Prof. Dr. Reinhold Kosfeld

Formelsammlung zur Multivariaten Statistik

1. Einleitung

Merkmale (Variablen): $X_1, X_2, \ldots, X_m$

Univariate Datenanalyse

Arithmetisches Mittel der Variablen X_j	$\bar{x}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}$
Varianz der Variablen X_j	$s_j^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$
Standardabweichung der Variablen X_j	$s_j = \sqrt{s_j^2}$

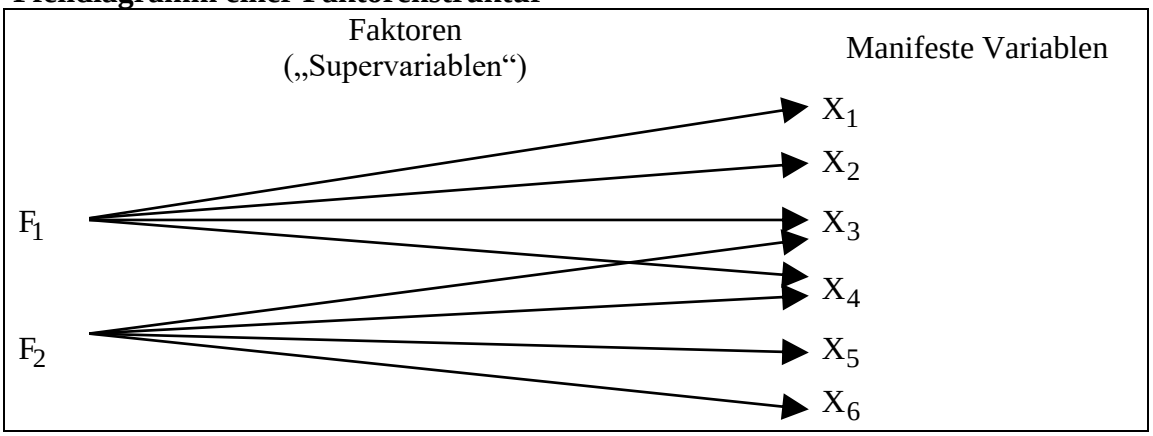
Bivariate Datenanalyse

Kovarianz Wertebereich: Menge der reellen Zahlen (keine Begrenzung)	$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \cdot \left(x_{ik} - \bar{x}_k\right)$
Korrelationskoeffizient (nach Bravais und Pearson) Wertebereich: $-1 \leq r_{jk} \leq 1$	$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \cdot \left(x_{ik} - \bar{x}_k\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2 \cdot \sum_{i=1}^{n} \left(x_{ik} - \bar{x}_k\right)^2}}$
Einfache Regressionsanalyse Regressionsfunktion: $\hat{x}_k = a + b \cdot x_j$	Kleinst-Quadrate-Schätzer: $\hat{b} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sum_{i=1}^{n} (x_{ij} - \bar{x})^2}$ $\hat{a} = \frac{\sum_{i=1}^{n} x_{ik}}{n} - \hat{b} \cdot \frac{\sum_{i=1}^{n} x_{ij}}{n}$ Determinationskoeffizient: $R^2 = r_{jk}^2$

2. Faktorenanalyse

2.1 Grundidee der Faktorenanalyse

Pfeildiagramm einer Faktorenstruktur



Beobachtungsmatrix X:

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

Standardisierung:

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

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:

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

Korrelation zwischen X_j und X_k:

$$r_{jk} = \frac{1}{n-1} \cdot [z_{1j}\ z_{2j} \cdots z_{nj}] \cdot \begin{bmatrix} z_{1k} \\ z_{2k} \\ \vdots \\ z_{nk} \end{bmatrix} = \frac{1}{n-1} \sum_{i=1}^{n} z_{ij} \cdot z_{ik}$$

Test über den Korrelationskoeffizienten

Null- und alternativhypothese bei zweiseitigem Test:

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

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

Prüfgröße des Tests:

$$T = \frac{r_{jk}}{\sqrt{1-r_{jk}^2}} \cdot \sqrt{n-2} \sim t_{n-2}\ .$$

Standardabweichung des Korrelationskoeffizienten (=Standardfehler):

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

2.2 Faktorenextraktion

• Faktorenanalytisches Modell

$X_1, X_2, \ldots, X_m$: manifeste (beobachtbare) Variablen
$F_1, F_2, \ldots, F_p$: gemeinsame Faktoren („Supervariablen“)

Strukturelle Gleichung für Merkmalsträger i und Variable X_j:

$$z_{ij} = a_{j1} \cdot f_{i1} + a_{j2} \cdot f_{i2} + \cdots + a_{mp} \cdot f_{ip} + e_{ij}$$

Matrixform der strukturellen Gleichung:

$$\underset{m\times n}{\mathbf{Z}'} = \underset{m\times p}{\mathbf{A}} \cdot \underset{p\times n}{\mathbf{F}'} + \underset{m\times n}{\mathbf{E}'}$$

$$\underset{mxp}{\begin{bmatrix} z_{11} & z_{21} & \cdots & z_{n1} \\ z_{12} & z_{22} & \cdots & z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1m} & z_{2m} & \cdots & z_{nm} \end{bmatrix}} = \underset{pxn}{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix}} \cdot \underset{mxn}{\begin{bmatrix} f_{11} & f_{21} & \cdots & f_{n1} \\ f_{12} & f_{22} & \cdots & f_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1p} & f_{2p} & \cdots & f_{np} \end{bmatrix}} + \underset{mxn}{\begin{bmatrix} e_{11} & e_{21} & \cdots & e_{n1} \\ e_{12} & e_{22} & \cdots & e_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1m} & e_{2m} & \cdots & e_{nm} \end{bmatrix}}$$

$F_1, F_2, \ldots, F_k$: gemeinsame Faktoren
$E_1, E_2, \ldots, E_m$: Einzelrestfaktoren

Fundamentaltheorem der Faktorenanalyse (= Zerlegung der Korrelationsmatrix):

$$\underset{m\times m}{\mathbf{R}} = \underset{m\times p}{\mathbf{A}} \cdot \underset{p\times m}{\mathbf{A}'} + \frac{1}{n-1} \cdot \underset{m\times n}{\mathbf{E}'} \cdot \underset{n\times m}{\mathbf{E}}$$

mxp-Matrix A: Matrix der Faktorladungen (Faktormatrix)

Faktormatrix (=Matrix der Faktorladungen):

$$\underset{mxp}{\mathbf{A}} = \begin{array}{c} \begin{array}{cccc} \text{Fakt. 1} & \text{Fakt. 2} & \dots & \text{Fakt. p} \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \vdots & a_{mp} \end{bmatrix} \end{array} \begin{array}{l} \text{Variable 1} \\ \text{Variable 2} \\ \vdots \\ \text{Variable m} \end{array}$$

$$= [\,\underset{\substack{\text{Faktor}-\\ \text{vektor 1}}}{\mathbf{a}_1} \quad \underset{\substack{\text{Faktor}-\\ \text{vektor 2}}}{\mathbf{a}_2} \quad \cdots \quad \underset{\substack{\text{Faktor}-\\ \text{vektor p}}}{\mathbf{a}_p}\,]$$

• Hauptkomponentenmethode

Ausgangspunkt: $\underset{m\times m}{\mathbf{R}} = \underset{m\times p}{\mathbf{A}} \cdot \underset{p\times m}{\mathbf{A}'}$ mit $p = m$

Bestimmung des 1. Faktors

gesucht: mx1-Vektor (**Ladungsvektor**) $\mathbf{a}_1$:

$$\underset{mx1}{\mathbf{a}_1} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} \begin{array}{l} \text{Ladung des Faktors } F_1 \text{ auf der Variablen } X_1 \\ \text{Ladung des Faktors } F_1 \text{ auf der Variablen } X_2 \\ \\ \text{Ladung des Faktors } F_1 \text{ auf der Variablen } X_m \end{array}$$

Approximation:

$$\underset{m\times m}{\mathbf{R}} \cong \underset{m\times 1}{\mathbf{a}_1} \cdot \underset{m\times 1}{\mathbf{a}_1'}$$

Eigenwert des 1. Faktors:

$$\lambda_1 = \sum_{j=1}^{m} a_{j1}^2 = \mathbf{a}_1' \cdot \mathbf{a}_1$$

Eigenwertproblem:

$$(\mathbf{R} - \lambda_1 \cdot \mathbf{I}) \cdot \mathbf{a}_1 = \mathbf{0}$$

Charakteristische Gleichung:

$$|\mathbf{R} - \lambda_1 \cdot \mathbf{I}| = 0$$

Beziehung zwischen dem Faktorvektor $\mathbf{a}_1$ und dem normiertem Eigenvektor $\mathbf{a}_1^n$ $[(\mathbf{a}_1^n)'\mathbf{a}_1^n = 1]$:

$$\mathbf{a}_1 = \mathbf{a}_1^n \cdot \sqrt{\lambda_1}$$

Bestimmung des 2. Faktors

Eigenwertproblem:

$$\left(\mathbf{R} - \lambda_2 \cdot \mathbf{I}\right) \cdot \mathbf{a}_2 = \mathbf{0}$$

Charakteristische Gleichung:

$$\left|\mathbf{R} - \lambda_2 \cdot \mathbf{I}\right| = 0$$

Eigenwert des 2. Faktors:

$$\lambda_2 = \sum_{j=1}^{m} a_{j2}^2 = \mathbf{a}_2' \cdot \mathbf{a}_2$$

Beziehung zwischen dem Faktorvektor $\mathbf{a}_2$ und dem normiertem Eigenvektor $\mathbf{a}_2^n$ $[(\mathbf{a}_2^n)' a_2^n = 1]$:

$$\mathbf{a}_1 = \mathbf{a}_1^n \cdot \sqrt{\lambda_1}$$

2.3 Bestimmung der Faktorenzahl

- Kumuliertes Varianzkriterium
- Scree-Test
- Kaiserkriterium

2.4 Das Kommunalitätenproblem

Anknüpfungspunkt**:** Fundamentaltheorem der Faktorenanalyse:

$$\underset{m\times m}{\mathbf{R}} = \underbrace{\underset{m\times p}{\mathbf{A}} \cdot \underset{p\times m}{\mathbf{A}'}}_{\mathbf{R}_h} + \underbrace{\frac{1}{n-1} \cdot \underset{m\times n}{\mathbf{E}'} \cdot \underset{n\times m}{\mathbf{E}}}_{\mathbf{R}^{Rest}}, \; p < m$$

Restkorrelationsmatrix $\mathbf{R}^{Rest}$:

$$\mathbf{R}^{Rest} = \frac{1}{n-1} \cdot \mathbf{E}' \cdot \mathbf{E} \qquad \text{oder} \qquad \underset{m\times m}{\mathbf{R}^{Rest}} = \begin{pmatrix} u_1^2 & \approx 0 & \dots & \approx 0 \\ 0 & u_2^2 & \dots & \approx 0 \\ \vdots & \vdots & \ddots & \vdots \\ \approx 0 & \approx 0 & \dots & u_m^2 \end{pmatrix}$$

u_j^2**:** Einzelrestvarianzen (merkmalsspezifische Varianzen)

Reproduzierte Korrelationsmatrix:

$$\tilde{\mathbf{R}}_h = \mathbf{A} \cdot \mathbf{A}'$$

Hauptdiagonalelemente von $\tilde{\mathbf{R}}_h$: Kommunalitäten h_j^2

$$h_j^2 = \sum_{k=1}^{p} \cdot a_{jk}^2 \quad \text{oder} \quad h_j^2 = 1 - u_j^2, \; j=1,2,\dots,m$$

• Hauptachsen-Faktorenanalyse (Faktorenanalyse i.e.S.)

Ausgangspunkt der Faktorextraktion: **reduzierte Korrelationsmatrix**

$$\underset{mxm}{\mathbf{R}_h} = \begin{bmatrix} h_1^2 & r_{12} & r_{13} & \cdots & r_{1m} \\ r21 & h_2^2 & r_{23} & \cdots & r2m \\ r_{31} & r_{32} & h_3^2 & \cdots & r_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & h_m^2 \end{bmatrix}$$

Schätzverfahren für Kommunalitäten

- Maximaler absoluter Korrelationskoeffizient

$$h_j^2 = \max\left\{\left|r_{jh}\right|\right\}$$

- Multipler Determinationskoeffizient

$$h_j^2 = R_j^2 \quad \text{mit} \quad R_j^2 = 1 - \frac{1}{r^{jj}}$$

r^{jj}: j-tes Diagonalelement der inversen Korrelationsmatrix $\mathbf{R}^{-1}$

- Iterationsverfahren

Unter Verwendung von Anfangsschätzern für die Kommunalität wird die Faktorextraktion iteriert bis Konvergenz eintritt oder eine Kommunalität den Wert 1 erreicht. Die sukzessive Verbesserung der Anfangsschätzer ergibt sich dadurch, dass am Ende einer jeden Iteration neue Kommunalitäten aus den aktuellen Ladungen der extrahierten Faktoren berechnet werden.

2.5 Das Rotationsproblem

Interpretationsproblem bei Ladungen im mittleren Bereich: $|0{,}3| \le a_{jk} \le |0{,}7|$

Rotationsproblem:
Drehung des Koordinatensystems, in dem die Variablenvektoren (= Vektoren, die die Faktorladungen jeweils einer Variablen enthalten = Zeilen der Faktormatrix) liegen („Faktorraum“), um eine **Einfachstruktur** zu erreichen.

Rechtwinklige Rotation (bei Rotation im mathematischen Uhrzeigersinn**):**

$$\underset{m\times p}{\mathbf{A}}^{RR} = \underset{m\times p}{\mathbf{A}} \cdot \underset{p\times p}{\mathbf{T}}$$

$\mathbf{A}^{RR}$: rechtwinklig rotierte Faktormatrix
T: Transformationsmatrix

Transformationsmatrix bei zwei gemeinsamen Faktoren:

$$\underset{2x2}{\mathbf{T}} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

Varimax-Methode

Varianz der quadrierten Ladungen des Faktors F_k:

$$s_{F_k}^2 = \frac{1}{m}\sum_{k=1}^{m}\left(a_{jk}^2 - \overline{a_k^2}\right)^2$$

(Streuung der quadrierten Ladungen des Faktors F_k um den Mittelwert seiner Ladungsquadrate)

mit $\overline{a_k^2} = \frac{1}{m}\sum_{j=1}^{m} a_{jk}^2$ (Arithm. Mittel der quadrierten Ladungen des Faktors F_k)

Varimax-Kriterium: $\text{VAR} = \frac{1}{p}\sum_{k=1}^{p} s_{F_k}^2$ → Max.!

<u>Anmerkung:</u> In der Maximierungsprozedur werden die Faktorladungen auf Wurzeln aus den Kommunalitäten bezogen, damit die Variablen das gleiche Gewicht haben (Normierung). Die Ladungsquadrate a_{jk}^2 sind daher durch die normierten Ladungsquadrate a_{jk}^2/h_j^2 zu ersetzen.

2.6 Bestimmung der Faktorwerte

„Matrizenmanipulation“ bei Hauptkomponentenmethode:

$$\underset{p\times n}{\mathbf{F}'_{KQ}} = \left(\mathbf{A}'\underset{p\times p}{\mathbf{A}}\right)^{-1}\underset{p\times m}{\mathbf{A}'}\underset{m\times n}{\mathbf{Z}'}$$

(KQ: Kleinst-Quadrate-Schätzer)

Faktorwerte-Koeffizientenmatrix: $(\mathbf{A}'\mathbf{A})^{-1}\cdot\mathbf{A}'$

Regressionsmethode bei Faktorenananalyse i.e.S.:

$$\mathbf{F}'_{REG} = \mathbf{A}'\cdot\mathbf{R}^{-1}\cdot\mathbf{Z}'$$

Faktorwerte-Koeffizientenmatrix: $\mathbf{A}'\cdot\mathbf{R}^{-1}$

3. Varianzanalyse

Varianzanalyse (ANOVA)

Einfaktorielle Varianzanalyse

Einfluss eines Faktors (=qualitative unabh. Variable) auf eine quantitative abh. Variable

Zweifaktorielle Varianzanalyse

Einfluss von zwei Faktoren (=qualitative unabh. Variablen) auf eine quantitative abh. Variable

ohne Interaktion | mit Interaktion

3.1 Einfaktorielle Varianzanalyse

Faktor A (qualitative unabh. Variable) → Abhängige Variable Y (quantitativ)

Modell der einfaktoriellen Varianzanalyse

Wert der abhängigen Variablen y_{ij} mit Doppelindex:
- 1. Index: i-te Untersuchungseinheit, $i=1,2,\ldots,n_j$,
- 2. Index: j-te Gruppe (Faktorstufe A_j), $j=1,2,\ldots,p$

Varianzanalytisches Modell:

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Gesamtmittel der Grundgesamtheit: μ

Mittelwerte der p Stufen des Faktors A in der Grundgesamtheit:

$$\mu_j = \mu + \alpha_j\,, \quad j = 1,2,\ldots,p$$

Effekt des Faktors A: $\alpha_j = \mu_{j\bullet} - \mu$

Störgröße: $\varepsilon_{ij} \sim N(0,\sigma^2)$

Erwartungswert der abhängigen Variablen y:

$$E(y_{ij}) = \mu + \alpha_j \quad \text{oder} \quad E(y_{ij}) = \mu_j$$

Nullhypothese $H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$

Stichprobenkenngrößen (Stichprobenäquivalente):

Mittelwerte der p Stufen des Faktors A:

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}\,,\ j=1,2,\ldots,p$$

Gesamtmittelwert:

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij}\,.$$

mit $\sum_{j=1}^{p} n_j = N$ (Gesamtzahl der Beobachtungswerte)

Stichprobenäquivalent des Erwartungswerts: $\bar{y}_j$

Residuum:

$$e_{ij} = y_{ij} - \bar{y}_j$$

Varianzanalytischer Test

Totale (gesamte) Abweichungsquadratsumme:

$$Q_T = \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}\right)^2 = \underbrace{\sum_{j=1}^{p} n_j \cdot \left(\bar{y}_j-\bar{y}\right)^2}_{Q_A} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}_j\right)^2}_{Q_R}$$

Mittlere Abweichungsquadratsumme zwischen den Gruppen:

$$MQ_A = \frac{Q_A}{p-1} = \frac{1}{p-1}\cdot\sum_{j=1}^{p} n_j \cdot \left(\bar{y}_j-\bar{y}\right)^2$$

Mittlere Abweichungsquadratsumme innerhalb der Gruppen (=Residualvarianz):

$$MQ_R = \frac{Q_R}{N-p} = \frac{1}{N-p}\cdot\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}_j\right)^2$$

Prüfgröße des F-Tests:

$$F = \frac{MQ_A}{MQ_R} = \frac{\frac{1}{p-1}\cdot\sum_{j=1}^{p} n_j \cdot \left(\bar{y}_j-\bar{y}\right)^2}{\frac{1}{N-p}\cdot\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(y_{ij}-\bar{y}_j\right)^2}$$

Verteilung der Prüfgröße F: F-Verteilung mit p-1 und N-p Freiheitsgaden

3.2 Zweifaktorielle Varianzanalyse

Zweifaktorielle Varianzanalyse

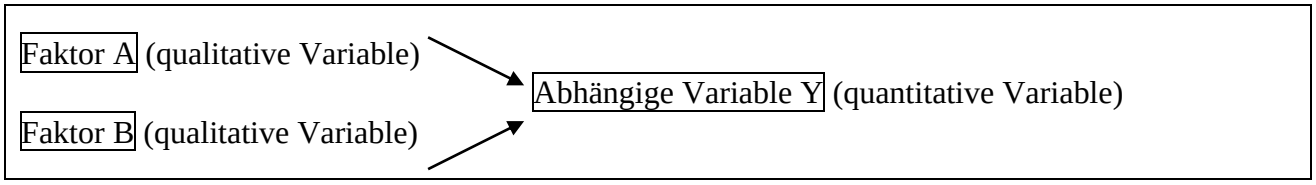
ohne Wechselwirkungen (ohne Interaktion)	**mit Wechselwirkungen** (mit Interaktion)
Einfluss der Faktoren A und B	Einfluss der Faktoren A und B und der Wechselwirkung zwischen A und B

Zweifaktorielle Varianzanalyse

Orthogonale Varianzanalyse gleiche Zellenbesetzung	Nonorthogonale Varianzanalyse ungleiche Zellenbesetzung

3.2.1 Orthogonale Varianzanalyse

A. Modell der zweifaktoriellen Varianzanalyse ohne Wechselwirkungen



Wert der abhängigen Variablen, y_{ijk}, mit Dreifachindex:

- 1. Index: i-te Untersuchungseinheit, $i=1,2,\ldots,n_{jk}$,
- 2. Index: j-te Ausprägung des Faktors A (Faktorstufe A_j), $j=1,2,\ldots,p$,
- 3. Index: k-te Ausprägung des Faktors B (Faktorstufe B_k), $k=1,2,\ldots,q$.

Varianzanalytisches Modell ohne Wechselwirkung:

$$y_{ijk} = \mu + \alpha_j + \beta_k + \varepsilon_{ijk} .$$

Gesamtmittel der Grundgesamtheit: μ

Mittelwerte der p Stufen des Faktors A in der Grundgesamtheit:

$$\mu_{j\bullet} = \mu + \alpha_j , \quad j = 1,2,\ldots,p$$

Mittelwerte der q Stufen des Faktors B in der Grundgesamtheit:

$$\mu_{\bullet k} = \mu + \beta_j , \quad j = 1,2,\ldots,p$$

Effekt des Faktors A: $\alpha_j = \mu_{j\bullet} - \mu$,

Effekt des Faktors B: $ß_k = \mu_{\bullet k} - \mu$

Störgröße: $\varepsilon_{ijk} \sim N(0,\sigma^2)$

Erwartungswert der abhängigen Variablen y:

$$E(y_{ijk}) = \mu + \alpha_j + \beta_k \quad \text{oder} \quad E(y_{ijk}) = \mu_{j\bullet} + \mu_{\bullet k} - \mu$$

Nullhypothesen:

$$H_0 \text{ (Faktor A)}: \quad \mu_{1\bullet} = \ldots = \mu_{j\bullet} = \ldots = \mu_{p\bullet} .$$

und

$$H_0 \text{ (Faktor B)}: \quad \mu_{\bullet 1} = \ldots = \mu_{\bullet k} = \ldots = \mu_{\bullet q} .$$

Stichprobenkenngrößen (Stichprobenäquivalente)**:**

Mittelwerte der p Stufen des Faktors A (Zeilenmittelwerte)**:**

$$\bar{y}_{j\bullet} = \frac{1}{q \cdot n} \cdot \sum_{k=1}^{q} \sum_{i=1}^{n} y_{ijk} \text{ , } j=1,2,\ldots,p$$

Mittelwerte der q Stufen des Faktors B (Spaltenmittelwerte):

$$\bar{y}_{\bullet k} = \frac{1}{p \cdot n} \cdot \sum_{j=1}^{p} \sum_{i=1}^{n} y_{ijk}$$

Gesamtmittelwert:

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{i=1}^{n} y_{ijk} \,.$$

mit $N = \sum_{j=1}^{p} \sum_{k=1}^{q} n_{jk} = p \cdot q \cdot n$ (Gesamtzahl der Beobachtungen)

Stichprobenäuivalent des Erwartungswerts

$$\bar{y}_{jk}^{erw} = \bar{y} + (\bar{y}_{j\bullet} - \bar{y}) + (\bar{y}_{\bullet k} - \bar{y}) \quad \text{oder} \quad \bar{y}_{jk}^{erw} = \bar{y}_{j\bullet} + \bar{y}_{\bullet k} - \bar{y}$$

Residuum:

$$e_{ijk} = y_{ijk} - \bar{y}_{jk}^{erw}$$

Varianzanalytischer Test

Totale (gesamten) Abweichungsquadratsumme**:**

$$Q_T = \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{i=1}^{n} (y_{ijk} - \bar{y})^2 =$$

$$\underbrace{n \cdot q \cdot \sum_{j=1}^{p} \left(\bar{y}_{j\bullet} - \bar{y}\right)^2}_{Q_A} + \underbrace{n \cdot p \cdot \sum_{k=1}^{q} \left(\bar{y}_{\bullet k} - \bar{y}\right)^2}_{Q_B} + \underbrace{\sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{i=1}^{n} \left(y_{ijk} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y}\right)^2}_{Q_R}$$

Mittlere Abweichungsquadratsumme des Faktors A:

$$MQ_A = \frac{Q_A}{p-1} = \frac{n \cdot q}{p-1} \sum_{j=1}^{p} \left(\bar{y}_{j\bullet} - \bar{y}\right)^2$$

Mittlere Abweichungsquadratsumme des Faktors A

$$MQ_B = \frac{Q_B}{q-1} = \frac{n \cdot p}{q-1} \sum_{k=1}^{q} \left(\bar{y}_{\bullet k} - \bar{y}\right)^2$$

Mittlere Abweichungsquadratsumme der Residuen (Residualvarianz):

$$MQ_R = \frac{Q_R}{N-(p+q-1)} = \frac{1}{N-(p+q-1)} \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{i=1}^{n} \left(y_{ijk} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y}\right)^2 \quad \text{oder}$$

$$MQ_R = \frac{Q_R}{N-(p+q-1)} = \frac{1}{N-(p+q-1)} \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{i=1}^{n} \left(y_{ijk} - \bar{y}_{ijk}^{erw}\right)^2$$

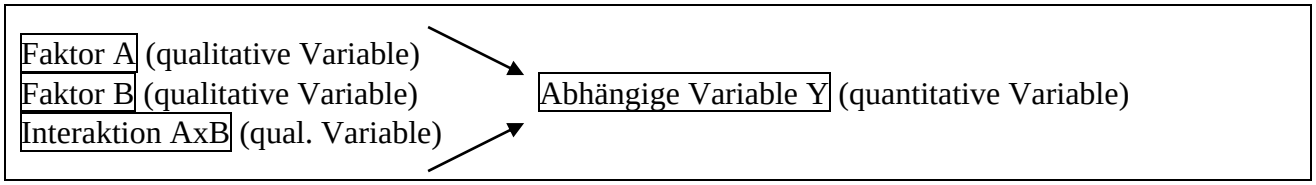
Prüfgrößen der F-Tests für die Faktoren A und B:

$$F_A = \frac{Q_A/(p-1)}{Q_R/[N-(p+q-1)]} = \frac{MQ_A}{MQ_R}$$

$$F_B = \frac{Q_B/(q-1)}{Q_R/[N-(p+q-1)]} = \frac{MQ_B}{MQ_R} \,.$$

Verteilung der Prüfgrößen F_A und F_B: F-Verteilung mit p-1 und N-(p+q-1) bzw. q-1 und N-(p+q-1) Freiheitsgraden.

B. Modell der zweifaktoriellen Varianzanalyse mit Wechselwirkungen



Varianzanalytisches Modell mit Wechselwirkung:

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

Interktionseffekt:

$$(\alpha\beta)_{jk} = \mu_{jk} - \mu_{j\bullet} - \mu_{\bullet k} + \mu$$

Störgröße: $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Erwartungswert der abhängigen Variablen y:

$$E(y_{ijk}) = \mu_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

Nullhypothese in Bezug auf die Interaktion AxB:

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = ... = (\alpha\beta)_{pq} = 0$$

Stichprobenkenngrößen (Stichprobenäquivalente)**:**

Mittelwertes $\bar{y}_{jk}$ der Kombination der Faktorstufen A_j und B_k (Zellenmittelwerte):

$$\bar{y}_{jk} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{ijk}$$

Stichprobenäquivalent des Interaktionseffekts:

$AB_{jk} = \bar{y}_{jk} - (\bar{y}_{j\bullet} - \bar{y}) - (\bar{y}_{\bullet k} - \bar{y}) - \bar{y}$ oder $AB_{jk} = \bar{y}_{jk} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y}$.

Stichprobenäuivalent des Erwartungswerts: $\bar{y}_{jk}$

Residuum:

$$e_{ijk} = y_{ijk} - \bar{y}_{jk}^{erw}$$

Varianzanalytischer Test

Totale (gesamten) Abweichungsquadratsumme**:**

$$Q_T = \sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{i=1}^{n}(y_{ijk} - \bar{y})^2 = \underbrace{n \cdot q \sum_{j=1}^{p}\left(\bar{y}_{j\bullet} - \bar{y}\right)^2}_{Q_A} + \underbrace{n \cdot p \sum_{k=1}^{q}\left(\bar{y}_{\bullet k} - \bar{y}\right)^2}_{Q_B}$$

$$+ \underbrace{n \sum_{j=1}^{p}\sum_{k=1}^{q}\left(\bar{y}_{jk} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y}\right)^2}_{Q_{A\times B}} + \underbrace{\sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{i=1}^{n}\left(y_{ijk} - \bar{y}_{jk}\right)^2}_{Q_R}$$

Mittlere Abweichungsquadratsumme der Interaktion AxB:

$$MQ_{A\times B} = \frac{Q_{A\times B}}{(p-1)(q-1)}$$

Mittlere Abweichungsquadratsumme der Residuen (Residualvarianz)**:**

$$MQ_R = \frac{Q_R}{N-p\cdot q} = \sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{i=1}^{n}\left(y_{ijk} - \bar{y}_{jk}\right)^2 / (N-p\cdot q)$$

Prüfgrößen für die Effekte von A, B und AxB:

$$F_A = \frac{Q_A/(p-1)}{Q_R/(N-p\cdot q)} = \frac{MQ_A}{MQ_R}$$ (F-verteilt mit p-1 und N-p·q Freiheitsgraden)

$$F_B = \frac{Q_B/(q-1)}{Q_R/(N-p\cdot q)} = \frac{MQ_B}{MQ_R}$$ (F-verteilt mit q-1 und N-p·q Freiheitsgraden)

$$F_{A\times B} = \frac{Q_{A\times B}/[(p-1)(q-1)]}{Q_R/(N-p\cdot q)} = \frac{MQ_{A\times B}}{MQ_R}$$ (F-verteilt mit (p-1)(q-1) und N-p·q Freiheitsgraden)

3.2.2 Nonorthogonale Varianzanalyse

Ungleichen Zellenhäufigkeiten (=Häufigkeiten der Kombinationen der Faktorstufen A_j mit B_k) n_{jk}: unbalanciertes Design

⇒ Faktoren sind nicht mehr orthogonal („korrelierte“ Faktoren)

Folge: Die separat berechneten Abweichungsquadratsummen summieren sich nicht mehr zur totalen Abweichungsquadratsumme Q_T

Bei der **zweifaktoriellen Varianzanalyse mit Wechselwirkung** gilt bei ungleichen Zellenhäufigkeiten n_{jk} nicht mehr

$Q_A + Q_B + Q_{AxB} + Q_R = Q_T$ (bei orthogonaler Varianzanalyse),

sondern

$Q_A + Q_B + Q_{AxB} + Q_R \neq Q_T$ (bei nonorthogonaler Varianzanalyse).

In dem Fall

$Q_A + Q_B + Q_{AxB} > Q_T - Q_R$

würde der Gesamteinfluss von A, B und AxB überschätzt werden, während er in dem Fall

$Q_A + Q_B + Q_{AxB} < Q_T - Q_R$

unterschätzt würde.

Schätzmethoden der nonorthogonalen Varianzanalyse

- **Hierarchische Methode**
  Die Haupteffekte, also die Faktoren A und Faktor B, werden sukzessive berechnet. Der Haupteffekt A enthält den mit B gemeinsamen Effekt, d.h. die Faktorstufen A_1, A_2, …, A_p werden jeweils zusätzlich nach den Stufen B_1, B_2, …, B_q des Faktors B untergliedert.
- **Experimentelle Methode**
  Die Haupteffekte werden gegenseitig korrigiert. Bei der Berechnung eines Haupteffekts werden vor- und nachstehende Faktoren berücksichtigt.
- **Regressionsmethode**
  Alle Effekte werden simultan auf ihren Einfluss untersucht. Entsprechend wird eine simultane Korrektur gemeinsamer Einflüsse vorgenommen.

4. Clusteranalyse

Klassifikation von Objekten

Arbeitsschritt 1	Quantifizierung der Ähnlichkeit von Objektpaaren			
	Wahl eines Ähnlichkeitsmaßes in Abhängigkeit vom Skalenniveau			
	Metrische Skala	Ordninalskala	Nominalskala	Gemischtes Skalenniveau

$\Downarrow$

Arbeitsschritt 2	Gruppierung der Objekte aufgrund ihrer Ähnlichkeit		
	Wahl eines Klassifikationsverfahrens		
	Arbeitsschritt 2a		Arbeitsschritt 2b
	Gruppenzahl nicht bekannt	$\Rightarrow$	Gruppenzahl bekannt
	Wahl und Einsatz eines hierarchischen Verfahrens		Wahl und Einsatz eines Partitionsverfahrens

4.1 Distanz- und Ähnlichkeitsmaße

Distanzmaße

Euklidische Distanz:

- aus originären Beobachtungswerten:

$$ED(i,j) = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}$$

- aus standardisierten Beobachtungswerten:

$$ED(i,j) = \sqrt{\sum_{k=1}^{m} \left(z_{ik} - z_{jk}\right)^2} \quad \text{mit } z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \; i = 1, 2, ..., n; \; k = 1, 2, ..., m$$

Quadrierte euklidische Distanz:

$$QED(i,j) = \sum_{k=1}^{m} \left(z_{ik} - z_{jk}\right)^2$$

City-Block-Distanz:

$$CBD(i,j) = \sum_{k=1}^{m} \left|z_{ik} - z_{jk}\right|$$

Änlichkeitskoeffizienten

A. Dichotome Merkmale

Zweiwertige Merkmale: Modalitäten hier mit 0 und 1 kodiert (≙binäre Kodierung)

Vierfeldertafel

Objekt i	Objekt j		Σ
	1	0	
1	a	b	a + b
0	c	d	c + d
Σ	a + c	b + d	a + b + c + d = m

Simple-Matching-Koeffizient (SMC):

$$SMC(i,j) = \frac{a+d}{m}$$

Jaccard-Koeffizient:

$$J(i,j) = \frac{a}{a+b+c}$$

RR-Koeffizient:

$$RR(i,j) = \frac{a}{m}.$$

Beziehung zwischen den Ähnlichkeitskoeffizienten:

$$RR(i,j) \le J(i,j) \le SMC(i,j)$$

B. Polytome Merkmale

Mehrkategorielle Merkmale, die auf einem nominalen oder ordinalen Skalenniveau gemessen werden können

Ähnlichkeitsmessung bei polytomen Merkmalen**:** auf der Basis der für dichotome Merkmale entwickelten Ähnlichkeitskonzepte nach Vornahme einer Dichotomisierung

Merkmale mit zwei Kategorien: 1 Dummy-Variable mit den Werten 0 und 1

Merkmale mit 3 Kategorien: 2 0,1-Dummy-Variablen

Merkmale mit p Kategorien: p-1 0,1-Dummy-Variablen

Gewicht der Dummy-Variablen: $1/(p-1)$

4.3 Hierarchische Klassifikationsverfahren

Agglomerativen Verfahren:

- Ausgehend von der feinsten Gruppierung einelementiger Cluster werden sukzessive die "ähnlichsten" Klassen bis hin zur gröbsten Gruppierung eines n-elementigen Clusters zusammengefasst,
- Gruppierung wird von Stufe zu Stufe heterogener, da zu den Clustern immer „entferntere“ Objekte hinzukommen

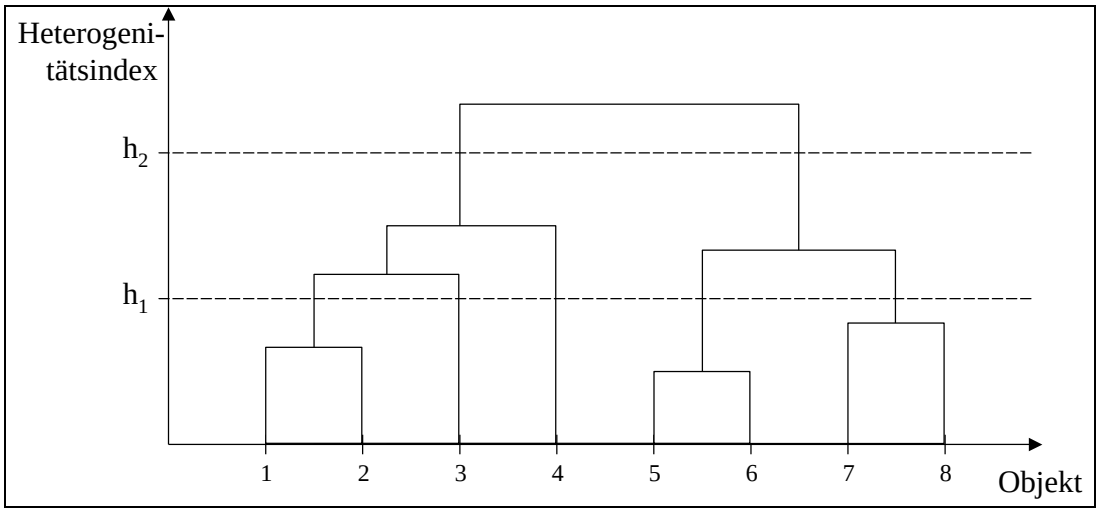
Divisive Verfahren:

Genau umgekehrte Vorgehensweise, d.h. von einem n-elementigen Cluster zum einelementigen Cluster

Ablauf einer hierarchischen Klassifikation (Agglomerationsverfahren)

Start: Feinste Partition (n einelementige Cluster)		
Berechnung der Ausgangsdistanz- (Ähnlichkeits-)Matrix		
Ermittlung der beiden Cluster mit der geringsten Distanz (größten Ähnlichkeit)	←	
Vereinigung der beiden Cluster mit der geringsten Distanz (größten Ähnlichkeit)		
Gibt es nur noch eine Gruppe (= n-elementiges Cluster)?	nein →	Neuberechnung der Distanzmatrix (Ähnlichkeitsmatrix)
↓ ja		
Ende		

Dendrogramm



• **Single-Linkage-Verfahren**

Skalenniveau: beliebig

Distanz zwischen zwei Clustern:

$$D(C_g, C_h) = \min\{d(i,j)\},\ i \in C_g,\ j \in C_h .$$

• **Complete-Linkage-Verfahren**

Skalenniveau: beliebig

$$D(C_g, C_h) = \max\{d(i,j)\},\ i \in C_g,\ j \in C_h$$

• **Average-Linkage-Verfahren**

Skalenniveau: beliebig

$$D(C_g, C_h) = \frac{1}{n_g \cdot n_h} \sum_{i \in C_g} \sum_{j \in C_h} d(i,j).$$

Fusionsregel der Verfahren Single-Linkage, Complete-Linkage und Average-Linkage:

$D(C_r, C_s) = \min\{D(C_g, C_h)\},\ g \neq h,\ \Rightarrow$ Fusion der Cluster C_r und C_s.

Ward-Verfahren

Skalenniveau: metrische Skala

Fusion von Clustern auf der Grundlage eines Varianzkriteriums

Streuung innerhalb des g-ten Clusters:

$$V_g = \sum_{k=1}^{m} \sum_{i \in C_g} \left(z_{ik} - \bar{z}_{gk}\right)^2 ,$$

mit $\bar{z}_{gk} = \frac{1}{n_g} \sum_{i \in C_g} z_{ik}$ (Mittelwert des k-ten Merkmals für das g-te Cluster)

Varianzkriterium (=Gesamtstreuung innerhalb der G Cluster einer gegebenen Partition):

$$V = \sum_{g=1}^{G} \sum_{k=1}^{m} \sum_{i \in C_g} \left(z_{ik} - \bar{z}_{gk}\right)^2$$

Bei Anwendung des **Varianzkriteriums der Ward-Methode** werden in jeder Stufe des Fusionsprozesses stets die **beiden Cluster fusioniert**, die zu einer **minimalen Erhöhung der Gesamtstreuung V** führen.

Bestimmung der Kriteriumsgröße V:

$$\Delta V(C_g \cup C_h) = \frac{n_g \cdot n_h}{n_g + n_h} \sum_{k=1}^{m} \left(\bar{z}_{gk} - \bar{z}_{hk}\right)^2$$

Auf jeder Stufe des Klassifikationsprozesses sind für alle Clusterpaare die Zuwächse ΔV zu berechnen. **Vereinigt** wird auf einer bestimmten Stufe jeweils das **Clusterpaar mit** dem **geringsten ΔV-Wert**.

Struktogramm: Koordinatensystem in dem Streuungszuwachs ΔV gegen die Clusterzahl abgetragen wird

4.4 Partitionierende Klassifikationsverfahren

Partitionierenden Verfahren:

- gegeben: eine Zerlegung der Objektmenge in G Cluster aus, die jedoch nicht als "optimal" angesehen wird;
- Verbesserung der Ausgangspartition – die etwa durch ein hierarchisches Klassifikationsverfahren bestimmt worden ist - durch Umgruppierung von Objekten;
- Anzahl der Cluster bleibt in jeder Stufe konstant (G Cluster).

Partitionsverfahren

Optimierende Austausverfahren (Hill-climbing methods)	**Iterative Minimal-Distanz-Verfahren**
Verbesserung einer Partition auf- grund eines Optimierungskriteriums	Verbesserung einer Partition durch Zuordnung eines Objekts zu einem Cluster mit einer geringeren Distanz zum Clusterschwerpunkt
- Varianzkriterium - Spurkriterium - Determinantenkriterium	- K-Means-Verfahren

K-Means-Verfahren

Clusterschwerpunkte (Zentroide): Komponenten die arithmetischen Mittelwerte (means) der Merkmale innerhalb der Gruppen sind:

$$\bar{\mathbf{z}}_g = \begin{pmatrix} \bar{z}_{g1} & \bar{z}_{g2} & \dots & \bar{z}_{gm} \end{pmatrix}', \; g = 1,2,\dots,G .$$

Komponenten der Zentroide $\bar{\mathbf{z}}_g$:

Arithmetische Mittelwerte (means) der Merkmale innerhalb der Gruppen

Distanzmaß für Abstände zwischen Objekten und Cluster: **quadrierten euklidische Distanz**

$$d(i, C_g) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{gk})^2 .$$

Wechselbedingung:

Ein Objekt i, das sich im Cluster C_g befindet, wird umgruppiert, wenn es ein Cluster C_h, $h \neq g$, mit der Eigenschaft

$$d(i, C_h) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{hk})^2 < d(i, C_g) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{gk})^2, h \neq g,$$

gibt.

Zuordnungsvorschrift:

Das Objekt i wird dem Cluster C_ℓ mit der Minimal-Distanz-Eigenschaft

$$d(i, C_\ell) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{\ell k})^2 = \min_g \{d(i, C_g)\} = \min_g \left\{ \sum_{k=1}^{m} (z_{ik} - \bar{z}_{gk})^2 \right\}$$

zugeordnet.

5. Diskriminanzanalyse

5.1 Idee der Diskriminanzanalyse

Diskrimination:

Trennung bestehender Gruppen aufgrund einer Reihe verfügbarer Merkmale (=Merkmalsvariablen)

Merkmalsvariablen: metrisch skaliert (auch dichotom, d.h. mit Ausprägungen 0 und 1)

Gruppen (1, 2,…, G): als Ausprägungen des nominalskalierten Merkmals Gruppe aufzufassen

Klassifikation:

Sind die Gruppen aufgrund der Merkmalsvariablen „optimal“ getrennt, lassen sich neue Objekte aufgrund ihrer Ausprägungen bei den Merkmalsvariablen einer der Gruppen zugeordnen

Zwei-Gruppen-Fall

Diskriminanzfunktion bei zwei Merkmalsvariablen X_1 und X_2:

$$D = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2$$

Diskriminanzkoeffizienten (=Diskriminanzgewichte): a_0, a_1 und a_2

Diskriminanzwerte (nach Einsetzen der Beobachtungswerte in die Diskriminanzfunktion):

- bei Objekten, deren Gruppenzugehörigkeit nicht bekannt ist

$$d_i = a_0 + a_1 \cdot x_{i1} + a_2 \cdot x_{i2}$$

- bei Objekten, deren Gruppenzugehörigkeit bekannt ist

$$d_{gi} = a_0 + a_1 \cdot x_{gi1} + a_2 \cdot xg_{i2}$$

x_{gij}: Beobachtungswerte des i-ten Objekts in der g-ten Gruppe beim Merkmal X_j,
d_{gi}: Diskriminanzwert des i-ten Objekts in der g-ten Gruppe

1. Gruppe: n_1 Objekte, 2. Gruppe: n_2 Objekte

Gruppe 1: $d_{11}, d_{12}, ..., d_{1n_1}$,

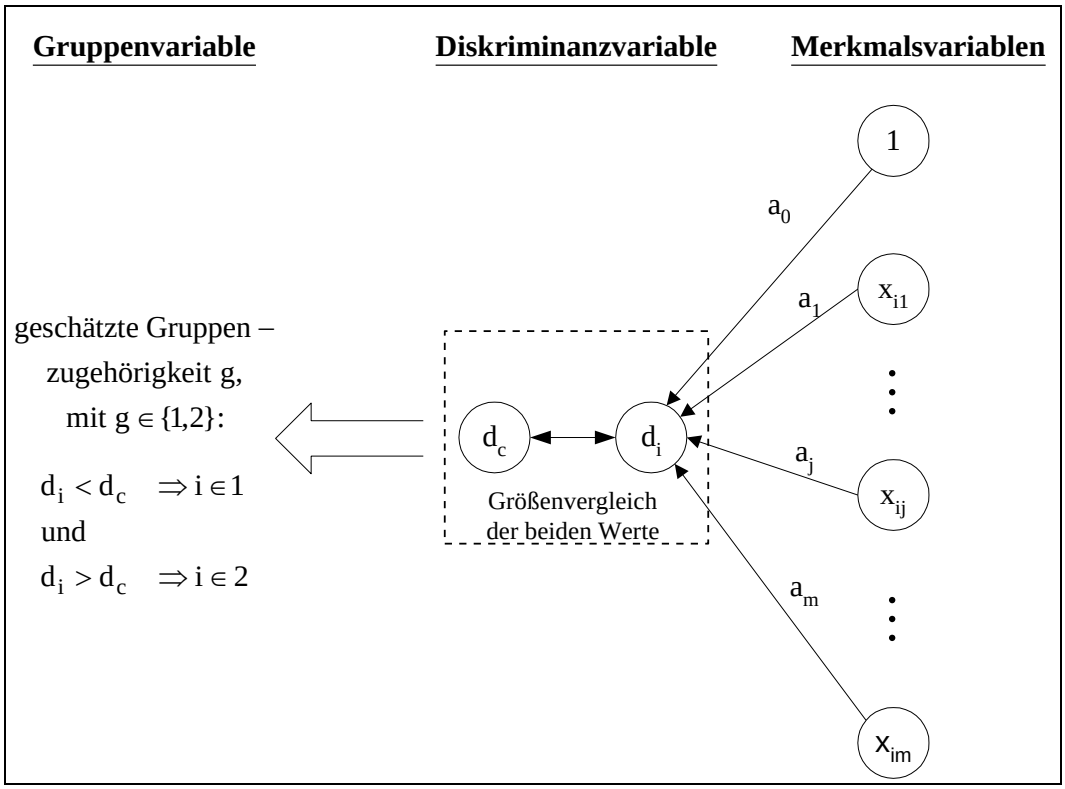
Gruppe 2: $d_{21}, d_{22}, ..., d_{2n_2}$

Trennpunkt d_c der beiden Gruppen auf der Diskriminanzachse:

$$d_c = \bar{d} = \frac{n_1 \cdot \bar{d}_1 + n_2 \cdot \bar{d}_2}{n_1 + n_2}$$

mit $\bar{d}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} d_{gi}$, g=1,2 (Gruppenmittelwerte auf der Diskriminanzachse)

Analytische Idee der Diskriminanzanalyse (Zwei-Gruppen-Fall)



a_j : Diskriminanzkoeffizient der j-ten Merkmalsvariablen
d_i : i-te Ausprägung der Diskriminanzvariablen D
d_c : kritischer Diskriminanzwert (Trennwert)
x_{ij}: Ausprägung der j-ten Merkmalsvariablen bei der i-ten Untersuchungseinheit

5.2 Lineare Diskriminanzfunktion

A. Allg. Darstellung des Zwei-Gruppen-Falls bei m Merkmalsvariablen

Diskriminanzfunktion:

- für neu zu klassifizierende Objekte

$$d_i = a_0 + a_1 \cdot x_{i1} + a_2 \cdot x_{i2} + \ldots + a_m \cdot x_{im}$$

- für bereits klassifizierte Objekte

$$d_{gi} = a_0 + a_1 \cdot x_{gi1} + a_2 \cdot x_{gi2} + \ldots + a_m \cdot x_{gim}$$

gesucht: **Diskriminanzkoeffizienten** a_0 und a_j (j=1,…,m)

Diskriminanzkriterium:

$$\lambda = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung innerhalb den Gruppen}}$$

Gesamtmittelwert auf der Diskriminanzachse:

$$\bar{d} = \frac{1}{n_1 + n_2} \cdot \sum_{g=1}^{2} \sum_{i=1}^{n_g} d_{gi}$$

Streuung zwischen den Gruppen:

$$B_d = \sum_{g=1}^{2} n_g \left(\bar{d}_g - \bar{d}\right)^2$$

Streuung innerhalb der Gruppen:

$$W_d = \sum_{g=1}^{2} \sum_{i=1}^{n_g} \left(d_{gi} - \bar{d}_g\right)^2$$

Maximierung des **Diskriminanzkriteriums**:

$$\lambda = \frac{B_d}{W_d} = \frac{\sum_{g=1}^{2} n_g \cdot \left(\bar{d}_g - \bar{d}\right)^2}{\sum_{g=1}^{2} \sum_{i=1}^{n_g} \left(d_{gi} - \bar{d}_g\right)^2} \rightarrow \text{Max!}$$

Das **Diskriminanzkrierium** λ ist im Hinblick auf die **Diskriminanzkoeffizienten** a_0, a_1, …, a_m zu maximieren, d.h. die Diskriminanzkoeffizienten sind so zu bestimmen, dass das Verhältnis aus der Streuung zwischen den Gruppen (B_d) und der Streuung innerhalb der Gruppen (W_d) maximal wird.

B. Diskriminanzfunktion und Diskriminanzkriterium in Matrixform

Diskriminanzfunktion:

$$\boxed{\mathbf{d} = \mathbf{a_0} + \mathbf{X} \cdot \mathbf{a}}$$

$$\text{mit } \mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}, \quad \mathbf{a}_0 = \begin{bmatrix} a_0 \\ a_0 \\ \vdots \\ a_0 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{nm} \end{bmatrix}.$$

Mittelwertvektor der Diskriminanzvariablen D:

$$\bar{\mathbf{d}} = \mathbf{a}_0 + \bar{\mathbf{X}} \cdot \mathbf{a}, \ , \ \bar{\mathbf{d}} = \begin{bmatrix} \bar{d} \\ \bar{d} \\ \vdots \\ \bar{d} \end{bmatrix}$$

mit der Mittelwertmatrix $\overline{X}$: $\underset{nxm}{\overline{\mathbf{X}}} = \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \end{bmatrix}$

Streuung zwischen den Gruppen:

$B_d = \mathbf{a}' \cdot \mathbf{B} \cdot \mathbf{a}$

Streuung innerhalb der Gruppen:

$W_d = \mathbf{a}' \cdot \mathbf{W} \cdot \mathbf{a}$

Streuungszerlegung:

T	=	**B**	+	**W**
Totale Streuungsmatrix		Streuungsmatrix zwischen den Gruppen		Streuungsmatrix innerhalb der Gruppen

Totale Streuungsmatrix T vom Typ mxm (hier für m=2):

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi1} - \overline{x}_1)^2 & \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi1} - \overline{x}_1)(x_{gi2} - \overline{x}_2) \\ \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi2} - \overline{x}_2)(x_{gi1} - \overline{x}_1) & \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi2} - \overline{x}_2)^2 \end{bmatrix}$$

Streuungsmatrix zwischen den Gruppen (hier für m=2)**:**

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \sum_{g=1}^{G=2} n_g \cdot (\overline{x}_{g1} - \overline{x}_1)^2 & \sum_{g=1}^{G=2} n_g \cdot (\overline{x}_{g1} - \overline{x}_1)(\overline{x}_{g2} - \overline{x}_2 \\ \sum_{g=1}^{G=2} n_g \cdot (\overline{x}_{g2} - \overline{x}_2)(\overline{x}_{g1} - \overline{x}_1) & \sum_{g=1}^{G=2} n_g \cdot (\overline{x}_{g2} - \overline{x}_2)^2 \end{bmatrix}.$$

Streuungsmatrix innerhalb der Gruppen (hier für m=2):

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi1} - \overline{x}_{g1})^2 & \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi1} - \overline{x}_{g1})(x_{gi2} - \overline{x}_{g2}) \\ \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi2} - \overline{x}_{g2})(x_{gi1} - \overline{x}_{g1}) & \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} (x_{gi1} - \overline{x}_{g1})^2 \end{bmatrix}$$

Gesamtmittel der Merkmalsvariablöen X_j:

$$\overline{x}_j = \frac{1}{n_1 + n_2} \cdot \sum_{g=1}^{G=2} \sum_{i=1}^{n_g} x_{gij}$$

Gruppenmittelwerte der Merkmalsvariablen X_j:

$$\overline{x}_{gj} = \frac{1}{n_g} \cdot \sum_{i=1}^{n_g} x_{gij}\text{ , g=1,2,…,G}$$

Maximierungsproblem:

$$\lambda = \frac{B_d}{W_d} = \frac{\mathbf{a}'\cdot\mathbf{B}\cdot\mathbf{a}}{\mathbf{a}'\cdot\mathbf{W}\cdot\mathbf{a}} \rightarrow \text{Max!}$$

gesucht: **Vektor der Diskriminanzkoeffizienten**, **a**, der das **Diskriminanzkriterium** λ maximiert

C. Lösungsverfahren

Eigenwertproblem:

$$\left(\mathbf{W}^{-1}\cdot\mathbf{B} - \lambda\cdot\mathbf{I}\right)\cdot\mathbf{a} = \mathbf{0}$$

gesucht: größter Eigenwert λ der Matrix $\mathbf{W}^{-1}\mathbf{B}$ mit dem zugehörigen Eigenvektor **a**

Charakteristische Gleichung:

$$\left|\mathbf{W}^{-1}\cdot\mathbf{B} - \lambda\cdot\mathbf{I}\right| = 0$$

Ausgangslösung: Eigenvektor **a** (vorzugsweise auf die Länge 1 normiert)

$$\mathbf{a}'\mathbf{a} = \sum_{j=1}^{m} a_j^2 = a_1^2 + a_2^2 + \ldots + a_m^2 = 1$$

unnormierte (rohe) Diskriminanzkoeffizienten: Komponenten a_j des Eigenvektors **a**:.

Bestimmung der normierten Diskriminanzfunktion

Normierung der Diskriminanzkoeffizienten, dass die gemeinsame (gepoolte) Varianz der Diskriminanzvariablen $(s_d^2)^{pool}$ gleich 1 wird: $\left(s_d^2\right)^{pool} \overset{!}{=} 1$

gepoolte Varianz der Diskriminanzvariablen: gewogenes arithmetisches Mittel der gruppen-spezifischen Varianzen:

$$\frac{1}{n-G}\cdot W_d \overset{!}{=} 1 \quad \Rightarrow \quad \tilde{\mathbf{a}}'\cdot\mathbf{W}\cdot\tilde{\mathbf{a}} \overset{!}{=} n-G$$

Ermittlung der normierten **Diskriminanzkoeffizienten** $\tilde{a}_j$ (aus beliebigem Lösungsvektor **a**):

$$\tilde{\mathbf{a}} = \gamma\cdot\mathbf{a}$$

Proportionalitätskonstante γ:

$$\gamma = \sqrt{\frac{n-G}{\mathbf{a}'\cdot\mathbf{W}\cdot\mathbf{a}}}$$

Vektor $\tilde{\mathbf{a}}$ der **normierten Diskriminanzkoeffizienten**:

$$\tilde{\mathbf{a}} = \sqrt{\frac{n-G}{\mathbf{a}'\cdot\mathbf{W}\cdot\mathbf{a}}}\cdot\mathbf{a}\,.$$

Konstante $\mathbf{a_0}$ der Diskriminanzfunktion:

$$\mathbf{a}_0 = -\overline{\mathbf{X}}\cdot\tilde{\mathbf{a}} \quad \text{mit} \quad \underset{n\times 1}{\overline{\mathbf{d}}} \overset{!}{=} \underset{n\times 1}{\mathbf{0}} \quad \text{(kritischer Diskriminanzwert } d_c \text{ wird 0)}$$

normierte Diskriminanzfunktion:

$\mathbf{d} = \mathbf{a}_0 + \mathbf{X} \cdot \tilde{\mathbf{a}}$.

Bestimmung der standardisierten Diskriminanzkoeffizienten

standardisierte Diskriminanzwerte: Mittelwert 0 und gepoolte Varianz 1

Standardisierte Diskriminanzkoeffizienten: $a_j^* = \tilde{a}_j \cdot s_j^{pool}$

mit $s_j^{pool} = \frac{1}{\sqrt{n-2}} w_{jj}^{1/2}$ (gepoolte Standardabweichung der Merkmalsvariablen X_j)

Vektor **a*** der **standardisierten Diskriminanzkoeffizienten**:

$$\mathbf{a}^* = \frac{1}{\sqrt{n-G}} \operatorname{diag}(\mathbf{W})^{1/2} \cdot \tilde{\mathbf{a}}$$

Relative Bedeutung der einzelnen Variablen X_j auf Grundlage der standard. Diskriminanz-koeffizienten:

$$PT_j = \left(\left| a_j^* \right| \Big/ \sum_{k=1}^{m} \left| a_k^* \right| \right) \cdot 100\%$$

Bestimmung der Strukturkoeffizienten

Korrelationen zwischen Merkmalsvariablen und der Diskriminanzvariablen (berücksichtigen Korrelationen zwischen den Merkmalvariablen)

Strukturkoeffizienten $r_{dx_j}^{pool}$:

$$r_{dx_j}^{pool} = \frac{s_{dx_j}^{pool}}{\sqrt{\left(s_j^2\right)^{pool}} \cdot \sqrt{\left(s_d^2\right)^{pool}}} = \frac{\frac{1}{n-G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(x_{ij} - \bar{x}_{gj}\right) \cdot \left(d_i - \bar{d}_g\right)}{\sqrt{\left(s_j^2\right)^{pool}} \cdot \sqrt{\left(s_d^2\right)^{pool}}}$$

Vektors $\mathbf{r}_{dx}$ der **Strukturkoeffizienten:**

$$\underset{m\times 1}{\mathbf{r}_{dx}^{pool}} = \underbrace{\frac{1}{\sqrt{n-G}}}_{\text{Skalar}} \cdot \underbrace{\operatorname{diag}(\mathbf{W})^{-1/2}}_{m\times m} \cdot \underset{m\times m}{\mathbf{W}} \cdot \underset{m\times 1}{\tilde{\mathbf{a}}}$$

Relative Bedeutung der Variablen X_j auf Grundlage der Strukturkoeffizienten:

$$PT_j = \left(\left| r_{dx_k}^{pool} \right| \Big/ \sum_{k=1}^{m} \left| r_{dx_k}^{pool} \right| \right) \cdot 100\%$$

5.3 Güte der Diskriminanzfunktion

Die Beurteilung der **Güte der Diskriminanzfunktion** bezieht sich auf ihre Aussagefähigkeit zur **Trennung der Gruppen**.

Basis für die Beurteilung der Güte der Diskriminanzfunktion:

- Maßzahlen (Kenngrößen),
- Signifikanztests,
- Klassifikationsergebnisse (Trefferquote)

A. Maßzahlen (Kenngrößen)

Diskriminanzkriterium (=**Eigenwert**) λ:

$$\lambda = \frac{B_d \quad \text{(Streuung der Diskriminanzwerte zwischen den Gruppen)}}{W_d \quad \text{(Streuung der Diskriminanzwerte innerhalb der Gruppen)}}$$

Eigenwert λ: kein normiertes Maß ($\lambda \geq 0$) → kein eigenständiges Gütemaß, jedoch Baustein von Gütemaßen

Verhältnis von erklärter Streuung (B_d) zur Gesamtstreuung (T_d) Diskriminanzvariablen D:

$$r_c^2 = \frac{B_d}{T_d} = \frac{B_d}{B_d + W_d}$$

mit $T_d = B_d + W_d$.

Kanonischer Korrelationskoeffizient:

$$r_c = \sqrt{\frac{\lambda}{\lambda + 1}}$$

Wilks Lambda:

$$\Lambda = \frac{1}{1 + \lambda}$$

Verhältnis von nicht erklärter Streuung (W_d) zur Gesamtstreuung (T_d) der Diskriminanzwerte (inverses Gütemaß):

$$\Lambda = \frac{W_d}{T_d}$$

Wertebereich: $0 \leq \Lambda \leq 1$

B. Signifikanztests

• Hotellings T^2-Test

Nullhypothese gleicher Mittelwertvektoren der Merkmalsvariablen im Zwei-Gruppen-Fall (G=2):

H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

Hotellings T^2:

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \cdot \left[\mathbf{S}^{pool} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2} \right]^{-1} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$= \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \cdot \mathbf{W}^{-1} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

mit $\mathbf{S}^{pool} = \frac{1}{n_1 + n_2 - 2} \cdot \mathbf{W}$

Prüfgröße:

$$F_0 = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2) \cdot m} \cdot T^2$$

Verteilung von F_0: F-Verteilung mit m und n_1+n_2-m-1 Freiheitsgraden

• χ^2-Test über Wilks Lambda (Bartlett-Approximation)

Nullhypothese gleicher Gruppenmittelwerte auf der Diskriminanzachse:

H_0: $\mu_1^d = \mu_2^d = ... = \mu_G^d$

μ_g^d: Mittelwert der Diskriminanzwerte der g-ten Gruppe in der Grundgesamtheit

Prüfgröße:

$$\chi_0^2 = -(n - \frac{m+G}{2} - 1) \cdot \ln\Lambda$$

Verteilung: approximativ χ^2-verteilt mit m(G-1) Freiheitsgraden

• χ^2-Test des multivariaten Wilks Lambda

bei mehr zwei Gruppen (G > 2): mehr als eine Diskriminanzfunktion

Anzahl der Diskriminanzfunktionen: q = min {G-1, m}

Zu den q Diskriminanzfunktionen gehören q Eigenwerte, die gemeinsam bei der Beurteilung der Unterschiedlichkeit der Gruppen zu berücksichtigen sind.

Multivariates Wilks Lambda:

$$\Lambda = \prod_{k=1}^{q} \Lambda_k = \prod_{k=1}^{q} \frac{1}{1 + \lambda_k}$$

univariate Lambdas: Λ_k, k=1,2,…,q

Verteilung von Λ (als χ_0^2-Prüfgröße): χ^2-Verteilung mit m(G-1) Freiheitsgraden