
Zusammenfassung

Das zweite Kapitel führt zunächst in die verschiedenen Möglichkeiten der univariaten Datenbeschreibung ein und widmet sich dann verschiedener Möglichkeiten der Deskription kategorialer (nominaler und ordinaler) Merkmale. Anhand einer Reihe von Beispielen wird das praktische Vorgehen zur Berechnung sowie die Interpretation folgender Statistiken beschrieben: Absolute und relative Häufigkeiten, absolute und relative Häufigkeitsverteilung und der Modus für nominale und ordinale Daten sowie die kumulierten absoluten und relativen Häufigkeitsverteilungen und der Median für ordinale Daten. Mit Balken-, Säulen- und Tortendiagramm werden außerdem verschiedene Möglichkeiten der grafischen Darstellung kategorialer Daten beschrieben.

Im folgenden Kapitel und in Kapitel vier befassen wir uns mit der eindimensionalen Beschreibung von Merkmalen.

2.1 Analysebereiche und Beschreibungskriterien von Merkmalen

Eindimensional bedeutet, dass man jeweils nur ein Merkmal für sich betrachtet. Diese univariaten (uni=eins; variate=nur eine Variable betreffend) Analysen bilden die Grundlage für später folgende bivariate Analysen, bei denen es um die Zusammenhänge zwischen zwei Variablen gehen soll.

Beispiel 2.1 Häufigkeit von Themen in Nachrichten

Vor dem Hintergrund der Konvergenzthese, die sich die Frage stellt, inwieweit sich die öffentlich-rechtlichen Fernsehsender den privaten Sendern angleichen, wurde in einer Inhaltsanalyseübung im SS 2012 am IPK eine Nachrichtenanalyse verschiedener Fernsehsender durchgeführt. Dabei wurden von sieben Sendern jeweils die Hauptnachrichten einer zufällig ausgewählten künstlichen Woche¹ nach verschiedenen Merkmalen inhaltsanalytisch erfasst. Unter anderem war dabei das *Thema* der Nachrichtenbeiträge von Interesse. Es wurde in insgesamt neun Hauptkategorien erfasst: 1=Politik, 2=Wirtschaft, 3=Gesellschaft, 4=Wissenschaft/Kultur, 5=Unfall/Katastrophe, 6=Kriminalität, 7=Human Interest/Buntes, 8=Sport, 9=Wetter. Weitere Variablen, die in Anlehnung an die Nachrichtenwerttheorie erhoben wurden, waren *Einfluss der Akteure*, über die berichtet wird, *Personalisierung* des Beitrags, *Reichweite* und *Aktualität* des Ereignisses und *kulturelle Nähe* des Ereignisortes. Die genauen Codieranweisungen finden Sie im Codierbuch im Anhang des Skripts.

Eine *eindimensionale Beschreibung* der in Beispiel 2.1 beschriebenen Untersuchung stellt zunächst dar, welche Art der Berichterstattung auf allen Sendern insgesamt oder bei einem einzelnen Sender vorgefunden wurde. Eine zweidimensionale Analyse weitete den Blickwinkel z. B. auf den Vergleich einzelner Sender oder die Zusammenhänge zwischen verschiedenen Themen und bestimmten Nachrichtenfaktoren.

► Analysebereiche

Eindimensional/univariat Nur ein Merkmal mit seinen Ausprägungen.

Zweidimensional/bivariat Zwei Merkmale werden gemeinsam untersucht/verglichen, um Zusammenhänge festzustellen.

Mehrdimensional/multivariat Mehrere Merkmale werden auf komplexe Zusammenhänge untersucht.

Zur Beschreibung von Daten bietet die Statistik verschiedene Werkzeuge an. Letztlich geht es dabei immer darum, ein begreifbares bzw. erkennbares Bild davon zu bekommen, wie es denn um ein Merkmal in seinen Ausprägungen in einer größeren Menge von Objekten bestellt ist.

¹ Der Zeitraum von Ende Februar (20.02) bis Anfang April (08.04.) wurde nach den jeweiligen Wochentagen sortiert. Aus jeder Gruppe wurde dann einer zufällig ausgewählt, so dass aus dem ganzen Zeitraum sieben Tage zufällig ausgewählt wurden und dabei sichergestellt war, dass jeder Wochentag einmal in der Stichprobe vertreten ist.

Statistiken zur Beschreibung von Daten kommen dafür immer dann zum Einsatz, wenn die Anzahl der untersuchten Objekte so groß ist, dass wir durch bloße Anschauung keinen Überblick mehr über die Situation bekommen können. Ab welcher Zahl dies genau der Fall ist, hängt von verschiedenen Aspekten wie der Art der Objekte, der Art des Merkmals und dem Erkenntnisinteresse ab. Ab etwa 10–15 Objekten ist auch bei einfachen Merkmalen wie *die Farben von Tassen* oder den *Arten von Obst* im Obstkorb das Erfassungsvermögen mit bloßem Auge schnell erschöpft.

Bei der Beschreibung von Merkmalen leiten uns letztendlich immer folgende Fragen:

- *Wie viele verschiedene Ausprägungen* des Merkmals liegen bei den Objekten vor?
- *Wie verteilen sich* die erfassten Objekte auf die einzelnen Ausprägungen des Merkmals?
- Zeigt sich eine *verallgemeinerbare Tendenz*, d. h. kann man über die Objekte z. B. Aussagen machen wie: „Meistens ist es....“ oder: „Die meisten sind....“, oder: „Am verbreitetsten ist...“?

Entsprechend dieser Fragen bietet die Statistik Werkzeuge zur Beschreibung der *Verteilung von Daten* an. Die *Verteilung* eines Merkmals gibt vor allem auf die ersten beiden oben genannten Fragen eine Antwort. Sie drückt aus, wie sich die Objekte auf die einzelnen Merkmalsausprägungen verteilen.

► **Verteilung** Beschreibt die Verteilung der Daten im Hinblick auf die verschiedenen vorkommenden Ausprägungen des interessierenden Merkmals.

Einen Eindruck der *zentralen Tendenz* geben die sogenannten *Lagemaße*. Sie verdichten das durch die Verteilung gewonnene Bild noch einmal und fokussieren auf die am meisten herausstechenden Ausprägungen.

► **Lagemaß** Statistischer Kennwert, der die zentrale Tendenz der Objekte im Hinblick auf das interessierende Merkmal ausdrücken soll.

Da bei dieser Fokussierung Informationen verloren gehen, stellt sich die Frage, wie aussagekräftig die Lagemaße für das Ganze sind. Diese Frage wird mit den *Streuungsmaßen* beantwortet. Sie geben Hinweise darauf, wie stark die Variation der Ausprägungen insgesamt ist.

► **Streuungsmaß** Statistischer Kennwert, der das Ausmaß der Abweichung der Daten von dem als zentrale Tendenz identifizierten Wert ausdrückt.

2.2 Häufigkeitsverteilung qualitativer bzw. kategorialer Merkmale

Die Erarbeitung der verschiedenen Verteilungen, Lagemaße und Streuungsmaße der beschreibenden Statistik beginnt in diesem Buch mit der Beschreibung von qualitativen Daten.

Wir erinnern uns daran: Als *qualitativ* werden solche Merkmale bezeichnet, die eine endliche Menge an Ausprägungen haben und bei denen die Daten keine Quantität/kein Ausmaß widerspiegeln, sondern eine Qualität als eine ganz bestimmte Eigenschaft der erfassten Objekte. Diese werden auch als kategoriale Merkmale bezeichnet. Merkmale mit einem *nominalen Skalenniveau* gehören eindeutig zu dieser Gruppe, ebenso Merkmale mit einem *ordinalen Skalenniveau*.²

Für qualitative Daten stellen Häufigkeiten eine gebräuchliche Beschreibung dar.

2.2.1 Absolute und relative Häufigkeiten von Ausprägungen

In Beispiel 2.1 wurde allen Nachrichtenbeiträgen eine Themenkategorie zugewiesen, d. h. die entsprechende Zahl zugeordnet. Das Skalenniveau der Messwerte ist nominal, da jedes Thema für sich steht und die Themenausprägungen keine Rangordnung darstellen. Die Ziffern repräsentieren damit nur Themenbezeichnungen. Eine einfache Liste von Zahlen, die die Merkmalsausprägungen eines Merkmals einer untersuchten Menge an Objekten darstellt, wird auch als *Urliste* bezeichnet.

► **Urliste** Ein Merkmal X werde an den n Merkmalsträgern einer Population bzw. einer Teilmenge (Stichprobe) davon gemessen. Die resultierenden Zahlen $x_1, \dots, x_n$ bezeichnen dann die Beobachtungswerte.

² Hinweis: Die verschiedenen Skalenniveaus wurden ebenso wie die die Verwendung der folgenden Begriffe Objekte (im Sinn statistischer Einheiten) Merkmal/Variable, Merkmalsausprägung/Werte, Daten, Grundgesamtheit, Stichprobe, Messung im letzten Kapitel eingeführt und werden als bekannt vorausgesetzt. Sie sollten also jetzt von diesen Begriffen eine Vorstellung haben und ihre Bedeutung für sich selbst in eigenen Worten ausdrücken können.

x_i bedeutet dabei die beim i -ten Objekt beobachtete Merkmalsausprägung der Variable X . Die Zahlenreihe der Länge n $(x_1, \dots, x_n)$ wird auch **Urliste** genannt (oder „ n -Tupel“).

Beispiel 2.1 Häufigkeit von Themen in Nachrichten (Fortsetzung)

Insgesamt wurden 745 Nachrichtenbeiträge untersucht, von denen 120 kein Thema zugeordnet werden konnte, weil es sich um prozedurale Elemente wie Trailer, Vorspann, Begrüßungen etc. handelt. Man erhält damit 625 verwertbare Daten.

Tabelle 2.1 zeigt einen Ausschnitt aus dem Datensatz von Beispiel 2.1. Der Übersicht halber wurde nur eine zufällig Auswahl vom Umfang $n=34$ aller Daten ausgewählt. Bei einer eindimensionalen Betrachtung sind keine weiteren Informationen von Interesse, so dass im Moment nur die Spalte des interessierenden Merkmals „Themenkategorie“ von Interesse ist. Das Merkmal *Themenkategorie* wird damit durch folgende Datenliste abgebildet:

1, 1, 7, 1, 1, 1, 3, 2, 4, 9, 2, 8, 1, 2, 2, 1, 5, 6, 2, 1, 7, 2, 1, 7, 1, 7, 6, 1, 9, 7, 5, 5, 1, 8

Es wurde in diesem Fall an 34 Objekten gemessen. Entsprechend gilt für Beispiel 2.1: Der Stichprobenumfang $n=34$, x_7 der Urliste hat den Wert 3 ($x_7=3$) und $x_{34}=8$.

Bei 34 Objekten wird deutlich, dass die Urliste an sich bereits nicht mehr auf einen Blick zu erfassen ist und dass der Einsatz statistischer Kennwerte hilfreich ist.

Eine gängige Möglichkeit, die Urliste zusammenzufassen, ist die Bildung von *absoluten Häufigkeiten*. Dazu ist es einmal notwendig, alle Merkmalsausprägungen, die vorkommen können bzw. die tatsächlich im Datensatz vorkommen, festzustellen und dann zu zählen, wie oft diese auftreten (Gl. 2.1). *Absolut* werden die Häufigkeiten genannt, weil Sie die gezählte Menge an Objekten dieser Ausprägung angeben.

► **relative und absolute Häufigkeiten** Sei mit $a_1, a_2, a_3, \dots, a_k$, $k < n$ die Menge der möglichen Merkmalsausprägungen bezeichnet, dann sind die Anzahl aller Werte, die mit einer Ausprägung a_j übereinstimmen,

$$\text{deren absolute Häufigkeiten } h(a_j) = h_j \quad (2.1)$$

und der Anteil dieser Werte an der Menge aller n Beobachtungen

$$\text{deren relative H\u00e4ufigkeiten} \quad f(a_j) = f_j = h_j / n \quad (2.2)$$

Durch Multiplikation mit 100 berechnet man aus den relativen H\u00e4ufigkeiten Prozentwerte.

Die relativen H\u00e4ufigkeiten dr\u00fccken die Anzahl in Relation zur Gesamtmenge aus – bei $n=400$ Objekten sind 10 gez\u00e4hlte Auspr\u00e4gungen eines Merkmals anderes zu interpretieren als bei $n=40$ Objekten.

Die **relative H\u00e4ufigkeit** gibt den Anteil einer Merkmalsauspr\u00e4gung an allen festgestellten Auspr\u00e4gungen an (Gl. 2.2). Greifen wir die Zahlen von oben wieder auf, so sind 10 von 400 ein Vierzigstel, also ein Anteil von 0,025. 10 von 40 dr\u00fccken dagegen einen Anteil von 0,25 aus.

Die im allgemeinen Sprachgebrauch h\u00e4ufiger verwendeten Prozentzahlen bekommen wir, wenn beim Anteilswert das Komma um zwei Stellen nach rechts verschoben wird – aus 0,025 werden 2,5 %, aus 0,25 werden 25 %.

Beispiel 2.1 H\u00e4ufigkeit von Themen in Nachrichten (Fortsetzung)

Unser Merkmal *Themenkategorie* kommt im Datensatz in folgenden Auspr\u00e4gungen vor: 1, 2, 3, 4, 5, 6, 7, 8, 9.

Damit gilt: $k=9$, es gibt also neun verschiedene Auspr\u00e4gungen die mit $a_1, a_2, a_3, \dots, a_9$ bezeichnet werden und $a_1=1, a_5=5$ usw.

Durch Ausz\u00e4hlen ergibt sich die absolute H\u00e4ufigkeit $h(a_1)=12$, die Kategorie 1=Politik kommt also 12 mal vor.

Entsprechend berechnet sich die relative H\u00e4ufigkeit $f(a_1)=h(a_1)/n=12/34=0,353$ (gerundet!). Der Anteil der Beitr\u00e4ge mit politischen Themen liegt also bei gut einem Drittel, bei 35,3 %.

Praktisch geht man zur Ermittlung von H\u00e4ufigkeiten eines Merkmals in einem Datensatz so vor, dass man z. B. von oben nach unten durch die Daten schaut und alle Werte mit der gesuchten Auspr\u00e4gung anstreicht und durchz\u00e4hlt.

Liegt die Tabelle in digitaler Form vor, dann ist der erste Schritt zu einer computergest\u00fctzten Ermittlung von H\u00e4ufigkeiten das Sortieren der Daten nach dem interessierenden Merkmal. Die H\u00e4ufigkeiten k\u00f6nnen dann ganz einfach ausgez\u00e4hlt werden bzw. wenn man alle Zellen mit der gleichen Auspr\u00e4gung markiert, dann kann man i. d. R. in der Statusleiste die Anzahl ablesen.

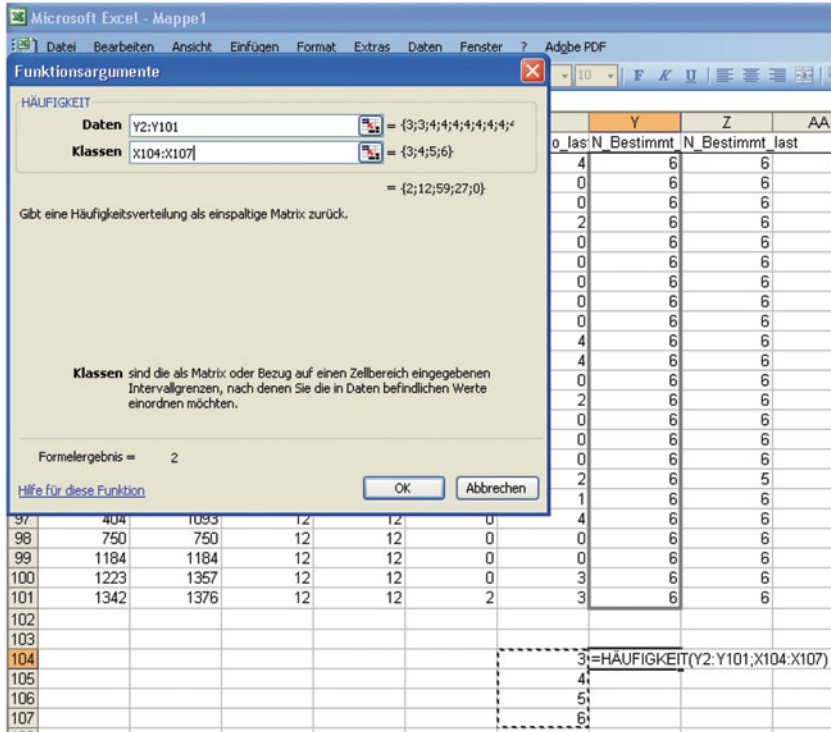


Abb. 2.1 Auszählen von Häufigkeiten mit Excel

Noch schneller geht es mit Hilfe der eigens dafür vom Programm angebotenen Funktion HÄUFIGKEIT:

Schritt 1: Geben Sie die möglichen Ausprägungen in einer Reihe untereinander ein.

Schritt 12: Gehen Sie nun in die Zelle neben der ersten Ausprägung, gehen Sie auf FORMELN → FUNKTION EINFÜGEN → STATISTIK → HÄUFIGKEIT und markieren Sie bei Daten den Datenbereich, den Sie auszählen wollen und bei Klassen die von Ihnen eingegebenen Ausprägungen (vgl. Abb. 2.1).

Schritt 3: Markieren Sie nun – von der Zelle aus beginnend, welche die Formel enthält – die Zellen neben den Klassen (Ausprägungen), drücken Sie F2 und anschließend Strg+Umsch+Enter. Die Tabelle ist fertig, sie enthält die absoluten Häufigkeiten.

2.2.2 Häufigkeitstabellen

Für Beispiel 2.1 wurde durch Auszählen ermittelt, wie oft die einzelnen Kategorien in den 34 Beiträgen auftreten. Üblicherweise erfolgt die Darstellung der Häufigkeiten aller vorliegenden Ausprägungen in einer Tabelle. Die Anordnung der Häufigkeiten ist dabei dem Forscher überlassen: Sie kann *zufällig* sein, sie kann *theoretisch* festgelegt werden, indem man z. B. mit den wichtigsten Themen beginnt oder sie kann sich *empirisch* ergeben, indem man mit der häufigsten oder der seltensten Kategorie beginnt.

Der Einfachheit halber wird die Tab. 2.1 *Themenhäufigkeit* zunächst nach den Ziffern der Kategorien geordnet. Daneben zum Vergleich eine Anordnung nach der Häufigkeit des Vorkommens.

Anhand Tab. 2.2 wird deutlich, dass bei qualitativen Merkmalen die Merkmalscodes wieder in Bezeichnungen überführt werden müssen, damit sich ein sinnvolles Bild ergibt.

Eine *Häufigkeitstabelle* gibt Auskunft darüber, wie sich die beobachteten Objekte insgesamt auf alle möglichen Ausprägungen verteilen. Rufen wir uns noch einmal die Definition einer Verteilung (vgl. 2.1) ins Gedächtnis, so wird deutlich dass eine solche Häufigkeitstabelle genau diesem Ziel entspricht. Dementsprechend wird sie auch Häufigkeitsverteilung genannt.

► **Absolute und relative Häufigkeitsverteilung** Eine Verteilung beschreibt, wie sich die erhobenen Werte der beobachteten Objekte innerhalb des durch die möglichen Ausprägungswerte vorgegebenen/eingegrenzten Wertebereichs verhält.

Danach bezeichnen

$h_1, \dots, h_k$ die absolute Häufigkeitsverteilung

$f_1, \dots, f_k$ die relative Häufigkeitsverteilung

Die Summe aller absoluten Häufigkeiten einer Verteilung ergibt n , die Summe aller relativen Häufigkeiten ergibt 1, die Gesamtprozent einer Verteilung sind immer 100 %.

Trifft dies auf eine Verteilung nicht zu, dann ist dies ein Hinweis darauf, dass ein Rechenfehler vorliegt. Ungültige Werte müssen in einer Tabelle entweder ausgewiesen werden, oder es werden die Stichproben entsprechend bereinigt. Damit ergibt sich dann auch eine geänderte Stichprobengröße und alles stimmt wieder zusammen.

Die Tab. 2.3 wurde so vom Statistikprogramm SPSS erzeugt. Sie stellt die Häufigkeitsverteilung der Themenkategorien dar und enthält sowohl die Beiträge, für die kein Thema feststellbar war, als auch weitere 5 fehlende Werte.

Tab. 2.1 Daten der Nachrichtenanalyse ($n = 34$)

Sender	Beitrag	Dauer_ sek	Ein- fluss	Perso- nalisie- rung	Reich- weite	Aktu- aliät	Dar- stel.- form	kult. Nähe	The- menka- tegorie
10	1007405	32	3	1	3	0	1	0	1
10	1007406	26	4	1	4	3	1	0	1
10	1007413	15	0	0	3	3	2	3	7
10	1026208	23	0	0	1	3	1	1	1
10	1028306	42	4	0	3	1	1	0	1
20	2001302	106	3	1	3	1	2	0	1
20	2007406	114	2	1	3	0	3	0	3
40	4001304	136	4	2	4	3	3	0	2
40	4007402	30	4	1	3	3	3	3	4
40	4009319	91	0	0	4	3	1	0	9
40	4026307	27	2	0	3	3	2	0	2
40	4026316	129	3	1	3	3	3	0	8
50	5001304	26	4	1	4	3	3	3	1
50	5007209	26	0	0	2	2	2	0	2
50	5028303	21	3	0	3	3	2	0	2
60	6007204	113	4	1	2	3	3	0	1
60	6007206	177	1	1	4	3	3	0	5
60	6009305	93	2	1	2	3	3	0	6
60	6026204	91	4	1	4	3	3	0	2
60	6026301	103	4	1	4	3	3	0	1
70	7007210	25	0	0	2	1	2	1	7
80	8001308	121	4	1	3	3	3	0	2
80	8007405	27	4	1	4	3	1	0	1
80	8007411	107	1	1	4	2	3	0	7
80	8009304	21	4	0	3	3	1	0	1
80	8009315	22	0	0	0	1	3	0	7
80	8026308	118	2	2	1	3	3	0	6
80	8028302	126	3	1	3	3	3	0	1
10	10010314	67	0	0	0	3	2	0	9
30	30260208	111	1	2	0	3	3	2	7
50	50070405	24	1	0	1	3	2	0	5
70	70010302	133	3	1	2	3	3	0	5

Tab. 2.1 (Fortsetzung)

Sender	Beitrag	Dauer_ sek	Ein- fluss	Perso- nalisie- rung	Reich- weite	Aktu- aliät	Dar- stel.- form	kult. Nähe	The- menka- tegorie
70	70010306	30	4	1	4	3	2	4	1
80	80260212	32	3	0	0	3	1	0	8

Tab. 2.2 Absolute Häufigkeitsverteilung des Merkmals Themenkategorie (Nachrichtenana- lyse $n = 34$)

a_j	$h(a_j)$	Themenkategorie a_j	$h(a_j)$	$f(a_j)$	Anteil in Prozent
1	12	Politik	12	0,353	35,3%
2	6	Wirtschaft	6	0,176	17,6%
3	1	Buntes	5	0,147	14,7%
4	1	Unfälle	3	0,088	8,8%
5	3	Kriminalität	2	0,059	5,9%
6	2	Sport	2	0,059	5,9%
7	5	Wetter	2	0,059	5,9%
8	2	Gesellschaft	1	0,029	2,9%
9	2	Kultur	1	0,029	2,9%
n	34	Gesamt	n=34	1,0	100%

Entsprechend werden in Tab. 2.3 zwei Prozentwerte ausgegeben, einmal die Gesamtprozent und einmal die gültigen Prozent³. Wollte man die 120 Fälle, für die kein Thema feststellbar ist, aus der Analyse ausschließen, dann muss man die relativen Werte entweder noch einmal von Hand berechnen oder bei SPSS die Ausprägung 10=kein Thema feststellbar als fehlenden Wert definieren und die Häufigkeitsverteilung neu berechnen. Dies wird im Folgenden für Beispiel 2.1 durchgeführt.

Beispiel 2.2: Häufigkeit von Themen in Nachrichten ($n = 750$) Welche Anteile ergeben sich für Tab. 2.3, wenn nur die vorhandenen Themen zugrunde gelegt werden?

Schritt 1: Stelle die neue Basis aller gültigen Werte fest: $h_{(\text{gültige Werte})} - h_{(\text{ungültige Werte})} - h_{(\text{kein Thema feststellbar})} = 750 - 5 - 120 = 625$ (vgl. Tab. 2.3).

Schritt 2: Berechne die relativen Häufigkeiten neu, indem jede absolute Häufigkeit der Tab. 2.3 durch die neue Basis geteilt wird, z. B. $f(a_1) = 186/625 = 0,2976$ (vgl. Gl. 2.2).

³ Die ebenfalls ausgegebenen kumulierten Prozent stellen für das Merkmal Themenkategorie keine sinnvolle Auswertung dar und können ignoriert werden.

Einführung in die Statistik für
Kommunikationswissenschaftler
Deskriptive und induktive Verfahren für das
Bachelorstudium
Uhlemann, I.A.
2015, XI, 242 S., Softcover
ISBN: 978-3-658-05768-8