Fachgebiet IT in Produktion und Logistik
Fakultät Maschinenbau, Technische Universität Dortmund

Recherche und Bewertung statistischer Zusammenhangsmaße für in Supply Chains anfallenden Daten

Fachwissenschaftliche Projektarbeit
vorgelegt von
Claudia Köster, B.Sc.

Matrikelnummer: 139857
Studiengang: Master of Science im Maschinenbau

Gutachterin: Dipl.-Inf. Anne-Antonia Scheidler

Ausgegeben am: 11.04.2014
Eingereicht am: 24.11.2014

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

In Supply Chains werden große Mengen von Daten ermittelt und abgespeichert. Die Grundlage für die automatische Untersuchung der so entstehenden Datenbestände bildet die Beschreibung der zwischen ihnen bestehenden Zusammenhänge. Die deskriptive Statistik stellt hierfür eine große Anzahl von Methoden zur Verfügung, die sich unter dem Begriff „Zusammenhangsmaße“ sammeln lassen. Die Aufgabe der vorliegenden Arbeit besteht darin, eine Auswahl solcher Zusammenhangsmaße zu recherchieren, gängige Kriterien zu ihrer Einteilung darzustellen und im Bezug auf Daten aus Supply Chains weitere Kritieren zu entwickeln, die dabei helfen, für den Sachzusammenhang geeignete Maße auszuwählen. Auf dieser Grundlage werden dann für das Anwendungsfeld passende Maße aus der ermittelten Auswahl empfohlen.

Im ersten Abschnitt werden zwei Sätze von Beispieldaten aus dem Bereich der Supply Chains betrachtet. Sie stammen aus am Fachgebiet durchgeführten Abschlussarbeiten und werden auf Spezifika untersucht, die für Auswahl von Zusammenhangsmaßen interessant sein könnten. Der Hauptteil beginnt mit einer Einordnung der Zusammenhangsmaße in die Gebiete der Statistik und Erläuterungen zu ihrer Zuordnung zu Skalenniveaus, sowie einigen Hinweisen zu ihrer Interpretation. Es folgt eine Darstellung von Zusammenhangsmaßen, die in der Fachliteratur mit verschiedenen Zielgruppen gefunden werden kann. Hierbei kann keine Vollständigkeit gewährleistet werden, daher beschränkt sich der Auflistung mit einer kurzen Charakterisierung jedes gefundenen Maßen und Hinweisen zu jeweils weiterführender Literatur. Im dann folgenden Kapitel werden die Kriterien zusammengetragen, nach denen im Anwendungsfeld passende Maße ausgewählt werden können. Gemäß dieser Kriterien werden die gefundenen Maße bewertet. Im letzten Kapitel werden die Befunde der Arbeit zusammen gefasst und Anknüpfungs- und Nutzungsmöglichkeiten aufgezeigt.

2 Betrachtung von Daten in Supply Chains

Im folgenden Kapitel werden die Begriffe „Daten“ und „Supply Chain“ definiert, um eine einheitliche sachliche Verständnisgrundlage zu schaffen. Es folgen einige Erläuterungen zu Begriffen, die im Kontext von Daten und Statistik gleichermaßen relevant sind. Dann werden die Beispieldaten, die als Grundlage für die Kategorisierungsmöglichkeiten von Daten in der Supply Chain dienen, vorgestellt.

2.1 Begriffsdefinitionen

Der Begriff „Supply Chain“ ist gängig für Wertschöpfungsketten, die sämtliche Fertigungs- und Absatzstufen von der Rohstoffgewinnung bis zum Konsumenten umfassen [Kla12]. Daten sind nach [FH11] abstrahierte und computergerecht aufgearbeitete Informationen. Die Informationen beschreibt er nach einem Zitat von Claude E. Shannon als alles, was Ungewissheit beseitigt. Wenn wir also Daten in einer Supply Chain betrachten, dann befassen wir uns damit, diese für uns gewisser und besser erfassbar zu machen.

Ein „Merkmal“ ist eine interessierende Eigenschaft eines Objektes, die erfasst und untersucht werden soll. Sie wird auch statistische Variable, Untersuchungsmerkmal, Untersuchungsvariable oder Erhebungsmerkmal genannt. Gelegentlich wird sie mit dem Begriff der Variable gleichgesetzt [FKP$^+$07]. „Merkmalsträger“ sind all die Objekte, an denen interessierenden Merkmale untersucht werden können. Die Merkmalsausprägungen, oder kurz „Ausprägungen“, stellen mögliche Werte dar, die ein Merkmal annehmen kann. Datensätze repräsentieren eine Ansammlung dieser Merkmalsausprägungen.

In diesem Zusammenhang sei auch auf die Bedeutung der Begriffe „univariat“, „bivariat“ und „multivariat“ hingewiesen, die in der Regel im Zusammenhang mit Analyseverfahren auftauchen, gelegentlich aber auch als adjektive für Merkmale verwendet werden [Wie13]. [Dud07] definiert multivariat als mehrere Variablen betreffend, während [CGHB$^+$91] ihm den Begriff „m-dimensional“ gleichsetzt. [Gra04] hält zudem fest, dass der Begriff erst in neuerer Zeit Verwendung findet. Die Begriffe uni- und bivariat werden zudem bei [Bor10], [EGH$^+$94] und [VB00] nicht explizit definiert aber in den oben genannten Zusammenhängen gemäß ihrer Vorsilbe als eine bzw. zwei Variablen betreffend genutzt und können analog zur Definition bei [CGHB$^+$91] auch als ein- bzw. zweidimensional übersetzt werden, wie [Bor10, S.183] grafisch zeigt.

2.2 Vorstellung der Beispieldaten

Rein technisch kann im Rahmen einer Supply Chain selbstverständlich jede Art von Daten auftreten. Um einen Überblick zu erhalten, was üblich ist und zu welchen Themenfeldern die vorhandenen Daten gehören, werden zwei studentische Arbeiten des Fachgebiets ITPL der TU Dortmund heran gezogen. [Guh14] beschäftigt sich mit den Daten, die in der Produktionsplanung und -steuerung verarbeitet werden, während [Hil14] eine Reihe von

Fragen stellt, die bei der strategischen Ausrichtung einer Supply Chain beantwortet werden müssen. Beide Quellen zusammen zeigen also Daten zu verschiedenen Aspekten von Supply Chains auf, sodass sie einen für die Zwecke dieser Arbeit ausreichenden Überblick geben.

Die Informationen, die in den beiden vorliegenden Arbeiten betrachtet werden, werden in den Arbeiten selbst nach inhaltlichen bzw. sachlichen Krtierien unterschieden: [Guh14] kategorisiert Stamm- und Bewegungsdaten, sowie Sachzusammenhänge bezüglich der Merkmalsträger. Da [Hil14] Fragen zur strategischen Ausrichtung stellt, decken diese eine ganz Reihe von Themengebieten ab: An welchem Standort werden wie viele Produktions-, Lager- und Distributionsstätten errichtet? Welche Kapazitäten sind vorzuhalten? Welche Teile werden eingekauft, welche werden gefertigt? Die Antworten auf diese Fragen lassen sich ebenfalls in Daten ablegen. Die Sortierung hier ist wieder inhaltlicher bzw. sachlicher Natur.

3 Statistische Zusammenhangsmaße

Im folgenden Kapitel werden ausgewählte Zusammenhangsmaße aus der Statistik vorgestellt. Die in dieser Arbeit betrachteten Maße lassen sich alle dem Gebiet der deskriptiven Statistik zuordnen, welches im ersten Abschnitt von den weiteren Gebieten der Statistik abgegrenzt wird. Abschnitt 3.2 enthält die im Kontext der Maßzahlen notwendigen Begriffsdefinitionen, bevor die einzelnen Maße genannt, beschrieben und eingeordnet werden. Sowohl bei den Gebieten der Statistik, den Skalenniveaus, wie auch den einzelnen Maßzahlen wird darauf geachtet, die Vielzahl der Synonyme darzustellen und eine Bezeichnung als Konvention im Rahmen dieser Arbeit kenntlich zu machen.

3.1 Einordnung

Die folgenden Abschnitten werden die Gebiete der Statistik und die verschiedenen existierenden Skalenniveaus vorstellen. Anschließend finden sich Hinweise zur Korrelation und Kausalität und der Definition von Fehlerreduktionsmaßen, die dabei helfen sollen, die anschließend aufgezählten Maße korrekt interpretieren und nutzen zu können.

3.1.1 Gebiete der Statistik

Die Literatur unterscheidet durchgängig drei Gebiete der Statistik, die verschiedene Methoden nutzen um jeweils andere Ziele zu erreichen. Einen Überblick über die zahlreichen in Verwendung befindlichen Synonyme für diese drei Felder bietet Abb. 3.1. Der jeweils zuoberst dargestellte Begriff wird im Rahmen der Arbeit verwendet.

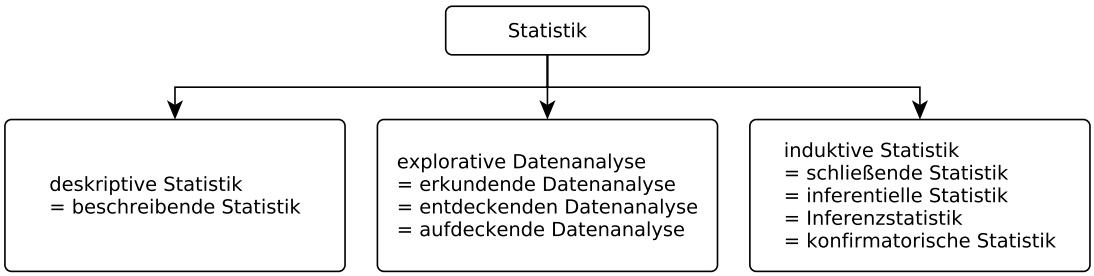

Abb. 3.1: Die Gebiete der Statistik mit ihren synonymen Bezeichnungen

Das erste Gebiet ist die *deskriptive Statistik*. Ihre Verfahren dienen der Erhebung, Aufbereitung und Auswertung von Daten und Beochbachtungsergebnissen [Gra04]. Ziel ist dabei die Systematisierung und anschließend die Gewinnung wesentlicher Informationen bezüglich der Gesamtheit der Daten und das Zeigen der Verhältnisse verschiedener Gesamtheiten zueinander [EGH+94]. Die Daten werden dabei in der Regel komprimiert, mittels weniger, aussagekräftiger Maßzahlen charakterisiert und graphisch und tabellarisch aufbereitet [FKP+07].

Die *explorative Datenanalyse* dient der Findung von Hypothesen und Modellen. Die verfügbare Datenmenge wird dafür systematisch oder versuchsweise umgestaltet, transformiert und reduziert. Inbesondere sollen dadurch „Strukturen, Muster und einfache (...) Zusammenhänge“ [EGH+94] bzw. „Strukturen und Besonderheiten in den Daten“ [FKP+07] aufgedeckt werden. Dazu werden modifizierte, erweiterte und verfeinerte Methoden der deskriptiven Statistik, sowie eine Vielzahl von graphischen Verfahren genutzt. Da durch die Verfahren Modelle und Hypothesen geliefert werden sollen, die dann noch Bestätigung durch die induktive Statistik benötigen oder die zu untersuchende Fragestellung genauer geklärt werden soll, werden vor der Analyse keine Modellannahmen getroffen [EGH+94, FKP+07]. [VB00] stellt einen Zusammenhang zwischen dem Data Mining und der explorativen Datenanalyse her. Hierin wird angemerkt, dass die Verfahren sich ähneln, die betrachteten Datenmengen aber beim Data Mining erheblich größer sind.

In der *induktiven Statistik* werden Feststellungen, die auf Stichproben basieren mittels mathematischer Methoden und Sätze auf die Grundgesamtheit übertragen [EGH+94, FKP+07, Gra04]. Hierzu werden Wahrscheinlichkeitsmodelle genutzt. Im Einzelnen liegen Methoden zur Auswahl von Stichproben, Hochrechnung, Schätzung und statistischen Tests vor [EGH+94]. Als Voraussetzungen für eine induktive Bearbeitung der Daten nennt [FKP+07] die sorgfältige Versuchsplanung, sowie deskriptive und explorative Analysen.

3.1.2 Arten von Skalenniveaus

Je nachdem, wie eine Ausprägung erfasst wird, ist sie einer *Skala* zuweisbar. Dies ist eine Art zur Einteilung von Merkmalsausprägungen [Gra04, S.220], die nach Gemeinsamkeiten von Objekten sucht und jeden diesbezüglich beobachteten Wert dann einem oder mehreren Zeichen zuordnet. Die Zeichen bilden die Skala, an denen auch ablesbar ist, welche Rechenoperationen sinnvoll am Objekt durchgeführt werden können. Diese Art der Einteilung macht sie für den Zweck der Zusammenhangserkennung wertvoll. Der Vorgang der Zuordnung von Werten zu Beobachtungen bei einem Merkmal wird als *Skalierung* oder Messung bezeichnet. Die einzelnen *Skalenniveaus* unterscheiden sich nach den Eigenschaften, die den betrachteten Ausprägungen zugeschrieben werden können. Unterschiede bestehen darin, welche Wertebereiche die Ausprägungen abdecken, ob man verschiedenen möglichen Ausprägungen eine sinnvolle Reihenfolge geben kann und in welchem Abstand sie auftauchen. Abb. 3.2 zeigt die Aufteilung des Begriffs Skala, wie sie unter anderem [Wie13, Gra04, Bor10] entnommen werden kann. In den unterschiedlichen Quellen werden dabei teilweise unterschiedliche Detaillierungsgrade angegeben. [CGHB+91] benennt neben den eindimensionalen Skalen, die in dieser Arbeit und in den meisten anderen Quellen ausschließlich betrachtet werden, noch die Möglichkeit, mehrdimensionale Skalen zu bilden, falls einem Objekt mehrere Zeichen zugeordnet werden. [FKP+07], [Hä87], [Bos97] und [Bos98] unterscheiden zwar Nominal- und Ordinalskala, die beiden metrischen Skalen bezeichnen sie aber nur mit dem zusammenfassenden Begriff Kardinalskala oder metrische Skala. Da im Zusammenhang mit dem Arbeitsthema eine große Anzahl von Begriffen in der Literatur zu finden ist, die für identische Sachverhalte stehen, ordnet Abbildung 3.2 alle gefundenen Begriffe ein und verdeutlicht synonyme Verwendungen durch ein Gleich-

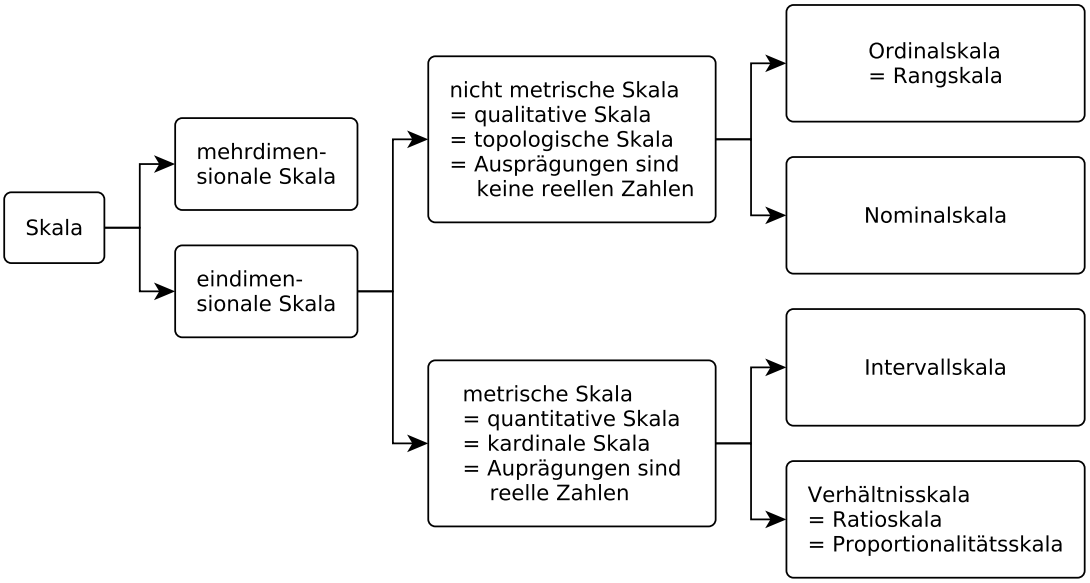

Abb. 3.2: Hierarchie der verschiedenen Skalenarten. Synonyme sind durch ein „=“ verbunden.

heitszeichen vor dem Begriff. Der in jedem Baumknoten zuerst genannte Begriff wird im Rahmen dieser Arbeit in der Regel für den Sachverhalt verwendet.

Folgt man dem in Abb. 3.2 gezeigten Baum von der Wurzel links weiter nach rechts, sind die Begriffe Skala, sowie ein- und mehrdimensionale Skala bereits erläutert worden. Die eindimensionalen Skalen werden weiterhin danach unterteilt, ob auf ihnen metrische und damit quantitative Werte angeordnet werden oder nicht metrische und damit qualitative Merkmale darauf zu finden sind. [Bor10] zeigt in seiner sehr ausführlichen Darstellung für die vier unten genannten Skalenarten jeweils eine Reihe von Axiomen, mit deren Hilfe die Zugehörigkeit eines Merkmals zu einer Skala auf seine Richtigkeit überprüft werden kann. Im Falle eines Zweifels an einer solchen Zuordnung, sei an dieser Stelle darauf verwiesen.

Die qualitativen Skalen weisen weniger mögliche Rechenoperationen auf, daher werden sie als geringerwertig angesehen: Die nominale Skala weißt Ausprägungen nur das Attribut der Unterschiedlichkeit zu. Es kann keine Aussage über Wertigkeiten der Ausprägungen zueinander getroffen werden. Beispiele hierfür ist die Nationalität oder das Geschlecht von Personen. Merkmale, die hier angesiedelt sind, können ausschließlich nach ihrer Häufigkeit untersucht werden [Wie13]. Erlaubte Rechenoperationen sind $=$ und $\neq$ [Gra04],[Run10].

Die ordinale Skala baut auf der nominalen auf, indem sie eine Rangfolge zwischen den Ausprägungen herstellt. Die Abstände zwischen den jeweils benachbart angeordneten Ausprägungen sind dabei nicht näher bestimmt und in der Regel als unterschiedlich anzunehmen. Beispiele für ordinale Werte sind Ergebnislisten bei Wettkämpfen oder die Altersreihenfolge von Geschwistern. Die Skala ermöglicht neben den Häufigkeitsanalysen, die schon für nominale Merkmale möglich war, die Untersuchung von Rangmerkmalen wie z.B. des Medians [Wie13]. Hinzukommende mögliche Rechenoperationen sind $>$ und $<$ [Gra04],[Run10].

Für die Kategorie der metrischen Skalen existieren wiederum zwei Vertreter. Auch hier gilt: Die Rechenoperationen, die für Merkmale niedrigerer Skalenniveaus durchführbar sind, sind grundsätzlich auch für Merkmale der höheren Niveaus anwendbar, werden

jedoch um jeweils spezifische weitere ergänzt. Die niedriger anzusiedelnde der beiden metrischen Skalen ist die Intervallskala: Sie ordnet unterschiedliche Merkmale in einer wertenden Reihenfolge an, deren Abstände definiert sind. Ab dieser Skala ist die Berechnung eines arithmetischen Mittels sinnvoll. Eine mögliche hier einzuordnende Größe ist die Temperatur in Grad Celsius [Wie13]. [Gra04] und [Run10] nennen als neue mögliche Operationen $+$ und $-$.

Zuletzt wird die Verhältnisskala eingeordnet. Zusätzlich zu den vorhergehenden Eigenschaften bieten Merkmale auf diesem Niveau einen fixen Nullpunkt, der sinnvolle Aussagen über die Verhältnisse zweier Ausprägungen zulässt. Als Beispiel wären hier Geschwindigkeit oder Lebensalter zu nennen. Da Verhältnisse zugelassen sind, nennen [Gra04] und [Run10] die Division und Multiplikation als erweiterte mögliche Rechenarten für Merkmale dieses Skalenniveaus.

3.1.3 Nomenklatur von Zusammenhangsmaßen

Die Begriffe *Zusammenhangsmaß* und *Assoziationsmaß* werden von [VB00] beide als Maßzahl, die über die Stärke eines Zusammenhangs von Merkmalen Auskunft gibt, definiert. Auch [HEK09] setzt die Begriffe gleich, spricht jedoch davon, dass hiermit die Abhängigkeit einer Kontingenztafel charakterisiert wird. [Gra04] schränkt ein, dass mindestens eine der betrachteten Größen nominalskaliert sein muss. Diese Einschränkung erscheint aufgrund der in Abschnitt 3.1.2 vorgestellten Abwärtskompabilität der einzelnen Skalenarten wenig sinnvoll zu sein. [Ben07] geht in einem einleitenden Abschnitt auf die Ungereimtheiten in der Bennenung der einzelnen Gruppen von Maßzahlen ein und schließt sich der Gruppe von Autoren an, welche die Begriffe Kontingenz, Assoziation und Korrelation als synonym betrachten. Die vorliegende Arbeit orientiert sich an dieser Ansicht und verwendet den grundsätzlich neutralen Begriff „Zusammenhangsmaß“. Durch eine *Maßzahl* wird allgemein ein bestimmter Sachverhalt aus einer Verteilung von Werten quantitativ charakterisiert [VB00, HEK09]. [EGH$^+$94] fügt hinzu, dass diese Zahl unter Beachtung der Zahlen- und Sachlogik ermittelt, berechnet oder geschätzt werden kann.

3.1.4 Kausalität von Korrelationen

Insbesondere bei den im Folgenden betrachteten asymmetrischen Maßen gilt es zu beachten, dass ein hoher Wert eines Zusammenhangsmaßes nicht damit gleichgesetzt werden darf, dass zwischen den betrachteten Größen ein *kausaler* Zusammenhang besteht. Es sollte immer im Hinterkopf behalten werden, dass dies nur eine Möglichkeit ist, die Variablen beispielsweise aber auch gemeinsam von einer dritten Größe abhängen könnten. Diese Tatsache sollte man sich gerade dann noch einmal vergegenwärtigen, wenn hohe auftretende Werte eine vorher gemachte Annahme scheinbar bestätigen [Bor10, S.159f.], [CK14, S.113], [FKP$^+$07, S.148ff.], [SH09, S.90].

3.1.5 Proportionale Fehlerreduktionsmaße

[VB00] verweist auf [Ben07], wenn es um das Thema der proportionalen Fehlerreduktionsmaße geht. Auch in anderer Literatur findet sich das Thema sehr selten, so z.B.

bei [Bor10, S.157f.], der ein Maß in gleicher Art beschreibt, aber den Begriff der *relativen Fehlerreduktion* benutzt. Außerdem existiert noch der Begriff „Maß der prädikativen Assoziation“ für diese Gruppe [Ben07]. Grundsätzlich sind die Maß so aufgebaut, dass sie eine Aussage dazu machen, um wie viel Prozent die Vorhersagegenauigkeit bezüglich einer Variable steigt, wenn Informationen über die Verteilung einer passenden unabhängigen Variable vorliegen. Diese Herangehensweise birgt den Vorteil, dass die Maße auf allen Skalenniveaus definiert werden können und jeweils eine vergleichbare Aussage erlauben. In der folgenden Aufzählung sind PRE-Maße (vom englischen *proportional reduction of error* abgeleitet) entsprechend gekennzeichnet. Für sie werden vier charakteristische Größen angegeben:

1. Welche Vorhersage über die abhängige Variable wird getroffen, wenn keine näheren Informationen zur Verteilung der Unabhängigen vorliegen?
2. Welche Vorhersage kann getroffen werden, sobald Informationen zu Verteilung der unabhängigen Variable vorliegt?
3. Wie wird der Begriff „Fehler“ im vorliegenden Zusammenhang definiert?
4. Wie ist die spezifische Ausprägung der allgemeinen PRE-Maßregel?

Die allgemeine Maßregel für PRE wird bei [Ben07, S.92] definiert als

$$PRE = \frac{(Fehler\ nach\ Regel\ 1) - (Fehler\ nach\ Regel\ 2)}{Fehler\ nach\ Regel\ 1} \tag{3.1}$$

$$= \frac{E_1 - E_2}{E_1} \tag{3.2}$$

3.2 Maße

Die im Folgenden betrachteten Maße können nur als eine Auswahl aus der Vielzahl der vielen statistischen Maßzahlen, die zur Beschreibung von Datenmengen existieren, angesehen werden. Analog zum bisherigen Vorgehen innerhalb der Arbeit und der Anordnung z.B. bei [FKP$^+$07] oder [Ben07] werden die Maße jeweils passend zur Skalierung der Daten, die mit Ihnen untersucht werden können, vorgestellt. Die Reihenfolge ist zu großen Teilen [Ben07] entnommen und betrachtet zunächst alle Maße für bivariate Verteilungen und im letzten Abschnitt Verfahren für multivariate Analysen. Vorausgehend wird eine allgemein Erklärung des Grundgedankens der proportionalen Fehlerreduktionsmaße gegeben, da diese Maße entlang aller Skalen auftauchen und so eine Wiederholung der Charakteristika vermieden werden kann.

3.2.1 Maße für zwei nominalskalierte Größen

Wie bereits in Abschnitt 3.1.2 beschrieben, kann man bei nominalskalierten Größen die Häufigkeiten des Auftretens verschiedener Ausprägungskombinationen betrachten und zur Auswertung heran ziehen. Auf dieser Grundlage wird in der Regel zunächst eine

	x_1	x_2	
y_1	a	b	$n_{1.}$
y_2	c	d	$n_{2.}$
	$n_{.1}$	$n_{.2}$	N

Tab. 3.1: 2x2-Tabelle mit gängigen Platzhaltern für die jeweiligen Zelleninhalte

Kontingenztafel erstellt. Vorgehen und Aufbau dieser Tafel lässt sich sehr kleinschrittig in [Ben07, S.69 ff.] nachvollziehen. Ebenfalls übersichtliche Darstellungen finden sich in [Kü01, S.300] und [FKP+07, S.109 ff.].

Tabelle 3.1 zeigt eine 2x2-Tabelle mit gängigen Platzhaltern für die jeweiligen Zelleninhalte. In der ersten Zeile und der ersten Spalte stehen x_i und y_j für Ausprägungen der Merkmale X und Y. Die Werte a bis d, auch h_{ij} genannt, in den mittleren Feldern stellen absolute oder relative Häufigkeiten dar, die für die jeweilige Kombination von Merkmalsausprägungen ermittelt wurde. Die Randhäufigkeiten werden mit $n_{j.}$ und $n_{.i}$ gekennzeichnet und N steht für die Summe der Randhäufigkeiten und damit je nach Häufigkeitsart für die Gesamtzahl der untersuchten Stichproben oder 100%. Die Tabelle dient als Referenz für die im Laufe der folgenden Abschnitte genannten mathematischen Darstellungen der Maße.

[Ben07, S.69ff.], [CK14, S.92ff.], [FKP+07, S.109ff.], [Ste13, S.47ff.]

Prozentsatzdifferenz

Die Prozentsatzdifferenz bietet ein einfaches Maß zum Vergleich konditionaler Häufigkeiten. Sie subtrahiert Prozentwerte voneinander und gibt daher an, wie viele Prozentpunkte der Unterschied zwischen den Zugehörigkeit zweier Ausprägungen der Variable X zu einer Ausprägung des Merkmals Y ausmacht. Sind bereits bedingte relative Häufigkeiten angegeben, so berechnet sie sich als Differenz der beiden zu vergleichenden Prozentwerte. Sind absolute Zahlen in der Tabelle angegeben, lautet die Formel

$$d\% = 100 \left(\frac{a}{a+c} - \frac{b}{b+d} \right) \tag{3.3}$$

$$= \frac{100(ad - bc)}{(a+c)(b+d)} \tag{3.4}$$

solange die Schichtung der Prozentwerte entlang des Merkmals Y vorgenommen wurde. Sein Wertebereich stellt eine Ausnahme dar, da er sich zwischen 0 und 100 bewegt, während sonst übliche Spannen zwischen 0 und 1 bzw. -1 und +1 liegen. Eine Umrechnung kann zwar einfach erfolgen, ist aber unüblich. Die Prozentsatzdifferenz gilt als einfach verständliches Maß, ist aber in seiner Anwendung auf größere Tabellen als 2x2 beschränkt, da dann mehr als nur eine Differenz berechnet werden muss, um alle Beziehungen abzubilden.

[Ben07, S.99ff.]

Relative Chance

Die relativen Chancen, auch Kreuzproduktverhältnis oder Odds-ratio, sind ein einfaches Maß, dass für eine 2x2-Tabelle zu einem Wert führt, sobald die Tabelle aber größer wird, mehrere Zahlenwerte beinhalten würde. Zunächst wird die Chance errechnet, in-

nerhalb einer Merkmalsausprägung x_i für das Merkmal Y zur Ausprägung y_1 oder y_2 zu gehören. Danach ist die relative Chance das Verhältnis der Chancen von x_1 und x_2.

$$\gamma(1,2|x_1) = \frac{a}{b} \tag{3.5}$$

$$\gamma(1,2|x_1,x_2) = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{3.6}$$

[FKP+07, S.119f.], [Ste13, S.81]

Die quadratische Kontingenz χ^2

Für die quadratische Kontingenz χ^2, auch χ^2-Koeffizient genannt, und gelegentlich mit dem Formelbuchstaben K bezeichnet wird die vorgefundene Besetzung der Tabellenzellen mit der Besetzung verglichen, die bei Unabhängigkeit der beiden Merkmale aufgrund der Randverteilung zu erwarten wäre. [Bor10, S.180] weist darauf hin, dass mit dem χ^2-Unabhängigkeitstest die Nullhypothese überprüft wird, dass zwei nominale Variablen voneinander unabhängig sind.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{h_{i\cdot}h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot}h_{\cdot j}}{n}} \tag{3.7}$$

In dieser Art und Weise kann der Wert χ^2 nicht als überzeugendes Zusammenhangsmaß angesehen werden, da es mit der Größe der betrachteten Tabelle proportional ansteigt und nicht auf Werte zwischen 0 und 1 oder -1 und +1 skaliert wird. Dadurch wird der sinnvolle Vergleich verschiedener Werte verhindert. Erwähnung findet das Maß in der Literatur und an dieser Stelle deshalb, weil es die Grundlage für viele Maße darstellt, die in folgenden Abschnitten erwähnt werden.

Gemäß [Ben07, S.121] ist die Interpretation der vielen auf χ^2 basierenden Maße schwierig, da sie scheinbar willkürlich zusammen gesetzt sind. Dies hat jedoch nicht verhindert, dass sie sich in den Anwendungsgebieten der Statistik großer Beliebtheit erfreuen, wie die lange Liste der Quellen für χ^2 und den von ihm abgeleiteten Maßen zeigt.

[Ben07, S.100ff.], [Bor10, S.138ff.], [CK14, S.97ff.], [FKP+07, S.122f.], [Wie13, Stichwort *Chi-Quadrat-Test*], [Ste13, S.51ff.], [EGH+94]

Φ-Koeffizient

Die naheliegendste Lösung, um dem proportionalen Wachstum von χ^2 entgegenzuwirken ist die Division mit N. Leider löst dies, sowie die anschließende Ziehung der Wurzel aus dem Term das Problem nur für 2x2-Tabellen.

$$\Phi^2 = \frac{\chi^2}{N} \tag{3.8}$$

$$\Phi = \sqrt{\frac{\chi^2}{N}} \tag{3.9}$$

Sobald diese Größe überschritten wird, können wieder Werte >1 erreicht werden. In diesen Grenzen ist der Φ-Koeffizient zwischen -1 und +1 ein mögliches Maß. [Ben07, S.111ff.], [Bor10, S.174], [FKP+07, S.140f.]

T nach Tschuprow

Ein Maß, das die Problematik von Φ lösen möchte ist das nach Tschuprow benannte T. Hier wird die Anzahl der Zeilen und Spalten hinzugenommen, um eine Steigerung des Wertes über 1 hinaus zu verhindern.

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(r-1)(c-1)}}} \tag{3.10}$$

Nachteil wiederum dieser Methode ist, dass der Wert 1 selbst in den Tabellen, die größer als 2x2 sind, nicht mehr erreicht werden kann, was die Einschätzung des Ergebnisses erschwert und dazu führt, dass T laut Aussage von [Ben07] keine Rolle in der Sozialforschung spielt. Da dieses Maß in keiner anderen gefundenen Quelle auftaucht, kann für andere Anwendungsgebiete der Statistik gleiches vermutet werden. [Ben07, S.113]

Cramers V

Das von Cramer vorgeschlagene Maß V verfolgt einen ähnlichen Ansatz wie T, nimmt jedoch nicht die Zahl der Zeilen und Spalten in der Rechnung auf, sondern nur den kleineren der beiden Werte. Dadurch wird es möglich, verschiedene Tabellen zu vergleichen. Eine andere Bezeichnung für das Maß ist „Cramer Index *CI*“.

$$V = \sqrt{\frac{\chi^2}{N\ min(r-1, c-1)}} \tag{3.11}$$

Quellen: [Ben07, S.113], [Bor10, S.180]

Kontingenzkoeffizient C und korrigierter Kontingenzkoeffizient C_{korr} nach Pearson

Pearsons C ist das älteste auf χ^2 basierende Maß und kann für beliebige Tabellengrößen berechnet werden. In einigen Quellen wie z.B. [EGH+94], [Ste13] und [FKP+07], wird als Formelbuchstabe K gewählt. Der Wert 0 wird hierbei angenommen, wenn keine Beziehung besteht, den oberen Wert 1 kann das Maß aber mathematisch nicht erreichen.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{3.12}$$

Der Maximalwert des Maßes ist definiert als C_{max}. Da dieser Maximalwert für Tabellen mir einer verschiedenen Anzahl von Reihen r und Spalten c variiert, ist ein sinnvoller

Vergleich zwischen den C-Werten der verschiedenen Tabellen nicht möglich. Um diesen Umstand zu korrigieren, wird C_{korr} definiert als

$$C_{korr} = \frac{C}{C_{max}} \tag{3.13}$$

$$mit\ C_{max} = \sqrt{\frac{k-1}{k}} \quad k = min(r, c) \tag{3.14}$$

Der Umstand, dass es sich hier um einen korrigierten Koeffizienten handelt, wird in einigen Quellen wie z.B. [CK14] und [FKP$^+$07] mit einem Stern, der hoch- oder tiefgestellt dem Formelbuchstaben angehängt wird, gekennzeichnet. Als Bezeichnung wird auch „normierter Kontingenzkoeffizient“ verwendet. Der Wertebereich leigt auch hier zwischen 0 und 1 und zeigt somit lediglich die Stärke des Zusammenhangs, nicht aber seine Richtung an. Eine eindeutige Interpretation des Maßes ist nur in Spezialfällen wie C = 1 gegeben. [Ben07, S.117ff.], [Bor10, S.180], [CK14, S.103], [Ste13, S.52]

Goodman und Krukals λ (PRE)

Dieses Maß wird auch als „Guttman's coefficient of (relative) predictability“ g bezeichnet [Ben07, 122]. Es ist je nach Ausprägung symmetrisch oder asymmetrisch, kennt keine Restriktion der Tabellengröße, nimmt Werte zwischen 0 und 1 an und bietet als PRE-Maß eine klare Interpretationsmöglichkeit, wie in Abschnitt 3.1.5 beschrieben wurde. Seine Spezifikationen als PRE-Maß lauten wie folgt:

1. Identifikation der einen Kategorie des abhängigen Merkmals mit den meisten summierten Einträgen anhand der Randhäufigkeiten (sog. Modalkategorie). Ohne weitere Informationen ist die sicherste Voraussage, die ohne Vorkenntnisse gemacht werden kann, die, alle abhängigen Variblen als Angehörigen dieser Kategorie vorherzusagen.

2. Für jede unabhängige Variable wird eine spezifische abhängige Modalkategorie vorhergesagt.

3. Jeder von der jeweiligen Vorhersage abweichender Wert ist ein Fehler.

4. λ_r und λ_c unterscheiden sich lediglich darin, welche Variable als abhängig und unabhängig angesehen wird, während λ_s die beiden Werte mittelt um so ein symmetrisches Maß zu erzeugen. Die jeweils spezifischen Formeln sind:

$$\lambda_r = \frac{\sum_{j=1}^{c} max(n_j) - max(n_{i\cdot})}{N - max(n_{i\cdot})} \tag{3.15}$$

$$\lambda_c = \frac{\sum_{i=1}^{r} max(n_i) - max(n_{\cdot j})}{N - max(n_{\cdot j})} \tag{3.16}$$

$$\lambda_s = \frac{\sum_{j=1}^{c} max(n_j) + \sum_{i=1}^{r} max(n_i) - max(n_{i\cdot} - max(n_{\cdot j}))}{2N - max(n_{i\cdot} - max(n_{\cdot j})} \tag{3.17}$$

1.	konkordante Paare	N_c
2.	diskonkordante Paare	N_d
3.	in X verknüpfte und in Y verschiedene Paare	T_x
4.	in Y verknüpfte und in X verschiedene Paare	T_y
5.	in X und Y verknüpfte Paare	T_{xy}

Tab. 3.2: Arten von Paarbeziehungen ordinalen Merkmalen und zugeordnete Formelzeichen

3.2.2 Maße für zwei ordinalskalierte Größen

Um Zusammenhänge zwischen zwei ordinalen Maßen zu überprüfen, sind zwei grundlegende Vorgehensweisen zu unterscheiden. Die erste Gruppe von vorgestellten Maßen zählt die Anzahl der gleich- und gegensinnig gerichteten Paare und evtl. auch die der sich nicht unterscheidenden und fasst diese in zu einer Zahl zusammen. Die verschiedenen Arten von Paaren werden als konkordant, diskonkordant und verknüpft oder nach der englischen Bezeichnung tied genannt. Zusätzlich zu diesen drei Arten gibt es noch solche Paare bei denen eine der beiden Variablen bei den untersuchten Merkmalsträgern verschieden ist und eine verknüpft. Tabelle 3.2 zeigt die verschiedenen Paararten und ihren in der Regel verwendeten Formelbuchstaben. Die Formeln 3.18 bis 3.20 zeigt den Zusammenhang zur Gesamtzahl der Paare in der Untersuchung.

$$N_{gesamt} = \sum N_c + N_d + T_x + T_y + T_{xy} \tag{3.18}$$

$$= \frac{N(N-1)}{2} \tag{3.19}$$

$$= \binom{N}{2} \tag{3.20}$$

Die zweite Gruppe von Maßen baut auf den Rangfolgen auf, die sich mit den Ergebnissen der beiden betrachteten Variablen bilden lassen. Abweichungen liegen vor, sobald ein Merkmalsträger innerhalb der Rangfolge zum Merkmal X an einem anderen Platz befindet als in der Rangfolge zum Merkmal Y. Von verknüpften Paaren ist analog zur ersten Methode auch hier von solchen Merkmalsträgern die Rede, die in beiden Folgen auf dem gleichen Rang liegen. Zu beachten bei dieser zweiten Möglichkeit ist, dass die eigentlich ordinalen Daten als intervallskaliert betrachtet werden, da mit diesen ja numerische Berechnungen durchgeführt werden. [FKP$^+$07] ordnet das Maß auch den Maßzahlen für metrische Variablen zu. Durch beide Vorgehensweisen lässt sich in der Regel nicht nur feststellen, welche Größe auf einer Skala von 0 bis 1 ein Zusammanhang hat, sondern auch in welche Richtung dieser Zusammanhang auf einer Skala von -1 für eine perfekt negative oder gegensinnige Beziehung über 0 für eine nicht vorhandene Beziehung bis hin zu +1 für eine perfekt positive oder gleichsinnige Beziehung weist. Ordinale Variablen werden oft genutzt, um als Indikatoren für schwerer messbare Umstände zu dienen. Aus diesem Grund kommt Ihnen häufig eine besondere Bedeutung zu [Ben07, S.139].

Kendalls τ_a, τ_b und τ_c

$$\tau_a = \frac{N_{konkordant} - N_{diskonkordant}}{\frac{N(N-1)}{2}} \tag{3.21}$$

$$\tau_b = \frac{N_{konkordant} - N_{diskonkordant}}{\sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}} \tag{3.22}$$

$$\tau_c = \frac{N_{konkordant} - N_{diskonkordant}}{\frac{1}{2}N^2 \left(\frac{min(r,c)-1}{min(r,c)}\right)} \tag{3.23}$$

τ_a bezieht ausschließlich die konkordanten und diskonkordanten Paare in einer Untersuchung mit ein und ist dementsprechend unempfindlich gegenüber verknüpften Paaren. Ihre Anzahl hat keinen Einfluss auf das Maß. Gibt es in einer Untersuchung keine verknüpften Paare, nehmen alle τ-Maße denselben Wert an. Dieser entspricht dann immer dem von τ_a. τ_b bezieht zwar verknüpfte Paare mit ein, kann die Extremwerte -1 und +1 aber nur erreichen, wenn die zu untersuchende Tabelle quadratisch ist. τ_c kann im Gegensatz dazu für alle Tabellengrößen angewendet werden und nimmt bei gleicher Datenausgangslage in der Regel den höchsten Wert der drei Maße an. Hierin liegt ein Grund für seine Beliebtheit in der Forschung, wenn hohe Zusammanhangswerte wünschenswert sind. Eine klare Interpretation dafür, was die Werte spezifisch ausdrücken, konnte nicht gefunden werden.

[Ben07, S.149ff.]

Goodman und Kruskals γ (PRE)

Goodman und Kruskals γ kann für beliebig große Tabellen berechnet werden und ist symmetrisch. Er beachtet lediglich konkordante und diskonkordante Paare, während Verknüpfungen ignoriert werden. Dadurch hat es in der Regel einen höheren Wert als τ_a und τ_b, wodurch es laut [Ben07, S.163] sehr schnell Popularität in den Sozialwissenschaften erhalten hat. Der Umstand führt auch dazu, dass es durch die Anzahl der Variablenausprägungen sehr stark beeinträchtigt wird und sich durch Zusammenfassen der Werte zu größeren Klassen sehr stark beeinflussen lässt. Es ist heute kaum noch gebräuchlich. Gammas vier Merkmale als PRE-Maß sind

1. Für jede Abhängige wird vorausgesagt, dass sie im Bezug auf ihre Unabhängige größer ist. Verknüpfte Paare werden außer acht gelassen.

2. Ist γ positiv, werden konkordante Paare vorhergesagt, ist γ negativ, werden diskonkordante Paare vorhergesagt. Verknüpfte Paare werden wiederum außer acht gelassen.

3. Ein Fehler liegt vor, wenn die vorhergesagte Paarbeziehung von der beobachteten abweicht.

4. Die Formel zur Berechnung ergibt sich wie Formel 3.24 bis 3.28 zeigen. Dadurch ergeben sich positive Werte, wenn mehr konkordante als diskonkordante Paare vorliegen und ein negativer Wert, wenn mehr diskonkordante Paare existieren.

$$E_1 = 0.5(N_c + N_d) \tag{3.24}$$

$$E_2 = min(N_c, N_d) \tag{3.25}$$

$$\gamma = \frac{E_1 - E_2}{E_1} \tag{3.26}$$

$$\gamma = \frac{0.5(N_c + N_d) - min(N_c, N_d)}{0.5(N_c + N_d)} \tag{3.27}$$

$$\gamma = \frac{N_c - N_d}{N_c + N_d}, \quad wenn\ N_c > N_d \tag{3.28}$$

Quellen: [Ben07, S.161ff.]

Maße nach Somer

Somer hat seine zwei asymmetrischen Maße d_{xy} und d_{yx} vorgestellt, die sowohl konkordante und diskonkordante, als auch verknüpfte Paare beachten. Der erste Index bezeichnet dabei immer die als abhängig betrachtete Variable. Somers erklärte die Zusammensetzung so, dass es sich hierbei im Prinzip um Goodman und Kruskals γ handele, jedoch um eine „Strafe“ für verknüpfte Paare erweitert. Selten wird auch eine symmetrisierte Variante d_s des Maßes genutzt. Es existieren für Somers Maße auch Interpretationen als PRE, die aber in [Ben07] nicht näher ausgeführt werden. Formel 3.32 zeigt den bestehenden Zusammenhang zwischen den Maßen von Somer und τ_b.

$$d_{xy} = \frac{N_c - N - d}{N_c + N_d + T_x} \tag{3.29}$$

$$d_{yx} = \frac{N_c - N - d}{N_c + N_d + T_y} \tag{3.30}$$

$$d_s = \frac{N_c - N_d}{N_c + N_d + \frac{1}{2}(T_y + T_x)} \tag{3.31}$$

$$\tau_b = \pm\sqrt{d_{xy} d_{yx}} \tag{3.32}$$

Quellen: [Ben07, S.168ff.]

Rangkorrelation nach Spearman

Die Rangkorrelation nach Spearman, abgekürzt r_s oder ρ, ist einer Vertreter der zweiten Gruppe von Maßen zur Beschreibung von Zusammenhängen zwischen zwei Ordinalmerkmalen. Sie ist mit der Produkt-Moment-Korrelation aus Abschnitt 3.2.3 identisch, wenn beide Merkmale dort die Werte 1 bis n annehmen, wie es in Rangreihen der Fall ist. Die erreichbaren Werte des Maßes liegen zwischen -1 und +1.

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \tag{3.33}$$

Die Größe d_i^2 steht für die Differenz der Rangplätze, die ein Merkmalsträger innerhalb der Rangfolgen für die Merkmale X und Y inne hat. Durch die Quadrierung werden große Abstände stark betont, auch verknüpfte Rangplätze erhöhen den Wert des Maßes. Ist die Zahl der Verknüpfungen niedrig werden sie in der Regel vernachlässigt, es gibt jedoch einige vorgeschlagene Korrekturen, auf die [Ben07] jedoch nicht näher eingeht. [Bor10,

S.179] und [SH09, S.88] geben hierfür einen Schwellwert von 20% an. Ab diesem Anteil von verknüpften Paaren sollte das Maß nicht mehr eingesetzt werden. Beide geben für diesen Fall eine korrigierte Berechnungsform an. Quellen: [Ben07, S.177ff.], [Bor10, S.178f.], [CK14, S.113ff.], [FKP+07, S.142ff.], [SH09, S.88f.], [Wie13, Stichwort *Korrelationskoeffizient*], [Ste13, S.57f.]

3.2.3 Maße für zwei metrisch skalierte Größen

Bei PRE-Maßen für metrisch skalierte Merkmale, stellt sich nicht mehr die Frage, ob die Vorhersage überhaupt von der Beobachtung abhängt, sondern in der Regel danach wie stark sie es tut. [Ben07, S.186ff.], [SH09, S.85] und [CK14, S.105f.] weisen in diesem Zusammenhang auf die Bedeutung der Streudiagramme, auch Scatterplots genannt, als erste Einschätzungsmöglichkeit hin. Er empfiehlt vor der Interpretation des reinen Zahlenwerts die ausführliche Konsultation dieses Werkzeuges. Auch ein zweidimensionales Histogramm wird z.B. von [FKP+07] als grafische Möglichkeit zur Ersteinschätzung genannt.

Kovarianz

Bei der (empirischen) Kovarianz wird zur Beschreibung eines linearen Zusammenhangs zwischen zwei intervallskalierten Merkmalen verwendet. Sie ist nicht Standardisiert, was es schwierig macht mehrere Kovarianzen von unterschiedlichen Merkmalskombinationen zu vergleichen, da diese dann in der Regel einen unterschiedlichen Maßstab haben. Das macht sie in vielen Fällen zu einem wenig geeigneten Maß. [CK14, S.106] bezeichnet sie im Zusammenhang mit dem Korrelationskoeffizienten nur als Hilfsgröße.

$$s_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.34}$$

Quelle: [Bor10, S.154ff.], [CK14, S.106ff.], [SH09, S.85f.], [Ste13, S.53]

Der Korrelationskoeffizient r bzw. r^2 nach Pearson (PRE)

Der Pearsonsche Produkt-Moment-Korrelationskoeffizient r wird auch als Pearsonscher Korrelationskoeffizient, empirischer Korrelationskoeffizient, Bravais-Pearson-Korrelationskoeffizient oder schlicht **der** Korrelationskoeffizient bezeichnet. Auch der Formelbuchstabe S wird für ihn verwendet. Er beschreibt sowohl Grad als auch Richtung der Linearität einer Beziehung. Seine Charakteristika als PRE-Maß sind die folgenden:

1. Ohne weitere Vorkenntnis wird Mittelwert vorhergesagt.
2. Mithilfe der Methode der kleinsten Quadrate wird eine (lineare) Regressionsrechnung durchgeführt und der jeweilige Regressionswert vorausgesagt.
3. Die Summen der quadrierten Abweichungen $E_1 = \sum(y_i - \bar{y})^2$ und $E_2 = \sum(y_i - y_i')^2$ sind die Fehlerdefinitionen
4. Der Koeffizient wird wie folgt bestimmt:

$$r^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - y_i')^2}{\sum(y_i - \bar{y})^2} \tag{3.35}$$

Der eigentliche Korrelationskoeffizient r kann erreicht werden, indem man die Wurzel aus diesem PRE zieht. Es gibt jedoch für das Maß auch eine eigene Definitionsformel, die hiervon unabhängig ist. Sie basiert auf den Standardabweichungen s_x und s_y wie in Formel 3.36 bis 3.38 gezeigt wird. Er wird dabei als Steigung der Regressionsgeraden interpretiert.

$$z_{xi} = \frac{x_i - \bar{x}}{s_x} \tag{3.36}$$

$$z_{yi} = \frac{y_i - \bar{y}}{s_y} \tag{3.37}$$

$$r = \frac{\sum z_{xi} z_{yi}}{N} \tag{3.38}$$

Quellen: [Ben07, S.185ff.], [Bor10, S.156ff.], [CK14, S.109ff.], [FKP+07, S.135ff.], [Run10, S.55], [SH09, S.87f.], [Wie13, Stichwort *Korrelationskoeffizient*], [Ste13, S.55ff.]

η und η^2 (PRE)

Der Koeffizient η ist anwendbar, sobald das als Abhängige betrachtete Merkmal mindestens intervallskaliert ist. Da nur numerische Werte dieser Größe in die Berechnung eingehen, kann das unabhängige Merkmal einer beliebigen Skala angehören. Es kann mit folgenden Eigenschaften als PRE-Maß interpretiert werden:

1. Ohne Vorkenntnis wird das arithmetische Mittel vorhergesagt
2. Mit Vorkenntnis wird das arithmetische Mittel für jede Kolonne, die aus den unabhängigen Merkmalsausprägungen gebildet wird, vorausgesagt
3. $E_1 = \sum(y_i - \bar{y})^2$ und $E_1 = \sum(y_i - \bar{y_j})^2$
4. Der Koeffizient wird wie folgt bestimmt:

$$\eta^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \bar{y_j})^2}{\sum(y_i - \bar{y})^2} \tag{3.39}$$

$$= \frac{\sum_{j=1}^{k} n_j(\bar{y_j} - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.40}$$

Quellen: [Ben07, S.228ff.]

Regressionsrechnung

Die Regressionsrechnung wird z.B. beim bereits vorgestellten Korrelationskoeffizienten nach Pearson genutzt und bietet auch für nicht lineare Zusammenhänge Möglichkeiten zur Bestimmung der Richtung und Güte. Die Methode der kleinsten Abstandsquadrate, die in Abschnitt 3.2.3 bereits skizziert wurde, wird dort immer wieder genutzt, um die Anpassung an verschiedenste kurvilineare Funktionen zu überprüfen.

[FKP+07, S.165], [SH09, S.99], [Wie13, Stichwort *Regression, nicht lineare*]

4 Einordnung und Bewertung der Maßzahlen im Hinblick auf die Daten der Supply Chain

Die Auswahl von Maßzahlen, die im vorhergehenden Kapitel 3 vorgestellt wurde, macht deutlich, dass im Laufe der Zeit eine Vielzahl von Maßzahlen für alle existierenden Skalenarten entwickelt wurden. Es ist auffällig, dass die erwähnten Maße zu einem Großteil sehr alt sind. Viele Originalarbeiten, auf die in der hier hauptsächlich genutzten Sekundärliteratur verwiesen wird, sind mehr als 30 Jahre alt [Ben07]. Die den Kurzbeschreibungen beigefügten Quellenangaben dienen sowohl dem Verweis auf weitergehende Informationen als auch dem Nachweis, dass es einige wenige Maße gibt, die für Nebenfächler verstärkt aufgegriffen werden. Es ist daraus offensichtlich zu ersehen, dass es einige wenige etablierte Maße gibt, auf die immer wieder zurück gegriffen wird, während andere kaum diskutiert werden. Insbesondere in den ausgewiesenen Lehrwerken wie z.B. [Bor10], [CK14], [FKP+07], [SH09] und [Ste13] wird pro Skala eine Berechnungsmöglichkeit vorgestellt ohne andere dem gegenüber zu stellen. Auffällig ist bei der Zuordnung der Maße zu den Skalenarten, dass zwar, wie in Abschnitt 3.1.2 beschrieben, in vielen Werken darauf geachtet wird, die Intervall- und die Verhältnisskala voneinander abzugrenzen, dies aber bei den Maßzahlen nicht mehr relevant ist.

Die oben dargelegte Recherche hat eine Reihe von Kriterien ergeben, die von den statistischen Lehrwerken selbst zur Einteilung der Maßzahlen genutzt wird. Daneben ergeben sich aus der Anwendung und dem spezifischen Anwendungsfeld unter Umständen weitere Gesichtspunkte, die bei der Auswahl des Maßes beachtet werden sollten. Im folgenden Abschnitt werden diese Kriterien aufgezählt. Im zweiten Abschnitt werden Maße aus Kapitel 3 aufgezeigt, die auf Grundlage dieser Kriterien für die Untersuchung von Supply-Chain-Daten in Frage kommen.

4.1 Mathematische Kriterien zur Auswahl statistischer Maßzahlen

Die folgenden vier Kriterien haben in der statistischen Literatur bei der Beschreibung, dem Vergleich und der Bewertung der Maßzahlen immer wieder Erwähnung gefunden:

1. Standardisierung: Das Ergebnis eines Maßes sollte entweder auf Werte zwischen 0 und 1 begrenzt sein, sodass eine Aussage über Stärke aber nicht über die Richtung des Zusammenhanges getroffen werden kann, oder auf Werte zwischen -1 und +1, sodass Aussagen über Stärke und Richtung möglich sind. Dies schließt auch ein, dass Maße für Stichprobenumfänge unterschiedlicher Größe vergleichbar sein sollten.

2. Die Symmetrie des Maßes muss beachtet werden, da hierdurch eine Aussage über die Abhängigkeit und Unabhängigkeit der beiden betrachteten Variablen impliziert werden.

3. Maße sollten möglichst auf Merkmale mit beliebig vielen Ausprägungen anwenbar sein, da z.B. die Beschränkung auf 2x2-Tabellen eine Vergröberung der Daten erfordert, die zu einer erheblichen Verzerrung der Ergebnisse führen kann.

4. Die Proportionalität eines Maßes hilft dabei, verschiedene Werte des gleichen Maßes in korrekte Beziehung zueinander zu setzen. Ist die lineare Proportionalität nicht gegeben darf z.B. nicht die Aussage getroffen werden, dass ein Zusammenhang von 0,8 doppelt so stark ist wie bei einem Wert von 0,4.

4.2 Bewertung der Maße im Hinblick auf die Supply Chain

Bedenkt man die in Abschnitt 2.2 erwähnten Merkmale, die im Rahmen einer Supply Chain anfallen können, so fällt auf, das diese von technischer Seite zunächst beliebig sind. Statistische Maße sind darauf ausgelegt, in jedem Sachzusammenhang zu funktionieren, sodass auch Supply-Chain-Daten keine besondere Herausforderung an sie stellen. Dieser Anspruch wird auch durch die große Bandbreite der Beispiele unterstrichen, die z.B. bei [Kü01] und [Bor10] behandelt wird. Es ist jedoch auffällig, das im Zusammenhang mit den Beispieldaten viele Merkmale genannt werden, die der nominalen oder ordinalen Skala zuzuordnen sind. Hierzu zählen zum Beispiel Fragen zu Standorten, Produkten und Adressbestandteilen. Insbesondere bei der strategischen Ausrichtung der Supply Chain sind viele Dinge festzulegen, die sich als nicht metrische Merkmale niederschlagen. Außerdem lässt sich feststellen, dass es kaum Merkmale gibt, die nur zwei Ausprägungen besitzen. Oft gibt es eine Reihe von Alternativen oder Werten, die als Ausprägung in Frage kommen. Aus diesen Umständen lassen sich die folgenden zwei sachbezogene Aussagen treffen, die einen Hinweis darauf geben, welche Maße man im Rahmen der Supply Chain benutzen sollte und welche nicht.

1. Maße für nicht metrische Merkmale sind für Anwendungen in der Supply Chain von besonderem Interesse, da sie häufig benötigt werden und über einen großen Sachbereich hinweg miteinander vergleichbar sein sollten.

2. Maße, die ausschließlich für Merkmale mit zwei Ausprägungen anwendbar sind, sind im Rahmen der Supply Chain eher ungeeignet, da solche Merkmale nicht häufig vorkommen. Wenn sie auftauchen, kann für sie ein allgemein anwendbares Maß herangezogen werden, sodass die Vergleichbarkeit innerhalb der Daten erhalten bleibt.

Führt man sich vor Augen, dass im Rahmen der strategischen und operativen Supply-Chain-Planung viele Personen mit unterschiedlichem wissenschaftlichen und nicht wissenschaftlichen Hintergrund zusammen arbeiten, erscheint es nachvollziehbar, dass die einfache inhaltliche Interpretierbarkeit der Maße eine hohe Bedeutung beim Auswahlprozess hat. In diesem Zusammenhang stachen bei der Recherche nicht die etablierten Maße für

die nicht metrische Skala heraus sondern eher die weniger beachteten PRE-Maße, die eine einfache, nachvollziehbare und den oben genannten Kriterien entsprechende Struktur besitzen. Bei den Maßen für metrische Merkmale, erscheint die Regressionsrechnung im Zusammenhang mit dem PRE-Prinzip eine gute Wahl zu sein, da sie wie im letzten Teil von Kapitel 3 angedeutet auch über lineare Zusammenhänge hinaus genutzt werden kann.

Anhand der genannten Kriterien ließen sich aus den drei Kategorien nominalskalierter, ordinalskalierter und metrisch skalierter Variablen beispielhaft die Maße „Goodman und Kruskals λ“, „d_s nach Somer“ und „η“ identifizieren, die den Anforderungen genügen würden. Hierdurch wäre gewährleistet, dass man Zusammenhänge zwischen einer Vielzahl von Daten in Supply Chains eindeutig interpretierbar untersuchen kann.

5 Zusammenfassung und Ausblick

Die vorliegende Arbeit gibt einen Überblick über Zusammenhangsmaße aus dem Bereich der deskriptiven Statistik und zeigt welche davon sich zur Untersuchung von Daten in Supply Chains besonders eignen. Dazu wurden Beispieldaten aus bestehenden Arbeiten zum Thema Datenstrukturen in Supply Chains herangezogen und unter dem Aspekt ihres statistischen Skalenniveaus betrachtet. Das jeweilige Niveau gibt einen ersten Hinweis, welche Maßzahlen zur näheren Beschreibung von Zusammenhängen in Frage kommen. Indem eine Auswahlliste verschiedener existierender Zusammenhangsmaße recherchiert wurde, konnten weitere Charakteristika herausgearbeitet werden, die bei der Auswahl eines geeigneten Maßes helfen. Hierzu zählen die Anzahl der Ausprägungen, die die betrachteten Merkmale annehmen können, sowie die Symmetrie des Maßes. In jedem Fall beachtet werden sollten die Standardisierung des Maßes, sowie eine möglichst gute Interpretierbarkeit auch durch nicht statistisch ausgebildete Personen.

Die Ausführungen bieten einen Überblick über die einzelnen Maße und Hinweise zur weiteren Vertiefung, falls das Maß als relevant identifiziert wurde. Der Umfang der Arbeit verhindert eine ausführlichere Darstellung. Insbesondere im Bereich der Regressionsrechnung für metrische Skalenniveaus gibt es Möglichkeiten zur Zusammenhangsüberprüfung, die weit über das hier vorgestellte hinaus gehen, wie in den Abschnitten zur nicht linearen Regression angedeutet wurde. Auch der Aspekt der Signifikanzüberprüfung wurde außen vor gelassen, um der eingehenden Beschäftigung mit rein deskriptiven Verfahren ausreichenden Platz einräumen zu können. Schnell kommt man im Zusammenhang mit Statistik und Daten zum Data Mining, das mit ausgewählten Methoden versucht, Zusammenhänge in großen Datenmengen aufzuspüren. All diese Aspekte sind mögliche Anknüpfungspunkte, zu denen man in der im Anhang aufgeführten Literatur genauere Informationen einholen kann. In diesem Sinne ist die Arbeit als erfolgreiche erste Orientierung in einem sehr großen Wissengebiet anzusehen.

Literaturverzeichnis

[Ben07] Benninghaus, Hans: *Deskriptive Statistik - Eine Einführung für Sozialwissenschaftler*. 11. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften | GWV Fachverlage GmbH, 2007

[Bor10] Bortz, Jürgen; Schuster, Christof (Hrsg.): *Statistik für Human- und Sozialwissenschaftler*. 7. Auflage. Springer-Verlag, 2010

[Bos97] Bosch, Karl: *Lexikon der Statistik*. 2. Auflage. R. Oldenbourg Verlag, 1997

[Bos98] Bosch, Karl: *Statistik-Taschenbuch*. 3. Auflage. R. Oldenbourg Verlag, 1998

[CGHB+91] Christoph, Gerd; Gillert, Heinz; Hahnewald-Busch, Andreas; et. al.; Müller, P. H. (Hrsg.): *Wahrscheinlichkeitsrechnung und mathematische Statistik*. 5. Auflage. Akademie Verlag, 1991

[CK14] Cramer, Erhard; Kamps, Udo; Kamps, Udo (Hrsg.): *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik*. 3. Auflage. Springer-Verlag, 2014

[Dud07] der Dudenredaktion, Wissenschaftlicher R. (Hrsg.): *Duden - Das große Fremdwörterbuch*. 4. Auflage. Dudenverlag, 2007

[EGH+94] Eckstein, Peter; Götze, Wolfgang; Hartl, Friedrich; Rönz, Bernd; Strohe, Hans G.; Rönz, Bernd (Hrsg.); Strohe, Hans G. (Hrsg.): *Lexikon Statistik*. 1. Auflage. Gabler Verlag, 1994

[FH11] Fischer, Peter; Hofer, Peter: *Lexikon der Informatik*. 15. Auflage. Springer-Verlag, 2011

[FKP+07] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: *Statistik - Der Weg zur Datenanalyse*. 6. Auflage. Springer-Verlag, 2007

[Gra04] Grabowski, Barbara; Walz, Guido (Hrsg.): *Lexikon der Statistik*. 1. Auflage. Spektrum Akademischer Verlag, 2004

[Guh14] Guhl, Philipp: *Erstellung eines konzeptuellen Datenbankschemas im Umfeld von Supply Chains*, Technischen Universität Dortmund, Fakultät Maschinenbau, Fachbereich ITPL, Diplomarbeit, 2014

[HEK09] Hartung, Joachim; Elpelt, Bärbel; Klösener, Karl-Heinz: *Lehr- und Handbuch der angewandten Statistik*. 15. Auflage. Oldenbourg, 2009

[Hil14] Hilpert, Kilian: *Einsatz maschineller Lernverfahren im Decision Support von Wertschöpfungsnetzwerken*, Technischen Universität Dortmund, Fakultät Maschinenbau, Fachbereich ITPL, Diplomarbeit, 2014

[Hä87] Härtter, Erich: *Wahrscheinlichkeitsrechnung, Statistik und mathematische Grundlagen*. 1. Auflage. Vandenhoeck & Ruprecht in Göttingen, 1987

[Kla12] Klaus, Peter; Krupp, Michael (Hrsg.); Krieger, Winfried (Hrsg.): *Gabler Lexikon Logistik*. 5. Auflage. Gabler Verlag, 2012

[Kü01] Kühlmeyer, Manfred: *Statistische Auswertungsmethoden für Ingenieure*. 1. Auflage. Springer-Verlag, 2001

[Run10] Runkler, Thomas A.: *Data Mining - Methoden und Algorithmen intelligenter Datenanalyse*. 1. Auflage. Vieweg+Teubner, 2010

[SH09] Sachs, Lothar; Hedderich, Jürgen: *Angewandte Statistik - Methoden mit R*. 13. Auflage. Springer-Verlag, 2009

[Ste13] Steland, Ansgar: *Basiswissen Statistik - Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*. 3. Auflage. Springer-Verlag, 2013

[VB00] Voss, Werner (Hrsg.); Buttler, Günther (Hrsg.): *Taschenbuch der Statistik*. 1. Auflage. Fachbuchverlag Leipzig, 2000

[Wie13] Wiesbaden, Springer F. (Hrsg.): *Kompaktlexikon Wirtschaftsmathematik und Statistik*. 1. Auflage. Springer Gabler, 2013

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Fachwissenschaftliche Projektarbeit selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort, Datum	Unterschrift