

Markus Tausendpfund et al.

Quantitative Analyseverfahren

Eine Einführung

Fakultät für
**Kultur- und
Sozialwissen-
schaften**

Das Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung und des Nachdrucks, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (Druck, Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung der FernUniversität reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m², weiß), hergestellt aus 100 % Altpapier.

Vorwort

In der quantitativen Sozialforschung wird zur Beschreibung von Daten und zur empirischen Überprüfung von Hypothesen auf statistische Verfahren zurückgegriffen. Wer eine (quantitative) Studie verstehen und kritisch bewerten möchte, der muss die grundlegenden Prinzipien, Anwendungsvoraussetzungen und auch Probleme der verwendeten statistischen Verfahren kennen. Für Sozialwissenschaftlerinnen und Sozialwissenschaftler sind deshalb elementare Kenntnisse dieser quantitativen Analyseverfahren unverzichtbar.

Für die Sozialwissenschaften stellt die Statistik eine zentrale Hilfswissenschaft dar. Während sich Statistiker – allgemeiner: Mathematikerinnen – häufig mit Beweisführung und der Weiterentwicklung mathematischer Algorithmen beschäftigen, steht für Studierende der Politikwissenschaft, Verwaltungswissenschaft und Soziologie das Kennenlernen und die praktische Anwendung statistischer Verfahren im Vordergrund. Im Mittelpunkt des Kurses steht das Verständnis quantitativer Analyseverfahren, mit denen Sozialwissenschaftlerinnen und Sozialwissenschaftler bei der Auseinandersetzung mit quantitativen Studien konfrontiert werden.

Der vorliegende Kurs behandelt vier Themenbereiche: Univariate, bivariate und multivariate Datenanalyse sowie Grundlagen der Inferenzstatistik. Das Kapitel zur univariaten Datenanalyse behandelt die Häufigkeitsverteilung einzelner Merkmale. Dabei werden Lage- und Streuungsmaße sowie Formmaße vorgestellt. Die bivariate Datenanalyse untersucht Zusammenhänge zwischen zwei Merkmalen. Dabei werden Kreuztabellen sowie wichtige Zusammenhangsmaße behandelt. Bei der multivariaten Datenanalyse werden mit der linearen und logistischen Regression zwei zentrale Analyseverfahren der Sozialwissenschaften vorgestellt, die den Einfluss mehrerer unabhängiger Variablen auf eine abhängige Variable schätzen können. Aus zeitlichen, finanziellen und forschungspraktischen Gründen dominieren in den Sozialwissenschaften Stichproben. Deshalb behandelt der vierte Teil des Kurses die Grundlagen der Inferenzstatistik, die Instrumente zur Verfügung stellt, um zu entscheiden, ob und wie empirische Befunde aus Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden dürfen.

Bei dem vorliegenden Kurs handelt es sich um eine erste Arbeitsversion, die in den folgenden Semestern aktualisiert und erweitert wird. Deshalb möchte ich Sie an dieser Stelle bitten, mich auf Fehler und Verbesserungsvorschläge hinzuweisen. In der Moodle-Lernumgebung des Moduls M1 „Quantitative Methoden der Sozialwissenschaften“ findet sich eine Errata-Liste zu dem Kurs. Außerdem werden dort auch Übungsaufgaben und Tests veröffentlicht, die die Auseinandersetzung mit den Inhalten des Kurses fördern sollen. Mittelfristig wird dieser Kurs den aktuell eingesetzten Kurs 33209 „Statistik“ ersetzen. Bitte beachten Sie deshalb auch die Hinweise in der Moodle-Lernumgebung.

Hagen, im Mai 2018

Markus Tausendpfund

Inhaltsverzeichnis

Abbildungsverzeichnis	VII
Tabellenverzeichnis	VIII
1 Univariate Datenanalyse	10
1.1 Häufigkeitstabelle	10
1.2 Lagemaße	14
1.2.1 Modus	14
1.2.2 Median	15
1.2.3 Arithmetisches Mittel	17
1.3 Streuungsmaße	20
1.3.1 Varianz	20
1.3.2 Standardabweichung	24
1.4 Formmaße	24
1.4.1 Schiefe	25
1.4.2 Wölbung	28
1.5 Variablen standardisieren (z-Transformation)	29
2 Bivariate Datenanalyse	31
2.1 Kreuztabellen	32
2.1.1 Relative Häufigkeiten und Prozentwerte	34
2.1.2 Gesamtprozente	35
2.1.3 Spaltenprozente	35
2.1.4 Zeilenprozente	36
2.1.5 Prozentsatzdifferenzen	37
2.1.6 Berechnung von Kreuztabellen	38
2.1.7 Hinweise und praktische Tipps	42
2.2 Cramér's V	43
2.2.1 Vorgehen bei der Berechnung von Cramér's V	45
2.2.2 Kontingenztabelle	46
2.2.3 Berechnung der Summenzahlen	47
2.2.4 Indifferenztabelle	47
2.2.5 Arbeitstabelle	49
2.2.6 Differenz von beobachteten und erwarteten Häufigkeiten	49
2.2.7 Quadrierung der Differenzen	50
2.2.8 Division der quadrierten Differenzen durch die erwarteten Häufigkeiten	51

2.2.9	Summe der Quotienten	51
2.2.10	Berechnung von Cramér's V	52
2.2.11	Interpretation von Cramér's V	52
2.3	Spearman's rho	53
2.3.1	Vorgehen bei der Berechnung von Spearman's rho	55
2.3.2	Bestimmung der Rangpositionen	56
2.3.3	Bestimmung der Differenzen zwischen den Rangpositionen	56
2.3.4	Bestimmung der quadrierten Differenzen	57
2.3.5	Bestimmung der Summe der quadrierten Differenzen	57
2.3.6	Berechnung von Spearman's rho	58
2.3.7	Interpretation von Spearman's rho	58
2.4	Pearson's r	59
2.4.1	Vorgehen bei der Berechnung von Pearson's r	61
2.4.2	Bestimmung der Produkte aus den Werten	62
2.4.3	Bestimmung der quadrierten Werte	63
2.4.4	Berechnung von Pearson's r	63
2.4.5	Interpretation von Pearson's r	64
3	Multivariate Datenanalyse	65
3.1	Einführung	65
3.2	Lineare Regression	67
3.2.1	Bivariate Regression	68
3.2.2	Multiple Regression	75
3.3	Logistische Regression	86
3.3.1	Bivariate Regression	87
3.3.2	Multiple Regression	90
4	Inferenzstatistik	96
4.1	Was ist das Problem?	96
4.2	Zentrale Konzepte der Inferenzstatistik	101
4.2.1	Zentraler Grenzwertsatz und Normalverteilung	101
4.2.2	Standardfehler	104
4.3	Schätzungsarten	110
4.3.1	Punktschätzung	110
4.3.2	Intervallschätzung	113
4.3.3	Berechnung der benötigten Fallzahl	121

4.3.4	Anwendungsprobleme in der Praxis.....	123
4.4	t-Test.....	125
4.4.1	Grundlagen.....	125
4.4.2	Durchführung eines t-Tests.....	128
5	Literatur.....	133

Abbildungsverzeichnis

Abbildung 1: Normalverteilung.....	25
Abbildung 2: Schiefe von Verteilungen.....	26
Abbildung 3: Empirische Verteilungen mit unterschiedlicher Schiefe.....	27
Abbildung 4: Wölbung.....	28
Abbildung 5: Streudiagramm mit Beispieldaten.....	70
Abbildung 6: Streudiagramm mit OLS-Regressionsgerade.....	72
Abbildung 7: Schematische Darstellung der vermuteten multivariaten Einflusstuktur.....	77
Abbildung 8: Bivariate logistische Regression (Beispieldaten).....	89
Abbildung 9: Grundgesamtheit und Stichprobe.....	96
Abbildung 10: Rückschluss von der Stichprobe auf die Grundgesamtheit.....	97
Abbildung 11: Wiederholte Ziehung von Zufallsstichproben.....	102
Abbildung 12: Normalverteilung.....	104
Abbildung 13: Abweichungen einzelner Stichprobenmittelwerte vom wahren Mittelwert.....	105
Abbildung 14: Stichprobenverteilungen bei unterschiedlicher Fallzahl.....	106
Abbildung 15: Ergebnisse des Politbarometers zu zwei Zeitpunkten (Angaben in Prozent).....	114
Abbildung 16: 95-Prozent-Konfidenzintervall.....	115
Abbildung 17: 99-Prozent-Konfidenzintervall.....	116
Abbildung 18: Fiktive Befragung zur Wahlentscheidung von 1000 Personen (in Prozent).....	118
Abbildung 19: 95-Prozent-Konfidenzintervalle (Stichprobengröße jeweils 1000 Personen).....	120

Tabellenverzeichnis

Tabelle 1: Interesse an Politik.....	10
Tabelle 2: Subjektive Schichtestufung	13
Tabelle 3: Lagemaße und Skalenniveau	14
Tabelle 4: Berechnung des Modus.....	14
Tabelle 5: Geschlecht	15
Tabelle 6: Berechnung des Medians (ungerade Fallzahl).....	16
Tabelle 7: Berechnung des Medians (gerade Fallzahl).....	16
Tabelle 8: Interesse an Politik.....	17
Tabelle 9: Berechnung des arithmetischen Mittels bei kleinen Fallzahlen	18
Tabelle 10: Lebenszufriedenheit	19
Tabelle 11: Mittelwerte und Ausreißer.....	19
Tabelle 12: Lebenszufriedenheit von zwei Gruppen	20
Tabelle 13: Arbeitstabelle für die Berechnung der Varianz (kleine Fallzahl)	22
Tabelle 14: Arbeitstabelle für die Berechnung der Varianz (große Fallzahl)	23
Tabelle 15: Variablen standardisieren	29
Tabelle 16: Wichtige Zusammenhangsmaße bei der bivariaten Datenanalyse.....	31
Tabelle 17: Kombinierte Häufigkeitsverteilung in einer Kreuztabelle	32
Tabelle 18: Beispiel mit zehn Murmeln (Urliste).....	33
Tabelle 19: Beispiel mit zehn Murmeln (Kreuztabelle)	34
Tabelle 20: Beispiel mit zehn Murmeln (Kreuztabelle mit Gesamtprozenten).....	35
Tabelle 21: Beispiel mit zehn Murmeln (Kreuztabelle mit Spaltenprozenten)	36
Tabelle 22: Beispiel mit zehn Murmeln (Kreuztabelle mit Zeilenprozenten).....	37
Tabelle 23: Interpretation von Prozentsatzdifferenzen	38
Tabelle 24: Geschlecht und Politikinteresse.....	38
Tabelle 25: Kreuztabelle mit absoluten Häufigkeiten (ohne Summenzahlen)	39
Tabelle 26: Kreuztabelle mit absoluten Häufigkeiten	40
Tabelle 27: Berechnung der Spaltenprozentage in einer Kreuztabelle (Beispiel 1).....	40
Tabelle 28: Berechnung der Spaltenprozentage in einer Kreuztabelle (Beispiel 2).....	41
Tabelle 29: Kreuztabelle mit absoluten Häufigkeiten und Spaltenprozenten	41
Tabelle 30: Kontingenztafel	44
Tabelle 31: Interpretation von Cramér's V	44
Tabelle 32: Geschlecht und politisches Interesse	46
Tabelle 33: Kontingenztafel	46
Tabelle 34: Berechnung der Summenzahlen	47
Tabelle 35: Indifferenztafel (Beispiel 1)	48
Tabelle 36: Indifferenztafel (Beispiel 2)	48
Tabelle 37: Vollständige Indifferenztafel.....	49
Tabelle 38: Arbeitstabelle	49
Tabelle 39: Berechnung der Differenz der beobachteten und erwarteten Häufigkeiten	50
Tabelle 40: Quadrierung der Differenz der beobachteten und erwarteten Häufigkeiten	50
Tabelle 41: Division der quadrierten Differenzen durch erwartete Häufigkeiten.....	51
Tabelle 42: Berechnung von Chi-Quadrat	51
Tabelle 43: Interpretation von Spearman's rho	54

Tabelle 44: Schulabschluss und politisches Interesse	55
Tabelle 45: Urliste.....	56
Tabelle 46: Rangpositionen bestimmen	56
Tabelle 47: Differenz der Rangpositionen	57
Tabelle 48: Rangpositionen quadriert	57
Tabelle 49: Summe der quadrierten Differenzen	57
Tabelle 50: Interpretation von Pearson's r	60
Tabelle 51: Urliste.....	61
Tabelle 52: Produkt berechnen	62
Tabelle 53: Berechnung der quadrierten Werte.....	63
Tabelle 54: Angaben zur Berechnung von Pearson's r	63
Tabelle 55: Unterschiedliche Bezeichnungen für Variablen der Regressionsanalyse	66
Tabelle 56: Fiktive Beispieldaten für bivariate Regression.....	69
Tabelle 57: Dummy-Kodierung für Familienstand.....	78
Tabelle 58: Bestimmungsfaktoren der Lebenszufriedenheit.....	80
Tabelle 59: Bestimmungsfaktoren der Lebenszufriedenheit (Regressionskoeffizienten)	84
Tabelle 60: Fiktive Beispieldaten für Wahlbeteiligung und Alter.....	88
Tabelle 61: Bestimmungsfaktoren der Wahlbeteiligung	92
Tabelle 62: Mittelwerte in Zufallsstichproben (Stichprobengröße jeweils 1000 Personen)	98
Tabelle 63: Mittelwerte von Zufallsstichproben (Stichprobengröße jeweils 1000 Personen).....	100
Tabelle 64: Vergleich zwischen Standardfehler und Standardabweichung	107
Tabelle 65: Mittelwerte von Zufallsstichproben	111
Tabelle 66: Erforderliche Stichprobengröße	122
Tabelle 67: Kritische t-Werte	127
Tabelle 68: Geschlecht und Nettoeinkommen.....	128
Tabelle 69: Urliste der Männer.....	129
Tabelle 70: Berechnung der Varianz der Männer	129
Tabelle 71: Urliste der Frauen	130
Tabelle 72: Berechnung der Varianz der Frauen	130
Tabelle 73: Angaben für beide Stichproben	131

1 Univariate Datenanalyse

Markus Tausendpfund



Bei der univariaten Datenanalyse wird ein Merkmal bzw. eine Variable betrachtet. Im ersten Schritt wird eine Häufigkeitstabelle erstellt, welche die absoluten und relativen Häufigkeiten der einzelnen Ausprägungen einer Variable ausweist (Diaz-Bone 2017; Völkl und Korb 2018). In der quantitativen Sozialforschung sind Forscherinnen häufig mit vielen Untersuchungsobjekten konfrontiert. Deshalb wird im zweiten Schritt die große Informationsmenge auf wichtige Informationen verdichtet. Dabei lassen sich Lage-, Streuungs- und Formmaße unterscheiden. Während Lage- maße (z.B. Modus, Median und arithmetisches Mittel) über das Zentrum einer Verteilung informieren, beschreiben Streuungsmaße (z.B. Varianz und Standardabweichung) die Variation eines Merkmals in einer Verteilung. Die Form einer Verteilung wird mit der Schiefe und der Wölbung charakterisiert. Die z-Transformation (z-Standardisierung) ermöglicht den Vergleich von Werten unterschiedlicher Verteilungen.

1.1 Häufigkeitstabelle

Eine Häufigkeitstabelle gibt eine Übersicht über die Merkmalsausprägungen (Kategorien) einer Variable und wie häufig jede einzelne Ausprägung vorkommt (Ludwig-Mayerhofer et al. 2014, S. 41-45; Diaz-Bone 2017, S. 35). Tabelle 1 zeigt die Häufigkeiten des Interesses an Politik, die in der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) wie folgt erfasst wird:

Wie stark interessieren Sie sich für Politik?

Die Frageformulierung erfasst das subjektive Interesse an Politik (van Deth 2013, S. 275). Als Antwortmöglichkeit können die Befragten bei der ALLBUS zwischen „sehr stark“, „stark“, „mittel“, „wenig“ und „überhaupt nicht“ wählen. In Tabelle 1 sind die einzelnen Antwortkategorien und die jeweiligen Häufigkeiten dargestellt.

Tabelle 1: Interesse an Politik

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte prozentuale Häufigkeit
sehr stark	425	0,122	12,2	12,2
stark	877	0,251	25,1	37,3
mittel	1437	0,412	41,2	78,5
wenig	564	0,162	16,2	94,7
überhaupt nicht	186	0,053	5,3	100,0
Gesamt	3490	1,0	100,0	

Quelle: ALLBUS 2016 (doi: 10.4232/1.12796). Eigene Berechnungen.

Vier Angaben zur Häufigkeit lassen sich in Tabelle 1 unterscheiden: die absolute Häufigkeit, die relative Häufigkeit, die prozentuale Häufigkeit und die kumulierte prozentuale Häufigkeit.

Die absolute Häufigkeit ist die Anzahl der Fälle (hier: Personen), mit der die jeweilige Kategorie auftritt. Bei der ALLBUS 2016 liegen insgesamt Angaben von 3490 Personen zum Interesse an Politik vor. 425 Personen haben die Antwortkategorie „sehr stark“ gewählt, 877 Befragte haben sich für die Antwort „stark“ entschieden, 1437 Personen interessieren sich nach eigenen Angaben „mittel“ für Politik, 564 Bürger wählten die Antwortoption „wenig“ und 186 Befragte haben die Kategorie „überhaupt nicht“ ausgewählt.

Absolute Häufigkeiten

Die absoluten Häufigkeiten werden häufig mit dem Buchstaben f (engl. frequency) abgekürzt. Um die einzelnen Kategorien einer Variablen zu unterscheiden wird meist der Index j verwendet. Mit f_j wird also die absolute Häufigkeit einer bestimmten Kategorie dargestellt: f_j ist die absolute Häufigkeit, mit der die Kategorie j vorliegt.

Tabelle 1 umfasst neben den absoluten Häufigkeiten auch die relativen Häufigkeiten, die als Anteilswerte ausgewiesen werden. Die relativen Häufigkeiten werden mit p (engl. proportion) abgekürzt. Die relative Häufigkeit einer Kategorie j ist p_j . Die relative Häufigkeit (p_j) ist definiert als die absolute Häufigkeit (f_j) dividiert durch die Fallzahl (n):

Relative Häufigkeiten

$$p_j = \frac{f_j}{n}$$

Für die Berechnung der relativen Häufigkeit wird die absolute Fallzahl einer Kategorie durch die Gesamtfallzahl geteilt. Für die Berechnung der relativen Häufigkeit der Kategorie „sehr stark“ wird die absolute Fallzahl dieser Kategorie (425) durch die Gesamtzahl (3490) dividiert:

$$p_{\text{sehr stark}} = \frac{425}{3490} = 0,122$$

Die Summe der relativen Häufigkeiten aller Kategorien ergibt immer 1.

Durch die Multiplikation der relativen Häufigkeit mit 100 werden die prozentualen Häufigkeiten der einzelnen Kategorien berechnet. Die Prozentwerte weisen die gleichen Informationen wie die relativen Häufigkeiten aus, es handelt sich nur um verschiedene Darstellungsformen (Diaz-Bone 2017, S. 36). Die als Prozentwert dargestellte relative Häufigkeit wird als prozentuale Häufigkeit bezeichnet. Die Prozentwerte lassen sich nach folgender Formel berechnen:

Prozentuale Häufigkeiten

$$p_j\% = \frac{f_j}{n} \cdot 100$$

Für die Berechnung der prozentualen Häufigkeit der Kategorie „wenig“ wird die absolute Häufigkeit (564) durch die Gesamtzahl (3490) dividiert. Das Ergebnis (0,162) wird anschließend mit 100 multipliziert.

$$p_j\% = \frac{564}{3490} \cdot 100 = 16,2 \text{ Prozent}$$

16,2 Prozent der Befragten geben an, sich wenig für Politik zu interessieren.

Kumulierte prozentuale Häufigkeit

In der letzten Spalte in Tabelle 1 werden die kumulierten Prozentwerte ausgewiesen. Dabei handelt es sich um eine schrittweise Addition (Kumulation) der Prozentwerte. Die kumulierten Prozente der Kategorie „stark“ (37,3 Prozent) ist die Addition der Prozentwerte in den Kategorien „sehr stark“ (12,2 Prozent) und „stark“ (25,1 Prozent). Alternativ können auch die absoluten Häufigkeiten der beiden Kategorien addiert ($425+877=1302$) und durch die Gesamtzahl (3490) dividiert werden. Das Ergebnis (0,373) entspricht der kumulierten relativen Häufigkeit. Durch die Multiplikation mit 100 werden die kumulierte prozentuale Häufigkeit bzw. die kumulierten Prozente berechnet (37,3 Prozent).

Die kumulierten Prozentwerte bieten bei Merkmalen ab ordinalem Skalenniveau eine anschauliche Interpretationsmöglichkeit. Der Wert von 37,3 Prozent in Tabelle 1 kann wie folgt interpretiert werden: 37,3 Prozent der Befragten haben mindestens ein starkes Interesse an Politik.

Häufigkeitstabelle mit Missing Values

In der empirischen Sozialforschung kommt es häufig zu fehlenden Werten (engl. Missing Values). Bei Befragungen antworten die Personen bei einzelnen Fragen beispielsweise mit „weiß nicht“ oder verweigern die Angabe (Item-Nonresponse). Gelegentlich handelt es sich auch um Fehler bei der Dateneingabe. Wenn eine Variable fehlende Werte aufweist, dann bietet sich eine Form der Prozentuierung an, bei der nur die gültigen Angaben berücksichtigt werden (Kuckartz et al. 2013, S. 38; Ludwig-Mayerhofer et al. 2014, S. 42-43).

Das Vorgehen wird mit der Frage nach der subjektiven Schichteinstufung illustriert. Mit diesem Erhebungsinstrument wird die wahrgenommene Einordnung der Befragten in eine Bevölkerungsschicht ermittelt (Kleining und Moore 1968; Noll 1999). In der ALLBUS 2016 wird dieses Konzept wie folgt erfasst:

Es wird heute viel über die verschiedenen Bevölkerungsschichten gesprochen. Welcher Schicht rechnen Sie sich selbst eher zu?

Die Befragten können dabei zwischen der „Unterschicht“, „Arbeiterschicht“, „Mittelschicht“, „Oberen Mittelschicht“ und „Oberschicht“ wählen. Neben diesen (gültigen) Angaben konnten die Personen auch mit „keiner der Schichten“, „keine Angabe“, „weiß nicht“ oder „verweigert“ antworten. Tabelle 2 informiert über die Häufigkeitsangaben dieser Variable in der ALLBUS 2016. In der Spalte „Häufigkeit“ werden die absoluten Häufigkeiten der Angaben dokumentiert. Insgesamt liegen Angaben von 3490 Befragten vor. Davon haben 3443 Personen gültige Angaben gemacht, 47 Antworten werden als fehlend gewertet (Missing Values). In der Spalte „in Prozent“ werden die prozentualen Häufigkeiten der einzelnen Antwortkategorien ausgewiesen. Zu diesem Zweck werden die Häufigkeiten der einzelnen Kategorien durch die Gesamtfallzahl (3490) dividiert und mit 100 multipliziert. Der Prozentwert der Mittelschicht ergibt sich wie folgt:

$$p_j\% = \frac{2037}{3490} \cdot 100 = 58,4 \text{ Prozent}$$

Für die Berechnung der gültigen Prozente wird nicht durch die Fallzahl aller Fälle geteilt, sondern durch die Anzahl der gültigen Fälle (3443):

$$p_j\% = \frac{2037}{3443} \cdot 100 = 59,2 \text{ Prozent}$$

Bei den kumulierten Prozenten in Tabelle 2 handelt es sich um die Addition der (gültigen) Prozentwerte der einzelnen Kategorien. Die kumulierten Prozente der Mittelschicht (86,1 Prozent) ergeben sich durch die Addition der gültigen Prozentwerte der Unterschicht (2,4 Prozent), der Arbeiterschicht (24,5 Prozent) und der Mittelschicht (59,2 Prozent). Insgesamt 86,1 Prozent der Befragten ordnen sich der Unter-, Arbeiter- oder Mittelschicht zu.

Tabelle 2: Subjektive Schichteinstufung

Kategorie	absolute Häufigkeit	in Prozent	gültige Prozent	kumulierte Prozent
Unterschicht	84	2,4	2,4	2,4
Arbeiterschicht	844	24,2	24,5	27,0
Mittelschicht	2037	58,4	59,2	86,1
Obere Mittelschicht	462	13,2	13,4	99,5
Oberschicht	16	0,5	0,5	100,0
Gesamt	3443	98,6	100,0	
keiner der Schichten	17	0,5		
Einstufung abgelehnt	1	0,0		
weiß nicht	17	0,5		
keine Angabe	13	0,4		
Gesamt	47	1,4		
Gesamt	3490	100,0		

Quelle: ALLBUS 2016 (doi: 10.4232/1.12796). Eigene Berechnungen.

Die Prozentangaben einer Häufigkeitstabelle können nur dann angemessen interpretiert werden, wenn erstens die absolute Häufigkeit eines Merkmals und zweitens die Art der Prozentuierung bekannt sind (Gehring und Weins 2009, S. 102-104).

Die Angabe von Prozentwerten ohne eine Information zur absoluten Häufigkeit ist irreführend, da es natürlich einen erheblichen Unterschied macht, ob die Prozentangabe auf Grundlage von 30 oder 3000 Fällen ermittelt wurde. Insbesondere bei kleinen Fallzahlen haben Veränderungen der absoluten Häufigkeiten starke Auswirkungen auf die Prozentangaben. Deshalb sollte eine Häufigkeitstabelle immer Angaben zur absoluten Häufigkeit enthalten. Bei weniger als 100 Fällen ($n < 100$) ist zu prüfen, ob die Prozentangabe inhaltlich sinnvoll interpretiert werden kann.

Neben der Fallzahl ist auch die Prozentuierungsbasis für eine gehaltvolle Interpretation wichtig. Es spielt natürlich eine Rolle, welche Kategorien als „ungültige Werte“ gewertet und wie viele Fälle bei der Berechnung der „gültigen Prozente“ ausgeschlossen werden.

