

HUMBOLDT UNIVERSITÄT ZU BERLIN
SCHOOL OF BUSINESS AND ECONOMICS
LADISLAUS VON BORTKIEWICZ CHAIR OF STATISTICS



Inhaltliche Analyse der Wichtigkeit der
Lehrveranstaltungen Statistik I & II an der Humboldt
Universität zu Berlin



Content analysis of the importance of courses „Statistics I & II“ at the Humboldt
University in Berlin

BACHELORARBEIT

ZUR ERKLÄRUNG DES AKADEMISCHEN GRADES BACHELOR OF SCIENCE (B.
Sc.) IN BETRIEBSWIRTSCHAFTSLEHRE

VORGELEGT VON

Xue Liu

(544357)

Prüfer: Prof. Dr. W. Härdle

Betreuer: Dr. S. Klinke

BERLIN, 21 OKTOBER 2013

Acknowledgement

Zunächst möchte ich mich an dieser Stelle bei all denjenigen bedanken, die mich während der Anfertigung dieser Bachelorarbeit unterstützt und motiviert haben.

Ganz besonders gilt dieses dank meinem Betreuer, Herrn Dr. Klinker, der es mir ermöglicht haben, an diesem interessanten Thema zu arbeiten. Seine freundliche und engagierte Betreuung hat mir ziemlich geholfen und seine Geduld und Ermutigung hat mich mein Bestes geben gelassen. Ich danke Ihnen herzlich.

Daneben gilt mein Dank meiner Freunde Qin und Hangyu, die mir zahlreiche Hilfe beim Verwenden des Softwares LATEX angeboten zu haben. Ich möchte auch an meine Kommilitonen Yolanda, Maryna und Ella danksagen. Ich danke euch für alle wir zusammen verbrachte schöne Zeit an Humboldt Universität zu Berlin.

Inhaltsverzeichnis

Abbildungsverzeichnis	iv
Tabellenverzeichnis	v
1 Einleitung	1
2 Datensatz	4
2.1 Datenquelle und Datenauswertung	4
2.2 Datenbereinigung	6
3 Methodik	13
3.1 t-Test für unabhängige Stichprobe	13
3.2 Hierarchische Clusteranalyse	15
3.3 Korrespondenzanalyse	18
4 Datenanalyse	23
4.1 Mittelwertvergleiche mit Hypothesentest	23
4.1.1 t-Test bezüglich des Merkmals <i>Geschlecht</i>	23
4.1.2 t-Test bezüglich des Merkmals <i>Muttersprache</i>	24
4.2 Häufigkeitsverteilung	25
4.2.1 Häufigkeitsverteilung bezüglich der statistischen Themen . . .	25
4.2.2 Häufigkeitsverteilung bezüglich der Vertiefungsgebiete	28

4.3	Cluster- und Korrespondenzanalyse des ersten Versuchs	31
4.4	Cluster- und Korrespondenzanalyse des zweiten Versuchs	37
5	Zusammenfassung	44
	Literaturverzeichnis	47
A	Anhang	48

Abbildungsverzeichnis

4.1	Dendrogramm mit Ward-Verfahren	33
4.2	Korrespondenzraum in zwei Dimensionen	37
4.3	Neues Dendrogramm mit Ward-Verfahren (ohne Fachgebiet 5)	38
4.4	Dendrogramm mit Complete-Linkage-Verfahren	39
4.5	Zeilenpunkte für Inhalte der statistischen Vorlesungen	41
4.6	Spaltenpunkte für Fachgebiete	42
4.7	Zeilen- und Spaltenpunkte	43

Tabellenverzeichnis

2.1	Häufigkeitsverteilung der Studenten an Universitäten (SPSS-Output)	7
2.2	Häufigkeitsverteilung der Studenten, die Statistik-Vorlesungen besucht haben oder nicht (SPSS-Output)	7
2.3	Identitätsregeln	9
2.4	Grad der Identität zwischen <i>Zufallsstichprobe</i> und ihren Unterbegriffen	9
2.5	Befragungsantworten	10
2.6	Befragungsantworten	10
3.1	Kontingenztafel	15
4.1	SPSS-Output des t-Tests mit der Variable Geschlecht	24
4.2	SPSS-Output des t-Tests mit der Variable Muttersprache	24
4.3	Relative Häufigkeit der Variablen in Fragenteil 1	25
4.4	Relative Häufigkeit der Variablen in Fragenteil 2	27
4.5	Häufigkeit der Themengebiete im Komplex 2	29
4.6	Anwendungshäufigkeit der Mengenlehre bei anderen Veranstaltungen	29
4.7	Summe der Male der Bewertung jeder Variable	30
4.8	Zuordnungsübersicht der Clusteranalyse (SPSS-Output)	33
4.9	Verteilung der Lehrveranstaltung in acht Fachgebieten	35
4.10	Auswertung der Trägheit der Dimensionen	36

1 Einleitung

Das Fach Statistik wird von Studenten häufig als schwierig empfunden, da es eine Vielzahl von Kompetenzen, einschließlich quantitative und grafische sowie mathematische Fähigkeiten erfordert. Doch eine wachsende Zahl von Menschen braucht Instrumente mit quantitativen Methoden und Studenten müssen statistische Fähigkeiten erwerben, weil sie mit mehr und mehr Datensets konfrontiert werden (vgl. Hårdler u. Klinke 1999, Abstract). In den Bachelorstudiengängen Betriebswirtschaftslehre und Volkswirtschaftslehre an der Humboldt Universität zu Berlin wurden Statistik-Vorlesungen I & II als Pflichtlehrveranstaltungen für alle Studierenden eingeführt, um diesen statistische Analysemethoden nahezubringen und zu anderen Wirtschaftsgebieten beizutragen. Beispiele sind: die Anwendung der empirischen Analysen und Modellierung von Finanzzeitreihen sowie die Agentenbasierte Modellierung für simulierte und reale Märkte in der Finanzwissenschaft; angewandte mathematische und statistische Methoden zur Risikomessung im Versicherungswesen und Bankensystem ¹; sowie die zahlreichen multivariaten Auswertungsverfahren im Marketingbereich. Deshalb ist es wichtig zu wissen, welche statistischen Themen für Studierende schwer zu erfassen sind und welche erforderlich sind, auch in Hinsicht auf andere Wirtschaftsgebiete. Damit beschäftigt sich die vorliegende Arbeit.

Diese Bachelorarbeit knüpft an die Arbeit „Notwendigkeit der Inhalte der Statistik-Vorlesungen für das Bachelor-Studium“ von Frau Anja Lorenz aus dem Jahr 2011 an. Frau Lorenz hat selbst den Fragebogen, der sich im Anhang befindet, konstruiert und insgesamt 92 Beobachtungen erhalten, von denen 57 Beobachtungen nach der Bereinigung der Daten verwendbar waren. Sie hat nur Häufigkeitstabellen benutzt, um zu herauszufinden, wie viele Studenten bzw. wie viele Prozent der Studenten die Meinung vertreten, dass die einzelnen Teile der Statistik-Lehrveranstaltungen notwendig für andere Vertiefungsgebiete sind, ansonsten hat Frau Lorenz die Daten nicht weiter ausgewertet.

In der vorliegenden Arbeit werden deshalb weitere Analysen hinzugefügt, um mehr

¹Vgl. *Statistik* in Wikipedia

Informationen aus den Daten zu erhalten und dadurch Methoden zur Verbesserung der Lehre zu erhalten. Die Vorgehensweise wird so durchgeführt, dass zum erstens die Plausibilität der Daten geprüft wird. Die ausführliche Diskussion findet in den Abschnitten 2.1 und 2.2 statt. Die Datenquelle, die zu analysierenden Variablen und Zielgruppen werden in Betracht gezogen und die fehlenden Werte sind auch zu bereinigen. Daran schließt sich der Teil der relevanten Theorie (3.1-3.3) an, mit einer kurzen Beschreibung des t-Tests, der Clusteranalyse und der Korrespondenzanalyse. Im t-Test werden die Mittelwerte der Bekanntheit bezüglich Geschlecht und Sprache verglichen. Hier wird erwartet, dass diese sich zwischen Männern und Frauen sowie Deutschen und Ausländern nicht unterscheiden, weil die zwei Faktoren keinen Einfluss auf die Kenntnisse der statistischen Inhalte ausüben sollten, solange die Befragten ihr Bachelor-Studium an HU abgeschlossen haben. Das Ergebnis kann zu einem Grad die Glaubwürdigkeit der Daten noch einmal bestätigen, falls die Nullhypothese nicht abgelehnt wird. In Abschnitt 4.2 wird dann die Häufigkeitsverteilung hinsichtlich der statistischen Begriffe und der Vertiefungsgebiete dargestellt. Es ist eigentlich das zuerst zu lösende Problem, welche Begriffe für andere Veranstaltungen unbedingt benötigt werden und welche nicht. Hier ist besonders die Untersuchung der unnötigen Begriffe erforderlich, damit die Lehrperson ihren Lehrplan dahingehend verändern kann. Das zweite Problem ist, den Zusammenhang zwischen statistischen Inhalten und Fachgebieten zu kennen, d.h., für welche Fachgebiete bzw. Schwerpunkte die einzelnen Statistik-Themen wichtig sind. Dafür ist die Korrespondenzanalyse geeignet: Nach Anwendung der Korrespondenzanalyse werden die isoliert stehenden Inhalte als unähnlich betrachtet und haben die nah beieinander liegenden Inhalte und Fachgebiete enge Beziehung. Um die Untersuchung zu erleichtern, wird zuerst die Clusteranalyse angewendet, weil bei beiden Variablen mehrere Elemente bestehen.

Wegen der Restriktion der Datensammlung und des Informationsvolumens können leider keine weitere Analysen und statistischen Methoden, z.B. Regressions- und Faktorenanalyse geeignet sein. Die Ziele der Arbeit sind, auf der einen Seite die Kompetenz der Untersuchung eines Themas und die Fähigkeit der Behandlung von statistischen Daten mithilfe von SPSS-Software aufzuweisen und auf der anderen

Seite empirische und hilfreiche Informationen anzubieten, um Statistik-Vorlesungen praxisorientierter zu machen und Lernerfolg zu erhöhen.

2 Datensatz

2.1 Datenquelle und Datenauswertung

Wie bereits erwähnt, wurden die Daten mit dem von Frau Anja Lorenz konstruierten Fragebogen erhoben. Die Zielgruppe der Befragung waren Master-Studierende der wirtschaftswissenschaftlichen Fakultät an der Humboldt Universität zu Berlin. Der Fragebogen ist eine direkte und schriftliche Befragung, der eine quantitative Datensammlung in kurzer Zeit ermöglicht. Eine schriftliche Befragung verursacht nur niedrige variable Kosten, erfordert keine Identifikation bzw. ist anonym und es gibt keinen Interviewer-Einfluss, Neben diesen Vorteilen gibt es jedoch auch unvermeidliche Nachteile. So ist häufig die Rücklaufquote gering, die Befragten möchten keine komplizierten Fragen beantworten und während der Umfrage gibt es auch keine Möglichkeit, komplizierter Fragen zu erklären². Dies alles konnte zu einem Mangel an Information unserer Stichprobe führen. Im Folgenden wird nun die Fragebogen-Konstruktion vorgestellt.

Der Fragebogen besteht aus insgesamt drei Komplexen: 1. Demografische Fragen; 2. Belegung der Vertiefungsgebiete; 3. Beurteilung der Statistik-Vorlesungen.

Im ersten Komplex wurden außer Geschlecht, Alter, Sprache und Studiengang noch interessante Fragen gestellt. Beispielsweise wurden alle Studenten, die bei dem Term „Haben Sie Statistik-Vorlesungen an der HU besucht?“ mit „Nein“ antworteten, von der Untersuchung ausgeschlossen, und durch die Fragen „Wie viele Semester haben Sie für Ihr Bachelor-Studium benötigt?“ und „Wie viel Zeit ist zwischen Ihrem Bachelorabschluss und dem Anfang Ihres Masterstudiengangs vergangen?“ wird versichert, dass der Bachelor-Abschluss der Befragten nicht zu lange zurückliegt und dass das dort erworbene Wissen noch präsent ist.

Im zweiten Komplex wurden 23 Vertiefungsgebiete genannt, wie *Marketing*, *Finanzwissenschaft*, *Operations Research* u-s-w. Die Befragten mussten ankreuzen, welche

²Vgl. die Lehrveranstaltung *Marktforschung* an Humboldt Universität zu Berlin, Kapitel 3.2: *Techniken der Datenerhebung und Aufbereitung*

davon sie belegt haben, um herauszufinden, welche Fächer relative enge Beziehung haben und welche häufig zusammen als Schwerpunkt belegt werden.

Im dritten Komplex wurden die Inhalte von Statistik I & II in Komponenten geteilt und die Befragten mussten jeweils zu diesen Komponenten auf drei Fragen antworten. Statistik I bestand aus acht Teilen, nämlich *Mengenlehre*, *Zufallsexperiment*, *Kombinatorik*, *Wahrscheinlichkeitsrechnung*, *Deskriptive Statistik*, *Zufallsvariable*, *Zweidimensionale Verteilung* und *Verteilungsmodelle*. Statistik II wurde in fünf Teile geteilt, nämlich *Zufallsstichprobe*, *Schätztheorie*, *Statistische Testverfahren*, *Regressionsanalyse* und *Zeitreihenanalyse*. Zu jedem Teil gab es noch zwei bis zehn Unterbegriffe, als Beispiele wurden *Mengen*, *Elemente*, *Teilmengen* und *Venn Diagramme* sowie *Mengenoperationen* zu *Mengenlehre* hinzugefügt. Die Befragten mussten zuerst angeben, ob sie diese Teile kennen oder nicht und danach überlegen, ob und für welche Vertiefungsgebiete der Teil wichtig ist.

Im dritten Komplex des Fragebogens gab es einige Design-Probleme, die korrigiert werden mussten:

- In der ursprünglichen Fragestellung wurde nicht deutlich gesagt, dass man Oberbegriffe oder Unterbegriffe einschätzen soll, sodass die meisten Befragten beides angekreuzt haben. Dann kommt die Frage: Welcher Teil soll eigentlich in Rücksicht genommen werden? Die ausführlichen Antworten von Unterbegriffen sind in der späteren Analyse nicht erwünscht, da zu viele Variable verbleiben, allein bei *Verteilungsmodelle* sind 10 verschiedene Verteilungen zu untersuchen. Exponentialverteilung wurde dennoch als unwichtig evaluiert und konnte in der Vorlesung auch nicht weggelassen werden, weil sie eine unentbehrliche Komponente von *Verteilungsmodelle* ist, deshalb war die partielle Untersuchung auch nicht so sinnvoll. Aber sind die Antworten plausibel, wenn man sich nur auf den Oberbegriff konzentrieren soll? Ein noch größeres Problem war die große Anzahl der fehlenden Werte der Oberbegriffe, die ziemlich mehr als Unterbegriffe sind. Die Auswahl der Variablen und die Behandlung der fehlenden Werte werden in 2.2.2 diskutiert.

- Im zweiten Fragenteil waren die Fragen „Wurde in einer anderen Veranstaltung benötigt“ und „Wurde in mehreren Veranstaltungen benötigt“ redundant, weil die Antwort von „eine oder mehrere“ für die Untersuchung unerheblich, d.h., die beiden Antworten wurden im Prozess der Datenanalyse kombiniert werden. Ein weiterer erwähnenswerter Punkt für die Verbesserung der Analyse ist, anstatt zu fragen, ob der Inhalt wichtig für andere Veranstaltungen ist, könnten Skalen als Hilfsmittel eingeführt werden, und die Frage „Wie wichtig ist der statistische Begriff für andere Veranstaltungen?“ wurde so gestellt und mit Zahlen von eins bis fünf von links nach rechts gewertet:

ganz wichtig Überhaupt unwichtig

1 2 3 4 5

Die Merkmalsprägungen der Ordinalskala unterliegen einer Rangordnung. Der Sprung von der Ordinalskala zur Intervall-Skala ist in den Sozialwissenschaften von großer Bedeutung, weil eine Vielzahl mehr oder weniger elaborierter Auswertungsverfahren (wie z.B. schon das arithmetische Mittel, die einfache Korrelation oder die Faktorenanalyse) „eigentlich“ intervallskalierte Variablen voraussetzt (Porst 2008, S71-73). Dadurch wird die Datenanalyse in dieser Arbeit erweitert.

Eine andere Einschätzung und Bearbeitung der Daten erfolgt in Abschnitt 2.2.

2.2 Datenbereinigung

Bevor die Daten mithilfe statistischer Methoden analysiert wurden, mussten sie zuerst bereinigt werden. Im letzten Abschnitt wurde schon die Kombination der zwei Terme „Wurde in einer anderen Veranstaltung benötigt“ und „Wurde in mehreren Veranstaltungen benötigt“ angewiesen, und jetzt musste noch über folgende zwei Fragen nachgedacht werden:

1. Sind die Daten bezüglich des Befragungsobjekts für weitere Untersuchung plausibel?

2. Welche Variablen werden zum Analysieren genommen und wie sehen die fehlenden Werte aus?

a) Zielgruppen

Für die Befragung wurden Master-Studenten als Zielgruppe ausgewählt, so war garantiert, dass die Befragten Inhalte von Statistik-Vorlesungen kennen und auch andere Vertiefungsgebiete belegt haben. Hier war jedoch zu berücksichtigen, dass die meisten dieser Master-Studierenden ihr Bachelor-Studium nicht an der HU abgeschlossen haben, da wegen der knappen Studienplätze nur wenige nach ihrem Abschluss an der HU bleiben können oder möchten. Die Tabellen 2.1 und 2.2 zeigen, dass von den Befragten nur 16,3% ihren Bachelor an der HU erworben haben und 35,9% die Statistik-Vorlesungen besucht haben.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Keine Angabe	3	3,3	4.3	4.3
	HU	15	16,3	21.7	26.1
	Andere Uni	51	55,4	73.9	100.0
	Gesamt	69	75,0	100.0	
Fehlend	System	23	25,0		
Gesamt		92	100,0		

Tabelle 2.1: Häufigkeitsverteilung der Studenten an Universitäten (SPSS-Output)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Ja	33	35,9	35.9	35.9
	Nein	59	64,1	64.1	100.0
	Gesamt	92	100,0	100.0	

Tabelle 2.2: Häufigkeitsverteilung der Studenten, die Statistik-Vorlesungen besucht haben oder nicht (SPSS-Output)

Die Befragungsobjekte waren also zu einem gewissen Grad nicht so geeignet. Eigentlich wären Bachelor-Studierende der HU geeigneter gewesen: Die Fragebögen hätten

in der Vorlesung „Ökonometrie“ und in einigen Seminaren verteilt werden können. Um „Ökonometrie“ zu besuchen, muss man mindestens im vierten Semester sein, die Besuchsvoraussetzungen sind die Kenntnisse aus Statistik I & II und die Seminare, die mehreren Restriktionen unterliegen, dürfen auch erst in höheren Semestern belegt werden.

Hier wurde trotzdem angenommen, dass die Studierenden, die ihre Statistik-Vorlesungen nicht an der HU besucht, aber ihr Studium in einem wirtschaftswissenschaftlichen Bereich absolviert haben, das nötige Wissen mitbringen, um die Fragen des Fragebogens zufriedenstellend zu beantworten. Erst mit dieser Annahme konnte die Analyse durchgeführt werden.

b) Variable

Die Bestimmungsregeln (Tabelle 2.3) waren, dass es ausreicht, einzelne Sachgebiete zu betrachten, falls deren Antworten mit ihren Unterbegriffen gut übereinstimmen. Nur wenn es einen sehr großen Unterschied dazwischen gibt, dann müssen alle Unterbegriffe jeweils analysiert werden, wobei zu viele Variable entstehen. Beispielsweise wurde *deskriptive Statistik* (Oberbegriff) zuerst mit *Laplace* (Unterbegriff) verglichen: Wenn im Fragebogen bei *deskriptive Statistik* im Fragenteil 2 „wurde in keiner Veranstaltung benötigt“ und bei *Laplace* „wurde in einer oder mehreren Veranstaltungen benötigt“ angekreuzt wurde, dann wurde dies als nicht identisch gezählt, sonstige Situationen wurden als identisch angesehen, dies gilt auch für fehlende Werte, d.h., wenn einmal „wurde in einer oder mehreren Veranstaltungen benötigt“ und einmal „keine Angabe“ ausgewählt wurde, war es als unwahrscheinlich anzunehmen, dass der/die Befragte diese Frage vergessen hatte oder sie nicht beantworten wollte, sondern halten sie für beide Situationen gleiche Ansicht und möchte nicht noch einmal ankreuzen.

Zeilenweise von der ersten Beobachtung anfangend gelesen, stellt sich heraus, dass es zahlreiche Leerzellen bei den Beobachtungen 31 bis 39, 67 bis 92 gab, in denen die Befragten nur Fragen im ersten Komplex beantworteten und im zweiten Komplex wurden alle Terme mit „Nein“ angekreuzt und im dritten Komplex wurden sogar

	keiner	einer oder mehreren	keine Angabe
Wurde in keiner Veranstaltung benötigt	✓	X	✓
Wurde in einer oder mehreren Veranstaltungen benötigt	X	✓	✓
Keine Angabe / Weiß ich nicht	✓	✓	✓

Tabelle 2.3: Identitätsregeln

keine Angaben gemacht. Solche Fälle fanden in der Untersuchung keine Anwendung und wurden direkt gelöscht. Für die übrigen 57 Beobachtungen kamen die oben erklärten Bestimmungsregeln zum Einsatz. Tabelle 2.4 zeigt beispielhaft die Anzahl und die Prozentsätze der Beobachtungen des Begriffs *Zufallsstichprobe*, deren Antworten von einzelnen Unterteilen mit ihm übereinstimmten. Die Tabellen zu den anderen Begriffen sind dem Anhang beigefügt ³.

		Frageteil 1	Frageteil 2
Zufallsstichprobe	Verteilung des Stichprobenmittelwertes	53 92,98%	57 100%
	Verteilung der Stichprobenanteilstwertes	52 91,23%	56 98,25%
	Verteilung der Stichprobenvarianz	54 94,74%	57 100%

Tabelle 2.4: Grad der Identität zwischen *Zufallsstichprobe* und ihren Unterbegriffen

Der Tabelle 2.4 lässt sich entnehmen, dass die Information von *Zufallsstichprobe* durch hohe Übereinstimmung (mehr als 90%) sehr repräsentativ sein kann, d.h., wenn nur sie allein als Variable analysiert wird, wird das Ergebnis weniger verzerrt. Dies passierte auch bei anderen Sachgebieten (siehe Anhang A.1). Die meisten zeigten mehr als 90% Übereinstimmung und einige zwischen 85% und 90%. Zu beachten war, dass Venn Diagramm beim Frageteil 1 nur 78,95% seinem Oberbegriff *Mengenlehre* und *Permutation* nur 80,70% ihrem Oberbegriff *Kombinatorik* entsprachen. Den Grund zeigt die folgende Tabelle (Tabellen 2.5 und 2.6):

³Anhang A.1.

	Mengenlehre	Venn Diagramm
Kenne ich und wurde behandelt	48 84,2%	37 64,9%
Wurde nicht behandelt	6 10,5%	6 10,5%
Kenne ich nicht	3 5,3%	14 24,6%

Tabelle 2.5: Befragungsantworten

	Kombinatorik	Permutation
Kenne ich und wurde behandelt	52 91,2%	46 80,7%
Wurde nicht behandelt	1 1,8%	1 1,8%
Kenne ich nicht	4 7,0%	10 17,5%

Tabelle 2.6: Befragungsantworten

Es war offensichtlich, dass den Befragten die Begriffe *Venn-Diagramm* und *Permutation* nicht so geläufig waren. Hier war zu vermuten, dass die Befragten nicht wussten, was hinter den Begriffen *Venn-Diagramms* und *Permutation* steht. Möglicherweise wurden diese Begriffe beim Lehren nicht deutlich herausgestellt, entweder wurden sie als nicht wichtig erachtet und während der Vorlesung zu schnell behandelt oder wegen ihrer wenigen Inhalte wurden sie einfach von Studenten ignoriert. In der folgenden Betrachtung wurden die Antworten auf diese beiden Fragen außer Acht lassen, da sie sehr weit entfernt von den anderen waren und das Ergebnis hätten verzerren können. Demnach kann geschlossen werden, dass es insgesamt 13 Variable waren, nämlich *Mengenlehre*, *Zufallsexperimente*, *Kombinatorik*, *Wahrscheinlichkeitsrechnung*, *Deskriptive Statistik*, *Zufallsvariablen*, *Zweidimensionale Verteilung*, *Verteilungsmodelle*, *Zufallsstichprobe*, *Schätztheorie*, *Statistische Testverfahren*, *Regressionsanalyse* und *Zeitreihenanalyse*.

c) Fehlende Werte

Eines der ersten Probleme, das nach der Dateneingabe und -bereinigung auftritt, sind fehlende Werte (engl. missing values). Sie treten auf, wenn ein Befragter eine Frage nicht beantworten konnte oder wollte.⁴

Grundsätzlich werden in der statistischen Analyse drei Typen von fehlenden Werten unterschieden:

- Missing Complete At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Man spricht von MCAR (zufällig fehlend), wenn das Fehlen des Wertes vollkommen zufällig ist, und mit keiner der abgefragten Variablen in Verbindung steht. MAR (abhängig fehlend) weist einen Zusammenhang zwischen gesuchter Frage A und einer anderen Frage B (die nicht zwangsläufig auch konkret gestellt sein muss) hin. Der problematischste Fall sind fehlende Werte, bei denen das Fehlen allein von der Höhe der Ausprägung abhängt, denn dieses Fehlen weist eine Struktur auf, die nicht rekonstruiert werden kann. Dies wird als MNAR bezeichnet (vgl. Speidel 2009, S5-S6).

Als Imputationsmethoden wurden drei Typen unterschieden: Fallausschuss, einfache Imputation und multiple Imputation. Hier in diesem Fall liegt der Fokus auf dem dritten Komplex, in dem es im Fragenteil 1 keine fehlenden Werte und in Fragenteil 2 zahlreiche fehlende Werte von der Form „weiß ich nicht“ und „keine Angabe“ gab. Die Nichtbeantwortungsfehler gehörten zu zwei Typen: der eine ist die verweigerte Antwort, das zu MNAR zählt und nicht reduziert werden kann, und der andere ist das Double Sampling, d.h., die Befragten haben *Mengenlehre* nicht angekreuzt, sondern jeweils ihre Unterbegriffe eingeschätzt und deshalb konnten die

⁴http://mars.wiwi.hu-berlin.de/mediawiki/mmstat_de/index.php/AQM-Behandlung_fehlender_Werte Erster Absatz

fehlenden Werte gemäß ihres Zusammenhangs ersetzt werden. Dies ist der Fall der MAR und wird mithilfe der multiplen Imputationsmethode in SPSS gelöst, beispielsweise hatte *Mengenlehre* 15 fehlende Werte vor der Imputation und setzte man ein *Mengenlehre* in die Option „Imputieren und als Einflussvariable verwenden“ und gleichzeitig *Mengen_ Elemente_ Teilmengen, Venn-Diagramm* und *Mengenoperation* in die Rolle „Nur als Einflussvariable verwenden“, erhält man fünf Datensätze und ein kombiniertes Ergebnis mit weniger fehlenden Werten.⁵

⁵Im Anhang A.2. befindet sich die Tabelle, die die Anzahl der fehlenden Werte vor und nach der Imputation ergibt.

3 Methodik

3.1 t-Test für unabhängige Stichprobe

Der t-Test ist eine Entscheidungsregel auf einer mathematischen Grundlage, mit deren Hilfe ein Unterschied zwischen den empirisch gefundenen Mittelwerten zweier Gruppen näher analysiert werden kann. Er liefert nur für intervallskalierte Daten zuverlässige Informationen. Deshalb gehört er zur Gruppe der parametrischen Verfahren.⁶

a) Voraussetzung

Für den t-Test gibt es drei mathematische Voraussetzungen:

1. Bei beiden Stichproben handelt es sich um einfache und unabhängige Zufallsstichproben.
2. Die beiden Grundgesamtheiten haben eine gleiche, jedoch unbekannte Varianz: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (Varianzhomogenität)
3. Entweder sind die Zufallsvariablen X_1 und X_2 in den Grundgesamtheiten normalverteilt, d.h. $X_1 \sim N(\mu_1, \sigma_1)$ und $X_2 \sim N(\mu_2, \sigma_2)$, oder die Stichprobenumfänge n_1 und n_2 sind genügend groß, dass der Zentrale Grenzwertsatz wirksam wird. (vgl. Rönz 2001, S 149)

Aus Monte-Carlo-Studien geht hervor, dass der t-Test auch bei einer Verletzung dieser Voraussetzungen noch zuverlässige Informationen liefert. Formal heißt das, der t-Test reagiert robust gegenüber den Verletzungen (vgl. Bortz et al. 2010, S122). „Dies gilt insbesondere, wenn gleich große Stichproben aus ähnlichen möglichst eingipflig-symmetrisch verteilten Grundgesamtheiten verglichen werden. Sind die Stichprobenumfänge sehr unterschiedlich, wird die Präzision des t-Tests nicht beeinträchtigt, solange die Varianzen gleich sind.“ (Bortz et al. S 122)

⁶Vgl. http://www.quantitative-methoden.de/Dateien/Auflage3/leseprobe_A3.pdf

b) Hypothese und Testergebnis

Die Nullhypothese lautet in folgendes:

$$H_0 : \mu_1 = \mu_2$$

bzw. äquivalent

$$H_0 : \mu_1 - \mu_2 = 0$$

Die Nullhypothese (H_0) nimmt an, dass die Mittelwertdifferenz zufällig entstanden ist.

Teststatistik

$$t = \frac{\text{empirische Mittelwertsdifferenz} - \text{theoretische Mittelwertsdifferenz}}{\text{geschätzter Standardfehler der Mittelwertsdifferenz}}$$

$$\text{Formal: } t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sigma_{x_1 - x_2}}$$

$$\text{wobei } \sigma_{x_1 - x_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\sigma_{x_1 - x_2}$: geschätzter Standardfehler der Mittelwertdifferenz

n_1 : Anzahl der Versuchspersonen bzw. Beobachtungen in Stichprobe 1

σ_1^2 : geschätzte Varianz der Population 1

n_2 : Anzahl der Versuchspersonen bzw. Beobachtungen in Stichprobe 2

σ_2^2 : geschätzte Varianz der Population 2

vereinfachte Formel: $t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sigma_{x_1 - x_2}}$, unter der Annahme, dass die Populationsmittelwerte der beiden zu vergleichenden Gruppen identisch sind. Die Prüfgröße ist mit $df = n_1 + n_2 - 2$ Freiheitsgraden verbunden. Die empirische Mittelwertdifferenz wird als signifikant bezeichnet, wenn die ermittelte Wahrscheinlichkeit des t-Werts kleiner ist als der angenommene α -Fehler bzw. der empirische t-Wert im Betrag größer ist als der Betrag des kritischen t-Werts. In diesem Fall ist die Wahrscheinlichkeit des empirischen t-Werts unter der Nullhypothese so klein, dass die Nullhypothese

abgelehnt wird.⁷

3.2 Hierarchische Clusteranalyse

Die Clusteranalyse ist ein heuristisches Verfahren und wird eingesetzt zur systematischen Klassifizierung der Objekte einer gegebenen Objektmenge. Die durch einen festen Satz von Merkmalen beschriebenen Objekte (Personen oder andere Untersuchungsobjekte) werden nach Maßgabe ihrer Ähnlichkeit in Gruppen (Cluster) eingeteilt, wobei die Cluster intern möglichst homogen und extern möglichst gut voneinander separierbar sein sollen (vgl. Bortz et al. 2010, S453). Es lässt sich verschiedene Verfahren in der Clusteranalyse zu unterscheiden, in dieser Arbeit wird nur hierarchische Clusteranalyse eingeführt.

Vorgehensweise

Schritt 1: Bestimmung der Ähnlichkeit

„Für jeweils zwei Personen bzw. Objekte werden die Ausprägungen der Beschreibungsmerkmale geprüft und die Unterschiede bzw. Übereinstimmungen durch einen Zahlenwert (Proximitätsmaß) gemessen.“ (Backhaus et al. 2008, S398) Die Proximitätsmaße bei binären Variablen steht in Tabelle 3.1:

Objekt 1 \ Objekt 2	1(Eigenschaft vorhanden)	2(Eigenschaft nicht vorhanden)
1(Eigenschaft vorhanden)	a	c
2(Eigenschaft nicht vorhanden)	b	d

Tabelle 3.1: Kontingenztafel

Anhand dieser Tabelle ist zu erkennen, wie oft die beiden verglichenen Objekte eine völlige Übereinstimmung aufweisen (a+d: beide haben das Merkmal oder beide haben das Merkmal nicht) und wie oft keine Übereinstimmung vorliegt (das eine Objekt weist das Merkmal auf und das andere nicht) (vgl. Herrmann et al. 2000,

⁷Vgl. http://www.quantitative-methoden.de/Dateien/Auflage3/leseprobe_A3.pdf

S347).

Die darauf aufbauend verschiedenen Koeffizienten sind folgendes:

- M-Koeffizient: $D = (a + d)/n$, wobei n die Anzahl der Merkmale repräsentiert
- Gewichteter M-Koeffizient: $D = \frac{\lambda(a+d)}{\lambda(a+d)+(1-\lambda)*(b+c)}$, wobei $0 < \lambda < 1$
- S-Koeffizient: $D=a/(a+b+c)$

Schritt 2: Auswahl der Fusionierungsalgorithmus

„Aufgrund der Ähnlichkeits- bzw. Distanzwerte werden die Fälle so zu Gruppen zusammengefasst, dass sich diejenigen Objekte oder Personen mit weitgehend übereinstimmend ausgeprägten Beschreibungsmerkmalen in einer Gruppe wiederfinden. Entsprechend der Vorschriften des Fusionierungsalgorithmus fasst die (agglomerative) Clusteranalyse die betrachteten Fälle solange zusammen, bis am Ende alle Fälle in einer einzigen Gruppe enthalten sind.“ (Backhaus et al. 2008, S398)

SPSS bietet insgesamt sieben verschiedene Fusionierungsmethoden an, nämlich Linkage zwischen den Gruppen, Linkage innerhalb der Gruppen, Nächstgelegener Nachbar, Entferntester Nachbar, Zentroid-Clustering, Median-Clustering und Ward-Methode. Da „die Ward-Methode zumindest für Ähnlichkeitsmaße, die sich als euklidische Distanzen interpretieren lassen, die besten Resultate erzielt“ (Bortz et al.2010, S460) und „im Vergleich zu anderen Algorithmen in den meisten Fällen sehr gute Partitionen findet und die Elemente 'richtig' den Gruppen zuordnet“ (Backhaus et al. 2008, S 430), wird sie im folgend ausführlich erzählt.

Ward-Verfahren

Das Verfahren von Ward (alternative Bezeichnungen: minimum variance method, error sum of squares method) ist eines der am weitesten verbreiteten Clusterverfahren. Es fasst die Gruppe so zusammen, dass die Varianz in den Gruppen möglichst gering ist, so dass die Gruppen intern möglichst homogen sind. Die Varianz wird dabei wie folgt berechnet:

$$V_g = \sum_{k=1}^{K_g} \sum_{j=1}^J (X_{kjg} - \bar{X}_{jg})^2$$

mit

X_{kjg} : Beobachtungswert der Variablen $j=1 \dots J$ bei Objekt k , für alle Objekte in $k=1 \dots K$ in Gruppe g

\bar{X}_{jk} : Mittelwert über die Beobachtungswerte der Variablen j in Gruppe g

Wie die anderen agglomerativen Verfahren beginnt auch Ward mit einer Gruppenzahl, die der Anzahl der zugrunde liegenden Objekte entspricht und fusioniert diese so lange zu größeren Gruppen, bis zum Schluss alle Objekte in einer Gruppe sind. Die Varianz der Gruppe ist zu Beginn des Verfahrens, wenn noch keine Gruppenbildung erfolgt ist, für jede "Gruppe" 0 und vergrößert sich dann zwangsläufig im Laufe des Verfahrens. Dabei strebt das Verfahren an, die Objekte so zusammenzufassen, dass die Varianz der neu gebildeten Gruppen möglichst klein ist. Hierzu ist es notwendig, hypothetische Gruppen zu bilden. Ausgehend von den bestehenden Gruppen werden alle Kombinationen zwei Gruppen zusammenzufassen durchgespielt. Für jede dieser hypothetischen Gruppen wird dann die Varianz ermittelt und diejenige Gruppe, die zur geringsten Varianz führt, gebildet. Im nächsten Schritt werden dann, ausgehend von den neuen Gruppen, wieder hypothetische Gruppen gebildet und die Kombination mit der dann geringsten Varianz verwirklicht, bis zum Schluss nur noch eine einzige Gruppe besteht, die alle Elemente enthält. Im Unterschied zu den vorher dargestellten Verfahren funktioniert das Verfahren von Ward nur mit Distanzmaßen und nicht mit Ähnlichkeitsmaßen (vgl. Herrmann et al. 2000, S358-S359).

Schritt 3: Bestimmung der optimalen Clusterzahl

"Anschließend ist zu entscheiden, welche Anzahl an Clustern die "beste" Lösung darstellt und im Ergebnis verwendet werden soll. Hier gilt es vor allem den Zielkonflikt zwischen Handhabbarkeit (geringe Clusterzahl) und Homogenitätsanforderung (große Clusterzahl) zu lösen." (Backhaus et al. 2008, S399)

Da in SPSS bisher keine Kriterien zur Bestimmung der optimalen Clusterzahl verfügbar sind, wird empfohlen ggf. auf alternative Programme wie S-Plus, R oder SAS zurückzugreifen, das ist aber hier unmöglich.

3.3 Korrespondenzanalyse

Die Korrespondenzanalyse ist als Strukturen-entdeckendes Verfahren einzuordnen und damit verwandt mit der Faktorenanalyse und der multidimensionalen Skalierung. Aufgrund ihrer Ähnlichkeit mit der Faktorenanalyse bzw. deren spezieller Form der Hauptkomponentenanalyse wird die Korrespondenzanalyse auch als „Hauptkomponentenanalyse mit kategorialen Daten“ bezeichnet. (vgl. Backhaus et al. 2008, S550)

Die Besonderheit der Korrespondenzanalyse unter den multivariaten Analyseverfahren liegt somit darin, dass sie die Spalten- und Zeilenstruktur einer Datenmatrix gleichzeitig untersuchen kann. Die Spalten- und Zeilenprofile der Matrix werden als Punkte in einem mehrdimensionalen Raum abgebildet. Um eine genau definierte und beabsichtigte grafische Interpretation der Datenstruktur vornehmen zu können, wird die Punktwolke auf einen zwei- oder dreidimensionalen Raum projiziert (vgl. Herrmann et al. 2000, S515). Die Objekte oder Merkmale werden von den Befragten als ähnlich wahrgenommen, wenn die betreffenden Punkte relativ dicht beieinander liegen und dagegen können die isolierten Positionen eine Unähnlichkeit hinweisen.

a) Vorteile und Nachteile

Vorteile ⁸

- Ein großer Vorteil besonders gegenüber der Faktorenanalyse besteht darin, dass hier kein bestimmtes Skalenniveau erforderlich ist und somit in der Praxis eine hohe Anwendungsbreite findet.

⁸Es gibt so viele Meinungen über Vor- und Nachteile in der Literatur. Hier werden Gedanken ausgewählt aus <http://mars.wiwi.hu-berlin.de/mediawiki/teachwiki/index.php/Korrespondenzanalyse>

- Weiterhin gibt es keine bestimmten Verteilungsannahmen.
- Gegenüber der Faktorenanalyse hat die Korrespondenzanalyse zudem den Vorteil, dass sie sowohl Objekt- als auch Merkmalsraum gleichzeitig abbilden kann.
- Für die Analyse spielt es keine Rolle, welche Elemente in den Zeilen und welche in den Spalten präsentiert werden. Eine Voraussetzung ist lediglich, dass die Daten nicht negativ sind.
- Ebenfalls für die Korrespondenzanalyse spricht, dass mit Hilfe dieses Tools qualitative Daten quantifiziert werden können.

Nachteile

- Problematisch bei der Korrespondenzanalyse ist die Schwierigkeit der Interpretierbarkeit, wodurch auch die Verbreitung dieser Analyse in der Praxis noch eher weniger gegeben ist. Durch die Schwierigkeit ist auch die Gefahr von Fehlinterpretation größer als bei vergleichbaren Verfahren, wie zum Beispiel der Faktorenanalyse.
- Zudem existiert eine Vielfalt von Variationen der Korrespondenzanalyse.

b) Vorgehensweise (vgl. Backhaus et al. 2008, S551-S554)

Schritt 1: Standardisierung der Daten

Die gemeinsamen Häufigkeiten einer (I*J)-Kreuztabelle (mit I Zeilen J Spalten) seien durch n_{ij} bezeichnet: n_{ij} = Häufigkeit der Merkmalskombination (i,j). Zuerst werden die beobachteten und erwarteten relativen Häufigkeiten bestimmt:

$$p_{i\cdot} = n_{i\cdot} / n \quad \text{mit} \quad n = \sum_i \sum_j n_{ij}$$

$$\text{und} \quad \hat{e}_{ij} = n_{i\cdot} * n_{\cdot j} \quad \text{mit} \quad n_{i\cdot} = \sum_{j=1}^J n_{ij} \quad \text{und} \quad n_{\cdot j} = \sum_{i=1}^I n_{ij}$$

In den Abweichungen der beobachteten Häufigkeiten von den erwarteten Häufigkeiten $p_{ij} - \hat{e}_{ij}$ zeigt sich der Informationsgehalt der Daten. Weichen die beobachteten

Werte nicht oder nur wenig von den erwarteten Werten ab, so erhalten sie auch keine oder nur wenig Information, denn sie können dann auch das marginalen Häufigkeiten der Kreuztabelle berechnet werden.

Dann sollen die relativen Häufigkeiten wie folgt standardisiert:

$$z_{ij} = \frac{p_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij}}}$$

Die Chi-Quadrat-Statistik $\chi^2 = n * \sum_{i,j} z_{ij}^2$ misst die Streuung der beobachteten Werte um die erwarteten Werte und kann somit als Maß für die in einer Kreuztabelle enthaltene Streuung oder Information angesehen werden. Aber ein Nachteil von Chi-Quadrat als Maß für die Streuung ist, dass es von der Höhe der Fallzahl der Daten abhängig ist, d.h. man erhält auch hohe Werte für Chi-Quadrat bei Daten mit niedriger Streuung, wenn nur die Fallzahl hinreichend groß ist. In der Kontingenzanalyse wird daher das durch die Fallzahl dividierte Chi-Quadrat verwendet. Es wird als Inertia (Trägheit) einer Kreuztabelle bezeichnet:

$$T = \chi^2/n = \sum_i \sum_j z_{ij}^2$$

Schritt 2: Extraktion der Dimensionen

Das Ziel des Schritts ist die Darstellung der Zeilen- und Spaltenkreuztabelle in einem gemeinsamen Raum (Korrespondenzraum) mit möglichst geringer Dimensionalität und gleichzeitig mit möglichst weniger Informationsverlust. Die maximale Anzahl der Dimensionen für den Korrespondenzraum ist K mit $K = \min(I, J) - 1$ und entspricht somit der maximalen Inertia.

Zur Gewinnung der Dimensionen bzw. der Koordinaten der Zeilen- und Spaltenelemente wird die Matrix mit den standardisierten Daten z_{ij} einer Singulärwertzerlegung unterzogen. Diese lässt sich in Matrixenschreibweise wie folgt darstellen:

$$Z = U * S * V'$$

Dabei bedeuten:

$Z = (z_{ij})$ einer (I x J) Matrix der standardisierten Daten

$U = (u_{ik})$ einer $(I \times K)$ Matrix der Zeilenelemente

$S = (s_{kk})$ einer $(K \times K)$ Diagonalmatrix mit den Singularwerten

$V = (v_{jk})$ einer $(J \times K)$ Matrix der Spaltenelemente

Die quadrierten Singulärwerte sind sog. Eigenwerte (Trägheitsgewichte) und liefern ein Maß für die Streuung (Information), die eine Dimension aufnimmt oder repräsentiert. Die Eigenwerte summieren sich zur Inertia: $T = \sum_k S_k^2$ und der Anteil des quadrierten Singulärwertes einer Dimension an der Inertia ergibt deren Eigenwertanteil: $EA_k = S_k^2/T$

Die erste Dimension nimmt einen maximalen Anteil der in den Daten vorhandenen Streuung (Information) auf und die zweite Dimension nimmt einen maximalen Anteil der noch verbleibenden Streuung auf, usw. „Um die 'günstigste' Anzahl der Dimensionen für die grafische Darstellung festzulegen, sind sogenannte Interpretationsregel heranzuziehen. Zuerst werden, wie auch bei der Hauptkomponentenanalyse, nur so viele Achsen extrahiert, bis dass durch Hinzunahme einer weiteren kein deutlicher Informationsgewinn mehr erzielt wird. Zweitens sollen aber so viele Achsen ausgewählt werden, dass insgesamt mindestens 80% der Gesamtvarianz erklärt werden“ (Herrmann et al. 2000, S532).

Schritt 3: Normalisierung der Koordinaten

Um aus den Matrizen U und V die endgültigen Koordinaten zu gewinnen, die die grafische Darstellung der Zeilen- und Spaltenelemente in einem gemeinsamen Korrespondenzraum (Biplot) ermöglichen, müssen diese noch normalisiert (reskaliert) werden. Dabei werden bei symmetrischer Normalisierung (klassische Form) die Zeilen- und Spaltenpunkte wie folgt bestimmt:

- Zeilenpunkte $r_{ik} = \frac{u_{ik} * \sqrt{s_k}}{\sqrt{p_i}}$
- Spaltenpunkte $c_{jk} = \frac{v_{jk} * \sqrt{s_k}}{\sqrt{p_j}}$

Daneben existieren weitere Formen der Normalisierung, wie z.B. die Prinzipal-Normalisierung

sowie zwei asymmetrische Formen, die Zeilen-Prinzipal- und die Spalten-Prinzipal-Normalisierung.

c) Interpretation

Zur inhaltlichen Interpretation der Dimension kann man sich an den Positionen der Zeilen- und Spaltenelemente im Korrespondenzraum orientieren. Im Koordinatensprung der graphischen Darstellung liegen alle Durchschnittsprofile, d.h. je weiter ein Profil vom Koordinatensprung entfernt liegt, desto stärker unterscheidet es sich vom Durchschnittsprofil. Darüber hinaus lässt sich bei der graphischen Darstellung in der Korrespondenzanalyse erkennen, wie ähnlich oder unähnlich verschiedene Objekte wahrgenommen werden. Je näher zwei Objekte beieinander liegen, desto ähnlicher werden diese wahrgenommen. Zudem werden die Merkmale selbst im Koordinatensystem dargestellt. Die Distanz zwischen einem Merkmal und einem Objekt birgt eine große Gefahr von Missinterpretationen, da hier in der Praxis fälschlicherweise oft die gleiche Interpretation angeführt wird wie bei den Distanzen zwischen einzelnen Objekten.⁹ Die konkrete empirische Interpretation wird in 4.4 ausführlich präsentiert.

⁹Vgl.<http://mars.wiwi.hu-berlin.de/mediawiki/teachwiki/index.php/>
Korrespondenzanalyse

4 Datenanalyse

4.1 Mittelwertvergleiche mit Hypothesentest

Das Ziel des Tests war es, zu beweisen, dass Mittelwerte der Variablen bezüglich des Geschlechts sowie der Muttersprache gleich sind; da eigentlich die beiden Merkmale keine große Rolle gespielt haben, wird Nullhypothese in der Regel nicht verworfen und dann wird die Reliabilität der Daten zu gewissem Grad überprüft.

4.1.1 t-Test bezüglich des Merkmals *Geschlecht*

Ob die Befragten die Themengebiete kennen, sollte nicht von ihrem Geschlecht beeinflusst werden, d.h., die Meinungen der weiblichen Befragten sind nicht stark unterschiedlich von denen der männlichen. Um diese Aussage zu stützen, wurde ein t-Test mit $H_0 : \mu_w = \mu_m$ & $H_1 : \mu_w \neq \mu_m$ in einem Konfidenzintervall von 95% eingeführt. Es war offensichtlich, dass nicht alle drei in 3.1.1 genannten Voraussetzungen erfüllt wurden, jedoch ist die Präzision des t-Tests nicht beeinträchtigt, solange die Varianzen gleich sind. Der Levene-Test ist ein Test für Varianzhomogenität. Üblicherweise verwirft man die Gleichheit der Varianzen, wenn der Levene-Test ein $p < 0,05$ ergibt, dabei handelte es sich hier um *Kombinatorik*, *deskriptive Statistik*, *Verteilungsmodelle* und *Schätztheorie* (siehe Ausbildung). In SPSS werden die Ergebnisse des Mittelwertvergleichs sowohl für den homogenen als auch für den heterogenen Fall ausgegeben. Wird die Irrtumswahrscheinlichkeit p in Betracht gezogen, sind alle ermittelten Wahrscheinlichkeiten des t-Wertes größer als 0,05%, das bedeutet, dass die empirische Mittelwertdifferenz nicht als signifikant bezeichnet wird.

Tabelle 4.1 zeigt die Auswertung der ersten zwei Variablen *Mengenlehre* und *Zufallsexperimente*. Die Abbildungen für die übrigen Variablen sind in Anhang A.3 dargestellt.

		Levene-Test der Varianzgleichheit		t-Test für Mittelwertvergleiche		
		F	Signifikanz	T	df	Signifikanz
Mengenlehre kennen	Varianzen sind gleich	0,536	0,467	0,534	55	0,596
	Varianzen sind nicht gleich			0,524	46,087	0,603
Zufallsexperimente kennen	Varianzen sind gleich	3,363	0,072	-0,911	55	0,366
	Varianzen sind nicht gleich			-0,854	36,941	0,398

Tabelle 4.1: SPSS-Output des t-Tests mit der Variable Geschlecht

4.1.2 t-Test bezüglich des Merkmals *Muttersprache*

Bei der Variablen Muttersprache wurde genauso verfahren wie bei der Variablen Geschlecht. Da es unter den Befragten mehrere Ausländer gab, stellte sich bei den Fragen, ob sie die Fachbegriffe kennen oder ob sie die Begriffe auf Deutsch kennen. Nach der Durchführung des t-Tests mit $H_0 : \mu_d = \mu_{nd}$ & $H_1 : \mu_d \neq \mu_{nd}$ wurde beschlossen, die Nullhypothesen auf jeden Fall nicht abzulehnen (siehe Tabelle 4.2). Die deutschen und nicht-deutschen Studenten hatten den gleichen Kenntnisstand, deshalb war zu vermuten, dass sie einen ähnlichen Ausbildungshintergrund besaßen, was zur Plausibilität der Untersuchung beiträgt.

		Levene-Test der Varianzgleichheit		t-Test für Mittelwertvergleiche		
		F	Signifikanz	T	df	Signifikanz
Mengenlehre kennen	Varianzen sind gleich	10,808	0,002	-1,773	55	0,82
	Varianzen sind nicht gleich			-1,472	22,567	0,155
Zufallsexperimente kennen	Varianzen sind gleich	8,850	0,004	-1,486	55	0,143
	Varianzen sind nicht gleich			-1,198	21,522	0,244

Tabelle 4.2: SPSS-Output des t-Tests mit der Variable Muttersprache

4.2 Häufigkeitsverteilung

4.2.1 Häufigkeitsverteilung bezüglich der statistischen Themen

Der dritte Komplex des Fragebogens bestand aus zwei Fragenteilen. In diesem Abschnitt sind die Häufigkeitsverteilungen der beiden Teile zu diskutieren.

Fragenteil 1	Wurde behandelt und kenne ich	Wurde nicht behandelt	Kenne ich nicht
Mengenlehre	84,2%	10,5%	5,3%
Zufallsexperimente	91,2%	3,5%	5,3%
Kombinatorik	91,2%	1,8%	7,0%
Wahrscheinlichkeitsrechnung	94,7%	1,8%	3,5%
Deskriptive Statistik	93,0%	0%	7,0%
Zufallsvariablen	94,7%	0%	5,3%
Zweidimensionale Verteilung	93,0%	0%	7,0%
Verteilungsmodell	94,7%	1,8%	3,5%
Zufallsstichprobe	86,0%	3,5%	10,5%
Schätztheorie	89,5%	3,5%	7,0%
Statistisches Testverfahren	93,0%	1,8%	5,3%
Regressionsanalyse	91,2%	3,5%	5,3%
Zeitreihenanalyse	80,7%	12,3%	7,0%

Tabelle 4.3: Relative Häufigkeit der Variablen in Fragenteil 1

Tabelle 4.3 zeigt, dass die meisten Begriffe mit einer Prozentzahl von größer als 90 sehr gut bekannt waren und behandelt wurden, wobei *Wahrscheinlichkeitsrechnung*,

Zufallsvariablen und *Verteilungsmodelle* zusammen auf dem ersten Platz standen. 10,5% der Befragten wiesen den Begriff *Mengenlehre* als „nicht behandelt“ aus, was nicht erstaunlich war, da dieser meist schon im Gymnasium und auch im Kurs „Mathematik I“ behandelt wird und deshalb nicht als neu empfunden wurde. Das Gleiche passierte auch mit dem Begriff *Zeitreihenanalyse*, bei dem 12,3% der Befragten der Meinung waren, dass er in der Vorlesung nicht vorkam. Die Ursache dafür ist wahrscheinlich, dass *Zeitreihenanalyse* wegen des Zeitproblems als letztes Kapitel der Vorlesung etwas vernachlässigt wurde. Der Begriff *Zufallsstichprobe* war am wenigstens bekannt, seine Unbekanntheitsquote erreichte 10,5. Bis jetzt können noch keine Vorschläge gemacht werden, welche statistischen Begriffe am wenigstens oder höchstens signifikant sind, da Fragenteil 2 auch ein wichtiger Indikator war und die Kombination der beiden Ergebnisse mehr Sinn hatte.

Nach der multiplen Imputation waren fehlende Werte in Fragenteil 2 großartig reduziert.¹⁰ Außergewöhnlich in Tabelle 4.4 ist, dass *Zufallsstichprobe* von 24,56% der Befragten und *Zeitreihenanalyse* von 19,30% der Befragten nicht eingeschätzt wurden, die deutlich höher als andere Variable waren, aber gerade mit dem Ergebnis von Fragenteil 1 übereinstimmten. Um zu untersuchen, welche Themengebiete verbleiben und welche verkürzt oder sogar gestrichen werden sollten, wurden zwei Kriterien angenommen: Prozentzahlen in erster Spalte > 20% oder Prozentzahlen in zweiter Spalte < 60%, das bedeutet: Variablen kamen zur Diskussion, die von mehr als 20% der Befragten für unnötig für andere Veranstaltungen gehalten oder von weniger als 60% der Befragten als notwendig angesehen wurden. Alle anderen Variable, nämlich *Zufallsexperimente*, *Deskriptive Statistik*, *Zufallsvariable* und *Verteilungsmodelle*, sollten auf jeden Fall in der Vorlesung bleiben. Dann war die Gruppe von *Wahrscheinlichkeitsrechnung*, *zweidimensionaler Verteilung* und *Regressionsanalyse* zu betrachten, deren erste Spalte 21,05%, aber deren zweite Spalte mehr als 60% war. Hier wurde mit großer Wahrscheinlichkeit beschlossen, solche Themengebiete weiter zu behandeln. Aber um sichere und genauere Aussagen zu treffen, sollten Vorschläge von relevanten Lehrpersonen eingeholt werden, dann wird eine Veränderung der Lehre aussagekräftiger. Die gleiche Entscheidung wurde auch bei *Mengenlehre*

¹⁰Die Reduzierung der fehlenden Werte wurde in 2.2 c) erklärt.

getroffen, bei der 22,81% der Studenten der Meinung waren, dass man sie in anderen Veranstaltungen nicht benötigt und 61,40% hatten eine gegenteilige Ansicht. Die Situation von *Kombinatorik* war eher schlechter, weil die Unbrauchbarkeitsquote deutlich höher war und die Nützlichkeitsquote auf weniger als 60% gesunken ist. Für die Entscheidung kann eine Nachforschung sehr hilfreich sein. Da die Nützlichkeitsquote von *Zufallsstichprobe* und *Zeitreihenanalyse* sehr niedrig war, wäre es möglich, ihre Inhalte zu kürzen. Aber die gemeinsame Eigenschaft von beiden war der signifikante Einfluss der fehlenden Werte, die das Ergebnis verfälschen können, demnach ist es auch schwer zu entscheiden, ob die Themen weiter in der Vorlesung behandelt werden sollen. Es lohnt sich, eine weitere Untersuchung durchzuführen und sich dabei an professionelle Personen zu wenden. Die letzte zu diskutierende Variable war statistisches Testverfahren. Hier gab es fast keinen Zweifel, das Thema zu streichen. 38,60% der Befragten fanden es unnötig, während nur 49,12% der Befragten nicht zu stimmten und fehlende Werte haben keine große Rolle gespielt.

Fragenteil 2	Wurde in keiner Veranstaltung benötigt	Wurde in einer oder mehreren benötigt	Kenne ich nicht	Weiß ich nicht
Mengenlehre	22,81%	61,40%	12,28%	3,51%
Zufallsexperimente	15,79%	64,91%	10,53%	8,77%
Kombinatorik	26,32%	57,89%	10,53%	5,26%
Wahrscheinlichkeitsrechnung	21,05%	68,42%	7,02%	3,51%
Deskriptive Statistik	15,79%	73,68%	7,02%	3,51%
Zufallsvariablen	19,30%	71,93%	5,26%	3,51%
Zweidimensionale Verteilung	21,05%	64,91%	10,53%	3,51%
Verteilungsmodell	19,30%	70,18%	7,02%	3,51%
Zufallsstichprobe	21,05%	45,61%	24,56%	8,77%
Schätztheorie	15,79%	57,89%	14,04%	12,28%
Statistisches Testverfahren	38,60%	49,12%	8,77%	3,51%
Regressionsanalyse	21,05%	66,67%	8,77%	3,51%
Zeitreihenanalyse	22,81%	52,63%	19,30%	5,26%

Tabelle 4.4: Relative Häufigkeit der Variablen in Fragenteil 2

4.2.2 Häufigkeitsverteilung bezüglich der Vertiefungsgebiete

Im Komplex 2 des Fragebogens wurden die Studenten gefragt, welche Vertiefungsgebiete sie belegt haben. Die Datenerfassung (Tabelle 4.5) ergab, dass *Externes Rechnungswesen*, *Internationales Management* und *Angewandte Mikroökonomie* am beliebtesten waren, sie wurden mehr als zehn Mal gewählt. Darauf folgten *Bank- und Börsenwesen*, *Marketing*, *Arbeitsmarktökonomie*, *Spieltheorie*, *Makroökonomie III* und *Versicherungs- und Risikomanagement*. Die Themengebiete *Angewandte Ökonometrie*, *Applied Statistics* und *Operations Research* wurden hier von sehr wenigen Studenten belegt, trotzdem besetzten sie einen großen Antwortteil von Komplex 3 in der Frage „Zu welchem Themengebiet wurde das statistische Wissen benötigt?“ Eigentlich ist das kein Widerspruch, da, wenn fehlende Werte von Komplex 2 zu betrachten waren, gefunden wurde, dass insgesamt 33 Befragte keine Antwort gegeben hatten. Es ist sehr wahrscheinlich, dass diese Befragten in Komplex 3 *Angewandte Ökonometrie*, *Applied Statistics* und *Operations Research* bevorzugt haben.

Themengebiete	Häufigkeit	Themengebiete	Häufigkeit	Themengebiete	Häufigkeit
Angewandte Makroökonomie	5	Externes Rechnungswesen	11	Öffentliche Finanzen in der Demokratie	5
Angewandte Mikroökonomie	10	Finanzwirtschaft	6	Operations Research	1
Angewandte Ökonometrie	2	Internationales Management	11	Spieltheorie	6
Applied Statistics	3	Internes Rechnungswesen	4	Theorie der Marktversagens	2
Arbeitsmarktökonomie	7	Makroökonomie III	5	Versicherungs- und Risikomanagement	5
Bank- und Börsenwesen	9	Marketing	8	Wettbewerbsstrategie	2

Betriebswirtschaftliche Steuerlehre	2	Mikroökonomie III	4	Wirtschaftsinformatik	4
Entrepreneurship	3	Monetary Economics	4	Andere	1

Tabelle 4.5: Häufigkeit der Themengebiete im Komplex 2

Tabelle 4.6 zeigt die Anwendungshäufigkeit des Begriffs *Mengenlehre* in anderen Veranstaltungen. Die Tabellen zu den übrigen zwölf Begriffen befinden sich in Anhang A.5. Fünf Studenten hielten *Mengenlehre* für wichtig für *Angewandte Ökonometrie*, während jeweils drei Studenten sie für wichtig für *Applied Statistics* und *Spieltheorie* hielten. Das auf dem ersten Platz stehende Andere könnte sich in anderen wirtschaftlichen Fächern oder Sozialwissenschaft befinden.

Mengenlehre	Häufigkeit
Andere	6
Angewandte Ökonometrie	5
Applied Statistics	3
Spieltheorie	3
Finanzwirtschaft	2
Angewandte Mikroökonomie	2
Mikroökonomie III	2
Externes Rechnungswesen	1
Marketing	1
Arbeitsmarktökonomie	1
Operations Research	1
Angewandte Makroökonomie	1
SUMME:	28

Tabelle 4.6: Anwendungshäufigkeit der Mengenlehre bei anderen Veranstaltungen

Wenn alle Begriffe zusammen berücksichtigt wurden, ergaben sich zwei bemerkenswerte Punkte. Erstens die Reihenfolge: Andere, *Angewandte Ökonometrie*, *Applied Statistics*, *Marketing* und *Spieltheorie* standen am meisten auf den ersten fünf Plätzen, woraus zu schließen war, dass Statistik-Vorlesungen in enger Verbindung mit solchen Themen stehen, besonders *Spieltheorie* und *Marketing*, die nicht zum Modul „Quantitative Methode“ zählen. Zweitens die Summen der Bewertungshäufigkeit (Tabelle 4.7): Der Indikator reflektiert mehr oder weniger auch die Wichtigkeit der statistischen Inhalte. Je weniger Bewertungen ein Begriff erhielt, desto unwichtiger war er, weil die Studenten ihn einem Fach zuordnen konnten. Die ersten vier kleinsten waren *Kombinatorik*, *Zufallsstichprobe*, *Zeitreihenanalyse* und *Statistisches Testverfahren*, die mit dem Ergebnis in 4.4 übereinstimmen. Dies kann weiter beweisen, dass die vier Begriffe infrage stehen und eine hohe Berücksichtigung in der Lehre finden sollten.

Statistische Begriffe	Summe
Kombinatorik	19
Zufallsstichprobe	19
Zeitreihenanalyse	25
Statistisches Testverfahren	27
Mengenlehre	28
Zufallsexperimente	30
Schätztheorie	39
Zweidimensionale Verteilung	40
Regressionsanalyse	44
Verteilungsmodelle	53
Zufallsvariablen	54
Wahrscheinlichkeitsrechnung	57
Deskriptive Statistik	58

Tabelle 4.7: Summe der Male der Bewertung jeder Variable

4.3 Cluster- und Korrespondenzanalyse des ersten Versuchs

Es wurden zweimal Analysen ausgeführt, beim ersten Mal wurden acht Gruppen von Vertiefungsgebieten während der Clusteranalyse gebildet und beim zweiten Mal gab es nur vier Gruppen. Der Vorteil des ersten Versuchs war der möglichst kleine Informationsverlust, aber die große Clusteranzahl konnte auch die Handhabbarkeit erschweren. Deshalb war der zweite Versuch der Analyse notwendig, wenn das erste Ergebnis nicht ideal war.

a) Clusteranalyse

Sowohl die statistischen Inhalte als auch die Themengebiete waren in den vorherigen Analysen separat, das bedeutet, dass jede Variable einzeln betrachtet wurde und ihr Zusammenhang bzw. ihre Korrelation noch nicht heranzuziehen waren. Anschließend war zu untersuchen, welche statistischen Begriffe hinsichtlich der Vertiefungsgebiete näher beieinander liegen, wobei das im empirischen Sinn zur Reihenfolge und Kombination der gelehrten Begriffe beitragen kann. Weil die Anzahl der Vertiefungsgebiete relativ groß war (24 insgesamt), war zuerst eine Klassifizierung hilfreich, um die weitere Untersuchung zu vereinfachen. Eigentlich war die Gruppenbildung von höherer Bedeutung, weil es auch eine interessante Frage für Lehrpersonen ist, welche Sachgebiete am häufigsten zusammen als Schwerpunkt belegt werden. Dafür handelt es sich um die Clusteranalyse, womit die Sachgebiete im Komplex 2 nach Maßgabe ihrer Ähnlichkeit in Gruppen (Cluster) eingeteilt wurden, wobei die Cluster intern möglichst homogen und extern möglichst gut voneinander separierbar sein sollten.

Vorgehensweise und Ergebnis

Mithilfe von SPSS wurden alle Vertiefungsgebiete als Variable gesetzt und die Ward-Methode war hier als Fusionierungsverfahren geeignet. Nach der üblichen Fallzahlstatistik wurde zunächst eine Zuordnungsübersicht ausgegeben (Tabelle 4.8), dieser konnten dann die Reihenfolge der Clusterbildung und die optimale Clusteranzahl entnommen werden. Die unter *zusammengeführte Cluster* angeordneten beiden Spalten sagten aus, dass im ersten Schritt die beiden Fälle 20 und 24 vereinigt werden

(also *Theorie des Marktversagens* und *Andere*); diese beiden Fächer hatten also die größte Ähnlichkeit bzw. den geringsten Abstand voneinander. Diese beiden Fälle bildeten daraufhin das Cluster 20, während die Nummer 24 im weiteren Verlauf der Zuordnungsübersicht nicht wieder auftrat. Im nächsten Schritt wurden 3 und 18 vereinigt (*Angewandte Ökonometrie* und *Operations Research*), danach 13 und 16 (*Makroökonomie III* und *Monetary Economics*) usw. Als Heterogenitätsmaß diente dem Ward-Verfahren dabei die Fehlerquadratsumme, deren Entwicklung in der Spalte „Koeffizienten“ aufgezeigt war. Hier war jedoch der Sprung, mit dem die optimale Clusteranzahl festzulegen ist, schwer zu identifizieren. Deswegen wurde noch das Dendrogramm (Abbildung 4.1) ausgegeben, welches eine Visualisierung des in der Zuordnungsübersicht wiedergegebenen Fusionierungsablaufs darstellte. Es identifizierte die jeweils zusammengefassten Cluster und die Werte des Koeffizienten bei jedem Schritt.¹¹ Nach zahlreichen Untersuchungen und ausführlichen Vergleichen wurde die Clusterzahl mit acht festgelegt.

Schritt	Zusammengeführte Cluster		Koeffizient	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	20	24	.500	0	0	6
2	3	18	1.000	0	0	9
3	13	16	1.500	0	0	8
4	22	23	2.500	0	0	5
5	12	22	3.500	0	4	10
6	4	20	5.000	0	1	13
7	15	17	6.500	0	0	13
8	5	13	8.000	0	3	14
9	3	8	9.500	2	0	12
10	7	12	11.500	0	5	18
11	10	21	14.000	0	0	19
12	1	3	16.750	0	9	16
13	4	15	19.650	6	7	20
14	2	5	22.650	0	8	17
15	11	14	26.150	0	0	22

¹¹Vgl. Achim Bühl: SPSS 14, Einführung in die moderne Datenanalyse, 2006 (Kapitel 20: Clusteranalyse)

16	1	19	30.200	12	0	18
17	2	6	34.400	14	0	21
18	1	7	38.711	16	10	19
19	1	10	43.645	18	11	20
20	1	4	50.763	19	13	22
21	2	9	57.896	17	0	23
22	1	11	68.111	20	15	23
23	1	2	85.333	22	21	0

Tabelle 4.8: Zuordnungsübersicht der Clusteranalyse (SPSS-Output)

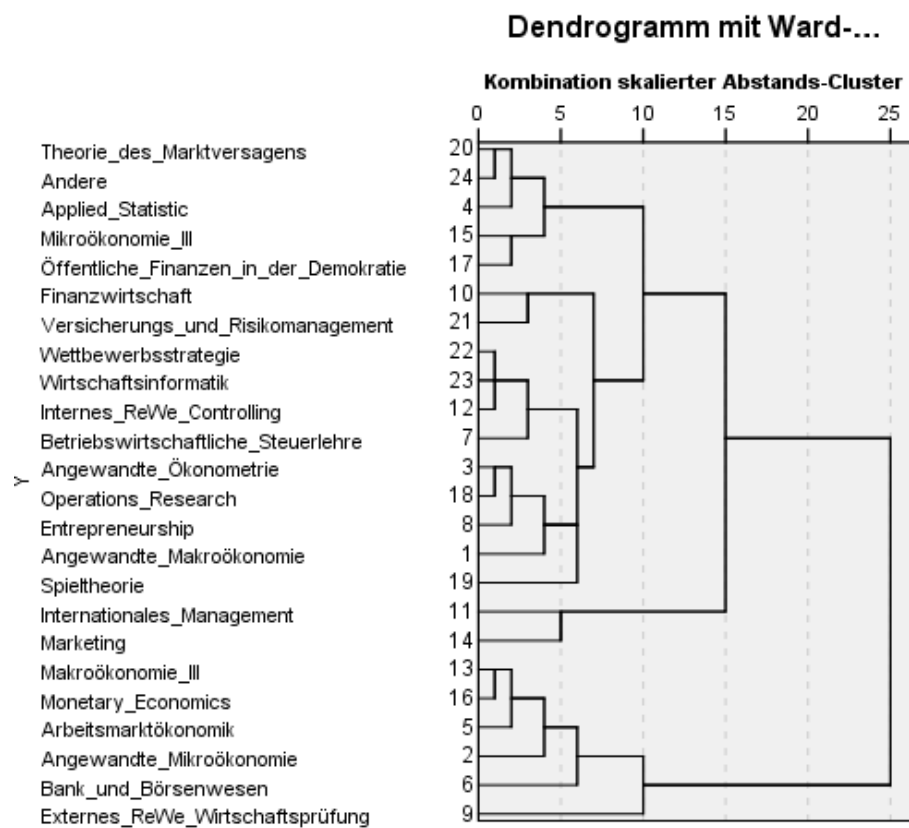


Abbildung 4.1: Dendrogramm mit Ward-Verfahren

Tabelle 4.9 zeigt die Verteilung der Cluster. Die Fächer von BWL, VWL und Quantitative Methoden standen in jedem Cluster gemischt, weil die Verteilung auf den von

den Studenten gegebenen Antworten basierte und nicht auf dem Studiengang. Dabei handelt es sich besonders um Fachgebiet 1, dessen fünf Komponenten aus verschiedenen Studiengängen kamen. Fachgebiet 2 und 8 waren verhältnismäßig einheitlich, im Gegensatz zu Fachgebiet 3 und 5. Es ergab einen Sinn, dass *Management* und *Marketing* (Fachgebiet 8) in enger Beziehung standen und *Applied Statistics* als Instrument für andere VWL-Fächern (Fachgebiet 3) dienten. Außerdem waren Fachgebiet 4, 6 und 7 relativ klein, obwohl 4 und 7 in Bezug zu *Finance* standen, wurden sie nicht in einem Cluster fusioniert. Die Verteilung der Cluster hatte inhaltlich wenig Sinn, die Gründe dafür waren, dass einerseits die kleine Stichprobe eine Verzerrung verursachte und andererseits die Studenten ihre Schwerpunkte nicht sehr fachlich und wenig gezielt erstellt hatten. Erfahrungsgemäß belegen Studenten Veranstaltungen nicht nur aus Interesse oder fachlichen Gründen, sondern auch, weil zum Beispiel bestimmte Professoren beliebt sind und manche Klausuren einfacher zu bestehen sind als andere.

Als Ergänzung steht im Anhang das Histogramm, wobei die Anteile des Fachgebiets jedes Begriffs beschrieben werden.

Fachgebiet 1	Angewandte Makroökonomie	Angewandte Ökonometrie	Entrepreneur- ship	Operations Research	Spiel- theorie
Fachgebiet 2	Angewandte Mikroökonomie	Arbeitsmarkt- ökonomie	Makroökonomie III	Monetary Economics	
Fachgebiet 3	Öffentliche Finanzen in der Demokratie	Theorie des Markt- versagens	Applied Statistics	Mikroökonomie III	Andere
Fachgebiet 4	Bank- und Börsenwesen				
Fachgebiet 5	Betriebs- wirtschaftliche Steuerlehre	Internes Rechnungs- wesen	Wettbewerbs- strategie	Wirtschafts- informatik	
Fachgebiet 6	Externes Rechnungs- wesen				
Fachgebiet 7	Versicherungs- und Risiko- management	Finanz- wirtschaft			

Fachgebiet 8	Inter- nationales Management	Marketing
--------------	------------------------------------	-----------

Tabelle 4.9: Verteilung der Lehrveranstaltung in acht Fachgebieten

b) Korrespondenzanalyse

Das Ziel der Korrespondenzanalyse bezüglich der acht Cluster der Vertiefungsgebiete und der 13 statistischen Begriffe war es, anhand des mehrdimensionalen Raumes zu erkennen, welche Begriffe nach ihren Anwendungen nah beieinander liegen und welche Begriffe in enger Verbindung mit welchen Fächern stehen. Die beiden Punkte sind von großer Bedeutung für Lehrpersonen und Studenten, denn so kann während der Vorlesungen darauf hingewiesen werden, für welche Veranstaltungen das jeweilige statistische Wissen benötigt wird, wodurch Studenten sich darauf mehr konzentrieren können. Beispielsweise müssen für den Kurs *Investmentanalyse und Portfoliomanagement* die Formelsammlung für Erwartungswert, Varianz und Korrelation bekannt sein, um Rendite und Risiko berechnen zu können. Deshalb würde es das Lernen erleichtern, wenn darauf schon in der statistischen Vorlesung hingewiesen würde. Zu beachten ist, dass nah beieinander liegende Begriffe in dem Korrespondenzraum nicht inhaltlich ähnlich sind, sondern nur ähnlicher Anwendung dienen, d.h., sie sind gemeinsam für ein Gebiet wichtig.

Die Korrespondenzanalyse wurde Mithilfe von SPSS durchgeführt, aber die Daten waren nicht erwartungsgemäß geeignet. Tabelle 4.10 zeigt, dass der Anteil der Trägheit der ersten zwei Dimensionen erst 67,6% betrug, was ziemlich viel niedriger war als gefordert. Außerdem ist die Abbildung 4.2 ein Chaos, weil fast alle Punkte sich zu einem Bereich aggregierten, der kaum interpretierbar war. Aber ein erwähnenswerter Punkt aus dem Ergebnis war, dass Fachgebiet 5 in weiterer Analyse ausgeschlossen werden konnte, da er allein weit entfernt von anderen Variablen lag und im Histogramm (siehe Anhang A.6) die geringste Häufigkeit (nur zweimal) besaß. Daraus wurde geschlossen, dass die Fächer *Betriebswirtschaftslehre, Internes*

Rechnungswesen, Wettbewerbsstrategie und *Wirtschaftsinformatik* aus Fachgebiet 5 kaum statistisches Wissen erfordern. Um weitere Entscheidungen für andere Variable zu treffen, wurde die Analyse des zweiten Versuchs durchgeführt (d.h. in vier Clustern).

Dimen- sion	Singulär- wert	Träg- heit	Chi- Quadrat	Sig.	Anteil der Trägheit		Singulärwert für Konfidenz		
					Bedingen	Kumuliert	Standard- abweichung	Korrelation	
								2	3
1	.218	.048			.420	.420	.043	.401	.003
2	.170	.029			.255	.676	.039		-.003
3	.134	.018			.157	.833	.044		
4	.094	.009			.078	.911			
5	.075	.006			.050	.961			
6	.056	.003			.028	.989			
7	.036	.001			.011	1.000			
Gesamt		.114	55.986	.992a	1.000	1.000			

Tabelle 4.10: Auswertung der Trägheit der Dimensionen

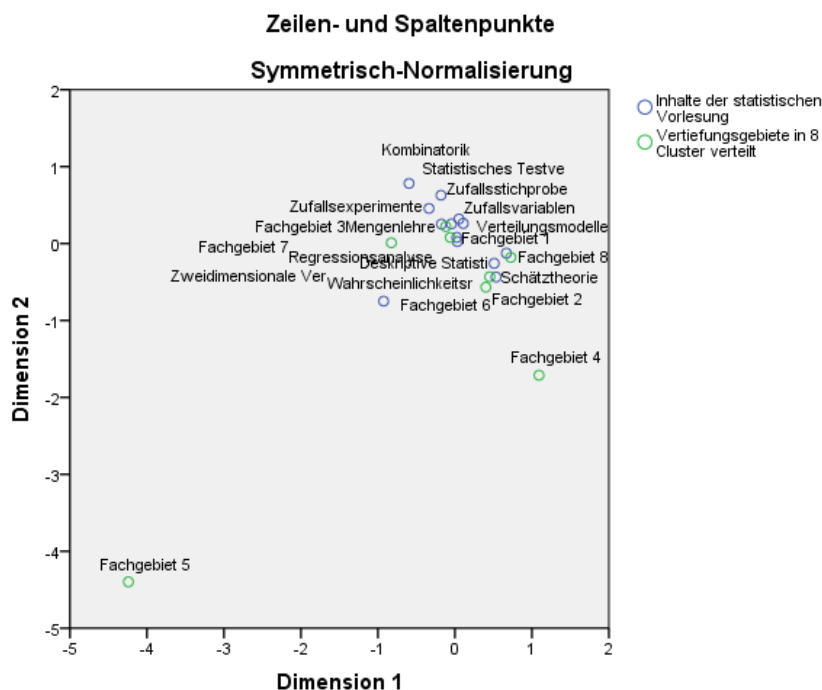


Abbildung 4.2: Korrespondenzraum in zwei Dimensionen

4.4 Cluster- und Korrespondenzanalyse des zweiten Versuchs

a) Clusteranalyse

Wie im letzten Abschnitt bereits erwähnt wurde, waren alle Komponenten in Fachgebiet 5 nicht mehr zu betrachten. Nach nochmaliger Durchführung der Clusteranalyse mit den übrigen Variablen erhielt man folgende Dendrogramme.

Abbildung 4.3 zeigt das Ergebnis des *Ward-Verfahrens* mit *Quadriertem Euklidischem Abstand*. Die Verteilung der vier Cluster war offensichtlich:

Cluster 1: Theorie des Marktversagens (18), Andere (20), Applied Statistic (4), Mikroökonomie III (13), Öffentliche Finanzen in der Demokratie (15)

Cluster 2: Finanzwirtschaft (9), Versicherungs- und Risikomanagement (19), Angewandte Ökonometrie (3), Operations Research (16), Entrepreneurship (7), Ange-

wandte Makroökonomie (1), Spieltheorie (17)

Cluster 3: Internationales Management (10), Marketing (12)

Cluster 4: Makroökonomie III (11), Monetary Economics (14), Arbeitsmarktökonomik (2), Bank- und Börsenwesen (6), Externes Rechnungswesen (8)

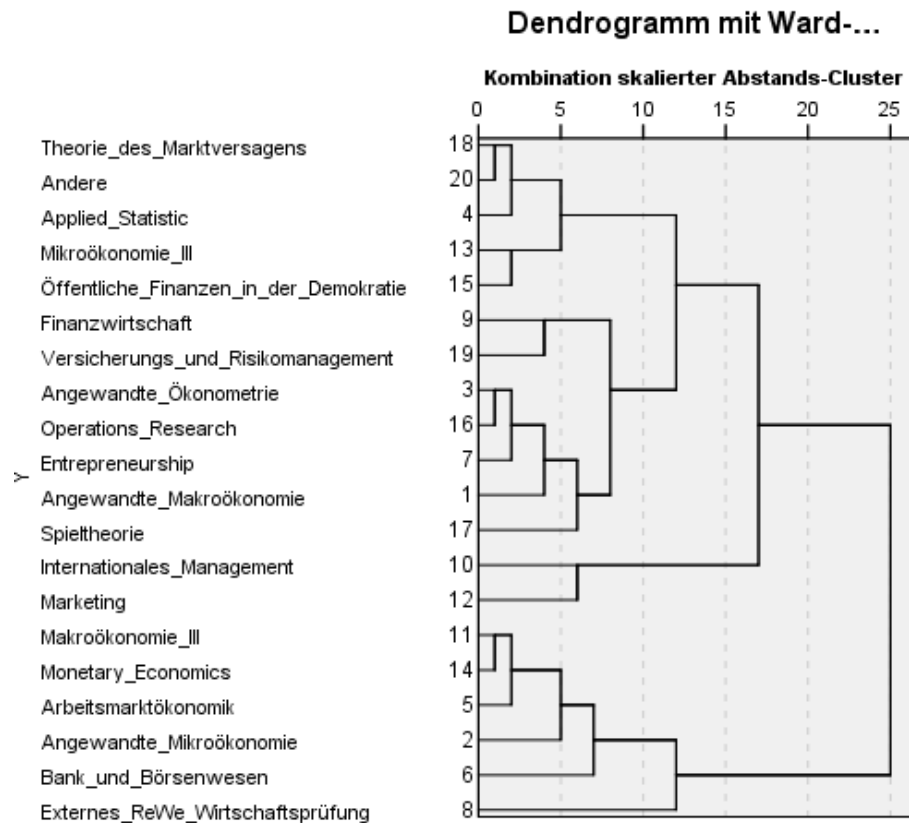


Abbildung 4.3: Neues Dendrogramm mit Ward-Verfahren (ohne Fachgebiet 5)

Die Abbildung 4.4 zeigt die Clusterverteilung von *Complete-Linkage*. Betrachtete man die Dendrogramm mithilfe der *Complete-Linkage-Methode* und des *Jaccard-Distanzmaßes*, dann bot sich auch hier nur eine 4-Cluster-Lösung an, sonst wird es zu viel. Hier wird geprüft, ob die 4-Cluster-Lösung mit *Ward* und *Complete-Linkage* die gleiche Lösung liefern. D.h., es wurde betrachtet, ob gleiche Fachgebiete in gleichen Clustern liegen. Wenn verschiedene Verfahren ähnliche Cluster-Lösungen ergeben, dann ist es mehr reliabel.

Hier wurden nur die Zahlen ausgenommen, um den Vergleich zu vereinfachen.

Cluster 1: 13, 15, 18, 20, 4

Cluster 2: 1, 7, 3, 16, 9,19

Cluster 3: 11, 14, 2, 5, 6, 8

Cluster 4: 10, 12, 17

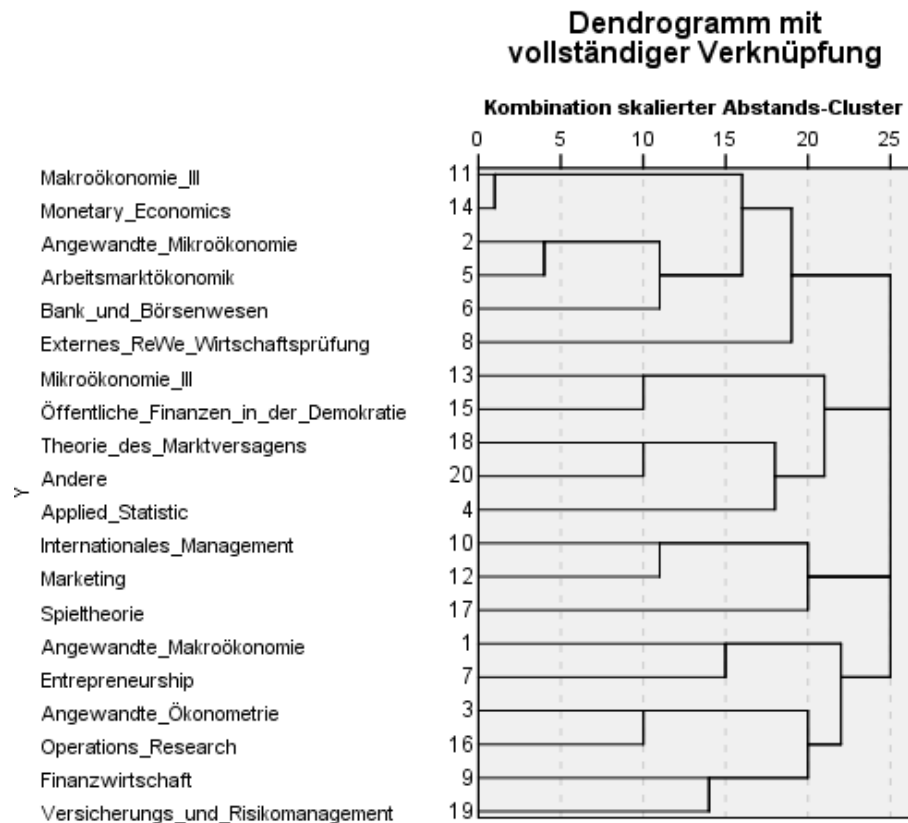


Abbildung 4.4: Dendrogramm mit Complete-Linkage-Verfahren

Die beiden Tabellen stimmten fast überein, was zeigte, dass die 4-Cluster-Lösung sehr zuverlässig ist. Die Erklärungsregel war dieselbe wie in 4.3.1, nämlich dass die Gruppenverteilung nicht auf inhaltlicher Ähnlichkeit basierte, sondern auf der Meinung der Studenten. Es wäre interessant gewesen, die Stabilität des Ergebnisses mithilfe von Bootstrapping zu untersuchen, wobei ein Teil der Beobachtungen zufällig auszuwählen gewesen wäre und die Clusteranalyse hätte in diesem Teil noch einmal durchgeführt werden müssen. Man würde dann schauen, ob sich die gleichen Cluster ergeben. Dies war aber nicht möglich, weil SPSS keine Implementation da-

für bietet. Man kann sich gegebenenfalls an R-Software wenden. In der folgenden Analyse wurde das Ergebnis des Ward-Verfahrens angewendet.

b) Korrespondenzanalyse

Bei der letzten Analyse, die auch mithilfe von SPSS ausgeführt wurde, wurde als Erstes die zu analysierende Kontingenztabelle¹ (Korrespondenztabelle genannt) ausgegeben, wobei auch die Randsummen ausgewiesen wurden. Die beiden nächsten Tabellen enthalten die Zeilen- und Spaltenprofile; dies sind die auf die Zeilen- bzw. Spaltensummen bezogenen relativen Häufigkeiten, beispielsweise ergeben die Zeilenprofile die relativen Häufigkeiten der 13 statistischen Begriffe in jedem Cluster. Der nächste Teil des Ausdrucks zeigt den Gesamtbetrag der Trägheit an und die Anteile, die den einzelnen Dimensionen zugeordnet werden können. Dimensionen mit nur geringem Anteil an der Trägheit konnten gegebenenfalls vernachlässigt werden. In diesem Fall wurden durch die ersten beiden Dimensionen 66,5 bzw. 20,8% der Trägheit erklärt, so dass die Darstellung in zwei Dimensionen gerechtfertigt war².

Danach wurde die Konfiguration der Zeilenpunkte (Abbildung 4.5), also die Konfiguration der beurteilten statistischen Inhalte, betrachtet. Auf der horizontalen Achse steuerten insbesondere die Begriffe *Kombinatorik* und *Deskriptive Statistik*, während die vertikale Achse sich dagegen als *Kombinatorik* und *Mengenlehre* bezeichnen lässt. Deutlich war zu erkennen, dass die Begriffe *Mengenlehre & Wahrscheinlichkeitsrechnung*, *Verteilungsmodelle & Zufallsvariablen & Zeitreihenanalyse* sowie *Zufallsstichprobe & Zweidimensionale Verteilung* eng beieinanderliegen, d.h. hinsichtlich der gegebenen Fachgebiete-Cluster ähnlich beurteilt werden. Hier ist noch einmal zu betonen, dass sie nicht inhaltlich ähnlich sind, sondern nur ähnliche Anwendung finden.

¹Alle in diesem Absatz genannten Tabellen sind in Anhang A.7 zu finden.

²Vgl. Achim Bühl: SPSS 16, Einführung in die moderne Datenanalyse, 2008 (Kapitel 25: Korrespondenzanalyse)

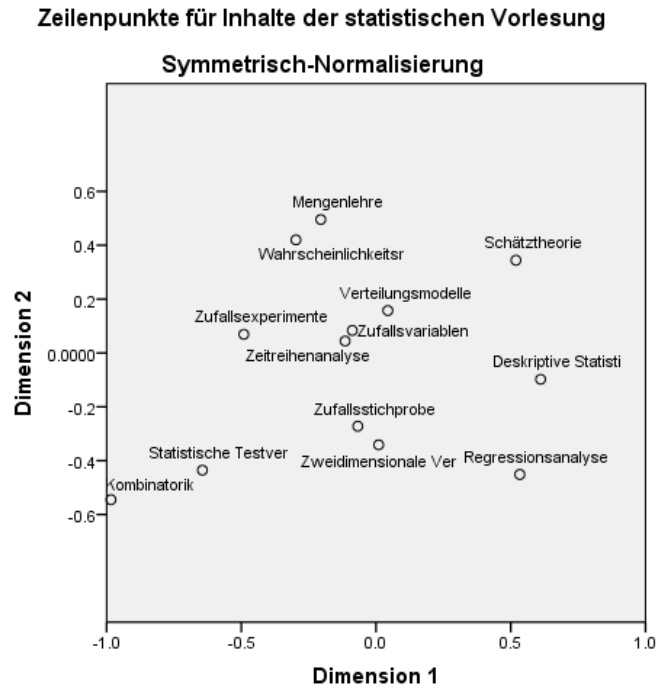


Abbildung 4.5: Zeilenpunkte für Inhalte der statistischen Vorlesungen

Die Konfiguration der Spaltenpunkte in Abbildung 4.6 sagt aus, dass die vier Cluster sehr weit entfernt voneinander stehen, d.h., die Cluster sind nicht sehr abhängig voneinander, wobei die Heterogenitätsregel zwischen den Gruppen in der Clusteranalyse noch einmal bestätigt wurde. Die erste Dimension wurde offensichtlich dominiert von Cluster 2 und Cluster 4, die zweite von Cluster 2 und Cluster 3, deren Abstand war allerdings nur etwas größer als der Abstand zwischen Cluster 1 und Cluster 4.

Um die Konfiguration der statistischen Inhalte besser deuten zu können, ist es entscheidend, beide Konfigurationen zusammen in einem Korrespondenzraum zu betrachten.

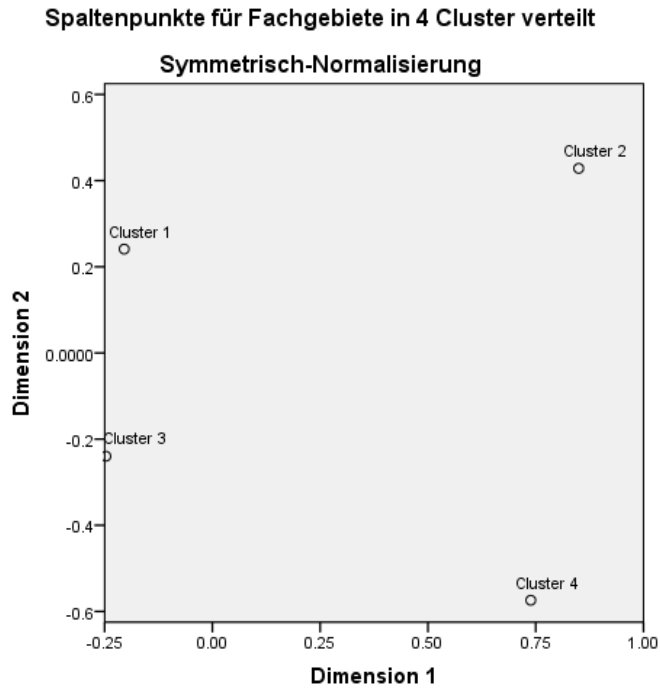


Abbildung 4.6: Spaltenpunkte für Fachgebiete

Abbildung 4.7 zeigt, dass Cluster 1 von mehreren statistischen Begriffen umgeben ist, während Cluster 2, 3 und 4 nur von ein bis zwei Begriffen begleitet sind. Cluster 1, das aus *Theorie des Marktversagens*, *Andere*, *Applied Statistics*, *Mikroökonomie III* und *Öffentliche Finanzen in der Demokratie* besteht, hat eine enge Beziehung mit *Mengenlehre*, *Wahrscheinlichkeitsrechnung*, *Verteilungsmodelle*, *Zufallsvariablen*, *Zeitreihenanalyse* und *Zufallsexperimente*. Das aus sieben Fachgebieten bestehende Cluster 2 (*Finanzwirtschaft*, *Versicherungs- und Risikomanagement*, *Angewandte Ökonometrie*, *Operations Research*, *Entrepreneurship*, *Angewandte Makroökonomie*, *Spieltheorie*) bezog sich gemäß der Untersuchung dennoch nur auf die *Schätztheorie*, was überrascht, da so viele Fachgebiete in Cluster 2 liegen. Relativ plausibel war, dass *Internationales Management* und *Marketing* (Cluster 3) sehr nah bei *zweidimensionale Verteilung* und *Zufallsstichprobe* lagen, weil *Marketing* sich stärker der Statistik bedient, um gesammelte Daten zu bearbeiten. Ähnlich wie Cluster 2 schnitt Cluster 4 mit recht hoher Entfernung von den anderen Begriffen am besten bei der *Regressionsanalyse* ab. Die Komponenten in Cluster 4 waren *Makroökonomie III*, *Monetary Economics*, *Arbeitsmarktökonomik*, *Bank- und Börsenwesen*

und *Externes Rechnungswesen*.

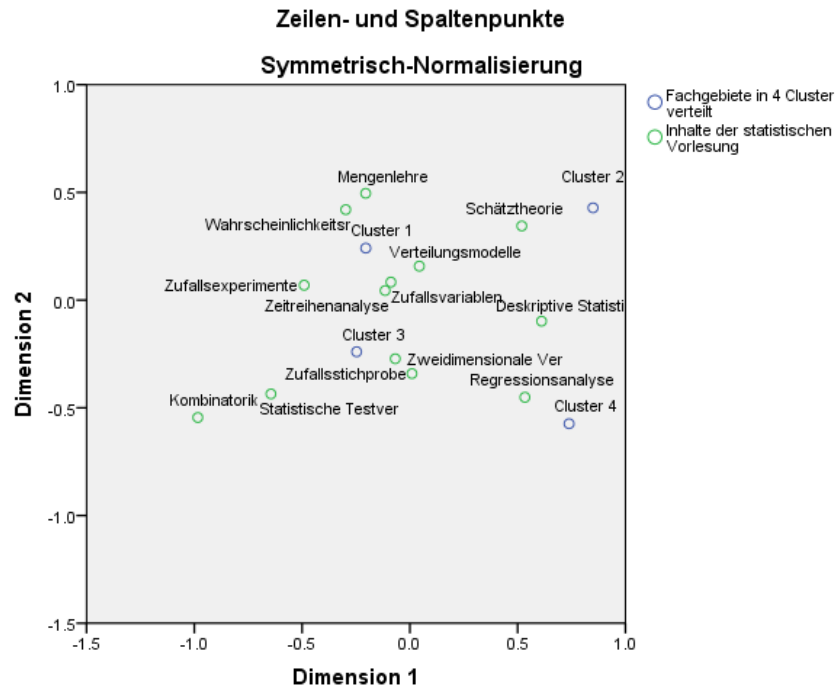


Abbildung 4.7: Zeilen- und Spaltenpunkte

Hiermit könnte das Missverständnis entstehen, dass Cluster 1 deutlich mehr Statistik benötigt als andere Cluster. Dies ist jedoch nicht der Fall, wenn man das Histogramm betrachtet. Das im Anhang befindliche Histogramm (Anhang A.6.) zeigt, dass die Cluster 1 und 3 fast gleiche Anteile der absoluten an Häufigkeiten besaßen, d.h., obwohl Cluster 3 nur bei zwei statistischen Begriffen eng lag, braucht es so viel Statistik wie Cluster 1.

Außerdem gab es noch drei weitere Begriffe. Bei *Kombinatorik* und *Statistische Testverfahren* konnte mit großer Wahrscheinlichkeit beschlossen werden, dass sie keinen anderen Veranstaltungen dienen. Zwar lag *Deskriptive Statistik* nicht nah bei einem Cluster, aber in der Mitte: man kann also nicht sagen, dass sie nicht relevant für andere Veranstaltungen sei. Eine andere mögliche Erklärung könnte sein, dass *Deskriptive Statistik* von allen hier genannten Veranstaltungen benötigt wird.

5 Zusammenfassung

Die Kernpunkte der vorliegenden Arbeit drücken sich in den folgenden Fragen aus:

- Welche Inhalte aus den Lehrveranstaltungen Statistik I & II werden am wenigsten benötigt?
- Für welche Vertiefungsgebiete ist das statistische Wissen wichtig?

Um diese Frage zu beantworten, wurden im ersten Schritt die Daten zu bereinigt, indem die Plausibilität der Daten geprüft und die fehlenden Werte in Kauf genommen wurden. Besonders relevant waren die Behandlung und Feststellung der Variablen (statistische Begriffe). Durch Ähnlichkeitsanalyse konnte geschlossen werden, dass die Unterbegriffe nur zur Ergänzung der Oberbegriffe dienen, demnach reicht es hier, 13 Variable zu untersuchen.

Im zweiten Schritt sollte das erst oben genannte Problem gelöst werden, die ausführliche Analyse dazu erfolgt in Abschnitt 4.2. Die Ergebnisse aus verschiedenen Quellen stimmten überein. Mithilfe der Häufigkeitsverteilung bezüglich statistischer Begriffe und Fachgebiete konnte man sicher sagen, dass bei *Kombinatorik*, *Zufallsstichprobe*, *Zeitreihenanalyse* und *Statistisches Testverfahren* mehr oder weniger Probleme entstehen. Sie wurden von den Befragten für nicht so wichtig für andere Veranstaltungen gehalten, und demnach können sie in Lehrveranstaltungen verkürzt behandelt oder sogar gestrichen werden. Aber um eine genaue Entscheidung darüber zu treffen, wäre es zusätzlich erforderlich, Vorschläge von relevanten Lehrpersonen einzuholen und weitere Nachforschungen anzustellen.

Im letzten Schritt wurde das zweite Problem erklärt. Obwohl die in Abschnitt 4.3 auf der 8-Cluster-Lösung basierende Analyse nicht erfolgreich war, konnte sie doch etwas zur Untersuchung beitragen, weil Fachgebiet 5 (*Betriebswirtschaftliche Steuerlehre*, *Internes Rechnungswesen*, *Wettbewerbsstrategie*, *Wirtschaftsinformatik*) ausgeschlossen war und die weitere Analyse dadurch erleichtert wurde. Die Analyse in Abschnitt 4.4 ergab hingegen, dass die Begriffe *Mengenlehre*, *Wahrscheinlichkeits-*

rechnung, Verteilungsmodelle, Zufallsvariable, Zeitreihenanalyse und Zufallsexperimente wichtig für die Veranstaltungen *Theorie des Marktversagens, Andere, Applied Statistics, Mikroökonomie III* und *Öffentliche Finanzen in der Demokratie* sind, und *Schätztheorie* wird allein für *Finanzwirtschaft, Versicherungs- und Risikomanagement, Angewandte Ökonometrie, Operations Research, Entrepreneurship, Angewandte Makroökonomie* und *Spieltheorie* benötigt. Außerdem sind *zweidimensionale Verteilung* und *Zufallsstichprobe* wichtig für *Internationales Management* und *Marketing*, und *Makroökonomie III, Monetary Economics, Arbeitsmarktökonomik, Bank- und Börsenwesen* und *Externes Rechnungswesen* stellen große Anforderung bezüglich *Regressionsanalyse. Deskriptive Statistik* wurde als überall notwendig und *Kombinatorik* und *Statistische Testverfahren* als nicht relevant für andere Veranstaltungen eingestuft.

Dank der für diese Untersuchung verwendeten Arbeit von Frau Lorenz stand bereits eine Datensammlung zur Verfügung, die für die vorliegende Arbeit genutzt werden konnte. Wegen der unvollständigen Konstruktion des Fragebogens und des zu geringen Umfangs der Stichprobe war die darauf folgende Analyse jedoch erschwert und nicht zufriedenstellend. Die meisten Variablen sind nominal oder sogar binär skaliert, dabei waren viele statistische Methoden nach Prüfung nicht geeignet und die einzelnen anwendbaren multivariaten Verfahren lieferten auch nicht so plausible und gerechtfertigte Ergebnisse wie erwartet. Die zu kleine Stichprobe enthielt so wenig Information, dass die Untersuchung unweigerlich verzerrt wurde. Als Empfehlung wurden einige Veränderungen für Komplex 3 genannt, beispielsweise alle Unterbegriffe zu löschen, um verzerrte und uneinheitliche Angaben zu vermeiden und auch die Zeit für die Befragten zu sparen; der zentrale Teil -die Wichtigkeit der Begriffe- wurde in Ratingskalen erstellt, um zahlreiche Auswertungsverfahren anwenden zu können. Ansonsten könnte die Frage „Welche anderen statistischen Themen benötigen Sie für Ihr Weiterstudium?“ als Inhaltsergänzung hinzugefügt werden. Außerdem wurden Master-Studierende als Zielgruppe für eine Untersuchung als nicht geeignet empfunden. Die Fragebögen hätten besser in entsprechenden Seminaren an Bachelor-Studierende verteilt werden können. Damit hätte man sichergestellt, dass die Befragten Statistik-Vorlesungen an der HU besucht und gleichzeitig auch

viele Vertiefungsgebiete belegt haben. Die in dieser Arbeit verwendeten Analysemethoden können als Referenz für weitere Untersuchungen dienen, dabei sollte dann jedoch eine hinreichende große Datenmenge vorhanden sein.

Literatur

- [1] Backhaus, K. ; Erichson, B. ; Plinke, W. ; Weiber, R., *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, 13. Auflage, Springer, Heidelberg, 2010
- [2] Bortz, J.; Schuster, C., *Statistik: Für Human- und Sozialwissenschaftler*, 7. Auflage, Springer, Heidelberg, 2010
- [3] Bühl, A., *SPSS 14. Einführung in die moderne Datenanalyse*, 10. Auflage Pearson Studium, München, 2006
- [4] Bühl, A., *SPSS 16. Einführung in die moderne Datenanalyse*, 11. Auflage Pearson Studium, München, 2008
- [5] Härdle, W.; Klinke, S., *Connected Teaching of Statistics*, DP 9924, SFB 373, Humboldt-University of Berlin, 1999, URL:<http://edoc.hu-berlin.de/series/sfb-373-papers/1999-24/PDF/24.pdf>
- [6] Herrmann, A., *Marktforschung*, 2. Auflage, Springer, Heidelberg, 2000
- [7] Porst, R., *Fragebogen*, 1. Auflage, Springer, Heidelberg, 2008
- [8] Rönz, B., *Skript zur Vorlesung Computergestützte Statistik I*, Humboldt Universität zu Berlin, 2001
- [9] Speidel, M., *Fehlende Werte*, Seminararbeit zum Zusatzkurs Amtliche Statistik, München, 2009, URL:<http://www.stat.uni-muenchen.de/~thomas/Home/Meldungen/Dateien/handout-fehlende%20Daten.pdf>,
- [10] Voß, H., *Einführung in LATEX*, 1. Auflage, Lehmanns Media, Berlin, 2012

A Anhang

A.1 Grad der Identität zwischen Ober- und Unterbegriffen

		Frageteil 1		Frageteil 2	
Mengenlehre	Mengen, Elemente, Teilmengen	57	100.00%	54	94.74%
	Venn Diagramme	45	78.95%	55	96.49%
	Mengenoperationen	56	98.25%	55	96.49%
Zufalls- experimente	Ereignisse, Elementarereignisse, Ereignisraum	57	100.00%	54	94.74%
	Relationen und Operationen von Ereignissen	55	96.49%	54	94.74%
Kombinatorik	Fakultät und Binomialkoeffizienten	54	94.74%	56	98.25%
	Permutation	46	80.70%	55	96.49%
	Variation	50	87.72%	55	96.49%
	Kombination	50	87.72%	55	96.49%
Wahrscheinlichkeits- rechnung	Laplace	54	94.74%	50	87.72%
	Wahrscheinlichkeitsbegriff	57	100.00%	51	89.47%
	Additionssatz, Multiplikationssatz	55	96.49%	55	96.49%

	Bedingte Wahrscheinlichkeit	57	100.00%	54	94.74%
	Totale Wahrscheinlichkeit	52	91.23%	52	91.23%
	Theorem von Bayes	51	89.47%	56	98.25%
	Baumdiagramme	55	96.49%	53	92.98%
Deskriptive Statistik	Skalierung	57	100.00%	55	96.49%
	diskrete und stetige Variablen	57	100.00%	57	100.00%
	Klassierung	57	100.00%	55	96.49%
	Hufigkeitsverteilung	55	96.49%	56	98.25%
	Lineare Interpolation	51	89.47%	56	98.25%
	Lageparameter	57	100.00%	54	94.74%
	Streuungsparameter	57	100.00%	56	98.25%
Zufalls- variablen	Wahrscheinlichkeitsdichte	57	100.00%	56	98.25%
	Wahrscheinlichkeitsfunktion	57	100.00%	53	92.98%
	Varianz einer Zufallsvariable	57	100.00%	55	96.49%
	Erwartungswert einer Zufallsvariable	57	100.00%	55	96.49%

	Verteilungsfunktion einer Zufallsvariable	57	100.00%	56	98.25%
Zwei-dimensionale Verteilung	Randverteilung	54	94.74%	56	98.25%
	bedingte Verteilung	55	96.49%	57	100.00%
	Unabhängige Variablen und Kovarianz	57	100.00%	57	100.00%
	Korrelation	57	100.00%	56	98.25%
	Kontingenz, Kontingenztafel	52	91.23%	55	96.49%
Verteilungsmodelle	Gleichverteilung	57	100.00%	56	98.25%
	Binominalverteilung	56	98.25%	55	96.49%
	Hypergeometrische Verteilung	52	91.23%	52	91.23%
	Poisson-Verteilung	55	96.49%	53	92.98%
	Exponentialverteilung	56	98.25%	55	96.49%
	Normalverteilung	57	100.00%	56	98.25%
	Zentrale Grenzwertsatz	55	96.49%	53	92.98%
	Chi-Quadrat-Verteilung	54	94.74%	53	92.98%
	t-Verteilung	55	96.49%	53	92.98%

	F-Verteilung	54	94.74%	54	94.74%
Zufallsstich- probe	Verteilung des Stichprobenmittelwertes	53	92.98%	57	100.00%
	Verteilung des Stichprobenanteilswertes	52	91.23%	56	98.25%
	Verteilung der Stichprobenvarianz	54	94.74%	57	100.00%
Schätztheorie	Schätzfunktionen	55	96.49%	55	96.49%
	Maximum-Likelihood- Methode	56	98.25%	56	98.25%
	Methode der kleinsten Quadrate	56	98.25%	56	98.25%
	Konfidenzintervall	55	96.49%	54	94.74%
Statistische Testverfahren	Testverfahren für den Erwartungswert der Grundgesamtheit	57	100.00%	56	98.25%
	Testverfahren für den Anteilswert π einer dichotomen Grundgesamtheit	53	92.98%	53	92.98%
	Testverfahren für die Differenz der Erwartungs- werte zweier Grundgesamtheiten	54	94.74%	54	94.74%
	Chi-Quadrat- Anpassungstest	54	94.74%	56	98.25%
	Chi-Quadrat- Unabhängigkeitstest	54	94.74%	55	96.49%
Regressions- analyse	Das Regressionsmodell	57	100.00%	56	98.25%
	Schätzung der Regressionsparameter	57	100.00%	56	98.25%

**Regressions-
analyse**

	Bestimmtheit (Gute) der Regression	57	100.00%	56	98.25%
	Statistischer Test der Regressionsparameter	57	100.00%	55	96.49%
	Konfidenzintervalle	57	100.00%	57	100.00%
Zeitreihen- analyse	Komponenten einer Zeitreihe	55	96.49%	56	98.25%
	Geometrisches Mittel	57	100.00%	54	94.74%
	Methode der gleitenden Durchschnitte	53	92.98%	54	94.74%
	Methode der kleinsten Quadrate	56	98.25%	54	94.74%
	Periodische Schwankungen	53	92.98%	54	94.74%
	Einschätzung der Gute des Zeitreihen	51	89.47%	55	96.49%

A.2 Anzahl der fehlenden Werte vor und nach der Imputation

	Vor der Imputation	Nach der Imputation
Mengenlehre	15 26.32%	9 15.79%
Zufallsexperimente	20 35.09%	11 19.30%
Kombinatorik	19 33.33%	9 15.79%
Wahrscheinlichkeitsrechnung	23 40.35%	6 10.53%
Deskriptive Statistik	19 33.33%	6 10.53%
Zufallsvariablen	12 21.05%	5 8.77%
Zweidimensionale Verteilung	22 38.60%	8 14.04%
Verteilungsmodelle	20 35.09%	6 10.53%
Zufallsstichprobe	23 40.35%	19 33.33%
Schätztheorie	18 31.58%	15 26.32%
Statistisches Verfahren	23 40.35%	7 12.28%
Regressionsanalyse	14 24.56%	7 12.28%
Zeitreihenanalyse	21 36.84%	14 24.56%

A.3 Ergebnisse des t-Tests beim Merkmal Geschlecht

Gruppenstatistiken

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Mengenlehre kennen	weiblich	33	.24	.502	.087
	männlich	24	.17	.565	.115
Zufallsexperimente kennen	weiblich	33	.09	.384	.067
	männlich	24	.21	.588	.120
Kombinatorik kennen	weiblich	33	.09	.384	.067
	männlich	24	.25	.676	.138
Wahrscheinlichkeitsrechnung kennen	weiblich	33	.06	.348	.061
	männlich	24	.13	.448	.092
Deskriptive Statistik kennen	weiblich	33	.06	.348	.061
	männlich	24	.25	.676	.138
Zufallsvariablen kennen	weiblich	33	.06	.348	.061
	männlich	24	.17	.565	.115
Zweidimensionale Verteilung kennen	weiblich	33	.12	.485	.084
	männlich	24	.17	.565	.115
Verteilungsmodelle kennen	weiblich	33	.00	.000	.000
	männlich	24	.21	.588	.120
Zufallsstichprobe kennen	weiblich	33	.21	.600	.104
	männlich	24	.29	.690	.141
Schätztheorie kennen	weiblich	33	.09	.384	.067
	männlich	24	.29	.690	.141
Statistisches Testverfahren kennen	weiblich	33	.09	.384	.067

	männlich	24	.17	.565	.115
Regressionsanalyse kennen	weiblich	33	.09	.384	.067
	männlich	24	.21	.588	.120
Zeitreihenanalyse kennen	weiblich	33	.21	.485	.084
	männlich	24	.33	.702	.143

		Levene-Test der Varianzgleichheit		t-Test für Mittelwertvergleiche		
		F	Signi- fikanz	T	df	Sig. (2-seitig)
Mengenlehre kennen	Varianzen sind gleich	.536	.467	.534	55	.596
	Varianzen sind nicht gleich			.524	46.087	.603
Zufalls- experimente kennen	Varianzen sind gleich	3.363	.072	-.911	55	.366
	Varianzen sind nicht gleich			-.854	36.941	.398
Kombinatorik kennen	Varianzen sind gleich	5.621	.021	-1.127	55	.265
	Varianzen sind nicht gleich			-1.038	33.756	.307
Wahrschein- lichkeits- rechnung kennen	Varianzen sind gleich	1.403	.241	-.610	55	.544
	Varianzen sind nicht gleich			-.587	41.811	.561
Deskriptive Statistik kennen	Varianzen sind gleich	8.386	.005	-1.381	55	.173
	Varianzen sind nicht gleich			-1.257	31.886	.218
Zufalls- variablen kennen	Varianzen sind gleich	3.172	.080	-.876	55	.385
	Varianzen sind nicht gleich			-.814	35.525	.421
Zwei- dimensionale Verteilung kennen	Varianzen sind gleich	.425	.517	-.326	55	.746

	Varianzen sind nicht gleich			- .318	44.967	.752
Verteilungsmodelle kennen	Varianzen sind gleich	21.303	.000	-2.042	55	.046
	Varianzen sind nicht gleich			-1.735	23.000	.096
Zufallsstichprobe kennen	Varianzen sind gleich	.803	.374	-.464	55	.645
	Varianzen sind nicht gleich			-.454	45.368	.652
Schätztheorie kennen	Varianzen sind gleich	8.394	.005	-1.401	55	.167
	Varianzen sind nicht gleich			-1.287	33.324	.207
Statistische Testverfahren kennen	Varianzen sind gleich	1.567	.216	-.603	55	.549
	Varianzen sind nicht gleich			-.568	38.010	.573
Regressionsanalyse kennen	Varianzen sind gleich	3.363	.072	-.911	55	.366
	Varianzen sind nicht gleich			-.854	36.941	.398
Zeitreihenanalyse kennen	Varianzen sind gleich	3.047	.086	-.772	55	.444
	Varianzen sind nicht gleich			-.729	38.393	.470

A.4 Ergebnisse des t-Tests beim Merkmal Muttersprache

Gruppenstatistiken

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Mengenlehre kennen	deutsch	39	.13	.409	.066
	nicht deutsch	18	.39	.698	.164
Zufallsexperimente kennen	deutsch	39	.08	.354	.057
	nicht deutsch	18	.28	.669	.158
Kombinatorik kennen	deutsch	39	.15	.540	.086
	nicht deutsch	18	.17	.514	.121
Wahrscheinlichkeitsrechnung kennen	deutsch	39	.08	.354	.057
	nicht deutsch	18	.11	.471	.111
Deskriptive Statistik kennen	deutsch	39	.10	.447	.072
	nicht deutsch	18	.22	.647	.152
Zufallsvariablen kennen	deutsch	39	.10	.447	.072
	nicht deutsch	18	.11	.471	.111
Zweidimensionale Verteilung	deutsch	39	.10	.447	.072
	nicht deutsch	18	.22	.647	.152
Verteilungsmodelle kennen	deutsch	39	.13	.469	.075
	nicht deutsch	18	.00	.000	.000
Zufallsstichprobe kennen	deutsch	39	.18	.556	.089
	nicht deutsch	18	.39	.778	.183
Schätztheorie kennen	deutsch	39	.13	.469	.075
	nicht deutsch	18	.28	.669	.158
Statistisches Testverfahren kennen	deutsch	39	.10	.447	.072

	nicht deutsch	18	.17	.514	.121
Regressionsanalyse kennen	deutsch	39	.13	.469	.075
	nicht deutsch	18	.17	.514	.121
Zeitreihenanalyse kennen	deutsch	39	.33	.662	.106
	nicht deutsch	18	.11	.323	.076

		Levene-Test der Varianzgleichheit		t-Test für Mittelwertvergleiche		
		F	Signi- fikanz	T	df	Sig. (2-seitig)
Mengenlehre kennen	Varianzen sind gleich	10.808	.002	-1.773	55	.082
	Varianzen sind nicht gleich			-1.472	22.567	.155
Zufalls- experimente kennen	Varianzen sind gleich	8.850	.004	-1.486	55	.143
	Varianzen sind nicht gleich			-1.198	21.522	.244
Kombinatorik kennen	Varianzen sind gleich	.009	.923	-.085	55	.933
	Varianzen sind nicht gleich			-.086	34.666	.932
Wahrschein- lichkeits- rechnung kennen	Varianzen sind gleich	.400	.530	-.304	55	.762
	Varianzen sind nicht gleich			-.274	26.221	.786
Deskriptive Statistik kennen	Varianzen sind gleich	2.613	.112	-.812	55	.420
	Varianzen sind nicht gleich			-.711	24.780	.484
Zufalls- variablen kennen	Varianzen sind gleich	.017	.896	-.066	55	.948
	Varianzen sind nicht gleich			-.065	31.597	.949
Zwei- dimensionale Verteilung kennen	Varianzen sind gleich	2.613	.112	-.812	55	.420

	Varianzen sind nicht gleich			-0.711	24.780	.484
Verteilungsmodelle kennen	Varianzen sind gleich	6.146	.016	1.154	55	.253
	Varianzen sind nicht gleich			1.707	38.000	.096
Zufallsstichprobe kennen	Varianzen sind gleich	4.757	.033	-1.161	55	.251
	Varianzen sind nicht gleich			-1.028	25.337	.314
Schätztheorie kennen	Varianzen sind gleich	3.494	.067	-0.974	55	.334
	Varianzen sind nicht gleich			-0.856	25.009	.400
Statistische Testverfahren kennen	Varianzen sind gleich	.775	.382	-0.480	55	.633
	Varianzen sind nicht gleich			-0.455	29.312	.652
Regressionsanalyse kennen	Varianzen sind gleich	.265	.609	-0.279	55	.781
	Varianzen sind nicht gleich			-0.270	30.532	.789
Zeitreihenanalyse kennen	Varianzen sind gleich	8.995	.004	1.347	55	.184
	Varianzen sind nicht gleich			1.702	54.745	.095

A.5 Häufigkeit der Vertiefungsgebiete jedes statistischen Themas

(Mengenlehre)	(Häufigkeit)
Andere	6
Angewandte Ökonometrie	5
Applied Statistics	3
Spieltheorie	3
Finanzwirtschaft	2
Angewandte Mikroökonomie	2
Mikroökonomie III	2
Externes Rechnungswesen	1
Marketing	1
Arbeitsmarktökonomie	1
Operations Research	1
Angewandte Makroökonomie	1
SUMME :	28

(Zufallsexperimente)	(Häufigkeit)
Andere	6
Spieltheorie	4
Angewandte Ökonometrie	3
Applied Statistics	3
Entrepreneurship	3
Marketing	2
Versicherungs- und Risikomanagement	2
Mikroökonomie III	2
Finanzwirtschaft	1
Makroökonomie III	1
Operations Research	1
Theorie der Marktversagens	1
Monetary Economics	1

SUMME : 30

(Kombinatorik)	(Häufigkeit)
Andere	6
Angewandte Ökonometrie	4
Applied Statistics	2
Marketing	1
Spieltheorie	1
Theorie der Marktversagens	1
Finanzwirtschaft	1
Versicherungs- und Risikomanagement	1
Operations Research	1
Mikroökonomie III	1
SUMME :	19

(Wahrscheinlichkeitsrechnung)	(Häufigkeit)
Andere	12
Spieltheorie	9
Angewandte Ökonometrie	7
Applied Statistics	4
Versicherungs- und Risikomanagement	4
Marketing	3
Mikroökonomie III	3
Finanzwirtschaft	2
Operations Research	2
Angewandte Mikroökonomie	2
Entrepreneurship	1
Theorie der Marktversagens	1
Entscheidungstheorie	1
Bank- und Börsenwesen	1
Externes Rechnungswesen	1

Makroökonomie III	1
Monetary Economics	1
Wettbewerbsstrategie	1
Entrepreneurship	1
SUMME :	57

(Deskriptive Statistik)	(Häufigkeit)
Andere	13
Angewandte Ökonometrie	8
Marketing	7
Applied Statistics	5
Bank- und Börsenwesen	3
Entrepreneurship	3
Spieltheorie	3
Makroökonomie III	2
Mikroökonomie III	2
Finanzwirtschaft	2
Monetary Economics	2
Econometric Methods	1
Versicherungs- und Risikomanagement	1
Externes Rechnungswesen	1
Angewandte Mikroökonomie	1
Angewandte Makroökonomie	1
Arbeitsmarktökonomie	1
Theorie der Marktversagens	1
Internationales Management	1
SUMME :	58

(Zufallsvariablen)	(Häufigkeit)
Andere	12
Angewandte Ökonometrie	10

Marketing	4
Applied Statistics	4
Spieltheorie	4
Entrepreneurship	3
Mikroökonomie III	3
Versicherungs- und Risikomanagement	3
Makroökonomie III	2
Monetary Economics	2
Econometric Methods	1
Finanzwirtschaft	1
Operations Research	1
Externes Rechnungswesen	1
Angewandte Makroökonomie	1
Theorie der Marktversagens	1
Internationales Management	1
SUMME :	54

(Zweidimensionale Verteilung)	(Häufigkeit)
Angewandte Ökonometrie	9
Andere	8
Applied Statistics	5
Marketing	4
Econometric Methods	2
Versicherungs- und Risikomanagement	2
Mikroökonomie III	2
Spieltheorie	2
Finanzwirtschaft	2
Theorie der Marktversagens	1
Bank- und Börsenwesen	1
Internationales Management	1
Monetary Economics	1

SUMME : 40

(Verteilungsmodelle)	(Häufigkeit)
Angewandte Ökonometrie	12
Andere	9
Marketing	6
Applied Statistics	5
Versicherungs- und Risikomanagement	4
Econometric Methods	3
Entrepreneurship	3
Spieltheorie	3
Mikroökonomie III	2
Finanzwirtschaft	1
Bank- und Börsenwesen	1
Theorie der Marktversagens	1
Angewandte Mikroökonomie	1
Externes Rechnungswesen	1
Angewandte Makroökonomie	1
SUMME :	53

(Zufallsstichprobe)	(Häufigkeit)
Angewandte Ökonometrie	5
Andere	4
Applied Statistics	3
Econometric Methods	2
Marketing	2
Entrepreneurship	1
Versicherungs- und Risikomanagement	1
Mikroökonomie III	1
SUMME :	19

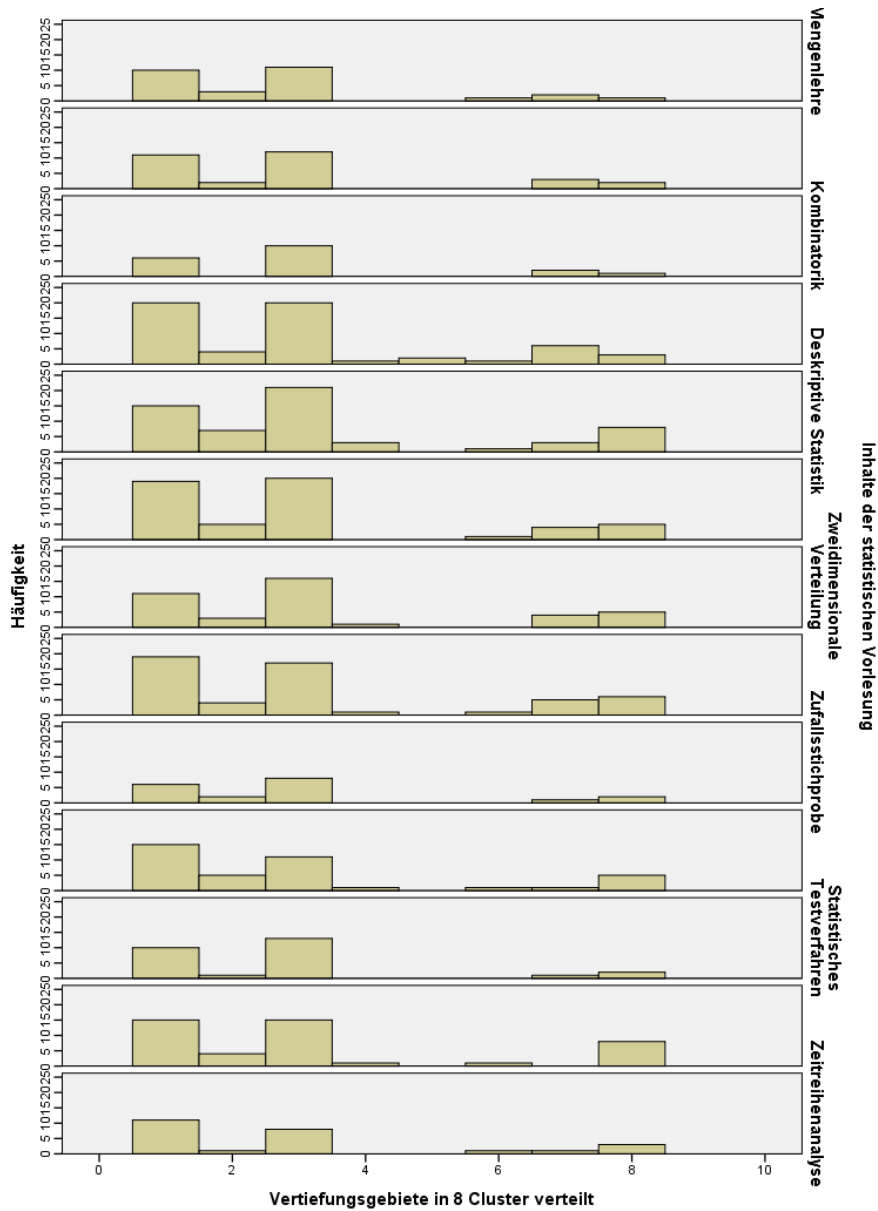
(Schätztheorie)	(Häufigkeit)
Angewandte Ökonometrie	11
Andere	7
Marketing	4
Applied Statistics	3
Econometric Methods	2
Entrepreneurship	2
Makroökonomie III	1
Externes Rechnungswesen	1
Angewandte Makroökonomie	1
Versicherungs- und Risikomanagement	1
Internationales Management	1
Bank- und Börsenwesen	1
Spieltheorie	1
Angewandte Makroökonomie	1
Monetary Economics	1
Mikroökonomie III	1
SUMME :	39

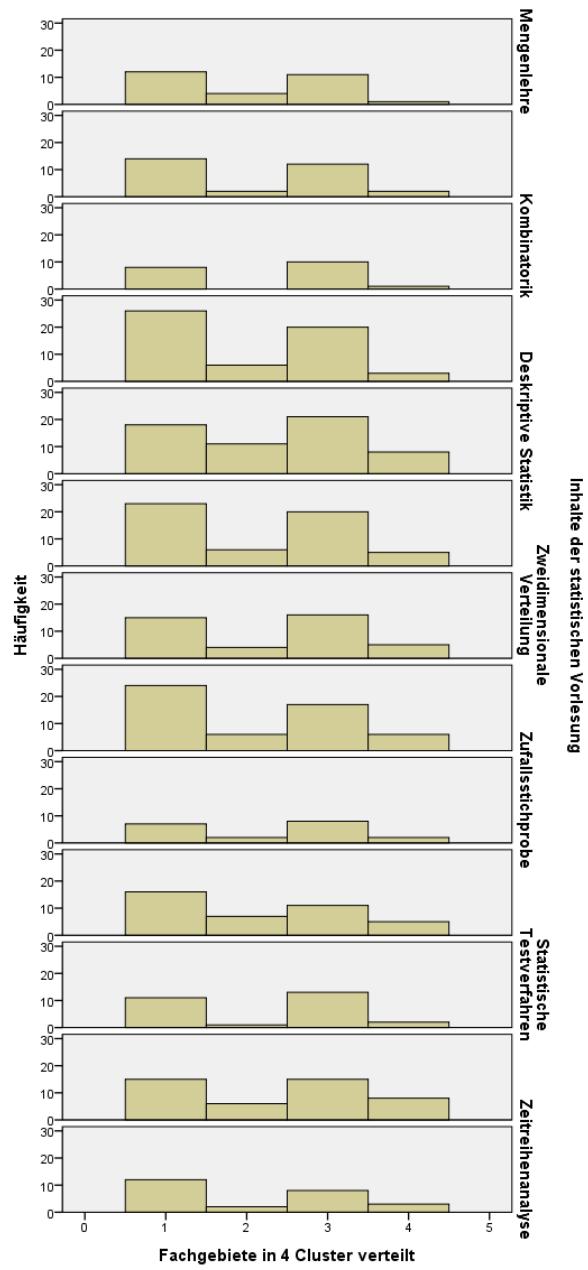
(Statistisches Testverfahren)	(Häufigkeit)
Angewandte Ökonometrie	8
Andere	8
Applied Statistics	4
Marketing	2
Econometric Methods	1
Finanzwirtschaft	1
Mikroökonomie III	1
Entrepreneurship	1
Spieltheorie	1
SUMME :	27

(Regressionsanalyse)	(Häufigkeit)
Angewandte Ökonometrie	12
Andere	10
Marketing	8
Applied Statistics	4
Econometric Methods	2
Entrepreneurship	2
Externes Rechnungswesen	1
Angewandte Makroökonomie	1
Bank- und Börsenwesen	1
Makroökonomie III	1
Monetary Economics	1
Mikroökonomie III	1
SUMME :	44

(Zeitreihenanalyse)	(Häufigkeit)
Angewandte Ökonometrie	9
Andere	4
Marketing	3
Applied Statistics	3
Angewandte Makroökonomie	1
Externes Rechnungswesen	1
Makroökonomie III	1
Finanzwirtschaft	1
Entrepreneurship	1
Mikroökonomie III	1
SUMME :	25

A.6 Häufigkeiten der verschiedenen Fachgebiet-Clusters jedes statistischen Themas





A.7 SPSS-Output der Korrespondenzanalyse

Zeilenprofile

Inhalte der statistischen Vorlesung	Fachgebiete in 4 Cluster verteilt				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Aktiver Rand
Mengenlehre	.429	.143	.393	.036	1.000
Zufallsexperimente	.467	.067	.400	.067	1.000
Kombinatorik	.421	.000	.526	.053	1.000
Wahrscheinlichkeitsrechnung	.473	.109	.364	.055	1.000
Deskriptive Statistik	.310	.190	.362	.138	1.000
Zufallsvariablen	.426	.111	.370	.093	1.000
Zweidimensionale Verteilung	.375	.100	.400	.125	1.000
Verteilungsmodelle	.453	.113	.321	.113	1.000
Zufallsstichprobe	.368	.105	.421	.105	1.000
Schätztheorie	.410	.179	.282	.128	1.000
Statistische Testverfahren	.407	.037	.481	.074	1.000
Regressionsanalyse	.341	.136	.341	.182	1.000
Zeitreihenanalyse	.480	.080	.320	.120	1.000
Masse	.409	.116	.371	.104	

Spaltenprofile

Inhalte der statistischen Vorlesung	Fachgebiete in 4 Cluster verteilt				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Aktiver Rand
Mengenlehre	.060	.070	.060	.020	.057
Zufallsexperimente	.070	.035	.066	.039	.061
Kombinatorik	.040	.000	.055	.020	.039
Wahrscheinlichkeitsrechnung	.129	.105	.110	.059	.112
Deskriptive Statistik	.090	.193	.115	.157	.118
Zufallsvariablen	.114	.105	.110	.098	.110
Zweidimensionale Verteilung	.075	.070	.088	.098	.081
Verteilungsmodelle	.119	.105	.093	.118	.108
Zufallsstichprobe	.035	.035	.044	.039	.039
Schätztheorie	.080	.123	.060	.098	.079
Statistische Testverfahren	.055	.018	.071	.039	.055
Regressionsanalyse	.075	.105	.082	.157	.090
Zeitreihenanalyse	.060	.035	.044	.059	.051
Masse	1.000	1.000	1.000	1.000	

Auswertung ³

Dimen- sion	Singulär- wert	Träg- heit	Chi- Quadrat	Sig.	Anteil der Trägheit		Singulärwert für Konfidenz	
					Bedingen	Kumuliert	Standard- abweichung	Korrelation
								2
1	.180	.032			.665	.665	.042	-.062
2	.101	.010			.208	.873	.044	
3	.079	.006			.127	1.000		
Gesamt		.049	23.970	.938a	1.000	1.000		

³a. 36 Freiheitsgrade

Übersicht über Zeilenpunkte ⁴

Fachgebiete in 4 Cluster verteilt	Masse	Wert in Dimension		Träg- heit	Beitrag				
		1	2		des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		Gesamt
					1	2	1	2	
Cluster1	.409	-.205	.241	.007	.095	.237	.435	.338	.773
Cluster2	.116	.850	.429	.019	.465	.212	.815	.116	.931
Cluster3	.371	-.247	-.240	.008	.126	.212	.507	.267	.774
Cluster4	.104	.738	-.574	.015	.314	.340	.674	.227	.901
Aktiver Gesamtwert	1.000			.049	1.000	1.000			

⁴Symmetrische Normalisierung

Übersicht über Spaltenpunkte ⁵

Fachgebiete in 4 Cluster verteilt	Masse	Wert in Dimension		Träg- heit	Beitrag				
		1	2		des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		Gesamt
					1	2	1	2	
Cluster1	.409	-.205	.241	.007	.095	.237	.435	.338	.773
Cluster2	.116	.850	.429	.019	.465	.212	.815	.116	.931
Cluster3	.371	-.247	-.240	.008	.126	.212	.507	.267	.774
Cluster4	.104	.738	-.574	.015	.314	.340	.674	.227	.901
Aktiver Gesamtwert	1.000			.049	1.000	1.000			
Gesamt									

⁵Symmetrische Normalisierung

Erklärung zur Urheberschaft

Hiermit erkläre ich, Xue Liu, dass ich die vorliegende Arbeit allein und unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe. Die Prüfungsordnung ist mir bekannt. Ich habe in meinem Studienfach bisher keine Bachelorarbeit eingereicht bzw. diese nicht endgültig bestanden.

Xue Liu,

Berlin, 21.10.2013