2. Faktorenanalyse

2.1 Grundidee der Faktorenanalyse

Arten eines statistischen Zusammenhangs zwischen den Variablen X_1 und X_2

a) Die Variable 2 (X_2) ist eine Funktion der Variablen 1 (X_1). So wird etwa in der Volkswirtschaftslehre unterstellt, dass die Investitionen I u.a. vom Marktzins i (X_1) abhängen.

$I = f(i)$, d.h. $X_1 \rightarrow X_2$

b) Umgekehrt kann die Variable 2 die abhängige und 1 die unabhängige Variable sein

$X_1 \leftarrow X_2$

c) Die Variablen 1 und 2 beeinflussen sich gegenseitig. Man spricht dann auch von wechselseitiger Abhängigkeit (Interdependenz). So ist der Konsum C eine Funktion des Einkommens Y; Y wird aber umgekehrt von der Höhe der privaten Konsumausgaben bestimmt. Es gilt

$C = f(Y)$ und $Y = g(C)$, d.h. $X_1 \leftrightarrow X_2$

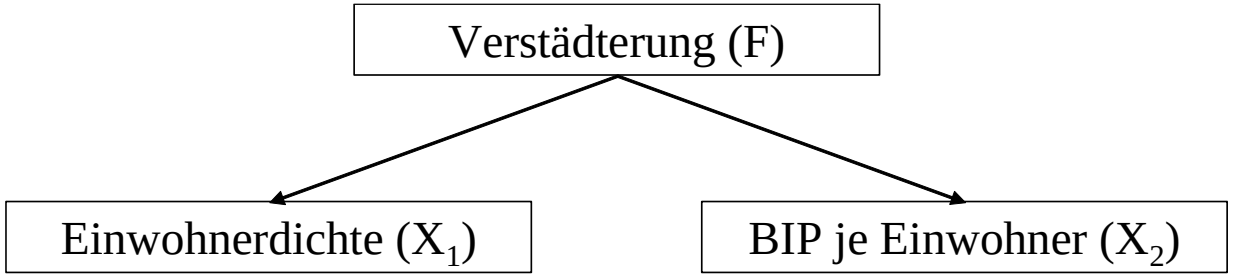
d) Eine dritte Variable bestimmt sowohl die Ausprägung der Variablen 1 als auch die der Variablen 2. So mag eine hohe Korrelation zwischen Geldmenge M und Beschäftigung B darauf zurückzuführen sein, dass sie beide im Konjunkturverlauf ähnliche Schwingungen aufweisen. Damit hat man $M = f(F_{KON})$ und $B = g(F_{KON})$ und entsprechend in graphischer Darstellung

$$F_{KON} \rightarrow X_1, \quad F_{KON} \rightarrow X_2$$

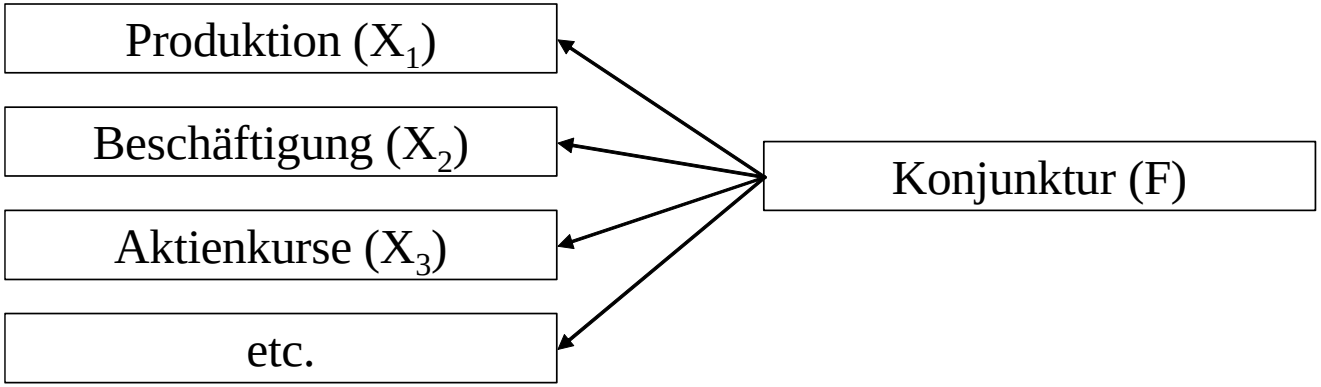
Der „Konjunkturfaktor“ F_{Kon} würde dann die hohe Korrelation zwischen M und B hervorrufen. Die **Faktorenanalyse** geht von diesem Erklärungsansatz aus.

Beispiel 2.1:

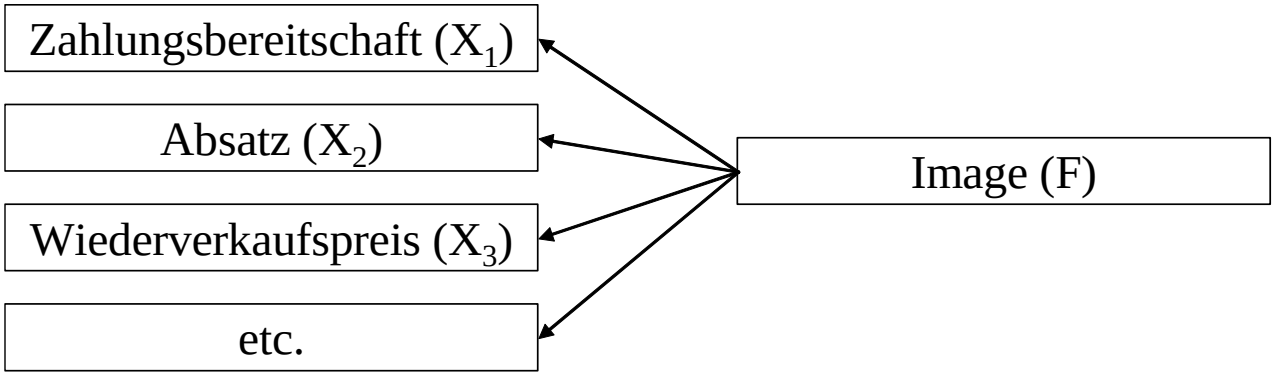
- Beide Variablen sind eine Funktion einer dritten Größe F, die wir als Faktor oder Supervariable bezeichnen wollen. In unserem Beispiel könnte F z.B. ein Verstädterungsgrad sein. Je höher der Verstädterungsgrad, umso größer sind ED und BIP.



- In der Volkswirtschaftslehre misst man die Konjunktur über verschiedene Indikatoren, so die Produktion, die Beschäftigung, die Aktienkurse etc.

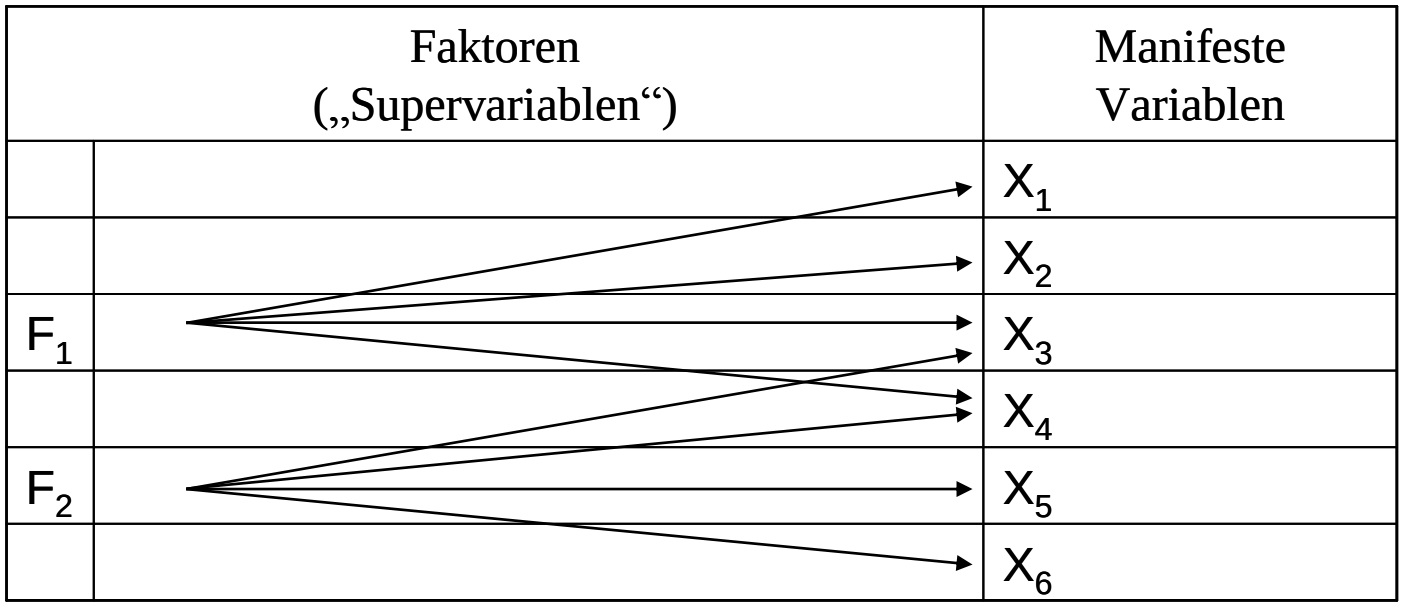


- Das Image eines Produktes ist ebenfalls nicht direkt messbar. Es kann unter Verwendung einiger X-Variablen abgebildet werden.



♦

Pfeildiagramm einer Faktorenstruktur



Hauptziel der Faktorenanalyse

Ableitung der hinter einer Menge manifester (beobachtbarer) Variablen $X_1, X_2, \ldots, X_m$ stehenden gemeinsamen Faktoren („Supervariablen“) $F_1, F_2, \ldots, F_p$ mit $p < m$

Faktoren („Supervariablen“)

- Nicht direkt beobachtbare Größen, die die Variablen beeinflussen
- Einige Variablen z.B. X_1 und X_2 werden nur vom Faktor F_1 beeinflusst oder X_5 und X_6, die nur vom Faktor F_2 bestimmt werden. Hinter anderen Variablen wie z.B. X_3 und X_4 stehen mit F_1 und F_2 mehrere Faktoren.

Anwendungszweck der Faktorenanalyse im 12-Regionen-Beispiel

Stehen hinter den regionalstatistischen Merkmalen bestimmte Strukturdimensionen, mit denen die zwölf Regionen charakterisiert werden können?

Beispiel 2.2: Datensatz aus 12 Regionen und 6 Merkmalen

In 12 Regionen seien im Rahmen einer regionalen Strukturanalyse sechs Merkmale beobachtet worden, die ihre sozioökonomische Struktur charakterisieren:

Merkmalsausprägungen für 12 Untersuchungsregionen und 6 Variablen

Region	X_1	X_2	X_3	X_4	X_5	X_6
A	212,4	20116	9,8	53,0	8,4	-0,7
B	623,7	24966	3,4	73,1	6,1	3,4
C	93,1	19324	23,6	47,9	12,3	-1,9
D	236,8	23113	8,7	66,8	8,7	2,0
E	412,0	23076	8,9	46,9	8,0	-3,1
F	566,7	24516	6,1	44,3	8,6	-3,0
G	331,9	22187	7,4	57,6	10,3	4,7
H	111,4	20614	16,3	63,8	13,9	5,2
I	489,0	25006	5,7	49,4	6,7	-2,6
J	287,4	23136	8,8	59,4	12,4	1,7
K	166,2	20707	14,1	74,0	13,0	3,6
L	388,1	23624	9,6	54,3	6,9	-0,4

Erläuterung:

Variable	Definition	Kurzbezeichnung
X_1	Bevölkerungsdichte (Einwohner je qkm)	ED
X_2	Bruttoinlandsprodukt (BIP) je Einwohner (€)	BIP
X_3	Anteil der Erwerbstätigen in der Landwirtschaft (Prozent)	EL
X_4	Wachstumsrate des BIP (in den letzten zehn Jahren)	WBIP
X_5	Allgemeine Geburtenziffer (Lebendgeborene je 1000 Einwohner)	GEB
X_6	Wanderungssaldo (Zu- minus Fortwanderungen je 1000 Einwohner)	WS

♦

Datensatz des 12-Regionen-Beispiels im SPSS-Dateneditor

Regionen - SPSS Daten-Editor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Fenster Hilfe

21 : ws

	region	ed	bip	el	wbip	geb	ws
1	A	212,4	20116	9,8	53,0	8,4	-,7
2	B	623,7	24966	3,4	73,1	6,1	3,4
3	C	93,1	19324	23,6	47,9	12,3	-1,9
4	D	236,8	23113	8,7	66,8	8,7	2,0
5	E	412,0	23076	8,9	46,9	8,0	-3,1
6	F	566,7	24516	6,1	44,3	8,6	-3,0
7	G	331,9	22187	7,4	57,6	10,3	4,7
8	H	111,4	20614	16,3	63,8	13,9	5,2
9	I	489,0	25006	5,7	49,4	6,7	-2,6
10	J	287,4	23136	8,8	59,4	12,4	1,7
11	K	166,2	20707	14,1	74,0	13,0	3,6
12	L	388,1	23624	9,6	54,3	6,9	-,4
13							

Datenansicht Variablenansicht

SPSS Prozessor ist bereit

Im SPSS-Dateneditor sind die Daten des 12-Regionen-Beispiel genauso angeordnet wie in der Datentabelle. Es handelt sich um eine **recheckige Anordnung** aus n=12 Zeilen und m=6 Variablen. Eine recheckige Anordnung von Größen wird in der linearen Algebra als **Matrix** bezeichnet. In der Statistik bezeichnet man rechteckige Anordnung der Daten in dieser Form als **Beobachtungsmatrix X**. Allgemein ist die nxm**-Beobachtungsmatrix X** durch

$$\underset{nxm}{\mathbf{X}} = \begin{array}{c} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \\ \begin{array}{cccc} \text{Vari-} & \text{Vari-} & \ldots & \text{Vari-} \\ \text{able 1} & \text{able 2} & & \text{able m} \end{array} \end{array} \begin{array}{c} \text{Merkmalsträger } 1 \\ \text{Merkmalsträger } 2 \\ \vdots \\ \text{Merkmalsträger } n \\ \\ \\ \end{array}$$

Standardisierte Beobachtungsmatrix und Korrelationsmatrix

Ausgangspunkt einer Faktorenanalyse ist die Bestimmung der **Korrelationsmatrix R**. Sie enthält die Korrelationen zwischen allen manifesten (beobachtbaren) Variablen, die bei einer Faktorenanalyse berücksichtigt werden. Zur Aufstellung der Korrelationsmatrix bietet es sich an, die vorliegenden Merkmalsausprägungen zunächst zu standardisieren. Eine **Standardisierung** einer Variablen geschieht dadurch, dass von der Ausprägung das arithmetische Mittel abgezogen und diese Differenz durch die Standardabweichung dividiert wird. Das Ergebnis sind neue standardisierte Variablen, die alle den Mittelwert 0 und die Standardabweichung 1 besitzen:

Standardisierung:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1,2,...,n;\ j = 1,2,...,m \tag{2.1}$$

z_{ij}: standardisierter Wert der Variablen j bei der Untersuchungseinheit i

Standardisierte Beobachtungsmatrix Z:

$$\underset{n \times m}{\mathbf{Z}} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}$$

Transponierte standardisierte Beobachtungsmatrix Z':

$$\underset{m\times n}{\mathbf{Z}'} = \begin{bmatrix} z_{11} & z_{21} & \cdots & z_{n1} \\ z_{12} & z_{22} & \cdots & z_{n2} \\ \vdots & \vdots & & \vdots \\ z_{1m} & z_{2m} & \cdots & z_{nm} \end{bmatrix}$$

Korrelationsmatrix R

(2.2) $$\underset{m\times m}{\mathbf{R}} = \frac{1}{n-1} \cdot \underset{m\times n}{\mathbf{Z}'} \cdot \underset{n\times m}{\mathbf{Z}}$$

• Eigenkorrelationen der Variablen

Auf der **Hauptdiagonale der Korrelationsmatrix R** stehen die Eigenkorrelationen r_{jj} der Variablen, die alle gleich 1 sein müssen: r_{jj} = 1 für alle j=1,2,…,m.

Eigenkorrelation r_{11} der Variablen X_1:

Faktor 1/(n-1) x Produkt der 1. Zeile von **Z**' (= 1. Spalte von **Z**) mit der 1. Spalte von **Z**

$$r_{11} = \frac{1}{n-1} \cdot [z_{11}\ z_{21} \cdots z_{n1}] \cdot \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix} = \frac{1}{n-1} \sum_{i=1}^{n} z_{i1}^2 = 1$$

• Korrelationen zwischen den Variablen

Die Nichthauptdiagonalelemente von **R** enthalten die Korrelationen r_{jk}, $j \neq k$, zwischen den Variablen.

Korrelation zwischen der Variablen X_1 und X_2:

Faktor 1/(n-1) x Produkt der 1. Zeile von **Z**' (= 1. Spalte von **Z**) mit der 2. Spalte von **Z**

$$r_{12} = \frac{1}{n-1} \cdot [z_{11}\ z_{21} \cdots z_{n1}] \cdot \begin{bmatrix} z_{12} \\ z_{22} \\ \vdots \\ z_{n2} \end{bmatrix} = \frac{1}{n-1} \sum_{i=1}^{n} z_{i1} \cdot z_{i2}$$

Wegen $r_{jk} = r_{kj}$ ist die Korrelationsmatrix **R** symmetrisch:

$$\underset{m \times m}{\mathbf{R}} = \begin{array}{c} \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r21 & 1 & r_{23} & \cdots & r2m \\ r_{31} & r_{32} & 1 & \cdots & r_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & 1 \end{bmatrix} \end{array} \begin{array}{l} \text{Korrel. von } X_1 \text{ mit allen anderen Var.} \\ \text{Korrel. von } X_2 \text{ mit allen anderen Var.} \\ \text{Korrel. von } X_3 \text{ mit allen anderen Var.} \\ \vdots \\ \text{Korrel. von } X_m \text{ mit allen anderen Var.} \end{array}$$

Exkurs: Herleitung des Korrelationskoeffizienten aus standardisierten Daten

Ausgangspunkt ist die Definition des Korrelationskoeffizienten nach Bravais und Pearson zwischen den Variablen X_j und X_k:

$$\text{(E1)} \quad r_{jk} = \frac{\sum (x_{ij} - \bar{x}_j) \cdot (x_{ik} - \bar{x}_k)}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2 \cdot \sum (x_{ik} - \bar{x}_k)^2}} \quad .$$

Die Summenbildung erfolgt hierin stets über alle i von 1 bis n. Die Varianzen der beiden Variablen X_j und X_k sind durch

$$s_j^2 = \frac{1}{n-1} \sum (x_{ij} - \bar{x}_j)^2 \quad \text{und} \quad s_k^2 = \frac{1}{n-1} \sum (x_{ik} - \bar{x}_k)^2$$

und ihre Standardabweichungen durch

$$s_j = \sqrt{\frac{1}{n-1} \sum (x_{ij} - \bar{x}_j)^2} \quad \text{und} \quad s_k = \sqrt{\frac{1}{n-1} \sum (x_{ik} - \bar{x}_k)^2}$$

gegeben. Wir multiplizieren jetzt die beiden Standardabweichungen s_j und s_k mit dem Faktor $\sqrt{n-1}$ und erhalten

$$\text{(E2)} \quad \sqrt{n-1} \cdot s_j = \sqrt{\sum (x_{ij} - \bar{x}_j)^2} \quad \text{und} \quad \sqrt{n-1} \cdot s_k = \sqrt{\sum (x_{ik} - \bar{x}_k)^2} \quad .$$

Mit (E2) können wir den Korrelationskoeffizienten (E1) in der Form

$$r_{jk} = \frac{\sum (x_{ij} - \bar{x}_j) \cdot (x_{ik} - \bar{x}_k)}{(n-1) \cdot s_j \cdot s_k} = \frac{1}{n-1} \sum \frac{x_{ij} - \bar{x}_j}{s_j} \cdot \frac{x_{ik} - \bar{x}_k}{s_k}$$

schreiben. Unter Verwendung der Standardisierung

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \text{und} \quad z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

erhalten wir schließlich die Formel

$$r_{jk} = \frac{1}{n-1} \sum z_{ij} \cdot z_{ik} \quad .$$

□

Berechnung der Korrelationsmatrix

Beispiel 2.3: Die für die Standardisierung benötigten arithmetischen Mittelwerte und Standardabweichungen können mit SPSS durch Wahl der Menüposition

ANALYSIEREN
    DESKRIPTIVE STATISTIKEN
        DESKRIPTIVE STATISTIKEN …

berechnet werden, in der die deskriptiven Statistiken „Mittelwert“, „Std.-Abweichung“ sowie „Minimum“ und „Maximum“ voreingestellt sind.

Nach Entfernen der Häkchen (durch Anklicken) bei den letzten beiden Items im Fenster „Deskriptive Statistik: Optionen“ gibt SPSS nach Betätigen der Schaltflächen „Weiter“ und „OK“ für die Variablen des 12-Regionen-Beispiels im SPSS-Viewer die Tabelle „Deskriptive Statistik“ aus:

Deskriptive Statistik

	N	Mittelwert	Standardabweichung
Einwohnerdichte	12	326,558	173,8854
BIP/Einwohner	12	22532,08	1940,683
Anteil der Erwerbstätigen in der Landwirtschaft (in %)	12	10,200	5,4734
Wachstum des BIP (in %)	12	57,542	10,0916
Geburten je 1000 Einwohner	12	9,608	2,6872
Fort- minus Zuwanderungen je 1000 Einwohner	12	,742	3,0684
Gültige Werte (Listenweise)	12		

Standardisierte Beobachtungsmatrix Z:

$$
\underset{12\times 6}{\mathbf{Z}} = \begin{array}{c}
\begin{array}{cccccc} \text{ED} & \text{BIP} & \text{EL} & \text{WBIP} & \text{GEB} & \text{WS} \end{array} \\
\begin{bmatrix}
-0{,}657 & -1{,}245 & -0{,}073 & -0{,}450 & -0{,}450 & -0{,}470 \\
1{,}709 & 1{,}254 & -1{,}242 & 1{,}542 & -1{,}306 & 0{,}866 \\
-1{,}343 & -1{,}653 & 2{,}448 & -0{,}955 & 1{,}002 & -0{,}861 \\
-0{,}516 & 0{,}299 & -0{,}274 & 0{,}917 & -0{,}338 & 0{,}410 \\
0{,}491 & 0{,}280 & -0{,}238 & -1{,}055 & -0{,}599 & -1{,}252 \\
1{,}381 & 1{,}022 & -0{,}749 & -1{,}312 & -0{,}375 & -1{,}219 \\
0{,}031 & -0{,}178 & -0{,}512 & 0{,}006 & 0{,}257 & 1{,}290 \\
-1{,}237 & -0{,}988 & 1{,}114 & 0{,}620 & 1{,}597 & 1{,}453 \\
0{,}934 & 1{,}275 & -0{,}822 & -0{,}807 & -1{,}082 & -1{,}089 \\
-0{,}225 & 0{,}311 & -0{,}256 & 0{,}184 & 1{,}039 & 0{,}312 \\
-0{,}992 & -0{,}940 & 0{,}713 & 1{,}631 & 1{,}262 & 0{,}932 \\
0{,}354 & 0{,}563 & -0{,}110 & -0{,}321 & -1{,}008 & -0{,}372
\end{bmatrix}
\end{array}.
$$

Der standardisierte Beobachtungswert der 1. Region für die 1. Variable, z_{11}, ergibt sich z.B. aus

$$
z_{11} = \frac{x_{11} - \bar{x}_1}{s_1} = \frac{212{,}4 - 326{,}558}{173{,}885} = -0{,}657
$$

Korrelationsmatrix **R**:

$$\boldsymbol{Z}=\begin{bmatrix} -0{,}657 & -1{,}245 & -0{,}073 & -0{,}450 & -0{,}450 & -0{,}470 \\ 1{,}709 & 1{,}254 & -1{,}242 & 1{,}542 & -1{,}306 & 0{,}866 \\ -1{,}343 & -1{,}653 & 2{,}448 & -0{,}955 & 1{,}002 & -0{,}861 \\ -0{,}516 & 0{,}299 & -0{,}274 & 0{,}917 & -0{,}338 & 0{,}410 \\ 0{,}491 & 0{,}280 & -0{,}238 & -1{,}055 & -0{,}599 & -1{,}252 \\ 1{,}381 & 1{,}022 & -0{,}749 & -1{,}312 & -0{,}375 & -1{,}219 \\ 0{,}031 & -0{,}178 & -0{,}512 & 0{,}006 & 0{,}257 & 1{,}290 \\ -1{,}237 & -0{,}988 & 1{,}114 & 0{,}620 & 1{,}597 & 1{,}453 \\ 0{,}934 & 1{,}275 & -0{,}822 & -0{,}807 & -1{,}082 & -1{,}089 \\ -0{,}225 & 0{,}311 & -0{,}256 & 0{,}184 & 1{,}039 & 0{,}312 \\ -0{,}992 & -0{,}940 & 0{,}713 & 1{,}631 & 1{,}262 & 0{,}932 \\ 0{,}354 & 0{,}563 & -0{,}110 & -0{,}321 & -1{,}008 & -0{,}372 \end{bmatrix}$$

$$\frac{1}{11}\cdot\underbrace{\begin{bmatrix} -0{,}657 & 1{,}709 & -1{,}343 & -0{,}516 & 0{,}491 & 1{,}381 & 0{,}031 & -1{,}237 & 0{,}934 & -0{,}225 & -0{,}992 & 0{,}354 \\ -1{,}245 & 1{,}254 & -1{,}653 & 0{,}299 & 0{,}280 & 1{,}022 & -0{,}178 & -0{,}988 & 1{,}275 & 0{,}311 & -0{,}940 & 0{,}563 \\ -0{,}073 & -1{,}242 & 2{,}448 & -0{,}274 & -0{,}238 & -0{,}749 & -0{,}512 & 1{,}114 & -0{,}822 & -0{,}256 & 0{,}713 & -0{,}110 \\ -0{,}450 & 1{,}542 & -0{,}955 & 0{,}917 & -1{,}055 & -1{,}312 & 0{,}006 & 0{,}620 & -0{,}807 & 0{,}184 & 1{,}631 & -0{,}321 \\ -0{,}450 & -1{,}306 & 1{,}002 & -0{,}338 & -0{,}599 & -0{,}375 & 0{,}257 & 1{,}597 & -1{,}082 & 1{,}039 & 1{,}262 & -1{,}008 \\ -0{,}470 & 0{,}866 & -0{,}861 & 0{,}410 & -1{,}252 & -1{,}219 & 1{,}290 & 1{,}453 & -1{,}089 & 0{,}312 & 0{,}932 & -0{,}372 \end{bmatrix}}_{\boldsymbol{Z}'}$$

$$\underbrace{\begin{bmatrix} 1 & 0{,}907 & -0{,}834 & -0{,}161 & -0{,}787 & -0{,}309 \\ 0{,}907 & 1 & -0{,}845 & -0{,}054 & -0{,}711 & -0{,}220 \\ -0{,}834 & -0{,}845 & 1 & -0{,}067 & 0{,}719 & 0{,}039 \\ -0{,}161 & -0{,}054 & -0{,}067 & 1 & 0{,}226 & 0{,}832 \\ -0{,}787 & -0{,}711 & 0{,}719 & 0{,}226 & 1 & 0{,}454 \\ -0{,}309 & -0{,}220 & 0{,}039 & 0{,}832 & 0{,}454 & 1 \end{bmatrix}}_{\boldsymbol{R}=\frac{1}{11}\cdot\boldsymbol{Z}'\boldsymbol{Z}}$$

Wir wollen beispielhaft die Berechnung des Korrelationskoeffizienten zwischen der Einwohnerdichte und dem BIP je Einwohner nachvollziehen:

$$r_{12}=\frac{1}{n-1}\sum_{i=1}^{n=12} z_{i1}\cdot z_{i2}$$

$$=\frac{1}{11}\cdot\left[(-0{,}657)\cdot(-1{,}245)+1{,}709\cdot 1{,}254+\ldots+0{,}354\cdot 0{,}563\right]=0{,}907$$

Mit SPSS lässt sich die Korrelationsmatrix über die Menüposition

ANALYSIEREN
KORRELATIONEN
BIVARIAT …
bestimmen:

Korrelationsmatrix

		Einwohnerdichte	BIP/ Einwohner	Anteil der Erwerbstätigen in der Landwirtschaft (in %)	Wachstum des BIP (in %)	Geburten je 1000 Einwohner	Fort- minus Zuwanderungen je 1000 Einwohner
Korrelation	Einwohnerdichte	1,000	,907	-,834	-,161	-,787	-,309
	BIP/Einwohner	,907	1,000	-,845	-,054	-,711	-,220
	Anteil der Erwerbstätigen in der Landwirtschaft (in %)	-,834	-,845	1,000	-,067	,719	,039
	Wachstum des BIP (in %)	-,161	-,054	-,067	1,000	,226	,832
	Geburten je 1000 Einwohner	-,787	-,711	,719	,226	1,000	,454
	Fort- minus Zuwanderungen je 1000 Einwohner	-,309	-,220	,039	,832	,454	1,000

Aus der **Korrelationsmatrix**: sind Faktoren („Supervariablen“) zu extrahieren“

Voraussetzung: Variablen müssen „Gemeinsamkeiten“ enthalten

„Gemeinsamkeiten“ der Variablen werden in der Statistik durch die **Korrelationen** erfasst

Korrelationen dürfen aber nicht nur zufällig von 0 abweichen, sondern sie müssen **substanziell** sein

$\Downarrow$

Signifikanztest der Korrelationskoeffizienten

♦

Test über den Korrelationskoeffizienten

Die Anwendung der Faktorenanalyse setzt voraus, dass „substanzielle“ Korrelationen zwischen den Variablen vorhanden sind. Bei rein zufälligen Beziehungen zwischen den Variablen erübrigt es sich, hinter den Variablen stehende Faktoren („Supervariablen“) zu identifizieren. „Substanziell“ wird in der Statistik als „signifikant“ interpretiert. Die berechneten **Korrelationskoeffizienten** r_{jk} zwischen X_j und X_k sind daher **auf Signifikanz** zu **testen**.

Nullhypothese: H_0: $\rho_{jk} = 0$
Alternativhypothese: H_1: $\rho_{jk} \neq 0$

ρ_{jk}: Korrelationskoeffizient zwischen X_j und X_k in der Grundgesamtheit

Eine Ablehnung der **Nullhypothese** bedeutet stets eine Annahme der **Alter nativhypothese,** da sie genau das Gegenteil behauptet. Die Nullhypothese eines Korrelationskoeffizienten ρ_{jk} in der Grundgesamtheit von 0 wird erst dann abgelehnt, wenn der berechnete Korrelationskoeffizient r_{jk} „substanziell“ oder „bedeutsam“, d.h. **signifikant** von 0 abweicht. Man sagt, der Korrelationskoeffizient r_{jk} ist signifikant, wenn seine Abweichung von 0 so groß ist, dass sie nicht mehr allein durch den Zufall erklärt werden kann. Bei einem solchen **Signifikanztest,** d.h. der Beurteilung wie groß die Abweichung von 0 sein darf, damit sie noch durch den Zufall erklärt werden kann, spielt die Streuung des Korrelationskoeffizienten (Standardabweichung) eine entscheidende Rolle. Lassen wir positive und negative Abweichungen zu, spricht man von einem **zweiseitiger Signifikanztest** vor.

Bei einem Signifikanztest ist stets ein **Signifikanzniveau** α zu wählen, dass die Wahrscheinlichkeit angibt, die Nullhypothese zu verwerfen, obwohl sie richtig ist (=**Irrtumswahrscheinlichkeit**) Die Ablehnung der Nullhypothese, wenn sie tatsächlich richtig ist, kennzeichnet den **Fehler 1. Art (**α-Fehler). Bei einem gebräuchlichen Signifikanzniveau von 5% (α=0,05) würde man auf dieser Grundlage in 95% aller Fälle die Nullhypothese korrekt annehmen, wenn sie tatsächlich richtig ist. In 5% aller Fälle würde man sie dagegen ablehnen, obwohl sie richtig ist (=Fehler 1. Art). Ein hohes Signifikanzniveau wird mit 1% beziffert. Ein niedriges Signifikanzniveau liegt bei 10%.

Aus der unten stehenden Übersicht geht hervor, dass bei einem statistischen Test stets zwei Fehler gemacht werden. Während der **Fehler 1. Art (α-Fehler)** kontrolliert werden kann, in dem man die Irrtumswahrscheinlichkeit vorgibt, ist dies beim **Fehler 2. Art** (**ß-Fehler**) nicht der Fall. Hierbei handelt es sich um den Fehler, den man begeht, wenn man die Nullhypothese beibehält, obwohl sie in Wirklichkeit falsch ist. Dieser Fehler hat die **Wahrscheinlichkeit ß**, der aber nur berechenbar ist, wenn man die Alternativhypothese konkretisiert.

Zustand Testentscheidung	H_0 wahr	H_0 falsch
H_0 wird abgelehnt	Entscheidung falsch (Fehler 1. Art) $P(H_1 \mid H_0) = \alpha$	Entscheidung richtig $P(H_1 \mid H_1) = 1 - ß$
H_0 wird beibehalten	Entscheidung richtig $P(H_0 \mid H_0) = 1 - \alpha$	Entscheidung falsch (Fehler 2. Art) $P(H_0 \mid H_1) = ß$

Prüfgröße des Tests:

(2.3) $$T_0 = \frac{r_{jk}}{\sqrt{1 - r_{jk}^2}} \cdot \sqrt{n-2} \sim t_{n-2}$$

Standardabweichung des Korrelationskoeffizienten (=Standardfehler):

$$(2.4) \qquad s(r_{jk}) = \frac{\sqrt{1-r_{jk}^2}}{\sqrt{n-2}}$$

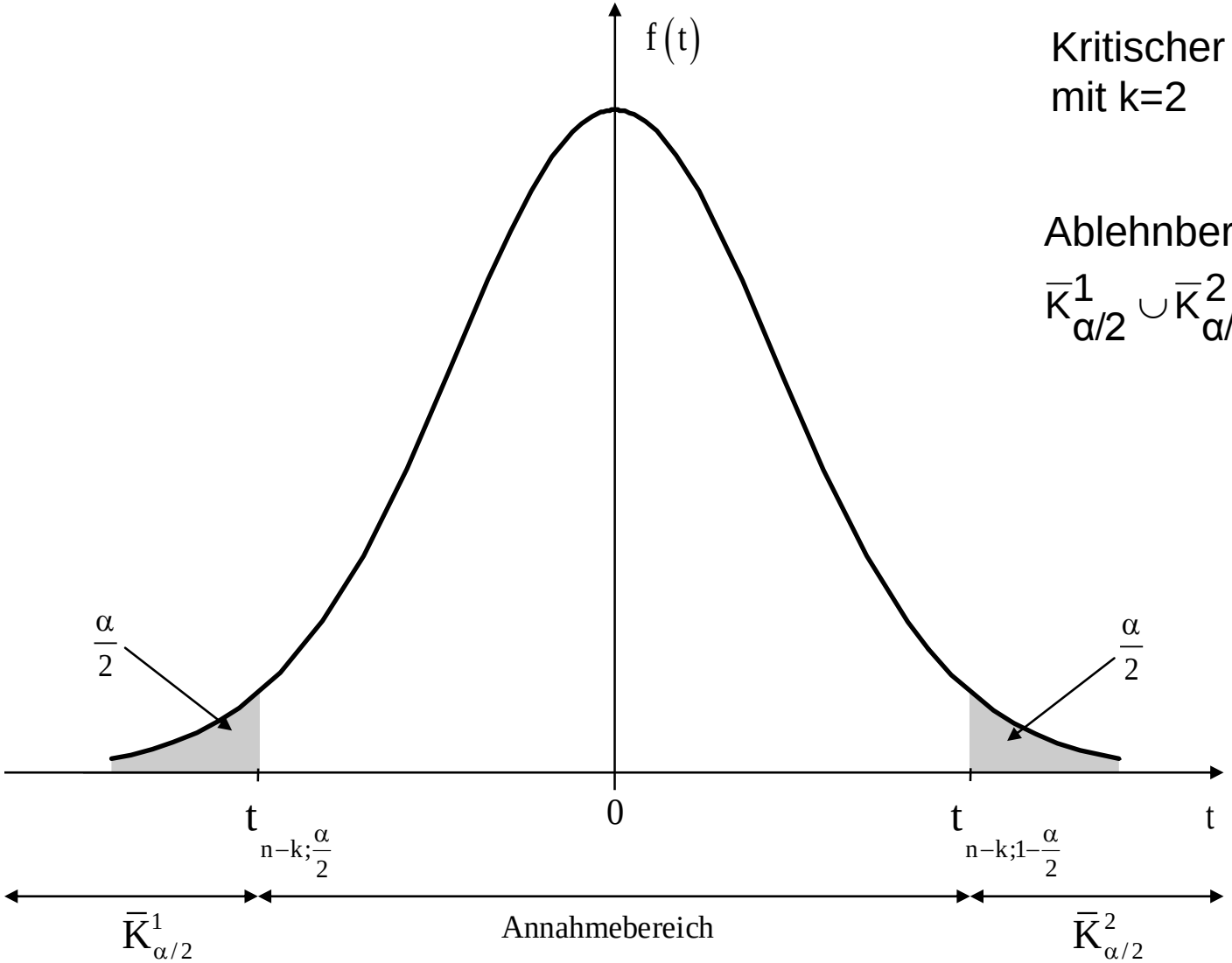
Bei normalverteilten Variablen ist die Teststatistik (Prüfgröße) t-verteilt mit n-2 Freiheitsgraden. Bei einem zweiseitigen Test ist der kritische Wert durch das (1-α/2)_ Quantil $t_{n-2;1-\alpha/2}$ gegeben. Die Nullhypothese einer fehlenden Korrelation zwischen den Variablen X_j und X_k wird bei einer Irrtumswahrscheinlichkeit α abgelehnt, wenn der realisierte Wert t der Prüfgröße absolut den kritischen Wert $t_{n-2;1-\alpha/2}$, übersteigt:

$|t| > t_{n-2;1-\alpha/2}$ → Nullhypothese verwerfen.

Der Wert der Prüfgröße liegt in diesem Fall im Ablehnbereich $\overline{K}^1_{\alpha/2} \cup \overline{K}^2_{\alpha/2}$

Falls der absolute Wert der Prüfgröße t dagegen kleiner oder gleich dem kritischen Wert $t_{n-2;1-\alpha/2}$ ist, liegt die Prüfgröße im Annahmebereich. Die Abweichung des empirischen Korrelationskoeffizienten r_{jk} ist dann unter Berücksichtigung seines Standardfehlers nicht so groß, als dass sie nicht auf den Zufall zurückgeführt werden könnte. Die Nullhypothese wird dann beibehalten.

Abbildung: Annahme- und Ablehnbereiche bei dem zweiseitigen Signifikanztest



Kritischer Wert $t_{n-k;1-\alpha/2}$
mit k=2 $(t_{n-k;\alpha/2} = -t_{n-k;1-\alpha/2})$

Ablehnbereiche

$\bar{K}^1_{\alpha/2} \cup \bar{K}^2_{\alpha/2}$: grau unterlegt

Bei einer Ablehnung der Nulhypothese wird der Zusammenhang zwischen den beiden Variablen X_j und X_k bei der vorgegebenen Irrtumswahrscheinlichkeit α als statistisch gesichert betrachtet.

Beispiel 2.4

Zwischen den beiden Variablen ED und BIP ist ein Korrelationskoeffizient $r_{ED,BIP}$ von 0,907 ermittelt worden. Um zu überprüfen, ob sich der berechnete Korrelationskoeffizient signifikant von 0 unterscheidet, bestimmen wird den Wert der Teststatistik:

$$t = \frac{r_{ED,BIP}}{\sqrt{1 - r_{ED,BIP}^2}} \cdot \sqrt{n-2}$$

$$= \frac{0{,}907}{\sqrt{1 - 0{,}823}} \cdot \sqrt{12-2} = 6{,}821.$$

Bei einem einseitigen und zweiseitigen Test ergeben sich für unterschiedliche α folgende kritischen Werte der t-Verteilung mit 10 Freiheitsgraden:

	•α	•1%	•5%	•10%
•Einseitiger Test	•$t_{10;1-\alpha}$	•2,764	•1,812	•1,372
•Zweiseitiger Test	•$1_{10;1-\alpha/2}$	•3,169	•2,228	•1,812

Einseitiger Test: Wenn die Richtung des Zusammenhangs a priori bekannt ist.
Zweiseitiger Test: Ohne A-priori-Kenntnis der Richtung des Zusammenhangs.

Bei allen gängigen Signifikanzniveaus α ist damit $t > t_{10;1-\alpha}$ bzw. $t > t_{10;1-\alpha/2}$, so dass die Nullhypothese $\rho_{ED,BIP} = 0$ abzulehnen und von einem statistisch signifikanten Zusammenhang zwischen beiden Variablen auszugehen ist.

Über die bereits bekannten Menüpunkte (ANALYSIEREN, KORRELATIONEN, BIVARIAT …) lassen wir erneut die Korrelationsmatrix von SPSS berechnen, wobei wir jetzt die Signifikanz mit ausweisen lassen:

Korrelationen und zweiseitiger Signifikanztest

Korrelationen

		Einwohnerdichte	BIP/ Einwohner	Anteil der Erwerbstätigen in der Landwirtschaft (in %)	Wachstum des BIP (in %)	Geburten je 1000 Einwohner	Fort- minus Zuwanderungen je 1000 Einwohner
Einwohnerdichte	Korrelation nach Pearson	1	.907**	-.834**	-.161	-.787**	-.309
	Signifikanz (2-seitig)		.000	.001	.617	.002	.328
	N	12	12	12	12	12	12
BIP/Einwohner	Korrelation nach Pearson	.907**	1	-.845**	-.054	-.711**	-.220
	Signifikanz (2-seitig)	.000		.001	.869	.010	.492
	N	12	12	12	12	12	12
Anteil der Erwerbstätigen in der Landwirtschaft (in %)	Korrelation nach Pearson	-.834**	-.845**	1	-.067	.719**	.039
	Signifikanz (2-seitig)	.001	.001		.836	.008	.904
	N	12	12	12	12	12	12
Wachstum des BIP (in %)	Korrelation nach Pearson	-.161	-.054	-.067	1	.226	.832**
	Signifikanz (2-seitig)	.617	.869	.836		.480	.001
	N	12	12	12	12	12	12
Geburten je 1000 Einwohner	Korrelation nach Pearson	-.787**	-.711**	.719**	.226	1	.454
	Signifikanz (2-seitig)	.002	.010	.008	.480		.138
	N	12	12	12	12	12	12
Fort- minus Zuwanderungen je 1000 Einwohner	Korrelation nach Pearson	-.309	-.220	.039	.832**	.454	1
	Signifikanz (2-seitig)	.328	.492	.904	.001	.138	
	N	12	12	12	12	12	12

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Unter „Signifikanz“ steht die Überschreitungswahrscheinlichkeit (p-Wert) des z-Wertes für den jeweiligen empirischen Korrelationskoeffizienten unter der Nullhypothese. Gibt man eine Irrtumswahrscheinlichkeit α vor, dann ist ein Korrelationskoeffizient signifikant, wenn der p-Wert kleiner als α ist. Bei einem gewöhnlichen Signifikanzniveau von 5% sind alle Korrelationen signifikant („substanziell“), deren p-Wert kleiner als 0,05 ist.