Glossar für statistische Analysetechniken[1]

Dimitri Prandner[1], Christopher Etter[2], Christoph Glatz[3] und Martin Ulrich[4]
[1] JKU – Johannes Kepler Universität Linz
[2] PLUS – Paris Lodron Universität Salzburg
[3] Universität Graz
[4] Universität des Saarlands
Kontakt: Dimitri.Prandner@jku.at

Abstract (dt.):
Das Glossar vermittelt Grundlagenwissen zu den statistischen Analysetechniken, die im Sammelband zum Einsatz kamen. Es gibt einen kurzen Überblick über zentrale Begriffe der Statistik, Analysetechniken, die zur Auswertung quantitativer Umfragen verwendet werden, und erklärt, wie einzelne Kennzahlen interpretiert werden können.
Schlüsselwörter: Univariate Datenanalyse, Bivariate Datenanalyse, Regressionsanalyse, Faktorenanalyse, Strukturgleichungsmodelle, Clusteranalyse

Abstract (eng.):
The glossary introduces basic knowledge about statistic procedures that were used in this edited volume. It gives a brief overview of central statistical terms, statistical techniques which are used in survey research and gives insights how to interpret statistical coefficients.
Keywords: Univariate data analysis, bivariate data analysis, regression analysis, factor analysis, structural equation modelling, cluster analysis

1 Grundlegende Informationen zum Glossar[2]

Dieser Sammelband bietet in 15 Kapiteln einen sozialwissenschaftlichen Einblick, wie die österreichische Bevölkerung die Corona-Krise erlebt hat. Die Erkenntnisse basieren auf Umfragedaten, die zwischen 2020 und 2022 im Rahmen der Values-in-Crisis-Studie in drei Onlineumfragen erhoben wurden. Es kam ein nicht-probabilistisches Quotenstichprobenverfahren zum Einsatz, um jeweils ca. 2000 Personen zu unterschiedlichen Themen zu befragen. Mehr als ein Drittel der Proband*innen nahmen an allen Umfragen teil und können deshalb auch im Sinne einer Längsschnittana-

1 Dimitri Prandner hat das Glossar koordiniert und ist korrespondierender Autor, die Listung der anderen Autoren erfolgt in alphabetischer Reihenfolge.
2 Dieser Text basiert in Teilen auf dem Glossar des Buches „Die österreichische Gesellschaft während der Corona Pandemie“ (Glatz et al.,2022).

lyse analysiert werden. Diese Form der Analyse ist in der Vergangenheit aufgrund von hohem Zeit- und Ressourcenaufwand selten durchgeführt worden (Bryman, 2016, 57). Während der Corona-Pandemie wurden in Österreich jedoch mit dem Austrian Corona Panel Project (ACPP) und der Values-in-Crisis-Studie (VIC) mehre Umfragestudien mit Panel-Charakter durchgeführt, die eine Längsschnittanalyse ermöglichen und es erlauben, Veränderungen in Einstellungen und Werten von einzelnen Personen zu beobachten.

Die Auswertung von Umfragedaten erfolgt in der Regel nach strikten Regeln, die sich in einem ersten Schritt hinsichtlich der Anzahl an benötigten Variablen unterscheiden lassen. So gibt es univariate, bivariate und multivariate Analysetechniken.

Die univariate Analyse konzentriert sich auf die Analysen der Informationen, die mittels einzelner Variablen gewonnen werden. Im Rahmen univariater Auswertungsverfahren wird die Häufigkeit bestimmter Antworten ermittelt, zentrale Tendenzmaße wie Durchschnitt oder Median berechnet und die Variation innerhalb einer einzelnen Variablen erfasst. Die bivariate Analyse hingegen untersucht, in welcher Beziehung die Informationen von zwei Variablen stehen. Dieser Ansatz erlaubt es, Zusammenhänge (Korrelationen) oder Unterschiede zwischen den Datenpunkten zweier Variablen aufzuzeigen. Bivariate Analyseverfahren helfen, Muster der Abhängigkeit oder Unabhängigkeit zwischen den Variablen festzustellen. In einer multivariaten Analyse werden dagegen Zusammenhänge zwischen drei oder mehr Variablen betrachtet. Hierbei werden fortgeschrittene statistische Techniken eingesetzt, um beispielsweise komplexe Beziehungen mittels Regressionsanalysen oder Strukturgleichungsmodellen zu erforschen, inhaltliche Dimensionen zu reduzieren und latente Strukturen aufzudecken. Diese Analyseform ermöglicht es, gleichzeitig mehrere Einflussfaktoren zu berücksichtigen und somit ein umfassenderes Verständnis der Daten zu erlangen.

In diesem Glossar werden wir Schlüsselbegriffe und -techniken beleuchten, die für diese Analyseformen zentral sind. Ziel des Glossars ist es, dass Sie die Begrifflichkeiten nachschlagen können, die in den inhaltlichen Kapiteln vorkommen, um die Informationen von Tabellen oder Grafiken nachvollziehen zu können.

2 Vom Fragebogen zur Datenanalyse: Fragen, Items und Variablen

Neben den grundlegenden Umfragedaten (sprich den Antworten der Interviewteilnehmer*innen) beinhalten Datensätze typischerweise zwei weitere Aspekte, nämlich Metadaten und Paradaten. Metadaten bieten generelle Informationen über die jeweilige Umfrage, z.B. über die Durchführungsmodalität, den Zeitraum der Umfrage, die Stichprobenziehung oder den inhaltlichen Schwerpunkt (Ruggles, 2018). Paradaten beinhalten Informationen, die sich im Verlauf der Durchführung ergeben, z.B. die Dauer der Beantwortung einer Frage, eine Selbstauskunft über den Interviewort (oder sogar GPS-Koordinaten), Informationen zu dem Gerät, mit dem eine Online-Umfrage ausgefüllt wurde oder auch, ob bei einem persönlichen Interview eine dritte Person anwesend war oder nicht (Kreuter, 2013; Prandner, 2019; West, 2011).

Diese Informationen können für unterschiedliche Analysen genutzt werden, wobei Paradaten meist der Qualitätskontrolle dienen (siehe Prandner & Seymer in diesem Band), während die Antworten der Teilnehmer*innen auf bestimmte Fragen (oder Items) für inhaltliche Analysen herangezogen werden.

All die Daten, die im Rahmen einer Umfrage erhoben werden, werden in der Regel in Zahlen übersetzt („codiert"), womit sie zu Variablen werden, welche für die statistische Analyse herangezogen werden konnen. Bei Umfragedaten ist dieser Prozess meist aufgrund von geschlossenen Fragen wenig komplex, weil bereits vorgefertigte Antwortkategorien bestehen. So stellt beispielsweise die Frage „Wie stolz sind Sie darauf, Bürger dieses Landes zu sein?" ein Item dar, aus dem sich z.B. die Variable „Nationalstolz" ableiten lässt, das mittels vier vorgefertigter Antwortkategorien „1" (sehr stolz) bis „4" (gar nicht stolz) beantwortet werden kann. Diese Kategorisierung wird als *Skalierung* bezeichnet und beträgt im Beispiel 1-4 (siehe Abbildung 1 und Tabelle 1).

Diese Daten der Umfrageteilnehmer*innen (der Stichprobe) werden dann für die gewünschten Analysen aufbereitet – z.B. zusammengefasst oder aufgeteilt – und anschließend mittels der gewünschten statistischen Verfahren analysiert. In der Regel wird das Ziel verfolgt, von der Stichprobe auf die Grundgesamtheit zu schließen. Bei statistischen Analysen wird dies üblicherweise mittels einer sogenannten Signifikanzprüfung durchgeführt (siehe Tabelle 1 „Signifikanz").

Bezüglich der Aussagekraft der Ergebnisse müssen jedoch immer der Hintergrund der Datenerhebung sowie die Qualität der Daten berücksich-

tigt werden, da potenzielle Verzerrungen die Ergebnisse verfälschen könnten.

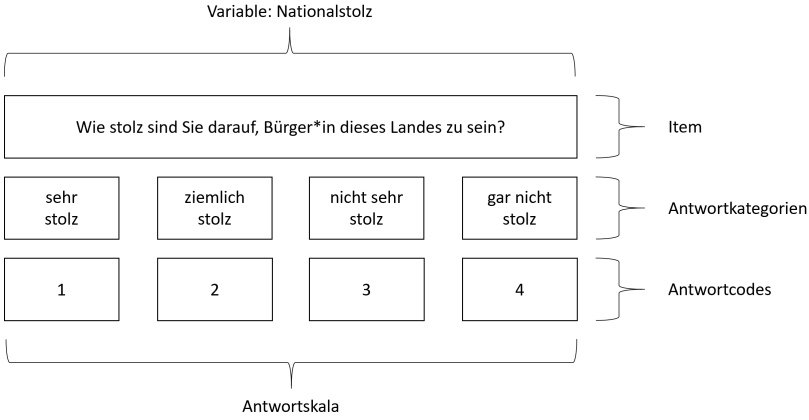

Abbildung 1: Beispielitem für die Grundbegriffe der statistischen Auswertung von Umfrageprogrammen

Tabelle 1: Grundbegriffe der statistischen Auswertung von Umfrageprogrammen

Antwortcodes	Numerischer Ausdruck der Antwortkategorien (z.B. „1“, „2“, „3“, „4“).
Antwortkategorien	Die Antwortmöglichkeiten (z.B. „sehr schlecht“, „schlecht“, „gut“, „sehr gut“).
Gewichtung	Um die Grundpopulation (hier die österreichische Bevölkerung) so gut wie möglich abzubilden, gibt es Gewichtungsvariablen im Datensatz, sodass die Stichprobe in bestimmten Merkmalen (üblicherweise Alter, Geschlecht, Bildung und Bundesland) der gewünschten Grundpopulation entspricht.
Items	Fragen in Umfrageprogrammen an die Teilnehmer*innen (Bsp. „Wie geht es Ihnen?“).
Metadaten	Metadaten beschreiben den Entstehungskontext und Hintergrund von Umfragedaten (z.B. Stichprobenverfahren, Zeitraum der Umfrage etc.)
Paradaten	Bei Umfragen entstehen neben konkret mittels Fragen erhobenen Umfragedaten auch sogenannte Paradaten durch den Erhebungsprozess an sich (z.B. Wie lange braucht eine Person, um die Umfrage auszufüllen? etc.)

Signifikanz (p-Wert < .05)	Die Signifikanz wird mittels des p-Wertes überprüft. Dieser beschreibt, nach dem Prinzip der Falsifizierbarkeit, die Wahrscheinlichkeit, dass es *keinen* Effekt bzw. Einfluss der jeweiligen Variable gibt. Je geringer der p-Wert, desto wahrscheinlicher ist daher die Annahme, dass es tatsächlich einen Effekt gibt (z.B. Unterschiede in der Impfbereitschaft zwischen älteren und jüngeren Personen etc.). Üblicherweise wird von einem signifikanten Effekt gesprochen, wenn die Wahrscheinlichkeit, dass kein Effekt in der Grundgesamtheit besteht, kleiner als 5% ist (p < .05).
Skalierung	Bandbreite der Antwortkategorien, in diesem Beispiel 1-4.
Umfragedaten	Die Daten, die durch die Antworten von Befragten auf Umfragefragen (Items) entstehen.
Variable	Abgeleitete und kodierte Information aus dem Item für die weiterführende Analyse (z.B. „aktuelles Wohlbefinden").

3 Begrifflichkeiten der univariaten und bivariaten Analyse

Einfache und einführende Analysetechniken sind meist uni- sowie bivariate Analysetechniken. Ein Beispiel wäre die Auswertung der Frage, wie viel Vertrauen man in die Wissenschaft in Österreich hat. Die Darstellung univariater Ergebnisse erfolgt häufig in Tabellen oder in Diagrammen (siehe Abbildung 2).

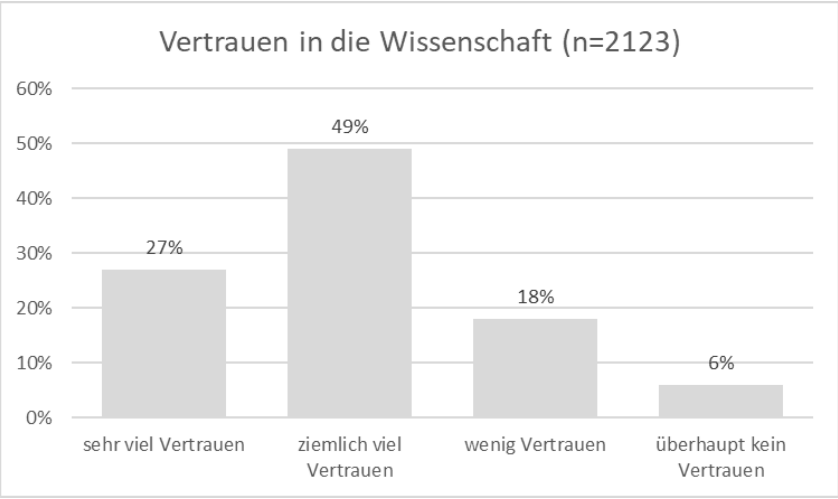

Abbildung 2: Zustimmung zur Frage, wie viel Vertrauen man in die Wissenschaft in Österreich habe (Datengrundlage: VIC 3. Welle, gewichtete Daten)

Eine bivariate Analyse betrachtet zwei Variablen gemeinsam, entweder mittels Zusammenhangs- oder Unterschiedsanalysen. Zusammenhangsanalysen befassen sich mit der Verbindung von zwei Variablen, beispielsweise ob ältere Personen eher mehr oder weniger Vertrauen in die Wissenschaft haben. Darstellen lassen sich solche Ergebnisse u.a. mittels Kreuztabellen, wobei eine Variable (in diesem Fall das Alter) in der Zeile und eine andere Variable (in diesem Fall das Vertrauen in die Wissenschaft) in der Spalte eingetragen wird (siehe Tabelle 2). Die jüngsten Teilnehmer*innen berichten am häufigsten davon, "sehr viel Vertrauen" aber auch "überhaupt kein Vertrauen" in die Wissenschaft zu haben, was auf eine polarisierte Gruppe hindeutet. Die Teilnehmer*innen mittleren Alters berichten generell von einem niedrigeren Vertrauen in die Wissenschaft als die jüngeren und älteren Teilnehmer*innen.

Tabelle 2: Kreuztabelle mit den Variablen "Alter" und "Vertrauen in die Wissenschaft"

	sehr viel Vertrauen	ziemlich viel Vertrauen	wenig Vertrauen	überhaupt kein Vertrauen
15-36 Jahre (n=670)	29,5%	44,9%	18,5%	7,0%
37-55 Jahre (n=641)	22,4%	50,4%	21,1%	6,1%
56+Jahre (n=805)	27,9%	52,0%	15,8%	4,3%
Gesamt (n=2116)	26,7%	49,1%	18,4%	5,8%

Anmerkung: Datengrundlage: VIC 3. Welle, gewichtete Daten

Wird dagegen eine Unterschiedshypothese überprüft, so analysiert man, ob sich bestimmte Gruppen in einem Merkmal unterscheiden. Hier sind häufig Lageparameter (wie Mittelwertsunterschiede) relevant. Ein einfaches Beispiel wäre die Frage: „Gibt es Unterschiede hinsichtlich der Impfbereitschaft zwischen Frauen und Männern?". Zur Beantwortung der Frage würden die Mittelwerte in der Befürwortung der Impfung zwischen Männern und Frauen verglichen sowie die Signifikanz überprüft werden. Zentrale Begriffe und Tests der univariaten und bivariaten Analyse sind in der Tabelle 3 dargestellt.

Tabelle 3: Grundbegriffe und Kennwerte der univariaten und bivariaten Analyse

Begriff	Erklärung
Univariate Analyse / Lageparameter	
Median	Der Wert, der eine Stichprobe in zwei gleich große Gruppen teilt. Wird z.B. die Impfbereitschaft von 1 (niedrig) bis 10 (hoch) erhoben, dann läge die Hälfte der Stichprobe bei einem Median von 5 unter diesem Wert, und die andere Hälfte darüber.
Mittelwert bzw. Durchschnitt	Durchschnittliche Ausprägung einer Variable in der jeweiligen Gruppe.
Mittelwertindex	Gebildeter Index aus dem Mittelwert mehrerer Variablen.
Schiefe-Wert	Erfasst, wie gleichmäßig sich die Stichprobe verteilt. Wenn beim oben genannten Beispiel (siehe Median) die meisten Teilnehmer*innen hohe Werte (z.B. 8, 9, 10) berichten, dann wäre diese Variable rechtssteil verteilt. Berichten die meisten Teilnehmer*innen dagegen geringe Werte, wäre die Variable linkssteil. Ist die Verteilung ausgeglichen, so entspricht der Median ungefähr dem Mittelwert.
Standard-abweichung	Durchschnittliche Abweichung vom Mittelwert in der jeweiligen Gruppe.
Z-Standardisierung	Transformierung von Variablen, sodass der Mittelwert 0 beträgt und die Standardabweichung 1. Diese Standardisierung ist sinnvoll, wenn Variablen mit unterschiedlichen Skalierungen miteinander verglichen werden.
Bivariate Analyse	
Kennwert zur Messung von Zusammenhangsanalysen	
Korrelations-koeffizient	Die Produkt-Moment-Korrelation (Pearson´s r) ist der zentrale Koeffizient zur Messung des Zusammenhangs zweier Variablen. Er nimmt Ausprägungen zwischen -1 und +1 an, je nachdem, ob es sich um einen indirekt proportionalen (negativen) Zusammenhang (je höher x, desto niedriger y) oder um einen direkt proportionalen (positiven) Zusammenhang (je höher x, desto höher y) handelt. Je weiter der Kennwert von 0 abweicht, desto stärker ist der Zusammenhang zweier Variablen. Je nach Skalenniveau gibt es weitere Zusammenhangsmaße, die angewendet werden können und die ähnlich interpretiert werden.
Tests zur Durchführung von Unterschiedsanalysen	
Chi²-Test	Test für Unterschiedshypothesen bei Variablen mit nominalen Skalenniveaus (= Antwortmöglichkeiten, die qualitativ unterschiedlich sind; z.B. Unterschiede nach Bundesland).
F-Test	Überbegriff mehrerer statistischer Tests, welche für die Signifikanzprüfung auf einer speziellen Wahrscheinlichkeitsverteilung, der F-Verteilung, beruhen. Ein klassisches Verfahren hierzu wäre die Varianzanalyse (siehe unten). Im Buch selbst wird der Begriff nicht explizit genutzt.
T-Test	Überbegriff mehrerer statistischer Tests, welche für die Signifikanzprüfung auf der t-Verteilung beruhen.
T-Test bei unabhängigen Stichproben	Überprüfung der Signifikanz zwischen zwei unabhängigen Gruppen mittels T-Test.

Begriff	Erklärung
T-Test bei verbundenen Stichproben	Überprüfung der Signifikanz zwischen zwei verbundenen Gruppen (z.B. dieselben Personen über zwei unterschiedliche Messzeitpunkte) mittels T-Test.
Varianzanalyse (einfaktoriell)	Varianzanalyse mit einer unabhängigen Variable, z.B. Mittelwertsvergleiche zwischen verschiedenen Altersgruppen.
Varianzanalyse (mehrfaktoriell)	Varianzanalyse mit mehreren unabhängigen Variablen, z.B. Mittelwertsvergleiche zwischen verschiedenen Alters- und Geschlechtergruppen.
Varianzanalyse (Mess-wiederholung)	Varianzanalyse bei Längsschnittdaten. Überprüft signifikante Änderungen in der jeweiligen Gruppe über mehrere Messzeitpunkte.
Welch-Test	Wird als Prüfinstrument eingesetzt, wenn sich die Standardabweichungen der Gruppen signifikant unterscheiden (robuster bzw. konservativer als T-Test).
Deskriptive (=beschreibende) Merkmale einer Stichprobe	
Effektstärke	Standardisierter Kennwert bei Unterschiedsanalysen mit einer Ausprägung von 0 bis 1, die Interpretation der Stärke des Effekts erfolgt in ähnlicher Form wie beim Korrelationskoeffizienten.

4 Erklärung und Begrifflichkeiten der multivariaten Analyse

In sozialwissenschaftlichen Erklärungsmodellen werden oftmals mehrere Variable berücksichtigt, um komplexere Sachverhalte erklären zu können. Dies wird als multivariate Analyse bezeichnet. Die Verfahren reichen von datenstrukturierenden Verfahren wie der Faktorenanalyse, um inhaltlich ähnliche Items zu bündeln, über lineare und logistische Regressionsmodelle zur Erklärung von bestimmten Phänomenen bis hin zu gruppenbildenden Klassifizierungsverfahren wie der Clusteranalyse.

4.1 Regressionsanalyse

Da zur Erklärung einer abhängigen Variablen meist mehrere Einflussfaktoren herangezogen werden müssen, werden in den (Sozial-)Wissenschaften häufig multivariate Analysen angewendet, um mehrere unabhängige Variable in ein Analysemodell zu integrieren[3]. Ein klassisches Verfahren zur Umsetzung dieser Idee stellt die multiple Regressionsanalyse dar. Diese Analyse erlaubt es, mehrere erklärende (oder unabhängige) Variablen in

3 Multivariate Analysen bieten außerdem die Möglichkeit, Interaktionen zwischen Variablen zu entdecken bzw. zu überprüfen. Auf diese Interaktionseffekte wird in diesem Glossar im Sinne der Übersichtlichkeit allerdings nicht weiter eingegangen.

ein Modell aufzunehmen, um eine zu erklärende (oder abhängige) Variable zu schätzen. Durch diese Art der Analyse kann man erkennen, wie sich die jeweilige unabhängige Variable auf die abhängige Variable auswirkt und man kann einzelne Effekte durch die Konstanthaltung der restlichen (Kontroll-)Variablen unverzerrt ausweisen.

So ist es beispielsweise realistisch, dass nicht nur das Alter die Impfbereitschaft erklärt, sondern auch andere Variablen wie beispielsweise das Geschlecht, die Bildung, und vieles mehr (siehe Abbildung 3). Demnach wäre es denkbar, dass formal niedriger gebildete Personen den Fortschritten der Wissenschaft und den allgemeinen Corona-Maßnahmen skeptischer gegenüberstehen und deshalb impfkritischer sind. Auch Frauen könnten im Vergleich zu Männern eine größere Impfskepsis aufweisen.

Im Beispiel der Abbildung 3 sieht man jeweils den Effekt des Alters, des Geschlechts und der Bildung auf die Impfbereitschaft unter Kontrolle der restlichen Variablen. Zentrale Kennwerte der multiplen Regressionsanalyse sind der Tabelle 4 zu entnehmen.

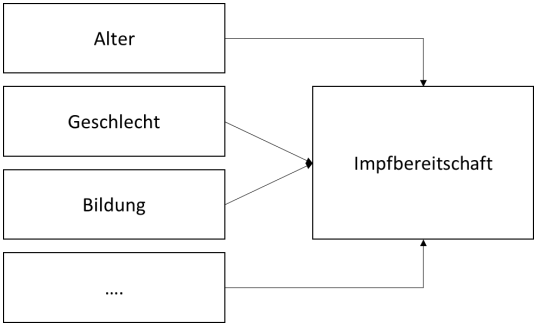

Abbildung 3: Multivariate Analyse (Beispiel mehrere Einflussfaktoren / additive Effekte)

Es gibt mehrere verschiedene Arten der Regressionsanalyse, welche sich hauptsächlich auf Basis des Skalenniveaus bzw. der Natur der Daten (Querschnitt vs. Längsschnitt) unterscheiden. Die klassische bzw. einfachste Form der Regressionsanalyse stellt die Ordinary Least Squares (OLS) Regressionsanalyse dar (oder lineare Regressionsanalyse). Darüber hinaus kommen in diesem Sammelband auch sequenzielle, ordinale und binär-logistische Regressionsanalysen zum Einsatz. Die zentralen Charakteristika der Verfahren sind jedoch ähnlich, auch wenn die Interpretation der Ergebnisse oftmals leicht unterschiedlich ist.

Tabelle 4: Zentrale Kennwerte der Regressionsanalyse

Begriff	Erklärung
(Regressions-) Konstante	Ausprägung der abhängigen Variable, wenn alle unabhängigen Variablen die Ausprägung „0“ aufweisen.
B-Wert	Die Veränderung der abhängigen Variable, wenn sich die unabhängige Variable um eine Einheit erhöht. Bei positiven B-Werten zeigt sich ein positiver Zusammenhang der Variablen (je höher die Ausprägung der Variable x, desto höher die Ausprägung der Variable y). Bei negativen B-Werten zeigt sich ein umgekehrter Zusammenhang.
Beta-Wert	Standardisierter B-Wert zwischen -1 und 1, sodass auch Variablen mit unterschiedlichen Skalierungen (z.B. 1-5 vs. 1-10) miteinander verglichen werden können. Die Interpretation erfolgt deckungsgleich zu den Korrelationskoeffizienten bzw. Effektstärken.
Dummyvariablen	Spezielle Form der unabhängigen Variablen, die mit „1“ und „0“ codiert sind. Typischerweise werden nominale Variablen mit mehr als 2 Ausprägungen oder ordinale Messungen für Regressionen in mehrere Dummyvariablen aufgespalten. So können Spezifikationsfehler – Regressionen setzen metrisches Skalenniveau voraus – minimiert werden und die Variablen können dennoch differenziert analysiert werden.
Multikollinearität	Eine Fehlerquelle in der Regressionsanalyse, die auftritt, wenn die Zusammenhänge zwischen den unabhängigen Variablen zu groß sind und dadurch eine wechselseitige Beeinflussung gegeben ist. Mittels verschiedener Koeffizienten kann Multikollinearität geprüft und somit auch kontrolliert werden.
R^2	Die gesamte Erklärungskraft (Varianzaufklärung) der unabhängigen Variablen für die abhängige Variable. Sie variiert zwischen 0 und 1 (bzw. 0% und 100%), wobei der Wert aussagt, welcher Anteil der abhängigen Variable durch die unabhängige(n) Variable(n) erklärt werden kann.
R^2 (korrigiertes)	Bei einem Regressionsmodell mit mehreren Variablen kommt es zu einer Verzerrung des R^2 Wertes, da sich dieser Wert durch das Hinzufügen weiterer unabhängigen Variablen nur vergrößern kann, nicht aber verkleinern. Das korrigierte R^2 berücksichtigt dieses Problem und sollte in der Regel bei Regressionsanalysen mit mehreren unabhängigen Variablen berichtet werden.
R^2 (Nagelkerke)	Alternativer (Pseudo-) R^2 Kennwert bei ordinalen, logistischen, binomialen und multinomialen Regressionsanalysen, da bei diesen Regressionsanalysen keine Varianzaufklärung berechnet werden kann. Der Wert zeigt an, um wie viel Prozent sich das Erklärungsmodell (mit unabhängigen Variablen) im Vergleich zum Nullmodell (ohne unabhängige Variablen) verbessert.
Odds Ratio (OR)	Wird u.a. bei ordinalen oder logistischen Regressionsanalysen anstatt des Beta- und B-Wertes berichtet. Erklärt wird bei diesen Regressionsanalysen, im Gegensatz zur klassischen (linearen) Regressionsanalyse, nicht eine kontinuierliche Änderung der abhängigen Variable (z.B. ein Zuwachs an Impfakzeptanz), sondern ein sprunghaftes Ereignis (z.B. Ich lasse mich impfen oder nicht). Die Odds Ratio schätzt dabei die Wahrscheinlichkeit des Eintretens dieses abhängigen Ereignisses auf Grundlage unabhängiger Variablen (UV). Bei der UV „Geschlecht“ (Mann = 0; Frau = 1) bedeutet eine OR von 1,10 beispielsweise, dass die Impfbereitschaft von Frauen im Vergleich zu Männern um das 1,10-fache erhöht ist (fiktives Beispiel).

4.2 Faktorenanalyse

Im Gegensatz zur Regressionsanalyse besteht das Ziel der Faktorenanalyse nicht in der Erklärung einer abhängigen Variablen, sondern darin, mehrere direkt erhobene Variablen zu einem oder mehreren Faktoren zu bündeln. Diese Faktoren stellen latente (und damit sehr messgenaue) Konstrukte dar, die mittels der direkt erhobenen Variablen geschätzt werden können. Diese geschätzten Faktorscores[4] können anschließend für weiterführende Analysen verwendet werden, beispielsweise als abhängige oder unabhängige Variablen in einer Regressionsanalyse. Dabei wird je nach Wissensstand zu den Themen, die mit den Variablen behandelt werden, zwischen sogenannten explorativen und konfirmatorischen Faktorenanalysen unterschieden.

Die explorative Faktorenanalyse basiert auf den Korrelationen der Variablen und der Annahme, dass diese Korrelation durch einen gemeinsamen latenten Faktor zustande kommt. Je besser eine Variable durch den jeweiligen Faktor erklärt werden kann, desto höher ist die Faktorladung dieser Variablen auf den Faktor. Dieser Vorgang ist vergleichbar mit einer Regressionsanalyse, bei der die Variablen die unabhängigen Variablen darstellen und die Faktorladung den Beta-Koeffizienten, während der Faktor die abhängige Variable bildet. Demnach gibt es auch hier eine Varianzaufklärung des Faktors durch die Variablen. Die Varianzaufklärung gibt an, wie gut dieser Faktor durch die auf ihn gebündelten Items erklärt werden kann. In der Regel werden Faktoren extrahiert, die mehr Information enthalten als eine der ursprünglichen Variablen; also einen sogenannten Eigenwert > 1 aufweisen. Für den Fall, dass sich dabei zwei oder mehrere Faktoren bilden, wird in der Regel eine (orthogonale oder schiefwinkelige) *Faktorrotation* durchgeführt, welche eine klare Zuordnung der Items zu den jeweiligen Faktoren erlaubt. Bei einer orthogonalen Rotation wird von einer Unabhängigkeit der Faktoren ausgegangen, diese stehen dann in keinem Zusammenhang zueinander. Bei der schiefwinkeligen Rotation geht man davon aus, dass die Faktoren miteinander in Zusammenhang stehen.

Während bei der explorativen Faktorenanalyse unbekannte Muster und Strukturen im Datensatz gesucht werden, geht es bei der konfirmatorischen Faktorenanalyse darum, eine vorab festgelegte Theorie oder Hypothese zu überprüfen. Dazu wird überprüft, wie beobachtbare Variablen mit den latenten Faktoren in Beziehung stehen. Ein Beispiel dafür kann sein,

4 Anstelle der Factorscores ist es auch möglich, den Mittelwert der dem Faktor zugehörigen Items zu bilden

dass man die latente Variable Institutionenvertrauen mittels fünf Items, die direkt das Vertrauen in die Regierung, das Parlament, die Justiz, das Bildungswesen und das Gesundheitswesen abbilden, messen möchte. Man legt für die konfirmatorische Faktorenanalyse fest, dass diese fünf Items den latenten Faktor beeinflussen, und formuliert mathematischen Gleichungen, die die erwarteten Beziehungen zwischen den latenten Faktoren und den beobachtbaren Variablen beschreiben. Diese Beziehungen werden dann getestet und es wird dargestellt, wie gut oder schlecht die Daten der einzelnen erhobenen Variablen die theoretischen Modellerwartungen erfüllen.

Statistische Maße wie der Chi²-Test, Comparative Fit Index (CFI) oder Root Mean Square Error of Approximation (RMSEA) geben an, ob das spezifizierte Modell empirisch nachgewiesen werden konnte und folglich eine hohe Modellgüte aufweist Diese Kennzahlen sind auch für Strukturgleichungsmodelle von Relevanz, die später beschrieben werden.

4.3 Reliabilitätsanalyse

Explorative Faktorenanalysen werden üblicherweise dann durchgeführt, wenn im vornherein nicht klar ist, wie viele verschiedene Dimensionen hinter einer Anzahl von Items verborgen sind. Wenn einzelne Itembatterien (=eine größere Anzahl an Items mit identischer Messskala) nur auf einen speziellen Faktor zurückgeführt werden können, dann kann die Messgenauigkeit dieses Faktors (= dieser Skala) geprüft werden. Sind die Items metrisch skaliert oder haben drei oder mehr Antwortmöglichkeiten, mit identischen Abständen zwischen den Antwortmöglichkeiten, wird die Reliabilität der Skala in der Regel mit dem Koeffizienten Cronbach´s Alpha (α) gemessen. Der Wert bildet die sogenannte interne Konsistenz der Skala ab. Sind die Items dagegen dichotom skaliert, das heißt mit nur zwei Antwortkategorien (z.B. „Ja“ und „Nein“), so wird üblicherweise der Reproduktionskoeffizient zur Prüfung der internen Konsistenz herangezogen (vgl. Tabelle 5). In der klassischen Literatur zur sozialwissenschaftlichen Statistik wird üblicherweise bei einem Wert > 0,7 von einer guten internen Konsistenz gesprochen, sprich, die Items messen das dahinterliegende Konstrukt mit ausreichender Präzision.

Tabelle 5: Zentrale Kennwerte der Faktorenanalyse & Reliabilitätsanalyse

Begriff	Erklärung
Faktorenanalyse	
KMO-Wert	Zeigt an, ob sich die Daten für die Verwendung einer Faktorenanalyse eignen. Werte gegen 1 deuten eine gute Eignung an, während Werte gegen 0 für eine schlechte Eignung stehen. Bei einem Wert unter 0.5 sollte von der Verwendung der Faktorenanalyse abgesehen werden.
Bartlett's Test auf Sphärizität	Testet die Annahme, dass es zwischen den verwendeten Variablen keinerlei Zusammenhänge gibt, welches gegen die Durchführung einer Faktorenanalyse sprechen würde. Ist der Test statistisch signifikant ($p < 0.05$), so liegen Zusammenhänge vor und die Analyse kann (mit gleichzeitiger Berücksichtigung des KMO-Wertes) durchgeführt werden.
Kommunalität	Gibt den Anteil der Varianz einer Variable an, für den die ermittelten Faktoren verantwortlich sind. Liegt der Wert bspw. Bei 0.6, so werden 60% der Varianz der Variable durch die Faktoren erklärt. Folglich bewegen sich die Werte im Bereich 0-1.
Faktorladungen	Geben an, wie stark eine Variable mit einem Faktor korreliert. Sie werden ähnlich wie Betakoeffizienten interpretiert und können daher zwischen –1 und +1 liegen. Werte gegen 0 verdeutlichen, dass es keine Korrelation zwischen Variable und Faktor gibt und diese somit kein "Teil" des Faktors ist. Werte gegen 1 stehen hingegen für hohe Korrelationen mit dem Faktor. Wenn Variablen größere Ladungen (etwa 0.3-0.4) auf mehreren Faktoren aufweisen, so spricht man von Querladungen. Diese können das Ermitteln trennscharfer Faktoren erschweren.
Eigenwert	Verdeutlicht, wie viel der gesamten Varianz aller Variablen durch einen Faktor erklärt werden kann. Liegt der Wert bei 1, so erklärt ein Faktor nicht mehr als eine einzige Variable, weshalb nur Faktoren mit einem Eigenwert >1 berücksichtigt werden. Der Begriff Eigenwert kommt in keinem Text dieses Buches explizit vor, ist aber für die Faktorenbildung zentral.
CFI/Comparative Fit Index	Wird im Kontext konfirmatorischer Faktorenanalysen als Gütekriterium verwendet und gibt an, wie gut das vorgeschlagene Modell auf die zugrundeliegenden Daten passt. Werte über 0.90 stehen für eine gute Eignung, Werte über 0,95 für eine sehr gute Eignung des Modells.
RMSEA/Root Mean Square Error of Approximation	Stellt ebenfalls einen Wert zur Überprüfung der Modellgüte bei konfirmatorischen Faktorenanalysen dar. Es wird meist vorgeschlagen, dass ein Wert <0.05 als gut einzustufen ist.
Reliabilitätsanalyse	
Cronbach´s Alpha (α)	Überprüfung der internen Konsistenz einer Skala, der Koeffizient zeigt die Zuverlässigkeit der Messung einer latenten Variable (eines zugrundeliegenden Faktors) auf. Anwendung bei metrischen Items bzw. auch identisch skalierten ordinalen Items (siehe Fließtext). Ähnlich dem KMO Wert gilt eine Reliabilität von 0,5 als mäßig, ab 0,7 kann von einer ausreichenden, ab 0,8 von einer guten Reliabilität berichtet werden.

4.4 Strukturgleichungsmodelle (SGM)

Strukturgleichungsmodelle (oder engl. SEM, Structural Equation Models) kommen dann zum Einsatz, wenn Hypothesen über die kausalen Zusammenhänge zwischen einer Vielzahl an Variablen getestet werden sollen. Während multivariate Regressionen nur den Einfluss von unabhängigen Variablen auf eine einzelne (Ziel-)Variable testen (siehe Abbildung 2), können SGMs komplexe Beziehungen zwischen einer Vielzahl an Variablen abbilden. Das schließt beispielsweise auch mehrere abhängige (Ziel-)Variablen gleichzeitig mit ein (Byrne, 2001, 3-4; Reinecke & Pöge, 2010, 775-776).

Zur Testung eines SEMs werden in einem ersten Schritt die Hypothesen über die Ursache- und Wirkungszusammenhänge zwischen den betrachteten Variablen aufgestellt. Diese werden dann in Folge in einem sogenannten Strukturmodell dargestellt. Darin ist jede der gemessenen Variablen als ein Kasten eingezeichnet. Zwischen diesen Kästen werden mit Pfeilen die vermuteten kausalen Zusammenhänge abgetragen. In Abbildung 4 wird z.B. die Hypothese aufgestellt, dass Alter, Geschlecht und Bildung sowohl die Impfbereitschaft als auch die Parteipräferenz der Befragten beeinflussen. Der Doppelpfeil zwischen der Impfbereitschaft und der Parteipräferenz signalisiert, dass eine Korrelation zwischen diesen Variablen angenommen wird.

SEMs erlauben es außerdem, sogenannte latente Variablen abzubilden. Der Begriff der latenten Variable beschreibt Einflüsse, die nicht direkt gemessen werden können, sondern nur anhand von konkreten Indikatoren indirekt erhoben werden kann. Depressivität kann z.B. nur über Fragen zu konkreten Symptomen gemessen werden (Gefühl von Niedergeschlagenheit, wenig Interesse an Aktivitäten, negative Gefühle nicht kontrollieren können etc.). Die Annahme ist dabei jeweils, dass die latente Variable die Antworten in den beobachteten Variablen bestimmt. Im Strukturmodell werden sie als Ellipsen dargestellt (siehe z.B. die Variable „politisches Vertrauen“ in Abbildung 4, die anhand des Vertrauens zu konkreten Institutionen erhoben wird). Diese latenten Variablen spielen v.a. für die konfirmatorische Faktorenanalyse eine wichtige Rolle (siehe Abschnitt zuvor) (Byrne, 2001, 4-5; Reinecke & Pöge, 2010, 777-778).

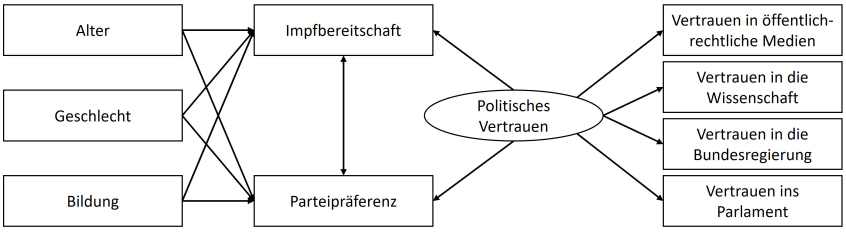

Abbildung 4: Beispielhaftes Strukturmodell

Nachdem die Hypothesen über die Beziehungen zwischen den latenten und beobachteten Variablen aufgestellt wurden, wird getestet, ob das Strukturmodell die Struktur in den Daten gut widerspiegelt. Um die Modellgüte – also die Passung des Modells an die tatsächlichen Daten – zu überprüfen, wird auf dieselben Kennzahlen wie bei der konfirmatorischen Faktorenanalyse zurückgegriffen (Chi², CFI, RMSEA) (Reinecke & Pöge, 2010, 783-784). Ist die Modellgüte nicht zufriedenstellend, wird das Strukturmodell entsprechend angepasst – entweder durch den Ausschluss von Variablen, die Aufnahme neuer Variablen, das Umkehren von Pfeilen im Strukturmodell, dem Einzeichnen noch fehlender Pfeile etc. So nähern sich die Annahmen über die Ursache- und Wirkungszusammenhänge zwischen den Variablen immer mehr der Realität der Daten an (Reinecke & Pöge, 2010, 783-784).

Latente Wachstumskurvenmodelle sind eine besondere Form von SGMs, die nur in Paneldaten zum Einsatz kommen. Sie sollen die Entwicklung der Werte, die Befragte in einer oder mehreren abhängigen Variablen erzielen, über die Zeit hinweg nachvollziehen. Die Annahme ist, dass hinter Veränderungen im Antwortverhalten latente Prozesse stecken, die nicht direkt messbar sind. Um diese im Strukturmodell abzubilden, werden zwei zusätzliche latente Variablen aufgenommen (siehe Abbildung 5). Zum einen hat jeder Befragte einen individuellen Startwert (der Wert in der ersten Welle, in der er befragt wurde), der alle Werte, die dieser Befragte in den folgenden Erhebungszeitpunkten erreicht, beeinflusst. Beispielsweise kann man davon ausgehen, dass die Lebenszufriedenheit in einem Jahr durch das Niveau im Vorjahr geprägt ist. Zum anderen werden die Werte der abhängigen Variablen durch die zeitliche Veränderung beeinflusst, die ebenfalls als latente Variable modelliert wird. Die latente Variable „Startwert“ nimmt auf die Werte der abhängigen Variable zu allen Messzeitpunkten denselben Einfluss (p), während die latente Variable der „zeitlichen Veränderung“ die

Werte ab dem zweiten Messzeitpunkt beeinflusst. Dieser Einfluss kann im berechneten Modell pro Messzeitpunkt variiert werden (q1 vs. q2 in Abbildung 5), damit unterschiedliche Entwicklungsverläufe modelliert werden können (Schmiedek & Wolff, 2010, 1017-1021). Nach der Berechnung wird, wie bei allen SGMs, die Modellgüte überprüft, um immer näher an die Realität, die sich in den Daten ausdrückt, heranzukommen (siehe oben).

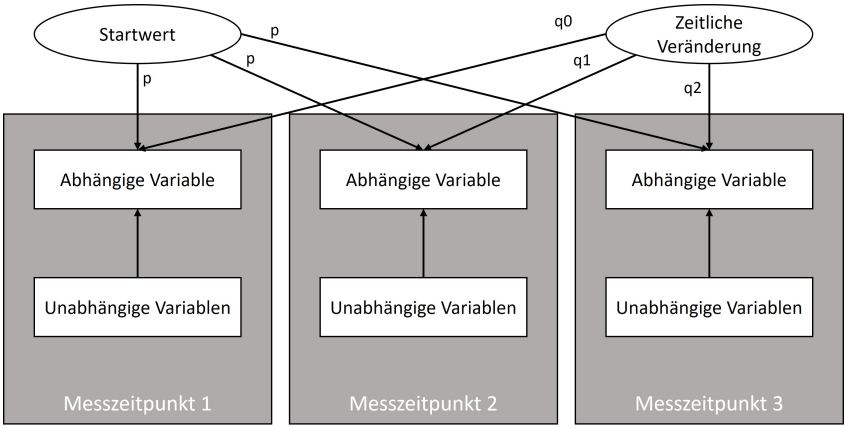

Abbildung 5: Vereinfachte Darstellung eines Latenten Wachstumskurvenmodells

4.5 Clusteranalyse

Die Clusteranalysen (CA) sind statistische Verfahren, die allesamt darauf ausgerichtet sind, die im Datensatz vorliegenden Untersuchungseinheiten zu klassifizieren. Somit ist es beispielsweise möglich, Personen, Organisationen oder auch ganze Länder auf Basis verschiedener Eigenschaften in Gruppen („Cluster") einzuteilen. Die Elemente einer Gruppe sollen sich dabei im Hinblick auf die gewählten Merkmale besonders ähnlich sein und sich gleichzeitig aber auch deutlich von den Elementen der anderen Gruppen unterscheiden (vgl. Bacher et al., 2010; König & Jäckle, 2017).

Um diesen Vorgaben gerecht zu werden, nutzen clusteranalytische Verfahren spezielle Maße für die Ähnlichkeit oder Unähnlichkeit der zugrundeliegenden Elemente (sogenannte „Proximitätsmaße"; z.B. die quadrierte euklidische Distanz). Diese können dann mithilfe verschiedener Algorithmen („Fusionsalgorithmen") dazu genutzt werden, um die Untersuchungs-

einheiten zu gruppieren (vgl. Wiedenbeck & Züll, 2010). Darüber hinaus können Clusteranalysen rein explorativ (also nur zum Zweck der Musterentdeckung) oder zur Bestätigung von Vorannahmen über etwaige Muster in den Daten verwendet werden (vgl. Bacher et al., 2010, König & Jäckle, 2017). Unterschieden wird daher auch zwischen Verfahren, bei denen die Gruppen schrittweise aus den Daten herausgebildet werden („hierarchische" CA), und jenen, bei denen die einzelnen Untersuchungsobjekte (auf Basis eines spezifischen Kriteriums) einer vorgegebenen Anzahl an Gruppen zugeordnet werden („partitionierende" CA) (vgl. Bacher et al., 2010).

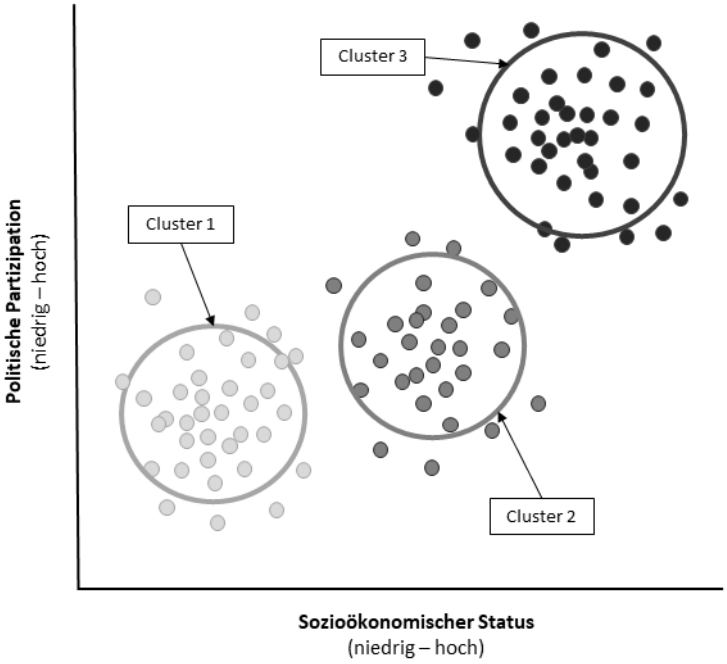

Abbildung 6: Fiktive Darstellung einer möglichen Clusterung

Eine der wohl wichtigsten Arten der Clusteranalyse ist das „k-Means"-Verfahren. Hier muss die Anzahl der Gruppen vorab festgelegt werden, woraufhin die Untersuchungsobjekte diesen zufällig zugeordnet werden. Anschließend werden die Gruppen neu berechnet, indem der Algorithmus

die Zugehörigkeit zu den Gruppen so bestimmt, dass diese intern möglichst homogen ausfallen (sodass also die Streuung innerhalb der Cluster minimiert wird) und sich nebenbei auch noch maximal voneinander unterscheiden (also eine möglichst große Streuung zwischen den Clustern vorliegt). Diese Schritte werden so lange wiederholt, bis es zu einer Optimierung der Gruppierung kommt, wobei die Elemente in den Gruppen so homogen wie möglich und die Gruppen selbst so unterschiedlich wie möglich sind. Da der verwendete Algorithmus auf den Mittelwerten basiert, müssen alle Variablen metrisches Skalenniveau aufweisen und das Verfahren ist anfällig für Verzerrungen durch Extremwerte (vgl. Bacher et al., 2010, König & Jäckle, 2017).

Die ebenfalls häufig und insbesondere bei hohen Fallzahlen angewandte Two-Step-Clusteranalyse benötigt hingegen keine Vorannahmen über die Anzahl der Cluster und kann metrische sowie auch kategoriale Variablen verarbeiten. Der zugrundeliegende Algorithmus nimmt die Gruppierung in zwei Schritten vor. Zuerst werden die Untersuchungseinheiten vorgruppiert (zu sogenannten „Präclustern"). Die Abstände zwischen den Präclustern werden schließlich dazu verwendet, um die endgültige Gruppierung vorzunehmen. Die tatsächliche Anzahl der Gruppen wird durch eine Maßzahl bestimmt, das „Bayes-Informationskriterium" (BIC). Dieses verändert sich mit jeder zusätzlichen Bildung eines Clusters, wobei letztendlich jene Gruppenanzahl gewählt wird, für die der BIC-Wert am niedrigsten ausfällt (vgl. Bacher et al., 2010; König & Jäckle, 2017).

Neben k-Means- und dem Two-Step-Verfahren gibt es auch noch weitere Arten der CA. Von hoher Bedeutung ist zudem die Latente Klassenanalyse (LCA), die sich gut für die Analyse nominalskalierter Daten eignet. Neben ihrem grundlegenden Anspruch der Klassifizierung haben Clusteranalysen gemeinsam, dass die durch sie gewonnenen Gruppen oft als unabhängige oder abhängige Variablen in weiterführenden Analysen (z.B. in Regressionsmodellen oder Varianzanalysen) verwendet werden (vgl. König & Jäckle, 2017). So können die Teilnehmer*innen einer Befragung beispielsweise auf Basis verschiedener Merkmale (z.B. Einkommen, Bildungsgrad, Werthaltungen etc.) in soziale Milieus unterteilt werden, während in einem nächsten Schritt analysiert wird, inwiefern sich diese Milieus im Hinblick auf ihre politischen Einstellungen unterscheiden.

Tabelle 6: Zentrale Konzepte und Verfahren der Clusteranalyse (nicht alle Begriffe werden im Buch explizit genannt, stellen aber den statistischen Hintergrund dar)

Begriff	Erklärung
Proximitätsmaße	Maßzahl für die Ähnlichkeit oder Distanz zwischen den zu gruppierenden Elementen. Welches Maß verwendet wird, hängt dabei vom Skalenniveau der gewählten Variablen ab. Ein häufig verwendetes Beispiel ist die (quadrierte) euklidische Distanz.
Fusionsalgorithmen	Bestimmen, nach welcher Logik die Gruppierung der Elemente vorgenommen wird. Unterschieden wird zwischen Verfahren, welche die Cluster schrittweise aus den Daten extrahieren ("hierarchisch") und solchen, die vorab bestehende Gruppierungen durch schrittweises Umsortieren der Elemente optimieren ("partitionierend").
K-Means-Verfahren	Eines der wichtigsten partionierenden Clusterverfahren für metrische Variablen. Arbeitet mit Mittelwerten und setzt voraus, dass die Anzahl der Gruppen vorab bekannt ist. Die Elemente werden diesen zufällig zugeordnet und anschließend so oft umsortiert, bis die Gruppen möglichst homogen sind.
Two-Step-Verfahren, auch 2-Step Verfahren	Ebenfalls häufig verwendetes Clusterverfahren. Benötigt keine Angabe der Gruppenanzahl und kann mit verschiedenen Skalenniveaus und hohen Fallzahlen verwendet werden. Die Clusterung erfolgt in zwei Schritten und die Anzahl an Clustern wird mathematisch durch die Berechnung des BIC bestimmt
Latente Klassenanalyse	Klassifizierungsverfahren, welches für die Analyse nominalskalierter Daten geeignet ist. Basiert ebenfalls auf einem schrittweisen Vorgehen, bei dem die Elemente den Gruppen auf Basis der größten Wahrscheinlichkeit zugeordnet werden, und ermöglicht auch eine Bestimmung der Clusteranzahl via BIC.
BIC, auch Bayes Information Criterion und Bayes-Informationskriterium	Gütekriterium für die Bewertung von statistischen Modellen. Im Gegensatz zum R^2 in der linearen Regression wird es nicht für sich interpretiert, sondern immer mit dem BIC-Wert alternativer Modelle verglichen. Kleinere Werte stehen dabei für ein besseres Modell.

Literatur

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Oldenbourg Wissenschaftsverlag.

Bryman, A. (2016). *Social Science Research Methods*. Oxford University Press.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS: basic concepts, applications, and programming*. Lawrence Erlbaum Associates, Inc. Publishers.

Glatz, C., Prandner, D. & Aschauer, W. (2022). Glossar für statistische Analysetechniken. In: W. Aschauer, C. Glatz, C & D. Prandner (Hrsg.), *Die österreichische Gesellschaft während der Corona-Pandemie*. Springer VS, 350-359.

König, P. D. & Jäckle, S. (2017). Clusteranalyse. In S. Jäckle (Hrsg.), *Neue Trends in den Sozialwissenschaften*. Springer Fachmedien, 51-84.

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In F. Kreuter (Hrsg.), *Improving Surveys with Paradata.*John Wiley & Sons, 1-9.

Prandner, D. (2019). Sozialer Survey Österreich – Methodik des Sozialen Survey Österreich 2016. In: J. Bacher, A. Grausgruber, A., M. Haller, F. Höllinger, D. Prandner & R. Verwiebe, (Hrsg.), *Sozialstruktur und Wertewandel in Österreich*. Springer VS, 515-531.

Reinecke, J. Pöge, A. (2010). Strukturgleichungsmodelle. In: C. Wolf & B. Henning (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse.*Springer Fachmedien, 775-804.

Ruggles, S. (2018). Metadata and Preservation. In: D. Vannette & Krosnick J.A. (Hrsg.), *The Palgrave Handbook of Survey Research*. Palgrave, 635-643.

Schmiedek, F. Wolff, J. K. (2010). Latente Wachstumskurvenmodelle. In: C. Wolf & B. Henning (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse*. Springer Fachmedien, 1017-1029.

West, Brady T. 2011. *Paradata in Survey Research*. Survey Practice 4 (4). https://doi.org/10.29115/SP-2011-0018 (Stand: 15.1.2024)

Wiedenbeck, M. & Züll, C. (2010). Clusteranalyse. In: C. Wolf & B. Henning (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse*. Springer Fachmedien, 525-552.