Business Data Mining: Statistische Grundlagen und ausgewählte Verfahren des Data Minings mit SPSS

1. Einführung und Grundlagen

1.1 Begriffsabgrenzungen und Anwendungsgebiete

1.2 Statistische Grundbegriffe

1.3 Datenerhebung

1.4 Datenaufbereitung

2. Multivariate Analysemethoden mit SPSS
(Mathematische Darstellung der Verfahren, Anwendungsgebiete und –voraussetzungen, SPSS-Beispiele)

2.1 Multiple Regressionsanalyse
(Modellformulierung, Schätzung der Regressionsfunktion, Prüfung von Funktion, Koeffizienten und Modellprämissen, Dummy-Variablen, Ergebnisinterpretation)

2.2 Mehrfaktorielle Varianzanalyse
(Modellformulierung und Modellannahmen, Streuungszerlegung, Kovarianzanalyse, Ergebnisinterpretation)

2.3 Hierarchisch-agglomerative Clusteranalyse
(Abgrenzung unterschiedlicher Verfahren, Variablenauswahl, Ähnlichkeits- und Distanzmaße, Fusionierungsalgorithmen, Bestimmung der Clusterzahl, Ergebnisinterpretation)

2.4 Weitere multivariate Verfahren im Überblick
(Diskriminanzanalyse, Faktorenanalyse, Conjoint-Analyse, Zeitreihenanalyse)

LITERATUR:

1) Backhaus, K./Erichson, B./Gensler, S./Weiber, R./Weiber, T.: Multivariate Analysemethoden – Eine anwendungsorientierte Einführung, 16. Auflage, Wiesbaden 2021.

Dieses Lehrbuch stellt die wichtigsten Analysemethoden, einschließlich der mathematisch-statistischen Grundlagen und Anwendungsgebiete, vor und zeigt ihren Einsatz mit dem Statistikprogrammpaket IBM SPSS Statistics. Zudem gibt es in den einzelnen Kapiteln zahlreiche weiterführende Literaturhinweise.

2) Brosius, F.: SPSS – Umfassendes Handbuch zu Statistik und Datenanalyse, 8. Auflage, Frechen 2018.

Dieses Lehrbuch ist vergleichbar zu Backhaus et al. mit etwas weniger Mathematik und mehr SPSS-Inhalten.

3) Han, J./Kamber, M./Pei, J.: Data Mining – Concepts and Techniques, 3. edition, Amsterdam et al. 2012.

Dieses englischsprachige Lehrbuch ist geeignet für Fortgeschrittene und geht teilweise deutlich über den Inhalt der Vorlesung hinaus.

Alle Bücher stehen auch als E-Book in der Bibliothek der DHBW Mannheim zur Verfügung.

Außerdem ist die Lektüre der in Moodle eingestellten Aufsätze zur Vorlesung obligatorisch.

1. Einführung und Grundlagen
1.1 Begriffsabgrenzungen und Anwendungsgebiete

zunehmende Digitalisierung (Scannerkassen, Transaktionen im Internet, soziale Medien, Activity Tracker, Wearables, Smart-Home-Systeme) → **Datenflut, Big Data**

Diese Datenflut erzeugt nur dann einen Mehrwert, wenn sie analysiert wird: **Daten** → **Informationen** → **Wissen**

Business Intelligence: „Sammelbegriff für den IT-gestützten Zugriff auf Informationen sowie die IT-gestützte Analyse und Aufbereitung dieser Informationen. Ziel dieses Prozesses ist es, aus dem im Unternehmen vorhandenen Wissen neues Wissen zu generieren. Bei diesem neu gewonnenen Wissen soll es sich um relevantes, handlungsorientiertes Wissen handeln, welches Managemententscheidungen zur Steuerung des Unternehmens unterstützt.“ (Lackes/Siepermann)

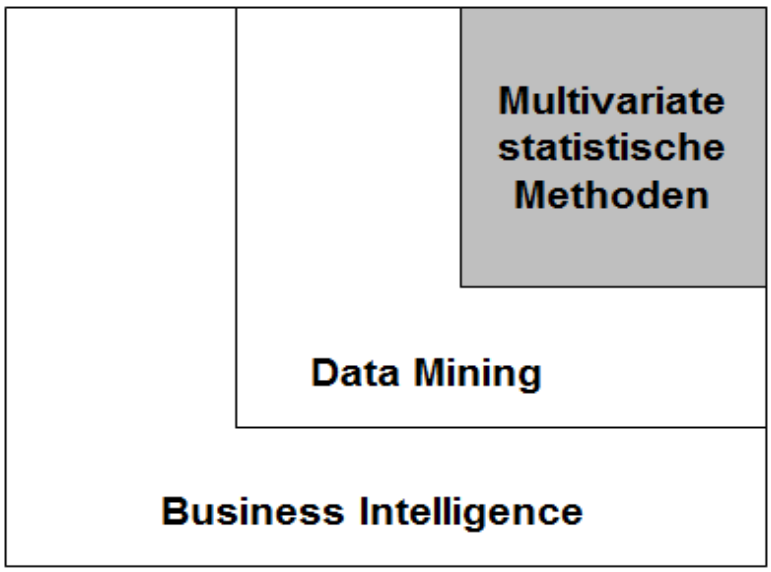
„**Business Intelligence (BI)** beschreibt die auf eine Unterstützung, Durchführung und Kontrolle betrieblicher Aktivitäten ausgerichtete Intelligenz (Einsicht) sowie die zu ihrer Erzielung eingesetzten Konzepte, Methoden und Informationssysteme.“ (Hummeltenberg)

Die Begriffe **Business Intelligence** und **Data Mining** werden nicht immer einheitlich abgegrenzt. In aller Regel wird **Data Mining als Teilgebiet von Business Intelligence** verstanden. BI umfasst zudem insbesondere noch die Bereiche Datenerfassung und Datenhaltung (z. B. Data Warehouse) sowie Wissensmanagement.

„**Data Mining** ist eine interdisziplinäre Wissenschaft, mit welcher mit verschiedenen hochkomplexen Verfahren und Methoden der Datenanalyse bisher unbekannte Informationen aus großen Datenbeständen in Datenbanken entdeckt werden können.“ (Schweizer)

„Unter **Data Mining** wird im weiteren Sinne der gesamte Prozess der Wissensentdeckung in großen Datenbeständen verstanden und im engeren Sinne nur die dabei verwendeten Analyseverfahren.“ (Chamoni)

Vor allem für Data Mining (i. w. S.) wird auch häufig der Begriff **Knowledge Discovery in Databases (KDD)** verwendet. Unter Data Mining (i. e. S.) werden verschiedene Datenanalysemethoden verstanden. Ein Großteil dieser Methoden sind **multivariate Verfahren aus der Statistik** (Regressionsanalyse, Varianzanalyse, Clusteranalyse, Diskriminanzanalyse, Faktorenanalyse usw.). Außerdem spielen Entscheidungsbaummodelle, maschinelles Lernen, neuronale Netze und Text Mining eine zentrale Rolle.



Ablauf eines Datenanalyseprozesses:

1) Aufgabendefinition:
Bestimmung der Problemstellung und Ableitung von Zielen

2) Datenerhebung/Datenverständnis:
Auswahl der relevanten Datenbestände (Primärstatistik versus Sekundärstatistik);
Datengrundlage oft Data Warehouse

3) Datenaufbereitung:
geeignetes Format;
Datensichtung über Tabellen/Grafiken (Qualität der Daten?);
Behandlung fehlender/falscher Werte sowie von Ausreißern

4) Auswahl und Anwendung von Data Mining-Methoden:
abhängig von inhaltlicher Fragestellung sowie Datenmenge und Skalierung (= **Data Mining im engeren Sinn**);
zum Großteil multivariate statistische Methoden

5) Interpretation der Ergebnisse:
Ausfiltern handlungsrelevanter Ergebnisse

6) Anwendung der Ergebnisse:
Bewertung und Umsetzung für Problemstellung

Diese Vorgehensweise orientiert sich an dem Standardmodell für Data Mining, dem **CRISP-DM** (Cross-industry standard process for data mining).

Zeitaufteilung:

In der Praxis erfordert das eigentliche Rechnen (Punkt 4.) nur etwa 10 Prozent der verwendeten Zeit. Der Großteil wird dagegen für die Datenerhebung, -auswahl und –aufbereitung verwendet.

häufigste Fehlerquellen:

1) unzureichende Datenqualität
(vor allem keine repräsentativen Daten)

2) Methoden werden in Zeiten einfach zu bedienender Softwareprogramme falsch angewendet oder falsch interpretiert.

Wichtige Anwendungsgebiete des Data Minings:

- Handel (z. B. CRM)
- Banken (z. B. Kreditvergabe)
- Versicherungen (z. B. Risikoanalyse)
- Pharmaindustrie (z. B. Medikamentenwirkungsforschung)
- Industrie (z. B. Fehleranalyse bei Fertigungsprozessen)
- Markt- und Meinungsforschung (z. B. politische Umfragen)
- Kriminalitätsbekämpfung (z. B. Rasterfahndung)
- Stadt-, Regional- und Länderanalysen in der VWL

1.2 Statistische Grundbegriffe

Statistische Einheit/Merkmalsträger:
Objekt, dessen Eigenschaften festgestellt werden sollen

Statistische Masse/Grundgesamtheit:
Gesamtheit der Merkmalsträger

Stichprobe/Teilgesamtheit:
Teil der Elemente der Grundgesamtheit

Panel:
Daten werden regelmäßig immer wieder bei der gleichen Stichprobe erhoben zwecks Längsschnittanalysen

Merkmal:
bestimmte messbare Eigenschaft eines Merkmalsträgers

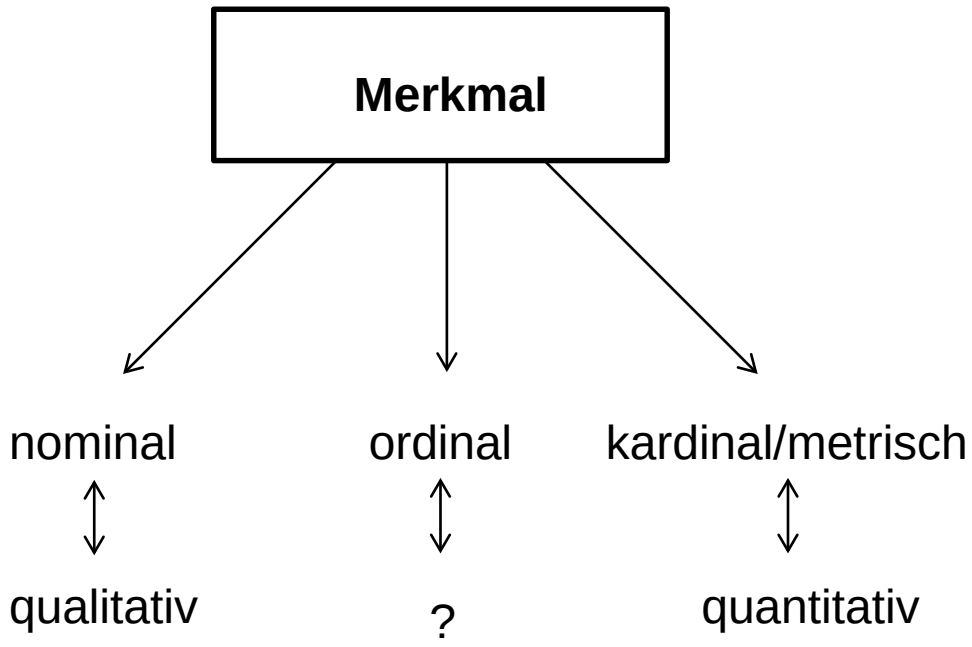
qualitative Merkmale:
nur kategoriale Merkmalsausprägung

quantitative Merkmale:
sinnvolle Zuordnung von reellen Zahlen möglich, Unterscheidung zwischen **diskret** und **stetig**

Nominalskala:
nur Verschiedenartigkeit kommt zum Ausdruck

Ordinalskala:
Verschiedenartigkeit und Rangordnung, **in der Praxis oft als kardinal und somit als quantitativ behandelt**

Kardinalskala/metrische Skala:
Verschiedenartigkeit und Rangordnung sowie Quantifizierbarkeit der Unterschiede mit Unterscheidung zwischen Intervallskala (willkürlicher Nullpunkt) und Verhältnisskala/Ratio-Skala (absoluter Nullpunkt)



Testverfahren:
Test- oder Prüfverfahren beruhen auf der Stichprobentheorie. Bei diesen Verfahren wird mit Hilfe einer **Zufallsstichprobe** geprüft, ob bestimmte Hypothesen über die Grundgesamtheit richtig oder falsch sind. Hierzu stellt man eine **Nullhypothese H_0** und eine **Alternativhypothese** auf. Es sind vier Fälle möglich:

Entscheidung aufgrund der Stichprobe	**wahrer Zustand der Grundgesamtheit, der allerdings unbekannt ist**	
	H_0 trifft zu	H_0 trifft nicht zu
H_0 wird nicht abgelehnt	richtige Entscheidung	β- Fehler (Fehler 2. Art)
H_0 wird abgelehnt	α – Fehler (Fehler 1. Art)	richtige Entscheidung

Es sollen nun möglichst **keine falschen Entscheidungen** getroffen werden, wobei bei statistischen Tests in aller Regel der **α-Fehler** im Mittelpunkt steht.

Für die Entscheidung werden geeignete **Prüfgrößen** festgelegt, **Signifikanzniveaus** vorgegeben (meist 1%, 5% oder 10%) und berechnet, ob die Prüfgrößen bestimmte, tabellierte Werte überschreiten.

In den meisten Statistikprogrammpaketen (auch in SPSS) wird zur Vereinfachung direkt eine **Wahrscheinlichkeit für den Fehler 1. Art** (α – Fehler) berechnet. Dieses berechnete Signifikanzniveau wird auch als **p-value** bezeichnet.

Es gibt zwei Möglichkeiten bei der Entscheidung:

Fall 1:

von SPSS berechnetes Signifikanzniveau (z. B. 2%)
≤
vom Nutzer festgelegtes Signifikanzniveau (z. B. 5%)

→ **Nullhypothese wird abgelehnt**

Fall 2:

von SPSS berechnetes Signifikanzniveau (z. B. 12%)
>
vom Nutzer festgelegtes Signifikanzniveau (z. B. 5%)

→ **Nullhypothese wird nicht abgelehnt**

1.3 Datenerhebung

Sofern nicht in ausreichendem Maße und in ausreichender Qualität Sekundärdaten vorliegen, müssen Primärdaten erhoben werden. Dies kann über eine **Vollerhebung oder Teilerhebung (= Stichprobe)** geschehen.

Vorteile einer Vollerhebung:

- Erfassung aller statistischen Einheiten mit großer regionaler und fachlicher Vielfalt und sehr umfang-reichen Auswertungsmöglichkeiten
- kein Stichprobenfehler
- Basis für Stichproben in der Folgezeit

Nachteile einer Vollerhebung:

- hoher Zeit- und Kostenaufwand
- Datenverarbeitungsproblematik trotz EDV
- in manchen Bereichen nicht durchführbar
- Ergebnisse können schlechter sein als bei einer Stich-probe, da sehr viele Interviewer benötigt werden, die dann oft nicht ausreichend geschult sind

Bei der Datenerhebung kommt es häufig zu **statistischen Fehlern**:

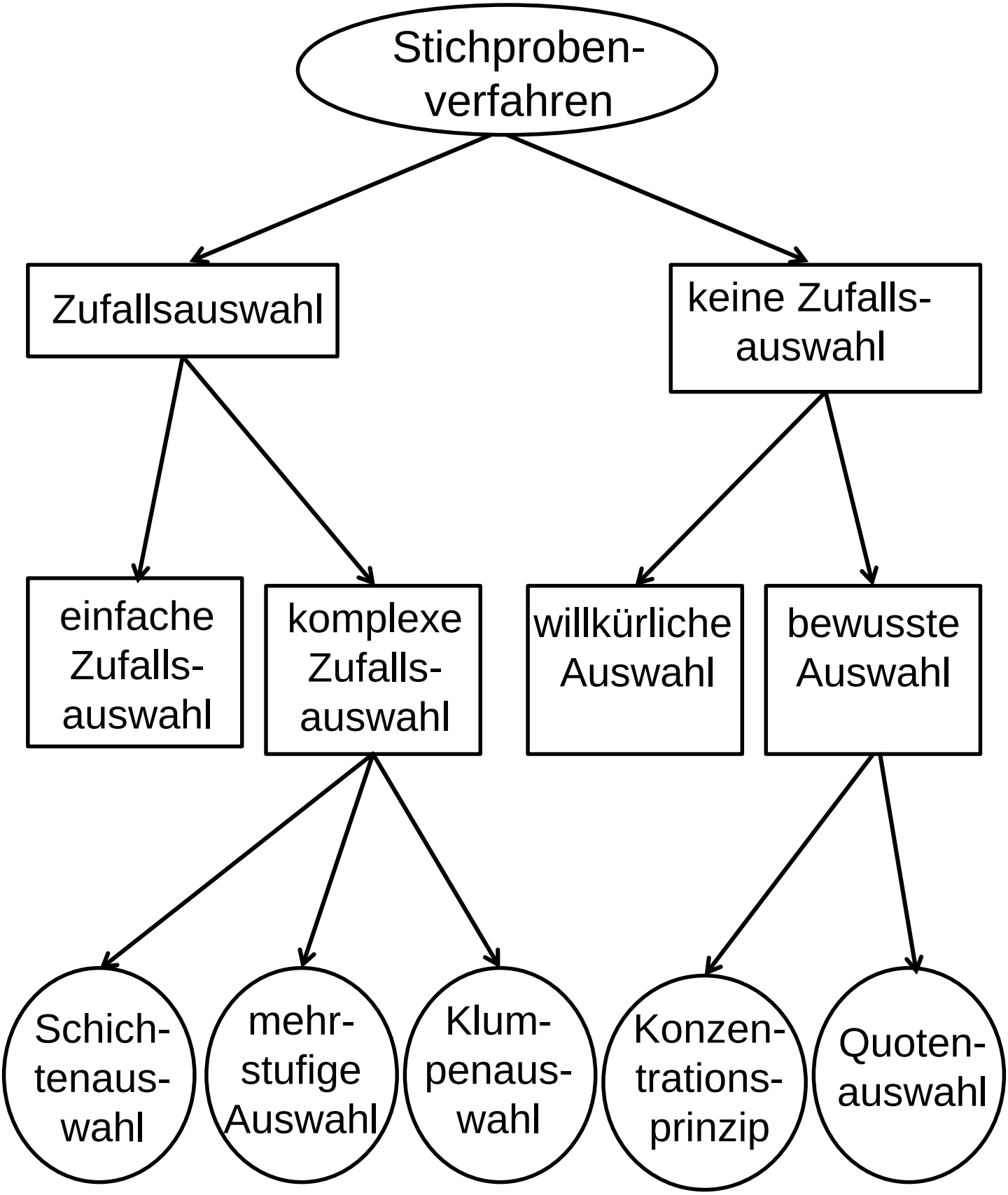
- **systematische (stichprobenfremde) Fehler**:
  Fehler, die sowohl in jeder Stichprobe als auch bei Vollerhebungen auftreten können

- **Stichprobenfehler**:
  resultieren aus **fehlender Repräsentativität** der Stichprobe (nicht möglich bei Vollerhebung);
  → Minimierung durch **geeignete Auswahlverfahren**

Ziel eines **Stichprobenverfahrens** ist es, eine möglichst große **Repräsentativität** herzustellen, um so die Ergebnisse auf die Grundgesamtheit übertragen zu können.
Oft werden Stichproben in zwei Teilstichproben geteilt.
→ **Lerndaten/Trainingsdaten** und **Testdaten**

Dies gelingt mit den verschiedenen **Verfahren der Zufallsauswahl** und mit Einschränkungen auch mit den **Verfahren der bewussten Auswahl**. Dagegen ist eine willkürliche Auswahl (Befragung „des Mannes auf der Straße“) unter statistischen Gesichtspunkten sehr problematisch.

Die Verfahren der Zufallsauswahl erlauben im Unterschied zu den Verfahren der bewussten Auswahl die Anwendung des gesamten **wahrscheinlichkeitstheoretischen Instrumentariums der Statistik** (z. B. Bestimmung von Konfidenzintervallen). Allerdings ist eine Zufallsauswahl **aus praktischen Gründen häufig nicht durchführbar**.


Stichproben-
verfahren
Zufallsauswahl
keine Zufalls-
auswahl
einfache
Zufalls-
auswahl
komplexe
Zufalls-
auswahl
willkürliche
Auswahl
bewusste
Auswahl
Schich-
tenaus-
wahl
mehr-
stufige
Auswahl
Klum-
penaus-
wahl
Konzen-
trations-
prinzip
Quoten-
auswahl

Verfahren der Zufallsauswahl:

1. einfache Zufallsauswahl:
Jeder Merkmalsträger der Grundgesamtheit hat die gleiche Chance, in die Stichprobe einbezogen zu werden. Dies entspricht letztendlich dem **Urnenmodell**.

2. Schichtenauswahl:
Die Grundgesamtheit wird in **Schichten** zerlegt, die in sich **möglichst homogen** und untereinander heterogen sind. Dann wird aus jeder Schicht eine Zufallsstichprobe gezogen und anschließend werden die Ergebnisse zusammengesetzt.

3. mehrstufige Auswahl:
Die Grundgesamtheit wird in einzelne in sich heterogene Klumpen zerlegt, die jeweils ein **verkleinertes Abbild der Grundgesamtheit** darstellen. Mittels Zufallsauswahl werden dann ein oder mehrere Klumpen bestimmt. Anschließend wird dann noch einmal innerhalb des Klumpens eine Zufallsauswahl durchgeführt.

4. Klumpenverfahren:
Dieses Verfahren verläuft zunächst ähnlich wie eine mehrstufige Auswahl. Allerdings werden dann innerhalb eines Klumpens alle Objekte befragt bzw. untersucht.

All diese Verfahren erlauben den Einsatz des **wahrscheinlichkeitstheoretischen Instrumentariums der Statistik**. Sie scheitern aber häufig daran, dass **keine vollständigen Listen** aller Objekte vorliegen.

Verfahren der bewussten Auswahl:

1. Auswahl nach dem Konzentrationsprinzip (Abschneideverfahren):
Die Stichprobe wird auf „besonders ins Gewicht fallende Fälle“ beschränkt.

2. Quotenauswahl:
Bei der Befragung werden bestimmte Auflagen = Quoten hinsichtlich der zu Befragenden in einem **Quotenplan** festgelegt, um damit eine möglichst hohe Repräsentativität zu erreichen. Der Interviewer kann dann aber seine Erhebungseinheiten selbst auswählen.

Die Verfahren der bewussten Auswahl sind meist **einfacher anzuwenden** als die Verfahren der Zufallsauswahl. Sie erlauben allerdings strenggenommen nicht den Einsatz des wahrscheinlichkeitstheoretischen Instrumentariums der Statistik. Unter bestimmten Anwendungsvoraussetzungen liefern sie aber trotzdem gute Ergebnisse.

Völlig ungeeignet ist dagegen eine **willkürliche Auswahl**, bei der ohne jede Vorüberlegungen irgendwelche (meist einfach zu erreichende) Personen befragt werden.

Die Aussagen zu den verschiedenen Stichprobenverfahren gelten nicht nur für Befragungen, sondern auch für andere Erhebungsformen.

Befragungstechniken:
Grundsätzlich können Daten auch aus Beobachtungen, Experimenten oder automatischer Erfassung (z. B. mit Scannerkassen oder über das Internet) gewonnen werden. Für die Datengewinnung hat aber die Befragung nach wie vor eine große Bedeutung.

Je nach Thematik kommen unterschiedliche Befragungsformen zum Einsatz. Man unterscheidet zwischen

- **mündlicher** Befragung
- **telefonischer** Befragung als Spezialform der mündlichen Befragung
- **schriftlicher** Befragung
- Befragung über das **Internet** (E-Mail oder Einstellen von Fragebögen) als Spezialform der schriftlichen Befragung

Nichtbeantwortungsproblem:
Bei Befragungen wird in aller Regel ein beachtlicher Teil der Befragten die Antwort verweigern. Dies ist vor allem dann problematisch, wenn zwischen der Fragestellung und denjenigen, die die Antwort verweigern, ein Zusammenhang besteht (→ **fehlende Repräsentativität**).

Mögliche **Gegenmaßnahmen** sind

- Pilotstudien
- Umformulieren/Weglassen bestimmter Fragen
- Wechseln der Befragungstechnik
- Erhöhung des Stichprobenumfangs
- Befragung von Ersatzpersonen (bei Quotenauswahl)

Stichprobenumfang:
Dieser hängt ab von der **gewünschten Genauigkeit**, der **Streuung** des gesuchten Merkmals in der Grundgesamtheit und der **Größe der Grundgesamtheit** (allerdings nur schwacher Einfluss). Je genauer das Ergebnis sein soll, je mehr das Merkmal streut und je größer die Grundgesamtheit ist, desto größer muss der Stichprobenumfang sein.

Er kann für die **Verfahren der Zufallsauswahl** über wahrscheinlichkeitstheoretische Berechnungen bestimmt werden. Dies geht bei den Verfahren der bewussten Auswahl nicht.

Allerdings ist zu berücksichtigen, dass die Streuung oft nicht bekannt ist bzw. dass bei einer Befragung Merkmale mit verschiedenen Streuungen erfragt werden.

In der Praxis bestimmen oft **Zeit und Kosten** den Stichprobenumfang.

Faustregel: Der Stichprobenumfang sollte auf keinen Fall unter 30 liegen. Bei seriösen Analysen und großen Grundgesamtheiten sollte der Stichprobenumfang auf alle Fälle zumindest im hohen dreistelligen Bereich liegen.

Beispiel: ZDF-Politbarometer der Forschungsgruppe Wahlen mit jeweils 1.000 bis 1.500 Befragten

Fragebogentheorie:
In aller Regel enthält ein Fragebogen **Identifikationsfragen** (z. B. Geschlecht, Alter), **Informationsfragen** (Gegenstand der Erhebung) und evtl. **Kontrollfragen** (Überprüfung der Konsistenz der Antworten: z. B. Fragen nach dem Zeitaufwand für den Arbeitsweg und Kontrollfrage nach der km-Zahl).

Überlegungen bei der Fragebogenerstellung:

- Länge des Fragebogens (Zusammenhang mit Nichtbeantwortungsproblem!)
- optische Aufmachung des Fragebogens (inkl. Begleitschreiben) → übersichtlich und seriös
- Reihenfolge der Fragen → systematisch und möglichst zunächst einfache Fragen
- Formulierung der Fragen → kurz und gut verständlich
- Neutralität der Fragen und Antwortmöglichkeiten
- offene, geschlossene oder als häufig guter Kompromiss halboffene Fragen
- Pilotstudie mit Testpersonen
- Hinweise auf Anonymität/rechtliche Grundlagen/Untersuchungsziel im Begleitschreiben

1.4 Datenaufbereitung

Die Aufbereitung der Daten ist in vielen Fällen der zeitaufwendigste Schritt im gesamten Data Mining-Prozess. Dabei sind verschiedene Teilschritte zu berücksichtigen:

1. Standarddatenformat erstellen:

Erstellung von Datentabellen für **Querschnittsanalysen** (verschiedene Merkmalsträger zu einem Zeitpunkt) oder **Längsschnittanalysen** (ein Merkmalsträger zu unterschiedlichen Zeitpunkten)

2. Datensichtung:

Datenanreicherung oder Datenreduktion notwendig?

erste **deskriptive Analysen** (Häufigkeitstabellen, Lageparameter, einfache Grafiken) zur Beurteilung der **Qualität der Daten**

3. Behandlung fehlender Merkmalswerte:

Weglassen (Datensätze oder Merkmale?) oder Auffüllen (durch Nacherhebung oder Schätzung) oder Codierung als „unbekannt“?

4. Behandlung falscher Merkmalswerte:

Welche Werte sind tatsächlich falsch und welche sind **Ausreißer**?

Orientierung an Wertebereichen, Plausibilitätsüberlegungen und Standardabweichungen

Korrigieren oder Ausschließen?

5. Datentransformationen:

Umskalierung, um bestimmte statistische Verfahren durchzuführen

Klassenbildung, um unzureichende Besetzung einzelner Werte zu verhindern und aussagekräftigere Ergebnisse zu bekommen

Normierung und Standardisierung, um Vergleichbarkeit unterschiedlicher Merkmale herzustellen

z. B. **z-Transformation**

$$z_i = \frac{x_i - \bar{X}}{S}$$

$$z_i = \frac{x_i - \mu}{\sigma}$$

2. Multivariate Analysemethoden mit SPSS
2.1 Multiple Regressionsanalyse

Während bei der **Korrelationsanalyse** die **Stärke des Zusammenhangs** zwischen unterschiedlichen Merkmalen überprüft wird, geht die **Regressionsanalyse** einen Schritt weiter. Bei ihr wird die **Art der funktionalen Abhängigkeit** berücksichtigt. So hat z. B. die Sonnenscheindauer einen Einfluss auf den Ernteertrag in der Landwirtschaft, nicht aber umgekehrt. In manchen Fällen sind die Abhängigkeiten nicht so eindeutig, sodass **lag-Strukturen** berücksichtigt werden müssen.

einfache Regressionsanalyse: nur eine erklärende (= unabhängige) Variable

multiple Regressionsanalyse: mehrere erklärende Variablen

Die Regressionsanalyse verlangt sowohl für die abhängige als auch für die unabhängigen Variablen **metrisches Skalenniveau** (Ausnahme: **Dummy-Variablen** = binäre unabhängige Variablen, mit denen auch nominal skalierte Merkmale dargestellt werden können). In den meisten Fällen geht man von **linearen Abhängigkeiten** aus.
→ multiple lineare Regressionsanalyse

Die Schätzung der Regressionsfunktion erfolgt meist über die KQ-Methode (**Methode der kleinsten Quadrate**). Ziel der Schätzung ist es, den (quadrierten) Abstand zwischen realisierten Werten und geschätzten Werten, also die **Restgrößen (= Residuen)**, möglichst gering zu halten. Je kleiner diese Abweichungen sind, desto „besser“ ist die Regressionsgerade.

Wichtigste **Ursachen für (große) Residuen** sind die fehlende Berücksichtigung wichtiger Einflussgrößen, das Vorhandensein von nicht-linearen Zusammenhängen sowie Mess- und Auswahlfehler (fehlende Repräsentativität der Daten).

einführendes Beispiel: 10 Lebensmittelläden in einer Stadt Besteht ein Zusammenhang zwischen der Verkaufsfläche X (in 1.000 qm) und dem Jahresumsatz Y (in Mio. Euro)?

x_i	0,5	0,9	1,1	1,5	1,2	1,4	1,6	0,8	1,0	0,4
y_i	3,0	5,1	5,5	7,3	6,2	7,0	8,1	4,9	6,1	3,2

$$\text{Jahresumsatz}_i = \beta_0 + \beta_1 \bullet \text{Verkaufsfläche}_i$$

→ Wenn die Verkaufsfläche um 1.000 qm steigt, dann erhöht sich der Umsatz um β_1 Millionen Euro.

Vorgehensweise bei einer Regressionsanalyse:

1. Modellformulierung
2. Schätzung der Regressionsfunktion
3. Prüfung der Regressionsfunktion
4. Prüfung der Regressionskoeffizienten
5. Prüfung der Modellprämissen

zu 1. **Modellformulierung:**

(ökonomisch) plausibler Ursache-Wirkungszusammenhang;
Streudiagramme als Entscheidungshilfe;
nicht zu viele Variablen berücksichtigen

zu 2. **Schätzung der Regressionsfunktion:**

KQ-Methode;
mehrere Schätzdurchgänge mit Berücksichtigung unterschiedlicher Variablenkombinationen;
plausible Koeffizienten (Vorzeichen, absolute Größe)?

zu 3. **Prüfung der Regressionsfunktion:**

Determinationskoeffizient = Bestimmtheitsmaß R^2:
Maßzahl für die Qualität einer Regression, bei der zwischen erklärten und nicht-erklärten Abweichungen getrennt wird;
Wertebereich von 0 bis 1, bei 1 perfekte Erklärung;
bei multipler Regression zum Vergleich korrigiertes R^2

F-Statistik:
Ist das geschätzte Modell allgemein auf die Grundgesamtheit übertragbar und somit signifikant?
Test der Nullhypothese $\beta_1 = \beta_2 = \ldots \beta_i = 0$ (keinerlei Zusammenhang);
Ausweis eines Signifikanzniveaus

Standardfehler der Schätzung:
misst mittleren Fehler der berechneten Regressionsfunktion; entspricht Standardabweichung der Residuen (Differenz aus Beobachtungswerten zu Prognosewerten);
sollte deutlich unter 20% der zu schätzenden Y-Werte liegen

zu 4. **Prüfung der Regressionskoeffizienten:**

t-Test:
ähnlich F-Test, aber Prüfung der einzelnen Koeffizienten;
Nullhypothese $\beta_j = 0$ (kein Zusammenhang zwischen der getesteten unabhängigen Größe und der abhängigen Größe);
Ausweis eines Signifikanzniveaus

Konfidenzintervall für die Regressionskoeffizienten:
Mit einer Wahrscheinlichkeit von x Prozent liegt der entsprechende Regressionskoeffizient für die Grundgesamtheit in einem bestimmten Intervall (besonders problematisch, wenn das Intervall den Wert 0 überdeckt).

standardísierte Regressionskoeffizienten:
Die Variable mit dem größten absoluten Wert hat den stärksten Einfluss auf die zu erklärende Größe.

zu 5. **Prüfung der Modellprämissen**:
Nur wenn die Annahmen des linearen Regressionsmodells (z. B. keine Multikollinearität, keine Autokorrelation) erfüllt sind, liefert die KQ-Methode gute Schätzergebnisse.

keine Multikollinearität: Zwischen den erklärenden Variablen darf keine lineare Abhängigkeit bestehen.

Folge von Multikollinearität: unplausible Schätzwerte

Messung über **Varianzinflationsfaktor VIF**;
Faustregel: sollte in allen Fällen << 10 sein

Lösung des Multikollinearitätsproblems: Entfernen von Variablen aus dem Modell

keine Autokorrelation: Residuen sollten nicht miteinander korreliert sein. Diese Modellverletzung spielt vor allem bei Zeitreihendaten eine Rolle. Bei Querschnittsdaten muss für die Überprüfung eine systematische Ordnung vorliegen.

Folge von Autokorrelation: Tests und Konfidenzintervalle nicht mehr vernünftig interpretierbar

Messung über **Durbin-Watson-Test** (Wertebereich 0 bis 4);
Faustregel: zwischen 1,5 und 2,5 in Ordnung (<< 1,5: positive Autokorrelation, >> 2,5: negative Autokorrelation)

Lösung des Autokorrelationsproblems: Berücksichtigung zusätzlicher Variablen oder andere funktionale Form der Regressionsgleichung

Exkurs: **Dummy-Variablen**

Generell versteht man unter einer Dummy-Variable eine **binäre Variable**, die nur die Werte 0 und 1 annimmt.

Anwendungsgebiete für Dummy-Variablen:

- Viele Probleme bei den Modellannahmen sind durch **Strukturbrüche** bedingt. Die Schätzung der linearen Regressionsgerade führt hier zu unbefriedigenden Ergebnissen. Als Lösung bietet sich der Einsatz einer Dummy-Variablen an, mit der sowohl **Niveauänderungen** als auch **Steigungsänderungen** modelliert werden können.

- Über eine solche Dummy-Variable können auch **nominal skalierte Merkmale** als erklärende Größen in die Regressionsanalyse aufgenommen werden. Hat die nominal skalierte Variable m unterschiedliche Ausprägungen, so müssen m-1 Dummy-Variablen verwendet werden. Die Zahl der Dummy-Variablen sollte nicht zu groß sein. Als **Alternative** zu einer Regressionsanalyse mit Dummy-Variablen bietet sich bei einer gemischten Skalierung der erklärenden Größen möglicherweise die **Kovarianzanalyse** an.

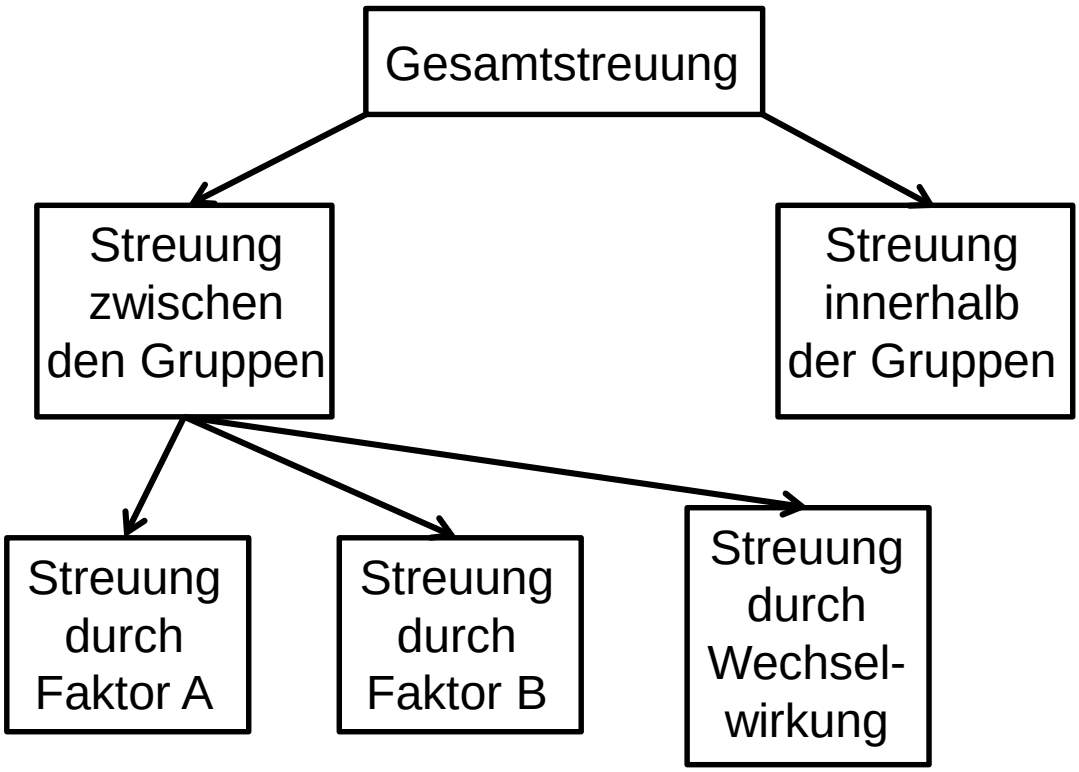
2.2 Mehrfaktorielle Varianzanalyse

Die Varianzanalyse untersucht wie die Regressionsanalyse den **Einfluss von unabhängigen Variablen auf abhängige Variablen**. Im Unterschied zur Regressionsanalyse besitzen die unabhängigen Variablen (= **Faktoren**) allerdings **nominales Skalenniveau**, während die abhängige Variable wie bei der Regressionsanalyse **metrisches Skalenniveau** hat.

Mit einem **ANOVA- Modell** (analysis of variance) könnte man z. B. untersuchen, ob das Geschlecht einen Einfluss auf die Lebenserwartung hat. Da in diesem Beispiel mit dem Geschlecht nur ein Faktor betrachtet wird, spricht man von einer **einfaktoriellen Varianzanalyse**.

Bei mehreren erklärenden Faktoren handelt es sich um eine **mehrfaktorielle Varianzanalyse** (z.B. Geschlecht, Raucher, Vegetarier).

Der Grundgedanke einer Varianzanalyse besteht in der **Zerlegung der Gesamtstreuung in eine Streuung innerhalb der Gruppen und in eine Streuung zwischen den Gruppen (= Faktorkombinationen)**. Die Streuungen werden über Varianzen gemessen und dann in Beziehung zueinander gesetzt. Über **F-Tests** wird die Signifikanz getestet. Bei einer mehrfaktoriellen Varianzanalyse können zusätzlich noch **Wechselwirkungen** (= Interaktionsbeziehungen) zwischen den erklärenden Größen auftreten.



Die Varianzanalyse ist ein relativ **robustes Verfahren**. Sind die unterschiedlichen Faktorkombinationen (= Gruppen) ungefähr **gleich häufig besetzt**, ist sie auch bei Verletzung der Modellannahmen anwendbar. Die Modellannahmen sind vor allem die **Normalverteilung** der Ausgangswerte sowie die **Varianzhomogenität** in den verschiedenen Gruppen. Sie müssen nur bei unterschiedlicher Besetzung der Faktorkombinationen überprüft werden.

Überprüfung der Normalverteilung:
Histogramm mit Normalverteilungskurve oder Kolmogorov-Smirnov-Test (Nullhypothese: Normalverteilung)

Überprüfung der Varianzhomogenität:
Levene-Test (Nullhypothese: Gleichheit der Fehlervarianzen der abhängigen Variable über alle Gruppen = Homogenität)

Vorgehensweise bei der Varianzanalyse:

1. Problemformulierung
2. Analyse der Abweichungsquadrate
3. Prüfung der statistischen Unabhängigkeit

(Überprüfung der Modellannahmen Normalverteilung und Varianzhomogenität nur bei ungleicher Besetzung der Faktorkombinationen)

zu 1. **Problemformulierung:**

sinnvoller (ökonomischer) Zusammenhang unter Berücksichtigung des Skalenniveaus und der Modellannahmen;

nur Haupteffekte oder auch Wechselwirkungen zwischen den erklärenden Größen?

zu 2. **Analyse der Abweichungsquadrate:**

Welcher Anteil der Gesamtstreuung wird durch die einzelnen Faktoren bzw. durch ihre Wechselwirkung erklärt?
→ Streuungszerlegung

Determinationskoeffizient = Bestimmtheitsmaß R^2:
Maßzahl für die Qualität der Varianzanalyse, bei der zwischen erklärten und nicht-erklärten Abweichungen getrennt wird;
Wertebereich von 0 bis 1, bei 1 perfekte Erklärung

zu 3. **Prüfung der statistischen Unabhängigkeit:**

Gibt es einen signifikanten Einfluss der einzelnen Faktoren auf das Untersuchungsmerkmal?

F-Tests:
Test der Nullhypothese, dass das Modell/der untersuchte Faktor keinen Einfluss auf die abhängige Variable hat;
Ausweis eines Signifikanzniveaus

Scheffe-Test:
paarweiser Mittelwertvergleichstest zur Feststellung, welche Faktorkategorien tatsächlich signifikant unterschiedliche Ergebnisse für das Untersuchungsmerkmal liefern;
Nullhypothese jeweils: keine Unterschiede;
Ausweis von Signifikanzniveau und Konfidenzintervall

Exkurs: **Kovarianzanalyse**

Bei manchen Fragestellungen ist es sinnvoll, zusätzlich zu den nominal skalierten Faktoren **metrisch skalierte unabhängige Variablen** zu berücksichtigen. Diese werden als **Kovariaten** bezeichnet. Man spricht dann von einer Kovarianzanalyse. Um den Einfluss dieser metrisch skalierten Variablen zu überprüfen, wird für diese in aller Regel eine Regressionsanalyse vorgeschaltet.

Eine Alternative zu der Kovarianzanalyse bei gemischtem Skalenniveau der erklärenden Variablen kann eine **Regressionsanalyse mit Dummy-Variablen** sein.

2.3 Hierarchisch-agglomerative Clusteranalyse

Unter dem Begriff Clusteranalyse fasst man verschiedene Verfahren zur **Gruppenbildung** zusammen. Aus einer heterogenen Gesamtheit von Objekten oder Personen sollen **homogene Teilmengen** gebildet werden. Die Mitglieder einer Gruppe sollen sich **möglichst ähnlich** sein, während zwischen den Gruppen deutliche Unterschiede bestehen sollen.

Der Begriff Clusteranalyse umfasst eine Vielzahl von unterschiedlichen Verfahrens- und Vorgehensweisen:

- **partitionierende Clusterverfahren**: bestimmte Clusterzahl und Gruppierung vorgegeben und dann Umgruppierung
- **hierarchisch-divisive Clusterverfahren**: beim Start des Verfahrens sind alle Objekte in einem Cluster, dann Zerlegung
- **hierarchisch-agglomerative Clusterverfahren** (in der Vorlesung!): beim Start bildet jedes Objekt einen eigenen Cluster, dann Zusammenfassen

Im **Unterschied zur Diskriminanzanalyse** werden die Gruppen erst im Laufe des Verfahrens aus den Merkmalsausprägungen der betrachteten Variablen gebildet. Clusteranalysen lassen sich bei unterschiedlichen Skalenniveaus durchführen. Allerdings sind die Vorgehensweisen je nach Skalenniveau etwas unterschiedlich.

Vorgehensweise bei der Clusteranalyse:

1. Variablenauswahl
2. Bestimmung der Ähnlichkeiten bzw. Unterschiede
3. Auswahl des Fusionierungsalgorithmus
4. Bestimmung der Clusterzahl
5. Interpretation der Cluster

zu 1. **Variablenauswahl:**

Die Variablen sollten für die Fragestellung auch relevant sein. Liegt zwischen einzelnen Merkmalen eine **sehr hohe Korrelation** (Korrelationskoeffizient > 0,9) vor, so kann es sinnvoll sein, eine dieser Variablen aus der Analyse auszuschließen. Die Informationen dieser Variablen werden dann auch durch die andere Variable geliefert. Eventuell kann es sich aber auch um eine gewünschte **Gewichtung** besonders wichtiger Informationen handeln.

zu 2. **Bestimmung der Ähnlichkeiten bzw. Unterschiede:**

Über **Proximitätsmaße** wird die Ähnlichkeit (**Ähnlichkeitsmaß**) bzw. die Distanz (**Distanzmaß**) zwischen den Objekten hinsichtlich der **ausgewählten Variablen** quantifiziert. Es gibt unterschiedliche Maße je nach Skalenniveau.

Bei **gemischter Skalierung** ist das niedrigste Skalenniveau maßgeblich. Höher skalierte Variablen müssen daher herabgestuft werden, was zu einem Informationsverlust führt.

Beispiel: metrisch skaliertes Merkmal Preis wird zum nominal skalierten Merkmal Preisklasse (teuer/billig)

Ein **Ähnlichkeitsmaß** für **binäre/nominal** skalierte Merkmale ist der **M-Koeffizient** (= Simple Matching-Koeffizient). Hierbei werden binäre Variablen und ihre Ausprägungen für die einzelnen Objekte verglichen. Die Zahl der Übereinstimmungen wird in Relation zu allen betrachteten Variablen gesetzt. Hat ein nominal skaliertes Merkmal n Ausprägungen, so werden zu seiner Darstellung n-1 binäre Variablen benötigt.

Ein wichtiges **Distanzmaß** für **metrisch** skalierte Merkmale ist die **quadrierte Euklidische Distanz**. Sie ergibt sich für zwei Objekte k und l und J Variablen nach folgender Formel:

$$d_{k,l}^{\,2} = \sum_{j=1}^{J} \left| x_{kj} - x_{lj} \right|^2$$

Im Unterschied zu vielen anderen Distanzmaßen werden große Unterschiede zwischen den Objekten bei einzelnen Variablen besonders stark gewichtet.

Bei der Bestimmung der Proximitätsmaße ist darauf zu achten, dass die Daten in vergleichbaren Maßeinheiten vorliegen (evtl. **Standardisierung** z.B. in z-Werte nötig).

$$z_{kj} = \frac{x_{kj} - \overline{x_j}}{s_j}$$

(z-Variable hat Mittelwert 0 und Standardabweichung 1)

zu 3. **Auswahl des Fusionierungsalgorithmus:**

Fragestellung: Nach welcher Methode sollen die verschiedenen Objekte zu Gruppen zusammengefasst werden?

Es gibt auch hier sehr unterschiedliche Möglichkeiten. Von großer Bedeutung sind die **hierarchisch-agglomerativen Verfahren**. Hierbei geht man zunächst davon aus, dass alle Objekte jeweils eine eigene Gruppe bilden und fasst diese dann nach und nach zusammen. Bei n Objekten hat man nach n-1 Schritten nur noch einen Cluster. Je nach Fusionierungsalgorithmus können beim Ablauf des Clusterverfahrens **unterschiedliche Gruppen** entstehen:

A. **Single-Linkage-Verfahren** = Nearest-Neighbour-Verfahren:
   kleinste Distanz bei Gruppenbildung relevant;
   entdeckt sehr gut Ausreißer

B. **Complete-Linkage-Verfahren** = Furthest-Neighbour-Verfahren:
   größte Distanz bei Gruppenbildung relevant

C. **Average-Linkage-Verfahren** = Linkage zwischen den Gruppen
   durchschnittliche Entfernung aller möglichen Fallpaare aus zwei Clustern ist für Gruppenbildung relevant

D. **Ward-Verfahren**:
   vereint die Objekte so zu Gruppen, dass sich die Varianzen innerhalb der Gruppen möglichst wenig erhöhen;
   häufig verwendet, aber problematisch bei Ausreißern;
   nur durchführbar mit Distanzmaßen

zu 4. **Bestimmung der Clusterzahl:**

Die betrachteten hierarchisch-agglomerativen Verfahren bilden zunächst bei n Objekten auch n Cluster und reduzieren die Clusterzahl dann in n-1 Schritten bis hin zu einem Cluster. Es gilt nun die **optimale Clusterzahl** zu bestimmen. Diese ist nicht nur von mathematisch-statistischen Kriterien, sondern häufig auch von der konkreten (ökonomischen) Fragestellung (z. B. bei Marketingaktionen) abhängig.

Bildet man **sehr wenige Cluster**, ist zwar die Gruppenzahl überschaubar, die Gruppen sind in sich aber sehr heterogen.

Bildet man **sehr viele Cluster**, so sind zwar die Gruppen in sich sehr homogen, aber ihre Anzahl ist zu unüberschaubar.

Unter statistischen Gesichtspunkten ist bei der Frage der Clusterzahl die **Entwicklung der Heterogenitätsmaße** beim Fortschreiten des Fusionierungsprozesses zu berücksichtigen. Große Sprünge deuten darauf hin, dass mit weiterer Fusionierung (= abnehmender Clusterzahl) die Heterogenität in den einzelnen Gruppen stark zunimmt.

In SPSS bietet sich als Hilfe für diese Fragestellung u. a. die grafische Lösung über ein **Dendrogramm** mit normiertem Heterogenitätsmaß an.

zu 5. **Interpretation der Cluster:**

Die gebildeten Gruppen sollten abschließend mittels deskriptiver Statistiken sinnvoll interpretiert werden, um geeignete Maßnahmen für die jeweilige Fragestellung auszuwählen.

2.4 Weitere multivariate Verfahren im Überblick

a) Diskriminanzanalyse:

Die Diskriminanzanalyse ist ein multivariates Verfahren zur **Analyse von Gruppenunterschieden** (z. B. bei Kreditwürdigkeitsprüfungen). Man hat zumindest zwei Gruppen (diejenigen, die ihren Kredit zurückgezahlt haben und diejenigen, die das nicht getan haben) und will nun anhand historischer Datenbestände wissen, ob sich diese Gruppen **signifikant voneinander unterscheiden** und **welche Variablen** zur Unterscheidung zwischen den Gruppen am besten geeignet sind. Wenn eine klare Unterscheidung gelingt, können anschließend **neue Elemente** (hier: neue Kreditnehmer) einer der Gruppen zugeordnet werden.

Die **Gruppen sind nominal skaliert**, während die **Merkmalsvariablen metrisch skaliert** sein müssen. Es wird somit die Abhängigkeit einer nominal skalierten Variablen (= Gruppe) von metrisch skalierten Variablen betrachtet. Somit unterscheidet sich die Diskriminanzanalyse auch hinsichtlich der Skalierung von der Regressions- und der Varianzanalyse.

Im Unterschied zur Clusteranalyse sind die **Gruppen schon vorgegeben**. Die Gruppierung ergibt sich meist aus dem Sachzusammenhang, kann aber auch aus einer vorgeschalteten Clusteranalyse stammen, bei der Gruppen erzeugt wurden.

Folgende **Anforderungen** sollten bei der Durchführung einer Diskriminanzanalyse berücksichtigt werden:

Zahl der Objekte/Fälle >> (Soll) Zahl der Variablen > (Soll) Zahl der Gruppen > (Muss) Zahl der Diskriminanzfunktionen

Vorgehensweise bei einer Diskriminanzanalyse:

1. Definition der Gruppen
   ergibt sich meist aus Fragestellung

2. Formulierung der Diskriminanzfunktion(en)
   Auswahl geeigneter metrisch skalierter Variablen nach (ökonomischen) Plausibilitätsüberlegungen

3. Schätzung der Diskriminanzfunktion(en)

4. Prüfung der Diskriminanzfunktion(en)
   mittels Klassifikationsmatrix (Trefferquote muss besser sein als bei zufälliger Zuordnung) und Kennzahlen wie Eigenwert oder Wilks Lambda

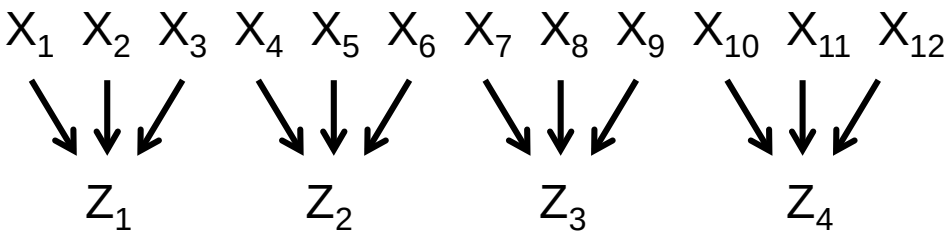
5. Prüfung der Merkmalsvariablen
   Welche Variablen sind besonders wichtig für die Gruppenbildung?

6. Klassifikation neuer Elemente
   nur sinnvoll, wenn Diskriminanzfunktion(en) ausreichend Erklärungswert haben

b) Faktorenanalyse:

Hinter der Faktorenanalyse verbirgt sich eine Reihe von Verfahren (z.B. Hauptkomponentenanalyse, Hauptachsen-analyse), bei denen es darum geht, viele einzelne (in aller Regel metrisch skalierte) Beobachtungsmerkmale zu einer kleinen Zahl hypothetischer Variablen zu verdichten. Es handelt sich somit um **Verfahren zur Datenreduktion**.

De facto geht es darum, aus sehr vielen Variablen X_i mittels Linearkombinationen einige wenige fiktive Variablen Z_i zu erzeugen, um damit dann z.B. eine Variable Y zu erklären. Die Zusammenhänge sollen dadurch übersichtlicher dargestellt werden.



→ zu viele X_i → zu unübersichtlich für weitere Analysen

→ Für stark korrelierte X-Variablen wird jeweils eine gemeinsame Hintergrundvariable Z (= **Faktor**) hergeleitet, die die verschiedenen X-Variablen repräsentiert.

$Z_1 = \alpha_1 \bullet X_1 + \alpha_2 \bullet X_2 + \alpha_3 \bullet X_3$

$Z_2 = \alpha_4 \bullet X_4 + \alpha_5 \bullet X_5 + \alpha_6 \bullet X_6$ usw.

Beispiele:

1) Bei der **Intelligenzmessung** werden Personen anhand vieler Tests zu unterschiedlichen Fragestellungen bewertet (z. B. Gruppierung von Symbolen, Erkennung von Ähnlichkeiten, Wortschatz, Satzbau, logische Ableitungen, mathematisches Grundverständnis usw.). Diese Vielzahl von Variablen kann z. B. auf die drei fiktiven Variablen bildliche, verbale und mathematische Fähigkeiten reduziert werden. Diese drei fiktiven Variablen erklären dann den IQ.

2) Bei der **Länderanalyse/Regionalanalyse/Städteanalyse** (z. B. für Förderprogramme oder Standortentscheidungen) werden häufig zahlreiche Daten erhoben (z. B. BIP/Kopf, Einkommen/Kopf, Steuereinnahmen, Zahl der Arbeitsplätze, Arbeitslosenquote, Ausbildungsquote, Alphabetisierung, Internetdichte, Mietpreise, Kinderbetreuungsmöglichkeiten, Säuglingssterblichkeit, Umweltqualität), um eine Charakterisierung vorzunehmen. Diese vielen Variablen können häufig auf wenige fiktive Variable wie wirtschaftliche Leistungsfähigkeit, soziales Umfeld usw. reduziert werden.

Probleme bei der Faktorenanalyse:

1) Informationsverlust, wenn aus mehreren Variablen eine fiktive Variable konstruiert wird

2) Interpretation der fiktiven Faktoren teilweise problematisch

3) sehr unterschiedliche Methoden der Faktorenanalyse, sodass nur für sehr erfahrene Anwender geeignet

c) Conjoint-Analyse:

Unter einer Conjoint-Analyse versteht man verschiedene Verfahren, die vor allem im Marketing bei der **Produktgestaltung**, der **Planung neuer Produkte** sowie der **Preisgestaltung** eingesetzt werden. Hierbei wird der Einfluss einzelner Merkmale auf den Gesamtnutzen des Produkts bestimmt.

Beispiele:

neues Pkw-Modell (Farbe, Sitzheizung, Airbag, Schiebedach, Bezug Sitze usw.)

neuer Laptop (Leistungsfähigkeit, Ausstattung mit Laufwerken/Anschlüssen, Garantieleistungen)

Vorgehensweise:

1. Auswahl der relevanten Eigenschaften
   - müssen realisierbar sein
   - keine KO-Kriterien

2. Festlegung eines Erhebungsdesigns
   Es werden **Stimuli** definiert (= Kombination von verschiedenen Eigenschaftsausprägungen).
   - Kombinationen sollten realistisch sein
   - Begrenzung auf max. 20 (→ **reduziertes Design**)

3. Bewertung der Stimuli
   Die Stimuli werden von Testpersonen (potenziellen Kunden) in eine Reihenfolge gebracht (→ **ordinale Reihung**).

4. Ermittlung der Teilnutzenwerte
   Aus der ordinalen Reihung werden dann **kardinale Teilnutzenwerte** für die verschiedenen Eigenschaftsausprägungen ermittelt. Es lassen sich dann Gesamtnutzenwerte für alle Stimuli ermitteln.

5. Aggregation der Nutzenwerte
   Zusammenfassung der Bewertung aller Befragungsteilnehmer

Probleme bei der Conjoint-Analyse:

1) Bei sehr vielen Eigenschaften mit einer großen Zahl von Merkmalsausprägungen wird die Beurteilung durch die Befragungsteilnehmer oft problematisch. Hat man bspw. sechs Eigenschaften mit jeweils drei unterschiedlichen Ausprägungen, so ergeben sich $3^6 = 729$ verschiedene Merkmalskombinationen. Hier muss mit **reduzierten Designs** (maximal 20 Varianten eines Produkts) gearbeitet werden, um nicht die Testpersonen zu überfordern, denn diese müssen die Varianten in eine Reihenfolge bringen. Das reduzierte Design sollte das vollständige Design möglichst gut repräsentieren.

2) Unter mathematisch-statistischen Gesichtspunkten ist die **Umrechnung ordinaler Reihenfolgen in kardinale Teilnutzenwerte** problematisch, denn es handelt sich um eine „verbotene Umskalierung von unten nach oben“. Dies ist nur zulässig, wenn man von gleichen Abständen bei der Reihenfolge ausgeht, also wenn der Abstand zwischen Platz 1 und 2 genauso groß ist wie der Abstand zwischen Platz 15 und 16.

d) Zeitreihenanalyse:

Unter der Zeitreihenanalyse wird eine Vielzahl von Verfahren zur Analyse der historischen **Entwicklung von Variablen im Zeitablauf** und ihrer Prognose zusammengefasst.

Beispiele:

- Wetter- bzw. Temperaturentwicklung in den letzten 100 Jahren
- Herzschlagentwicklung eines Komapatienten über 24 Stunden
- Entwicklung des Umsatzes oder Gewinns eines Unternehmens über 40 Quartale
- Entwicklung der Verkaufszahlen eines Produkts über 40 Quartale
- Entwicklung von Aktienkursen, Aktienindizes oder Wechselkursen über 52 Wochen (→ **Charttechnik**)

Komponenten von ökonomischen Zeitreihen:

- Trendkomponente (sehr langfristig)
- Konjunkturkomponente (mehrjährig)
- Saisonkomponente (unterjährig)
- Zufallskomponente (singuläre Einflüsse)

Um zu erkennen, welche Komponenten von Bedeutung sind und das richtige Zeitreihenmodell (z.B. linear oder exponentiell) auszuwählen, sollte die Zeitreihe zunächst **grafisch** dargestellt werden.

Ausgewählte Verfahren der reinen Zeitreihenanalyse:

1. **Regressionsanalyse** mit Zeit als erklärender Variable

   $$Y_t = \beta_0 + \beta_1 \bullet t$$

   Das Regressionsmodell kann um weitere Variablen erweitert werden. Eine Prognose ist möglich.

2. **Methode der gleitenden Durchschnitte**
   rein vergangenheitsbezogene Durchschnittsbildung und damit Glättung über mehrere Perioden

3. **Exponentielles Glätten**
   Vergangenheitsdaten erhalten mit zunehmender Aktualität ein höheres Gewicht. Eine Prognose ist möglich.

4. **Box-Jenkins-Modelle** (ARIMA-Modelle)
   Zeitreihe wird aus sich selbst und der Entwicklung der Störgrößen erklärt. Eine Prognose ist möglich.

Bei einer **reinen Zeitreihenanalyse** wird die Entwicklung der Zeitreihe nur aus ihrer historischen Entwicklung erklärt. Andere Einflussfaktoren werden nicht berücksichtigt. Dies ist **aus ökonomischer Sicht oft unbefriedigend**, da daraus keine Handlungsempfehlungen abgeleitet werden können.

Beispiel:
Charttechnik versus Fundamentalanalyse zur Erklärung der Entwicklung eines Aktienkurses

Übungsaufgaben:

1) Welche der folgenden Aussagen sind richtig (R), welche sind falsch (F)?
   Ein kardinal bzw. metrisch skaliertes Merkmal
   a) sind Stundenlöhne.
   b) sind Schulnoten.
   c) ist gleichzeitig ein qualitatives Merkmal
   d) ist gleichzeitig ein quantitatives Merkmal
   e) ermöglicht sowohl Aussagen zur Verschiedenartigkeit und Rangordnung als auch eine Quantifizierung der Unterschiede zwischen den Merkmalsausprägungen.

2) Welche der folgenden Aussagen sind richtig (R), welche sind falsch (F)?
   Bei einem statistischen Test
   a) ist der wahre Zustand der Grundgesamtheit unbekannt.
   b) kommt es zu einem α – Fehler, wenn die Nullhypothese nicht abgelehnt wird, obwohl sie falsch ist.
   c) kommt es zu einem β – Fehler, wenn die Nullhypothese nicht abgelehnt wird, obwohl sie falsch ist.
   d) wird aufgrund einer Stichprobe überprüft, ob die Nullhypothese richtig oder falsch ist.
   e) sollte die ausgewählte Stichprobe aus einer Quotenauswahl stammen.

3) Was verstehen Sie unter einer Regressionsanalyse?
   Gehen Sie auch darauf ein,
   - was mit diesem Verfahren bezweckt werden soll,
   - wo es in der Praxis Anwendung findet,
   - wie die grundsätzliche Vorgehensweise ist.

4) Ein großes Warenhaus verfügt aus „dunklen Quellen“ über eine umfangreiche Datensammlung über seine Kunden. Zudem liegen natürlich auch Daten zu den Lieferanten und zur eigenen Unternehmensentwicklung vor. All diese Daten möchte das Warenhaus mittels multivariater statistischer Verfahren auswerten. Welches Verfahren ist jeweils geeignet?

a) Bildung von homogenen Kundengruppen mittels der Merkmale Alter, Einkommen, Vermögen und Konsumausgaben des letzten Jahres

b) Zuordnung neuer Kunden auf Basis der Merkmale und Gruppeneinteilung aus Teilaufgabe a.

c) Analyse der Kundenausgaben im Warenhaus in Abhängigkeit vom Alter, Einkommen und Vermögen der Kunden

d) Analyse der Kundenausgaben im Warenhaus in Abhängigkeit vom Schulabschluss, Familienstand und Geschlecht

e) Bildung von homogenen Kundengruppen mittels der Merkmale Geschlecht, Schulabschluss und Familienstand

f) Analyse der Umsatzentwicklung in den letzten Jahren in Abhängigkeit von den Investitionsausgaben und den Werbeausgaben

g) Bildung homogener Klassen von Lieferanten über die Merkmale Umsatz, Zahl der bezogenen Produkte und Alter der Geschäftsbeziehung

5) Interpretieren Sie die beigefügten SPSS-Ausdrucke („BeispieleÜbungsaufgaben.pdf“) zu den verschiedenen multivariaten Analyseverfahren. Es handelt sich um unterschiedliche Daten für 32 verschiedene Länder zu den Themenfelder Arbeitsmarkt, Fußball und Bevölkerung.

a) Regressionsanalyse (Ausdrucke 1-2):
Die Jugenderwerbslosenquote wird erklärt durch die Erwerbslosenquote und das BIP je Einwohner.

b) Regressionsanalyse (Ausdrucke 3-4):
Die Fußballerquote (prozentualer Anteil der Fußballspieler an der Bevölkerung) wird erklärt durch die Zahl der Vereine, die Weltranglistenpunkte und die Anzahl der WM-Teilnahmen.

c) Varianzanalyse (Ausdrucke 5-7):
Die Fußballerquote (prozentualer Anteil der Fußballspieler an der Bevölkerung) wird erklärt durch den Kontinent und die WM-Spielbilanz bei allen WMs (deutlich positiv, ausgeglichen, deutlich negativ).

d) Clusteranalyse (Ausdrucke 8-12):
Klassifizierung der Länder nach Bevölkerungsdaten (Anteil unter 15 Jahren, Anteil über 65 Jahren, Geburten je Frau, Lebenserwartung Männer, Lebenserwartung Frauen)