5.3 Güte der Diskriminanzfunktion

Eine Beurteilung der **Güte der Diskriminanzfunktion** bezieht sich auf ihre Aussagefähigkeit zur **Trennung der Gruppen**. Für eine solche Beurteilung können zum einen

- Maßzahlen (Kenngrößen) und
- Signifikanztests

herangezogen werden. Zum anderen kann sie auch durch eine Bewertung der

- Klassifikationsergebnisse,

d.h. der Trefferquote, erfolgen, die sich aus der mit Hilfe der Diskriminanzfunktion geschätzten und der tatsächlichen Gruppenzugehörigkeit ergibt.

A. Maßzahlen (Kenngrößen)

Da das **Diskriminanzkriterium** (=**Eigenwert**) λ, das als Verhältnis von erklärter zu nicht erklärter Streuung interpretiert werden kann,

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

kein normiertes Maß ($\lambda \geq 0$) ist, kommt es nicht unmittelbar zur Interpretation der Güte der Trennfähigkeit in Betracht. Der Eigenwert λ kann jedoch als Baustein eines Gütemaßes verwendet werden, dem eine ähnliche Interpretation zukommt wie dem Determinationskoeffizienten (Bestimmtheitsmaß) in der Regressionsanalyse.

Ausgangspunkt ist dabei das Verhältnis von erklärter Streuung (Bd) zur Gesamtstreuung (T_d) Diskriminanzvariablen D:

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

mit

$$T_d = B_d + W_d.$$

Erweitert wir diesen Ausdruck mit dem Faktor $1/W_d$, dann erhalten wir

$$r_c^2 = \frac{B_d/W_d}{B_d/W_d + 1} = \frac{\lambda}{\lambda + 1}.$$

Obwohl r_c^2 bereits auf das Intervall [0; 1] normiert ist, wird in der Diskriminanzanalyse üblicherweise die Wurzel von r_c^2 als Gütemaß herangezogen:

(5.29) $$r_c = \sqrt{\frac{\lambda}{\lambda + 1}},$$

die als **kanonischer Korrelationskoeffizient** bezeichnet wird. Ein hoher Wert des kanonischen Korrelationskoeffizienten r_c weist auf eine gute Trennung der durch die Diskriminanzfunktion gebildeten Gruppen hin.
Das gebräuchlichste Kriterium zur Prüfung der Trennfähigkeit der aus den Merkmalsvariablen gebildeten Diskriminanzfunktion ist **Wilks Lambda**,

(5.30) $$\Lambda = \frac{1}{1 + \lambda},$$

das das Verhältnis von nicht erklärter Streuung (W_d) zur Gesamtstreuung (T_d) der Diskriminanzwerte wiedergibt:

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

Wilks Lambda ist ein **inverses Gütemaß**. Je größer der Eigenwert λ, um so kleiner ist Λ. Ein niedriger Wert von Wilks Lambda zeigt daher eine geringe unerklärte Streuung („Reststreuung“) und damit eine gute Trennung der Gruppen an. Wilks Lambda ist ein normiertes Maß, das stets zwischen 0 und 1 liegt.

Beispiel 5.5:

Bei der Diskriminanzanalyse der beiden Käufergruppen ($n_1 = 16$, $n_2 = 14$) hatten wir einen **Eigenwert** λ von 0,3658 berechnet. Man erhält den Eigenwert λ ebenfalls aus dem Verhältnis aus B_d und W_d. Hierzu berechnen wir die Streuung zwischen den Gruppen auf der Diskriminanzachse,

$$B_d = \tilde{\mathbf{a}}'\mathbf{B}\tilde{\mathbf{a}} = \begin{bmatrix} -0{,}1653 & 0{,}1398 \end{bmatrix} \begin{bmatrix} 91{,}99 & -110{,}86 \\ -110{,}86 & 133{,}60 \end{bmatrix} \begin{bmatrix} -0{,}1653 \\ 0{,}1398 \end{bmatrix}$$

$$= \begin{bmatrix} -30{,}7042 & 37{,}0024 \end{bmatrix} \begin{bmatrix} -0{,}1653 \\ 0{,}1398 \end{bmatrix} = 10{,}2483$$

und die Streuung innerhalb der Gruppen auf der Diskriminanzachse:

$$W_d = \tilde{\mathbf{a}}'\mathbf{W}\tilde{\mathbf{a}} = \begin{bmatrix} -0{,}1653 & 0{,}1398 \end{bmatrix} \begin{bmatrix} 1014{,}81 & 600{,}06 \\ 600{,}06 & 1433{,}37 \end{bmatrix} \begin{bmatrix} -0{,}1653 \\ 0{,}1398 \end{bmatrix}$$

$$= \begin{bmatrix} -83{,}8597 & 101{,}1195 \end{bmatrix} \begin{bmatrix} -0{,}1653 \\ 0{,}1398 \end{bmatrix} = 27{,}9985$$

Mit dem Streuungsverhältnis B_d/W_d lässt sich der Eigenwerte $\lambda = 0{,}3658$ – bis auf Rundungsabweichungen – reproduzieren:

$$\lambda = \frac{B_d}{W_d} = \frac{10{,}2483}{27{,}9985} = 0{,}3660 \approx 0{,}3658$$

Setzen wir $\lambda = 0{,}3658$ in (5.29) ein, erhalten wir einen **kanonischen Korrelations-koeffizienten** in Höhe von

$$r_c = \sqrt{\frac{\lambda}{1+\lambda}} = \sqrt{\frac{0{,}3658}{1{,}3658}} = 0{,}518.$$

Die mittelstarke kanonische Korrelation ist mit der Aussage verbunden, das die Dis-kriminanzfunktion einen Anteil von $r_c^2 = 0{,}518^2 = 0{,}268$ der gesamten Streu-ung der Diskriminanzwerte erklärt.

Wilks Lambda nimmt hier den Wert

$$\Lambda = \frac{1}{1+\lambda} = \frac{1}{1{,}3658} = 0{,}7322$$

an. Der Anteil der nicht erklärten Streuung beträgt damit 73,22%. □

B. Signifikanztests

• Hotellings T^2-Test

Hotellings T^2-Test ist das multivariate Analogon des univariaten Tests der Mittelwert-differenzen

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{oder} \quad H_0: \mu_1 = \mu_2$$

zweier unabhängiger Stichproben (Zwei-Stichproben-Test) bei gleichen Varianzen $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in den beiden Grundgesamtheiten ((Varianzhomogenität).

Quadriert man die t-verteilte Prüfgröße des Tests,

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2}{S^{pool} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}} \sim t_{n_1+n_2-2}$$

univariater Zwei – Gruppen – Fall ,

mit $$S^{pool} = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}$$

erhält man eine Statistik, die F-verteilt ist mit 1 Zähler- und n_1+n_2-2 Nennerfreiheits-graden:

$$T_0^2 = \frac{(\overline{x}_1 - \overline{x}_2)^2}{\left(s^2\right)^{pool} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}} = (\overline{X}_1 - \overline{X}_2) \cdot \left[\left(S^2\right)^{pool} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}\right]^{-1} \cdot (\overline{X}_1 - \overline{X}_2) \sim F_{1;n_1+n_2-2}$$

Die Differenz der Mittelwerte $\bar{x}_1 - \bar{x}_2$ wird im multivariaten Fall entsprechend durch die Differenz der gruppenspezifischen Mittelwertvektoren $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ ersetzt. Damit erhält man für den multivariaten Zwei-Gruppen-Fall die **Hotellings T²**

(5.31)
$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \cdot \left[\mathbf{S}^{pool} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2} \right]^{-1} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$
$$= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \cdot \left[\frac{\mathbf{W}}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2} \right]^{-1} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$
$$= \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \cdot \mathbf{W}^{-1} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

mit dem multivariate Pendant zur gepoolten Varianz im univariaten Fall, die aus den beiden gruppenspezifischen (Stichproben-) Varianzen besteht:

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

Die Prüfgröße der Nullhypothese H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ des Hotellings T^2-Tests,

(5.32)
$$F_0 = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2) \cdot m} \cdot T^2, \qquad \text{multivariater Zwei-Gruppen-Fall}$$

folgt einer F-Verteilung mit m und n_1+n_2-m-1 Freiheitsgraden.

Beispiel 5.6:

In dem Beispiel 5.3 zur Diskriminanzanalyse der beiden Käufergruppen der Marken A und B hatten wir die gruppenspezifischen Mittelwerte der beiden Variablen Einkommen (X_1) und Alter (X_2) berechnet, die durch die Mittelwertvektoren $\bar{\mathbf{x}}_1$ und $\bar{\mathbf{x}}_2$ wiedergegeben werden:

Gruppe 1

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{12} \end{bmatrix} = \begin{bmatrix} 30{,}563 \\ 34{,}938 \end{bmatrix}$$

Gruppe 2

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} \bar{x}_{21} \\ \bar{x}_{22} \end{bmatrix} = \begin{bmatrix} 34{,}071 \\ 30{,}714 \end{bmatrix}$$

Aus Beispiel 5.4 ist die inverse Streuungsmatrix innerhalb der Gruppen bekannt:

$$\mathbf{W}^{-1} = \begin{bmatrix} 0{,}00130958 & -0{,}00054824 \\ -0{,}00054824 & 0{,}00092717 \end{bmatrix},$$

Die Nullhypothese der Gleichheit der Mittelwertvektoren in den Grundgesamtheiten beider Gruppen, H_0: $\boldsymbol{\mu}_1$ und $\boldsymbol{\mu}_2$, soll mit Hotellings T^2-Test geprüft werden.

Mit den Stichprobenumfängen $n_1 = 16$ und $n_2 = 14$ erhält man für Hotellings T^2 den Wert

$$T^2 = \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2} \cdot (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{W}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$= \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2} \cdot [(\bar{x}_{11} - \bar{x}_{21}) \quad (\bar{x}_{12} - \bar{x}_{22})] \cdot \begin{bmatrix} w^{11} & w^{12} \\ w^{21} & w^{22} \end{bmatrix} \cdot \begin{bmatrix} \bar{x}_{11} - \bar{x}_{21} \\ \bar{x}_{12} - \bar{x}_{22} \end{bmatrix}$$

$$T^2 = \frac{16 \cdot 14 \cdot (16+14-2)}{16+14} \cdot [(30{,}563 - 34{,}071) \quad (34{,}938 - 30{,}714)]$$

$$\cdot \begin{bmatrix} 0{,}00130958 & -0{,}00054824 \\ -0{,}00054824 & 0{,}00092717 \end{bmatrix} \cdot \begin{bmatrix} 30{,}563 - 34{,}071 \\ 34{,}938 - 30{,}714 \end{bmatrix}$$

$$= 209{,}0667 \cdot [-3{,}508 \quad 4{,}224] \cdot \begin{bmatrix} 0{,}00130958 & -0{,}00054824 \\ -0{,}00054824 & 0{,}00092717 \end{bmatrix} \cdot \begin{bmatrix} -3{,}508 \\ 4{,}224 \end{bmatrix}$$

$$= 209{,}0667 \cdot [-3{,}508 \quad 4{,}224] \cdot \begin{bmatrix} -0{,}006909772 \\ 0{,}005839592 \end{bmatrix} = 209{,}0667 \cdot 0{,}04890592 = 10{,}2246$$

Prüfgröße: $$F_0 = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2) \cdot m} \cdot T^2 = \frac{16+14-2-1}{(16+14-2) \cdot 2} \cdot 10{,}2246$$

$$= 0{,}4821429 \cdot 10{,}2246 = 4{,}9297$$

Kritischer Wert (α=0,05): $F_{m;n1+n2-m-1;1-\alpha} = F_{2;16+14-2-1;0,95} = F_{2;27;0,95} = 3{,}37$

$F_{2;20;0,95} = 3{,}49$ und $F_{2;30;0,95} = 3{,}32$: Differenz 17 Einheiten · 0,3 = 5,1 Einheiten

$F_{2;27;0,95} = F_{2;30;0,95} = 3{,}32 + 0{,}05 = 3{,}37$

Testentscheidung: $F_0 = 4{,}93 > F_{2;27;0,95} = 3{,}37 \Rightarrow H_0$ verwerfen
(auf Unterschiedlichkeit der Mittelwertvektoren der beiden Gruppen $\boldsymbol{\mu}_1$ und $\boldsymbol{\mu}_2$ geschlossen)

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

Wilks Lambda bildet die Grundlage verschiedener Signifikanztests in der Diskriminanzanalyse. Ein **Standardsignifikanztest** in der Diskriminanzanalyse bezieht sich auf die Nullhypothese **gleicher Gruppenmittelwerte** auf der **Diskriminanzachse**:

$$H_0: \quad \mu_1^d = \mu_2^d = \ldots = \mu_G^d$$

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

Nach einer Approximation von Bartlett ist die **Prüfgröße (Teststatistik)** des **Tests auf Gleichheit der Gruppenmittelwerte**,

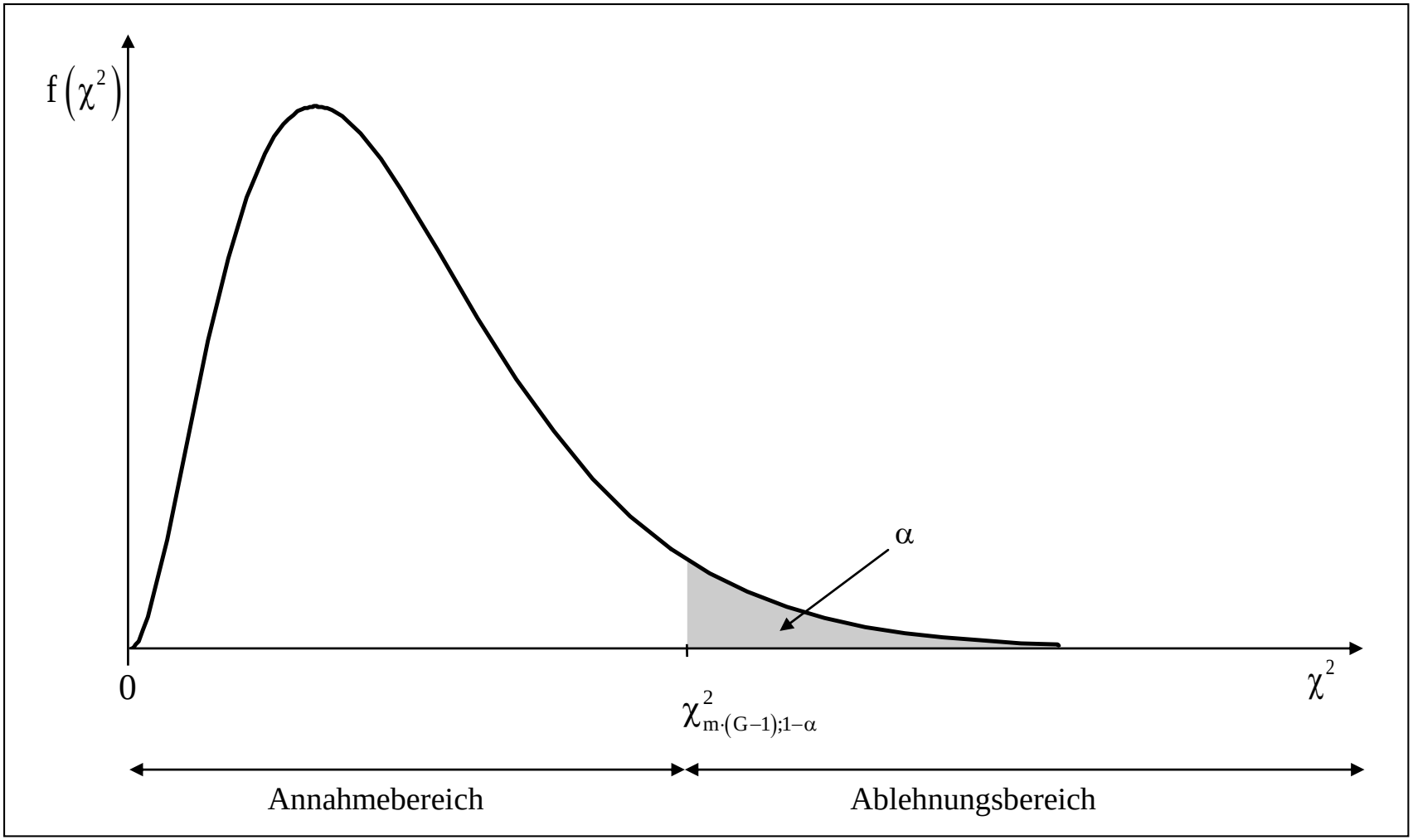
$$\text{(5.33)} \quad \chi_0^2 = -(n - \frac{m+G}{2} - 1) \cdot \ln\Lambda$$

approximativ χ^2-verteilt mit m(G-1) Freiheitsgraden. Die Approximation ist umso besser, je größer n im Vergleich zu G+m ist.

Bei einem Signifikanzniveau (Irrtumswahrscheinlichkeit) α wird die Nullhypothese H_0 verworfen, wenn der Wert der Prüfgröße χ^2 das (1-α)-Quantil der χ^2-Verteilung mit m(G-1) Freiheitsgraden überschreitet:

$$\chi_0^2 > \chi^2_{m(G-1);1-\alpha} \Rightarrow H_0 \text{ ablehnen}$$

Abbildung 5.5: Annahme- und Ablehnungsbereich des χ^2-Tests



Bei einer Ablehnung der Nullhypothese H_0 sind die Gruppenunterschiede auf der Diskriminanzachse statistisch gesichert.

Sofern mehr als zwei Gruppen vorhanden sind (G > 2), lassen sich mehr als eine Diskriminanzfunktion bestimmen. Allgemein gibt es

$$q = \min \{G-1, m\}$$

Diskriminanzfunktionen. Zu den q Diskriminanzfunktionen gehören q Eigenwerte, die gemeinsam bei der Beurteilung der Unterschiedlichkeit der Gruppen berücksichtigt werden müssen. An die Stelle des univariaten Wilks Lambda tritt dann das **multivariate Wilks Lambda**, das man aus dem Produkt der univariaten Lambdas Λ_k erhält:

$$(5.34) \quad \Lambda = \prod_{k=1}^{q} \Lambda_k = \prod_{k=1}^{q} \frac{1}{1+\lambda_k}.$$

Das multivariate Lambda von Wilk folgt (als χ^2-Prüfgröße) einer χ^2-Verteilung mit m(G-1) Freiheitsgraden.

Beispiel 5.6:

Mit dem in Beispiel 5.4 berechneten Wert von Wilks Lambda 0,7322 können wir die Prüfgröße des **Signifikanztests** [Gl. (5.33)] auf **Gleichheit der Mittelwerte** der beiden Gruppen auf der **Diskriminanzachse**,

$$H_0: \quad \mu_1^d = \mu_2^d \ ,$$

bestimmen (n=30, m=2, G=2):

$$\chi_0^2 = -(n - \frac{m+G}{2} - 1) \cdot \ln\Lambda = -(30 - \frac{2+2}{2} - 1) \cdot \ln(0{,}7322)$$
$$= (-27) \cdot (-0{,}3117) = 8{,}416$$

Bei einem Signifikanzniveau α von 5% ermitteln wir einen kritischen Wert von

$$\chi^2_{m(G-1);0,95} = \chi^2_{2;0,95} = 5{,}99$$

Damit lautet die Testentscheidung:

$$(\chi_0^2 = 8{,}416) > \chi^2_{2;0,95} = 5{,}99) \Rightarrow H_0 \text{ ablehnen.}$$

Die Diskrimination der beiden Käufergruppen durch die Merkmalsvariablen Einkommen und Alter ist damit statistisch gesichert. □

5.4 Klassifikation

5.4.1 Klassifikationskonzepte

Ziel der Klassifikation:
Zuordnung eines neuen zusätzlichen Untersuchungsobjektes zu einer Gruppe aufgrund der metrisch skalierten Variablen, d.h. die Prognose der unbekannten Ausprägung der nominalskalierten Gruppenvariablen

Fragestellung:
- In welche Gruppe ist ein neues Element, dessen Gruppenzugehörigkeit nicht bekannt ist, aufgrund der Merkmalsausprägungen seiner metrisch skalierten Variablen einzuordnen?

Klassifikationskonzepte:
- Distanzkonzept,
- Wahrscheinlichkeitskonzept,
- Konzept der Klassifikationsfunktionen,
- Konzept des kritischen Diskriminanzwertes.

Methodisch betrachten wir die beiden ersten Konzepte, bei denen entweder
- die metrisch skalierten Ausgangsvariablen

oder
- die zuvor berechneten Diskriminanzvariablen

verwendet werden können

Ermittlung der **Zuordnungs- bzw. Klassifikationsregeln** anhand eines **Lerndatensatzes**

Lerndatensatz: Untersuchungseinheiten, bei denen nicht nur die Ausprägungen der metrischen Variablen bekannt sind, sondern auch die der nominalen Gruppenvariablen.

Interne Analyse: Untersuchungseinheiten des Lerndatensatzes werden klassifiziert (Trefferquote ermittelt → Hinweis auf Güte der Diskriminanzfunktion)

Externe Analyse: Klassifikation neuer Objekte bei denen lediglich die Ausprägungen der metrisch skalierten Variablen bekannt sind

Tabelle 5.1: Vergleich der Klassifikationskonzepte

		Klassifikationskonzepte			
		Distanz-Konzept	Wahr-scheinlich-keits-Konzept	Klassifizie-rungs-Funktionen	Kritischer Diskrimi-nanzwert
Mögliche Variab-len-grund-lage	Ausgangsvariablen	ja	ja	ja	nein
	Diskriminanzvariablen	ja	ja	nein	ja
zusätz-liche Aspekte / Erweite-rungen	Berücksichtigung unter-schiedlicher A-priori-Wahrscheinlichkeiten	nein	ja	ja	ja
	Berücksichtigung unter-schiedlicher Kosten der Fehlklassifikation	nein	ja	nein	ja
	Berücksichtigung un-gleicher Streuungen in den Gruppen	ja	ja	nein	nein
	Unterdrückung irrelevanter Diskriminanzfunktionen	ja	ja	nein	ja

5.4.2 Distanzkonzept

Messung der Distanz bzw. Ähnlichkeit zweier Objekte A und B:
Vergleich der beiden individuellen Merkmalsprofile

Distanzmaße (aus Clusteranalyse bekannt):
Quantifizierung der Ähnlichkeit oder Distanz von Objekten aufgrund ihrer Ausprägungen bei m Merkmalsvariablen

Fragestellung bei **Klassifikationsverfahren der Diskriminanzanalyse**:

Wie gut passt ein Objekt zu den verschiedenen Gruppen, d.h. in welchem Ausmaß stimmt das individuelle Merkmalsprofil eines Objektes x_i mit dem durchschnittlichen Merkmalsprofil der zur Gruppe g gehörenden Untersuchungsobjekte $\bar{\mathbf{x}}_g$ (Gruppenzentroid) überein?

Wenn bei der Berechnung der Distanzen zwischen Objekt und Gruppenzentroid die Maßeinheiten und Standardabweichungen der Variablen sowie die **Korrelationen** zwischen den Variablen berücksichtigt werden sollen, kann **nicht** die **euklidische Distanz** verwendet werden.

Ein Maß, das diese Forderungen durch Einbeziehung der Inversen der gruppenspezifischen Varianz-Kovarianz-Matrizen $\mathbf{S}_g^{-1}$ erfüllt, ist die **Mahalanobis-Distanz**. Unter der Annahme gleicher gruppenspezifischen Varianz-Kovarianz-Matrizen in der Grundgesamtheit kann hierzu vereinfachend die **Inverse der gepoolten Varianz-Kovarianz-Matrix** $\left(\mathbf{S}^{pool}\right)^{-1}$ verwendet werden.

● **Mahalanobis-Distanz auf Basis der m Merkmalsvariablen**

Für ein Untersuchungsobjekt i wird nun auf Basis der Ausgangsvariablen die **Mahalanobis-Distanz** zur Gruppe g wie folgt berechnet:

(5.35) $$MD(i,g) = \underbrace{(x_i - \bar{x}_g)'}_{1\times m} \cdot \underset{m\times m}{\left(S^{pool}\right)^{-1}} \cdot \underbrace{(x_i - \bar{x}_g)}_{m\times 1}$$

Klassifikationsregel:
Zuordnung des Objekts i zu der Gruppe, zu der es die geringste Distanz aufweist.

● **Mahalanobis-Distanz auf Basis der p Diskriminanzvariablen**
- bei p Diskriminanzfunktionen:

(5.36) $$MD_d(i,g) = \underbrace{(d_i - \bar{d}_g)'}_{1\times p} \cdot \underset{p\times p}{\left(S_d^{pool}\right)^{-1}} \cdot \underbrace{(d_i - \bar{d}_g)}_{p\times 1}$$

- im Zwei-Gruppen-Fall:

(5.37) $$MD_d(i,g) = \frac{(d_i - \bar{d}_g)^2}{\left(s_d^2\right)^{pool}}$$

Aufgrund der Normierung der Diskriminanzkoeffizienten $(s_d^2)^{pool} = 1$ gilt:

(5.38) $$MD_d(i,g) = QED_d(i,g) = (d_i - \bar{d}_g)^2$$

Die Mahalanobis-Distanzen auf Basis der Diskriminanzvariablen entsprechen also den quadrierten euklidischen Distanzen $QED_d(i,g)$.

Beispiel 5.7: Für die Region A ergab sich beispielsweise ein Diskriminanzwert von $d_A = -1{,}836$. Aus allen Regionen, die zur Gruppe 1 = ländlicher Raum (Gruppe 2 = städtischer Raum) gehören, kann nun ein gruppenspezifischer Diskriminanzmittelwert berechnet werden:

$$\bar{d}_1 = \frac{1}{3} \cdot (d_C + d_H + d_K) = \frac{1}{3} \cdot (6{,}320 + 4{,}046 + 4{,}399) = 4{,}922$$

und

$$\begin{aligned}\bar{d}_2 &= \frac{1}{9} \cdot (d_A + d_B + d_D + d_E + d_F + d_G + d_I + d_J + d_L) \\ &= \frac{1}{9} \cdot (-1{,}836 - 1{,}179 - 1{,}325 - 1{,}717 - 2{,}125 - 1{,}532 - 3{,}558 + 0{,}004 - 1{,}496) \\ &= -1{,}641.\end{aligned}$$

Damit können die quadrierten Euklidischen Distanzen für die Region A zur Gruppe 1 (ländlicher Raum) und zur Gruppe 2 (städtischer Raum) berechnet werden.

$$QED_d(A,1) = (-1{,}836 - 4{,}922)^2 = 45{,}6646338,$$

$$QED_d(A,2) = (-1{,}836 - (-1{,}641))^2 = 0{,}03812753.$$

Wir untersuchen eine weitere Region namens „neu“ und wollen wissen, ob sie eher der Gruppe 1 oder der Gruppe 2 zugeordnet werden kann. Bekannt sind die Ausprägungen der metrisch skalierten Merkmalsvariablen:

Tabelle: Prognosedaten des Rechenbeispiels

Region	**Merkmalsvariablen**						**Gruppierungs-variable** g
	X_1	X_2	X_3	X_4	X_5	X_6	
neu	200,0	18020	6,8	53,0	11,0	2,5	unbekannt

Für die neue Region wird zunächst der zugehörige Diskriminanzwert ermittelt:

$$d_{neu} = a_0 + x_{neu}' \cdot \tilde{a}$$

$$= -15{,}277 + \begin{bmatrix} 200{,}0 & 18020 & 6{,}8 & 53{,}0 & 11{,}0 & 2{,}5 \end{bmatrix} \cdot \begin{bmatrix} 0{,}011 \\ -0{,}000 \\ 0{,}583 \\ 0{,}130 \\ 0{,}448 \\ -0{,}092 \end{bmatrix} = -2{,}300 .$$

Damit können die quadrierten Euklidischen Distanzen für die neue Region zur Gruppe 1 (ländlicher Raum) und zur Gruppe 2 (städtischer Raum) berechnet werden:

$$QED_d(neu,1) = (-2{,}300 - 4{,}922)^2 = 52{,}159$$

$$QED_d(neu,2) = (-2{,}300 - (-1{,}641))^2 = 0{,}435$$

Die Distanzen, die sich bei interner und externer Analyse auf Basis der Diskriminanzvariable ergeben, sind in der nachfolgenden Tabelle dargestellt.

Demnach würden also die Regionen C,H und K der Gruppe 1 (ländlicher Raum) und die Regionen A,B,D,E,F,G,I,J und L der Gruppe 2 (städtischer Raum) zugeordnet. Die neue Region wird der Gruppe 2 zugeordnet.

Tabelle: Berechnung der quadrierten Mahalanobis-Distanzen auf Basis der Diskriminanzvariablen

Name	Quadrierte Euklidische Distanzen bzw. quadrierte Mahalanobis-Distanzen $QED_d(i, g)$		Vorhergesagte Gruppenzugehörigkeit	Tatsächliche Gruppenzugehörigkeit
	Gruppe 1	Gruppe 2		
A	45,665	0,038*	2	2
B	37,222	0,213*	2	2
C	1,956*	63,374	1	1
D	39,024	0,099*	2	2
E	44,077	0,006*	2	2
F	49,656	0,235*	2	2
G	41,652	0,012*	2	2
H	0,768*	32,333	1	1
I	71,905	3,676*	2	2
J	24,186	2,704*	2	2
K	0,273*	36,480	1	1
L	41,190	0,021*	2	2
Neu	52,159	0,435*	2	Unbekannt

* minimaler Distanzwert
Trefferquote: 100%

5.4.3 Wahrscheinlichkeitskonzept

Das **Wahrscheinlichkeitskonzept** behandelt die Klassifikation als ein statistisches Entscheidungsproblem

Zuordnung der einzelnen Objekte zu einer der Gruppen erfolgt auf der Grundlage des **Bayesschen Theorems:**

<u>gesucht</u>:
- Bedingte Wahrscheinlichkeit $P(g|\mathbf{x}_i)$, dass eine Untersuchungseinheit i mit dem Merkmalsvektor $\mathbf{x}_i$ der Gruppe g angehört (=**A-posteriori-Wahrscheinlichkeit**)

<u>gegeben</u>:
- **A-priori-Wahrscheinlichkeit** $P_i(g)$ der Gruppenzugehörigkeit der Untersuchungseinheit i,
- **Bedingte Wahrscheinlichkeit** $P(\mathbf{x}_i|g)$ mit der man bei gegebener Gruppenzugehörigkeit den Merkmalsvektor $\mathbf{x}_i$ erhält

Zuordnungswahrscheinlichkeit nach dem **Bayes-Theorem**:

(5.39) $$P(g \mid x_i) = \frac{P(x_i \mid g) \cdot P_i(g)}{\sum_{h=1}^{G} P(x_i \mid h) \cdot P_i(h)} \qquad (g = 1, \ldots, G)$$

In der Regel werden statt der individuellen nur **gruppenspezifische A-priori-Wahrscheinlichkeiten** berücksichtigt. In diesem Falle gilt:

(5.40) $$P_i(g) = P(g) \qquad (i = 1, \ldots, n) \quad (g = 1, \ldots, G)$$

● **Bestimmung der gruppenspezifischen A-priori-Wahrscheinlichkeiten P(g)**:

Zufallsstichprobe oder repräsentative Beurteilungsstichprobe	Keine Informationen über Gruppenanteile in der Grundgesamtheit
(5.41) $P(g) = \frac{n_g}{n} \qquad (g = 1, \ldots, G)$	(5.42) $P(g) = \frac{1}{G} \qquad (g = 1, \ldots, G)$

Ggf. können auch externe Quellen (z.B. amtliche Statistik, Marktforschung) zur Schätzung der A-priori-Wahrscheinlichkeiten herangezogen werden.

● **Berechnung der bedingten Wahrscheinlichkeiten P($\mathbf{x}_i$|g)**:

Die Berechnung der bedingten Wahrscheinlichkeiten P($\mathbf{x}_i$|g) erfolgt unter den Annahmen, dass die **metrisch skalierten Ausgangsvariablen** in jeder Gruppe **normalverteilt** sind und **gleiche gruppenspezifische Varianz-Kovarianz-Matrizen** vorliegen.

Für ein Objekt i mit dem Merkmalsvektor $\mathbf{x}_i$ erhält man dann unter der Voraussetzung, dass Objekt i zur Gruppe g gehört, die (bedingte) Dichtefunktion der multivariaten Normalverteilung:

(5.43) $$f(x_i \mid g) = (2\pi)^{-m/2} \cdot \left|S^{pool}\right|^{-1/2} \cdot \exp\left(\frac{(x_i - \bar{x}_g)' \cdot (S^{pool})^{-1} \cdot (x_i - \bar{x}_g)}{2}\right)$$

Unter der Normalverteilungsannahme müssen die bedingten Wahrscheinlichkeiten P($\mathbf{x}_i$|g) bei der Berechnung A-posteriori-Wahrscheinlichkeiten P(g|$\mathbf{x}_i$) nach Formel (5.39) durch die bedingten Dichten f(g|$\mathbf{x}_i$) ersetzt werden.

Unter Verwendung der **individuellen A-priori-Wahrscheinlichkeiten** $P_i(g)$ erhält man die **Zuordnungswahrscheinlichkeiten** aus

$$(5.44) \quad P(g \mid x_i) = \frac{\exp(-MD(i,g)/2) \cdot P_i(g)}{\sum_{h=1}^{G} \exp(-MD(i,h)/2) \cdot P_i(h)} \qquad (g = 1, \ldots, G)$$

wenn man beachtet, dass die konstanten Faktoren $(2\pi)^{-m/2}$ und $|\mathbf{S}\text{pool}|^{-1/2}$ im Zähler und Nenner stehen und sich daher herauskürzen.

Bei **gleichen A-priori-Wahrscheinlichkeiten** lassen sich die **Zuordnungswahrscheinlichkeiten** vereinfacht aus

$$(5.45) \quad P(g \mid x_i) = \frac{\exp(-MD(i,g)/2)}{\sum_{h=1}^{G} \exp(-MD(i,h)/2)} \qquad (g = 1, \ldots G)$$

berechnen.

Verhältnis zwischen Wahrscheinlichkeits- und Distanzkonzept:

Die Gleichungen (5.44) und (5.45) zeigen die **enge Kohärenz** zwischen dem **Wahrscheinlichkeits- und Distanzkonzept** auf. Bei gleichen A-priori-Wahrscheinlichkeiten erfolgt die Zuordnung der Objekte auch beim Wahrscheinlichkeitskonzept letztlich aufgrund ihrer Distanzen zu den Gruppen.

- **Bestimmung der A-posteriori-Wahrscheinlichkeiten $P(g|x_i)$ auf der Basis der Diskriminanzvariablen**

Im Zwei-Gruppen-Fall ist die Wahrscheinlichkeit $P(g|d_i)$, dass ein Objekt i mit dem Diskriminanzwert d_i der Gruppe g entstammt, durch

$$(5.46)\quad P(g \mid d_i) = \frac{P(d_i \mid g) \cdot P_l(g)}{\sum_{h=1}^{2} P(d_i \mid h) \cdot P_l(h)} \qquad (g = 1, 2)$$

gegeben.

Es wird nun angenommen, dass die Diskriminanzvariable in jeder Gruppe normalverteilt ist und dass gleiche gruppenspezifische Varianzen vorliegen. Damit lässt sich die bedingte Wahrscheinlichkeit $P(d_i|g)$ über eine Transformation der auf Basis der Diskriminanzvariablen berechneten quadrierten Mahalanobis-Distanzen $MD_d(i,g)$ berechnen. Die **Zuordnungswahrscheinlichkeit** lautet damit:

$$(5.47)\quad P(g \mid d_i) = \frac{\exp(-MD_d(i,g)/2) \cdot P_l(g)}{\sum_{h=1}^{2} \exp(-MD_d(i,h)/2) \cdot P_l(h)} \qquad (g = 1, 2)$$

bzw. bei gleicher A-priori-Wahrscheinlichkeit

$$(5.48)\quad P(g \mid d_i) = \frac{\exp(-MD_d(i,g)/2)}{\sum_{h=1}^{2} \exp(-MD_d(i,h)/2)} \qquad (g = 1,2).$$

Ein Objekt i wird stets derjenigen Gruppe g zugeordnet, für die die Zuordnungswahrscheinlichkeit $P(g|d_i)$ am größten ist

Beispiel 5.8:

Unter der Annahme gleicher A-priori-Wahrscheinlichkeiten erhält man über (5.46) beispielsweise für die Region A die Zuordnungswahrscheinlichkeiten:

$$P(1 \mid d_A) = \frac{e^{(-45{,}665/2)}}{e^{(-45{,}665/2)} + e^{(-0{,}038/2)}} = 0{,}000000 \quad \text{und}$$

$$P(2 \mid d_A) = \frac{e^{(-0{,}038/2)}}{e^{(-45{,}665/2)} + e^{(-0{,}038/2)}} = 1{,}000000.$$

Die folgende Tabelle zeigt die Ergebnisse für die 12 Untersuchungsregionen mit bekannter Gruppenzugehörigkeit und die neue Region.

Tabelle: Berechnung der Zuordnungswahrscheinlichkeiten auf Basis der Diskriminanzvariablen

Name	$P(g \mid d_i) = \frac{\exp(-MD_d(i,g)/2)}{\sum_{h=1}^{2} \exp(-MD_d(i,h)/2)}$		Vorhergesagte Gruppenzugehörigkeit	Tatsächliche Gruppenzugehörigkeit
	Gruppe 1	Gruppe 2		
A	0,000000	1,000000*	2	2
B	0,000000	1,000000*	2	2
C	1,000000*	0,000000	1	1
D	0,000000	1,000000*	2	2
E	0,000000	1,000000*	2	2
F	0,000000	1,000000*	2	2
G	0,000000	1,000000*	2	2
H	1,000000*	0,000000	1	1
I	0,000000	1,000000*	2	2
J	0,000022	0,999978*	2	2
K	1,000000*	0,000000	1	1
L	0,000000	1,000000*	2	2
neu	0,000000	1,000000*	2	unbekannt

Trefferquote: 100% * maximale Zuordnungswahrscheinlichkeit ♦

5.4.4 Kritischer Diskriminanzwert

Beim Distanzkonzept wird ein Objekt i – bei Verwendung der Diskriminanzvariablen – in diejenige Gruppe g eingeordnet, bezüglich derer die Distanz zwischen Objekt und Gruppenmittel (Zentroid) minimal wird. Dies ist äquivalent damit, ob das Objekt unterhalb oder oberhalb eines kritischen Diskriminanzwertes d_c liegt.

Im **Zwei-Gruppen-Fall** kann unter der Voraussetzung einer **gleichen Anzahl von Objekten in den Gruppen** das **ungewogene arithmetische Mittel** der gruppenspezifischen Diskriminanzmittelwerte als **kritischer Diskriminanzwert** verwendet werden:

(5.49) $$d_c = \frac{\bar{d}_1 + \bar{d}_2}{2}.$$

Enthalten die Gruppen unterschiedlich viele Objekte, ist das **gewogene arithmetische Mittel** zu bilden:

(5.50) $$d_c = \frac{n_1 \cdot \bar{d}_1 + n_2 \cdot \bar{d}_2}{n_1 + n_2}.$$

Wir haben die Konstante a0 der Diskriminanzfunktion derart bestimmt, dass der Gesamtmittelwert der Diskriminanzwerte $\bar{d}$ gleich null wird. Der kritische Diskriminanzwert (5.48) entspricht genau diesem Gesamtmittelwert:

(5.51) $$d_c = 0.$$

Tabelle 5.2: Zuordnungsregeln bei Verwendung des kritischen Diskriminanzwertes

Fall	Vergleich des Diskriminanzwertes der i-ten Untersuchungseinheit mit dem kritischen Diskriminanzwert	Vorhergesagte Gruppenzugehörigkeit, wenn $\bar{d}_1 < \bar{d}_2$ gilt	Vorhergesagte Gruppenzugehörigkeit, wen $\bar{d}_2 < \bar{d}_1$ gilt
I	$d_i < d_c$	1	2
II	$d_i > d_c$	2	1
III	$d_i = d_c$	keine eindeutige Zuordnung	keine eindeutige Zuordnung

Beispiel 5.9:

Die gruppenspezifischen Diskriminanzmittelwerte $\bar{d}_1 = 4{,}922$ und $\bar{d}_2 = -1{,}641$ hatten wir bereits ermittelt. Damit beträgt der kritische Diskriminanzwert:

$$d_c = \frac{n_1 \cdot \bar{d}_1 + n_2 \cdot \bar{d}_2}{n_1 + n_2} = \frac{3 \cdot 4{,}922 + 9 \cdot (-1{,}641)}{12} = 0.$$

Da hier $\bar{d}_1 > \bar{d}_2$ gilt, werden alle Objekte, deren Diskriminanzwerte größer (kleiner) als der kritische Diskriminanzwert in Höhe von null sind, der Gruppe 1 (Gruppe 2) zugeordnet. Die untenstehende Tabelle zeigt die entsprechenden Klassifikationsergebnisse bei einer internen und externen Analyse.

Tabelle: Klassifikation über den kritischen Diskriminanzwert

Name	**Diskriminanzwerte** d_i	**Vorhergesagte Gruppenzugehörigkeit**	**Tatsächliche Gruppenzugehörigkeit**
A	-1,836	2	2
B	-1,179	2	2
C	6,320	1	1
D	-1,325	2	2
E	-1,717	2	2
F	-2,125	2	2
G	-1,532	2	2
H	4,046	1	1
I	-3,558	2	2
J	0,004	1**	2
K	4,399	1	1
L	-1,496	2	2
neu	-2,300	2	unbekannt

** falsche Vorhersage

Diskriminanzwert der Region A: $d_A = -1{,}836$ < 0 → Zuordnung von A zur Gruppe 2
Diskriminanzwert der „neuen“ Region: < 0 → Zuordnung der „neuen“ Region zur Gruppe 2

Auswertung der internen Analyse:
11 der 12 Untersuchungsregionen, d.h. 11/12 = 91,7% der Fälle sind richtig zugeordnet worden → hoher Prozentsatz der richtig klassifizierten Objekte deutet auf eine relativ hohe Güte der Diskriminanzfunktion hin, d.h. Gruppen lassen sich gut durch die metrisch skalierten Ausgangsvariablen erklären. ♦

5.4.5 Beurteilung der Klassifikation

Ist die **Gruppenzugehörigkeit** der klassifizierten Objekte **bekannt**, kann unabhängig vom verwendeten Klassifikationsverfahren durch **Vergleich** der **vorhergesagten mit der tatsächlichen Gruppenzugehörigkeit** eine **Beurteilung der Klassifikation** erfolgen. Der **Prozentsatz richtig klassifizierter Objekte** ist eine **Maßzahl für die Klassifikationsgüte**.

● **Resubstitutions-Methode**
Gesamter Lerndatensatz wird klassifiziert und anschließend die Trefferquote bestimmt
Diese Trefferquote ist in der Regel vermittelt jedoch ein zu optimistisches Bild über die Güte der Klassifikation, da sie auf Basis derselben Daten ermittelt wird, die auch in die Berechnung der Klassifikationsvorschriften eingehen.

●**Kreuzvalidierung**
Um das Problem der Überschätzung der Trefferquote zu vermeiden, sollte eine Kreuzvalidierung vorgenommen werden.
Ermittlung der Trefferquote bei der **Kreuzvalidierung**:

a) Klassifikation der Objekte einer **zusätzlichen Stichprobe** als eine **Kontrollstichprobe**, bei der die Gruppenzugehörigkeiten bekannt sind, die aber nicht der Konstruktion der Klassifikationsregel zugrunde liegen

b) Aufteilung der **vorhandenen Stichprobe** in eine **Lernstichprobe** und eine **Kontrollstichprobe**

Spezialfall von b): **Leave-one-out-Methode**

Hierbei wird ein Klassifikationsverfahren bei gegebener Größe n des ursprünglichen Datensatzes jeweils auf n-1 Untersuchungseinheiten angewendet und für die n-te Untersuchungseinheit getestet. Die Lernstichprobe besteht folglich aus n-1 Objekten, wobei das nicht berücksichtigte Objekt klassifiziert wird. Die Klassifikation eines Objektes basiert also auf den übrigen n-1 Objekten und wird n-mal, d.h. für jedes Objekt, durchgeführt. Dabei wird jedes Mal eine neue Klassifikationsregel bestimmt.

Übertragbarkeit der ermittelten Trefferquote auf die Grundgesamtheit:

Bei Verwendung einer 10-fachen Kreuzvalidierung kann die Trefferquote hinreichend genau geschätzt werden. Die **Leave-one-out-Method**e liefert eine fast unverzerrte Schätzung der Trefferquote. Generell ist die Anzahl der zufällig richtig klassifizierten Objekte bei der Interpretation der Trefferquoten zu berücksichtigen.

Beispiel 5.10:

Das Verfahren der Kreuzvalidierung eignet sich für jedes Klassifikationskonzept und ist unabhängig davon, ob die Klassifikation auf Basis der Ausgangsvariablen oder der Diskriminanzvariablen durchgeführt wurde. Der Einfachheit halber wollen wir das **Verfahren der Kreuzvalidierung** nur anhand der Klassifikation auf Basis der **metrisch skalierten Ausgangsvariablen** demonstrieren. Es soll im Folgenden die **Leave-one-out-Methode** zugrunde gelegt werden.

Wir wollen die bei der **Leave-one-out-Methode** notwendigen Modifikationen beispielhaft für die **Region A** darstellen. Das Merkmalsprofil dieser Region wird weiterhin durch $x_A = [212{,}4 \quad 20116 \quad 9{,}8 \quad 53 \quad 8{,}4 \quad -0{,}7]'$ beschrieben. Da wir die Region A klassifizieren wollen, enthält die Lernstichprobe nur die restlichen Regionen. Unter Ausschluss der Region A ergibt sich dabei die *veränderte* [(n–1)×m]-Datenmatrix **X** zu:

$$\underset{11\times 6}{\mathbf{X}} = \begin{array}{c} \begin{bmatrix} 623{,}7 & 24966 & 3{,}4 & 73{,}1 & 6{,}1 & 3{,}4 \\ 93{,}1 & 19324 & 23{,}6 & 47{,}9 & 12{,}3 & -1{,}9 \\ 236{,}8 & 23113 & 8{,}7 & 66{,}8 & 8{,}7 & 2{,}0 \\ 412{,}0 & 23076 & 8{,}9 & 46{,}9 & 8{,}0 & -3{,}1 \\ 566{,}7 & 24516 & 6{,}1 & 44{,}3 & 8{,}6 & -3{,}0 \\ 331{,}9 & 22187 & 7{,}4 & 57{,}6 & 10{,}3 & 4{,}7 \\ 111{,}4 & 20614 & 16{,}3 & 63{,}8 & 13{,}9 & 5{,}2 \\ 489{,}0 & 25006 & 5{,}7 & 49{,}4 & 6{,}7 & -2{,}6 \\ 287{,}4 & 23136 & 8{,}8 & 59{,}4 & 12{,}4 & 1{,}7 \\ 166{,}2 & 20707 & 14{,}1 & 74{,}0 & 13{,}0 & 3{,}6 \\ 388{,}1 & 23624 & 9{,}6 & 54{,}3 & 6{,}9 & -0{,}4 \end{bmatrix} \end{array} \begin{array}{c} B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \\ J \\ K \\ L \end{array}$$

Unter Ausschluss der Region A verändern sich auch die gruppenspezifischen Mittelwerte der metrisch skalierten Variablen. Man erhält die folgenden Spaltenvektoren:

$$\bar{\mathbf{x}}_1 = [123{,}567 \quad 20215{,}000 \quad 18{,}000 \quad 61{,}900 \quad 13{,}067 \quad 2{,}300]' \quad \text{und}$$

$$\bar{\mathbf{x}}_2 = [416{,}95 \quad 23703 \quad 7{,}325 \quad 56{,}475 \quad 8{,}4625 \quad 0{,}3375]'.$$

Die einzelnen Elemente des Spaltenvektors des gruppenspezifischen Mittelwerte berechnen sich über:

$$\bar{x}_{1j} = \frac{1}{3} \cdot (x_{Cj} + x_{Hj} + x_{Kj}) \quad \text{bzw.}$$

$$\bar{x}_{2j} = \frac{1}{8} \cdot (x_{Bj} + x_{Dj} + x_{Ej} + x_{Fj} + x_{Gj} + x_{Ij} + x_{Jj} + x_{Lj}).$$

Beispielsweise gilt:

$$\bar{x}_{1ED} = \frac{1}{3} \cdot (93{,}1 + 111{,}4 + 166{,}2) = 123{,}567 \quad \text{bzw.}$$

$$\bar{x}_{2ED} = \frac{1}{8} \cdot (623{,}7 + 236{,}8 + 412 + 566{,}7 + 331{,}9 + 489 + 287{,}4 + 388{,}1) = 416{,}95$$

Damit hat die [(n–1)×m]-Matrix der Gruppenmittelwerte bei Kreuzvalidierung jetzt folgendes Aussehen:

$$\underset{11\times 6}{\overline{\mathbf{X}}}^{grup} = \begin{bmatrix} 123{,}567 & 20215{,}000 & 18{,}000 & 61{,}900 & 13{,}067 & 2{,}300 \\ 123{,}567 & 20215{,}000 & 18{,}000 & 61{,}900 & 13{,}067 & 2{,}300 \\ 123{,}567 & 20215{,}000 & 18{,}000 & 61{,}900 & 13{,}067 & 2{,}300 \\ 416{,}950 & 23703{,}000 & 7{,}325 & 56{,}475 & 8{,}463 & 0{,}338 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 416{,}950 & 23703{,}000 & 7{,}325 & 56{,}475 & 8{,}463 & 0{,}338 \end{bmatrix}$$

Mit diesen Angaben lässt sich die modifizierte [m×m]-Matrix für die Streuung der m Merkmalsvariablen in den Gruppen ermitteln, die ebenfalls unter Ausschluss der Untersuchungsregion A berechnet wird:

$$\underset{6\times 6}{\mathbf{W}} = \begin{bmatrix} 130583{,}227 & 834082{,}000 & -1946{,}820 & -200{,}980 & -1246{,}268 & -738{,}915 \\ 834082{,}000 & 8516308{,}000 & -18585{,}800 & 17785{,}300 & -8930{,}700 & -3287{,}500 \\ -1946{,}820 & -18585{,}800 & 80{,}735 & -169{,}095 & 8{,}508 & -39{,}148 \\ -200{,}980 & 17785{,}300 & -169{,}095 & 1033{,}535 & 6{,}893 & 252{,}948 \\ -1246{,}268 & -8930{,}700 & 8{,}508 & 6{,}893 & 31{,}585 & 19{,}551 \\ -738{,}915 & -3287{,}500 & -39{,}148 & 252{,}948 & 19{,}551 & 92{,}899 \end{bmatrix}$$

Für die Mahalanobis-Distanzen der Region A gilt dann bei Kreuzvalidierung:

$$MD(A,1) = (\mathrm{x}_A - \bar{\mathrm{x}}_1)' \cdot (\mathrm{S}^{pool})^{-1} \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_1) = (n-1-G) \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_1)' \cdot \mathrm{W}^{-1} \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_1) = 92{,}483 \quad \text{und}$$

$$MD(A,2) = (\mathrm{x}_A - \bar{\mathrm{x}}_2)' \cdot (\mathrm{S}^{pool})^{-1} \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_2) = (n-1-G) \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_2)' \cdot \mathrm{W}^{-1} \cdot (\mathrm{x}_A - \bar{\mathrm{x}}_2) = 49{,}551$$

Diese Prozedur wird nun für jede einzelne Region durchgeführt. Für die Mahalanobis-Distanzen der Region B muss dann wieder eine *neue* [11×6]-Ausgangsmatrix **X** ermittelt werden, die alle Regionen außer B enthält. Insgesamt erhält man für die 12 Untersuchungsregionen bei Kreuzvalidierung die in der unten stehenden Tabelle dargestellten **Mahalanobis-Distanzen**.

Tabelle: Berechnung der Mahalanobis-Distanzen auf Basis der Merkmalsvariablen bei Kreuzvalidierung

Name	Mahalanobis- Distanzen MD(i,g) bei Kreuzvalidierung		Vorhergesagte Gruppenzugehö-rigkeit	Tatsächliche Gruppenzugehö-rigkeit
	Gruppe 1	Gruppe 2		
A	92,483	49,551*	2	2
B	59,634*	73,914	1**	2
C	59,177*	215,199	1	1
D	41,646	13,289*	2	2
E	41,252	1,771*	2	2
F	61,724	11,140*	2	2
G	50,431	19,298*	2	2
H	12,831*	32,979	1	1
I	144,232	14,489*	2	2
J	32,542*	39,201	1**	2
K	21,517*	36,430	1	1
L	40,198	4,844*	2	2

Trefferquote: 10/12 = 83,3%

* minimale Distanz

** falsch klassifizierter Fall

Aus den gerade **berechneten Distanzen** lassen sich relativ einfach die zugehörigen **Zuordnungswahrscheinlichkeiten** ermitteln, die aber zu **keinen anderen Zuordnungen** führen, so dass wir auf ihre Darstellung verzichten.

Wir hatten oben mit der **Resubstitutions-Methode** die Trefferquoten nur auf Basis der Diskriminanzvariablen berechnet. Eine entsprechende Berechnung auf Basis der Merkmalsvariablen hätte allerdings ebenfalls zu einer **Trefferquote von 100 %** geführt. Wie die vorherige Tabelle zeigt, verändern sich die Klassifikationsergebnisse bei Kreuzvalidierung (hier: Leave-one-out-Methode). Die ursprünglich mit der Resubstitutions-Methode berechnete Trefferquote von 100 % ist gegenüber der bei der **Kreuzvalidierung** ermittelten **Trefferquote von 83,3 %** optimistisch überhöht.

♦