Multivariate Analyseverfahren in der Marktforschung

Dipl.-WiInf.(FH) Christian Reinboth

- Skript zur Vertiefungsrichtung Marktforschung -

- Sommersemester 2006 -

Hochschule Harz
Fachbereich Wirtschaftswissenschaften

Statistical Analysis: Mysterious, sometimes bizarre, manipulations performed upon the collected data of an experiment in order to obscure the fact that the results have no generalizable meaning for humanity. Commonly, computers are used, lending an additional aura of unreality to the proceedings.

There was this statistics student who, when driving his car, would always accelerate hard before coming to any junction, whizz straight over it , then slow down again once he'd got over it. One day, he took a passenger, who was understandably unnerved by his driving style, and asked him why he went so fast over junctions. The statistics student replied, "Well, statistically speaking, you are far more likely to have an accident at a junction, so I just make sure that I spend less time there."

A mathematician and a non-mathematician are sitting in an airport hall waiting for their flight to go. The non has terrible flight panic. "Hey, don't worry, it's just every 10000th flight that crashes." "1:10000? So much? Then it surely will be mine!" "Well, there is an easy way out. Simply take the next plane. It's much more probable that you go from a crashing to a non-crashing plane than the other way round. So you are already at 1:10000 squared."

A stats major was completely hung over the day of his final exam. It was a True/False test, so he decided to flip a coin for the answers. The stats professor watched the student the entire two hours as he was flipping the coin...writing the answer...flipping the coin...writing the answer. At the end of the two hours, everyone else had left the final except for the one student. The professor walks up to his desk and interrupts the student, saying: "Listen, I have seen that you did not study for this statistics test, you didn't even open the exam. If you are just flipping a coin for your answer, what is taking you so long? The student replies bitterly (as he is still flipping the coin): " Shhh! I am checking my answers!"

Two statisticians were travelling in an airplane from LA to New York. About an hour into the flight, the pilot announced that they had lost an engine, but don't worry, there are three left. However, instead of 5 hours it would take 7 hours to get to New York. A little later, he announced that a second engine failed, and they still had two left, but it would take 10 hours to get to New York. Somewhat later, the pilot again came on the intercom and announced that a third engine had died. Never fear, he announced, because the plane could fly on a single engine. However, it would now take 18 hours to get to new York. At this point, one statistician turned to the other and said, "Gee, I hope we don't lose that last engine, or we'll be up here forever!"

A famous statistician would never travel by airplane, because he had studied air travel and estimated the probability of there being a bomb on any given flight was 1 in a million, and he was not prepared to accept these odds. One day a colleague met him at a conference far from home. "How did you get here, by train?" "No, I flew" "What about your the possibiltiy of a bomb?" Well, I began thinking that if the odds of one bomb are 1:million, then the odds of TWO bombs are (1/1,000,000) x (1/1,000,000). This is a very, very small probability, which I can accept. So, now I bring my own bomb along!"

Inhaltsverzeichnis

A Einführung

I Zu diesem Manuskript

Ziel dieses Manuskripts ist die Vermittlung grundlegender Kenntnisse im Bereich der Anwendung multivariater statistischer Analyseverfahren in der Marktforschung und die Unterstützung der Studentinnen und Studenten bei der Vorbereitung auf die Fachprüfungen. Dabei werden lediglich die Inhalte eines der beiden Fachsemester – explorative Datenanalyse, multiple Regression, Varianzanalyse und Faktorenanalyse – im Detail betrachtet.

Auf ständige, den Lesefluss unterbrechende Hinweise auf Literaturquellen wurde zugunsten der Lesbarkeit verzichtet. Dennoch sind sämtliche verwendeten Quellen korrekt aufgeführt, und zwar erstens in ihrer Vollständigkeit im Literaturverzeichnis im Anhang und zweitens kapitelweise am Ende jedes Abschnitts unter der Überschrift „Weiterführende Literatur". Diese Angaben sind daher zugleich als Quellenangaben für das jeweilige Kapitel aufzufassen.

Mein Dank gilt Prof. Dr. Lammers, nicht nur für viele spannende Vorlesungen, sondern auch für die Chance, seine Vertiefungsrichtung für ein Semester zu übernehmen. Des weiteren danke ich Mirjam Scholz und Jana Busch, die sich mit Korrekturhinweisen und Verbesserungsvorschlägen in die Entstehung der finalen Fassung eingebracht haben, sowie Joachim Verhagen für seine kleine Sammlung statistischer Pointen.

Helfen Sie bitte mit, dieses Manuskript besser zu machen: Sollten Sie Fehler bemerken oder Verbesserungsvorschläge haben, dann freue ich mich über Ihre Hinweise, die Sie mir per E-Mail an creinboth@hs-harz.de zukommen lassen können.

Christian Reinboth, 27.08.2006

II Inhalte dieses Manuskripts

1 Was ist Marktforschung?

Die Marktforschung ist der Bestandteil des Unternehmens, der dieses mit der Außenwelt verbindet. Sie wird unter anderem betrieben, um Grundwissen über die verschiedenen Märkte[1] zu schaffen, auf denen ein Unternehmen agiert, Unsicherheiten über zukünftige Zustände zu beseitigen, den Informationsaustausch zwischen Unternehmen und Außenwelt zu verbessern und um dem Controlling Daten für die Überprüfung von Zielvorgaben, beispielsweise in der Markenpflege, zu liefern.

Der Marktforschungsprozess lässt sich dabei in fünf wesentliche Phasen unterteilen. In der Literatur finden sich zwar auch noch andere Darstellungen, diese lassen sich aber immer wieder auf das fünfphasige Modell zurückführen, welches auch als „5-D-Modell“ bezeichnet wird. Auf eine ausführliche Beschreibung der fünf Phasen kann im Rahmen dieses Skripts getrost verzichtet werden, die tabellarische Darstellung dient lediglich dazu, einen gewissen Bezugsrahmen für die Anwendung der multivariaten Analyseverfahren aufzuspannen, die in der vierten Phase – der Datengewinnungsphase – zum Einsatz kommen.

Phase	*Inhalte*
Phase 1: Definitionsphase	• Formulierung des Informationsbedarfs • Auswahl des Forschungsansatzes
Phase 2: Designphase	• Bestimmung der Informationsquellen • Bestimmung der Erhebungsmethoden
Phase 3: Datengewinnungsphase	• Durchführung von Befragungen und Experimenten • Auswahl, Ansprache, Motivation und Kontrolle
Phase 4: Datenanalysephase	• Datenbereinigung und Kodierung • Datenauswertung und Ergebnisinterpretation
Phase 5: Dokumentationsphase	• Erstellung des Forschungsberichts • Präsentation der Ergebnisse

Abbildung 1: Die 5 D's der Marktforschung

1 Die wichtigsten Märkte sind hier der Finanz-, der Arbeits-, der Beschaffungs-, und natürlich der Absatzmarkt, an dem die angebotenen Güter und Dienstleistungen letztendlich verkauft werden.

2 Warum sollte man sich mit Marktforschung beschäftigen?

Neben der großen persönlichen Bereicherung, die man durch die intensive Beschäftigung mit den in der Marktforschung zum Einsatz kommenden statistischen Methoden erfährt, ist es natürlich das vorrangige Ziel jedes Studenten, nach Beendigung des Studiums einen möglichst sicheren und gut bezahlten Arbeitsplatz zu ergattern. Detaillierte Kenntnisse im Bereich der Marktforschung sind hier von Vorteil, da es sich bei der Marktforschung um einen Wirtschaftszweig handelt, der in den letzten Jahren kontinuierlich gewachsen ist und damit eine der wenigen stabilen Wachstumsbranchen im europäischen Markt darstellt.

Insbesondere im Zuge der immer wichtiger werdenden Online-Marktforschung ist es wichtig, neben grundlegenden Kenntnissen über statistische Verfahren auch über wesentliche Methodenkenntnisse zu verfügen, da die Umsetzung von traditionellen „Offline-Befragungen" in der Online-Welt mit vielen Einschränkungen und Problemen behaftet ist. Aber nicht nur online, auch offline sind theoretische Kenntnisse unabdingbar, wenn man saubere und wissenschaftlich korrekte Marktforschung betreiben will.

Dieses Wissen, der Spaß an der Arbeit mit Zahlen, sowie die beruflichen Chancen, die sich guten Marktforscherinnen und Marktforschern momentan eröffnen, sind dann hoffentlich auch Ansporn genug, die Motivation über die nächsten 200 Seiten aufrecht zu erhalten.

3 Welche Inhalte werden in diesem Skript vermittelt?

Ziel dieses Skriptes ist eine auf die Bedürfnisse des Praktikers (und des sich auf die Fachprüfungen vorbereitenden Studenten) zugeschnittene Darstellung von vier wichtigen multivariaten Analyseverfahren, die in der Marktforschung zum Einsatz kommen: der explorativen Datenanalyse, der multiplen Regression, der Varianzanalyse und der Faktorenanalyse. Neben einigen theoretischen Grundlagen wird dabei insbesondere auf die Umsetzung der Verfahren mit der bekannten Statistiksoftware SPSS und die Interpretation der SPSS-Ergebnistabellen eingegangen. Nachfolgend findet sich eine Kurzdarstellung der einzelnen Verfahren zur ersten Orientierung.

Explorative Datenanalyse

Die explorative Datenanalyse ist in der Regel das erste Verfahren, dem zur Analyse vorliegende Daten unterzogen werden und zudem auch das einzige Verfahren, das im Verlauf beinahe jeder Analyse zum Einsatz kommt. Ziel der explorativen Datenanalyse ist es, dem Marktforscher einen ersten Überblick über die vorliegenden Daten zu ermöglichen, sowie nach bestimmten Auffälligkeiten und Regelmäßigkeiten in der Verteilung eines Merkmals oder mehrerer Merkmale zu suchen. Die Kernfrage der explorativen Datenanalyse lässt sich daher auch so formulieren: Was ist an der Verteilung eines Merkmals bemerkenswert?

Um diese Frage beantworten zu können, kann anhand verschiedener statistischer Lagemaße wie dem arithmetischen Mittel, dem Median oder dem Modus bzw. statistischer Streumaße wie der Varianz, der Spannweite oder dem IQR versucht werden, einen ersten Einblick in die Lage und Verteilung des vorliegenden Merkmals zu erhalten. Dieser Einblick kann durch eine Vielzahl grafischer Darstellungsformen ergänzt werden, wobei zwischen den Möglichkeiten für die grafische Darstellung diskreter und stetiger Daten zu unterscheiden ist.

Doch nicht nur der Gesamtüberblick ist von Bedeutung. Mitunter sind es gerade einzelne Werte, die das Ergebnis weiterführender Analyseverfahren entscheidend beeinflussen können. Hier sind vor allem die Ausreißer von Bedeutung, einzelne Werte, die besonders groß oder klein ausfallen und nicht zum Rest der Verteilung zu gehören scheinen. Ebendiese Ausreißer müssen gefunden, analysiert und für die weitere Analyse entweder beibehalten oder aus dem Datensatz entfernt werden. Aber auch fehlende Werte, also leere Zellen in der SPSS-Datentabelle, können für den Marktforscher von Interesse sein, da sie auf mögliche Probleme während der Datenerhebung hinweisen. Daher muss auch das Auftreten fehlender Werte entsprechend festgestellt, analysiert und bewertet werden. All dies geschieht im Rahmen der explorativen Datenanalyse.

Zuguterletzt ist vor der Anwendung weiterführender Analyseverfahren wie der multiplen Regression oder der Varianzanalyse häufig zu überprüfen, ob die vorliegenden Daten gewisse Grundvoraussetzungen erfüllen, wie beispielsweise das Vorliegen einer Normalverteilung in

der Grundgesamtheit, gleichbleibende Varianzen in verschiedenen Untergruppen einer Merkmalsverteilung (die sogenannte Homoskedastizität) oder das Auftreten von linearen Zusammenhängen zwischen zwei Variablen. Für die Überprüfung dieser Voraussetzungen liegen verschiedene Testverfahren und -kriterien vor, die ebenfalls zum Methodenspektrum der explorativen Datenanalyse gezählt werden können.

Ein festgelegter „Ablaufplan“, der bei jeder explorativen Datenanalyse einzuhalten wäre, existiert nicht. Je nach Art der Daten und nach den Voraussetzungen der eingeplanten weiteren Analyseverfahren sind andere Kennwerte, andere grafische Darstellungsformen und andere Testverfahren anzuwenden. Der Marktforscher bestimmt im Rahmen einer explorativen Datenanalyse sein Vorgehen daher bis zu einem gewissen Grad selbst, wobei ihm ein beträchtlicher Methodenvorrat zur Verfügung steht. Lediglich die Ausreißeranalyse und die Analyse der fehlenden Werte sind aus naheliegenden Gründen stets durchzuführen.

Multiple Regression

Bei der multiplen Regression handelt es sich um ein äußerst vielseitiges, strukturenprüfendes Verfahren, dass für den Marktforschungs-Studenten von besonderer Relevanz ist, da es von allen multivariaten Analyseverfahren in der Praxis am häufigsten eingesetzt wird.

Das Ziel des Verfahrens ist die Analyse von Beziehungen zwischen einer abhängigen Variablen und einer (univariater Fall) oder mehreren (multivariater Fall) unabhängigen Variablen. Die multiple Regressionsanalyse wird in der Praxis hauptsächlich für die Beschreibung und Erklärung von Zusammenhängen und die Durchführung von Prognosen eingesetzt.

Eine typische Fragestellung der Regressionsanalyse könnte beispielsweise lauten: Hängt die Absatzmenge eines bestimmten Produkts von den Ausgaben für die Qualitätssicherung, den Ausgaben für die Werbung oder bzw. und der Anzahl der Verkaufsstellen ab? Wenn dies der Fall ist, wie stark stellen sich die jeweiligen Zusammenhänge dar? Wie wird sich die Absatzmenge entwickeln, wenn bestimmte Ausgaben erhöht oder gesenkt werden?

Da der Informationswert eines multiplen Regressionsmodells sehr hoch ist, bringt die Re-

gressionsanalyse unter allen im Rahmen dieses Skriptes betrachteten Analyseverfahren die meisten und detailliertesten Voraussetzungen mit sich. Nur selten lassen sich diese in der Praxis vollständig erfüllen und häufig liegt es in der Verantwortung des Marktforschers, auf der Basis seiner Erfahrungen und seiner Methodenkenntnisse eine Entscheidung über Abbruch oder Fortsetzung des Verfahrens zu treffen.

Varianzanalyse

Die Varianzanalyse gehört ebenso wie die Regressionsanalyse zu den strukturenprüfenden Verfahren und dient der Feststellung von Mittelwertsunterschieden einer abhängigen Variablen zwischen zwei oder mehr durch verschiedene Faktoren definierten Gruppen von Merkmalsträgern. Das Vorliegen solcher Mittelwertsunterschiede weist auf einen Zusammenhang zwischen den gruppenbildenden Faktoren (den unabhängigen Variablen) und der abhängigen Variablen hin, bei der die Unterschiede auftreten.

Eine typische Fragestellung der Varianzanalyse könnte daher lauten: Haben verschiedene Unterrichtsmethoden einen Einfluss auf die Prüfungsergebnisse. Wenn alle Schülergruppen, die mit den unterschiedlichen Methoden unterrichtet wurden, die gleiche Prüfung ablegen müssen, finden sich dann Unterschiede zwischen den einzelnen Gruppen bezüglich der durchschnittlichen Prüfungsergebnisse in verschiedenen Kategorien und Unterkategorien?

Solche und ähnliche Fragestellungen werden in der Varianzanalyse beantwortet. Das mathematische Prinzip hinter dem Verfahren ist vergleichsweise einfach: Es wird getestet, ob die Varianz der abhängigen Variablen innerhalb der durch die unabhängigen Variablen gebildeten Gruppen größer ist als zwischen diesen. Das Verhältnis von gruppeninterner und gruppenexterner Varianz zur Gesamtvarianz ermöglicht eine Aussage darüber, ob sich die einzelnen Gruppen bezüglich der abhängigen Variablen wirklich voneinander unterscheiden – mit anderen Worten, es wird überprüft, ob die Einteilung der Gruppen anhand der unabhängigen Variablen „gerechtfertigt“ ist.

Generell ist dabei zwischen der ANOVA – der Analysis of Variance mit nur einer abhängigen Variablen – und der MANOVA – der Multivariate Analysis of Variance mit minde-

stens zwei abhängigen Variablen – zu unterscheiden. Im Rahmen dieses Manuskripts wird hauptsächlich auf die ANOVA eingegangen, die aus der Sicht des für die Fachprüfungen lernenden Studenten sicherlich das relevantere Verfahren ist, während die MANOVA nur am Rande betrachtet wird.

Faktorenanalyse

In der Marktforschung hat man es häufig mit recht komplexen Untersuchungsgegenständen und Sachverhalten zu tun. Begriffe wie „Nutzen“ oder „Qualität“ lassen sich nicht durch eine einzige Variable ausdrücken. Würde man Personen nach der „Qualität“ eines Produktes befragen, kann man sich leicht vorstellen, dass die erste Person unter „Qualität“ die Haltbarkeit versteht, die zweite den Begriff aber mit dem „Preis-Leistungs-Verhältnis“ verbindet. Hochgradig komplexe Sachverhalte können daher nicht in simplen Variablen erhoben werden. Statt dessen ist oft ein ganzes Bündel von Variablen notwendig, um beispielsweise die „Qualität“ eines Produktes abbilden zu können: Zuverlässigkeit, Sicherheit, Haltbarkeit, Preis-Leistungs-Verhältnis, Lieferzeit....

Aber dennoch gibt es sie ja, die komplexen Hintergrundvariablen, die Hintergrundfaktoren, und oft kann es für eine marktforscherische Analyse auch hilfreich sein, den umgekehrten Weg zu gehen, und einmal zu analysieren, auf welche „großen Nenner“, auf welche komplexen Hintergrundfaktoren sich eine Vielzahl an betrachteten Variablen reduzieren lässt. Dies geschieht im Rahmen der Faktorenanalyse.

Die Faktorenanalyse gehört somit zu den dimensionsreduzierenden Verfahren. Ihr Ziel ist die Reduktion eines ganzen Bündels von Variablen und Einflussgrößen auf wenige, wesentliche Hintergrundfaktoren, in denen sich aber dennoch sämtliche Zusammenhänge und Abhängigkeiten zwischen diesen Einzelvariablen wiederspiegeln sollen. Solche Hintergrundfaktoren erleichtern das Verständnis und vereinfachen die weitere Analyse – und sie helfen auch, solche Variablen zu identifizieren, die für den untersuchten Sachverhalt irrelevant sind. Der Marktforscher kann also zunächst einmal alle interessant erscheinenden Variablen erheben und dann anschließend über die Faktorenanalyse herausfinden, welche dieser Variablen wirklich in sein Modell gehören.

4 Welche Kenntnisse aus Statistik I&II werden benötigt?

Welche Vorkenntnisse aus Statistik I & II sind notwendig, um den vorliegenden Stoff nachvollziehen zu können?

Da die Inhalte der Vertiefungsrichtung wesentlich auf den in Statistik I & II geschaffenen Grundlagen aufbauen, ist es in jedem Fall von Vorteil, noch mit dem Stoff vertraut zu sein. Insbesondere die statistischen Grundlagen und wesentlichen Grundbegriffe können im Rahmen dieses Skriptes nicht mehr im Detail wiederholt werden. Aber auch Kenntnisse weiterführender statistischer Verfahren wie der Berechnung von Korrelationskoeffizienten oder Konfidenzintervallen oder die Grundideen hinter statistischen Tests sind für das Verständnis der hier behandelten multivariaten Verfahren von Vorteil.

An dieser Stelle einen kompletten Überblick der Inhalte dieser beiden Semester geben zu wollen, würde den Rahmen dieses Skriptes natürlich bei weitem sprengen. Statt dessen sei auf eine kurze Liste von Schlagworten verwiesen, von denen zumindest die meisten Ihnen noch vertraut erscheinen sollten:

- Konfidenzintervalle
- Empirische Verteilungsfunktion
- Diskrete und stetige Verteilungen
- Absolute und relative Häufigkeiten
- Statistische Tests (t-Test, F-Test...)
- Korrelationskoeffizienten (Bravais-Pearson...)
- Statistische Lagemaße (Mittel, Median, Modus....)
- Kombinatorik (Permutation, Variation, Kombination)
- Dispersionsparameter (Varianz, Standardabweichung....)
- Statistische Grundbegriffe (Grundgesamtheit, Stichprobe...)
- Skalenniveaus (Nominalskala, Ordinalskala, Kardinalskala)

Falls Sie mit diesen Begriffen nichts mehr anzufangen wissen, seien Ihnen die beiden Bücher von Prof. Lammers sowie die zwei nachfolgenden Werke empfohlen.

Bleymüller, J.; Gehlert, G. & Gülicher, H. (2000). Statistik für Wirtschaftswissenschaftler. München: Verlag Vahlen

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (1999). Statistik. Der Weg zur Datenanalyse (2. Aufl.). Berlin: Springer.

III Weiterführende Literatur

Bleymüller, J.; Gehlert, G. & Gülicher, H. (2000). Statistik für Wirtschaftswissenschaftler. München: Verlag Vahlen

Dannenberg, M. & Barthel, S. (2004). Effiziente Marktforschung. Frankfurt/Wien: Wirtschaftsverlag Carl Ueberreuther

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (1999). Statistik. Der Weg zur Datenanalyse (2. Aufl.). Berlin: Springer.

Koch, J. (1997). Marktforschung - Begriffe und Methoden. München: R.Oldenbourg Verlag

B Explorative Datenanalyse

I Einführung

Zu Beginn jeder Datenanalyse ist es sinnvoll, sich zunächst einen Überblick über die vorliegenden Daten bzw. Verteilungen zu machen. Diesem Zweck dient die explorative Datenanalyse, die dem Marktforscher ein umfangreiches Sortiment an Kennwerten, Grafiken und Analysemethoden zur Verfügung stellt.

Zur allgemeinen Erstanalyse gehört insbesondere die Darstellung von Lage und Verteilung der Werte – anhand verschiedener Kennwerte und Grafiken lassen sich so Auffälligkeiten in den Daten feststellen. Zu den für die Marktforschung relevanten Lagemaßen gehören das arithmetische Mittel, der Median und die Perzentile sowie der Modalwert, zu den relevanten Streumaßen die Spannweite, der Interquartilsabstand, die Varianz und die Standardabweichung. Ergänzend zu diesen Kennwerten lassen sich verschiedene grafische Darstellungen zur Suche nach Auffälligkeiten einsetzen, darunter Balken-, Kreis- und Stabdiagramme, Stem-and-Leaf-Plots, Histogramme und Box-Plots.

Die Ausreißeranalyse, also die Identifikation von und der Umgang mit extrem großen oder kleinen Werten in der Verteilung, gehört ebenfalls zum Repertoire der explorativen Datenanalyse. Zu klären ist unter anderem, inwiefern Ausreißer auf Fehler bei der Erhebung und Speicherung der Daten, außergewöhnliche Umstände oder aber eventuell untersuchungsgefährdende Mängel im Erhebungsdesign zurückzuführen sind. Verzerren die Ausreißer die Ergebnisse der Datenanalyse oder haben sie auf diese keinen Einfluss? Inwiefern ist es möglich, sie für die Durchführung weiterer Analyseverfahren aus dem Datensatz zu entfernen oder als fehlende Werte zu kennzeichnen?

Ebenfalls in den Bereich der explorativen Datenanalyse fällt die Analyse und der Umgang mit sogenannten fehlenden Werten, d.h. nicht vorhandenen Werten im Datensatz. Ebenso wie Ausreißer können auch fehlende Werte ein Hinweis auf Fehler bei der Durchführung der Erhebung oder Mängel im Erhebungsdesign sein. Von Interesse ist vor allem die Frage, ob die

Wahrscheinlichkeit des Auftretens von fehlenden Werten gewissen Regelmäßigkeiten folgt oder nicht, also ob fehlende Werte zufällig auftreten oder im Zusammenhang mit einer für die Untersuchung relevanten Größe stehen. Dies kann beispielsweise dann der Fall sein, wenn bei der Frage nach der Höhe des Einkommens die Bezieher niedriger Einkommen in stärkerem Maße die Auskunft verweigern, was selbstverständlich dazu führen muss, dass aus den tatsächlich erhobenen Daten falsche Rückschlüsse, beispielsweise auf die Höhe des Durchschnittseinkommens der Befragten, gezogen werden.

Eine der Aufgaben, die ebenfalls dem Bereich der explorativen Datenanalyse zugeordnet werden, ist die Überprüfung bestimmter Voraussetzungen für weiterführende Analyseverfahren. So ist es beispielsweise für einige der nachfolgend dargestellten Verfahren von Bedeutung, ob die Verteilung von Daten in der Grundgesamtheit einer Normalverteilung folgt, oder ob bezüglich einer erhobenen Größe in verschiedenen Untergruppen der Grundgesamtheit Homoskedastizität, d.h. Gleichheit der Varianzen, vorliegt. Diesbezügliche Tests werden zumeist im Rahmen der einzelnen Verfahren gesondert durchgeführt, deren theoretischer Hintergrund sowie deren praktische Durchführung werden aber in diesem Abschnitt erschöpfend dargestellt, so dass in nachfolgenden Abschnitten nur noch zurückverwiesen werden muss.

Kurz angerissen wird noch die Erstellung sogenannter Dummy-Variablen, Variablen, die immer dann eingesetzt werden, wenn das intendierte Verfahren metrisch skalierte Werte voraussetzt, aber dennoch nominalskalierte oder ordinalskalierte Werte in die Berechnung eingehen sollen, wobei eine solche Verletzung der Voraussetzungen selbstverständlich nur unter besonderen Umständen empfohlen werden kann.

Anzumerken sei noch, dass nie explizit festgelegt wurde, welche Kennzahlen, Grafiken und Tests bzw. Untersuchungen zu einer explorativen Datenanalyse „dazugehören“. Statt dessen sind in der Praxis durch den Marktforscher die je nach Art der Daten und nach den Voraussetzungen der im Anschluss geplanten Analysen geeigneten Elemente und Verfahren aus dem Repertoire der explorativen Datenanalyse auszuwählen. Weitestgehender Konsens besteht aber darin, dass die Untersuchung eventuell vorliegender Ausreißer oder fehlender Werte sicherheitshalber in jedem Fall durchgeführt werden sollte.

II Lagemaße

1 Was sind Lagemaße?

Lagemaße geben Auskunft über das Zentrum einer Verteilung. Sie vermitteln einen Eindruck von der Höhe sowie zum Teil auch von der Verteilung der Werte. Im Rahmen der explorativen Datenanalyse spielen das arithmetische Mittel, der Median, der Modus und die Perzentilwerte eine Rolle, auf die im Zusammenhang mit dem Median kurz eingegangen wird.

2 Das arithmetische Mittel

```
The average statistician ist just plain mean.
```

Das arithmetische Mittel ist das mit weitem Abstand bekannteste statistische Lagemaß, weswegen es auch als Standardmittelwert bezeichnet wird. Es ist ausschließlich für metrisch skalierte Daten berechenbar (Intervallskala, Verhältnisskala).

Hierbei ist zu beachten, dass SPSS grundsätzlich auch das arithmetische Mittel aus nominalskalierten und ordinalskalierten Daten berechnet – es ist also rein softwaretechnisch auch möglich, durchschnittliche Telefonnummern oder ähnlich sinnlose Kennwerte zu berechnen. Hier sind daher die methodischen Kenntnisse des Anwenders gefragt, der das Skalenniveau erkennen und entscheiden muss, ob die Berechnung des arithmetischen Mittels sinnvoll vorgenommen werden kann. Dies gilt in besonderem Maße auch für weiterführende Berechnungen, die auf dem arithmetischen Mittel aufbauen oder dieses mit einschließen.

Liegen von einem metrischen Merkmal x insgesamt n Werte vor, berechnet sich das arithmetische Mittel wie folgt:

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

Es wird also die Gesamtsumme aller Merkmalsausprägungen gebildet und durch die Anzahl der Merkmalsausprägungen geteilt. Die Gesamtsumme aller Abweichungen (der einzelnen Merkmalsausprägungen) vom arithmetischen Mittel beträgt daher stets Null, was, wie wir nachfolgend sehen werden, einen Einfluss auf die Berechnung der Streuung um das arithmetische Mittel hat.

Das arithmetische Mittel ist nicht robust, d.h. sehr empfindlich gegenüber Ausreißern. Wird zusammen mit drei Dutzend Normalverdienern auch ein Millionär befragt, wird dessen Verdienstangabe den „Durchschnittsverdienst“ deutlich nach oben treiben. Die in diesem Abschnitt ebenfalls dargestellte Ausreißeranalyse ist daher unbedingt durchzuführen, wenn im weiteren Verlauf mit dem arithmetischen Mittel gearbeitet werden soll.

Des weiteren ist zu beachten, dass das arithmetische Mittel nur dann sinnvoll interpretiert werden kann, wenn auch Informationen über die Streuung der Werte vorliegen. Stellt sich beispielsweise bei einer Untersuchung heraus, dass das Durchschnittseinkommen bei 2.500 € liegt, so ist festzustellen, ob die Einkommenswerte generell in der Nähe des Durchschnittswertes liegen, oder zu beiden Richtungen stark abweichen. Die Beachtung der Streuung ist insbesondere für Vergleiche von Mittelwerten von Interesse, damit aus der relativen Übereinstimmung oder der relativen Abweichung zweier Werte keine falschen Schlüsse gezogen werden.

Die Bedeutung der Streuung für die Interpretation des Mittelwertes zeigt sich an folgendem, zu Beispielzwecken stark simplifizierten Fall der Lebenserwartung im Mittelalter: Wie allgemein bekannt ist, wurden die Menschen des Mittelalters „im Durchschnitt“ nicht so alt wie heutige Einwohner westlicher Staaten, das „Durchschnittsalter“ lag um 1500 bei etwa 32 Jahren. Daraus ließe sich eventuell der Schluss ziehen, dass Menschen jenseits der 30 schon froh waren, ein hohes Alter erreicht zu haben und mit 40 oder gar 50 als wahre Methusalems gegolten haben mussten – dem ist allerdings nicht so. Auch um 1500 gab es 70- oder 80jährige Menschen, die keineswegs eine Seltenheit waren, auch wenn ihr Gesamtanteil an der Bevölkerung sicher deutlich geringer war, als dies heute der Fall ist. Der Grund für die niedrige durchschnittliche Lebenserwartung ist vor allem in der hohen Kindersterblichkeit zu suchen, auch wenn die verbesserten hygienischen Bedingungen sowie die medizinischen Fort-

schritte wie viele andere Faktoren einen Anteil an der Entwicklung haben. Letzten Endes senkt aber eine hohe Kindersterblichkeit stets die durchschnittliche Lebenserwartung, so dass vermutlich nur die wenigsten Menschen des 16. Jahrhunderts mit oder um die 32 verstarben. Dies zeigt, dass der Standardmittelwert stets mit Sorgfalt und unter Berücksichtigung der Streuung sowie anderer Faktoren, wie beispielsweise der Ausreißer oder auch fehlender Werte zu interpretieren ist.

Bei der Interpretation des arithmetischen Mittels sollte der Marktforscher außerdem stets im Hinterkopf haben, wie sich dieses berechnet. So lässt sich anhand der oben dargestellten Formel leicht nachweisen, dass die Mehrheit der Deutschen überdurchschnittlich viele Augen im Kopf hat: Da nämlich unter 80.000.000 Bundesbürgern auch etwa 20.000 Einäugige leben, ergibt sich bei der Berechnung der durchschnittlichen Augenzahl folgendes:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{80000000}\sum_{i=1}^{80000000} (79980000 * 2 + 20000 * 1) = 1{,}9975$$

Da die meisten Deutschen aber immer noch zwei Augen haben...die Schlussfolgerungen seien jedem selbst überlassen. Festzustellen bleibt aber, dass das arithmetische Mittel nur dann sinnvoll und richtig interpretiert werden kann, wenn man es nicht einfach als „Durchschnitt“ hinnimmt, sondern während der Interpretation nie vergisst, wie der Wert berechnet wird bzw. welche Effekte auftreten können. Insbesondere ist es wichtig, den Datensatz vor der Berechnung des Standardmittels auf eventuelle Ausreißer zu überprüfen, ein Verfahren, welches weiter unten noch im Detail dargestellt wird.

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert
Gehalt	474	15.750	135.000	34.419,57
Gültige Werte (Listenweise)	474			

Abbildung 2: Ausgabe des arithmetischen Mittels in SPSS

3 Der Median

Beim sogenannten Median handelt es sich um den Wert, der in der Mitte der geordneten Verteilung liegt. Seine Berechnung setzt lediglich ordinalskalierte Daten voraus, d.h. nur für nominalskalierte Daten kann kein Median gebildet werden, da sich diese per Definition nicht in eine geordnete Reihenfolge bringen lassen.

Bei einer ungeraden Anzahl von Werten wird einfach der mittlere Wert der geordneten Reihe gewählt:

$$x_{med} = \frac{x_{(n+1)}}{2}$$

Bei einer geraden Anzahl von Werten wird das arithmetische Mittel der beiden zentralen Werte berechnet:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

Der Median ist äußerst robust, d.h. er wird von Ausreißern nicht beeinflusst. Der weiter oben befragte einzige Millionär liegt am Rand der Verteilung. Verzehnfacht sich sein Einkommen, hat dies keinen Einfluss auf die Lage des mittleren Wertes und damit auf den Median der Verteilung, während sich das arithmetische Mittel deutlich verlagern würde.

Der Median ist zugleich auch das 50%-Perzentil. Gemeinsam mit dem 25%-Perzentil und dem 75%-Perzentil gehört er zu den sogenannten Quartilen, die eine in der Praxis häufig durchgeführte Aufteilung der Verteilung in vier Wertebereiche ermöglichen. Die Quartile werden unter anderem zur Konstruktion der Box-Plots benötigt, die ebenfalls in diesem Abschnitt dargestellt werden.

Statistiken

Anfangsgehalt

N	Gültig	474
	Fehlend	0
Mittelwert		17.016,09
Median		15.000,00
Minimum		9.000
Maximum		79.980

Abbildung 3: Ausgabe des Medians in SPSS

4 Der Modus

Beim Modus handelt es sich um den in den vorliegenden Daten am häufigsten auftretenden Wert, bei klassierten Daten um die Klassenmitte der Klasse, welche die meisten Fälle auf sich vereint (nur bei gleichbreiten Klassen möglich). Der Modus ist somit das am einfachsten zu berechnende Lagemaß und zugleich das einzige Lagemaß, welches auch für nominalskalierte Daten sinnvoll angegeben werden kann.

Die Berechnung des Modus ist in der Regel nur für diskrete Daten sinnvoll, da sich die Werte bei stetigen Daten kaum wiederholen (die Punktwahrscheinlichkeit liegt bei einer stetigen Verteilung ja sogar per Definition bei Null), so dass es keine häufig in der Verteilung auftretenden Werte gibt. Soll der Modus auch für stetige Daten gebildet werden, so ist eine Aufteilung der Daten in (gleichbereite) Klassen vorzunehmen und anschließend der Modus für klassierte Daten zu bilden.

Die Vorteile des Modus liegen vor allem in seiner einfachen Berechnung und der Tatsache, dass er auch für niedrigstskalierte Daten noch Gültigkeit besitzt. Nachteilig sind dagegen der vergleichsweise geringe Informationsgehalt und die teils schlechte Interpretierbarkeit des Wertes, denn sinnvoll ist der Modus im Grunde nur dann zu interpretieren, wenn ein einfaches, klares Maximum vorliegt – die Verteilung also unimodal ist (im Balkendiagramm beispielsweise zu erkennen an einem deutlich hervorragenden einzelnen Balken).

Handelt es sich dagegen um eine multimodale Verteilung und kommen mehrere Werte mit der gleichen Häufigkeit darin vor, kann auch SPSS den Modus nicht mehr eindeutig ermitteln

und gibt aus der Reihe der möglichen Modi einfach den Wert aus, der in der Urliste an oberster Stelle steht. Sobald eine solche multimodale Verteilung vorliegt ist der Modus daher nicht mehr eindeutig zu interpretieren, auch wenn es marginale Unterschiede in den Häufigkeiten gibt und sich ein Wert als mathematisch eindeutiger Modus festlegen lässt.

Statistiken

Anfangsgehalt

N	Gültig	474
	Fehlend	0
Mittelwert		17.016,09
Median		15.000,00
Modus		15.000
Minimum		9.000
Maximum		79.980
Perzentile	25	12.450,00
	50	15.000,00
	75	17.617,50

Abbildung 4: Ausgabe des Modus in SPSS

5 Lagemaße und Verteilungsform

Berechnet man alle drei der hier dargestellten Lagemaße, so lassen sich aus deren Vergleich noch zusätzliche Informationen zur Form der Verteilung gewinnen. Sind alle drei Lagemaße in etwa identisch, handelt es sich um eine symmetrische Verteilung, ist das arithmetische Mittel am größten, gefolgt von Median und Modus so ist die Verteilung linkssteil bzw. rechtsschief, im umgekehrten Fall ist sie rechtssteil bzw. linksschief. Dieser Zusammenhang wird in der nachfolgenden Tabelle übersichtsweise wiedergegeben.

Verhältnis der Lagemaße zueinander	***Form der vorliegenden Verteilung***
$\bar{x} \approx x_{med} \approx x_{mod}$	Symmetrische Verteilung
$\bar{x} > x_{med} > x_{mod}$	Linkssteile, rechtsschiefe Verteilung
$\bar{x} < x_{med} < x_{mod}$	Rechtssteile, linksschiefe Verteilung

Tabelle 1: Lagemaße und Verteilungsform

III Streuungsmaße

1 Was sind Streuungsmaße?

Zusammen mit den Lagemaßen vermitteln die Streuungsmaße dem Analytiker einen ersten Eindruck von einer Verteilung, indem sie zeigen, wie stark die Werte einer Verteilung um deren Zentrum (in der Regel angegeben durch das arithmetische Mittel) streuen. Für die explorative Datenanalyse sind Spannweite, Varianz und Standardabweichung sowie der Interquartilsabstand von Bedeutung, der zwar nicht unmittelbar durch SPSS berechnet werden kann, aber in die Konstruktion der Box-Plots einfließt.

2 Die Spannweite

Die Spannweite wird gebildet, indem die Differenz zwischen dem größten (Maximum) und dem kleinsten (Minimum) Wert in der vorliegenden Verteilung berechnet wird. Sie ist als Maß für die Streuung absolut ungenügend, da ihr Informationsgehalt recht niedrig ist (es handelt sich ja nur um den Abstand zweier Werte aus allen insgesamt vorliegenden Werten, d.h. es fließen nur sehr wenige der vorhandenen Informationen in die Berechnung ein) und sie zudem extrem stark durch Ausreißer beeinflusst wird und die wahre Streuung daher nur verzerrt wiedergibt.

Dies ist leicht vorstellbar: Gibt es in der Verteilung auch nur einen Ausreißer, so wird dieser entweder als größter oder als kleinster Wert in die Berechnung der Spannweite mit eingehen, gibt es Ausreißer an beiden Enden, so wird die Spannweite nur noch durch die Ausreißer bestimmt, was gegen ihre Zuverlässigkeit spricht. Ebenso wie das arithmetische Mittel ist die Spannweite daher äußerst unrobust: Der bereits beispielhaft verwendete einzige befragte Millionär wird die Spannweite des Merkmals „Einkommen“ sogar ganz erheblich vergrößern.

Statistiken

Gehalt

N	Gültig	474
	Fehlend	0
Spannweite		119.250
Minimum		15.750
Maximum		135.000

Abbildung 5: Ausgabe der Spannweite in SPSS

3 Der Interquartilsabstand

Der Interquartilsabstand (IQR = Inter Quartile Range) ist der Abstand zwischen dem oberen (75%) und dem unteren (25%) Quartil. Da beide Quartile nicht von Ausreißen beeinflusst werden können, ist der Interquartilsabstand im Gegensatz zur Spannweite ein äußerst robustes Streuungsmaß. Er zeigt an, in welchem Wertebereich die 50% der Werte liegen, die sich zu gleichen Teilen um den Median als Zentrum der Verteilung anordnen.

Der Interquartilsabstand wird von SPSS nie unmittelbar ausgegeben, geht aber in die Konstruktion der in der Praxis äußerst beliebten Box-Plots ein, die weiter unten noch im Detail vorgestellt werden.

4 Varianz und Standardabweichung

Varianz und Standardabweichung sind die mit Abstand gebräuchlichsten und aussagekräftigsten Streuungsmaße. Da sich die Standardabweichung aus der Varianz ergibt, wird diese zuerst berechnet, und zwar als Summe der quadrierten Abweichungen der Einzelwerte vom arithmetischen Mittel, geteilt durch die Gesamtzahl aller Werte:

$$S^2 = \frac{1}{(N)} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Wie man an der Formel erkennen kann, handelt es sich im Grunde um nichts weiter als

das arithmetische Mittel der Abweichungen eben von jenem arithmetischen Mittel. Die Quadrierung der Abweichungen ist erforderlich, damit sich die positiven und die negativen Abweichungen nicht gegenseitig aufheben – denn die Summe aller Abweichungen vom arithmetischen Mittel ist ja, wie bereits oben beschrieben, stets Null.

Wer die von SPSS ausgegebenen Werte anhand der oben angegebenen Formel „per Hand" nachrechnet, wird leichte Abweichungen feststellen, die zum einen auf Rundungen und zum anderen darauf zurückzuführen sind, dass SPSS statt der Varianz die Stichprobenvarianz berechnet, statt der Gesamtzahl der Werte also die Freiheitsgrade der Verteilung (N-1) in die Berechnung eingehen:

$$S^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Betrachtet man beide Formeln genauer, so wird deutlich, dass die Varianz abnimmt, je dichter die beobachteten Einzelwerte am arithmetischen Mittel liegen. Würden alle Werte genau dem arithmetischen Mittel entsprechen ergäbe sich eine Varianz von Null – in einem solchen Fall wäre ja aber auch keinerlei Streuung zu beobachten. In diesem Sinne ist die Varianz eine geeignete Kennzahl für die Streuung der Werte um das Zentrum der Verteilung.

Statistiken

Gehalt

N	Gültig	474
	Fehlend	0
Standardabweichung		17.075,661
Varianz		291578214,453
Minimum		15.750
Maximum		135.000

Abbildung 6: Ausgabe von Varianz und Standardabweichung in SPSS

Neben der Varianz lässt sich bei SPSS auch die Standardabweichung mit ausgeben. Wieso wird nun dieses zusätzliche Streuungsmaß noch benötigt?

Da für die Berechnung der Varianz die Abweichungen der beobachteten Einzelwerte vom arithmetischen Mittel quadriert werden, wird die Varianz auch in quadrierten Einheiten wiedergegeben, also beispielsweise Stunden², €² oder $². Da ein solcher Wert nur sehr schwer zu interpretieren ist und seine Bedeutung sich dem Laien nur unter Mühen erschließt, wird in der Regel noch die Standardabweichung als positive Quadratwurzel der Varianz berechnet. Dieses Streuungsmaß ist dann wieder „richtig dimensioniert" und wesentlich leichter zu interpretieren – man betrachte nur das obige Beispiel. Mit der Angabe, dass die Gehälter im Schnitt 291.578.214 €² um das Durchschnittsgehalt streuen, lässt sich unmittelbar kaum etwas anfangen. Die Aussage, dass die Standardabweichung 17.075 € beträgt ist dagegen leichter nachvollzieh- und damit auch interpretierbar.

IV Grafische Darstellungsformen univariater Daten

1 Möglichkeiten der grafischen Darstellung

Welcher Vorteil ergibt sich aus der grafischen Darstellung von Verteilungen?

Nun, nicht umsonst lautet eine gängige Weisheit in der modernen Pädagogik „Bilder sind Schüsse ins Gehirn" und auch dass ein Bild „mehr sagt als 1000 Worte" ist eine altbekannte Tatsache. Insbesondere Personen mit einem ausgeprägten visuellen Verstand können grafische Darstellungen der Lage, der Streuung und der Form von Verteilungen besser nachzuvollziehen als die reine Angabe einiger der weiter oben besprochenen Lage- und Streuungsmaße. Aus diesem Grund gehören Grafiken in jeden statistischen Bericht und in jede Ergebnispräsentation – durch sie werden Verteilungen und Analyseerkenntnisse direkt greifbar und verständlich.

Zudem können einige Grafiken dem geübten Auge des Analytikers wertvolle Hinweise darauf liefern, wie in der weiteren Analyse zu verfahren ist: Gibt es Ausreißer und wenn ja, wo liegen diese? Sind die Werte in der Grundgesamtheit in etwa normalverteilt? Welcher Wert tritt in einer diskreten Verteilung am häufigsten auf? All dies sind Fragen, die durch die richtige Grafik unmittelbar aufgeklärt werden können.

Bei der Betrachtung des äußerst umfassenden grafischen Repertoires zur Darstellung univariater Daten, welches SPSS dem Benutzer zur Verfügung stellt, ist zwischen diskreten Merkmalen mit wenigen Ausprägungen und stetigen Merkmalen mit vielen Ausprägungen zu unterscheiden. Eine Schnellübersicht der am häufigsten verwendeten Darstellungsformate findet sich in der nachfolgenden Grafik.

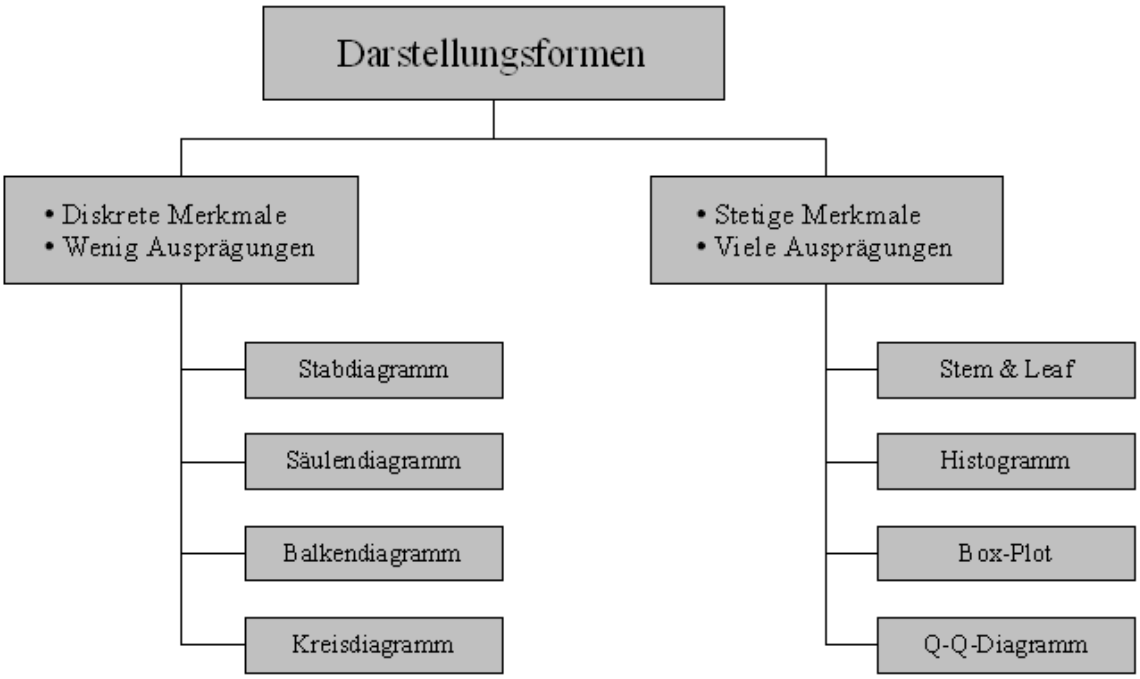

Abbildung 7: Grafische Darstellung univariater Daten

2 Säulen- und Balkendiagramme

Eine bekannte Form der grafischen Darstellung sind die Säulen- oder Balkendiagramme. Sie eignen sich primär für diskrete Merkmale mit einer geringen Anzahl an Ausprägungen. Liegt dagegen eine Verteilung mit vielen Ausprägungen vor, ergeben sich zu viele Säulen bzw. Balken, die vom Betrachter nicht mehr eindeutig interpretiert werden können. Sollen stetige Merkmale dennoch in einem Säulen- oder Balkendiagramm dargestellt werden, so sind die Daten entsprechend zu klassieren.

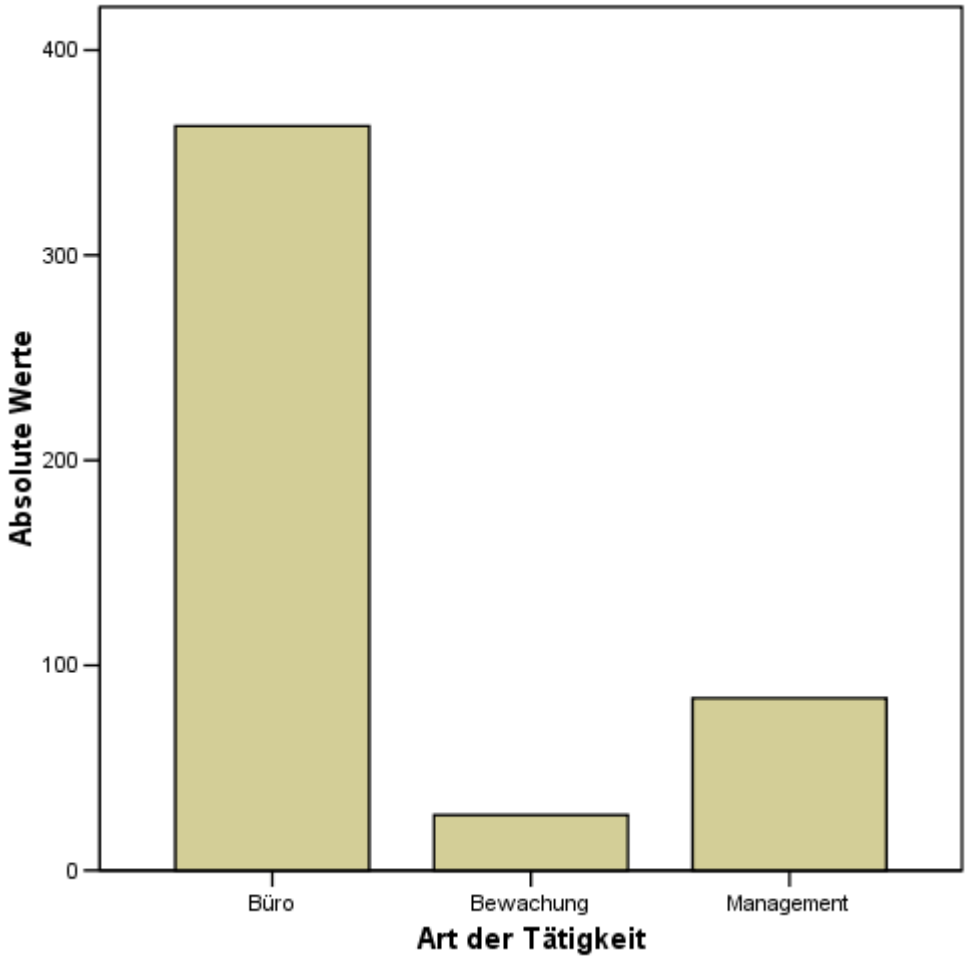

Abbildung 8: Balkendiagramm einer unimodalen, diskreten Verteilung in SPSS

3 Kreisdiagramme

Ebenso wie die Säulen- und Balkendiagramme sind auch die Kreisdiagramme primär für die grafische Darstellung diskreter Merkmalsverteilungen mit nur wenigen Ausprägungen geeignet, da sie sonst zu unübersichtlich ausfallen. Auch hier ist bei stetigen Merkmalen eine Klassierung der Daten nötig, damit das Diagramm sinnvoll interpretiert werden kann.

SPSS ermöglicht – dies gilt ebenfalls für die Säulen- und Balkendiagramme – die grafische Darstellung sowohl der absoluten als auch der relativen Häufigkeiten, wobei letzteres zu empfehlen ist, wenn Verteilungen miteinander verglichen werden sollen.

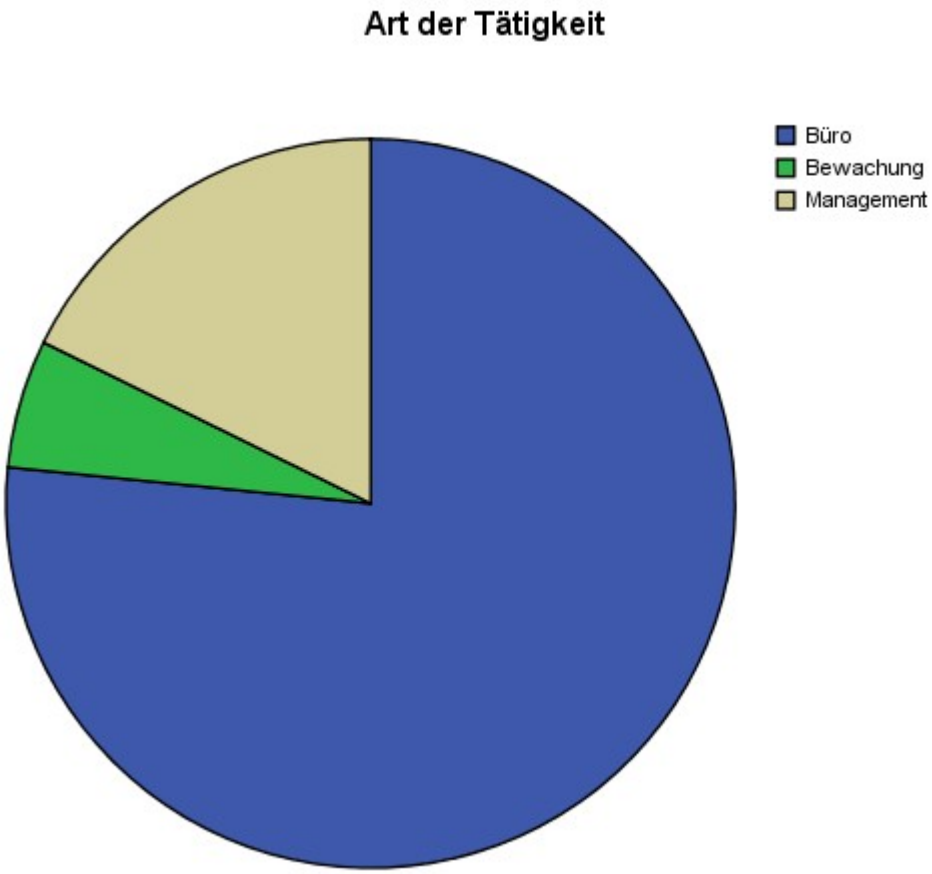

Abbildung 9: Kreisdiagramm einer unimodalen, diskreten Verteilung in SPSS

4 Histogramme

Ein Histogramm stellt die Häufigkeitsverteilung der Werte einer intervallskalierten Variablen dar. Dabei wird von den nach der Größe geordneten Daten ausgegangen, die in n Klassen aufgeteilt werden, welche theoretisch nicht die gleiche Breite besitzen müssen. Um die verzerrungsfreie Überlagerung eines Histogramms mit einer Vergleichsverteilung zu ermöglichen (eine Form des Verteilungsvergleichs auf die weiter unten beim Test auf Normalverteilung noch im Detail eingegangen werden wird), erstellt SPSS Histogramme stets mit gleichbreiten Klassen. Über jeder dieser Klassen wird ein Rechteck konstruiert, dessen Flächeninhalt sich proportional zur absoluten bzw. relativen Häufigkeit der jeweiligen Klasse verhält, wobei die

Darstellung der relativen Häufigkeiten gängiger ist.

Das Histogramm eignet sich primär für die Darstellung von stetigen Merkmalen mit einer größeren Anzahl an Ausprägungen. Bei der Erstellung in SPSS ist neben der Möglichkeit, eine Vergleichsverteilung direkt in das Histogramm einzubinden weiterhin beachtenswert, dass maximal 21 Klassen gebildet werden können.

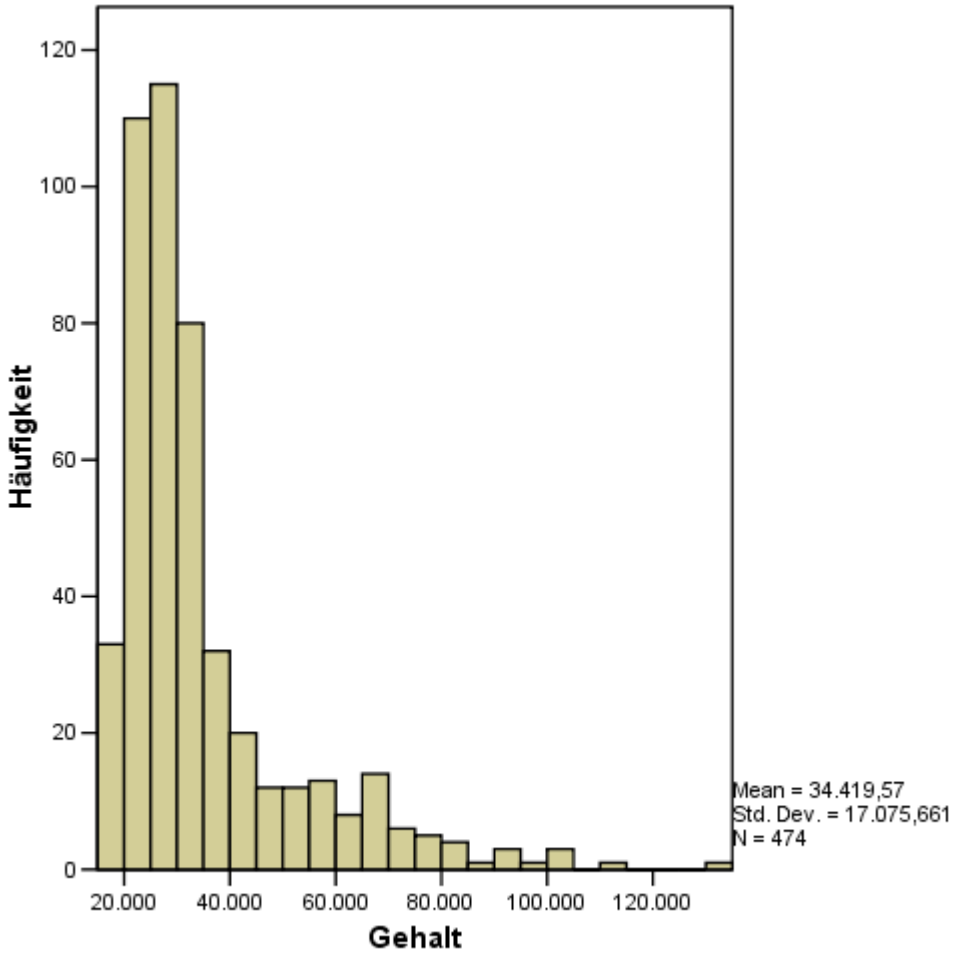

Abbildung 10: Histogramm mit gleichbreiten Klassen in SPSS

5 Stem-and-Leaf-Plots

Stem-and-Leaf-Plots (auch als Stamm-Blatt-Diagramme oder Stengel-Blatt-Diagramme bezeichnet) eignen sich ebenfalls zur Darstellung stetiger Merkmale. Der große Vorteil dieser

Darstellungsform besteht darin, dass die Originaldaten (bis zu einer gewissen Genauigkeit, da eventuell gerundet werden muss) noch aus der Grafik heraus gelesen werden können. Dies ermöglicht eine bessere Vergleichbarkeit von Verteilungen und gestattet zudem die bereits erwähnte Wiederherstellung der Originaldaten aus dem Diagramm heraus, die ausschließlich beim Stem-and-Leaf-Plot möglich ist.

Ein Stem-and-Leaf-Plot ist ähnlich aufgebaut wie ein seitlich gekipptes Histogramm. Es besteht aus einem Stamm, der sich aus der ersten Ziffer der in Klassen eingeteilten Werte zusammensetzt, sowie aus Blättern, die aus den auf die zweite Ziffer gerundeten restlichen Zahlenwerten bestehen. Sehr große oder kleine Zahlen werden bei SPSS als Extremwerte ausgewiesen, des weiteren wird noch die Stammbreite (stem width) und die Gültigkeit der einzelnen Blätter mit ausgegeben, da bei sehr vielen Werten ein Blatt auch für mehrere Fälle (cases) Gültigkeit haben kann.

Der nachfolgende Stem-and-Leaf-Plot gibt eine Verteilung mit folgenden Werten (in geordneter Reihenfolge) wieder: 11, 11, 11, 12, 12, 13, 14, 15, 17, 17, 22, 22, 24, 33, 33, 33, 34, 35, 38, 38, 41, 42, 49, 49, 49, 49, 124, 212

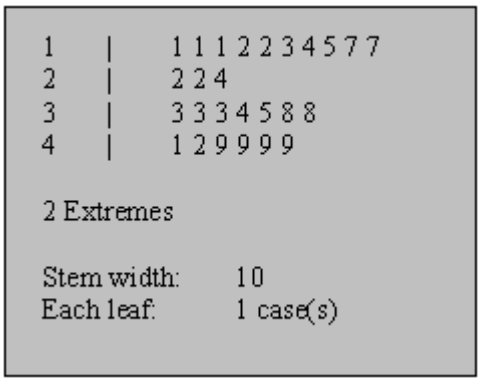

Abbildung 11: Einfacher Stem-and-Leaf-Plot

Die beiden größten Werte werden als Extremwerte ausgewiesen, die restlichen Zahlen lassen sich aus dem Diagramm ablesen. Der Darstellungsvorteil ist offensichtlich: im 40er-Stamm kann mit bloßem Auge erkannt werden, dass die meisten Werte dicht an der 50 liegen. Hätten sich statt 41, 42, 49, 49, 49, 49 die Zahlen 41, 41, 41, 42, 42, 43 in der Verteilung befunden, so wäre die deutlich andere Verteilung der Werte innerhalb der Klasse beispielsweise

in einem Histogramm oder einem Balkendiagramm nicht zum Ausdruck gekommen. Nur der Stem-and-Leaf-Plot gestattet es dem Analytiker, diesen Unterschied festzustellen. Dies ist einer der Gründe, warum sich der Stem-and-Leaf-Plot insbesondere im angloamerikanischen Raum so großer Beliebtheit erfreut und dort in vielen Lehrbüchern, aber auch öffentlichen Unternehmensbilanzen oder Veröffentlichungen von Regierungsseite zu finden ist.

Ein weiterer Grund für die Beliebtheit des Stem-and-Leaf-Plots liegt in der Möglichkeit, zwei Verteilungen auf übersichtliche Art und Weise gegenüberzustellen und sie direkt miteinander zu vergleichen. Einzige Voraussetzungen hierfür sind eine gleiche Stammbreite sowie die identische Gültigkeit der einzelnen Blätter. Sind diese Voraussetzungen erfüllt, lassen sich zwei Verteilungen an einem gemeinsamen Stamm darstellen, wobei die Blätter jeder Verteilung in eine andere Richtung „wachsen“.

Datensatz A		Datensatz B
88832	1	1112234577
21	2	224
954433	3	3334588
43321	4	129999
3 Extremes		2 Extremes
	Stem width:	10
	Each leaf:	1 case(s)

Abbildung 12: Stem-and-Leaf-Plot zweier Verteilungen

6 Box-Plots

Box-Plots gehören zu den wichtigsten Darstellungsformen in der explorativen Datenanalyse. Sie bieten einen direkten Verteilungsüberblick und eignen sich insbesondere zum Verteilungsvergleich. Dabei zeigen sie sowohl die Lage (die Quartile lassen sich unmittelbar aus dem Median ablesen) als auch die Streuung (IQR und Spannweite sind ebenfalls unmittelbar zu erkennen) der Werte und ermöglichen zudem die einfache Identifikation von Ausreißern.

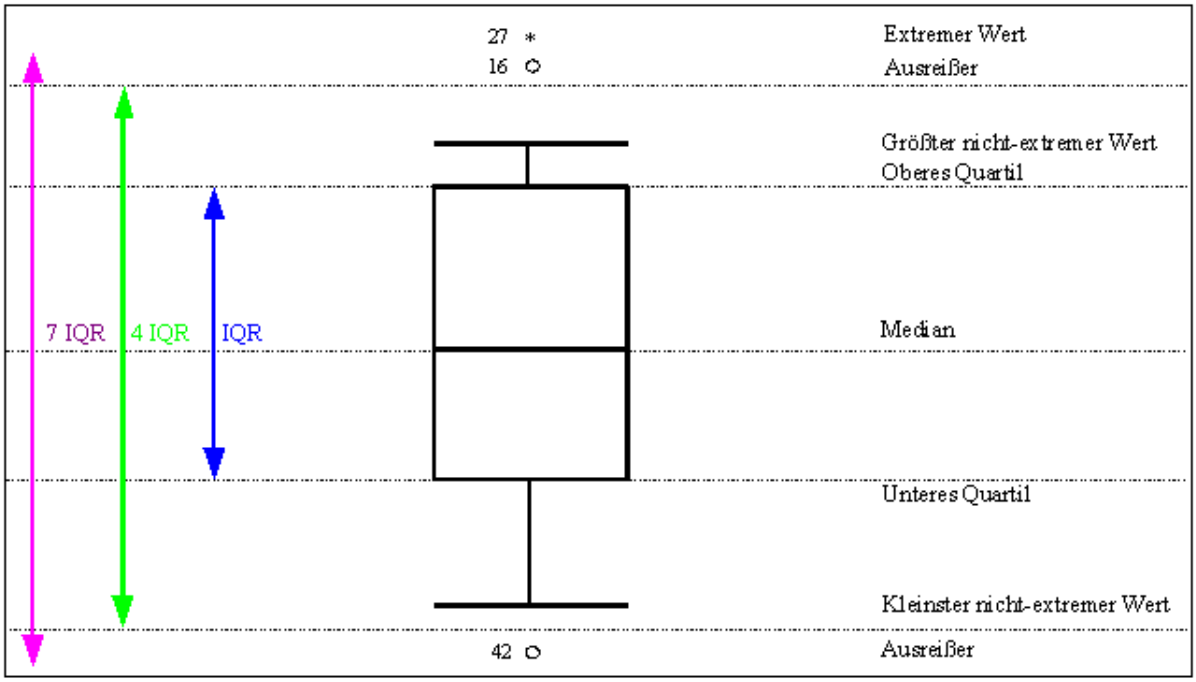

Abbildung 13: Genereller Aufbau eines erweiterten Box-Plots

Den generellen Aufbau eines sogenannten erweiterten Box-Plots verdeutlicht man sich am besten anhand der oben stehenden Grafik. Die simplifizierte Version des Box-Plots, der sogenannte einfache Box-Plot, weicht in seinem Aufbau vom erweiterten Box-Plot ab und wird aufgrund seiner geringen Bedeutung in der Praxis im Rahmen dieses Skriptes nicht weiter beachtet.

Die Box, das zentrale Rechteck des Box-Plots, verläuft vom oberen (75%) Quartil zum unteren (25%) Quartil, das mittlere (50%) Quartil, welches auch als Median bekannt ist, wird deutlich erkennbar in die Box eingezeichnet. Der Median liegt keineswegs immer in der Mitte der Box – seine Lage hängt von der Form der Verteilung ab, die somit ebenfalls direkt aus dem Box-Plot abgelesen werden kann (mehr dazu weiter unten). Da die Box zwischen dem oberen und dem unteren Quartil verläuft, entspricht ihre Länge auch genau dem IQR.

Nun wird ein Abstand von jeweils 1,5 IQR auf die obere und die untere Kante der Box „aufgerechnet“, so dass sich ein Feld mit einer Gesamtlänge von 4 IQR ergibt. Zwei Werte, der größte und der kleinste in der Verteilung real beobachtete Wert, die noch in diesem Bereich von 4 IQR liegen, bilden nun die Grenzpunkte für den oberen und den unteren Zaun des

Box-Plots, die jeweils durch eine Linie mit der Box verbunden werden. Zu beachten ist unbedingt, dass die Zäune nicht an der Grenze von +/- 1,5 IQR um die beiden Enden der Box liegen, sondern dort, wo der größte bzw. der kleinste Wert der Verteilung innerhalb dieser beiden Abstände liegen. Nur in dem unwahrscheinlichen Fall, dass ein empirisch beobachteter Wert mit einem der Grenzwerte übereinstimmt, würde ein Zaun an diesem Punkt verlaufen.

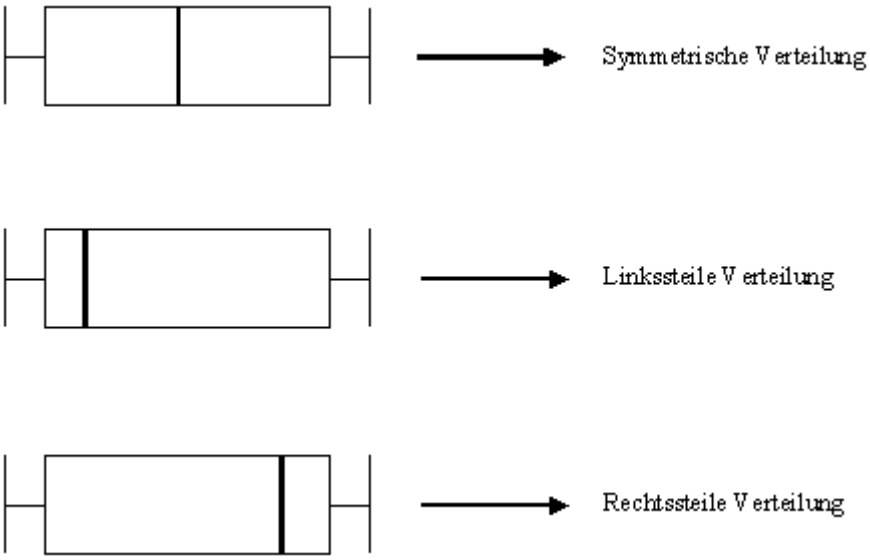

Abbildung 14: Lage des Medians und Verteilungsform

Alle Werte die außerhalb der Zäune liegen sind Ausreißer, wobei der erweiterte Box-Plot noch zwischen Ausreißern und Extremwerten – extremen Ausreißern – unterscheidet. Als Ausreißer werden alle Werte gekennzeichnet, die innerhalb eines Abstands von + 1,5 IQR vom oberen Zaun bzw. - 1,5 IQR vom unteren Zaun liegen. Sie werden mit einem Kreis und der Nummer des entsprechenden Datensatzes gekennzeichnet. Alle Werte außerhalb dieses nun auf 7 IQR angewachsenen Bereichs werden als Extremwerte ebenfalls mit der Nummer des Datensatzes sowie einem Sternchen gekennzeichnet. Die Kennzeichnung mit der Datensatznummer macht den Box-Plot zu einem der beliebtesten Instrumente für die Identifikation von Ausreißern, da der Analytiker direkt am Box-Plot ablesen kann, welche Datensätze der genaueren Überprüfung bedürfen. Es ist allerdings anzumerken, dass die Definition des Begriffs „Ausreißer“, die hier bei der Konstruktion des Box-Plots zur Anwendung kommt, keineswegs allgemeingültig ist (siehe weiter unten).

Sollen mehrere Verteilungen oder aber mehrere überschneidungsfreie (disjunkte) Untergruppen innerhalb einer einzigen Verteilung, beispielsweise Männer und Frauen, grafisch miteinander verglichen werden, so ist es möglich, Box-Plots einander gegenüberzustellen.

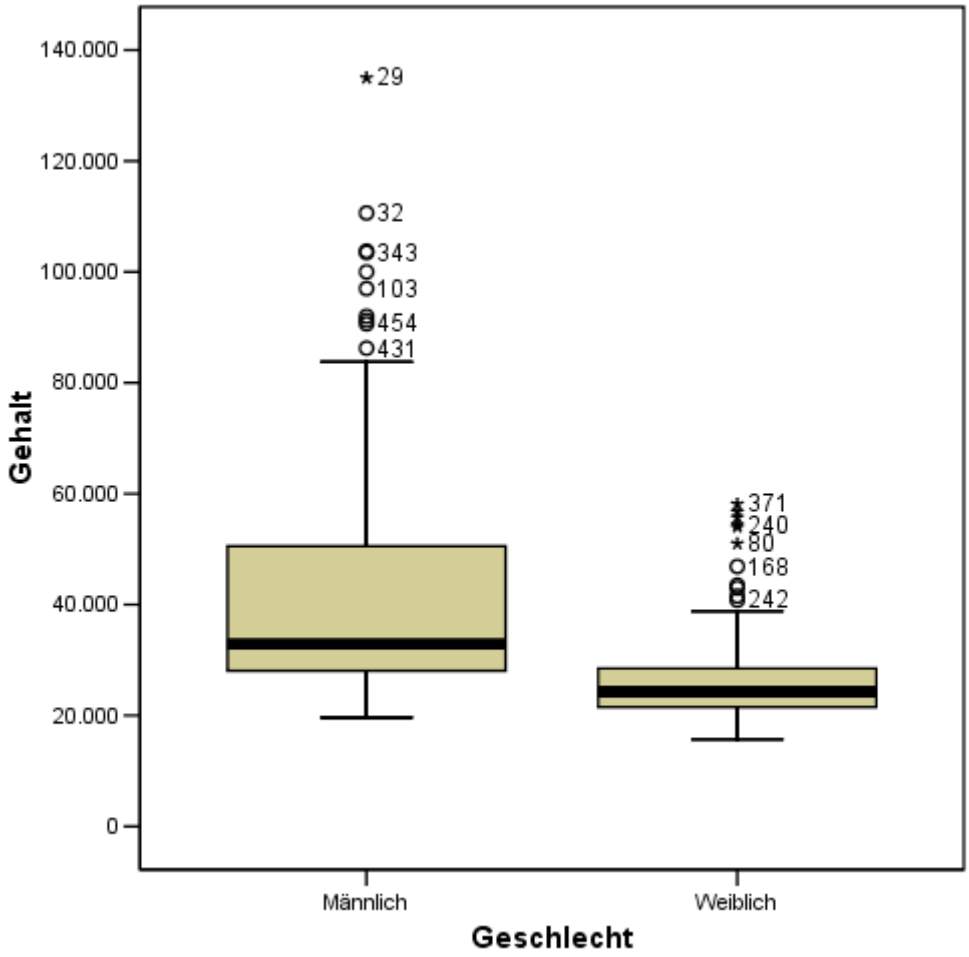

Abbildung 15: Vergleich zweier Box-Plots in SPSS

Weitergehende Vergleiche sind über die sogenannten gruppierten Box-Plots möglich. Hier erfolgt eine Aufteilung anhand mehr als nur eines Merkmals. Beispiel: Die Verteilung der Gehälter aus dem obigen Box-Plot soll nun nicht nur für die Gruppen der Männern und Frauen getrennt betrachtet werden, es soll zudem auch noch in Personen unterschieden werden, die einer Minderheit angehören. Dadurch ergeben sich insgesamt vier verschiedenen Gruppen, die in einem gruppierten Box-Plot dargestellt werden können.

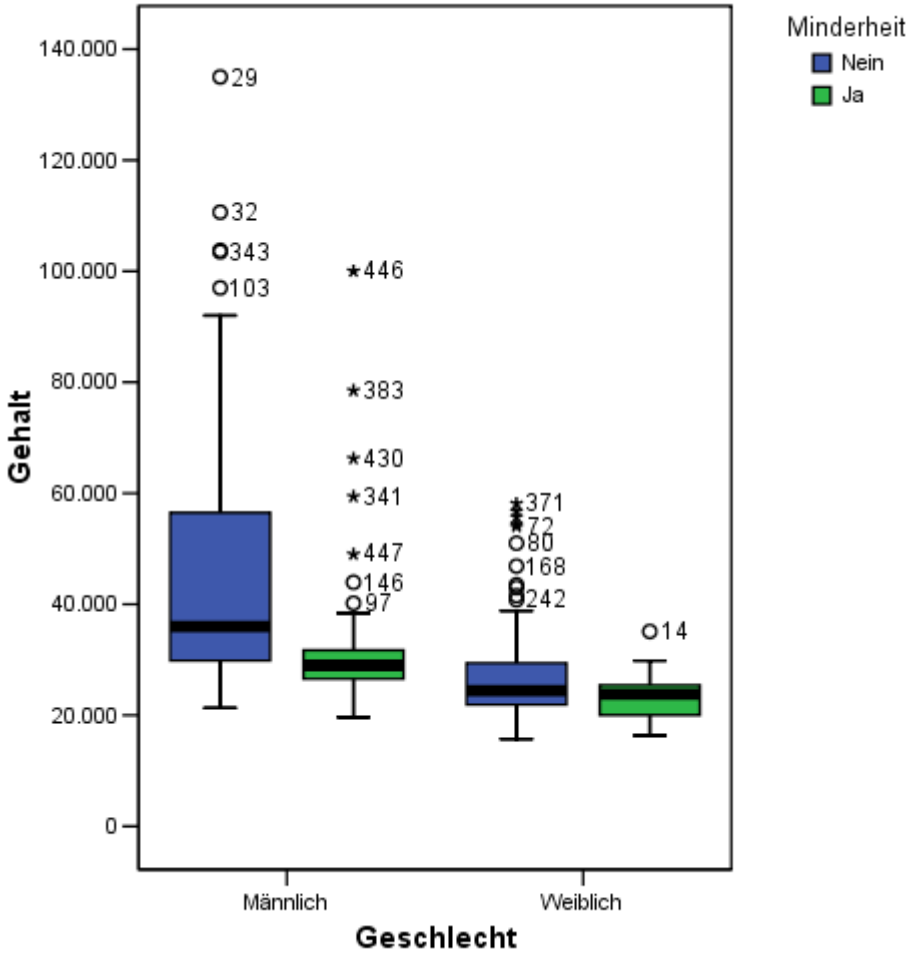

Abbildung 16: Gruppierter Box-Plot mit vier Gruppen in SPSS

7 P-P-Diagramme

Die P-P-Diagramme sowie die „artverwandten“ Q-Q-Diagramme gestatten den schnellen Vergleich einer Verteilung mit einer Testverteilung. Sie kommen beispielsweise dann zum Einsatz wenn zu klären ist, inwiefern ein vorliegender Datensatz normalverteilte Werte aufweist (siehe weiter unten).

Ein P-P-Diagramm trägt die kumulierten Häufigkeiten der beobachteten Werte (dargestellt durch einzelne Punkte) gegen die zu erwartenden kumulierten Häufigkeiten einer perfekt verlaufenden Vergleichsverteilung (dargestellt durch eine diagonale Linie) ab. In den meisten Fällen ist dies eine Normalverteilung, da die Prüfung auf das Vorliegen einer Normalvertei-

lung zu vielen multivariaten Analyseverfahren dazugehört, es kann aber auch gegen jede beliebige andere Verteilung abgetragen werden, beispielsweise gegen eine Student-Verteilung (T-Verteilung) oder eine F(isher)-Verteilung. Je stärker sich die Verteilung der real aufgetretenen Stichprobenwerte und die Vergleichsverteilung ähneln, desto stärker stimmen die empirischen mit den erwarteten kumulierten Häufigkeiten überein – zu erkennen am mehr oder weniger diagonalen Verlauf des Diagramms. Bei einer perfekten Übereinstimmung von empirischen Werten und theoretischer Verteilung, die in der Praxis allerdings nicht zu erwarten ist, liegen sämtliche Punkte auf der diagonalen Linie.

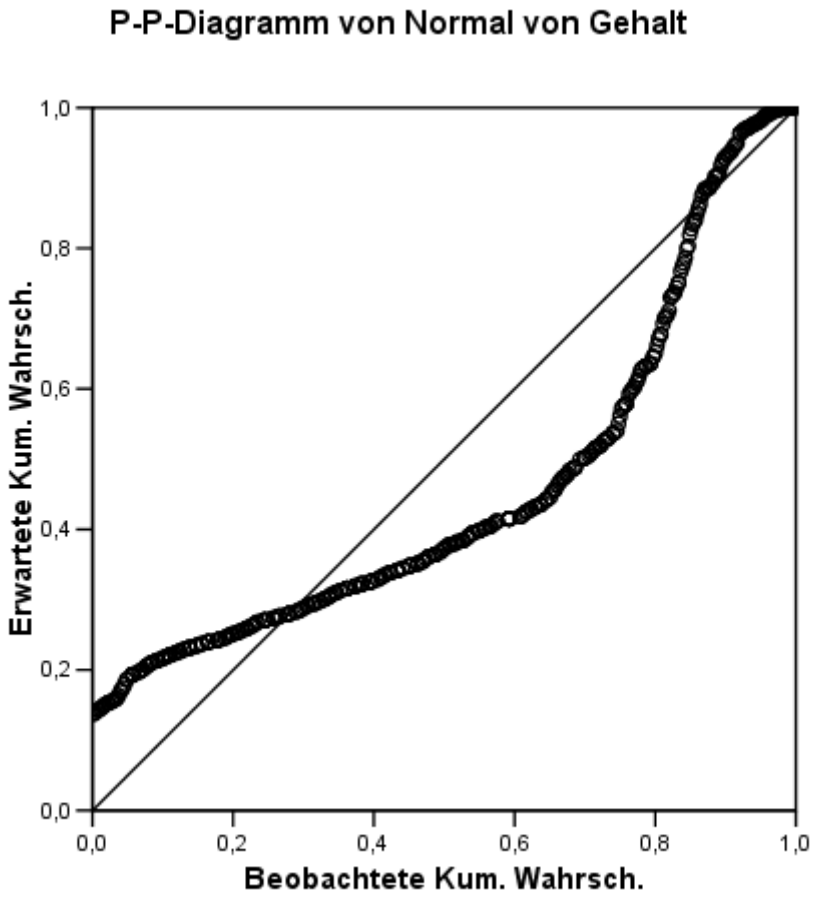

Abbildung 17: P-P-Diagramm in SPSS

Zusätzliche zum P-P-Diagramm kann mit SPSS auch ein sogenanntes trendbereinigtes P-P-Diagramm erstellt werden, bei dem die beobachteten kumulierten Häufigkeiten nicht gegen die erwarteten kumulierten Häufigkeiten, sondern gegen die Abweichungen der beobachteten

von den erwarteten kumulierten Häufigkeiten abgetragen werden. Auch hier kann neben der Normalverteilung noch gegen andere Vergleichsverteilungen abgetragen werden.

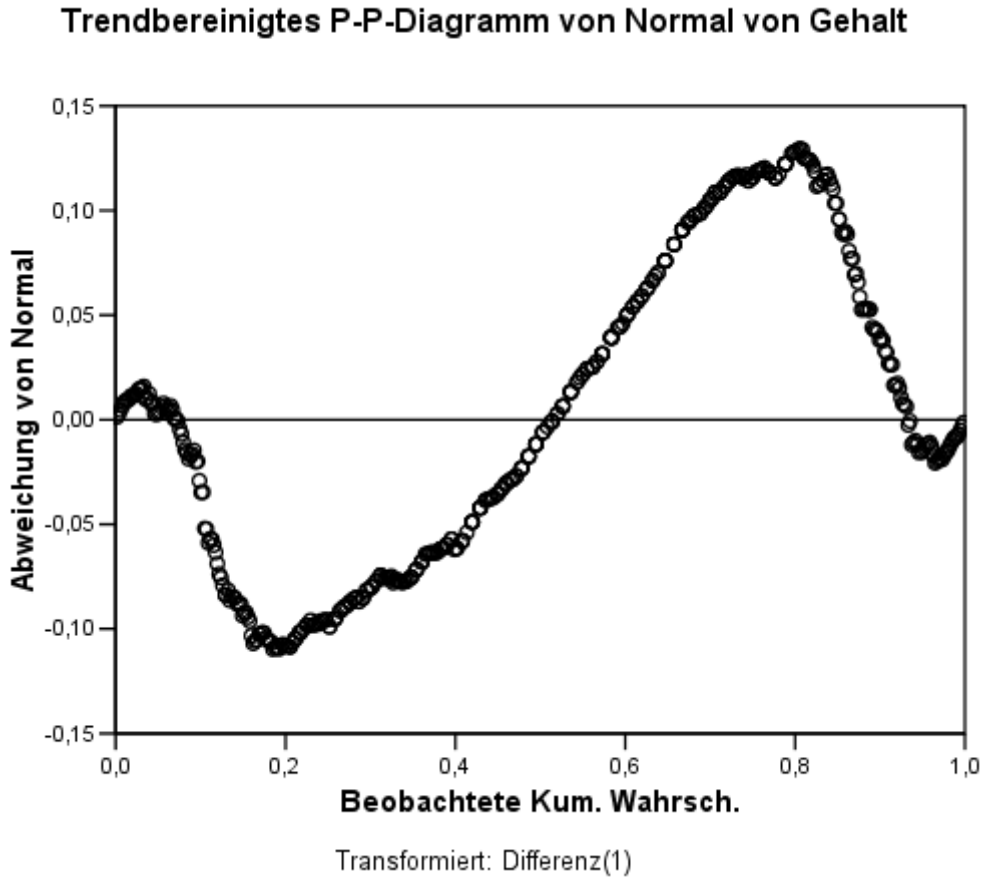

Abbildung 18: Trendbereinigtes P-P-Diagramm in SPSS

8 Q-Q-Diagramme

Q-Q-Diagramme dienen wie P-P-Diagramme dem visuellen Vergleich einer vorliegenden Verteilung mit einer Referenzverteilung wie der Normalverteilung oder der Fischer-Verteilung. Im Gegensatz zum P-P-Diagramm werden im Q-Q-Diagramm nicht die beobachteten und die erwarteten kumulierten Häufigkeiten gegenübergestellt, sondern die direkt beobachteten und die bei Vorliegen der Vergleichsverteilung zu erwartenden Werte. Wie im P-P-Diagramm kennzeichnen auch im Q-Q-Diagramm stärkere oder einem Muster folgende Abweichungen der Punkte vom Verlauf der Diagonalen Abweichungen der beobachteten von den erwartenden Werten und sind als Indiz dafür zu werten, dass die beobachteten Merkmalswerte

nicht aus einer in der Grundgesamtheit der Vergleichsverteilung folgenden Verteilung entstammen.

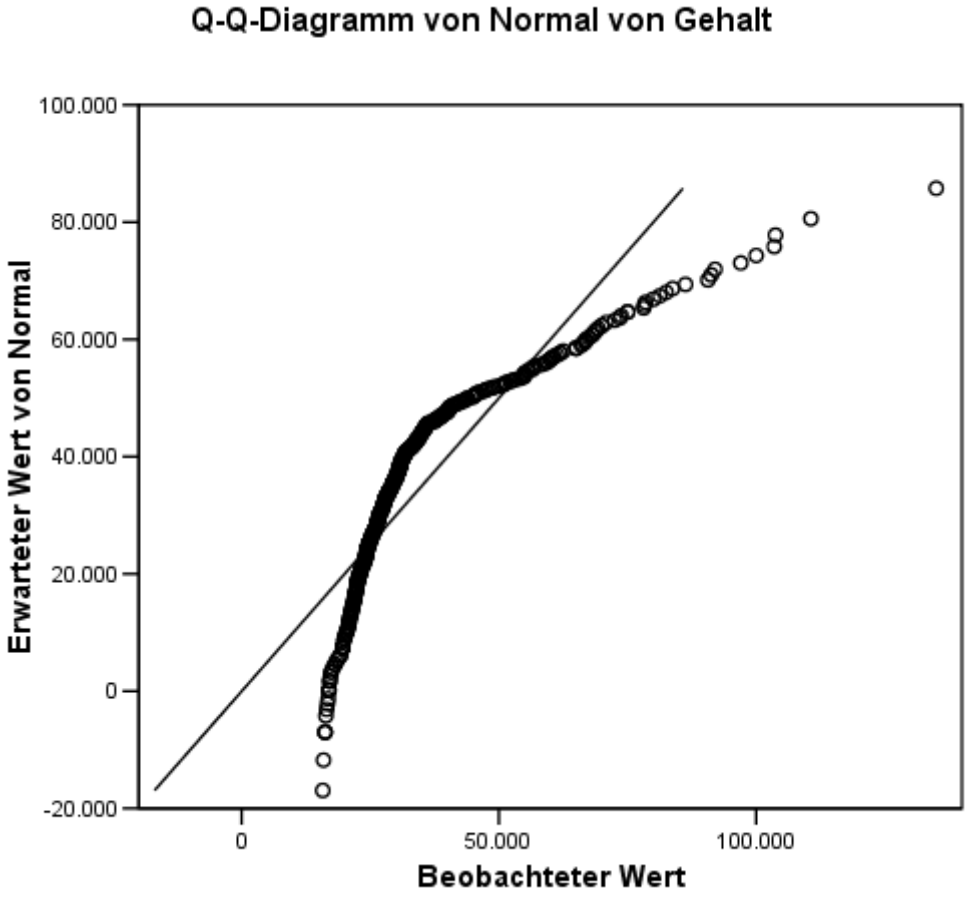

Abbildung 19: Q-Q-Diagramm in SPSS

Zusätzlich zum Q-Q-Diagramm kann analog zum P-P-Diagramm mit SPSS auch ein trendbereinigtes Q-Q-Diagramm erstellt werden, bei dem die beobachteten Werte nicht mit den erwarteten Werten, sondern mit den Abweichungen der beobachteten von den erwarteten Werten dargestellt werden. Ebenso wie bei dem trendbereinigten P-P-Diagramm kann der Analytiker in SPSS auch bei der Darstellung des trendbereinigten Q-Q-Diagramms aus einer ganzen Reihe von möglichen Vergleichsverteilungen wählen.

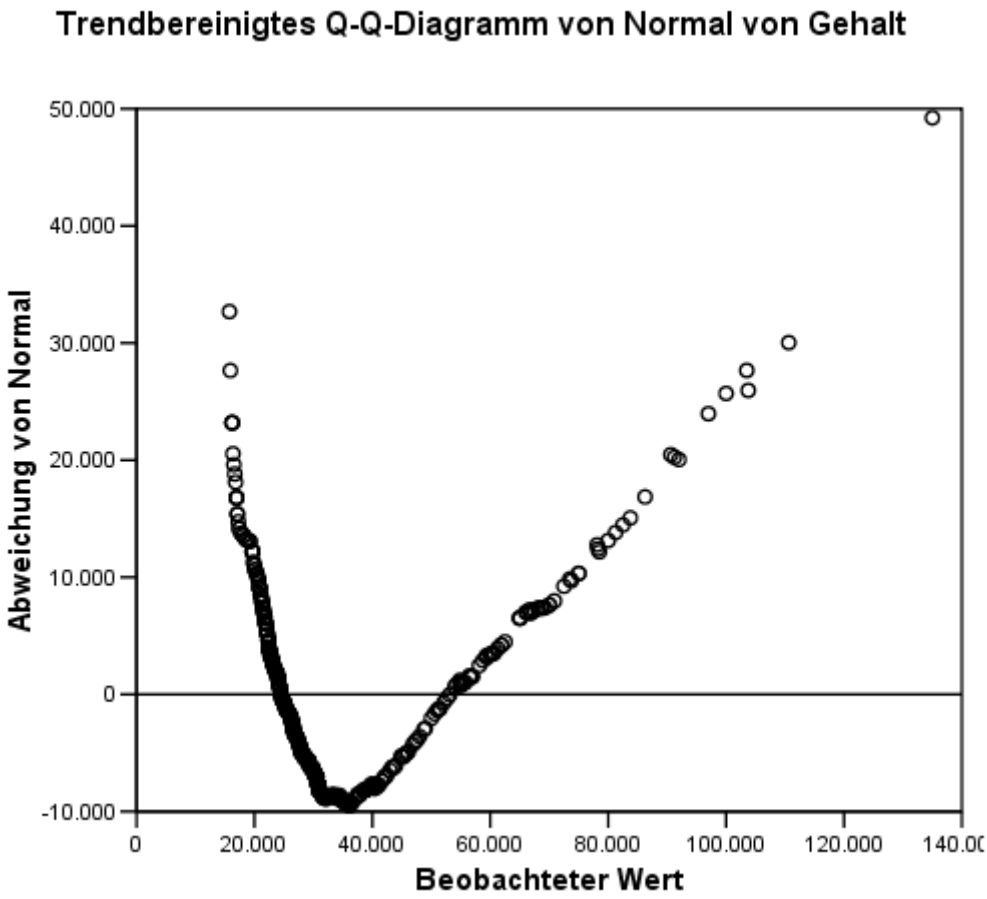

Abbildung 20: Trendbereinigtes Q-Q-Diagramm in SPSS

V Grafische Darstellungsformen multivariater Daten

Bei der grafischen Darstellung multivariater Daten ist zwischen bivariaten Darstellungen – also Darstellungen von zwei Variablen in einer Grafik – und multivariaten Darstellungen – Darstellungen von drei oder mehr Variablen in einer Grafik – zu unterscheiden. Eine Besonderheit stellt das 3-D-Streudiagramm dar, welches nur für die Darstellung von drei Variablen geeignet ist.

Grafische Darstellungen mit vier Dimensionen – die vierte Dimension kann als zeitliche Dimension durch eine Bewegung im Diagramm dargestellt werden – finden sich in SPSS nicht.

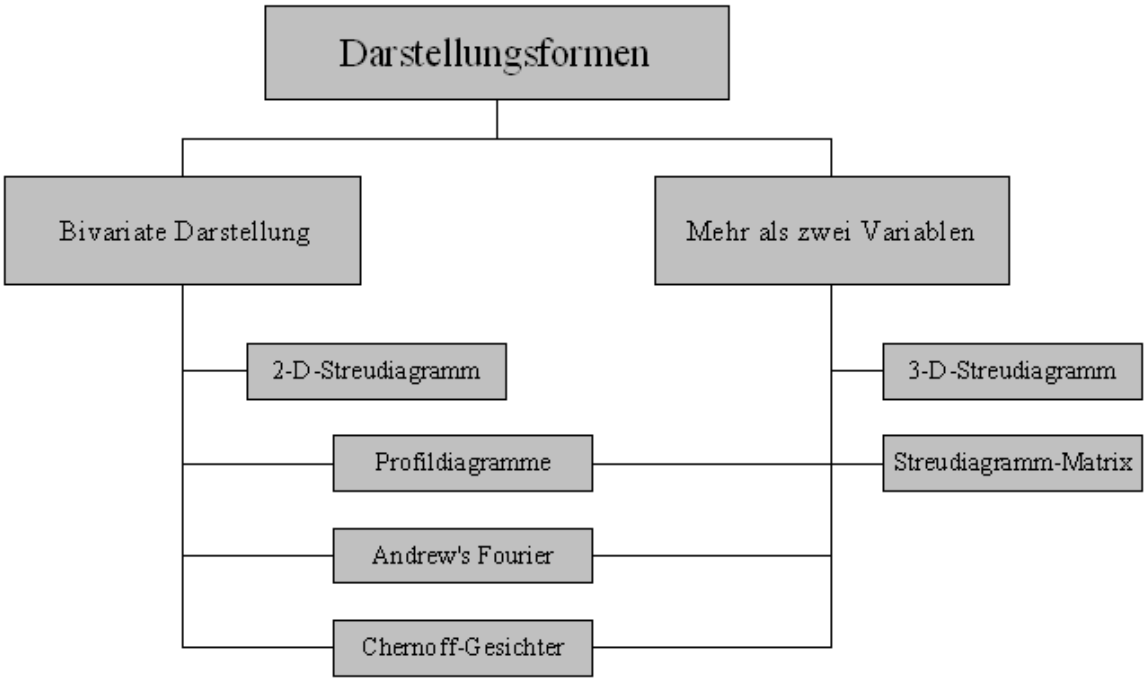

Abbildung 21: Grafische Darstellung bi- und multivariater Daten

1 Streudigaramme

Streudiagramme stellen die gemeinsame Verteilung der Werte von zwei Variablen (bzw. drei Variablen im 3D-Streudiagramm) dar, indem die entsprechenden Werte der Variablen gegeneinander abgetragen werden. Die Lage und Verteilung der Werte ermöglicht Rückschlüsse auf mögliche Zusammenhänge zwischen den dargestellten Variablen.

Beispiel: Treten in der Tendenz große Werte der einen Variablen gepaart mit großen Werten der anderen Variablen auf, so kann ein positiver Zusammenhang vermutet werden (beispielsweise bei Werbeausgaben und Verkaufszahlen). Bei der Interpretation von Streudiagrammen ist stets zu berücksichtigen, dass ein gefundener Zusammenhang nicht in eine bestimmte Richtung interpretiert werden kann. Aus einem Streudiagramm ist daher nicht abzulesen, ob die Variable A die Variable B beeinflusst, ob der umgekehrte Zusammenhang vorliegt, ob eine dritte Variable beide Variablen beeinflusst oder ob ein sogenannter Scheinzusammenhang besteht (das Beispiel mit der Geburtenrate und der Storchquote aus Statistik I dürfte ja noch bekannt sein).

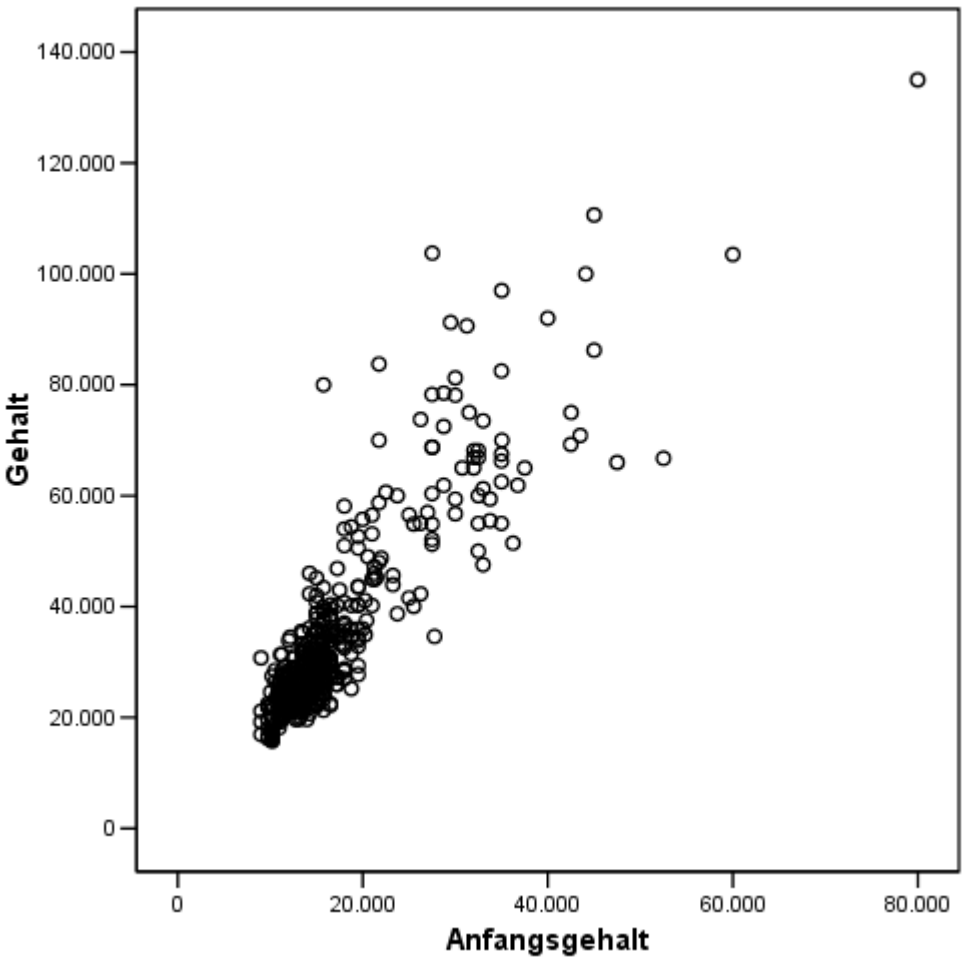

Abbildung 22: Streudiagramm zweier Variablen in SPSS

2 Streudiagramm-Matrix

Liegt ein multivariater Fall vor, d.h. sollen für mehrere Variablenpaare jeweils die gemeinsamen Verteilungen dargestellt werden, ist statt einer Reihe bivariater Streudiagramme oder schwer interpretierbarer 3D-Streudiagramme ein gemeinsames Streudiagramm in Form einer sogenannten Streudiagramm-Matrix sinnvoll.

Eine Streudiagramm-Matrix gestattet den schnellen Überblick über die Vielzahl aller möglichen Paarverteilungen und ermöglicht das rasche Auffinden symmetrischer oder anderweitig auffälliger Einzel-Streudiagramme. Diese können im Anschluss einzeln und detaillierter dargestellt und vom Analytiker eingehender untersucht werden. Jedes mögliche Streu-

diagramm taucht genau zweimal in der Matrix auf, und zwar einmal oberhalb und einmal unterhalb der Hauptdiagonalen. Dabei sind jeweils die Achsen der Diagramme miteinander vertauscht, d.h. es wird beispielsweise einmal Gehalt gegen Alter abgetragen und einmal Alter gegen Gehalt.

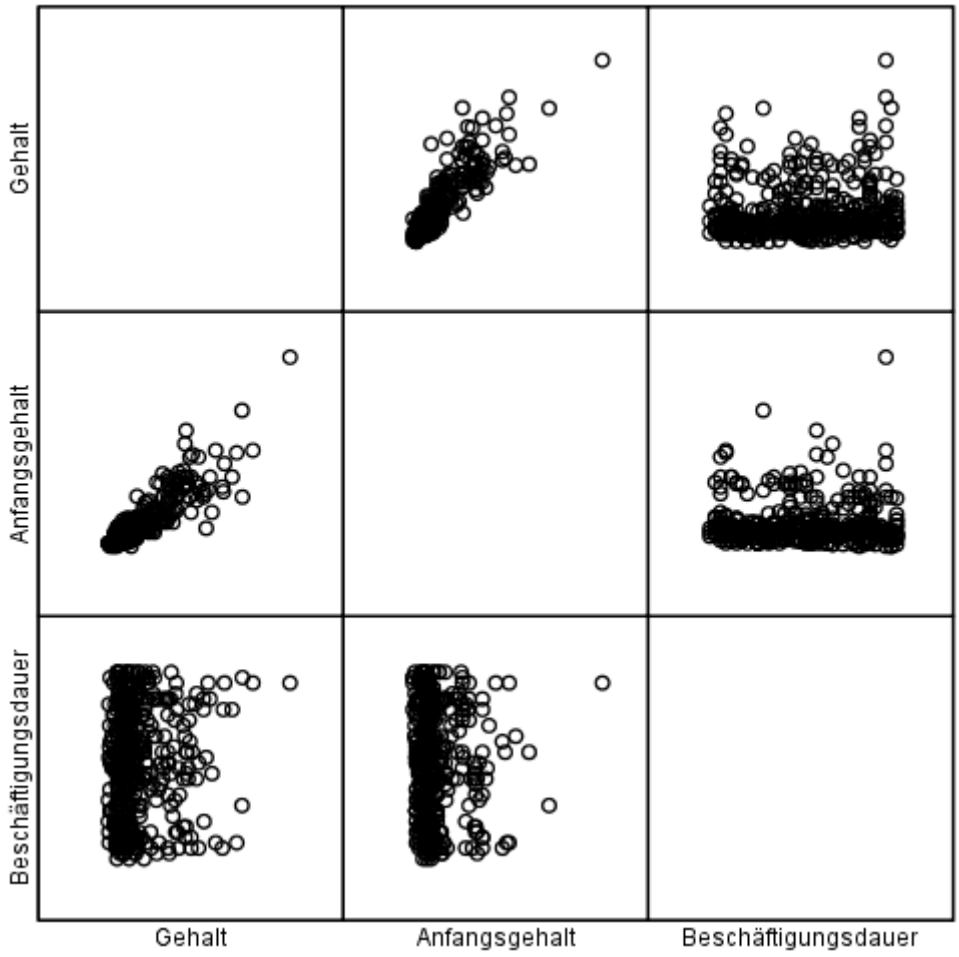

Abbildung 23: Streudiagramm-Matrix in SPSS

VI Ausreißeranalyse

1 Einführung

Bei einem Ausreißer handelt es sich, wie bereits weiter oben dargestellt, um einen gemessenen, erhobenen oder experimentell gefundenen Wert, der nicht den Erwartungen entspricht

bzw. nicht zu den restlichen Werten der Verteilung passt. Es existiert keine klare Regel für die eindeutige Identifikation von Ausreißern, so dass an dieser Stelle nicht angegeben werden kann, ab welchem „Schwellenwert“ ein Wert als Ausreißer zu bezeichnen ist. Die bei der Konstruktion des Box-Plots geltende Unterscheidung in Ausreißer und Extremwerte außerhalb eines Bereichs von 4 bzw. 7 IQR ist also keineswegs als allgemeingültig zu betrachten, auch wenn sie einen guten Richtwert darstellt. Letzten Endes ist es aber stets eine Entscheidung des Analytikers, welche Werte als Ausreißer gekennzeichnet werden.

Für das Auftreten von Ausreißern gibt es drei mögliche Ursachen:

Es wäre möglich, dass der Ausreißer durch einen verfahrenstechnischen Fehler verursacht wurde, beispielsweise einen Fehler bei der Dateneingabe (130 statt 13), einen Fehler beim Codieren der Daten oder aber einen technischen Ausfall bei der EDV-Datenspeicherung. Solche Ausreißer können immer auftreten und unter Umständen sogar wieder rückgängig gemacht werden, wenn sich der „echte“ Wert noch wiederherstellen lässt.

Der Ausreißer könnte auch einfach nur von einem ungewöhnlichen Wert herrühren, der so real bei der Erhebung aufgetreten ist, und damit auch erklärt werden kann. Der schon bei den Lagemaßen als Beispiel herangezogene einzige befragte Millionär in einer Gruppe von Normalverdienern wäre genau ein solcher Fall. Solche Fälle können unter Umständen darauf hindeuten, dass die Befragung falsch angelegt oder durchgeführt wurde, da ein wichtiges Selektionskriterium nicht bedacht wurde und daher nun Merkmalsträger in der Stichprobe gelandet sind, die eigentlich gar nicht untersucht werden sollten. Allein schon aus diesem Grund müssen die Ausreißer vor jeder weiterführenden Analyse gründlich untersucht werden – sie könnten einen Hinweis auf mangelnde Repräsentativität der Stichprobe geben.

Schlußendlich könnte der Ausreißer auch einen „echten“ und ungewöhnlichen Wert kennzeichnen, der durch den Forscher nicht erklärt werden kann.

Generell ist noch zwischen univariaten und multivariaten Ausreißern zu unterscheiden. Bei univariaten Ausreißern handelt es sich um einen einzelnen außergewöhnlich hohen oder niedrigen Wert eines bestimmten erhobenen Merkmals – hier kann wieder der einsame und

vermutlich versehentlich befragte Millionär als Beispiel herangezogen werden, der im Datensatz auch schnell zu erkennen ist. Die Identifikation eines multivariaten Ausreißers ist dagegen komplizierter, denn hier handelt es sich um einen Datensatz, der mehrere für sich genommen normale Merkmalsausprägungen aufweist, die aber in ihrer Kombination äußerst ungewöhnlich sind. Ein Beispiel hierfür wäre eine 80jährige Frau, die über einen Internetanschluss verfügt. Weder 80jährige noch Personen mit Internetanschluss sind in einer Studie der Allgemeinbevölkerung eine große Seltenheit, diese Frau aber ganz sicher, da die beiden Merkmalsausprägungen üblicherweise nicht in Kombination zu erwarten sind. In den nachfolgenden Abschnitten wird hauptsächlich auf univariate Ausreißer eingegangen.

2 Identifikation von Ausreißern

Es bieten sich mehrere Methoden an, mit denen der Marktforscher Ausreißer im Datensatz aufspüren kann. Am mühsamsten ist dabei die manuelle Durchsicht des gesamten Datensatzes, die bei umfangreicheren Untersuchungen auch irgendwann unmöglich wird. Effizienter ist da bereits die visuelle Identifikation anhand eines Box-Plots oder eines Streudiagramms, wobei der Box-Plot den zusätzlichen Vorteil aufweist, dass die Identifikationsnummern der auffälligen Datensätze gleich mit angegeben werden. Beide Darstellungsformen wurden bereits im Detail weiter oben beschrieben.

Extremwerte

			Fallnummer	Wert
Gehalt	Größte Werte	1	29	135.000
		2	32	110.625
		3	18	103.750
		4	343	103.500
		5	446	100.000
	Kleinste Werte	1	378	15.750
		2	338	15.900
		3	411	16.200
		4	224	16.200
		5	90	16.200

Abbildung 24: Extremwerttabelle in SPSS

Eine weitere Möglichkeit ist die Ausgabe einer sogenannten Extremwerttabelle in SPSS. Diese Tabelle listet die fünf größten und die fünf kleinsten Variablenwerte für ein bestimmtes erhobenes Merkmal auf. Dabei ist zu beachten, dass die hier aufgeführten Werte (a) nicht unbedingt alle Ausreißer sein müssen (wenn die Variable weniger als fünf Ausreißer in beide Richtungen aufweist) und (b) nicht alle möglichen Ausreißer in der Tabelle zu finden sein müssen (wenn die Variable mehr als fünf Ausreißer in mindestens eine der beiden Richtungen aufweist). Damit dürfte klar sein, dass die Extremwerttabelle die Anzahl und Lage der Ausreißer nur unter bestimmten Voraussetzungen korrekt wiedergibt – die Identifikation solcher Werte über den Box-Plot ist also klar vorzuziehen.

3 Der Leverage-Effekt

Warum ist es nun so wichtig, sich mit Ausreißern zu beschäftigen? Welche Gefahr könnten sie möglicherweise für die Richtigkeit und Genauigkeit einer Analyse darstellen?

Nun, soweit es beispielsweise die Berechnung des arithmetischen Mittels betrifft ist die Gefahr offensichtlich: ein oder mehrere Ausreißer können das Mittel in eine bestimmte Richtung „ziehen“ und so seine Aussagekraft erheblich schwächen. Es gibt allerdings auch „subtilere“ Probleme, die von Ausreißern ausgelöst werden können. Eines davon ist als der sogenannte Leverage-Effekt bekannt, und soll hier näher betrachtet werden.

Aus der Statistik I ist die lineare Regressionsanalyse bekannt, bei der durch eine „Wolke“ von Messpunkten mit mehr oder weniger deutlichem linearen Trend eine Regressionsgerade gelegt wird, die möglichst viel Streuung erklären soll (mehr zur linearen Regression und zum Gütemaß R^2 im nächsten Abschnitt). Ist ein deutlicher linearer Trend vorhanden, so ergibt sich eine hohe Streuungsaufklärung und damit ein brauchbares Regressionsmodell.

Bemerkenswerterweise kann aber schon ein einziger Ausreißer, wenn er an der „richtigen“ Stelle liegt, das Ziel der Streuungsaufklärung vollständig unterlaufen, indem er die Regressionsgerade in eine bestimmte Richtung „zieht“ und damit deren Erklärungswert deutlich verringert.

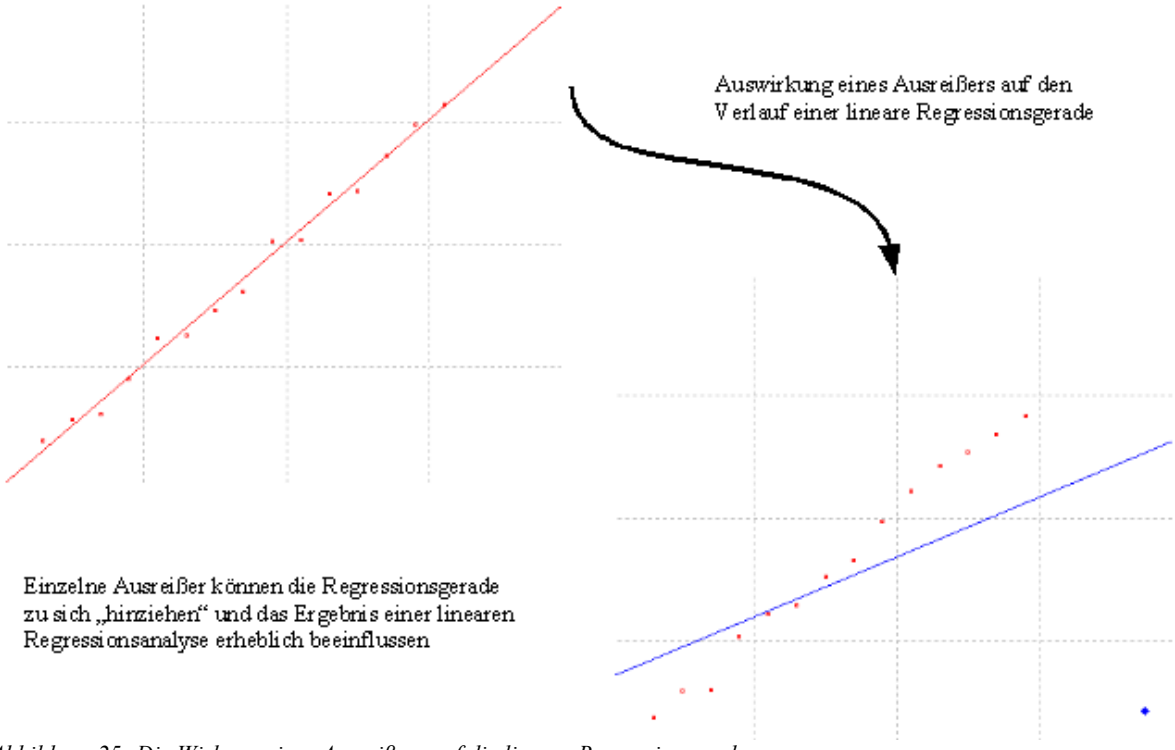

Abbildung 25: Die Wirkung eines Ausreißers auf die lineare Regressionsanalyse

Dieser Effekt macht deutlich, wieso die Analyse der Ausreißer vor der Durchführung einer weiterführenden Analyse, wie beispielsweise der linearen Regressionsanalyse, ein unbedingtes Muss für den Marktforscher ist.

4 Der Umgang mit Ausreißern

Die entscheidende Frage der Ausreißeranalyse lautet: Wie ist mit den aufgefundenen Ausreißern (und Ausreißer werden sich bei nahezu jeder praktischen Analyse finden) umzugehen? Generell existieren drei Möglichkeiten: Den Ausschluss aus der Analyse, den Eingang in die Analyse oder die Kennzeichnung als fehlenden Wert. Welche dieser Möglichkeiten gewählt wird, hängt von verschiedenen inhaltlichen Überlegungen ab.

Von ganz entscheidender Bedeutung ist die Frage, wie die Ausreißer überhaupt zustande gekommen sind. Handelt es sich um „harmlose“, durch Eingabe- oder Speicherfehler verurs-

achte Ausreißer, dann lassen diese sich eher entfernen oder als fehlende Werte kennzeichnen. Handelt es sich um reale Werte muss untersucht werden, inwiefern Fehler im Erhebungsdesign oder eine fehlerhafte Eingrenzung der Grundgesamtheit für das Auftreten der Werte verantwortlich gemacht werden können und was dies über die Tauglichkeit der Gesamtheit der erhobenen Daten für die weitere Analyse aussagt.

Ebenfalls von Bedeutung ist die Überlegung, welchen Einfluss die Ausreißer auf die weitere Analyse haben, was in erster Linie von der Art der geplanten Auswertung abhängt. Wie bereits betrachtet, können das arithmetische Mittel und die lineare Regressionsanalyse ganz erheblich durch Ausreißer beeinflusst werden. Bei der Berechnung des Medians oder dem Answer-Tree-Verfahren spielen die Ausreißer dagegen keine Rolle.

Vor einer Entscheidung zugunsten der Entfernung von Werten ist auf jeden Fall noch zu überprüfen, welcher Datenverlust durch die Löschung oder Kennzeichnung als fehlende Werte entsteht. Sinkt die Zahl der verfügbaren Datensätze unter das zum Fortfahren notwendige Niveau, so ist von der Entfernung der Ausreißer abzusehen. Dies unterstreicht die Notwendigkeit einer ausreichenden Stichprobengröße n, da der Wegfall einiger Werte niemals die Fortsetzung einer Analyse bedrohen sollte.

VII Fehlende Werte

1 Einführung

Unter fehlenden Werten sind im Datensatz fehlende Werte zu verstehen, die beispielsweise das Ergebnis nicht ausgefüllter Felder in einem Fragebogen sein können. Solche leeren Felder sind bei Personenbefragungen häufig darauf zurückzuführen, dass die betreffende Person die Antwort als zu persönlich oder die Frage als zu intrusiv betrachtet und die Antwort daher verweigert hat. Typischerweise wird es beispielsweise bei Fragen zum Einkommen, zum Körper oder zum Sexualverhalten eine hohe Anzahl fehlender Werte geben.

Fehlende Werte können aber auch auf andere Ursachen zurückzuführen sein. Sie sind ge-

nerell dann ein Problem für die Integrität der durchgeführten Untersuchung, wenn ein Zusammenhang zwischen der Wahrscheinlichkeit des Fehlens eines Wertes und einem untersuchten Sachverhalt zu vermuten ist, sich die fehlenden Werte also nicht zufällig verteilen. Dies kann beispielsweise bei der Frage nach dem Einkommen der Fall sein, wenn zu vermuten wäre, dass Personen mit niedrigem Einkommen verstärkt die Auskunft verweigern würden. Die Verteilung der fehlenden Werte wäre in diesem Beispiel nicht zufällig, die Masse der fehlenden Angaben hätte nämlich im unteren Einkommensbereich gelegen. In einem solchen Fall würde sich das Durchschnittseinkommen aufgrund der überproportional nicht angegebenen Niedrigeinkommen nach oben verzerren – die Auswertung der Daten würde also durch die fehlenden Werte verfälscht.

Anders verhält es sich, wenn die Werte beliebig fehlen, daher ist bei der Untersuchung fehlender Werte in erster Linie zu klären, inwiefern die fehlenden Werte zufällig auftreten oder ob ein Muster erkennbar ist. Desweiteren stellt sich die Frage, wie viele Werte überhaupt fehlen dürfen, damit eine sinnvolle Auswertung der Daten noch möglich erscheint.

Bezüglich des Umgang mit fehlenden Werten bieten sich drei Möglichkeiten an: Entweder es werden ausschließlich vollständige Werte zur weiteren Auswertung zugelassen, es erfolgt ein variablenweiser bzw. fallweiser Ausschluss oder aber die fehlenden Werte werden induktiv oder statistisch ersetzt.

Für eine dieser Möglichkeiten wird sich der Marktforscher auf jeden Fall entscheiden müssen, denn mit fehlenden Werten ist bei jeder marktforscherischen Untersuchung in der Praxis zu rechnen – und das Problem der fehlenden Daten kann nicht einfach ignoriert werden.

2 Zufälligkeitsgrade

Das Fehlen von Werten kann im Wesentlichen auf drei Ursachen zurückgeführt werden: Zunächst einmal können einfache Fehler für das Fehlen verantwortlich sein, wie sie in jeder Erhebung auftreten können. Vorstellbar sind beispielsweise Eingabefehler, bei denen etwa

Buchstaben in einem Zahlenfeld eingegeben werden. Aber auch Codierungs- und Übertragungsfehler bei der Eingabe oder Speicherung der Daten können zu leeren Feldern im Datensatz führen.

Bedenklicher sind da schon fehlende Werte, die auf ungenaue Fragen bei der Personenerhebung zurückzuführen sind. Ein Nicht-Akademiker wird kaum in der Lage sein, die Frage nach der Studienrichtung zu beantworten, ebenso wenig wie ein Arbeitsloser die Zufriedenheit mit seiner Arbeitsstelle auf einer Skala von 1 bis 10 beantworten kann. Sind viele fehlende Werte im Datensatz auf solche Probleme zurückzuführen, so ist der Fragebogen für weitere Befragungen unbedingt zu überarbeiten. Für statistische Einheiten nicht relevante Daten, wie eben die besagte Studienrichtung beim Nicht-Akademiker, können als benutzerdefiniert fehlende Werte deklariert werden (siehe weiter unten).

Die dritte der möglichen Ursachen verdient die besondere Aufmerksamkeit des Marktforschers. Gemeint sind die bereits oben angesprochenen Aktionen des Befragten, beispielsweise das Vergessen von Angaben, widersinnige Angaben wie beispielsweise die Eintragung einer Studienrichtung bei gleichzeitiger Angabe, nie studiert zu haben, die Nichtauskunftsfähigkeit oder aber das direkte Verweigern einer Antwort – womit wir wieder bei den Fragen zu Einkommen, Körper oder Sexualverhalten wären.

Insbesondere aufgrund der letzten Ursache stellt sich natürlich stets die Frage, ob fehlende Werte zufällig auftreten oder ob sich bestimmte Muster und Zusammenhänge erkennen lassen. Dabei wird in drei Zufälligkeitsgrade unterschieden: MCAR, MAR und NRM. Der Zufälligkeitsgrad ist wiederum entscheidend für die Frage, ob fehlende Werte ausgeschlossen oder ersetzt werden können.

MCAR = Missing completely at random

Auf dieser Stufe tritt das Fehlen von Werten vollkommen zufällig auf, d.h. die Wahrscheinlichkeit des Fehlens einzelner Werte steht in keinerlei Zusammenhang mit irgendwelchen anderen Größen. Es ist somit kein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variable Y mit der Variable Y selbst (Beispiel: niedrige Einkommen wer-

den häufig nicht angegeben) oder mit einer anderen Variablen X (Beispiel: Frauen verweigern tendenziell häufiger Angaben zu ihrem Körpergewicht) feststellbar.

MAR = Missing at random

Das Auftreten von fehlenden Werten steht auf dieser Stufe zumindest teilweise im Zusammenhang mit einer anderen erhobenen Variablen. Es ist kein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variablen Y mit der Variable Y selbst feststellbar, wohl aber ein Teilzusammenhang mit einer anderen Variablen X.

NRM = Nonrandom missing

Hier folgt das Auftreten von fehlenden Werten ganz klaren Gesetzmäßigkeiten, eine Zufälligkeit ist vollkommen auszuschließen. Es kann entweder ein Zusammenhang zwischen dem Auftreten von fehlenden Werten der Variablen Y und der Variablen Y selbst oder auch einer anderen Variablen X oder auch beides vorliegen, d.h. das Auftreten eines fehlenden Wertes ist vollständig durch eine andere Variable oder die Variable selbst vorhersagbar.

3 Umgang mit fehlenden Werten

In Abhängigkeit vom Zufälligkeitsgrad (siehe oben) lassen sich drei Methoden zum Umgang mit fehlenden Werten anwenden: Der sogenannte complete case approach, der Ausschluss von Fällen oder Variablen oder das induktive bzw. statistische Ersetzen von Werten.

Beim complete case approach (CCA) werden ausschließlich die vollständigen Fälle für die weitere Analyse verwendet – alle Fälle mit mindestens einem einzigen fehlenden Datensatz werden aus dem Datensatz entfernt. Diese Methode kann nur zum Einsatz kommen, wenn zufällig fehlende Daten (MCAR) vorliegen. Außerdem ist darauf zu achten, dass die Stichprobe durch die Entfernung der Fälle nicht zu klein ausfällt und damit die Interpretation der Daten unmöglich wird. Bei entsprechend großen Stichproben und Vorliegen von MCAR ist der complete case approach zu empfehlen.

Das Ziel des Ausschlusses von Fällen oder Variablen ist die Verringerung des Gesamtanteils fehlender Werte. Der Marktforscher muss hier zwischen dem Datenverlust durch den Verlust von Daten und den Vorteilen aus der Reduktion fehlender Werte abwägen. Diese Vorgehensweise ist vor allem bei nicht zufällig auftretenden Werten (MAR, NRM) zu empfehlen, wobei der Ausschluss fallweise oder paarweise erfolgen kann.

Liegen metrische Daten vor, so besteht unter bestimmten Voraussetzungen auch noch die Möglichkeit, fehlende Werte über verschiedene induktive und statistische Verfahren zu ersetzen. Eine solche Ersetzung lässt sich nur durchführen, wenn klare Regelmäßigkeiten in den vorhandenen Daten erkennbar sind. Sie bringt stets die Gefahr mit sich, dass vorhandene Regelmäßigkeiten verstärkt werden und dass der Marktforscher die Daten in späteren Analysen so behandelt, als seien sie vollständig – die Angabe, inwiefern Daten durch welche Methode ersetzt wurden, hat daher Bestandteil jedes Untersuchungsberichts zu sein.

Ausschlussverfahren

Bei den Ausschlussverfahren ist, wie bereits oben dargestellt, in den fallweisen Ausschluss und den paarweisen Ausschluss zu unterscheiden.

Entscheidet sich der Marktforscher für den fallweisen Ausschluss, so kann er einzelne Fälle aus dem Datensatz entfernen, die besonders viele fehlende Werte aufweisen. Ausgeschlossen wird jeweils der komplette Datensatz – dadurch werden bestimmte Asymmetrien ausgeschlossen, die beim paarweisen Ausschluss entstehen können (siehe weiter unten), es geht aber auch relevantes Datenmaterial verloren (gültige Werte aus den jeweiligen Fällen) und der Stichprobenumfang nimmt ab.

Alternativ kann auch ein paarweiser Ausschluss durchgeführt werden. Dabei wird mit den gültigen Werten eines Falls weitergearbeitet, auch wenn dieser fehlende Werte aufweist. Hier bleiben also alle Fälle erhalten und auch der Stichprobenumfang sinkt nicht. Bei multivariaten Analysen kann sich allerdings von Variable zu Variable die Berechnungsgrundlage ändern, was zu Asymmetrien führt. Ein Beispiel: Angenommen, 90 von 100 Befragten hätten ihr er-

stes Gehalt angegeben, aber nur 45 von 100 ihr jetziges Gehalt. Sollen nun die Durchschnittsgehälter zu Karrierebeginn und aktuell miteinander vergleichen werden, würden die beiden Werte beim paarweisen Ausschluss auf unterschiedlicher Basis berechnet – die direkte Vergleichbarkeit ist damit zu bezweifeln. Um solche Probleme zu vermeiden, ist der fallweise Ausschluss das weitaus häufiger angewandte Ausschlussverfahren.

Ersatzwertverfahren

Bei den Ersatzwertverfahren ist in die nichtmathematischen, iterativen und die mathematischen, statistischen Verfahren zu unterscheiden.

Werden die fehlenden Werte ohne Rechnungen und auf der Basis von Informationen ersetzt, die über die Stichprobe vorliegen, so spricht man von induktiven Ersatzwertverfahren. Dazu zählen beispielsweise Nachfassaktionen oder auch Nachbeobachtungen, wobei letztere aber bei Zufallsstichproben nicht machbar sind, ohne die Repräsentativität zu gefährden. Alternativ lassen sich auch Konstanten, beispielsweise aus einer externen Quelle oder einer früheren Studie als Ersatzwerte für fehlende Werte verwenden.

Treten fehlende metrische Werte komplett zufällig auf (MCAR), lassen sie sich auch statistisch ersetzen. Ein einfaches Verfahren zum statistischen Ersatz ist beispielsweise der Mittelwertersatz: Ein fehlender Wert wird durch den Mittelwert der entsprechenden Variablen ersetzt. Dabei sind unterschiedliche Variationen möglich, zum Beispiel der Einsatz des arithmetischen Mittels oder des Medians der Nachbarpunkte, die Berechnung eines Zeitreihen-Mittelwerts (wo Zeitreihen-Daten vorliegen) oder die lineare Interpolation. Solche Formen des Mittelwertersatzes sind leicht anzuwenden, können aber die Verteilung der Daten, die Varianz der Variablen und eventuell auftretende Korrelationen in den Daten verzerren.

Lässt sich für die gültigen Werte ein deutlicher linearer Trend ermitteln, können die fehlenden Werte auch durch die entsprechenden Werte aus einem linearen Trendmodell ersetzt werden. Hierbei handelt es sich um eine Sonderform des statistischen Werteersatzes, welches zwei wesentliche Nachteile aufweist: Die Varianz der Variablen verringert sich in jedem Fall und lineare Trends werden erheblich verstärkt – unabhängig davon, wie stark sie in der

Grundgesamtheit tatsächlich ausgeprägt sind. Sämtliche weiterführenden Analyseverfahren, die auf linearen Trends aufbauen oder diese untersuchen (wie beispielsweise die multiple Regressionsanalyse) können mit derartig aufbereiteten Daten keinesfalls mehr durchgeführt werden.

VIII Prüfung auf Normalverteilung

1 Einführung

Die Gauß- oder Normalverteilung ist die wichtigste kontinuierliche Wahrscheinlichkeitsverteilung in der Statistik. Der Graph der zugehörigen Dichtefunktion ist als die Gaußsche Glockenkurve bekannt:

$$f(x)=\frac{1}{(\sigma\sqrt{(2\pi)})}e^{((\frac{-1}{2}\frac{(x-\mu)}{\sigma})^2)}$$

Die Normalverteilung ist durch drei charakteristische Eigenschaften bestimmt:

- Die Dichtefunktion ist glockenförmig und symmetrisch
- Der Erwartungswert, der Median und der Modus sind identisch
- Die normalverteilte Zufallsvariable weist eine unendliche Spannweite auf

Viele statistische Verfahren in der Marktforschung setzen voraus, dass Variablen in der Grundgesamtheit normalverteilt sind. Es ist daher häufig zu prüfen, ob von dem Vorliegen einer solchen Verteilung ausgegangen werden kann, wobei ein näherungsweises Vorliegen häufig für die Fortsetzung der Analyse ausreichend ist.

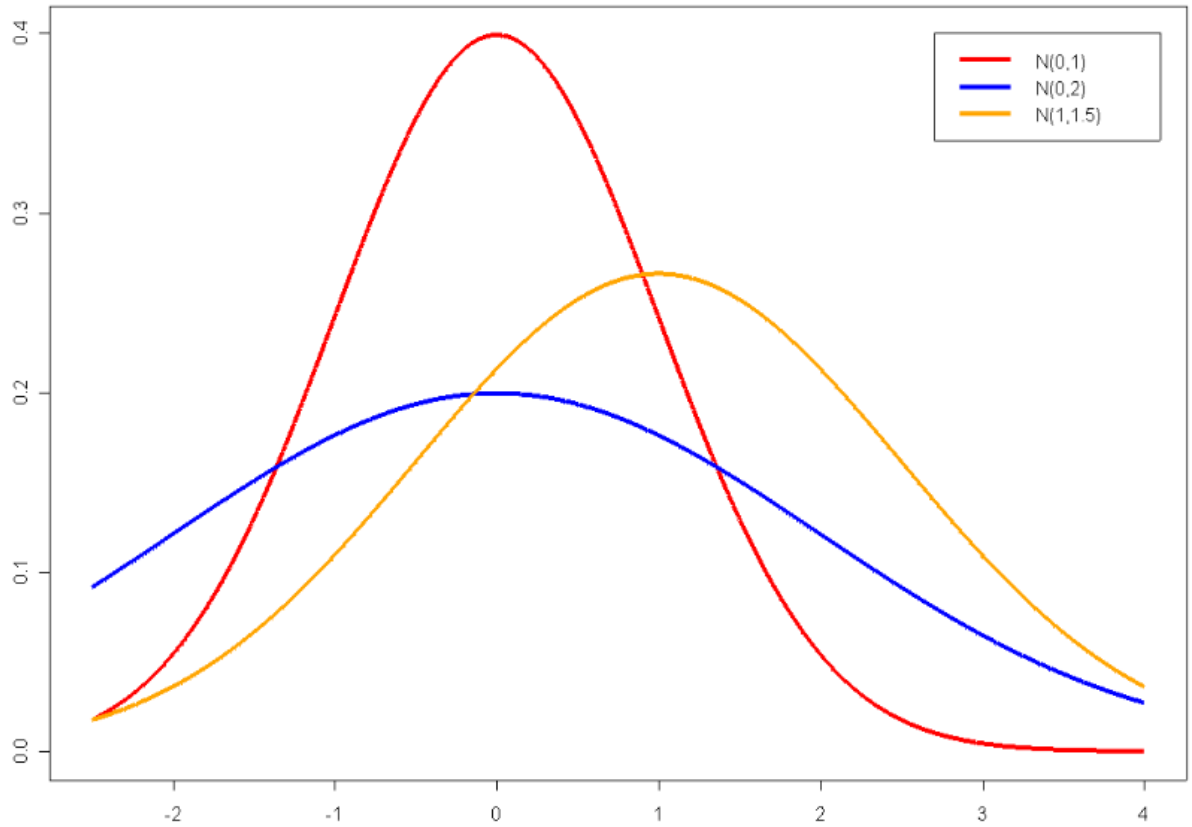

Abbildung 26: Dichtefunktion von normalverteilten Zufallsgrößen

Inwiefern die Normalverteilungsprüfung zu den Verfahren der explorativen Datenanalyse gehört, sei an dieser Stelle einmal dahingestellt. Sie wird ebenso wie die Prüfung auf Homoskedastizität im Rahmen dieses Manuskripts im Kapitel zur explorativen Datenanalyse vorgestellt, um sie überhaupt einem Kapitel zuordnen zu können. In allen anderen Kapiteln, in denen die Normalverteilungsprüfung oder die Prüfung auf Homoskedastizität eine Rolle spielen, kann dann auf die entsprechenden Abschnitte zurückverwiesen werden, um Wiederholungen zu vermeiden.

SPSS bietet dem Analytiker eine ganze Reihe von Prüfmöglichkeiten an, von denen die wichtigsten hier vorgestellt werden sollen: Die grafische Prüfung anhand des Histogramms, des Q-Q- oder auch des P-P-Diagramms bzw. ihrer trendbereinigten Varianten und die statistische Prüfung mit dem Kolmogorov-Smirnoff-Anpassungstest.

2 Grafische Prüfung

Die beliebteste Form der grafischen Prüfung auf Normalverteilung ist die Prüfung anhand eines Histogramms mit eingeblendeter Normalverteilungskurve. Wie bereits in Abschnitt IV dargestellt wurde, spiegeln die Balken des Histogramms die komplette Breite der Wertebereiche wieder. Da zudem für leere Wertebereiche ein Freiraum ausgegeben wird, kommt im Histogramm die gesamte empirische Verteilung der Variablen zum Ausdruck. Dies ermöglicht den direkten Vergleich mit einer eingezeichneten theoretischen Verteilung wie beispielsweise der Normalverteilung.

Je schwächer der Balkenverlauf dem Verlauf der eingeblendeten Normalverteilungskurve folgt, desto eher ist davon auszugehen, dass keine Normalverteilung vorliegt. Dabei ist zu beachten, dass es sich um eine Prüfung, aber keinen Test handelt. Dies bedeutet, dass in die Grafik ausschließlich die vorliegenden Werte aus der Stichprobe einfließen, die ja nicht zwangsweise die Verteilungsverhältnisse in der Grundgesamtheit optimal abbilden, sondern aufgrund von Zufallseffekten auch abweichende Verhältnisse aufweisen können. Einen Test auf Vorliegen einer Normalverteilung kann nur anhand eines „echten“ statistischen Tests wie des Kolmogorov-Smirnoff-Anpassungstests erfolgen, nicht aber anhand einer grafischen Prüfung – es sei denn, es liegen die Daten einer Vollerhebung vor.

Handelt es sich jedoch um Daten aus einer Stichprobe, was in der Praxis in der Regel der Fall sein wird, so stellen die grafischen Prüfungen lediglich einen Indikator dafür da, inwiefern ein Test überhaupt sinnvoll erscheint. Eine Ausnahme bilden Prüfungen auf eine Normalverteilung von Residuen (Abweichungen der erhobenen Werte von den prognostizierten Werten, siehe nächstes Kapitel) in der multiplen Regressionsanalyse, da hier im Grunde eine „Vollerhebung“ der Residuen vorliegt.

Alternativ zum Histogramm können auch die ebenfalls bereits in Kapitel 3 der explorativen Datenanalyse vorgestellten Q-Q-Diagramme, trendbereinigten Q-Q-Diagramme, P-P-Diagramme und trendbereinigten P-P-Diagramme zur grafischen Prüfung auf Normalverteilung oder andere Verteilungsformen eingesetzt werden.

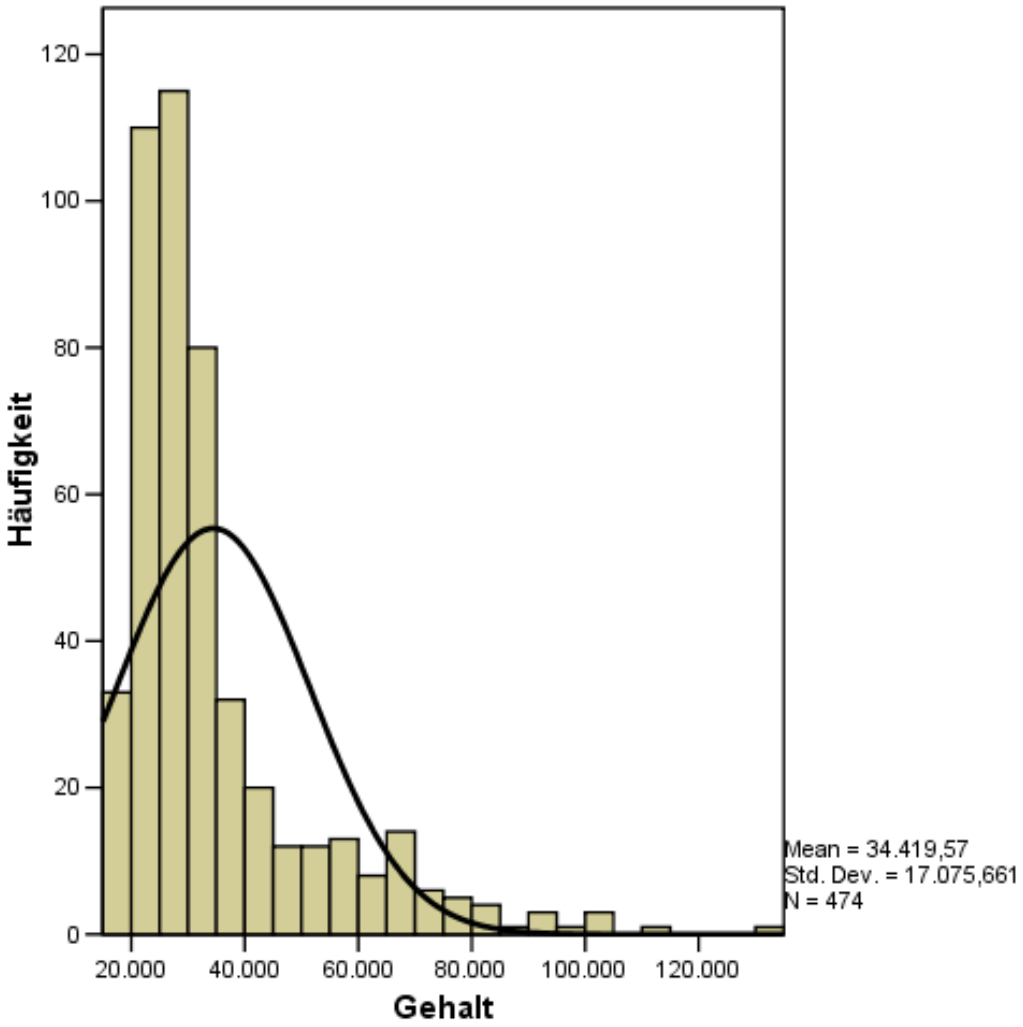

Abbildung 27: Histogramm mit eingezeichneter Normalverteilungskurve

Die Einschränkungen, die für die grafische Prüfung mittels des Histogramms gelten, sind ebenso für die weiteren Formen der grafischen Prüfung relevant.

3 Kolmogorov-Smirnoff-Anpassungstest

Die Prüfung auf das Vorliegen einer Normalverteilung erfolgt idealerweise mit einem Anpassungstest, wie beispielsweise dem Kolmogorov-Smirnoff-Anpassungstest. Ein solcher Test ist die einzige Möglichkeit, das Vorliegen einer Verteilung statistisch „sauber“ nachzuweisen. Ist also eine Voraussetzung für die weitere Analyse betroffen, sollte aufgrund der oben bereits betrachteten Einschränkungen bei grafischen Prüfungen immer ein solcher Anpassungstest

durchgeführt werden – die alleinige Interpretation eines Diagramms durch den Marktforscher ist subjektiv und insbesondere in Zweifelsfällen nicht ausreichend.

Der hier betrachtete Kolmogorov-Smirnoff-Anpassungstest (KSA) arbeitet mit der kumulierten empirischen Verteilung und der kumulierten erwarteten Referenzverteilung, in diesem Fall also der kumulierten Normalverteilung. Die maximale Differenz zwischen den beiden Verteilungen wird zur Berechnung der Prüfgröße Z nach Kolmogorov-Smirnoff verwendet, mit der dann aus einer Tabelle der für einen Stichprobenumfang n kritische Wert für die maximale Differenz bei einem durch den Marktforscher festgelegten Signifikanzniveau abgelesen werden kann.

Kolmogorov-Smirnov-Anpassungstest

		Gehalt
N		474
Parameter der Normalverteilung[a,b]	Mittelwert	34.419,57
	Standardabweichung	17.075,661
Extremste Differenzen	Absolut	,208
	Positiv	,208
	Negativ	-,143
Kolmogorov-Smirnov-Z		4,525
Asymptotische Signifikanz (2-seitig)		,000

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Abbildung 28: Kolmogorov-Smirnoff-Anpassungstest auf Normalverteilung in SPSS

Die Nullhypothese H0 des KSA in SPSS lautet: Die Werte der untersuchten Variablen sind in der Grundgesamtheit normalverteilt. Ausgegeben wird, wie bei allen statistischen Tests in SPSS, die Wahrscheinlichkeit für einen Fehler beim Zurückweisen der Nullhypothese (die sogenannte Irrtumswahrscheinlichkeit). Je größer diese Wahrscheinlichkeit ausfällt, desto eher ist von einer tatsächlichen Normalverteilung der Werte auszugehen.

IX Prüfung auf Varianzgleichheit

1 Einführung

Viele statistische Verfahren setzen voraus, dass die Varianzen einer Variablen innerhalb verschiedener Fallgruppen identisch oder zumindest näherungsweise identisch sind, so beispielsweise diverse Signifikanztests oder Vergleiche zweier arithmetischer Mittel. Liegt eine Gleichheit der Varianzen vor, so ist von Homoskedastizität die Rede, sind die Varianzen statt dessen ungleich, von Heteroskedastizität.

Ähnlich wie die Prüfung auf Normalverteilung kann auch die Prüfung auf Homoskedastizität anhand diverser grafischer Darstellungen oder mittels eines statistischen Tests erfolgen. Die grafische Prüfung auf Homoskedastizität erfolgt in der Regel mittels des bereits weiter oben vorgestellten gruppierten Box-Plots oder mittels eines Streudiagramms. Es gelten die bereits für die grafische Prüfung auf Normalverteilung dargestellten Einschränkungen bezüglich der Ergebnisinterpretation bei Prüfungen in Stichprobendaten.

Als statistischer Test kommt der Levene-Test in Frage, der im folgenden Unterabschnitt noch im Detail betrachtet wird. Auf eine eingehendere Darstellung der grafischen Prüfung wird im Rahmen dieses Manuskripts verzichtet.

2 Levene-Test

Mit dem Signifikanztest nach Levene wird die Nullhypothese H0 überprüft, dass die Varianz einer Variablen in der Grundgesamtheit in allen Gruppen homogen ist.

Der Test arbeitet mit dem F-Wert als statistischem Prüfmaß mit bekannter Verteilung. Es wird getestet, mit welcher Wahrscheinlichkeit die in der Stichprobe beobachteten Abweichungen in den Varianzen auftreten können, wenn in der Grundgesamtheit eine völlige Varianzgleichheit herrscht. Diese Wahrscheinlichkeit wird durch SPSS ausgewiesen und stellt gleichermaßen die Irrtumswahrscheinlichkeit bei einer Ablehnung der H0 dar. Eine geringe

Wahrscheinlichkeit weist daher auf eine Varianzungleichheit hin, da H0 in diesem Fall problemlos abgelehnt werden kann.

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	1,1E+11	2	5,46E+10	898,947	,000[a]
	Residuen	2,9E+10	471	60785788		
	Gesamt	1,4E+11	473			

a. Einflußvariablen : (Konstante), Ausbildung (in Jahren), Anfangsgehalt

b. Abhängige Variable: Gehalt

Abbildung 29: Levene-Test auf Varianzgleichheit in SPSS

X Arbeit mit Dummy-Variablen

Viele statistische Analyseverfahren setzen ein metrisches Skalenniveau voraus, so beispielsweise die multiple Regressionsanalyse. Sollen nun nominalskalierte Variablen in eine solche Analyse einfließen, so müssen sogenannte Dummy-Variablen gebildet werden. Die Bildung dieser Variablen wird zum Abschluss dieses Kapitels betrachtet, auch wenn sie nicht direkt zur explorativen Datenanalyse gehört.

Bei Dummy-Variablen handelt es sich um binäre Variablen, also um Variablen, die nur die Werte 0 und 1 annehmen können. Eine dichotome Variable (eine Variable mit lediglich zwei Ausprägungen) lässt sich durch eine einfache Transformation leicht in eine Dummy-Variable überführen: Liegt eine festgelegte Ausprägung vor, nimmt die Variable den Wert 1 an, liegt sie dagegen nicht vor, so nimmt die Variable den Wert 0 an.

Dazu ein Beispiel: Ein Marktforscher untersucht die Auswirkungen unterschiedlicher Verpackungsdesigns auf das Kaufverhalten, wobei auch die Farbgestaltung der Verpackung als Einflussgröße miteinbezogen werden soll. Die Dummy-Variable q1 nimmt nun für rote Verpackungen den Wert 1, für nicht-rote Verpackungen den Wert 0 an. Liegen nur zwei mögliche Ausprägungen vor (beispielsweise rot und grün), so lassen diese sich in einer einzigen Dummy-Variable abbilden. Für weitere Farben lassen sich weitere Dummy-Variablen definieren,

so dass auch nicht-dichotome Sachverhalte ausgedrückt werden können.

Dabei ist zu beachten, dass eine nominale Variable mit n Ausprägungen stets durch n-1 Dummy-Variablen abgebildet werden kann, da sich der Wert für die „letzte“ Ausprägung aus den anderen Dummy-Variablen ergibt. Soll beispielsweise neben Rot und Grün noch die Farbe Gelb betrachtet werden, sind dafür nur zwei statt drei Dummy-Variablen erforderlich, da bei q1 = 0 und q2 = 0 klar ist, dass q3 = 1 also die Verpackung gelb wäre.

Dummy-Variablen sind vorsichtig und mit Verstand einzusetzen. Es ist nicht sinnvoll, zehn nominalskalierte Variablen in Dummy-Variablen umzuwandeln und damit in ein Analyseverfahren einzusteigen, welches metrisch skalierte Daten voraussetzt. Neun metrisch skalierte Variablen dagegen durch eine fachlich sinnvolle Dummy-Variable zu ergänzen, ist methodisch als angemessener zu betrachten. Die Information, ob Dummy-Variablen verwendet wurden und, wenn ja, wie sich diese zusammensetzen, gehört in jeden Analysebericht.

XI Weiterführende Literatur

Bleymüller, J.; Gehlert, G. & Gülicher, H. (2000). Statistik für Wirtschaftswissenschaftler. München: Verlag Vahlen

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (1999). Statistik. Der Weg zur Datenanalyse (2. Aufl.). Berlin: Springer.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). Multivariate data analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Heiler, S. & Michels, P. (1994). Deskriptive und Explorative Datenanalyse. München: Ol-

denbourg.

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

C Multiple Regression

I Einführung

1 Hintergründe der Regressionsanalyse

Die multiple Regressionsanalyse ist das flexibelste und in der Praxis sowohl in der Markt- als auch in der Sozialforschung am häufigsten eingesetzte multivariate Analyseverfahren. Sie dient der Untersuchung der Beziehung zwischen einer abhängigen und einer oder mehrerer unabhängigen Variablen. Die Regressionsanalyse kann unter anderem in der Ursachenanalyse (Darstellung und Quantifizierung von Wirkungszusammenhängen) und in der Prognostik (Prognose der Werte der abhängigen Variablen) zum Einsatz kommen.

Ein einfaches Beispiel: Wie verändert sich die Absatzmenge eines Konsumguts (abhängige Variable) bei Veränderungen am Produktpreis, an den Werbeausgaben oder der Anzahl der Vertreterbesuche pro potentiellem Kunden (unabhängige Variablen)? Sind lineare Zusammenhänge zu vermuten, dann ist eine multiple Regressionsanalyse hier das Idealverfahren zur Untersuchung des Sachverhalts.

Ein Ergebnis einer Regressionsanalyse ist stets die Regressionsfunktion:

- $Y = f(x)$ >> einfache Regression
  (eine abhängige und eine unabhängige Variable)
- $Y = f(x1, x2, x3, ... xn)$ >> multiple Regression
  (eine abhängige und mehrere unabhängige Variablen)

Einen besonderen Problemfall stellen die sogenannten interdependenten Beziehungen dar. So ist beispielsweise bei einer Untersuchung von Bekanntheitsgrad und Absatzmenge eines Produkts nicht unbedingt klar, ob der Bekanntheitsgrad die Absatzmenge beeinflusst (bekannte Produkte werden häufiger gekauft) oder umgekehrt (Produkte, die zu großen Stückzahlen im Umlauf sind, sind auch bekannter). Solche interdependenten Systeme lassen sich nicht in

einer einzigen Regressionsfunktion erfassen, sondern nur in Mehrgleichungsmodellen. Sie werden im Rahmen dieses Manuskripts nicht weiter beachtet.

2 Exkurs: Korrelation und Kausalität

Zum Einstieg in die Regressionsanalyse sei noch einmal explizit darauf hingewiesen, dass es sich um ein strukturprüfendes Verfahren und nicht um ein strukturentdeckendes Verfahren handelt. Es kann nicht dazu eingesetzt werden, nach Kausalitäten zu suchen, sondern lediglich, ein vorhandenes, auf Kausalitäten aufbauendes Modell zu überprüfen.

Warum ist dieser Unterschied wichtig? Als unerfahrener Marktforscher ist man oft (und bei der Regressionsanalyse trifft dies im besonderen Maße zu) versucht, Korrelationen auch gleich kausal zu interpretieren, oder grundsätzlich davon auszugehen, dass Korrelationen von Kausalitäten herrühren. Korrelation und Kausalität sind aber mitnichten dasselbe. Es handelt sich um vollkommen unterschiedliche Konzepte, die gerade bei der Regressionsanalyse nicht durcheinandergebracht werden sollten.

Eine Korrelation kann grundsätzlich auf eine Kausalität hindeuten, sie ist eine notwendige, aber noch keine hinreichende Bedingung für das Vorliegen von Kausalität. Liegt beispielsweise eine Korrelation zwischen den Variablen A und B vor, so könnte man vermuten, dass die Variable A die Variable B beeinflusst. Es könnte aber genauso gut sein, dass die Variable B die Variable A beeinflusst – das reine Auftreten einer Korrelation und der Grad derselben gestatten noch keinen definitiven Rückschluss auf Kausalzusammenhänge. Schließlich wäre es auch möglich, dass eine Scheinkorrelation vorliegt – die Variablen A und B würden in diesem Falle gar nicht wirklich direkt miteinander korrelieren, sondern mit einer dritten Variable C.

3 Der Ablauf der Regressionsanalyse

Die Regressionsanalyse lässt sich in vier wesentliche Arbeitsschritte unterteilen.

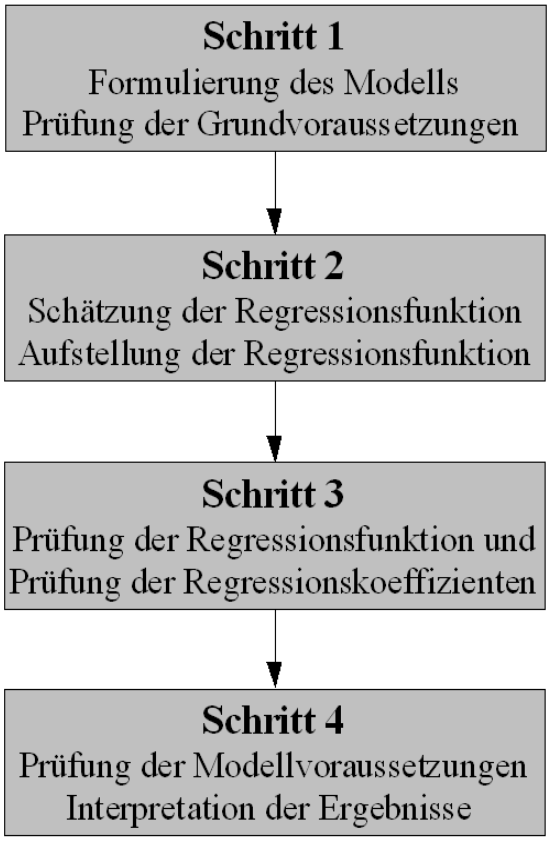

Abbildung 30: Der Ablauf der Regressionsanalyse

Im ersten Arbeitsschritt muss zunächst dass zu untersuchende Modell bestimmt werden, insbesondere sind die abhängige und die unabhängige(n) Variable(n) festzulegen, wobei hier fachliche Überlegungen im Vordergrund stehen müssen. Außerdem sind verschiedene Grundvoraussetzungen bezüglich des Skalenniveaus und des vermuteten Kausalgeflechts zu überprüfen.

Im zweiten Schritt werden dann die Regressionskoeffizienten anhand der Methode der kleinsten Quadrate berechnet, anschließend kann dann mit den Regressionskoeffizienten die Regressionsfunktion – das eigentliche Ergebnis der Regressionsanalyse – aufgestellt werden.

Bevor man die Regressionskoeffizienten und die Regressionsfunktion inhaltlich interpretieren kann, ist im dritten Schritt zu prüfen, ob (a) die gefundene Funktion als Ganzes die abhängige Variable gut erklären kann und (b) welchen Beitrag die einzelnen unabhängigen Variablen zum Gesamtmodell leisten.

Im vierten Schritt ist dann noch zu prüfen, inwieweit die Modellprämissen eingehalten wurden, insbesondere ob keine Autokorrelation der Residuen und keine Multikollinearität vorliegt. Ist das gefundene Modell valide, kann es inhaltlich interpretiert werden, sind dagegen die Voraussetzungen grob verletzt worden, so kann es auch in diesem letzten Schritt noch zu einem Abbruch der Regressionsanalyse bzw. zu einer Verwerfung der bisherigen Erkenntnisse kommen.

II Formulierung des Regressionsmodells

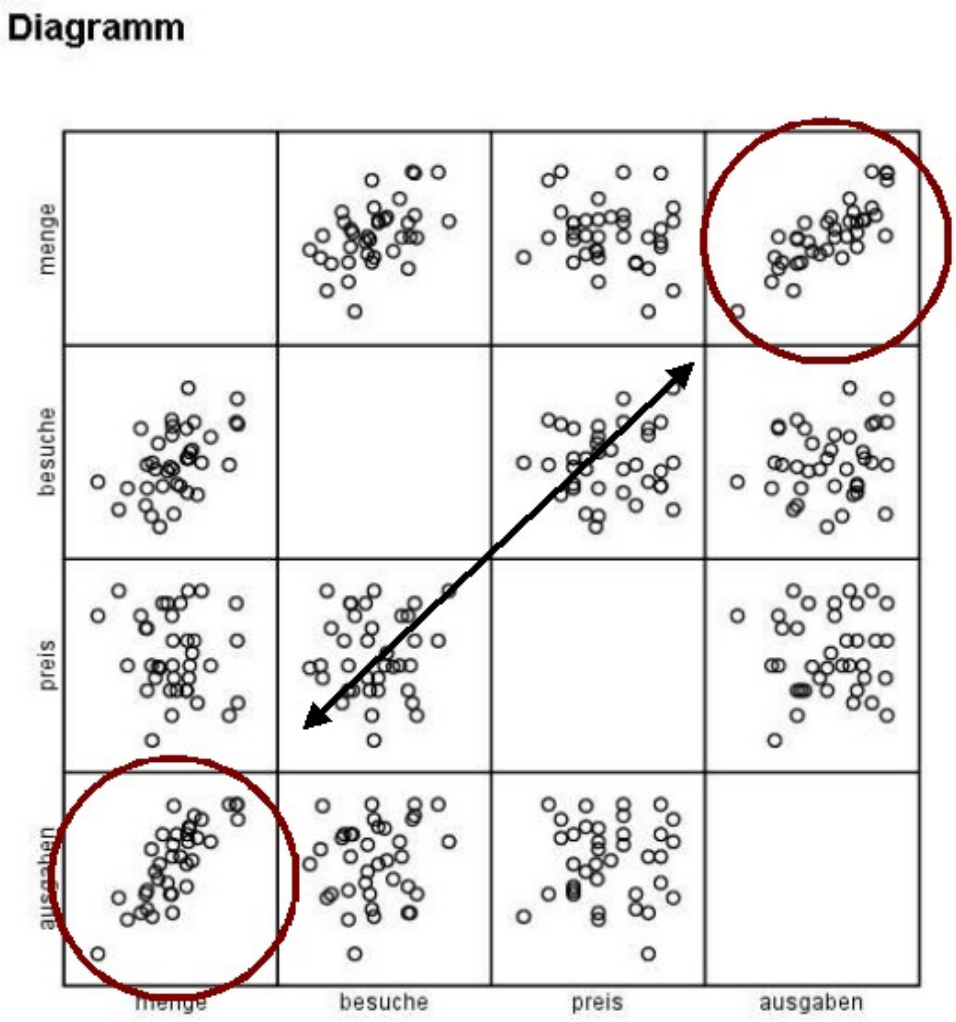

Abbildung 31: Aufdeckung eines möglichen linearen Zusammenhangs im Matrixdiagramm

Wie bereits oben dargestellt wurde, handelt es sich bei der Regressionsanalyse um ein strukturprüfendes Verfahren – sie dient insbesondere nicht der Aufdeckung unbekannter Zusammenhänge. Das zu untersuchende Regressionsmodell muss daher bereits vor Beginn der Analyse auf der Basis von Sachinformationen gebildet werden. Es sollte so konstruiert sein, dass von einer möglichst vollständigen Abbildung der Ursache-Wirkungs-Beziehung ausgegangen werden kann. Idealerweise sollte dieses Modell nicht vom Marktforscher allein, sondern unter Zuhilfenahme eines entsprechenden Fachexperten formuliert werden.

Als Hilfe bei der Auffindung der für die Aufnahme ins Modell geeigneten Variablen eignen sich Streudiagramme für univariate Fälle bzw. Matrixdiagramme für multivariate Fälle. Ein linearer Zusammenhang – und nur ein solcher kann im Rahmen der Regressionsanalyse untersucht werden – ist im Streu- oder Matrixdiagramm immer dann zu vermuten, wenn die Punkte im Diagramm eng um eine gedachte Linie streuen.

III Prüfung der Analysevoraussetzungen

1 Generelle Analysevoraussetzungen

Die Analysevoraussetzungen sind im Wesentlichen in zwei Gruppen zu unterteilen: Die Gruppe der Grundvoraussetzungen, die vor dem Beginn der eigentlichen Regressionsanalyse im ersten Schritt überprüft werden müssen und die Gruppe der übrigen Modellvoraussetzungen, die erst im Anschluss an die eigentliche Regressionsanalyse im vierten Schritt überprüft werden können.

Zu den Grundvoraussetzungen gehören:

- Das Kausalgeflecht (abhängige und unabhängige Variablen) muss bekannt sein oder vermutet werden. Die Regressionsanalyse dient ausschließlich der Prüfung von Zusammenhängen und nicht deren Auffinden.
- Der Zusammenhang zwischen abhängiger und unabhängigen Variablen muss linear sein. Ein quadratischer, logarithmischer oder sonstwie anders gearteter Zusammen-

hang kann mit der Regressionsanalyse nicht sinnvoll geprüft werden.

- Alle verwendeten Variablen müssen metrisch skaliert sein, da das Standardmittel in die Rechnungen einfließt. Für die unabhängigen Variablen lassen sich gegebenenfalls auch Dummy-Variablen verwenden.

Die übrigen Modellvoraussetzungen umfassen:

- Unabhängige Variablen dürfen nicht untereinander korrelieren (Multikollinearität)
- Die standardisierten Residuen (durch das Modell nicht erklärte Abweichungen):
  - müssen näherungsweise normal verteilt sein
  - müssen die gleiche Varianz aufweisen (Homoskedastizität)
  - dürfen nicht untereinander korrelieren (Autokorrelation)

2 Transformation nichtlinearer Variablen

Das lineare Regressionsmodell dient nicht der Bestimmung der optimalen Kurvenanpassung in allen Fällen. Es setzt einen linearen Zusammenhang zwischen abhängigen und unabhängigen Variablen voraus. Dies bedeutet allerdings nicht, dass nichtlineare Zusammenhänge keinesfalls in die Analyse einfließen dürfen: Liegen solche Zusammenhänge vor, ist die Transformation einzelner Variablen möglich.

Dazu ein Beispiel: Bei Wachstumsprozessen kommt es häufig vor, dass sich die unabhängige Variable linear, die abhängige aber exponentiell verändert (beispielsweise bei der Umweltbelastung durch bestimmte Schadstoffe). Bei einer solchen zeitgebundenen exponentiellen Entwicklung lässt sich der Zusammenhang zwischen der Umweltbelastung (abhängige Variable) und der Zeit (unabhängige Variable) als exponentielle und damit nichtlineare Gleichung darstellen. Wird diese nicht für die Regressionsanalyse geeignete Gleichung nun aber logarithmiert, so ergibt sich ein linearer Zusammenhang, der eine Regressionsanalyse gestattet.

Vorsicht: In diesem Fall bilden die logarithmierten Werte für die Umweltbelastung die ab-

hängige Variable. Dies ist bei der Interpretation der Ergebnisse unbedingt zu beachten.

IV Schätzung der Regressionsfunktion

1 Grundprinzip der Schätzung

Das Grundprinzip dieser Schätzung, die das Herzstück der Regressionsanalyse darstellt, sei hier der Übersichtlichkeit halber am Beispiel der einfachen linearen Regression mit einer abhängigen und einer unabhängigen Variablen beschrieben.

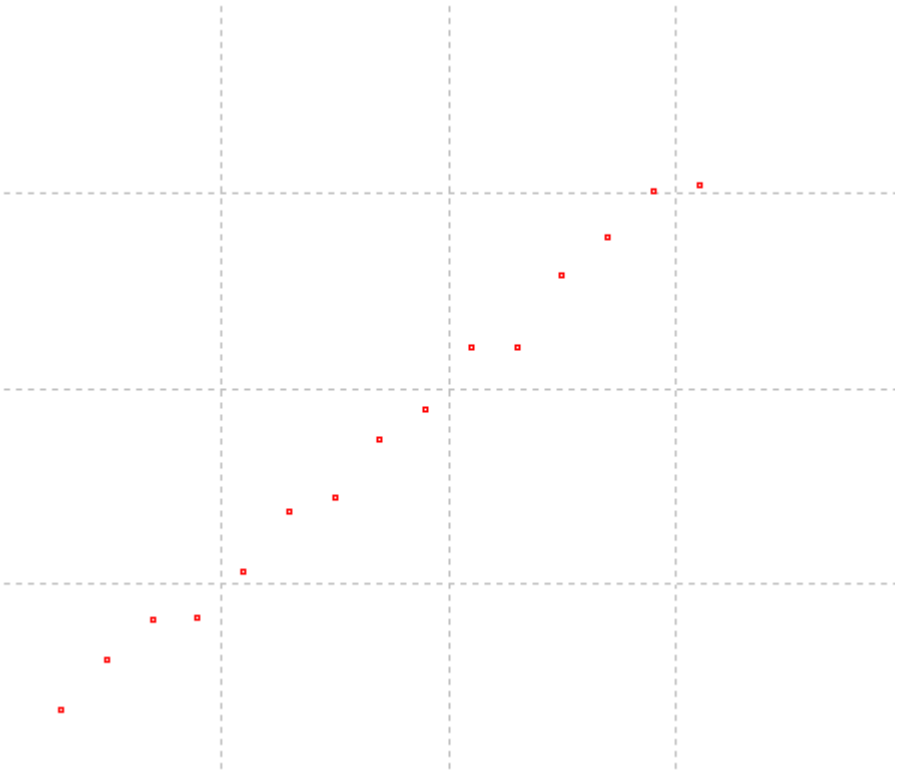

Abbildung 32: Linearer Zusammenhang im Streudiagramm

Der im Beispieldiagramm dargestellte Zusammenhang zwischen den beiden Variablen ist keineswegs perfekt, ein linearer Trend ist jedoch klar zu erkennen. Es kommen nun theoretisch mehrere Regressionsgeraden in Frage, mit denen sich dieser Zusammenhang mathematisch beschreiben ließe.

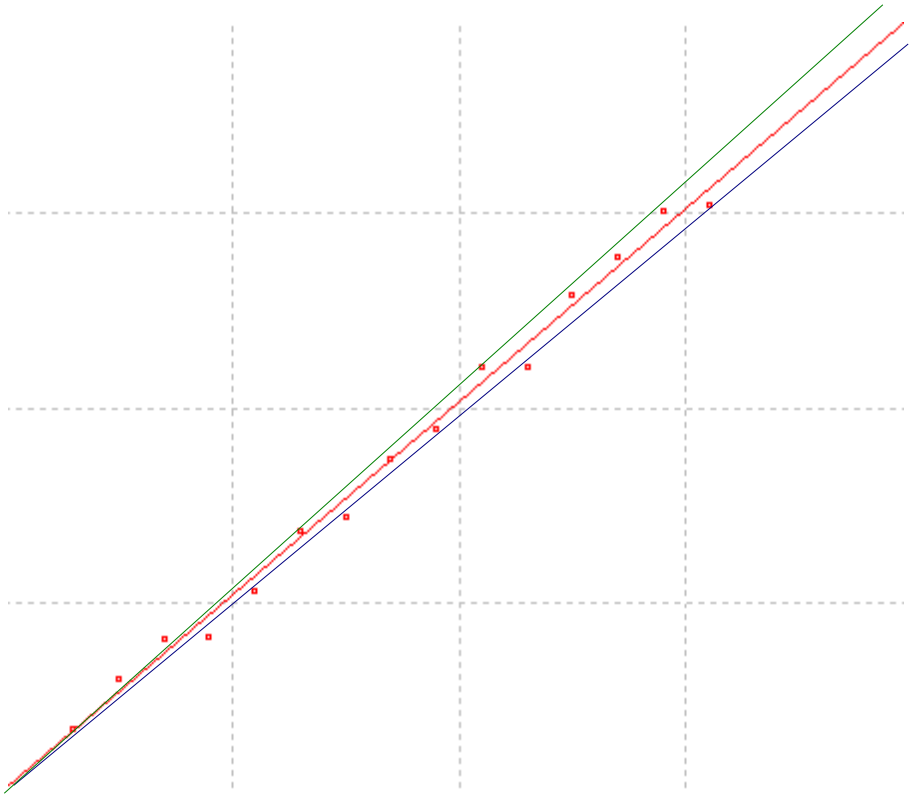

Abbildung 33: Mögliche lineare Regressionsgeraden im Streudiagramm

Die entscheidende Frage lautet nun: Welche dieser möglichen Geraden beschreibt den linearen Zusammenhang zwischen beiden Variablen am besten? Mit bloßem Auge ist dies ganz offensichtlich nicht zu erkennen, wobei es zudem sehr viel mehr mögliche Geradenverläufe gäbe. Es muss also ein mathematischer Ansatz gefunden werden, mit dem sich die bestmögliche Regressionsgerade ermitteln lässt.

2 Auswahl einer Geraden

Einen Ansatzpunkt für eine solche Rechnung kann man direkt in der Grafik erkennen: Die Abstände zwischen den „echten“, beobachteten Punkten und der Gerade an derselben Stelle, also den durch die Regressionsfunktion geschätzten Punkten. Da die Punkte auf beiden Seiten der Gerade liegen, könnte man ja vielleicht diejenige Gerade auswählen, bei der sich die positiven und negativen Abstände gegenseitig aufheben – der durchschnittliche Schätzfehler bei

dieser Graden würde dann Null betragen, was ja ein gutes Auswahlkriterium sein könnte.

Betrachten wir einmal, zu welcher Regressionsgeraden dieses Kriterium führt. Der senkrechte Abstand eines beliebigen Punktes i zur Graden

$$Y = a + b * X$$

berechnet sich als:

$$e_i = Y_i - (a + b * X_i)$$

Für eine Beispielerhebung mit 100 Fällen kann i die Werte von 1 bis 100 annehmen. Die Summe aller Abstände berechnet sich in diesem Fall dann als:

$$\sum e_i = \sum (Y_i - (a + b * X_i)) = \sum Y_i - 100 * a - b * \sum X_i$$

Soll die Gesamtsumme aller Abstände Null sein (durchschnittlicher Schätzfehler von Null), dann gilt:

$$\sum Y_i - 100 * a - b * \sum X_i = 0$$

Und schließlich dividiert durch die Zahl der Beobachtungen:

$$(\frac{\sum Y_i}{100}) - a - b * (\frac{\sum X_i}{100}) \rightarrow \bar{Y} - a - b * \bar{X} = 0$$

Diese Rechnung führt unweigerlich zu dem Schluss, dass die Summe aller Abstände stets Null beträgt, wenn gilt:

$$\bar{Y} = a + b * \bar{X}$$

Dies bedeutet, dass alle Geraden, die durch die beiden Mittelpunkte verlaufen, als ideale Geraden selektiert werden würden. Das Problem hierbei ist, dass es deutlich mehr als eine Gerade gibt, welche diese Bedingungen erfüllt, so dass dieses Kriterium offenbar nicht zu einer eindeutigen Geradenauswahl führt. Dazu kommt noch, dass die Steigung der Geraden hier vollkommen irrelevant wäre – jede Gerade, welche durch beide Mittelpunkte verläuft wäre gleich gut zur Schätzung der Punkte geeignet[1].

Die hier skizzierte Vorgehensweise ist daher für die Ermittlung der optimalen Regressionsgerade ungeeignet. Der Gedanke, die Abstände zwischen den wahren Punkten und den geschätzten Punkten zu nutzen, um den optimalen Geradenverlauf zu identifizieren, ist aber keineswegs verkehrt. Genau dieser Gedanke bildet die Grundlage für die sogenannte „Methode der kleinsten Quadrate“, dem „wirklich“ in der Regressionsanalyse verwendeten Verfahren zur Ermittlung der Regressionsgerade.

3 Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate oder auch Methode zur Minimierung der Summe der Abweichungsquadrate, ist bereits aus der in der Statistik I besprochenen linearen Regressionsanalyse bekannt, die nichts anderes als den einfachst möglichen Fall der Regressionsanalyse darstellt, nämlich die Regression mit einer abhängigen und einer unabhängigen Variablen.

Wie bereits oben erwähnt, arbeitet auch diese Methode mit den senkrechten Abständen der real erhobenen Werte von der Regressionsgeraden. Die Abstände werden jedoch quadriert, so dass sämtliche negativen Vorzeichen wegfallen. Auf diese Weise wird eine Kompensation der negativen und der positiven Abstände vermieden, wodurch eine ähnliche Problematik wie bei der oben diskutierten Methode umgangen wird. Gesucht wird nun nach derjenigen Regressionsgeraden, bei der die Summe der quadrierten Abweichungen minimal ist.

1 Die ausführliche Diskussion dieser „falschen“ Methode zur Ermittlung der Regressionsgeraden findet sich bei Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Durch die Umformung der Zielfunktion:

$$\sum_{k=1}^{K} e_k^2 = \sum_{k=1}^{k} [\, y_k - a + b * x_k]^2 \rightarrow min\,!$$

erhält man die Parameter der Regressionsfunktion:

$$b = \frac{(I(\sum x_I * y_k) - \sum x_I * \sum y_I)}{(I(\sum x_k^2) - \sum x_k)^2}$$

und:

$$a = \bar{y} - b * \bar{x}$$

Die Gleichung der Regressionsgraden im Einfaktoren-Fall lautet:

$$Y = a * b + X$$

Die Gleichung der Regressionsgraden im Mehrfaktoren-Fall lautet dementsprechend:

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \ldots + b_j * X_j + \ldots + b_J * X_J$$

Die Berechnung der Regressionsparameter erfolgt analog zum Einfaktoren-Fall.

4 Aufstellung der Regressionsgleichung

Mit der Methode der kleinsten Quadrate lassen sich, wie oben gezeigt, die Konstante und die Regressionskoeffizienten der Regressionsgleichung berechnen. Diese Arbeitsschritte nimmt SPSS dem Marktforscher ab und gibt die Regressionsgleichung mehr oder weniger direkt aus. Bei der Arbeit mit SPSS ist unbedingt zu beachten, dass immer eine Regressionsglei-

chung berechnet wird, unabhängig davon, ob die Vorbedingungen für die Regressionsanalyse erfüllt sind.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B		Kollinearitätsstatistik	
		B	Standardfehler	Beta			Untergrenze	Obergrenze	Toleranz	VIF
1	(Konstante)	311,219	170,379		1,827	,077	-35,032	657,470		
	besuche	9,513	1,816	,396	5,238	,000	5,822	13,205	,978	1,023
	ausgaben	,550	,055	,752	9,945	,000	,438	,662	,978	1,023

a. Abhängige Variable: menge

Abbildung 34: Ausgabe der Koeffizienten-Tabelle in SPSS

Die Regressionsgleichung lässt sich aus der Spalte „nicht standardisierte Koeffizienten“ in der Koeffizienten-Tabelle ablesen. Im Beispielfall ergibt sich die Regressionsgleichung:

Y (Verkaufsmenge) = 311,219 + 9,513 * Besuchszahl + 0,55 * Werbeausgaben

5 Regressions- und Beta-Koeffizienten

Häufig ist es interessant festzustellen, welchen Einfluss die einzelnen unabhängigen Variablen auf die abhängige Variable ausüben. Welche der Variablen beeinflusst Y also am stärksten, welche am geringsten.

Zur Beantwortung dieser Frage ist der einfache Vergleich der Korrelationskoeffizienten nicht ausreichend. Da die erklärenden Variablen in unterschiedlichen Dimensionen (wie z.B. Werbeausgaben in Euro und absolute Anzahl der Kundenbesuche im Beispielfall) vorliegen können, werden auch die Koeffizienten in unterschiedlichen Dimensionen ausgegeben. Eine Änderung der Dimensionen (beispielsweise Prozent- statt Absolutwerte) hat demnach einen unmittelbaren Einfluss auf den Koeffizienten – aber natürlich keinen Einfluss auf den Erklärungsgehalt der Variablen an sich. Aus diesem Grund ist der direkte Vergleich der Regressionskoeffizienten zur Klärung der Frage nach der Bedeutung der Variablen unzulässig.

Die Lösung besteht in der Berechnung der standardisierten Beta-Koeffizienten, die in SPSS unmittelbar in der Spalte neben den standardisierten Koeffizienten in der Koeffizienten-

Tabelle ausgegeben werden.

Die Beta-Koeffizienten werden berechnet, indem vor dem Beginn der Regressionsanalyse alle Variablen einer Z-Transformation unterzogen werden. Alternativ lassen sie sich über die nachfolgende Formel auch direkt aus den standardisierten Regressionskoeffizienten berechnen:

$$beta_i = b_i * (\frac{s_{xi}}{s_y})$$

mit:	betai	=	Beta-Koeffizient der unabhängigen Variablen i
	bi	=	Regressionskoeffizient der unabhängigen Variablen i
	sxi	=	Standardabweichung der unabhängigen Variablen i
	sy	=	Standardabweichung der abhängigen Variablen Y

V Prüfung der Regressionsfunktion

1 Einführung

Da sich mit SPSS, wie bereits oben dargestellt, immer eine Regressionsfunktion berechnen lässt, stellt sich die Frage nach deren Güte. Wie gut wird die abhängige Variable durch das aufgestelle Regressionsmodell erklärt?

Zur Feststellung dieser sogenannten Anpassungsgüte bieten sich drei Kennwerte an:

- das Bestimmtheitsmaß R^2 und das korrigierte R^2 (bei multivariaten Modellen)
- der Standardfehler der Schätzung
- die F-Statistik

Am gängigsten ist dabei die Berechnung von R^2 und korrigiertem R^2, die auch als allgemeine Gütemaße der Regressionsanalyse bezeichnet werden.

2 R^2 und korrigiertes R^2

Die Regressionsgerade gibt Zusammenhänge, die nicht perfekt linear sind[1], auch nicht perfekt wieder. Es ist daher mit der Regressionsfunktion in der Regel nicht möglich, alle Veränderung der abhängigen Variablen Y durch die unabhängigen Variablen zu erklären. Ein Teil der Streuung der abhängigen Variablen wird daher durch das Modell erklärt werden, ein anderer Teil wird unaufgeklärt bleiben.

Das Verhältnis von erklärter Streuung zur Gesamtstreuung ist ein gutes Maß für die Güte des Regressionsmodells. Die Residuen werden quadriert, damit sich positive und negative Abweichungen nicht gegenseitig aufheben. Aus dem Verhältnis von erklärter Streuung zu Gesamtstreuung ergibt sich das Gütemaß R^2:

- TSS = Total Sum of Squares = Summe aller quadrierten Abweichungen
- ESS = Explained Sum of Squares = Summe aller erklärten quadrierten Abweichungen
- RSS = Residual Sum of Squares = Summe aller nicht erklärten quadrierten Abweichungen

Die Relation von erklärter Streuung zu Gesamtstreuung wird mit R^2 bezeichnet:

$$R^2 = \frac{ESS}{TSS}$$

R^2 gibt also den Anteil der erklärten Streuung an der Gesamtstreuung an und drückt damit die Güte der Anpassung der Regressionsgerade an die Lage der Werte aus. R^2 ist als prozentualer Wert zu verstehen und liegt daher stets zwischen Null und Eins. Wird R^2 gleich Eins, so wird die gesamte Streuung durch das Regressionsmodell aufgeklärt – es besteht also ein perfekter linearer Zusammenhang. Je kleiner R^2 ausfällt, desto stärker weicht der vorliegende Fall

1 Das Auftreten solcher Zusammenhänge ist in der Praxis nicht zu erwarten, vor allem, da die Daten für die Regressionsanalyse meist aus einer Zufallsstichprobe stammen, bei der sich noch Zufallseffekte ergeben würden, selbst wenn in der Grundgesamtheit ein perfekter linearer Zusammenhang vorläge.

von diesem Zusammenhang ab.

Vorsicht: R^2 ist lediglich ein Maß für die Stärke eines linearen Zusammenhangs, nicht aber für andere Zusammenhänge[1].

Zusätzlich zu R^2 wird von SPSS noch das korrigierte R^2 berechnet. Wieso ist dieses zusätzliche Gütemaß noch erforderlich? Gibt R^2 die Güte des Regressionsmodells nicht mit ausreichender Genauigkeit wieder?

Das Problem mit R^2 ist, dass die Aufnahme zusätzlicher erklärender Variablen (also unabhängiger Variablen) nie zu einer Verschlechterung von R^2 führt. Besteht gar kein Zusammenhang zwischen der neuen unabhängigen Variablen und der abhängigen Variablen bleibt R^2 unverändert. Besteht auch nur ein minimaler Zusammenhang oder ein Scheinzusammenhang, steigt R^2 leicht an. In keinem Fall aber kann R^2 sich verschlechtern.

Dies kann dazu führen, dass der Marktforscher beliebig viele unabhängige Variablen geradezu wahllos in das Regressionsmodell aufnimmt. Es ergibt sich ein hohes R^2 und damit ein vermeintlich gutes Regressionsmodell. Die prognostizierten Werte werden jedoch mit steigender Zahl der unabhängigen Variablen unzuverlässiger. Daher sollte man keine Variablen zur Minimalsteigerung von R^2 ins Regressionsmodell aufnehmen.

Zur Entscheidung der Frage, ob der zusätzliche Erklärungsgehalt einer weiteren unabhängigen Variablen die Zunahme an prognostischer Unsicherheit rechtfertigt, kann das korrigierte R^2 herangezogen werden.

Die Berechnungsvorschrift für das korrigierte R^2 lässt sich aus der für R^2 herleiten:

$$R^2 = \frac{ESS}{TSS}$$

1 Bei R^2 handelt es sich auch um den quadrierten Bravais-Pearson-Korrelationskoeffizienten, der bereits aus der Statistik I bekannt sein sollte. Auch er gibt die Stärke eines linearen Zusammenhangs wieder.

Da sich TSS aus RSS und ESS zusammensetzt, lässt sich R^2 auch berechnen als:

$$R^2 = \frac{(TSS - RSS)}{TSS} = \frac{TSS}{TSS} - \frac{RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Das korrigierte R^2 berechnet sich dann als:

$$R^2_{korr} = 1 - \frac{(RSS/(n-k))}{(TSS/(n-1))}$$

Wird nun eine zusätzliche erklärende Variable hinzugefügt, ergeben sich zwei gegenläufige Effekte: RSS verringert sich oder bleibt gleich, wodurch sich das korrigierte R^2 entweder erhöht oder gleich bleibt. Der Wert für k erhöht sich um Eins, wodurch sich das korrigierte R^2 verringert. Welcher der beiden Effekte überwiegt, entscheidet darüber ob das korrigierte R^2 durch die Hinzunahme der erklärenden Variable ansteigt oder absinkt – je nachdem, sollte diese Variable dann ins Modell übernommen werden oder nicht.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken					Durbin-Watson-Statistik
					Änderung in R-Quadrat	Änderung in F	df1	df2	Änderung in Signifikanz von F	
1	,900[a]	,810	,799	170,29361	,810	72,503	2	34	,000	1,898

a. Einflußvariablen : (Konstante), ausgaben, besuche

b. Abhängige Variable: menge

Abbildung 35: Modellzusammenfassung der Regressionsanalyse in SPSS

Sowohl die Werte für R^2 als auch für das korrigierte R^2 werden von SPSS in der Modelzusammenfassungs-Tabelle ausgegeben, gemeinsam mit dem Bravais-Pearson-Korrelationskoeffizienten und dem Durbin-Watson-Koeffizienten, der weiter unten betrachtet werden wird. In unserem Beispielfall ergibt sich ein recht hoher Wert von 0,810 für R^2 – insgesamt werden also 81% der Gesamtstreuung durch das Regressionsmodell aufgeklärt. Da der Wert für das korrigierte R^2 nicht weit vom R^2-Wert abweicht, kann außerdem geschlussfolgert werden, dass alle ins Modell aufgenommenen unabhängigen Variablen ihre Daseinsberechtigung haben.

3 Standardfehler der Schätzung

Die Residuen, die Abweichungen der beobachteten Werte von den durch die Regressionsfunktion prognostizierten Werte, können sowohl positiv als auch negativ ausfallen, liegen im Durchschnitt jedoch stets bei Null[1]. Es stellt sich jedoch die Frage, ob die prognostizierten Werte in der Nähe der wahren Werte liegen oder stark von diesen abweichen – in beiden Fällen würde sich im Durchschnitt eine Abweichung von Null ergeben[2]. Theoretisch zumindest vorstellbar sind gewaltige Abweichungen in beide Richtungen, die sich lediglich im Durchschnitt neutralisieren. Inwiefern dies der Fall ist, kann mit dem R^2 und dem korrigierten R^2 festgestellt werden. Alternativ ist auch die Berechnung des Standardfehlers der Schätzung möglich, die ebenfalls auf den Residuen basiert.

Um zu vermeiden, dass sich die positiven und die negativen Abweichungen gegenseitig aufheben, werden sie, analog zur Methode der kleinsten Quadrate zunächst quadriert. Die Summe der quadrierten Residuen wird anschließend durch die Anzahl der Beobachtungswerte geteilt. Dadurch wird die sich ergebende Kennzahl unabhängig von der Stichprobengröße:

$$\frac{\sum e_i^2}{n}$$

Da der Mittelwert der Residuen gleich Null ist, gilt ebenfalls:

$$\frac{\sum e_i^2}{n} = \frac{\sum (e_i^2 - \bar{e})}{n} \quad \text{(Varianz der Residuen)}$$

Aus methodischen Gründen, die an dieser Stelle nicht weiter erläutert werden sollen, wird in SPSS nicht durch n sondern durch n abzüglich der erklärenden Variablen dividiert. Es ergibt sich der folgende Term:

1 Wäre dies nicht der Fall, so wäre die Quadrierung der Residuen im Rahmen der Methode der kleinsten Quadrate überflüssig.

2 Wie im Abschnitt zum Thema Lagemaße bereits festgestellt wurde, beträgt die Gesamtsumme aller Abweichungen vom arithmetischen Mittel ohnehin stets Null.

$\frac{\sum e_i^2}{(n-k)}$ (Beachte: Auch die Konstante gehört zu den erklärenden Variablen)

Die Quadratwurzel dieses Terms ergibt die Standardabweichung der Residuen, die auch als Standardfehler der Schätzung bezeichnet wird:

$$\sqrt{(\frac{\sum e_i^2}{(n-k)})}$$

Der Standardfehler der Schätzung ist ein Maß für die Anpassungsgüte der Regressionsgleichung. Er ist vergleichbar mit R^2 und korrigiertem R^2 und ist auch inhaltlich ähnlich zu interpretieren.

4 F-Statistik

Wie bereits gezeigt wurde, geben R^2 und korrigiertes R^2 Auskunft über die Anpassung der Regressionsgeraden an die beobachteten Werte. Es stellt sich aber daher auch die Frage, ob das Regressionsmodell auch über die Stichprobenwerte hinaus Gültigkeit besitzt. Ein geeignetes Prüfkriterium hierfür bildet die F-Statistik, in die neben der Streuungszerlegung auch der Umfang der Stichprobe eingeht. Die Prüfung der Regressionsfunktion mit der F-Statistik basiert auf gänzlich anderen Überlegungen als die Prüfung mittels R^2, korrigiertem R^2 und Standardfehler der Schätzung, auch wenn sie inhaltlich das gleiche Ziel verfolgt.

Die Regressionsfunktion der Stichprobe lässt sich darstellen als:

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \ldots + b_j * X_j + \ldots + b_J * X_J$$

Sie ist die Realisation der „wahren“ Regressionsfunktion:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_j * X_j + \ldots + \beta_J * X_J + u$$

Die neue Variable u wird auch als Störgröße bezeichnet. Sie repräsentiert alle zufälligen Einflüsse außerhalb der betrachteten Variablen. Die Störgröße selbst kann nicht beobachtet werden, sie zeigt sich aber in den Residuen.

Durch den Einfluss von u wird Y zu einer Zufallsvariablen, ebenso wie die Schätzwerte der Regressionsparameter. Würde man eine neue Stichprobe ziehen, würden sich jeweils andere Regressionsparameter ergeben. Bei wiederholten Stichproben schwanken diese Parameter um die „wahren" Regressionsparameter in der Grundgesamtheit, also die Regressionskoeffizienten, die sich bei einer Vollerhebung zeigen würden.

Nun ist die Grundannahme der Regressionsanalyse ja, dass es einen kausalen Zusammenhang zwischen der abhängigen und den unabhängigen Variablen gibt. Besteht ein solcher Zusammenhang tatsächlich, können diese „wahren" Regressionsparameter unmöglich Null sein.

Zur Überprüfung dieser Annahme wird das Regressionsmodell mit Hilfe des F-Tests varianzanalytisch untersucht. Die Nullhypothese H0 dieses Tests lautet:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_J = 0$$

Diese Nullhypothese besagt im Klartext, dass kein wirklicher Zusammenhang zwischen der abhängigen und den unabhängigen Variablen besteht – alle „wahren" Regressionskoeffizienten in der Grundgesamtheit sind daher gleich Null. Lässt sich diese Nullhypothese nicht mit einer entsprechend geringen Irrtumswahrscheinlichkeit verwerfen, so ist das Regressionsmodell offensichtlich nutzlos.

Um einen solchen F-Test durchzuführen genügt es, einen empirischen Wert aus der bekannten F-Verteilung zu berechnen und diesen mit einem (tabellierten) kritischen Wert zu vergleichen.

$$F_{(m,n)} = \frac{\frac{\chi^2_m}{m}}{\frac{\chi^2_n}{n}}$$ (F-verteilte Größe mit m und n Freiheitsgraden)

Bei Gültigkeit der H0 ist ein F-Wert von Eins zu erwarten. Je stärker nun der F-Wert von Eins abweicht, desto größer ist die Wahrscheinlichkeit, dass H0 unzutreffend ist. Bei entsprechend deutlichen Abweichungen kann H0 verworfen und die Schlußfolgerung gezogen werden, dass in der Grundgesamtheit mindestens ein „wahrer" Regressionskoeffizient ungleich Null existiert.

Vorsicht: Es kann nichts darüber ausgesagt werden, welche Regressionskoeffizienten ungleich Null sind, also welche der unabhängigen Variablen tatsächlich in das Modell gehören. Die einzige Aussage, die sich aus dem F-Test ergeben kann, ist dass es mit großer Wahrscheinlichkeit mindestens eine modellrelevante unabhängige Variable geben muss. Die Signifikanz der einzelnen Variablen ist im Anschluss an die Prüfung der Regressionsfunktion noch mittels der Prüfung der Regressionskoeffizienten durchzuführen, vorausgesetzt der F-Test wurde bereits im Vorfeld signifikant.

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4205166	2	2102583,0	72,503	,000[a]
	Residuen	985997,0	34	28999,913		
	Gesamt	5191163	36			

a. Einflußvariablen : (Konstante), ausgaben, besuche

b. Abhängige Variable: menge

Abbildung 36: Ergebnisse des F-Tests für den Beispielfall der Regressionsanalyse

Für den Beispielfall ergibt sich übrigens eine so geringe Irrtumswahrscheinlichkeit für die Ablehnung von H0, dass SPSS diese mit näherungsweise Null angibt. Die Nullhypothese kann hier also getrost verworfen werden. Schlussfolgerung: Entweder von den Werbeausga-

ben oder der Anzahl der Vertreterbesuche geht ein „wahrer“ auch in der Grundgesamtheit vorhandener Einfluss auf die Verkaufsmenge aus – es ist nur noch nicht klar, ob von beiden unabhängigen Variablen oder, wenn nein, von welcher der beiden.

VI Prüfung der Regressionskoeffizienten

1 Einführung

Konnte im Verlauf der Prüfung der Regressionsfunktion festgestellt werden, dass das Regressionsmodell als solches von Relevanz ist, so stellt sich nun noch die Frage nach der Relevanz der einzelnen Regressionskoeffizienten. Gehören also alle im Regressionsmodell untergebrachten unabhängigen Variablen auch wirklich in dieses Modell?

Zur Feststellung der Güte der Regressionskoeffizienten eignen sich zwei Kriterien:

- ein T-Test an jedem der Regressionskoeffizienten
- die Konfidenzintervalle um die Regressionskoeffizienten

2 T-Test der Regressionskoeffizienten

Wird die Nullhypothese des F-Test verworfen, so ist, wie im vorangegangenen Abschnitt dargestellt, mit großer Wahrscheinlichkeit davon auszugehen, dass mindestens einer der „wahren“ Regressionskoeffizienten in der Grundgesamtheit signifikant wird. Damit steht allerdings keineswegs fest, dass alle „wahren“ Regressionskoeffizienten der unabhängigen Variablen signifikant sind, lediglich, dass es mindestens einer ist. Es ist daher geboten, einen identischen Test für jeden einzelnen Regressionskoeffizienten im Modell durchzuführen. Ein geeignetes Prüfkriterium für einen solchen Test findet sich in der t-Statistik. Die t-Statistik folgt der aus der Statistik II bereits bekannten t-Verteilung (oder auch Student-Verteilung) um den Mittelwert Null.

Der empirische t-Wert einer unabhängigen Variable wird berechnet, indem deren Regres-

sionskoeffizient durch deren Standardfehler dividiert wird:

$$t_{emp} = \frac{(b_j - \beta_j)}{s_{bj}}$$

Die Nullhypothese H0 lautet: Der „wahre“ Regressionskoeffizient in der Grundgesamtheit beträgt Null. Bei Gültigkeit dieser Nullhypothese ist auch ein t-Wert von Null oder zumindest annähernd Null zu erwarten[1]. Weicht nun der empirische t-Wert stark von Null ab, so ist es unwahrscheinlich, dass H0 korrekt ist. In diesem Fall kann die H0 verworfen werden, woraus wiederum die Schlussfolgerung zu ziehen ist, dass der „wahre“ Regressionskoeffizient ungleich Null sein muss, dass also in der Grundgesamtheit ein „echter“ Zusammenhang zwischen der abhängigen und der jeweiligen getesteten unabhängigen Variablen besteht.

SPSS gibt die empirischen t-Werte in der Koeffizienten-Tabelle gemeinsam mit der bereits aus anderen statistischen Tests in SPSS bekannten Irrtumswahrscheinlichkeit bei Ablehnung der H0 aus. Je kleiner also dieser Wert ausfällt, desto eher kann H0 ohne das Risiko eines großen Fehlers verworfen werden.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B		Kollinearitätsstatistik	
		B	Standardfehler	Beta			Untergrenze	Obergrenze	Toleranz	VIF
1	(Konstante)	311,219	170,379		1,827	,077	-35,032	657,470		
	besuche	9,513	1,816	,396	5,238	,000	5,822	13,205	,978	1,023
	ausgaben	,550	,055	,752	9,945	,000	,438	,662	,978	1,023

a. Abhängige Variable: menge

Abbildung 37: Ausgabe der Koeffizienten-Tabelle in SPSS

3 Konfidenzintervalle um die Koeffizienten

Mit den aus der Statistik II hinlänglich bekannten Konfidenzintervallen (Vertrauensbereichen) lässt sich die Lage eines Parameters in der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit umreißen.

1 Es gilt immerhin zu beachten, dass durch die Art der (Zufalls-)Stichprobe noch weitere Effekte auftreten können, die das Zustandekommen von genau Null selbst im Falle eines „wahren“ Regressionskoeffizienten von Null unwahrscheinlich machen.

SPSS gibt in der Koeffizienten-Tabelle automatisch die 95%-Konfidenzintervalle um die Regressionskoeffizienten und die Konstante mit aus, wobei letzteres nicht weiter beachtet werden muss. Die 95%-Konfidenzintervalle um die Regressionskoeffizienten zeigen an, zwischen welchen beiden Grenzwerten der „wahre“ Regressionskoeffizient in der Grundgesamtheit mit 95%iger Sicherheit liegt. Zu beachten sind jeweils der Abstand zwischen Ober- und Untergrenze sowie die Vorzeichen beider Werte.

Ist das Konfidenzintervall um einen dieser Regressionskoeffizienten besonders breit, muss die auf seiner Basis geschätzte Regressionsfunktion als unsicher betrachtet werden. Zu beachten ist, dass die Koeffizienten in teils unterschiedlichen Dimensionen gemessen werden – bei Geldbeträgen von mehreren tausend Euro ist ein breiteres Konfidenzintervall zu erwarten als bei einem Prozentwert.

Diese sehr subjektive Betrachtung der Intervallbreite gibt bedauerlicherweise oft wenig Auskunft über die tatsächliche Relevanz der Regressionskoeffizienten. Viel aufschlussreicher ist das Auftreten bzw. das Nicht-Auftreten eines Vorzeichenwechsels innerhalb des Konfidenzintervalls. Worauf kann ein solcher Vorzeichenwechsel hindeuten?

Liegt der “wahre“ Regressionskoeffizient einer Beispieluntersuchung mit 95%iger Sicherheit zwischen +5 und +10, so mag er in Wirklichkeit 6 oder 7 betragen – auf die Art des Einflusses der unabhängigen Variablen hat dies keine Auswirkung, sondern lediglich auf die Stärke. Im ersten Fall erhöht sich der Wert der abhängigen Variablen um 6, wenn sich der Wert der unabhängigen Variablen um 1 erhöht, im zweiten Fall erhöht sich der Wert der abhängigen Variablen unter gleichen Umständen um 7. In beiden Fällen ist der Einfluss positiv – der Wert der abhängigen Variablen steigt an, wenn auch der Wert der unabhängigen Variablen ansteigt.

Liegt aber der „wahre“ Regressionskoeffizient derselben Beispieluntersuchung mit 95%iger Sicherheit zwischen -5 und +5, so mag er in Wirklichkeit -1 oder +1 betragen – dies aber hat ganz erhebliche Auswirkung auf das Regressionsmodell und die Interpretation des Regressionskoeffizienten, denn einmal steigt der Wert der abhängigen Variablen an und ein-

mal sinkt er ab. Ein Vorzeichenwechsel im Konfidenzintervall kann darauf hindeuten, dass ein berechneter Regressionskoeffizient mit einer nicht zu vernachlässigenden Wahrscheinlichkeit ein anderes Vorzeichen aufweisen könnte als der „wahre“ Regressionskoeffizient in der Grundgesamtheit. In einem solchen Fall würde sich der Einfluss der Variablen gewissermaßen „umkehren“ und das gesamte Regressionsmodell würde einen vollkommen falschen Eindruck von den Zusammenhängen vermitteln.

Aus diesem Grund sind Regressionskoeffizienten, deren Konfidenzintervalle einen solchen Vorzeichenwechsel aufweisen, mit Vorsicht zu interpretieren. Idealerweise ist ein Fachexperte zu konsultieren, der zu einer sicheren Aussage darüber gelangen kann, ob der Einfluss der entsprechenden unabhängigen Variablen auf die abhängige Variable ein positiver oder ein negativer sein muss. Nicht zulässig dagegen ist die Vergrößerung der Irrtumswahrscheinlichkeit des Konfidenzintervall. Durch eine solche Vergrößerung (beispielsweise von 95% auf 90%) schrumpft das Intervall und verliert gegebenenfalls den Vorzeichenwechsel – ein solches Vorgehen ist aber als manipulativ und unter dem Gesichtspunkt der methodischen Sorgfalt unangemessen zu betrachten.

VII Prüfung der Modellvoraussetzungen

1 Einführung

Im vierten und letzten Schritt der Regressionsanalyse sind noch die bereits Eingangs erwähnten weiteren Modellvoraussetzungen zu überprüfen. Können sie alle als gegeben betrachtet werden, so ist das berechnete Modell als gültig zu betrachten und kann aus fachlicher Sicht interpretiert werden. Dabei ist zu unterteilen in Tests an den Residualgrößen und Tests an der abhängigen und den unabhängigen Variablen.

Tests an den Residualgrößen:

- Test auf Normalverteilung der Residualgrößen
- Test auf Varianzgleichheit der Residualgrößen

- Test auf Autokorrelation der Residualgrößen

Tests an den abhängigen und unabhängigen Variablen:

- Test auf linearen Zusammenhang
- Test auf Multikollinearität

2 Test auf Normalverteilung der Residualgrößen

Wie Tests auf Normalverteilung funktionieren, wurde bereits im Zusammenhang mit der explorativen Datenanalyse umfassend dargestellt, und muss daher an dieser Stelle nicht weiter betrachtet werden. Erwähnt seien nur noch einmal die verschiedenen Methoden, die dabei zur Anwendung kommen können:

- Grafische Prüfung anhand eines Histogramms mit Normalverteilungskurve
- Grafische Prüfung anhand eines Q-Q- oder P-P-Diagramms
- Statistische Prüfung mit dem Kolmogorov-Smirnoff-Anpassungstest

3 Test auf Varianzgleichheit der Residualgrößen

Auch die Homoskedastizitätsprüfung wurde im Abschnitt zur explorativen Datenanalyse bereits ausführlich erläutert. Es sei daher an dieser Stelle nur kurz auf die verschiedenen Methoden hingewiesen, mit denen eine Varianzgleichheit festgestellt werden kann:

- Grafische Prüfung mit gruppierten Box-Plots oder Streudiagrammen
- Statistische Prüfung mit dem Levene-Test

4 Test auf Autokorrelation der Residualgrößen

Was ist unter dem Begriff der Autokorrelation zu verstehen? Als Autokorrelation bezeichnet man eine Systematik in den Residuen, insbesondere Zusammenhänge zwischen nebenein-

anderliegenden Residualgrößen. Beispiel: Auf große positive Residuen folgen regelmäßig große negative Residuen – eine deutliche Systematik ist erkennbar.

Autokorrelation kann generell immer dort auftreten, wo die einzelnen Fälle nicht zufälllig angeordnet sind. Dies ist beispielsweise bei Zeitreihenanalysen der Fall, wo die Fälle zeitlich geordnet vorliegen. Das Auftreten von Autokorrelationen kann auf zwei mögliche Probleme mit dem Regressionsmodell hindeuten: Entweder eine erklärungsrelevante Variable wurde nicht in das Modell aufgenommen oder aber es liegt ein nicht-linearer Zusammenhang, beispielsweise ein quadratischer Zusammenhang vor.

Abgesehen von diesen Problemen birgt die Autokorrelation noch eine weitere Gefahr: Sie führt dazu, dass die Standardfehler der Regressionskoeffizienten zu gering eingeschätzt werden. Die Ergebnisse der Signifikanztests können damit nicht mehr als zuverlässig betrachtet werden, als Folge davon werden die Regressionskoeffizienten als signifikanter bewertet, als sie es tatsächlich sind – ihr Einfluss auf die abhängige Variable wird also überschätzt.

Wie man sieht, wirft das Auftreten von Autokorrelation aus verschiedenen Gründen Fragen bezüglich der Interpretationsfähigkeit einer Regressionsfunktion auf – bei starker Autokorrelation ist sogar das ganze Modell als ungültig zu betrachten. Aus diesem Grund muss vor der Modellinterpretation nach Autokorrelationen in den Residualgrößen gesucht werden – dies geschieht in SPSS mittels des Durbin-Watson-Tests auf Autokorrelation.

Der Ergebniswert des Durbin-Watson-Tests, der Durbin-Watson-Koeffizient, kann Werte zwischen 0 und 4 annehmen. Je näher der Wert des Koeffizienten dabei an 2 liegt, desto geringer ist das Ausmaß der Autokorrelation. Dagegen deuten Werte deutlich unter 2 auf eine positive Autokorrelation, Werte deutlich über 2 auf eine negative Autokorrelation hin. Als generelle Faustregel für die Interpretation des Koeffizienten kann folgendes festgehalten werden: Durbin-Watson-Werte zwischen 1,5 und 2,5 sind akzeptabel, Werte unter 1 oder über 3 deuten definitiv auf Autokorrelation hin. Dazwischen existiert eine interpretatorische Grauzone, in der die Entscheidung über die Fortsetzung der Regressionsanalyse beim Marktforscher selbst liegt.

Der Durbin-Watson-Test kann nur unter zwei Voraussetzungen durchgeführt werden: Die Regressionsfunktion muss einen konstanten Term enthalten (diesen baut SPSS aber automatisch ein) und die abhängige Variable im Regressionsmodell darf nicht zeitverzögert auch als unabhängige Variable verwendet werden. Dies kann zum Beispiel bei Zeitreihenanalysen der Fall sein, wenn der Wert einer Variablen aus der vorangegangenen Periode als erklärender Wert für eben diese Variable in der aktuellen Periode herangezogen wird (Beispiel: Schätzung der Ozonwerte für August aus den Ozonwerten für Juli).

Als Einschränkung ist noch festzuhalten, dass mit dem Durbin-Watson-Test lediglich die sogenannten Autokorrelationen der 1. Ordnung identifiziert werden können. Eine solche Autokorrelation liegt dann vor, wenn direkt benachbarte Residuen miteinander verknüpft sind. In einigen Fällen, beispielsweise bei quartalsweise erhobenen Daten, ist jedoch auch eine Autokorrelation der 4. Ordnung denkbar. Besteht eine solche Möglichkeit, sollte neben dem Durbin-Watson-Test auch noch der Wallis-Test auf Autokorrelationen 4. Ordnung durchgeführt werden.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken					Durbin-Watson-Statistik
					Änderung in R-Quadrat	Änderung in F	df1	df2	Änderung in Signifikanz von F	
1	,900[a]	,810	,799	170,29361	,810	72,503	2	34	,000	1,898

a. Einflußvariablen : (Konstante), ausgaben, besuche

b. Abhängige Variable: menge

Abbildung 38: Modellzusammenfassung der Regressionsanalyse in SPSS

Der Durbin-Watson-Koeffizient wird von SPSS in der Modellzusammenfassungs-Tabelle gemeinsam mit den Werten für R^2 und korrigiertes R^2 ausgegeben. In unserem Beispielfall liegt er dicht am Idealwert von 2, da außerdem keine Voraussetzungen verletzt wurden, ist aus dem Ergebnis zu schlussfolgern, dass keine Autokorrelation der Residualgrößen vorliegt.

5 Test auf linearen Zusammenhang

Das Vorliegen eines linearen Zusammenhangs zwischen der abhängigen und den einzelnen unabhängigen Variablen ist eine essentielle Voraussetzung für die Durchführung einer Regressionsanalyse. Viele Autoren ordnen daher den Test auf einen solchen linearen Zusammenhang auch in der ersten Phase der Regressionsanalyse ein, wo ja anhand von Streudiagrammen und Matrixdiagrammen bereits nach linearen Zusammenhängen gesucht wird. Alternativ ließe sich auch der aus der Statistik I bekannte Bravais-Pearson-Korrelationskoeffizient berechnen, der die Stärke eines linearen Zusammenhangs (und nur eines linearen) wiedergibt.

6 Test auf Multikollinearität

Der letzte Schritt der Regressionsanalyse vor der fachlichen Interpretation des Modells[1] ist der Test auf Multikollinearität der unabhängigen Variablen, der auch als Kollinearitätsdiagnostik bekannt ist.

Was ist unter Multikollinearität zu verstehen? Multikollinearität liegt dann vor, wenn zwei oder mehr der unabhängigen Variablen in einem Regressionsmodell nicht nur mit der abhängigen Variablen, sondern auch untereinander korrelieren. Würden in unserem Beispielfall die Kosten für Kundenbesuche zu den Werbeausgaben gerechnet werden, dann würden diese beiden unabhängigen Variablen untereinander korrelieren – denn je mehr Kundenbesuche desto höher die Werbeausgaben. Tritt so eine Situation auf, lässt sich nicht mehr feststellen, zu welchen Teilen eine Veränderung der abhängigen Variablen auf die eine oder die andere der beiden korrelierenden unabhängigen Variablen zurückführen ist.

Die Ausnahme bildet eine perfekte Multikollinearität, also einen perfekten Zusammenhang zwischen zwei unabhängigen Variablen. Liegt eine solche perfekte Multikollinearität vor, kann die Regressionsanalyse mathematisch gar nicht erst durchgeführt werden, solange sich beide Variablen noch im Modell befinden. SPSS schließt in solchen Fällen automatisch eine der beiden Variablen aus und weist in der Ausgabe auf die entdeckte Multikollinearität

1 Diese ist in der Regel nicht mehr durch den Marktforscher allein zu leisten, daher wird auf diesen Arbeitsschritt im Rahmen dieses Manuskripts nicht detaillierter eingegangen.

hin. Perfekte Multikollinearitäten sind daher relativ ungefährlich – imperfekte Multikollinearitäten können dagegen zum Problem werden.

Bei Vorliegen einer imperfekten Multikollinearität lässt sich die Regressionsanalyse mathematisch wie gehabt durchführen. Es ergibt sich zwar ein unverzerrtes R^2, die Berechnung der Regressionskoeffizienten und damit auch der Beta-Koeffizienten liefert jedoch unzuverlässige Ergebnisse. Zu befürchten ist, dass der Koeffizient und auch der Einfluss bei einer der beiden Variablen über- und bei der anderen unterschätzt wird. Der gemeinsame Einfluss beider Variablen auf die abhängigen Variable wird so noch korrekt ausgewiesen, bezüglich der Verteilung dieses Einflusses gelangt der Marktforscher aber zu falschen Schlussfolgerungen.

Unter SPSS bieten sich drei Möglichkeiten an, um unabhängige Variablen auf Multikollinearität zu überprüfen:

- Erstellung einer Korrelationmatrix
- Berechnung von Toleranz und Varianzinflationsfaktor
- Berechnung der Varianzanteile

Erstellung einer Korrelationsmatrix

In einer Korrelationsmatrix wird der Bravais-Pearson-Korrelationskoeffizient für jede mögliche Kombination aus abhängiger und unabhängiger sowie unabhängiger Variablen untereinander ausgegeben. Zeigt sich hier ein hoher Korrelationskoeffizient zwischen zwei unabhängigen Variablen, liegt eine Multikollinearität vor und eine der Variablen sollte dann konsequenterweise noch aus dem Modell ausgeschlossen werden.

Denkbar ist aber auch das Auftreten von paarweisen Korrelationen zwischen Variablenkombinationen (multiple Korrelation) anstatt der einfachen linearen Korrelation zwischen zwei Einzelvariablen. Da solche Formen der Multikollinearität nicht in der Korrelationsmatrix erkannt werden können, sind weitere Tests auf Multikollinearität erforderlich.

Korrelation der Koeffizienten[a]

Modell			ausgaben	besuche
1	Korrelationen	ausgaben	1,000	-,148
		besuche	-,148	1,000
	Kovarianzen	ausgaben	,003	-,015
		besuche	-,015	3,299

a. Abhängige Variable: menge

Abbildung 39: Korrelationsmatrix in SPSS

Im vorliegenden Beispielfall liegt keiner der „kritischen" Korrelationskoeffizienten nahe Eins, es liegt also kein Anhaltspunkt für Multikollinearität vor.

Berechnung von Toleranz und Varianzinflationsfaktor

Die Toleranz ist definiert als die Differenz von Eins und dem multiplen Korrelationskoeffizienten. Fällt sie sehr klein aus, ist dies als Hinweis auf eine Multikollinearität in den Daten zu werten. Als Faustregel für die Interpretation kann gelten: Toleranzen unterhalb von 0,1 legen den Verdacht auf Multikollinearität nahe, Toleranzen unterhalb von 0,01 können als sicherer Beweis für eine Multikollinearität gewertet werden.

SPSS berechnet zusätzlich zur Toleranz auch noch den sogenannten Varianzinflationsfaktor (VIF), der aber lediglich den Kehrwert der Toleranz wiedergibt. Entsprechend der Faustregel für die Interpretation der Toleranz kann bezüglich des Varianzinflationsfaktors festgestellt werden: VIF-Werte oberhalb von 10 legen den Verdacht auf Multikollinearität nahe, VIF-Werte oberhalb von 100 können als sicherer Beweis für eine Multikollinearität gewertet werden.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B		Kollinearitätsstatistik	
		B	Standardfehler	Beta			Untergrenze	Obergrenze	Toleranz	VIF
1	(Konstante)	311,219	170,379		1,827	,077	-35,032	657,470		
	besuche	9,513	1,816	,396	5,238	,000	5,822	13,205	,978	1,023
	ausgaben	,550	,055	,752	9,945	,000	,438	,662	,978	1,023

a. Abhängige Variable: menge

Abbildung 40: Ausgabe der Koeffizienten-Tabelle in SPSS

SPSS gibt sowohl die Toleranz als auch den Varianzinflationsfaktor in der Koeffizienten-Tabelle gemeinsam mit den Regressionskoeffizienten, den Beta-Koeffizienten und den Konfidenzintervallen um die Regressionskoeffizienten aus. Im hier verwendeten Beispielfall sprechen die Ergebnisse nicht gegen die Verwendbarkeit des Regressionsmodells. Sowohl die Toleranz als auch der Varianzinflationsfaktor sind weit von den in den Faustregeln genannten „kritischen Werten“ entfernt. Eine weitere Prüfung, wie beispielsweise die Berechnung der Varianzanteile, ist nach diesem Ergebnis statistisch gesehen überflüssig.

Berechnung der Varianzanateile

Als (etwas umständliche und daher auch weniger verbreitete) Alternative zur Berechnung der Toleranz bzw. des Varianzinflationsfaktors lassen sich noch die Varianzen der jeweiligen Regressionskoeffizienten in Komponenten zerlegen und den Eigenwerten zuordnen. Die Summe aller Komponenten beträgt für jeden Regressionskoeffizienten wieder gleich Eins – es handelt sich also um eine Anteilszerlegung.

Wenn nun derselbe Eigenwert die Varianz mehrerer Regressionskoeffizienten in hohem Maße erklärt, so deutet dies auf einen Zusammenhang der beiden Variablen hin, also auf eine Multikollinearität. Zur Erleichterung der Interpretation berechnet SPSS auch die sogenannten Konditionsindizes der Eigenwerte, wobei Konditionsindizes zwischen 10 und 30 auf eine mittlere und Konditionsindizes über 30 auf eine starke Multikollinearität hindeuten.

VIII Weiterführende Literatur

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). Multivariate Analysemethoden (10. Aufl.). Berlin: Springer.

Bleymüller, J.; Gehlert, G. & Gülicher, H. (2000). Statistik für Wirtschaftswissenschaftler. München: Verlag Vahlen

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (1999). Statistik. Der Weg zur Datenanalyse (2. Aufl.). Berlin: Springer.

Götze, W., Deutschmann, C. & Link, H. (2002). Statistik. München: Oldenbourg.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). Multivariate data analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

D Varianzanalyse

I Einführung

1 Hintergründe der Varianzanalyse

Zur Hinführung auf die Grundidee hinter der Varianzanalyse soll dieser Beispielversuch dienen: In fünf Schulklassen der gleichen Ausbildungsstufe werden parallel zueinander fünf verschiedene Unterrichtskonzepte eingesetzt, anschließend wird der Lernerfolg in einem gemeinsamen Test gemessen. Die entscheidende Frage lautet: haben sich die Unterrichtskonzepte signifikant auf den Lernerfolg ausgewirkt – lassen sich also Unterschiede zwischen den beiden Gruppen feststellen? Besonders interessant wären hier Unterschiede zwischen den Testergebnis-Mittelwerten der unterschiedlichen, durch die Unterrichtskonzepte gebildeten Schülergruppen.

Die Varianzanalyse untersucht also die Wirkung einer oder mehrerer unabhängiger Variablen, der sogenannten Faktoren, auf eine oder mehrere abhängige Variablen. Abhängige Variablen müssen dabei intervallskaliert sein, für die Faktoren ist das nominale Skalenniveau ausreichend. Sie testet für Fälle mit mehr als zwei Gruppen (ansonsten lässt sich vereinfachend auch ein T-Test durchführen, wie weiter unten noch gezeigt werden wird), inwiefern signifikante Mittelwertunterschiede vorliegen. Der Varianzanalyse liegen daher die folgenden Hypothesen zugrunde:

- Nullhypothese H0: Alle „wahren“ Mittelwerte der Grundgesamtheit sind gleich
- Alternativhypothese Ha: Mindestens zwei „wahre“ Mittelwerte unterscheiden sich

Die Varianzanalyse ist das bedeutendste Verfahren für die Auswertung von Experimenten, wobei das Wirkungsmodell (welche Variablen sind abhängig, welche unabhängig?) im Voraus bekannt sein muss – was bei Experimenten ja in der Regel auch der Fall ist.

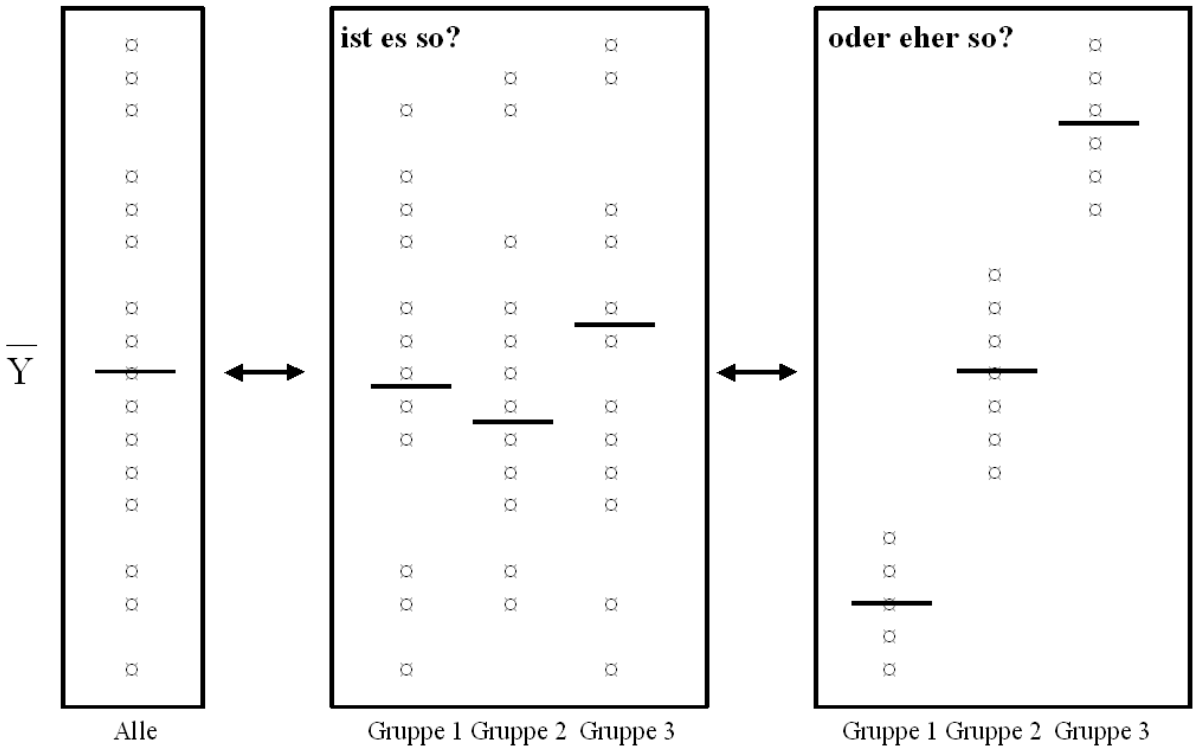

Abbildung 41: Die visuelle Grundidee der Varianzanalyse

2 Der T-Test als Alternative

Der T-Test bei zwei Gruppen

Um Mittelwerte (gemeint ist hier natürlich das Standardmittel, also das arithmetische Mittel) miteinander zu vergleichen, kann bekanntermaßen auch der T-Test (aus Statistik II) eingesetzt werden. Wieso also die wesentlich aufwendigere Varianzanalyse? Ein kurzer Rückblick auf den Ablauf des T-Tests beantwortet diese Frage.

Der T-Test verwendet als Prüfkriterium einen Wert aus der t- oder Student-Verteilung. Dieser Wert lässt sich aus den Stichprobenwerten berechnen und kann anschließend mit dem theoretischen t-Wert unter bestimmten Annahmen verglichen werden. Standardmäßig wird beim T-Test von der Annahme ausgegangen, dass die „wahren“ Mittelwerte einer Variablen in zwei Gruppen in der Grundgesamtheit identisch sind, dass also die unabhängige Variable, durch welche die Gruppen gebildet werden, keinen signifikanten Einfluss auf die betrachtete abhängige Variable ausübt.

Der t-Wert berechnet sich aus den Stichprobenwerten als

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{(\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}})}$$

Sind beide Gruppenmittelwerte identisch, ergibt sich ein t-Wert von Null. Hierbei ist zu beachten, dass der t-Wert aus den Stichprobenwerten und nicht aus den Daten einer Vollerhebung berechnet wird. Mittelwerte, Varianzen und andere statistische Größen aus einer Stichprobe stimmen aber höchstens zufällig exakt mit den „wahren" Werten in der Grundgesamtheit überein – in der Regel werden sie, in Abhängigkeit von der Sorgfalt der Messung oder der Genauigkeit des Erhebungsdesigns mehr oder weniger stark von diesen „wahren" Werten abweichen. Dies ist jedem Marktforscher klar, denn würde man in irgendeiner Untersuchung erneut eine Stichprobe ziehen, würden sich andere Stichprobenwerte und damit auch andere Rechenergebnisse einstellen. Bei der Ziehung mehrerer Stichproben ist zu erwarten, dass die erhobenen Varianzen und Mittelwerte also mehr oder weniger stark um die „wahren" Werte herum streuen.

Dieser Gedanke ist wiederum auch auf die t-Werte übertragbar, da sich bei unterschiedlichen Stichproben aus derselben Grundgesamtheit ebenfalls unterschiedliche t-Werte ergeben dürften. Die Schlussfolgerung, die daraus gezogen werden kann ist, dass eine Übereinstimmung der „wahren" Mittelwerte in der Grundgesamtheit durch die Berechnung des t-Wertes nicht unmittelbar erkannt werden muss, da eben mit den besagten Schwankungen in den t-Werten zu rechnen ist. Da die Verteilung von t bekannt ist, lässt sich aber errechnen, mit welcher Wahrscheinlichkeit der t-Wert um ein bestimmtes Ausmaß von dem Wert abweicht, den t bei einer perfekten Übereinstimmung von Stichprobe und Grundgesamtheit angenommen hätte.

Der T-Test ist daher durchaus ein probates und wissenschaftliches Mittel zum Vergleich zweier Mittelwerte. Wie sieht es aber aus, wenn mehr als zwei Gruppen existieren und daher auch mehr als zwei Mittelwerte miteinander verglichen werden müssen?

Der T-Test bei drei und mehr Gruppen

Liegt eine solche Situation vor, lassen sich theoretisch doch auch zwei oder mehr T-Tests hintereinander durchführen. Wieso also der Rückgriff auf die komplexere Varianzanalyse?

Wie bei jedem statistischen Test wird auch beim T-Test mit einer Irrtumswahrscheinlichkeit α gerechnet. Diese kann durch den Marktforscher frei festgelegt werden, liegt aber üblicherweise entweder bei 0,05 oder auch bei 0,01. Bei der Durchführung einer Reihe von T-Tests kommt es nun zu einer sogenannten α-Fehlerinflation, also einer Potenzierung der ursprünglichen Irrtumswahrscheinlichkeit. Wie ist dies zu erklären?

Angenommen, die Irrtumswahrscheinlichkeit für eine beliebige Reihe von T-Tests wird auf 0,05 festgelegt. Die Wahrscheinlichkeit dafür, dass ein Vergleich nun lediglich zufällig signifikant wird, also die Wahrscheinlichkeit für einen sogenannten α-Fehler, liegt somit bei 0,05 oder 5%. Bei mehreren Vergleichen erhöht sich diese Irrtumswahrscheinlichkeit dramatisch. Nach nur 28 durchgeführten Vergleichen ist sie bereits auf 76,2% gestiegen – die Wahrscheinlichkeit ist also recht gross, schon mindestens einen fehlerhaften Vergleich in der Reihe zu haben. Der Grund dafür wird deutlich, dass eine Irrtumswahrscheinlichkeit von 5% auch aussagt, dass die Nicht-Irrtumswahrscheinlichkeit für einen T-Test bei 95% liegt. Führt man zwei hintereinander durch, liegt die Wahrscheinlichkeit, dass beide fehlerfrei verlaufen sind schon nicht mehr bei 95% sondern bei $95\%^2 = 90{,}25\%$. T-Tests werden daher generell für Vergleiche von mehr als zwei Gruppen als ungeeignet eingestuft.

3 Der Ablauf der Varianzanalyse

Die Varianzanalyse lässt sich in drei wesentliche Arbeitsschritte unterteilen.

Im ersten Arbeitsschritt muss dass für den restlichen Analyseverlauf unterstellte Erklärungsmodell aus abhängigen und unabhängigen Variablen formuliert werden. Daneben existiert eine ganze Reihe methodischer Voraussetzungen (z.B. Skalenniveau, Homoskedastizität, Verteilungsform), die vor dem eigentlichen Beginn der Analyse überprüft werden müssen.

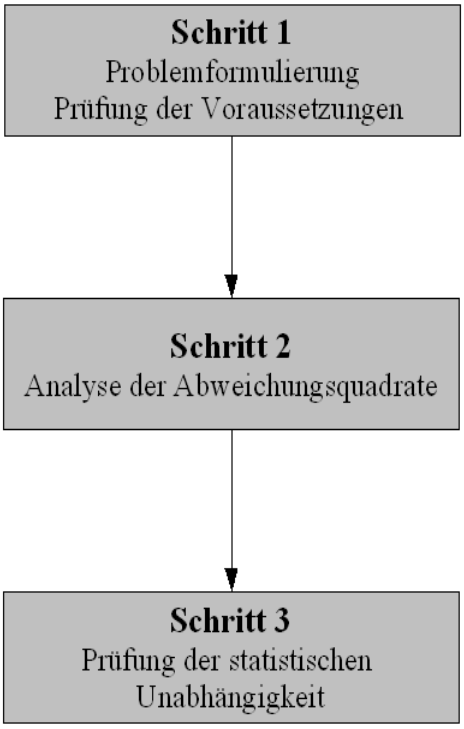

Abbildung 42: Der Ablauf der Varianzanalyse

Im zweiten Arbeitsschritt, dem Hauptschritt der Varianzanalyse, wird die im Modell auftretende Gesamtvarianz in die Varianz innerhalb der Gruppen und die Varianz zwischen den Gruppen zerlegt. Das Verhältnis dieser beiden Werte zueinander liefert Aufschluss bezüglich des Erklärungsgehalts der Faktoren.

Finden sich signifikante Unterschiede zwischen den Gruppenmittelwerten ist im dritten und finalen Arbeitsschritt noch zu überprüfen, ob sich diese Unterschiede auf Zufallseffekte während der Stichprobenziehung zurückführen lassen, oder ob es sich um „echte“ Unterschiede handelt, die auch in der Grundgesamtheit auftreten. Dies geschieht zum einen mit dem bereits aus der Regressionsanalyse bekannten F-Test und zum anderen anhand einer Auswahl aus einer ganzen Reihe möglicher sogenannter Post-Hoc-Tests.

4 Formen der Varianzanalyse

Es sind verschiedene Szenarien aus dem Umfeld der Marktforschung vorstellbar, bei denen eine Varianzanalyse von Nutzen wäre. So könnte man beispielsweise untersuchen, welche

Wirkung verschiedene Formen der Werbung (z.B. Anzeigen in Zeitschriften, Plakate, Radiowerbespots etc.) auf die Verkaufszahlen eines bestimmten Produktes haben. Die abhängige Variable ist in diesem Fall die Anzahl der verkauften Einheiten oder der Gesamtumsatz, die unabhängige Variable – der Faktor – die durchgeführten Werbemaßnahmen. Es liegt eine einfaktorielle Varianzanalyse vor.

Von Interesse könnte auch die Untersuchung der Wirkung zweier Faktoren auf den Verkauf sein, nämlich der Verpackung einer Ware und der Plazierung im Supermarktregal und zwar sowohl isoliert als auch gemeinsam. Da hier zwei Faktoren in die Varianzanalyse eingehen, handelt es sich um eine zweifaktorielle Varianzanalyse.

Man erkennt bereits, dass es keine „einheitliche“ Varianzanalyse gibt, sondern verschiedene Formen und zwar in Abhängigkeit von der Anzahl sowohl der abhängigen als auch der unabhängigen Variablen.

Zahl der AV	***Zahl der UV***	***Bezeichnung der Verfahren***	
1	1	Einfaktorielle VA	ANOVA
1	2	Zweifaktorielle VA	
1	3	Dreifaktorielle VA	
1	...		
>= 2	>= 1	Mehrdimensionale VA	MANOVA

Tabelle 2: Formen der Varianzanalyse

Wie die Tabelle zeigt, ist insbesondere in die ANOVA (= Analysis of Variance) – die Varianzanalyse mit nur einer abhängigen Variablen – und die MANOVA (= Multivariate Analysis of Variance) – die Varianzanalyse mit mehr als einer abhängigen Variablen – zu unterscheiden.

II Prüfung der Analysevoraussetzungen

1 Skalenniveaus der verwendeten Variablen

Alle abhängigen Variablen müssen auf jeden Fall metrisch skaliert sein. Der Grundgedanke der Varianzanalyse basiert auf dem Vergleich der Mittelwerte einer Variablen in zwei oder mehr Gruppen, wobei der Standardmittelwert, also das arithmetische Mittel, gemeint ist. Wie in Kap. 2 dargestellt, müssen zur Berechnung des arithmetischen Mittels metrisch skalierte Daten vorliegen, darum kann man auch in der Varianzanalyse nicht ohne sie auskommen.

Die unabhängigen Variablen werden dagegen nur für die Gruppeneinteilung benötigt – hier werden keine Mittelwerte gezogen oder verglichen. Sie können daher auch nominalskaliert sein und sind es in der Regel auch. Da metrisch skalierte Merkmale stetig sind und sich daher häufig in einer Vielzahl von Merkmalsausprägungen ausdrücken, eignen sie sich meist wenig zur Einteilung von Gruppen, da zuviele Gruppen mit zu wenigen Fällen entstehen würden. Liegen solche Merkmale vor und werden sie für die Gruppeneinteilung benötigt, so ist eine Klassierung des Merkmals sinnvoll, um eine künstliche Diskretisierung herbeizuführen.

2 Vermutete Wirkungszusammenhänge

Die Varianzanalyse gehört zu den strukturprüfenden Verfahren und nicht zu den strukturentdeckenden Verfahren. Es muss bereits zu Beginn der Analyse feststehen, welche Variablen abhängig und welche unabhängig sein sollen – das Wirkungsmodell, welches während der Varianzanalyse unterstellt wird, muss also von Anfang an vorliegen. Im Zweifelsfalle sollte es ein Fachexperte zur jeweiligen Thematik sein, der dieses Wirkungsmodell entwirft, der Statistiker als Nicht-Experte kann dann dieses Modell anschließend unter objektiven Bedingungen durch eine Varianzanalyse testen.

3 Verschiedenheit der Faktoren

Alle im Modell verwendeten Faktoren – also alle unabhängigen Variablen – müssen eigenständige Einflussgrößen der abhängigen Variablen darstellen. So kann es beispielsweise einen Faktor „Verpackungstyp“ geben, der verschiedene Papier- und Plastikverpackungen als Ausprägungen aufweist, aber keine Faktoren „Papierverpackung“ und „Plastikverpackung“ mit jeweils unterschiedlichen Ausprägungen, da für jeden Fall im Modellsystem eine Ausprägung je Faktor vorhanden sein muss (dies schließt das Auftreten von fehlenden Werten aber nicht aus). Eine Aufspaltung eines einzelnen Faktors wie „Verpackung“ in mehrere Faktoren, wie hier am Beispiel demonstriert, ist in der Varianzanalyse unzulässig.

4 Normalverteilung der Grundgesamtheit

Die verwendeten Daten müssen per Zufallsstichprobe oder methodisch sauber durchgeführtem Experiment aus einer normalverteilten Grundgesamtheit entnommen werden. Die Normalverteilung aller (!) Abstufungsgruppen ist vor dem Einstieg in eine Varianzanalyse zu prüfen – es muss also nicht nur die abhängige Variable „Umsatz“ insgesamt normalverteilt sein, diese Variable muss sich auch in den durch die Faktoren wie „Werbeausgaben“ oder „Verpackungstyp“ gebildeten Untergruppen normal verteilen.

Die Prüfung auf Normalverteilung kann anhand eines Kolmogorov-Smirnoff-Anpassungstests oder grafisch (Q-Q-Diagramm, P-P-Diagramm, Histogramm) erfolgen. Sämtliche Möglichkeiten werden im Kapitel zur explorativen Datenanalyse im Detail dargestellt, weswegen im Folgenden nicht weiter auf die Normalverteilungsprüfung eingegangen werden wird.

5 Varianzgleichheit in den Fallgruppen

Um die Mittelwerte einer Variable in verschiedenen Gruppen überhaupt miteinander vergleichen zu können, muss die Varianz dieser Variablen in allen Gruppen etwa gleich sein. Da genau dies in der Varianzanalyse geschieht, ist die Gleichverteilung der Varianz (Homoskeda-

stizität) der abhängigen Variablen in allen durch die Faktoren gebildeten Untergruppen eine wesentliche Voraussetzung dieses Analyseverfahrens, die in jedem Fall zu Beginn der Varianzanalyse zu überprüfen ist.

Die Prüfung auf Varianzgleichheit kann anhand eines Levene-Tests oder grafisch (Streudiagramm, Box-Plot) erfolgen. Sämtliche Möglichkeiten werden im Kapitel zur explorativen Datenanalyse im Detail dargestellt, weswegen im Folgenden nicht weiter auf die Prüfung auf Varianzgleichheit eingegangen werden wird.

III Analyse der Abweichungsquadrate

1 Beispielfall

Zur Verdeutlichung der Analyse der Abweichungsquadrate soll nachfolgend dieser Beispielfall dienen: Ein großer Filmverleih möchte erfahren, ob sich die Verwendung unterschiedlicher Plakatdesigns (romantisch, modern, schwarz-weiß...) für einen Film signifikant auf den Verkauf von Kinokarten auswirkt. Dazu werden nun an vier verschiedenen Kinos vier Plakatversionen für den selben Film ausgehängt und die Besucherzahlen des ersten Tages für die jeweils fünf Tagesvorstellungen der Kinos erfasst. Das Resultat sind vier Teilstichproben mit jeweils fünf Beobachtungswerten.

Aus allen Beobachtungswerten werden nun die Gruppenmittelwerte (jedes Kino stellt hier eine eigene Gruppe dar) und der Gesamtmittelwert errechnet. Gibt es keinerlei Unterschiede bezüglich der durchschnittlichen Besucherzahlen zwischen den einzelnen Gruppen, dann kann man zu Recht vermuten, dass die Plakatdesigns keinen oder nur einen sehr geringen Einfluss auf die Verkaufszahlen besitzen. Kann aber im Umkehrschluss schon aus dem Vorliegen von Mittelwertsunterschieden auf einen Effekt bezüglich der Plakate geschlossen werden? Können nicht auch zufällige Vorgänge zu solchen Unterschieden führen? Dies wollen wir mit Hilfe einer einfaktoriellen ANOVA erkunden, wobei die Erfüllung sämtlicher Voraussetzungen hier einmal unterstellt wird.

		Vst. 1	*Vst. 2*	*Vst. 3*	*Vst. 4*	*Vst. 5*	*Mittel*
Kino 1	Plakat 1	23	28	31	24	19	25
Kino 2	Plakat 2	44	51	41	46	39	44,2
Kino 3	Plakat 3	22	18	15	23	41	23,8
Kino 4	Plakat 4	35	41	39	27	34	35,2
Gesamt							32,05

Tabelle 3: Beispielfall: Besucherzahlen aus fünf Kinos und fünf Vorstellungen

ONEWAY deskriptive Statistiken

Besucherzahl

	N	Mittelwert	Standardabweichung	Standardfehler	95%-Konfidenzintervall für den Mittelwert		Minimum	Maximum
					Untergrenze	Obergrenze		
1,00	5	25,0000	4,63681	2,07364	19,2426	30,7574	19,00	31,00
2,00	5	44,2000	4,65833	2,08327	38,4159	49,9841	39,00	51,00
3,00	5	23,8000	10,13410	4,53211	11,2169	36,3831	15,00	41,00
4,00	5	35,2000	5,40370	2,41661	28,4904	41,9096	27,00	41,00
Gesamt	20	32,0500	10,45529	2,33787	27,1568	36,9432	15,00	51,00

Abbildung 43: Der Beispielfall für die Varianzanalyse in SPSS

2 Streuungszerlegung

Nach der Einteilung in abhängige und unabhängige Variablen – in diesem Fall gehen wir davon aus, dass die Besucherzahlen vom Plakatdesign abhängig gemacht werden können – erfolgt die Gruppenbildung nach Faktorstufen. Für die Werte der abhängigen Variablen wird dabei in jeder Gruppe gesondert der Mittelwert ausgewiesen. Die entscheidende Frage lautet nun: Unterscheiden sich all diese Mittelwerte auch in der Grundgesamtheit signifikant voneinander, oder sind alle bei diesen Stichprobenwerten festellbaren Unterschiede lediglich auf einfache Zufallseffekte zurückzuführen?

Wenn sich die im Modell nicht erfassten Einflüsse (Wetterlage, Stattfinden von anderen Events in der Umgebung etc.) für alle vier Kinos und alle fünf Vorstellungen bis auf zufällige Abweichungen gleich stark auswirken bzw. nicht existent sind (ceteris paribus-Gedanke), ist zu schlussfolgern, dass die Abweichungen der Mittelwerte voneinander sich auf den Einfluss der Plakatdesigns zurückführen lassen. Der Erwartungswert für die Anzahl der Kinobesucher in einem beliebigen Kino zu einer beliebigen Vorstellung läge bei 32,05, wenn die Plakate

keine Rolle spielen würden. Geht man dagegen von einem Einfluss der Plakate aus, so ergeben sich für die vier Kinos unterschiedliche Erwartungswerte, nämlich 25, 44,2, 23,8 und 35,2. Die Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert sind nun durch den Einfluss des Faktors zu erklären, die Abweichungen der einzelnen gemessenen Werte vom Gruppenmittelwert sind dagegen auf zufällige Einflüsse zurückzuführen.

Die Gesamtabweichung lässt sich daher in zwei Komponenten zerlegen (hier spricht man auch von der sogenannten Streuungszerlegung): die (durch die Faktorstufen) erklärte Abweichung und die (durch die Faktorstufen) nicht erklärte Abweichung.

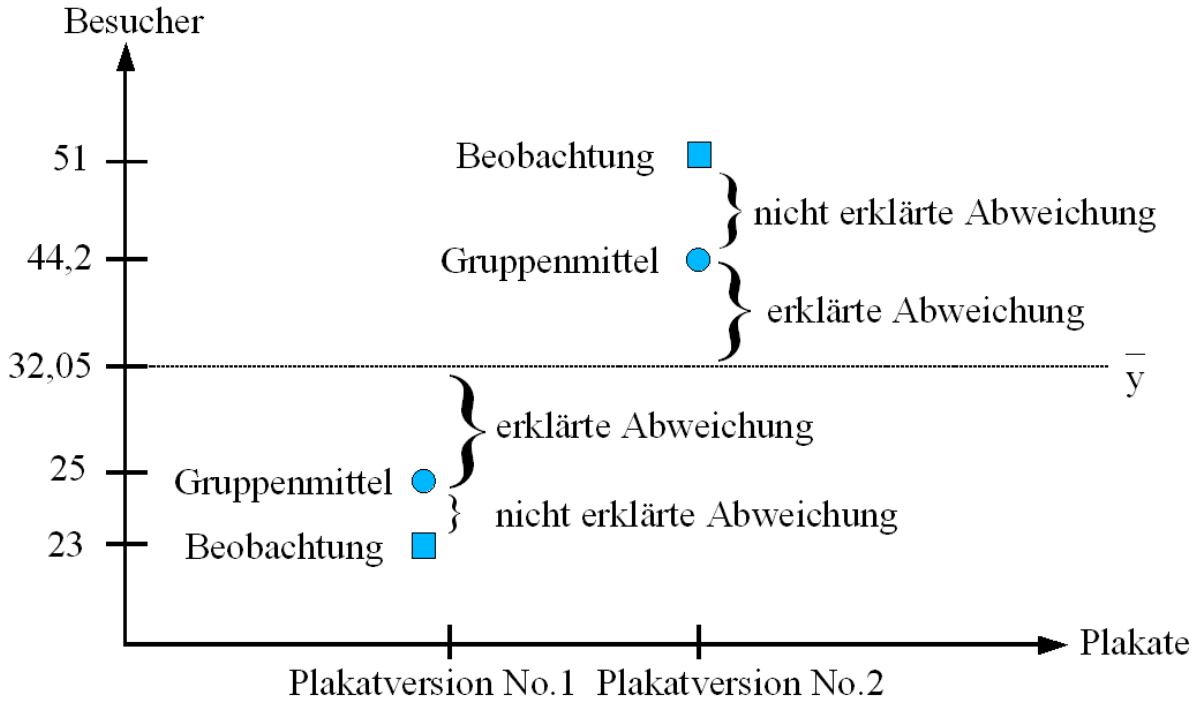

Abbildung 44: Grafische Streuungszerlegung im Beispielfall der Varianzanalyse

Diese Streuungszerlegung, die Aufteilung der Streuung in erklärte und nicht erklärte Abweichung, lässt sich nicht nur grafisch darstellen, sondern auch für jeden einzelnen erhobenen Wert rechnerisch ermitteln. Beispielhaft betrachten wir den Wert für die erste Vorstellung im ersten Kino: 23.

Die Summe der quadrierten (um den bereits bekannten positiv-negativ-Ausgleichseffekt zu umgehen) Gesamtabweichungen SSt (SSt = Total Sum of Squares) setzt sich also zusam-

men aus der Summe der quadrierten Abweichungen zwischen den Faktorstufen SSb (SSb = Sum of Squares between) und der Summe der quadrierten Abweichungen innerhalb der Faktorstufen SSw (SSw = Sum of Squares within): SSt = SSb + SSw.

Für den von uns betrachteten Wert beträgt die Summe der quadrierten Gesamtabweichungen SSt = (23-32,05)² = 81,9025, die Summe der quadrierten Abweichungen zwischen den Faktorstufen SSb = (25-32,05)² = 49,7025 (die durchschnittliche Besucherzahl des ersten Kinos am Stichtag lag bei 25), die Summe der quadrierten Abweichungen innerhalb der Faktorstufen SSw = (23-25)² = 4. Die Gesamtwerte für SSt, SSb und SSw lassen sich nach diesem Muster errechnen.

	SSt	SSb	SSw
Plakat No.1	(23-32,05)²=81,9025 (28-32,05)²=16,4025 (31-32,05)²=1,1025 (24-32,05)²=64,8025 (19-32,05)²=170,30	(25-32,05)²=49,7025 (25-32,05)²=49,7025 (25-32,05)²=49,7025 (25-32,05)²=49,7025 (25-32,05)²=49,7025	(23-25)²= 4 (28-25)²=9 (31-25)²=36 (24-25)²=1 (19-25)²=36
Plakat No. 2	(51-32,05)²=359,1025 (44-31,05)²=142,8025 (41-32,05)²=80,1025 (46-32,05)²=194,6025 (39-32,05)²=48,3025	(44,2-32,05)²=147,6225 (44,2-32,05)²=147,6225 (44,2-32,05)²=147,6225 (44,2-32,05)²=147,6225 (44,2-32,05)²=147,6225	(51-44,2)²=46,24 (44-44,2)²=0,04 (41-44,2)²=10,24 (46-44,2)²=3,24 (39-44,2)²=27,04
Plakat No. 3	(22-32,05)²=101,0025 (18-32,05)²=197,4025 (15-32,05)²=290,7025 (23-32,05)²=81,9025 (41-32,05)²=80,1025	(23,8-32,05)²=68,0625 (23,8-32,05)²=68,0625 (23,8-32,05)²=68,0625 (23,8-32,05)²=68,0625 (23,8-32,05)²=68,0625	(22-23,8)²=3,24 (18-23,8)²=33,64 (15-23,8)²=77,44 (23-23,8)²=0,64 (41-23,8)²=295,84
Plakat No. 4	(35-32,05)²=8,7025 (41-32,05)²=80,1025 (39-32,05)²=48,3025 (27-32,05)²=25,5025 (34-32,05)²=3,8025 SSt =2076,95	(35,2-32,05)²=9,9225 (35,2-32,05)²=9,9225 (35,2-32,05)²=9,9225 (35,2-32,05)²=9,9225 (35,2-32,05)²=9,9225 SSb =1376,55	(35-35,2)²=0,04 (41-35,2)²=33,64 (39-35,2)²=14,44 (27-35,2)²=67,24 (34-35,2)²=1,44 SSw =700,4

Abbildung 45: Rechnerische Streuungszerlegung im Beispielfall der Varianzanalyse

Zusammenfassend lässt sich festhalten: die Gesamtstreuung wird in zwei additive Komponenten zerlegt. Die erklärte Abweichung ist auf den Einfluss der Faktoren auf die abhängige Variable zurückzuführen, die nicht erklärte Abweichung wird durch unbekannte äußere oder zufällige Einflüsse verursacht.

Betrachtet man die Streuungszerlegung im Detail, so fällt auf, dass die Quadratsummen größer werden, je mehr Beobachtungswerte in die Berechnung eingehen. Dies bedeutet, dass SSt, SSb und SSw direkt von der Stichprobengröße abhängig sind und darum für sich genommen nicht aussagekräftig sind. Denn: der Einfluss der Faktoren auf die unabhängige Variable fällt oder wächst nicht mit der Größe der Stichprobe – wir überprüfen ja nicht die Zusammenhänge in der Stichprobe, sondern die „wahren" Zusammenhänge in der Grundgesamtheit, die mit Sicherheit nicht von der Größe unserer Stichprobe abhängig sind.

Gesamtabweichung	= erklärte Abweichung	+ nicht erklärte Abweichung
Summe der quadrierten Gesamtabweichungen	= Summe der quadrierten Abweichungen zwischen den Faktorstufen	+ Summe der quadrietern Abweichungen innerhalb der Faktorstufen
$\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk}-\bar{y})^2 =$	$\sum_{g=1}^{G}K(\bar{y}_g-\bar{y})^2$	$+ \sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk}-\bar{y}_g)^2$
SSt(otal)	= SSb(etween)	+ SSw(ithin)
SS = „sum of squares"		

Abbildung 46: Zusammenfassende Darstellung der Streuungszerlegung

Um eine aussagefähige Größe für die Streuung zu erhalten, werden die Werte durch die Anzahl der Freiheitsgrade geteilt. Es ergibt sich die Varianz, die von der konkreten Anzahl der Beobachtungswerte unabhängig ist. Diese empirische Varianz ist auch als mittlere quadrierte Abweichung MSS (MSS = Mean Sum of Squares) definiert.

$$Varianz\,(MSS) = \frac{SS}{(Zahl\,der\,Beobachtungen - 1)}$$

3 Exkurs: Freiheitsgrade

Die Freiheitsgrade geben die Anzahl von Größen eines Systems an, die bei einem feststehenden arithmetischen Mittel unabhängig voneinander variiert werden können. Ihre Bedeutung erklärt sich daraus, dass die Schätzung von Parametern in der Statistik stets eng mit den zur Verfügung stehenden Informationen verbunden ist. Die Anzahl an Informationen für die

Schätzung entspricht der Anzahl der Freiheitsgrade.

Wie lässt sich der Begriff des Freiheitsgrades nun definieren?

Man stelle sich folgenden Beispielfall vor: Eine Verteilung besteht aus den fünf Werten 1, 1, 2, 3 und 3. Das arithmetische Mittel dieser Verteilung liegt bei 2. Wenn die erste Zahl von 1 auf 2 geändert würde und die zweite Zahl von 1 auf 0, so läge das arithmetische Mittel immer noch bei 2 – die erste Zahl der geordneten Verteilung kann also frei verändert werden ohne das arithmetische Mittel zu verändern, solange auch die anderen Zahlen frei verändert werden können. Dies lässt sich bis zum letzten Wert der Verteilung fortsetzen – dieser kann dann allerdings nicht mehr frei festgelegt werden, wenn ein bestimmtes arithmetisches Mittel noch erreicht werden soll. Die Beispielverteilung hätte also fünf Werte und vier Freiheitsgrade – vier Werte die unter der Vorbedingung eines feststehenden arithmetischen Mittels noch frei festgelegt werden können.

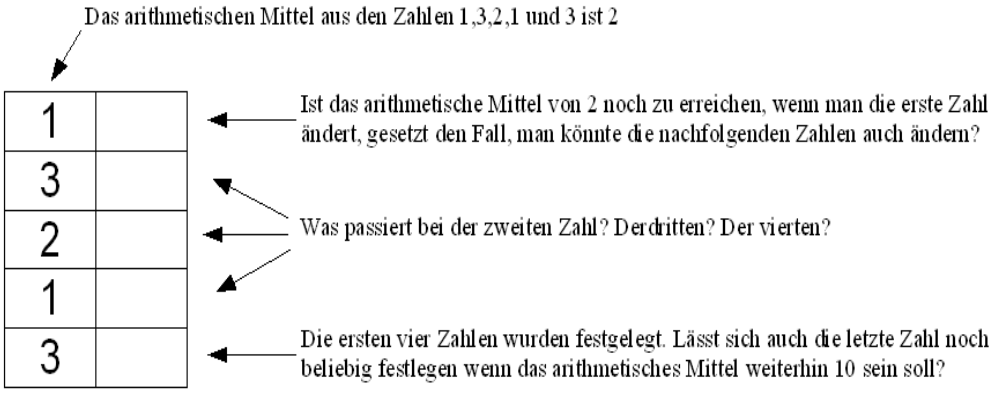

Abbildung 47: Definition des Begriffs der Freiheitsgrade

Merksatz: Die Freiheitsgrade geben die maximale Anzahl an Werten in einer Verteilung an, die beliebig geändert werden können, ohne dass sich das arithmetische Mittel der Verteilung ändert.

4 Zerlegung der Freiheitsgrade

Die Freiheitsgrade in unserem Beispielfall können analog zur Gesamtstreuung zerlegt werden. Insgesamt gibt es mit den 4 Kinos 4 Faktorstufen und 5 Beobachtungen pro Kino am Erhebungstag – dies bedeutet, dass eine Verteilung mit 19 Freiheitsgraden für die Berechnung der mittleren quadratischen Gesamtabweichung MSt (MSt = Mean Sum of Squares total) vorliegt. Auf der Ebene der 4 Faktorstufen können dann nur 3 frei variiert werden, was uns zu 3 Freiheitsgraden für die Berechnung der MSb (MSb = Mean Sum of Squares between) bringt. Auch die mittlere quadratische Abweichung innerhalb der Gruppen MSw (MSw = Mean Sum of Squares within) kann berechnet werden, wenn die Anzahl der Freiheitsgrade – 16 bei 4 von 5 frei variierbaren Beobachtungen pro Faktorstufe – feststeht.

Mittlere quadratische (Gesamt-)Abweichung („mean sum of squares“)	$MS_t = \frac{SS_t}{(G*K-1)}$	$MS_t = \frac{2076,95}{(4*5-1)} = \frac{2076,95}{19} = 109,313$
Mittlere quadratische Abweichung zwischen den Faktorstufen	$MS_b = \frac{SS_b}{(G-1)}$	$MS_b = \frac{1376,55}{(4-1)} = \frac{1376,55}{3} = 458,85$
Mittlere quadratische Abweichung innerhalb der Faktorstufen	$MS_w = \frac{SS_w}{(G*(K-1))}$	$MS_w = \frac{700,4}{(4*(5-1))} = \frac{700,4}{16} = 43,775$

Abbildung 48: Zerlegung der Freiheitsgrade im Beispielfall der Varianzanalyse

Wäre nun die mittlere quadratische Abweichung innerhalb der Faktorstufen MSw gleich Null, dann würde die mittlere quadratische Gesamtabweichung MSt ausschließlich durch den Einfluss der Faktoren erklärt werden. Je stärker MSw von Null abweicht, desto geringer muss also der Einfluss der Faktoren auf die abhängige Variable sein: Von Interesse ist also das Verhältnis von MSb zu MSw.

In unserem Beispielfall mit den Filmplakaten übertrifft der Wert von MSb den Wert von MSw bei weitem. Daraus kann geschlussfolgert werden, dass ein Einfluss der unabhängigen Variable – in diesem Fall die unterschiedliche Gestaltung der Plakate – auf die abhängige Variable – den Verkauf an Kinokarten – vorliegt.

5 Berechnung der Effektstärke

Ein gängiges Maß für die Stärke des Gesamteffekts ist das multiple Eta². Je näher das multiple Eta² an Eins liegt, desto größer ist der durch die Faktoren erklärte Anteil der Streuung an der Gesamtstreuung – und umso stärker ist der Gesamteffekt zu bewerten.

Das multiple Eta² berechnet sich aus :

$$Eta^2 = \frac{SS_b}{SS_t} = \frac{(Summe\,der\,quadrierten\,Abweichungen\,zwischen\,den\,Faktorstufen)}{(Summe\,der\,quadrierten\,Gesamtabweichungen)}$$

Beim Filmplakate-Beispielfall ergibt sich ein multiples Eta² von 0,6628. Dies bedeutet, dass 66,28% der Gesamtstreuung durch den Faktor „Plakatdesign“ aufgeklärt werden – ein durchaus beachtlicher Anteil, der auf einen existenten Einfluss des Faktors auf die abhängige Variable hinweist.

IV Prüfung der statistischen Unabhängigkeit

1 Varianzanalytischer F-Test

Aufgrund der Tatsache, dass die Daten für eine Varianzanalyse in der Regel aus einer Zufallsstichprobe stammen, können sich Unterschiede zwischen den Gruppenmittelwerten in der Stichprobe auch rein zufällig ergeben. Mit dem varianzanalytischen F-Test kann im Anschluss an die Streuungszerlegung überprüft werden, ob die gefundenen Effekte auch tatsächlich signifikant sind.

Die Nullhypothese dieses Tests lautet: Sämtliche „wahren“ Mittelwerte der abhängigen Variablen aller Faktorstufen in der Grundgesamtheit sind identisch. Ist dies der Fall, so besteht zwischen den durch die Faktoren gebildeten Gruppen bezüglich der abhängigen Variablen keinerlei Unterschied – die im Verlauf des zweiten Schritts der Varianzanalyse gefundenen Mittelwertunterschiede wären dann nur ein Zufallsergebnis.

Wie wird der varianzanalytische F-Test durchgeführt? Zunächst einmal ist festzustellen, dass jede in der Stichprobe durch die Faktorstufen gebildeten Gruppen als eine eigene, unabhängige Stichprobe betrachtet werden kann, wenn es sich bei der Ausgangsstichprobe um eine Zufallsstichprobe handelt (zwischen den Gruppen kann es keine Überschneidungen geben). Auf der Grundlage dieser „Einzelstichproben" kann nun, ausgehend von der oben dargestellten Varianzzerlegung, die Gesamtvarianz der Grundgesamtheit auf zwei verschiedene Arten geschätzt werden, nämlich mittels der Varianz innerhalb der Gruppen aus der Stichprobe oder mittels der Varianz zwischen den Gruppen aus der Stichprobe. Beide Stichprobenvarianzen führen zu zwei verschiedenen Schätzungen der „wahren" Varianz in der Grundgesamtheit.

Gilt die Nullhypothese (sämtliche „wahren" Mittelwerte der abhängigen Variablen aller Faktorstufen sind identisch), so müssten nun beide Schätzungen zum gleichen Ergebnis führen. Liegt dagegen auch in der Grundgesamtheit ein Einfluss des Faktors auf die abhängige Variable vor, so unterscheiden sich die Ergebnisse beider Schätzungen signifikant voneinander. Grund dafür ist, dass in diesem Fall die Einzelstichproben tatsächlich aus unterschiedlichen Grundgesamtheiten stammen, wenn man die abhängige Variable als Seperationskriterium verwenden würde. Die Varianz innerhalb der Gruppen ist dann in jedem Fall ein guter Schätzwert für die „wahre" Varianz in der Grundgesamtheit, die Varianz zwischen den Gruppen ist dagegen nur dann ein guter Schätzwert, wenn kein Einfluss der Faktoren vorliegt. Sind also beide Varianzen in etwa gleich, so spricht dies für die Richtigkeit der Nullhypothese und damit eine Einflusslosigkeit des Faktors, unterscheiden sich die beiden Werte dagegen signifikant, so ist von einem Einfluss des Faktors auszugehen.

Die Prüfung der Nullhypothese erfolgt im Rahmen des varianzanalytischen F-Tests. Von wesentlicher Bedeutung ist hier der Quotient aus der Varianz zwischen den Gruppen und der Varianz innerhalb der Gruppen, der die Testgröße F bildet.

$$F = \frac{s_b^2}{s_w^2}$$

Aus der bekannten F-Verteilung lässt sich unter Berücksichtigung der Freiheitsgrade für beide Varianzschätzungen die Wahrscheinlichkeit für das Auftreten des berechneten F-Wertes

unter Gültigkeit der Nullhypothese berechnen. Bei Gültigkeit der H0 ist ein F-Wert von nahe Eins zu erwarten, je weiter der Wert abweicht, desto unwahrscheinlicher ist es, dass in der Grundgesamtheit in der Tat kein Zusammenhang zwischen abhängiger Variable und Faktoren besteht.

SPSS gibt die Irrtumswahrscheinlichkeit des Tests aus, also die Wahrscheinlichkeit, mit welcher der Marktforscher einen Fehler macht, wenn er die Nullhypothese verwirft. Ist diese Irrtumswahrscheinlichkeit niedrig (oft werden 0,05 bzw. 0,01 als kritische Werte verwendet), kann die Nullhypothese problemlos verworfen werden. Die Schlussfolgerung aus einer solchen Ablehnung der H0 lautet dann: Mindestens zwei der „wahren" Faktorstufenmittelwerte sind mit großer Wahrscheinlichkeit nicht identisch, daher ist von einem Zusammenhang zwischen der abhängigen Variable und den Faktoren auszugehen, wobei noch nicht geklärt ist, welche der Faktoren bzw. Faktorstufen signifikant werden (dies muss im Anschluss an die Verwerfung der H0 im varianzanalytischen F-Test anhand einer Auswahl diverser Post-Hoc-Tests ermittelt werden).

ONEWAY ANOVA

Besucherzahl

	Quadrats umme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	1376,550	3	458,850	10,482	,000
Innerhalb der Gruppen	700,400	16	43,775		
Gesamt	2076,950	19			

Abbildung 49: Ergebnis des varianzanalytischen F-Tests

In unserem Beispielfall mit den Kinoplakaten liegt die Irrtumswahrscheinlichkeit bei der Ablehnung der Nullhypothese so dicht an Null, das SPSS den Wert auf Null abrundet. Die Nullhypothese kann daher problemlos verworfen werden, zwischen mindestens zwei „wahren" Faktorstufenmittelwerten bestehen mit großer Sicherheit Unterschiede in der Grundgesamtheit. Interpretatorisch bedeutet dieses Ergebnis, dass es bezüglich der durchschnittlichen Besucherzahlen zwischen mindestens zwei der fünf betrachteten Kinos einen deutlichen Unterschied gibt und sich somit mindestens zwei der fünf Plakatversionen in ihrer Wirkung auf die Besucherzahlen unterscheiden. Damit ist allerdings noch nicht bewiesen, dass alle Plakate, also alle Faktorstufen, signifikant werden.

2 Post-Hoc-Tests

Diese Überprüfung erfolgt im Rahmen der Post-Hoc-Tests, die sinnvollerweise nur dann durchgeführt werden sollten, wenn der F-Test zuvor mindestens einen signifikanten Mittelwertsunterschied ergeben hat. Handelt es sich bei der Varianzanalyse um eine ANOVA mit nur einem Faktor und lediglich zwei Faktorstufen, so kann bei signifikantem F-Test der Post-Hoc-Test vollständig übersprungen werden, da in einem solchen Fall nur ein einziger Mittelwertsunterschied vorliegt.

Post-Hoc-Tests ermöglichen also die Feststellung, welche der „wahren" Faktorstufenmittelwerte sich nach einem signifikant gewordenen F-Test nun tatsächlich voneinander unterscheiden, bzw. aus welchen Faktorstufen gegebenenfalls homogene Untergruppen gebildet werden können.

Dabei ist zwischen Paarvergleichstests und Spannweitentests zu unterscheiden. Ein Paarvergleichstest testet die Mittelwertdifferenzen aller möglichen Faktorstufenpaare auf Signifikanz, während mit einem Spannweitentest umgekehrt nach nicht signifikanten Mittelwertdifferenzen zur Bildung homogener Untergruppen gesucht wird. Üblicherweise kommen in der Varianzanalyse Paarvergleichstests zum Einsatz, wobei der konservative Scheffé-Test der gebräuchlichste ist.

Wie bereits oben erwähnt, ist der am häufigsten angewandte Post-Hoc-Test der sogenannte Scheffé-Test – bei allen anderen Testverfahren handelt es sich um Spezialtests, die im Rahmen dieses Manuskripts nicht weiter beachtet werden. Der Scheffé-Test ist vergleichsweise robust gegenüber Verletzungen seiner Voraussetzungen (insbesondere die der Linearität des Zusammenhang) und testet äußerst konservativ – der Marktforscher ist also mit den Ergebnissen des Scheffé-Tests sozusagen „auf der sicheren Seite". Dieser Vorteil kann aber auch zum interpretatorischen Nachteil werden, da eine Situation auftreten kann, in welcher der Scheffé-Test keinen einzigen signifikanten Mittelwertsunterschied ausweist, obwohl zuvor ein F-Test ergeben hat, dass mit großer Wahrscheinlichkeit zwischen mindestens zwei Faktorstufenmit-

telwerten ein signifikanter Unterschied besteht. In solchen Fällen liegt es am Marktforscher, ob in der Folge ein weniger konservativer Post-Hoc-Test eingesetzt oder die Varianzanalyse als solche noch einmal in Frage gestellt wird.

Paarvergleichstests	***Spannweitentests***
• Scheffé	• Student-Newman-Keuls
• LSD	• Duncan
• Bonferroni	• Tukey-B
• Sidak	• Waller-Duncan
• Tukey	• F-Test nach Ryan-Einot-Gabriel-Welsh
• GT2 Hochberg	• Q-Test nach Ryan-Einot-Gabriel-Welsh
• Gabriel	
• Dunnett	
• Tamhane T2	
• Dunnett D3	
• Games-Howell	
• Dunnett C	

Tabelle 4: Übersicht der Post-Hoc-Tests für die Varianzanalyse

Mehrfachvergleiche

Abhängige Variable: Besucherzahl
Scheffé-Prozedur

(I) Plakatversion	(J) Plakatversion	Mittlere Differenz (I-J)	Standardfehler	Signifikanz	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
1,00	2,00	-19,2000*	4,18450	,003	-32,2437	-6,1563
	3,00	1,2000	4,18450	,994	-11,8437	14,2437
	4,00	-10,2000	4,18450	,158	-23,2437	2,8437
2,00	1,00	19,2000*	4,18450	,003	6,1563	32,2437
	3,00	20,4000*	4,18450	,002	7,3563	33,4437
	4,00	9,0000	4,18450	,242	-4,0437	22,0437
3,00	1,00	-1,2000	4,18450	,994	-14,2437	11,8437
	2,00	-20,4000*	4,18450	,002	-33,4437	-7,3563
	4,00	-11,4000	4,18450	,099	-24,4437	1,6437
4,00	1,00	10,2000	4,18450	,158	-2,8437	23,2437
	2,00	-9,0000	4,18450	,242	-22,0437	4,0437
	3,00	11,4000	4,18450	,099	-1,6437	24,4437

*. Die mittlere Differenz ist auf der Stufe .05 signifikant.

Abbildung 50: Ergebnisse des Scheffé-Tests im Beispielfall der Varianzanalyse

V Die zweifaktorielle Varianzanalyse

1 Einführung

Während wir bislang ausschließlich die einfaktorielle Varianzanalyse betrachtet haben, soll im letzten Abschnitt dieses Kapitels nun die zweifaktorielle Varianzanalyse als Beispiel für die mehrfaktorielle Varianzanalyse im Allgemeinen betrachtet werden. Der einzige rechnerische Unterschied zwischen der zweifaktoriellen und der drei-, vier- oder fünffaktoriellen Varianzanalyse liegt ohnehin in der steigenden Komplexität – abgesehen davon, ändert sich am Ablauf des Verfahrens durch das Hinzufügen weiterer Faktoren nichts. Daher ist es vollkommen ausreichend, die zweifaktorielle Varianzanalyse stellvertretend für alle mehrfaktoriellen Varianzanalysen zu betrachten.

Wie es bereits der Bezeichnung zu entnehmen ist, untersucht die zweifaktorielle Varianzanalyse die Effekte zweier unabhängiger Größen auf eine abhängige Variable. Dabei kann der kombinierte Einfluss beider Faktoren ebenso untersucht werden wie der isolierte Einfluss jedes Faktors sowie eine mögliche Interaktion zwischen diesen. Im Vergleich zur einfaktoriellen Varianzanalyse, bei der lediglich der isolierte Einfluss eines Faktors auf die abhängige Variable betrachtet wurde, steigt also die Komplexität des untersuchten Szenarios und damit auch die Anzahl der durchzuführenden Tests und Rechenoperationen. Viele Elemente der einfaktoriellen Varianzanalyse bleiben aber auch gleich, so dass im Folgenden nur die Zusätze und Erweiterungen im Detail dargestellt werden.

2 Haupt- und Interaktionseffekte

Wie bereits festgestellt, ist bei der zwei- bzw. der mehrfaktoriellen Varianzanalyse im Gegensatz zur einfaktoriellen Varianzanalyse noch in Haupt- und Interaktionseffekt zu unterscheiden, also zwischen dem isolierten und dem kombinierten Einfluss der Faktoren auf die abhängige Variable. Der mehrfaktoriellen Varianzanalyse liegt ein linear-additives Modell zugrunde – der Gesamteffekt setzt sich additiv aus Haupt- und Interaktionseffekten zusammen.

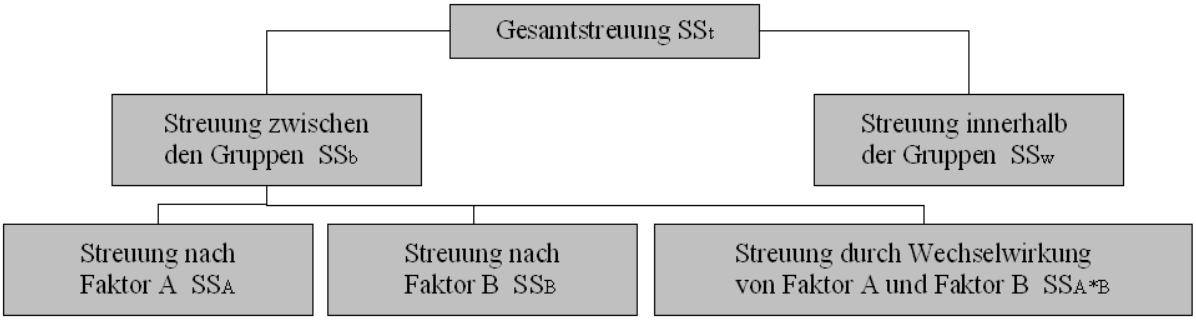

Abbildung 51: Streuungszerlegung in der zweifaktoriellen Varianzanalyse

Die Streuungszerlegung gestaltet sich analog zur einfaktoriellen Varianzanalyse, nur dass neben der Streuung innerhalb des zweiten Faktors auch noch die Streuung hinzukommt, die der Wechselwirkung der beiden Faktoren zuzurechnen ist. Es ergeben sich also zwei statt nur einer neuen Streuungskomponente durch die Aufnahme eines weiteren Faktors in das Modell. Ähnlich stellt sich die Entwicklung bei der Aufnahme eines dritten Faktors dar: In diesem Fall gibt es dann schon acht Streuungskomponenten – die Komplexität der Varianzanalyse steigt also rapide mit der Anzahl der Faktoren.

Gesamtstreuung	$SS_t = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{k=1}^{K}(Y_{ghk} + \bar{Y})^2$	$\underline{Y}ghk$ = Beobachtungswert $\overline{Y}$ = Gesamtmittelwert
Haupteffekte Streuung nach Faktor A & Streuung nach Faktor B	$SS_A = H * K * \sum_{g=1}^{G}(Y_g - \bar{Y})^2$ $SS_B = G * K * \sum_{g=1}^{G}(Y_h - \bar{Y})^2$	G = Zahl der Ausprägungen von A H = Zahl der Ausprägungen von B K = Zahl der beobachteten Elemente Yg = Zeilenmittelwert Yh = Spaltenmittelwert
Interaktionseffekt Interaktion zwischen den Faktoren A und B	$SS_{(A*B)} = K * \sum_{g=1}^{G}\sum_{h=1}^{H}(\overline{Y}_{gh} - \hat{Y}_{gh})^2$ $mit: \hat{Y} = \overline{Y}_g + \overline{Y}_h - \bar{Y}$	G = Zahl der Ausprägungen von A H = Zahl der Ausprägungen von B K = Zahl der beobachteten Elemente $\overline{Y}gh$ = Mittelwert in Zelle (g,h) $\hat{Y}gh$ = Schätzwert in Zelle (g,h)
Streuung innerhalb der Gruppen	$SS_w = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{k=1}^{K}(Y_{ghk} - \overline{Y}_{gh})^2$	$\underline{Y}ghk$ = Beobachtungswert $\overline{Y}gh$ = Mittelwert in Zelle (g,h)

Abbildung 52: Streuungszerlegung in der zweifaktoriellen Varianzanalyse (2)

3 Prüfung der statistischen Unabhängigkeit

Da die zweifaktorielle Varianzanalyse abgesehen von der anderen Streuungszerlegung wie die einfaktorielle Varianzanalyse abläuft, gibt es keinen Grund, das Verfahren hier noch einmal ausführlich darzustellen. Es soll aber noch auf die neuen Tests im vierten Schritt der Varianzanalyse, der Prüfung auf statistische Unabhängigkeit, hingewiesen sowie, im nächsten Unterabschnitt, auf die Interpretation der Interaktionsdiagramme eingegangen werden.

Während im Rahmen der einfaktoriellen Varianzanalyse ein einzelner Test (varianzanalytischer F-Test) ausreicht, um das Gesamtmodell zu testen[1] und anschließend über eine Reihe von Post-Hoc-Tests noch die einzelnen Faktorstufenmittelwerte auf Differenzen überprüft werden konnten, sind in der zweifaktoriellen Varianzanalyse drei separate Sets von Tests erforderlich, bevor mit den Post-Hoc-Tests begonnen werden kann.

Test des Gesamteffekts auf Signifikanz

Zunächst einmal wird, analog zur einfaktoriellen Varianzanalyse, die Nullhypothese getestet, dass alle Faktorstufenmittelwerte (aller Faktoren!) identisch sind und es demzufolge keine signifikanten, durch die Faktoren bestimmten Mittelwertunterschiede zwischen den einzelnen Gruppen gibt. Auch hier kann die Prüfgröße der F-Verteilung entnommen werden.

Test der Haupteffekte A und B

Wurde beim vorangegangenen Test ein signifikanter Gesamteffekt festgestellt, so ist noch nicht klar, ob dieser durch einen oder beide Faktoren des zweifaktoriellen Modells ausgelöst wurde. Der Test des Gesamteffekts wird ja bereits signifikant, wenn es nur einen Mittelwertunterschied zwischen zwei Faktorstufen der Grundgesamtheit gibt[2], die ja durchaus zum gleichen Faktor gehören können. Daher ist ein separater Test beider Haupteffekte notwendig, wobei bereits feststeht, dass mindestens einer der beiden signifikant werden muss. Die Nullhypo-

1 Mit dem varianzanalytischen F-Test konnte die Nullhypothese getestet werden, dass alle Faktorstufenmittelwerte in der Grundgesamtheit identisch sind – dass es also keine Mittelwertunterschiede zwischen den einzelnen Gruppen gibt, womit die Fortführung der Varianzanalyse überflüssig wäre.

2 bzw., wenn von der Existenz einer solchen Mittelwertdifferenz zwischen zwei Faktorstufenmittelwerten mit einer hohen Wahrscheinlichkeit ausgegangen werden kann.

these dieses Tests lautet wieder: Alle Faktorstufenmittelwerte (diesmal innerhalb A oder B) sind identisch. Die Prüfgröße kann ebenfalls wieder der F-Verteilung entnommen werden.

Test auf Interaktionseffekte

Anschließend ist noch auf Interaktionseffekte zu testen – inwiefern existiert also eine Interaktion zwischen den beiden Faktoren. Die H0 dieses Tests lautet konsequenterweise: Die Mittelwerte sämtlicher Interaktionsstufen sind identisch, daher liegt keine Interaktion zwischen den Faktoren vor.

Kann diese Nullhypothese verworfen werden, ist von dem Vorhandensein signifikanter Interaktionseffekte auszugehen, die in einem zusätzlichen fünften Schritt der mehrfaktoriellen Varianzanalyse nun noch interpretiert werden müssen.

4 Interpretation der Interaktionseffekte

Liegen signifikante Interaktion zwischen den Faktoren vor, wird die Interpretation der Haupteffekte erheblich erschwert, unter Umständen bis zu einem Punkt, an dem keiner der Haupteffekte überhaupt noch isoliert interpretiert werden kann. Der Einfluss eines interagierenden Faktors kann im Grunde nur adäquat dargestellt werden, wenn gleichzeitig stets auch der andere Faktor betrachtet wird.

Interaktionsdiagramme helfen bei der komplexen Interpretation der Interaktionseffekte, indem sie diese grafisch veranschaulichen. In ihnen werden sämtliche Gruppenmittelwerte, getrennt nach Faktoren, gegeneinander abgetragen. Es ergeben sich zwei Verlaufsformen:

- Parallele Verläufe = Es liegt keine Interaktion zwischen den Faktoren vor. In diesem Fall sind die Haupteffekte problemlos isoliert interpretierbar und ergeben in ihrer Summe den Gesamteffekt. Das Nichtvorhandensein von Interaktionseffekten sollte sich aber auch schon im Test auf Interaktionseffekte zeigen – die Erstellung von Interaktionsdiagrammen ist im Falle eines nicht-signifikanten Tests im Grund überflüssig.

- Nichtparallele Verläufe = Es liegen Interaktionen zwischen den Faktoren vor. In diesem Fall können die Haupteffekte eventuell nicht mehr inhaltlich interpretiert werden. Die Art der Interaktion ergibt sich aus der Art des Verlaufs, wobei in drei mögliche Verlaufsformen zu unterscheiden ist:

  - Ordinale Interaktion = Die Linienzüge weisen in beiden Diagrammen den gleichen Trend auf und schneiden sich daher nicht. In solchen Fällen sind beide Haupteffekte auch isoliert noch sinnvoll interpretierbar.
  - Hybride Interaktion = Die Linienzüge schneiden sich nur in einem der beiden Diagramme. In diesem Fall ist nur der Haupteffekt inhaltlich interpretierbar, dessen Linienzüge sich nicht kreuzen.
  - Disordinale Interaktion = Die Linienzüge schneiden sich in beiden Diagrammen. In diesem Fall kann keiner der Haupteffekte mehr inhaltlich interpretiert werden.

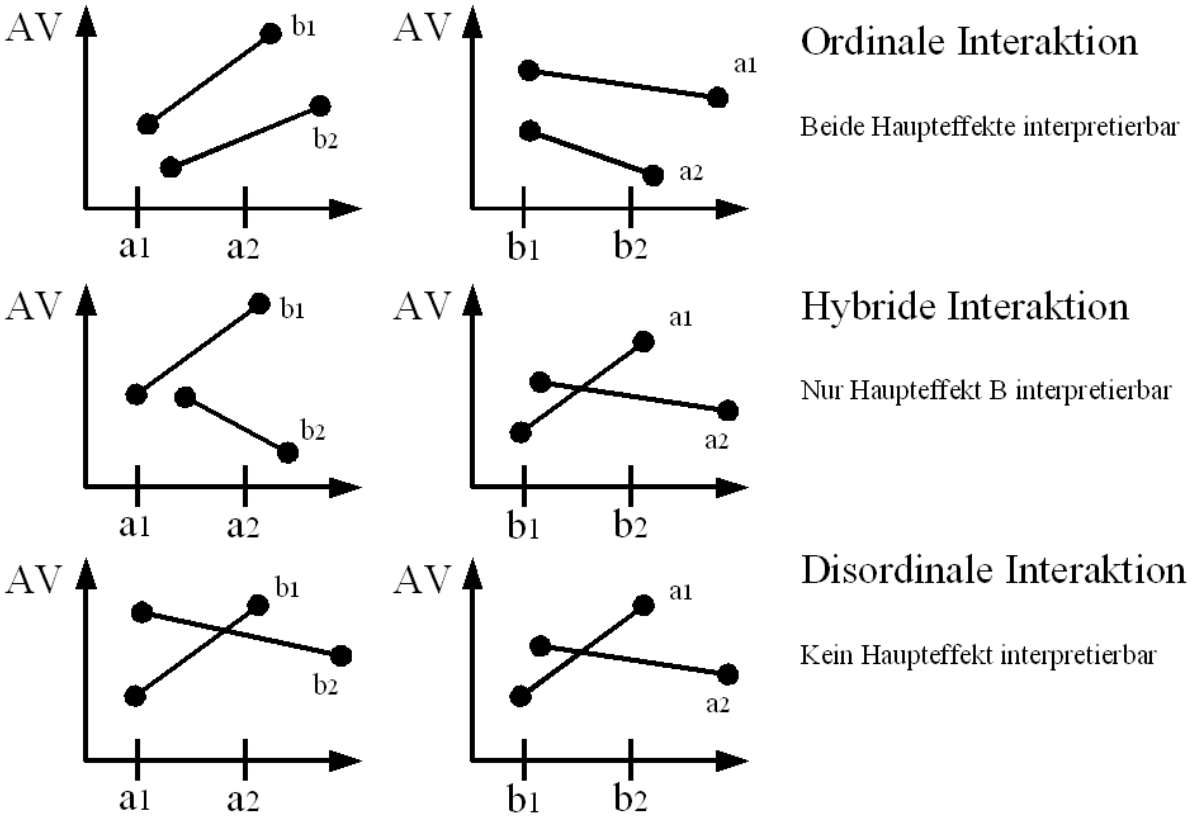

Abbildung 53: Interpretation von Interaktionsdiagrammen

VI Weiterführende Literatur

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). Multivariate Analysemethoden (10. Aufl.). Berlin: Springer.

Bortz, J. (1999). Statistik für Sozialwissenschaftler (5. Aufl.) Berlin: Springer.

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz.

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

E Faktorenanalyse

I Einführung

1 Hintergründe der Faktorenanalyse

Für viele marktforscherische Fragestellungen ist die Untersuchung des Wirkungszusammenhangs zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen von Bedeutung. Existiert nur eine geringe Anzahl unabhängiger Variablen, so lassen sich problemlos Korrelationsanalysen oder Regressionsanalysen durchführen, wie bereits oben dargestellt. Für eine Vielzahl von Fragestellungen ist dies ausreichend, beispielsweise für physikalische oder technische Untersuchungen. Existieren dagegen sehr viele unabhängige Variablen – und dies ist bei komplexen Untersuchungen aus dem Bereich der Markt- oder Sozialforschung keineswegs eine Seltenheit – dann wird die Auswertung wesentlich komplizierter. Dazu kommt, dass dem Marktforscher häufig verborgen bleibt, welche dieser vielen unabhängigen Variablen wirklich unabhängig voneinander zum Erklärungsmodell beitragen.

Die Faktorenanalyse wird eingesetzt, um aus einer großen Menge von Variablen voneinander unabhängige Beschreibungs- und Erklärungsfaktoren zu extrahieren. Dies führt nicht nur zu einer Vereinfachung bei der Auswertung durch die Reduktion der Variablen auf komplexere Hintergrundfaktoren, sondern erlaubt es dem Marktforscher auch, zunächst einmal wahllos eine große Menge an interessant erscheinenden Variablen zu erheben, und dann im Zuge der Faktorenanalyse alle irrelevanten Merkmale wieder auszuschließen.

2 Was ist unter einer Faktorenanalyse zu verstehen?

Die Faktorenanalyse gehört zu den strukturen-entdeckenden Verfahren – ihr Ziel ist also die Aufdeckung von Zusammenhängen zwischen verschiedenen Variablen. Eine vorausgehende Aufteilung der Variablen in abhängige und unabhängige Variablen ist daher bei diesem Verfahren überflüssig.

Die Faktorenanalyse kommt immer dann zum Einsatz, wenn die Bündelung von Variablen von methodischem Interesse ist. Dies ist insbesondere dann oft der Fall, wenn, wie oben bereits dargestellt, eine sehr große Anzahl von Merkmalen zu einer Fragestellung erhoben wurde. Hier stellt sich die Frage, ob es möglich ist, die Vielzahl von Variablen auf einige wenige zentrale Faktoren zu reduzieren. Das Ziel ist also die Verdichtung von Informationen, bzw. die Identifikation erklärungsrelevanter Variablen.

Dazu ein Beispiel: Es lassen sich viele Variablen vorstellen, mit denen man die technische Beschaffenheit eines Autos abbilden könnte: PS-Zahl, Höchstgeschwindigkeit, maximale Beschleunigung, Drehzahl etc. Sicher wäre es praktisch, diese Vielzahl von Variablen auf wenige Hintergrundfaktoren zusammenfassen zu können, beispielsweise Leistung, Sicherheit und Preisklasse. Dies geschieht im Rahmen der Faktorenanalyse.

3 Explorativ oder konfirmatorisch?

Bevor wir mit der Betrachtung der Faktorenanalyse beginnen, muss noch festgehalten werden, dass es zwei teils sehr unterschiedliche Formen der Faktorenanalyse gibt, die sich sowohl von der Zielsetzung als auch vom interpretatorischen Ansatz deutlich voneinander unterscheiden: Die explorative und die konfirmatorische Faktorenanalyse.

Angenommen, zu Beginn einer Untersuchung liegen keinerlei Informationen über mögliche Zusammenhänge zwischen den betrachteten Variablen vor. Der Marktforscher, der sich auf die Suche nach Strukturen im Datensatz begibt, nutzt die Faktorenanalyse explorativ, als rein strukturen-entdeckendes Verfahren zur Hypothesengenerierung. Solche Faktorenanalysen werden als explorative Faktorenanalysen bezeichnet.

Ein anderer Fall liegt vor, wenn es zu Beginn einer Untersuchung zumindest bereits vage Vorstellungen über mögliche Zusammenhänge und Strukturen in den Daten gibt. Der Marktforscher vermutet das Vorhandensein verschiedener Faktoren und hat auch bereits eine Vorstellung darüber, welche der betrachteten Variablen möglicherweise welchem Faktor zugeord-

net werden sollen. Die Faktorenanalyse dient in diesem Fall als Instrument zu Hypothesenüberprüfung. Solche Faktorenanalysen werden als konfirmatorische Faktorenanalysen bezeichnet.

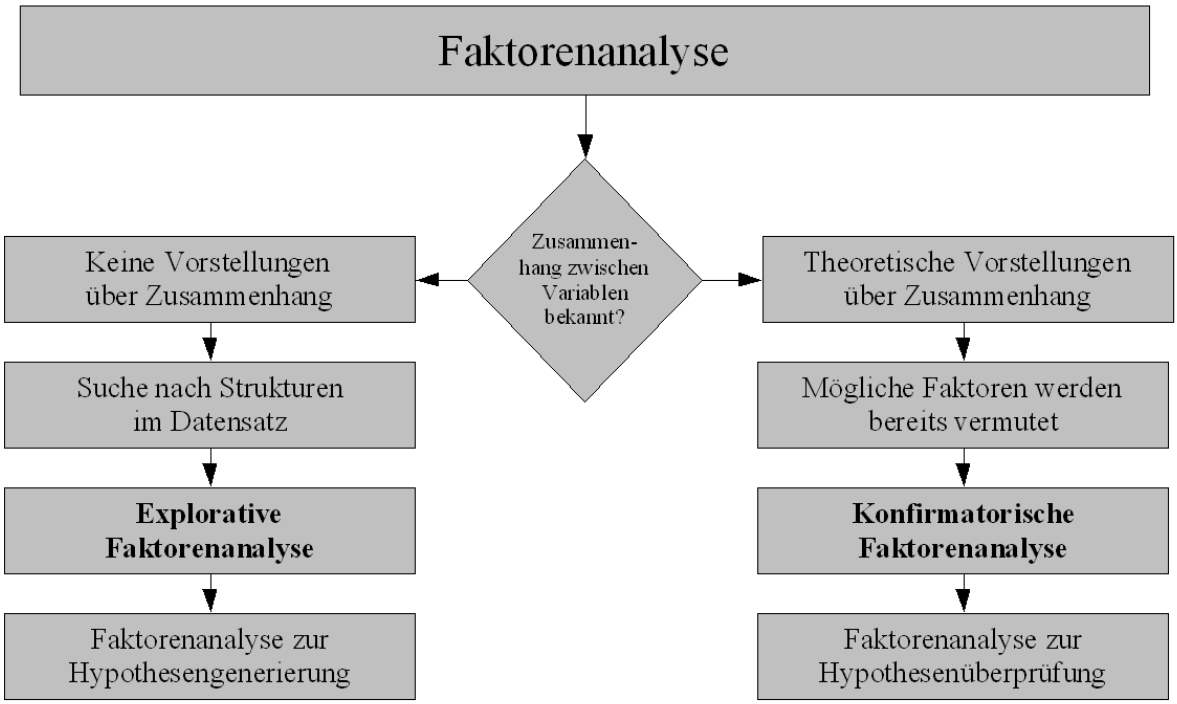

Abbildung 54: Explorative und konfirmatorische Faktorenanalysen

4 Genereller Zielkonflikt der Faktorenanalyse

Das Ziel der Faktorenanalyse ist, wie bereits dargestellt, die Reduktion vieler Variablen auf mehr oder weniger komplexe Hintergrundfaktoren. Dabei kommt es zu einem Zielkonflikt, der sich durch das komplette Verfahren zieht und der an dieser Stelle kurz beleuchtet werden soll.

Werden für die vielen Variablen nur wenige Hintergrundfaktoren gebildet oder „extrahiert“, wie der Fachterminus lautet, dann hat die Faktorenanalyse einen großen Beitrag zur Reduktion der Variablen und damit zur Vereinfachung des Sachverhalts erbracht. Die Reduktion vieler Variablen auf wenige Hintergrundfaktoren ist genau der Grund, aus dem die Fakto-

renanalyse überhaupt betrieben wird – je mehr Variablen sich also auf wenige Faktoren reduzieren lassen, umso größer ist der analytische Nutzen der Faktorenanalyse. Wie sich jeder versierte Marktforscher leicht vorstellen kann, bringt eine geringe Anzahl extrahierter Faktoren bei einer Vielzahl von Variablen aber auch einen großen Informationsverlust mit sich – umso weniger Faktoren extrahiert werden, umso größer ist dieser Informationsverlust und damit auch die Unsicherheit des gesamten Modells.

Nun ließe sich dieser Informationsverlust natürlich verringern, indem einfach mehr Faktoren extrahiert werden – damit unterwandert man allerdings das eigentliche Ziel, nämlich die bereits angesprochene Reduktion und Vereinfachung. Im Extremfall könnte man aus 20 Variablen auch 20 Faktoren extrahieren und hätte einen Informationsverlust von Null – aber auch eine Vereinfachung und damit einen Nutzen von Null.

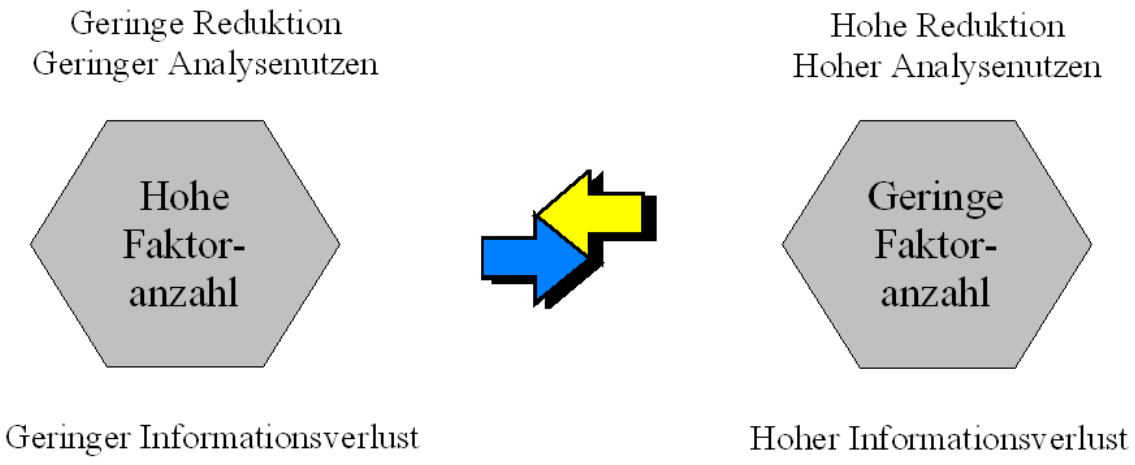

Abbildung 55: Zielkonflikt der Faktorenanalyse

Dieser Zielkonflikt – möglichst wenig Faktoren bei gleichzeitig möglichst geringem Informationsverlust zu extrahieren – prägt die Faktorenanalyse maßgeblich, und wir werden später noch sehen, dass der Marktforscher im Verlauf jeder Faktorenanalyse an einen Punkt gelangt, an dem er die „goldene Mitte“ zwischen beiden Zielen finden und die Anzahl der zu extrahierenden Faktoren bestimmen muss – diese Aufgabe nimmt einem SPSS nämlich keineswegs ab.

5 Ablauf einer Faktorenanalyse

Jede Faktorenanalyse läuft in vier wesentlichen Schritten ab.

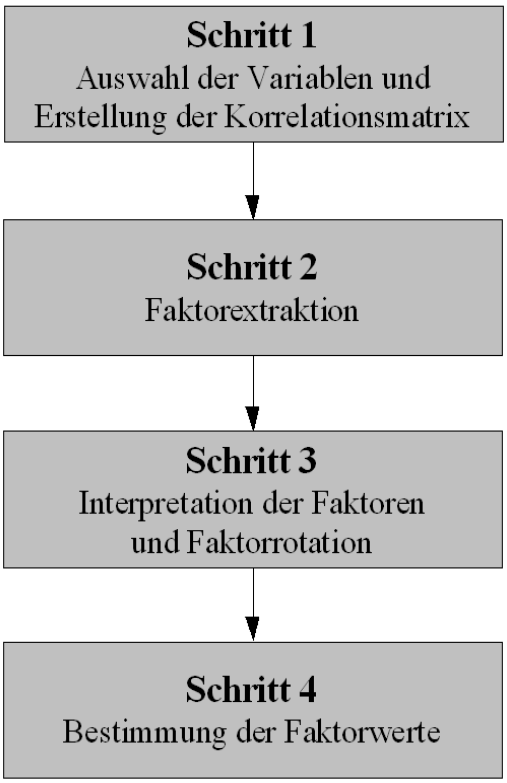

Abbildung 56: Der Ablauf der Faktorenanalyse

Im ersten Schritt sind alle Variablen auszuwählen, die in die Faktorenanalyse eingehen sollen. Für alle selektierten Variablen wird anschließend die sogenannte Korrelationsmatrix erstellt. Der Korrelationsmatrix lässt sich entnehmen, welche Variablen in der weiteren Analyse unberücksichtigt bleiben sollen, da sie mit den übrigen Variablen nur minimal korrelieren und somit sicher keinem gemeinsamen Hintergrundfaktor zugeordnet werden können.

Der zweite Schritt, die Faktorextraktion, wird auch als „Ziehen der Faktoren“ bezeichnet. Aufgrund verschiedener statistischer Kennzahlen kann in dieser Stufe entschieden werden, ob das gefundene Faktorenmodell geeignet ist, um die vorliegenden Variablen auf Hintergrundfaktoren zurückzuführen.

Die im zweiten Schritt extrahierten Faktoren sind in der Regel nur sehr schwer oder auch gar nicht zu interpretieren. Um die Ergebnisinterpretation zu erleichtern, werden die Faktoren im dritten Schritt einer speziellen Transformation unterzogen, die als Faktorrotation bezeichnet wird.

Im vierten und letzten Schritt wird ermittelt, welche Werte die untersuchten Variablen hinsichtlich der extrahierten und rotierten Faktoren annehmen. Dies dient der inhaltlichen Interpretation der Faktoren: Welche Variablen sind welchen Faktoren zuzuordnen und wie gut erklären die extrahierten Faktoren die betrachteten Variablen insgesamt?

II Erstellung und Eignung der Korrelationsmatrix

1 Erstellung der Korrelationsmatrix

Wie bereits eingangs dargestellt, sind Faktoren als „hinter den Variablen stehende Größen“ zu begreifen. Sie repräsentieren damit den Zusammenhang zwischen verschiedenen der betrachteten Ausgangsvariablen. Dieser Zusammenhang kann durch eine Korrelationsrechnung messbar gemacht werden. Korrelationen, dies ist (hoffentlich) noch aus der Statistik I bekannt, zeigen den Grad des Zusammenhangs zwischen Variablen, wodurch diese im Sinne der Faktorenanalyse als „bündelungsfähig“ oder „nicht bündelungsfähig“ identifiziert werden können.

Die Korrelationsmatrix für die Faktorenanalyse wird anhand des Bravais-Pearson-Korrelationskoeffizienten berechnet.

$$r_{(x1,x2)} = \frac{\left(\sum_{k=1}^{K}\left((x_{k1}-\bar{x}_1)*(x_{k2}-\bar{x}_2)\right)\right)}{\left(\sqrt{\left(\sum_{k=1}^{K}(x_{k1}-\bar{x}_1)^2\right)*\left(\sum_{k=1}^{K}(x_{k2}-\bar{x}_2)^2\right)}\right)}$$

Vor der Berechnung der Korrelationsmatrix sind gegebenenfalls noch die Ausgangsdaten

zu standardisieren. Dadurch werden die Korrelationsrechnung und die Rechenschritte der Faktorenanalyse erleichtert und die Interpretation der Ergebnisse vereinfacht. Außerdem ist die Standardisierung die einzige Möglichkeit, Variablen mit unterschiedlichen Maßeinheiten vergleichbar zu machen – und mit dem Auftreten unterschiedlicher Maßeinheiten und Dimensionen ist ja bei entsprechend vielen Variablen durchaus zu rechnen.

Das Standardisierungsverfahren ist als Z-Standardisierung bereits aus der Statistik II bekannt: Es wird die Differenz zwischen Mittelwert und Beobachtungswert einer Variablen gebildet und durch die Standardabweichung dividiert. Dadurch ist sichergestellt, dass der neue Erwartungswert Null und die neue Standardabweichung Eins ist – dass also die standardisierte Variable einer Standardnormalverteilung[1] folgt.

Korrelationsmatrix[a]

		Hubraum (cu. inches)	PS	Gewicht (lbs.)	Beschleunigung von 0 auf 100 km/h (sec.)	Anzahl der Zylinder
Korrelation	Hubraum (cu. inches)	1,000	,899	,934	-,562	,952
	PS	,899	1,000	,865	-,709	,844
	Gewicht (lbs.)	,934	,865	1,000	-,438	,896
	Beschleunigung von 0 auf 100 km/h (sec.)	-,562	-,709	-,438	1,000	-,528
	Anzahl der Zylinder	,952	,844	,896	-,528	1,000
Signifikanz (1-seitig)	Hubraum (cu. inches)		,000	,000	,000	,000
	PS	,000		,000	,000	,000
	Gewicht (lbs.)	,000	,000		,000	,000
	Beschleunigung von 0 auf 100 km/h (sec.)	,000	,000	,000		,000
	Anzahl der Zylinder	,000	,000	,000	,000	

a. Determinante = 7,980E-04

Abbildung 57: Korrelationsmatrix in SPSS

Im Beispielfall – verschiedene technische Merkmale von Automobilen – lassen sich unmittelbar diverse Korrelationen zwischen den Variablen erkennen. Jede der Variablen korreliert mit mindestens einer anderen Variablen – es können also zunächst einmal keine Variablen aus der Analyse ausgeschlossen werden.

1 Bei der Standardnormalverteilung handelt es sich um eine Normalverteilung eben mit einem Erwartungswert von Null und einer Standardabweichung von Eins. Sie wird daher auf oft als N(0/1)-Verteilung bezeichnet.

2 Eignung der Korrelationsmatrix

Die generelle Eignung der Ausgangsdaten für die Faktorenanalyse spiegelt sich in der Korrelationsmatrix wieder. Eine Überprüfung der Eignung anhand verschiedener Prüfkriterien ist anzuraten. Insgesamt stehen sechs Prüfkriterien zur Auswahl:

- Prüfung der Variablen auf Normalverteilung
- Überprüfung des Signifikanzniveaus der Korrelationen
- Analyse der Struktur der Inversen der Korrelationsmatrix
- Durchführung eines Bartlett-Tests auf Spherizität
- Analyse der Anti-Image-Kovarianz-Matrix
- Überprüfung des Kaiser-Meyer-Olkin-Kriteriums

Nicht alle diese Kriterien (die im wesentlichen das gleiche aussagen) müssen vor der Weiterführung der Analyse zwingend überprüft werden. Anzuraten ist aber die Überprüfung anhand mehr als nur eines Kriteriums. Insbesondere das Signifikanzniveau der Korrelationen und das Kaiser-Meyer-Olkin-Kriterium sollten beachtet werden.

Im Folgenden werden alle Kriterien ausführlicher betrachtet, abgesehen von der Prüfung auf Normalverteilung, die bereits im Rahmen der explorativen Datenanalyse ausführlich dargestellt wurden. Zur Erinnerung: Bei der Prüfung auf Normalverteilung können verschiedene Methoden zum Einsatz kommen:

- Grafische Prüfung anhand eines Histogramms mit Normalverteilungskurve
- Grafische Prüfung anhand eines Q-Q- oder P-P-Diagramms
- Statistische Prüfung mit dem Kolmogorov-Smirnoff-Anpassungstest

3 Signifikanzniveaus der Korrelationen

Der Signifikanzwert gibt die Wahrscheinlichkeit wieder, mit welcher der Marktforscher bei der Annahme einer zuvor formulierten Nullhypothese einen Irrtum begeht. Die Nullhypo-

these H0 besagt in diesem Fall, dass in der Grundgesamtheit kein Zusammenhang zwischen den Variablen existiert (der Bravais-Pearson-Korrelationskoeffizient r liegt bei Null). Ergeben sich sehr niedrige Signifikanzwerte nahe Null bedeutet dies, dass der Marktforscher mit einer Wahrscheinlichkeit von nahezu 0% einen Fehler begeht, wenn er die Nullhypothese verwirft – in einem solchen Fall ist also von einem signifikanten Zusammenhang zwischen den Variablen auszugehen.

Korrelationsmatrix[a]

		Hubraum (cu. inches)	PS	Gewicht (lbs.)	Beschleunigung von 0 auf 100 km/h (sec.)	Anzahl der Zylinder
Korrelation	Hubraum (cu. inches)	1,000	,899	,934	-,562	,952
	PS	,899	1,000	,865	-,709	,844
	Gewicht (lbs.)	,934	,865	1,000	-,438	,896
	Beschleunigung von 0 auf 100 km/h (sec.)	-,562	-,709	-,438	1,000	-,528
	Anzahl der Zylinder	,952	,844	,896	-,528	1,000
Signifikanz (1-seitig)	Hubraum (cu. inches)		,000	,000	,000	,000
	PS	,000		,000	,000	,000
	Gewicht (lbs.)	,000	,000		,000	,000
	Beschleunigung von 0 auf 100 km/h (sec.)	,000	,000	,000		,000
	Anzahl der Zylinder	,000	,000	,000	,000	

a. Determinante = 7,980E-04

Abbildung 58: Korrelationsmatrix in SPSS (2)

Für unseren Beispielfall ist festzustellen, dass sich sämtliche Korrelationen zwischen den Variablen mit einer Wahrscheinlichkeit von nahezu 100% von Null unterscheiden. Es spricht daher nichts dagegen, alle Variablen in die weitere Faktorenanalyse zu übernehmen.

4 Struktur der Inversen der Korrelationsmatrix

Die Eignung einer Korrelationsmatrix für die Faktorenanalyse lässt sich auch an der Struktur der Inversen erkennen. Es ist davon auszugehen, dass die Daten dann für die weitere Analyse geeignet sind, wenn die Inversen eine Diagonalmatrix bilden, also die nicht-diagonalen Elemente der inversen Korrelationsmatrix relativ nahe bei Null liegen.

Dabei ist zu beachten, dass kein mathematisches Kriterium dafür existiert, wie stark oder wie häufig die nicht-diagonalen Elemente von Null abweichen dürfen, ohne dass die Eignung

der Daten für die weitere Analyse in Frage gestellt werden muss – diese Entscheidung bleibt somit dem Marktforscher überlassen.

Inverse Korrelationsmatrix

	Hubraum (cu. inches)	PS	Gewicht (lbs.)	Beschleunigung von 0 auf 100 km/h (sec.)	Anzahl der Zylinder
Hubraum (cu. inches)	20,440	-3,978	-6,242	,549	-10,218
PS	-3,978	9,396	-4,174	3,306	1,347
Gewicht (lbs.)	-6,242	-4,174	10,403	-2,551	-1,208
Beschleunigung von 0 auf 100 km/h (sec.)	,549	3,306	-2,551	2,764	,434
Anzahl der Zylinder	-10,218	1,347	-1,208	,434	10,903

Abbildung 59: Struktur der Inversen der Korrelationsmatrix in SPSS

In unserem Beispielfall zeigen sich diverse mehr oder weniger starke Abweichungen von Null. Sie sind aber weder auffallend groß, noch lässt sich irgendein Muster in den Abweichungen erkennen, so dass die Interpretation der Inversen hier unklar ausfällt. In solchen Fällen kann nur dazu geraten werden, noch andere Kriterien für die Eignung der Korrelationsmatrix zu konsultieren.

5 Bartlett-Test auf Spherizität

Mittels des Bartlett-Test auf Spherizität (test of sphericity) wird die Nullhypothese H0 überprüft, dass alle (!) Variablen der Grundgesamtheit, aus der die untersuchte Stichprobe stammt, untereinander unkorreliert sind.

Dies würde implizieren, dass sich die in der Korrelationsmatrix erkennbaren Korrelationen allesamt auf Zufallseffekte bei der Stichprobenziehung zurückführen lassen, während in der Grundgesamtheit kein „realer“ Zusammenhang zwischen den Variablen besteht. Träfe diese Nullhypothese zu, wäre der Datensatz für eine Faktorenanalyse vollkommen ungeeignet, da er, wenn er keine unkorrelierten Variablen enthält, sicher auch keine Variablen enthält, die sich in irgendeiner Form auf gemeinsame Hintergrundfaktoren zurückführen lassen könnten.

Der Bartlett-Test fußt auf zwei Voraussetzungen:

- Die untersuchten Variablen sind in der Grundgesamtheit normalverteilt
- Die Prüfgröße folgt näherungsweise einer Chi²-Verteilung

Die erste Voraussetzungen wurde bereits eingangs erwähnt, und gilt als allgemein sinnvoll für Daten, die mittels einer Faktorenanalyse untersucht werden sollen. Sie kann anhand diverser bereits ausführlich dargestellter Methoden überprüft werden (Histogramm, Kolmogorov-Smirnoff...). Die zweite Voraussetzung kann durch den Test selbst festgelegt werden (indem die Prüfgröße einfach der Chi²-Verteilung entnommen wird), sie bringt aber mit sich, dass der Wert der Prüfgröße auch von der Stichprobengröße abhängig ist – dies ist bei der Interpretation der Testergebnisse zu beachten.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,793
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	2821,259
	df	10
	Signifikanz nach Bartlett	,000

Abbildung 60: Bartlett-Test auf Spherizität in SPSS

SPSS gibt die Ergebnisse des Bartlett-Tests gleichzeitig mit dem Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin aus, welches weiter unten noch im Detail betrachtet wird. In unserem Beispielfall kann der mit 2821,259 sehr hohe Chi²-Wert mit einer Wahrscheinlichkeit von nahezu 100% nur dann zustandekommen, wenn mindestens zwei (aber nicht zwangsweise alle) Variablen in der Grundgesamtheit auch miteinander korrelieren.

Wichtig: Der Bartlett-Test erlaubt keinerlei Rückschlüsse auf die Signifikanz der einzelnen Korrelationen. Ein hoher Chi²-Wert ist keineswegs dahingehend zu interpretieren, dass alle Korrelationen, die der Korrelationsmatrix entnommen werden können, auch in der Grundgesamtheit signifikant werden. Um dies zu überprüfen, ist für jeden Korrelationskoeffizienten

ein eigener Signifikanztest durchzuführen – wie bereits oben im Zusammenhang mit dem Signifikanzniveau der Korrelation gezeigt.

6 Anti-Image-Kovarianz-Matrix

Dem Anti-Image liegt folgende Idee (nach Guttmann) zugrunde: Wenn zwei Variablen miteinander korrelieren, lässt sich die Varianz jeder der beiden Variablen wenigstens teilweise durch die andere Variable erklären. Je stärker diese Korrelation ist, desto größer ist der Anteil an Varianz, der durch die Korrelation erklärt werden kann. Solange der Zusammenhang zwischen beiden Variablen aber nicht perfekt ist, gibt es auch immer noch einen unerklärbaren Varianzanteil. Nach dieser Logik lässt sich die Gesamtvarianz einer korrelierenden Variablen also aufteilen in:

- einen durch die korrelierende Variable erklärbaren Teil (das Image)
- einen durch die korrelierende Variable nicht erklärbaren Teil (das Anti-Image)

Schlussfolgerung: Ein Variablenpaar mit einem niedrigen Anti-Image-Wert weist eine starke Korrelation auf.

Bei der Faktorenanalyse ist zu beachten, dass stets mehr als zwei Variablen betrachtet werden[1] und jede dieser Variablen mit jeder anderen Variablen im Datensatz korrelieren kann. Daher sind nicht die einfachen Korrelationen (wie in der Korrelationsmatrix) sondern die partiellen Korrelationen zu beachten. Eine partielle Korrelation ist die Korrelation zwischen zwei Variablen bei Ausschaltung aller anderen Variablen.

Das Anti-Image eines Variablenpaars lässt sich in diesem Zusammenhang also begreifen als der Teil der Varianz einer Variablen, der sich nicht durch die korrelierende Variable erklären lässt, wenn zugleich der Einfluss aller übrigen Variablen ausgeschaltet wird. Variablenpaare sind dann für die Faktorenanalyse geeignet, wenn ihre Anti-Image-Werte möglichst gering ausfallen.

1 Die Reduktion von nur zwei oder auch drei oder vier Variablen auf Hintergrundfaktoren würde definitiv nur wenig Sinn ergeben. Ziel der Faktorenanalyse ist immerhin die Reduktion der Komplexität.

Idealerweise ergibt sich für die Anti-Image-Kovarianz-Matrix in SPSS sogar eine Diagonalmatrix, wobei in der Realität und vor allem bei Vorliegen von Daten aus einer Zufallsstichprobe mit einer perfekten Diagonalmatrix kaum gerechnet werden kann. Es stellt sich daher die Frage, wann das Kriterium der Diagonalmatrix zumindest näherungsweise erfüllt ist. Dziuban & Shirkey schlagen vor, dass der Anteil an nicht-diagonalen Elementen ungleich Null in jedem Fall unter 25% liegen sollte, wobei ungleich Null als > 0,09 definiert wird.

Anti-Image-Matrizen

		Hubraum (cu. inches)	PS	Gewicht (lbs.)	Beschleunigung von 0 auf 100 km/h (sec.)	Anzahl der Zylinder
Anti-Image-Kovarianz	Hubraum (cu. inches)	4,892E-02	-2,07E-02	-2,935E-02	9,714E-03	-4,585E-02
	PS	-2,071E-02	,106	-4,270E-02	,127	1,315E-02
	Gewicht (lbs.)	-2,935E-02	-4,27E-02	9,613E-02	-8,870E-02	-1,065E-02
	Beschleunigung von 0 auf 100 km/h (sec.)	9,714E-03	,127	-8,870E-02	,362	1,441E-02
	Anzahl der Zylinder	-4,585E-02	1,315E-02	-1,065E-02	1,441E-02	9,172E-02
Anti-Image-Korrelation	Hubraum (cu. inches)	,797[a]	-,287	-,428	7,301E-02	-,684
	PS	-,287	,799[a]	-,422	,649	,133
	Gewicht (lbs.)	-,428	-,422	,813[a]	-,476	-,113
	Beschleunigung von 0 auf 100 km/h (sec.)	7,301E-02	,649	-,476	,662[a]	7,908E-02
	Anzahl der Zylinder	-,684	,133	-,113	7,908E-02	,842[a]

a. Maß der Stichprobeneignung

Abbildung 61: Anti-Image-Kovarianz-Matrix in SPSS

Wichtig: Bei SPSS werden in der Anti-Image-Kovarianz-Matrix nicht die partiellen Korrelationskoeffizienten, sondern deren invertierte negative Werte ausgewiesen.

7 Kaiser-Meyer-Olkin-Kriterium

Kaiser, Meyer & Olkin entwickelten auf der Basis der oben dargestellten Anti-Image-Kovarianz-Matrix eine leicht zu interpretierende Prüfgröße, mit der das Problem der komplexen Beurteilung der Matrix nach Dziuban & Shirkey umgangen wird. Diese Prüfgröße wird als KMO-Kriterium oder MSA (measure of sampling adequacy) bezeichnet.

Das KMO-Kriterium gibt an, in welchem Umfang die Variablen in der Grundgesamtheit miteinander korrelieren. Es ist somit ein geeigneter Indikator dafür, ob mit der Faktorenanalyse fortgefahren werden sollte oder nicht. Der KMO-Wert liegt dabei stets zwischen 0 und 1.

Er kann für einzelne Variablenpaare ebenso wie für die gesamte Korrelationsmatrix berechnet werden. Zur Interpretation des KMO-Werts schlagen Kaiser und Rice das folgende Schema vor:

MSA >= 0,9	Marvelous („erstaunlich“)
MSA >= 0,8	Meritorious („verdienstvoll“)
MSA >= 0,7	Middling („ziemlich gut“)
MSA >= 0,6	Mediocre („mittelmäßig“)
MSA >= 0,5	Miserable („kläglich“)
MSA < 0,5	Unacceptable („untragbar“)

Tabelle 5: Interpretation des Kaiser-Meyer-Olkin-Kriteriums

Eine Korrelationsmatrix als Ganzes ist also dann für eine Faktorenanalyse geeignet, wenn der KMO-Wert wenigstens oberhalb von 0,5 liegt, wobei Werte oberhalb von 0,8 auf jeden Fall wünschenswert sind.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,793
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	2821,259
	df	10
	Signifikanz nach Bartlett	,000

Abbildung 62: Bartlett-Test auf Spherizität in SPSS

In unserem Beispielfall erhalten wir einen KMO-Wert von knapp 0,8, können also die Faktorenanalyse problemlos fortsetzen.

III Faktorextraktion

1 Fundamentaltheorem der Faktorenanalyse

Nachdem die Eignung der Variablen für die Faktorenanalyse im vorangegangenen Schritt bestätigt wurde, werden nun im nächsten Schritt die Faktoren extrahiert. Die Grundannahme hinter der Extraktion, wie auch hinter der gesamten Faktorenanalyse, ist dabei, dass jeder Wert einer Ausgangsvariablen sich als Linearkombination hypothetischer Faktoren beschreiben lässt. Dieser Zusammenhang lässt sich mathematisch so formulieren:

$$X_{kj} = a_{j1} * p_{k1} + a_{j2} * p_{k2} + \ldots + a_{jQ} * p_{kQ}$$

Standardisiert man diese Werte, lautet dieser Ausdruck wie folgt:

$$Z_{kj} = a_{j1} * p_{k1} + a_{j2} * p_{k2} + \ldots + a_{jQ} * p_{kQ} = \sum_{q=1}^{Q} a_{jq} * p_{kq}$$

Die Faktorladung zeigt dabei die Stärke des Zusammenhangs zwischen Faktor und Variablen. In Matrixschreibweise lässt sich dieser standardisierte Ausdruck formulieren als:

$$Z = P * A'$$

Die standardisierte Datenmatrix Z wird also, wie eingangs angeführt, als Linearkombination verschiedener Faktoren dargestellt. Ausgehend von dieser mathematisch ausgedrückten Grundannahme lässt sich nun das Fundamentaltheorem von Thurstone herleiten:

$$R = A * C * A'$$

Dieser Ausdruck lässt sich für untereinander unabhängige Faktoren vereinfachen zu:

$$R = A * A'$$

Die Grundaussage dieses Fundamentaltheorems lässt sich mit Worten wie folgt formulieren: Die Korrelationsmatrix R lässt sich in ihrer Gänze durch die Faktorladungen A und die Korrelationen zwischen den Faktoren C reproduzieren. Für voneinander unabhängige (also unkorrelierte) Faktoren entspricht C einer Einheitsmatrix. Da die Multiplikation einer Matrix mit einer Einheitsmatrix immer zur Ausgangsmatrix zurückführt, kann an dieser Stelle gekürzt werden – es ergibt sich die vereinfachte Form des Fundamentaltheorems.

Wichtig: Die vereinfachte Form des Fundamentaltheorems setzt sowohl die Unabhängigkeit der Faktoren untereinander als auch das Vorhandensein besagter Linearverknüpfung voraus.

Fassen wir also noch einmal zusammen. Das vereinfachte Fundamentaltheorem lässt sich wie folgt herleiten: Wir gehen aus von der Grundannahme der Faktorenanalyse:

$$X_{kj} = a_{j1} * p_{k1} + a_{j2} * p_{k2} + \ldots + a_{jQ} * p_{kQ}$$

Bei standardisierten Werten stellt sich dieser Ausdruck wie folgt dar:

$$Z_{kj} = a_{j1} * p_{k1} + a_{j2} * p_{k2} + \ldots + a_{jQ} * p_{kQ} = \sum_{q=1}^{Q} a_{jq} * p_{kq}$$

Reduziert man diese Annahme auf Matrix-Schreibweise, so ergibt sich: $Z = P * A'$

Die Korrelationsmatrix R lässt sich bei standardisierten Daten aus der Datenmatrix Z ermitteln und stellt sich wie folgt dar:

$$R = \frac{1}{(K-1)} * Z' * Z$$

Da Z im Rahmen der Faktorenanalyse als Linearkombination durch P * A' beschrieben wird, lässt sich dieser Ausdruck wie folgt umstellen:

$$R = \frac{1}{(K-1)} * (P * A')' * (P * A')$$

Lösen wir nun die Klammern auf und folgen den Multiplikationsregeln für Matrizen, so erhalten wir folgenden Term:

$$R = A * (\frac{1}{(K-1)}) * P' * P * A'$$

Da die Daten ja bereits im zweiten Schritt dieser Herleitung standardisiert wurden, lässt sich der Teilausdruck

$$\frac{1}{(K-1)} * P' * P$$

auch als Korrelationsmatrix C der Faktoren bezeichnen. Substituiert man C für den angesprochenen Teilausdruck, so erhält man das Fundamentaltheorem von Thurstone:

$$R = A * C * A'$$

Nimmt man an, dass die Faktoren allesamt untereinander unkorreliert sind, so entspricht die Korrelationsmatrix der Faktoren einer Einheitsmatrix (ausschließlich Null-Korrelationen und eine Diagonale mit perfekten Eigenkorrelationen A auf A etc.). Da die Multiplikation einer Matrix mit der Einheitsmatrix immer wieder zu dieser Matrix führt, kann der Ausdruck noch etwas gekürzt werden, und wir erhalten:

$$R = A * A'$$

2 Grafische Interpretation der Faktoren

Der Informationsgehalt einer Korrelationsmatrix lässt sich auch grafisch in einem Vektor-Diagramm darstellen. Aus der Interpretation solcher Diagramme sollte noch bekannt sein, dass zwei Vektoren dann als linear unabhängig voneinander gelten, wenn sie im Diagramm senkrecht (orthogonal) zueinander stehen. Sind die Vektoren (und damit die hinter den Vektoren stehenden Variablen) dagegen auf irgendeine Weise miteinander korreliert, so drückt sich dies grafisch in einem Winkel aus.

Dazu ein Beispiel: Eine lineare Korrelation mit einem Bravais-Pearson-Korrelationskoeffizienten von r = 0,5 drückt sich im Vektor-Diagramm als ein Winkel von genau 60° aus. Wie lässt sich dieser Winkel berechnen? Es zeigt sich, dass der grafische Ausdruck für einen Korrelationskoeffizienten der Cosinus ist, und der Cosinus eines 60°-Winkels exakt 0,5 beträgt. Analog zu dieser Überlegung lassen sich nun auch die Winkel für zwei charakteristische Sonderfälle in der Korrelationsmatrix ermitteln: Die perfekte Korrelation (r = 1,0) und die perfekte Unabhängigkeit (r = 0,0).

Der Cosinus eines 90°-Winkels beträgt exakt 0,0. Aus diesem Grund stehen voneinander unabhängige Vektoren, wie bereits eingangs erwähnt, immer senkrecht zueinander. Ein gegenteiliges Bild ergibt sich bei der perfekten Korrelation: Der Cosinus eines 0°-Winkels beträgt genau 1,0 – Vektoren, die einen perfekten linearen Zusammenhang aufweisen, liegen daher unmittelbar übereinander.

Einige Korrelationskoeffizienten/Cosinus-Werte und die dazugehörigen Gradangaben finden sich in der nachfolgenden Tabelle.

Grad	***Cosinus/r***	***Grad***	***Cosinus/r***
0	1,00	10	0,98
20	0,94	30	0,87
40	0,77	50	0,64
60	0,50	70	0,34
80	0,17	90	0,00

Tabelle 6: Korrelationskoeffizienten und Lage im Vektor-Diagramm

Grundsätzlich gilt natürlich, dass die Zahl der zur Darstellung benötigten Dimensionen unmittelbar an die Zahl der Variablen im Datensatz gebunden ist: Je mehr Variablen im Datensatz vorliegen, desto mehr Dimensionen werden für die grafische Darstellung benötigt.

Das Ziel der Faktorenanalyse ist, wie bereits oben erläutert, die Reduktion der Komplexität. Bezogen auf die grafische Darstellung mittels der Vektoren bedeutet dies, dass das durch die Korrelationskoeffizienten wiedergegebene Verhältnis der Variablen untereinander in einem Raum mit möglichst wenig Dimensionen dargestellt werden soll. Die Zahl der benötigten Achsen dieser Darstellung entspricht dann der Zahl der gefundenen Faktoren.

Dazu ein Beispiel: Zwei Variablen, die eine Korrelation von r = 0,5 aufweisen, stehen sich als Vektoren A und B im 60°-Winkel gegenüber.

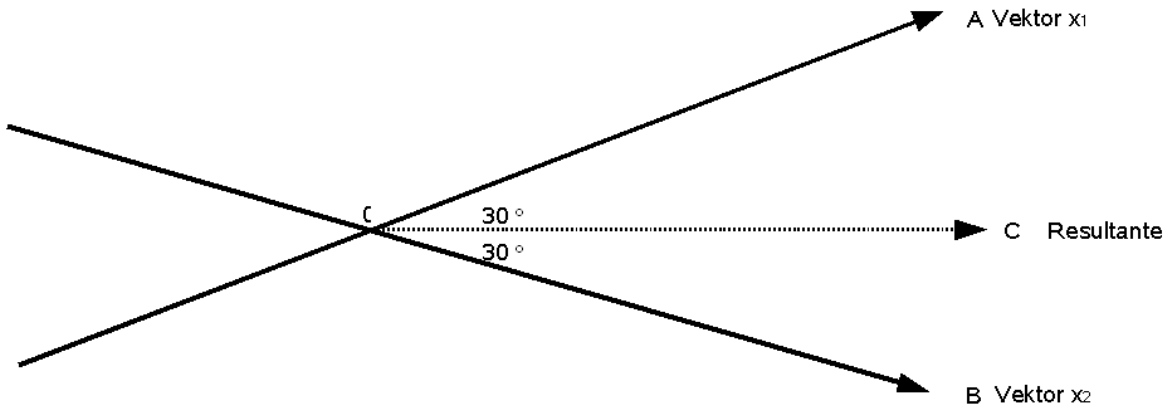

Abbildung 63: Zwei Variablen als Vektoren im Vektor-Diagramm

Der eingezeichnete zusätzliche Vektor C – die sogenannte Resultante – ist eine zusammenfassende (faktorielle) Beschreibung der beiden anderen Vektoren. Die beiden neu entstehenden 30°-Winkel geben den Zusammenhang zwischen der Resultante – unserem Faktor – und den beiden Ausgangsvariablen wieder. Sie repräsentieren gleichwohl die Korrelationskoeffizienten zwischen den Variablen und dem Faktor. Diese Korrelationskoeffizienten werden auch als Faktorladungen bezeichnet (cos 30° = 0,87), und werden später noch eine bedeutende

Rolle bei der Interpretation des gefundenen Modells spielen.

Wie lassen sich nun aber Vektoren (Faktoren) finden, die zusammenfassend für die übrigen Vektoren (Variablen) stehen? Die Bildung des ersten Faktors ist relativ einfach: Er ergibt sich aus dem Schwerpunkt aller durch die Variablen gebildeten Vektoren. Da eine der Voraussetzungen des hier betrachteten Faktorenmodells ist, dass die Faktoren voneinander unabhängig sein sollen, also nicht miteinander korrelieren dürfen, und da völlige Unabhängigkeit, wie bereits festgestellt, sich in einer senkrechten Stellung der Vektoren zueinander manifestiert, steht der zweite Faktor dann genau rechtwinklig zum ersten Faktor, der dritte wieder rechtwinklig zu beiden etc. Erklären die extrahierten Faktoren die Variablen restlos, ist die Summe der Ladungsquadrate jeder Variablen gleich Eins.

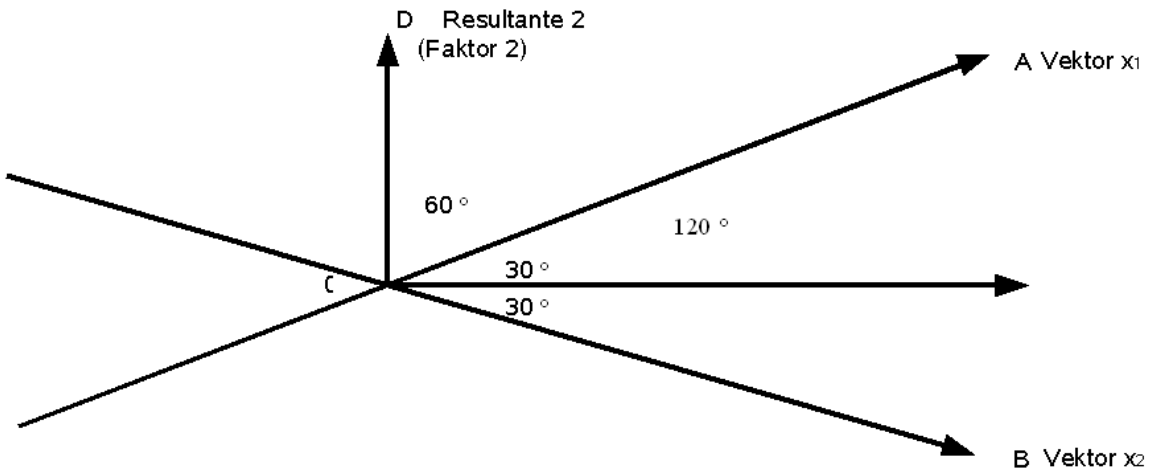

Abbildung 64: Konstruktion eines zweiten, unabhängigen Faktors

3 Bestimmung der Kommunalitäten

Das erklärte Ziel der Faktorenanalyse ist es, die vorhandenen Variablen auf eine geringere Zahl von Faktoren aufzuteilen. Es ist niemandem damit gedient, wenn 8 Variablen auf 8 Faktoren geladen werden, wenn aber 8 Variablen auf 3 Faktoren geladen werden, dann liegt bereits eine deutliche Reduktion der Komplexität vor.

In der Praxis tritt nun das Problem auf, dass nicht die gesamte Varianz der Variablen durch die extrahierten Faktoren erklärt wird, wenn man deren Zahl einschränkt (siehe auch Ausführungen zum Zielkonflikt der Faktorenanalyse). Es verbleibt eine Restvarianz, die durch andere, nicht extrahierte Faktoren oder auch durch Messfehler und Zufallseffekte verursacht wird. Dabei gilt: Je mehr Faktoren im Modell extrahiert werden, desto mehr Varianz wird insgesamt durch diese Faktoren erklärt. Der Teil der Gesamtvarianz der durch alle extrahierten Faktoren erklärt wird, wird in der Faktorenanalyse als Kommunalität bezeichnet.

Das (gekürzte) Fundamentaltheorem ist aufgrund dieser unbestimmbaren Einflüsse um eine Unbekannten-Komponente U zu erweitern:

$$R = A * A' + U$$

In U fließen sowohl die spezifische Varianz als auch die potentiellen Messfehler ein (die sogenannten Einzelrestfaktoren).

An dieser Stelle kommt der Marktforscher ins Spiel, der die Kommunalität, und damit auch den Anteil der nicht erklärbaren Varianz, selbst schätzen muss. Legt er beispielsweise eine Kommunalität von 0,7 fest, so bedeutet dies, dass er vermutet, dass insgesamt 70% der Ausgangsvarianz durch gemeinsame Faktoren erklärt werden können.

Neben fachlichen Überlegungen spielt vor allem die Variablenanzahl bei der Schätzung der Kommunalitäten eine große Rolle. Je größer nämlich die Anzahl der Variablen im Modell ist, umso unwichtiger ist die exakte Schätzung der Kommunalitäten. Der Grund dafür ist, dass bei einer steigenden Anzahl an Variablen der prozentuale Anteil der diagonalen Matrixelemente in der Korrelationsmatrix immer weiter abnimmt. In einer 2x2-Matrix machen diese diagonalen Elemente noch 50% aus, in einer 100x100-Matrix dagegen nur noch 1%. Wie man sich leicht vorstellen kann, hat eine fehlerhafte Einschätzung im letzteren Fall wesentlich geringere negative Auswirkungen als im ersten.

Wie kann der Marktforscher nun aber zu einem Schätzwert für die Kommunalität gelangen? In der Praxis sind heute vor allem zwei Verfahren der Kommunalitätenschätzung von

Bedeutung.

Erste Möglichkeit: Der Marktforscher geht einfach davon aus, dass die gesamte Varianz aller Ausgangsvariablen durch die Faktoren erklärt werden kann. In diesem Fall ist die Summe der Kommunalitäten stets gleich Eins, da keine Einzelrestfaktoren auftreten. Eine explizite Schätzung der Kommunalitäten im Rahmen der Faktorenanalyse findet also in diesem Fall gar nicht statt.

Zweite Möglichkeit: Aufgrund verschiedener inhaltlicher Überlegungen wird ein Schätzwert für die Kommunalitäten vorgegeben. Der Vorgabewert für diese Schätzung ist häufig der höchste quadrierte Korrelationskoeffizient aus der Korrelationsmatrix. Grund dafür ist, dass die Faktoren in ihrer Gesamtheit mindestens den gleichen Erklärungsbeitrag liefern wie die höchste vorgefundene Korrelation, meist liefern sie jedoch deutlich mehr. Wird dieser Wert zur Schätzung der Kommunalitäten verwendet, fällt diese daher in der Regel zu niedrig aus – es handelt sich also um ein konservatives Schätzverfahren. Abweichende Schätzverfahren sind aber ebenfalls denkbar.

Die Art der Kommunalitätenschätzung wirkt sich unmittelbar auf die Wahl des Faktorextraktionsverfahrens aus. Auch hier wird in zwei wesentliche Verfahren unterschieden: Die Hauptachsenanalyse, bei der sich die Varianz stets in Kommunalitäten und Einzelrestvarianz aufteilt und die Hauptkomponentenanalyse, bei der die Varianz vollständig durch die Faktoren erklärt wird. Die Wahl des Faktorextraktionsverfahrens beeinflusst wiederum die Interpretation der gewonnenen Ergebnisse.

4 Die Hauptachsenanalyse

Der Hauptachsenanalyse liegt die Annahme zugrunde, dass sich die Varianz jeder Ausgangsvariablen stets in Kommunalitäten und Einzelrestvarianz aufteilt. Der Marktforscher muss hier also eine Schätzung bezüglich der Höhe der Kommunalitäten abgeben. Diese kann entweder auf inhaltlichen und fachlichen Überlegungen basieren oder sich aus einem Iterationsprozess ergeben, der ebenfalls Bestandteil der Hauptachsenanalyse ist.

Das Ziel der Hauptachsenanalyse ist die inhaltliche Erklärung der Varianzen der Variablen in Höhe der Kommunalitäten durch die Faktoren. Aus diesem Grund ist die Hauptachsenanalyse das richtige Verfahren, wenn die inhaltliche Interpretation der Faktoren im Vordergrund steht – also immer dann, wenn eine kausale Interpretation gefragt ist.

Die entscheidende Frage der Hauptachsenanalyse lautet: Wie lässt sich die Ursache bezeichnen, die für hohe Ladungen der Variablen auf diesen Faktor verantwortlich ist?

5 Die Hauptkomponentenanalyse

Der Hauptkomponentenanalyse liegt die Annahme zugrunde, dass die Varianz jeder Ausgangsvariablen vollständig durch die Faktoren erklärt werden kann, die Kommunalität also bei Eins liegt. Wie wir aus den vorangegangenen Betrachtungen wissen, liegt die Kommunalität dann bei Eins, wenn genauso viele Faktoren extrahiert werden, wie Variablen im Modell sind – dann allerdings hat die Faktorenanalyse keinerlei Sinn mehr. Werden dagegen weniger Faktoren extrahiert, sinkt auch die Kommunalität.

Im bewussten Verzicht auf Informationen zur Herbeiführung eines brauchbaren Modells spiegelt sich der bereits oben dargestellte Zielkonflikt der Faktorenanalyse wieder. Das Ziel der Hauptkomponentenanalyse ist demzufolge auch die möglichst umfassende Reproduktion der Zusammenhänge im Datensatz mit einer möglichst geringen Anzahl von Faktoren.

Aus diesem Grund wird in der Hauptkomponentenanalyse nicht in Kommunalitäten und Einzelrestvarianzen unterschieden, die Interpretation der Faktoren kann demzufolge dann auch nicht mehr kausal interpretiert werden.

Die entscheidende Frage der Hauptkomponentenanalyse lautet daher: Wie lassen sich die auf einen Faktor hochlandenden Variablen durch einen Sammelbegriff (Komponente) zusammenfassen?

Die Vorgehensweise der Hauptkomponentenanalyse wurde oben bereits kurz umrissen: Zunächst wird der erste Faktor (auch als erste Hauptkomponente bezeichnet) so bestimmt, dass durch ihn ein möglichst großer Teil der Gesamtvarianz erklärt wird. Der zweite Faktor wird dann so bestimmt, dass er orthogonal zum ersten Faktor steht (also unkorreliert ist) und gleichzeitig einen möglichst großen Teil der verbliebenen Restvarianz erklärt. Auf diese weise lassen sich theoretisch so lange Faktoren ziehen, bis ein Faktor auf jede beobachtete Variable kommt, wobei in diesem Fall auch die Gesamtvarianz vollständig erklärt werden würde.

Werden also n Variablen durch n Faktoren dargestellt, kann die Varianz komplett aufgeklärt werden. Dies liefe aber, wie bereits mehrfach erläutert, dem eigentlichen Ziel der Varianzanalyse zuwider. Werden dagegen weniger als n Faktoren extrahiert, wird ein Teil der Varianz nicht durch das Modell erklärt. Volkswirtschaftlich betrachtet liegt hier also eine Trade-off-Situation zwischen dem Grad der Dimensionsreduktion und der Genauigkeit des Modells vor – dieser Aspekt wurde bereits unter dem Stichwort des Zielkonflikts der Faktorenanalyse im Detail beleuchtet.

Es ist an dieser Stelle dem Marktforscher überlassen zu entscheiden, welche Faktoren in das Modell aufgenommen werden sollen und welche ausgeschlossen werden können. Dabei erscheint es logisch, solche Faktoren mit einem hohen Erklärungsgehalt aufzunehmen und solche Faktoren mit einem niedrigen Erklärungsgehalt auszuschließen. Hier bieten sich verschiedene Entscheidungskriterien an, die nachfolgend näher betrachtet werden sollen.

6 Bestimmung der Faktoranzahl

Zur Bestimmung der Anzahl der zu extrahierenden Faktoren existieren keine allgemeinverbindlichen Vorschriften. Ein Stück weit ist hier also die subjektive Entscheidung des Marktforschers gefragt, der aber auf sechs verschiedene Entscheidungskriterien zurückgreifen kann:

- Fortsetzung der Extraktion, bis xy% der Varianz erklärt sind (vorher festlegen!)
- Fortsetzung der Extraktion solange Anzahl der Faktoren < halbe Variablenanzahl

- Extraktion aller inhaltlich noch sinnvoll interpretierbaren Faktoren
- Extraktion von genau n Faktoren (Anzahl zuvor nach fachlichen Aspekten bestimmt)
- Kaiser-Kriterium (der Standard in SPSS)
- Screeplot/Scree-Test

Wie man leicht sehen kann, sind nur die letzten beiden Kriterien mathematischer Natur, so dass ausschließlich diese nachfolgend noch näher betrachtet werden sollen. Daraus ist aber nicht der Schluss zu ziehen, dass sie auch „wichtiger“ oder gar „richtiger“ sind. Der Marktforscher kann an dieser Stelle der Faktorenanalyse vielmehr jedes der Kriterien, eine Kombination aus diesen oder auch eine eigene subjektive Entscheidung als Richtlinie für die Bestimmung der Faktoranzahl auswählen.

Kaiser-Kriterium

Das Kaiser-Kriterium ist eines der einfachsten Entscheidungskriterien. Es besagt, dass alle Faktoren mit einem Eigenwert oberhalb von Eins extrahiert und alle anderen Faktoren verworfen werden. Bei den Eigenwerten handelt es sich um die Summe aller quadrierten Faktorladungen eines Faktors über alle Variablen. Sie können als Indikator für die durch den jeweiligen Faktor erklärte kombinierte Varianz aller Variablen betrachtet werden.

Um die Einschätzung dieser Eigenwerte zu erleichtern, werden alle Variablen einer Z-Transformation unterzogen (dies wurde weiter oben schon einmal erläutert). Jede Variable weist dann hinterher einen Erwartungswert von Null und eine Standardabweichung von Eins auf. Daraus folgt, dass die Gesamtstreuung von n Variablen ebenfalls n beträgt.

Ein Faktor mit einem Eigenwert von mehr als Eins erklärt daher mehr als eine ganze Variable und kann extrahiert werden. Dagegen erklärt ein Faktor mit einem Eigenwert von weniger als Eins auch weniger, als eine einzelne Variable erklären würde, würde man sie direkt als Faktor extrahieren. Solche Faktoren tragen nicht wesentlich zum Erklärungsgehalt des Modells bei und können daher nach dem Kaiser-Kriterium ignoriert werden.

Erklärte Gesamtvarianz

Komponente	Anfängliche Eigenwerte			Rotierte Summe der quadrierten Ladungen		
	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %
1	1,741	43,534	43,534	1,642	41,057	41,057
2	1,131	28,283	71,818	1,230	30,760	71,818
3	,580	14,489	86,306			
4	,548	13,694	100,000			

Extraktionsmethode: Hauptkomponentenanalyse.

Die Eigenwerte der Faktoren werden von SPSS in der Tabelle für die erklärten Gesamtvarianzen ausgegeben. Für unseren Beispielfall zeigt sich, dass die ersten beiden Faktoren extrahiert werden können – der dritte Faktor hat nur noch einen Eigenwert von 0,58 und erklärt damit nicht mehr viel Gesamtvarianz (diese liegt insgesamt bei 4).

Screeplot/Scree-Test

Beim Scree-Test handelt es sich um einen grafischen Test zur Bestimmung der optimalen Faktoranzahl. Die Eigenwerte der möglichen Faktoren werden in einem Diagramm in absteigender Reihenfolge angeordnet und miteinander verbunden, wodurch eine sich asymptotisch der Abszisse annähernde Punktlinie ergibt.

An der Stelle mit der größten Differenz zwischen zwei Eigenwerten ist ein deutlicher Knick in dieser Linie auszumachen (der sogenannte elbow). Der letzte Punkt links dieses Knicks bestimmt die Anzahl der zu extrahierenden Faktoren. Faktoren rechts vom Knick werden als Geröll (scree) bezeichnet und tragen offensichtlich wenig zum Erklärungsgehalt des Modells bei. Die Bezeichnung rührt daher, dass man dem Screeplot eine gewisse Ähnlichkeit mit einem Geländeprofil bei Hanglage nachsagt – dort, wo die Steilheit des Hangs endet und das Geröll und der Schutt beginnen endet synonym auch die Extraktion der Faktoren.

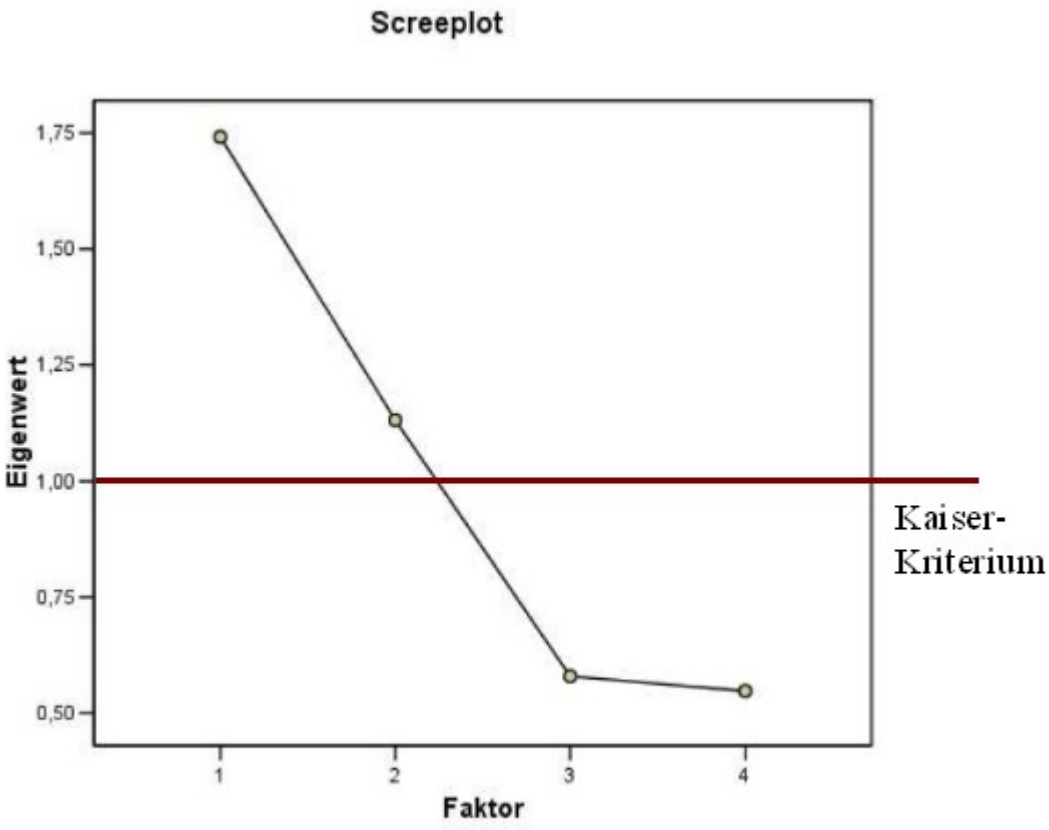

Abbildung 66: Scree-Plot in SPSS

Das Verfahren liefert nicht immer eine eindeutige Lösung, beispielsweise dann, wenn die Differenzen zwischen den Eigenwerten nur gering ausfallen. Eine subjektive Entscheidung des Marktforschers ist hier also nicht zu umgehen. Praxisüblich ist die Ausgabe des Scree-Plots zur visuellen Unterstützung einer Entscheidung nach dem Kaiser-Kriterium, da beide Entscheidungsverfahren in der Regel gleich enden.

Dies ist auch in unserem Beispielfall zu erkennen: Der deutliche Knick liegt zwischen dem zweiten und dem dritten Punkt – es sollten also nur die ersten beiden Faktoren extrahiert werden. Genau diese Schlussfolgerung ergab sich auch bereits aus der Betrachtung des Kaiser-Kriteriums.

IV Interpretation der Faktoladungen und Faktorrotation

1 Interpretation der Faktorladungen

Im Anschluss an die Extraktion der Faktoren sind diese noch entsprechend zu interpretieren. Der erste Schritt hierzu ist die Analyse der zu diesem Zeitpunkt noch unrotierten Faktorladungen.

Bei der Analyse der Faktorladungen ist vor allem zu beachten, dass die inhaltliche Interpretation der Faktoren dem Marktforscher umfassende Sachkenntnis des untersuchten Gegenstands abverlangt – gute Methodenkenntnisse alleine befähigen also noch nicht zur inhaltlichen Interpretation. Zudem sind die Unterschiede zwischen der Hauptachsenanalyse und der Hauptkomponentenanalyse zu beachten: Wurde eine Hauptachsenanalyse durchgeführt, lassen sich die Faktoren kausal interpretieren, wurde dagegen auf das Verfahren der Hauptkomponentenanalyse zurückgegriffen, so ist dies nicht der Fall.

Komponentenmatrix[a]

	Komponente	
	1	2
iq	,793	-,104
sprache	,584	-,650
mathe	,396	,817
englisch	,783	,177

Extraktionsmethode: Hauptkomponentenanalyse.

a. 2 Komponenten extrahiert

Abbildung 67: Komponentenmatrix in SPSS

Weichen wir einmal vom Automobil-Beispiel ab und betrachten die hier abgebildete Komponentenmatrix für verschiedene Testergebnisse aus dem schulischen Bereich. Es ist zu erkennen, dass IQ und Englisch hauptsächlich durch den ersten Faktor erklärt werden und Mathematik durch den zweiten, während Sprache etwa gleich gut auf beide Faktoren lädt. Sie wäre daher auch inhaltlich beiden Faktoren zuzuordnen – dies erschwert eine sinnvolle Interpretation der Faktoren. Um solchen Problemen aus dem Weg zu gehen, können die extra-

hierten Faktoren noch einer Transformation, der sogenannten Faktorrotation unterzogen werden, welche die Interpretation des finalen Ergebnisses teils erheblich erleichtert.

2 Faktorrotation

Wie bereits dargestellt, hat der Marktforscher bei der Interpretation der Faktorladungen teils einen erheblichen Spielraum, da er beispielsweise selbst entscheiden kann, wie stark eine Variable mindestens auf einen Faktor laden kann (oder sollte), um diesem fest zugeordnet zu werden. Als grundsätzliche Regel ließe sich festlegen, dass Faktorladungen ab 0,5 aufwärts automatisch zu einer Zuordnung führen sollen. Wenn aber eine Variable mit jeweils mehr als 0,5 auf mehrere Faktoren lädt, müsste sie dann konsequenterweise auch mehreren Faktoren zugeordnet werden. In solchen Fällen wäre das entstehende Modell nicht mehr sinnvoll zu interpretieren.

Um eine Lösung für dieses Problem zu finden, kehrt man zu den Grundüberlegungen der Faktorenanalyse zurück: Die beobachteten Variablen sind Ausdruck komplexer Hintergrundfaktoren. Die Beziehungen der einzelnen Variablen zu diesen Hintergrundfaktoren zeigt sich dann an den jeweiligen Faktorladungen, wobei große Faktorladungen eine große Bedeutung und geringe Faktorladungen eine geringe Bedeutung des Faktors für die jeweilige Variable anzeigen. Ein Faktor ist immer dann einfach zu interpretieren, wenn die auf ihn ladenden Variablen untereinander homogen sind, er ist dagegen sehr viel schwerer zu interpretieren, wenn er mit sehr vielen oder gar allen Variablen mehr oder weniger stark korreliert (also viele Variablen mehr oder weniger stark auf ihn laden).

Eine solche Situation ist aber, wie wir oben gesehen haben, nach der Extraktion der Faktoren nicht unwahrscheinlich. Die Faktoren können aber einer als Rotation bezeichneten Transformation unterworfen werden, welche deren Interpretation erheblich erleichtert.

Der Grundgedanke der Rotation ergibt sich aus der Darstellung der Faktoren im Vektoren-Diagramm. Rotiert man die Koordinatenachsen dieses Diagramms in ihrem Ursprung, lassen sich die Faktorladungen besser auf die Faktoren verteilen.

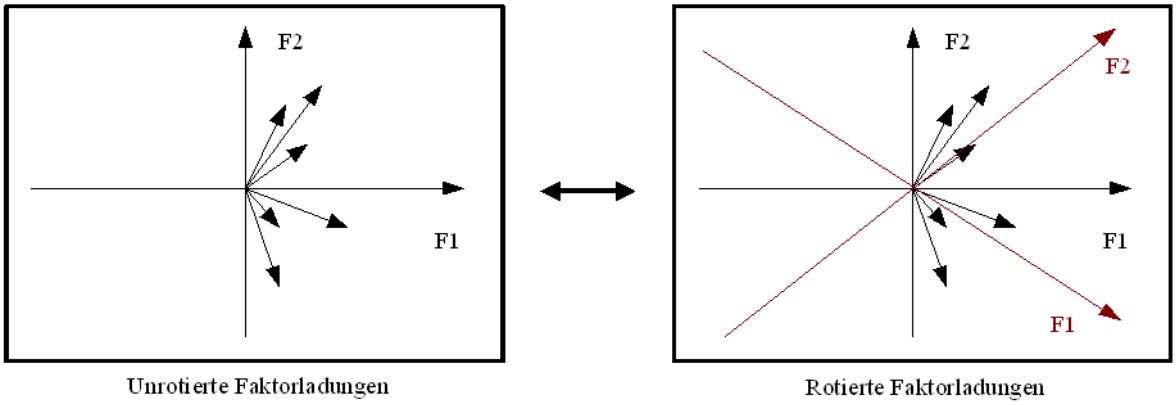

Abbildung 68: Prinzip der Faktorrotation

Dabei kann in zwei Rotationsmethoden unterschieden werden: Die orthogonale (rechtwinklige) Rotation und die oblique (schiefwinklinge) Rotation. Das Ergebnis jeder Rotation ist eine verbesserte Zuordnung der einzelnen Variablen zu den Faktoren. Dabei verändern sich durch die Rotation sowohl die Faktorladungen als auch die Eigenwerte, nicht aber die Kommunalitäten – die Aussagekraft einer Hauptachsenanalyse (und nur hier macht eine Faktorrotation wirklich Sinn, da ja die Faktoren bei der Hauptkomponentenanalyse ohnehin nicht inhaltlich interpretiert werden können) wird durch die Rotation des Koordinatenkreuzes in keinster Weise verzerrt. Alle Formen der Rotation sind daher nicht als Änderungen am Modell sondern lediglich als Nachoptimierungen zu verstehen.

Die Rotation hat sich als so praxistauglich erwiesen, dass in der Praxis zumeist nur noch die rotierte Faktorladungsmatrix überhaupt inhaltlich interpretiert wird.

3 Orthogonale Rotation

Bei der orthogonalen (rechtwinkligen) Rotation geht man davon aus, dass die Faktoren nicht untereinander korrelieren und ihre Vektoren daher stets senkrecht zueinander stehen. Sämtliche Faktorachsen bleiben daher während der Rotation ebenfalls im rechten Winkel zueinander.

SPSS bietet dem Marktforscher drei orthogonale Rotationsmethoden zur Auswahl, wobei

die Varimax-Methode in der Praxis die gebräuchlichste ist:

- Varimax-Methode. Das Ziel dieser Methode ist die Vereinfachung der Interpretation der Faktoren. Die Achsen werden so rotiert, dass sich die Anzahl der Variablen mit hohen multiplen Faktorladungen reduziert.
- Quartimax-Methode. Das Ziel dieser Methode ist die Vereinfachung der Interpretation der Variablen. Die Achsen werden so rotiert, dass jede Variable durch möglichst wenig Faktoren erklärt wird (idealerweise nur durch einen einzigen).
- Equimax-Methode. Diese Methode stellt einen Kompromiss zwischen Varimax- und Quartimax-Methode dar, der in der Praxis aber nur selten zum Einsatz kommt.

4 Oblique Rotation

Bei der obliquen (schiefwinklingen) Rotation wird im Gegensatz zur orthogonalen Rotation davon ausgegangen, dass die Faktoren durchaus untereinander korrelieren. Aus diesem Grund können sich die Winkel zwischen den Faktoren während der Rotation beliebig verschieben. Dies resultiert in einer wesentlich besseren Aufteilung der Faktorladungen auf die Faktoren (da die Rotationsmethode sehr viel mehr Spielraum lässt), untergräbt aber einen der Grundgedanken der Faktorenanalyse: Die Unabhängigkeit der Faktoren voneinander.

SPSS bietet dem Marktforscher zwei oblique Rotationsmethoden zur Auswahl:

- Direktes Oblimin. Auf der Basis von inhaltlichen Überlegungen kann hier der Grad der Schiefwinkligkeit vorgegeben werden. Gebräuchlichste oblique Rotationsmethode.
- Promax. Der optimale Grad an Schiefwinkligkeit wird bei dieser Methode durch ein Iterationsverfahren bestimmt. Es kommt in der Regel nur bei sehr umfangreichen Stichproben zur Anwendung.

V Bestimmung und Interpretation der Faktorwerte

1 Einführung

Im Anschluss an die Extraktion und die Rotation der Faktoren stellt sich im letzten Schritt der Faktorenanalyse noch die Frage, welche Werte die untersuchten Objekte bezüglich der Faktoren annehmen. Die Analyse dieser Fragestellung wird auch als Problem der Bestimmung der Faktorenwerte bezeichnet.

Kehren wir hierzu noch einmal zu unserem Beispielfall zurück, bei dem Personen zu verschiedenen Eigenschaften von Automobilen befragt wurden. Die Faktoranalyse ergab, dass sich die verschiedenen Eigenschaften von Automobilen durch Faktoren darstellen lassen. So gehören PS-Zahl, Drehmoment, Höchstgeschwindigkeit etc. Zum Faktor „Technik“ usw. Interessant ist nun die Frage, wie die verschiedenen Automobiltypen hinsichtlich der Faktoren beurteilt worden: Welche Marke ist stark bei „Sicherheit“, welche bei „Technik"“?

2 Bestimmung der Faktorwerte

Zur Bestimmung der Faktorwerte kehren wir zur Basis-Funktion der Faktorenanalyse zurück:

$$Z = P * A'$$

Zur Bestimmung der Faktorwerte ist diese Gleichung nach P aufzulösen. Dies lässt sich durch die Multiplikation von rechts mit der inversen Matrix erreichen:

$$Z * A'^{-1} = P * A' * A'^{-1}$$

Da $A' * A'^{-1}$ definitionsgemäß die Einheitsmatrix E ergibt, folgt:

$$P = Z * A'^{-1}$$

Diese einfache Inversion ist für nicht-quadratische Faktormuster von A nicht möglich. In diesem Fall ist von rechts mit A zu multiplizieren:

$$Z * A = P * A' * A$$

Die Matrix $A' * A$ ist quadratisch und daher invertierbar:

$$Z * A * (A' * A)^{-1} = P * (A' * A) * (A' * A)^{-1}$$

Da $(A' * A) * (A' * A)^{-1}$ per Definition der Einheitsmatrix entspricht, ergibt sich:

$$P = Z * A * (A' * A)^{-1}$$

Zur Lösung dieser Gleichung sind gegebenenfalls Schätzverfahren anzuwenden (beispielsweise Bartlett oder Anderson-Ruth).

3 Interpretation der Faktorwerte

Faktorwerte können generell positiv oder negativ ausfallen und auch näherungsweise dicht bei Null liegen. Sie werden, wie oben dargestellt, unter Verwendung aller Faktorladungen aus der rotierten Faktorladungsmatrix berechnet. Auch kleine Faktorladungen haben daher einen Einfluss auf die Größe der Faktorwerte.

Die Faktorwerte sind hinsichtlich ihres mathematischen Wertes zu interpretieren:

- Negative Faktorwerte deuten darauf hin, dass das Objekt bezüglich des betrachteten Faktors und im direkten Vergleich mit den anderen Objekten unterdurchschnittlich

ausgeprägt ist.

- Positive Faktorwerte deuten darauf hin, dass das Objekt bezüglich des betrachteten Faktors und im direkten Vergleich mit den anderen Objekten überdurchschnittlich ausgeprägt ist.
- Faktorwerte nahe Null deuten darauf hin, dass das Objekt bezüglich des betrachteten Faktors und im direkten Vergleich mit den anderen Objekten durchschnittlich ausgeprägt ist.

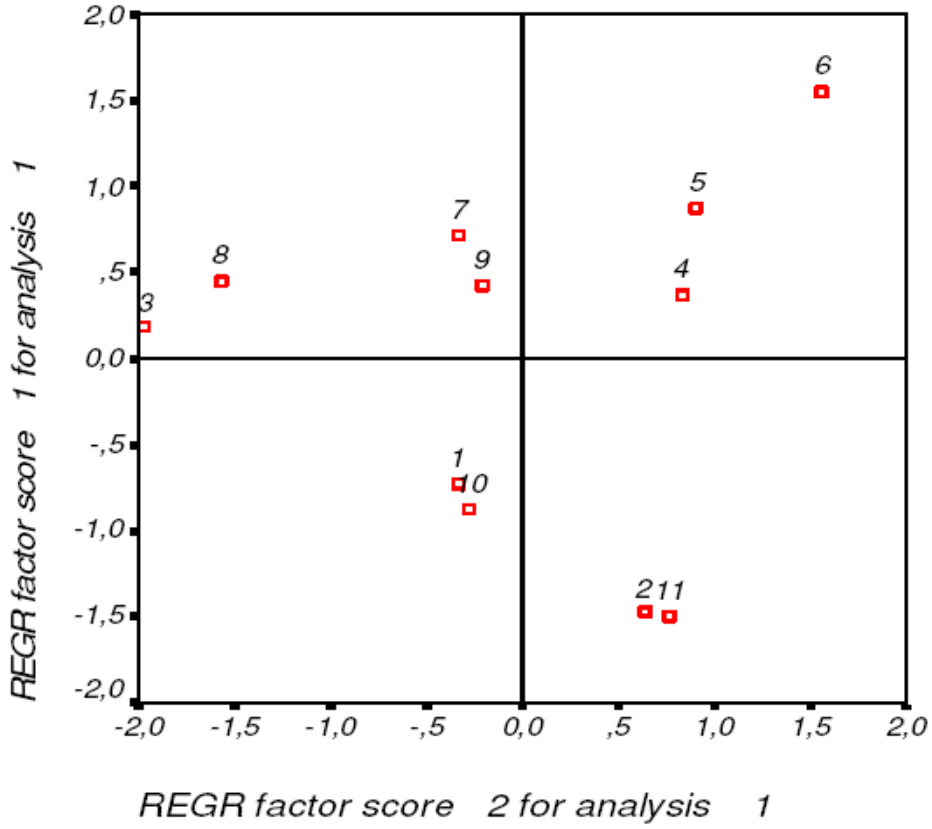

Abbildung 69: Grafische Darstellung der Faktorwerte in SPSS

VI Weiterführende Literatur

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). Multivariate Analysemethoden (10. Aufl.). Berlin: Springer.

Bortz, J. (1999). Statistik für Sozialwissenschaftler (5. Aufl.) Berlin: Springer.

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz

Götze, W., Deutschmann, C. & Link, H. (2002). Statistik. München: Oldenbourg.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). Multivariate data analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

F Weitere Analyseverfahren

I Clusteranalyse

1 Grundlagen

Das Ziel der Clusteranalyse ist die Unterteilung einer Menge von Objekten in Gruppen – die sogenannten Cluster. Die einem Cluster zugeordneten Objekte sollen sich dabei möglichst ähnlich sein (homogen), die unterschiedlichen Clustern zugeordneten Objekte sollen sich dagegen möglichst stark voneinander unterscheiden (heterogen). Die Besonderheit des Verfahrens ist, dass mehrere Merkmale parallel zueinander zur Clusterbildung herangezogen werden können, sich die Betrachtung der „Ähnlichkeit“ oder „Unähnlichkeit“ von Objekten also über mehrere Dimensionen erstreckt.

Genau darin liegt dann auch die theoretische Herausforderung des Verfahrens: Die „Ähnlichkeit“ von Objekten muss genau genug gemessen werden können, um zu einer Einteilung in Cluster zu gelangen. Es wird ein Verfahren benötigt, mit dem sich eine Kombination aus Merkmalsausprägungen für ein Objekt mit einer ebensolchen Kombination für ein anderes Objekt vergleichen lässt.

Das bestmögliche Verfahren hierfür ist die hierarchische Clusteranalyse. Es erfordert nur geringfügige Voraussetzungen und ist daher fast in jeder Situation anwendbar – der Informationsgehalt des Ergebnisses ist dafür aber auch sehr gering, zumindest verglichen mit dem anderer multivariater Analyseverfahren.

2 Methodik

Bei der Clusteranalyse kommt die sogenannte „Methodik des hierarchischen Agglomerierens“ zum Einsatz. Diese Methodik umfasst mehrere Schritte, die mehrfach in einer Schleife durchlaufen werden.

Zunächst wird jedes vorhandene Objekt als einzelner Cluster betrachtet. Die beiden Cluster – auf dieser Stufe also noch die beiden Einzelobjekte – zwischen denen die geringste Distanz besteht – die sich also am ähnlichsten sind – werden miteinander vereinigt. Diese Vereinigung reduziert also die Anzahl der insgesamt vorhandenen Cluster um Eins. Für die noch vorhandenen Cluster werden anschließend erneut alle möglichen Distanzen berechnet und es kommt wieder zu einer Vereinigung der Objekte mit dem geringsten Abstand (gemessen werden muss also die Distanz zwischen zwei Einzelobjekten, die Distanz zwischen zwei Clustern und die Distanz zwischen Clustern und Einzelobjekten). Dieses Verfahren wird so lange fortgesetzt, bis alle Objekte in einem einzigen, großen Cluster vereinigt sind.

Dieser „Megacluster“ ist natürlich nicht als Endergebnis der Clusteranalyse zu betrachten. Statt dessen werden nun die Teilschritte untersucht, die zur Bildung dieses Megaclusters geführt haben. Die Teilstufe mit der am sinnvollsten erscheinenden Clusterung kann dann als Endergebnis selektiert werden, wobei hier viel interpretatorisches Geschick des Marktforschers gefragt ist.

3 Voraussetzungen

Für eine sinnvolle Clusteranalyse muss die Ähnlichkeit zweier Objekte mathematisch quantifizierbar sein. Metrisch skalierte Merkmale können daher problemlos in eine Clusteranalyse einfließen, ordinal und nominal skalierte Merkmale müssen dagegen als Dummy-Variablen codiert werden. Die gleichzeitige Verwendung metrischer und nichtmetrischer Merkmale in einer Clusteranalyse ist gestattet.

Zu bedenken ist auch, dass in unterschiedlichen Dimensionen skalierte Merkmale zu einer Ergebnisverzerrung führen. Wird beispielsweise eine Clusteranalyse mit den Merkmalen Alter und Einkommen durchgeführt, so werden sich bei den Einkommenswerten größere Abstände zwischen den Objekten zeigen. Dies wiederum hätte zur Folge, dass das Einkommen die Wertung der Ähnlichkeit/Unähnlichkeit viel stärker beeinflussen würde als das Alter. Um dieses Problem zu umgehen, können die Werte vor der Durchführung einer Clusteranalyse standardi-

siert werden – in der Regel durch die aus der Statistik II bereits bekannte Z-Transformation. Eine solche Transformation darf aber nicht durchgeführt werden, wenn keine unterschiedlich dimensionierten Werte vorliegen – hier muss der Marktforscher also aufpassen.

Die letzte Voraussetzung besteht darin, dass die zur Clusterbildung genutzten Merkmale nicht untereinander korrelieren sollten. Der Grund dafür ist, dass solche Korrelationen die Bedeutung der betroffenen Merkmalsgruppen verstärken, was leicht eine inhaltliche Fehlinterpretation der gefundenen Cluster nach sich zieht.

4 Distanzmaße

Distanzmaße dienen der Bestimmung der Distanz zwischen zwei Einzelobjekten. Die Grundlage für die Distanzmessung bildet die sogenannte Minkowski-Metrik:

$$d_k, l = \left(\sum_{j=1}^{J} |x_{kj} - x_{lj}|^r\right)^1 / r$$

Die entscheidende Größe in dieser Gleichung ist die Minkowski-Konstante r, mit der die Art der Minkowski-Metrik festgelegt wird. Dabei ist zu beachten, dass es sich bei Minkowski-Metriken grundsätzlich um Unähnlichkeitsmaße handelt, d.h. je größer das Maß, desto unähnlicher sind sich die Objekte.

Für metrische Daten sind vor allem vier Distanzmaße von Bedeutung:

- Euklidischer Abstand (bei r = 2)
- Quadrierter euklidischer Abstand
- Block-Distanz/Manhattan-Distanz (bei r=1)
- Tschebyscheff-Distanz

Für nichtmetrische Daten existieren sechs wesentliche Distanzmaße:

- Dice
- Jaccard
- Tanimoto
- Kulczynski
- Russel and Rao
- Simple Matching

5 Clustermethoden

Die Clustermethoden dienen der Bestimmung der Distanz zwischen zwei Clustern oder einem Cluster und einem Einzelobjekt. Hier sind fünf Methoden üblich, die nachfolgend jeweils kurz betrachtet werden sollen.

Linkage zwischen den Gruppen: Hier werden alle Paare konstruiert, die aus jedem der beiden Cluster je ein Objekt enthalten. Für jedes dieser Paare wird anhand der Distanzmaße auf die übliche Art die Distanz bestimmt. Das arithmetische Mittel aller aufgetretenen Distanzen wird dann als Distanz zwischen den Clustern gewertet.

Linkage innerhalb der Gruppen: Hier werden alle Paare konstruiert, die sich insgesamt aus den Objekten beider Cluster bilden lassen – also auch Paare, bei denen beide Objekte im gleichen Cluster liegen. Analog zum vorangegangenen Verfahren wird wieder das arithmetische Mittel aller aufgetretenen Distanzen als Distanz zwischen den Clustern gewertet.

Nächster Nachbar: Es wird das Paar aus Objekten beider Cluster gesucht, welches die kürzeste Distanz aufweist. Diese Distanz wird dann als Distanz zwischen den Clustern gewertet.

Entferntester Nachbar: Es wird das Paar aus Objekten beider Cluster gesucht, welches die größte Distanz aufweist. Diese Distanz wird dann als Distanz zwischen den Clustern gewertet.

Zentroid-Clustering: Hier entspricht die Distanz zwischen den Clustern der Distanz zwischen den beiden Objekten, die sich aus den jeweiligen arithmetischen Mitteln aller Objekte

in den einzelnen Clustern errechnen.

6 Auswertung und Clusterfindung

Die Ergebnisse einer Clusteranalyse werden dem Marktforscher in Form von zwei tabellarischen Darstellungen und zwei Grafiken präsentiert, deren Interpretation nachfolgend kurz betrachtet werden soll.

Die Distanzmatrix zeigt die Distanzwerte sämtlicher möglichen Einzelobjekt-Paare und damit gewissermaßen die Ausgangssituation vor dem ersten Schritt der Clusteranalyse. Sie ist symmetrisch aufgebaut, d.h. jeder Wert ist zweifach vorhanden, weshalb es ausreicht, die halbe Distanzmatrix zu betrachten.

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	12	,068	0	0	6
2	3	6	,097	0	0	3
3	3	11	,164	2	0	6
4	2	7	,524	0	0	10
5	10	13	,809	0	0	8
6	3	5	,828	3	1	8
7	1	8	1,166	0	0	11
8	3	10	1,316	6	5	10
9	14	15	1,329	0	0	12
10	2	3	4,175	4	8	11
11	1	2	5,630	7	10	12
12	1	14	8,489	11	9	13
13	1	9	12,077	12	0	14
14	1	4	37,427	13	0	0

Abbildung 70: Agglomerationstabelle in SPSS

Die Agglomerationstabelle zeigt den schrittweisen Verlauf der Clusteranalyse. In der ersten, obersten Stufe bildet jedes Objekt noch einen einzelnen Cluster, in der letzten, untersten Stufe sind alle Objekte in einem Megacluster vereint. Jede Zeile der Tabelle beschreibt somit

genau einen Schritt der Clusterbildung (welcher Cluster bzw. welches Objekt wurde mit welchem anderen Cluster oder Objekt vereinigt?), wobei zu erkennen ist, dass die Distanzwerte (Koeffizienten) zwischen den vereinigten Objekten – also die bei der Vereinigung zu überbrückende Distanz – im Verlauf des Verfahrens beständig ansteigt. Dies ist typisch für die Clusteranalyse, da ja auch immer unähnlichere Objekte vereinigt werden, je näher das Verfahren der Bildung des Megaclusters kommt. Ein großer Sprung in den Koeffizienten kann als Hinweis darauf betrachtet werden, dass die finale Clusterbildung zwischen diesen beiden Vereinigungsstufen enden sollte.

Das Eiszapfendiagramm stellt den Prozess der Clusterbildung grafisch dar und gibt somit die gleichen Informationen wieder, wie die oben angesprochene Tabelle der Agglomerationsschritte – allerdings bei besserer Übersichtlichkeit. Von oben nach unten gelesen beschreibt also auch dieses Diagramm den schrittweisen Ablauf der Clusteranalyse, wobei sich in der untersten Stufe die Ausgangssituation und in der obersten Stufe der Megacluster findet.

Einen ähnlichen Informationsgehalt besitzt auch das Dendrogramm, welches wie das Eiszapfendiagramm den Ablauf der Clusteranalyse grafisch darstellt. Das Dendrogramm vereinigt dabei wichtige Informationen aus der Tabelle der Agglomerationsschritte (Entfernung der Cluster oder Einzelobjekte voneinander zum Zeitpunkt der Vereinigung) mit der grafischen Übersichtlichkeit des Eiszapfendiagramms, weshalb es diejenige Ergebnisgrafik ist, anhand der in der Regel die finale Einteilung der Objekte in die Cluster vollzogen wird. Da die Abstände zwischen den Objekten oder Clustern bei der Vereinigung direkt erkennbar sind, kann der Marktforscher sofort ablesen, ob es sich um eine Vereinigung noch relativ homogener oder bereits relativ heterogener Objekte oder Cluster handelt.

Zu beachten ist, dass in keinem Fall eine einzige, „richtige“ Clustereinteilung existiert, sondern in der Regel mehrere Clustereinteilungen Sinn machen. Bei der finalen Entscheidung sind sowohl die Erfahrung und das interpretatorische Geschick des Marktforschers als auch gegebenenfalls die Unterstützung durch entsprechende Fachexperten gefragt.

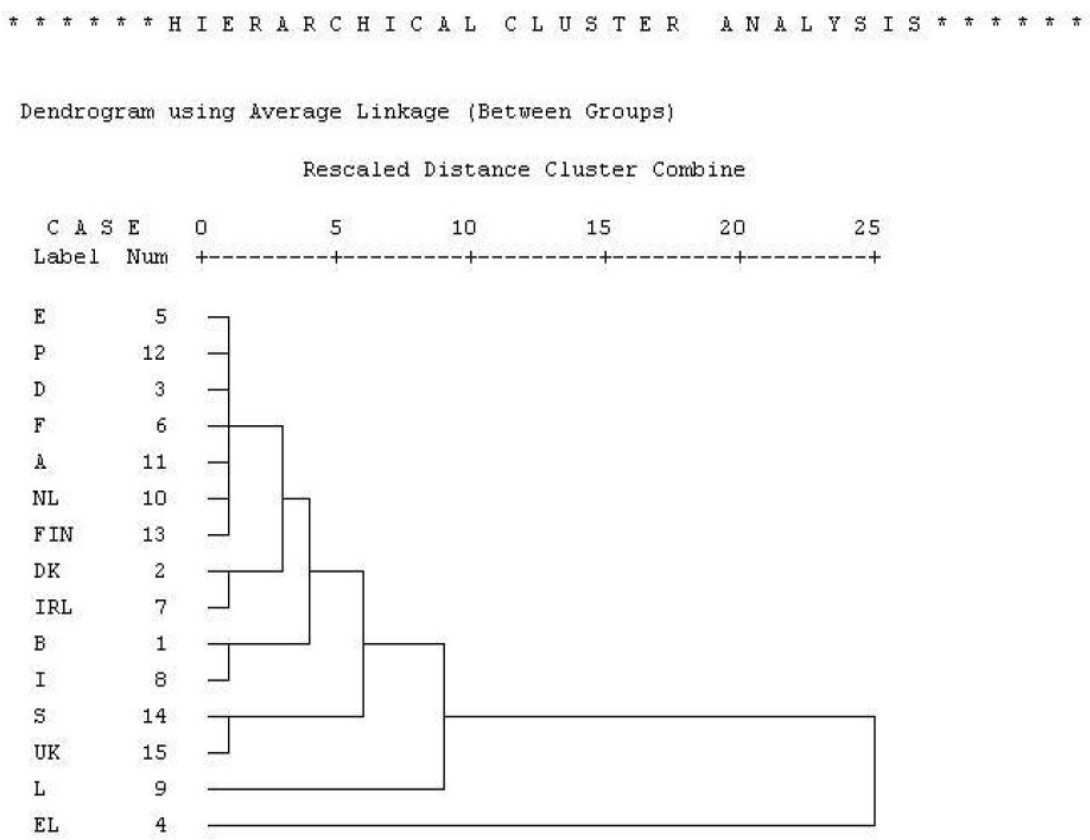

Abbildung 71: Answer Tree in SPSS

II Answer-Tree-Verfahren

1 Grundlagen

Ziel des Answer-Tree-Verfahrens ist die Unterteilung einer Population in mehrere Teilpopulationen anhand eines baumartigen Klassifikationssystems, womit es sich in der Durchführung und im Ergebnis, nicht aber in der Zielsetzung von der Clusteranalyse abhebt.

Das Ausgangsmodell des Answer-Tree-Verfahrens besteht aus einer abhängigen Variablen, der sogenannten Zielvariablen (beispielsweise der Anzahl der Käufer eines Produktes) und einer Reihe von unabhängigen Variablen, den sogenannten Prediktoren (beispielsweise diverse demographische Merkmale wie Alter, Geschlecht, Bildungsstand und Einkommen).

Die Population soll nun anhand der Prediktoren in Teilpopulationen aufgespalten werden,

die sich wiederum bezüglich der Zielvariablen signifikant voneinander unterscheiden sollen, d.h. es werden anhand der Prediktoren und hinsichtlich der Zielvariablen Gruppen gebildet, die intern möglichst homogen und extern möglichst heterogen ausfallen.

Das Verfahren, welches in der Praxis häufig für Webetests oder in der Produktforschung eingesetzt wird, bringt keinerlei Voraussetzungen mit sich, alle verwendeten Variablen, sowohl die Zielvariable als auch die Prediktoren, können nominalskaliert sein (und sind es auch meistens).

2 Verfahrensablauf

Das Answer-Tree-Verfahren läuft in vier wesentlichen Schritten ab. Zunächst müssen die Zielvariable und die Prediktorvariablen identifiziert und ein entsprechendes Modell aufgestellt werden. Dabei sind – wie bereits oben angesprochen – außer der inhaltlichen Logik des Modells keinerlei speziellen Voraussetzungen zu beachten – Answer Tree ist das einzige voraussetzungsfreie Verfahren überhaupt.

Im zweiten Schritt – dem sogenannten Merging – werden dann Prediktorgruppen zusammengefasst, die keine bedeutsamen Unterschiede aufweisen. Dadurch verkleinert sich die Zahl der im Splitting – dem vierten und letzten Schritt – zu berücksichtigenden Prediktorgruppen.

Zwischen dem Merging und dem Splitting steht noch die Bonferroni-Korrektur. Das Ziel dieses Korrekturverfahrens ist die Unterdrückung der Alpha-Fehlerinflation (analog zur ANOVA). Der Rechenschritt wird durch das SPSS-Answer-Tree-Modul automatisch durchgeführt und wird im Rahmen dieser Kurzdarstellung nicht weiter beachtet.

Im letzten Schritt – dem Splitting – wird die für das Answer-Tree-Verfahren charakteristische Baumstruktur durch die Zerlegung der Population anhand der aus dem Merging übriggebliebenen Prediktorgruppen vorgenommen.

3 Merging

Während des Mergings werden die einzelnen, durch die Prediktoren entstehenden Untergruppen auf Unterschiede bezüglich der Zielvariablen geprüft. Das Prüfverfahren wird dabei in Abhängigkeit vom Skalenniveau der Prediktoren selektiert: Ist es metrisch, so erfolgt die Prüfung anhand des aus der Varianzanalyse bekannten F-Tests, ist es ordinal kann über das Y-Verknüpfungsmodell oder den Likelihood-Quotient-Test geprüft werden und bei nominalskalierten Prediktoren kommt der aus der Statistik I bekannte Chi² zum Einsatz.

Für das Merging stehen drei bzw. vier verschiedene Algorithmen zur Auswahl:

- CHAID (Chi-Squared Automatic Interaction Detector) / Exhaustive CHAID
- C & RT (Classifications and Regressions Tree)
- QUEST (Quick and Unbiased Statistical Tree)

Im Folgenden werden nur noch CHAID and Exhaustive CHAID betrachtet, die in der Praxis die größere Rolle spielen.

Wie läuft nun das Merging an sich ab? Betrachten wir dies am Beispiel eines Chi²-Tests, also beim Vorliegen nominalskalierter Prediktoren: Die Nullhypothese H0 dieses Tests lautet, dass in der Grundgesamtheit kein signifikanter Zusammenhang zwischen Prediktorvariable und Zielvariable besteht (also Chi² = 0). Der von SPSS für den Chi²-Test berechnete Wert p gibt die Wahrscheinlichkeit dafür an, beim Verwerfen dieser Nullhypothese einen Fehler zu begehen. Ist also p groß, kann H0 nicht verworfen werden und es ist vom Nicht-Vorliegen eines Zusammenhangs in der Grundgesamtheit auszugehen, d.h. es wird eine interne Homogenität vermutet. Dies bedeutet, dass die beiden untersuchten Gruppen zu einer einzigen Gruppe zusammengefasst werden können – der als Merging bezeichnete Vorgang. Auf diese Weise werden nun alle möglichen Kombinationspaare untersucht und gegebenenfalls gemergt.

Worin besteht nun aber der Unterschied zwischen CHAID und Exhaustive CHAID?

Kommt der CHAID-Algorithmus zum Einsatz, wird ein kritischer Wert (α-merge) festge-

legt, wobei 0,05 und 0,01 übliche Werte sind. Mit diesem kritischen Wert werden die jeweiligen p-Werte aus den Chi²-Tests verglichen und gegebenenfalls ein Merging durchgeführt. Das Merging der Kategorien wird solange fortgesetzt, bis der p-Wert der Tests für die verbliebenen Kategorien nicht mehr den kritischen Wert überschreitet Es wird also bis zu einem Punkt gemergt, an dem die verbliebenen Teilpopulationen bezüglich der Prediktoren so unterschiedlich sind, dass kein weiteres Merging mehr stattfinden sollte

Anders verhält es sich beim Exhaustive CHAID-Algorithmus, bei welchem auf einen kritischen Wert vollständig verzichtet wird. Statt dessen wird die Merging-Phase so lange fortgesetzt, bis nur noch zwei Kategorien übrig bleiben. Anschließend werden die kumulierten p-Werte der einzelnen Vereinigungsebenen betrachtet, wobei nach der Ebene mit dem niedrigsten Wert gesucht wird. Logischerweise ist dies stets die Ebene mit der größten Heterogenität zwischen den Gruppen, so dass mit dieser Gruppenaufteilung ins Splitting gegangen werden kann.

Die Anzahl der auf Unterschiede zu testenden Gruppen ist dabei vom Skalenniveau der Prediktorvariablen abhängig: Liegen ordinalskalierte und metrisch skalierte Prediktoren vor, so dürfen nur „benachbarte“ Kategorien gemergt werden, bei nominalskalierten Prediktoren ist dagegen das Merging aller Kategorien gestattet.

4 Splitting

Während des Splittings werden die sogenannten Knoten aufgeteilt, bis bestimmte Abbruchbedingungen erfüllt sind. Durch die Aufteilung ergibt sich die für das Answer-Tree-Verfahren charakteristische Baumstruktur.

Dabei beginnt man mit dem Gesamtknoten – einem einzelnen Knoten der Größe n, welcher die gesamte Population enthält. Aus diesem Knoten wird nun der Prediktor mit dem kleinsten p-Wert zum Splitting herausgesucht, wobei die Nullhypothese H0 auch in dieser Phase des Answer-Tree-Verfahrens besagt, dass in der Grundgesamtheit keinerlei Zusammenhänge zwischen Prediktorvariable und Zielvariable bestehen (also $Chi^2 = 0$). Im Gegensatz

zum Merging wird aber im Splitting nicht nach besonders großen p-Werten sondern nach besonders kleinen p-Werten gesucht, denn große p-Werte deuten auf homogene Gruppen hin (kein Zusammenhang feststellbar), während kleine p-Werte auf heterogene Gruppen hinweisen (Zusammenhang feststellbar).

Dieses Verfahren wird nun so lange fortgesetzt, bis eine der drei möglichen Abbruchbedingungen erfüllt ist:

- Es finden sich keinerlei signifikanten Unterschiede zwischen den Gruppen mehr (dies kann analog zum CHAID-Verfahren mit einem kritischen Wert geprüft werden).
- Die Tiefe des Baumes hat den festgelegten Höchstwert erreicht (dieser Höchstwert kann durch den Marktforscher anhand inhaltlicher Überlegungen festgelegt werden).
- Die minimale Größe der teilbaren Knoten wurde erreicht.

Während des Splittings ist besonders auf das Auftreten fehlender Werte zu achten. Diese können – wie bereits oben dargestellt – auf Probleme hinweisen und sollten daher ohnehin grundsätzlich untersucht werden. Ein Answer Tree sollte nur erstellt werden, wenn die fehlenden Werte zufällig auftreten (MCAR). Es ergibt sich eine eigene Prediktorkategorie (missing), die nur bei nominal skalierten Merkmalen mit anderen Prediktorkategorien gemergt werden darf.

5 Interpretation des Baumes

Weiter unten findet sich beispielhaft ein Ausschnitt aus einem Answer-Tree-Ergebnisbaum. Er zeigt zwei Finalknoten – zwei Knoten ohne weitere Verzweigungen – aus einer Untersuchung über die gefühlte Bedrohung durch Atomkraftwerke.

Es lässt sich erkennen, dass Knoten 3 während des Splittings noch einmal nach dem Prediktormerkmal „Geschlecht“ aufgespalten wurde – offenbar gab es zwischen den beiden Teilpopulationen bezüglich der Zielvariablen – der besagten gefühlten Bedrohung – noch signifikante Unterschiede. Eine weitere Teilung der Knoten 10 und 11 kann aufgrund der Erreichung

einer der drei Abbruchbedingungen nicht mehr möglich gewesen sein.

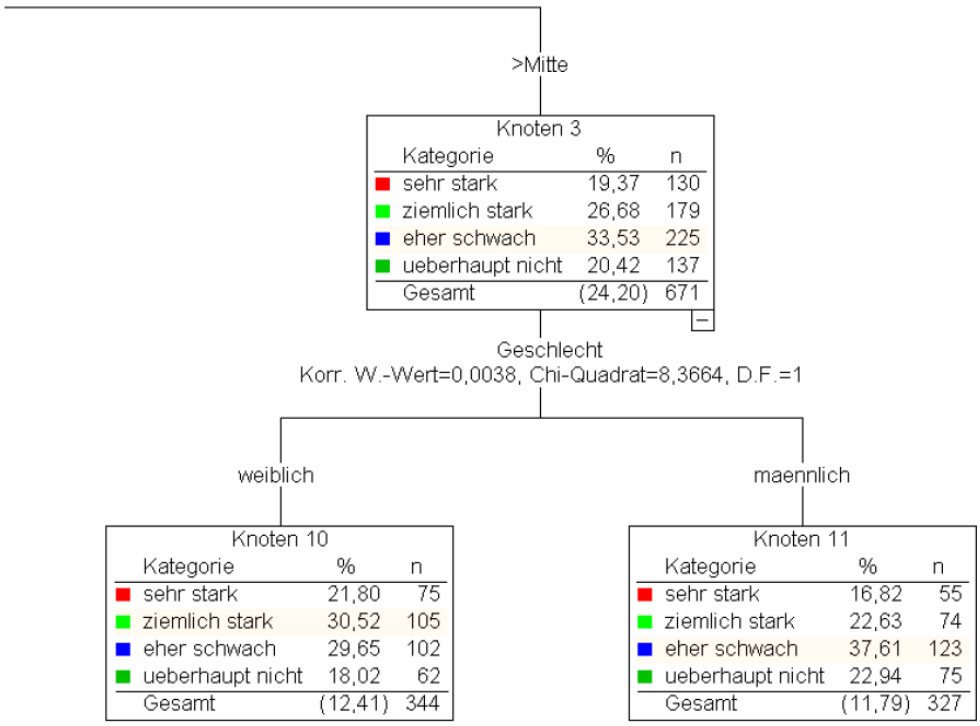

Abbildung 72: Answer Tree in SPSS

III Conjoint Analysis

1 Einführung

Die Grundidee hinter der Conjoint Analysis ist, dass der vom Kunden subjektiv wahrgenommene Produktnutzen – auch als Produktwert bezeichnet – sich aus der Summe der Teilnutzenwerte der einzelnen Merkmale und Merkmalsausprägungen des Produktes ergibt (beispielsweise Farbe, Geschmack, Verpackung etc.). Das Ziel der Conjoint Analysis ist es, die Beiträge dieser Merkmale und Merkmalsausprägungen zum Gesamtnutzen festzustellen und so beispielsweise das optimale Produkt zu ermitteln oder die Veränderungen am Markt bei Änderungen an bestehenden Produkten zu prognostizieren.

Die Idee, dass sich der Produktnutzen aus den Teilnutzenwerten der Merkmale und Merk-

malsausprägungen zusammensetzt, wird auch als linear-additives Teilwertmodell bezeichnet. Liegt ein solches Modell vor, so verlaufen die Präferenzwirkungen der Merkmale nicht in bestimmte Richtungen (wie beispielsweise beim Zuckeranteil in der Limonade, der die geschmackliche Bewertung bis zu einem gewissen Punkt linear steigen lässt), statt dessen sind unterschiedliche Präferenzwirkungen für unterschiedliche Ausprägungen denkbar (wie beispielsweise bei Farbe, Form oder Geschmack eines Lebensmittelprodukts).

Vom Vorliegen eines solchen linear-additiven Teilwertmodells muss für das zu untersuchende Produkt ausgegangen werden können, was längst nicht bei jedem Produkt der Fall ist. So ist das Modell bereits verletzt, wenn die Merkmale ihren Teilnutzen nicht unabhängig voneinander beeinflussen, sondern sich gegenseitig beeinflussen. Auch darf es keinerlei K.O.-Kriterien geben, d.h. alle negativen Merkmalswirkungen müssen auch kompensierbar sein. So darf es beispielsweise bei der Untersuchung eines Schokoladenprodukts nicht vorkommen, dass eine bestimmte Geschmacksrichtung, die einem Verbraucher nicht zusagt, in jedem Fall zum Nichtkauf des Produkts führt – ist dies dennoch der Fall, spricht man von einer Verletzung des kompensatorischen Beurteilungsprozesses.

Neben der Integrität des linear-additiven Teilwertmodells und des kompensatorischen Beurteilungsprozesses sind weitere Voraussetzungen zu betrachten: Verwendete Kombinationen aus Merkmalsausprägungen müssen technisch realisierbar sein, die Anzahl der relevanten Merkmale und Merkmalsausprägungen muss begrenzbar sein und es sollten nur solche Merkmale betrachtet werden, die für die Beurteilung relevant sind.

2 Verfahrensablauf

Die Conjoint Analysis läuft in sechs wesentlichen Schritten ab. Zunächst einmal ist die Fragestellung festzulegen, wobei qualitative Pretests dazu eingesetzt werden können, relevante Merkmale und Merkmalsausprägungen zu identifizieren und deren Zahl so von vornherein zu beschränken.

Anschließend sind die in der Analyse zu verwendenden Merkmale und Merkmalsauspräa-

gungen zu beschreiben. Dabei ist auch darauf zu achten, dass, wie bereits oben dargestellt, alle Kombinationen aus Merkmalsausprägungen technisch und praktisch umsetzbar sind.

Im dritten Schritt ist das Conjoint-Verfahren auszuwählen. Der Marktforscher kann sich hier zwischen verschiedenen traditionellen Verfahren (Full Profile, Trade Off, Logit/Probit) und neuen Verfahren (Hybrid Conjoint Analysis, Adaptive Conjoint Analysis, Choice-Based Conjoint Analysis) entscheiden, wobei nachfolgend nur noch Full Profile, ACA und CBC betrachtet werden.

Die Erhebung der Daten erfolgt im nächsten Schritt. Neben der Erhebung im Labor mit Hilfe von Karten, die von den Befragten in eine Präferenzrangfolge gebracht werden müssen, hat sich in den letzten Jahren vor allem die Befragung am Computer in der Praxis durchgesetzt. Der Einsatz von Software zur Befragung bringt eine Menge an Vorteilen mit sich, darunter die Möglichkeit, die Zahl der Vergleiche während des Befragungsverlaufs anhand vorangegangener Antworten einzuschränken.

Liegen die erhobenen Daten vor, lassen sich im vorletzten Schritt der Conjoint Analysis noch die bereits angesprochenen Teilnutzenwerte ermitteln. Dadurch lässt sich unter anderem feststellen, welche Merkmale für den Gesamtnutzen besonders wichtig sind und welche Merkmalsausprägungen bei den Befragten auf besonderen Anklang stoßen.

Im finalen Schritt erfolgt die weitere Analyse der erhobenen Daten. Einige der am Markt erhältlichen Conjoint-Softwarepakete (wie Sawtooth SMRT) ermöglichen beispielsweise die Durchführung von Marktsimulationen, basierend auf den Conjoint-Ergebnissen. So kann der Marktforscher, anhand eines aus den Erhebungsdaten berechneten Basecase, feststellen, ob es bei Veränderungen an einem Produkt zu Käuferwanderungen kommt oder wie der Markt auf die Einführung eines neuen Produkts reagieren würde. Insbesondere diese weiterführenden Analysemöglichkeiten haben in den letzten Jahren stark dazu beigetragen, der Conjoint Analysis in der Marktforschungs-Praxis zu einem der meistverwendeten Analyseverfahren überhaupt werden zu lassen.

3 Verfahrensansätze

Grundsätzlich ist bei der Conjoint Analysis zwischen zwei Verfahrensansätzen zu unterscheiden, und zwar den kompositionellen Verfahren und den dekompositionellen Verfahren. Der in der Fachliteratur ebenfalls auftauchende dritte Verfahrensansatz – die hybriden Verfahren – umfasst Verfahren, die eine Mischung aus den beiden grundsätzlichen Verfahrensansätzen darstellen, wie beispielsweise das CBC.

Im Rahmen eines kompositionellen Conjoint-Verfahrens werden die Teilnutzenwerte der einzelnen Merkmalsausprägungen getrennt berechnet. Auf der Basis dieser Werte wird dann additiv, im oben bereits dargestellten linear-additiven Teilwertmodell, der Gesamtnutzen eines Produktes mit einer bestimmten Kombination aus Merkmalsausprägungen berechnet.

Im Rahmen eines dekompositionellen Conjoint-Verfahrens werden den Befragten keine einzelnen Merkmale und Merkmalsausprägungen zur Bewertung vorgelegt. Statt dessen werden ausschließlich ganzheitliche Produktkonzepte präsentiert und deren Gesamtnutzen ermittelt. Auf der Basis dieses Gesamtnutzens wird dann, in einer Umkehrung des ursprünglichen linear-additiven Teilwertmodells versucht, die einzelnen Teilnutzenwerte zu extrahieren.

4 Full Profile, ACA und CBC

Nachfolgend werden drei der bekannteren (und daher meist auch klausurrelevanteren) Conjoint-Verfahren näher betrachtet: Full Profile, Adaptive Conjoint Analysis und Choice Based Conjoint Analysis.

Der sogenannte Full Profile-Ansatz ist das älteste Conjoint-Verfahren überhaupt. Der Befragte muss dabei ganzheitliche Produktkonzepte in eine Präferenzrangfolge bringen – es handelt sich also um ein dekompositionelles Verfahren. Der Full Profile-Ansatz verlangt, dass sämtliche möglichen Kombination von Merkmalsausprägungen vom Befragten bewertet werden. Aus diesem Grund ist das Full Profile-Verfahren nur durchführbar, wenn die Zahl der Merkmale und Merkmalsausprägungen gering ist – als Richtwert kann von 2-3 Merkmalen

mit jeweils 2-3 Ausprägungen ausgegangen werden. Wesentlich mehr Merkmale oder Merkmalsausprägungen führen zu einer für normalkonzentrierte Befragte nicht mehr überblickbaren Vielfalt von zu bewertenden Alternativen. Lässt sich diese Zahl mit Blick auf den zu untersuchenden Fall nicht bis zu besagtem Limit verringern, ist auf ein anderes Conjoint-Verfahren zurückzugreifen, wie beispielsweise auf die Adaptive Conjoint Analysis.

Bei der Adaptive Conjoint Analysis (ACA) handelt es sich um ein computergestütztes Conjoint-Verfahren, welches von der US-amerikanischen Firma Sawtooth entwickelt wurde. Es ist das heute am weitesten verbreitete Conjoint-Verfahren und vereint kompositionelle und dekompositionelle Elemente in sich, ist also den hybriden Conjoint-Verfahren zuzuordnen. Im kompositionellen Teil der ACA selektieren und bewerten die Befragten einzelne Merkmale und auch Merkmalsausprägungen, im dekompositionellen Teil müssen sie sich dann zwischen verschiedenen möglichen Produkten entscheiden, die von der Software anhand der Ergebnisse des kompositionellen Teils designed worden sind. Dies hat den Vorteil, dass nicht mehr alle möglichen Alternativen bewertet werden müssen, weshalb sich mit der ACA auch Studien mit bis zu 30 Merkmalen mit bis zu 9 Ausprägungen durchführen lassen – wesentlich mehr als beim Full Profile-Ansatz verwertbar sind.

Das jüngste aller Conjoint-Verfahren ist das sogenannte Choice Based Conjoint (CBC). Es basiert als einziges Conjoint-Verfahren nicht auf dem weiter oben dargestellten linear-additiven Teilwertmodell und der Befragte wird an keiner Stelle zu Präferenzen, Merkmalsbewertungen oder Kaufwahrscheinlichkeiten befragt. Statt dessen wird er mit einem realistischen Kaufszenario konfrontiert: Aus zwischen 2-8 ganzheitlichen und vollständig beschriebenen Produkten kann eines ausgewählt oder die Entscheidung getroffen werden, keines der Produkte zu kaufen (man spricht hier auch von der sogenannten Non-Option). Die dem Probanden präsentierte Auswahlsituation ist beim CBC daher als wesentlich realistischer zu bewerten als bei jedem anderen Conjoint-Verfahren. Dazu kommt noch, dass CBC das einzige Conjoint-Verfahren ist, bei dem Interaktionseffekte zwischen den einzelnen Merkmalen berücksichtigt und (über das sogenannte Conditional Pricing) auch der Preis als ein an andere Merkmale gekoppeltes Zusatzmerkmal in die Analyse eingebunden werden kann. Nachteilig ist dagegen, dass ähnlich wie beim Full Profile-Ansatz nicht mehr als 12 Kombination bewertet werden können und die einzelnen Teilnutzenwerte nicht mehr separat analysiert werden kön-

nen, da das Verfahren mit aggregierten Daten arbeitet.

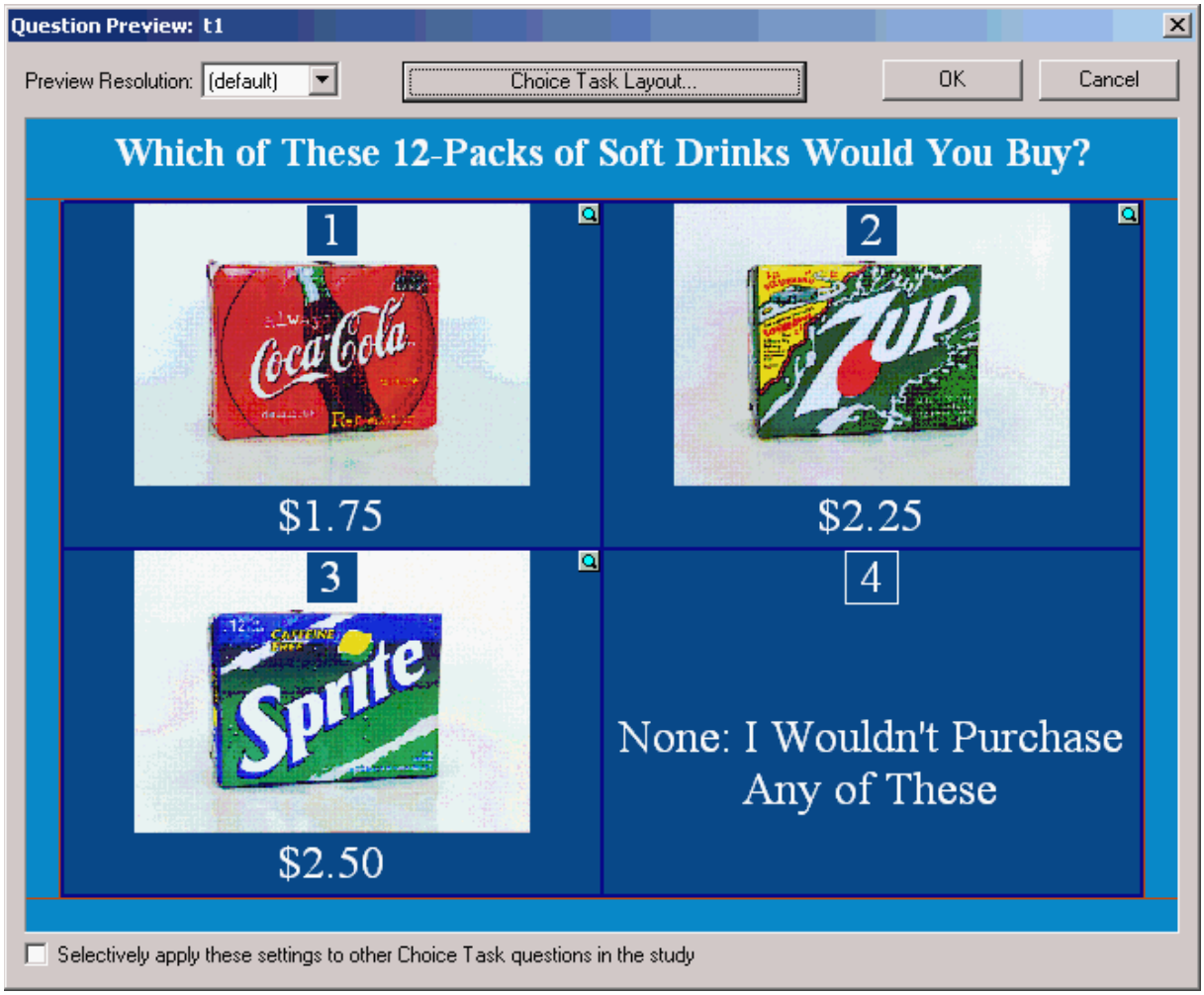

Abbildung 73: Bewertung ganzheitlicher Produktkonzepte im CBC

5 Segmentspezifische Analyse

In der Praxis wird bei der Befragung von potentiellen Kunden aus verschiedenen Bevölkerungs- und Einkommensschichten häufig der Fall auftreten, dass die Gruppe insgesamt hinsichtlich ihrer Präferenzstruktur nicht homogen, sondern heterogen ist. Durch einfache Aggregation der Teilnutzenwerte gehen in einer solchen Situation wertvolle Informationen verloren, da sich starke positive und starke negative Abweichungen ausgleichen – die Streuung der Bewertungen ist also so groß, dass sie das Ergebnis der Conjoint Analysis in Frage stellen.

Die Lösung für dieses Problem besteht darin, die Aggregation der Daten stets nur mit Daten aus Gruppen mit relativ homogener Präferenzstruktur durchzuführen. Um dies zu erreichen, sind die Daten so zu segmentieren, dass sich besagte Gruppen ergeben. Dies kann mittels einer a-priori-Segmentierung oder einer a-posteriori-Segmentierung geschehen.

Bei einer a-priori-Segmentierung erfolgt die Einteilung der Gruppen im Vorfeld der eigentlichen Conjoint Analysis auf der Basis demographischer Merkmale (wie beispielsweise Alter, Geschlecht oder Einkommensgruppe). Eine solche Segmentierung ist nur dann sinnvoll, wenn es einen Zusammenhang zwischen den hierfür verwendeten Merkmalen und der untersuchten Präferenzstruktur gibt, da ansonsten das Ziel der homogenen Gruppenbildung verfehlt wird. Die Auswahl der demographischen Merkmale sollte daher unter Hinzuziehung eines Fachexperten erfolgen, wenn eine Entscheidung zugunsten der a-priori-Segmentierung gefällt wurde.

Die a-posteriori-Segmentierung erfolgt dagegen erst nach Abschluss der Conjoint Analysis auf der Basis der gesammelten Testwerturteile. Zusätzlich erhobene demographische Merkmale können dann noch verwendet werden, um die so entstehenden Gruppen zu beschreiben. Die a-posteriori-Segmentierung ist das in der Conjoint Analysis übliche Segmentierungsverfahren. Zur Aufteilung der Gruppen wird in der Regel eine Clusteranalyse verwendet – in diesem Zusammenhang ist oft auch von der sogenannten Benefitsegmentierung die Rede.

6 Verfahrensprobleme

Abschließend sollen noch drei der im Zusammenhang mit der Conjoint Analysis häufig anzutreffende Verfahrensprobleme kurz betrachtet werden.

Das erste dieser Probleme ist auch unter der Bezeichnung Level-Effekt bekannt und kann auch im Zusammenhang mit anderen Erhebungsverfahren auftreten[1]. Aufgrund psychologischer Gründe, die an dieser Stelle nicht weiter betrachtet werden sollen, werden Merkmale

1 Sowohl bei der Erstellung von Papierfragebögen als auch bei Internet-basierten Befragungen kann der Level-Effekt eine Rolle spielen. Bei experimentellen oder automatischen Erhebungsverfahren, die keine direkte oder gar keine Beteiligung von menschlichen Probanden erfordern, kann der Effekt dagegen nicht auftreten.

mit vielen Ausprägungen von manchen Probanden subjektiv als wichtiger empfunden. Dies hat zur Folge, dass die relative Wichtigkeit eines Merkmals und damit dessen Einfluss auf den Gesamtnutzen eines Produkts mit der Anzahl der diesem Merkmal zugeordneten alternativen Merkmalsausprägungen steigt. Als Folge davon werden Merkmale über- und unterbewertet und der Marktforscher erhält ein aussageschwaches Endergebnis. Die einzige Möglichkeit, dieses Problem zu umgehen, besteht in der Verteilung etwa gleichvieler Merkmalsausprägungen auf die einzelnen Merkmale – eine solche „Gleichverteilung" ist daher stets anzustreben und im Zweifelsfalle dem Informationsverlust vorzuziehen, der duch das Weglassen von Merkmalsausprägungen entsteht.

Das zweite hier betrachtete Problem wird in der Fachsprache als Positionseffekt bezeichnet. Dieser Effekt, der ebenfalls einen pychologischen Hintergrund hat, bewirkt, dass Merkmale, die bei Vergleichen stets an erster Stelle genannt werden, von manchen Probanden analog zum Level-Effekt als subjektiv bedeutender eingeschätzt werden, was wiederum zur Über- und Unterbewertung von Merkmalen und einem verzerrten Ergebnis führt. Die Lösung dieses Problems besteht in der Randomisierung – also der zufälligen Anordnung der Merkmale – bei solchen Vergleichen durch die verwendete Erhebungssoftware. Durch die Randomisierung wird der Positionseffekt umgangen, allerdings gestalten sich auch die Vergleiche für die Probanden unübersichtlicher und wirken unprofessioneller gestaltet – hier ist also zwischen zwei möglichen negativen Auswirkungen abzuwägen.

Als letztes Verfahrensproblem der Conjoint Analysis sollen die häufig viel zu subjektiven Beschreibungen betrachtet werden. Werden subjektive Beschreibungen (wie beispielsweise „knusprig") anstelle objektiver Beschreibungen (wie beispielsweise „mit Cornflakes") eingesetzt, so kann dies die Ergebnisse einer Analyse schwer verfälschen. Grund dafür ist, dass Probanden angesichts subjektiver Beschreibungen ihre eigenen Vorstellungen entwickeln – knusprig mag ja von jedem Befragten für sich anders definiert werden, während jeder unter der Zugabe von Cornflakes wohl in etwa das gleiche verstehen dürfte. Für jede Conjoint Analysis gilt daher, dass die Beschreibungen von Merkmalen und insbesondere Merkmalsausprägungen stets möglichst objektiv gestaltet werden sollen.

Eine Ausnahme bilden besondere Sachverhalte, beispielsweise auf technischer Ebene. So

ist die Angabe der Durchschnittstemperatur von Friteusen als rein metrischer Wert eine besonders objektive Darstellung – ein solches Merkmal dürfte aber von den meisten Probanden kaum als bedeutsam wahrgenommen werden und beeinflusst damit die Kaufentscheidung nicht. Die eher subjektive Angabe, ab welcher Durchschnittstemperatur dagegen von einem höheren Krebsrisiko auszugehen ist, und inwiefern eine zur Bewertung anstehende Friteuse diese Temperatur über- oder unterschreitet wird von den Probanden sehr wohl als kaufwirksames Merkmal wahrgenommen werden – in solchen Fällen sind subjektivere Darstellungen zugelassen um das Merkmal für die Probanden verständlich zu gestalten und auch „Nicht-Eingeweihten“ eine Entscheidung zu ermöglichen.

IV Weiterführende Literatur

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). Multivariate Analysemethoden (10. Aufl.). Berlin: Springer.

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

G Anhang

Literaturverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Stichwortverzeichnis

Linksammlung

Literaturverzeichnis

Bleymüller, J.; Gehlert, G. & Gülicher, H. (2000). Statistik für Wirtschaftswissenschaftler. München: Verlag Vahlen

Dannenberg, M. & Barthel, S. (2004). Effiziente Marktforschung. Frankfurt/Wien: Wirtschaftsverlag Carl Ueberreuther

Brosius, F. (2002). SPSS 11. Bonn: mitp-Verlag

Diehl, J.M. & Staufenbiel, T. (2002). Statistik mit SPSS Version 10 +11. Eschborn: Klotz

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (1999). Statistik. Der Weg zur Datenanalyse (2. Aufl.). Berlin: Springer.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). Multivariate data analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Heiler, S. & Michels, P. (1994). Deskriptive und Explorative Datenanalyse. München: Oldenbourg.

Janssen, J. & Laatz, W. (2003). Statistische Analyse mit SPSS für Windows (4. Aufl.). Berlin: Springer.

Koch, J. (1997). Marktforschung - Begriffe und Methoden. München: R.Oldenbourg Verlag

Abbildungsverzeichnis

Tabellenverzeichnis

Stichwortverzeichnis

Linksammlung

Allgemeine Statistik

Statistisches Bundesamt	http://www.destatis.de
Statistisches Landesamt Sachsen-Anhalt	http://www.stala.sachsen-anhalt.de
Bundesverband Deutscher Markt- und Sozialforscher	http://www.bvm.de

Forschung & Lehre

Projekt „Neue Statistik“ der FU Berlin	http://www.neuestatistik.de
Das Online-Statistiklabor	http://www.statistiklabor.de
Rice Virtual Lab in Statistics	http://www.ruf.rice.edu/~lane/

Fragebögen, Interviews & Gruppendiskussionen

Qualitative Research Consultants Association	http://www.qrca.com
American Association for Pulblic Opinion Research	http://www.aapor.de
Council of American Survey Research Organizers	http://www.casro.org

Online-Marktforschung

Arbeitsgemeinschaft Sozialwissenschaftlicher Institute	http://www.gesis.org/asi/
Deutsche Gesellschaft für Online-Forschung	http://www.dgof.de
Forschungsgruppe Wahlen	http://www.forschungsgruppe.de
Quirk's Market Research Review	http://www.quirks.com
Marketagent Online-Marktforschung	http://www.marketagent.com