Monte Carlo Simulation Practice

A replication of analyses presented in Muthén & Muthén (2002)

Adam Garber

1/28/2020

All analyses are conducted in `R/RStudio` with data & models estimated in `Mplus` using the package `MplusAutomation`

This tutorial closely follows the concepts and model syntax presented here:

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural equation modeling, 9(4), 599-620.

The associated Mplus syntax can be found here: http://www.statmodel.com/examples/penn.shtml

This tutorial relies heavily on the functionality provided by the R package `MplusAutomation`:

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. Structural equation modeling: a multidisciplinary journal, 25(4), 621-638.

Create an R-Project:

Download repository here: https://github.com/garberadamc/SIM-REPLICATION

This project contains the following 4 sub-folders:

a. "figures"
b. "mplus_files"
c. "mplus_bias"
d. "mplus_tune"

IF package rhdf5 does not load then run:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("rhdf5")
```

```
library(tidyverse)
library(MplusAutomation)
library(rhdf5)
library(here)
library(glue)
library(psych)
library(gt)
```

loading packages...

Practicing Monte Carlo Simulation

Factors that influence minimum sample size requirements:

- size of the model (number of parameters)
- distribution of the variables (skew, kurtosis, multi-modal..)
- amount of missing data & pattern of missingness
- reliability of the variables
- strength of the relations among the variables

Simulation purpose: "This article focuses on parameter estimates, standard errors, coverage, and power assuming correctly specified models. Misspecified models can also be studied in the Mplus Monte Carlo framework but are not included here." - Muthen & Muthen (2002)

CFA Model example:

- 2 factors
- 10 indicators (5 per factor)
- 31 free parameters & 24 *df*
- factor loadings = 0.8 (freely estimated in the model)
- residual variances = 0.36 (error)
- factor variances = 1 (fixed)
- reliability of factor = 0.64 = .8^2(1) /.8^2(1 + 0.36)

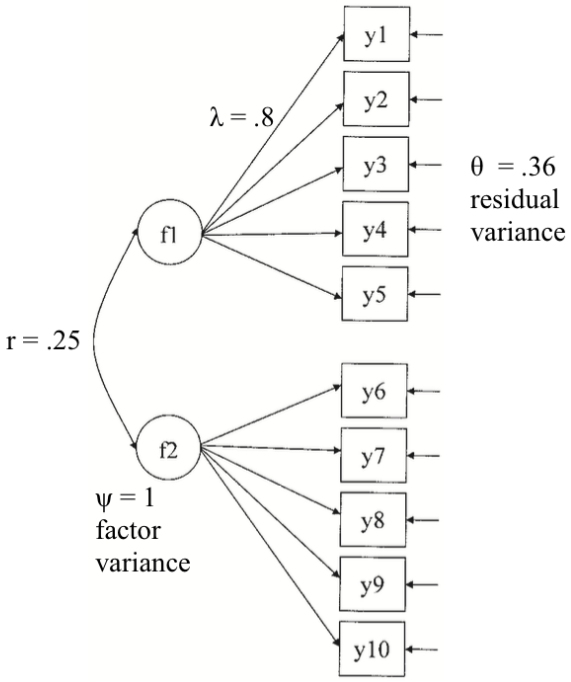


Figure 1. CFA model example. Picture adapted from, Muthen & Muthen 2002

Monte Carlo conditions:

1. normally distributed continuous factor indicators without missing data
2. normally distributed continuous factor indicators with missing data
3. non-normal continuous factor indicators without missing data
4. non-normal continuous factor indicators with missing data

Missing data:

"... all participants have data on y1, y2, y3, y4, and y5, and 50% of the participants have data on y6, y7, y8, y9, and y10" (Muthen & Muthen, 2002).

Non-normal data (how to create variables with skew & kurtosis):

Muthen & Muthen (2002):

"non-normal data are created using a mixture of two normal sub-populations or classes of individuals. Normal data are generated for two classes that have different means and variances for the factor indicators. The combined data are analyzed as though they come from a single population."

[...]

"The first step is to generate data for two classes such that the combination of the data from the two classes **has the desired skewness and kurtosis**."

[...]

"For the CFA model with non-normal data, Class 1, the outlier class, contains 12% of the participants and Class 2 contains the remaining 88%. Only the factor indicators for the second factor are non-normal. Therefore, the Class 1 mean for the second factor is chosen to be 15 and the variance 5, as compared to the Class 2 mean and variance of zero and 1. The resulting population univariate skewness for variables y6 through y10 is 1.2. The resulting population univariate kurtosis for variables y6 through y10 ranges from 1.5 to 1.6."

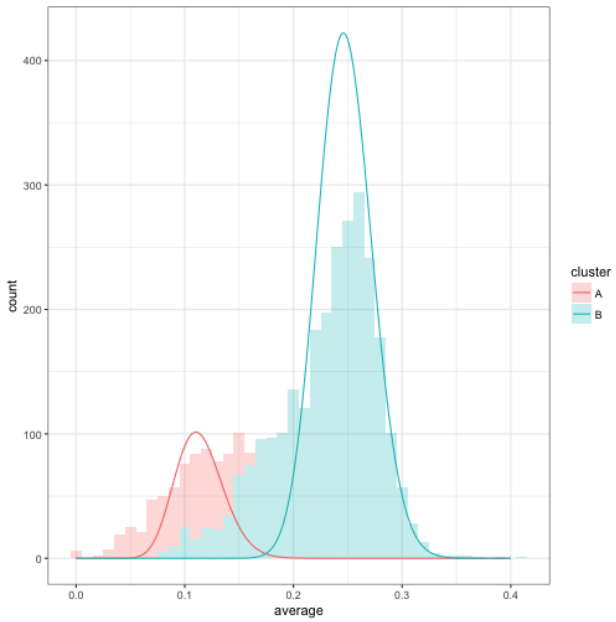


Figure 2. Two class model. What are the data generating processes that might result in skew?

Question: How many data generating processes can we think of that will result in non-normality?

Simulation 0 - tuning skew exericse

- "The second step is to run the analysis with one replication and a large sample to obtain approximate population values for the one class model (e.g., achieve factor indicator reliabilities of 0.64)." (Muthen & Muthen, 2002).
- We will just do 1 replication with an N-size of 100,000 (exploit law of large numbers!)

```
cfa_tune  <- lapply(1:5, function(k) {
  cfa_0  <- mplusObject(

  TITLE = "CFA 1 - non-normal, no missing",

  MONTECARLO =
    sprintf("NAMES ARE y1-y10;
```

```
    NOBSERVATIONS = 100000;
    NREPS = 1;
    SEED = 53487;
    CLASSES = C(1);
    GENCLASSES = C(2);
    SAVE = cfa0_%d.sav;",k),

  ANALYSIS =
    "TYPE = MIXTURE;
    ESTIMATOR = MLR;",

  MODELPOPULATION =
    glue("%OVERALL%
    f1 BY y1-y5*.8;
    f2 BY y6-y10*.8;
    f1@1 f2@1;
    y1-y5*.36 y6-y10*9;
    f1 WITH f2*.95;
    [C#1@-{k*.5}];   ! parameter we will tune to adjust the size of the outliers

    %C#1%            ! outlier class

    [f1@0 f2@15];    ! means (factor 2 set to 15 to tune skewness & kurtosis)
    f1@1 f2@5;       ! variances (factor 2 set to 5  to tune skewness & kurtosis)

    %C#2%            ! majority class

    [f1@0 f2@0];
    f1@1 f2@1;"),

  MODEL =
    "%OVERALL%
    f1 BY y1-y5*.8;
    f2 BY y6-y10*4;
    f1@1 f2@1;
    y1-y5*.36 y6-y10*9;
    f1 WITH f2*.20;

   [y6-y10*1.42];" ,

  OUTPUT = " SAMPSTAT TECH9;")

cfa_0_fit <- mplusModeler(cfa_0,
                          dataout=here("mplus_tune", "cfa0_demo_sim.dat"),
                          modelout=sprintf(here("mplus_tune", "%d_cfa0_demo_sim.inp"), k),
                          check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

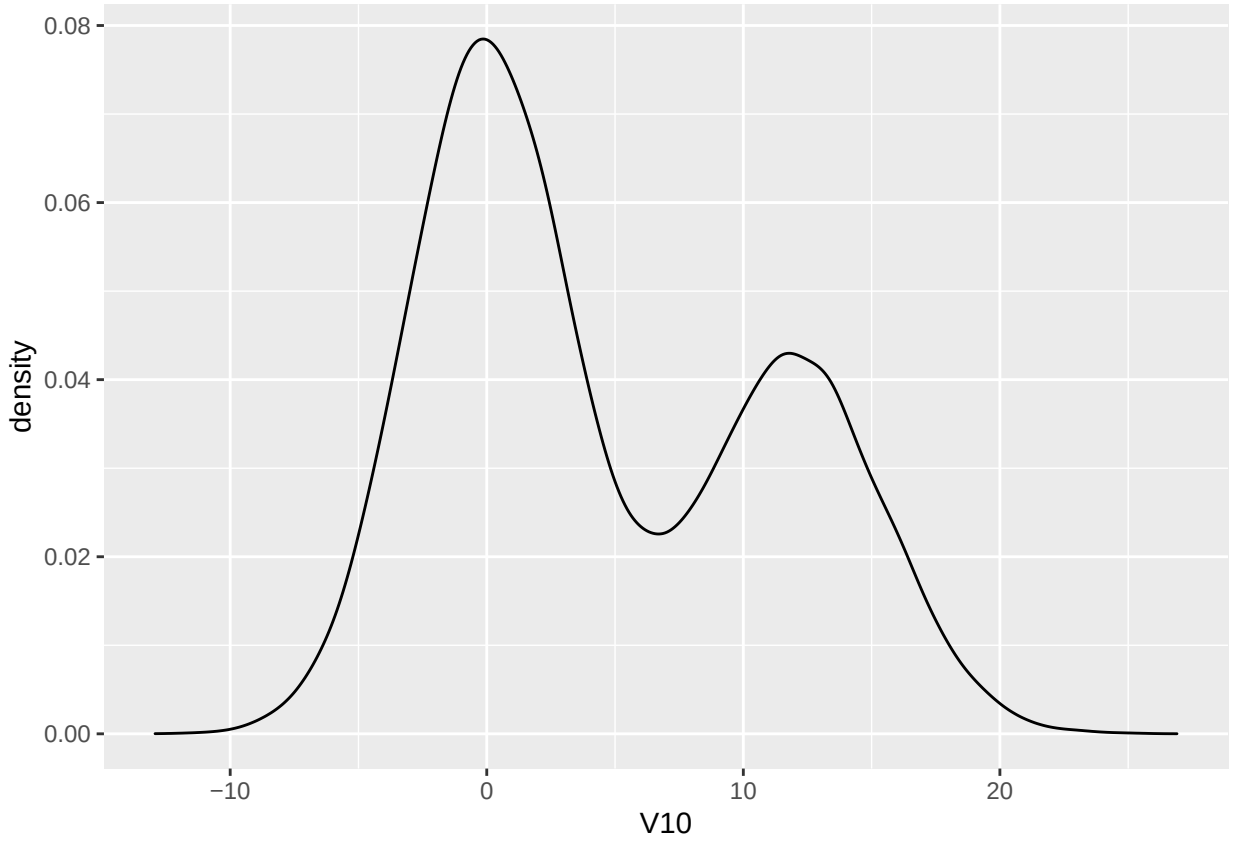
Read in simulated data, tabulate, plot distribution shape

```
library(gt)

data0_1 <- read.delim(here("mplus_tune", "cfa0_1.sav"), sep = "",
```

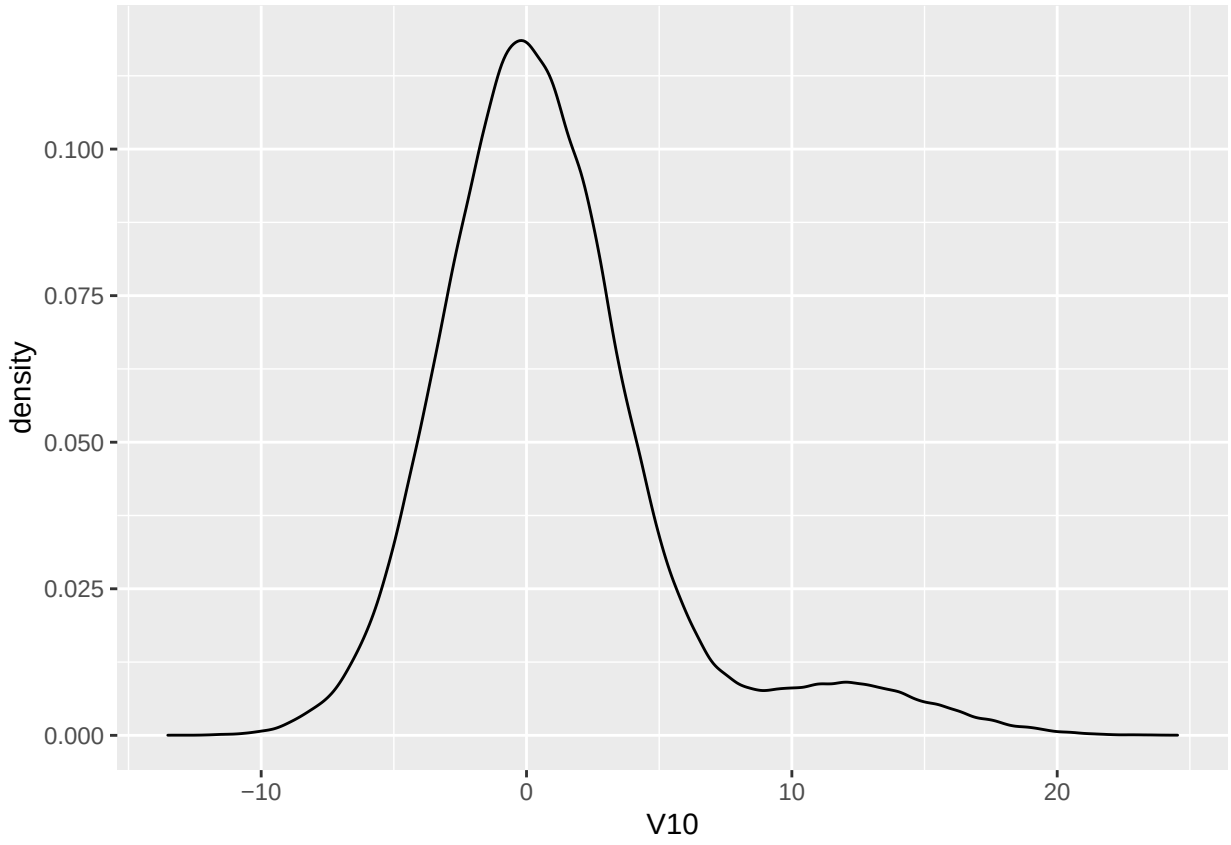
```
                    header = FALSE,
                    na.strings = "999.000000")

data0_2 <- read.delim(here("mplus_tune", "cfa0_5.sav"), sep = "",
                    header = FALSE,
                    na.strings = "999.000000")

describe(data0_1) %>% gt()
```

vars	n	mean	sd	median	trimmed	mad	min	max	range
1	1e+05	0.0005885588	1.0014277	0.0021225	0.0009827837	1.001229	-4.462478	4.489732	8.952210
2	1e+05	0.0029076816	1.0012728	0.0040230	0.0029815934	1.002304	-4.017530	4.568498	8.586028
3	1e+05	0.0029345232	1.0008448	-0.0003880	0.0016696903	1.004249	-4.067287	4.292325	8.359612
4	1e+05	-0.0012094374	1.0023827	-0.0052255	-0.0027470112	1.006132	-4.555642	4.313627	8.869269
5	1e+05	0.0032116140	0.9993024	-0.0000500	0.0025562450	1.000367	-4.655969	4.828608	9.484577
6	1e+05	4.5400706202	6.6517297	2.6562820	4.2192362366	7.048305	-13.657281	27.604618	41.261899
7	1e+05	4.5458945877	6.6747814	2.6611575	4.2336710184	7.145709	-12.918963	26.903231	39.822194
8	1e+05	4.5348804461	6.6673899	2.6363260	4.2139381264	7.074098	-12.434250	27.135822	39.570072
9	1e+05	4.5693057612	6.6829918	2.6483670	4.2505847502	7.122006	-13.386879	28.177935	41.564814
10	1e+05	4.5434661292	6.6702913	2.6473195	4.2245783391	7.085001	-12.928998	26.902138	39.831136
11	1e+05	1.6213600000	0.4850506	2.0000000	1.6517000000	0.000000	1.000000	2.000000	1.000000

```
data0_1 %>%  ggplot(aes(x=V10)) +
  geom_density()
```



```r
data0_2 %>%  ggplot(aes(x=V10)) +
  geom_density()
```



Criteria to assess level of bias

- "The first criterion is that parameter and standard error biases do not exceed 10% for any parameter in the model."
- "The second criterion is that the standard error bias for the parameter for which power is being assessed does not exceed 5%"
- "The third criterion is that coverage remains between 0.91 and 0.98."
- "Once these three conditions are satisfied, the sample size is chosen to keep power close to 0.80."

Simulation 1

- Normally distributed
- No missing
- Sample size (n) = 150
- Number of repetitions = 10,000

```
cfa_1  <- mplusObject(

  TITLE = "CFA 1 - normal, no missing",

  MONTECARLO =
    "NAMES ARE y1-y10;
     NOBSERVATIONS = 150;
     NREPS = 10000;
     SEED = 53487;
     CLASSES = C(1);
     GENCLASSES = C(1);
     SAVE = cfa1.sav;",

  ANALYSIS =
    "TYPE = MIXTURE;
     ESTIMATOR = ML; ! when normal MLR simplifies to ML",

  MODELPOPULATION =
    "%OVERALL%           !
     f1 BY y1-y5*.8;     !  factor loadings = .8 (average?)
     f2 BY y6-y10*.8;    !
     f1@1 f2@1;          !  factor variances = 1 (fixed)
     y1-y10*.36;         !  residual variances = .36
     f1 WITH f2*.25;     !  factor correlation = .25
    ",

  MODEL =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y10*.36;
     f1 WITH f2*.25;" ,

  OUTPUT = "TECH9;")

cfa_1_fit <- mplusModeler(cfa_1,
                          dataout=here("mplus_files", "cfa1_demo_sim.dat"),
                          modelout=here("mplus_files", "cfa1_demo_sim.inp"),
                          check=TRUE, run = TRUE, hashfilename = FALSE)
```

Example of some results from simulation 1 (Y1):

- Parameter bias (columns 1 & 2): $(.8 - .7979)/.8 = .0026$
- Standard error bias (columns 2 & 3): $(.0706 - .0699)/.0706 = 0.0099$
- Coverage (column 6): .947
- Power(column 7): 1.0

Simulation 2

- Normally distributed
- **missing set to 50% for y6 - y10**

```
cfa_2  <- mplusObject(

  TITLE = "CFA 2 - normal, missing (50%)",

  MONTECARLO =
    "NAMES ARE y1-y10;
     NOBSERVATIONS = 175;
     NREPS = 10000;
     SEED = 53487;
     CLASSES = C(1);
     GENCLASSES = C(1);
     PATMISS = y6 (.5) y7 (.5) y8 (.5) y9 (.5) y10 (.5);
     PATPROB = 1;
     SAVE = cfa2.sav;",

  ANALYSIS =
    "TYPE = MIXTURE MISSING;
     ESTIMATOR = ML; ! when normal MLR simplifies to ML",

  MODELPOPULATION =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y10*.36;
     f1 WITH f2*.25;",

  MODEL =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y10*.36;
     f1 WITH f2*.25;" ,

  OUTPUT = "PATTERNS TECH9;")

cfa_2_fit <- mplusModeler(cfa_2,
                          dataout=here("mplus_files", "cfa2_demo_sim.dat"),
                          modelout=here("mplus_files", "cfa2_demo_sim.inp"),
                          check=TRUE, run = TRUE, hashfilename = FALSE)
```

Simulation 3

- **Non-normally distributed**
- No missing

```
cfa_3  <- mplusObject(

  TITLE = "CFA 3 - non-normal, no missing",

  MONTECARLO =
    "NAMES ARE y1-y10;
     NOBSERVATIONS = 265;
     NREPS = 10000;
     SEED = 53487;
     CLASSES = C(1);
     GENCLASSES = C(2);
     SAVE = cfa3.sav;",

  ANALYSIS =
    "TYPE = MIXTURE;
     ESTIMATOR = MLR;",

  MODELPOPULATION =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.95;
     [C#1@-2];

     %C#1%         ! outlier class (size = 12%)

     [f1@0 f2@15]; ! means (facotr 2 set to 15 to tune skewness & kurtosis)
     f1@1 f2@5;    ! variances (facotr 2 set to 5  to tune skewness & kurtosis)

     %C#2%         ! majority class (size = 88%)

     [f1@0 f2@0];
     f1@1 f2@1;",

  MODEL =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*4;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.20;

   [y6-y10*1.42];" ,

  OUTPUT = "TECH9;")

cfa_3_fit <- mplusModeler(cfa_3,
                          dataout=here("mplus_files", "cfa3_demo_sim.dat"),
                          modelout=here("mplus_files", "cfa3_demo_sim.inp"),
                          check=TRUE, run = TRUE, hashfilename = FALSE)
```

Simulation 4a

- **Non-normally distributed**
- **missing set to 50% for y6 - y10**

```
cfa_4  <- mplusObject(

  TITLE = "CFA 4 - non-normal, missing (50%)",

  MONTECARLO =
    "NAMES ARE y1-y10;
     NOBSERVATIONS = 315;
     NREPS = 10000;
     SEED = 53487;
     CLASSES = C(1);
     GENCLASSES = C(2);
     PATMISS = y6 (.5) y7 (.5) y8 (.5) y9(.5) y10 (.5);
     PATPROB = 1;
     SAVE = cfa4.sav;",

  ANALYSIS =
    "TYPE = MIXTURE;
     ESTIMATOR = MLR;",

  MODELPOPULATION =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.95;
     [C#1@-2];
     %C#1%
     [f1@0 f2@15];
     f1@1 f2@5;
     %C#2%
     [f1@0 f2@0];
     f1@1 f2@1;",

  MODEL =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*4;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.20;
     [y6-y10*1.42];" ,

  OUTPUT = "PATTERNS TECH9;")

cfa_4_fit <- mplusModeler(cfa_4,
```

```
                          dataout=here("mplus_files", "cfa4_demo_sim.dat"),
                          modelout=here("mplus_files", "cfa4_demo_sim.inp"),
                          check=TRUE, run = TRUE, hashfilename = FALSE)
```

```
cfa4 <- read.delim(here("mplus_files", "cfa4.sav"), sep = "",
                   header = FALSE,
                   na.strings = "999.000000")

describe(cfa4) %>% gt()
```

vars	n	mean	sd	median	trimmed	mad	min	max	range	
1	315	-0.01179597	1.0279552	-0.0946230	0.003296929	1.0189598	-3.267687	2.970156	6.237843	-0.1
2	315	-0.01671664	0.9967482	0.0026150	0.007926889	0.9662238	-3.005738	2.647759	5.653497	-0.1
3	315	-0.01757470	1.0015924	0.0245540	-0.024351909	1.1127595	-2.332704	2.793124	5.125828	0.0
4	315	-0.05658550	0.9963691	-0.0263060	-0.054610577	1.0255530	-2.846280	2.866616	5.712896	-0.0
5	315	-0.06728217	0.9697836	-0.0381010	-0.051657806	0.9561035	-3.245657	2.823360	6.069017	-0.1
6	168	0.63516383	4.2201999	0.2773905	0.199561118	2.8957142	-7.537531	16.745489	24.283020	1.2
7	155	1.45862913	5.0283412	0.8493480	1.007651136	4.5227944	-11.340349	18.045953	29.386302	0.7
8	150	1.37581445	4.6393863	0.6389525	0.897577058	3.4766503	-7.371015	18.302114	25.673129	1.0
9	153	1.31932019	4.6393365	0.5498650	0.806264854	3.3762360	-7.774050	20.451471	28.225521	1.1
10	153	1.40275725	5.0186248	0.6205400	0.751691740	3.9618927	-9.297217	18.320741	27.617958	1.1
11	315	1.89841270	0.3025855	2.0000000	1.996047431	0.0000000	1.000000	2.000000	1.000000	-2.0
12	315	1.00000000	0.0000000	1.0000000	1.000000000	0.0000000	1.000000	1.000000	0.000000	

view characteristics of simulated data

Simulation 4b

- explore biased outputs
- vary sample size to see changes in bias

```
# testing & tuning

cfa_bias  <- lapply(1:5, function(k) {
  cfa_004  <- mplusObject(

  TITLE = "CFA 1 - non-normal, no missing",

  MONTECARLO =
    glue("NAMES ARE y1-y10;
     NOBSERVATIONS = {315-k*25}; ! vary sample size
     NREPS = 10000;
     SEED = 53487;
     CLASSES = C(1);
     GENCLASSES = C(2);
```

```
    SAVE = cfa004_{k}.sav;"),

  ANALYSIS =
    "TYPE = MIXTURE;
     ESTIMATOR = MLR;",

  MODELPOPULATION =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*.8;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.95;
     [C#1@-2];

     %C#1%         ! outlier class

     [f1@0 f2@15]; ! means (factor 2 set to 15 to tune skewness & kurtosis)
     f1@1 f2@5;    ! variances (factor 2 set to 5  to tune skewness & kurtosis)

     %C#2%         ! majority class

     [f1@0 f2@0];
     f1@1 f2@1;",

  MODEL =
    "%OVERALL%
     f1 BY y1-y5*.8;
     f2 BY y6-y10*4;
     f1@1 f2@1;
     y1-y5*.36 y6-y10*9;
     f1 WITH f2*.20;

    [y6-y10*1.42];" ,

  OUTPUT = " SAMPSTAT TECH9;")

cfa_004_fit <- mplusModeler(cfa_004,
                            dataout=here("mplus_bias", "cfa004_demo_sim.dat"),
                            modelout=sprintf(here("mplus_bias", "%d_cfa004_demo_sim.inp"), k),
                            check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

End of simulation practice
