

*Preconditions
and
Check of Preconditions*

Introduction

This course unit concentrates on the preconditions for the statistical investigations; these are preconditions that must be given for achieving **valid results**.

Violations of preconditions can lead to the **termination** of the computations in some cases; but sometimes the program does not detect a violation.

Some violations of preconditions lead to results that signify very **bad model fit** although the model is good.

In such a case the computed statistics are **not correct / invalid**

Outline

1. The model

- 1.1 The model: identification of complete model**
- 1.2 The model: identification of model of measurement**
- 1.3 The model: fixations**
- 1.4 The model: specification of relationships**

2. The data

- **2.1 Properties of ideal data**
- **2.2 Typical problems**

3. The check of preconditions

1. The model

1.1 The model: identification of **complete model**

The degree of freedom provided by the program corresponds to what is achieved by using the formula listed below.

Formula:

$$\text{d.f.} = s - t$$

$$s = [n(n+1)] / 2$$

*The degree of freedom must **be larger** than zero!*

1. The model

1.2 The model: identification of **model of measurement**

In this case the user must compute the degree of freedom.

Formula:

$$\text{d.f.} = s - t$$

$$s = [n (n+1)] / 2$$

*The degree of freedom must **be at least** zero!
Larger than zero is better!*

1. The model

1.3 The model: fixations

Are the fixations correct?

- variance of a latent variable: 1 (but not zero)
- factor loadings: 1 or equal sizes
(or zero)
other values
- residual variances: 0 or equal sizes

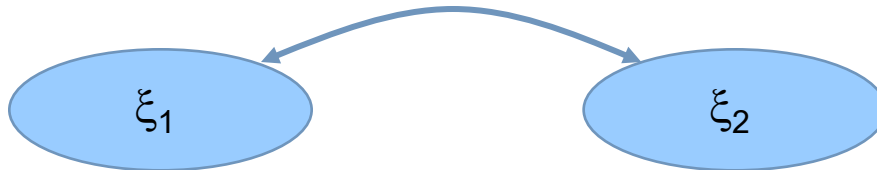
1. The model

1.3 The model: specification of relationships

Is the to-be-investigated relationship specified correctly?

- is the relationship a correlation?

... or correlate the corresponding latent variables:



1. The model

1.3 The model: specification of relationships

Is the to-be-investigated relationship specified correctly?

- is it a directional effect?

... define the latent variables as independent (exogenous) and dependent (endogenous) latent variables:



... and check whether the corresponding ψ parameter is free for estimation.

2. The data

2.1 The properties of ideal data

- They show continuous scale (interval level)
- There is a sample of sufficient size

The required size varies as a function of

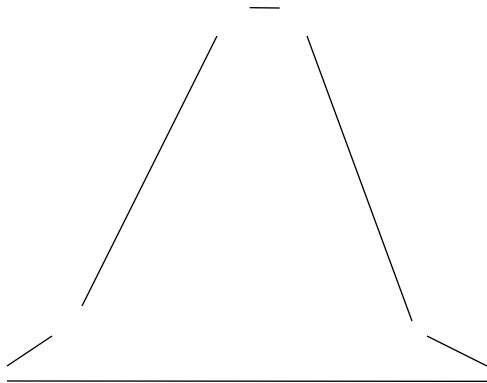
- the relationship of the numbers of participants and manifest variables
- the selected estimation method
- the type of input

In many studies (e.g. studies using cognitive data) $N \geq 200$ is sufficient.

2. The data

2.1 The properties of ideal data

- They show continuous scale (interval level scale)
- There is a sample of sufficient size
- Each random variable follows a normal distribution

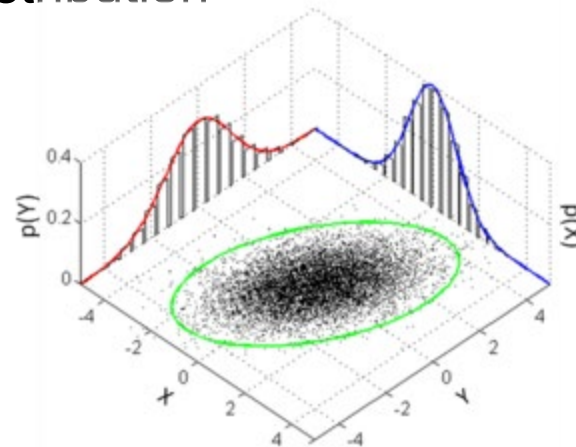


2. The data

2.1 The properties of ideal data

- They show continuous scale (interval level)
- There is a sample of sufficient size
- Each random variable follows a normal distribution
- There is multivariate normal distribution

e.g. a bivariate



2. The data

2.1 The properties of ideal data

- They show continuous scale (interval level)
- There is a sample of sufficient size
- Each random variable follows a normal distribution
- There is multivariate normal distribution
- The input matrix is positive definite

M: data matrix

v: any vector

$$\mathbf{v}^T \mathbf{M} \mathbf{v} > 0$$

2. The data

2.1 The properties of ideal data

- They show continuous scale (interval level)
- There is a sample of sufficient size
- Each random variable follows a normal distribution
- There is multivariate normal distribution
- The input matrix is positive definite
- The variables do show linear relationships

2. The data

2.2 Typical problems

- The scale level differs from what is expected (see course unit on data).

In psychological research data are typically ...

- binary data
- dichotomous data
- ordered categorical data (Likert data)
- ordinal Data
- frequencies
- (rarely) data showing intervall level scale

No
problem


2. The data

2.2 Typical problems

- The scale differs from what is expected (see course unit on data).

In psychological research data are typically ...

- 1 {
 - binary data
 - dichotomous data
- 2 {
 - ordered categorical data (Likert data)
 - ordinal Data

Use  ¹ tetrachoric correlation or probability-based covariances

Use  ² polychoric correlation

2. The data

2.2 Typical problems

- The scale differs from what is expected (see course unit on data).

In psychological research data are typically ...

- 1 {
 - binary data
 - dichotomous data
- 2,3 {
 - ordered categorical data (Likert data)
 - ordinal Data

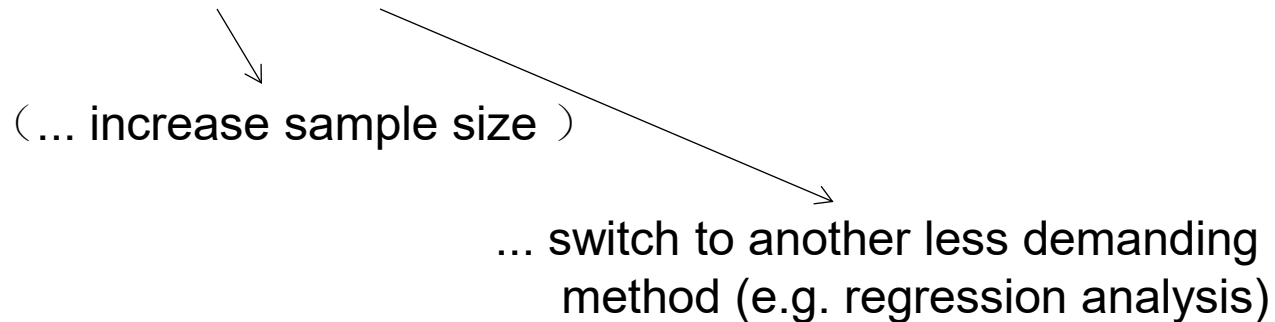
³ if there are six or more than six categories / six different values, ...

➡ treat the data as if they were continuous data

2. The data

2.2 Typical problems

- The scale differs from what is expected (see course unit on data).
- The sample is too small

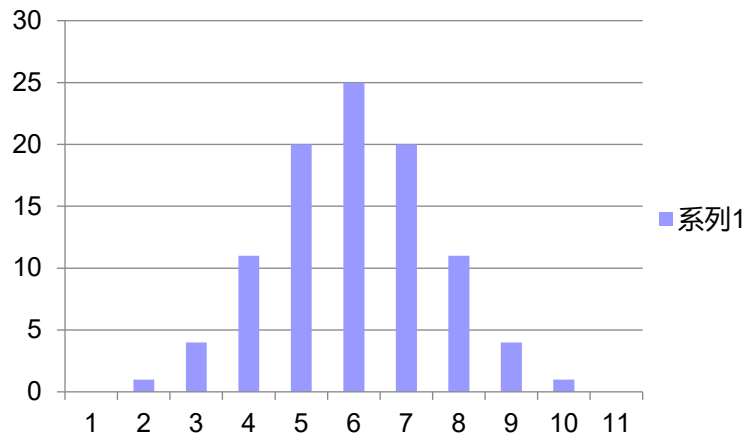


2. The data

2.2 Typical problems

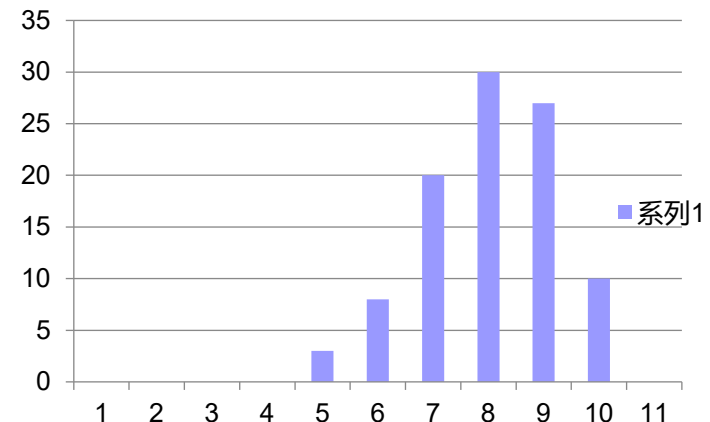
- **No normal distribution. Big problem!**

Normal distribution



Skewness=0.017

Distorted normal distribution



Skewness=-0.378

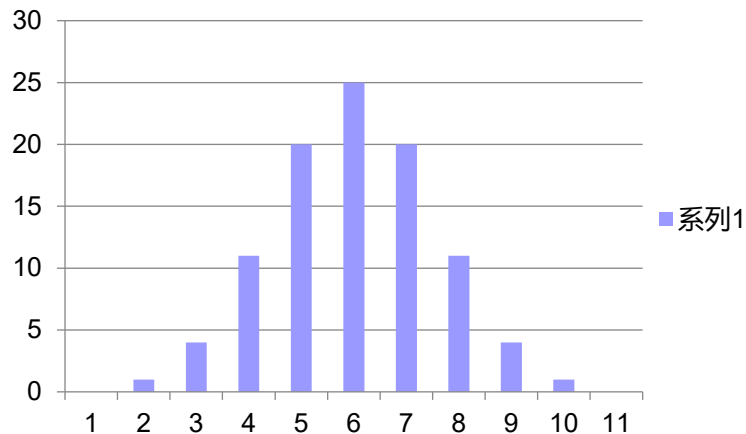
The variance becomes smaller!

2. The data

2.2 Typical problems

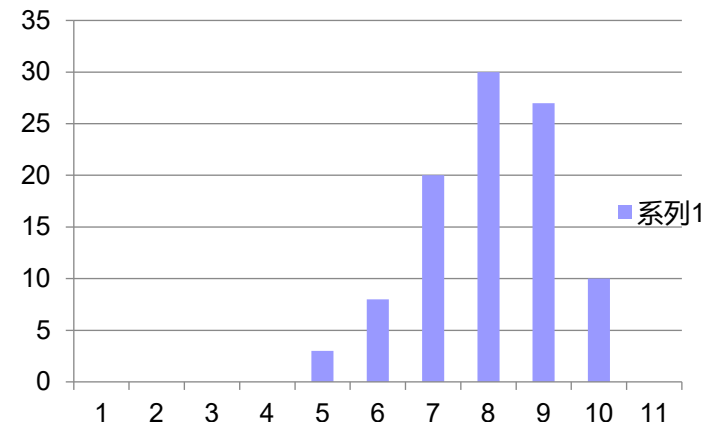
- **No normal distribution. Big problem!**

Normal distribution



Mean: 6 Variance: 2.4

Distorted normal distribution



Mean: 8 Variance: 1.5

2. The data

2.2 Typical problems

- **No normal distribution. Big problem!**

*Transformationen of distribution is **no more** recommended!*

Instead select an estimation method that compensates for the deviation from normality (see next course unit)

(Finney, DiStefano, & Kopp, 2016)

... so-called robust estimation methods need to be selected
(e.g. the Satorra-Bentler corrected maximum likelihood estimation method)

2. The data

2.2 Typical problems

- The scale differs from the ideal (see course unit on data).
- The sample is too small
- The data do not follow the normal distribution
- There is **no** multivariate normal distribution

.... try robust estimation

2. The data

2.2 Typical problems

- Matrix is not positive definite (= terminates computations unless ...).

i.e. there is multi-**collinearity**

(... eventually it is overcome by the program using the so-called ***ridge procedure***)

e.g. $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$... is not positive definite (e.g. **not** >0)

2. The data

2.2 Typical problems

- Matrix is not positive definite.

e.g. $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$... is not positive definite

Demonstration:

$$\begin{aligned} \begin{bmatrix} -1 & 1 \end{bmatrix} &\times \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \end{bmatrix} \times \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -2 < 0 \end{aligned}$$

2. The data

2.2 Typical problems

- Matrix is not positive definite. Use **Ridge Procedure**: an example

... the covariance matrix

$$\mathbf{M}_{\text{cov}} = \begin{bmatrix} \text{cov}_{11} & . & . & . & \text{cov}_{1p} \\ \text{cov}_{21} & \text{cov}_{22} & & & . \\ . & & . & & . \\ . & & & . & . \\ \text{cov}_{p1} & . & . & . & \text{cov}_{pp} \end{bmatrix}$$

the **ridge matrix**

$$\mathbf{C} = \begin{bmatrix} c & 0 & . & . & 0 \\ 0 & c & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ 0 & . & . & . & c \end{bmatrix}$$

the ridge-adjusted matrix

$$\mathbf{M}_{\text{ridge_cov}} = \begin{bmatrix} \text{cov}_{11} + c & . & . & . & \text{cov}_{1p} \\ \text{cov}_{21} & \text{cov}_{22} + c & & & . \\ . & & . & & . \\ . & & & . & . \\ \text{cov}_{p1} & . & . & . & \text{cov}_{pp} + c \end{bmatrix}$$

2. The data

2.2 Typical problems

- Matrix is not positive definite.

e.g.
$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix}$$

Demonstration:

$$\begin{bmatrix} -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -2$$

$$\begin{bmatrix} -1 & 1 \end{bmatrix} \times \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix} \times \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 16 \quad (> 0)$$

2. The data

2.2 Typical problems

Practice

- Find out whether the following matrix is positive definite:

$$\begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}$$

Correct ?

Incorrect ?

Outcome: -2

2. The data

2.2 Typical problems

- Matrix is not positive definite.
- Variables do not show linear relationships.

.... reject the model

or

.... check whether the consideration of a moderator variable solves the problem

3. The check of preconditions

3. The check of preconditions

- It is to be checked whether the preconditions of 1.1, 1.2 and 2.1 are given (i.e. regarding the model and regarding obvious data characteristics)

It remains to check characteristics that are not obvious!

3. The check of preconditions

- It is to be checked whether the preconditions regarding *normality* are given
- Especially the shape of distribution needs to be checked ...
 - skewness statistic
 - z test for finding out whether the distribution of individual variables is symmetric
 - Kolmogorov-Smirnov test (for individual variables)
 - Mardia test (for a set of variables / a matrix;
sometimes it is demanded)

3. The check of preconditions

- It is to be checked whether the preconditions regarding *normality* are given
- Especially the shape of distribution needs to be checked ...
 - ... there may be several several peaks
(suggesting a mixture of subsample showing different distributions)

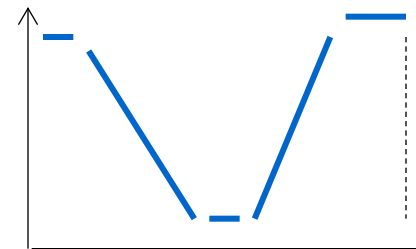
... subsample should be investigated separately



3. The check of preconditions

- It is to be checked whether the preconditions regarding *normality* are given
- Especially the shape of distribution needs to be checked ...

... there is an indication of a binary scale

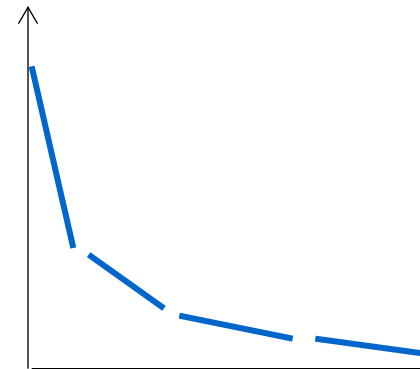


3. The check of preconditions

- It is to be checked whether the preconditions regarding *normality* are given
- Especially the shape of distribution needs to be checked ...

... there may be an indication of a Poisson distribution

... is not suited for the
computation of correlations or
covariances



QUESTIONS REGARDING COURSE UNIT 9

- If the data are ordered-categorical, what is the minimum number of categories for treating these data as continuous?
- Which estimation method has to be selected if the data do not follow the normal distribution?
- What has to be done, if the sample size is too small?
- What means „*not* positive definite“?