THE DATA AND THEIR PROPERTIES:

COVARIANCES AND CORRELATIONS

Outline

1. The procedure
2. The requirements regarding data
3. The characteristic properties of real data
4. The typical data problems
5. Covariances and correlations as input

Outline

1. **The procedure**
2. The requirements regarding data
3. The characteristic properties of real data
4. The typical data problems
5. Covariances and correlations as input

THE PROCEDURE

- CFA and SEM combine features of different statistical approaches for investigating the structur of data.
- One of these approaches has grown out of the *analysis of the covariance matrix (AoC)* (Jöreskog, 1970)

THE PROCEDURE

- ***Analysis of the covariance matrix (AoC)*** (Jöreskog, 1970) amounts to contrasting …

- … the empirical covariance matrix (**S**) - and -
- … the model-implied matrix ($\boldsymbol{\Sigma}$)

THE PROCEDURE

- ***Analysis of the covariance matrix (AoC)*** (Jöreskog, 1970) amounts to contrasting …

- … the empirical covariance matrix (**S**) - and -
- … the model-implied matrix (Σ)

- … for the purpose of this comparison the parameters included in Σ must be specified; the whole of them is represented by θ

… so that $\Sigma (\theta)$

THE PROCEDURE

- ***Analysis of the covariance matrix (AoC)*** (Jöreskog, 1970) amounts to contrasting …

- … the empirical covariance matrix (**S**) - and -
- … the model-implied matrix ($\boldsymbol{\Sigma}$)

- … the difference $\mathbf{d(S,\Sigma(\theta))}$ is determined
- … it has to be as small as possible → $\mathbf{F(S,\Sigma(\theta))}$

fitting function

THE PROCEDURE

- ***Analysis of the covariance matrix (AoC)*** (Jöreskog, 1970) amounts to contrasting …

This means that the **input** to AoC is …

- … a ***covariance matrix*** (CM) - or -
- … a ***correlation matrix*** (KM) - or –
- … the matrix including the raw data (in this case the computer mostly transforms the data in CM or KM)

THE PROCEDURE

- ***Analysis of the covariance matrix (AoC)*** (Jöreskog, 1970) amounts to contrasting …
- … a ***covariance matrix*** (CM) and a ***correlation matrix*** (KM)

… the ***model-fit approach***

(recent denotation)

Outline

1. The procedure
2. **The requirements regarding data**
3. The characteristic properties of data
4. The typical data problems
5. Covariances and correlations as input

THE REQUIREMENTS REGARDING DATA

- The requirements regarding data are determined by the model of measurement: …. *that is what we know*

$X = \mu + \Lambda\xi + \delta$ respectively $Y = \mu + \Lambda\eta + \varepsilon$

continuous, normally distributed

continuous, normally distributed

… assures a high degree of accuracy

… is a useful statistical assumption

THE REQUIREMENTS REGARDING DATA

- The requirements regarding data are determined by the model of measurement:

i.e. X and Y must be continuous and normally distributed!

THE REQUIREMENTS REGARDING DATA: *SUPPLEMENT*

- Why is it necessary that there is correspondence

Answer: the *estimation* of the parameters included in Σ is based on assumptions; *x* and *y* must comply with these assumptions

THE REQUIREMENTS REGARDING DATA: *SUPPLEMENT*

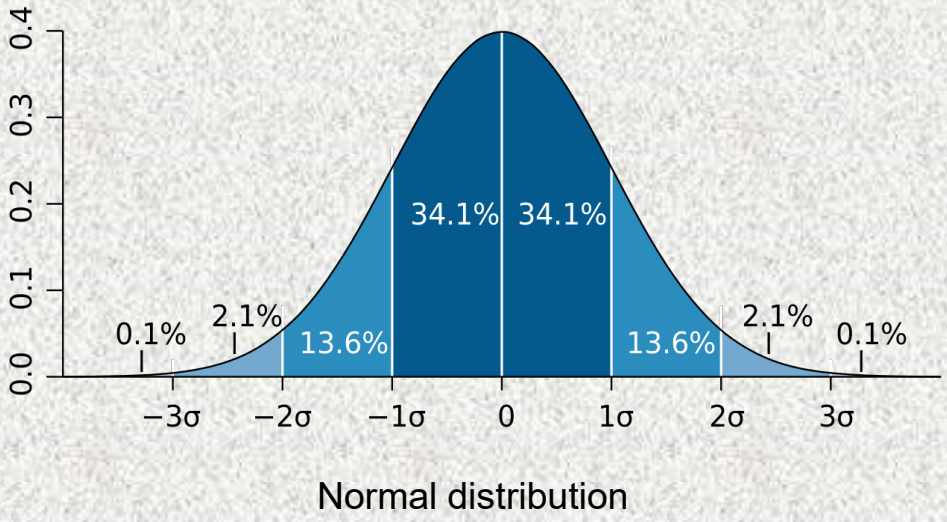
- What are the consequences of the restricting assumption?
  - real data may show violations of assumptions
  - … distorted distributions
  - … inappropriate scales

THE REQUIREMENTS REGARDING DATA

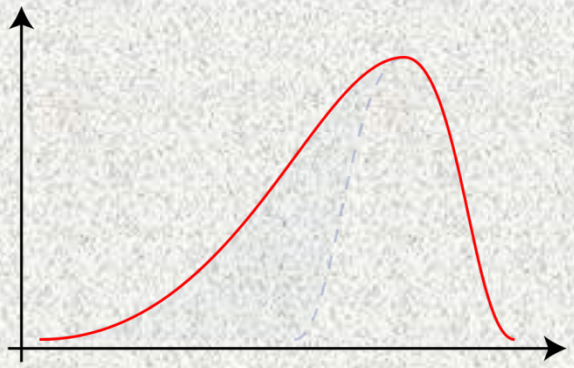
What means a violation of the distribution?

An example

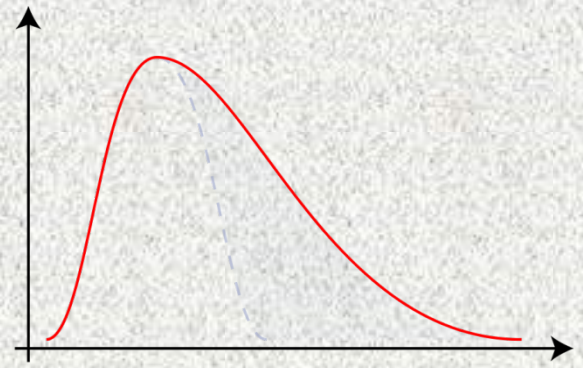
Assume that ✗ is skewed instead of normally distributed.



Normal distribution



Negative Skew



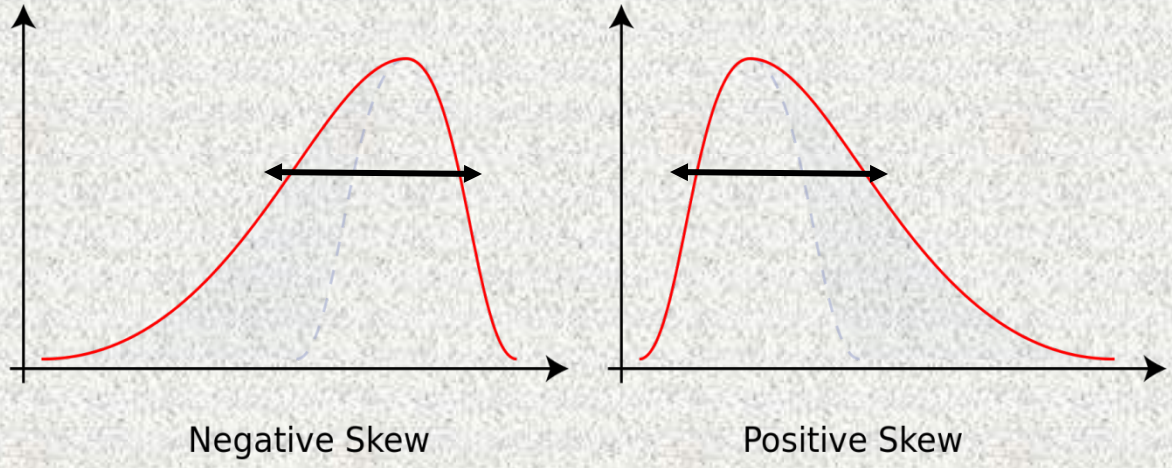
Positive Skew

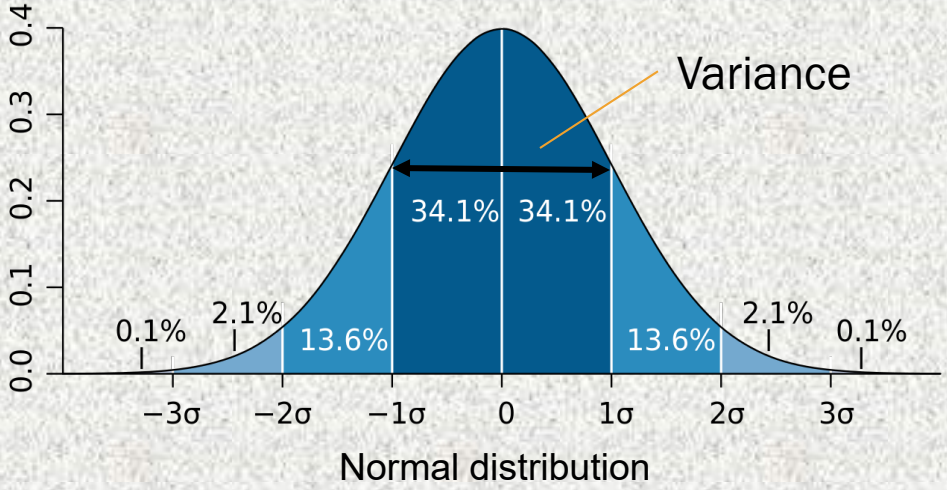
THE REQUIREMENTS REGARDING DATA

What means a violation of these assumptions?

An example

Assume that ***X*** is skewed instead of normally distributed.





... skewness ***diminishes*** the variance!!!

THE REQUIREMENTS REGARDING DATA

What means a violation of these assumptions?

An example

Assume that x shows to be platykurdic (= it is broader than normel)

… now the variance of the variable is larger than otherwise!!!

THE REQUIREMENTS REGARDING DATA

What are the consequences of a violation of assumptions?

An example

- assume a set of items measuring the same construct but ...
  - some data show *skewness*, some the *normal* distribution, some are *platykurdic*

- the data are investigated by the same latent variable of the same model

THE REQUIREMENTS REGARDING DATA

What are the consequences of a violation of assumptions?

An example

- assume a set of items measuring the same construct but …
  - some data show *skewness*, some the *normal* distribution, some are *platykurdic*

- the observed (distorted) covariances lead to (distorted) factor loadings:

$cov(x_i,x_j) \longrightarrow \lambda_i, \lambda_j \; (\longrightarrow \; \sigma_{ij} = \lambda_i \times \lambda_j)$

THE REQUIREMENTS REGARDING DATA

What are the consequences of a violation of assumptions?

An example

- assume a set of items measuring the same construct but ...
  - some data show *skewness*, some the *normal* distribution, some are *platykurdic*

- the data are investigated by the same latent variable of the same model

- as a consequence, the true variance of some items is ...
  - underestimated
  - overestimated

- *the reproduction of parts of the covariance matrix may be flawed*:

$cov(x_i,x_j)$ *may deviate from* $\sigma_{ij} = \lambda_i \times \lambda_j$

THE REQUIREMENTS REGARDING DATA

What are the consequences of a violation of assumptions?

… a number of method studies show that violations of the distributional assumptions lead to problems regarding model fit and correctness of parameter estimation

… e.g. :

Lai, K. (2018). Estimating standardized SEM parameters given nonnormal data and incorrect model: methods and comparisons. *Structural Equation Modeling, 25(4)*, 1-21. doi: 10.1080/10705511.2017.1392248

West, S. G., Finch, J. F., & Curran, P.J. (1995). Structural equation models with non-normal variables: problems and remedies. In R. Hoyle (Ed.), Structural Equation Modeling: Concepts, Issues, and Applications (pp. 56-75). SAGE.

THE REQUIREMENTS REGARDING DATA

- In sum: deviations from the requirements are likely to lead to …

  - deviations in parameter estimation

  - problems in reproducing the empirical covariance matrix

Outline

1. The procedure
2. The requirements regarding data
3. **The characteristic properties of data**
4. The typical data problems
5. Covariances and correlations as input

THE CHARACTERISTIC PROPERTIES OF DATA: SCALE

- In psychological research data are typically …
  - binary data
  - dichotomous data
  - categorical data
  - ordered categorical data
  - ordinal Data
  - frequencies
  - (rarely) data showing intervall scale

THE CHARACTERISTIC PROPERTIES OF DATA: SCALE

Search for **one example** for each type of data, i.e. for ...

- binary data	?	male / female
- dichotomous data	?	poor / rich
- categorical data	?	apples / pears / oranges
- ordered categorical data	?	ratings like „agree fully“ ...
- ordinal Data	?	grades
- frequencies	?	hours on computer
- (rarely) data showing intervall scale	?	age

THE CHARACTERISTIC PROPERTIES OF DATA: SCALE - EXAMPLES

Search for **one example** for each type of data, i.e. for ...

- binary data	male / female
- dichotomous data	poor / rich
- categorical data	apples / pears / oranges
- ordered-categorical data (Likert data)	ratings like „agree fully“ ...
- ordinal Data	grades
- frequencies	hours on computer
- (rarely) data showing intervall scale	age

THE CHARACTERISTIC PROPERTIES OF DATA: SCALE

- In psychological research data are typically …
  - binary data
  - dichotomous data
  - ordered categorical data (Likert data)
  - ordinal Data
  - frequencies
  - (rarely) data showing intervall scale

Only these types get usually accepted as continuous!

THE CHARACTERISTIC PROPERTIES OF DATA: DISTRIBUTION

- In psychological research, data typically show …

- (rarely) a normal distribution 正态分布

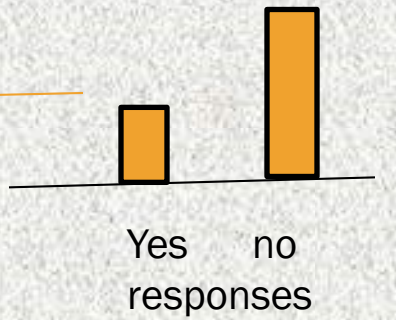
- a distribution that is similar to the normal distribution (e.g. a skewed distribution 偏态分布)

- the binomial distributions (二项分布)

- not clearly identifiable distribution

fits to the requirements!

… can be modified to fit requirement



Outline

1. The procedure
2. The requirements regarding data
3. The characteristic properties of data
4. **The typical data problems**
5. Covariances and correlations as input

THE TYPICAL DATA PROBLEMS

- **Necessary mathematical operations that are *not really allowed***

  ⟶ computation of variances方差 and covariances协方差

  e.g. gender data have to be investigated (this means binary data) – the researcher computes the arithmetic mean 算术平均数!

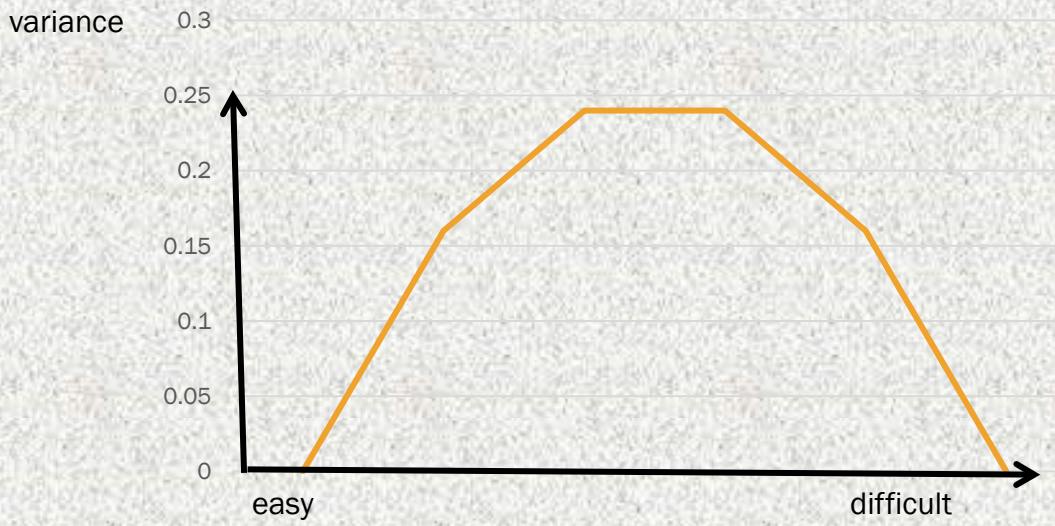
Wrong! The arithmetic mean is only appropriate for continuous data!

THE TYPICAL DATA PROBLEMS

- Variances and covariances are computed using mathematical operations that are *not appropriate*
- The observed variances and covariances may deviate from the expected variances and covariances because of …

- … characteristic dependencies (e.g. there may be a dependency of variances (and covariances) on the item difficulties 难度 [mean]as in binary data)

Illustration: variance – difficulty - dependency



THE TYPICAL DATA PROBLEMS

- Variances and covariances are computed using mathematical operations that are *not really appropriate*
- The observed variances and covariances deviate from the expected variances and covariances because of …

- … characteristic dependencies (e.g. there may be a dependency of variances and covariances on the item difficulties as in binary data)

- … skewness of the distribution (or another distributional irregularity e.g. kurtosis, more than one peak)

Skewness 偏态 ⟶ distorted variances 方差减小

Outline

1. The procedure
2. The requirements regarding data
3. The characteristic properties of data
4. The typical data problems
5. **Covariances and correlations as input**

COVARIANCES AND CORRELATIONS AS INPUT

- Covariances and *Pearson correlations* based on the *products of moments* usually are expected as **input**

  - the pre-condition is that they are computed from continuous and normally distributed data

... in order to overcome restrictions regarding scale and distribution, special coefficients have been developed!

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

- Covariances and *Pearson correlations* based on the *products of moments* usually are expected as **input**

  - Problem: *the data are binary / dichotomous*

- Compute tetrachoric correlations 四分相关

  - Problem: *the data are ordered-categorical (or ordinal)*

- Compute polychoric correlations 多分相关

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

- Covariances and *Pearson correlations* based on the *products of moments* usually are expected as **input**

  - Problem: *the data are binary*

- Compute tetrachoric correlations

  - Problem: *the data are ordered-categorical (or ordinal)*

- Compute polychoric correlations

  - Problem: *the data are ordered-categorical (or ordinal) with more than 6 categories*

- Treat the data as ***continuous***

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

The computation of tetrachoric and polychoric correlations includes the computation of ***thresholds*** *临界值* (that is necessary for the transformation into the normal distribution) :

i.e. the thresholds establish the relationship to continuous and normally distributed parameters

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

The computation of tetrachoric and polychoric correlations includes the computation of ***thresholds*** (that is necessary for the adaptation to the normal distribution) :

i.e. the thresholds establish the *relationship to* continuous and normally distributed parameters. They serve as basis for the computation of correlations that are in line with the requirements of the model of measurement!

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

The computation of tetrachoric and polychoric correlations
includes the computation of ***thresholds*** with respect to the normal distribution …

There is a major disadvantage:

in the marginal areas of the normal distribution the thresholds are usually not very accurate (i.e. in data obtained by very easy or very difficult items. Therefore, **very large samples are necessary** for achieving accurate estimates)

- A consequence is that in very easy or very difficult items *very high correlations* can be observed (that is bad because the other correlations may be low)!
- Another consequence is that the input matrix may not be *positive definite* 正定的.

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

- Covariances based on product moments and Pearson correlations usually are expected / used as input

  In binary data

- Compute tetrachoric correlations

- Alternatively: compute ***probability-based covariances*** ***概率协方差***

  (Schweizer, Ren, & Wang, 2015)

  (… additionally needs a *link transformation* or *robust estimation*)

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

The formula for computing the probability-based covariance:

$$\text{cov}(X_i,X_j) = \Pr(X_i=1 \wedge X_j=1) - \Pr(X_i=1)\times\Pr(X_j=1)$$

with X_i und X_j as binary variables.

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement: *an example*

The probability-based covariance of *handedness* and *arts interest*:

Handedness	Arts interest
right	yes
left	no
left	no
right	no
right	yes
left	no
right	yes

$$\text{cov}(X_i,X_j) = \Pr(X_i=1 \wedge X_j=1) - \Pr(X_i=1)\Pr(X_j=1)$$

Pr(handedness=right) = 4/7 = 0.571

Pr(arts interest=yes) = 3/7 = 0.428

Pr((handedness=right) and (arts interest=yes)) = 3/7 = 0.428

cov(handeness,arts interest) = 0.428 – 0.244 = 0.184

r = 0.184 / (0.493 x 0.493) = 0.417

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

A supplement

.....

Computing the probability-based covariance performs only the step from binary to continuous.

→ A *link transformation* is additionally necessary (... is described in another course unit)

(Link transformations are used for relating variables showing different distributions to each other).

... alternatively ..

→ *Robust estimation* is additionally necessary

COVARIANCES AND CORRELATIONS AS INPUT: *THE SCALE*

- Covariances based on product moments and Pearson correlations

.....

- further options

- .... what ever correlation together with *robust WLS estimation*

WLS is a correction method 修正方法 (that means that it is not a real solution to the problem)

COVARIANCES AND CORRELATIONS AS INPUT: *THE DISTRIBUTION*

If the data show skewness, it is necessary to select a special way of parameter estimation

- Robust estimation according to Satorra-Bentler
- DWLS estimation

Summary and brush up:

1. The foundations

… remember: the method amounts to the comparison of empirical and model-based covariance matrices

2. The requirements regarding data

… remember: manifest and latent variables must show the same properties

3. The characteristic properties of data

… remember: important properties are scale and distribution

4. The typical data problems

… remember: data frequently do not show the desirable properties so that special adaptation may be required

5. Covariances and correlations as input

… remember: the selection of special types of input and estimation method helps to overcome the problems

QUESTIONS REGARDING COURSE UNIT 3

- Which data properties are desirable in CFA and SEM?
- Which correlation type should be selected for investigating binary data?
- What data type requires the use of polychoric correlations?
- Which data type can be used for representing males and females?

LITERATURE

Basic:

- Kline, R. b. (2011). *Principles and practices of structural equation modeling* (3rd edition) (Chapter 1:Introduction). New York, NY: The Guilford Press.
- Schweizer, K., Ren, X., & Wang, T. (2015). A comparison of confirmatory factor analysis of binary data on the basis of tetrachoric correlations and of probability-based covariances: a simulation study. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative Psychology Research* (pp. 273-292). Heidelberg: Springer.