Hochschule für Angewandte Wissenschaften Hamburg

*Hamburg University of Applied Sciences*

# Einsatz der linearen Diskriminanzanalyse als Alternative zur Baumanalyse – methodische Evaluierung und Durchführung am Beispiel einer Erhebung zum Erfolg von Kinofilmen

# Bachelor-Thesis

An der Hochschule für angewandte Wissenschaften Hamburg

Fakultät Wirtschaft und Soziales

Studiengang technische BWL/Marketing

vorgelegt von

Johannes Tauscher

[redacted]

18. September 2017

1. Gutachter: Prof. Stefan Tuschl
2. Gutachterin: Prof. Dr. Elke Hörnstein

# Inhalt

# 1. Einleitung

Die Rolle und Bedeutung von angewandter Statistik und quantitativen Methoden ist in den letzten Jahren in vielen Anwendungsgebieten immer wichtiger geworden. Durch die vorangehende Digitalisierung und Ausstattung der Umwelt mit vielseitiger Sensorik können in vielen Bereichen immer einfacher große Datenmengen gesammelt werden. Doch nur ein geringer Teil davon wird tatsächlich verwertet. Gute Datenanalysen setzen fundierte Fachkenntnisse, hohe Zahlenaffinität und Analysegeschick voraus, so zumindest die Anforderungen der freien Wirtschaft[1]. Dabei sind Erkenntnisse und Zusammenhänge, die anhand statistischer Methoden aus erhobenen Daten hervorgebracht werden, häufig inhaltlich wertvoll, bringen Überlegungen und Entscheidungsprozesse voran oder ermöglichen das Ableiten von ganz neuen Perspektiven auf gegebene Sachverhalte. Methoden, Werkzeuge, Verfahren und die dazugehörige Software sollten einem breiten Publikum nicht vorenthalten sein. Nicht nur in der Wissenschaft - auch ökonomische und gesellschaftspolitische Aktivitäten sind heutzutage größtenteils datengesteuert[2]. Dabei stehen Anwendern der wissenschaftlichen Statistik bereits eine Vielzahl von Werkzeugen und Methoden zur Verfügung. Die Beherrschung dieser Werkzeuge und Methoden setzt häufig allerdings fundierte Vorkenntnisse voraus, die etwa durch ein entsprechendes Studium erlangt werden müssen. Durch die wachsende Relevanz von angewandter Statistik und quantitativen Methoden ist einem breitgefächerten Publikum einen verständlichen Zugang zu den verschiedenen Themen, Methoden und Anwendungen in diesem Wissenschaftsgebiet zu ermöglichen, eine entscheidende Aufgabe der Wissenschaft selbst.

Übergeordnetes Ziel dieser Bachelorarbeit ist daher dem sich für angewandte Statistik und quantitative Methoden interessierenden Leser bzw. der Leserin die lineare Diskriminanzanalyse als eine Methode zur Auswertung bestimmter Datenstrukturen näher zu bringen. Die Arbeit soll den Leser bzw. die Leserin letzten Endes befähigen, die einzelnen Schritte der Methode nachvollziehen zu können und sie sogar selbst auf Fragestellungen in seinem Wissenschafts-, Arbeits- und Aufgabenumfeld anzuwenden. Darüber hinaus soll er anhand des Methodenvergleichs in die Lage versetzt werden, eine sachgerechte Wahl zwischen dem Verfahren der Diskriminanzanalyse und dem der Baumanalyse für künftige Analyseaufgaben zu treffen.

## 1.1. Inhaltlichen Ziele

Diese wissenschaftliche Arbeit stellt, erst theoretisch und dann anwendungsorientiert mit SPSS anhand eines Datensatzes zu verschiedenen Erfolgsfaktoren von Kinofilmen, das Basiskonzept der linearen Diskriminanzanalyse vor und vergleicht schließlich sowohl die Durchführungen als auch die Ergebnisse einer Diskriminanzanalyse mit denen einer Baumanalyse. Dabei soll sich herausstellen, ob und in welchem Fall die Diskriminanzanalyse eine Alternative zur Baumanalyse ist.

*„Die Diskriminanzanalyse ist ein multivariates Verfahren zur Analyse von Gruppenunterschieden. Sie ermöglicht es, die Unterschiedlichkeit von zwei oder mehreren Gruppen hinsichtlich einer Mehrzahl von Variablen zu untersuchen“ ( Backhaus, 2016,* 13. Auflage S.216)

## 1.2. Aufbau der Thesis

Zunächst werden in Kapitel 2 statistische Grundbegriffe definiert und relevanten Methoden und Begriffe vorgestellt. Dies schafft die Grundlage, um den theoretischen Ausführungen der Grundlagen der Diskriminanzanalyse in Kapitel 3 der Arbeit und der Vorstellung der Baumanalyse in Kapitel 6.1 zu folgen. Um das Konzept der Diskriminanzanalyse vorzustellen, wird mit der Vorgehensweise gearbeitet, die Klaus Backhaus et al. in „Multivariate Analysemethoden“ (13. Auflage) vorstellt. Die Inhalte des relevanten Kapitels dienen dem Theorieteil dieser Arbeit strukturell als Vorlage und werden allgemein als Leitfaden zur typischen Durchführung einer Diskriminanzanalyse angenommen.

[1] https://www.prospects.ac.uk/job-profiles/data-analyst
[2] http://www.uni-bielefeld.de/(en)/zest/studiengang_sw.html

Im Hauptteil, dem Kapitel 4, wird das Verfahren mit der Statistiksoftware SPSS anhand eines Datensatzes zu Kinofilmen beispielhaft durchgeführt.

Ein übergeordnetes Element bei der Vorstellung des Verfahrens wird dabei außerdem sein, die zugrundeliegenden Überlegungen der exemplarischen Diskriminanzanalyse in Kapitel 4.3 an sich darzustellen. Die Durchführung einer Diskriminanzanalyse kann verschiedene Ziele haben. Das Ziel der Diskriminanzanalyse anhand des Datensatzes, der in dieser Arbeit zur Vorführung des Verfahrens dient, wird eine Untersuchung des Sachverhalts sein, ob eine vor- oder frühzeitige Erfolgsprognose in dem ersten Eintrittsmarkt eines Kinofilms anhand von Merkmalen möglich ist, die bereits vor der öffentlichen Aufführung des Films feststehen oder zeitlich sehr früh, beispielsweise durch Prescreenings[3], erhoben werden können. Hierfür wird der Forschungsstand verschiedener Wissenschaftler und deren Untersuchungen vom Einfluss von verschiedenen Faktoren auf den Erfolg von Kinofilmen vorgestellt, der von Michel Clement zusammengefasst und 2004 in der M&K, 52. Ausgabe veröffentlicht wurde. Anhand der Überlegungen in der Kombination mit der Diskriminanzanalyse ist es je nach Ergebnis der Analyse möglich, künftige, noch nicht klassifizierte Filme einer Klasse wie „erfolgreich“ oder „nicht erfolgreich“ zuzuweisen. Die Ergebnisse der Analyse des Beispieldatensatzes geben Rückschlüsse über die Effektstärke der einzelnen Prädiktoren[4]. Hieraus resultiert, dass für künftige Markteinführungen der Filme die betrachteten Prädiktoren eventuell als Stellschrauben angesehen werden können, die den möglichen Erfolg beeinflussen. Die Ergebnisse der Analyse und die Evaluation der beiden statistischen Verfahren werden im Schlussteil der Arbeit verarbeitet und eine Empfehlung für Einsatzgebiete und Anwendungsszenarien ausgesprochen.

Die Ergebnisse dieser Auswertung sind lediglich untergeordnetes Ziel und werden nur beiläufig in Kapitel 5 dargestellt, um einen beispielhaften Einblick in Interpretation der Resultate zu liefern.

Anschließend wird in Kapitel 6 die Entscheidungsbaumanalyse kurz dargestellt und ebenfalls exemplarisch anhand der Überlegungen durchgeführt, die bei der Diskriminanzanalyse aufgestellt wurden. Es folgt in Kapitel 6.4 ein tabellarischer Vergleich, welcher die Überleitung für das vergleichende Fazit und einen wissenschaftlichen Ausblick in Kapitel 7 ist.

## 1.3. Abgrenzung

Die Arbeit richtet sich an Interessierte, Anwender und Anwenderinnen quantitativer Methoden, die mit allgemeinen Grundkonzepten der angewandten Statistik vertraut sind, Basiswissen mit der SPSS Anwendung vorweisen und ihre Methodenkompetenzen um die Durchführung einer Diskriminanzanalyse mit SPSS erweitern möchten.

Um dies zu erreichen, wird das Verfahren der Diskriminanzanalyse erklärt und anschließend, unterfüttert mit Bildschirmaufnahmen aus der SPSS Software, schrittweise durchgeführt. Es werden Stück für Stück Voraussetzungen und Überlegungen, die der Analyse zugrunde liegen, beschrieben und anhand der einzelnen Software-Menu-Objekte Inputs und Outputs der SPSS Software sowie ihre sachliche Bedeutung anhand des Beispieldatensatzes entschlüsselt.

Im Laufe der Vorstellung der statistischen Verfahren wird weniger über die Mathematik des statistischen Verfahrens gesprochen als viel mehr über die Anwendung des Verfahrens und dessen Voraussetzungen. Für die mathematische Herleitung und Beweise bestimmter Elemente wird auf geeignete Literatur und wissenschaftliche Veröffentlichungen verwiesen. Es ist nicht die Absicht dieser Arbeit dem Leser das Verfahren und dessen einzelne Schritte mathematisch herzuleiten, sondern ihn zu befähigen selbst eine Diskriminanzanalyse durchführen und die allgemeine Vorgehensweise nachvollziehen zu können.

[3] Aufführungen vor einer Filmpremiere vor einem Testpublikum zu Marktforschungszwecken
[4] erklärende Faktoren, beschreibende Variablen

# 2. Definition wichtiger Begriffe und Verfahren

In diesem Kapitel werden umfangreich relevante statistische Schlüsselbegriffe, Tests und Verfahren definiert.

## 2.1. Schlüsselbegriffe

2.1.1. **a-priori-Wahrscheinlichkeit**: Die a-priori-Wahrscheinlichkeit ist die Wahrscheinlichkeit, dass eine Beobachtung einer Gruppe angehört, bevor die Daten erfasst werden[5].

2.1.2. **a-posteriori-Wahrscheinlichkeit**: Die a-posteriori-Wahrscheinlichkeit ist die Wahrscheinlichkeit, dass eine Beobachtung einer Gruppe angehört, nachdem ein Modell geschätzt wurde, welches die Wahrscheinlichkeit, dass eine Beobachtung einer Gruppe angehört beeinflusst. In der Terminologie der statistischen Entscheidungstheorie werden die Klassifizierungswahrscheinlichkeiten als A-posteriori-Wahrscheinlichkeiten bezeichnet[6].

2.1.3. **Skalenniveau**[7]: Das Skalenniveau oder auch Messniveau gibt wieder, was für eine Art von messbarer Eigenschaft die Ausprägungen einer Variable wiedergibt.

- **Nominalskala**: Unterschiedliche Merkmalsausprägungen ohne Rangordnung werden als nominalskaliert bezeichnet.

  *Beispiele:*
  - *Farben: grün, gelb, blau (usw.)*
  - *Familienstand: ledig, verheiratet, geschieden (usw.)*

- **Metrische Skala**: Wenn sowohl die Rangordnung als auch die Abstände zwischen Merkmalsausprägungen bestimmbar sind, dann liegen metrisch skalierte Variablen vor.

  *Beispiele:*
  - *Umsätze: 100T€ sind 40T€ mehr als 60T€*
  - *Temperaturen: 40°C sind doppelt so viel wie 20°C*

- **Ordinalskala**: Liegt eine Rangordnung mit natürlicher Reihenfolge zwischen verschiedenen Merkmalsausprägungen vor, so spricht man von ordinal skalierten Variablen.

  *Beispiele:*
  - *Schulnoten: Note 1,2,3 ... bis 6*
  - *Güteklassen: sehr gut, gut, mittel, schlecht, sehr schlecht*

  *(können für Rechenoperationen in Intervallskalen[8] mit metrischen Werten überführt werden)*

2.1.4. **Grundgesamtheit**[9]**:** Die Grundgesamtheit, auch Population, ist die Menge an Objekten, deren Merkmale untersucht werden können. Eine **Stichprobe** ist ein Teil einer Grundgesamtheit; ihre Größe ist immer begrenzt. Informationen, die anhand der Grundgesamtheit gewonnen werden, werden Parameter genannt. Durch das Betrachten von Stichproben kann auf die Grundgesamtheit geschlossen werden, allerdings nennt man Informationen Schätzwerte.

[5] vgl. Backhaus 2016, S. 247
[6] vgl. Backhaus 2016, S. 250
[7] vgl. Kapitel 2.2 Mittag 2016, 18 ff., Krol et al. 20.05
[8] Skala mit gleichgroßen Abständen ohne natürlichen Nullpunkt
[9] vgl. Mittag 2016, S. 16

2.1.5. **Irrtumswahrscheinlichkeit p**[10]: Viele statistische Verfahren testen, ob eine bestimmte Verteilung (z.B. T-Verteilung, Chi-Quadrat-Verteilung, F-Verteilung) der Testergebnisse der statistischen Tests mit einer bestimmten Wahrscheinlichkeit auf zufälliger Basis zustande gekommen ist oder ob diese vom Zufall abweicht. Hierfür werden zwei Hypothesen aufgestellt:

- **Nullhypothese**: die Konstellation der Dichtefunktionen ist zufälliger Natur, es gibt keine nicht-zufälligen Unterschiede in den Stichproben
- **Alternativhypothese**: die Konstellation der Dichtefunktionen ist keiner zufälligen Natur, es gibt nicht-zufällige Unterschiede in den Stichproben.

  Das Ablehnen der einen Hypothese führt zum nicht-Ablehnen der anderen und vice versa.

Die Irrtumswahrscheinlichkeit p gibt wieder, wie hoch die Wahrscheinlichkeit ist, dass man sich *irrt,* die Nullhypothese abzulehnen, also zu einer verkehrten Schlussfolgerung anhand des vorliegenden Tests zu kommen, und fälschlicherweise die Alternativhypothese annimmt. „Irrtümlicherweise hat man Ergebnis A abgelehnt und sich für B entschieden, obwohl Ergebnis A richtig gewesen wäre." Diesen Sachverhalt nennt man **Fehler 1. Art** oder auch **alpha-Fehler**. Die Wahrscheinlichkeit einen solchen Fehler zu begehen, kann über die Festlegung eines Signifikanzniveaus kontrolliert werden. Das **Signifikanzniveau**, die erforderliche Deutlichkeit[11] eines Ergebnisses, kann je nach Wissenschaftsbereich oder Anwendungsgebiet variieren. Meist wird es bei 0.05 festgelegt. Das bedeutet, dass p-Werte unterhalb des Schwellwerts signifikant sind und Werte darüber nicht signifikant.
Der p-Wert wird von der Statistiksoftware SPSS als Testergebnis ausgegeben. Basierend auf ihm wird die Entscheidung für oder gegen die Nullhypothese getroffen.

| p-Wert | Bedeutung | Schlussfolgerung |
|---|---|---|
| $p \leq 0{,}05$ | signifikant | Nullhypothese wird abgelehnt, Alternativhypothese wird nicht abgelehnt |
| $p > 0{,}05$ | nicht signifikant | Nullhypothese wird nicht abgelehnt, Alternativhypothese wird abgelehnt |

*Tabelle 2-1 – Erklärung der p-Werte*

2.1.6. **Klassierung:** Bei einer Klassierung (Gruppierung, Klasseneinteilung) werden Merkmalsausprägungen zusammengefasst und in getrennte Gruppen eingeteilt. Die Gruppen, auch Klassen genannt, fassen benachbarte Merkmalsausprägungen zusammen (Intervalle)[12].

2.1.7. **Statistische (Häufigkeits-)Verteilung**[13]: Die Verteilung der Ergebnisse einer Messung kann als Häufigkeitsverteilung angesehen werden. Mathematisch gesehen ist eine Häufigkeitsverteilung eine Dichtefunktion, die zu jedem vorgekommenen Wert angibt, wie häufig dieser Wert vorgekommen ist. Je nach Ausprägung der Lageparameter können Häufigkeitsverteilungen verschiedene Verteilungsgestalten annehmen.

[10] Bühl 2016, 176 f.; Skript: Tuschl 2016, Hypothesenprüfung
[11] Latein significans (dt. deutlich)
[12] Krol et al. 20.05
[13] vgl. Kapitel 4 in Mittag 2016, Fantapié Altobelli 2017, S. 233

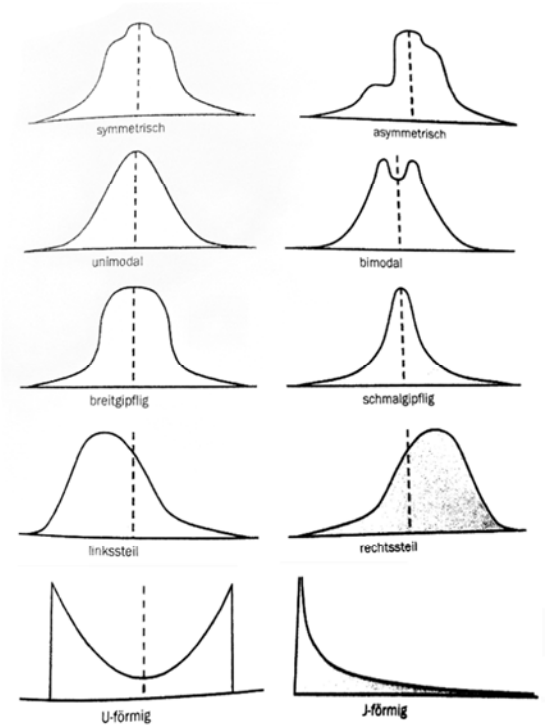



*Abbildung 1 - ausgewählte idealtypische Formen von Häufigkeitsverteilungen, Quelle: S.233 Marktforschung, 3. Aufl., Altobelli*

2.1.8. **Normalverteilung[14]:** Als wichtigste Voraussetzung für die Durchführung von vielen statistischen Verfahren gilt neben der Forderung nach einem bestimmten Skalenniveau die annähernde Normalverteilung der Daten. Normalverteilungen sind theoretische Verteilungen, die von dem Mathematiker Gauß untersucht wurden und deshalb auch als "Gauß'sche Glockenkurven" bezeichnet werden. Es gibt unendlich viele Normalverteilungen. Allen gemeinsam ist es, dass sie symmetrisch um den Mittelwert sind, "glockenförmig" sind und asymptotisch gegen 0 laufen. Die Schiefe ist somit immer Null. Außerdem gilt, dass etwa 68% aller Messwerte innerhalb von $1 * Standardabweichung\ \sigma$, 95% innerhalb von zwei $2 * \sigma$ und 99.7% aller Werte innerhalb von $3 * \sigma$ von $Mittelwert\ \mu$ liegen.

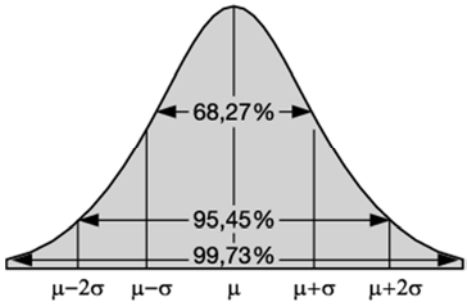



*Abbildung 2 - Normalverteilung, Quelle: Gabler Wirtschaftslexikon*

2.1.9. **SPSS[15]:** SPSS ist eine Software, die 1967 von zwei Studenten erdacht wurde, um statistische Datenanalysen von einem Computer erledigen zu lassen. Damals zur Verfügung stehende Programme erwiesen sich als ungeeignet und bereiteten den beiden Studenten Frustration, weswegen sie „Statistical Package for Social Science" inklusive eigener Steuersprache, der Syntax, entwickelten. Später etablierte sich das Kürzel SPSS. Seither hat sich das Programm stetig weiterentwickelt und gilt mittlerweile als das weltweit verbreitetste Anwendersystem zur

[14] Universität Zürich, http://www.methodenberatung.uzh.ch/de/datenanalyse/deskuniv.html
[15] Bühl 2016, S. 38

statistischen Datenanalyse. Auf Grund der vielseitigen Einsatzmöglichkeiten und der weiten Verbreitung entschied man sich dazu, dem Kürzel eine neue Bedeutung zu verleihen: „Superior Performance Software System“. Der vollständige Name des Programms setzte sich im Gegensatz zur Abkürzung „SPSS“ bei den Anwendern jedoch nie durch.

Das Produkt und die Marke „SPSS“, welche 2009 von IBM, einem weltweit führenden Soft- und Hardware Konzern, aufgekauft wurde, werden stetig weiterentwickelt und Funktionen aktualisiert, optimiert und ergänzt. Inzwischen ist die Version IBM SPSS 25 erhältlich[16]. *(In dieser Arbeit wird die Version SPSS 19 verwendet. Für den Funktionsumfang der Software für das Verfahren der linearen Diskriminanzanalyse, mit der sich die Arbeit auseinandersetzt, spielt die Aktualität der Version keine wichtige Rolle.)*

2.1.10. **Kovarianz[17]:** Die Kovarianz misst die durchschnittliche Übereinstimmung in der Streuung von zwei Merkmalen. Sie ist Ausdruck der Stärke des linearen Zusammenhangs. Ist die Kovarianz positiv, dann gehen kleine Werte der einen Variablen überwiegend einher mit kleinen Werten der anderen Variable und gleichfalls für große Werte. Für eine negative Kovarianz ist das genau umgekehrt.

2.1.11. **Kovarianz-Matrix:** Eine Kovarianz-Matrix ist die quadratische und symmetrische Matrix aus allen möglichen Kombinationen ihrer Kovarianzen. Die Spiegelachse der Kovarianz-Matrix bildet die Varianzen ab, weshalb die Kovarianz-Matrix auch Kovarianz-Varianz-Matrix genannt wird.

2.1.12. **Varianz $s^2$:** Ein Maß für die Streuung der Messwerte[18]. Sie bezeichnet die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert, dividiert durch die Freiheitsgrade, also um 1 verminderte Anzahl der Messwerte $(n-1)$.

$$s^2 = \frac{1}{n-1} \cdot \sum_{1-i}^{n} Quadrierte\ Abweichung\ aller\ Messwerte\ m_i vom\ Mittelwert\ \mu_i$$

*(Formel 2:1 - Varianz)*

$$Standardabweichung = \sigma = \sqrt{s^2} = \sqrt{(m_i - \mu_i)^2}$$

*(Formel 2:2 - Standardabweichung)*

2.1.13. **Verteilung (Distribution)[19]:** *Die Schiefe* ist ein Maß für die Abweichung einer Häufigkeitsverteilung von einer normalverteilten Grundgesamtheit und kann zum Test auf diesen Unterschied benutzt werden. *Kurtosis* beschreibt die Breite des Gipfels der Verteilung. Bei beiden Kennwerten deutet der Wert "Null" auf eine Normalverteilung hin.

2.1.14. **Zusammenhangsmaße[20]:** In der Statistik wird der Zusammenhang zwischen zwei statistischen Variablen mit verschiedenen Zusammenhangsmaßen (Koeffizienten) gemessen. Statistische Zusammenhänge werden auch Korrelation genannt. Es gibt verschiedene Verfahren, um Rückschlüsse über mögliche Korrelationen zu gewinnen.

## 2.2. Erwähnenswerte Verfahren und statistische Tests

[16] Bühl, SPSS 23, S.38f
[17] vgl. Kapitel 9.2 in Mittag 2016
[18] Krol et al. 20.05
[19] Krol et al. 20.05
[20] vlg. Kapitel 9 von Mittag 2016

2.2.1. **Box's M Test [21]:** Statistischer Test, der die Gleichheit (Homogenität) der Kovarianz-Varianz-Matrizen der klassenunabhängigen Variablen von der abhängigen Variable testet.

2.2.2. **Clusteranalyse**: Exploratives Verfahren zum Ermitteln von natürlichen Gruppierungen von Objekten. Ziel der Clusteranalyse ist es, eine heterogene Gesamtheit von Objekten anhand relevanter Merkmale in Gruppen (Cluster) einzuteilen. Dabei sollen die klassifizierten Objekte innerhalb der Gruppe möglichst ähnlich und die Gruppen untereinander möglichst unähnlich sein[22].

2.2.3. **Dependenzanalysen**: Unter Dependenzanalysen werden Verfahren verstanden, mit denen Strukturen überprüft werden sollen. Zu diesen Verfahren gehören beispielsweise die Varianzanalyse, Regressionsanalyse, Diskriminanzanalyse, t-Tests, Chi-Quadrat-Tests etc. Ziel der Analysen ist es, Abhängigkeiten (Dependenzen) zwischen abhängigen und unabhängigen Variablen zu untersuchen. Entsprechend werden die Variablen daher vor Anwendung der entsprechenden Methode in abhängige (zu erklärende) und unabhängige (erklärende) Variablen aufgeteilt. Das Gegenteil von Dependenzanalysen sind Interpendenzanalysen. Sie untersuchen die Wechselwirkungen von Variablen untereinander.

2.2.4. **Klassifikationsverfahren:** Klassifikationsverfahren sind Verfahren der Datenanalyse, die das Ziel haben, Objekte einer Population anhand ihrer Merkmalsausprägungen in Teilgruppen aufzusplittern. Dieser Vorgang kann in zwei Schritte eingeteilt werden[23]. Im ersten Schritt, der Lernphase oder Trainingsphase, wird aus einer vorgegebenen Stichprobe ein Modell erzeugt, das die Daten innerhalb dieser Stichprobe beschreibt. Es wird dabei angenommen, dass jedes Objekt der Stichprobe einer bestimmten Klasse angehört. Da für jedes Objekt des Lernstichprobees die Klassenzugehörigkeit bereits zur Verfügung steht (oder festgelegt wurde), ist dieser erste Schritt als überwachtes Lernen (engl. „supervised learning“) bekannt. Im Gegensatz dazu ist beim unüberwachten Lernen (engl. „clustering“) die Klassenzugehörigkeit unbekannt.

Typischerweise wird das gelernte Modell in Form von Klassifikationsregeln, Entscheidungsbäumen oder mathematischen Formeln festgelegt. Diese Klassifikationsregeln können sowohl dazu verwendet werden, die Gruppen anhand ihrer Merkmale zu untersuchen, als auch die Klassenzugehörigkeit von zukünftigen Datensätzen vorherzusagen.

Der zweite Schritt ist die Anwendung des Modells zur Klassifikation. Um ein realistisches Szenario zu simulieren, werden die restlichen Objekte dazu verwendet, die Vorhersagegenauigkeit des Modells abzuschätzen. Anhand jedes dieser restlichen Objekte wird die bekannte Klassenzugehörigkeit mit der vom Modell vorhergesagten Klassenzugehörigkeit verglichen. Sofern das Modell als *akzeptabel* eingeschätzt wird, kann es anschließend prognostizierend zur Klassifikation von noch nicht klassifizierten Objekten eingesetzt werden.

2.2.5. **Lineare und nichtlineare Modelle:** Ein Modell ist dann linear, wenn die Elemente aus einer Konstanten und Produkten eines Parameters mit einer Prädiktorvariable bestehen. Eine lineare Gleichung ist das Konstrukt der additiven Aneinanderreihung der Ergebnisse dieser Elemente. Die Basisform einer linearen Gleichung sieht wie folgt aus:

2.2.6.

$$Funktionswert = Konstante_c + Parameter_1 \cdot Prädiktor_1 + Parameter_2 \cdot Prädiktor_2 \ldots + Parameter_n \cdot Prädiktor_n$$

[21] IBM Knowledge Center, Box's M Test
[22] IBM Knowledge Center, Clusteranalyse; Skript: QM im Marketing – Prof. Dr. Stefan Tuschl
[23] Moritz Duhme, Ansätze zur Konstruktion von Entscheidungsbäumen

*(Formel 2:3 – Basisform einer linearen Gleichung)*

Nichtlinear ist ein Modell immer dann, wenn die Kriterien für ein lineares Modell gebrochen werden, also in den Fällen, wo die Aneinanderreihung der Ergebnisse dieser Elemente durch nicht-additive Weise geschieht. Für diese Form der Gleichung gibt es viele Formen.

2.2.7. **Multivariate Verfahren:** Die multivariaten Verfahren sind durch die Untersuchung der Beziehungen zwischen drei und mehr Variablen gekennzeichnet. Davon abzugrenzen sind univariate Verfahren, bei denen die Merkmalsausprägungen einer einzigen Variable betrachtet werden und bivariate Verfahren, bei denen die Beziehung zwischen genau zwei Variablen untersucht wird.
Es gibt zwei Arten von multivariaten Verfahren:

- **explorative, strukturenentdeckende Verfahren**: Unter strukturenentdeckenden Analysemethoden versteht man solche Verfahren, deren Ziel die Entdeckung von bisher unbekannten Zusammenhängen zwischen Variablen oder Objekten ist.
- **konfirmatorische, strukturenprüfende Verfahren**: Die strukturprüfenden Verfahren überprüfen die Zusammenhänge zwischen Variablen, wenn über den Zusammenhang vor Anwendung der Analyse bereits Hypothesen existieren.

2.2.8. **Verteilungsfreie und verteilungsabhängige Verfahren:** In der Statistik existieren eine Vielzahl von Testverfahren, die sich in verteilungsfreie und verteilungsgebundene Prüfverfahren einteilen lassen. Verteilungsgebundene Prüfverfahren (auch: parametrische Tests) setzen Normalverteilung der betrachteten Variablen voraus. Verteilungsfreie Prüfverfahren hingegen (auch: nichtparametrische Tests) kommen ohne Normalverteilungsvorausetzung aus. Je nach Gegenstand der Prüfung lassen sich statistische Tests danach unterscheiden, ob sie Parameter einer Verteilung oder eine Verteilung als Ganzes überprüfen. Parameter sind typischerweise Lageparameter, wie der Mittelwert, oder Streuungsparameter, wie die Varianz, wohingegen beim Test einer gesamten Verteilung geprüft wird, ob die Verteilung der gemessenen Werte einer theoretischen Verteilung folgt.[24]

2.2.9. **Varianzanalyse[25]:** Die Varianzanalyse ist eine der allgemeinsten statistischen Analysemethoden. Sie ist ein Mittelwerttest für mehrere Stichproben. Die Varianz der zusammengefasst betrachteten Gruppen wird mit der Varianz innerhalb der einzelnen Gruppen in Beziehung gesetzt. Die abhängige Variable muss intervallskaliert sein; die unabhängige Variable nominalskaliert. Das Ziel ist die Klärung der Frage, ob sich die Mittelwerte einer oder mehrerer abhängiger Variablen für Gruppen von Fällen verursacht durch eine unabhängige Variable (ANOVA) oder mehrere unabhängige Variablen (MANOVA), signifikant unterscheiden. Zu unterscheiden sind:

- ANOVA (Analysis of Variance):
  - einfaktorielle Varianzanalyse: Einfluss einer abhängigen Variablen auf eine unabhängige Variable; es werden Rückschlüsse auf die Grund-Gesamtheit gemacht
  - mehrfaktorielle Varianzanalyse: Einfluss mehrerer abhängiger Variablen auf eine unabhängige Variable
- MANOVA (Multiple Analysis of Variance):
  - mehrfaktorielle Varianzanalyse: Einfluss mehrerer abhängiger Variablen auf mehrere unabhängige Variablen *(die mehrdimensionale Varianzanalyse ist keine Hintereinander-Ausführung von mehrfaktoriellen Varianzanalysen, da die abhängigen Variablen auch untereinander voneinander abhängen können.)*

[24] Altobelli, S.236
[25] vgl. Kapitel 3 von Backhaus 2016

# 3. Einführung in die lineare Diskriminanzanalyse

Im folgenden Kapitel soll die lineare Diskriminanzanalyse näher dargestellt werden. Ausgehend davon wird die Diskriminanzanalyse schließlich anhand eines Datensatzes zur Veranschaulichung exemplarisch durchgeführt. Zweck der Diskriminanzanalyse ist die Untersuchung der Beziehung zwischen einer nicht-metrisch skalierten abhängigen Variable mit metrischen unabhängigen Variablen. Bei Vorliegen einer derartigen Datenstruktur besteht die Aufgabenstellung der Diskriminanzanalyse darin, Gruppenunterschiede zu untersuchen, indem versucht wird, die Zugehörigkeit von ausgewählten Objekten zu a-priori vorgegebenen Gruppen anhand ihrer Ausprägungen bei zwei oder mehr metrisch skalierten Merkmalen zu erklären sowie zu prognostizieren.

## 3.1. Was ist die Diskriminanzanalyse und wozu dient sie?

Die Diskriminanzanalyse ist eine dependanzanalytische, strukturprüfende Methode der multivariaten Statistik[26], welche die Abhängigkeit zwischen nominal-skalierten Klassen oder Gruppen innerhalb einer Gesamtheit anhand der Unterschiede ihrer metrisch-skalierten Variablen untersucht. Anders als bei strukturentdeckenden (explorativen) Verfahren, wird dabei a priori[27] die Existenz von Klassen oder Gruppen innerhalb der Gesamtheit vorausgesetzt[28]. Für die betreffende Untersuchung werden die Merkmalsvariablen, die die unterschiedlichen charakteristischen Merkmale der verschiedenen Teilgruppen abbilden, in einer mathematischen Funktion derart kombiniert, dass der ermittelte Funktionswert eine Aussage über die Gruppenzugehörigkeit der Objekte ermöglicht.[29]
Den entscheidenden Grundgedanken[30] zur linearen Diskriminanzanalyse, welche in dieser Arbeit vorgestellt wird, hat R. A. Fisher 1936 in „The use of multiple measurements in taxonomic problems“ veröffentlicht:

Wenn zwei oder mehr Populationen anhand verschiedener Merkmale gemessen würden, galt besonderes Interesse derjenigen linearen Funktion dieser Messungen, durch die die Teilpopulationen am besten voneinander unterschieden werden. Fragen bezüglich der Genauigkeit des statistischen Prozesses würden ebenfalls diskutiert werden, schreibt R.A. Fisher in der Einleitung seiner wissenschaftlichen Veröffentlichung und führt anschließend anhand von Beobachtungen von verschiedenen Pflanzen sein Analyseverfahren durch[31].

> *“When two or more populations have been measured in several characters, $x_1$,…, $x_s$, special interest attaches to certain linear functions of these measurements by which the populations are best discriminated. (…) Questions connected with the precision of the process employed will also be discussed.”*

*Textausschnitt 1 - The use of multiple measurements in taxonomic problems, 1936, R.A. Fisher*

Dabei weist er auf einige mathematische Zusammenhänge hin, die sich später als die Basis der linearen Diskriminanzanalyse etablieren werden. Besonders die Überlegung, eine lineare Trennfläche als Trennkriterium zu nutzen, welche die Varianz der Messungen berücksichtigt, spielen eine große Rolle für die anhaltende Relevanz des Ansatzes. Dabei wird über das Trennkriterium der Schwellenwert für die Diskriminanzregel gewonnen. Je nachdem, ob der Wert des zu klassifizierenden Objekts größer oder kleiner als der Schwellenwert ist, wird das Objekt der dazugehörigen Gruppe zugeordnet.

Bis heute werden immer wieder neue Formen der Diskriminanzanalyse in die Diskussion eingebracht bzw. bestehende Formen weiterentwickelt. Der traditionelle, lineare Ansatz von Fisher ist allerdings auch heute

[26] Vgl. 2.2.7 und 2.2.3
[27] deutsch: *„von vornherein“*
[28] Vgl. Malhotra 2007, 5. Auflage
[29] Handbuch der Sozialwissenschaften
[30] Backhaus 2016, S. 217
[31] Fisher 1936

noch weit verbreitet, was vor allem auf seine gute Performance und sein relativ einfaches mathematisches Gerüst zurückzuführen sein dürfte.[32]

### 3.1.1. Untersuchungsziele der Diskriminanzanalyse

Die Anwendung der Diskriminanzanalyse kann generell drei Untersuchungsziele haben[33].

Der *diagnostische Ansatz* fokussiert sich auf die Untersuchung der Unterschiede der Gruppenmittelwerte und Standardabweichungen der Gruppen, die auf Grund von den ausgewählten Prädiktoren auftreten.

Der *prognostische Ansatz* verfolgt die Überlegung, noch nicht klassifizierte Fälle anhand ihrer Merkmalsvariablen mit Hilfe der Diskriminanzfunktion mit einer gewissen Wahrscheinlichkeit voraussagend einer Klasse oder Gruppe zuweisen zu wollen.

Der dritte Ansatz verfolgt die *Reduzierung der Dimension des Analyseproblems* durch die Identifikation derjenigen Merkmale, die die größte Bedeutung für die Erklärung der Gruppenunterschiede bzw. Gruppenzugehörigkeiten haben bei gleichzeitiger „Entwichtung“[34] derjenigen Merkmale, die keinen Informationsgehalt zum Analyseproblem beitragen.

Ergebnisse und Prognosen anhand der Analyse können dann zur Entwicklung von zielgruppenspezifischen Maßnahmen dienen, um übergeordnete Ziele zu erreichen.

### 3.1.2. Methodische Grundüberlegungen und Voraussetzungen

Durch die lineare Kombination der Prädiktoren wird eine neue metrisch skalierte Merkmalsvariable erzeugt, welche die bestimmte Eigenschaft haben soll, die Objekte anhand der unterschiedlichen Ausprägungen der Prädiktoren zu teilen. Die neue Merkmalsvariable, die Diskriminanzvariable, ist das Ergebnis einer Dimensionsreduktion durch eine Funktion, die die Ausprägungen der Merkmalsvariablen pro untersuchtem Objekt in gewichteter Form aufsummiert[35]. Sie wird daher auch als *kanonische*[36] *Variable* bezeichnet und die Diskriminanzfunktion *kanonische Diskriminanzfunktion*[37]. Die Gewichtung findet dabei anhand der Einflussstärken der einzelnen Merkmalsvariablen statt. Dabei liegt das Augenmerk darin, die Dimensionen in einer solchen Weise zu reduzieren, dass die Informationen der dazu verwendeten Prädiktoren hinsichtlich angenommener Gruppenmittelwertunterschiede in Abhängigkeit einer Gruppierungsvariablen nicht verloren gehen. Die Gewichtungen der erklärenden Merkmalsvariablen werden auch Diskriminanzkoeffizienten genannt. Sie sorgen dafür, dass die gewünschten Eigenschaften der Diskriminanzvariable optimiert sind und aus der Funktion resultierende Werte die optimale Separation der Gruppen erreicht.

[32] Wolf und Best 2010, S. 495.
[33] Wolf und Best 2010, S. 496., vgl. Hair et al., vgl. Backhaus 2016, S. 216
[34] im Sinne von Entkräftigung, weniger gewichtet
[35] Wolf und Best 2010, S. 497
[36] „kanonisch“: Die kanonische Einbettung ist eine mathematische Funktion, die eine Teilmenge in ihre Grundmenge einbettet. Wird auch Inklusionsabbildung genannt (kurz: Inklusion)
[37] Backhaus 2016, S. 221; Müller 2015/2016

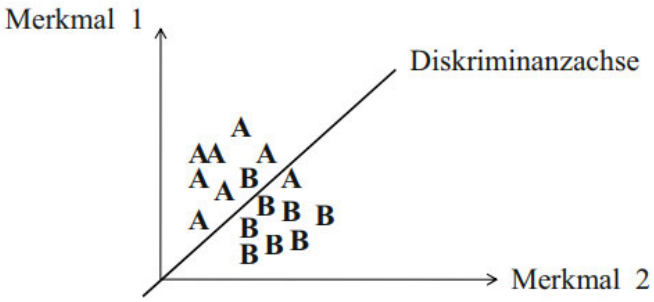



*Abbildung 3: Grafische Darstellung der Separation der Gruppen im Zwei-Gruppen-Fall, Quelle: Wolf und Best 2010, S. 498*

Geometrisch ausgedrückt handelt es sich bei *einer* Diskriminanzfunktion um eine Achse, der Diskriminanzgeraden, welche die Eigenschaft hat, die zu untersuchenden Objekte an einer Stelle in ihre Untergruppen zu trennen (siehe Abbildung 3: Grafische Darstellung der Separation der Gruppen im Zwei-Gruppen-Fall, Quelle: Wolf und Best 2010, S. 498). Dabei befinden sich die Objekte in dem Merkmalsraum, welcher durch die dazugehörigen Merkmalsvariablen aufgespannt wird. Jede Merkmalsvariable definiert eine Achse des Raumes. Im Mehr-Gruppen-Fall können *mehrere* Diskriminanzfunktionen erzeugt werden, sodass aus der Achse ein Raum bzw. ein Hyperraum wird. Die durch die Diskriminanzfunktionen erzeugten Geraden werden auch als Diskriminanzachsen bezeichnet, deren Steigung durch das Verhältnis der entsprechenden Diskriminanzkoeffizienten definiert ist. Für eine Separation der Gruppen wird jedes Objekt über seinen Diskriminanzwert auf jeder Diskriminanzachse abgetragen. Anschließend wird der Mittelwert für jede der Gruppen bestimmt. Die einzelnen Objekte werden dann jener Gruppe zugeordnet, zu deren Mittelwert sie die geringste Distanz aufweisen. Sollen mehr als zwei Gruppen untersucht werden, so können anhand mehrerer Gruppen-Zentroide mehrere Diskriminanzfunktionen bestimmt werden. Am Ende ist jedes Objekt als die Kombination der Diskriminanzwerte der Diskriminanzfunktionen darstellbar. Der Merkmalsraum besitzt gegenüber dem Diskriminanzraum immer eine höhere Dimension als der Diskriminanzraum[38]. Die Diskriminanzanalyse, deren Durchführung gleichzeitig einen mindestens eindimensionalen Diskriminanzraum *(eine lineare Funktion = eine Achse = eine Dimension = 1-dimensionaler Raum, zwei lineare Funktionen = zwei Achsen = zwei Dimensionen = 2-dimensionaler Raum und so weiter)* erzeugt, kann deshalb auch als eine Transformation der Ausgangsdaten in einen niedriger dimensionierten Raum aufgefasst werden. Distanzen zu den Gruppenmittelwerten der mehrdimensionalen Räume lassen sich dann nur noch über multivariate Proximitätsmaße[39], wie das Mahalanobis-Distanzmaß, berechnen.

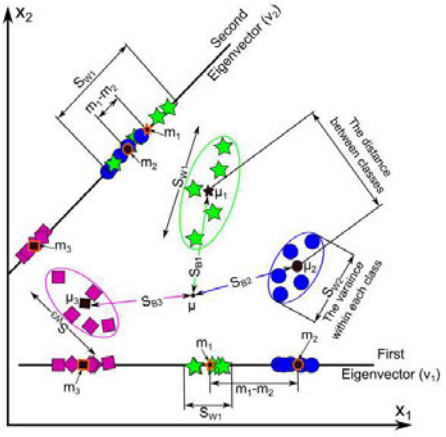



*Abbildung 4 - Ein Visueller Vergleich zweier reduzierter Dimensionen anhand der Betrachtung eines Falls mit 3 Gruppen, Quelle: (Tharwat et al.)*

[38] Wolf und Best 2010
[39] Oberbegriff für eine Kategorie von Maßen, die Ähnlichkeiten und Unterschiede ausdrücken

Die Abbildung zeigt den Mehr-Gruppen-Fall in einem Merkmalsraum der Merkmale $x_1$ und $x_2$ mit drei Gruppen, die farblich (blau, grün, lila) und symbolisch (Kreis, Stern, Viereck) voneinander abgegrenzt sind. Ihre Gruppen-Zentroide im Raum sind $\mu_1$, $\mu_2$ und $\mu_3$. Durch die Diskriminanzanalyse wurden zwei Diskriminanzachsen (Eigenvector) $v_1$ und $v_2$ ermittelt. Auf ihnen sind die Objekte des Merkmalsraum abgetragen. Es wurden dadurch aus den zwei Dimensionen (Merkmalsraum $x_1|x_2$) eine jeweils neue Dimension geschaffen, $v_1$ bzw. $v_2$. Die neuen Diskriminanzdimensionen, $v_1$ und $v_2$, sind hinsichtlich der der Eigenschaft optimiert, eine Gruppe optimal von den anderen Gruppen zu separieren. Hierfür wird ein Distanzmaß, welches in Kapitel 3.2.3 näher beschrieben wird, durch die Maximierung die Streuung zwischen den Gruppen $m_1$, $m_2$ und $m_3$ bei gleichzeitiger Minimierung der Streuung in den Gruppen optimiert. Eine abgetragene Dimension sollte nicht durch eine weitere erklärt werden können, sondern zusätzliche Informationen für die entstehende Dimension bieten. Eine hohe Korrelation der erklärenden Merkmalsvariablen untereinander sollten folglich vermieden werden. Außerdem dienen Dimensionen keinem erklärenden Zweck, in denen die Objekte der Gruppen zu sehr beieinander liegen und dabei zu sehr ineinander streuen. Signifikante Gruppenmittelwertunterschiede der erklärenden Variablen sollten daher vorhanden sein.

Das Prozedere der Diskriminanzanalyse unterliegt demnach folgenden Voraussetzungen:

Als erstes muss eine Stichprobe oder Gesamtheit vorliegen, die in zwei oder mehr disjunkte Teilpopulationen (Teilmengen, Untergruppen) zerlegbar ist. Jedes Objekt darin gehört genau einer Gruppe an. Des Weiteren dürfen keine fehlenden Werte innerhalb der Merkmalsausprägungen der Objekte vorkommen. Auch die Gruppenzugehörigkeiten für jedes Objekt müssen bekannt sein. Der Stichprobenumfang sollte dabei mindestens doppelt so groß sein, wie die Anzahl der Merkmalsvariablen, die für die Analyse verwendet werden. Die Anzahl der erklärenden Merkmalsvariablen sollte dabei die Anzahl der Gruppen immer mindestens um 1 übersteigen. Hat die Gruppierungsvariable zwei Ausprägungen, spricht man von einer dichotomen Variable oder dem Zwei-Gruppen-Fall, liegen mehr Ausprägungen vor, handelt es sich um eine polytome Variable oder den sogenannten Mehr-Gruppen-Fall der Diskriminanzanalyse[40]. Von der Anzahl der Gruppen hängt auch die Anzahl der möglichen Diskriminanzfunktionen ab. Im Zwei-Gruppen-Fall kann durch die lineare Kombination der Prädiktoren nur eine Diskriminanzfunktion erzeugt werden.

Zusammengefasst lässt sich sagen, dass die Diskriminanzanalyse ein verteilungsabhängiges (parametrisches) Verfahren ist. Das heißt, dass bestimmte Parameter der für sie verwendeten metrischen Variablen vorliegen sollten. Hierzu gehören gleiche Kovarianzen der Prädiktoren in den abhängigen Klassen (Homogenität der Kovarianzen bzw. Varianzhomogenität der Gruppen). Außerdem sollen die Gruppen vollständig durch die Mittelwerte und Kovarianzen beschrieben werden, wofür eine multivariate Normalverteilung ein typisches Beispiel ist. Bezüglich der Verteilungsannahmen ist das Verfahren jedoch recht robust[41] (leichte Abweichungen sind zulässig). Des Weiteren können Ausreißer und Extremwerte die Wirkungsbeziehung zwischen den Prädiktoren und ihre Dichtefunktion beeinflussen, weshalb sie von der Betrachtung ausgeschlossen werden sollten. Wirkungsbeziehungen zwischen den erklärenden Variablen sollten nach Möglichkeit ebenfalls vermieden werden. Verwendet man einen Prädiktor, der teilweise durch einen anderen erklärt wird, so wird dem Prädiktor dadurch eine höhere Gewichtung zugewiesen. Es ist jedoch keine zwingend notwendige Voraussetzung, dass keine Korrelationen der erklärenden Variablen untereinander vorliegen, sondern viel mehr eine Empfehlung hinsichtlich der Interpretierbarkeit des Modells.

### 3.1.3. Abgrenzung zur quadratischen Diskriminanzanalyse und zu anderen ähnlichen statistischen Verfahren

Der Unterschied zwischen der linearen Diskriminanzanalyse und der quadratischen Diskriminanzanalyse ist die wegfallende Voraussetzung für gleiche Kovarianz-Varianz-Matrizen bei der quadratischen Diskriminanzanalyse. Es ist also keine statistische Normalverteilung der klassenspezifischen Beobachtung nötig. Das Verfahren der

[40] Wolf und Best 2010, S. 497
[41] Backhaus 2016, S. 219

quadratischen Diskriminanzanalyse ist also verteilungsfrei im Gegensatz zur linearen Diskriminanzanalyse, bei dem die multivariate Normalverteilung der Prädiktoren als Voraussetzung vorliegen sollte. Die lineare Diskriminanzanalyse ist demnach ein verteilungsabhängiges Verfahren. Die Trennfläche der quadratischen Diskriminanzanalyse ist außerdem nicht linear. Je nach Kombination verschieden vieler Parameter nimmt sie eine elliptische, parabolische oder hyperbolische Form an.

Welches Verfahren geeigneter ist, hängt unter anderem mit der Datengrundlage zusammen. Wenn viele Prädiktoren in ein Modell zur Erklärung der abhängigen Variable gefügt werden, steigt die Gesamtstreuung rapide an. Dagegen ist die quadratische Diskriminanzanalyse recht robust, allerdings zu Kosten eines hohen Schätzfehlers. Dieser könnte über eine große Lernstichprobe kompensiert werden. Liegt ein eher kleiner Stichprobenumfang und dementsprechend ein kleiner Lernstichprobe vor, ist die Verwendung der lineare Diskriminanzanalyse deutlich genauer und sinnvoller, denn mit ihr können die kleinen Varianzen sachgemäßer erklärt werden.

Gegenüber anderen vergleichbaren Verfahren der multivariaten Statistik sind eine Reihe methodischer Unterschiede hervorzuheben.

Bei einer metrischen Regressionsanalyse wird die Abhängigkeit einer metrisch skalierten Variablen von metrischen unabhängigen Variablen betrachtet. Die Diskriminanzanalyse dagegen betrachtet metrisch skalierte unabhängige Variablen und deren Einfluss auf eine nominalskalierte abhängige Variable.

Die logistische Regression untersucht den Zusammenhang zwischen der Wahrscheinlichkeit bestimmter Ereignisse und metrisch skalierten Prädiktoren mit Hilfe einer Maximum Likelihood-Schätzung[42]. Gerade für den Zwei-Gruppen-Fall besitzt die logistische Reggression bis auf das Schätz-Verfahren eine große Ähnlichkeit zur Diskriminanzanalyse[43].

Eine Varianzanalyse unterscheidet sich von der Diskriminanzanalyse, da die Dependanzen der Variablen verdreht sind. Die Varianzanalyse untersucht den Einfluss von nominalskalierten unabhängigen Variablen auf eine oder mehrere metrisch skalierte abhängige Variablen. Eine multiple Varianzanalyse (MANOVA) untersucht dabei den Einfluss mehrerer nominal skalierter Faktoren auf mehrere abhängige metrische Variablen. Das heißt, es werden im Gegensatz zur Diskriminanzanalyse auch Wechselwirkungen der nominalen Variablen zueinander untersucht.

Die Clusteranalyse wird im Zusammenhang mit der Diskriminanzanalyse ebenfalls des Öfteren erwähnt. Allerdings ist die Clusteranalyse ein explorativ, taxonomisches Verfahren, welches auf der Grundlage der Daten Gruppierungen strukturentdeckend erzeugt, während die Diskriminanzanalyse strukturprüfend a-priori vorhandene Gruppen untersucht und prüft, ob diese anhand bestimmter Merkmalsvariablen zu beschreiben und zu unterscheiden sind.[44]

## 3.2. Das Modell der Diskriminanzanalyse

Zu Beginn einer Diskriminanzanalyse muss das Untersuchungsdesign spezifiziert werden. Dies beinhaltet die Festlegung der Zielsetzung, die zu untersuchenden Gruppen sowie die Überprüfung der Gültigkeit der zugrundeliegenden Prämissen. Im Anschluss erfolgt die Bestimmung der Diskriminanzfunktion und des Diskriminanzkriteriums, welches zur Schätzung der Koeffizienten und für die Überprüfung der Trennfähigkeit der Funktion zwischen den Gruppen eingesetzt wird. Bei ausreichender Güte lassen sich die gewonnenen Funktionen dazu verwenden, die Gruppenzugehörigkeit neuer Objekte zu prognostizieren und die bekannten Gruppenzugehörigkeiten näher zu charakterisieren. Nachfolgend wird das methodische Vorgehen bei der Durchführung einer Diskriminanzanalyse für den linearen Fall ausführlicher anhand von folgenden Schritten *(in Anlehnung an Backhaus, 2016)* erläutert:

[42] Parametrisches Schätzverfahren der multivariaten Statistik, bei dem Prädiktoren die Wahrscheinlichkeit eines Ereignisses prognostizieren sollen
[43] vgl. Backhaus et al. 2016, S. 218 ff
[44] Müller 2015/2016

1. Definition der Gruppen
2. Formulierung der Diskriminanzfunktion
3. Schätzung der Diskriminanzfunktion
4. Prüfung der Diskriminanzfunktion
5. Prüfung der Merkmalsvariablen
6. Klassifikation neuer Elemente

### 3.2.1. Schritt 1: Definition der Gruppen

Im ersten Schritt muss eine kategoriale Variable mit $\geq 2$ Ausprägungen definiert werden, anhand derer die Diskriminanzanalyse durchgeführt wird. Die Diskriminanzanalyse ist ein strukturprüfendes Verfahren und kein taxonomisches Verfahren, wie beispielsweise das Clusterverfahren[45]. Das heißt, dass das Vorhandensein von Gruppen oder Klassen vorausgesetzt (a priori) und diese nicht erst durch das Verfahren erschaffen werden[46]. Gibt es eine solche Variable nicht, werden hypothetisch anhand eines der Untersuchung zugrundeliegenden theoretischen Modells oder einer im sachlogischen Zusammenhang mit dem Ziel der Untersuchung stehenden taxonomischen Analyse die Objekte einer Gruppe zugewiesen. Die Definition der Gruppen ist dabei gleichzeitig die Festlegung der Anzahl der Gruppen, anhand derer die Diskriminanzanalyse durchgeführt werden soll. Beim Diskriminanzanalyseverfahren unterscheidet man zwischen dem Zwei-Gruppen-Fall, bei dem die unabhängige Variable zwei Kategorien oder Gruppen aufweist, und dem Mehr-Gruppe-Fall, bei dem die unabhängige Variable n > 2 Kategorien oder Gruppen aufweist. Dabei ist zu beachten, dass die Anzahl der Gruppen G die Anzahl der Prädiktoren nicht übersteigen sollte.

### 3.2.2. Schritt 2: Formulierung der Diskriminanzfunktion

Die Formulierung der Diskriminanzfunktion erfordert eine Auswahl geeigneter Prädiktoren. Diese erfolgt zunächst hypothetisch und liegt einem theoretischen Modell oder einem sachlogischen Zusammenhang zugrunde[47].

Das Prinzip der Diskriminanzanalyse ist, mehrere Variablen bei minimalem Informationsverlust durch eine kanonische Linearkombination zu einer einzigen zusammenzufassen. Dies erfolgt über die Diskriminanzfunktion.

Das allgemeine lineare Diskriminanzmodell Y hat folgende Form:

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \dots b_n \cdot x_n$$

(Formel 3:1 – Allgemeines lineares Diskriminanzmodell)

Mit:

Y = abhängige Diskriminanzvariable
b0 = Konstante
bn = Koeffizient für unabhängige Variable bzw. Gewichtung
xn = unabhängige, erklärende Variable bzw. Prädiktor

[45] vgl. 2.2.2
[46] Backhaus 2016
[47] vgl. Backhaus 2016, S. 220

Zu sehen ist eine lineare Funktion mit der abhängigen Diskriminanzvariablen $Y$ und den Prädiktoren $x_n$. Die Koeffizienten $b_n$ heißen Diskriminanzkoeffizienten und stellen die Gewichtungen der Prädiktoren dar. Die Koeffizienten kann man auch als „Bedeutung eines Merkmals zur Erklärung von Gruppenunterschieden" deuten[48]. $b_0$ ist ein konstantes Glied und hat keine inhaltliche Bedeutung. Die Konstante bewirkt lediglich eine Skalenverschiebung, um beispielsweise den Gesamtmittelwert auf den Wert Null zu normieren.

Die Diskriminanzfunktion erzeugt durch die Verkettung von $n$-Zufallsvariablen eine neue metrische Variable aus einem $n$-dimensionalen Merkmalsraum. Dieser Sachverhalt verdeutlicht, dass die Diskriminanzanalyse unter anderem zur Reduktion von Dimensionen verwendet werden kann. Dabei ist vorausgesetzt, dass alle Prädiktoren linear-kombiniert multivariat normalverteilt sind. Eine multivariate Verteilung ist genau dann eine multivariate Normalverteilung, wenn alle Linearkombinationen der Komponenten univariate Normalverteilungen sind. Mit anderen Worten liegt eine multivariate Normalverteilung vor, sofern alle Prädiktoren normalverteilt sind. [49] Dies lässt sich durch verschiedene statistische Tests prüfen und gegebenenfalls durch die Transformation der Daten korrigieren. Das Ergebnis der Funktion, der Diskriminanzwert, ist nicht gleichzusetzen mit der nominalskalierten abhängigen Variable, die die Gruppenzugehörigkeit ausdrückt. Aus dem Diskriminanzwert wird allerdings die Gruppenzugehörigkeit abgeleitet[50].

Es wird angenommen, dass die ausgewählten Prädiktoren Unterschiede der in *Schritt 1* definierten Gruppen signifikant erklären können. Dieser Sachverhalt wird schließlich in *Schritt 5* getestet, um gegebenenfalls ungeeignete, nicht signifikante Prädiktoren aus der Funktion zu entfernen und/oder neue hinzuzufügen, wodurch sich *Schritt 2 bis 5* wiederholen.

### 3.2.3. Schritt 3: Schätzen der Diskriminanzfunktion

Nach der Auswahl der Prädiktoren müssen deren Gewichtungen ermittelt werden. Die Bestimmung der Diskriminanzkoeffizienten beruht auf der Grundüberlegung, dass eine bestmögliche Gruppentrennung dann gegeben ist, wenn einerseits die Gruppenmittelwerte der Diskriminanzwerte möglichst weit auseinanderliegen und andererseits die gruppeninternen Diskriminanzwerte möglichst gering um ihren jeweiligen Gruppenmittelwert streuen.

Die Überlegung, den Abstand der Gruppenmittelwert zueinander zu maximieren, würde vernachlässigen, dass Gruppen, die stark streuen, bei gleicher Distanz der Gruppenmittelwerte nicht so gut voneinander diskriminiert werden können, wie Gruppen mit geringerer Streuung. Deshalb optimiert man hinsichtlich eines diese Eigenschaften berücksichtigenden Diskriminanzkriteriums. Das Diskriminanzkriterium errechnet sich aus den Varianzen zwischen den Gruppen der abhängigen Variable und den Varianzen innerhalb der Gruppen der abhängigen Variable. Dabei sollen die Varianzen zwischen den Gruppen möglichst groß und die Varianzen innerhalb der Gruppen möglichst klein sein. Mit anderen Worten sollen die Gruppen möglichst heterogen und alle Objekte innerhalb einer Gruppe möglichst homogen sein.

$$\frac{Varianz\ zwischen\ den\ Gruppen}{Varianz\ innerhalb\ der\ Gruppen} = \Gamma$$

*(Formel 3:2 - Berechnung des Diskriminanzmaß)*

Mit:

$\Gamma =$ Diskriminanzmaß

[48] Müller 2015/2016
[49] Jörg Rahnenführer, Multivariate Verfahren
[50] Müller 2015/2016

Hieraus ergibt sich für $\Gamma$ ein mathematisches Optimierungsproblem. Für die mathematische Vorgehensweise zur Maximierung des Diskriminanzkriteriums wird Backhaus 2016, S. 274 empfohlen. Entscheidend ist, dass bei der Maximierung des Diskriminanzkriteriums als Ergebnis der sogenannte Eigenwert $\gamma$ der Diskriminanzfunktion erzeugt wird. Wichtig ist außerdem, dass je nachdem für wie vielen Gruppen ein Diskriminanzmaß maximal werden kann, durch andersartige Schätzung der Koeffizienten so viele Maximalwerte des Diskriminanzkriteriums erzeugt werden können, wie es Gruppen $-1$ gibt und damit ebenfalls weitere Diskriminanzfunktionen. *„(Jede weitere) … Diskriminanzfunktion wird so ermittelt, dass sie einen maximalen Anteil derjenigen Streuung erklärt, die nach Ermittlung der ersten Diskriminanzfunktion als Rest verbleibt. Da die erste Diskriminanzfunktion so ermittelt wurde, dass ihr Eigenwert und damit ihr Erklärungsanteil maximal wird, kann der Erklärungsanteil der zweiten Diskriminanzfunktion (bezogen auf die gesamte Streuung) nicht größer sein. Entsprechend wird jede weitere Diskriminanzfunktion so ermittelt, dass sie jeweils einen maximalen Anteil der verbleibenden Reststreuung erklärt.“* (Backhaus 2016, S.237) Der Eigenwert $\gamma$ kann auch als Verhältnis von erklärter zu nicht erklärter Streuung interpretiert werden. Ein Eigentwert von 2 signalisiert beispielsweise, dass der Anteil der erklärten Streuung doppelt so hoch ist, wie der Anteil der nicht erklärten Streuung. Dabei entscheidet die Höhe des Eigenwerts über die relative Wichtigkeit der jeweiligen Schätzung der Koeffizienten und der zugehörigen Diskriminanzfunktion.

Ein Maß hierfür ist der sogenannte Eigenwertanteil (erklärter Varianzanteil):

$$A_i = \frac{\gamma_i}{\gamma_1 + \gamma_2 + \cdots \gamma_i}$$

*(Formel 3:3 – Berechnung des relativen Eigenwertanteils)*

Mit:

$\gamma_i = \text{Eigenwert}_i$
$A_i = \text{relativer Eigenwertanteil}$
$i = \text{Anzahl der Diskriminanzfunktionen}$

$A_i$ gibt den Anteil der durch die $i$-te Diskriminanzfunktion erklärten Streuung an. Die Eigenwertanteile summieren sich zu 1, während die Eigenwerte selbst auch größer als 1 sein können. Die diskriminatorische Bedeutung der ermittelten Diskriminanzfunktionen nimmt in der Regel sehr schnell ab. Empirische Erfahrungen zeigen, dass man auch bei großer Anzahl von Gruppen und Merkmalsvariablen meist mit zwei Diskriminanzfunktionen auskommt. Dies hat unter anderem den Vorteil, dass sich die Ergebnisse leichter interpretieren und auch graphisch darstellen lassen.

### 3.2.4. Schritt 4: Prüfung der Diskriminanzanalyse

Grundsätzlich kann man die Güte der Diskriminanzfunktion durch zwei Art und Weisen prüfen. Einerseits kann man anhand der bereits bekannten Gruppenzugehörigkeiten abgleichen, mit welcher Genauigkeit die Diskriminanzvariable trennt. Andererseits kann man auf das bereits beschriebene Diskriminanzkriterium zurückgreifen, um die Diskriminanzbeiträge der einzelnen Prädiktoren zu beurteilen.

Ersteres lässt sich über einer Klassifikationsmatrix und den entsprechenden Trefferquoten, also dem Verhältnis von falsch und richtig klassifizierten Objekten, schnell feststellen. Eine Diskriminanzfunktion hat dann einen Nutzen, wenn die Trefferquote durch sie höher ist, als durch den reinen Zufall, also derjeniger Zuordnung der Objekte, die Grundsätzlich am wahrscheinlichsten zu erwarten wäre.

Dabei kommt es zu einem Stichprobeneffekt, wenn die Diskriminanzanalyse auf Basis derselben Stichprobe berechnet wird, an der ihre Güte getestet wird. Da die Diskriminanzfunktion immer so ermittelt wird, dass die

Trefferquote in der verwendeten Stichprobe maximal wird, ist bei der Anwendung auf eine andere Stichprobe mit einer niedrigeren Trefferquote zu rechnen. Daher sollte man vor der Schätzung der Diskriminanzfunktion seine Datengrundlage zufällig in zwei Stichproben teilen, eine Lern- und eine Kontrollstichprobe. Anhand der Lernstichprobe wird die Diskriminanzfunktion geschätzt, welche schließlich anhand der Kontrollstichprobe getestet wird. Dadurch lässt sich eine realistische Trefferquote gewinnen.

Eine zweite Art und Weise, die Güte der Diskriminanzfunktion zu testen, ist mit dem Eigenwert verbunden, also dem Maximalwert des Diskriminanzkriteriums und dem Anteil der erklärbaren Streuung der jeweiligen Diskriminanzfunktion. Dadurch, dass die erklärte und nicht erklärte Streuung beliebige positive Werte annehmen kann, kann der Eigenwert ebenfalls beliebige Werte annehmen. Es stellt sich also die Frage, ob ein besonders hoher Eigenwert eine besonders hohe Güte der Diskriminanzfunktion zur Folge hat oder wie hoch der Eigenwert überhaupt sein muss, damit die Funktion überhaupt Trennkraft besitzt. Durch die Beliebigkeit des Ausmaßes des Eigenwerts als Testgröße, ist der Eigenwert als Maß für die Güte so nicht geeignet. Erst die Normierung auf eine festgelegte Skala, beispielsweise von Null bis Eins, erlaubt es, dann aus diesem Wert Schlüsse zu ziehen. Hierzu gibt es verschiedene Quotienten. Der bekannteste von ihnen ist Wilks Lambda.

#### *3.2.4.1. Wilks Lambda*

Wilks Lambda ist ein inverses Gütemaß, d.h. kleinere Werte bedeuten höhere Trennkraft der Diskriminanzfunktion und umgekert. Wilks Lambda berechnet sich wie folgt:

$$\Lambda = \frac{1}{1+\gamma} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}}$$

*(Formel 3:4 – Berechnung von Wilks Lambda)*

Mit:

$\Lambda =$ Wilks Lambda
$\gamma =$ Eigenwert

Dieses Wilks Lambda stellt ein Maß für die Unterschiedlichkeit der Gruppen dar. Je geringer sein Wert ist, desto homogener sind die Gruppen und desto größer ist der Unterschied zwischen den Gruppen. Die Bedeutung von Wilks Lambda liegt darin, dass es sich in eine probabilistische Variable transformieren lässt und damit Wahrscheinlichkteitsaussagen über die Unterschiedlichkeit von Gruppen erlaubt. Dies erlaubt statistische Signifikanzprüfung und damit auch die Signifikanzprüfung der Diskriminanzfunktion[51]. Wilks Lambda nach der Transformation liefert eine Variable, die angenähert wie Chi-Quadrat verteilt ist. Die Nullhypothese, dass sich die Gruppen nicht unterscheiden, wird durch den Signifikanztest anhand des Chi-Quadrats und der Freiheitsgrade getestet. Je kleiner Wilks Lambda wird, desto größer wird Chi-Quadrat. Der entsprechende p-Wert ist dabei ein empirisches Signifikanzniveau in Abhängigkeit der Chi-Quadrat Verteilung.

$H_0$: Die beiden Gruppen unterscheiden sich nicht.

$H_1$: Die beiden Gruppen unterscheiden sich.

Der Signikanztest über Wilks' Lambda gibt nur eine Aussage über die mit den hergeleiteten Diskriminanzfunktionen erzielte Trennung insgesamt, bzw. über die relative Diskriminanzkraft der einzelnen Funktionen. Er beinhaltet auch den Test der Nullhypothese, dass sich die Gruppen nicht unterscheiden. Es wird jedoch keine Aussage darüber gemacht, wie gut die Funktionen die einzelnen Gruppen trennen.

[51] Backhaus 2016, S. 241

#### 3.2.4.2. Kanonische Korrelation

Ein zweites Maß stellt der kanonische Korrelationskoeffizient dar. Dieser Quotient ist *kein* inverses Gütemaß. Er berechnet sich wie folgt:

$$c = \sqrt{\frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}}} = \sqrt{\frac{\gamma}{\gamma + 1}}$$

*(Formel 3:5 – Berechnung kanonischer Korrelationskoeffizient)*

Mit:

$c$ = kanonische Korrelationskoeffizient
$\gamma$ = Eigenwert

Je größer der kanonische Korrelationskoeffizient auf einer Skala von 0 bis 1 ist, desto höher ist der Anteil der erklärten Streuung an der Gesamtstreuung der Diskriminanzwerte und desto größer ist folglich die Trennkraft der Diskriminanzfunktion. Ist der kanonische Korrelationskoeffizient beispielsweise 0.76, erklärt er 76% der Gesamtstreuung der Diskriminanzwerte.

Der quadrierte kanonische Korrelationskoeffizien ist dabei sachlich identisch mit der Beurteilungsgröße des Bestimmtheitsmaßes ($r^2$), welches eine wichtige Rolle in der der Regressionsanalyse spielt. Wilks Lambda und der quadrierte kanonische Korrelationskoeffizient ergeben aufsummiert Eins und sind komplementär zueinander.[52]

### 3.2.5. Schritt 5: Prüfung der Merkmalsvariablen

Wie bereits angesprochen, steigen mit Hinzunahme jeder Merkmalvariablen die Dimensionen und damit die Gesamtvarianz des Modells an. Um dies zu vermeiden, ist es bei der linearen Diskriminanzanalyse von Vorteil, Merkamalsvaraiblen mit geringer Trennfähigkeit und entsprechend geringer Erklärkraft von Gruppenunterschieden auszuschließen. Dies reduziert die Komplexität und die Gesamtvarianz (erklärte Streuung + nicht erklärte Streuung) der abhängigen Variable, vermindert Redundanzen und steigert die Interpretierbarkeit der Einflüsse der erklärenden Variablen.

Über die Diskriminanzkoeffizienten kann die Trennfähigkeit der dazugehörigen Merkmalsvariablen ausfindig gemacht werden. Da die Merkmalsvariablen häufig auf verschiedenen Skalen gemessen werden, verwendet man zur Bestimmung der Bedeutung der Merkmalsvariablen allerdings die standardisierten Diskriminanzkoeffizienten. Diese berechnen sich, indem man die Diskriminanzkoeffizienten mit der Standardabweichung der betreffenden Merkmalsvariablen multipliziert.

$$b_i^* = b_i \cdot s_i$$

*(Formel 3:6 – Berechnung des standardisierten Diskriminanzkoeffizienten)*

Mit:

$b_i^*$ = standardisierter Diskriminanzkoeffizient von Merkmalsvariable $i$
$b_i$ = Diskriminanzkoeffizient von Merkmalsvariable $i$
$s_i$ = Standardabweichung von Merkmalsvariable $i$

Danach kann am Betrag der standardisierten Diskriminanzkoeffizienten kann ihre relative Bedeutung für die Diskriminierung identifiziert werden. Dies kann durch einen prozentualen Wirkungskoeffizienten zum Ausdruck gebracht werden, indem man folgende Berechnung ausführt:

[52] Prof. Dr. Wolfgang Müller, Diskriminanzanalyse

$$b_n^r = \frac{b_n^*}{b_1^* + b_2^* \ldots + b_i^*} \cdot 100$$

*(Formel 3:7 - Berechnung der relativen Bedeutung eines standardisierten Diskriminanzkoeffizienten)*

Mit:

$b_i^*$ = standardisierter Diskriminanzkoeffizient von Merkmalsvariable $i$
$b_n^r$ = relative Bedeutung des Diskriminanzkoeffizient von Merkmalsvariable $n$
$s_i$ = Standardabweichung von Merkmalsvariable $i$
$n$ = Wert von 1 bis $i$
$i$ = Anzahl der Prädiktoren

Zur Unterscheidung von den standardisierten Diskriminanzkoeffizienten, werden die Koeffizienten in der Diskriminanzfunktion auch als nicht-standardisierte Diskriminanzkoeffizienten bezeichnet. Zur Berechnung von Diskriminanzwerten müssen immer die nicht-standardisierten Diskriminanzkoeffizienten verwendet werden.[53]

Wenn mehrere Diskriminanzfunktionen und somit verschiedene Diskriminanzkoeffizienten vorliegen, kann die diskriminatorische Bedeutung einer Merkmalsvariablen auch bezüglich aller Diskriminanzfunktionen beurteilt werden. Hierfür addiert man alle Diskriminanzkoeffizienten gewichtet anhand des Eigenwertanteils der dazugehörigen Diskriminanzfunktion. Man erhält auf diese Weise den mittleren Diskriminanzkoeffizienten.

$$\overline{b_j} = \sum_{k=1}^{K} \left| b_{jk}^* \right| \cdot EA_k$$

*(Formel 3:8 - Berechnung der relativen Wichtigkeit eines standardisierten Diskriminanzkoeffizienten bei mehreren Diskriminanzfunktionen)*

Mit:

$b_{jk}^*$ = standardisierter Diskriminanzkoeffizient für Merkmalsvariable $j$ bezueglich Diskriminanzfunktion $k$

$EA_k$ = Eigenwertanteil der Diskriminanzfunktion $k$

$\overline{b_j}$ = mittlerer standardisierter Diskriminanzkoeffizient $j$

### 3.2.6. Schritt 6: Klassifizierung neuer Elemente

Für die Klassifizierung von neuen Elementen existieren grundsätzlich drei verschiedene Konzepte: Klassifizierung mittels Distanzen, mittels Klassifizierungsfunktion und über Wahrscheinlichkeitsberechnungen.

Die Klassifizierungsfunktion bietet die Möglichkeit einer Gruppenzuordnung der Objekte anhand ihrer Merkmalsausprägungen. Durch die Diskriminanzanalyse wird jeder Gruppe eine eigene Funktion zugewiesen. In diese Funktion werden die jeweiligen Merkmalsausprägungen des Objekts eingetragen. Das Objekt gehört der Gruppe an, bei der dessen Funktionswert maximal ist.

Anhand des Distanzkonzepts lassen sich Elemente klassifizieren, indem ihre Distanzen zu den Gruppen-Zentroiden ausgerechnet werden. Die geringste Distanz zu einer Gruppe bedeutet eine entsprechende Klassifikation. Im zwei Gruppen-Fall ist der kritische Distanzwert die Mitte der Gruppenmittelwerte. Mathematisch wird dieses Konzept bei einem Mehr-Gruppen-Fall deutlich komplexer. Hier kann die Distanz entweder durch (quadrierte) euklidische Abstände oder die Mahalanobis-Distanz ermittelt werden. Beide Verfahren zielen ebenfalls auf das minimieren des Abstands des Elements zu den Gruppenmittelwerten und

[53] Backhaus 2016

können ein Ergebnis erzielen, welches das Element eindeutig einer Gruppe zuweist. Für die mathematische Herleitung dieses Konzepts wird an dieser Stelle auf Backhaus, 2016, S.277 verwiesen.

Auf dem Distanzkonzept beruht außerdem die Wahrscheinlichkeitsberechnung. Sie macht die Klassifizierung zum statistischen Entscheidungsproblem. Das Konzept besitzt dabei die größte Flexibilität, ist aber, laut Backhaus et. al., für Nicht-Statistiker eher schwerer verständlich[54]. In SPSS kommt das Wahrscheinlichkeitskonzept zum Einsatz.

Voraussetzung für die Wahrscheinlichkeitsberechnung ist das Kennen der a-priori Wahrscheinlichkeiten, also derjenigen Wahrscheinlichkeiten, die erwartungsgemäß rein zufällig eintreten würden, gäbe es keine Diskriminanzfunktion. Durch diese Betrachtung lässt sich beispielsweise berücksichtigen, dass in der Realität eine bestimmte Gruppe häufiger oder weniger häufig als eine andere vorkommt. Über alle Gruppen hinweg müssen sich die Wahrscheinlichkeiten zu 1 addieren.

Die Klassifizierungsregel bei diesem Konzept ist, dass ein Element derjenigen Gruppe zugeordnet wird für das es die größte a-postpriori-Wahrscheinlichkeit besitzt. Diese Wahrscheinlichkeiten errechnen sich über den Satz von Bayes. Eine mathematische Herleitung ist ebenfalls Backhaus, 2016, S.278 zu entnehmen.

Die Güte des Wahrscheinlichkeitskonzepts lässt sich über einen statistischen Test feststellen. Durch Box's M lässt sich feststellen, ob die Vorrausetzung der Annahmen für das Klassifizieren mit dem Wahrscheinlichkeitskonzept, nämlich gleiche Streuung (Kovarianzen der Merkmalsvariablen) vorliegen. Die Annahme lässt sich mittels eines F-Tests auf Signifikanz untersuchen. Dabei deuten niedrige Signifikanzwerte auf eine ungleiche Streuung hin.

$$H_0 = \text{Kovarianzen bzw. Streuung der Gruppen ist gleich}$$

$$H_1 = \text{Kovarianzen bzw. Streuung der Gruppen ist ungleich}$$

Hierdurch werden ungleiche Streuungen in den Gruppen berücksichtig und entsprechend bewertet. In Kombination mit dem Theorem von Bayes lassen sich so individuelle Wahrscheinlichkeiten der Gruppenzugehörigkeit für jedes Objekt feststellen. Eine mathematische Herleitung ist ebenfalls (Backhaus 2016), S. 250ff zu entnehmen.

# 4. Exemplarische Durchführung der Diskriminanzanalyse mit SPSS

In Anlehnung an das Kapitel 3 „Einführung in die Diskriminanzanalyse" soll nun exemplarisch eine Diskriminanzanalyse inklusive theoretischen Überlegungen zur Untersuchungsgrundlage anhand von Erfolgsfaktoren zu Kinofilmen und den Voraussetzungsprüfungen der parametrisch statistischen Verfahren mit der Software IBM SPSS 19 vorgeführt werden. Es ist im Rahmen der exemplarischen Diskriminanzanalyse unmöglich, die vielfältigen Möglichkeiten im Einzelnen darzustellen. Es wird sich folglich daher auf die im Zusammenhang stehenden und relevanten Bereiche konzentriert, die bereits anhand der theoretischen Grundlagen zur Diskriminanzanalyse vorgestellt wurden.

## 4.1. Datengrundlage

Bei der Arbeit mit dem Datenmaterial handelt sich um eine Sekundärforschung[55], also der Auswertung von bereits vorhandenen Daten unter einem speziellen Blickwinkel. Der Datensatz, an dem die Diskriminanzanalyse vorgeführt wird, bildet gemessene Beobachtungen anhand von 1257 Kinofilmen ab, die zwischen den Jahren

[54] Backhaus 2016, S. 249
[55] vgl. Fantapié Altobelli 2017, S.45

2002 und 2006 in deutschen Kinos gelaufen sind. Quellen der Daten sind der Verband für Filmverleiher e.V. (VdF), Mediabiz.de, IMdB.de, Umfragen an Kinobesuchern und Branchenexperten nicht bekannter Quelle, Cinema.de, BoxofficeMojo.com, the-numbers.com, MediaCom, EDI Nielsen, Oscars.com, insidekino.com und die Hompages der Fußball-Europa- und -Weltmeisterschafts Wettkämpfe. Der Urheber der Datensammlung ist nicht bekannt.

Der Datensatz umfasst insgesamt 106 Messungen mit unterschiedlichen Merkmalsausprägungen und 1257 Merkmalsträger. 80 Messungen mit Merkmalsausprägungen gehören zu vier übergeordneten Messungen, die jeweils über bis zu 20 Wochen erfolgten. Die Besucherzahlen (@n-WOBesucher), Anzahl der Kinoleinwände „Screens“, auf denen der Film in Deutschland (@n-WOKopien) und den USA (@n-USWOKopien) lief, sowie den Umsatz, den die Filme in den USA erbrachten (@n-WOUSUmsatz). Diese vier Variablen kann man unter **Variablen zur Erfolgsmessung** zusammenfassen.

Die verbleibenden 26 Variablen lassen sich in Produktvariablen, Promotionsvariablen, Distributionsvariablen, eine Preisindikatorvariable und weitere, nicht näher spezifizierte Variablen aufteilen.

Zu **Produktvariablen** gehören die nominale Variable „Herkunft des Films“ (COUNTRY) auf Länderbasis, die ordinalen Variablen Besetzung (ACTORS) und Regie (DIRECTOR), welche durch eine repräsentative Umfrage jeweils von 0 (min.) bis 5 (max.) beurteilt wurden und nominal binär, ob der Film eine Fortsetzung (SEQUEL) ist oder nicht.

Zu den **Promotionsvariablen** gehören die Erfolgserwartung (ANTICIPATE), die in einer metrischen Variablen durch die Kombination von drei standardisierten Variablen Film Budget, Marketing Budget und Anzahl der Vorführungen in der Ausstrahlungswoche zusammensetzt wurde. Des Weiteren liegt eine gewichtete metrische Variable vor, die Aufschluss darüber gibt, für wie viele Preise der Film nominiert wurde und wie viele Preise der Film gewonnen hat (AWARDS). Die Gewichtung berücksichtigt die Wichtigkeit der Nominierungen und Preise. Nähere Informationen hierüber sind nicht bekannt. Darüber hinaus gibt es eine nominale binäre Variable Kinotipp (CIN_TIP), welche beschreibt, ob der Film als Tipp auf einem der einflussreichsten deutschen Kinoportalen (Cinema.de) erschienen oder nicht ist. Eine weitere ordinale Variable Filmkritik (CIN_CRITIC) zeigt auf, was für eine Kritik der Film von den Usern eben dieses Portals von 1 (max.) bis 5 (min.) erhalten hat.

Zu den **Distributionsvariablen** gehören Wettbewerbsindikatorvariablen, wie die bereits gelaufene Dauer von konkurrierenden Filmen (COMP_ONG) sowie die Anzahl von ähnlichen Filmen gewichtet mit der bereits gelaufenen Dauer (COM_REV). Ebenfalls gibt eine nominale binäre Variable an, ob der Film während einer Fußball WM oder EM lief (EM_WM). Eine weitere nominale Variable beschreibt den Namen des Distributors (DISTRIBUTOR). Schließlich gibt eine metrische Variable in Prozentangaben Aufschluss über den Marktanteil des jeweiligen Distributors im jeweiligen Jahr (DIS_POWER).

**Preisindikatorvariable** ist die metrische Variable Laufzeit des Films in Minuten (MIN).

**Weitere nicht näher spezifizierte Variablen** sind das Genre (GENRE), die Altersfreigabeklassifikation (FSK), ob der Film einer Buchvorlage (BOOK) folgt oder nicht und schließlich die Erscheinungsdaten des Filmes für den deutschen (Release_Germany) und den nordamerikanischen Markt (Release_US), sowie die metrische Variable Anzahl der beobachteten Wochen (Reported_Weeks).

Die letzte übrige Variable ist eine Identifikationsvariable, die jedem Objekt chronologisch nach Datum der Veröffentlichung des Films in Deutschland eine individuelle Fallzahl zuweist (ID).

## 4.2. Einschätzung der Datenqualität

Die Datengrundlage für Untersuchungen ist umfangreich und die Möglichkeit, aus der Betrachtung verschiedener Aspekte der Daten neue Erkenntnisse zu generieren, vielseitig. Essentielle und allgegenwärtige

Voraussetzung dabei ist die Qualität der Daten[56]. Diese gilt es vor der weiteren Arbeit mit den Daten einzuschätzen.

Es ist in keiner Weise dokumentiert, ob und wie bereits an dem Datensatz gearbeitet wurde. Es gibt auch keinen vollständigen Testdatensatz über dessen Abgleich man Abweichungen feststellen könnte. Es liegt jedoch der Kodierungsplan vor. Laut diesem fehlt im Datensatz die Variable „Season" komplett, welche die Wochen des Jahres mit einem Index nach ihrer „Saisonalität" von 1 (min.) bis 100 (max.) gewichtet, analog zum Vorgehen von Elberse und Eliashberg 2003. Des Weiteren wurde zur Berechnung der Variablen „Anticipate" die Variable „Advertising" verwendet, welche das Marketing Budget in Euro beschreibt, die sich ebenfalls nicht im Datensatz finden lässt.

Darüber hinaus ist die Vollständigkeit der Daten wie folgt zu beurteilen[57]: Wie bereits angemerkt handelt es sich bei der Datensammlung mit hoher Wahrscheinlichkeit um einen Auszug einer größeren Datenbank, also einer Stichprobe. Hierfür spricht die individuelle Bezeichnung der Fälle durch chronologische Zuweisung einer Fallzahl (ID) beginnend bei 3048 bis 4551 abhängig vom Erscheinungsdatum des Films in deutschen Kinos. Es kann angenommen werden, dass es sich um einen Ausschnitt bzw. einer zeitabhängigen Stichprobe handelt. Innerhalb dieses Ausschnitts befinden sich 1257 Objekte. Angefangen bei ID 3048 bis ID 4551 müssten unter der Annahme, dass die IDs chronologisch vergeben wurden, 1503 Fälle zu finden sein. Es kann also davon ausgegangen werden, dass 246 Fälle bereits vom Datensatz ausgeschlossen wurden. Das Fehlen dieser 246 Objekte könnte ein Indiz dafür sein, dass die Datengrundlage bereits beispielsweise um einige Fälle mit Ausreißern oder Fälle mit fehlenden Merkmalen bereinigt wurde. Da die Objekte vollständig und irreversibel aus dem Datensatz entfernt wurden, sind die tatsächlichen Gründe hierfür nicht mehr nachzuvollziehen.

Die übrigen 1257 Objekte sind vollständig hinsichtlich ihrer Einheitlichkeit und weisen keine doppelten Werte oder mehrfache Datenzeilen auf mit folgenden Anmerkungen bezüglich fehlender Werte (Missings)[58]:

Bei einer Vielzahl von Objekten fehlen Messungen für verschiedene Variablen. Hierfür kann es verschiedene Gründe geben. Es ist denkbar, dass fehlende Werte zufällig und zusammenhangslos auf Grund einer fehlerhaften Erhebung oder Datenbankpflege entstanden sind („missing completley at random" auch MCAR). Ebenfalls ist es möglich, dass Werte fehlen, die nur teilweise zufällig nicht vorhanden sind (MAR). Dies liegt oft an einem nicht-geeigneten Umfrage- oder Datenerhebungsdesign. Es könnten allerdings auch Werte fehlen, dessen Fehlen einem bestimmten Muster zugrunde liegt („not missing at random" auch NMAR). Diese Muster können identifiziert werden.

Betrachtet man die Variable „Release_US", die das Datum wiedergibt, wann der Film im nordamerikanischen Markt veröffentlicht wurde, stellt man fest, dass 716 Einträge und 541 fehlende Fälle vorliegen. Bei 43% der beobachteten Objekte fehlt das Datum für den Eintritt in den nordamerikanischen Markt. Hier scheint ein Muster vorzuliegen. In allen Fällen, wo ein Objekt einen fehlenden Wert bei der Variable „Release_US" aufweist, weist das Objekt ebenso den Wert „0" für jede weitere Beobachtung auf, die am nordamerikanischen Markt hätte gemessen werden müssen. Dieser Sachverhalt kann als Erklärung für die eingetragenen „0" Werte für die Variablen US Umsatz Woche 1 bis Woche 20 sowie US Kopien Woche 1 bis Woche 20 bei 539 von 541 Objekten angesehen werden und folgt somit einer klaren Gesetzmäßigkeit. Hieraus lässt sich schlussfolgern, dass Objekte mit fehlender Beobachtung für die Variable „Release_US" nicht zufällig fehlen, sondern deshalb, weil sie dort nie, oder zumindest nicht bevor die Datengrundlage geschaffen wurde, im Kino erschienen sind (NMAR). Bei zwei Objekten mit fehlendem Wert bei der Variable „Release_US" wurden dennoch Daten für US Umsatz und US Kopien gemessen. Bei diesen beiden Objekten scheint das Datum aus anderweitigen Gründen (MAR) zu fehlen.

Bei den Messungen der Objekte hinsichtlich der Variable „Budget" sind ebenfalls 520 von 1267 Werte nicht vorhanden. Eine Abhängigkeit von einer oder mehreren anderen Variablen dieser fehlenden Werte lässt sich nicht eindeutig ausfindig machen. Es fällt jedoch auf, dass Budgetinformationen eher von denjenigen Distributoren zur Verfügung gestellt werden, die einen höheren relativen Marktanteil aufweisen.

[56] vgl. Schendera 2007
[57] vgl. Schendera 2007
[58] vgl. Schendera 2007

Auffällig ist außerdem, dass die Variable „Reported Weeks“ nicht als Indikator für die Anzahl der Wochen, in denen Beobachtungen gemessen wurden, genutzt werden kann. Das Maximum dieser Variable liegt bei 41 und das Minimum bei 1 (Wochen). Es könnte angenommen werden, dass die Anzahl der Reported Weeks sich aus der Anzahl der Wochen ergibt, in denen für die Variablen 1-20 Wochen Besucher DE, Umsatz US und Kopies DE sowie US Daten erhoben wurden. Dieser Sachverhalt wird durch die Daten nicht wiedergespiegelt. Eine weitere Annahme, dass die Variable aussagt, wie lange der Film in Kinos lief, kann ebenfalls verworfen werden. Keine der Variablen, die über bis zu 20 Wochen erhoben wurden, spiegeln diesen Sachverhalt wieder. Über die Aussagekraft dieser Variable ist nichts weiter bekannt, daher hat sie keinen deutbaren Informationsgehalt und kann aus der Betrachtung ausgeschlossen werden.

Als letztes ist anzumerken, dass bei vielen Messungen, die über Wochen hinweg passierten, im Datenblatt häufig eine „0“ eingetragen ist. Diese Werte fehlen nicht, sondern kommen als eingetragener, vorhandener Wert dadurch zu Stande, dass ein Film, der nur 5 Wochen lang in den Kinos läuft, ab der 6. Woche folglich keinen anderen Wert als „0“ mehr für die Variablen haben kann. Entsprechend füllen Nullen die Datenzeile gänzlich, wenn beispielsweise ein Film und dessen Beobachtungen für den nordamerikanischen Markt betrachtet werden, der gar nicht in Amerika lief. Je nach Art des statistischen Verfahrens kann es sinnvoll sein, diese Werte als „benutzerdefiniert fehlend“ aus einer Betrachtung auszuschließen, beispielsweise um eine nicht zufällige Häufung dieser Ausprägung in einem Verteilungstest zu vermeiden. Sollte dem so sein, wird es in dieser Arbeit an entsprechender Stelle erwähnt werden.

Es ist für weitere Überlegungen wichtig, solche Variablen auszuwählen, die sachlich nachvollziehbar sind und Daten wiedergeben, die möglichst vollständig sind. Dadurch ist gewährleistet, dass die Datengrundlage eine hohe Qualität aufweist, um entsprechend hochwertige Aussagen aus den Daten gewinnen zu können.

Eine Bewertung der Ausreißer und Extremwerte innerhalb der Datengrundlage wird anhand derjenigen Variablen erfolgen, die als Merkmalsvariablen für die Diskriminanzanalyse ausgewählt werden.

## 4.3. Exemplarische Definition der Gruppen

Im Datensatz gibt es noch keine kategoriale Variable, die die Filme in disjunkte Gruppen unterteilt, die Aufschlüsse über den Erfolg oder Misserfolg des Films geben. Da dies zu prognostizieren, das übergeordnete Ziel der Diskriminanzanalyse werden soll, muss eine solche erfolgsdefinierende Variable erzeugt werden. Grundsätzlich eignen sich hierfür taxonomische Verfahren, wie eine Clusteranalyse. In dieser Arbeit wird jedoch keine Clusteranalyse genutzt, um die Klassifizierung zu erzeugen. Stattdessen wird einer theoretischen Überlegung zugrundeliegend univariat ein Merkmal gewählt, welches Rückschlüsse über den Erfolg des Films erlaubt. Es wird ab einem Grenzwert per Definition festgelegt, welcher Gruppe ein Objekt zugewiesen wird. Hierfür muss erst einmal definiert werden, was genau unter „Erfolg“ und „Misserfolg“ von Kinofilmen verstanden wird.

Wenn in der Arbeit von Erfolg von Kinofilmen gesprochen wird, wird auf den objektiv messbaren wirtschaftlichen Erfolg abgezielt. Erfolg ist hier nicht im Sinne des Kinobesuchers als der subjektiv wahrgenommene Erfolg der Umsetzung des Films seinerseits verbunden mit einer positiven Bewertung des Kinoerlebnisses zu verstehen. Eine solche Art von Betrachtung ist durchaus möglich, jedoch nicht Gegenstand dieser Untersuchung.

Nach Henning-Thurau muss konzeptionell zwischen verschiedenen wirtschaftlichen Erfolgsebenen unterschieden werden[59]. Einnahmen, die durch Vorführungen des Kinobetriebs geschehen, bilden dabei das Basismedium und übernehmen eine Signalfunktion für nachfolgende Verwertungsschritte des Films, wie Umsätze durch Video und TV. Untersuchungen bezüglich der Einspielergebnisse im nordamerikanischen Markt und Einnahmen aus dem dortigen Videoverleih weisen eine hohe Korrelation (0.67) auf[60]. Hierauf stützt sich

[59] Thurau-Henning und Wruck 2000
[60] vgl. Thurau-Henning 2004

die Annahme, dass, wenn ein Film im Kino erfolgreich war, er schließlich auch bei späteren Verwertungsschritten erfolgreich sein wird. Über den tatsächlichen wirtschaftlichen Erfolg über das Kinoergebnis im US-nordamerikanischen Markt hinaus liegen keine weiteren Daten vor, die diese Annahme zusätzlich stützen könnten.

Im exemplarischen Datensatz kann der wirtschaftliche Erfolg anhand der näheren Betrachtung des Umsatzes beurteilt werden, der durch die Kinobesucher an den sogenannten Boxoffices, den Kinokassen, erzeugt wird. Der Datensatz liefert mit den 20 Variablen der Wochenumsätze von Woche 1 bis 20 im US Markt einige finanzielle Erfolgsindikatoren. Diese Variablen können verwendet werden, um ein Bild des finanziellen Filmerfolgs für den nordamerikanischen Markt zu erzeugen. Dabei läuft nicht jeder Film 20 Wochen bzw. erzeugt 20 Wochen lang Umsätze. Dennoch kann die Kumulation der vorhandenen Ausprägungen als Gesamtumsatz des Kinofilms im nordamerikanischen Markt verstanden werden.

Laut Thurau ist der nordamerikanische Markt für 40% der globalen Kinokassenumsätze verantwortlich[61]. Für Auswirkungen des nordamerikanischen Kinoeinspielergebnisses für ausländische Märkte ermittelten Henning-Thurau/Wruck eine Korrelation von 0.54, die für einzelne Filmgenres sogar bis zu 0.85 betrug[62]. Es wird also deutlich, dass die Umsätze des nordamerikanischen Markts von enormer Bedeutung für den allgemeinen Erfolg eines Kinofilms sind.

Die Kosten werden durch die Variablen „Budget“, welche die Filmproduktionskosten darstellen und die Variable „Marketingbudget“ dargestellt, welches in der Filmindustrie üblicherweise 50% der Produktionskosten beträgt[63]. Die Gesamtkosten einer Produktion ergeben sich aus dem Produktionsbudget multipliziert mit dem Faktor 1.5.

Die Kosten stehen den Umsätzen entgegen. Anhand der Gegenüberstellung der Variablen können die Filme bereits nach bis zu 20 Wochen Laufzeit im nordamerikanischen Markt frühzeitig auf ihre *Rentabilität* geprüft werden. Spielt der Film bereits nach 20 Wochen Gewinne ein, also übersteigen die Umsätze die Produktionsgesamtkosten, wird er als *„erfolgreich“* klassifiziert. Erzielt er keinen Deckungsbeitrag, wird er als *„nicht erfolgreich*“ eingestuft. Deckt der Film im seltenen Fall lediglich seine Kosten ohne dabei Gewinne zu erzielen, ist der Film zwar erfolgsneutral, wird jedoch ebenfalls als *„nicht erfolgreich“* bewertet.

Filme, die weder durch ihr Ergebnis im nordamerikanischen Markt ihre eigenen Kosten decken können, noch der Erfolg durch spätere Verwertungsstufen approximativ in Aussicht steht, werden als „nicht erfolgreich“ bewertet.

*Klassifizierungsregeln anhand des Grenzwerts sind folglich:*

erfolgreich:

$$a\text{: Umsatz nach US Laufzeit} - \text{Produktionsbudget } x\ 1.5\ >\ 0$$

nicht erfolgreich:

$$b\text{: Umsatz nach US Laufzeit } x\ 2.5 - \text{Produktionsbudget } x\ 1.5\ <=\ 0$$

Um einen Mehr-Gruppen-Fall zu konstruieren kann eine dritte Gruppe hergeleitet werden. Die Filme, welche durch spätere Verwertungsstufen wirtschaftlich wahrscheinlich erfolgreich sein werden, werden in dieser Betrachtung als *„nicht erfolgreich mit Risiko“* eingeschätzt. *„Erfolgreich mit Risiko“* ist ein Film dann, wenn seine Umsatzerlöse gemessen an der Laufzeit im US Markt multipliziert mit dem Faktor 2.5 die Gesamtkosten der Produktion übersteigen. Diese Annahme wird anhand der Aussage getroffen, dass etwa 40% der globalen Kinokassenumsätze bereits im US-nordamerikanischen Markt erschlossen werden. Es ist anzunehmen, dass

[61] vgl. Thurau-Henning 2004
[62] vgl. Thurau-Henning 2004
[63] Thurau

Filme bis auf Ausnahmen einem internationalen Publikum vorgeführt werden sowie, dass die wirtschaftliche Verwertung über weitere Maßnahmen stattfindet. Der grundsätzliche Erfolg im US Markt wird somit als Trend und Indikator für künftigen wirtschaftlichen Erfolg bewertet. Es wird also deutlich, dass die Umsätze des US-nordamerikanischen Markts von enormer Bedeutung für den allgemeinen Erfolg eines Kinofilms sind.

*Zusätzliche Klassifizierungsregeln für den Mehr-Gruppen-Fall anhand des Grenzwerts sind folglich:*

erfolgreich mit Risiko:

$c$: Gewinn nach US Laufzeit $\leq 0$

$d$: Umsatz nach US Laufzeit $x$ 2.5 – Produktionsbudget $x1.5 > 0$

## 4.4. Auswahl geeigneter Prädiktoren für die Lineare Diskriminanzanalyse

Heutzutage gehen Filmstudios erhebliche finanzielle Risiken ein, wenn sie in ein neues Filmprojekt investieren. Alleine das Produktionsbudget der Filme übersteigt dabei des Öfteren die US $ 50-Millionen Grenze, potentielle Marketing und Merchandisekosten noch nicht mit einbegriffen. In Einzelfällen übertreffen die Produktionsbudgets sogar US $ 200 Millionen Dollar. Der dritte Teil von Walt Disneys „Fluch der Karibik – am Ende der Welt“ von 2005 führt derzeit im Jahr 2017 noch immer die Spitze der teuersten Filme aller Zeiten mit einem Produktionsbudget von etwa US $ 350 Millionen an[64]. Dabei variiert der wirtschaftliche Erfolg dieser Art von Investitionen stark. Laut vorherigen Analysen erreichen von 10 Filmen nur 6 bis 7 einen Return on Invest durch die Kinokassen[65]. Welche Faktoren den Erfolg von Kinofilmen ausmachen, wurde bereits mehrfach empirisch untersucht[66]. Klare Ergebnisse werden laut einer Analyse des Forschungsstands in diesem Gebiet von Michel Clement aus dem Jahr 2004 jedoch nicht geliefert. Nach Clement vermuten „manche Autoren […] daher gar, dass die chaotische Natur des Filmgeschäfts eine fundierte theoretische und empirische Analyse nicht zulässt“ (Clement, 2004). Clement selbst fasst verschiedene Ansätze einiger führender Wissenschaftler auf diesem Gebiet zusammen und gibt eine Empfehlung, welche Faktoren bei künftigen Betrachtungen genauer untersucht werden sollten. Dieser Empfehlung wird in dieser Arbeit Folge geleistet und am Forschungsstand zu den Erfolgsfaktoren von Kinofilmen von Clement angeknüpft.

**Forschungsstand nach Michel Clement, 2004:**

- „Die Anzahl der Screens determiniert den Box-Office-Erfolg. Der Box-Office-Erfolg determiniert die Anzahl der Screens in der folgenden Periode. Die Anzahl der Screens ist der mit Abstand wichtigste Einfluss auf den Erfolg des Films. Wenn Screens nicht berücksichtigt werden, so nimmt die Stärke des Einflusses anderer Variablen stark zu und verzerrt so das Ergebnis. Die Relevanz von Marketingmaßnahmen gegenüber Kinobetreibern wird durch dieses Ergebnis gestützt.
- Je höher das Produktionsbudget, desto höher ist der Box-Office, jedoch nicht unbedingt der Gewinn. Wenn der Film keine aufwändigen Effekte aufweist, dann sind die Zahlungen an Stars und Regisseure zumeist die höchsten Posten im Budget. Das Mitwirken von Star-Schauspielern und -Regisseuren beeinflusst den Box-Office positiv, belastet jedoch den Gewinn, weil die Produktionskosten steigen. Auch weisen Serien einen höheren Box-Office aus, jedoch bei steigenden Produktionskosten für die späteren Folgen. Dies liegt vor allem an den höheren Gagen, die dann von den Schauspielern eingefordert werden.
- Einschränkende Altersbegrenzungen reduzieren den Markt und somit den BoxOffice. Damit wird auch deutlich, dass bestimmte Genres eine geringere Wahrscheinlichkeit haben, ein Kassenerfolg zu werden, denn ein Erotik- oder Horrorstreifen wird im Vergleich zu Kinderfilmen immer einem kleineren Gesamtmarkt gegenüberstehen.

[64] http://www.insidekino.com/TOPoderFLOP/TOPBudgetAllTime.htm
[65] Jedidi et al. 1998
[66] Clement 2004

- Wenn der Film von einem Major-Distributor vermarktet wird, dann erreicht er mehr Screens und somit mehr Box-Office. Je mächtiger der Distributor ist, desto bessere Verhandlungsspielräume besitzt er bei der Aufteilung des Umsatzes und erzielt damit einen höheren Gewinn.
- Filme unterliegen starken saisonalen Effekten. Filme, die in den Hochsaisonphasen starten, haben trotz des starken Wettbewerbs eine höhere Wahrscheinlichkeit, erfolgreich zu sein. Das Startdatum des Films wird als einer der zentralen Marketing Instrumente der Studios angesehen. Der Wettbewerb ist am stärksten in den Hauptsaisonphasen (in den USA: Weihnachten und Sommer).
- Je höher das Werbebudget, desto höher der Box-Office, jedoch nicht unbedingt der Gewinn.
- Kritiker haben in späten Phasen des Lebenszyklus keinen Einfluss auf den BoxOffice.
- Bei einer sequenziellen Verwertung des Films in internationalen Märkten kann der Erfolg des Films in den USA als Prognose für Europa verwendet werden, es sei denn es handelt sich um einen sehr speziell auf das US-Publikum zugeschnittenen Film (z. B. einen Baseball-Film). Es lassen sich keine eindeutigen Aussagen z. B. bezüglich der Wirkung der Herkunft eines Filmes oder von Komplementärgütern, Trailern, Filmpreisen oder Mundpropaganda festhalten.“ (Clement, 2004)

Kinofilme stellen eine besondere Art von Gut dar. Sie sind ein Hybrid aus Produkt und Dienstleistung. Das Erstellen des Films bis zum fertigen Filmmedium ist der Herstellprozess eines Produkts, wohingegen der Konsum des Films durch den Kinobesuch eines Konsumenten als Dienstleistung bewertet werden muss[67]. Die Entscheidung zum Konsum unterliegt möglicherweise externen Faktoren, wie Saisonalität oder dem Distributionsgrad eines Films. Einige Variablen konkretisieren sich bis zur Fertigstellung des Films bereits vor der Veröffentlichung, wie etwa das Genre, die Besetzung oder der Regisseur. Andere hingegen ergeben sich erst durch den Konsum des Films durch die Audienz des Publikums, wie quantitative Kennzahlen über den Besuch, Umsatz oder qualitative Bewertungen des Erlebnisses. Umrahmt werden diese Faktoren von externen Einflüssen, ob der Film zu einem günstigen oder ungünstigen Zeitpunkt veröffentlicht wird bedingt durch Wettbewerb oder Saisonalität[68]. Um in der Lage zu sein, den Erfolg eines Kinofilms frühzeitig zu prognostizieren, kann auf Variablen zurückgegriffen werden, die bereits vor der Uraufführung des Films feststehen. Nun ist es jedoch typisch für die Filmbranche, dass Filme zuerst im US-nordamerikanischen Markt ausgestrahlt und erst später über die Ländergrenzen hinweg distribuiert werden. Dies erlaubt es, einer Analyse zur Prognose des Erfolgs eines Films auch Variablen zu verwenden, die anhand des US Markts im Sinne eines Testmarkts erhoben werden konnten, um Rückschlüsse auf künftige Markteinführungen in weiteren Ländern für den Kinofilm zu erlangen.

Einige der von Clement genannten Sachverhalte können anhand des Datensatzes aufgegriffen werden. Es folgt ein Abgleich der vorhandenen Merkmalsvariablen mit den Überlegungen von Clement, um zu definieren, anhand welcher metrischen Variablen die Diskriminanzanalyse durchgeführt werden soll:

Die wohl wichtigste Variable, um den Erfolg eines Kinofilms zu bestimmen, ist laut Clement die Anzahl der Screenings, also wie viele Vorführungen der Film erzielen konnte. Die über 20 Wochen gesammelten beobachteten Merkmale werden kumuliert und als eine neue Variable „USScreens“ gespeichert. Sie gibt inhaltlich wieder, wie oft der Film über 20 Wochen in Kinos in Amerika gelaufen ist. Eng verknüpft mit dieser Variablen ist auch der Umsatz „USUmsatz“, den der Film über 20 Wochen erzielen konnte. Es wird angenommen, dass Filme, die oft ausgestrahlt wurden, eher erfolgreich sind.

Das Produktionsbudget ist verbunden mit teuren, guten und bekannten Schauspielern und Regisseuren. Bekanntere Schauspieler haben laut Thurau eine Markenwirkung, der die Konsumenten vertrauen[69]. Dementsprechend sind Filme mit bekannten Schauspielern häufiger erfolgreich, so die Annahme. Um Wechselwirkungen zwischen den einzelnen unabhängigen Variablen zu vermeiden, sollten Merkmalsvariablen gewählt werden, die nicht miteinander korrelieren. Zumindest sollte man sich darüber im Klaren sein, dass eine Variable, die durch eine zweite zu einem gewissen Anteil erklärt werden kann, die Effekte beider Variablen eine höhere Gewichtung im GesamtModell haben werden. Das Produktionsbudget ist auch deshalb interessant, weil

[67] (Hennig-Thurau & Wruck, 2000)
[68] (Jedidi, Krider, & Weinberg, 1998)
[69] Thurau-Henning und Wruck 2000

typischerweise das Marketingbudget vom Produktionsbudget abhängt[70]. Deshalb sollte „Budget“ nicht aus der Betrachtung ausgeschlossen werden. „ACTORS“ und „DIRECTOR“ können ebenfalls in das Modell eingeschlossen werden, unter der Prämisse, dass die Korrelationen zu überprüfen sind. Es wird angenommen, dass Filme mit einem hohen Produktionsbudget eher erfolgreich sind.

Zurzeit hat die Variable „FSK“ die möglichen Ausprägungen „0“, „6“, „12“, „16“ und „18“. Den nominalen Ausprägungen liegt eine rein sachliche Interpretation ohne echte Rangordnung zugrunde. Laut Clement schränkt man den potentiellen Markt mit zunehmender Altersbegrenzung ein, was sich negativ auf den Filmerfolg auswirken kann. Es wird angenommen, dass altersunbeschränkte Filme ein Marktpotential von 100 % haben. Laut der „United States Age Structure“ der Homepage indexmundi.com, die sich auf aktuelle Daten der CIA beruft, wohnen zurzeit etwa 60 Millionen Menschen zwischen 0 und 14 Jahren in Amerika. Weitere etwa 20 Millionen Menschen sind zwischen 15 und 19. Laut dieser Statistik verläuft die demographische Struktur der Bevölkerung für die Jahre 0-25 recht stabil[71]. Auf dieser Grundlage wird geschätzt, dass 4 Millionen je Jahr Altersbeschränkung mehr vom maximalen Marktpotential verloren gehen. Zwischen 2002 und 2004 hatte Amerika rund 300 Millionen Einwohner. Hieraus lassen sich folgende Prozentsätze ermitteln:

| FSK | Maximales Marktpotential in % | altersbeschränkt für Mio. Menschen (kumuliert) |
|---|---|---|
| 0 | 100 | 0 |
| 6 | 92 | 24 |
| 12 | 84 | 48 |
| 16 | 78,7 | 64 |
| 18 | 76 | 72 |

*Tabelle 4-1 – Einschätzung des maximalen Marktpotentials anhand der Altersklassifikation*

Diese Schätzung vernachlässigt, dass es vermutlich Unterschiede im Konsumverhalten innerhalb der Altersstruktur gibt, aber erlaubt die qualitative, nominale in eine Art „ordinale metrisch“ skalierte Variable „Marktpotential“ umzuwandeln. Es wird angenommen, dass das Marktpotential einzuschränken sich negativ auf den Erfolg eines Films auswirkt.

Die Variable „DIS_POWER“ gibt den Marktanteil der jeweiligen Distributoren an. Laut Clement begünstigt eine starke Position des Distributors den Filmerfolg, was eine weitere Annahme darstellt.
Da der Einfluss von Kritiken laut Clement in späteren Phasen verschwindet und sich nicht auf das Box-Office Ergebnis auswirkt, wird angenommen, dass die Variable „CIN_CRITIC“ keinen Einfluss auf das Modell haben wird und somit nicht weiter zur Erklärung des Filmerfolgs beiträgt. Diese Annahme kann ebenfalls überprüft werden.

| Variablenname | Funktion | Bedeutung |
|---|---|---|
| Rentabilität | Abhängige Variable | Erzielt der Film nach 20 Wochen Gewinne (ja = 1 /nein = 2) |
| DIS_POWER | Prädiktor | Marktanteil des Filmdistributors (in %) |
| CIN_CRITIC | Prädiktor | Bewertung des Films durch Kritiken von (0 bis 5) |
| Marktpotential | Prädiktor | Einschränkung des Marktes durch Altersbeschränkungen (in %) |
| ACTORS | Prädiktor | Bewertung der Besetzung (von 0 bis 5) |
| DIRECTOR | Prädiktor | Bewertung des Regisseurs (von 0 bis 5) |
| BUDGET | Prädiktor | Produktionsbudget (in US$) |
| USUmsatz | Prädiktor | Umsatz der Kinovorstellungen nach 20 Wochen (in US$) |
| USScreens | Prädiktor | Anzahl der Kinovorstellungen nach 20 Wochen |

[70] Thurau-Henning 2004
[71] http://www.indexmundi.com/united_states/age_structure.html

*Tabelle 4-2 – Ausgewählte Variablen für die Diskriminanzanalyse*

Aus diesem Set von möglichen Einflüssen auf die abhängige Variable können mögliche Leitfragen der Untersuchung formuliert werden:

- Haben Marktanteil des Distributors, Kinokritiker, Verringerungen des Marktpotentials auf Grund von Altersbeschränkungen, die Besetzung, der Regisseur, das Budget, der Umsatz im nordamerikanischen Markt und die Anzahl der Filmvorführung signifikanten Einfluss auf den Erfolg von Kinofilmen?
- Beeinflusst die Alterbeschränkung eines Films seinen Erfolg negativ?
- Sind die Filmvorführungen tatsächlich die wichtigsten Erfolgsindikatoren?
- Ist das lineare Modell zur Prognose des allgemeinen Erfolgs von Filmen gebräuchlich?

## 4.5. Holdout Verfahren: Einteilung in Lern- und Kontrollstichprobe

Ein Datensatz wird bei dem Holdout-Verfahren in der Regel in zwei disjunkte Teildatensätze aufgeteilt. Der erste Teil besteht aus den Schätzdaten und wird zur Schätzung der Modellparameter verwendet. Der zweite Teil ist der Prüfdatensatz oder das Holdout Sample. Zu Beginn der Analyse sollte die gesamte Stichprobe zunächst in zwei weitere Stichproben eingeteilt werden. Etwa 30 % der gesamten Stichprobe sollte die Lernstichprobe darstellen, mit welcher das Modell zur Klassifikation erstellt wird, welches anhand der anderen 70% der gesamten Stichprobe getestet wird, um „overfitting“[72] zu vermeiden. Für gute Klassifikationsmodelle mit der Diskriminanzanalyse ist ein angemessener Stichprobenumfang sehr wichtig. Es sollten zu jeder Merkmalsvariable mindestens 20 Beobachtungen vorliegen. Hair et al. [73] definieren sogar 100 Objekte als angemessenen Stichprobenumfang für das Modell. An mindestens weiteren 200 Objekten sollte das Modell schließlich getestet werden. Dabei sollte beachtet werden, dass das Verhältnis der Ausprägungen der abhängigen Gruppierungsvariable in beiden Stichproben in etwa gleich sein sollte. Befinden sich entsprechend des exemplarischen Beispiels in der Gesamtstichprobe 200 erfolgreiche und 500 nicht erfolgreiche Filme, sollte in der 30 % Trainingsstichprobe entsprechend das Verhältnis stimmen und 60 erfolgreiche und 150 nicht erfolgreiche Filme vorliegen[74].

Sollte der gesamte Stichprobenumfang nicht dazu geeignet sein, ihn in zwei Teile zu zerlegen, kann die Diskriminanzanalyse dennoch anhand der gesamten Stichprobe durchgeführt werden. Allerdings sollte man sich dessen bewusst sein, dass „overfitting“ stattfinden kann und das Modell bei einer künftigen Anwendung eine entsprechend geringere Treffergenauigkeit aufweisen könnte.

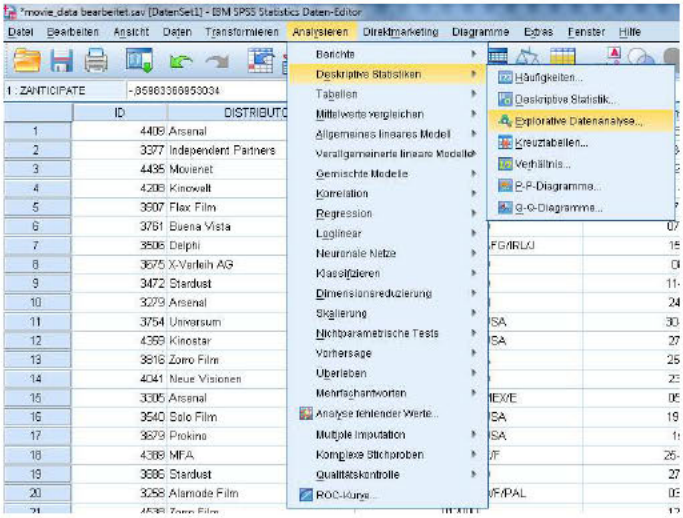
## 4.6. Parametrische Voraussetzungsprüfungen

Anhand eben definierter Prädiktoren soll in SPSS die Diskriminanzanalyse durchgeführt werden. Doch bevor dies geschieht, müssen die Variablen auf multivariate Normalverteilung, Varianz-, und Kovarianzhomogenität der Gruppen getestet, gegebenenfalls transformiert und um Ausreißer und Extremwerte bereinigt werden, um Voraussetzungsfehler für die verschiedenen statistischen Tests des Verfahrens auszuschließen.

### 4.6.1. Test der (multivariaten) Normalverteilung

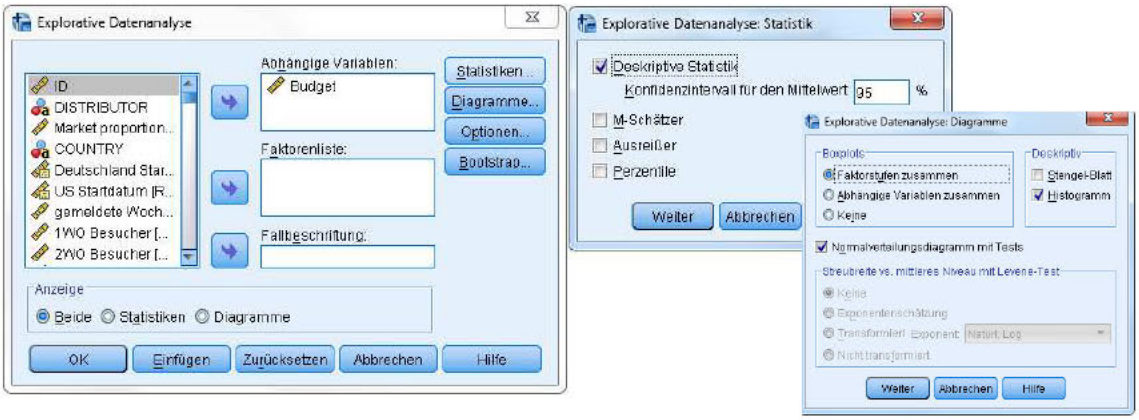
[72] „overfitting“ bedeutet, dass eine zu starke Anpassung des Modells an die vorliegenden Daten stattfindet. Dies verhindert, dass das Modell aussagekräftig auf andere Stichproben oder die Grundgesamtheit angewandt werden kann.
[73] vgl. Hair et al. 19.02
[74] vgl. Kapitel Diskriminanzanalyse von Hair et al. 19.02

Die meisten wahrscheinlichkeitsabhängigen statistischen Verfahren beruhen auf der Häufigkeitsverteilung der Normalverteilung. Um die Normalverteilung von Variablen zu testen, kann man sich verschiedenen Prozeduren in SPSS bedienen. Eine einfache von vielen Methoden, um die Normalverteilung zu prüfen, stellt die Explorative Datenanalyse dar.




*SPSS Bildschirmfoto 4-1 - Analysieren, Deskriptive Statistik, Explorative Datenanalyse ...*

Die Explorative Datenanalyse erlaubt es, sich sowohl über die Diagramme die Histogramme als auch über Statistiken die Streuungsparameter der ausgewählten Merkmalsvariablen ausgeben zu lassen. Des Weiteren bewirkt das Häkchen bei „Normalverteilungsdiagramm mit Tests“ SPSS zwei Signifikanzprüfungen auf Normalverteilung durchführen zu lassen. Darüber hinaus lässt sich über die Streuungsparameter errechnen, ob von einer Normalverteilung ausgegangen werden kann und über die Histogramme und Boxplotts der Sachverhalt grafisch per Augenmaß eingeschätzt werden.




*SPSS Bildschirmfoto 4-2 - Explorative Datenanalyse, Auswahlfelder: Statistik und Diagramme*

**Deskriptive Statistik**

| | | | Statistik | Standardfehler |
|---|---|---|---|---|
| Budget | Mittelwert | | 33532206,35 | 2539598,962 |
| | 95% Konfidenzintervall des Mittelwerts | Untergrenze | 28524978,40 | |
| | | Obergrenze | 38539434,30 | |
| | 5% getrimmtes Mittel | | 29513112,72 | |
| | Median | | 22000000,00 | |
| | Varianz | | 1,322E15 | |
| | Standardabweichung | | 36361523,51 | |
| | Minimum | | 0 | |
| | Maximum | | 210000000 | |
| | Spannweite | | 210000000 | |
| | Interquartilbereich | | 39740000 | |
| | Schiefe | | 1,718 | ,170 |
| | Kurtosis | | 3,406 | ,338 |

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| Budget | ,178 | 205 | ,000 | ,818 | 205 | ,000 |

a. Signifikanzkorrektur nach Lilliefors

*SPSS Bildschirmfoto 4-3 - Tests auf Normalverteilung für die Variable "Budget"*

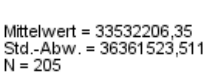
Der Teil des Outputs, der für die Normalverteilung von Interesse ist, ist am Boden der Deskriptiven Statistik Tabelle zu finden. Über die Schiefe und Kurtosis kann in Abhängigkeit des jeweiligen Standardfehlers auf eine Normalverteilung geschlossen werden. Faustregel für *keine* Normalverteilung sind Beträge von >0.8 der Schiefe und Beträge von >3 der Kurtosis. Der Quotient aus der Kurtosis sowie der Schiefe und deren Standardfehler kann ebenfalls als Test auf Normalverteilung verwendet werden. Man kann die Normalverteilung ausschließen, wenn einer der Quotienten unter -2 oder über +2 liegt.[75]

$\frac{1.718}{0.170} = 10.11$; folglich kann durch die Ausprägung der Schiefe angenommen werden, dass die Normalverteilung ausgeschlossen werden kann.

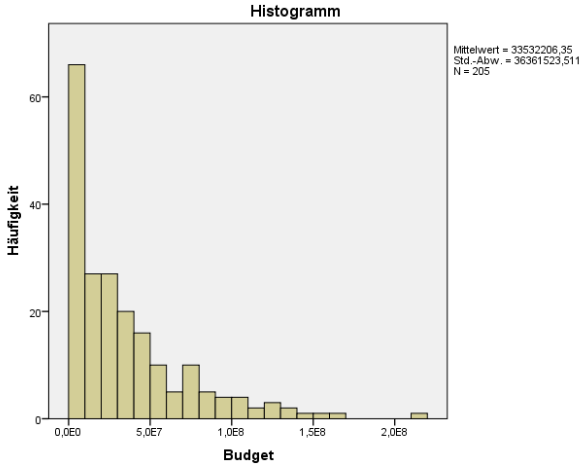
$\frac{3.406}{0.338} = 10.08$ ; folglich kann durch die Ausprägung der Kurtosis ebenfalls angenommen werden, dass die Normalverteilung ausgeschlossen werden kann.

Darunter sind im Output die zwei „Tests auf Normalverteilung“, die die Nullhypothese prüfen, ob die Daten nicht von einer Normalverteilung abweichen und umgekehrt der Alternativhypothese, dass die Daten einer Normalverteilung entsprechen. Signifikanzwerte $< 0.05$ sprechen für Abweichungen von einer Normalverteilung.

Im Fall des Merkmals „Budget“ sind die Tests auf Normalverteilung beide signifikant. Die Nullhypothese muss abgelehnt werden. Die Daten entsprechen keiner Normalverteilung. Die positive Schiefe spricht für eine linkssteile (rechtschiefe) Verteilung der Daten. Die positive Kurtosis ist Indikator für eine Schmalgipfligkeit der Verteilung. Dies zeigt auch das Histogramm des Merkmals, welches die Häufigkeitsverteilung der Datenpunkte anhand eines Balkendiagrams darstellt.




[75] https://www.ibm.com/support/knowledgecenter/




*SPSS Bildschirmfoto 4-4 - Histogramm der Dichteverteilung der Variable Budget*

Die Untersuchung auf Normalverteilung muss für jede der ausgewählten Prädiktoren erfolgen.

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| CIN_CRITIC | ,244 | 205 | ,000 | ,880 | 205 | ,000 |
| USUmsatz | ,272 | 205 | ,000 | ,613 | 205 | ,000 |
| DIS_POWER | ,176 | 205 | ,000 | ,871 | 205 | ,000 |
| Budget | ,178 | 205 | ,000 | ,818 | 205 | ,000 |
| Marktpotential | ,292 | 205 | ,000 | ,846 | 205 | ,000 |
| USScreens | ,166 | 205 | ,000 | ,872 | 205 | ,000 |
| DIRECTOR | ,321 | 205 | ,000 | ,546 | 205 | ,000 |
| ACTORS | ,115 | 205 | ,000 | ,935 | 205 | ,000 |

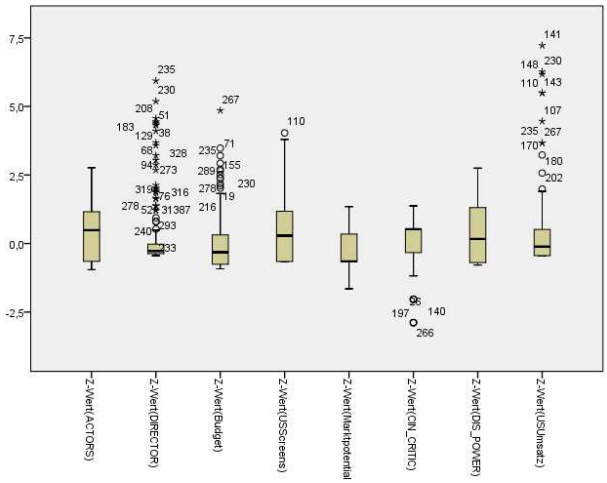
a. Signifikanzkorrektur nach Lilliefors

*SPSS Bildschirmfoto 4-5 - Tests auf Normalverteilung für alle Prädiktoren*

Der Test auf Normalverteilung ist für beide Test-Varianten bei allen Variablen signifikant. Keine der vorliegenden Variablen ist normalverteilt.

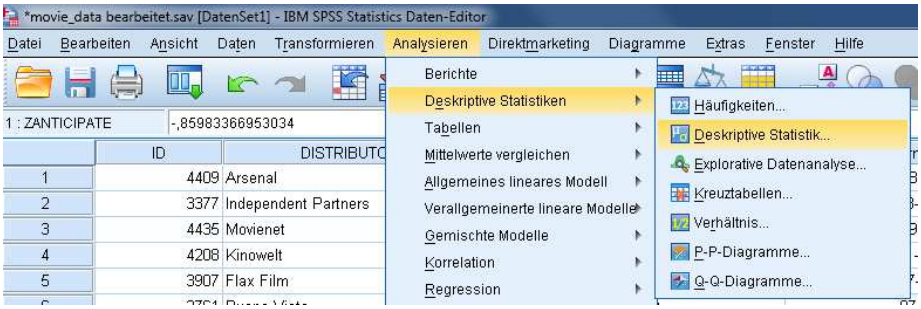
Über die Boxplotts (SPSS Bildschirmfoto 4-6) kann man sich ebenfalls einen Eindruck verschaffen, wie die Verteilung der Daten aussieht. Normalverteilte Variablen haben symmetrische Boxplotts, d.h. die 0.25- und 0.75-Quartilgrenzen sind gleich weit vom Median entfernt und die Whiskers etwa gleich lang. Lässt man sich für die vorliegenden Variablen die Boxplotts ihrer standardisierten[76] Form ausgeben, stellt man fest, dass viele Ausreißer (Kreise;○) und Extremwerte (Sterne;★) die Dichteverteilung beeinflussen.

[76] für „Standardisieren einer Variablen“ siehe SPSS Bildschirmfoto 4-8 und (Formel 4:1 – z-Transformation zum Standardisieren von Variablen)



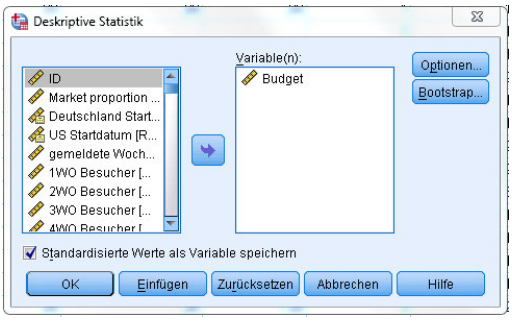

*SPSS Bildschirmfoto 4-6 - Gruppiertes Boxplottdiagramm zum Vergleich der Dichteverteilungen*

Also sollte als nächstes die Ausreißer- und Extremwert-Analyse erfolgen. Hierfür gibt es viele Verfahren. Ein sehr gängiges und einfaches Verfahren ist, die betrachteten Variablen zu standardisieren und Werte, die weiter als ein bestimmter Faktor der Standardabweichung vom Mittelwert entfernt liegen, als Ausreißer und Extremwerte zu deklarieren und anschließend von der weiteren Untersuchung auszuschließen. Dies ist daher sehr einfach, da die Standardabweichung bei standardisierten Variablen 1 und der Mittelwert 0 ist. Folglich können Fälle ausgeschlossen werden, die Ausprägungen der betrachteten Merkmalsvariablen aufweisen, die größer sind als der entsprechende Faktor. Die einfachste Art und Weise, aus Merkmalsvariablen ihre standardisierte Form zu erzeugen, ist über die Deskriptive Statistik.




*SPSS Bildschirmfoto 4-7 - Analysieren, Deskriptive Statistiken, Deskriptive Statistk..*

Im nächsten Schritt werden die Merkmalsvariablen in die Variablenspalte übergeführt, die für die Betrachtung relevant sind. Durch das setzen des Häkchens bei „Standardisierte Werte als Variable abspeichern“




*SPSS Bildschirmfoto 4-8 - Deskriptive Statistik, Auswahl der Variablen*

erzielt man, dass SPSS den folgenden Rechenvorgang automatisch ausführt und die Ergebnisse in einer neuen Variable abspeichert:

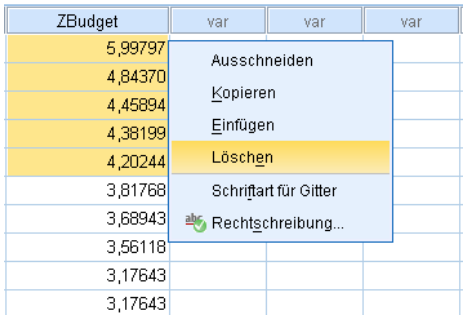
$$Z - Wert_i = \frac{Variable_i - \mu_i}{\sigma_i}$$

*(Formel 4:1 – z-Transformation zum Standardisieren von Variablen)*

Mit:

$$\text{Z} - \text{Wert}_j = \text{transformierte, standardisierte Variable}_j$$
$$\mu = \text{Mittelwert}$$
$$\sigma = \text{Standardabweichung}$$

Man spricht bei dieser Prozedur auch von der sogenannten z-Transformation. Da Variablen häufig verschiedene Messniveaus oder physikalische Dimensionen, wie Euro, Kilogramm, Kilometer oder ähnliches aufweisen, kann man die Ausprägungen nicht miteinander vergleichen. Die z-Transformation bereinigt die Variablen um ihre Skalenniveaus und reduziert die Verteilung der Ausprägungen immer auf denselben Wertebereich mit 0 als Mittelwert und 1 als Standardabweichung. Dadurch werden Variablen mit unterschiedlichen Skalenniveaus vergleichbar. Der zweite Vorteil, den diese Transformation mit sich bringt, ist, dass bei angenommener Normalverteilung immer etwa 68% aller Messwerte innerhalb von einer Standardabweichung, 95% innerhalb von zwei Standardabweichungen und 99.7% aller Werte innerhalb von drei Standardabweichungen liegen. Werte, die über 3 Standardabweichungen vom Mittelwert entfernt liegen, sind mit sehr hoher Wahrscheinlichkeit Ausreißer und Extremwerte und repräsentieren nicht den statistischen Normalfall. Sie verzerren Rückschlüsse, die auf die Grundgesamtheit getroffen werden und können daher von den Untersuchungen ausgeschlossen werden. Bei keiner Normalverteilung der Häufigkeiten, wie soeben getestet wurde, kann unter Beachtung des Theorems von Tschebyscheff[77] angenommen werden, dass zumindest 94% der Daten innerhalb von 4 Standardabweichungen liegen und daher bei jeder anderen, nicht normalverteilten Häufigkeitsverteilung der Daten Ausprägungen über 4 hinaus als Ausreißer deklariert und entfernt werden können. Bevor man irreversible Änderungen am Datensatz vornimmt, sollte man die Originaldatei sichern, denn künftige Analysen sind eventuell nicht Ausreißer empfindlich oder betrachten andere Merkmale und könnten von dem Informationswert der zusätzlichen Objekte profitieren. Diese Vorgehensweise erfolgt für jede Merkmalsvariable.




*SPSS Bildschirmfoto 4-9- Entfernen der Mittelwerte*

Da Ausreißer und Extremwerte die Verteilung der Daten und Lageparameter verfälschen können, muss nach ihrer Bereinigung der Test auf Normalverteilung erneut durchgeführt werden. Hierfür wird erneut die bereits vorgestellte Deskriptive Statistik aufgerufen und inspiziert.

[77] Theorem von Chebyshev, http://www.statistics4u.info/fundstat_germ/cc_chebyshev.html

**Deskriptive Statistik**

| | | | Statistik | Standardfehler |
|---|---|---|---|---|
| Budget | Mittelwert | | 30316848,48 | 2319646,945 |
| | 95% Konfidenzintervall des Mittelwerts | Untergrenze | 25742031,38 | |
| | | Obergrenze | 34891665,58 | |
| | 5% getrimmtes Mittel | | 26637599,01 | |
| | Median | | 20000000,00 | |
| | Varianz | | 1,055E15 | |
| | Standardabweichung | | 32475057,23 | |
| | Minimum | | 0 | |
| | Maximum | | 160000000 | |
| | Spannweite | | 160000000 | |
| | Interquartilbereich | | 36653500 | |
| | Schiefe | | 1,648 | ,174 |
| | Kurtosis | | 2,705 | ,346 |

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| Budget | ,175 | 196 | ,000 | ,821 | 196 | ,000 |

a. Signifikanzkorrektur nach Lilliefors

*SPSS Bildschirmfoto 4-10 – Tests auf Normalverteilung*

In der Betrachtung der erneuten Auswertung der Variable „Budget“ ändert das Bereinigen der Ausreißer und Extremwerte nicht viel. Quotienten der Schiefe und Kurtosis und ihrer Standardfehler liegen nach wie vor oberhalb der Grenzwertebeträge von 2 vor, sodass die Normalverteilungsannahme verworfen werden kann. Hierfür sprechen auch die Tests auf Normalverteilung. Um eine solche nicht-normalverteilte Variable für Analysen, die die Normalverteilung voraussetzen, verwenden zu können, kann eine lineare Skalentransformation vorgenommen werden. Eine Form der Skalentransformation ist die z-Transformation zum Standardisieren von Variablen. Skalentransformationen können aber auch verwendet werden, um die Dichtefunktion einer Verteilung eine andere Form annehmen zu lassen. Skalentransformationen geschehen durch Veränderungen anhand einer oder mehreren arithmetischen Konstanten. Diese Veränderungen bewirken nicht, dass sich die linearen Beziehungen der Datenpunkte zu einander verändern. Transformierte Merkmalsausprägungen verlieren ihre Skalen und können alleinstehend ohne Weiteres, etwa das Rückwärtsrechnen der vorgenommenen Transformation, sachlich nicht mehr interpretiert werden. Die Transformation der Daten bewirkt zusätzlich, dass lineare Beziehungen zwischen den Variablen, wie die Korrelation, sich verändert. Um dies zu vermeiden, muss man die entsprechende Transformation an jeder ins Verhältnis gesetzten Variablen durchführen. Verzerrte Verteilungsfunktionsformen können so in (annähernde) Normalverteilungsform überführt werden[78]. Dabei gibt es für verschiedene Verzerrungen verschiedene Transformationsansätze. Hier einige Beispiele:

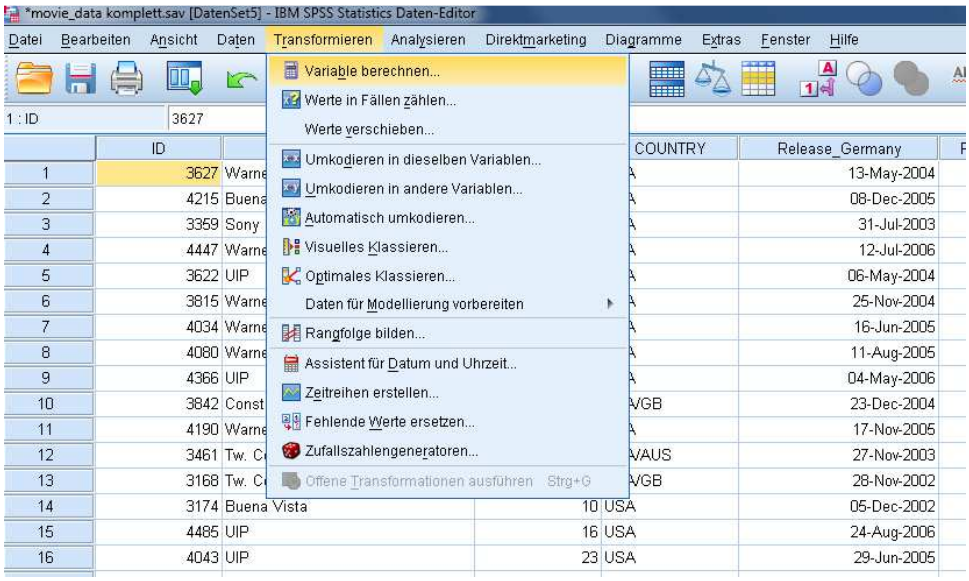
| Art der Transformation | Formel | Transformation |
|---|---|---|
| Z-Transformation | $\frac{X-\mu}{\sigma}$ | überführt (beliebig) normalverteilte Variablen in eine Standardnormalverteilung |
| Reziproke Transformation (Inverse) | $\frac{1}{X}$ | stark linkssteile Dichtefunktionen werden symmetrischer |
| Logarithmische Transformation | $ln(X)$ | linkssteile Dichtefunktionen werden symmetrischer |
| Quadratische Transformationen | $X^n$ | (stark) rechtssteile Dichtefunktionen werden symmetrischer |
| Wurzeltransformation | $\sqrt{X}$ | stark rechtssteile Dichtefunktionen werden symmetrischer |

[78] Nothnagel und Berlin 05.10; https://statistikguru.de/

*Tabelle 4-3:verschiedene Transformationsansätze*

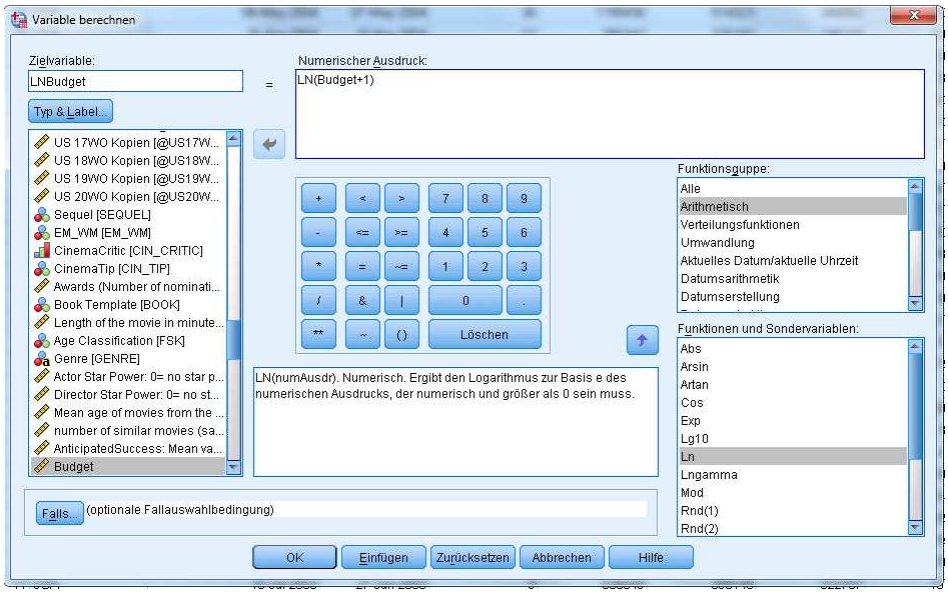
Entscheidend ist, dass beliebige Kombinationen von Transformationen an einer Variablen angewandt werden können, solange man die Schritte dokumentiert, sie konsequent an jedem in eine Beziehung gebrachten Merkmals durchführt und in der Lage ist die entsprechende Umkehrfunktion zu formulieren, um die Originalwerte und deren zusätzlichen Informationsgehalt nicht zu verlieren.

Da die Variable „Budget“ spitzgipflig linkssteil ist, kann man in diesem Fall versuchen eine logarithmische Transformation mit dem natürlichen Logarithmus anzuwenden. Hierfür wird die „Variable berechnen“-Funktion in SPSS genutzt.




*SPSS Bildschirmfoto 4-11 – Transformieren, Variable berechnen…*

Über das Auswahlfeld der Funktionsgruppen gelangt man zu möglichen Rechenoperatoren. Hierunter ist auch der natürliche Logarithmus „ln“ zu finden. Bei Variablen mit Merkmalsausprägungen $\leq 0$ ist bei der logarithmischen Transformation zu beachten, dass man vorher mit das $Minimum\ der\ Variable + 1$ addiert, um alle Ausprägungen größer als 0 werden zu lassen. Die Variable Budget besitzt laut der bereits betrachteten Deskriptiven Statistik ihr Minimum bei 0. Um dieser Regel gerecht zu werden, addiert man konstant 1 über alle Fälle der Variablen „Budget“ und transformiert anschließend mit dem natürlichen Logarithmus. Das sieht in SPSS wie folgt aus:




*SPSS Bildschirmfoto 4-12 - Variable berechnen…*

Nach dem definieren der Zielvariablen schließt man den Vorgang ab und testet die neue Variable über die explorative Datenanalyse erneut auf Normalverteilung.

**Deskriptive Statistik**

| | | | Statistik | Standardfehler |
|---|---|---|---|---|
| LnBudget | Mittelwert | | 16,1768 | ,19403 |
| | 95% Konfidenzintervall des Mittelwerts | Untergrenze | 15,7942 | |
| | | Obergrenze | 16,5595 | |
| | 5% getrimmtes Mittel | | 16,5470 | |
| | Median | | 16,8112 | |
| | Varianz | | 7,379 | |
| | Standardabweichung | | 2,71648 | |
| | Minimum | | ,00 | |
| | Maximum | | 18,89 | |
| | Spannweite | | 18,89 | |
| | Interquartilbereich | | 2,10 | |
| | Schiefe | | -3,910 | ,174 |
| | Kurtosis | | 19,461 | ,346 |

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| LnBudget | ,182 | 196 | ,000 | ,620 | 196 | ,000 |

a. Signifikanzkorrektur nach Lilliefors

*SPSS Bildschirmfoto 4-13 – erneute Tests auf Normalverteilung*

Wieder stellt man anhand der bekannten Testwerte fest, dass keine Normalverteilung vorliegt. Um auf der sicheren Seite zu sein, sollte man eine solche Variable dementsprechend nicht in einer linearen Diskriminanzanalyse verwenden, da sie die Verteilungsvoraussetzungen verletzt. Alternativ könnte man auf ein verteilungsfreies Verfahren umstellen, wie die Quadratische Diskriminanzanalyse, wo eine derartige Variable keine Analysevoraussetzungen verletzt.

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| ACTORS | ,112 | 146 | ,000 | ,950 | 146 | ,000 |
| DIRECTOR | ,328 | 146 | ,000 | ,551 | 146 | ,000 |
| Budget | ,157 | 146 | ,000 | ,864 | 146 | ,000 |
| USScreens | ,080 | 146 | ,022 | ,956 | 146 | ,000 |
| Marktpotential | ,282 | 146 | ,000 | ,854 | 146 | ,000 |
| CIN_CRITIC | ,246 | 146 | ,000 | ,886 | 146 | ,000 |
| DIS_POWER | ,149 | 146 | ,000 | ,898 | 146 | ,000 |
| USUmsatz | ,139 | 146 | ,000 | ,888 | 146 | ,000 |

a. Signifikanzkorrektur nach Lilliefors

*SPSS Bildschirmfoto 4-14 – erneute Tests auf Normalverteilung für alle Prädiktoren*

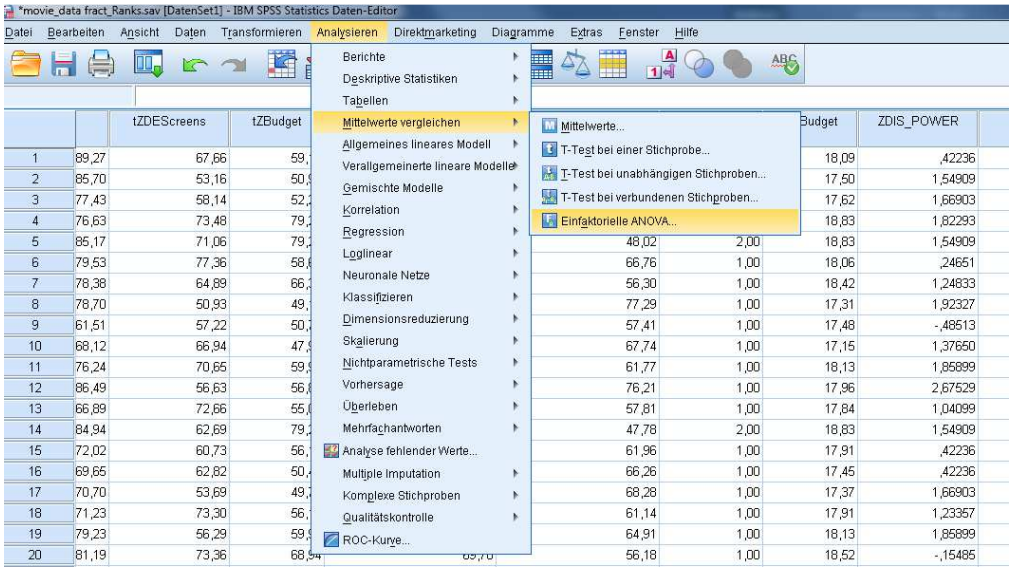
Auch die anderen Variablen verletzen die Annahme der Normalverteilung, wie hier an den Tests auf Normalverteilung zu sehen ist. Die Teststatistiken sind signifikant. Also kann bei keiner Variablen die Nullhypothese der Normalverteilung angenommen werden.

Für die weitere *reine exemplarische* Erläuterung der linearen Diskriminanzanalyse wird die Normalverteilung aller Prädiktoren an dieser Stelle einfachheitshalber unterstellt. Zudem ist das Verfahren der linearen Diskriminanzanalyse diese Voraussetzung betreffend insgesamt recht robust.

### 4.6.2. Test der Homogenität der Varianzen

Als nächstes soll in Abhängigkeit der Gruppierungsvariablen festgestellt werden, ob die Varianzen der Gruppen gleich sind. Dieser Test unterliegt üblicherweise ebenfalls der Voraussetzung der Normalverteilung, sowie der Unabhängigkeit der Stichproben. Unabhängigkeit der Stichproben liegt dann vor, wenn die Wahrscheinlichkeit einer bestimmten Merkmalsausprägung in jeder Stichprobe dieselbe ist. Abhängige Stichproben würden beispielsweise durch einen Faktor abhängig sein, wie dem Erheben der Stichproben zu zwei verschiedenen Zeitpunkten mit unterschiedlichen Einflussfaktoren. Die Stichprobe würde in diesem Fall von den Einflussfaktoren abhängen. Da die Gruppierungsvariable festgelegt wurde, die die ursprüngliche Stichprobe in zwei oder mehr Gruppen aufteilt, kann der Schluss gezogen werden, dass die Stichproben von keinen nicht zufälligen Einflussfaktoren abhängen.

In SPSS wird die Homogenität der Varianzen anhand des Levene-Tests festgestellt. Dieser lässt sich durch die „Einfaktorielle ANOVA „im Reiter „Mittelwerte vergleichen“ aufrufen.




*SPSS Bildschirmfoto 4-15 – Analysieren, Mittelwerte vergleichen, Einfaktorielle AnOVA …*

Die Überlegung zur Auswahl der Prädiktoren für die Diskriminanzanalyse hat ergeben, dass die Variablen „US Kopien“ und „Budget“ vermutlich den größten Einfluss auf den Erfolg haben. Dadurch, dass „Kinovorführungen“ und „Budget in US$“ als Werte nicht miteinander verglichen werden können, müssen die Variablen transformiert werden. Das gilt auch für die anderen Prädiktoren. Hierfür wird die bereits erwähnte z-Transformation angewandt. Sie ändert nichts an den linearen Beziehungen der Variablen untereinander.

Im Auswahlfenster der einfaktoriellen ANOVA werden nun relevante Prädiktoren dem „Abhängige Variablen“-Feld hinzugefügt. Als „Faktor“ wird die Variable eingefügt, die der Diskriminanzanalyse als abhängige Gruppierungsvariable zugrunde liegt. In diesem Fall ist es die binäre Variable Rentabilität_nach_US, die die Objekte, wie vorher definiert, in gewinn- und nicht-gewinnbringende Kinofilme unterteilt. Als nächstes müssen die Statistiken definiert werden, die sich hinter dem Options-Button verbergen. Hinter der Angabe „Test auf Homogenität der Varianzen“ verbirgt sich der Levene-Test. Dieser testet die Annahme (Nullhypothese), dass die Varianzen der Variablen in Abhängigkeit ihrer Gruppierung gleich sind. Der Levene-Test ist gegenüber der Verletzung der Annahme der Normalverteilung der Daten robust.


*SPSS Bildschirmfoto 4-16 – Einfaktorielle ANOVA, Auswahlfenster: Optionen*

Durch das Bestätigen der Inputs errechnet SPSS in der Ausgabe den „Test der Homogenität der Varianzen“.

**Test der Homogenität der Varianzen**

| | Levene-Statistik | df1 | df2 | Signifikanz |
|---|---|---|---|---|
| Z-Wert(ACTORS) | ,135 | 1 | 194 | ,713 |
| Z-Wert(DIRECTOR) | ,070 | 1 | 194 | ,791 |
| Z-Wert(Budget) | 15,114 | 1 | 194 | ,000 |
| Z-Wert(USScreens) | 2,408 | 1 | 144 | ,123 |
| Z-Wert(Marktpotential) | 1,561 | 1 | 194 | ,213 |
| Z-Wert(CIN_CRITIC) | 2,397 | 1 | 194 | ,123 |
| Z-Wert(DIS_POWER) | ,165 | 1 | 194 | ,685 |
| Z-Wert(USUmsatz) | 10,969 | 1 | 144 | ,001 |

*SPSS Bildschirmfoto 4-17 – Test der Homogenität der Varianzen*

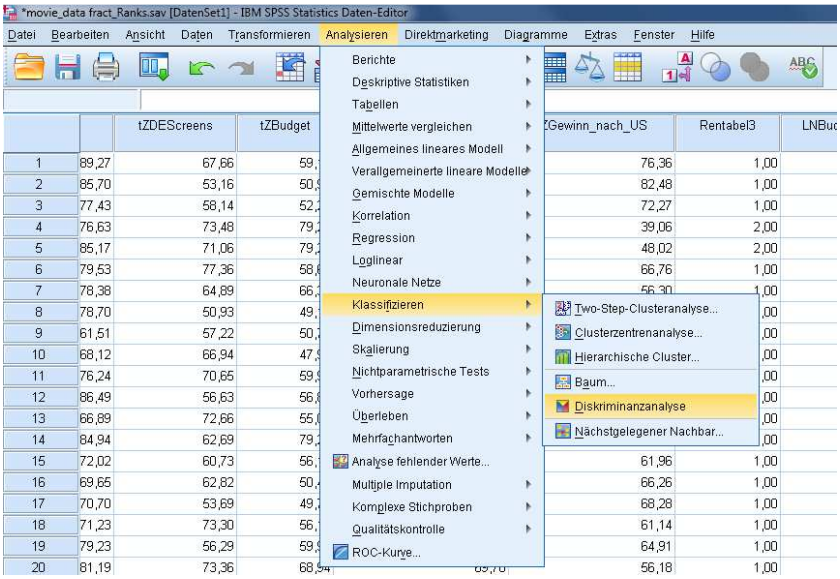
Die Auswertung der Tabelle zeigt, dass die Levene-Teststatistik für die transformierten Variablen „ACTORS“, „DIRECTOR“, „USScreens“, „Marktpotential“, „CIN_CRITIC“ und „DIS_POWER“ in Abhängigkeit der Gruppierungsvariable nicht signifikant ist. Für diese Variablen kann die Nullhypothese nicht abgelehnt werden. Bei diesen Variablen liegt eine Homogenität der Varianzen der Gruppen vor oder mit anderen Worten signifikante Gruppenmittelwertunterschiede der abhängigen Variablen.

Anders müssen die übrigen Variablen „Budget“, „USUmsatz“ bewertet werden. Wie an SPSS Bildschirmfoto SPSS Bildschirmfoto 4-17 – Test der Homogenität der Varianzenzu sehen ist, ist der Levene-Test für diese Variablen signifikant, was bedeutet, dass die Nullhypothese abgelehnt wird. Stattdessen wird von der Alternativhypothese ausgegangen, die besagt, dass keine Homogenität der Varianzen in Abhängigkeit der Gruppierungsvariablen vorliegt. Variablen, für die der Levene-Test ein signifikantes Ergebnis ausgibt, sollten als Prädiktor für die Diskriminanzfunktion ausgeschlossen werden.

### 4.6.3. Test der Homogenität der Kovarianz-Varianz-Matrizen der Gruppen

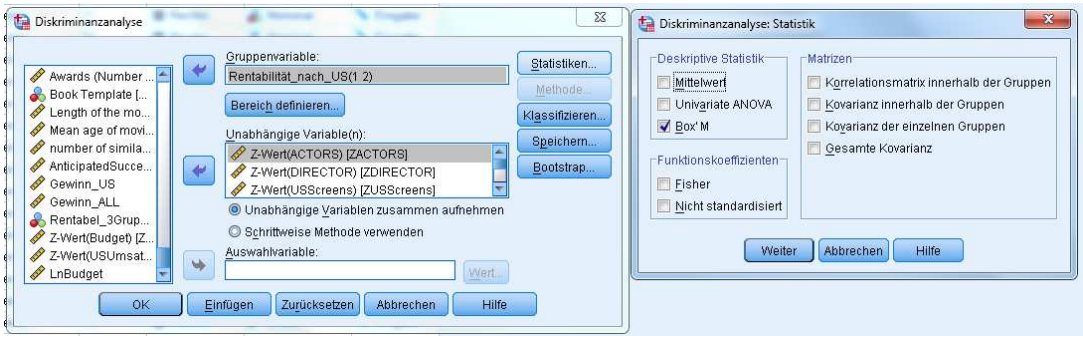
Als nächstes werden die Kovarianz-Varianz-Matrizen der Gruppen auf Homogenität überprüft. Grundsätzlich prüft der Box’s M Test diesen Sachverhalt. Die Nullhypothese des Tests ist, dass gleiche gruppenspezifische Kovarianzen wie die der Grundgesamtheit vorliegen. Die Prüfung dieser Hypothese unterliegt der Voraussetzung, dass die Merkmalsvariablen multivariat normalverteilt sind. Dieser Annahme bezüglich reagiert der Box’s M Test extrem sensibel. Außerdem sollte ein großer Stichprobenumfang vorliegen, denn die Toleranz (alpha-Fehler) für diesen sehr sensiblen Test liegt bei 0.001 (strengeres Signifikanzniveau). Zu beachten ist, dass der Test bei Hinzunahme oder Ausschluss eines Prädiktors vom Modell jedes Mal erneut durchgeführt werden muss, da in der Kovarianz-Varianz-Matrix die Prädiktoren alle in einer Beziehung stehen. Daher ist eine vorgeschaltete Varianzanalyse der einzelnen Variablen sinnvoll, um sich einen Überblick über entsprechende Abweichungen zu machen.

Der Box M Test ist Inhalt der Diskriminanzanalyse selbst. Er kann durch den Menüpunkt „Klassifizierung“ und „Diskriminanzanalyse“ aufgerufen werden.




*SPSS Bildschirmfoto 4-18 - Analysieren, Klassifizieren, Diskriminanzanalyse*

Die „Gruppenvariable“ ist die bereits bekannte Rentabilitätsvariable. Als „Unabhängige Variablen“ werden die zu testenden Prädiktoren ausgewählt, bei denen vorher eine vorliegende Varianzhomogenität bestätigt wurde. Über die Statistiken kann man im Bereich der Deskriptiven Statistik den Box’s M-Test aufrufen. Durch „Weiter“ und „Ok“ erhalten wir den Output des Tests.




*SPSS Bildschirmfoto 4-19 – Diskriminanzanalyse, Auswahlfenster:Statistik*

Im Output des Tests findet man unter der Überschrift „Box-Test auf Gleichheit der Kovarianz-Matrizen“ zwei Tabellen. Eine weitere Tabelle erscheint nur in dem speziellen Fall, wenn der Box’s M-Test durch einen ebenfalls durchgeführten eigenen Toleranztest eine Variable als „durchgefallen“ bewertet. Diese Variable(n) schließt Box’s M automatisch aus der Berechnung aus.

**Log-Determinanten**

| Rentabilität_nach_US | Rang | Log-Determinante |
|---|---|---|
| 1,00 | 6 | -1,168 |
| 2,00 | 6 | -1,077 |
| Gemeinsam innerhalb der Gruppen | 6 | -,964 |

Die Ränge und natürlichen Logarithmen der ausgegebenen Determinanten sind die der Gruppen-Kovarianz-Matrizen.

**Textergebnisse**

| | | |
|---|---|---|
| Box-M | | 21,431 |
| F | Näherungswert | ,972 |
| | df1 | 21 |
| | df2 | 54851,305 |
| | Signifikanz | ,495 |

Testet die Null-Hypothese der Kovarianz-Matrizen gleicher Grundgesamtheit.

*SPSS Bildschirmfoto 4-20 - Box' M Test nach Homogenität der Varianzen*

Weitere Tabellen sind vorerst nicht relevant. Der Box’s M Test berechnet die Determinanten der Varianz-Kovarianz-Matrix für die Gruppen und vergleicht die natürlichen Logarithmen der Werte miteinander. Bereits bei einer Signifikanz von 0,001 würde man den Box-M Testwert als nicht signifikant bewerten, was zur nicht-ablehnung der Nullhypothese führen würde. In diesem Fall zeigt der Box’s M Test eine Signifikanz von 0,495 an. Das ist ein nicht signifikantes Testergebnis. Die Nullhypothese kann nicht abgelehnt werden. Homogenität der Kovarianzen der Gruppen der Prädiktoren liegt vor.

### 4.6.4. Zusammenfassung der Voraussetzungsprüfungen

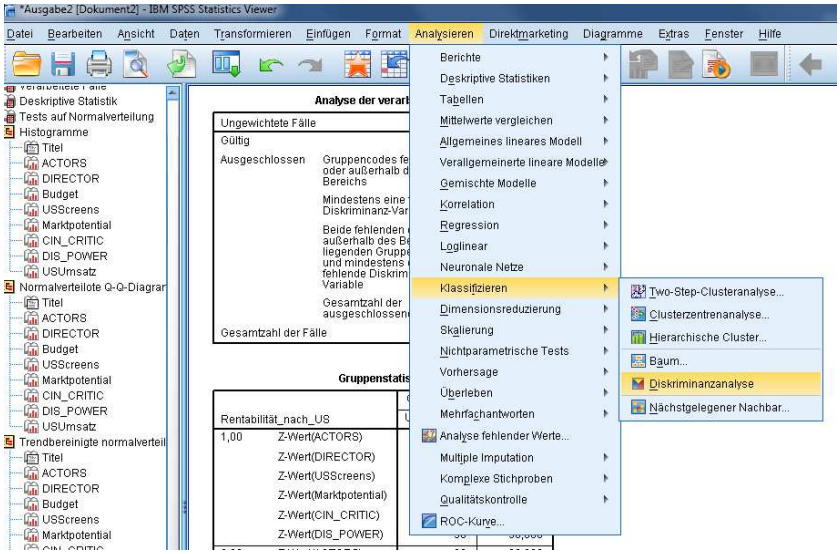
Von allen in Frage kommenden Variablen verletzen nur zwei Variablen die Voraussetzungen, nämlich die über 20 Wochen kumulierte Umsätze „USUmsatz“ und das Produktionsbudget für den Film „Budget“. Übrige metrische Prädiktoren für Schätzung der Diskriminanzfunktion sind: „ACTORS“, „DIRECTOR“, „USScreens“, „Marktpotential“, „CIN_CRITIC“, „DIS_POWER“. Mit diesem Ergebnis kann mit der eigentlichen linearen Diskriminanzanalyse begonnen werden.

| Variablenname | Funktion | Bedeutung |
|---|---|---|
| Rentabilität | Abhängige Variable | Erzielt der Film nach 20 Wochen Gewinne (ja = 1 /nein = 2) |
| „DIS_POWER | Prädiktor | Marktanteil des Filmdistributors (in %) |
| CIN_CRITIC | Prädiktor | Bewertung des Films durch Kritiken von (0 bis 5) |
| Marktpotential | Prädiktor | Einschränkung des Marktes durch Altersbeschränkungen (in %) |
| ACTORS | Prädiktor | Bewertung der Besetzung (von 0 bis 5) |
| DIRECTOR | Prädiktor | Bewertung des Regisseurs (von 0 bis 5) |
| ~~BUDGET~~ | ~~Prädiktor~~ | ~~Produktionsbudget (in US$)~~ |
| ~~USUmsatz~~ | ~~Prädiktor~~ | ~~Umsatz der Kinovorstellungen nach 20 Wochen (in US$)~~ |
| USScreens | Prädiktor | Anzahl der Kinovorstellungen nach 20 Wochen |

*Tabelle 4-4 – Ausschluss der Prädiktoren, die Voraussetzungsannahmen verletzen*

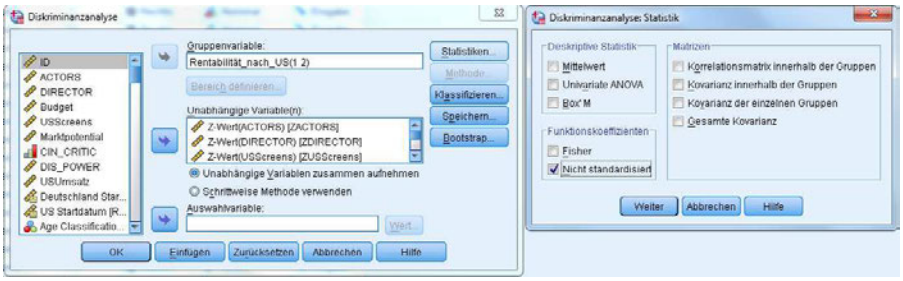
## 4.7. Durchführung der linearen Diskriminanzanalyse mit SPSS

Um die Schätzung der Diskriminanzfunktion anhand der festgelegten Prädiktoren durchzuführen, ruft man über „Analysieren“ unter dem Reiter „Klassifizieren“ erneut die „Diskriminanzanalyse“ auf.




*SPSS Bildschirmfoto 4-21 - Analysieren, Klassifizieren, Diskriminanzanalyse*

Um die Diskriminanzfunktionskoeffizienten für die Klassifizierungsfunktion des Modells ausgegeben zu bekommen, muss „nicht standardisiert“ bei den „Funktionskoeffizienten“ ausgewählt werden.

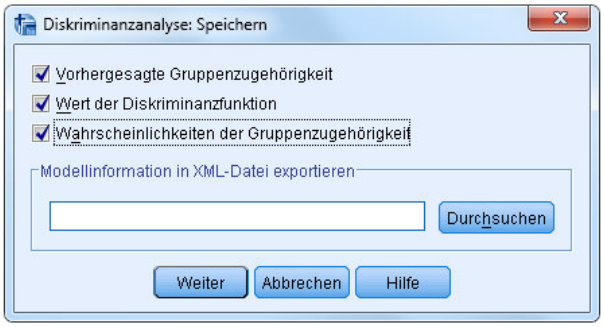



*SPSS Bildschirmfoto 4-22 – Diskriminanzanalyse, Auswahlfenster: Statistik*

Des Weiteren sollte man unter den Klassifizieren-Optionen bei einer Stichprobe mit ungleichen Gruppengrößen die a-priori-Wahrscheinlichkeiten „aus der Gruppengrößen berechnen“ lassen. Unter „Anzeige“ hat man die

Möglichkeit, sich einige Klassifikationsergebnisse im Output anzeigen zu lassen. Die zusammenfassende Tabelle ist die Klassifikationsmatrix, anhand der die Güte des Modells bewertet werden kann. Für den zwei Gruppen Fall zeigt SPSS als Diagramm untereinander die gruppenspezifischen Histogramme an. Erst ab einer Gruppengröße von > 2 hat man die Möglichkeit, sich über die Diagramme ein Streudiagramm der einzelnen und kombinierten Gruppendiskriminanzwerte und Gruppen-Zentroide ausgeben zu lassen. Das Territorien Diagramm gibt grafische Rückschlüsse über die linearen Trennachsen im Diskriminanzraum. Über das Optionsfeld „Speichern“ kann angegeben werden, dass man die Wahrscheinlichkeitswerte der Gruppenzugehörigkeit als Variable speichern möchte. Durch die entstehenden Variablen hat man die Möglichkeit, sich den Sachverhalt manuell zu visualisieren. Abgespeichert sollten ebenfalls die vorhergesagte Gruppenzugehörigkeit und der dazugehörige Diskriminanzwert.


*SPSS Bildschirmfoto 4-23 – Diskriminanzanalyse, Auswahlfenster: Klassifiziere*




*SPSS Bildschirmfoto 4-24 – Diskriminanzanalyse, Auswahlfenster: Speichern*

Durch das Bestätigen der Eingaben gelangt man zum Output.

## 4.8. Interpretation der Diskriminanzanalyse anhand des Outputs von SPSS

Hinter der Analyse der verarbeiteten Fälle findet man in der Gruppenstatistik Informationen zur Häufigkeitsverteilung der Gruppen und über die gültigen Werte, die im Modell verwendet werden. Für dieses Modell stehen zur Schätzung der Diskriminanzkoeffizienten 146 Objekte zur Verfügung. 58 davon gehören zur Gruppe 1 „erfolgreich“ und 88 zur Gruppe 2 „nicht erfolgreich“. Das spiegelt die zuvor kennengelernte These wieder, dass 6 bis 7 von 10 Filmen erfolgreich sind und deutet auf ein realistisches Klassifikationsszenario hin.

**Gruppenstatistik**

| Rentabilität_nach_US | | Gültige Werte (listenweise) Ungewichtet | Gewichtet |
|---|---|---|---|
| 1,00 | Z-Wert(ACTORS) | 58 | 58,000 |
| | Z-Wert(DIRECTOR) | 58 | 58,000 |
| | Z-Wert(USScreens) | 58 | 58,000 |
| | Z-Wert(Marktpotential) | 58 | 58,000 |
| | Z-Wert(CIN_CRITIC) | 58 | 58,000 |
| | Z-Wert(DIS_POWER) | 58 | 58,000 |
| 2,00 | Z-Wert(ACTORS) | 88 | 88,000 |
| | Z-Wert(DIRECTOR) | 88 | 88,000 |
| | Z-Wert(USScreens) | 88 | 88,000 |
| | Z-Wert(Marktpotential) | 88 | 88,000 |
| | Z-Wert(CIN_CRITIC) | 88 | 88,000 |
| | Z-Wert(DIS_POWER) | 88 | 88,000 |
| Gesamt | Z-Wert(ACTORS) | 146 | 146,000 |
| | Z-Wert(DIRECTOR) | 146 | 146,000 |
| | Z-Wert(USScreens) | 146 | 146,000 |
| | Z-Wert(Marktpotential) | 146 | 146,000 |
| | Z-Wert(CIN_CRITIC) | 146 | 146,000 |
| | Z-Wert(DIS_POWER) | 146 | 146,000 |

*SPSS Bildschirmfoto 4-25 – Gruppenstatistik der Diskriminanzanalyse*

### 4.8.1. Interpretation der Eigenwerte

Zunächst werden die errechneten Eigenwerte des Modells angegeben.

**Eigenwerte**

| Funktion | Eigenwert | % der Varianz | Kumulierte % | Kanonische Korrelation |
|---|---|---|---|---|
| 1 | ,282[a] | 100,0 | 100,0 | ,469 |

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

*SPSS Bildschirmfoto 4-26 - Eigenwerte*

Im Zwei-Gruppen-Fall wird lediglich eine Funktion erzeugt und daher auch nur ein Eigenwert. Für den Mehr-Gruppen-Fall entstehen weitere Funktionen mit verschiedenen Eigenwerten. Die Rangfolge der Beträge der Eigenwerte gibt dabei an, wie wichtig die jeweilige Schätzung ist. SPSS gibt den größten, wichtigsten Eigenwert als erstes an und setzt entsprechend fort.
Neben einem Eigenwert steht der Prozentsatz der Varianz. Dieser Wert gibt an, wie wichtig diese Diskriminanzfunktion im Verhältnis zu weiteren Diskriminanzfunktionen ist. In diesem Fall gibt es nur eine Schätzung, weswegen 100% der Varianz durch die Funktion abgedeckt werden. Ganz rechts in der Tabelle findet man ein weiteres Gütemaß, die kanonische Korrelation, welches die wechselseitige Abhängigkeit von mehrdimensionalen Variablen mit einer abhängigen Varialben abbilden kann. Es prüft in welchem Maß Deckungsgleichheit der Strukturen der Gruppen vorliegt, sozusagen die Korrelation der Variablenblöcke zwischen den Gruppen. Der Wert von 0.469 weist auf eine mittelmäßige Deckungsgleichheit der durch die Diskriminanzfunktion gebildeten Gruppen hin. Im Zwei-Gruppen-Fall gleicht die quadrierte kanonische Korrelation dem Bestimmheitsmaß $r^2$. Wenn ein Zusammenhang zwischen einer oder mehreren unabhängigen Variablen und einer abhängigen Variablen erklärt werden soll, gibt $r^2$ gibt an, wie gut die unabhängigen Variablen geeignet sind, die Varianz der abhängigen zu erklären. Wilks Lambda und die quadrierte kanonische Korrelation sind komplementär zueinander und ergeben immer 1. In diesem Fall ist die quadrierte kanonische Korrelation 0.219, womit die Wilks Lambda Statistik der Funktion von 0.78 hergeleitet werden kann (Abgleich SPSS Bildschirmfoto 4-27 - Wilks Lambda).
Die wichtigste Angabe ist jedoch der Eigenwert. Der Eigenwert von 0.282 wurde in der Schätzung der Koeffizienten als optimales Diskriminanzmaß verwendet. Da er sich aus dem Verhältnis der Varianz in den Gruppen und der Varianz zwischen den Gruppen errechnet, bedeutet der niedrige Betrag eine relativ hohe Varianz innerhalb der Gruppen im Verhältnis zur Varianz zwischen den Gruppen. Das kann zwei Ursachen haben. Entweder es befinden sich in den Gruppen viele Objekte, die keine ähnlichen Merkmalsausprägungen aufweisen und daher stark um den jeweiligen Gruppencentroiden streuen oder die Gruppencentroide unterscheiden sich nicht sehr voneinander, da die Varianz zwischen den Gruppen, sehr niedrig ist. Es lässt sich jedoch nicht eindeutig über den Eigenwert sagen, ob die Funktion eine hohe Güte hat. Ein niedriger Wert deutet jedoch darauf hin, dass Objekte nicht eindeutig einer Gruppe zugewiesen werden können.

### 4.8.2. Interpretation Wilks Lambda

Aufschluss über die Güte der Trennfähigkeit gibt SPSS durch das inverse Gütemaß Wilks Lambda an.

**Wilks' Lambda**

| Test der Funktion(en) | Wilks-Lambda | Chi-Quadrat | df | Signifikanz |
|---|---|---|---|---|
| 1 | ,780 | 35,007 | 6 | ,000 |

*SPSS Bildschirmfoto 4-27 - Wilks Lambda*

Wilks Lambda gibt zwischen 0 und 1 die Güte der Trennfähigkeit der durch die Diskriminanzfunktion erzeugten Gruppen wieder. Dabei ist die Trennfähigkeit auf Grund der Mittelwertunterschiede der Gruppen besser, je näher sich Wilks Lambda dem Wert 0 nähert und umgekehrt. Der Testwert von 0.78 impliziert eine

mittelmäßige bis schwache Trennfähigkeit der Gruppen. SPSS führt außerdem den Chi-Quadrat-Test[79] mit der Nullhypothese durch, dass sich die betrachteten Gruppen nicht unterscheiden beziehungsweise, dass die Mittelwerte der Gruppen gleich sind. Das Testergebnis ist signifikant. Die Nullhypothese muss abgelehnt werden. Die Gruppen unterscheiden sich. Das Modell kann grundsätzlich verwendet werden, um die Gruppen zu diskriminieren.

### 4.8.3. Interpretation der standardisierten Diskriminanzkoeffizienten

Nach Wilks Lambda werden die standardisierten kanonischen Koeffizienten der Diskriminanzfunktion ausgegeben.

**Standardisierte kanonische Diskriminanzfunktionskoeffizienten**

| | Funktion |
|---|---|
| | 1 |
| Z-Wert(ACTORS) | -,472 |
| Z-Wert(CIN_CRITIC) | ,326 |
| Z-Wert(DIS_POWER) | -,385 |
| Z-Wert(DIRECTOR) | ,036 |
| Z-Wert(USScreens) | 1,257 |
| Z-Wert(Marktpotential) | -,027 |

*SPSS Bildschirmfoto 4-28 - Standardisierte kanonische Diskriminanzfunktionskoeffizienten*

Über die standardisierten Diskriminanzkoeffizienten kann die relative Bedeutung der Prädiktoren für die Diskriminanzfunktion ermittelt werden. Die Formel (Formel 3:7 - Berechnung der relativen Bedeutung eines standardisierten Diskriminanzkoeffizienten) hierfür ist in Kapitel 3 zu finden.

| Variable | Berechnung der relativen Bedeutung der mittleren Diskriminanzkoeffizienten | Relative Bedeutung in % |
|---|---|---|
| ACTORS | $\frac{.472}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 17.1 |
| CIN_CRITIC | $\frac{,326}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 11.8 |
| DIS_POWER | $\frac{.385}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 14.0 |
| DIRECTOR | $\frac{.036}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 1.3 |
| USScreens | $\frac{1.257}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 45.7 |
| Marktpotential | $\frac{.27}{.472 + .326 + .385 + .036 + 1.257 + .27}$ | 9.8 |

*Tabelle 4-5 – Berechnung der relativen Bedeutung der standardisierten Diskriminanzkoeffizienten*

### 4.8.4. Interpretation der Struktur-Matrix

Die Struktur-Matrix offenbart die Korrelationen zwischen den Prädiktoren und der Diskriminanzfunktion. Auf diese Weise ist es mehreren Diskriminanzfunktionen möglich, Korrelationen zu vergleichen und zu sehen, wie

[79] vgl. Kapitel 3.2.4.1

eng ein Prädiktor mit einer bestimmten Funktion verbunden ist. Hierdurch lässt sich die jeweilige Diskriminanzfunktion durch bestimmte, wichtige Prädiktoren charakterisieren, interpretieren und benennen. Im Allgemeinen werden Prädiktoren mit einer Korrelation von $> 0.3$ als wichtig betrachtet.

**Struktur-Matrix**

| | Funktion |
|---|---|
| | 1 |
| Z-Wert(USScreens) | ,779 |
| Z-Wert(CIN_CRITIC) | ,182 |
| Z-Wert(DIS_POWER) | ,144 |
| Z-Wert(Marktpotential) | ,115 |
| Z-Wert(ACTORS) | -,045 |
| Z-Wert(DIRECTOR) | -,022 |

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

*SPSS Bildschirmfoto 4-29 – Struktur-Matrix*

Da in diesem Beispiel nur eine Funktion ausgegeben wird, kann auch nur eine Funktion anhand der Prädiktoren charakterisiert werden. Diese Funktion wird durch ihre starke Abhängigkeit von dem Prädiktor „USScreens“ charakterisiert. Diesen Sachverhalt spiegelt auch die relative Bedeutung der Prädiktoren wieder. Die dazugehörige Diskriminanzfunktion könnte demnach einen Namen wie „Leinwandtrennfunktion“ erhalten, um bei künftigen Beschreibungen der verschiedenen Diskriminanzfunktionen eindeutige Bezeichnungen zu verwenden.

### 4.8.5. Interpretation der kanonischen Diskriminanzkoeffizienten und der Funktionen bei Gruppen-Zentroiden

Nach der Struktur-Matrix werden die nicht-standardisierten kanonischen Diskriminanzkoeffizienten ausgegeben.

**Kanonische Diskriminanzfunktionskoeffizienten**

| | Funktion |
|---|---|
| | 1 |
| Z-Wert(ACTORS) | -,466 |
| Z-Wert(CIN_CRITIC) | ,372 |
| Z-Wert(DIS_POWER) | -,356 |
| Z-Wert(DIRECTOR) | ,032 |
| Z-Wert(USScreens) | 1,416 |
| Z-Wert(Marktpotential) | -,028 |
| (Konstant) | ,183 |

Nicht-standardisierte Koeffizienten

*SPSS Bildschirmfoto 4-30 - Kanonische Diskriminanzfunktionskoeffizienten*

Anhand dieser Koeffizienten gestaltet sich die Klassifizierungsfunktion (lineare Diskriminanzfunktion):

$$Diskr.wert = 0.183 - 0.466 \cdot ACTORS + 0.372 \cdot CIN\ \ CRITIC - 0.356 \cdot DIS\ \ POWER + 0.032 \cdot DIRECTOR + 1.416 \cdot USScreens - \ 0.028 \cdot Marktpotential$$

Anhand dieser Funktion wird jedes Objekt den Gruppen zugeordnet. Die Klassifizierungsregel im zwei Gruppenfall ergibt sich aus dem Mittelwert der Gruppencentroiden. Dieser kann durch die nächste Tabelle, „Funktion bei den Gruppen-Zentroiden“ errechnet werden.

**Funktionen bei den Gruppen-Zentroiden**

| | Funktion |
|---|---|
| Rentabilität_nach_US | 1 |
| 1,00 | ,649 |
| 2,00 | -,428 |

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

*SPSS Bildschirmfoto 4-31 - Funktionen bei Gruppen-Zentroiden*

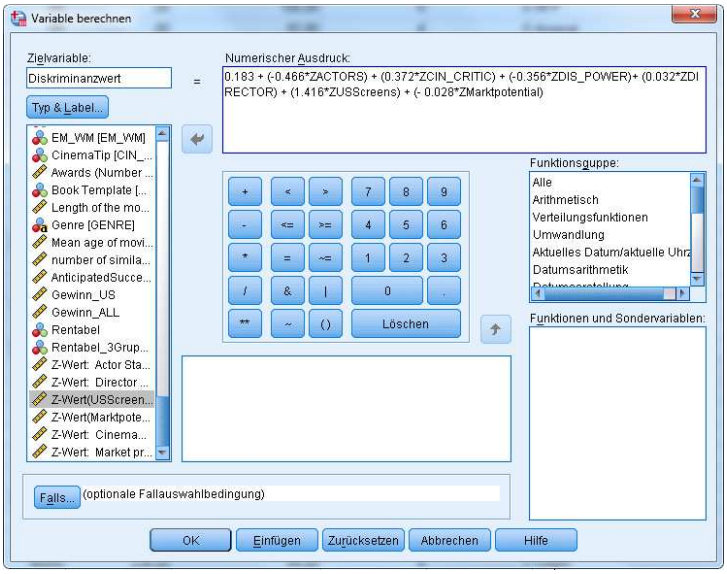
Grenzwert der Distanzentscheidungsregel im $2-$ Gruppen $-$ Fall:

$$\text{Mittelwert} = \frac{0.649-0.428}{2} = 0.1105$$

Diskriminanzfunktionswerte für Objekte, die über 0.1105 liegen, werden, demnach der Gruppe 1: „rentable“ zugeordnet und Werte, die unterhalb von 0.1105 liegen, dementsprechend der Gruppe 2: „nicht rentabel“. Durch diese Zuordnung werden die Prädiktoren und ihre Gewichtung interpretierbar. Laut der Diskriminanzfunktion beeinflussen eine starke Besetzung sowie ein größerer Marktanteil des Distributors den Filmerfolg negativ. Diese Interpretation erscheint auf den ersten Blick nicht sehr logisch. Gemeinsam haben diese beiden Prädiktoren immerhin eine relative Bedeutung von 26.9%. Der Rest des Modells deckt sich mit anfänglich aufgestellten Annahmen. Das Modell gibt außerdem einen negativen Einfluss eines schrumpfenden Marktpotentials wieder. Dem Modell zufolge scheinen Filme mit geringer Altersklassifizierung weniger Erfolg zu versprechen als Filme mit vorhandener Altersklassifizierung. Die Variable „USScreens“ hat die höchste diskriminatorische Bedeutung, sowohl betragsmäßig als auch relativ mit 45.7%. Je mehr Vorstellungen eines Films innerhalb von 20 Wochen vorgeführt werden, desto eher wird er der Gruppe „erfolgreich“ zugewiesen. Welcher Direktor für den Film verantwortlich ist, hat in diesem Modell fast gar keine Bedeutung. Um das Modell zu optimieren, könnte man Prädiktoren mit einer solch geringen relativen Bedeutung aus dem Modell ausschließen. Das lässt die Gesamtstreuung der abhängigen Variablen sinken, Redundanzen verringern sich und Effekte der übrigen Prädiktoren können eindeutiger interpretiert werden.

## 4.9. Anwendung des Diskriminanzmodells mit zwei Gruppen

Um nun die Güte der Klassifizierung des Modells zu testen, ruft man die zweite Stichprobe auf. Da die Diskriminanzfunktion anhand standardisierter Variablen geschätzt wurde, müssen die verwendeten Prädiktoren erneut standardisiert werden, bevor aus ihnen die Diskriminanzwerte der Objekte berechnet werden können. Ist dies, geschehen ruft man unter „Transformieren“ die Funktion „Variable berechnen“ auf. Hier trägt man die gesamte Diskriminanzfunktion ein, also die Konstante, die geschätzten Koeffizienten und den dazugehörigen Prädiktor. Als Zielvariable erhält man dann die Diskriminanzwerte der jeweiligen Objekte.

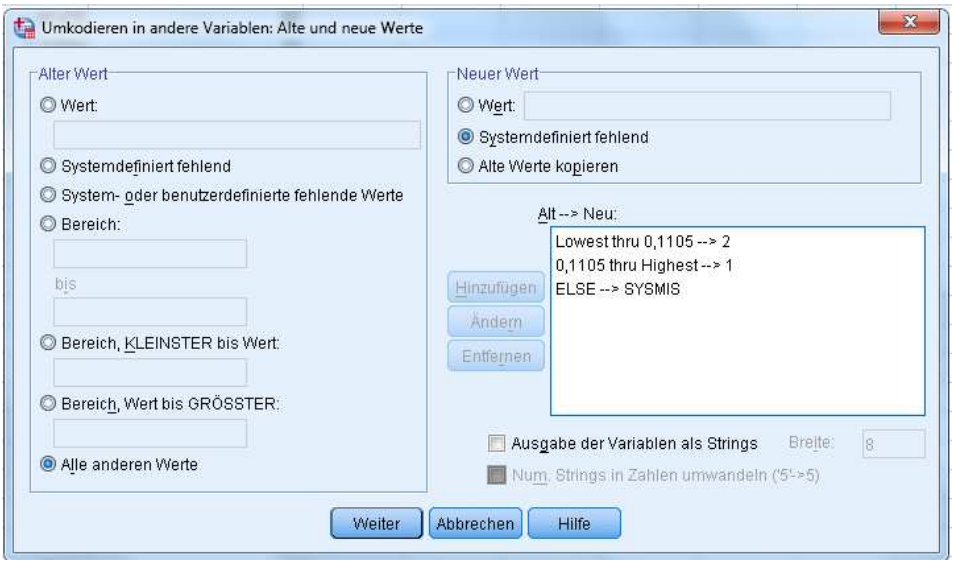



*SPSS Bildschirmfoto 4-32 – Diskriminanzwert über „Variable berechnen“ für neue Fälle berechnen*

Das Zuordnungsprozedere anhand einer Klassifizierungsfunktion beim Zwei-Gruppen-Fall besagt, dass das Objekt anhand seines Diskriminanzfunktionswerts, der entweder oberhalb oder unterhalb des Mittelwerts der Gruppenmittelwerte liegt, der entsprechenden Gruppe zugeordnet wird. Um Objekte nun den Gruppen zuzuordnen, kann im nächsten Schritt über „Transformieren“ und „Umkodieren in andere Variable“ eine Klassifikation anhand von entsprechenden Regeln erfolgen. Hierfür wählt man als numerische Variable die zuvor errechnete Variable „Diskriminanzwert“, trägt einen neuen Namen für die Ausgabevariable ein, in diesem Fall „vorhergesagte_Gruppe“, und bestätigt mit „Ändern“. Als nächstes bestimmt man die Regeln, die für die Umkodierung gelten sollen über „Alte und neue Werte“.


*SPSS Bildschirmfoto 4-33 – Gruppenzugehörigkeit im Zwei-Gruppen-Fall über Umkodieren festlegen*

Um aus Diskriminanzwerten Gruppenzugehörigkeiten zu machen, bestimmt man, dass die kleinsten Werte bis zum Mittelwert der „nicht erfolgreich“ Gruppe, also Gruppe 2 zugeordnet werden und die größten Werte bis zum Mittelwert der Gruppe „erfolgreich“, also Gruppe 1. Ein Objekt, dessen Diskriminanzwert exakt auf dem Mittelwert liegt, kann nicht eindeutig klassifiziert werden. Indem definiert wird, dass alle anderen Fälle die Ausprägung „Systemdefiniert fehlend“ erhalten, werden diese Sonderfälle aus der Betrachtung eliminiert.




*SPSS Bildschirmfoto 4-34 – Alte und Neue Werte zur Bestimmung der Gruppenzugehörigkeit*

SPSS berechnet nun eine neue Variable mit den entsprechenden prognostizierten Gruppenzugehörigkeiten.

### 4.9.1. Interpretation der Ergebnisse und Bewertung der Klassifikationsgüte

Über eine Kreuztabelle, die unter „Analysieren“ und „Deskriptive Statistik“ zu finden ist, lassen sich als Spalten und Zeilen die bekannte Gruppenzugehörigkeit und die vorhergesagte Gruppenzugehörigkeit abbilden.


*SPSS Bildschirmfoto 4-35 – Kreuztabelle als Klassifikationsmatrix*

Als Output erhält man eine Übersicht der verarbeiteten Fälle und die Klassifikationsmatrix in Form einer Kreuztabelle. Hieraus lässt sich die Güte der Trennfähigkeit der Diskriminanzfunktion ableiten.

**Verarbeitete Fälle**

| | Fälle | | | | | |
|---|---|---|---|---|---|---|
| | Gültig | | Fehlend | | Gesamt | |
| | N | Prozent | N | Prozent | N | Prozent |
| Rentabel * vorhergesagte_Gruppe | 737 | 100,0% | 0 | ,0% | 737 | 100,0% |

**Rentabel * vorhergesagte_Gruppe Kreuztabelle**

Anzahl

| | | vorhergesagte_Gruppe | | Gesamt |
|---|---|---|---|---|
| | | 1,00 | 2,00 | |
| Rentabel | 1,00 | 196 | 25 | 221 |
| | 2,00 | 208 | 308 | 516 |
| Gesamt | | 404 | 333 | 737 |

*SPSS Bildschirmfoto 4-36 – Klassifikationsmatrix*

Die Zeilen geben die tatsächlichen und die Spalten die prognostizierten Gruppenzugehörigkeiten wieder. In der Hauptdiagonalen stehen die Fallzahlen, die korrekt klassifiziert wurden. Dieser Sachverhalt lässt sich auch auf den Mehr-Gruppen-Fall anwenden. Insgesamt klassifiziert das Modell 504 von 737 Gruppenzugehörigkeiten korrekt, also 68.38%. Die übrigen Felder geben die verkehrten Klassifizierungen wieder und ergeben ein Fehlklassifikationsrisiko von 31.62%. Um die Klassifikationsfähigkeit der Diskriminanzfunktion anhand der Klassifizierungsmatrix richtig beurteilen zu können, müssen die Trefferquoten mit denjenigen verglichen werde, die durch eine rein zufällige Zuordnung der Objekte eingetroffen wäre, also mit der a-priori-Wahrscheinlichkeit,

die sich aus der Größe der Gruppen ableiten lässt[80]. Die a-priori-Wahrscheinlichkeiten sind 70%, dass ein Film nicht rentabel ist und 30%, dass ein Film rentabel ist (vgl. 4.3). In 88.7% der Fälle schafft es das Modell einen tatsächlich rentablen Film auch als rentabel zu klassifizieren und irrt sich hinsichtlich der Klassifikation eines nicht rentablen Films der eigentlich rentabel ist nur in 11.3% der Fälle. Rentable Filme scheinen korrekt vom Modell auch als solche identifiziert zu werden. Allerdings klassifiziert das Modell grundsätzlich deutlich mehr Filme in die Kategorie „rentabel". 208 Objekte der insgesamt 737 Objekten werden dieser Kategorie fälschlicherweise zugeordnet. Von allen als „rentabel" vorhergesagten Filmen sind tatsächlich nur 48.5% wirklich rentabel. Das ist um 18.5 Prozentpunkte höher, als die a-priori-Wahrscheinlichkeit. Das Risiko sich anhand des Modells zu hinsichtlich der Rentabilität irren, ist jedoch mit 51.5% enorm. Das Modell kann nicht gut zwischen nicht rentablen Filmen und rentablen Filmen unterscheiden. Der generelle Diskriminanzerfolg des Modells unterscheidet sich nicht außerordentlich von den a-priori-Wahrscheinlichkeiten. Somit ist es nicht geeignet, Gruppenunterschiede zu erklären und den Erfolg von Filmen, und zwar besser als der Zufall es könnte, zu prognostizieren.

*ANMERKUNG: Auf Grund eines Speicherfehlers während des Verfassens der Thesis, sind Teststichprobe und Trainingsstichprobe nicht ordnungsgemäß erhalten geblieben. Daher wurde das Modell an der Gesamtpopulation und nicht an der Trainingsstichprobe getestet. Es kann davon ausgegangen werden, dass die Ergebnisse der Klassifikationsgüte ähnlich und sogar etwas schlechter ausfallen wären, denn das Modell wurde anhand von 30% von 70% des gesamten Datenumfangs optimiert und nun an 100% getestet. Dies hat zur Folge, dass das Modell für 30% der vorliegenden Daten „optimal" ist, was die Ergebnisse etwas verzerrt.*

## 4.10. Der Mehr-Gruppen-Fall anhand der schrittweisen Diskriminanzanalyse

Grundsätzlich spielt sich der Mehr-Gruppen-Fall sehr ähnlich ab, wie der Zwei-Gruppen-Fall. Unterschiedlich ist, dass die Gruppenzuordnung nicht mehr über die einfache Distanz zu den Gruppen-Zentroiden berechnet werden kann, sondern über eine komplexere multivariate Methode, da nun die Distanzen eines Objekts zu mehreren Gruppen-Zentroiden gleichzeitig bewertet werden müssen. Hierfür kann beispielsweise das euklidische oder Mahalanobis Proximitätsmaß genutzt werden. Dieses Distanzkonzept ist um Wahrscheinlichkeiten erweiterbar. Die Berechnung hierfür ist sehr umfangreich und komplex - jedoch standardmäßig in SPSS eingebettet. Auf S. 246 (Backhaus 2016) kann hierzu mehr gefunden werden. Zusätzlich zum Mehr-Gruppen-Fall soll nun außerdem die schrittweise Prozedur der Diskriminanzanalyse mit SPSS dargestellt werden.

### 4.10.1. Festlegung der dritten Gruppe

Wie in Kapitel 4.3 beschrieben, kann aus dem Datensatz eine dritte Gruppe, die der „vermutlich in Zukunft rentabel werdenden" Filme, gebildet werden. Hierfür wird eine neue Gruppierungsvariable berechnet, die dementsprechend 3 Gruppen anhand der Grenzwerte aus Kapitel 4.3 erzeugt. Über eine weitere Lernstichprobe wird ein neues Klassifizierungsmodell geschätzt, nun mit drei Gruppen und zwei Diskriminanzfunktionen.

### 4.10.2. Exemplarische Durchführung der schrittweisen Diskriminanzanalyse für drei Gruppen

Bei der schrittweisen Methode schließt SPSS automatisch Variablen aus, die nicht signifikant zur Erklärung der Gruppenunterschiede beitragen. Das ist etwas bequemer, als selbst die relativen Bedeutungen der jeweiligen

[80] Backhaus 2016, S. 239

Prädiktoren zu errechnen und zu bewerten. Hierfür bezieht SPSS schrittweise denjenigen Prädiktor in das Modell ein, der ein vorher definiertes Diskriminanzmaß[81] maximiert, also einen hohen Anteil an erklärter Streuung gegenüber nicht erklärter Streuung zur Gesamtvarianz des Modells beiträgt. Eine Kritik an dieser Vorgehensweise ist, dass die Optimierung des Modells zu sehr in den Vordergrund rückt, anstelle der sachlogischen Idee, die hinter der Analyse stehen sollte. Beachtet man jedoch, dass ein theoretisches Konzept der Auswahl der Prädiktoren zugrunde liegen sollte, kann man diese recht bequeme Methode durchaus als Alternative, auch für den Zwei-Gruppen-Fall, nutzen.

### 4.10.3. Voraussetzungsannahmen im Mehr-Gruppen-Fall

Voraussetzungen bleiben dieselben, wie beim Zwei-Gruppen-Fall. Multivariate Normalverteilung der Prädiktoren, Homogenität der Kovarianz-Varianz-Matrizen, Gruppenmittelwertunterschiede anhand der erklärenden Variablen und geringe Korrelationen der erklärenden Variablen, um die Effekte der erklärenden Variablen auf die Gruppenmittelwertunterschiede eindeutiger erklären zu können. Die exemplarische Diskriminanzanalyse am Zwei-Gruppen-Fall wurde sehr *nach Lehrbuch* vorgeführt. Man sollte darauf achten, dass keine Voraussetzungsannahmen verletzt werden, jedoch ist das Verfahren recht robust. Deshalb werden in diesem Modell die Variablen „USUmsatz“ und „Budget“ nicht ausgeschlossen. Vorab werden die Variablen über die z-Transformation standardisiert. Dies stellt sicher, dass etwa gleiche Kovarianz-Varianz-Matrizen vorliegen. Die multivariate Normalverteilungsvoraussetzung wird verletzt, dieser Sachverhalt jedoch erneut ignoriert.

| Variablenname | Funktion | Bedeutung |
|---|---|---|
| Rentabilität_3 | Abhängige Variable | Erzielt der Film Wochen Gewinne (nach 20 Wochen = 1, mit Risiko später = 2, nein = 3) |
| „DIS_POWER | Prädiktor | Marktanteil des Filmdistributors (in %) |
| CIN_CRITIC | Prädiktor | Bewertung des Films durch Kritiken von (0 bis 5) |
| Marktpotential | Prädiktor | Einschränkung des Marktes durch Altersbeschränkungen (in %) |
| ACTORS | Prädiktor | Bewertung der Besetzung (von 0 bis 5) |
| DIRECTOR | Prädiktor | Bewertung des Regisseurs (von 0 bis 5) |
| BUDGET | Prädiktor | Produktionsbudget (in US$) |
| USUmsatz | Prädiktor | Umsatz der Kinovorstellungen nach 20 Wochen (in US$) |
| USScreens | Prädiktor | Anzahl der Kinovorstellungen nach 20 Wochen |

*Tabelle 6 - Variablenauswahl für den Mehr-Gruppen-Fall*

### 4.10.4. Interpretation und Auswertung des Outputs

[81] [81] in SPSS gibt es weitere Maße, hinsichtlich derer optimiert werden kann. Bei allen wird das Maß durch das Hinzunehmen weiterer erklärenden Variablen minimiert oder optimiert. Wilks Lambda: Testwert minimieren; Unerklärte Varianz: Anteil minimieren; Mahalanobis Distanz: Distanz minimieren; Kleinster F-Wert: Maximierung der F-Verteilung anhand der Mahalanobisdistanz zwischen den Objekten einer Gruppe; Raos‘ V: Messwert für Gruppenmittelwertunterschiede, welches maximiert wird.

**Analyse der verarbeiteten Fälle.**

| Ungewichtete Fälle | | N | Prozent |
|---|---|---|---|
| Gültig | | 227 | 58,8 |
| Ausgeschlossen | Gruppencodes fehlend oder außerhalb des Bereichs | 0 | ,0 |
| | Mindestens eine fehlende Diskriminanz-Variable | 0 | ,0 |
| | Beide fehlenden oder außerhalb des Bereichs liegenden Gruppencodes und mindestens eine fehlende Diskriminanz-Variable | 159 | 41,2 |
| | Gesamtzahl der ausgeschlossenen | 159 | 41,2 |
| Gesamtzahl der Fälle | | 386 | 100,0 |

*SPSS Bildschirmfoto 37 – Analyse der verarbeiteten Fälle im Mehr-Gruppen-Fall*

Da die Variable „Budget“ sehr viele Missings aufweist, fehlen 41.2% der Stichprobe, um das Modell zu schätzen. 227 Objekte und 8 Prädiktoren erlauben nichtsdestotrotz eine Schätzung ausreichender Güte, um sie an einer größeren Stichprobe zu testen.

**Gleichheitstest der Gruppenmittelwerte**

| | Wilks-Lambda | F | df1 | df2 | Signifikanz |
|---|---|---|---|---|---|
| Z-Wert(USScreens) | ,490 | 116,466 | 2 | 224 | ,000 |
| Z-Wert(Marktpotential) | ,994 | ,692 | 2 | 224 | ,502 |
| Z-Wert: CinemaCritic | ,994 | ,645 | 2 | 224 | ,526 |
| Z-Wert: Market proportion | ,869 | 16,933 | 2 | 224 | ,000 |
| Z-Wert(Budget) | ,919 | 9,829 | 2 | 224 | ,000 |
| Z-Wert(USUmsatz) | ,628 | 66,251 | 2 | 224 | ,000 |
| Z-Wert: Actor Star Power | ,879 | 15,437 | 2 | 224 | ,000 |
| Z-Wert: Director Star Power | ,976 | 2,733 | 2 | 224 | ,067 |

*SPSS Bildschirmfoto 38 - Gleichheitstest der Gruppenmittelwert im Mehr-Gruppen-Fall*

Laut Gleichheitstest der Gruppenmittelwerte, diesmal nach Wilks Lambda, kann für 5 von 8 Variablen die Nullhypothese der Gleichheit der Gruppenmittelwerte abgelehnt werden. Andersherum weisen alle Variablen, außer „CIN_CRITIC“, „Marktpotential“ und „DIRECTOR“, mindestens in einer Gruppe Mittelwertunterschiede gegenüber den anderen Gruppen auf. Die drei Variablen, die keine Informationen zu Gruppenmittelwertunterschieden liefern, können aus der linearen Kombination ausgeschlossen werden. Die größten Gruppenmittelwertunterschiede hat die Variable „USScreens“ mit einem Wilks Lambda von 0.49.

**Gemeinsam Matrizen innerhalb der Gruppen**

| | | Z-Wert (USScreens) | Z-Wert (Budget) | Z-Wert (USUmsatz) | Z-Wert: Actor Star Power | Z-Wert: Director Star Power | Z-Wert: Market proportion |
|---|---|---|---|---|---|---|---|
| Korrelation | Z-Wert(USScreens) | 1,000 | ,794 | ,799 | ,463 | ,209 | ,405 |
| | Z-Wert(Budget) | ,794 | 1,000 | ,781 | ,399 | ,260 | ,328 |
| | Z-Wert(USUmsatz) | ,799 | ,781 | 1,000 | ,408 | ,372 | ,298 |
| | Z-Wert: Actor Star Power | ,463 | ,399 | ,408 | 1,000 | ,339 | ,351 |
| | Z-Wert: Director Star Power | ,209 | ,260 | ,372 | ,339 | 1,000 | ,155 |
| | Z-Wert: Market proportion | ,405 | ,328 | ,298 | ,351 | ,155 | 1,000 |

*SPSS Bildschirmfoto 39 - Korrelationsübersicht der Prädiktoren im Mehr-Gruppen-Fall*

Als nächstes folgt im Output die Korrelationsmatrix. Hier sind die Wechselwirkungen bzw. Beziehungen der einzelnen Prädiktoren miteinander abgebildet. Eine extrem hohe Korrelation liegt beispielsweise zwischen Budget und USScreens (0.794), Budget und USUmsatz (0.781) sowie USScreens und USUmsatz vor. Die Variablen

scheinen recht Deckungsgleich zu sein. Falls jede von ihnen in das Modell aufgenommen wird, wirkt ihre Gewichtung übermäßig. Niedrige bis mittlere Korrelationen sind vertretbar.

**Log-Determinanten**

| Rentabel_3 | Rang | Log-Determinante |
|---|---|---|
| 1,00 | 4 | -2,286 |
| 2,00 | 4 | -4,276 |
| 3,00 | 4 | -8,053 |
| Gemeinsam innerhalb der Gruppen | 4 | -3,271 |

Die Ränge und natürlichen Logarithmen der ausgegebenen Determinanten sind die der Gruppen-Kovarianz-Matrizen.

**Textergebnisse**

| | | |
|---|---|---|
| Box-M | | 544,582 |
| F | Näherungswert | 26,436 |
| | df1 | 20 |
| | df2 | 98049,271 |
| | Signifikanz | ,000 |

Testet die Null-Hypothese der Kovarianz-Matrizen gleicher Grundgesamtheit.

*SPSS Bildschirmfoto 40 - Box M Test im Mehr-Gruppen-Fall*

Der Box M Test ist signifikant. Der p-Wert liegt bei 0.000. Die Nullhypothese der Homogenität der Varianzen muss abgelehnt werden. Das liegt sehr wahrscheinlich einerseits an dem geringen Stichprobenumfang und der geringen Anzahl an Fällen in den jeweiligen Gruppen als auch an der Verletzung der Normalverteilungsvoraussetzung für diesen Test. Der Sachverhalt wird jedoch ignoriert und mit vorhandenen Gruppen und Prädiktoren weitergemacht.

**Aufgenommene/Entfernte Variablen[a,b,c,d]**

| Schritt | Aufgenommen | Wilks-Lambda | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Exaktes F | | | |
| | | Statistik | df1 | df2 | df3 | Statistik | df1 | df2 | Signifikanz |
| 1 | Z-Wert (USScreens) | ,490 | 1 | 2 | 224,000 | 116,466 | 2 | 224,000 | ,000 |
| 2 | Z-Wert (Budget) | ,321 | 2 | 2 | 224,000 | 85,269 | 4 | 446,000 | ,000 |
| 3 | Z-Wert (USUmsatz) | ,270 | 3 | 2 | 224,000 | 68,358 | 6 | 444,000 | ,000 |
| 4 | Z-Wert: Actor Star Power: 0= no star power, 5=max star power (sum of 12 cinema visitors and two industry experts inclusive confidence based weighted mean of van Bruggen et al. 2002) | ,257 | 4 | 2 | 224,000 | 53,699 | 8 | 442,000 | ,000 |

Bei jedem Schritt wird die Variable aufgenommen, die das gesamte Wilks-Lambda minimiert.

a. Maximale Anzahl der Schritte ist 12.
b. Maximale Signifikanz des F-Werts für die Aufnahme ist .05.
c. Minimale Signifikanz des F-Werts für den Ausschluß ist .10
d. F-Niveau, Toleranz oder VIN sind für eine weitere Berechnung unzureichend.

*SPSS Bildschirmfoto 41 - Schrittweise Methode im Mehr-Gruppen-Fall*

Bei der nächsten Tabelle „Aufgenommene/Entfernte Variablen“ unterscheidet sich die schrittweise Diskriminanzanalyse zum ersten Mal. Anstelle davon, dass die Schätzung simultan für alle Variablen gleichzeitig geschieht, sucht das Verfahren über ein ausgewähltes Maß, in diesem Fall Wilks Lambda bzw. der F-Wert, die wichtigsten Prädiktoren und schließt sie in das Modell ein. Im ersten Schritt wird die Diskriminanzfunktion mit einer Variablen geschätzt, dann mit zwei und so weiter. Die wichtigste, am meisten erklärende Variable mit dem größten F-Wert (116.466) ist „USScreens“ . Die zweite in die Schätzung aufgenommene Variable ist „Budget“. Im ersten Schritt ist das Wilks Lambda der Funktion durch einen Prädiktor 0.490. Das Gütemaß ist inverse, was bedeutet, dass es optimiert wird, je kleiner es wird. Im zweiten Schritt verringert sich der Wert durch eine

weitere erklärende Variable um 0.169. Das Gütemaß wird besser. Auch durch das hinzunehmen des dritten Prädiktors „USUmsatz“ sinkt Wilks Lambda. Diesmal um 0.101. Schließlich erzielt ein letzter Prädiktor noch eine minimale Verbesserung von Wilks Lambda um 0.013. Jeder Prädiktor hat dabei signifikanten Einfluss auf die Erklärung der Gruppenmittelwertunterschiede (p-Wert = 0.000). Damit steht das finale Modell. Die restlichen Prädiktoren wurden nicht mit in das Modell integriert, weil sie einen F-Grenzwert unterschritten haben, anhand dessen festgestellt wird, ob das Maß durch den Prädiktor weiter hinreichend optimiert wird. Der F-Grenzwert kann manuell angepasst werden, um den Prädiktor Ausschluss und Einschluss zu beeinflussen.

Aus den Prädiktoren werden zwei Funktionen geschätzt. Funktion 1 und Funktion 2.

**Eigenwerte**

| Funktion | Eigenwert | % der Varianz | Kumulierte % | Kanonische Korrelation |
|---|---|---|---|---|
| 1 | 2,158[a] | 90,3 | 90,3 | ,827 |
| 2 | ,231[a] | 9,7 | 100,0 | ,434 |

a. Die ersten 2 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

**Wilks' Lambda**

| Test der Funktion(en) | Wilks-Lambda | Chi-Quadrat | df | Signifikanz |
|---|---|---|---|---|
| 1 bis 2 | ,257 | 302,159 | 8 | ,000 |
| 2 | ,812 | 46,323 | 3 | ,000 |

*SPSS Bildschirmfoto 42 - Eigenwerte und Wilks Lambda im Mehr-Gruppen-Fall*

Die erste Funktion hat einen vergleichsweise hohen Eigenwert. Dabei kann 90.3% der Varianz bereits durch die Funktion 1 abgedeckt werden. Die erste Funktion hat also eine deutlich höhere Fähigkeit die abhängige Variable zu erklären, als die zweite. Funktion 2 wirkt eher ergänzend. Mit einem deutlich geringeren Eigenwert von 0.231 ist sie weniger wichtig und deckt auch nur für 9.7% der Varianz ab. Als nächstes kann Wilks Lambda interpretiert werden. Durch die erste Funktion werden die Gruppenmittelwertunterschiede am besten erklärt. Sie trennt die Gruppen sehr gut und signifikant voneinander. Die zweite Funktion ist auch signifikant und trägt ebenfalls, wenn auch nur sehr geringfügig, zur Unterscheidung von Gruppenmittelwertunterschieden bei.

**Standardisierte kanonische Diskriminanzfunktionskoeffizienten**

| | Funktion | |
|---|---|---|
| | 1 | 2 |
| Z-Wert(USScreens) | 1,376 | ,818 |
| Z-Wert(Budget) | -1,294 | ,713 |
| Z-Wert(USUmsatz) | ,500 | -1,376 |
| Z-Wert: Actor Star Power: 0= no star power, 5=max star power (sum of 12 cinema visitors and two industry experts inclusive confidence based weighted mean of van Bruggen et al. 2002) | -,161 | ,485 |

*SPSS Bildschirmfoto 43 - standardisierte kanonische Diskriminanzkoeffizienten im Mehr-Gruppen-Fall*

Anhand der standardisierten kanonischen Diskriminanzkoeffizienten kann abgelesen werden, wie hoch die Bedeutung der einzelnen Prädiktoren der jeweiligen Diskriminanzfunktion des gesamten Modells ist. (Formel 4:2)

Relative Gewichtung der Koeffizienten in %

| Variable | Funktion 1 | Funktion 2 |
|---|---|---|
| USScreens | 37.3 | 2.3 |
| Budget | 35.1 | 2.0 |
| Umsatz | 13.6 | 4.0 |
| Actors | 4.3 | 1.4 |
| Gesamt | 90.3 | 9.7 |

*Tabelle 7 – relative Gewichtung der Koeffizienten im Mehr-Gruppen-Fall*

Durch die kanonischen Diskriminanzkoeffizienten erhält man die eigentlichen Diskriminanzfunktionen des Modells:

**Kanonische Diskriminanzfunktionskoeffizienten**

| | Funktion | |
|---|---|---|
| | 1 | 2 |
| Z-Wert(USScreens) | 1,871 | 1,112 |
| Z-Wert(Budget) | -1,343 | ,740 |
| Z-Wert(USUmsatz) | ,524 | -1,442 |
| Z-Wert: Actor Star Power: 0= no star power, 5=max star power (sum of 12 cinema visitors and two industry experts inclusive confidence based weighted mean of van Bruggen et al. 2002) | -,165 | ,497 |
| (Konstant) | -,792 | -,176 |

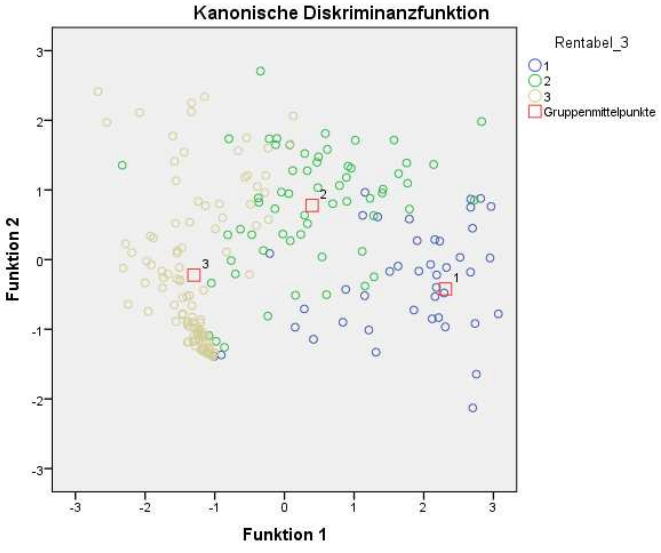
Nicht-standardisierte Koeffizienten

*SPSS Bildschirmfoto 44 - Kanonische Diskriminanzkoeffizienten des Mehr-Gruppen-Falls*

Funktion 1: $-0.792 + 1.871 \cdot USScreens - 1.343 \cdot Budget + 0.524 \cdot USUmsatz - 0.165 \cdot ACTORS$

Funktion 2: $-0.176 + 1.112 \cdot USScreens + 0.740 \cdot Budget - 1.442 \cdot USUmsatz + 0.497 \cdot ACTORS$

Diese können als Achsen orthogonal aufgespannt werden, um eine Diskriminanzfläche zu erzeugen. Die Diskriminanzwerte der zwei Diskriminanzfunktionen der Objekte sind ihre jeweiligen Positionen bzw. Koordinaten innerhalb dieser Fläche.




*SPSS Bildschirmfoto 45 - Kombiniertes Diagramm der Diskriminanzgruppenwerte im 3 Gruppen-Fall*

Bei dieser Darstellung ist gut zu sehen, dass sich die Gruppen-Zentroide deutlich voneinander unterscheiden. Die zugewiesenen Objekte überschneiden sich wenig. Gruppe 3 hat eine größere Ansammlung von Objekten bei (-1|-1.3). Ansonsten haben die Gruppen alle eine recht hohe Streuung. Die Streuung ist teilweise so hoch, dass Elemente der Gruppen zwischen den Elementen anderer Gruppen zu finden sind. Diese Fälle werden wahrscheinlicher Fehlklassifiziert, je weiter sie sich vom eigenen Gruppenmittelpunkt entfernen.

**Klassifizierungsergebnisse[a]**

| | | | Vorhergesagte Gruppenzugehörigkeit | | | |
|---|---|---|---|---|---|---|
| | | Rentabel_3 | 1,00 | 2,00 | 3,00 | Gesamt |
| Original | Anzahl | 1,00 | 43 | 7 | 3 | 53 |
| | | 2,00 | 6 | 45 | 10 | 61 |
| | | 3,00 | 0 | 16 | 97 | 113 |
| | % | 1,00 | 81,1 | 13,2 | 5,7 | 100,0 |
| | | 2,00 | 9,8 | 73,8 | 16,4 | 100,0 |
| | | 3,00 | ,0 | 14,2 | 85,8 | 100,0 |

a. 81,5% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

*SPSS Bildschirmfoto 46 - Klassifikationsmatrix im 3 Gruppen-Fall*

Über die Klassifikationsmatrix erhält man Einblick in die Güte des Modells. Insgesamt hat das Modell 81.5% der Fälle richtig klassifiziert. Das ist beachtlich. Die a-priori Wahrscheinlichkeiten der Gruppen wurden diesmal nicht aus der Gruppengröße berechnet, sondern für alle Gruppen gleich auf 33.33% bestimmt. Jeder Fall hat also die gleiche Wahrscheinlichkeit rein zufällig einer Klasse zugeordnet zu werden. In 81% der Fälle ist der frühe Erfolg der Filme richtig klassifiziert. Das stellt eine Verbesserung der a-prior-Wahrscheinlichkeit von 47.67% dar. Filme, die eventuell rentabel werden, also die der zweiten Gruppe, trennt das Modell auch sehr gut.73.8% werden vom Modell der richtigen Gruppe zugewiesen. Am besten jedoch erkennt das Modell nicht rentable Filme. Sie können mit einer Wahrscheinlichkeit von 85.8% richtig vorhergesagt werden. Das Modell hat statistisch gesehen anwendungspotential. Jedoch möglicherweise praktisch keine Bedeutung. Die Variablen „USUmsatz“ und „USBudget“ waren für die anfängliche Gruppierung in „erfolgreich“, „erfolgreich mit Risiko“ und „nicht erfolgreich“ die essentiellen Faktoren. Es ist dementsprechend nicht verwunderlich, wenn diese Variablen die Mittelwertunterschiede gut erklären können, weil sie die Unterschiede im Kern selbst überhaupt erst hergeleitet haben. Des Weiteren sind die Variablen, die zur Erklärung der Mittelwertunterschiede verwendet wurden, sind stark miteinander korreliert. Die einzige Variable, die in diesem Modell auch von praktischer Bedeutung ist, ist „ACTORS“ und trägt am wenigsten zur Erklärung der Gruppenmittelwertunterschiede bei. Das Modell hat keine praktische Bedeutsamkeit.

### 4.10.5. Klassifikation neuer Elemente im 3-Gruppen-Fall mit SPSS

Die einfachste Art und Weise, das Modell an einer anderen Stichprobe anzuwenden, ist anhand der Klassifizierungsfunktionskoeffizienten nach Fisher. Diese kann man in der „Statistiken“-Oberfläche in SPSS auswählen. Hierdurch erscheint im Output von SPSS eine Tabelle, die individuelle Koeffizienten für jede Gruppe ausgibt.

**Klassifizierungsfunktionskoeffizienten**

| | Rentabel_3 | | |
|---|---|---|---|
| | 1,00 | 2,00 | 3,00 |
| Z-Wert(Budget) | -4,673 | -1,215 | ,323 |
| Z-Wert(USUmsatz) | 2,065 | -,665 | -,112 |
| Z-Wert: Actor Star Power: 0= no star power, 5=max star power (sum of 12 cinema visitors and two industry experts inclusive confidence based weighted mean of van Bruggen et al. 2002) | -,300 | ,611 | ,393 |
| Z-Wert(USScreens) | 5,419 | 3,172 | -1,115 |
| (Konstant) | -6,005 | -2,323 | -1,289 |

Lineare Diskriminanzfunktionen nach Fisher

*SPSS Bildschirmfoto 47 - Lineare Diskriminanzfunktion nach Fisher*

Unter Voraussetzung gleicher a-priori-Wahrscheinlichkeiten[82], lässt sich über diese Koeffizienten für jede Gruppe eine Klassifikationsfunktion formulieren:

Gruppe 1 – rentabel nach US Markt:

$$F1 = -6.005 + (-4.673 \cdot ZBudget) + (2.065 \cdot ZUSUmsatz) + (-0.300 \cdot ZACTORS)$$

Gruppe 2 – rentabel mit Risiko:

$$F2 = -6.005 + (-1.215 \cdot ZBudget) + (-0.665 \cdot ZUSUmsatz) + (0.611 \cdot ZACTORS)$$

Gruppe 3 - nicht rentabel:

$$F3 = -6.005 + (0.323 \cdot ZBudget) + (-0.112 \cdot ZUSUmsatz) + (-0.300 \cdot ZACTORS)$$

Über diese Funktionen können durch die SPSS Funktion „Variable berechnen“ drei neue Variablen erstellt werden. Im nächsten Schritt müssen diese miteinander verglichen werden. Dabei ist der höchste Funktionswert $F1, F2$ oder $F3$ des Objekts gleichzeitig die zugewiesene Gruppenzugehörigkeit des Objekts. Eine entsprechende Gruppierungsvariable kann ebenfalls über „Variable berechnen“ ausgerechnet werden.

# 5. Zusammenfassung der Ergebnisse der exemplarischen Diskriminanzanalysen

Die Datenqualität für eine Durchführung der linearen Diskriminanzanalyse ist als „mittelmäßig geeignet“ zu bewerten. Obwohl eine Vielzahl an metrischen Variablen vorliegt, können nur wenige auf Grund von hohen Korrelationen untereinander oder fehlendem sachlichen Zusammenhang in die Überlegung mit einbezogen werden. Außerdem ist keine der erklärenden Variablen strenggenommen normalverteilt (vgl. 4.6.1). Die Variablen „Budget“ und „US Umsatz“, von denen mit der höchste sachliche Interpretationsgehalt angenommen wurde, wurden von der Analyse ausgeschlossen, da sie laut der Varianzanalyse keine signifikanten Gruppenmittelwertunterschiede aufweisen (vgl. 4.6.2). In die Analyse wurden ebenfalls vier ordinale Variablen aufgenommen (vgl. 4.4), obwohl das Verfahren für stetige metrische Merkmale ausgelegt ist. Eine der ordinalen Merkmalsvariablen, Marktpotential, wurde aus einer nominalen Variable mit nur 4 möglichen Ausprägungen hergeleitet und hat dementsprechend einen geringen Informationsgehalt. Übrig bleiben zwei „echte“ metrische Variablen „USScreens“ und „DIS_POWER“. Mit einer nicht signifikanten Varianzhomogenität von p 0.123 (SPSS Bildschirmfoto 4-17) ist der Beitrag der Variable „USScreens“, die Unterschiede der abhängigen Gruppen insgesamt zu erklären, recht gering. Dafür ist die Varianz der Variable zu hoch und die Homogenität der Gruppenvarianzen zu gering. Dafür spricht auch der geringe Eigenwert der geschätzten Diskriminanzfunktion von 0.282 (siehe SPSS Bildschirmfoto 4-26). Dennoch hat der Faktor „USScreens“ in dieser Funktion insgesamt relativ die höchste Trennfähigkeit mit 45.7% (vgl. Tabelle 4-5). „DIS_POWER“ hat hingegen zwar deutlichere homogene Varianzen (p Wert: 0,685), trägt jedoch relativ betrachtet deutlich weniger (10.9%) zur Separation der Gruppen durch die Diskriminanzfunktion bei. Die Diskriminanzfunktion ist Alles im Allem als nicht ausreichend fähig Gruppenmittelwertunterschiede anhand der ausgewählten Prädiktoren zu erklären, einzuschätzen. Hierfür spricht das Gütemaß Wilks Lambda mit nur 0.78 (SPSS Bildschirmfoto 4-27) und ein entsprechendes Ergebnis des Klassifikationserfolgs anhand der Klassifikationsmatrix am Kontrolldatensatz, welches zeigt, dass der Gesamterfolg des Modells nicht weit von den a-priori-Wahrscheinlichkeiten abweicht und die Voraussage von erfolgreichen Filmen nur mit hohen Fehklassifikationskosten verbunden wäre (vgl. 4.9.1).

[82] Bei ungleichen a-priori-Wahrscheinlichkeiten wird der natürliche Logarithmus der a-priori-Wahrscheinlichkeit der entsprechenden Gruppenzugehörigkeit zum Funktionswert F1, F2 und F3 addiert. Bps: 10% a-priori für Gruppe 1 → $F1 + (ln\ 0.1) = b_0 + b_1 \cdot x_1 \dots b_n \cdot x_n$ , 40% a-priori-W. für Gruppe 2 → $F2 + (ln\ 0.4) = b_0 + b_1 \cdot x_1 \dots b_n \cdot x_n$ , 50% a-priori-W. für Gruppe 2 → $F3 + (ln\ 0.5) = b_0 + b_1 \cdot x_1 \dots b_n \cdot x_n$. Dabei ist die Summe der a-priori-Wahrscheinlichkeiten immer 1.

Die Leitfragen der Untersuchung (vgl. 4.4) können daher folglich nicht sicher anhand des zugrundeliegenden Modells bewertet werden. Laut des vorliegenden Modells haben die Variablen „Budget“ und „USUmsatz“ keine Auswirkungen auf die abhängigen Variablen. Obgleich grundsätzlich anhand von Wilks Lambda die Trennfähigkeit der Diskriminanzfunktion statistisch als signifikant bewertet werden kann (vgl. 4.8.2), lässt die Auswertung der Güte der Diskriminanzfunktion die Schlussfolgerung zu, dass die Ausprägungen der restlichen erklärenden Variablen auf die abhängige Variable die Gruppen nicht hinreichend gut voneinander unterscheiden können. Den im Modell größten Einfluss auf die Erklärung der Mittelwertunterschiede hat dabei die Variable USScreens, mit 45.7%, gefolgt von der Variable ACTORS mit 17.1%. Der Regisseur des Films beeinflusst laut diesem Modell den Erfolg am wenigsten (1.3%).

Unter Annahme der Gültigkeit dieses Modells, beeinflusst eine Altersbeschränkung den Erfolg des Films tatsächlich negativ. Die relative Bedeutung des erklärenden Faktors ist 9.8%. Das Vorzeichen des nicht-standardisierten (kanonischen) Koeffizienten ist negativ (vgl. 4.8.5), während der Gruppenmittelwert der Gruppe „rentabel“ im Gegensatz zur Gruppe „nicht rentabel“ positiv ist (vgl. SPSS Bildschirmfoto 4-31). Das bedeutet, dass ein geringeres Marktpotential (in %) sich entsprechend geringer negativ auf den Gesamtwert der Diskriminanzfunktion auswirkt.

Laut des Modells lässt sich die Frage, ob die Filmvorführungen tatsächlich die wichtigsten Erfolgsindikatoren sind, mit „ja“ beantworten. Der standardisierte Funktionskoeffizient ist mit 1.257 betragsmäßig am größten und hat dadurch den größten und einen positiven Einfluss auf den Wert der Diskriminanzfunktion (vgl. SPSS Bildschirmfoto 4-28).

Die Antwort auf die letzte Leitfrage ist, dass das Modell des 2-Gruppen-Falls nicht dazu geeignet ist, den Erfolg eines Films durch die gewählten Merkmalsvariablen zu prognostizieren.

Diese Aussage trifft für das Modell, welches für den Mehr-Gruppen-Fall geschätzt wurde, ebenfalls zu. Allerdings nicht, weil die Treffsicherheit des Modells nicht hinreichend ausfällt. Im Gegenteil: durch „overfitting“ auf Grund des Verwendens bestimmter erklärender Variablen, die ursprünglich zur Generierung der Gruppen verwendet wurden, verliert das Modell seine praktische Bedeutsamkeit (vgl. 4.10.4).

Das Beispiel am Fall mit 3 Gruppen (vgl. 4.10) verdeutlicht außerdem, dass es durchaus sinnvoll sein kann, die Prädiktoren für sein Modell manuell auszuwählen. Automatisierte, schrittweise Methoden optimieren immer hinsichtlich statistischer Vorteile. Wenn man etwa weiß, welche Prädiktoren sachlogisch von Bedeutung sein könnten, dann sollte man die erklärenden Variablen selbst auswählen. Falls eine hohe Anzahl an erklärenden Variablen zur Verfügung steht, kann die schrittweise Methode verwendet werden, um explorativ diejenigen Prädiktoren ausfindig zu machen, die besonders hohe Erklärungskraft haben. In einem zweiten Schritt sollte dann überlegt werden, ob diese auch in einen entsprechenden logischen Zusammenhang mit der abhängigen Variablen gebracht werden können.

# 6. Methodenvergleich: Diskriminanzanalyse und Entscheidungsbäume

Im folgenden Kapitel sollen das statistische Verfahren der Diskriminanzanalyse und der Entscheidungsbaumanalyse miteinander verglichen werden. Hierfür werden zunächst die Grundlagen der Entscheidungsbaumanalyse wiedergegeben. Das Verfahren wird dabei nicht so detailliert wiedergegeben, wie zuvor die Diskriminanzanalyse, grundlegendes jedoch verständlich gemacht. Danach wird anhand derselben Überlegungen zu Erfolgsfaktoren von Kinofilmen exemplarisch eine Entscheidungsbaumanalyse durchgeführt und die Ergebnisse der Diskriminanzanalyse und die der Entscheidungsbaumanalyse verglichen.

## 6.1. Allgemeiner Überblick über die Entscheidungsbaumanalyse

Die Entscheidungsbaumanalyse ist ein exploratives Verfahren zur Identifizierung von und Klassifizierung in Untergruppen, die sich hinsichtlich der Verteilung der abhängigen Variablen möglichst stark voneinander

unterscheiden[83]. Hierfür wird anhand eines Datensatzes ein baumbasiertes Klassifizierungsmodell erstellt, welches die Objekte verschiedener Merkmalsvariablen anhand einer abhängigen Variablen einer Untergruppe zuordnet. Wie durch den prognostischen Ansatz der Diskriminanzanalyse, können auf Basis der dadurch entstehenden Entscheidungsregeln noch nicht klassifizierte Objekte den Untergruppen des Entscheidungsbaumes zugewiesen werden[84].

Entscheidungsbäume zeichnen sich durch die leichte Nachvollziehbarkeit des Verfahrens sowie die einfache Interpretation der – meist graphisch wiedergegebenen – Ergebnisse aus. Im Laufe des Verfahrens werden schrittweise immer feinere Untergruppen der Datenbasis gebildet, wobei die nächste Unterteilung jeweils anhand des Merkmals erfolgt, das an dieser Stelle eine bestmögliche Trennung der Zielgrößen-Klassen erlaubt. Im Gegensatz zur Diskriminanzanalyse ist dieser Ansatz „monothetisch“, d.h., der p-dimensionale Raum der Variablen wird in einem Schritt immer nur bzgl. einer Koordinate weiter aufgeteilt. Dies kann wesentlich unflexibler sein, als beliebige Schnitt-Hyperebenen zuzulassen, ist dafür aber meist auch wesentlich interpretierbarer[85], da die Reihenfolge der Schritte und in den Prozess der Generierung der Entscheidungsregeln deutlich einfacher eingesehen und gegebenenfalls auch eingegriffen werden kann. Zielsetzung ist die Bildung weitgehend homogener Gruppen auf der untersten Baumebene[86]. Die im letzten Schritt entstehenden Klassen, auch Endknoten genannt, sind durch die mit ihr darüber verästelten Entscheidungsregeln charakterisiert. Anschließend wird der Baum jedoch auf eine sinnvolle Größe „zurückgestutzt“ (sog. „pruning“), um die Überanpassung des Modells an die Trainingsdaten (sog. „overfitting“) zu reduzieren und damit die Prognosegüte des Modells auf neuen Daten zu erhöhen[87].

## 6.2. Vorgehensweise der Entscheidungsbaumanalyse

Um die Trennfähigkeit der Merkmalsvariablen zu prüfen, gibt es verschiedene statistische Ansätze, die sich hauptsächlich in der Art und Weise unterscheiden, nach welchem Testkriterium sie die Untergruppen ausfindig machen[88]. In dieser Arbeit wird der Vorgang vorgestellt, der auf dem CHAID Algorithmus beruht.

### 6.2.1. Trennen oder Verbinden: CHAID Algorithmus

Für die Baumanalyse sind nur ordinale oder nominale Prädiktoren zulässig. Metrische Merkmale werden automatisch in eine ordinale Form transformiert, bevor der Algorithmus beginnen kann. Die CHAID Prozedur erzeugt nichtbinäre Bäume, das heißt aus Trennungen können mehr als nur zwei Unterknoten resultieren.

Die CHAID (Chi-squared Automatic Interaction Detectors) Prozedur prüft alle zu bewertenden Prädiktoren über den Chi-Quadrat-Test[89]. Als erstes wird überprüft, ob die Kategorien eines Prädiktors zusammengefasst werden können[90]. Dies ist dann der Fall, wenn die Kategorien des Prädiktors in Abhängigkeit der Zielvariablen keine signifikanten Unterschiede aufweisen. Kategorien können nur dann zusammengefasst werden, wenn es davon mehr als zwei gibt. Bei nominalen Prädiktoren können so beliebige Kategorien miteinander fusionieren, wohingegen bei ordinal skalierten Prädiktoren nur benachbarte Kategorien zusammengefasst werden können. Die dann übrigbleibenden Kategorien können im Sinne der Baumanalyse als homogene Teilgruppen verstanden werden. So resultiert für jeden Prädiktor häufig ein überarbeiteter Satz von Kategorien.
Sobald keine Kategorien der verschiedenen Prädiktoren mehr zusammengefasst werden können, wird erneut durch den Chi-Quadrat Test geprüft, welche der noch nicht verwendeten Prädiktoren einen signifikanten

[83] Tuschl 2016
[84] vgl. IBM – Decision Trees
[85] Dr. Martin Mächler, Statistisches Data-Mining, ETH Zürich, Okt. 2016
[86] Grundlagen des CRM: Strategie, Geschäftsprozesse und IT-Unterstützung
[87] Hippner 2011
[88] Ansätze zur Konstruktion von Entscheidungsbäumen, Moritz Duhme, Studienarbeit
[89] IBM
[90] Tuschl 2016

Einfluss auf die abhängige Variable haben. Haben mehrere Prädiktoren einen Einfluss, wird derjenige Prädiktor als Trennvariable bzw. Splitvariable definiert, der die höchste Signifikanz (kleinster p-Wert bei höchstem Chi-Quadrat) aufweist.
Die CHAID Prozedur ist an der Stelle beendet, wenn kein Prädiktor mehr einen signifikanten Einfluss auf die abhängige Zielvariable vorweisen kann. Die Prozedur kann ebenfalls durch vorher festgelegte Parameter beendet werden, wie dem Unterschreiten einer Mindestanzahl an Objekten im nächsten Knoten, die minimale Größe eines Knotens ist unterschritten oder einer anderen Abbruchregel. Die letzten Knoten werden Endknoten genannt. Anhand des Knotenverlaufs können rückwärts die Entscheidungsregeln für die spezifischen Endknoten abgelesen werden. Die Entscheidungsregeln spezifizieren die jeweiligen Charaktereigenschaften der durch die Endknoten entstehenden Teilpopulationen, auch finale Segmente genannt. Dabei gehören Objekte in den Endknoten der Kategorie der abhängigen Variablen an, dessen Ausprägung im jeweiligen Endknoten am häufigsten vorkommt.

### 6.2.2. Pruning

Das englische Wort „Pruning“ bedeutet im deutschen „Beschneiden“, „Kürzen“ oder „Schnitt“[91]. Wortwörtlich soll der entstandene Entscheidungsbaum durch das „pruning“ auf eine sinnvollere Form zurückgestutzt werden. Er wird so verkleinert, dass die Genauigkeit des Baumes optimiert wird. Durch das Entdecken von Anomalien wie Ausreißern, können die entsprechenden Verästelungen eines Baumes entfernt werden.
Bei dem Pruning gibt es zwei verschiedene Ansätze, das Prepruning und das Postpruning:

- Prepruning: der Baum wird bei seiner Konstruktion rechtzeitig angehalten und dadurch „vorzeitig“ gestutzt.
- Postpruning wird zuerst ein Entscheidungsbaum vollständig aufgebaut, und anschließend wieder zurechtgestutzt.

Beides passiert durch die Vorgabe von Regeln, wie etwa einer maximalen Anzahl von Knoten, minimalen Anzahl von Fällen in einem neuen Knoten oder minimalen Anzahl von Fällen in einem Knoten, um ihn erneut zu trennen. Wenn entschieden wird, die Partitionierung an einem Knoten zu stoppen, wird bei diesem Ansatz der Knoten zu einem Endknoten.

## 6.3. Ergebnisse der Baumanalyse anhand des Datensatzes zu Erfolgsfaktoren von Kinofilmen

Anhand der gleichen Prädiktoren, die schon für die Diskriminanzanalyse verwendet wurden, soll nun eine Baumanalyse stattfinden. Der einfachheitshalber wird an dieser Stelle kein Holdout-Verfahren verwendet, um Test- und Trainingsstichproben zu ermitteln. Angelehnt an die vorherige Analyse, sind die Prädiktoren für das Modell „ACTORS“, „DIRECTOR“, „USScreens“, „Marktpotential“, „DIS_POWER und „CRITIC“. SPSS berechnet von selbst, welche Prädiktoren am besten geeignet sind, um die Untergruppen zu erzeugen. In der Modellzusammenfassung wird außerdem aufgezeigt, welchen zusätzlichen Prepruning-Regeln der Baum unterliegt. Der Baum hat eine maximale Tiefe von 3 Stufen, es sollen immer mindestens 100 Fälle in einem Knoten sein, um ihn ein weiteres Mal aufzuteilen und sobald ein Knoten weniger als 50 Fälle in ihm hätte, wird er nicht entstehen, um overfitting zu vermeiden.

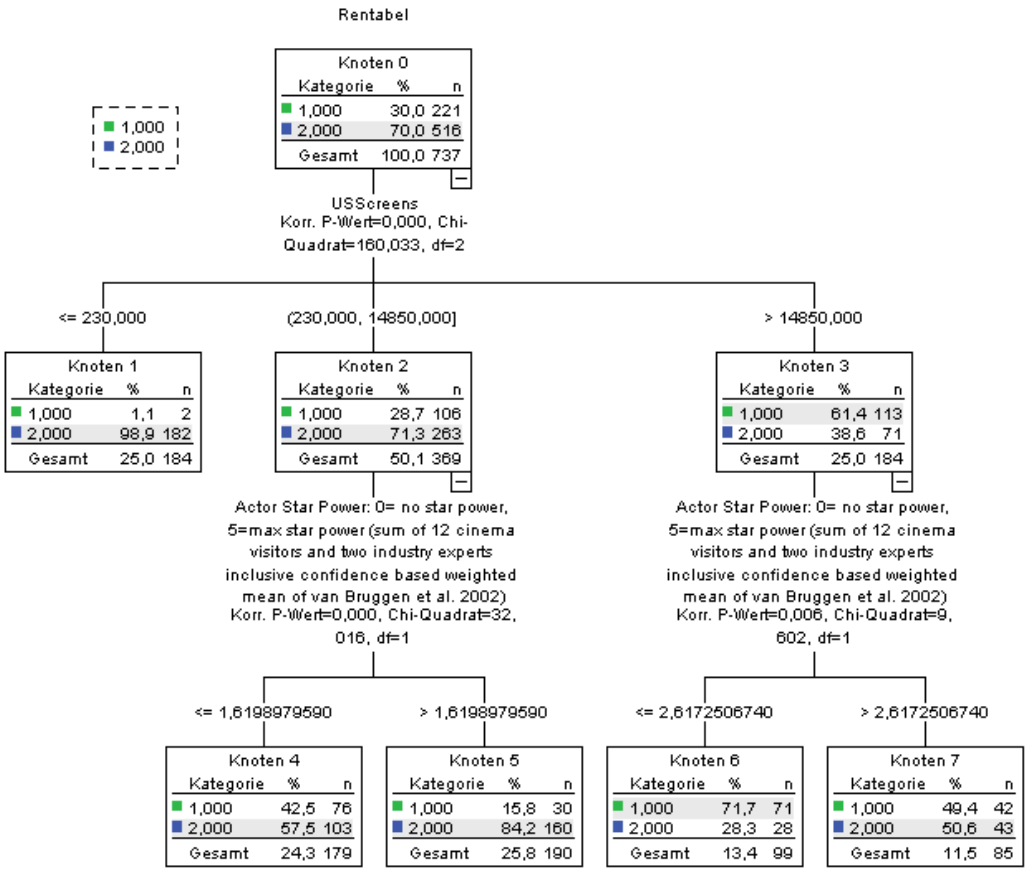
[91] https://www.dict.cc/englisch-deutsch/pruning.html

**Modellzusammenfassung**

| | | |
|---|---|---|
| Spezifikationen | Aufbaumethode | CHAID |
| | Abhängige Variable | Rentabel |
| | Unabhängige Variablen | ACTORS, DIRECTOR, USScreens, Marktpotential, CIN_CRITIC, DIS_POWER |
| | Validierung | Keine |
| | Maximale Baumtiefe | 3 |
| | Mindestanzahl der Fälle im übergeordneten Knoten | 100 |
| | Mindestanzahl der Fälle im untergeordneten Knoten | 50 |
| Ergebnisse | Aufgenommene unabhängige Variablen | USScreens, ACTORS |
| | Anzahl der Knoten | 8 |
| | Anzahl der Endknoten | 5 |
| | Tiefe | 2 |

*SPSS Bildschirmfoto 6-1 – Modellzusammenfassung des Entscheidungsbaumverfahrens*

Im Ergebnisbereich der Modellzusammenfassung kann man sehen, welche Variablen als für am geeignetsten befunden wurden, nämlich „USScreens“ und „ACTORS“. Darüber hinaus kann man ablesen, dass der Baum eine Tiefe von 2 hat und es insgesamt 5 finale Untergruppen gibt, die sich anhand ihrer Merkmale deutlich voneinander unterscheiden.




*SPSS Bildschirmfoto 6-2 – exemplarischer Entscheidungsbaum anhand Erfolgsfaktoren von Kinofilmen*

Zu sehen ist ein Entscheidungsbaum, der 737 Fälle in 6 Unterknoten aufteilt. 221 Fälle davon gehören Gruppe 1 “rentabel“ und 516 der Gruppe 2 „nicht rentabel“ an. Die wichtigste Splitvariable ist „USScreens“. Mit einem p-Wert von 0.000, also sehr hoher Signifikanz, muss die Nullhypothese, dass der Prädiktor keinen signifikanten Einfluss auf die abhängige Variable hat, abgelehnt werden. Doch bevor die Variable als für am geeignetsten

befunden wurde, wurden automatisch geeignete Intervalle erzeugt, um benachbarte Ausprägungen zusammenzufassen. Da die Variable metrisch ist, hat SPSS die Intervalle automatisch erzeugt. Über die Festlegung der Kriterien hat man ebenfalls die Möglichkeit zu bestimmen, wie viele Intervalle der Algorithmus maximal ausfindig machen soll oder mit anderen Worten, wie viele Äste aus einem Knoten maximal entstehen sollen. Der Standardwert von SPSS beträgt 10. Falls ein Baum zu unübersichtlich wird, kann man ihn über die Intervalle stutzen. SPSS konfiguriert den Baum hinsichtlich der korrekten Vorhersage optimal, weswegen eine hohe Baumbreite zwar eine höhere Trefferquote erzielt aber auch „overfitting" bedeutet und eine schlechtere Interpretierbarkeit. Deswegen wurde die Anzahl der maximalen Intervalle auf 4 begrenzt. Dabei wurde jede Ausprägung „kleiner oder gleich 230" zu einem Intervall zusammengefasst, Ausprägungen „zwischen 230 und 14850" und Ausprägungen „die größer als 14850" sind. Die Zerteilung erzeugt einen Endknoten, Knoten 1, und zwei Knoten, Knoten 2 und 3, die weiter zerteilt werden. Der Endknoten Knoten 1 beinhaltet eine extrem homogene Gruppe (98,9%) von Fällen, die zu Gruppe 2 gehören. Ihn charakterisiert die Variable „ZUSScreens" anhand der Regel, dass ein Film nur weniger als oder 230 Vorführungen innerhalb der ersten 20 Wochen erzielen konnte. Hat ein Objekt diese Merkmalsausprägungen, kann man ihn mit einer Wahrscheinlichkeit von 98,9% in die Gruppe 2 "nicht rentabel" einordnen. Knoten 2 und 3 werden erneut – beide - durch die Variable „ACTORS" gesplittet. An dieser Stelle ist wichtig zu erwähnen, dass SPSS für jeden Knoten erneut jeden Prädiktor auf Trennfähigkeit überprüft. Es kann durchaus sein, dass Filme, die häufig vorgestellt wurden, also beispielsweise zum Intervall „größer als 14850" gehören, durch eine andere Variable besser aufgeteilt werden können, als Filme, die weniger Vorführungen erzielen konnten. Dieser Sachverhalt wird von dem CHAID Algorithmus berücksichtigt. Nichtsdestotrotz ist der wichtigste Prädiktor zum Splitten der Knoten 2 und 3 in beiden Fällen „ACTORS". Die Variable ist ebenfalls metrisch. SPSS hat hier für Knoten 2 Intervalle von „kleiner oder gleich 1.6", und „größer als 1.6" festgelegt und für Knoten 3 „kleiner oder gleich 2.6" und „größer als 2.6" festgelegt. Sowohl aus Knoten 2 als auch aus Knoten 3 entstehen zwei Endknoten. Knoten 4 und 5 aus Knoten 2 und Knoten 6 und 7 aus Knoten 3. Objekte in Endknoten 4 charakterisiert eine eher schlechte Besetzung bei einer breitgefächerten möglichen Anzahl von Kinovorführungen bis 14850. Objekte, die in diesen Knoten fallen, können nicht gut genau einer Klasse zugewiesen werden. Von 179 Fällen gehören 76 zu Gruppe 1 und 103 zu Gruppe 2. 76 (42.5%) Fälle werden also durch die Zuweisung des Knotens als „Prognoseknoten zu Gruppe 2" verkehrt klassifiziert. Das sind bereits etwa ein Drittel aller rentablen Filme (76 von 221), die durch diesen Endknoten verkehrt klassifiziert werden. Knoten 5 ist deutlich genauer. Ihn charakterisieren eine eher schlechte bis hin zur besten Besetzung bei einer breitgefächert möglichen Anzahl von Kinofilmen. Filme mit diesen Merkmalsausprägungen sind in 84.2% nicht rentabel. Knoten 6 ist der einzige Knoten, der dazu verwendet werden kann, rentable Filme zu prognostizieren. 71 von 99 Filmen dieses Knotens gehören der Gruppe 1 an. Dieser Knoten hat eine Trefferquote von 71.7%. Ihn charakterisieren eine schlechte bis mittlere Besetzung und überdurchschnittlich viele Filmvorführungen ab 14850. Knoten 7 wird durch mittlere bis top Besetzung charakterisiert, bei überdurchschnittlich vielen Filmvorführungen. Durch diese Charakterisierung kann man die Gruppen nicht sonderlich gut auseinanderhalten. Es befinden sich 42 Objekte der Gruppe 1 und 43 Objekte der Gruppe 2 in ihm. 49.4% der rentablen Filme werden durch diesen Knoten als nicht rentabel eingestuft.

Auch hier kann man die Klassifizierungsmatrix zur Analyse der Ergebnisse verwenden. Sie wird von SPSS mit im Output des Baumverfahrens ausgeworfen.

**Klassifikation**

| | Vorhergesagt | | |
|---|---|---|---|
| Beobachtet | 1,00 | 2,00 | Prozent korrekt |
| 1,00 | 71 | 150 | 32,1% |
| 2,00 | 28 | 488 | 94,6% |
| Gesamtprozentsatz | 13,4% | 86,6% | 75,8% |

Aufbaumethode: CHAID
Abhängige Variable: Rentabel

*SPSS Bildschirmfoto 6-3 – Klassifikationsmatrix des Entscheidungsbaum-Modells*

In der Hauptdiagonalen findet man die richtig vorhergesagten Fälle. Der Baum hat eine Trefferquote von 75.8%. Von besonderem Interesse ist allerdings, ob das Modell den Erfolg eines Films prognostizieren kann, weswegen die Zeile „Gruppe 1"-Beobachtet und der Wert „Prozent korrekt" von übergeordneter Bedeutung

sind. Dieser Wert unterscheidet sich nicht viel (2.1%) von der a-priori-Wahrscheinlichkeit (30%), die sich aus den Gruppengrößen der Stichprobe ableiten lässt. Des Weiteren weist das Modell ebenfalls hohe Fehlklassifikationskosten auf, wenn man die Opportunitätskosten als solche versteht. Sie würden dadurch anfallen, dass man 67.9% der Filmprojekte nicht weiterverfolgen würde, obwohl sie sich als rentabel herausstellen würden. Das Modell eignet sich insgesamt nicht, um den Erfolg von Kinofilmen vorherzusagen.

## 6.4. Tabellarischer Vergleich: Diskriminanzanalyse und Baumanalyse

| Merkmal | Lineare Diskriminanzanalyse | Entscheidungsbaumanalyse |
|---|---|---|
| Leitfrage | Welche Variablen können Objektgruppen optimal voneinander unterscheiden? | Welche Objektgruppen entstehen mit was für Eigenschaften? |
| Verfahrenstyp | Dependanzanalyse | Dependanzanalyse |
| Parameterabhängigkeit | verteilungsabhängig | verteilungsfrei |
| Voraussetzungsannahmen | Verteilungsabhängig, multivariate Normalverteilung, geringe Korrelation, gruppengleiche Kovarianz-Varianz-Matrizen | nicht verteilungsabhängig, keine Voraussetzungen |
| multivariates Verfahren | konfirmatorisch, strukturprüfend | taxonomisch, strukturentdeckend |
| abhängige Variable | nominal | nominal |
| metrische Variablen als Prädiktor | Ja | Ja (aber nur zusammengefasst in Klassen bzw. Intervallen) |
| ordinale Variablen als Prädiktor | nicht optimal | Ja |
| nominale Variablen als Prädiktor | Nein, aber Transformation je nach Sachverhalt möglich | Ja |
| Daten- und Stichprobenumfang | min. 20 pro Merkmalsausprägung zur Schätzung des Modells ($n \geq 20$), mindestens ein Prädiktor mehr als Anzahl der Gruppen | $\geq 2$ Gruppen und mindestens eine abhängige Variable |
| „overfitting“ | durch zu große Teststichprobe | durch zu komplexen Baum, durch zu große Teststichprobe |
| „overfitting“-Vorkehrungen | Lern- und Kontrollstichproben | Post- und Prepruning, Lern- und Kontrollstichproben |
| Einschätzung der Interpretierbarkeit und Komplexität | eher komplex, da viele Voraussetzungsannahmen und fundierte Vorkenntnisse zur Interpretation der Gütekriterien notwendig sind | sehr einfach und verständlich, da Ergebnisse visualisiert und intuitiv interpretierbar |
| Hypothesenprüfungen | Levene-Test, Box M, Wilks Lambda, Chi-Quadrat-Test | Chi-Quadrat-Test |
| Anzahl der abhängigen Gruppen | Nicht relevant, kann bei entsprechender Datengrundlage für beliebig viele Gruppen angewandt werden. Es ist jedoch empfehlenswert, dass besonders unterschiedliche Gruppen betrachtet werden, wie zum Beispiel entgegengesetzte Extremfälle (Heavy-User vs. Infrequent- | gering, wird sehr unübersichtlich und nicht interpretierbar durch zu vielen Gruppen, da die Endknoten nur einer Gruppenzugehörigkeit entsprechen |

| | | |
|---|---|---|
| | User) | |
| Datenqualität | Fehlende Fälle werden nicht berücksichtig, Ausreißer verzerren die Verteilungsfunktionen und sollten ausgeschlossen werden | Fehlende Fälle werden nicht berücksichtig, Ausreißer unempfindlich, da metrischen Merkmale zusammengefasst werden |
| Art der Klassifikation | Über eine oder mehrere lineare Diskriminanzfunktionen | Über eine Klassifikationsregel je Endknoten |
| Funktionswert | eindeutiger Diskriminanzwert für jedes Element | geschätzt anhand von Entscheidungsregeln und Wahrscheinlichkeiten der Knoten, kein eindeutiger Wert für jedes Element |
| Visualisierung | Merkmalsraum (oder Achse) mit Gruppenzugehörigkeiten, Betrachtung einzelner Fälle möglich | Verästelter Entscheidungsbaum, differenzierte Betrachtung einzelner Elemente nicht möglich |
| Anwendbarkeit | für komplexe metrische Dependanzanalysen anhand verschiedener Gruppen sehr gut geeignet | für schnelle, oberflächliche Analysen, um Muster und andere Anomalien anhand der Daten zu entdecken |

*Tabelle 6-1: Tabellarischer Vergleich zwischen der Diskriminanzanalyse und dem Entscheidungsbaumverfahren*

# 7. Zusammenfassendes Fazit

Die exemplarische Durchführung der Analysen und der tabellarische Vergleich haben gezeigt, dass die Methoden einige Gemeinsamkeiten aber auch Unterschiede haben. Die Ergebnisse der Entscheidungsbaumanalyse sind sehr einfach zu interpretieren. Gerade für Anwender eines nicht-analytischen Hintergrunds. Es werden wenig statistische Vorkenntnisse vorausgesetzt, um das Verfahren durchzuführen. Es kann jede Art von Variablenskalierung für die Betrachtung von Gruppenunterschieden verwendet werden. Die grafische Darstellungsweise der Ergebnisse ist intuitiv und sehr leicht nachvollziehbar. Das Verfahren eignet sich außerdem, um Strukturen zu erkennen. Beispielsweise können besonders wichtige, in der Baumanalyse früh als Splitvariable verwendete, unabhängige Prädiktoren für nachfolgende Dependanzanalysen verwendet werden. Im Umkehrschluss macht diese Eigenschaft das Entscheidungsbaumverfahren darüber hinaus zu einer „Wunderwaffe“, nicht aussagekräftige Prädiktoren aus einer Analyse auszuschließen. Das jedoch, kann die Diskriminanzanalyse mit SPSS dank der Funktion des schrittweisen Auswählens der geeignetsten erklärenden Variable zum Modell jedoch auch. Hierdurch erhält die Diskriminanzanalyse ebenfalls einen explorativen Charakter. Metrische Merkmale werden bei der Baumanalyse zu Intervallen zusammengefasst. Dadurch ist das Verfahren relativ Ausreißer unempfindlich. Allerdings werden lineare Beziehungen der Merkmalsvariablen untereinander in keiner Weise berücksichtig und durch die Gruppierung der metrischen Ausprägungen zu Intervallen werden einige Informationen nicht mit in die Analyse einbezogen. Hier ist die Diskriminanzanalyse entscheidend genauer. Wechselwirkungen werden beachtet und durch entsprechende Gewichtungen der volle Informationsgehalt der metrischen Variablen genutzt. Das erlaubt präzise Gruppenzugehörigkeitsbestimmungen anhand individueller Diskriminanzwerte für jedes durch eine Diskriminanzfunktion geschätzte Element. Die Genauigkeit kommt jedoch zum Preis der Komplexität des Verfahrens und dessen Voraussetzungen. Koeffizienten können schnell durch parametrische Voraussetzungsfehler verkehrte Werte annehmen und die Diskriminanzfunktion dadurch verzerren. In Anwendungsbereichen wie beispielsweise der Medizin, zum Erkennen versteckter Muster, ist das Verfahren

daher zwar sehr wertvoll, die Durchführung selbst allerdings auch sehr sorgsam und gewissenhaft zu tätigen. Die Interpretierbarkeit der Ergebnisse setzt ebenfalls voraus, dass man sich methodisch mit dem Verfahren auseinandergesetzt hat. Die Möglichkeit Ergebnisse der Diskriminanzanalyse zu visualisieren, ist vielseitig - muss jedoch manuell erfolgen. Auch dies setzt Fachkenntnisse voraus. Auch andere Voraussetzungen, wie die Normalverteilung der Variablenwerte, welche bei realen Erhebungen nur selten vorliegt, stehen der einfachen Anwendbarkeit in der Praxis im Weg. Hiermit hat die Baumanalyse keine Probleme. Das Verfahren ist verteilungsfrei.

Zusammenfassend kann gesagt werden, dass die Baumanalyse bei Analysen geringen Umfangs und geringer Komplexität ein sehr geeignetes Mittel ist, rasch Ergebnisse zu erhalten, die sehr gut und intuitiv interpretierbar sind. Ohne viel Aufwand versteht man als Anwender schnell, wie die statistische Methode abläuft und anzuwenden ist. Allerdings hat die Baumanalyse auch ihre Nachteile. Sie kann keine Wechselwirkungen der Variablen untereinander betrachten und stößt an ihre Grenzen, wenn zu viele Gruppen in Abhängigkeit stehen. Hier ist der Diskriminanzanalyse kein Limit gesetzt. Bei komplexen Dependanzanalysen können, bei entsprechender Datengrundlage, beliebig viele Gruppen und Prädiktoren verarbeitet werden. Dabei liefert die Diskriminanzanalyse unter Beachtung der Voraussetzungsannahmen statistisch optimierte Ergebnisse für jedes individuelle Objekt, die die Erstellung eines präzisen Modells zur Prognose noch nicht klassifizierter Objekte und zur Erklärung der Gruppenmittelwertunterschiede zulassen. Hieraus lassen sich exakte zielgruppenspezifische Maßnahmen formulieren, um die Gruppenzugehörigkeit eines Objekts zu beeinflussen.

Obgleich die lineare Diskriminanzanalyse bei Verletzungen der Voraussetzungsannahmen recht robust ist, gibt es einige verteilungsfreie Alternativmethoden. In künftigen Arbeiten kann auf dem Wissensstand dieser Arbeit aufgebaut und moderne, von der linearen Methode abweichende Diskriminanzanalysen können vorgestellt werden. Untersuchungsfragen könnten die Untersuchung der Präzision der verschiedenen Verfahren am gleichen Datensatz sein oder eine weitere anwendungsorientierte Einführung in eine nicht-lineare Form der Diskriminanzanalyse. Des Weiteren wurde bei Weitem noch nicht der gesamte Funktionsumfang der SPSS-Software hinsichtlich des Analyseverfahrens erläutert. Auf diese Arbeit aufbauende Untersuchungen können weiter in die Umfänge der Analysemethode mit SPSS vordringen und Funktionen darstellen, wie das ergebnisorientierte visualisieren eines der Diskriminanzwerte und – funktionen im Merkmalsraum.

# 8. Verzeichnisse

## 8.1. Literaturverzeichnis

Backhaus, Klaus (2016): Multivariate Analysemethoden. Berlin: Springer Gabler.

Bühl, Achim (2016): SPSS 23. Einführung in die moderne Datenanalyse. 15., aktualisierte Auflage. Hallbergmoos: Pearson (st - scientific tools).

Clement, Michel (2004): Erfolgsfaktoren von Spielfilmen im Kino. Eine Übersicht der empirischen betriebswirtschaftlichen Literatur. In: *M&K* 2 (52), S. 250–271. Online verfügbar unter https://doi.org/10.5771/1615-634x-2004-2-250.

Fantapié Altobelli, Claudia (2017): Marktforschung. Methoden, Anwendungen, Praxisbeispiele. 3., vollständige überarbeitete Auflage. Konstanz, München: UVK Verlagsgesellschaft mbH; UVK/Lucius (UTB Betriebswirtschaftslehre, 8342).

Fisher, R. A. (1936): The Use of Multiple Measurements in Taxonomic Problems.

Hair, Joseph F.; Black, William C.; Babin, Barry J.; Anderson, Rolph E. (19.02): Multivariate Data Analysis. Upper Saddle River, NJ [u.a.]: Pearson Prentice Hall.

Hippner, Hajo (2011): Grundlagen des CRM. Wiesbaden: Springer Fachmedien. Online verfügbar unter http://gbv.eblib.com/patron/FullRecord.aspx?p=749220.

IBM: IBM SPSS Decision Trees 20.

Jedidi, Kamel; Krider, Robert E.; Weinberg, Charles B. (1998): Clustering at the Movies. In: *Marketing Letters* 1998, S. 393–405.

Krol, Bianca; Lübke, Karsten; Ökonomie und Management, FOM-Hochschule für (20.05): Wörterbuch Statistik.

Malhotra, Naresh K. (2007): Marketing research. An applied orientation. 5. ed., internat. ed. Upper Saddle River, NJ: Pearson Prentice Hall (Pearson education international).

Mittag, Hans-Joachim (2016): Statistik. Berlin: Springer Spektrum.

Müller, Wolfgang (2015/2016): Marketing Analytics. Diskriminanzanalyse. Dortmund.

Nothnagel, Michael; Berlin, Humboldt-Universität zu (05.10): Klassifikationsverfahren der Diskriminanzanalyse.

Schendera, Christian F. G. (2007): Datenqualität mit SPSS. München: Oldenbourg. Online verfügbar unter http://www.oldenbourg-link.com/isbn/9783486710694.

Tharwat, Alaa; Gaber, Tarek; Ibrahim, Abdelhameed; Hassanien, Aboul Ella: Linear discriminant analysis: A detailed tutorial.

Thurau-Henning, Thorsten (2004): Spielfilme als Anlageobjekt. Die Höhe des Filmbudgets als Grundlage für Investitionsentscheidungen. "If you make an American film with a beginning, a middle and an end with a budget less than five million dollars, you must be an idiot to lose money.". In: *zfbf* (56), S. 171–188.

Thurau-Henning, Thorsten; Wruck, Oliver (2000): Warum wir ins Kino gehen: Erfolgsfaktoren von Kinofilmen. In: *Marketing ZPF* (3), S. 241–256.

Tuschl, Stefan (2016): QM im Marketing. Hamburg. PDF.

Wolf, Christof; Best, Henning (2010): Handbuch der sozialwissenschaftlichen Datenanalyse. Hg. v. Henning Best Christof Wolf. Wiesbaden: VS Verlag für Sozialwissenschaften.

## 8.2. Tabellenverzeichnis



## 8.3. SPSS Bildschirmfotoverzeichnis

## 8.4. Abbildungsverzeichnis



## 8.5. Formelverzeichnis



## 8.6. Textausschnitts-Verzeichnis