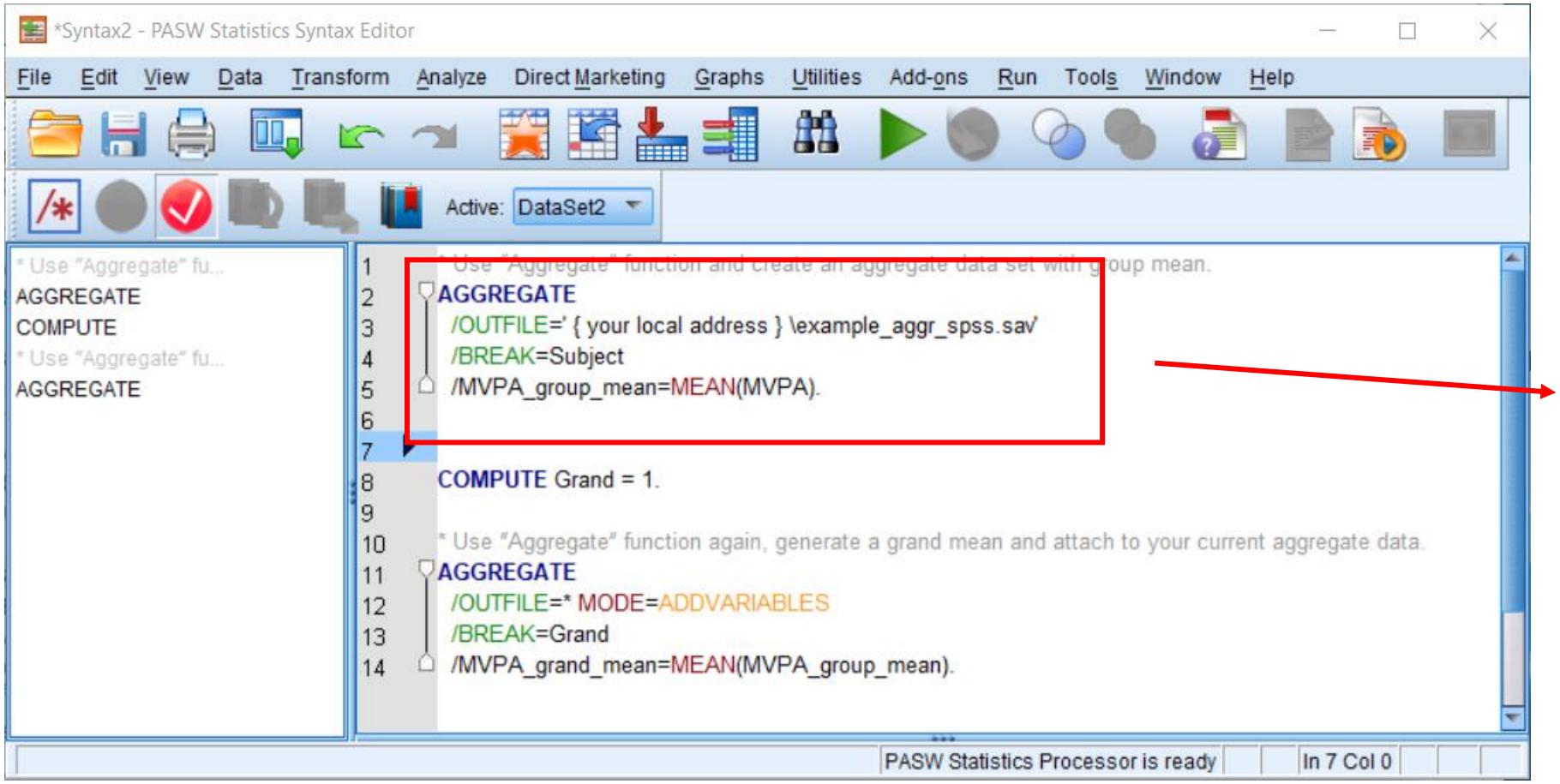
Main idea is to allow everyone to have an “equal voice” in the data set and calculate an unbiased estimate of the grand mean.

1. Two-stage aggregate method
2. Weighting approach

Two-stage aggregate method (SPSS)

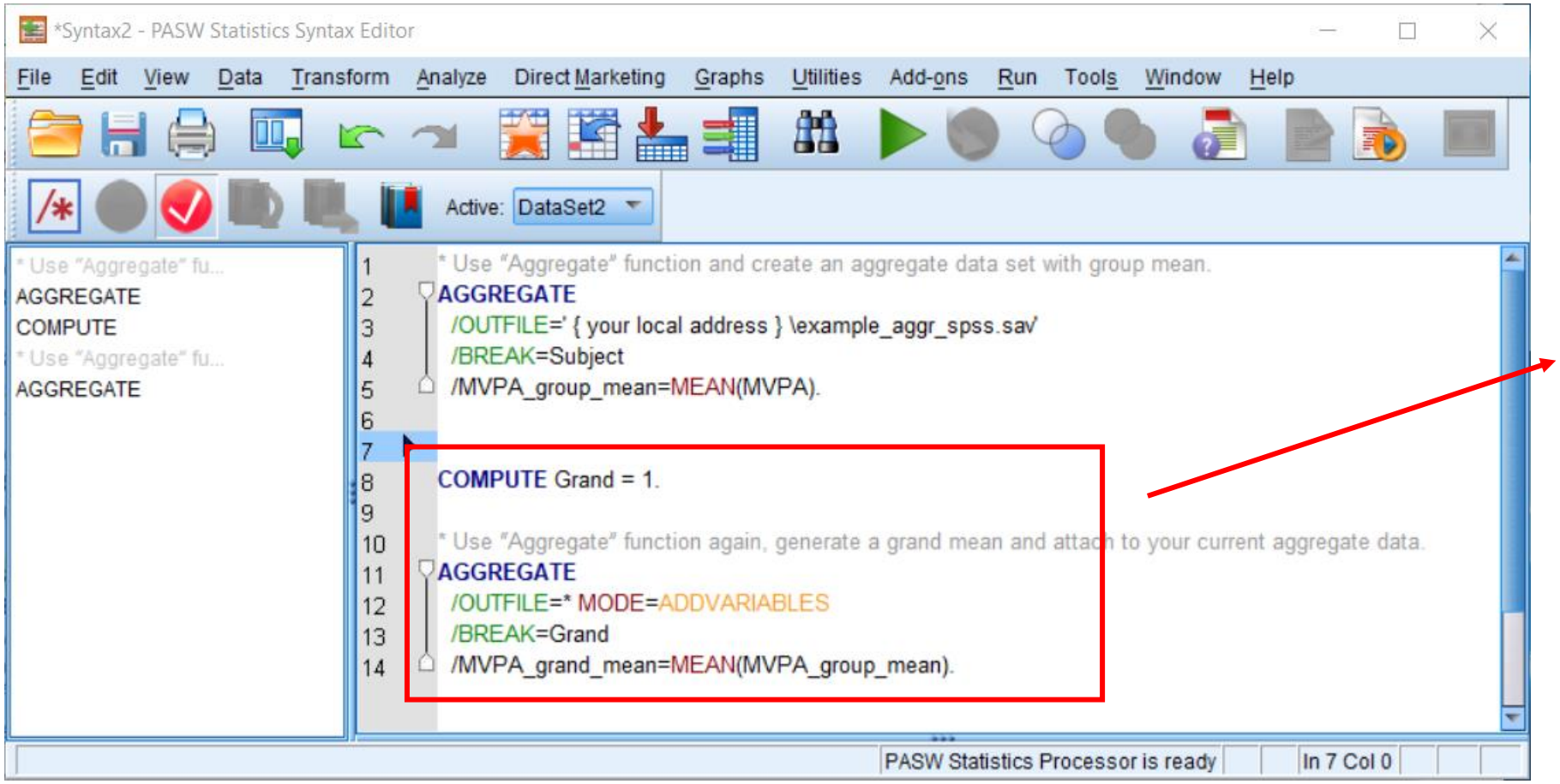
- Aggregate method is to obtain the grand meaning from changing/aggregating data structure.
- In the new data, every subject just has an aggregated observation.
- By changing the data structure to the higher level (subject level), we could calculate the grand mean directly.

Two-stage aggregate method (SPSS)



Use "Aggregate" function and create an aggregate data with group mean.

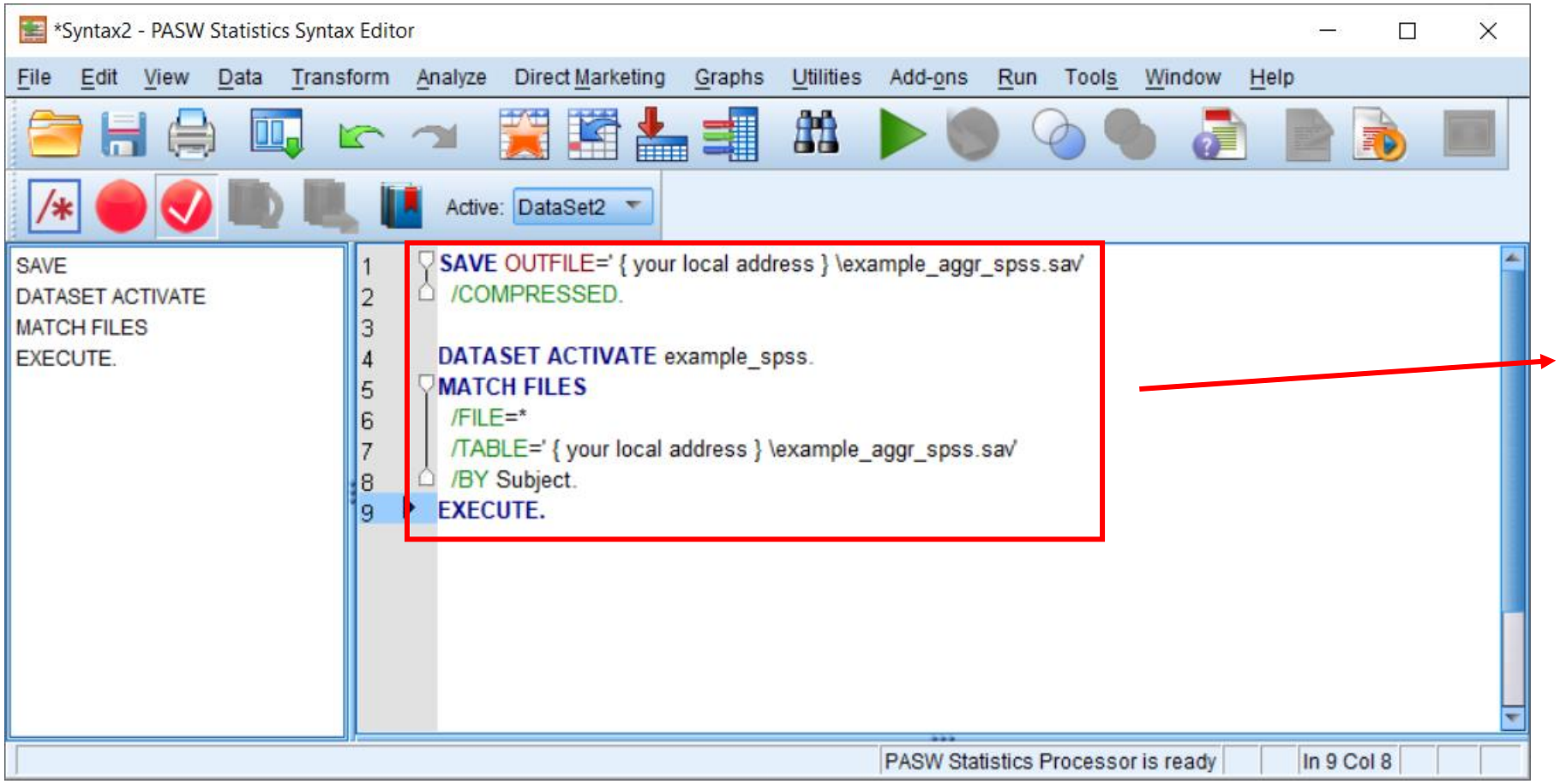
Two-stage aggregate method (SPSS)



Compute a grand group in advance so that you can aggregate the data by the whole group.

Use “Aggregate” function again and generate a grand mean.

Two-stage aggregate method (SPSS)



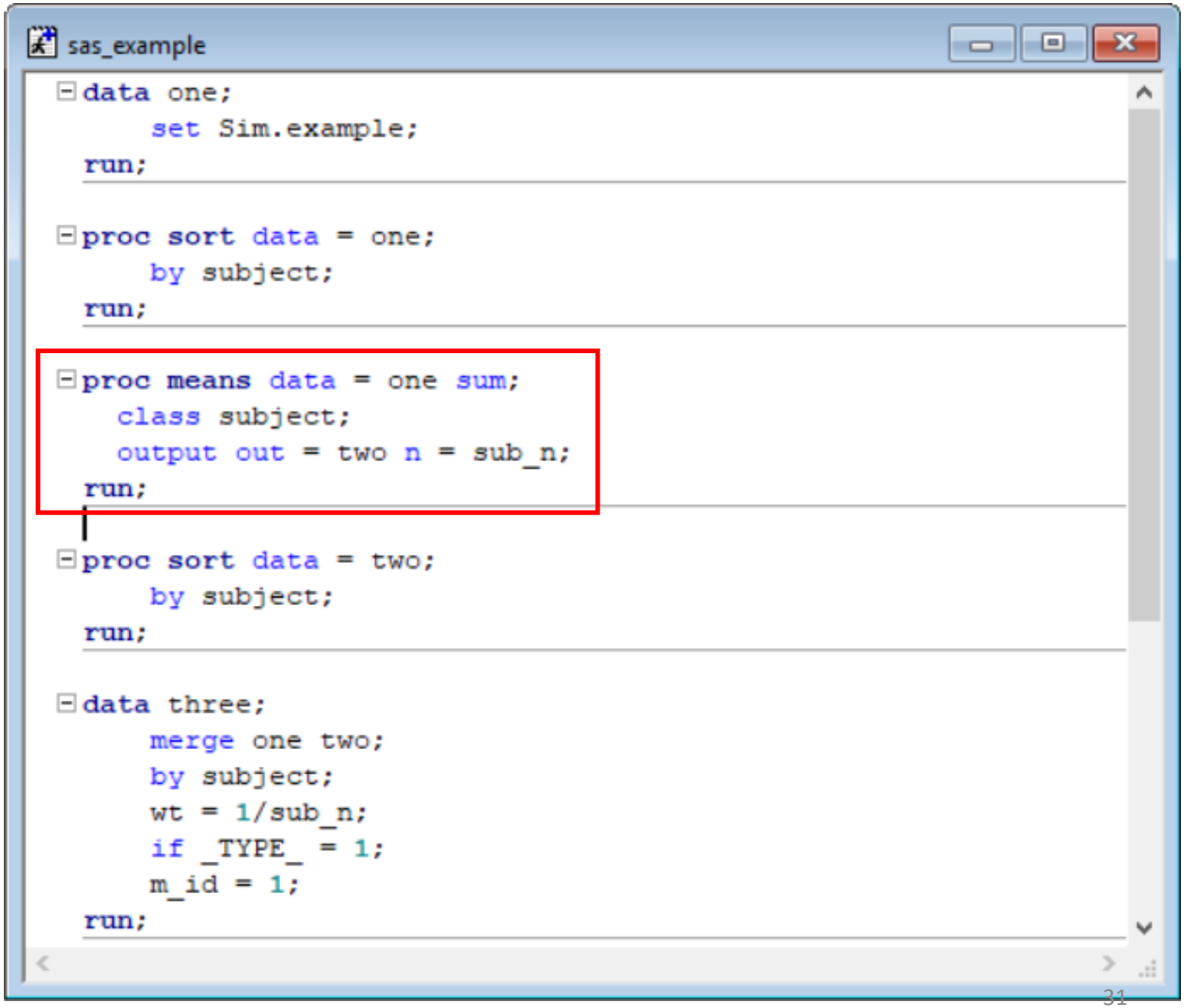
Merge the aggregate data set with the main data set later so you could calculate BSV.

Weighting (SAS)

- Weighting is a method that we give every subject an equal voice by reversing the sampling fraction – the probability of ending up in the sample/data.
- We will apply “normalized weights” or “standardized weights”.
- In this case, the sum of weights in the data set equals the size of the sample at subject level.
- The idea of weighting is to calculate the grand mean under the same data structure/format. However, the trick is that the estimate of grand mean is adjusted by the weights.

Weighting (SAS)

Use “Means” function and generate a new data set with count number by subject.



Weighting (SAS)

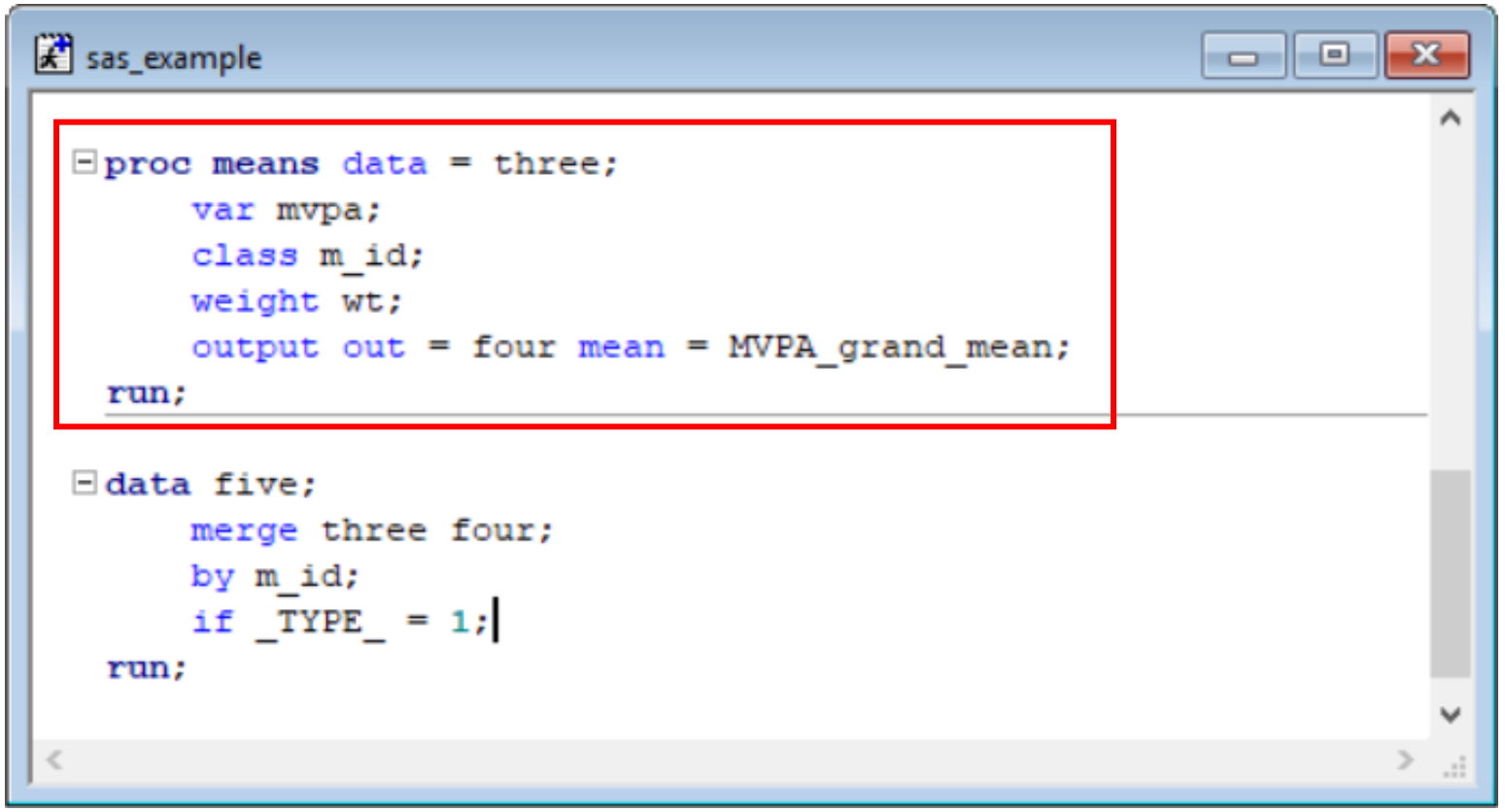
Calculate weight (wt).
Weights are the inverse of the observation count.

sas_example

```
data one;
     set Sim.example;
run;

proc sort data = one;
     by subject;
run;

proc means data = one sum;
   class subject;
   output out = two n = sub_n;
run;

proc sort data = two;
     by subject;
run;

data three;
     merge one two;
     by subject;
     wt = 1/sub_n;
     if _TYPE_ = 1;
     m_id = 1;
run;
```

Weighting (SAS)

Estimate the grand mean by factoring weights into account and save it as a new variable.



Tips

- To generate grand mean variable needs to use “merge” function.
- It is important to make sure that the data sets we would like to merge have the same key/index variable to match and the variable has been sorted before merging.
- The way I use SPSS to do “two-stage aggregate method” and SAS to do “weighting approach” is just an example. Actually, SPSS can do weighting and SAS can do two-stage aggregate method as well. The method and software are all interchangeable.

When do we need WS and BS decomposition

- The intraclass correlation (ICC) is a common measure of WS and BS effects.

$$ICC = \frac{Var_{between}}{Var_{between} + Var_{within}}$$

- Ranges from

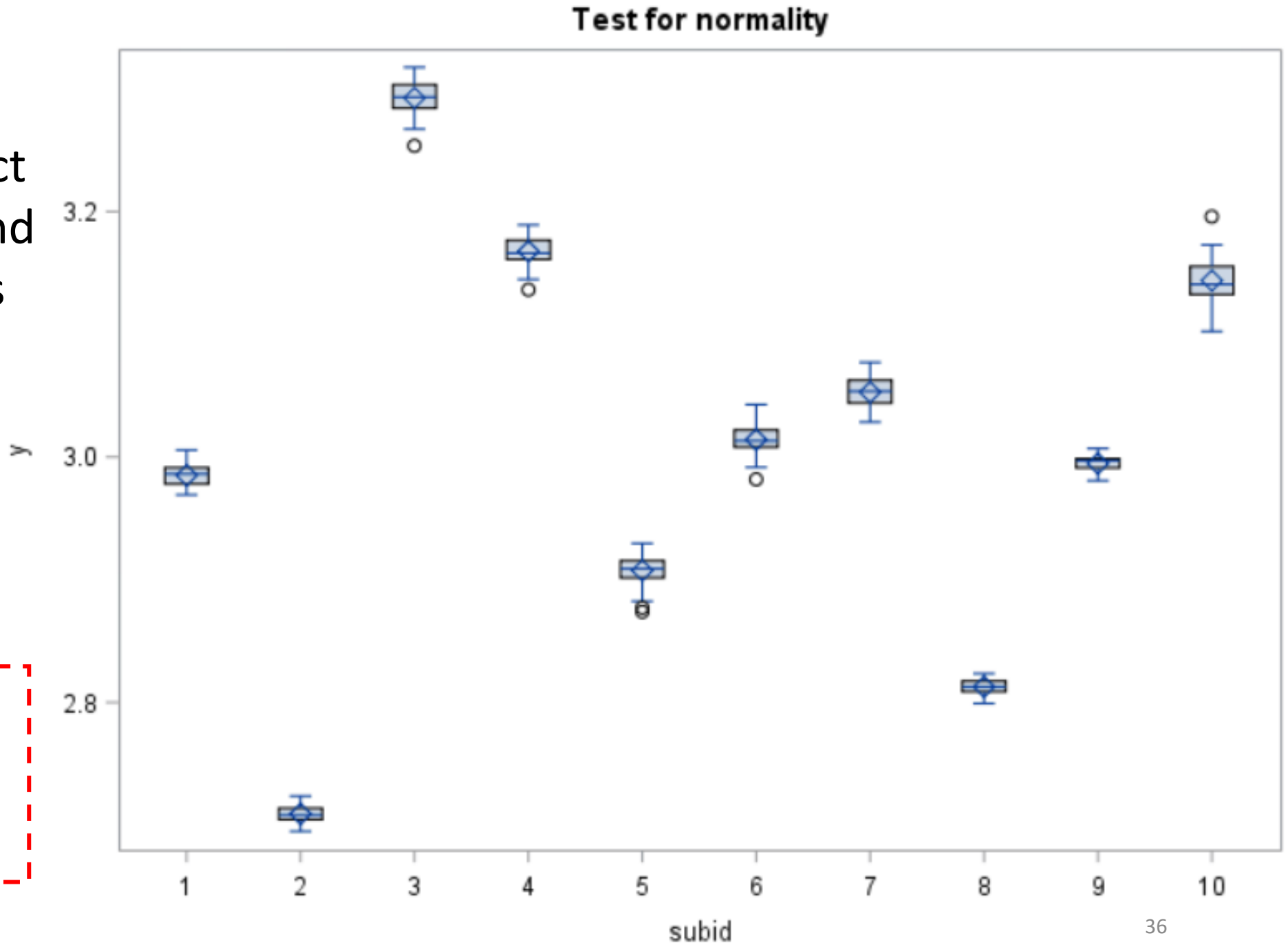
- Zero: each subject is a microcosm.

to

- One: subjects are very different between each other.

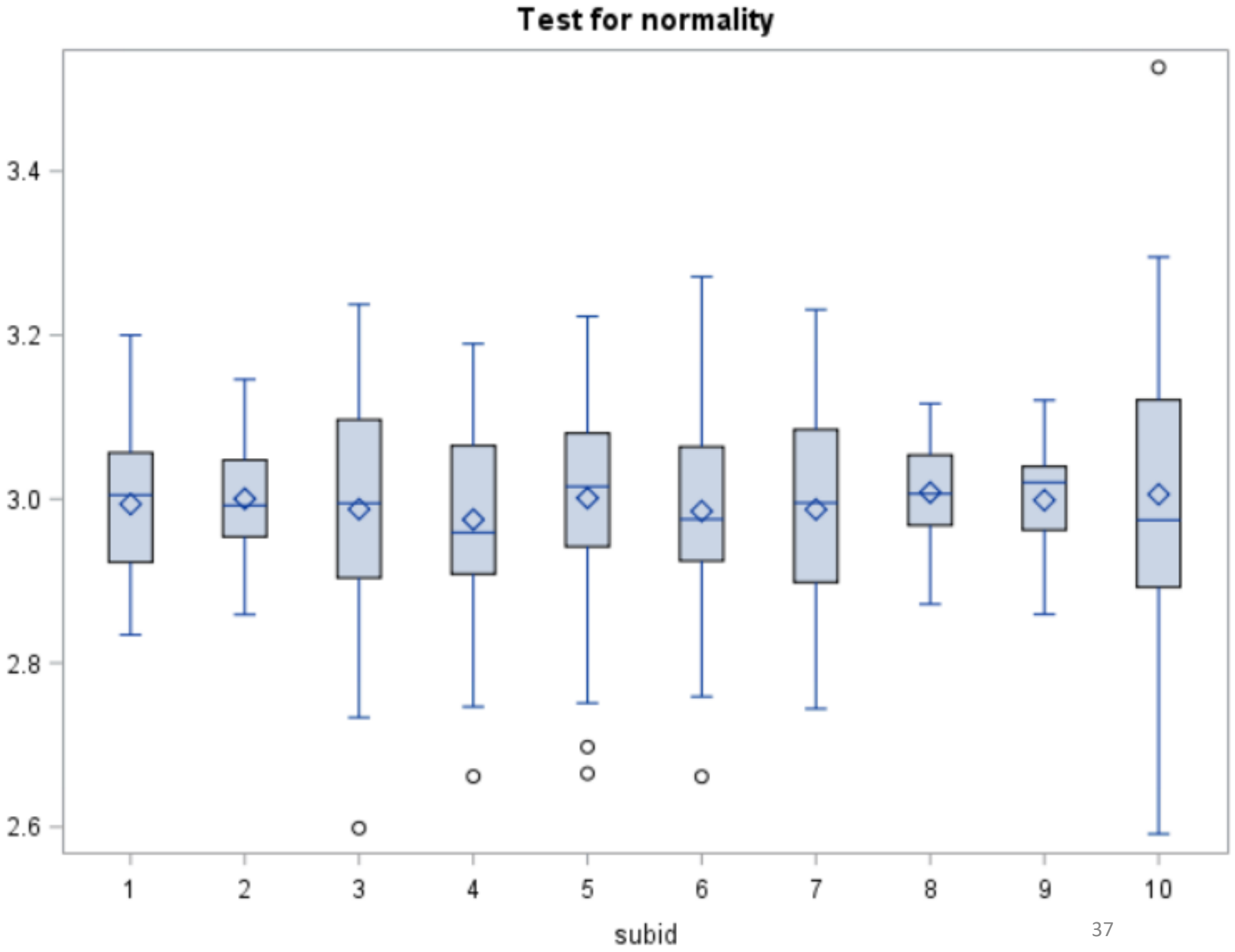
- The subject effect is very strong, and the model needs to control BS effect.

Data Simulation
Grand Mean = 3.0
ICC = 0.99



- Each subject is a microcosm of population.
- The BS and WS-decomposition has no effect on statistical analysis.

Data Simulation
Grand Mean = 3.0
ICC = 0.01



Take home message

1. Need to check data structure and the association between the quantity of observations and the variables of interest.
2. WSV and BSV are very sensitive to the data. Be sure to clean the data before doing data analyses/processing.
3. Use proper methods to treat everyone equally when calculating grand mean and BSV.
4. Use ICC to check if WS and BS decomposition is a better option for the statistical model.

Reference

- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in medicine*, *31*(27), 3328-3336.
- Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *american Journal of political Science*, 218-237.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99-114.