8. Methoden der klassischen multivariaten Statistik

8.1. Darstellung von Daten

Voraussetzungen auch in diesem Kapitel:

- Grundgesamtheit (Datenraum) Ω von Objekten (Fällen, Instanzen), denen *J*-Tupel von Merkmalsausprägungen zugeordnet sind
- "multivariate Statistik": mehrere Merkmale (Variablen; Attribute), d.h. $J > 1$
- Merkmale können nominal, ordinal, Intervall- oder Ratio-skaliert sein (siehe Kap. 1)

"Beschreibende Statistik" beschreibt die Daten durch einzelne Werte, durch übersichtliche Tabellen oder durch grafische Darstellungen. (im Folgenden teilw. nach Schwenker 2004)

Absolute und relative Häufigkeit:

Das Merkmal X habe J verschiedene mögliche Merkmalswerte, die wir mit $\zeta_1, \ldots, \zeta_J$ bezeichnen.

Absolute Häufigkeit von ζ_j ist n_j die Anzahl der Daten mit dem Wert ζ_j für $j = 1, \ldots, J$.

Relative Häufigkeit von ζ_j ist $f_j = \frac{n_j}{n}$ der Anteil der Daten mit dem Wert ζ_j für $j = 1, \ldots, J$

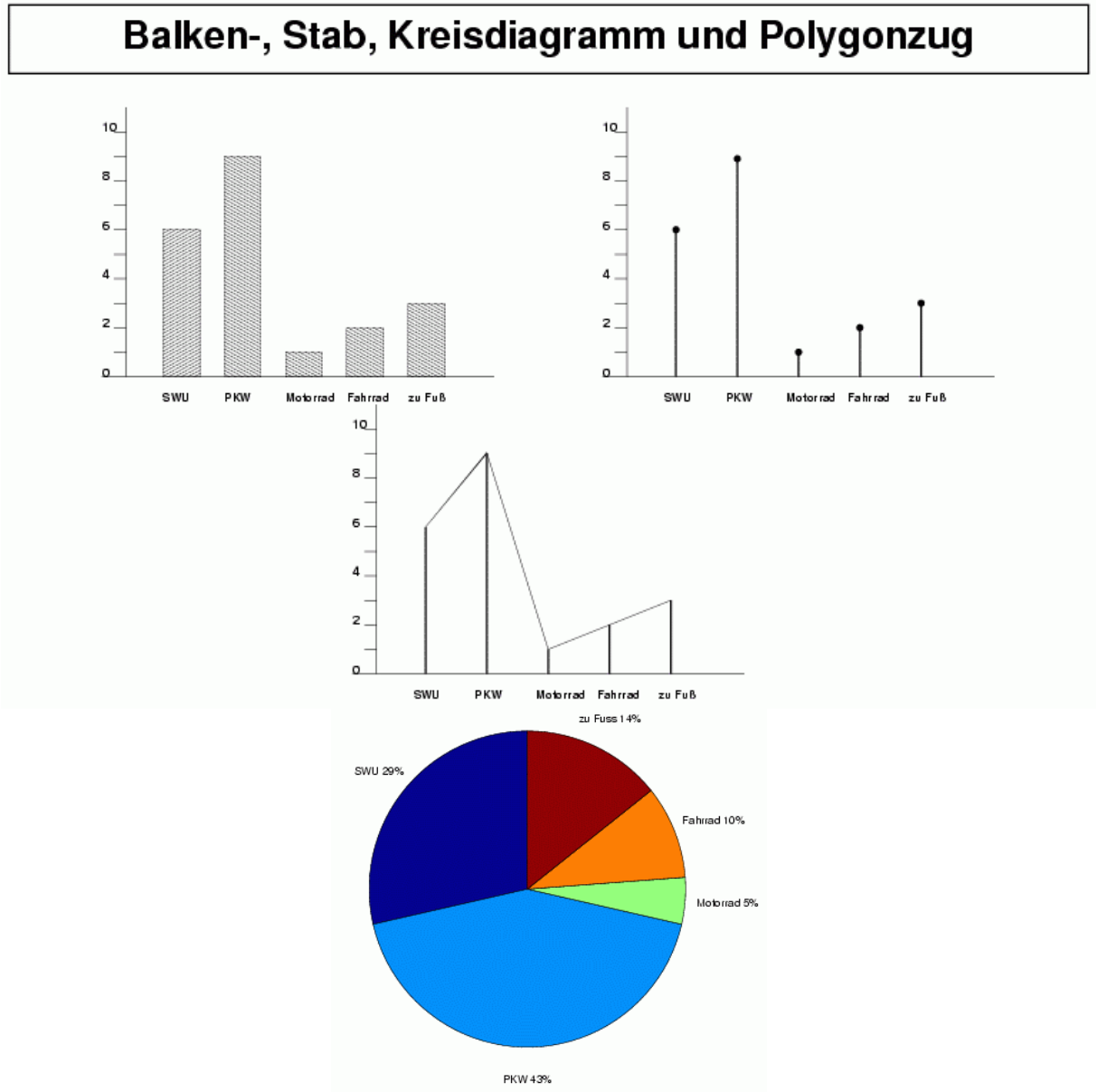
Es gilt dabei offenbar

$$\sum_{j=1}^{J} n_j = n \quad \text{und} \quad \sum_{j=1}^{J} f_j = 1$$

Darstellung durch Tabellen der absoluten oder relativen Häufigkeiten für jedes Merkmal und jede Ausprägung:

- beschreibt den Datensatz bei Vernachlässigung der Zusammenhänge zwischen Merkmalen

grafische Darstellungen bei 1 Merkmal ("univariater" Fall):



Der Modalwert

Die Ausprägung ζ_j heißt **Modus** oder **Modalwert**, falls $n_j \geq n_k$ für alle $k = 1, \ldots, J$. Der Modus ist nicht eindeutig bestimmt, d.h. die können Daten mehrere Modi aufweisen.

anschaulich: "Gipfel" im Balkendiagramm

Verallgemeinerung des Balkendiagramms: Histogramm

Histogramme

Angenommen es liegt wiederum ein metrisch skaliertes Merkmal vor mit den Daten: $x_1, x_2, ...x_n$. Ferner nehmen wir an, dass n groß ist, so dass die einzelnen Daten x_i keine Rolle spielen. Für solche Anwendungen bietet sich eine Darstellung der Daten durch **Histogramme** an.

Histogramm-Darstellung

- Die Werte des Merkmals werden in Intervallen („Klassen“) K_j zusammengefaßt.
- Einzeldaten x_i kommen nicht mehr vor.
- Anzahl n_j der Daten je Klasse K_j werden angegeben.

Mittels der Klassengrenzen

$$a_1 < b_1 = a_2 < b_2 = a_3 < \cdots = b_{J-1} = a_J < b_J$$

werden J Klassen festgelegt durch

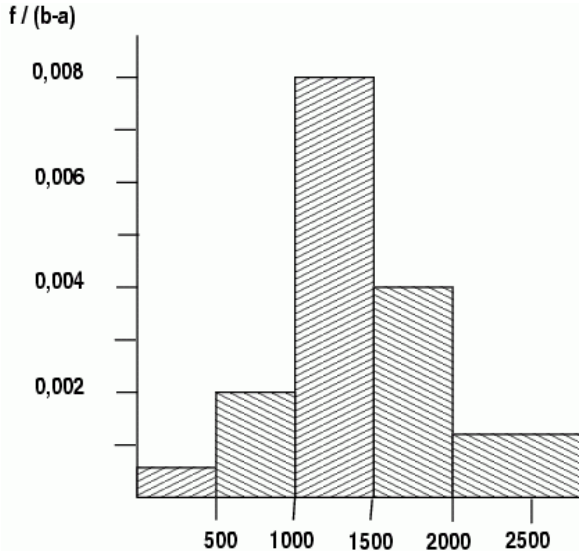
$$K_j := [a_j, b_j) \text{ für } j = 1, \ldots, J-1 \quad \text{und} \quad K_J = [a_J, b_J]$$

Aus den eigentlichen Daten

$$x_1, \ldots, x_n$$

werden die Klassen K_j mit Häufigkeiten n_j, wobei $n_j = K_j \cap \{x_1, \ldots, x_n\}$ ist.

Trägt man die empirischen Daten über den Klassen (also über den Intervallen) ab, so entsteht ein *Histogramm*.



Probleme, die bei der Histogrammbildung aufkommen:

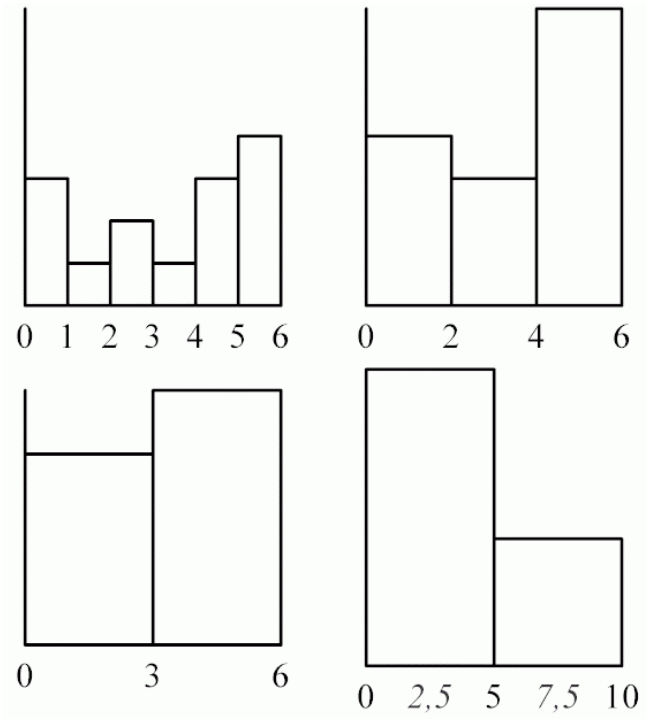
- Wie viele Klassen sind für die vorliegenden Daten erforderlich ? Ist $J = 5$ ausreichend?
- Eine sehr grobe Faustregel:

$$J \approx \begin{cases} 10 \log_{10} n & \text{falls } n > 1000 \\ \sqrt{n} & \text{sonst} \end{cases}$$

- Sollen die Klassen (Intervalle) jeweils gleich lang sein?
- Kann man sich auf endliche Unter- und Obergrenzen der Klassen beschränken?

Innerhalb der Klassen K_j nimmt man eine Gleichverteilung der Daten an.

- Abhängigkeit des Histogramms von der gewählten Klasseneinteilung
- ungünstig gewählte Klasseneinteilung kann falsche Strukturen in den Daten vortäuschen
  Beispiel:



Deshalb in vielen Fällen vorzuziehen:

Die (empirische bzw. kumulative) Verteilungsfunktion

Das Merkmal X sei nun (mindestens) ordinalskaliert, d.h. es gibt eine natüliche Ordung der Merkmalsausprägungen (dies sind oBdA reelle Zahlen also $\in \mathbb{R}$).

$$x_1, \ldots, x_n$$

seien die Daten der Urliste.

Als **(empirische) Verteilungsfunktion** der Daten bezeichnet man die Funktion $F : \mathbb{R} \to [0, 1]$ definiert durch

$$F(x) = \frac{|\{i \ : \ x_i \leq x\}|}{n} = \sum_{j \in \{r \mid \zeta_r \leq x\}} f_j$$

$F(x)$ ist also der Anteil der Daten x_i mit der Eigenschaft: $x_i \leq x$.

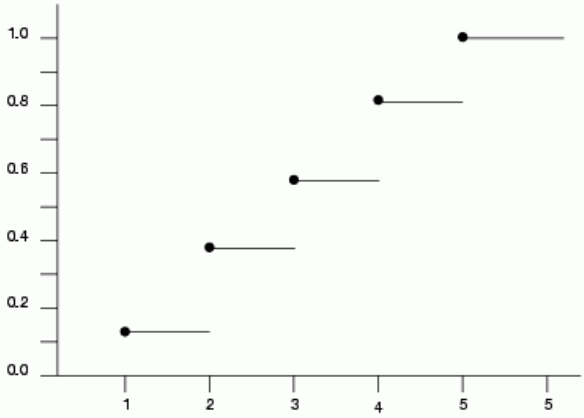
Die Verteilungsfunktion ergibt sich also direkt aus den relativen Häufigkeiten.

Beispiel: Ergebnisse einer Klausur mit 16 Teilnehmern sind gegeben durch die Urliste

$$3, 4, 2, 1, 2, 4, 5, 5, 2, 1, 4, 5, 3, 3, 2, 4$$

Hieraus ergibt sich die empirische Verteilungsfunktion dieser Daten:

ζ_j	n_j	f_j in %	$F(\zeta_j)$ in %
1	2	12,50	12,50
2	4	25,00	37,50
3	3	18,75	56,25
4	4	25,00	81,25
5	4	18,75	100,00

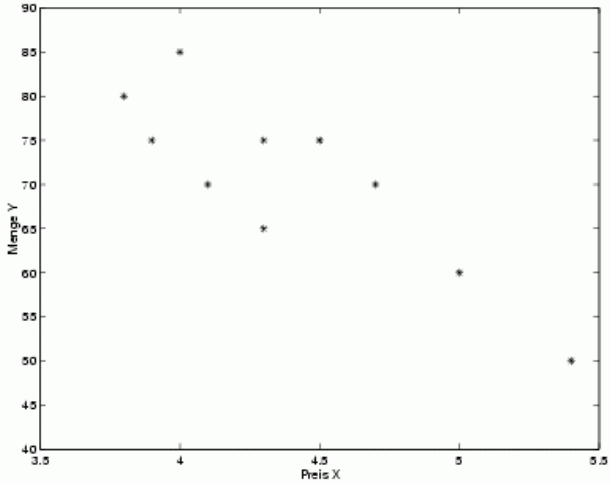


Verteilungsfunktionen sind monoton wachsend, d.h. $F(x_1) \leq F(x_2)$ für $x_1 < x_2$ und rechtsseitig stetige **Treppenfunktion**:

Vorteil der kumulativen Verteilungsfunktion:
Unabhängigkeit von Klassierungsannahmen

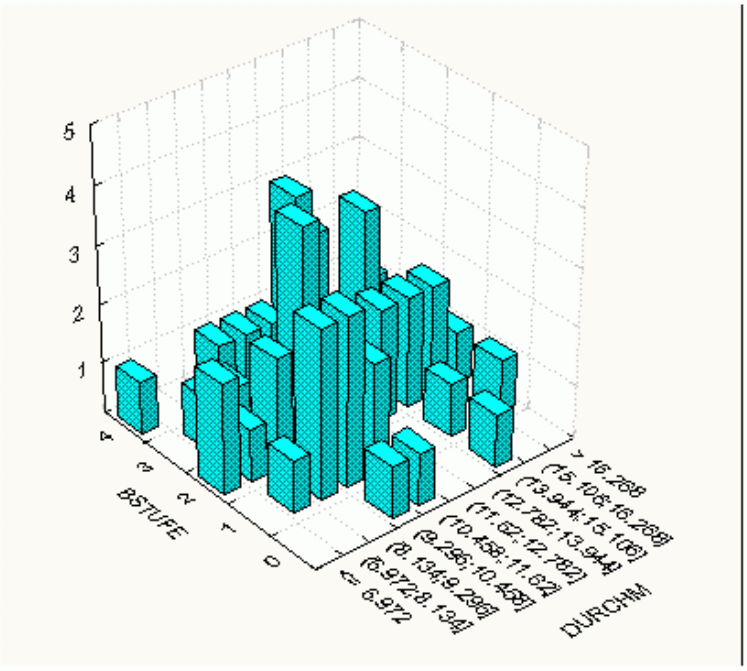
Multivariater Fall: mehrere Merkmale

Sind zwei Merkmale X_i und X_j metrisch skaliert, veranschaulicht man sich die Daten in einem **Streudiagramm** (scatterplot)



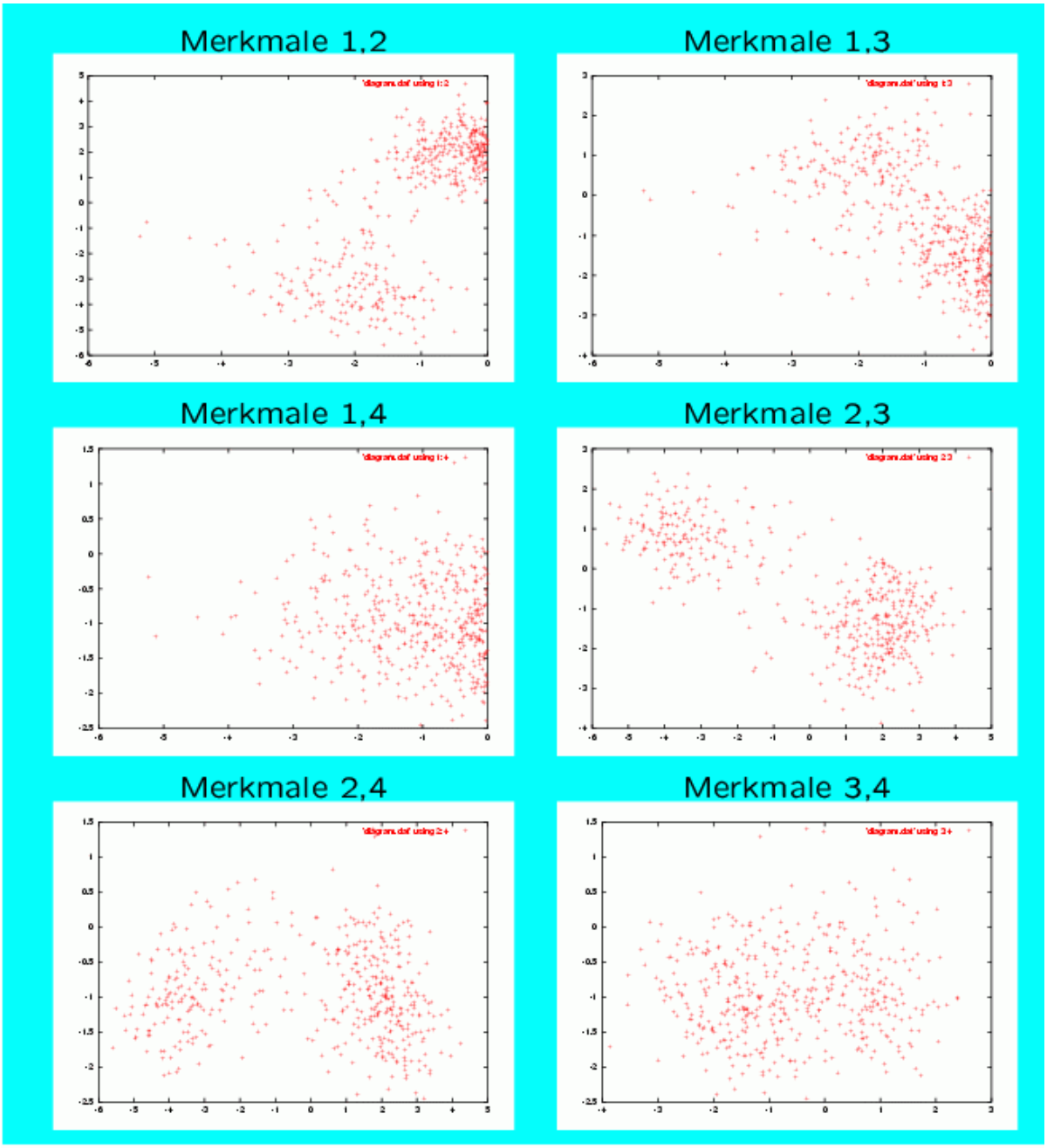
Wie hängen X und Y von einander ab? Vermutung: *Höhere Preise entsprechen geringeren Mengen*.

in Analogie zum univariaten Histogramm:
bivariates Histogramm (3D-Grafik)



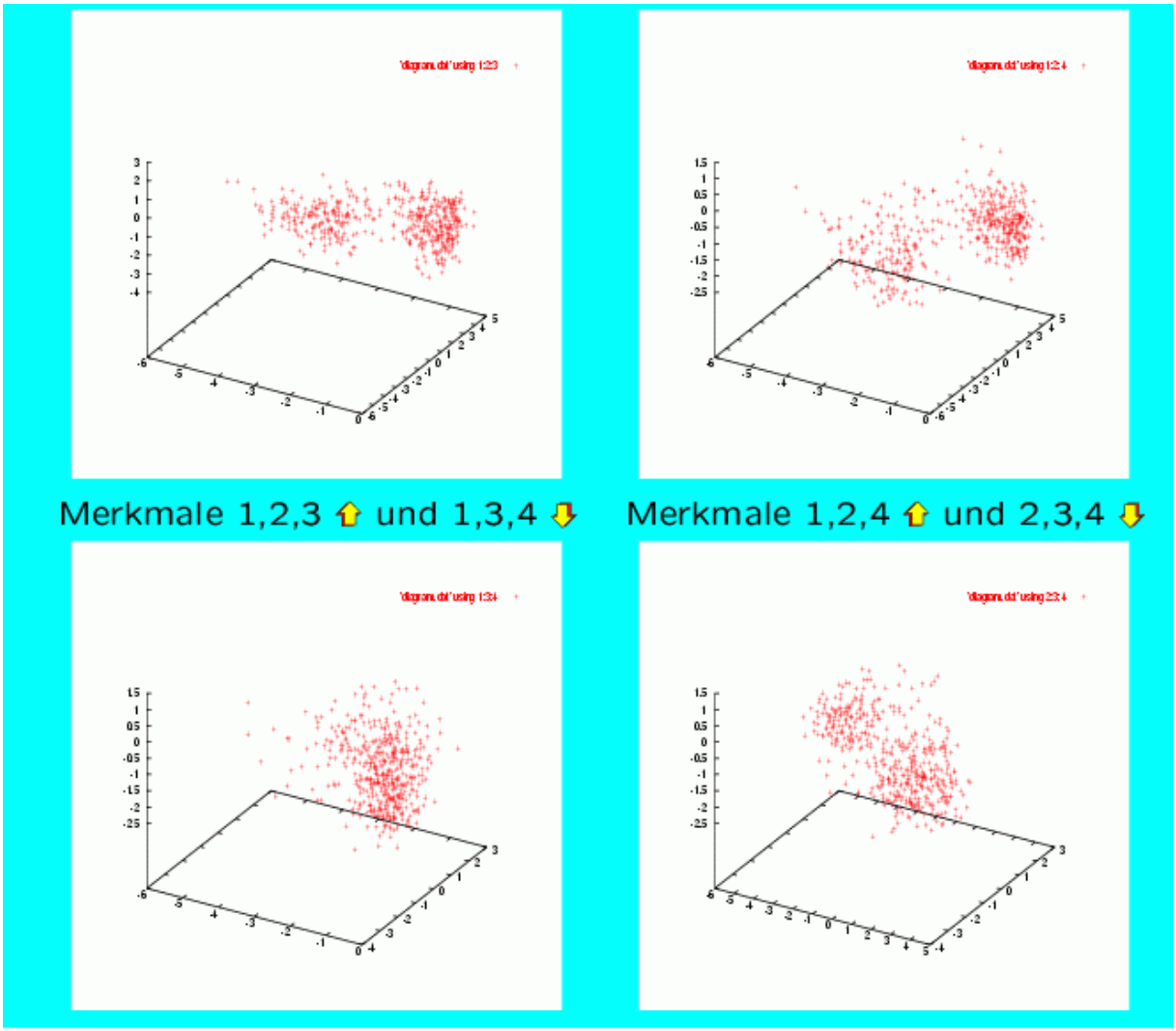
(gleiche Nachteile wie bei univariaten Histogrammen: Informationsverluste, abhängig von der gewählten Klassierung)

bei mehr als 2 Merkmalen: alle Kombinationen in einzelnen Scatterplots erfassen



Es gibt $\binom{N}{2} = (N-1)\cdot N/2$ viele 2D-Streudiagramme!

- oder: 3D-Scatterplots verwenden



Es gibt $\binom{N}{3} = (N-2)(N-1)N/6$ viele 3D-Streudiagramme!

Mehr als 3 Merkmale sind nicht gleichzeitig in einer Grafik als Scatterplot darstellbar

⇒ Fragestellung der *mehrdimensionalen Skalierung*:

lässt sich eine niedrigdimensionale Darstellung für eine hochdimensionale Datenmenge (Attributzahl *J* groß) finden, so dass die Strukturen in den Daten weitgehend erhalten bleiben?

QUELLDATENSATZ	ZIELDATENSATZ
$\omega^x = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\} \subset \mathbb{R}^N$	$\omega^y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\} \subset \mathbb{R}^M$
$r_{ij}^x = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$	$r_{ij}^y = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|$

mit der Dimension $M << N$

GESUCHT: niederdimensionale Punktmenge aus ω^y mit geringstmöglicher Abweichung der paarweisen Distanzen:

$$r_{ij}^x \approx r_{ij}^y \qquad (\forall i, j)$$

Fehlermaße

- **Absoluter quadratischer Fehler**

$$\varepsilon_a(\boldsymbol{R}^x, \boldsymbol{R}^y) = \frac{1}{\sum_{i<j}(r_{ij}^x)^2} \sum_{i<j} \left(r_{ij}^y - r_{ij}^x\right)^2$$

- **Relativer quadratischer Fehler**

$$\varepsilon_r(\boldsymbol{R}^x, \boldsymbol{R}^y) = \sum_{i<j} \left(\frac{r_{ij}^y - r_{ij}^x}{r_{ij}^x}\right)^2$$

- Kompromiß: der **Sammon-Fehler**

$$\varepsilon_s(\boldsymbol{R}^x, \boldsymbol{R}^y) = \frac{1}{\sum_{i<j}(r_{ij}^x)^2} \sum_{i<j} \frac{\left(r_{ij}^y - r_{ij}^x\right)^2}{r_{ij}^x}$$

(aus Schukat-Talamazzini 2002)

Tabellarische Darstellungen multivariater Datensätze:

Häufigkeits- und Kontingenztafeln

Zwei Merkmale X und Y seine beliebig skaliert. Für X seien die möglichen Merkmalsausprägungen $\xi_1,\ldots,\xi_J$ und für Y seien $\eta_1,\ldots,\eta_K$ die Merkmalsausprägungen.

Sei nun die Urliste in Form einer $n \times 2$ Datenmatrix gegeben, so lässt sich hieraus eine sogenannte **Häufigkeitstabelle** erstellen.

Es sei dabei n_{jk} die Anzahl der Datenpaare (x_i, y_i) mit $x_i = \xi_j$ und $y_i = \eta_k$.

$$n_{j\cdot} = \sum_{k=1}^{K} n_{jk} \quad \text{und} \quad n_{\cdot k} = \sum_{j=1}^{J} n_{jk}$$

sind die **absoluten Randhäufigkeiten** von ξ_j bzw. η_k.

		Y				
		η_1	η_2	$\cdots$	η_k	$\sum$
X	ξ_1	n_{11}	n_{12}	$\cdots$	n_{1K}	$n_{1\cdot}$
	ξ_2	n_{21}	n_{22}	$\cdots$	n_{2K}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	ξ_J	n_{J1}	n_{J2}	$\cdots$	n_{JK}	$n_{J\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K}$	n

$$\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = \sum_{j=1}^{J} n_{j\cdot} = \sum_{k=1}^{K} n_{\cdot k} = n$$

Die Randhäufigkeiten $n_{\cdot 1},\ldots,n_{\cdot K}$ beziehen sich offenbar nur auf das Merkmal Y und die Randhäufigkeiten $n_{1\cdot},\ldots,n_{J\cdot}$ nur auf das Merkmal X.

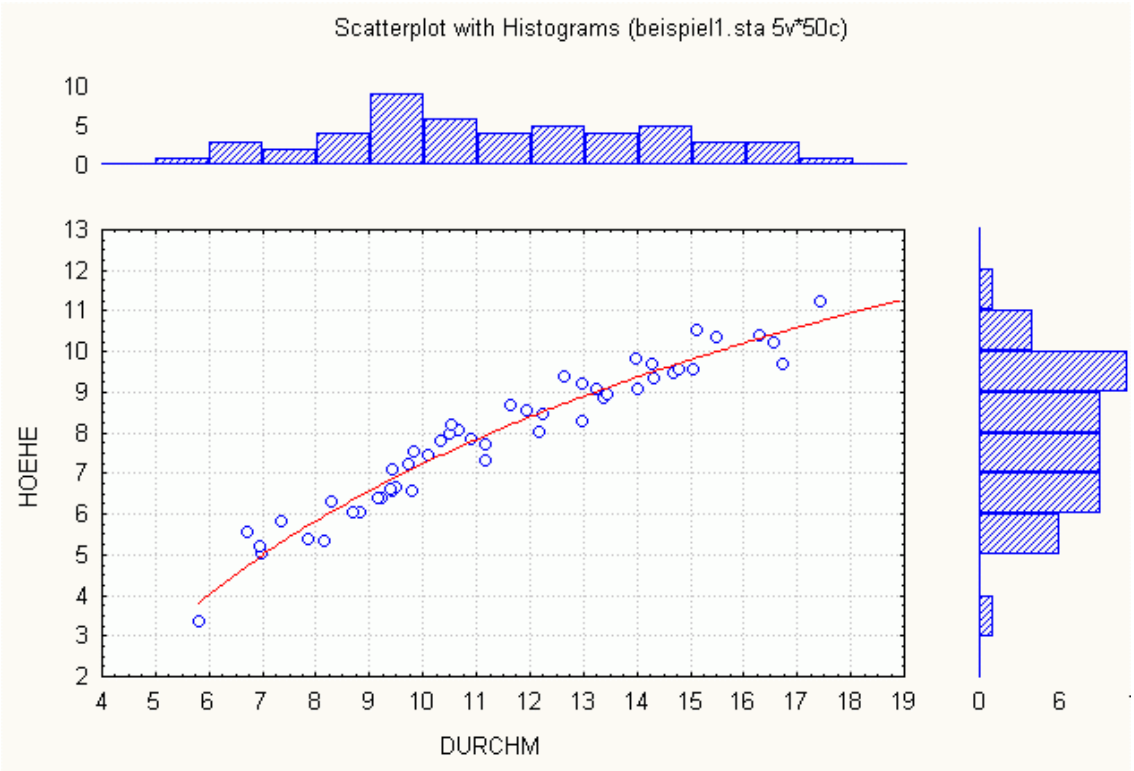
Statt der absoluten Häufigkeiten kann man auch die **Kontingenztafel** mit relativen Häufigkeiten aufstellen.

		Y				
		η_1	η_2	$\cdots$	η_k	$\sum$
X	ξ_1	f_{11}	f_{12}	$\cdots$	f_{1K}	$f_{1\cdot}$
	ξ_2	f_{21}	f_{22}	$\cdots$	f_{2K}	$f_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	ξ_J	f_{J1}	f_{J2}	$\cdots$	f_{JK}	$f_{J\cdot}$
	$\sum$	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot K}$	1

$f_{1\cdot}, \ldots, f_{J\cdot}$ heißt Randverteilung von X

$f_{\cdot 1}, \ldots, f_{\cdot K}$ heißt Randverteilung von Y

Die Randverteilungen können wieder durch Histogramme oder kumulative Verteilungsfunktionen dargestellt werden.



Scatterplot für 2 Merkmale mit Randhistogrammen (je eines für jedes Merkmal) und mit Anpassungskurve (hier eine Logarithmusfunktion).

Bedingte Verteilungen:

Von den gemeinsamen relativen Häufigkeiten zu unterscheiden sind die sogenannten **bedingten relativen Häufigkeiten**

Für festes $k \in 1, \ldots, K$ und $j = 1, \ldots, J$ ist

$$f_{j|Y=\eta_k} = \frac{f_{jk}}{f_{\cdot k}}$$

die **bedingte relative Häufigkeit** von ξ_j unter der Bedingung $Y = \eta_k$.

Sie stellt die relative Häufigkeit des Wertes ξ_j in der Teilmenge der Objekte dar, die in der Variablen Y den Wert η_k haben.

$$f_{1|Y=\eta_k}, \ldots, f_{J|Y=\eta_k}$$

heißt die **bedingte Verteilung** von X unter der Bedingung $Y = \eta_k$.

Analog ist

$$f_{k|X=\xi_j} = \frac{f_{jk}}{f_{j\cdot}}$$

die **bedingte relative Häufigkeit** von η_k unter der Bedingung $X = \xi_j$.

Die **bedingte Verteilung** von Y unter der Bedingung $X = \xi_j$ ist gegeben durch:

$$f_{1|X=\xi_j}, \ldots, f_{K|X=\xi_j}$$

Aus den bedingten relativen Häufigkeiten für Y unter der Bedingung $X = \xi_j$ und den absoluten Randhäufigkeiten von X können die gemeinsamen absoluten Häufigkeiten η_{jk} bestimmt werden, es gilt:

$$n_{jk} = f_{k|X=\xi_j} n_{j\cdot} = \frac{n_{jk}}{n_{j\cdot}} n_{j\cdot} \quad \text{und} \quad n_{jk} = f_{j|Y=\eta_k} n_{\cdot k} = \frac{n_{jk}}{n_{\cdot k}} n_{\cdot k}$$

Auch die bedingten Verteilungen können durch Histogramme oder kumulative Verteilungsfunktionen veranschaulicht werden.

8.2. Kenngrößen der beschreibenden Statistik

Lokalisations- (Lage-) Maße:

Lagemaße

Für metrisch skalierte Daten $x_1, \ldots, x_n$ ist das **arithmetische Mittel** $\bar{x}$ definiert durch

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

das am weitesten verbreitete Lagemaß

Weiteres Lokalisationsmaß: *Modus* (Modalwert) –

- schon für nominal skalierte Attribute definiert
- aber nicht immer eindeutig bestimmt

Quantile

Weiterer wichtiger Begriff zur Beschreibung von Daten ist der des **Quantils**.

Mit Hilfe der empirischen Verteilungsfunktion F definieren wir für $0 < p < 1$:

$$\tilde{x}_p = \min_{x \in \mathbb{R}} \{x \ : \ F(x) \geq p\}$$

$\tilde{x}_p$ ist der kleinste Wert $x \in \mathbb{R}$ mit der Eigenschaft: $F(x) \geq p$.

$\tilde{x}_p$ wird als p-Quantil der Daten bezeichnet.

$\tilde{x}_p$ ist der kleinste Wert $x \in \mathbb{R}$ so dass $p * 100$ % der Daten $\leq x$ sind.

$Q : [0, 1] \to \mathbb{R}$ mit $p \to \tilde{x}_p$ heißt die **Quantilfunktion**.

Quantilfunktion und empirische Verteilungsfunktion enthalten die gleiche Information über die Daten (Spiegelung an der Winkelhalbierenden!).

Quantile könne auch direkt aus den Daten berechnet werden, also ohne Bestimmung der empirischen Verteilungsfunktion.

Hierzu seien die Daten in der Urliste bereits aufsteigend sortiert:

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$$

Dann ist für $p \in (0,1)$

$$\tilde{x}_p = \begin{cases} x_{np+1} & \text{falls } np \text{ ganzzahlig} \\ x_{[np+1]} & \text{sonst} \end{cases}$$

hierbei ist $[x]$ der ganzzahlige Anteil von x.

Einige Quantile haben besondere Namen:

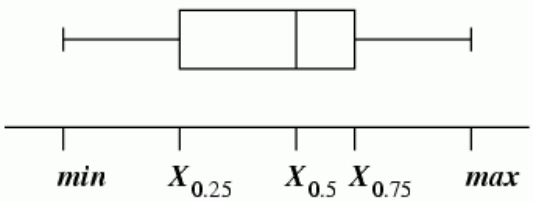
Median	$\tilde{x}_{\frac{1}{2}}$
Quartile	$\tilde{x}_{\frac{1}{4}}, \tilde{x}_{\frac{2}{4}}, \tilde{x}_{\frac{3}{4}}$
Quintile	$\tilde{x}_{\frac{1}{5}}, \tilde{x}_{\frac{2}{5}}, \tilde{x}_{\frac{3}{5}}, \tilde{x}_{\frac{4}{5}}$
Dezile	$\tilde{x}_{\frac{1}{10}}, \tilde{x}_{\frac{2}{10}}, \ldots, \tilde{x}_{\frac{8}{10}}, \tilde{x}_{\frac{9}{10}}$
Perzentile	$\tilde{x}_{\frac{1}{100}}, \tilde{x}_{\frac{2}{100}}, \ldots, \tilde{x}_{\frac{98}{100}}, \tilde{x}_{\frac{99}{100}}$

Quantile sind offensichtlich gut zu interpretieren und nützlich um große Datenmengen mit vielen verschiedenen Werten zu charakterisieren.

- $\tilde{x}_{\frac{1}{2}}$, der Median, ist der Wert der die unteren 50% von den oberen 50% der Daten trennt.
- $\tilde{x}_{\frac{1}{4}}, \tilde{x}_{\frac{1}{2}}, \tilde{x}_{\frac{3}{4}}$ teilen die Daten in vier Blöcke, die jeweils 25 Prozent der Daten umfassen. Zwischen $\tilde{x}_{\frac{1}{4}}$ und $\tilde{x}_{\frac{3}{4}}$ – dem unteren und oberen Quartil – liegen die mittleren 50% der Daten.

Analog sind Quintile, Dezile und Perzentile zu interpretieren.

Kastendiagramm/Boxplot



Minimum und Maximum.

Median innerhalb der Box.

1. und 3. Quartil beschreiben die Lage der Box.

Streuungsmaße (*Dispersionsmaße*):

Varianz/Standardabweichung

$x_1, \ldots, x_n$ seien metrisch skaliert (mindestens intervallskaliert)

- **Varianz und Standardabweichung** sind am gebräuchlichsten

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{(Varianz)}$$

durch Wurzelziehen ergibt sich

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{(Standardabweichung)}$$

Gelegentlich findet man auch $\frac{1}{n}$ statt $\frac{1}{n-1}$ Normalisierung.
Allerdings wird hierdurch die Varianz/Standardabweichung systematisch unterschätzt!

Eigenschaften von Varianz/Streuung:

1. $s^2 \geq 0$ und $s \geq 0$. Denn es gilt: $s = 0 \Leftrightarrow s^2 = 0 \Leftrightarrow x_1 = \cdots = x_n = \bar{x}$
2. Durch Umformung erhält man leicht (bei $\frac{1}{n}$ Normierung): $s^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$
3. Nach einer affinen Transformation $y_i := ax_i + b$ der Daten x_i gilt (bei $\frac{1}{n}$ Normierung): $s_y^2 = a^2 s_x^2$ bzw. $s_y = |a| s_x$

- **Mittlere absolute Abweichung vom Median**

$$d := \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}_{1/2}|$$

- **Ginis mittlere Differenz**

$$\Delta := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

d und Δ sind in geringerem Maße von Ausreißern betroffen als s^2.
Denn es werden nicht die quadrierten Abstände, sondern nur die gewöhnlichen Abstände in diesem Streuungsmaß eingehen.

- **Quartilabstand**

$$Q := \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

ist der Quartilabstand.
Q ist die Spanne in der die mittleren 50% der Daten liegen (siehe Boxplot).
Q ist besonders robust gegenüber Ausreißern.

- **Spannweite/Range**

$$R = \max_i x_i - \min_i x_i$$

heißt die Spannweite.
R ist besonders empfindlich gegenüber Ausreißern.

- **Variationskoeffizient**

$$V = \frac{s}{\bar{x}} \quad \text{für } \bar{x} > 0$$

drückt die Standardabweichung in „Mittelwerteinheiten“ aus.

Kenngrößen für multivariate Daten (Betrachtung mehrerer Attribute zugleich):

Kovarianz

Zur Herleitung einer **Zusammenhangsmaßzahl** für zwei Merkmale X und Y bilden wir die arithmetischen Mittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{und} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

und die Varianzen

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{und} \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Wir definieren ferner die **Kovarianz** s_{xy} durch:

$$s_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Eigenschaften der Kovarianz:

- Die Kovarianz s_{xy} kann negativ sein (Varianz ist stets ≥ 0).

- Es gilt $s_{xy} = s_{yx}$

- Durch einen Punkt $(x_i, y_i) \in \mathbb{R}^2$ und den Schwerpunkt $(\bar{x}, \bar{y}) \in \mathbb{R}^2$ wird offenbar ein Rechteck aufgespannt, dessen Flächeninhalt F_i gegeben ist durch:

$$F_i = |(x_i - \bar{x}) \cdot (y_i - \bar{y})|$$

- $(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$, so liegt der Punkt (x_i, y_i) im 1. oder 3. Quadranten (bzgl. $(\bar{x}, \bar{y})$ als Nullpunkt)

- $(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$, so liegt (x_i, y_i) im 2. oder 4. Quadranten

- $s_{xy} > 0$, so haben X und Y die gleiche Tendenz.

- $s_{xy} < 0$, so haben X und Y die entgegengesetzte Tendenz

- Die Kovarianz ist lage-invariant und linear. Für die Transformation

$$(x_i, y_i) \mapsto (\hat{x}_i, \hat{y}_i)$$

$$\hat{x}_i := ax_i + b, \quad \text{und} \quad \hat{y}_i = cx_i + d \quad \text{mit} \quad a, b, c, \in \mathbb{R}$$

gilt dann $s_{\hat{x}\hat{y}} = acs_{xy}$, denn

$$s_{\hat{x}\hat{y}} = \frac{1}{n-1} \sum_{i=1}^{n} (ax_i + b - (a\bar{x} + b))(cy_i + d - (c\bar{y} + d)) = acs_{xy}$$

- Offenbar ist s_{xy} nicht normiert und kann beliebige reelle Werte annehmen.

Kovarianzmatrix: $\Sigma = (s_{xy})_{x,y \text{ Merkmale}}$

(die Diagonalelemente s_{xx} sind die Varianzen.)

Korrelationskoeffizienten

Normierung der Kovarianz durch die Standardabweichungen geben den **Korrelationskoeffizienten**:

$$r_{xy} := \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

1. Es gilt $r_{xy} = r_{yx}$, da $s_{xy} = s_{yx}$

2. Für $\hat{x}_i = ax_i + b$ und $\hat{y}_i = cy_i + d$ mit $a, b, c, \in \mathbb{R}$ und $a \cdot c \neq 0$ gilt:

   $$r_{\hat{x}\hat{y}} = \frac{ac}{|a||c|} r_{xy}$$

   Dies folgt aus den bekannten Eigenschaften der Kovarianz und der Standardabweichung. Es ist offensichtlich, dass sich $r_{\hat{x}\hat{y}}$ und r_{xy} nur um das Vorzeichen des Produkts der Skalierungskonstanten a und c unterscheiden.
   Es gilt:

   - $ac > 0$, dann ist $r_{\hat{x}\hat{y}} = r_{xy}$
   - $ac < 0$, dann ist $r_{\hat{x}\hat{y}} = -r_{xy}$

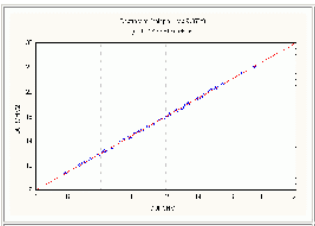
3. Es ist $r_{xy} \in [-1, 1]$

4. $r_{xy} = \pm 1$, gdw. $y_i = ax_i + b$ für alle $i = 1, \ldots, n$ gilt, d.h. ein exakter linearer Zusammenhang zwischen den Merkmalen X und Y besteht.

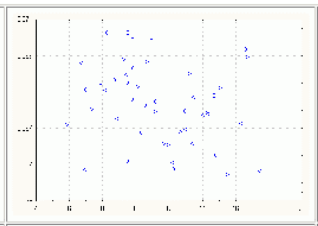
   - $r_{xy} = 1$, gdw. $a > 0$ und $b \in \mathbb{R}$ mit $y_i = ax_i + b$ für alle $i = 1, \ldots, n$.
   - $r_{xy} = -1$, gdw. $a < 0$ und $b \in \mathbb{R}$ mit $y_i = ax_i + b$ für alle $i = 1, \ldots, n$.

r_{xy} heißt auch *Pearsonscher Korrelationskoeffizient*. Auch: *R*.

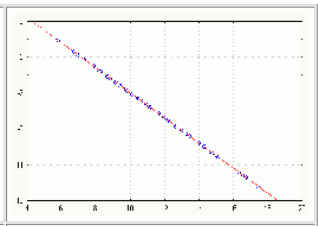
Korrelationsmatrix $K = (r_{xy})_{x,y}$ Merkmale



$R = +1$



R ist ungefähr 0, kein Zusammenhang



$R = -1$

Unabhängigkeit zweier Attribute (Merkmale):

Zwei Variablen (Merkmale) X und Y heißen **deskriptiv unabhängig**, wenn gilt:

$$n_{jk} = \frac{n_{j\cdot} n_{\cdot k}}{n} \quad \text{für alle} \quad j = 1, \ldots, J \text{ und } k = 1, \ldots, K$$

Folgende Aussagen sind dazu äquivalent:

1. $f_{jk} = f_{j\cdot} f_{\cdot k}$ für alle $j = 1, \ldots, J,$ und $k = 1, \ldots, K$
2. $f_{j\cdot} = f_{j|Y=\eta_1} = f_{j|Y=\eta_2} = \cdots = f_{j|Y=\eta_K}$ für alle $j = 1, \ldots, J$
3. $f_{k\cdot} = f_{k|X=\xi_1} = f_{k|X=\xi_2} = \cdots = f_{k|X=\xi_J}$ für alle $k = 1, \ldots, K$

Zusammenhangsmaße:

X und Y seien deskriptiv unabhängige Variablen, dann gilt :

$$s_{xy} = 0$$

d.h. X und Y sind unkorreliert.

$$\begin{aligned} s_{xy} &= \frac{1}{n-1} \sum_{j=1}^{J} \sum_{k=1}^{K} (\xi_j - \bar{x})(\eta_k - \bar{y}) n_{jk} \\ &= \frac{1}{n-1} \sum_{j=1}^{J} (\xi_j - \bar{x}) n_{j\cdot} \; \frac{1}{n-1} \sum_{k=1}^{K} (\eta_k - \bar{y}) n_{\cdot k} \\ &= 0 \end{aligned}$$

Damit ist auch $r_{xy} = 0$. Die Umkehrung ist falsch.

r_{xy} ist ein Maß für den **linearen Zusammenhang** zweier Merkmale X und Y. Andere Formen des funktionalen Zusammenhangs erfasst r_{xy} nicht!

Zusammenhangsmaß für ordinal skalierte Daten

Die Anwendung des Korrelationskoeffizienten ist nicht zulassig, da $\bar{x}, \bar{y}, {s_X}^2, s_Y^2$ und s_{XY} für ordinal skalierte Merkmale nicht sinnvoll berechnet werden können.

Idee: Die daten x_i, y_i für $i = 1, \ldots, n$ werden durch ihre **Ränge** $R_X(x_i)$ und $R_Y(y_i)$ ersetzt und der Korrelationskoeffizient für die Ränge bestimmt.

Seien die Daten $x_1, x_2, \ldots, x_n$ paarweise verschieden, dann ist $R_X(x_i) = r$, wenn x_i in der aufsteigend geordneten Folge der x_i-Werte an der r-ten Stelle steht.

Der **Rangordnungskoeffizient von Spearman** ist nun definiert durch:

$$r_{Sp} = \frac{\sum_{i=1}^{n}(R_X(x_i) - \bar{R}_X)(R_Y(y_i) - \bar{R}_Y)}{\left(\sum_{i=1}^{n}(R_X(x_i) - \bar{R}_X)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n}(R_X(x_i) - \bar{R}_X)^2\right)^{\frac{1}{2}}}$$

Sind die x_i und die y_i jeweils paarweise verschieden, so gilt die vereinfachte Formel:

$$r_{Sp} = 1 - \frac{6\sum_{i=1}^{n}(R_X(x_i) - R_Y(y_i))^2}{n(n^2-1)}$$

Kommen Datenwerte mehrfach vor (sog. Bindungen), so werden Durchschnittsränge gebildet.

Beispiel:

$x_1 = 3.7$, $x_2 = 3.9$, $x_3 = 3.1$ und $x_4 = 3,7$.
Dann ist $R_X(x_2) = 4$, $R_X(x_3) = 1$ und wegen $x_1 = x_4 = 3.7$ entfallen die Ränge 2 und 3 und werden durch einen Durchschnittsrang realisiert
$R_X(x_1) = R_X(x_4) = 2.5$

Zusammenhangsmaß für nominal skalierte Daten

Daten seien in Form einer Kontingenztafel gegeben. Die Variablen sind deskriptiv unabhängig, falls gilt

$$n_{jk} = \frac{n_{j\cdot} n_{\cdot k}}{n} \quad \text{für alle} \quad j = 1, \ldots, J \text{ und } k = 1, \ldots, K$$

Maß für die Abweichung von der Unabhängigkeit ist

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j\cdot} n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot} n_{\cdot k}}{n}} = n\left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{n_{jk}^2}{n_{j\cdot} n_{\cdot k}} - 1\right)$$

$\chi^2 = 0$, gdw. X und Y deskriptiv unabhängig sind.
χ^2 ist nicht normiert! Statt χ^2 verwendet man den **Kontingenzkoeffizienten**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n} \frac{\min\{J, K\}}{\min\{J, K\} - 1}} \in [0, 1]$$

Es gilt: $C = 0$ gdw. $\chi^2 = 0$ gdw. X und Y deskriptiv unabhängig.

(Schwenker 2004)

8.3. Statistische Tests

Hypothesentests

Ziel: Anhand empirischer Daten über ein Merkmal soll eine Entscheidung herbeigeführt werden zwischen zwei konträren Aussagen.

Beispiel:

a. Einstichprobenproblem (1-sample problem):

Hat ein neuer Datensatz A einen mittleren Wert μ des betrachteten Attributs, der größer ist als der (bisher für dieses Attribut angenommene) Wert μ_0? Entscheidung anhand der Ergebnisse $x_1, \ldots, x_n$ von n Messungen.

b. Zweistichprobenproblem (2-sample problem):

Hat der Datensatz A im Mittel einen größeren Attributwert als der Datensatz B?
Entscheidung anhand der Ergebnisse $x_1, ..., x_{n_A}$ von Datensatz A und $x_1, ..., x_{n_B}$ von Datensatz B.

Nullhypothese (null hypothesis)

H_0: $\mu \leq \mu_0$

bzw. $\mu_A \leq \mu_B$ (2 Stichproben).

Alternativhypothese (alternative hypothesis)

H_A: $\mu > \mu_0$

bzw. $\mu_A > \mu_B$ (2 Stichproben).

Vorgehensweise:

Ergibt die Empirie einen *ungewöhnlich* großen Stichprobenmittelwert $\bar{x}$ (d.h. $\bar{x}$ im *Ablehnungsbereich*, kritischer Bereich), also ein Ergebnis, das unter der Nullhypothese H_0: $\mu \leq \mu_0$ sehr *unwahrscheinlich* wäre, so vermutet (schließt) man, daß H_0 *wohl falsch* ist, d.h. man entscheidet sich für H_A.

Bei diesem Schluß kann man sich irren:

Fehler 1. Art (type I error) :

Ablehnung von H_0 zu Unrecht,
also obwohl H_0 zutrifft.
Fehlerwahrscheinlichkeit α
(wird vorgegeben, z.B. $\alpha = 0{,}05$).
α = "*Risiko* I", *Signifikanzniveau*;

$1-\alpha$ heißt *statistische Sicherheit*.

Liegt $\bar{x}$ dagegen im *Annahmebereich* von H_0, ist $\bar{x}$ also nicht zu groß, so wird H_0 als richtig angenommen.

Auch hier kann es zum Irrtum kommen:

Fehler 2. Art (type II error) :

Annahme von H_0 zu Unrecht,
also obwohl H_A zutrifft.
Fehlerwahrscheinlichkeit β , hängt ab vom
wahren Erwartungswert μ . β = "*Risiko* II".
(1–β heißt *Schärfe des Tests*.)

(Beachte: Entsprechendes gilt für die Nullhypothesen $\mu \geq \mu_0$ oder $\mu = \mu_0$, die gegen H_A: $\mu < \mu_0$ bzw. $\mu \neq \mu_0$ geprüft werden.)

Übersicht:

	Entscheidung für:		
Realität:		H_0	H_A
	H_0 stimmt	richtige Entscheidung W'keit 1–α	Fehler 1. Art W'keit α
	H_A stimmt	Fehler 2. Art W'keit β	richtige Entscheidung W'keit 1–β

Allgemein gilt: Verkleinerung von α führt zu Vergrößerung von β ("Antagonismus"). Einzige Möglichkeit, β bei festem α zu verkleinern: Stichprobenumfang *n* vergrößern.

Festlegung des Ablehnungs- und des Annahmebereichs:

Man berechnet aus den Beobachtungen $x_1, ..., x_n$ eine Prüfgröße (*Teststatistik*) $T(x_1, ..., x_n)$, bei der ein großer Wert eher für H_A und ein kleinerer eher für H_0 spricht. Außerdem legt man einen kritischen Wert $c_{1-\alpha}$ fest, so daß

$$P(T > c_{1-\alpha}) \leq \alpha$$

gilt, wenn H_0 zutrifft. Dann wird entschieden:

$$T > c_{1-\alpha} \Rightarrow H_0 \text{ wird abgelehnt}$$

$$T \leq c_{1-\alpha} \Rightarrow H_0 \text{ wird akzeptiert.}$$

(Beim einseitigen Einstichproben-Gauß-Test (H_0: $\mu \leq \mu_0$) ist z.B. die Prüfgröße $T = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$ (= standardisierter Stichprobenmittelwert) und der kritische Wert $c_{1-\alpha} = u_{1-\alpha}$ = (1–α)-Quantil der Standard-Normalverteilung (Gaußverteilung).

WICHTIG:

In Statistik-Programmpaketen wie STATISTICA wird bei Hypothesentests der Wert *p* angegeben:

p = *P*(Teststatistik *T* >= aktueller Wert von *T*)
unter der Annahme der Nullhypothese.

p gibt also das maximale α-Niveau an, für das die Nullhypothese gerade noch akzeptiert werden kann:

$p < \alpha$: **H_0 ablehnen**

$p >= \alpha$: **H_0 beibehalten.**

Je kleiner *p* ist, desto signifikanter ist die Abweichung von der Nullhypothese!

Beispiel: Der 2-Stichproben-t-Test

Anwendung:
Entscheidung der Frage, ob zwei Stichproben, bei denen jeweils 1 Merkmal gemessen wird, denselben Mittelwert aufweisen (bzw. ob der eine Mittelwert größer als der andere ist).
Voraussetzungen:
Das gemessene Merkmal muss als *normalverteilt* angenommen werden (d.h. approximierbar durch eine Gaußverteilung); die Varianz dieser Normalverteilung ist unbekannt, aber in beiden Stichproben gleich.

Zweistichproben-t-Test (two-sample t-test)

$X_1, ..., X_{n_A}$ seien stochastisch unabh. $N(\mu_A, \sigma^2)$-Verteilungen, $Y_1, ..., Y_{n_B}$ seien stochastisch unabh. $N(\mu_B, \sigma^2)$-Verteilungen (entsprechend den beiden Stichproben). σ sei unbekannt, aber identisch bei beiden Verteilungen.

S_A, S_B seien die empirischen Standardabweichungen für X bzw. Y, $\bar{X}$, $\bar{Y}$ die empirischen arithmetischen Mittelwerte.

H_0: $\mu_A = \mu_B$, d.h. die beiden Verteilungen sind identisch.

Unter dieser Annahme ist die Prüfgröße

$$T = \sqrt{\frac{n_A n_B (n_A + n_B - 2)}{n_A + n_B}} \cdot \frac{\bar{X} - \bar{Y}}{\sqrt{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}}$$

t-verteilt mit $n_A + n_B - 2$ Freiheitsgraden.

(*t-Verteilung*, auch Student-Verteilung: tabellierte Verteilung, Wahrscheinlichkeitsdichte

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

darin ist n ein Parameter, die "Anzahl der Freiheitsgrade".)

$$|T| > t_{n_A+n_B-2,\ 1-\frac{\alpha}{2}} \Rightarrow H_0 \text{ wird abgelehnt.}$$

Analog kann man auch einseitig testen, d.h.
H_0: $\mu_A \leq \mu_B$, dann Ablehnung von H_0 bei

$$T > t_{n_A+n_B-2,\,1-\alpha}$$

Dabei ist $t_{n,q}$ das q-Quantil der t-Verteilung mit n Freiheitsgraden.

8.4 Die Varianzanalyse (Analysis of Variance; ANOVA)

Verallgemeinerung der Fragestellung des 2-Stichproben-t-Tests

Ziel:
Unterschiede zwischen einzelnen (Teil-) Datensätzen feststellen
statt 2 Stichproben jetzt also *p* Stichproben

Voraussetzungen:

- die für die Unterscheidung verwendeten Attribute haben numerische Werte als Ausprägungen
- diese sind in den Teil-Datensätzen jeweils normalverteilt mit gleicher Varianz (Varianzhomogenität)

die einzelnen Teil-Datensätze heißen auch *Stufen*, *Faktorstufen*, *Varianten* oder *Behandlungsarten*

Gesamtzahl der Daten in allen Stufen: $N = n_1 + \ldots + n_p$

Zeilen- (Stufen-) Mittelwerte $\bar{x}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{ik}$

Gesamt-Mittelwert $\bar{x} = \frac{1}{N}\sum_i\sum_k x_{ik} = \frac{1}{N}\sum_{i=1}^{p} n_i\bar{x}_i$

Zerlegung der Abweichung eines Beobachtungswertes vom Mittelwert:

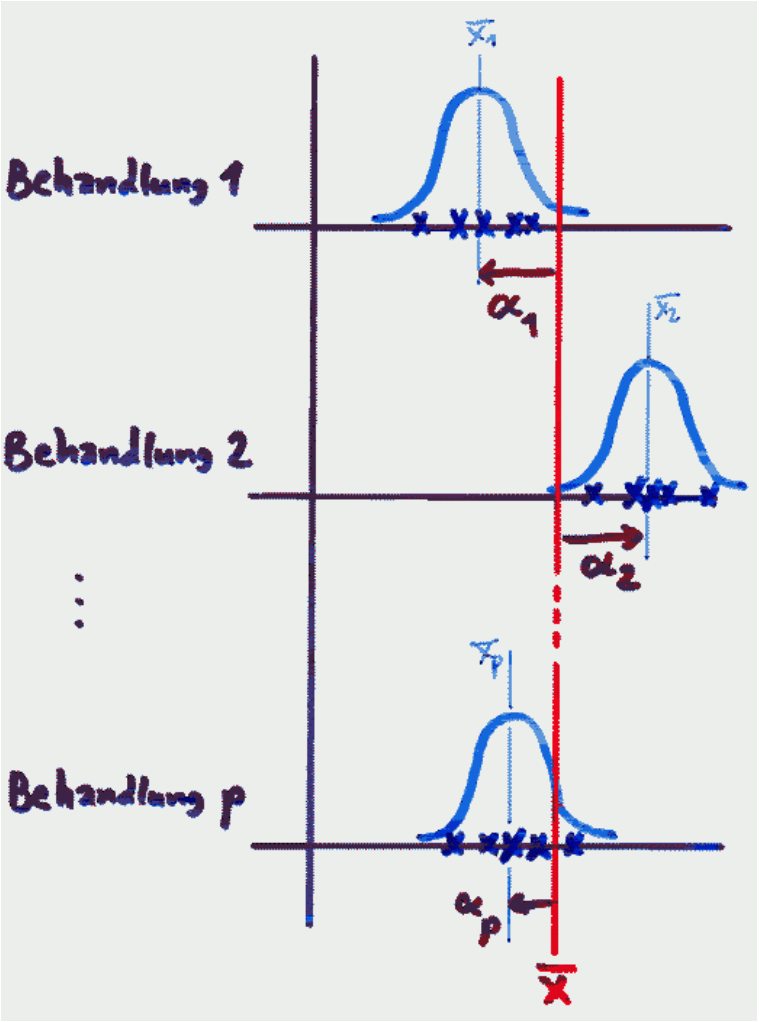
$$x_{ik} - \bar{x} = (\bar{x}_i - \bar{x}) + (x_{ik} - \bar{x}_i)$$
$$= \alpha_i + \varepsilon_{ik}$$

$\alpha_i = \bar{x}_i - \bar{x}$ = Abweichung des *i*-ten Stufenmittels vom Gesamtmittelwert
= Effekt des Einflussfaktors A (*i*-te Behandlungsart)

$\varepsilon_{ik} = x_{ik} - \bar{x}_i$ = Abweichung des Beobachtungswertes

vom zugehörigen Stufenmittel = "Restfehler". Beachte: Die Restfehler innerhalb einer Zeile summieren sich zu Null:

$$\sum_{k=1}^{n_i} \varepsilon_{ik} = \sum_{k=1}^{n_i} x_{ik} - n_i\bar{x}_i = n_i\bar{x}_i - n_i\bar{x}_i = 0$$



$x_{ij} = \bar{x}_i + \varepsilon_{ij}$
$= \bar{x} + \alpha_i + \varepsilon_{ij}$
$\varepsilon_{ij} \sim N(0; \sigma)$

Entsprechende Zerlegung der Summe der Abweichungs-quadrate
(damit auch Zerlegung der Varianz → "*Varianzanalyse*"):

$$SQ_{total} = \sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x})^2 = \sum_{i=1}^{p}\sum_{k=1}^{n_i}(\alpha_i + \varepsilon_{ik})^2$$

$$= \sum\sum \alpha_i^2 + 2\sum\sum \alpha_i \varepsilon_{ik} + \sum\sum \varepsilon_{ik}^2$$

$$= \sum_{i=1}^{p} n_i \alpha_i^2 + 2\sum_{i=1}^{p} \alpha_i \left(\sum_{k=1}^{n_i} \varepsilon_{ik} \right) + \sum_{i=1}^{p}\sum_{k=1}^{n_i} \varepsilon_{ik}^2$$

$= SQ_{Effekt} + 0 + SQ_{Fehler}$

also $SQ_{total} = SQ_{Effekt} + SQ_{Fehler}$,

der Anteil von SQ_{Effekt} misst den Einfluss des Faktors A auf die Gesamtstreuung.

$SQ_{Effekt} = SS_{effect}$ (sum of squares),

$SQ_{Fehler} = SS_{error}$ im Englischen.

Zur Vergleichbarkeit muss man normieren, d.h. durch p–1 bzw. durch N–p dividieren:

$FG_{Effekt} = df_{effect} = p$–1 = Anzahl der Freiheitsgrade "zwischen den Stufen"

$FG_{Fehler} = df_{error} = N$–$p$ = Anzahl der Freiheitsgrade "innerhalb der Stufen"
(df = ***d***egrees of ***f***reedom)

$MQ_{Effekt} = MS_{effect} = \frac{SQ_{Effekt}}{p-1}$ = mittlere Quadratsumme zwischen den Stufen,

$MQ_{Fehler} = MS_{error} = \frac{SQ_{Fehler}}{N - p}$ = mittlere Quadratsumme innerhalb der Stufen (***m***ean sum of ***s***quares).

Test auf Signifikanz der Mittelwertunterschiede:

Voraussetzung: Die Restfehler ε_{ik} sind stochastisch unabhängig und normalverteilt mit Mittelwert 0 und identischer Varianz σ^2 (*Varianzhomogenität*).

Dann folgt die Prüfgröße $F := \frac{MQ_{Effekt}}{MQ_{Fehler}}$ einer F-Verteilung mit den Parametern $p-1$ und $N-p$

(Quantile tabelliert bzw. in Statistik-Software abrufbar).

Nullhypothese H_0: $\mu_1 = \mu_2 = ... = \mu_p$, d.h. kein Effekt des Faktors *A*,
dagegen H_A: wenigstens 2 der μ_i unterscheiden sich.

$F > F_{p-1,\, N-p,\, 1-\alpha}$ (Quantil der F-Verteilung):

H_0 ablehnen.

Varianzanalyse bei mehreren Einflussfaktoren

multifactorial analysis of variance, MANOVA

bei 2 Faktoren *A*, *B* hat man:
- Effekt des Faktors *A*
- Effekt des Faktors *B*
- evtl. "Synergie-Effekt" von *A* und *B*, d.h. "Wechselwirkungs-Effekt" *AB*
- Restfehler.

$$x_{ijk} = \overline{x} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Entsprechend zerlegt man die Summe der Abweichungsquadrate SQ_{total} in 4 Komponenten SQ_A, SQ_B, SQ_{AB} und SQ_{Fehler} und berechnet analog zur einfaktoriellen Analyse die Werte MQ und F (jeweils gesondert für A, B, AB).

$F >$ krit. Wert der F-Verteilung

bedeutet einen signifikanten Einfluss des entsprechenden Faktors bzw. der Wechselwirkung.

Analog bei mehr als 2 Faktoren (mit entsprechend mehr Wechselwirkungsmöglichkeiten: AB, AC, BC, ABC...).

8.5. Die Diskriminanzanalyse

(vgl. Hartung & Elpelt 1999, S. 240 ff.)

Ziel: Klassifikation
Zuordnung von Instanzen zu Klassen
unter Verwendung einer Lernstichprobe

Voraussetzungen:

- alle Attribute sind numerische Größen
- und sind *normalverteilt* mit in allen Klassen gleicher Kovarianzmatrix

Grundgesamtheit der Objekte: Ω
p normalverteilte Attribute $Y_1, \ldots, Y_p$
m Klassen (Teilpopulationen): $\Omega = K_1 \cup K_2 \cup \ldots \cup K_m$
Erwartungswertvektor der Attribute in K_i: $\mu^{(i)}$
Kovarianzmatrix der Attribute: Σ (in allen K_i gleich)

$\mu^{(1)}, \ldots, \mu^{(m)}$ und Σ sind in der Regel unbekannt und müssen geschätzt werden.
Grundlage: Lernstichproben vom Umfang n_i $(i = 1, \ldots, m)$ aus den m Klassen. Dabei sei $n = n_1 + \ldots + n_m > p$.

y_{ik} sei der beobachtete Attributvektor am k-ten Objekt aus der i-ten Klasse

$\Rightarrow$ $\mu^{(i)}$ wird geschätzt durch

$$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik} \qquad (i = 1, \ldots, m)$$

Σ wird geschätzt durch

$$S = \frac{1}{n-m} \sum_{i=1}^{m} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)(y_{ik} - \bar{y}_i)^T .$$

Die Matrix $S_e = (n-m) \cdot S$ nennt man <u>Fehlermatrix</u>

und $S_h = \sum_{i=1}^{m} n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T$ mit $\bar{y} = \frac{1}{n} \sum_{i=1}^{m} \sum_{k=1}^{n_i} y_{ik}$

die <u>Hypothesenmatrix</u>.

Für die Klassifikation neuer Objekte wird die <u>lineare Fishersche Diskriminanzfunktion</u> verwendet:

Ein neues Objekt mit Attributvektor y wird der Klasse K_i zugeordnet, wenn für alle $j \neq i$ gilt:

$$h_{ij}(y) = (\bar{y}_i - \bar{y}_j)^T \cdot S^{-1} \cdot y - \frac{1}{2} (\bar{y}_i - \bar{y}_j)^T \cdot S^{-1} \cdot (\bar{y}_i + \bar{y}_j) > 0.$$

(Bei der Berechnung kann man verwenden: $h_{ij}(y) = -h_{ji}(y)$.)

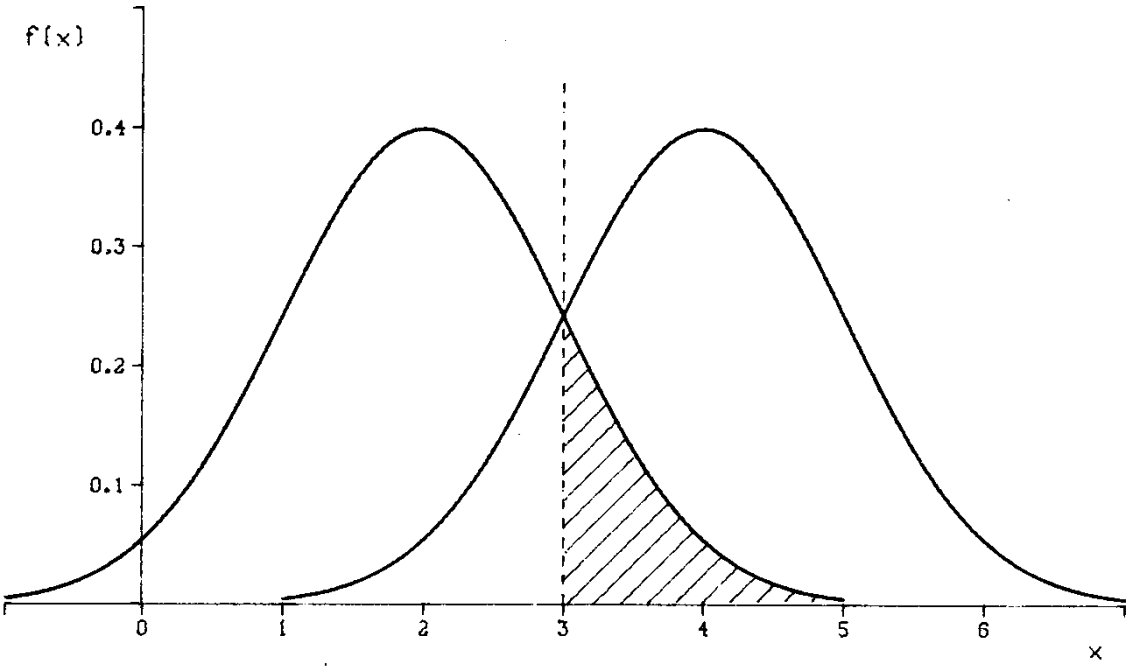


Abb.1: Veranschaulichung der Diskrimination zwischen zwei Populationen aufgrund eines N(2;1) - bzw. N(4;1) - verteilten Merkmals; die schraffierte Fläche entspricht der Wahrscheinlichkeit dafür, daß ein Objekt aus der ersten Population falsch klassifiziert wird

Ein Beispiel (aus Hartung & Elpelt 1999, S. 247 ff.):

Tab.3: Länge (L) und Breite (B) der Kelch- und Blütenblätter von je 50 Pflanzen dreier Irisarten

k	Iris setosa y_{1k}^T				Iris versicolor y_{2k}^T				Iris virginica y_{3k}^T			
	Kelch		Blüte		Kelch		Blüte		Kelch		Blüte	
	L	B	L	B	L	B	L	B	L	B	L	B
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
8	5.0	3.3	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
11	5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
12	4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	2.1
13	4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
14	4.3	3.0	1.1	0.1	6.1	2.9	4.1	1.4	5.7	2.5	5.0	2.0
15	5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
16	5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
17	5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
18	5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
19	5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
20	5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
21	5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
22	5.1	3.7	1.3	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
23	4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
24	5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
25	4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
26	5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
27	5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
28	5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
29	5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	1.6
30	4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
31	4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
32	5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
33	5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
34	5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
35	4.9	3.1	1.5	0.2	5.4	3.0	4.3	1.5	6.1	2.6	5.6	1.4
36	5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
37	5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
38	4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
39	4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
40	5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
41	5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
42	4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
43	4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
44	5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
45	5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
46	4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
47	5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
48	4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
49	5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
50	5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Tab.3 enthält die $p = 4$ Messungen von Länge und Breite der Kelch- und Blütenblätter von je $n = n_i = 50$ Pflanzen der $m = 3$ Irisarten setosa, versicolor und virginica;vgl. R.A. Fisher (1936): The use of multiple measurements in taxonomic problems, Ann. Eugen. 7, S.179-188.

Betrachtet man die Daten aus Tab.3 als Lernstichprobe, so kann mittels der Diskriminanzanalyse eine beliebige Irispflanze einer der drei Arten zugeordnet werden. Als Kriterium für die Diskrimination sollen hier nur *Länge und Breite des Kelchblattes* herangezogen werden.

Kommen für die Klassifizierung nur die Arten iris setosa und versicolor in Frage (*Zweigruppenfall*), so werden die Mittelwertvektoren durch

$$\bar{y}_1 = (5.006, 3.428)^T \quad \text{bzw.} \quad \bar{y}_2 = (5.936, 2.770)^T$$

geschätzt. Der Schätzer für die gemeinsame Kovarianzmatrix der Populationen setosa und versicolor ist

$$S = \frac{1}{100-2}((n_1-1)S_1 + (n_2-1)S_2)$$

$$= \frac{1}{98}\left(\begin{pmatrix} 5.9682 & 4.7628 \\ 4.7628 & 6.8992 \end{pmatrix} + \begin{pmatrix} 12.7939 & 4.0915 \\ 4.0915 & 4.7285 \end{pmatrix}\right)$$

$$= \frac{1}{98}\begin{pmatrix} 18.7621 & 8.8543 \\ 8.8543 & 11.6277 \end{pmatrix} .$$

Mittels dieser Schätzer ergibt sich wegen

$$S^{-1} = \begin{pmatrix} 8.153 & -6.209 \\ -6.209 & 13.156 \end{pmatrix} ,$$

daß eine neue Irispflanze mit Länge y_1 und Breite y_2 des Kelchblattes in die Population iris setosa eingeordnet wird, falls gilt:

$$-11.673y_1 + 14.431y_2 + 19.141 > 0 ,$$

und in die Art iris versicolor klassifiziert wird, falls gilt:

$$-11.673y_1 + 14.431y_2 + 19.141 < 0 .$$

Kann eine neu hinzukommende Irispflanze auch noch der Art iris virginica angehören (*Dreigruppenfall*), so muß auch die Lernstichprobe für iris virginica aus Tab.3 zur Schätzung der Parameter herangezogen werden. Mit

$$\bar{y}_3 = \begin{pmatrix} 6.588 \\ 2.974 \end{pmatrix} \qquad \text{und} \qquad S_3 = \begin{pmatrix} 0.3963 & 0.0919 \\ 0.0919 & 0.1019 \end{pmatrix}$$

ergibt sich der Schätzer für die gemeinsame Kovarianzmatrix Σ von Länge und Breite der Kelchblätter der drei Irisarten zu

$$S = \frac{1}{147}(50\,S_1 + 50\,S_2 + 50\,S_3) = \begin{pmatrix} 0.2650 & 0.0927 \\ 0.0927 & 0.1154 \end{pmatrix} .$$

Weiterhin ergibt sich

$$\bar{y} = \frac{1}{150} \sum_{i=1}^{3} \sum_{k=1}^{50} y_{ik} = \begin{pmatrix} 5.843 \\ 3.057 \end{pmatrix} .$$

Wegen

$$S^{-1} = \begin{pmatrix} 5.248 & -4.216 \\ -4.216 & 12.052 \end{pmatrix}$$

wird bei der Verwendung der linearen Fisherschen Diskriminanzfunktion eine neue Irispflanze mit Länge y_1 und Breite y_2 des Kelchblattes dann wie folgt diskriminiert. Mit

$$h_{12}(y) = -\ 7.655y_1 + 11.851y_2 + \ 5.154 \quad ,$$
$$h_{13}(y) = -10.216y_1 + 12.141y_2 + 20.359 \quad ,$$
$$h_{23}(y) = -\ 2.562y_1 + \ 0.290y_2 + 15.210$$

gehört sie zur Art

iris setosa , falls $h_{12}(y) > 0$ und $h_{13}(y) > 0$,

iris versicolor, falls $h_{12}(y) < 0$ und $h_{23}(y) > 0$,

iris virginica , falls $h_{13}(y) < 0$ und $h_{23}(y) < 0$.

Güte / Signifikanz der Trennung der Klassen (Diskrimination)?

Trennmaß $T^2(Y_1,\dots,Y_p) = \text{Spur}(S_h \cdot S_e^{-1})$.

(Spur einer quadratischen Matrix = Summe der Hauptdiagonalelemente der Matrix.)

Gilt $T^2(Y_1,\dots,Y_p)=0 \Rightarrow$ die p Attribute ermöglichen keine Trennung zwischen den Klassen

Je größer T^2, desto besser die Diskrimination.

Diskrimination signifikant zum Niveau α $\Leftrightarrow$

$$T^2(Y_1,\dots,Y_p) > c_{HL,1-\alpha}(p, n-m, m-1).$$

Dabei ist $c_{HL,1-\alpha}(p,n-m,m-1)$ das $(1-\alpha)$-Quantil der Hotelling-Lawley-Verteilung (tabelliert).

Bei 2 Klassen ($m=2$) lässt sich dieses Quantil durch ein Quantil der (geläufigeren) F-Verteilung approximieren:

$$T^2(Y_1,\dots,Y_p) > \frac{p}{n-p-1} F_{p,n-p-1;1-\alpha}.$$

Weiteres Ziel: Reduktion der Zahl der für die Trennung verwendeten Attribute.

→ dabei:
möglichst geringe Verschlechterung des Trennmaßes.

Wenn q Merkmale berücksichtigt werden sollen:
- Berechne T^2 für jede q-Auswahl der p Attribute
- Verwende diejenigen q Attribute zur Diskrimination, für die T^2 maximal ist.

Nachteil: S_h und S_e^{-1} müssen jedes Mal neu berechnet werden
→ zu großer Aufwand, wenn p groß!

→ Alternative:

<u>Schrittweise Methode der Unentbehrlichkeit</u>

(q dann am Anfang noch nicht fest)

Sei $(s_{jl})_{j,l=1,\dots,p} = S = \frac{1}{n-n} S_e$,

$(t_{jl})_{j,l=1,\dots,p} = S^{-1}$,

$A = \left[\bar{y}_1 - \bar{y}, \bar{y}_2 - \bar{y}, \dots, \bar{y}_m - \bar{y} \right]$,

$(b_{ji})_{j=1,\dots,p,\ i=1,\dots,m} = B = S^{-1} \cdot A$.

Die <u>Unentbehrlichkeit</u> U_j für das j-te Attribut $(j = 1,\dots,p)$ sei:

$$U_j = \frac{1}{(n-m) \cdot t_{jj}} \sum_{i=1}^{m} n_i b_{ji}^2 .$$

Wenn das l-te Merkmal die kleinste Unentbehrlichkeit hat und eliminiert wird, so ist das Trennmaß für die restlichen:

$$T^2(Y_1, \ldots, Y_{l-1}, Y_{l+1}, \ldots, Y_p) = T^2(Y_1, \ldots, Y_p) - U_l \quad .$$

Soll noch ein weiteres Attribut eliminiert werden, wiederholt man die Berechnung mit den aktualisierten Matrizen S und A für die restlichen Attribute.

ausführliches Beispiel: Hartung & Elpelt 1999, S. 253-257.