

Markus Tausendpfund
Simone Abendschön

Quantitative Analyseverfahren. Eine Einführung

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



FernUniversität in Hagen

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 Statistikerinnen – allgemeiner: Mathematikerinnen – häufig mit der Beweisführung und der Weiterentwicklung mathematischer Algorithmen beschäftigen, stehen 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 und Unterschiede zwischen zwei Merkmalen (Mittelwertvergleiche). 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.

In der Moodle-Lernumgebung des Moduls M1 „Quantitative Methoden der Sozialwissenschaften“ findet sich eine Errata-Liste zu dem Kurs. Außerdem werden dort Videos und Übungsaufgaben veröffentlicht, die die Auseinandersetzung mit den Inhalten des Kurses fördern sollen. Für die kritische Durchsicht des Kurses sind wir Christian Cleve und Daniel Saar sehr dankbar.

Über Hinweise auf Fehler, Kommentare und Verbesserungsvorschläge freuen wir uns. Senden Sie Ihre Kommentare bitte an Markus.Tausendpfund@fernuni-hagen.de. Vielen Dank.

Hagen, im Juni 2020

Markus Tausendpfund und Simone Abendschön

Inhaltsverzeichnis

Abbildungsverzeichnis	VI
Tabellenverzeichnis	VII
1 Einführung	9
1.1 Einordnung im Forschungsprozess	9
1.2 Grundgesamtheit und Stichprobe	12
1.3 Klassifikationen von Variablen	14
2 Univariate Datenanalyse	17
2.1 Häufigkeitstabelle	17
2.2 Lagemaße	21
2.2.1 Modus	21
2.2.2 Median	22
2.2.3 Arithmetisches Mittel	24
2.3 Streuungsmaße	27
2.3.1 Varianz	27
2.3.2 Standardabweichung	31
2.4 Formmaße	31
2.4.1 Schiefe	32
2.4.2 Wölbung	35
2.5 Variablen standardisieren (z-Transformation)	36
2.6 Grafische Darstellungen	38
2.6.1 Säulen- und Balkendiagramm	38
2.6.2 Kreisdiagramm	39
2.6.3 Histogramm	40
2.6.4 Boxplot	42
3 Bivariate Datenanalyse	44
3.1 Kreuztabellen	45
3.2 Zusammenhangsmaße für nominale Merkmale	53
3.3 Zusammenhangsmaße für ordinale Merkmale	59
3.4 Zusammenhangsmaße für metrische Merkmale	63
3.5 Eta-Quadrat für metrische und nominale Merkmale	72
3.6 Zusammenfassung	77
4 Multivariate Datenanalyse	78
4.1 Einführung	78

4.2	Lineare Regression	80
4.2.1	Bivariate Regression	81
4.2.2	Multiple Regression	88
4.3	Logistische Regression.....	99
4.3.1	Bivariate Regression	100
4.3.2	Multiple Regression	103
5	Inferenzstatistik	109
5.1	Was ist das Problem?	109
5.2	Zentrale Konzepte der Inferenzstatistik.....	114
5.2.1	Zentraler Grenzwertsatz und Normalverteilung	114
5.2.2	Standardfehler	117
5.3	Schätzungsarten	123
5.3.1	Punktschätzung	123
5.3.2	Intervallschätzung	126
5.3.3	Berechnung der benötigten Fallzahl	134
5.3.4	Anwendungsprobleme in der Praxis	136
5.4	Statistisches Testen	138
5.4.1	Allgemeine Vorgehensweise bei einem Signifikanztest	140
5.4.2	Alpha- und Beta-Fehler	143
5.4.3	t-Test.....	144
6	Literatur	159

Abbildungsverzeichnis

Abbildung 1: Idealtypischer Ablauf eines quantitativen Forschungsprojekts	10
Abbildung 2: Grundgesamtheit und Stichprobe.....	13
Abbildung 3: Normalverteilung	32
Abbildung 4: Schiefe	33
Abbildung 5: Empirische Verteilungen mit unterschiedlicher Schiefe	34
Abbildung 6: Wölbung.....	35
Abbildung 7: Säulendiagramm des Interesses an Politik (in Prozent, n = 3490)	39
Abbildung 8: Balkendiagramm des Interesses an Politik (absolute Häufigkeiten, n = 3490)	39
Abbildung 9: Zweitstimmen bei der Bundestagswahl 2017 (in Prozent)	40
Abbildung 10: Histogramm des Alters (absolute Häufigkeiten, n = 3486)	41
Abbildung 11: Elemente eines Boxplots.....	42
Abbildung 12: Boxplot der Interviewdauer (n = 3479)	43
Abbildung 13: IQ und Testergebnis beim räumlichen Denken – Streudiagramm	64
Abbildung 14: Weitere Arten des Zusammenhangs von zwei Merkmalen	65
Abbildung 15: Nettoeinkommen und Lebenszufriedenheit – Streudiagramm.....	70
Abbildung 16: Streudiagramm	83
Abbildung 17: Streudiagramm mit OLS-Regressionsgerade.....	85
Abbildung 18: Schematische Darstellung der vermuteten multivariaten Einflusstruktur.....	90
Abbildung 19: Streudiagramm mit Regressionskurve	102
Abbildung 20: Grundgesamtheit und Stichprobe.....	109
Abbildung 21: Rückschluss von der Stichprobe auf die Grundgesamtheit	110
Abbildung 22: Wiederholte Ziehung von Zufallsstichproben	115
Abbildung 23: Normalverteilung	117
Abbildung 24: Abweichungen einzelner Stichprobenmittelwerte vom wahren Mittelwert.....	118
Abbildung 25: Stichprobenverteilungen bei unterschiedlicher Fallzahl.....	119
Abbildung 26: Ergebnisse des Politbarometers zu zwei Zeitpunkten (in Prozent)	126
Abbildung 27: 95-Prozent-Konfidenzintervall	127
Abbildung 28: 99-Prozent-Konfidenzintervall	128
Abbildung 29: Fiktive Befragung zur Wahlentscheidung von 1000 Personen (in Prozent)	131
Abbildung 30: 95-Prozent-Konfidenzintervalle (Stichprobengröße jeweils 1000 Personen).....	133
Abbildung 31: Schätzen und Testen im Vergleich	139
Abbildung 32: t-Verteilung und Normalverteilung	146
Abbildung 33: Verschiedene t-Verteilungen	147
Abbildung 34: Varianten des t-Tests.....	147
Abbildung 35: Einseitiger und zweiseitiger t-Test.....	149

Tabellenverzeichnis

Tabelle 1: Zulässige Rechenoperationen in Abhängigkeit vom Skalenniveau	15
Tabelle 2: Interesse an Politik	17
Tabelle 3: Subjektive Schichteinstufung	20
Tabelle 4: Lagemaße und Skalenniveau	21
Tabelle 5: Berechnung des Modus	21
Tabelle 6: Geschlecht	22
Tabelle 7: Berechnung des Medians (ungerade Fallzahl)	23
Tabelle 8: Berechnung des Medians (gerade Fallzahl)	23
Tabelle 9: Interesse an Politik	24
Tabelle 10: Berechnung des arithmetischen Mittels bei kleinen Fallzahlen	25
Tabelle 11: Berechnung des arithmetischen Mittels bei großen Fallzahlen	26
Tabelle 12: Mittelwerte und Ausreißer	26
Tabelle 13: Lebenszufriedenheit von zwei Gruppen	27
Tabelle 14: Arbeitstabelle für die Berechnung der Varianz (kleine Fallzahl)	29
Tabelle 15: Arbeitstabelle für die Berechnung der Varianz (große Fallzahl)	30
Tabelle 16: Variablen standardisieren	37
Tabelle 17: Bivariate Zusammenhangsmaße in Abhängigkeit vom Skalenniveau	44
Tabelle 18: Urliste – Abendliche Bibliotheksnutzung und Studiengang (n = 9)	46
Tabelle 19: Kreuztabelle – Abendliche Bibliotheksnutzung und Studiengang (n = 9)	46
Tabelle 20: Abendliche Bibliotheksnutzung und Studiengang – Zeilenprozentage (n = 100)	47
Tabelle 21: Abendliche Bibliotheksnutzung und Studiengang – Spaltenprozentage (n = 100)	48
Tabelle 22: Abendliche Bibliotheksnutzung und Studiengang – Gesamtprozentage (n = 100)	48
Tabelle 23: Politisches Interesse und Geschlecht (Spaltenprozentage)	49
Tabelle 24: Schulabschluss und elterlicher Bildungshintergrund (Spaltenprozentage)	52
Tabelle 25: Schulabschluss und elterlicher Bildungshintergrund (Zeilenprozentage)	53
Tabelle 26: Politisches Interesse und Geschlecht (beobachtete Häufigkeiten) – Kontingenztafel	54
Tabelle 27: Berechnung der erwarteten Häufigkeiten	54
Tabelle 28: Politisches Interesse und Geschlecht (erwartete Häufigkeiten) – Indifferenztafel	55
Tabelle 29: Arbeitstabelle zur Berechnung von Chi-Quadrat	56
Tabelle 30: Interpretation von Cramer's V	58
Tabelle 31: Interpretation von Spearman's Rho	60
Tabelle 32: Soziale Schicht und Gesundheitszustand	61
Tabelle 33: Arbeitstabelle zur Berechnung von Spearman's Rho	62
Tabelle 34: IQ und Testergebnis beim räumlichen Denken – Urliste	63
Tabelle 35: Arbeitstabelle zur Berechnung der Kovarianz	66
Tabelle 36: Interpretation von Pearson's r	67
Tabelle 37: Arbeitstabelle zur Berechnung von Pearson's r	68
Tabelle 38: Nettoeinkommen und Lebenszufriedenheit – Urliste	69
Tabelle 39: Arbeitstabelle zur Berechnung von Pearson's r	71
Tabelle 40: Zwischenergebnisse zur Berechnung von Pearson's r	71
Tabelle 41: Interpretation von Eta-Quadrat	74
Tabelle 42: Migrationshintergrund und politisches Wissen	74
Tabelle 43: Arbeitstabelle Migrationshintergrund und politisches Wissen	75

Tabelle 44: Arbeitstabelle Migrationshintergrund (Nein) und politisches Wissen	76
Tabelle 45: Arbeitstabelle Migrationshintergrund (Ja) und politisches Wissen	76
Tabelle 46: Unterschiedliche Bezeichnungen für Variablen der Regressionsanalyse	79
Tabelle 47: Bivariate lineare Regression mit Lebenszufriedenheit und Einkommen	82
Tabelle 48: Dummy-Kodierung für Familienstand	91
Tabelle 49: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 1)	93
Tabelle 50: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 2)	97
Tabelle 51: Bivariate logistische Regression mit Wahlbeteiligung und Alter	101
Tabelle 52: Bestimmungsfaktoren der Wahlbeteiligung	105
Tabelle 53: Mittelwerte in Zufallsstichproben (Stichprobengröße jeweils 1000 Personen)	111
Tabelle 54: Mittelwerte von Zufallsstichproben (Stichprobengröße jeweils 1000 Personen)	113
Tabelle 55: Vergleich zwischen Standardfehler und Standardabweichung	120
Tabelle 56: Mittelwerte von Zufallsstichproben	124
Tabelle 57: Erforderliche Stichprobengröße	135
Tabelle 58: Fehlerarten beim Hypothesentest	143
Tabelle 59: Lebenszufriedenheit von Frauen und Männern	150
Tabelle 60: Kritische Werte der t-Verteilung	152
Tabelle 61: Lebenszufriedenheit von West- und Ostdeutschen	153
Tabelle 62: Zufriedenheit mit der Demokratie	155
Tabelle 63: Beispieldaten für die Berechnung eines t-Tests bei abhängigen Stichproben	156

1 Einführung

Markus Tausendpfund

Vorschau



Dieses Kapitel macht Sie mit den Grundlagen der quantitativen Datenanalyse vertraut.¹ Nach der Einordnung der Phase „Datenanalyse“ innerhalb des Forschungsprozesses werden die Begriffe „Grundgesamtheit“ und „Stichprobe“ erläutert. Bei einer empirischen Studie werden meist Aussagen über größere Gruppen angestrebt (z.B. die wahlberechtigte Bevölkerung in Deutschland). Allerdings liegen in den meisten Studien keine Informationen über alle Elemente dieser Gruppe vor, sondern nur über eine (zufällige) Auswahl dieser Gruppe. Die Gruppe, über die eine Aussage gemacht werden soll, wird als Grundgesamtheit oder Population bezeichnet. Die Gruppe, über die empirische Informationen vorliegen, wird als Stichprobe bezeichnet. Diese Begriffe werden knapp erläutert und es werden die Voraussetzungen skizziert, unter denen Befunde einer Stichprobe auf die zugehörige Grundgesamtheit übertragen werden können. Abschließend werden typische Klassifikationen von Variablen vorgestellt. Dabei liegt der Fokus auf dem Skalenniveau von Variablen, da das Skalenniveau eine wichtige Voraussetzung für die Anwendung bestimmter Analyseverfahren ist.

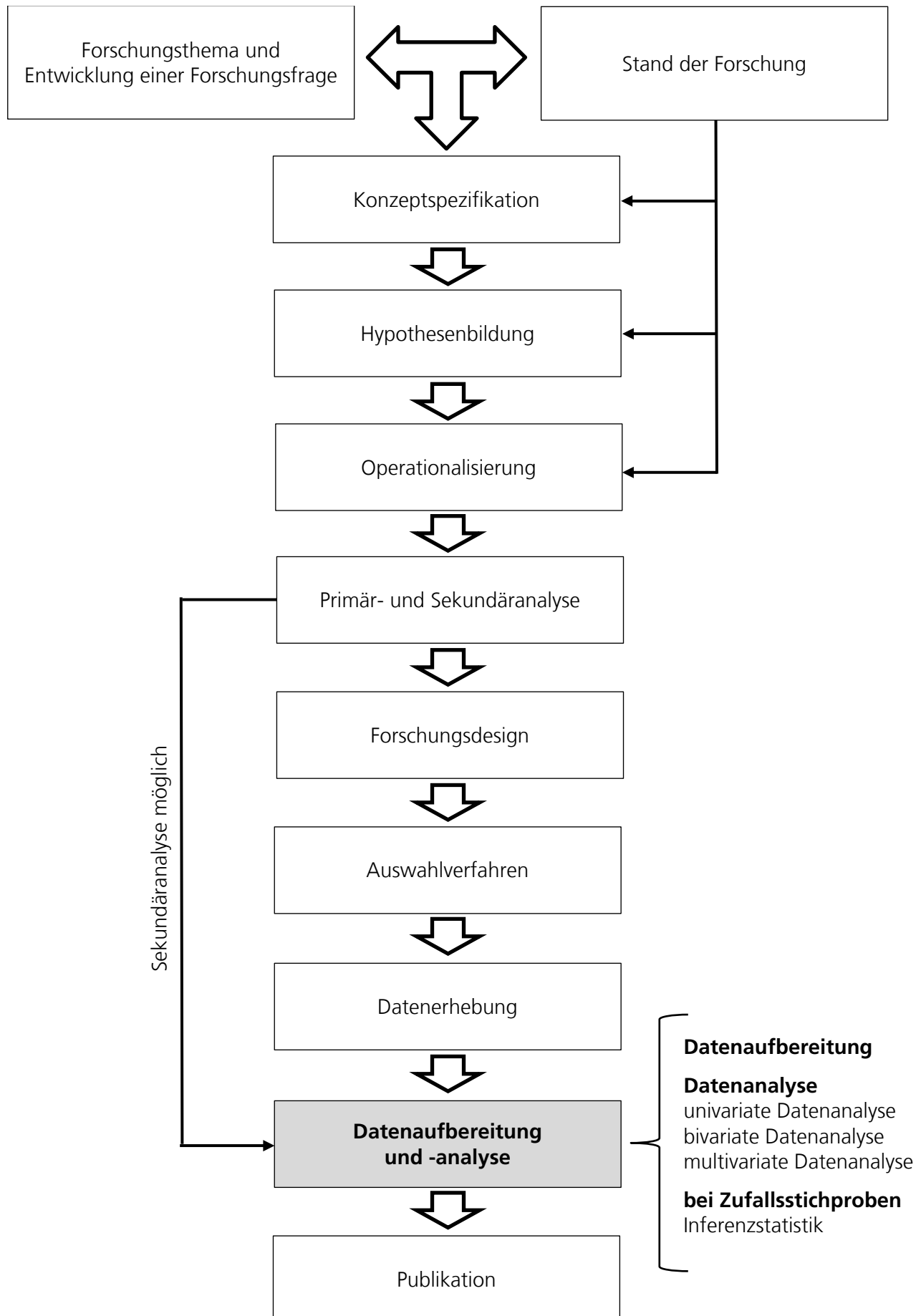
1.1 Einordnung im Forschungsprozess

Die quantitativen Analyseverfahren werden häufig mit dem quantitativen Forschungsprozess gleichgesetzt. Quantitativ arbeitende Sozialwissenschaftlerinnen nutzen statistische Analyseverfahren, um die theoretisch formulierten Hypothesen empirisch zu überprüfen. Sicherlich ist die Anwendung statistischer Analyseverfahren ein zentrales Merkmal des quantitativen Forschungsprozesses, aber die quantitative Datenanalyse sollte nicht isoliert betrachtet werden.

Vor der Datenanalyse bzw. Anwendung quantitativer Analyseverfahren müssen empirische Sozialforscher wichtige vorgelagerte Entscheidungen treffen, die unmittelbare Auswirkungen auf die empirischen Befunde haben. Wie Abbildung 1 zeigt, stehen die Festlegung eines Forschungsthemas und die Entwicklung einer geeigneten Forschungsfrage am Beginn eines Forschungsprojekts. Auf dieser Grundlage werden die zentralen Konzepte identifiziert und theoretisch geklärt, ehe inhaltvolle Hypothesen formuliert und valide Operationalisierungen dieser Konzepte entwickelt werden. Diese Phasen in einem Forschungsprozess erfolgen in intensiver Auseinandersetzung mit dem existierenden Forschungsstand. Nur wer den Forschungsstand zu seinem Forschungsthema kennt, kann eine inhaltvolle Forschungsfrage entwickeln. Die Auseinandersetzung mit der Fachliteratur ist aber auch für die Konzeptspezifikation und die Entwicklung von Hypothesen erforderlich. Schließlich ist auch bei der „Übersetzung“ theoretischer Konzepte in empirische Indikatoren ein Überblick über existierende Operationalisierungen notwendig.

¹ Ausschließlich aus Gründen der besseren Lesbarkeit wird in diesem Kurs nicht durchgängig eine geschlechterneutrale Sprache verwendet. Männliche, weibliche und genderneutrale Formen wechseln sich in diesem Kurs zufallsverteilt ab. Mit den Bezeichnungen sind jeweils alle Geschlechter gemeint.

Abbildung 1: Idealtypischer Ablauf eines quantitativen Forschungsprojekts



Quelle: Eigene Darstellung

Kein Analyseverfahren kann die intensive Auseinandersetzung mit dem existierenden Forschungsstand ersetzen. Ungeeignete Konzeptspezifikationen, schwammige Hypothesen oder auch ungültige Operationalisierungen führen zwangsläufig zu schlechten Daten und kein Analyseverfahren der Welt kann aus schlechten Daten valide empirische Befunde machen. Deshalb: Die Anwendung bzw. Durchführung quantitativer Analyseverfahren kann nur dann zu belastbaren empirischen Befunden führen, wenn die vorgelagerten Phasen erfolgreich bearbeitet wurden.

Wenn für die Bearbeitung einer Forschungsfrage und die Überprüfung der Hypothesen bereits geeignetes Datenmaterial existiert (z.B. ALLBUS), dann können die Phasen „Forschungsdesign“, „Auswahlverfahren“ und „Datenerhebung“ übersprungen werden. In einem solchen Fall führt die Sozialwissenschaftlerin eine Sekundäranalyse durch. Es werden existierende Daten genutzt, um die Forschungsfrage zu bearbeiten. Falls keine geeigneten Daten zur Verfügung stehen, bietet sich eine Primäranalyse an. Bei einer Primäranalyse werden neue Daten erhoben, um die Forschungsfrage zu beantworten.

Die Phase „Datenaufbereitung und -analyse“ umfasst in der Regel mehrere Zwischenschritte (Stein 2014, S. 150; Tausendpfund 2018b, S. 50-51). Zunächst müssen die im Rahmen der Datenerhebung gesammelten empirischen Informationen systematisch in einen Datensatz aufgenommen werden (Kromrey et al. 2016, S. 217-218). Die Variablen müssen beschriftet und ein Codebuch muss angelegt werden (z.B. Lück und Baur 2011; Lück und Landrock 2014; Tausendpfund 2018b, S. 291-297). Bei der Arbeit mit qualitativ hochwertigen Sekundärdaten (z.B. ALLBUS) stehen meist „fertige“ Datensätze zur Verfügung. Insbesondere bei der eigenständigen Dateneingabe, aber auch bei der Arbeit mit Sekundärdaten, sind Fehlerkontrollen (z.B. Eingabefehler) und Plausibilitätstests erforderlich.

Datenaufbereitung und -analyse

Vor der eigentlichen Datenanalyse müssen Variablen häufig verändert oder neu erstellt werden. Dieser Prozess wird häufig als Datenmodifikation oder Datentransformation bezeichnet (Fromm 2011; Kohler und Kreuter 2017, S. 91-130; Tausendpfund 2018a, S. 62-91). Dabei wird die Kodierung von Variablen angepasst, einzelne Subgruppen werden gebildet oder es werden auf Basis der verfügbaren Informationen auch neue Variablen erstellt. Das Verändern und das Erstellen neuer Variablen dauert häufig länger als die eigentliche Datenanalyse. Eine sorgfältige Durchführung der einzelnen Schritte ist dabei eine Voraussetzung für die Gültigkeit der anschließenden Analysen.

Bei der anschließenden Datenanalyse lassen sich meist vier Schritte unterscheiden, die auch im Mittelpunkt dieses Kurses stehen: die univariate, die bivariate und die multivariate Datenanalyse sowie die Inferenzstatistik.

Die univariate Datenanalyse befasst sich mit einzelnen Variablen. In einem ersten Schritt werden die absoluten und relativen Häufigkeiten der einzelnen Ausprägungen einer Variable in Tabellen oder Grafiken dargestellt. In der quantitativen Sozialforschung sind wir allerdings in der Regel mit vielen Untersuchungsobjekten konfrontiert. Deshalb wird in einem zweiten Schritt die Informationsmenge von mehreren tausend Beobachtungen auf wenige Kennzahlen verdichtet. Dabei lassen sich Lage-, Streuungs- und Formmaße unterscheiden. Während Lagemaße über das Zentrum einer Verteilung informieren, beschreiben Streuungsmaße

Univariate Datenanalyse

die Variation eines Merkmals in einer Verteilung. Mit Schiefe und Wölbung kann die Form einer Verteilung charakterisiert werden.

Bivariate Datenanalyse

Bei der bivariaten Datenanalyse werden immer genau zwei Variablen in Beziehung gesetzt (z.B. Bildung und Einkommen). Bivariate Analyseverfahren werden genutzt, um Zusammenhänge oder Unterschiede zwischen zwei Merkmalen zu untersuchen und Hypothesen empirisch zu überprüfen. Dafür nutzen wir Kreuztabellen und Zusammenhangsmaße. Kreuztabellen (engl. crosstabs) sind eine einfache und anschauliche Möglichkeit, um die Beziehung zwischen zwei Merkmalen in den Blick zu nehmen. Neben absoluten Häufigkeiten können auch die Anteile der einzelnen Häufigkeiten (Anteile) berechnet werden. Die Stärke einer Beziehung zwischen zwei Merkmalen (z.B. Bildung und Einkommen) kann mit Zusammenhangsmaßen – sogenannten Koeffizienten – charakterisiert werden. Die bekanntesten Zusammenhangsmaße sind sicherlich Cramér's V, Spearman's rho und Pearson's r.

Multivariate Datenanalyse

Mit bivariaten Analyseverfahren wird der Zusammenhang zwischen zwei Variablen untersucht. In der Realität können Merkmale wie Einkommen oder Wahlbeteiligung aber nicht durch eine Variable „erklärt“ werden, sondern in der Regel muss eine Vielzahl an Variablen gleichzeitig berücksichtigt werden. Mit der Regressionsanalyse steht ein sehr mächtiges Analyseinstrument zur Verfügung, um den Einfluss einer oder mehrerer unabhängiger Variablen auf eine abhängige Variable zu schätzen. Mit der linearen und logistischen Regression werden in diesem Kurs zwei multivariate Analyseverfahren vorgestellt, die in den Sozialwissenschaften häufig verwendet werden.

Inferenzstatistik

Die univariate, bivariate und multivariate Datenanalyse haben das Ziel, die Verteilung von Variablen zu beschreiben und Zusammenhänge zwischen zwei oder mehr Variablen zu untersuchen. Diese Datenanalyse basiert in der Regel auf Stichproben. Das heißt, es liegen nicht von allen Untersuchungsobjekten einer Grundgesamtheit empirische Informationen vor, sondern nur von einer (zufälligen) Auswahl. Die Inferenzstatistik beschäftigt sich mit der Frage, ob und wie Befunde von Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden können (siehe auch Abschnitt 1.2).

Abbildung 1 soll verdeutlichen, dass die Datenanalyse bzw. die Anwendung statistischer Analyseverfahren immer nur eine Phase in einem sozialwissenschaftlichen Projekt darstellt. Empirische Befunde „sprechen“ niemals für sich selbst, sondern sind immer eingebunden in eine sozialwissenschaftliche Forschungsfrage. Ohne theoretische Vorarbeiten (z.B. Entwicklung von Hypothesen) kann eine quantitative Analyse nicht zielorientiert erfolgen. Deshalb muss eine quantitative Datenanalyse immer an den (theoretischen) Forschungsprozess zurückgekoppelt werden, um in Publikationen empirisch interessante und vor allem valide Schlussfolgerungen ziehen zu können.

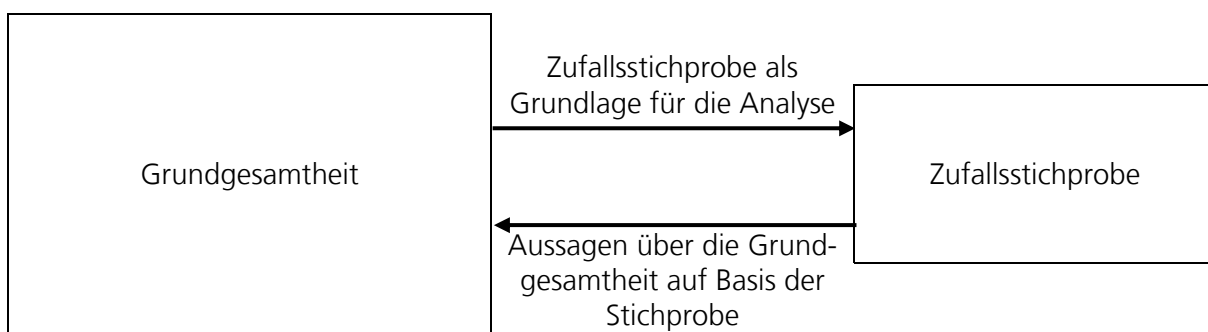
1.2 Grundgesamtheit und Stichprobe

Aus zeitlichen, finanziellen und forschungspraktischen Gründen dominieren in den Sozialwissenschaften Stichproben (Tausendpfund 2018b, S. 207-210). Bei empirischen Studien werden in der Regel nicht alle Elemente der Grundgesamtheit untersucht, sondern nur eine zufällige Auswahl

dieser Elemente. Ein Beispiel: Bei der Bundestagswahl 2017 waren nach Angaben des Bundeswahlleiters 61.688.485 Personen wahlberechtigt. Bei einer Analyse des Wahlverhaltens bei der Bundestagswahl bilden diese Personen die Grundgesamtheit. Bei einer Vollerhebung würden empirische Informationen aller Untersuchungsobjekte der Grundgesamtheit erhoben. Die Kosten der Datenerhebung und die Dauer der Erhebung sprechen allerdings gegen eine solche Vollerhebung. Für die Analyse des Wahlverhaltens (z.B. im Rahmen der German Longitudinal Election Study) wird deshalb auch keine Vollerhebung angestrebt, sondern lediglich eine Zufallsstichprobe realisiert.

In Abbildung 2 wird der Zusammenhang zwischen Grundgesamtheit und Stichprobe illustriert. Im Rahmen eines Forschungsprojekts sollen Aussagen über die Grundgesamtheit gemacht werden. In vielen Fällen ist allerdings eine Vollerhebung nicht möglich. Deshalb wird eine Zufallsstichprobe realisiert, die als Grundlage für empirische Analysen dient. Für die Berechnung einfacher Lage- und Streuungsmaße (univariate Datenanalyse), die Untersuchung von Zusammenhängen zwischen zwei Merkmalen (bivariate Datenanalyse) sowie die Schätzung von Regressionsmodellen (multivariate Datenanalyse) werden die Daten der Stichprobe genutzt.

Abbildung 2: Grundgesamtheit und Stichprobe



Quelle: Eigene Darstellung

Bei Zufallsstichproben sind allerdings Stichprobenfehler unvermeidlich. Der Mittel- oder Anteilswert einer Stichprobe wird vom wahren Mittel- oder Anteilswert der Grundgesamtheit abweichen. Ein Beispiel: In der Stichprobe wird ein mittleres Alter von 45,2 Jahren ermittelt. Dieses mittlere Alter wird vom mittleren Alter in der Grundgesamtheit abweichen. Das mittlere Alter in der Grundgesamtheit ist möglicherweise 45,1 Jahre oder 45,3 Jahre, aber vermutlich nicht 45,2 Jahre. Diese Abweichung wird als Stichprobenfehler bezeichnet.

Die Inferenzstatistik stellt uns Instrumente bereit, um zu entscheiden, ob und wie empirische Befunde aus Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden dürfen. Die Grundlagen der Inferenzstatistik und ihre Instrumente werden im Kapitel „Inferenzstatistik“ behandelt. Die grundsätzliche Frage, ob Stichprobenergebnisse auf die Grundgesamtheit übertragen werden dürfen, begegnet uns allerdings bereits bei der Darstellung der univariaten, bivariaten und multivariaten Datenanalyse. Wir werden entsprechende Fragen an den erforderlichen Stellen knapp beantworten und ggf. auf das ausführliche Kapitel am Ende dieses Kurses verweisen.

An dieser Stelle möchten wir auf zwei häufige Fehler hinweisen, die wir im Zusammenhang mit Stichproben immer wieder beobachten. Erstens: Die Anwendung der Inferenzstatistik setzt eine



Zufallsstichprobe voraus. Nur bei einer Zufallsstichprobe kann innerhalb statistischer Fehlergrenzen ein Befund auf die Grundgesamtheit übertragen werden. Zweitens: Bei einem sogenannten Signifikanztest (dabei handelt es sich um ein Instrument der Inferenzstatistik) wird geprüft, ob ein in der Stichprobe gefundener Zusammenhang (sehr) wahrscheinlich auch in der Grundgesamtheit existiert. Ein Befund wird als signifikant bezeichnet, wenn er mit großer Sicherheit von der Stichprobe auf die Grundgesamtheit übertragen werden kann. Signifikant bedeutet aber nicht, dass es sich um einen wichtigen oder starken Zusammenhang zwischen zwei Merkmalen handelt.

1.3 Klassifikationen von Variablen

Eine Variable ist ein sozialwissenschaftliches Merkmal mit mindestens zwei Ausprägungen. Das Geschlecht, der allgemeinbildende Schulabschluss oder auch das politische Interesse einer Person sind Beispiele für sozialwissenschaftliche Variablen. Sozialwissenschaftliche Merkmale bzw. Variablen können nach verschiedenen Kriterien klassifiziert werden. Wir unterscheiden vier Kriterien: Skalenniveau, diskrete und stetige Variablen, dichotome und polytome Variablen sowie manifeste und latente Variablen.

Verschiedene Skalenniveaus

Eine wichtige Voraussetzung für die Anwendung bestimmter Analyseverfahren ist das Skalenniveau der Variable bzw. des Merkmals. In den Sozialwissenschaften werden meist die Skalenarten von Stevens (1946) verwendet, der vier Skalenniveaus unterscheidet: Nominal-, Ordinal-, Intervall- und Ratioskala. Intervall- und Ratioskalen werden auch metrische Skalen genannt (Tausendpfund 2018b, S. 119-124). Das jeweilige Skalenniveau bestimmt die zulässigen Rechenoperationen. Je höher das Skalenniveau ist, desto mehr Rechenoperationen sind möglich.

Das nominale Skalenniveau ist das niedrigste Skalenniveau. Können die Ausprägungen eines Merkmals lediglich im Hinblick auf Gleichheit oder Ungleichheit verglichen werden, liegt ein nominales Skalenniveau vor (Gehring und Weins 2009, S. 43-47). Ein Beispiel für eine nominal skalierte Variable ist das Geschlecht. In vielen sozialwissenschaftlichen Datensätzen wird der Ausprägung „weiblich“ die Ziffer 1 und der Ausprägung „männlich“ die Ziffer 2 zugeordnet. Aber diese Zuordnung ist eine Konvention. Man könnte auch 1 für „männlich“ und 2 für „weiblich“ verwenden. Bei einer nominalskalierten Variable stellen die Ziffern lediglich eine Kennzeichnung dar, die nicht richtig oder falsch, sondern allenfalls mehr oder weniger sinnhaft ist. Die Möglichkeiten der quantitativen Datenanalyse bei nominalskalierten Variablen sind daher begrenzt.

Das ordinale Skalenniveau ist das nächsthöhere Skalenniveau. Bei einer ordinalskalierten Variable können die verschiedenen Ausprägungen einer Variable in eine Rangfolge gebracht werden. Beispiele für ordinalskalierte Variablen sind der Schulabschluss oder auch das politische Interesse. Die allgemeine Hochschulreife ist ein höherer Schulabschluss als die Mittlere Reife und die Mittlere Reife ist ein höherer Abschluss als ein Hauptschulabschluss. Ein „sehr starkes“ Interesse für Politik ist ein größeres Interesse als ein „mittleres“ Interesse für Politik. Bei einer ordinalskalierten Variable können zwar die einzelnen Ausprägungen in eine Rangfolge gebracht werden, aber die Abstände zwischen den Ausprägungen (z.B. Abstand zwischen „Hauptschulabschluss“ und „Mittlere Reife“ sowie zwischen „Mittlere Reife“ und „Allgemeine Hochschulreife“) sind nicht gleich. Über die

Abstände zwischen den Ausprägungen von ordinalskalierten Variablen sind daher keine Aussagen möglich.

Pseudometrische Variablen

In der Praxis werden ordinale Variablen ab etwa fünf Ausprägungen häufig als pseudometrische Variable behandelt. Neben der Mindestanzahl von fünf geordneten Ausprägungen ist allerdings entscheidend, dass angenommen wird, dass die Abstände zwischen den Ausprägungen gleich sind (Baur 2011; Faulbaum et al. 2009, S. 26; Urban und Mayerl 2018, S. 14).

Variablen sind intervallskaliert, wenn deren Ausprägungen nicht nur in eine Rangfolge gebracht werden können, sondern auch die Abstände zwischen den Ausprägungen sinnvoll interpretiert werden können. Ein Beispiel ist die Temperaturmessung in Celsius. Der Abstand zwischen 15 und 20 Grad Celsius ist genau so groß wie der Abstand zwischen 20 und 25 Grad Celsius (jeweils fünf Grad Celsius). Intervallskalen besitzen allerdings keinen natürlichen Nullpunkt. Der Nullpunkt bei der Celsius-Skala wurde lediglich unter pragmatischen Gesichtspunkten gewählt; auch Temperaturen im negativen Bereich der Celsius-Skala sind immer noch eine „Temperatur“. Bei einer Intervallskala sind die Abstände zwischen den Merkmalsausprägungen interpretierbar, aber es können keine Verhältnisse berechnet werden.

Bei einer Ratioskala (auch Verhältnisskala genannt) existiert ein natürlicher (echter) Nullpunkt. Die Temperaturmessung in Kelvin erfolgt auf einer Ratioskala, da bei 0 Kelvin keine Temperatur (keine Bewegungsenergie) mehr feststellbar ist. Auch das Einkommen und das Alter sind Beispiele für ratioskalierte bzw. verhältnisskalierte Variablen. Dabei können nicht nur die Abstände zwischen zwei Ausprägungen, sondern auch die Verhältnisse von zwei Ausprägungen interpretiert werden. Ein Einkommen von 5000 Euro ist doppelt so hoch wie ein Einkommen von 2500 Euro. Eine 60-jährige Person ist doppelt so alt wie eine 30-jährige Person.

In Tabelle 1 sind die zulässigen Rechenoperationen in Abhängigkeit vom Skalenniveau dokumentiert. Wie Tabelle 1 zeigt, steigt mit dem Skalenniveau auch die Anzahl der möglichen Rechenoperationen. Bei einem nominalskalierten Merkmal können die Ausprägungen nur ausgezählt werden, bei einem ordinalskalierten Merkmal können die Ausprägungen in eine Reihenfolge gebracht werden. Bei intervallskalierten Variablen können Differenzen, bei ratioskalierten Variablen auch Verhältnisse gebildet werden.

Tabelle 1: Zulässige Rechenoperationen in Abhängigkeit vom Skalenniveau

	Auszählen	ordnen	Differenz bilden	Quotienten bilden
Nominalskala	Ja	Nein	Nein	Nein
Ordinalskala	Ja	Ja	Nein	Nein
Intervallskala	Ja	Ja	Ja	Nein
Ratioskala	Ja	Ja	Ja	Ja

Quelle: Mittag (2017, S. 20)

! Die Kenntnis des Skalenniveaus einer Variable ist eine wichtige Voraussetzung für die Wahl eines geeigneten Analyseverfahrens. Je höher das Skalenniveau einer Variable ist, desto mehr (und leistungsfähigere) Analyseverfahren stehen der Sozialwissenschaftlerin zur Verfügung. Die Kenntnis des Skalenniveaus einer Variable ist wichtig, um bei der Datenanalyse nur die zulässigen Analyseverfahren auszuwählen. Viele statistische Verfahren sind nur zulässig, wenn die Variable mindestens intervallskaliert ist bzw. als pseudometrisch behandelt werden kann.

Diskrete und stetige Variablen

Die Einteilung als diskrete oder stetige Variable basiert auf der Anzahl der möglichen Ausprägungen. Eine diskrete Variable ist eine Variable, die nur endlich viele Ausprägungen oder höchstens „abzählbar“ unendlich viele verschiedene Ausprägungen besitzt (Diaz-Bone 2018, S. 22; Mittag 2017, S. 18). Bei einer diskreten Variable sind keine Zwischenwerte zwischen zwei aufeinander folgenden Ausprägungen möglich. Beispiele für diskrete Variablen sind der Familienstand einer Person, die Anzahl der Fachsemester oder auch die Kinderzahl einer Familie. Bei diesen Variablen sind Zwischenwerte wie 5,6 Fachsemester oder 2,3 Kinder keine möglichen Ausprägungen. Eine stetige Variable ist dadurch gekennzeichnet, dass auch Zwischenwerte möglich sind. Typische Beispiele für stetige Variablen sind Zeit- und Größenangaben, aber auch monetäre Größen wie Einkommen oder Mietpreise. In der Praxis wird bei solchen Merkmalen aber nur eine begrenzte Anzahl an Nachkommastellen erfasst, beispielsweise werden bei Größenangaben meist nur zwei Nachkommastellen angegeben. Grundsätzlich sind allerdings auch mehr Nachkommastellen möglich.

Dichotome und polytome Variablen

Eine diskrete Variable, die nur eine geringe Anzahl an Ausprägungen hat, wird als kategoriale Variable bezeichnet (Diaz-Bone 2018, S. 23). Hat eine kategoriale Variable nur zwei mögliche Ausprägungen, dann handelt es sich um eine dichotome Variable. Typische Beispiele für dichotome Variablen sind der Tabakkonsum oder auch die Wahlbeteiligung, bei denen nur die Ausprägungen „Ja“ und „Nein“ möglich sind. Eine diskrete Variable mit mehreren Ausprägungen wird als polytome Variable bezeichnet. Ein Beispiel für eine polytome Variable ist die Zugehörigkeit bzw. Nicht-Zugehörigkeit zu einer Religionsgemeinschaft mit den Ausprägungen „römisch-katholische Kirche“, „evangelische Kirche (ohne Freikirchen)“, „evangelische Freikirche“, „eine andere christliche Religionsgemeinschaft“, „eine andere, nicht-christliche Religionsgemeinschaft“ und „keine Religionsgemeinschaft“.

Manifeste und latente Variablen

Schließlich lassen sich auch manifeste und latente Variablen unterscheiden. Bei manifesten Variablen handelt es sich um Merkmale, die direkt beobachtbar sind. Eine manifeste Variable ist beispielsweise das Geschlecht oder die Haarfarbe einer Person. Dagegen handelt es sich bei latenten Variablen um Merkmale, die sich der direkten Beobachtung entziehen. Latente Variablen sind beispielsweise Intelligenz, Einstellungen wie die Zufriedenheit mit der Demokratie oder auch das soziale Vertrauen. Für eine empirische Untersuchung müssen latente Variablen erst „beobachtbar“ gemacht werden. Dieser Vorgang wird als Operationalisierung bezeichnet (Tausendpfund 2018b, S. 107-137).

2 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 2018; 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 Lagemaße (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.



2.1 Häufigkeitstabelle

Eine Häufigkeitstabelle gibt eine Übersicht über die Merkmalsausprägungen (Kategorien) einer Variable und zeigt, wie häufig jede einzelne Ausprägung vorkommt (Ludwig-Mayerhofer et al. 2014, S. 41-45; Diaz-Bone 2018, S. 35). Tabelle 2 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). Bei ihrer Antwort können die Befragten bei der ALLBUS zwischen „sehr stark“, „stark“, „mittel“, „wenig“ und „überhaupt nicht“ wählen. In Tabelle 2 sind die einzelnen Antwortkategorien und die jeweiligen Häufigkeiten dargestellt.

Tabelle 2: Interesse an Politik

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte prozentuale Häufigkeit
sehr stark	427	0,122	12,2	12,2
stark	882	0,253	25,3	37,5
mittel	1429	0,409	40,9	78,5
wenig	556	0,159	15,9	94,4
überhaupt nicht	196	0,056	5,6	100,0
Gesamt	3490	1,000	100,0	

Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).

Vier Angaben zur Häufigkeit lassen sich in Tabelle 2 unterscheiden: die absolute Häufigkeit, die relative Häufigkeit, die prozentuale Häufigkeit und die kumulierte prozentuale Häufigkeit.

Absolute Häufigkeiten

Die absolute Häufigkeit ist die Anzahl der Fälle (hier: Personen), bei der die jeweilige Kategorie auftritt. Bei der ALLBUS 2016 liegen insgesamt Angaben von 3490 Personen zum Interesse an Politik vor. 427 Personen haben die Antwortkategorie „sehr stark“ gewählt, 882 Befragte haben sich für die Antwort „stark“ entschieden, 1429 Personen interessieren sich nach eigenen Angaben „mittel“ für Politik, 556 Bürger wählten die Antwortoption „wenig“ und 196 Befragte haben die Kategorie „überhaupt nicht“ ausgewählt.

Die absoluten Häufigkeiten werden oftmals 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.

Relative Häufigkeiten

Tabelle 2 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):

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

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 (427) durch die Gesamtzahl (3490) dividiert:

$$p_{\text{sehr stark}} = \frac{427}{3490} \approx 0,122$$

Die Summe der relativen Häufigkeiten aller Kategorien ergibt immer 1.

Prozentuale Häufigkeiten

Durch die Multiplikation der relativen Häufigkeit mit 100% werden die prozentualen Häufigkeiten der einzelnen Kategorien berechnet. Die Prozentsätze weisen die gleichen Informationen wie die relativen Häufigkeiten aus, es handelt sich nur um verschiedene Darstellungsformen (Diaz-Bone 2018, S. 36). Die als Prozentsatz dargestellte relative Häufigkeit wird als prozentuale Häufigkeit bezeichnet. Die Prozentsätze lassen sich mit folgender Formel berechnen:

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

Für die Berechnung der prozentualen Häufigkeit der Kategorie „wenig“ wird die absolute Häufigkeit (556) durch die Gesamtzahl (3490) dividiert. Das gerundete Ergebnis (0,159) wird anschließend mit 100% multipliziert.

$$p_j \% = \frac{556}{3490} * 100\% \approx 15,9 \text{ Prozent}$$

15,9 Prozent der Befragten geben an, sich wenig für Politik zu interessieren.

In der letzten Spalte in Tabelle 2 werden die kumulierten Prozentsätze ausgewiesen. Dabei handelt es sich um eine schrittweise Addition (Kumulation) der Prozentsätze. Die kumulierten Prozente der Kategorie „stark“ (37,5 Prozent) kommen durch die Addition der Prozentsätze in den Kategorien „sehr stark“ (12,2 Prozent) und „stark“ (25,3 Prozent). Alternativ können auch die absoluten Häufigkeiten der beiden Kategorien addiert ($427+882=1309$) und durch die Gesamtzahl (3490) dividiert werden. Das Ergebnis (0,375) entspricht der kumulierten relativen Häufigkeit. Durch die Multiplikation mit 100 werden die kumulierte prozentuale Häufigkeit bzw. die kumulierten Prozente berechnet (37,5 Prozent).

Kumulierte prozentuale Häufigkeiten

Bei Merkmalen, die mindestens ordinal skaliert sind, bieten die kumulierten Prozentsätze eine anschauliche Interpretationsmöglichkeit. Der Wert von 37,5 Prozent in Tabelle 2 kann wie folgt interpretiert werden: 37,5 Prozent der Befragten haben mindestens ein starkes Interesse an Politik.

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).

Häufigkeitstabelle mit Missing Values

Das Vorgehen wird anhand 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). Im 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“ und „weiß nicht“ antworten oder die Aussage verweigern. Tabelle 3 informiert über die Häufigkeitsangaben dieser Variable im ALLBUS 2016. In der Spalte „Häufigkeit“ werden die absoluten Häufigkeiten der Angaben dokumentiert. Insgesamt liegen Angaben von 3490 Befragten vor. Davon haben 3446 Personen gültige Angaben gemacht, 44 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 Prozentsätze der Mittelschicht ergibt sich wie folgt:

$$p_j\% = \frac{2002}{3490} * 100\% = 57,4 \text{ Prozent}$$

Für die Berechnung der gültigen Prozente wird nicht durch die Gesamtzahl aller Fälle geteilt, sondern durch die Anzahl der gültigen Fälle (3446):

$$p_j\% = \frac{2002}{3446} * 100\% = 58,1 \text{ Prozent}$$

Bei den kumulierten Prozentsätzen in Tabelle 3 handelt es sich um die Addition der (gültigen) Prozentsätze der einzelnen Kategorien. Die kumulierten Prozentsätze der Mittelschicht (87,5 Prozent) ergeben sich durch die Addition der gültigen Prozentsätze der Unterschicht (2,9 Prozent), der Arbeiterschicht (26,5 Prozent) und der Mittelschicht (58,1 Prozent). Insgesamt 87,5 Prozent der Befragten ordnen sich der Unter-, Arbeiter- oder Mittelschicht zu.

Tabelle 3: Subjektive Schichteinstufung

Kategorie	absolute Häufigkeit	in Prozent	gültige Prozent	kumulierte Prozent
Unterschicht	101	2,9	2,9	2,9
Arbeiterschicht	912	26,1	26,5	29,4
Mittelschicht	2002	57,4	58,1	87,5
Obere Mittelschicht	417	11,9	12,1	99,6
Oberschicht	14	0,4	0,4	100,0
Gesamt	3446	98,7	100,0	
keiner der Schichten	18	0,5		
Einstufung abgelehnt	1	0,0		
weiß nicht	14	0,4		
keine Angabe	11	0,3		
Gesamt	44	1,3		
Gesamt	3490	100,0		

Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).



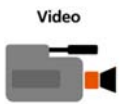
Die Prozentangaben einer Häufigkeitstabelle können nur dann angemessen interpretiert werden, wenn erstens die absolute Häufigkeit eines Merkmals und zweitens die Art der Prozentuierung bekannt sind (Gehring und Weins 2009, S. 102-104).

Die Angabe von Prozentsätzen 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 Prozentsätze“ ausgeschlossen werden.

2.2 Lagemaße

In der quantitativen Sozialforschung sind Forscherinnen in der Regel mit großen Fallzahlen konfrontiert. Bei der univariaten Datenanalyse wird diese Informationsmenge meist auf wenige Kenngrößen verdichtet. Dabei beschreiben Lagemaße das Zentrum bzw. typische Werte einer Verteilung. Zu den wichtigsten Lagemaßen in den Sozialwissenschaften zählen Modus, Median und arithmetisches Mittel (Völkl und Korb 2018, S. 56; Diaz-Bone 2018, S. 44-47).



Die Zulässigkeit der Berechnung eines Lagemaßes ist vom Skalenniveau der Variable abhängig. Je höher das Skalenniveau ist, desto mehr Lagemaße lassen sich berechnen. Wie Tabelle 4 zeigt, setzt die Berechnung des arithmetischen Mittels mindestens eine intervallskalierte Variable (z. B. Alter) voraus. Für die Bestimmung des Medians muss die Variable mindestens ordinalskaliert sein (z. B. politisches Interesse). Der Modus kann bereits für nominalskalierte Variablen angegeben werden (z. B. Geschlecht).

Tabelle 4: Lagemaße und Skalenniveau

	Nominalskala	Ordinalskala	ab Intervallskala
Modus	Ja	Ja	Ja
Median	Nein	Ja	Ja
Arithmetisches Mittel	Nein	Nein	Ja

Quelle: Eigene Darstellung

2.2.1 Modus

Der Modus, auch Modalwert genannt, ist der am häufigsten vorkommende (gültige) Wert in einer Verteilung. Der Modus wird mit dem Buchstaben h (Diaz-Bone 2018, S. 47) oder mit \hat{x} (x mit Punkt, siehe Völkl und Korb 2018) abgekürzt. Der Modus muss nicht berechnet werden, sondern kann aus einer Tabelle einfach abgelesen werden. In Tabelle 5 wird die „Berechnung“ des Modus an drei Beispielen illustriert. Es liegen jeweils Angaben von neun Befragten (x_1 bis x_9) zu drei Variablen vor.

Tabelle 5: Berechnung des Modus

Beispiel	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	\hat{x}
Geschlecht	1	1	1	2	2	2	2	1	1	1
Alter	20	45	45	30	30	25	25	40	25	25
Familienstand	1	1	2	2	1	3	3	3	4	1 und 3



Quelle: Eigene Darstellung

In der ersten Zeile ist das Geschlecht der Befragten mit den Werten 1 und 2 kodiert. Dabei ist Mann mit 1 und Frau mit 2 kodiert. Insgesamt liegen Angaben von fünf Männern und vier Frauen

vor. Der am häufigsten vorkommende Wert ist 1. Deshalb ist der Modus dieser Variable der Wert 1, da mehr Männer als Frauen befragt wurden.

In der zweiten Zeile wurde das Alter der neun Personen erfasst. Ein Befragter ist 20 Jahre, drei Befragte sind 25 Jahre, zwei Personen sind 30 Jahre, zwei Befragte sind 45 Jahre und ein Befragter ist 45 Jahre alt. Der häufigste Wert ist 25. Der Modus des Alters ist 25.

In der dritten Zeile wird der Familienstand der Personen erfasst. Dabei wird zwischen verheiratet (1), geschieden (2), ledig (3) und andere (4) unterschieden. Am häufigsten werden verheiratet (1) und ledig (3) genannt. Der Modus ist 1 und 3. Wenn zwei Ausprägungen einer Variable gleich häufig vorkommen, dann spricht man auch von einer bimodalen Verteilung.

In einer Häufigkeitstabelle kann der Modus direkt abgelesen werden. Tabelle 6 zeigt die Häufigkeiten der Variable „Geschlecht“ in der ALLBUS 2016. Offensichtlich wurden in der ALLBUS 2016 mehr Männer als Frauen befragt. Der Modus ist deshalb 1 (Mann).

Tabelle 6: Geschlecht

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte Prozent
Mann (1)	1770	0,507	50,7	50,7
Frau (2)	1720	0,493	49,3	100,0
Gesamt	3490	1,000	100,0	

Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).

Der Modus ist ein einfaches Maß der univariaten Datenanalyse. Der Informationsgehalt ist allerdings eher gering, da er sich nur auf einen einzigen Wert der Verteilung bezieht (Völkl und Korb 2018, S. 72).

2.2.2 Median

Der Median, auch Zentralwert genannt, ist der „mittlere“ Wert einer geordneten Verteilung. Er wird meist mit \tilde{x} (lies: x-Schlange oder x-Tilde) abgekürzt und setzt eine mindestens ordinalskalierte Variable voraus. Der Median teilt eine Verteilung in der Mitte, so dass 50 Prozent der Werte kleiner oder gleich dem Median und 50 Prozent der Werte größer oder gleich dem Median sind. Neben dem ordinalen Skalenniveau ist eine Voraussetzung für die Berechnung des Medians, dass die Werte bereits in aufsteigender Reihenfolge sortiert sind (Ludwig-Mayerhofer et al. 2014, S. 62-65; Völkl und Korb 2018, S. 72-77; Diaz-Bone 2018, S. 45-46).



In Tabelle 7 sind die Angaben von elf Befragten zum politischen Interesse dargestellt. Dabei deuten höhere Zahlen auf ein größeres politisches Interesse hin. In der oberen Zeile von Tabelle 7 sind die Angaben der Befragten noch unsortiert. Vor der Berechnung des Medians müssen die Angaben der Reihe nach sortiert werden. Der Median ist dann der Wert, der in der Mitte der Verteilung liegt. Bei elf Befragten ist der Wert an der sechsten Position der Median. In unserem Beispiel ist der Median 3.

Tabelle 7: Berechnung des Medians (ungerade Fallzahl)

Position	1	2	3	4	5	6	7	8	9	10	11
unsortiert	1	5	3	2	3	1	5	3	3	4	2
sortiert	1	1	2	2	3	3	3	3	4	5	5

Quelle: Eigene Darstellung

Bei einer ungeraden Fallzahl ist der mittlere Wert eines geordneten Datensatzes eindeutig bestimmt. Bei einer geraden Fallzahl gibt es allerdings zwei Werte, die die Mitte des Datensatzes repräsentieren. Tabelle 8 enthält die Angaben von zehn Befragten. Der Median findet sich zwischen der fünften und sechsten Position. In unserem Beispiel ist der Median auch bei einer ordinalen Variable eindeutig bestimmbar, da sich die Werte an den beiden Positionen nicht unterscheiden ($\tilde{x} = 3$). Bei einer intervallskalierten Variable kann der Median auch bestimmt werden, wenn sich die Werte unterscheiden. In einem solchen Fall wird der Mittelwert der beiden zentralen Werte berechnet (Diaz-Bone 2018, S. 46; Völkl und Korb 2018, S. 75).

Tabelle 8: Berechnung des Medians (gerade Fallzahl)

Position	1	2	3	4	5	6	7	8	9	10
unsortiert	1	5	3	2	3	1	5	3	3	4
sortiert	1	1	2	2	3	3	3	3	4	5

Quelle: Eigene Darstellung

Bei einer ungeraden Fallzahl wird der Median wie folgt bestimmt:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Bei einer geraden Fallzahl wird der Median wie folgt bestimmt:

$$\tilde{x} = \frac{1}{2} * (x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)})$$

In beiden Fällen muss der Datensatz zunächst sortiert werden. Bei einer ungeraden Fallzahl ist der Median der Wert, der in der Mitte des Datensatzes liegt. Bei einer geraden Fallzahl ergibt sich der Median aus den beiden zentralen Werten eines Datensatzes.

In der Forschungspraxis kann der Median meist auf Basis einer Häufigkeitstabelle abgelesen werden, die im Idealfall bereits die kumulierten Prozentwerte ausweist. Ansonsten müssen die kumulierten Prozentwerte zunächst berechnet werden (Völkl und Korb 2018, S. 73). In Tabelle 9 ist erneut die Häufigkeitsverteilung des Interesses an Politik auf Basis der ALLBUS 2016 dokumentiert. Die Spalte „Kategorie“ enthält neben den Antwortvorgaben jetzt auch die numerische Kodierung. Die Angabe „sehr stark“ entspricht dem Wert 1, die Antwort „stark“ dem Wert 2 usw. Insgesamt gibt es (gültige) Angaben von 3490 Personen. Da eine gerade Fallzahl vorliegt, gibt es keinen Wert, der genau in der Mitte liegt. Der Median liegt also zwischen den Positionen 1745 und 1746.



Tabelle 9: Interesse an Politik

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte Prozent
sehr stark (1)	427	0,122	12,2	12,2
stark (2)	882	0,253	25,3	37,5
mittel (3)	1429	0,409	40,9	78,5
wenig (4)	556	0,159	15,9	94,4
überhaupt nicht (5)	196	0,056	5,6	100,0
Gesamt	3490	1,0	100,0	

Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).

Für die Berechnung des Medians können Sie jetzt die Werte der Befragten ermitteln, die an Position 1745 und 1746 liegen. Alternativ – und einfacher – kann der Median auch anhand der Spalte „kumulierte Prozent“ ermittelt werden. Der Median ist der Wert, bei dem die kumulierten Prozente 50 Prozent erreichen oder übersteigen (Ludwig-Mayerhofer et al. 2014, S. 63; Völkl und Korb 2018, S. 73). Der Median des Interesses an Politik ist 3.

2.2.3 Arithmetisches Mittel



Das bekannteste Lagemaß einer Verteilung ist das arithmetische Mittel, das mit \bar{x} (lies: x quer) abgekürzt wird. Das arithmetische Mittel begegnet uns auch im Alltag, umgangssprachlich wird es als Mittelwert oder Durchschnittswert bezeichnet (Völkl und Korb 2018, S. 77-82; Diaz-Bone 2018, S. 45; Benninghaus 2007, S. 45-49). Das arithmetische Mittel kann allerdings nur für Daten berechnet werden, die mindestens intervallskaliert sind oder als pseudometrisch behandelt werden können. Im Alltag wird das arithmetische Mittel häufig auch bei ordinalen Daten (z.B. Schulnoten) berechnet. Dies ist formal allerdings nicht korrekt, da bei ordinalen Daten die Abstände zwischen den Werten nicht gleich sind.

Das arithmetische Mittel wird berechnet, indem zunächst alle (gültigen) Werte addiert und die Summe anschließend durch die Anzahl der (gültigen) Fälle dividiert wird. Formal ist das arithmetische Mittel definiert als die Summe der Messwerte geteilt durch ihre Anzahl (Benninghaus 2007, S. 45):

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$



In Tabelle 10 sind drei Beispiele für die Berechnung des arithmetischen Mittels bei kleinen Fallzahlen aufgeführt. Für die Berechnung müssen die Angaben der (fiktiven) Personen addiert und durch die Fallzahl (9) dividiert werden. Das arithmetische Mittel für das Alter der neun Befragten in Tabelle 10 wird wie folgt berechnet:

$$\bar{x} = \frac{38 + 22 + 17 + 52 + 37 + 58 + 31 + 45 + 60}{9} = 40$$

Tabelle 10: Berechnung des arithmetischen Mittels bei kleinen Fallzahlen

Beispiel	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	\bar{x}
Alter	38	22	17	52	37	58	31	45	60	40
Einkommen	2000	1500	1700	2500	1000	1700	2600	3200	1800	2000
Berufsjahre	10	2	2	25	15	30	5	15	40	16

Quelle: Eigene Darstellung

In der Forschungspraxis sind wir in der Regel allerdings mit deutlich größeren Fallzahlen konfrontiert. Die Addition aller Werte ist dann eine aufwändige Angelegenheit. Bei großen Fallzahlen wird daher eine Formel verwendet, bei der die Messwerte nicht einzeln aufsummiert, sondern in Gruppen berechnet werden. Statt „10 + 10 + 10“ wird dann „3 * 10“ gerechnet.

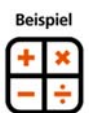
$$\bar{x} = \frac{n_1 * x_1 + n_2 * x_2 + n_3 * x_3 + \dots + n_j * x_j}{n} = \frac{\sum_{j=1}^j n_j * x_j}{n}$$

Als Beispiel dient die Berechnung des arithmetischen Mittels der Lebenszufriedenheit auf Basis der ALLBUS 2016. Im ALLBUS wird die Lebenszufriedenheit mit folgender Frage erfasst:

Und jetzt noch eine allgemeine Frage. Wie zufrieden sind Sie gegenwärtig – alles in allem – mit Ihrem Leben?

Als Antwort können die Personen eine Zahl von 0 bis 10 wählen, wobei 0 „ganz und gar unzufrieden“ und 10 „ganz und gar zufrieden“ bedeutet. Mit den Zahlen dazwischen kann die Antwort abgestuft werden. Die Lebenszufriedenheit wird als pseudometrische Variable behandelt, da sie mehr als fünf Ausprägungen hat und die Abstände zwischen den Ausprägungen gleich sind.

Tabelle 11 zeigt die Häufigkeitstabelle der Lebenszufriedenheit auf Grundlage der ALLBUS 2016. Für die Berechnung des arithmetischen Mittelwerts wird die Häufigkeit der einzelnen Gruppen mit dem jeweiligen Wert multipliziert. Die Ergebnisse der einzelnen Gruppen werden anschließend addiert und durch die Gesamtfallzahl dividiert.



$$\bar{x} = \frac{15 * 0 + 10 * 1 + 15 * 2 + 72 * 3 + 78 * 4 + 268 * 5 + 220 * 6 + 478 * 7 + 1095 * 8 + 758 * 9 + 479 * 10}{3488}$$

$$\bar{x} = \frac{0 + 10 + 30 + 216 + 312 + 1340 + 1320 + 3346 + 8760 + 6822 + 4790}{3488}$$

$$\bar{x} = \frac{26946}{3488} = 7,73$$

Für die Berechnung des arithmetischen Mittels bei großen Fallzahlen bietet sich eine Erweiterung der Häufigkeitstabelle an. Tabelle 11 wurde für diesen Zweck um eine (grau markierte) Spalte erweitert, in der die Multiplikation der einzelnen Gruppen ($n_j * x_j$) dokumentiert wird. Diese Vorgehensweise ist übersichtlicher und deutlich weniger fehleranfällig. Das (arithmetische) Mittel der Lebenszufriedenheit liegt bei 7,7. Der Modus und der Median der Lebenszufriedenheit sind jeweils 8.

Tabelle 11: Berechnung des arithmetischen Mittels bei großen Fallzahlen

Kategorie	absolute Häufigkeit	in Prozent	kumulierte Prozent	$n_j * x_j$
0 (ganz unzufrieden)	15	0,4	0,4	$15 * 0 = 0$
1	10	0,3	0,7	$10 * 1 = 10$
2	15	0,4	1,1	$15 * 2 = 30$
3	72	2,1	3,2	$72 * 3 = 216$
4	78	2,2	5,4	$78 * 4 = 312$
5	268	7,7	13,1	$268 * 5 = 1340$
6	220	6,3	19,4	$220 * 6 = 1320$
7	478	13,7	33,1	$478 * 7 = 3346$
8	1095	31,4	64,5	$1095 * 8 = 8760$
9	758	21,7	86,3	$758 * 9 = 6822$
10 (ganz zufrieden)	479	13,7	100,0	$479 * 10 = 4790$
Gesamt	3488	100,0		26946

Daten: ALLBUS 2016. Eigene Berechnungen

Der Vorteil des arithmetischen Mittels ist, dass bei der Berechnung alle verfügbaren Informationen ausgeschöpft werden. Allerdings ist das arithmetische Mittel – insbesondere bei kleinen Fallzahlen – auch besonders sensibel für Extremwerte. Als Illustration dienen die Altersangaben von neun (fiktiven) Personen in zwei Gruppen (siehe Tabelle 12). Das Alter von jeweils acht Personen in den beiden Gruppen ist identisch, lediglich das Alter der ältesten Person unterscheidet sich. In Gruppe 1 ist der älteste Befragte 35 Jahre alt, in Gruppe 2 ist der älteste Befragte 70 Jahre alt. Das arithmetische Mittel des Alters der Gruppe 1 liegt bei 26,67 Jahren und bei Gruppe 2 bei 30,56 Jahren. Eine Person bzw. das Alter einer Person hat also einen deutlichen Effekt auf das Durchschnittsalter der zweiten Gruppe.

Tabelle 12: Mittelwerte und Ausreißer

Beispiel (Alter)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	\bar{x}
Gruppe 1	20	25	25	25	25	25	30	30	35	26,67
Gruppe 2	20	25	25	25	25	25	30	30	70	30,56

Quelle: Eigene Darstellung



Während das arithmetische Mittel durch Extremwerte beeinflusst wird, gelten Modus und Median als relativ robust gegenüber Ausreißern bzw. Extremwerten. Bei beiden Gruppen in Tabelle 12 liegen Modus und Median bei 25. Ein Wert bzw. der Ausreißer hat keinen Einfluss auf Modus und Median.

2.3 Streuungsmaße

Lagemaße informieren über das Zentrum bzw. typische Werte einer Verteilung, sie geben keinen Hinweis auf die Streuung der Variablenwerte (Benninghaus 2007, S. 51-62; Diaz-Bone 2018, S. 48-56; Völkl und Korb 2018, S. 87-116). In Tabelle 13 wird die (fiktive) Lebenszufriedenheit von insgesamt 22 Personen dokumentiert, die jeweils auf einer Skala von 0 „ganz unzufrieden“ und 10 „ganz zufrieden“ angegeben wurde. 11 dieser Befragten leben in einer Stadt und 11 der Personen leben auf dem Land. Bevor Sie weiterlesen: Berechnen Sie bitte einmal Modus, Median und arithmetisches Mittel der Lebenszufriedenheit für die beiden Gruppen.

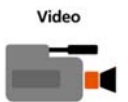


Tabelle 13: Lebenszufriedenheit von zwei Gruppen

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁
Stadt	6	7	7	7	7	7	7	7	7	7	8
Land	1	2	7	7	7	7	8	8	10	10	10



Quelle: Eigene Darstellung

Haben Sie die Lagemaße für beide Gruppen berechnet? In beiden Gruppen sind Modus, Median und arithmetisches Mittel jeweils 7. Die Verteilung der Werte in der Stadt-Gruppe ist zwar auf den ersten Blick homogener (gleichartiger), aber dies wird durch die Lagemaße nicht dokumentiert. Mit anderen Worten: Trotz gleicher Lagemaße können Verteilungen von Merkmalen (hier: Lebenszufriedenheit) unterschiedlich sein. Bei einer überschaubaren Anzahl an Untersuchungsobjekten ist diese Variation unmittelbar ersichtlich, bei größeren Datensätzen ist dies allerdings nicht mehr erkennbar.

Deshalb sind für die Beschreibung einer Verteilung neben Lagemaßen auch Streuungsmaße erforderlich, die die Variation einer Verteilung abbilden. Wichtige Streuungsmaße in den Sozialwissenschaften sind Varianz und Standardabweichung, die ab intervallskalierten Variablen berechnet werden können.

2.3.1 Varianz

Die Varianz ist definiert als die durchschnittliche quadrierte Abweichung vom arithmetischen Mittel (Völkl und Korb 2018, S. 94). Eine Varianz von 0 zeigt an, dass überhaupt keine Streuung vorliegt. Je größer die Varianz ist, desto stärker streuen die einzelnen Werte einer Verteilung um das arithmetische Mittel (Schendera 2015, S. 130).

Für die Berechnung der Varianz wird für jeden Wert die Abweichung vom arithmetischen Mittel bestimmt. Eine Abweichung ist positiv, wenn der Wert über dem arithmetischen Mittel liegt. Falls der Wert kleiner als das arithmetische Mittel ist, ist die Abweichung negativ. Die Summe aller Abweichungen vom Mittelwert ist per Definition 0. Damit sich bei der Addition der einzelnen Abweichungen positive und negative Abweichungen nicht gegenseitig aufheben, werden diese quadriert. Die quadrierten Abweichungen aller Werte werden anschließend addiert.

! Abschließend muss die Summe der quadrierten Abweichungen noch durch die Fallzahl dividiert werden. Dabei ist zu berücksichtigen, ob Daten einer Vollerhebung (z.B. alle Personen der Grundgesamtheit) oder einer Stichprobe (z.B. zufällige Auswahl von Personen der Grundgesamtheit) vorliegen. Wir müssen deshalb zwischen der empirischen und der korrigierten Varianz unterscheiden (Mittag 2017, S. 69-70; siehe auch Weins 2010, S. 70; Ludwig-Mayerhofer et al. 2014, S. 74-76).

Liegen Daten einer Vollerhebung vor, dann wird die sogenannte empirische Varianz (s^2) berechnet. In diesem Fall wird die Summe der quadrierten Abweichungen einfach durch die Fallzahl dividiert. Formal ausgedrückt:

Empirische Varianz:
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Bei Stichproben wird die sogenannte korrigierte Varianz (s^{*2}) berechnet. Dabei wird die Fallzahl vor der Division um 1 verringert. Diese Modifikation bietet bei Stichprobendaten bestimmte Vorteile, die bei der Inferenzstatistik noch dargestellt werden.² Statistikprogramme wie SPSS oder Stata berechnen die korrigierte Varianz.

Korrigierte Varianz:
$$s^{*2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Die Berechnung der empirischen und der korrigierten Varianz unterscheidet sich also nur dadurch, dass bei der korrigierten Varianz nicht durch n , sondern durch $n - 1$ dividiert wird. Bei kleiner Fallzahl (z.B. $n = 10$) sind Unterschiede bzw. unterschiedliche Ergebnisse sichtbar (Weins 2010, S. 70), bei einer großen Fallzahl sind die Unterschiede zwischen der empirischen und der korrigierten Varianz praktisch vernachlässigbar.

! Für die händische Berechnung der Varianz wird empfohlen, eine Arbeitstabelle anzulegen und die Berechnung schrittweise durchzuführen. Dabei sind folgende Schritte zu beachten:

1. Das arithmetische Mittel (\bar{x}) wird berechnet.
2. Für jeden Merkmalswert wird die Abweichung vom arithmetischen Mittel berechnet: $x_i - \bar{x}$
3. Diese Abweichung wird quadriert: $(x_i - \bar{x})^2$
4. Die quadrierte Abweichung wird für alle Fälle aufsummiert: $\sum_{i=1}^n (x_i - \bar{x})^2$

² Für die interessierte Leserin: Wird für die Berechnung der Varianz bei einer Zufallsstichprobe anstelle des Terms $n-1$ durch die Fallzahl (n) dividiert, dann wird die Varianz in der Grundgesamtheit unterschätzt. Das Ausmaß der Verzerrung ist von der Fallzahl abhängig. Je größer die Fallzahl, desto geringer die Verzerrung (Kühnel und Krebs 2007, S. 235-237).

5. Für die empirische Varianz wird die Summe der quadrierten Abweichungen durch die Fallzahl (n) dividiert. Für die korrigierte Varianz wird die Summe der quadrierten Abweichungen durch die Fallzahl minus 1 dividiert.

Zur Illustration wird die empirische Varianz der Lebenszufriedenheit der elf (fiktiven) Befragten auf dem Land berechnet. In Tabelle 14 sind die Angaben der Befragten in der Spalte „Wert“ eingetragen. Nach der Berechnung des arithmetischen Mittels (7) wird für jeden Befragten die Abweichung berechnet. Die Abweichung wird quadriert, da sich sonst negative und positive Abweichungen aufheben würden. Die quadrierten Abweichungen werden schließlich aufsummiert (90).



Tabelle 14: Arbeitstabelle für die Berechnung der Varianz (kleine Fallzahl)

ID	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$\sum_{i=1}^n (x_i - \bar{x})^2$
1	1	$(1 - 7)$	-6^2	36
2	2	$(2 - 7)$	-5^2	25
3	7	$(7 - 7)$	0^2	0
4	7	$(7 - 7)$	0^2	0
5	7	$(7 - 7)$	0^2	0
6	7	$(7 - 7)$	0^2	0
7	8	$(8 - 7)$	1^2	1
8	8	$(8 - 7)$	1^2	1
9	10	$(10 - 7)$	3^2	9
10	10	$(10 - 7)$	3^2	9
11	10	$(10 - 7)$	3^2	9
Gesamt	77			90

Quelle: Eigene Darstellung

Für die Berechnung der empirischen Varianz wird die Summe der quadrierten Abweichungen durch die Fallzahl (11) dividiert. Formal:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{90}{11} \approx 8,18$$

Die empirische Varianz der Lebenszufriedenheit der elf Befragten liegt bei 8,18. Das Ergebnis lässt sich als durchschnittliche quadrierte Abweichung vom arithmetischen Mittel interpretieren (Völkl und Korb 2018, S. 94). Diese Interpretation ist abstrakt und nicht sonderlich zugänglich. Die Varianz ist allerdings ein erforderlicher Zwischenschritt für die Berechnung der Standardabweichung, die dieselbe Maßeinheit wie die Ursprungsvariable aufweist (hier: Punkte). Für die Berechnung der Standardabweichung muss lediglich die Wurzel aus der Varianz gezogen werden.



Zuvor wird die Berechnung der (korrigierten) Varianz bei einer großen Fallzahl illustriert. Im Unterschied zu der Berechnung bei einer kleinen Fallzahl (Tabelle 14) müssen bei einer großen Fallzahl die quadrierten Abweichungen mit der Häufigkeit der Merkmalsausprägung multipliziert und anschließend aufsummiert werden. In der ersten Spalte in Tabelle 15 sind die Antwortkategorien dargestellt, in der zweiten Spalte ist die absolute Häufigkeit abgebildet, mit der die jeweilige Antwortkategorie gewählt wurde (ein Beispiel: 1090 Befragte haben die Antwortkategorie 8 gewählt). In der dritten Spalte $(x_i - \bar{x})$ wird die Differenz zwischen dem dieser Kategorie zugeordneten Wert und dem arithmetischen Mittel gebildet. Diese Differenz wird in der vierten Spalte quadriert. In der fünften Spalte wird die quadrierte Differenz mit der entsprechenden absoluten Häufigkeit (n_j) multipliziert. Die Produkte für die einzelnen Antwortkategorien werden addiert.

Tabelle 15: Arbeitstabelle für die Berechnung der Varianz (große Fallzahl)

Kategorie	absolute Häufigkeit	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_j * (x_i - \bar{x})^2$
0	15	$(0 - 7,7)$	59,29	889,35
1	10	$(1 - 7,7)$	44,89	448,90
2	15	$(2 - 7,7)$	32,49	487,35
3	72	$(3 - 7,7)$	22,09	1590,48
4	78	$(4 - 7,7)$	13,69	1067,82
5	268	$(5 - 7,7)$	7,29	1953,72
6	220	$(6 - 7,7)$	2,89	635,80
7	478	$(7 - 7,7)$	0,49	234,22
8	1095	$(8 - 7,7)$	0,09	98,55
9	758	$(9 - 7,7)$	1,69	1281,02
10	479	$(10 - 7,7)$	5,29	2533,91
Gesamt	3488			11221,12

Quelle: Eigene Darstellung

Für die Berechnung der korrigierten Varianz wird die Summe der quadrierten Abweichungen (11221,12) durch $n - 1$ dividiert:

$$s^{*2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{11221,12}{3488 - 1} = \frac{11221,12}{3487} = 3,22$$

Die korrigierte Varianz der Lebenszufriedenheit liegt bei 3,22. Das Ergebnis lässt sich analog zu Tabelle 14 als die durchschnittliche quadrierte Abweichung vom arithmetischen Mittel interpretieren. Diese Interpretation ist weiterhin nicht sonderlich zugänglich. Durch die Quadrierung der Abweichungen vom arithmetischen Mittel hat die Varianz zudem eine andere Einheit als das zugrundeliegende Merkmal (hier: subjektive Lebenszufriedenheit, die auf einer Skala von 0 bis 10 erfasst wurde). Bei der Beschreibung empirischer Verteilungen spielt die Varianz deshalb nur eine untergeordnete Rolle. Sie ist ein (notwendiger) Zwischenschritt auf dem Weg zu einem einfacheren und zugänglicheren Streuungsmaß: der Standardabweichung.

2.3.2 Standardabweichung

Die Standardabweichung ist die Wurzel aus der Varianz. Die Standardabweichung kann als durchschnittliche Variabilität der Werte um das arithmetische Mittel interpretiert werden (Schendera 2015, S. 131). Etwas zugänglicher formulieren es Völkl und Korb (2018, S. 94-95), die die Standardabweichung als durchschnittliche Abweichung vom arithmetischen Mittel definieren. Ein kleiner Wert deutet auf eine geringe Streuung der Werte in der Verteilung hin, ein großer Wert auf eine große Streuung der Werte in der Verteilung. Beträgt die Standardabweichung 0, dann liegt überhaupt keine Streuung vor.

Für die Berechnung der empirischen Standardabweichung wird die Wurzel aus der empirischen Varianz gezogen. Übertragen auf das Beispiel in Tabelle 14 wird die Wurzel aus 8,18 gezogen. Die empirische Standardabweichung liegt bei 2,86 Punkten. Die durchschnittliche Abweichung der Lebenszufriedenheit der Bürger, die auf dem Land leben, liegt somit bei 2,86 Punkten.



Empirische
Standardabweichung:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{90}{11}} = \sqrt{8,18} = 2,86$$

Für die Berechnung der korrigierten Standardabweichung wird die Wurzel aus der korrigierten Varianz gezogen. Übertragen auf das Beispiel in Tabelle 15 wird die Wurzel aus 3,22 gezogen. Die korrigierte Standardabweichung der Lebenszufriedenheit liegt bei 1,79 Punkten.

Korrigierte
Standardabweichung:

$$s^* = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{11.221,12}{3488-1}} = \sqrt{\frac{11.221,12}{3487}} = \sqrt{3,22} = 1,79$$

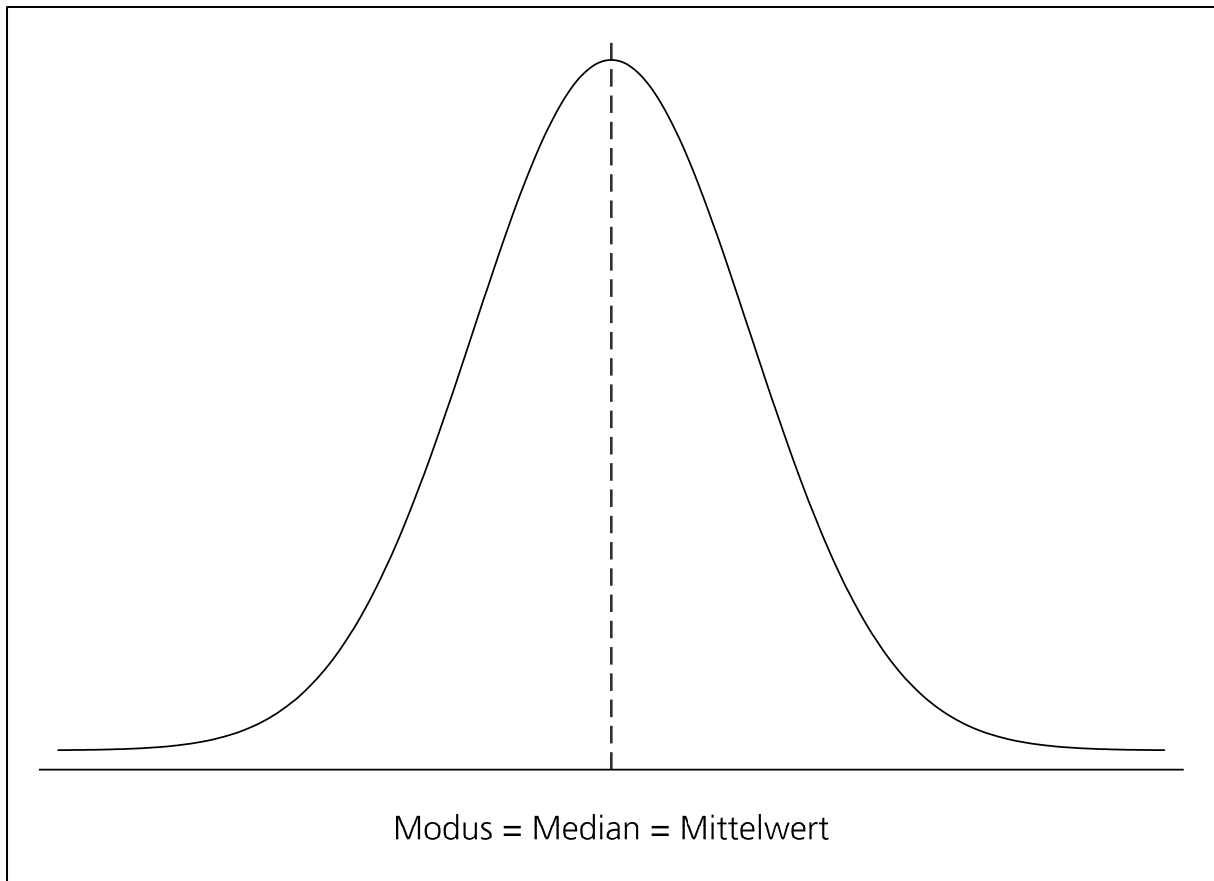
2.4 Formmaße

Mit Schiefe (engl. skewness) und Wölbung (engl. kurtosis) wird die Form einer Verteilung beschrieben (Kühnel und Krebs 2007, S. 101-103; Ludwig-Mayerhofer et al. 2014, S. 82-84; Schendera 2015, S. 133-135; Völkl und Korb 2018). Beide Maße beschreiben dabei die Abweichung einer Verteilung von der Normalverteilung (siehe Abbildung 3).

Bei der Normalverteilung handelt es sich um eine wichtige Verteilung der Inferenzstatistik, die nach dem Mathematiker Carl Friedrich Gauß häufig auch als Gauß-Verteilung bezeichnet wird. Die Normalverteilung lässt sich wie folgt charakterisieren: Die Verteilung ist symmetrisch, das heißt in der Mitte befinden sich die meisten Werte. Die Häufigkeiten der Werte nehmen links und rechts gleichermaßen – also wieder symmetrisch – vom arithmetischen Mittel ab (Ludwig-Mayerhofer et al. 2014, S. 77-80; Völkl und Korb 2018, S. 117-120). Durch die symmetrische Form der Normalverteilung sind Modus, Median und das arithmetische Mittel identisch.

! Die Kenntnis der Schiefe und Wölbung einer Verteilung ist wichtig, da bestimmte statistische Verfahren eine normalverteilte Variable voraussetzen. Bei den meisten Verfahren sind geringe Abweichungen zwar unproblematisch, aber bei starken Abweichungen müssen die Variablen vor der Analyse transformiert werden.

Abbildung 3: Normalverteilung



Quelle: Eigene Darstellung

Empirische Verteilungen können horizontal und/oder vertikal von der Normalverteilung abweichen. Die Schiefe ist ein Maß für die horizontale (waagerechte) Abweichung einer Verteilung von der Normalverteilung. Der höchste Punkt einer Verteilung befindet sich nicht mehr in der Mitte, sondern links oder rechts von der Mitte. Die Wölbung ist ein Maß für die vertikale (senkrechte) Abweichung einer Verteilung von der Normalverteilung. Sie informiert, ob eine Verteilung breit- oder schmalgipflig ist.

Für die Berechnung der Schiefe und Wölbung wurden Maßzahlen entwickelt. Sind die Werte jeweils 0, dann ist die Abweichung von der Normalverteilung gleich 0 (Schendera 2015, S. 133).

2.4.1 Schiefe

Die Schiefe ist ein Maß für die horizontale Abweichung einer Verteilung von der Normalverteilung. In Abbildung 4 sind drei typische Verteilungen dargestellt. Die mittlere Grafik in Abbildung 4 zeigt eine symmetrische Verteilung. Die Häufigkeiten der Werte nehmen symmetrisch von der Mitte ab.

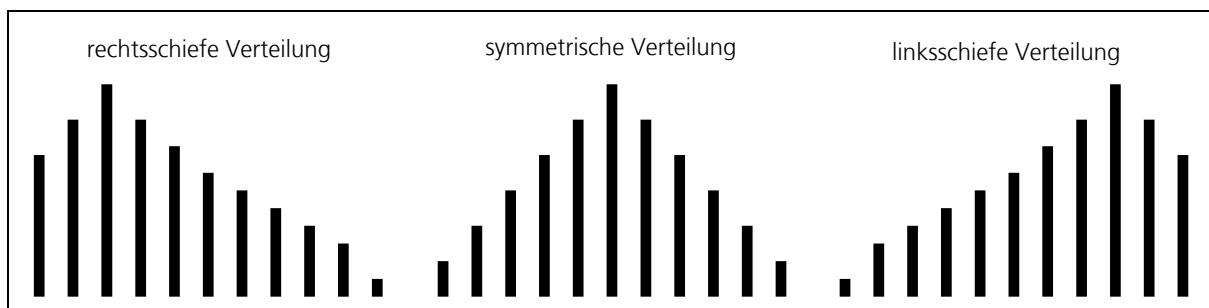
Bei der linken Grafik in Abbildung 4 handelt es sich um eine sogenannte rechtsschiefe Verteilung. Die meisten Werte bzw. der häufigste Wert befinden bzw. befindet sich auf der linken Seite der Verteilung. Nach rechts nehmen die Werte immer mehr ab. Anstelle der Bezeichnung „rechtsschief“ wird auch der Begriff „linksgipflig“ verwendet (Völkl und Korb 2018, S. 118).

Verwirrende Begriffsvielfalt

Die Bezeichnungen „rechtsschief“ und „linksschief“ sind nicht sehr zugänglich, da der „Gipfel“ der Verteilung jeweils entgegengesetzt liegt. Weiß (2013, S. 67) nutzt die Bezeichnungen „rechtsgipflig“ und „linksgipflig“, die eindeutig den Ort der meisten Werte beschreiben.

Bei der rechten Grafik in Abbildung 4 handelt es sich um eine sogenannte linksschiefe Verteilung. Die meisten Werte bzw. der häufigste Wert befinden bzw. befindet sich auf der rechten Seite der Verteilung. Nach links nehmen die Werte immer mehr ab. Anstelle der Bezeichnung „linksschief“ wird auch der Begriff „rechtssteil“ verwendet (Völkl und Korb 2018, S. 120).

Abbildung 4: Schiefe



Quelle: Eigene Darstellung

Für die Berechnung der Schiefe wird zunächst die Abweichung jedes Werts (x_i) vom arithmetischen Mittel (\bar{x}) gebildet. Diese Differenz wird durch die Standardabweichung (s_x) dividiert und anschließend mit 3 potenziert. Durch dieses Vorgehen werden erstens Werte, die weiter vom arithmetischen Mittel entfernt sind, stärker gewichtet und zweitens bleiben negative Abweichungen durch die dritte Potenz erhalten. Abschließend wird die Summe durch die Fallzahl (n) dividiert. Formal:

$$\text{Schiefe} = \frac{\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^3}{n}$$

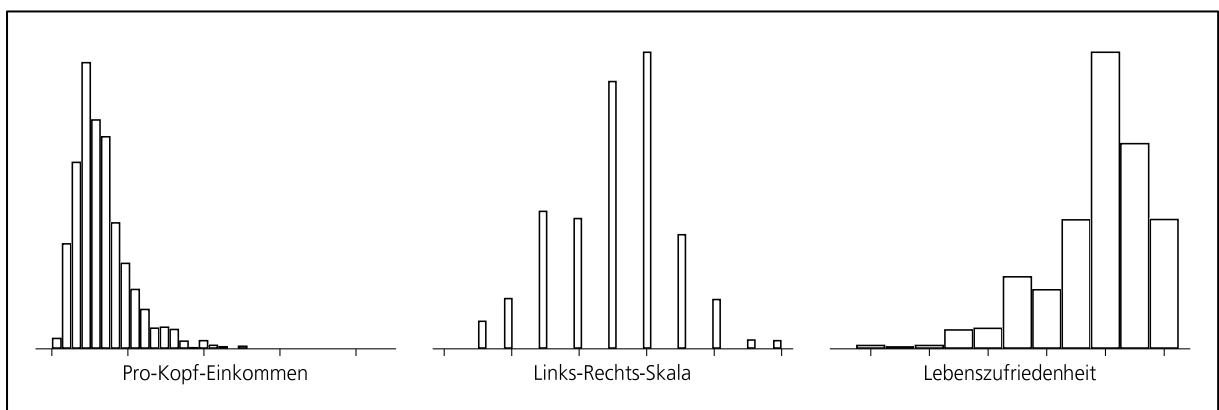
Die Schiefe einer Verteilung wird heute nicht mehr händisch berechnet. Diese Aufgabe übernimmt ein Statistikprogramm wie SPSS, Stata oder R. Zur angemessenen Beschreibung einer Verteilungsform muss allerdings der Schiefe-Koeffizient korrekt interpretiert werden. Drei Fälle werden unterschieden:

- Schiefe < 0: linksschiefe Verteilung (rechtsgipflig)
- Schiefe = 0: symmetrische Verteilung
- Schiefe > 0: rechtsschiefe Verteilung (linksgipflig)

Nach Schendera (2015, S. 134) werden Schiefe-Werte, deren Beträge größer oder gleich 1 sind, als deutliche Abweichung vom symmetrischen Verlauf der Normalverteilung interpretiert.

In Abbildung 5 sind drei Beispiele für empirische Verteilungen mit unterschiedlicher Schiefe dargestellt. Ein typisches Beispiel für eine linksgipflige Verteilung (rechtsschief) ist das Pro-Kopf-Einkommen. Viele Bürgerinnen und Bürger haben ein geringes Einkommen, wenige Menschen ein hohes Einkommen. Mit einem Skewness-Wert von 2,18 kann die Einkommensverteilung als asymmetrisch bezeichnet werden. Ein Beispiel für eine rechtsgipflige Verteilung (linksschief) ist die subjektive Lebenszufriedenheit, die in der ALLBUS auf einer Skala von 0 bis 10 erfasst wird. Viele Personen sind mit ihrem Leben zufrieden, wenige Personen sind (sehr) unzufrieden. Mit dem Skewness-Wert von $-1,18$ ist die Schiefe zwar geringer ausgeprägt als beim Pro-Kopf-Einkommen, aber gemäß der Konvention von Schendera (2015, S. 134) als deutliche Abweichung vom symmetrischen Verlauf der Normalverteilung zu werten. Relativ symmetrisch ist die Verteilung auf der Links-Rechts-Skala, die in der empirischen Sozialforschung genutzt wird, um die ideologische Position einer Person zu erfassen (Warwick 2002). Der Wert der Schiefe liegt mit $-0,16$ nur gering unter dem Referenzwert von 0. Das negative Vorzeichen deutet auf eine rechtsgipflige (linksschiefe) Verteilung hin.

Abbildung 5: Empirische Verteilungen mit unterschiedlicher Schiefe



Daten: ALLBUS 2016. Eigene Berechnungen

Fechner'sche Lageregel

Eine einfache Möglichkeit, die Schiefe einer Verteilung ohne Berechnung des entsprechenden Formmaßes und ohne grafische Darstellung zu bestimmen, bietet die sogenannte Fechner'sche Lageregel (Völkl und Korb 2018, S. 120-122; Kosfeld et al. 2016, S. 132-134).

- Bei symmetrischen Verteilungen weisen die drei Lagemaße Modus, Median und arithmetisches Mittel denselben Wert auf (siehe Abbildung 3).
- Eine Verteilung ist rechtsschief (linksgipflig), wenn der Modus kleiner als der Median und der Median kleiner als das arithmetische Mittel ist ($\text{Modus} < \text{Median} < \text{Arithmetisches Mittel}$).

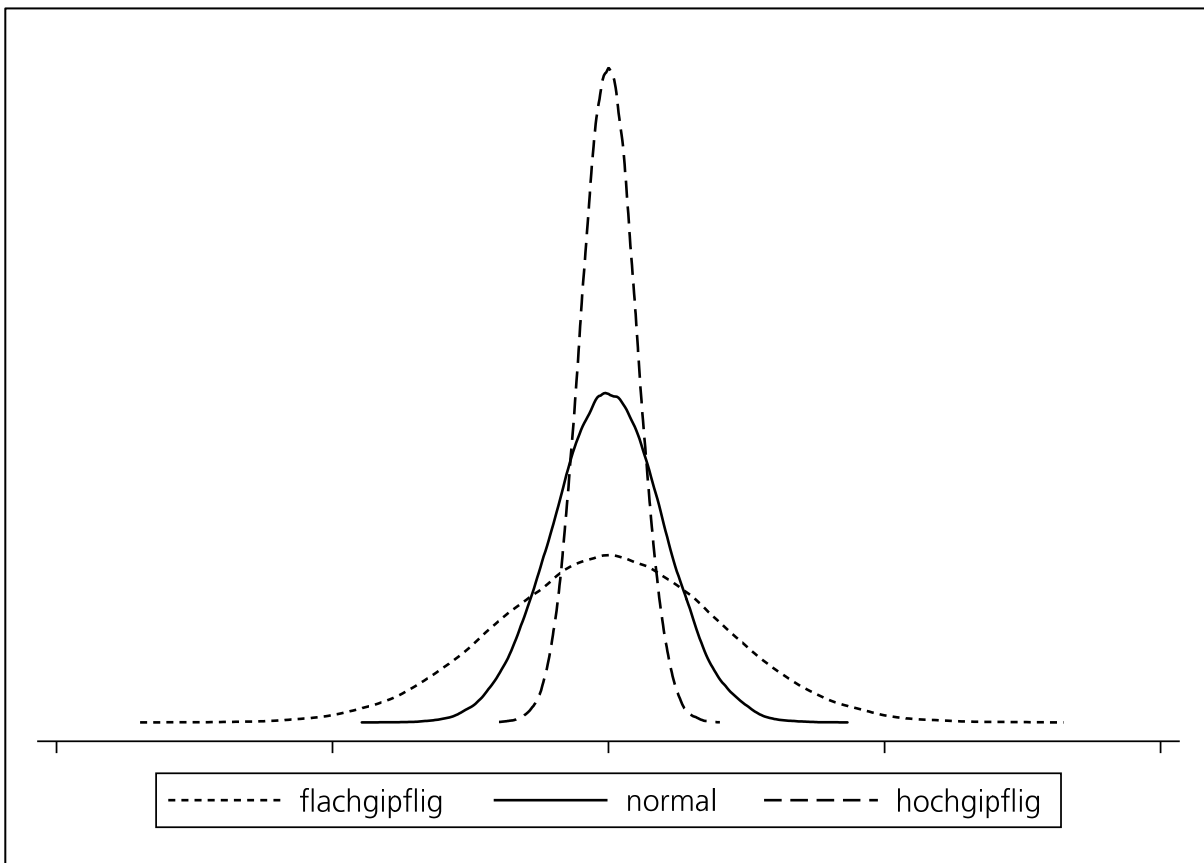
- Eine Verteilung ist linksschief (rechtsgipflig), wenn der Modus größer als der Median und der Median größer als das arithmetische Mittel ist ($\text{Modus} > \text{Median} > \text{Arithmetisches Mittel}$).

Völkl und Korb (2018, S. 121) weisen allerdings darauf hin, dass es sich bei der Fechner'schen Lageregel um eine Daumenregel handelt, die nicht immer korrekt ist.

2.4.2 Wölbung

Die Wölbung (Kurtosis) einer Verteilung ist das Maß ihrer vertikalen Abweichung von der Normalverteilung. In Abbildung 6 werden die Normalverteilung und zwei typische vertikale Abweichungen dargestellt. Die Normalverteilung (durchgezogene Linie) dient als Referenzwert der Wölbung. Eine Verteilung kann einerseits flacher (flachgipflige Verteilung) oder andererseits steiler verlaufen (hochgipflige Verteilung).

Abbildung 6: Wölbung



Quelle: Eigene Darstellung

Bei einem flacheren Verlauf liegen weniger Werte in der Mitte und mehr Werte an den Rändern der Verteilung. Eine solche Verteilung wird als flach- oder breitgipflig bezeichnet. Bei einem steileren Verlauf liegen mehr Werte in der Mitte und weniger Werte an den Rändern der Verteilung. Eine solche Verteilung wird hochgipflig oder auch schmalgipflig genannt (Schendera 2015, S. 134-135; Völkl und Korb 2018, S. 124-126).

Die Formel für die Berechnung der Wölbung unterscheidet sich in zwei Punkten von der Formel für die Berechnung der Schiefe. Erstens ist der Exponent 4 statt 3, zweitens gibt es einen Korrekturfaktor 3. Der Korrekturfaktor ist erforderlich, da die Normalverteilung einen Referenzwert von 3 hat. Durch die Korrektur wird das Maß auf 0 zentriert und ermöglicht einfache Vergleiche. Für die Berechnung wird auf folgende Formel zurückgegriffen:

$$\text{Wölbung} = \frac{\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^4}{n} - 3$$



Auch die Wölbung einer Verteilung wird heute nicht mehr händisch berechnet. Zur Unterscheidung der Verteilungsformen muss der entsprechende Koeffizient allerdings angemessen interpretiert werden. Drei Fälle werden unterschieden:

- Kurtosis < 0: flachgipflige Verteilung
- Kurtosis = 0: Normalverteilung
- Kurtosis > 0: hochgipflige Verteilung

Bei einem Kurtosis-Wert von 0 entspricht die Verteilung exakt der Wölbung einer Normalverteilung. Werte kleiner als 0 deuten auf einen flacheren Verlauf und Werte größer als 0 auf eine steilere Verteilung hin (Ludwig-Mayerhofer et al. 2014, S. 83).

2.5 Variablen standardisieren (z-Transformation)

In der Forschungspraxis sind wir häufig mit unterschiedlichen Skalierungen (Messeinheiten) konfrontiert. Die Lebenszufriedenheit könnte beispielsweise nicht mit einer 11-Punkt-Skala (Werte von 0 bis 10), sondern auch mit einer 7-Punkt-Skala (Werte von 1 bis 7) erfasst werden. Unterschiedliche Skalierungen erschweren erheblich den Vergleich von zwei Messwerten. Wir wollen dieses Problem an einem Beispiel illustrieren.



Beispiel

Lisa und Bart haben an zwei verschiedenen Leistungstests teilgenommen. Lisa hat 45 Punkte erzielt, Bart 60 Punkte. Bart hat absolut also mehr Punkte erreicht als Lisa und ist entsprechend stolz auf seinen Erfolg. Doch war seine Leistung wirklich besser als die Leistung von Lisa? Ohne weitere Informationen können wir diese Frage nicht beantworten. Lisa hat zwei wichtige Informationen zu den beiden Leistungstests recherchiert: das arithmetische Mittel und die Standardabweichung. Wie Tabelle 16 zeigt, liegt das arithmetische Mittel bei Lisas Leistungstest bei 25 und die Standardabweichung bei 10. Das arithmetische Mittel des Leistungstests, an dem Bart teilgenommen hat, liegt bei 50 und die Standardabweichung beträgt 25. Die Werte von Lisa (45) und von Bart (60) liegen jeweils über dem arithmetischen Mittel. Beide haben überdurchschnittliche Leistungen erzielt. Ein direkter Vergleich ist allerdings nicht möglich, da sich die Verteilungen der beiden Leistungstests unterscheiden.

Tabelle 16: Variablen standardisieren

Person	Wert (x_i)	Arithmetisches Mittel (\bar{x})	Standardabweichung (s_x)
Lisa	45	25	10
Bart	60	50	25

Quelle: Eigene Darstellung

Um Werte unterschiedlicher Verteilungen vergleichen zu können, müssen diese standardisiert werden. Diese Standardisierung wird in der Fachliteratur als z-Transformation (Völkl und Korb 2018, S. 110-116; Ludwig-Mayerhofer et al. 2014, S. 80-81) oder auch als z-Standardisierung (Diaz-Bone 2018, S. 64-65) bezeichnet.

Bei der z-Standardisierung wird jeder Wert einer Verteilung – in unserem Fall die Punktzahl jedes Teilnehmenden – in einen z-Wert transformiert. Nach der Transformation hat jede Person einen ursprünglichen Wert (x-Wert) und einen standardisierten Wert (z-Wert). Während die x-Werte nicht vergleichbar sind, sind die z-Werte über die Ursprungsverteilung hinaus vergleichbar. Für die Berechnung des z-Werts wird folgende Formel verwendet:

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Bei der Berechnung des z-Werts lassen sich zwei Schritte unterscheiden: Zentrierung und Normierung (Diaz-Bone 2018, S. 64-65; Völkl und Korb 2018, S. 110-111). Im ersten Schritt wird die Variable zentriert. Dafür wird von jedem x-Wert das entsprechende arithmetische Mittel abgezogen ($x_i - \bar{x}$). Dadurch liegt das neue arithmetische Mittel der Variable bei 0. Für Personen, die einen x-Wert unterhalb des arithmetischen Mittels haben, ergibt sich ein negativer Wert. Für Personen, die einen x-Wert oberhalb des arithmetischen Mittels haben, ergibt sich ein positiver Wert. Der zweite Schritt wird als Normierung bezeichnet. Dabei wird der zentrierte x-Wert durch die Standardabweichung (s_x) dividiert. Dadurch wird die ursprüngliche Verteilung gestaucht oder gestreckt. Die Standardabweichung der neuen Verteilung beträgt daher immer 1.

Zur Illustration berechnen wir an dieser Stelle die z-Werte von Lisa und Bart. Für diesen Zweck müssen die Werte aus Tabelle 16 in die Formel eingetragen werden.

$$z_{\text{Lisa}} = \frac{45 - 25}{10} = 2$$

$$z_{\text{Bart}} = \frac{60 - 50}{25} = 0,4$$

Der z-Wert von Lisa beträgt 2, der z-Wert von Bart liegt bei 0,4. Wie können diese beiden Werte jetzt verglichen werden? Durch die z-Transformation liegen die Werte von Lisa und Bart nicht mehr in der ursprünglichen Maßeinheit vor, sondern in einem Vielfachen der Standardabweichungen.

„Das heißt, ein z-Wert gibt an, wie viele Standardabweichungen ein Wert über oder unter dem Mittelwert liegt. Z-Werte können positiv oder negativ sein. Ein negativer z-



Wert bedeutet, dass der interessierende Datenpunkt kleiner ist als der Mittelwert, ein positiver z-Wert, dass der Datenpunkt größer ist als der Mittelwert. Die Größe der z-Werte wiederum gibt Aufschluss darüber, wie breit sie um den Mittelwert streuen. Ein kleiner Betrag des z-Werts weist darauf hin, dass er in der Nähe des Mittelwerts liegt und damit in einen Bereich fällt, in dem sich ein Großteil der Fälle einer Verteilung befindet. Ein hoher Betrag des z-Werts dagegen gilt als außergewöhnlich und kann ein Hinweis auf einen Ausreißer sein.“ (Völkl und Korb 2018, S. 111)

Übertragen auf unser Beispiel hat Bart zwar eine überdurchschnittliche Leistung erzielt, weil sein z-Wert positiv ist. Seine Leistung ist allerdings relativ zu Lisa schlechter, da Lisa einen z-Wert von 2 hat. Ihre Leistung ist zwei Standardabweichungen größer als der Durchschnittswert. Durch die Standardisierung können die Testergebnisse von Bart und Lisa direkt verglichen werden.

2.6 Grafische Darstellungen

Für die Darstellung statistischer Informationen bieten sich häufig Diagramme an, die Häufigkeiten oder Anteile eines Merkmals übersichtlicher und schneller erfassbar präsentieren können (z.B. Gehring und Weins 2009, S. 110-117; Degen 2010; Kuckartz et al. 2013, S. 42-53; Diaz-Bone 2018, S. 39-44). Mit dem Säulen-, Balken- und Kreisdiagramm sowie dem Histogramm und dem Boxplot stellen wir wichtige grafische Darstellungsformen vor.

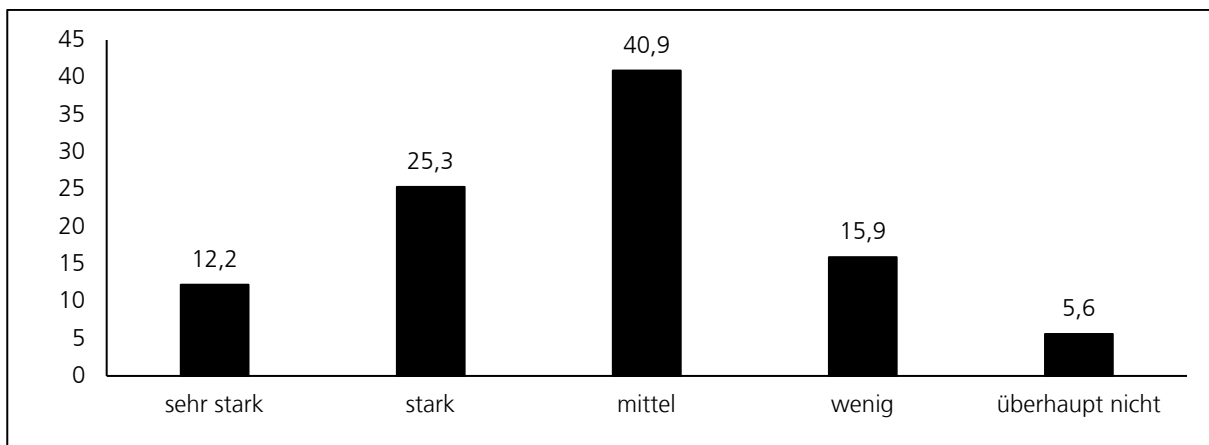
2.6.1 Säulen- und Balkendiagramm

Säulen- und Balkendiagramme sind Darstellungsformen für Merkmale mit wenigen Ausprägungen. Sie eignen sich für nominale und ordinale Variablen. Die Häufigkeiten oder Anteile der einzelnen Kategorien werden durch Säulen bzw. Balken gleicher Breite dargestellt. Die Höhe der Säulen bzw. die Länge der Balken entspricht der absoluten oder relativen Häufigkeit der jeweiligen Merkmalsausprägung der jeweiligen Kategorie (Kühnel und Krebs 2007, S. 64).

Wie unterscheiden sich Säulen- von Balkendiagrammen? Bei einem Säulendiagramm werden die einzelnen Kategorien durch stehende (vertikale) Rechtecke dargestellt, bei einem Balkendiagramm durch liegende (horizontale) Balken (Gehring und Weins 2009, S. 110). Säulendiagramme können unkompliziert in Balkendiagramme transformiert werden. Dazu muss die Abbildung nur um 90 Grad gedreht werden.

Abbildung 7 zeigt das Säulendiagramm des Interesses an Politik auf Grundlage der ALLBUS 2016. Die einzelnen Säulen entsprechen den Antwortkategorien. Dargestellt ist die prozentuale Häufigkeit der einzelnen Antwortmöglichkeiten. 40,9 Prozent der Befragten haben angegeben, sich „mittel“ für Politik zu interessieren.

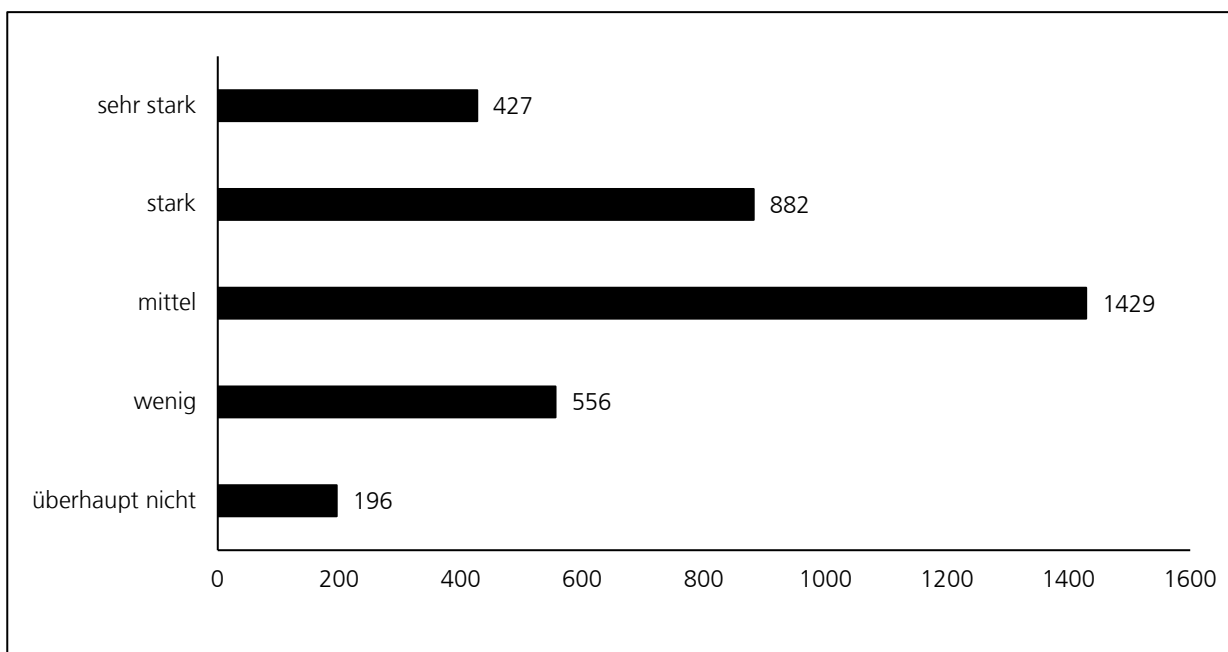
Abbildung 7: Säulendiagramm des Interesses an Politik (in Prozent, n = 3490)



Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).

Abbildung 8 informiert über das subjektive Interesse an Politik in Form eines Balkendiagramms. Anstelle der relativen Häufigkeiten in Prozent werden die absoluten Häufigkeiten der einzelnen Antwortmöglichkeiten dargestellt.

Abbildung 8: Balkendiagramm des Interesses an Politik (absolute Häufigkeiten, n = 3490)



Daten: ALLBUS 2016. Eigene Berechnungen (ohne Designgewicht).

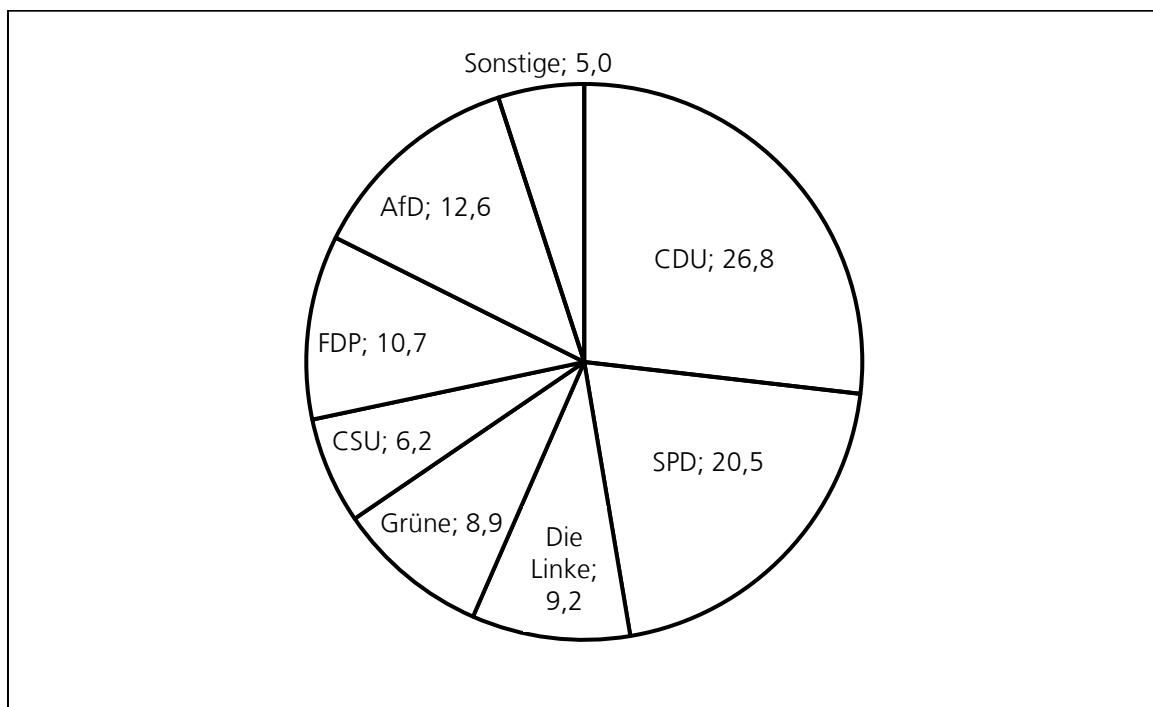
2.6.2 Kreisdiagramm

In den Medien und in Präsentationen von Unternehmensberatungen werden häufig Kreisdiagramme genutzt (häufig auch Kuchen- oder Tortendiagramme genannt), um Merkmale mit wenigen Ausprägungen darzustellen. Die Häufigkeiten oder Anteile einer Kategorie werden durch verschieden große Kreissegmente dargestellt. In einer dreidimensionalen Darstellung entsprechen die einzelnen Kreissegmente Tortenstücken. Einzelne Kreiselemente lassen sich auch hervorheben.

! In der sozialwissenschaftlichen Literatur werden Kreisdiagramme eher kritisch gesehen (Kuckartz et al. 2013, S. 48; Diaz-Bone 2018, S. 44). Die Darstellung von Häufigkeiten oder Anteilen in Kreesegmenten oder Tortenstücken ist schwierig zu interpretieren, weil der Betrachter Winkel und Flächen interpretieren muss (Cleveland 1994, S. 262-264). Deshalb kann die Darstellung leicht über die tatsächlichen Häufigkeiten bzw. Anteile der einzelnen Kategorien täuschen (Kühnel und Krebs 2007, S. 63). Insbesondere die dreidimensionale oder perspektivische Gestaltung eines Kreisdiagramms erschwert die Interpretation (Diaz-Bone 2018, S. 44). Deshalb wird von dieser Darstellungsform eher abgeraten (Kohler und Kreuter 2017, S. 181). Deutlicher formuliert es Plümper (2012, S. 173): „In keinem vernünftigen Journal werden Tortendiagramme abgedruckt.“

Abbildung 9 zeigt ein Kreisdiagramm der Zweitstimmenanteile bei der Bundestagswahl 2017. Die Interpretation wird erschwert, wenn die Häufigkeiten oder Anteile ähnlich groß sind. Dies trifft in Abbildung 9 insbesondere auf die kleineren Parteien zu. Ohne die Angabe der Prozente wäre der Unterschied zwischen Grünen und Linkspartei praktisch nicht zu erkennen. In den meisten Fällen bietet ein Balken- oder Säulendiagramm eine verständlichere Präsentation der Daten als ein Kreisdiagramm.

Abbildung 9: Zweitstimmen bei der Bundestagswahl 2017 (in Prozent)



Quelle: Bundeswahlleiter (<https://bundeswahlleiter.de/bundestagswahlen/2017/ergebnisse.html>)

2.6.3 Histogramm

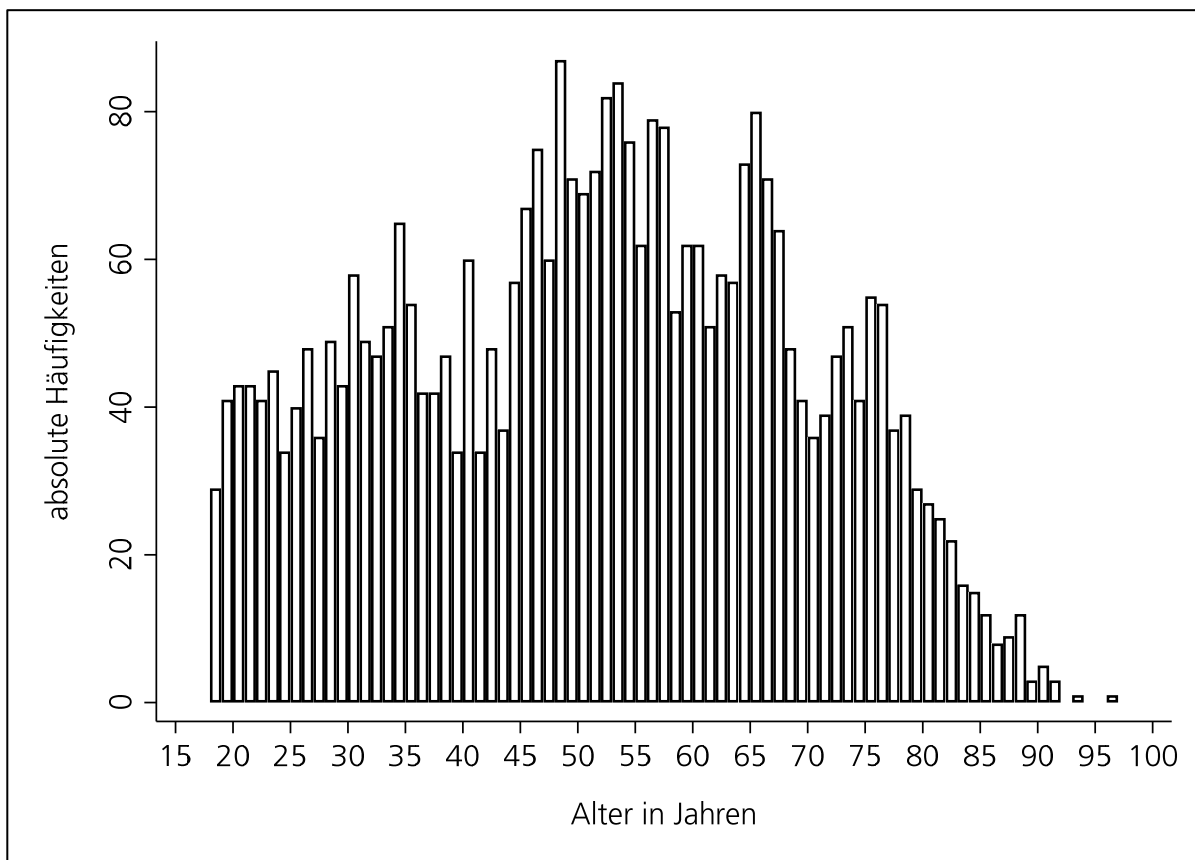
Histogramme sind Darstellungen für metrische Variablen mit vielen Ausprägungen bzw. vielen Gruppen. Der auffälligste Unterschied zu Säulen- und Balkendiagrammen ist, dass die Säulen eines Histogramms unmittelbar aneinander angrenzen (Gehring und Weins 2009, S. 113). Für die Säulen gilt das Prinzip der Flächentreue: Die Fläche über den Klassen (das Produkt aus Säulenhöhe und Säulenbreite) ist proportional zu den absoluten bzw. relativen Häufigkeiten (Kühnel und Krebs

2007, S. 56-60; Gehring und Weins 2009, S. 113). Deshalb bieten Histogramme einen informativen Eindruck der empirischen Häufigkeitsverteilung. Symmetrie, Schiefe und Steilheit einer Verteilung können leicht erkannt werden.

Häufig werden Histogramme für gruppierte Daten verwendet (z.B. Altersgruppen oder Einkommensklassen). Für die Mindestanzahl der Klassen gibt es zahlreiche Empfehlungen. Bei bis zu 100 Beobachtungswerten sind mindestens zehn Klassen, bei 1000 Beobachtungen mindestens 13 Klassen und bei 10000 Beobachtungen mindestens 16 Klassen zu bilden (Degen 2010, S. 99; siehe für weitere Vorschläge auch Schnell 1994, S. 21-25). Abbildung 10 zeigt ein Histogramm des Merkmals „Alter“ in der ALLBUS 2016. Die einzelnen Säulen repräsentieren jeweils eine Klasse (Alter in Jahren).

Bei der Konstruktion eines Histogramms empfiehlt Degen (2010, S. 100) die Bildung gleich breiter (äquidistanter) Klassen. Durch eine identische Klassenbreite können die Häufigkeiten leichter verglichen werden. Grundsätzlich sind aber auch unterschiedliche Klassenbreiten möglich. Dann müssen allerdings die Flächeninhalte der Rechtecke verglichen werden. Diese sind deutlich schwieriger zu interpretieren.

Abbildung 10: Histogramm des Alters (absolute Häufigkeiten, $n = 3486$)



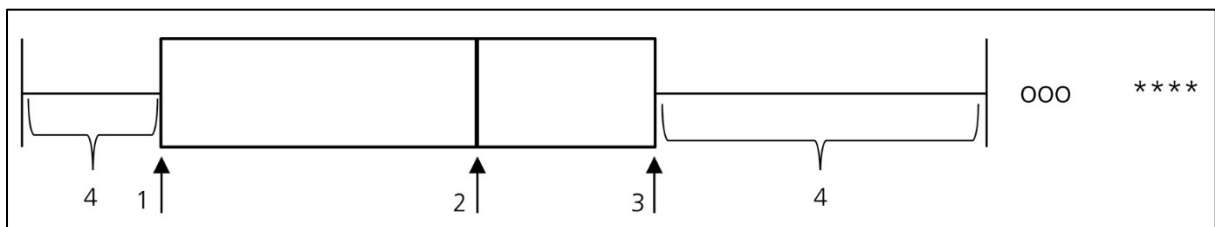
Daten: ALLBUS 2016. Eigene Berechnungen

2.6.4 Boxplot

Eine wichtige Darstellungsform für metrische Merkmale ist der Boxplot, der oft auch als Box-and-Whisker-Plot bezeichnet wird und von John W. Tukey vorgestellt wurde (Tukey 1977; Degen 2010, S. 95-98).

Abbildung 11 sind die charakteristischen Elemente eines Boxplots zu entnehmen. Ein Boxplot besteht aus einem Rechteck (Box), das die mittleren 50 Prozent der Beobachtungswerte umfasst. Linie 1 repräsentiert das untere Quartil (25 Prozent der Beobachtungen) und Linie 3 das obere Quartil (75 Prozent der Beobachtungen). Linie 2 zeigt den Median an (50 Prozent der Beobachtungen). Der Bereich vom unteren zum oberen Quartil wird auch als Interquartilsabstand bezeichnet.

Abbildung 11: Elemente eines Boxplots

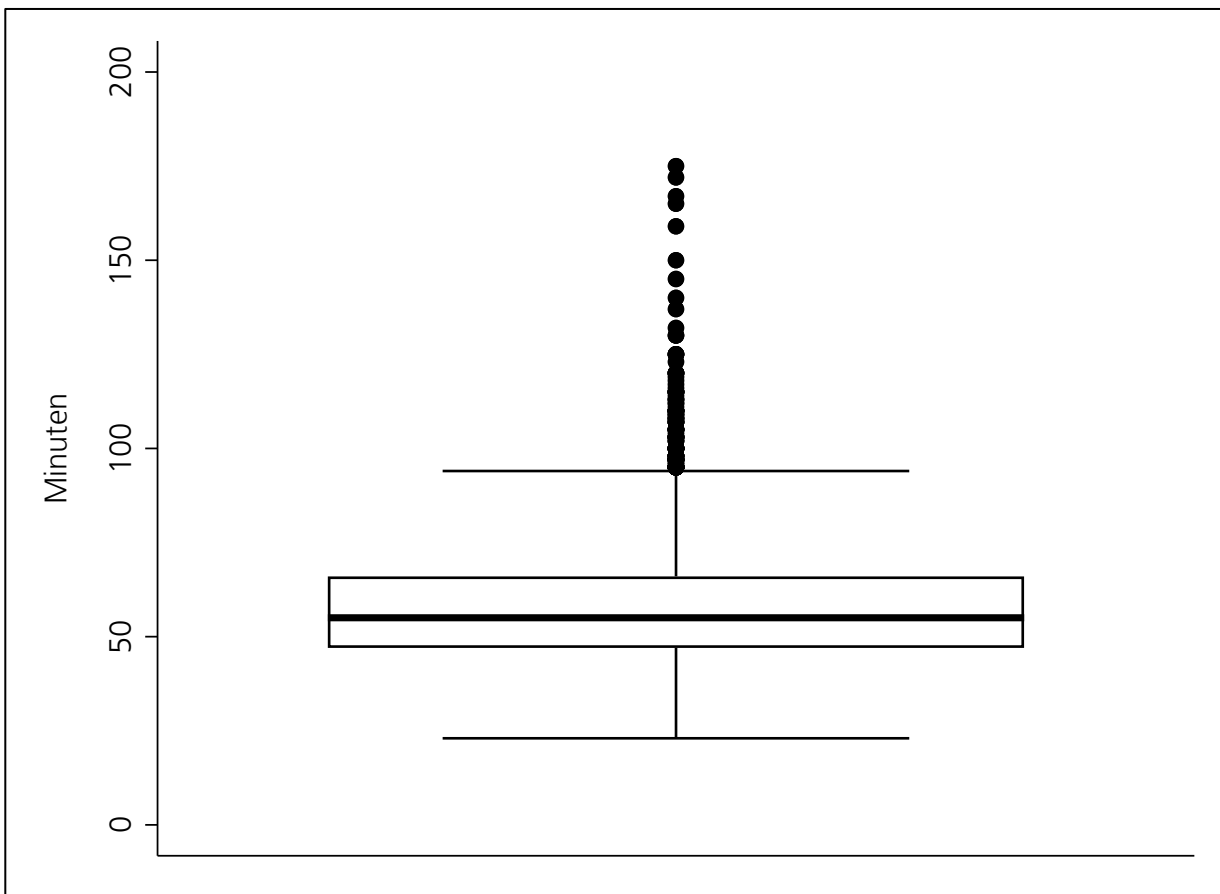


Quelle: Eigene Darstellung

Durch die Antennen (4) werden die Werte außerhalb der Box dargestellt. Diese Antennen werden auch Whisker (Barthaare) genannt. Daher kommt die englische Bezeichnung als Box-and-Whisker-Plot. Die Länge der Whiskers wird auf maximal das 1,5-Fache des Interquartilsabstands beschränkt. Gibt es keine Werte außerhalb der Grenze, dann wird die Länge durch den minimalen und maximalen Datenpunkt festgelegt. Beobachtungen, die außerhalb der Whiskers liegen, gelten als auffällige Datenpunkte. Das Statistikprogramm SPSS unterscheidet dabei zwischen Ausreißern, die mit einem Kreis markiert werden, und Extremwerten, die mit einem Stern gekennzeichnet werden. Als Extremwerte gelten Beobachtungen, die mehr als das Dreifache des Interquartilsabstands von den Rändern der Box entfernt liegen (Degen 2010, S. 96-97). Das Statistikprogramm Stata verzichtet auf diese Differenzierung.

Abbildung 12 zeigt einen Boxplot der Interviewdauer in der ALLBUS 2016. Der Median der Verteilung – also der „mittlere“ Wert einer geordneten Verteilung – liegt bei 55 Minuten. Das 25-Prozent-Quartil liegt bei 47 Minuten, das 75-Prozent-Quartil bei 66 Minuten. 50 Prozent der Interviews dauerten also zwischen 47 und 66 Minuten. Mit den Punkten sind die Ausreißer markiert. Das längste Interview dauerte 175 Minuten.

Abbildung 12: Boxplot der Interviewdauer (n = 3479)



Daten: ALLBUS 2016. Eigene Berechnungen

3 Bivariate Datenanalyse

Simone Abendschön



Nachdem sich das letzte Kapitel ausführlich mit der Beschreibung und Analyse einzelner Variablen beschäftigt hat, geht es in diesem Kapitel darum, zwei Merkmale in ihrem gemeinsamen Auftreten zu betrachten. Ein häufiges Ziel der empirischen Sozialforschung ist es, relevante Zusammenhänge bzw. Unterschiede zwischen zwei Merkmalen aufzudecken. Beispiele dafür sind Zusammenhänge zwischen dem Bildungshintergrund der Eltern und schulischem Erfolg ihrer Kinder oder, ob die individuelle Wahlbeteiligung mit dem finanziellen Einkommen zusammenhängt. Um solche bivariaten Zusammenhänge zu überprüfen, stehen verschiedene Analysemöglichkeiten zur Verfügung.

In diesem Abschnitt werden die bekanntesten Instrumente der bivariaten Datenanalyse vorgestellt. Dazu wird zunächst die Kreuztabelle (auch Kontingenztafel genannt) eingeführt, die ein Werkzeug der deskriptiven Statistik darstellt und zwei Merkmale kombiniert. Neben der Einführung in die Konstruktion und Interpretation von Kreuztabellen wird auch gezeigt, dass auf Basis von Kreuztabellen bereits erste Rückschlüsse auf Hypothesen möglich sind, die dann mit weiteren Analysen ergänzt werden können.

Begleitend zu einer deskriptiven Betrachtung lassen sich sogenannte Zusammenhangs- oder Assoziationsmaße berechnen, das sind statistische Parameter wie Phi und Cramer's V^3 , die eine Aussage zum Zusammenhang zwischen den beiden betrachteten Variablen erlauben. Diese Parameter werden auch als Kontingenz-, Assoziations- und Korrelationskoeffizienten bezeichnet (Benninghaus 2007, S. 67). Wie aus Tabelle 17 hervorgeht, spielt das Messniveau der beiden Variablen eine zentrale Rolle bei der Wahl des richtigen bzw. aussagekräftigen Zusammenhangsmaßes.

Tabelle 17: Bivariate Zusammenhangsmaße in Abhängigkeit vom Skalenniveau

Skalenniveau	nominal	ordinal	metrisch
nominal	Cramer's V (Phi) Lambda C	Cramer's V Lambda C	Eta-Koeffizient
ordinal	Cramer's V Lambda C	Spearman's Rho Kendalls Tau A, B, C gamma	Spearman's Rho Kendalls Tau A, B, C gamma
metrisch	Eta-Koeffizient	Spearman's Rho Kendalls Tau A, B, C gamma	Pearson's r

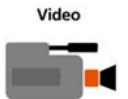
Quelle: Eigene Darstellung

³ Da das Maß auf den schwedischen Statistiker Harald Cramér zurückgeht, lautet die korrekte Schreibweise eigentlich Cramér's V. Viele Statistikbücher nutzen allerdings die vereinfachte Schreibweise Cramer's V.

Liegen nominale Merkmale vor, dann bietet sich beispielsweise Cramer's V an. Bei ordinalen Variablen kann Spearman's Rho verwendet und bei metrischen Variablen kann auf Pearson's r zurückgegriffen werden. Die bekanntesten und am häufigsten genutzten Zusammenhangsmaße sind sicherlich Cramer's V, Spearman's Rho, Pearson's r und Eta (Benninghaus 2007; Gehring und Weins 2009, S. 148-176; Weins 2010, S. 73-88), die in den folgenden Abschnitten vorgestellt werden.

3.1 Kreuztabellen

Die Kreuztabellenanalyse stellt in der sozialwissenschaftlichen Datenanalyse ein klassisches bivariates Instrument zur Betrachtung von Zusammenhängen zweier Merkmale dar (Diaz-Bone 2018, S. 70-77; Westle 2018, S. 329-334). Sie ermöglicht eine übersichtliche und kompakte Darstellung der Häufigkeiten der interessierenden Merkmale, in der auch prozentuale Anteile berechnet werden können. Die Erstellung von Kreuztabellen empfiehlt sich grundsätzlich als erster Schritt der bivariaten Analyse, wenn die Skalenvoraussetzungen der betreffenden Variablen vorhanden sind.



Mit Hilfe von Kreuztabellen lassen sich die kombinierten Häufigkeiten mindestens zweier Variablen darstellen. Deren gemeinsames Auftreten wird in Tabellenform abgebildet. Technisch gesprochen handelt es sich bei Kreuztabellen um Matrizen, in deren Zellen die beobachteten (absoluten und relativen) Häufigkeiten der Fälle der Stichprobe abgetragen werden. Allein schon aus Gründen der Übersichtlichkeit ist es daher plausibel, dass sich zur Darstellung in Kreuztabellen vor allem *kategoriale* (nominale) bzw. teilweise auch *ordinale* Variablen eignen, letztere vor allem dann, wenn sie über wenige Ausprägungen verfügen. Metrische Variablen hingegen, die über eine Vielzahl von potenziellen Ausprägungen verfügen, eignen sich per se nicht (siehe dazu den Abschnitt über Pearson's r). Allerdings lassen sich metrische Variablen häufig in wenige Kategorien (z.B. Alters- oder Einkommensgruppen) zusammenfassen. In dieser gruppierten Form kann dann ebenfalls die Kreuztabellenanalyse genutzt werden.



An einem Beispiel soll das Prinzip der Kreuztabelle deutlich werden. Eine Universitätsbibliothek überlegt, ob sie ihr Serviceangebot in den Abendstunden ausweiten soll. Es soll dabei herausgefunden werden, welche Zielgruppe insbesondere angesprochen werden soll. Unterscheiden sich BA- und MA-Studierende in der Nutzung des Abendangebots der Unibibliothek? Dazu wurden 100 Studierende befragt, die angeben sollten, ob sie die Öffnungszeiten am Abend nutzen oder nicht und ob sie in einem BA- oder einem MA-Studiengang eingeschrieben sind. Tabelle 18 zeigt die beobachteten Merkmale für neun der 100 Studierenden.



Soll nun in kompakter Form abgebildet werden, wie sich die Nutzungspräferenzen auf die beiden Studiengänge verteilen, kann man aus den Angaben aus Tabelle 18 eine Kreuztabelle in ihrer einfachsten Form erstellen. Eine sogenannte 2x2-Tabelle oder Vierfeldertafel wie sie in Tabelle 19 zu sehen ist, stellt die beiden beobachteten dichotomen Merkmale kombiniert in einer Kreuztabelle dar. Dabei findet sich in der Zeile die Antwort auf die Frage, ob eine abendliche Nutzung erfolgt, und in der Spalte die Studiengangszugehörigkeit. Kreuztabellen verfügen also über k Zeilen und l Spalten, was bedeutet, dass eine Kreuztabelle $k \times l$ Zellen umfasst. Im einfachen Beispiel aus Tabelle 19 gibt es $2 \times 2 = 4$ Zellen. Als Zeilenvariable, hier „Nutzung“, bezeichnet man die Variable, deren Merkmalsausprägungen die Zeilen der Kreuztabelle bilden. Als Spaltenvariable

bezeichnet man die Variable, deren Ausprägungen die Spalten der Kreuztabelle vorgeben (hier „Studiengang“).

Tabelle 18: Urliste – Abendliche Bibliotheksnutzung und Studiengang (n = 9)

ID	Studiengang	Nutzung
1	BA	Nein
2	MA	Ja
3	MA	Nein
4	BA	Nein
5	BA	Ja
6	MA	Ja
7	MA	Ja
8	BA	Nein
9	MA	Ja

Quelle: Eigene Darstellung

Die abgetragenen Zahlenwerte in den Zellen von Tabelle 19 sind Auszählungen der Kombinationen, die sich aus Tabelle 18 ergeben. Wir sehen in den vier Zellen die absoluten Häufigkeiten. Drei Studierende des BA-Studiengangs nutzen das Abendangebot nicht, während dies vier Studierende des MA-Studiengangs tun. Aus dem Beispiel von Tabelle 19 ist ersichtlich, dass man Zeilen- und Spaltenvariable auch hätte vertauschen können – die kombinierten Häufigkeiten ändern sich dabei nicht, sondern lediglich die Form der Darstellung.

Tabelle 19: Kreuztabelle – Abendliche Bibliotheksnutzung und Studiengang (n = 9)

Nutzung \ Studiengang	Studiengang		
	BA	MA	Gesamt
Ja	1	4	5
Nein	3	1	4
Gesamt	4	5	9

Quelle: Eigene Darstellung

Die unterste Zeile sowie die Spalte ganz rechts von Tabelle 19 beinhalten die sogenannten Randhäufigkeiten. Als Randhäufigkeiten bezeichnet man den rechten und unteren „Rand“ einer Kreuztabelle. Diese Informationen, die sich auch durch univariate Häufigkeitsauszählungen der beiden Merkmale getrennt voneinander herausbekommen lassen, sind allgemein deskriptiver Natur. So lässt sich ablesen, dass insgesamt fünf Studierende eine Abendnutzung bejahen, während vier Befragte dies verneinen. In der untersten Zeile wiederum können wir ablesen, dass vier BA-Studierende bzw. fünf MA-Studierende befragt wurden und wir insgesamt über neun Beobachtungseinheiten verfügen.

Im Beispiel wurden die absoluten Häufigkeiten der Merkmalskombinationen eingetragen. In Kreuztabellen lassen sich diese auch um relative Häufigkeiten ergänzen bzw. um eine prozentuale Anschauung des gemeinsamen Auftretens beider Variablen. Da nicht immer davon ausgegangen werden kann, dass sich die Beobachtungseinheiten in allen Zellen ähnlich sind, lassen sich die einzelnen Zellen durch die Ergänzung von prozentualen Häufigkeiten besser miteinander vergleichen. Die Prozentuierung innerhalb einer Kreuztabelle kann dabei aus drei unterschiedlichen Perspektiven erfolgen und wiedergegeben werden. Jede dieser drei Möglichkeiten kann unterschiedliche Fragen beantworten. Die folgenden drei Tabellen zeigen dies wiederum am Beispiel der abendlichen Bibliotheksnutzung, nun auf Basis der gesamten fiktiven Stichprobe von 100 befragten Studierenden.

Relative Häufigkeiten

Wenn wir, wie in Tabelle 20 dargestellt, die Zeilenprozent betrachten, dann können wir beispielsweise die Frage beantworten, wie viel Prozent der Abendnutzer MA-Studierende sind (77 %). Das heißt, wir berechnen die bedingten Anteile innerhalb der Variable „Nutzung“. Wie aus der rechten Spalte ersichtlich, addieren sich alle Zellen einer Zeile bei einer Zeilenprozentuierung auf 100 %.

Zeilenprozent

Tabelle 20: Abendliche Bibliotheksnutzung und Studiengang – Zeilenprozent (n = 100)

Nutzung \ Studiengang	Studiengang		
	BA	MA	Total
Ja	13 23%	43 77%	56 100%
Nein	17 39%	27 61%	44 100%
Total	30 30%	70 70%	100 100%

Quelle: Eigene Darstellung

Will man dagegen zum Beispiel wissen, wie viel Prozent der BA-Studierenden das Abendangebot nutzen, benötigen wir die Spaltenprozent, die uns den Anteil innerhalb des jeweiligen Studiengangs angeben. Aus Tabelle 21 lässt sich ablesen, dass 43 Prozent der Befragten im BA die Abendstunden nutzen, wohingegen dies bei 61 Prozent der MA-Studierenden der Fall ist.

Spaltenprozent

Tabelle 21: Abendliche Bibliotheksnutzung und Studiengang – Spaltenprozentage (n = 100)

Nutzung \ Studiengang	Studiengang		
	BA	MA	Total
Ja	13 43%	43 61%	56 56%
Nein	17 57%	27 39%	44 44%
Total	30 100%	70 100%	100 100%

Quelle: Eigene Darstellung

Gesamtprozente Eine letzte Möglichkeit der Prozentuierung zeigt Tabelle 22 mit der Angabe der sogenannten Gesamtprozente und besteht darin, die einzelnen absoluten Häufigkeiten pro Zelle auf die Gesamtzahl der Befragten zu beziehen und den bedingten Anteil der Zelle im Hinblick auf alle Beobachtungseinheiten zu berechnen. Damit kann beispielsweise Auskunft über die Frage gegeben werden, wie viel Prozent aller Befragten abends in die Bibliothek gehen und gleichzeitig im BA-Studium (13%) oder MA-Studium (43%) eingeschrieben sind (Tabelle 22).

Tabelle 22: Abendliche Bibliotheksnutzung und Studiengang – Gesamtprozente (n = 100)

Nutzung \ Studiengang	Studiengang		
	BA	MA	Total
Ja	13 13%	43 43%	56 56%
Nein	17 17%	27 27%	44 44%
Total	30 30%	70 70%	100 100%

Quelle: Eigene Darstellung

Da vor dem Hintergrund der unterschiedlichen Perspektiven der Prozentuierung die Interpretation einer Kreuztabelle nicht immer intuitiv zu erfassen ist, gibt es in der empirischen Sozialforschung eine sinnvolle Konvention bezüglich der Erstellung von Kreuztabellen. Diese ist insbesondere dann relevant, wenn vorab kausale Annahmen aufgestellt wurden, wir also den Einfluss einer (unabhängigen) Variable auf eine (abhängige) Variable untersuchen möchten. In einem solchen Fall wird konventionell in der Spalte die unabhängige Variable und in der Zeile die abhängige Variable abgebildet. Als Basis der Prozentuierung wird dann die unabhängige Variable gewählt, womit die Spaltenprozente ausgegeben und interpretiert werden müssen. Selbstverständlich würde ein umgekehrtes Vorgehen mit der abhängigen Variable in der Spalte, der unabhängigen Variable in der Zeile und der Angabe der Zeilenprozente zu denselben Ergebnissen führen. Dieses Vorgehen entspricht aber nicht der sozialwissenschaftlichen Konvention.

An einem weiteren Beispiel soll die (konventionelle) Erstellung und Interpretation einer Kreuztabelle deutlich gemacht werden. Auf der Basis von sozialwissenschaftlicher Forschungsliteratur und Sozialisationstheorien, lässt sich vermuten, dass Männer stärker an Politik interessiert sind als Frauen (Westle und Schoen 2002; van Deth 2004; aber: Westle 2009). Auf Basis des ALLBUS-Datensatzes 2016 lässt sich diese Hypothese mit einer bivariaten Kreuztabelle überprüfen.



In Tabelle 23 sind in den Spalten das Geschlecht und in den Zeilen das abgefragte politische Interesse auf einer fünfstufigen Ordinalskala von „sehr stark“ bis „überhaupt nicht“ zu finden. Diese Einteilung auf Zeilen und Spalten wurde entlang der Konvention und auf Basis der Hypothese „Das Geschlecht hat einen Einfluss auf das politische Interesse“ vorgenommen. Geschlecht stellt dabei die unabhängige Variable und das politische Interesse die abhängige Variable dar. In den Zeilen ist in den einzelnen Zellen nun oben die absolute Häufigkeit, also die Anzahl Männer bzw. Frauen, die sich für eine Ausprägung des politischen Interesses entschieden haben, abgetragen. Demnach haben 311 männliche Befragte und 116 weibliche Befragte geantwortet, dass sie über ein sehr starkes politisches Interesse verfügen. Da die Anzahl der Männer und Frauen im ALLBUS annähernd gleich ist, lassen sich in diesem Fall tatsächlich bereits die absoluten Häufigkeiten miteinander vergleichen. Offenbar interessieren sich wesentlich mehr Männer als Frauen sehr stark für Politik. Allerdings ist die absolute Häufigkeit für verschiedene zu vergleichende Gruppen in vielen Stichproben häufig nicht ähnlich groß, sodass man ausgehend von den absoluten Häufigkeiten in jedem Fall die prozentualen Häufigkeiten ermitteln sollte. Um aussagekräftige Werte zu erhalten, mit Hilfe derer die Hypothese überprüft werden kann, werden dabei die Spaltenprozentage, das sind die bedingten Anteilswerte für das jeweilige Geschlecht, berechnet, also die Prozentangaben, die sich auf die Spaltensummen in der untersten Zeile beziehen. Das heißt, es wird ermittelt, wie viel Prozent der Männer bzw. der Frauen jeweils die fünf verschiedenen Ausprägungen des politischen Interesses angegeben haben.

Tabelle 23: Politisches Interesse und Geschlecht (Spaltenprozentage)

Politisches Interesse \ Geschlecht	Geschlecht		Gesamt
	Männliche Befragte	Weibliche Befragte	
Sehr stark	311 17,6%	116 6,7%	427 12,2%
Stark	537 30,3%	345 20,1%	882 25,3%
Mittel	634 35,8%	795 46,2%	1429 40,9%
Wenig	207 11,7%	349 20,3%	556 15,9%
Überhaupt nicht	81 4,6%	115 6,7%	196 5,6%
Gesamt	1770 100,0%	1720 100,0%	3490 100,0%

Daten: ALLBUS 2016. Eigene Berechnungen

Um die vorab formulierte Hypothese zu überprüfen, werden nun die jeweiligen Spaltenprozentente miteinander verglichen. Auf dieser Basis lässt sich bereits sehen, dass sich die Hypothese eines Zusammenhangs erhärtet. Während beispielsweise fast 18 Prozent der Männer angeben, „sehr stark“ an Politik interessiert zu sein, behaupten das nur fast 7 Prozent der weiblichen Befragten von sich selbst, was eine sogenannte Prozentsatzdifferenz von über 10 Prozentpunkten ausmacht. Am anderen Ende der Interessensskala dagegen findet sich ein höherer Prozentsatz an weiblichen als an männlichen Befragten, die antworten „wenig“ oder „überhaupt nicht“ an Politik interessiert zu sein. Damit ist bei einer konventionell erstellten Kreuztabelle der zeilenweise Vergleich zwischen den Spaltenprozenten der unabhängigen Variablen ein einfacher Weg, um einen Anhaltspunkt bezüglich möglicher Zusammenhänge zwischen zwei Merkmalen zu erhalten.

Prozente und Prozentpunkte

Prozente machen immer relative Angaben, drücken also ein Verhältnis aus. Die Zahl vor einem Prozentzeichen (%) wird als Prozentsatz bezeichnet, der vom Verhältnis unabhängig, also absolut zu verstehen ist. Vergleicht man Prozentsätze miteinander, werden keine relativen Prozentangaben verwendet, da dies zu Missverständnissen führen kann. Stattdessen kommt der Begriff „Prozentpunkte“ zum Einsatz, der den absoluten Unterschied zwischen zwei Prozentsätzen angibt. Das heißt, in Bezug auf obiges Beispiel muss es heißen, dass der Anteil an Männern unter den stark an Politik Interessierten um knapp 10 Prozentpunkte höher ist als der Anteil an Frauen in dieser Kategorie des Politikinteresses. Er ist aber nicht um 10 Prozent höher, denn das würde einem Prozentsatz der Männer von knapp über 22 Prozent entsprechen (20,1 plus 10% von 20,1).

Bei der Berechnung der sogenannten Prozentsatzdifferenz wird die zeilenweise Differenz von zwei Spaltenprozenten berechnet. Damit kann sie theoretisch Werte zwischen -100 und $+100$ Prozentpunkten annehmen. Unterscheiden sich die Spaltenprozente in einer Zeile (zwischen den Ausprägungen einer unabhängigen Variable) nicht (weisen sie also den Wert 0 auf) oder nur kaum, lässt sich davon ausgehen, dass kein Zusammenhang besteht. Als Anhaltspunkt für die Interpretation von festgestellten Prozentsatzdifferenzen gilt (Kühnel und Krebs 2007, S. 319):

- Differenzen unter 5 Prozentpunkten sind kaum interpretierbar und weisen auf keinen Zusammenhang hin,
- Differenzen ab 5 bis unter 10 Prozentpunkten weisen auf einen schwachen Zusammenhang hin,
- Differenzen ab 10 bis unter 25 Prozentpunkten weisen auf einen mittelstarken Zusammenhang hin,
- Differenzen von 25 und mehr Prozentpunkten weisen auf einen starken Zusammenhang hin.

Bei Anwendung dieser Faustregeln sollte auf die Anzahl der Beobachtungseinheiten in den einzelnen Zellen geachtet werden. Mindestens 15 Fälle pro Zelle werden benötigt, um eine robuste Aussage treffen zu können.

Das Beispiel aus Tabelle 23 aufgreifend, lässt sich zusätzlich interpretieren, dass es auf Basis der ALLBUS-Stichprobe offenbar einen mittelstarken Zusammenhang zwischen dem Geschlecht und

dem (sehr) starken Interesse an Politik gibt. Bezüglich des fiktiven Bibliotheksbeispiels lässt sich auf Basis von Tabelle 21 zeigen, dass der Studiengang einen mittelstarken Zusammenhang mit der Nutzung des Abendangebots der UB aufweist.

Sobald man über eine kausale Hypothese verfügt, das heißt, sobald man eine Hypothese bezüglich einer angenommenen Zusammenhangsstruktur der beiden betrachteten Merkmale aufgestellt hat, sollte die Kreuztabelle konventionell erstellt werden. Zwei letzte Hinweise sollen in diesem Zusammenhang jedoch gegeben werden:

- Zum Zweck einer allgemeinen deskriptiven Darstellung zweier Variablen, zwischen denen kein Zusammenhang vermutet wird (z.B. im ALLBUS das Geschlecht der Befragten pro Bundesland) kann es dagegen Sinn machen, auch weitere bedingte Anteile in Form von Zeilen- und Gesamtprozentuierung in die Kreuztabelle mit aufzunehmen. In einem solchen Fall sollte den Lesern aber genau beschrieben werden, welche Aussagen man damit treffen möchte und die Tabelle muss vollständig und verständlich interpretiert werden.
- Der zweite Punkt betrifft den Fall, dass sich die Forscherin noch in einer explorativen Datensichtungsphase befindet und noch keine Hypothesen erstellt wurden. Dabei sind natürlich verschiedene Anzeigen auch jenseits der Konvention für den „eigenen Gebrauch“ in Ordnung. Allerdings sollte man sich bewusst sein, dass es das Lesen und Verstehen einer Kreuztabelle deutlich erleichtert, wenn zum einen konventionell gearbeitet wird und zum anderen die Leserinnen beim Verständnis der Kreuztabelle an die Hand genommen werden.

An einem weiteren Beispiel soll dies verdeutlicht werden: Die soziologische Bildungsforschung hat wiederholt darauf hingewiesen, dass in Deutschland der Bildungserfolg vom sozialen Status und damit auch der Bildung der Eltern der Herkunftsfamilie abhängig ist (Becker 2017). Insbesondere Personen, die einen elterlichen Migrationshintergrund aufweisen, waren und sind in der Bildungsteilnahme benachteiligt (z.B. Stanat 2003; Dollmann 2017). Das folgende (fiktive) Beispiel möchte die beiden Merkmale daher gemeinsam betrachten und kreuztabelliert nun eine dichotome Variable „elterlicher Bildungshintergrund“ mit „Schulabschlüssen der Kinder“, um folgende Hypothese zu überprüfen: „Kinder, deren Eltern über einen niedrigen Bildungshintergrund verfügen, haben tendenziell niedrigere Bildungsabschlüsse als Kinder deren Eltern einen hohen Bildungshintergrund aufweisen“. Tabelle 24 ist konventionell erstellt: Die abhängige Variable „Schulabschluss“ steht in den Zeilen, die unabhängige Variable „Elterlicher Bildungshintergrund“ steht in den Spalten. Damit können nun die Spaltenprozente miteinander verglichen werden.



Tabelle 24: Schulabschluss und elterlicher Bildungshintergrund (Spaltenprozent)

Bildungshintergrund der Eltern Schulabschluss der Kinder	Bildungshintergrund der Eltern		
	Niedrig	Hoch	Gesamt
Kein Abschluss	25 13,0%	34 1,5%	59 2,3%
Hauptschulabschluss	59 30,6%	523 22,5%	582 23,1%
Mittlere Reife	54 28,0%	780 33,5%	834 33,1%
(Fach-)Hochschulreife	55 28,5%	990 42,5%	1045 41,5%
Gesamt	193 100,0%	2327 100,0%	2520 100,0%

Quelle: Eigene Darstellung

Es zeigt sich, dass sich insbesondere in den Kategorien „Kein Abschluss“ sowie „(Fach-)Hochschulreife“ die Spaltenprozente deutlich voneinander unterscheiden. So haben rund 13 Prozent der Befragten mit niedrigem elterlichen Bildungshintergrund keinen Schulabschluss, während dies auf Befragte mit einem hohen elterlichen Bildungshintergrund für knapp 1,5 Prozent zutrifft. Beim höchsten Bildungsabschluss, der (Fach-)Hochschulreife, lässt sich dagegen eine Prozentsatzdifferenz von knapp 14 Prozentpunkten zugunsten der Befragten ohne Migrationshintergrund feststellen. Auf Grundlage dieses Vergleichs lässt sich die Hypothese, dass ein niedriger elterlicher Bildungshintergrund tendenziell mit einem geringeren Bildungserfolg der Kinder zusammenhängt, (vorläufig) bestätigen. Wir konstatieren einen mittleren Zusammenhang.

! Achtet man nicht darauf, Spaltenprozente zu berechnen und zu vergleichen, sondern berechnet für die Kreuztabelle die Zeilenprozente, zeigt sich in Tabelle 25 ein anderes Bild. Wer nun die Angaben zwischen den Spalten in einer Zeile miteinander vergleicht und interpretiert, begeht möglicherweise einen Fehler, denn mit dieser Art der Darstellung kann beispielsweise lediglich eine Aussage darüber getroffen werden, wie viel Prozent der Befragten mit Hauptschulabschluss über einen niedrigen elterlichen Bildungshintergrund verfügen (10,1%). Auf keinen Fall können Schlussfolgerungen getroffen werden wie beispielsweise, dass 57,6 Prozent der Befragten mit hohem elterlichen Bildungshintergrund ohne Migrationshintergrund keinen Schulabschluss haben und somit knapp 15 Prozentpunkte mehr als bei Befragten, deren Eltern über einen niedrigen Bildungsabschluss verfügen.

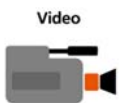
Tabelle 25: Schulabschluss und elterlicher Bildungshintergrund (Zeilenprozente)

Bildungshintergrund der Eltern Schulabschluss der Kinder	Bildungshintergrund der Eltern		
	Niedrig	Hoch	Gesamt
Kein Abschluss	25 42,4%	34 57,6%	59 100%
Hauptschulabschluss	59 10,1%	523 89,9%	582 100%
Mittlere Reife	54 6,5%	780 93,5%	834 100%
(Fach-)Hochschulreife	55 5,3%	990 94,7%	1045 100%
Gesamt	193 7,7%	2327 92,3%	2520 100%

Quelle: Eigene Darstellung

3.2 Zusammenhangsmaße für nominale Merkmale

Mit der im letzten Abschnitt im Rahmen von konventionell gestalteten Kreuztabellen kennengelernten Prozentsatzdifferenz lassen sich bereits erste Schlüsse auf mögliche Zusammenhänge zwischen zwei Variablen ziehen. Diese lassen sich mit dem Vergleich der sogenannten erwarteten und beobachteten Häufigkeiten weiter erhärten. Auf einem Abgleich zwischen erwarteten und beobachteten Häufigkeiten basiert auch das Assoziationsmaß Chi-Quadrat (Benninghaus 2007, S. 104-121).



Die absoluten Häufigkeiten, die in den einzelnen Zellen abgebildet sind, stellen die sogenannten beobachteten Häufigkeiten dar. Tabelle 26 zeigt die beobachteten Häufigkeiten der Merkmale „Geschlecht“ und „Politisches Interesse“ in einer Kontingenztafel (bzw. einer Kreuztabelle ohne prozentuale Anteile).

**Beobachtete
Häufigkeiten**

Wenn es keinen Zusammenhang zwischen den beiden Variablen gibt, dann lässt sich erwarten, dass diese beobachteten Häufigkeiten auch den erwarteten Häufigkeiten entsprechen. Dies lässt sich im Kontext der Kreuztabelle mit Hilfe einer sogenannten Indifferenztabelle darstellen. Eine Indifferenztabelle stellt die kombinierte Verteilung zweier Variablen dar, die erwartet wird, wenn es zwischen den beiden Merkmalen keinen Zusammenhang oder Unterschied gibt – das heißt, wenn zwischen den beiden betrachteten Variablen statistische Unabhängigkeit vorliegt. Um das Beispiel von Tabelle 26 aufzugreifen: Wenn es keinen Zusammenhang zwischen Geschlecht und politischem Interesse gibt, dann lässt sich erwarten, dass die Anzahl von Männern und Frauen gemäß ihres Anteils an der gesamten Stichprobe in jeder Kategorie des politischen Interesses mit den beobachteten Häufigkeiten deckungsgleich ist.

Tabelle 26: Politisches Interesse und Geschlecht (beobachtete Häufigkeiten) – Kontingenztafel

Politisches Interesse \ Geschlecht	Geschlecht	
	Männliche Befragte	Weibliche Befragte
Sehr stark	311	116
Stark	537	345
Mittel	634	795
Wenig	207	349
Überhaupt nicht	81	115

Daten: ALLBUS 2016 (n=3490). Eigene Berechnungen

Erwartete Häufigkeiten

Die erwarteten Häufigkeiten können unter Einbezug der Randhäufigkeiten ermittelt werden. Es werden die Zeilensumme und die Spaltensumme der entsprechenden Zelle miteinander multipliziert und dann durch den Umfang aller Beobachtungen geteilt. Am bereits erwähnten Beispiel des gemeinsamen Auftretens der Merkmale Geschlecht und politisches Interesse werden in Tabelle 27 für zwei Zellen beispielhaft die erwarteten Häufigkeiten berechnet. Mit gängigen Statistikprogrammen können die erwarteten Häufigkeiten auch mit ausgegeben werden. Im ersten Schritt werden die Randhäufigkeiten der Kontingenztafel berechnet. So erhält man durch Addieren der Zellenhäufigkeiten der ersten Spalte die Spaltensumme der ersten Spalte (siehe Tabelle 27).

Tabelle 27: Berechnung der erwarteten Häufigkeiten

Politisches Interesse \ Geschlecht	Geschlecht		Gesamt
	Männliche Befragte	Weibliche Befragte	
Sehr stark	$1770 * 427 / 3490 = 216,56$		427
Stark			882
Mittel		$1720 * 1429 / 3490 = 704,26$	1429
Wenig			556
Überhaupt nicht			196
Gesamt	1770	1720	3490

Daten: ALLBUS 2016 (n = 3490). Eigene Berechnungen

Dies ist die Anzahl der männlichen Befragten. Summiert man die entsprechenden Zellenhäufigkeiten in den Zeilen, so erhält man die Zeilensummen. Insgesamt 1429 Befragte haben beispielsweise angegeben, über ein mittleres politisches Interesse zu verfügen. In zwei Zellen von Tabelle 27 ist nun exemplarisch dargestellt, wie die erwarteten Häufigkeiten konkret berechnet werden. So würde man – wenn die Merkmale „Geschlecht“ und „Politisches Interesse“ statistisch unabhängig voneinander sind – aufgrund der beobachteten Randhäufigkeiten der Ausprägungen „männlich“ und „sehr stark“ erwarten, dass 216,56 männliche Befragte über ein sehr starkes politisches Interesse verfügen (1770 multipliziert mit 427 und anschließend dividiert durch die Fallzahl der Stichprobe 3490). 704,26 weibliche Befragte, die mittel an Politik interessiert sind, würden auf Basis des vorgestellten Rechenwegs erwartet (1720 multipliziert mit 1429 und dividiert durch 3490). Für die weiteren Zellen werden auf diesem Weg ebenfalls die erwarteten Häufigkeiten berechnet und in Tabelle 28 eingetragen. Gibt man in einer Kreuztabelle die berechnete erwartete Anzahl an, spricht man auch von einer sogenannten Indifferenztabelle.

Tabelle 28: Politisches Interesse und Geschlecht (erwartete Häufigkeiten) – Indifferenztabelle

Geschlecht Politisches Interesse	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	216,56	210,44	427
Stark	447,32	434,68	882
Mittel	724,74	704,26	1429
Wenig	281,98	274,02	556
Überhaupt nicht	99,40	96,60	196
Gesamt	1770	1720	3490

Daten: ALLBUS 2016 (n = 3490). Eigene Berechnungen

Vergleicht man nun beobachtete mit erwarteten Häufigkeiten, also Tabelle 26 mit Tabelle 28, zeigt sich, dass beim starken und sehr starken politischen Interesse die erwartete Anzahl bei den männlichen Befragten deutlich unter der tatsächlich beobachteten Anzahl liegt, während das bei den Frauen umgekehrt ist. Bei wenigem und gar keinem politischen Interesse dagegen zeigt sich das umgekehrte Bild. Hier ist die beobachtete Anzahl bei den weiblichen Befragten höher als die erwartete Anzahl. Das heißt, basierend auf einem Abgleich der beobachteten mit den erwarteten Häufigkeiten können wir ebenfalls schlussfolgern, dass keine statistische Unabhängigkeit zwischen den Variablen „Geschlecht“ und „Politisches Interesse“ besteht und Frauen sich offensichtlich weniger für Politik interessieren als dies Männer tun.

Berechnung von Chi-Quadrat

Auf Basis des Vergleichs zwischen erwarteten und beobachteten Häufigkeiten wird das sogenannte Chi-Quadrat-Assoziationsmaß berechnet. Zunächst werden von den beobachteten Häufigkeiten die erwarteten Häufigkeiten subtrahiert, man erhält die sogenannten Residuen (Kuckartz et al. 2013, S. 265). Diese Differenzen werden dann quadriert. Dadurch erhalten alle Werte ein positives Vorzeichen. Anschließend wird durch die erwarteten Häufigkeiten dividiert. Chi-Quadrat ist die Summe dieser Quotienten. Formal ergibt sich:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(f_{b(ij)} - f_{e(ij)})^2}{f_{e(ij)}}$$

$f_{b(ij)}$ = beobachtete Häufigkeit in der i-ten Zeile und j-ten Spalte

$f_{e(ij)}$ = erwartete Häufigkeit in der i-ten Zeile und j-ten Spalte

k = Anzahl der Zeilen

m = Anzahl der Spalten

Für die Berechnung von Chi-Quadrat bietet es sich an, eine Arbeitstabelle zu erstellen. In Tabelle 29 sind die einzelnen Schritte dargestellt. Zunächst werden die Residuen berechnet – dabei werden die erwarteten von den beobachteten Häufigkeiten abgezogen. Im Anschluss werden für jede Zelle die Residuen quadriert.

Tabelle 29: Arbeitstabelle zur Berechnung von Chi-Quadrat

f_b	f_e	$f_b - f_e$	$(f_b - f_e)^2$	$\frac{(f_b - f_e)^2}{f_e}$
311	216,56	94,44	8918,91	41,18
537	447,32	89,68	8042,50	17,98
634	724,74	-90,74	8233,75	11,36
207	281,98	-74,98	5622,00	19,94
81	99,40	-18,40	338,56	3,41
116	210,44	-94,44	8918,91	42,38
345	434,68	-89,68	8042,50	18,50
795	704,26	90,74	8233,75	11,69
349	274,02	74,98	5622,00	20,52
115	96,60	18,40	338,56	3,50
Chi-Quadrat				190,46

Quelle: Eigene Darstellung

In der letzten Spalte von Tabelle 29 sind der vorletzte und der letzte Arbeitsschritt zu sehen. Zunächst werden die quadrierten Differenzen durch die erwarteten Häufigkeiten geteilt. Danach werden die Quotienten aufsummiert. Diese Summe entspricht dem gesuchten Chi-Quadrat-Wert, der von der Anzahl der Beobachtungseinheiten und der Dimension, also „Größe“ bzw. Anzahl der Zellen, der Kreuztabelle beeinflusst wird.

Nimmt Chi-Quadrat den Wert Null an, besteht kein Unterschied zu den erwarteten Häufigkeiten. Mit ansteigendem Wert nimmt auch die empirische Abhängigkeit der beiden Merkmale zu. Der im Beispiel berechnete Wert von 190,46 unterscheidet sich deutlich von 0 und weist ebenfalls darauf hin, dass keine statistische Unabhängigkeit der beiden Merkmale gegeben ist.

Allerdings hat Chi-Quadrat den Nachteil, dass es von der Fallzahl, also der Größe der Stichprobe, abhängig ist. Daher wurden auf Basis der Berechnung von Chi-Quadrat weitere Assoziationsmaße entwickelt, die auch vergleichend eingesetzt werden können. Zwei Assoziationsmaße sind wichtig: phi und darauf aufbauend Cramer's V.

Assoziationsmaße: Phi und Cramer's V

Für zwei dichotome Merkmale, also 2x2-Tabellen (Kreuztabellen, die über zwei Zeilen und zwei Spalten verfügen), wird häufig der Phi-Koeffizient genutzt. Hier wird der Chi-Quadrat-Wert durch die Anzahl der Beobachtungen geteilt und danach die Wurzel gezogen:

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

2x2-Tabellen bzw. Zusammenhänge zwischen zwei dichotomen Merkmalen stellen einen Sonderfall dar. Eine mathematische Weiterentwicklung für Kreuztabellen, die mehr als vier Zellen umfassen, ist Cramer's V. Dieser Assoziationskoeffizient für zwei diskrete und nominalskalierte Variablen hat, wie Phi auch, zudem den Vorteil, dass er Werte zwischen 0 und 1 annimmt und daher gut interpretierbar ist.⁴ Sowohl der Wert von Cramer's V als auch der Wert des Phi-Koeffizienten können damit vergleichend interpretiert werden (z.B. Kühnel und Krebs 2007, S. 355-356). Der Wert 0 bedeutet dabei empirische Unabhängigkeit der beiden Variablen, während 1 die vollständige Abhängigkeit der Merkmale abbildet. Um Cramer's V zu berechnen, muss Chi-Quadrat dividiert werden durch das Produkt aus dem Umfang der Beobachtungen und der kleinsten Abmessung der Kontingenztafel. Aus diesem Quotienten wird schließlich die Wurzel gezogen, um Cramer's V zu erhalten (Kühnel und Krebs 2007, S. 329-335):

$$V = \sqrt{\frac{\chi^2}{n * (M - 1)}}$$

Die Interpretation von Cramer's V berücksichtigt (in der Regel) nur die Stärke des Zusammenhangs, nicht aber dessen Richtung. Da es sich um nominale Variablen handelt, ist es allerdings auch nicht möglich, eine Aussage über die Richtung des Zusammenhangs zu treffen, also inwiefern es sich um einen positiven oder negativen Zusammenhang zwischen den beiden Merkmalen handelt. Bei nominalskalierten Merkmalen gibt es keine Rangfolge (wie bei einer ordinalen Skala) oder interpretierbare Abstände zwischen Werten (wie bei einer metrischen Skala). Damit muss ein Ansteigen des Werts einer Variablen nicht zwingend eine inhaltliche Steigerung bedeuten, wie die Beispiele „Geschlecht“ oder „Familienstand“ zeigen.

⁴ Für die Berechnung des Phi-Koeffizienten findet sich in der Literatur eine alternative Berechnungsformel, bei der Phi auch negative Werte annehmen kann. Die Interpretation des Vorzeichens als positive oder negative Beziehung setzt allerdings ordinale Variablen voraus (Kühnel und Krebs 2007, S. 336-338; Mittag 2017, S. 123-124).

Zusammenhänge zwischen nominalen Merkmalen werden als Wenn-dann-Hypothesen formuliert. Beispielsweise würde man formulieren: „Wenn jemand im Master-Studium eingeschrieben ist, dann nutzt er das Abendangebot der Bibliothek.“ Daher empfiehlt es sich auch immer, ergänzend zur Berechnung von Cramer's V eine Kreuztabelle zu erstellen und die beobachteten und erwarteten Häufigkeiten in Augenschein zu nehmen, um die „Substanz“, die Art oder Qualität des Zusammenhangs erfassen zu können.

Interpretation des Zusammenhangs

Es gibt Faustregeln zur Interpretation der Stärke des Zusammenhangs vom Phi-Koeffizienten und Cramer's V, die in Tabelle 30 vorgestellt werden (z.B. Gehring und Weins 2009, S. 152; Kühnel und Krebs 2007, S. 356):

Tabelle 30: Interpretation von Cramer's V

Wert von Cramer's V (V) bzw. Betrag von Phi ($ \Phi $)	Interpretation
$\leq 0,05$	kein Zusammenhang
$> 0,05$ bis $\leq 0,10$	sehr schwacher Zusammenhang
$> 0,10$ bis $\leq 0,20$	schwacher Zusammenhang
$> 0,20$ bis $\leq 0,40$	mittelstarker Zusammenhang
$> 0,40$ bis $\leq 0,60$	starker Zusammenhang ⁵
$> 0,60$	sehr starker Zusammenhang

Quelle: Eigene Darstellung

Generell ist Cramer's V ein Zusammenhangsmaß für zwei nominalskalierte Merkmale (z.B. Familienstand und Wahlabsicht). Allerdings bietet sich Cramer's V auch an, um den Zusammenhang zwischen einem nominalskalierten Merkmal und einem ordinalskalierten Merkmal (z.B. Geschlecht und formale Bildung) bzw. einem metrischskalierten Merkmal⁶ (z.B. Geschlecht und Einkommen) zu untersuchen, da bei unterschiedlichen Messniveaus ein Zusammenhangsmaß gewählt werden muss, welches für das niedrigere Messniveau geeignet ist (Weins 2010).

Am bereits bekannten Beispiel von Geschlecht und politischem Interesse, in dem das politische Interesse ordinalskaliert ist, soll nun die Berechnung von Cramer's V vorgestellt werden. Wie beim Chi-Quadrat-Wert müssen auch hier zunächst die Randhäufigkeiten und die erwarteten Häufigkeiten ermittelt werden. Das heißt, die Arbeitsschritte bis zur Berechnung des Chi-Quadrat-Werts sind identisch. Ist Chi-Quadrat berechnet, muss lediglich der Nenner in der Formel zur Berechnung von Cramer's V bestimmt werden. Dabei entspricht n dem Umfang der Beobachtungen bzw. Fälle, welcher multipliziert wird mit M minus 1. M bezeichnet dabei die kleinere Abmessung der Kontingenztafel, welche im vorliegenden Beispiel aus 5 Zeilen und 2 Spalten besteht. M ist also 2.

⁵ Kuckartz et al. (2013, S. 93) sprechen allerdings erst ab einem Cramer's V von größer oder gleich 0,7 von einem starken Zusammenhang.

⁶ Für die Analyse des Zusammenhangs zwischen einem nominal- und einem metrischskalierten Merkmal eignet sich auch Eta (siehe Abschnitt 3.5).

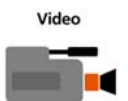
Mit einem ermittelten Chi-Quadrat von 190,46, einem n von 3490 und einem M von 2 ergibt sich eingesetzt in obige Formel für Cramer's V :

$$V = \sqrt{\frac{190,46}{3490 * (2 - 1)}} = 0,23$$

Es ergibt sich somit ein Cramer's V von 0,23, was einem mittelstarken Zusammenhang entspricht.

3.3 Zusammenhangsmaße für ordinale Merkmale

Liegen zwei ordinalskalierte Variablen vor, wie beispielsweise das politische Interesse auf einer Ordinalskala sowie der formale Bildungsabschluss, dann lassen sich ebenfalls Zusammenhänge illustrieren und berechnen. Je nach Anzahl der Ausprägungen der Merkmale bietet sich nach wie vor eine Kreuztabelle zur deskriptiven Darstellung an. Im Gegensatz zu Cramer's V , welches keine Aussage zur Richtung des Zusammenhangs geben kann, lässt sich mit Spearman's Rho allerdings auf ein Maß zurückgreifen, welches positive und negative Korrelationen bestimmen sowie die Stärke des Zusammenhangs angeben kann.



Der Rangkorrelationskoeffizient Spearman's Rho ist ein normiertes Maß für die Bestimmung eines Zusammenhangs zwischen zwei (mindestens) ordinalskalierten Merkmalen (Benninghaus 2007, S. 177-184). Die Voraussetzung des ordinalen Messniveaus ist wichtig, da der Rangkorrelationskoeffizient darauf basiert, dass die Merkmalsausprägungen zweier Merkmale in jeweils geordnete Rangfolgen gebracht werden. Anschließend werden nicht ihre einzelnen Merkmalsausprägungen paarweise zueinander in Beziehung gesetzt, sondern die jeweiligen Rangpositionen der gepaarten Merkmalsausprägungen. Die (vereinfachte) Formel zur Berechnung von Spearman's Rho lautet wie folgt:

Spearman's Rho

$$r_{SP} = 1 - \frac{6 * \sum_{i=1}^n d_i^2}{n * (n^2 - 1)} \text{ wobei } d_i = rg(x_i) - rg(y_i)$$

Dabei kann der Rangkorrelationskoeffizient nach Spearman Werte zwischen -1 und $+1$ annehmen. Der Wert 0 zeigt dabei keinen Zusammenhang an und ein positiver Koeffizient weist auf einen gleichgerichteten bzw. positiven Zusammenhang hin. Betrachtet man beispielsweise die beiden ordinalen Merkmale „Schulbildung“ und „Politisches Interesse“, dann würde man einen positiven Rangkorrelationskoeffizienten wie folgt interpretieren: Je höher der formale Bildungsabschluss einer Person ist, desto höher ist auch ihr politisches Interesse. Ein negatives Vorzeichen weist dagegen auf einen gegenläufigen bzw. negativen Zusammenhang zwischen zwei Merkmalen hin. Mit steigendem Bildungsgrad sinkt das politische Interesse, wäre dann die beispielbezogene Interpretation. Der Wert -1 steht für einen perfekten negativen Zusammenhang und $+1$ für einen perfekten positiven Zusammenhang.

Auch bei diesem Assoziationsmaß stehen wiederum Faustregeln für die Interpretation der Stärke des Zusammenhangs zur Verfügung, die vom Betrag des Rangkorrelationskoeffizienten abhängen. Tabelle 31 enthält Angaben für die Interpretation der Stärke des Zusammenhangs (Kühnel und Krebs 2007, S. 404-405; Kuckartz et al. 2013, S. 213).

Tabelle 31: Interpretation von Spearman’s Rho

Rangkorrelationskoeffizient ($ r_{SP} $)	Interpretation
$\leq 0,05$	kein Zusammenhang
$> 0,05$ bis $\leq 0,20$	schwacher Zusammenhang
$> 0,20$ bis $\leq 0,50$	mittelstarker Zusammenhang
$> 0,50$ bis $\leq 0,70$	starker Zusammenhang
$> 0,70$	sehr starker Zusammenhang

Quelle: Eigene Darstellung

An einem Beispiel wird die einfache Berechnung von Spearman’s Rho vorgestellt. Anschließend wird auf Probleme im Kontext großer Stichproben und Alternativen zum Rangkorrelationskoeffizienten nach Spearman verwiesen.

Für sozialwissenschaftliche Untersuchungen spielt das Konzept „Soziale Schicht“ eine wichtige Rolle. Ob politische Einstellungen, Gesundheitszustand oder kulturelle Präferenzen analysiert werden, der soziale Status bzw. die soziale Schicht trägt häufig einen wesentlichen Bestandteil zur Erklärung der verschiedenen Merkmale bei (Pollak 2016). Soziale Schicht wird daher auch in Bevölkerungsumfragen häufig erfasst und in vielen soziologischen und politikwissenschaftlichen Analysen verwendet.

Soziale Schicht

Der Begriff „soziale Schicht“ hat keine einheitliche und anerkannte Definition im sozialwissenschaftlichen Kontext. Konkurrierende bzw. ergänzende Begriffe sind „Klasse“ oder „soziale Milieus“. Als konsensual kann allerdings gelten, dass soziale Schichten gesellschaftliche Gruppen bezeichnen, die sich durch eine bestimmte Kombination von sozioökonomischen Merkmalen konstituieren. Pollak (2016) weist insbesondere auf den zentralen Stellenwert von berufsbezogenen Merkmalen (z.B. beruflicher Status, Einkommen und Arbeitsweise) sowie der Bildung hin. Auch Scheuch und Daheim (1961), die als eine der ersten deutschen Sozialwissenschaftler dieses Konzept auch empirisch erfasst haben, verstehen soziale Schicht als eine Kombination aus Bildung, Beruf und Verdienst. Den meisten sozialen Schichtkonzepten ist zudem gemeinsam, dass sie die unterschiedlichen Schichten vertikal anordnen. Die einfachste Differenzierung ist dabei sicherlich ein dreistufiges Modell aus Unter-, Mittel- und Oberschicht, welches durch feinere Abstufungen (z.B. obere Mittelschicht) noch ergänzt werden kann.

Bei der empirischen Erfassung der Schichtzugehörigkeit lassen sich Indikatoren unterscheiden, die einerseits die Schichtzugehörigkeit von Befragten objektiv zu erfassen suchen und die andererseits Personen selbst ihre Schichtzugehörigkeit subjektiv vornehmen lassen. Bei der objektiven Messung werden vor allem Kombinationen aus beruflichen Faktoren (z.B. Verdienst, Berufsstatus) und formaler Bildung genutzt. Die sogenannte subjektive Schichtzugehörigkeit wird meist mit einer Frage gemessen, die die Befragten bittet, anhand von unterschiedlichen Schichten eine Selbstzuweisung vorzunehmen.



Die soziale Schicht hat Auswirkungen auf verschiedene Aspekte des Lebensstils, politische Einstellungen, Konsumverhalten, Ernährung und Freizeitgestaltung. Dies lässt sich zum einen durch zur Verfügung stehende materielle Mittel erklären und zum anderen auch durch psychosoziale Faktoren, die mit der sozialen Schicht zusammenhängen (Richter und Hurrelmann 2007). Basierend auf der Hypothese, dass mit einer höheren sozialen Schicht auch ein besserer allgemeiner Gesundheitszustand einher geht (Lampert et al. 2016), wird im folgenden (fiktiven) Beispiel auf der Basis von fünf Befragten ein möglicher Zusammenhang zwischen den beiden ordinalskalierten Merkmalen „soziale Schicht“ und „Gesundheitszustand“ mit dem Rangkorrelationskoeffizienten Spearman's Rho untersucht. Dabei wurden die beiden Merkmale jeweils mit fünf Ausprägungen gemessen.⁷ Die soziale Schicht wurde mit der subjektiven Schichtzugehörigkeit erfasst, bei der sich die Befragten selbst einer der folgenden fünf Schichten zuordnen sollten: Unterschicht (1), Arbeiterschicht (2), Mittelschicht (3), obere Mittelschicht (4) oder Oberschicht (5). Der Gesundheitszustand wird ebenfalls durch einen fünfstufigen Indikator erfasst: Sehr schlecht (1), schlecht (2), zufriedenstellend (3), gut (4) oder sehr gut (5). Tabelle 32 zeigt die Merkmale für fünf Befragte. In Klammern stehen dabei die vergebenen Werte zur Erfassung, die in diesem Beispiel auch den Rangplätzen entsprechen.

Tabelle 32: Soziale Schicht und Gesundheitszustand

ID	Soziale Schicht	Gesundheitszustand
1	Mittelschicht (3)	Schlecht (2)
2	Arbeiterschicht (2)	Sehr schlecht (1)
3	Oberschicht (5)	Sehr gut (5)
4	Obere Mittelschicht (4)	Gut (4)
5	Unterschicht (1)	Zufriedenstellend (3)

Quelle: Eigene Darstellung

Diese Angaben werden nun im ersten Schritt zur Berechnung von Spearman's Rho in Rangpositionen überführt. Im vorliegenden Beispiel entsprechen die Ränge den x- bzw. y-Werten. Tabelle 33 zeigt die vergebenen Rangpositionen. Zur Berechnung des Rangkorrelationskoeffizienten nach der oben vorgestellten Formel müssen nun die Differenzen d_i zwischen den Rangpositionen $rg(x_i)$ und $rg(y_i)$ der beiden Variablen „Soziale Schicht“ (x-Variable) und „Gesundheitszustand“ (y-Variable) bestimmt werden.

⁷ Eine fünfstufige ordinale Erfassung der beiden Merkmale findet sich z.B. auch im ALLBUS 2016.

Tabelle 33: Arbeitstabelle zur Berechnung von Spearman's Rho

ID	Soziale Schicht (rg(x _i))	Gesundheitszustand (rg(y _i))	d _i = rg(x _i) – rg (y _i)
1	3	2	1
2	2	1	1
3	5	5	0
4	4	4	0
5	1	3	–2

Quelle: Eigene Darstellung

Eingesetzt in die Formel ergibt sich dann:

$$r_{SP} = 1 - \frac{6 * \sum_{i=1}^n d_i^2}{n * (n^2 - 1)} = 1 - \frac{6 * (1^2 + 1^2 + 0^2 + 0^2 + (-2)^2)}{5 * (5^2 - 1)} = 1 - \frac{6 * 6}{120} = 1 - 0,3 = 0,7$$

Das heißt, der Rangkorrelationskoeffizient Spearman's Rho zwischen sozialer Schicht und Gesundheitszustand liegt im (fiktiven) Beispiel bei 0,7. Dies entspricht einem starken Zusammenhang zwischen den beiden Merkmalen. Auf Basis der vorgenommenen Analyse lässt sich im Sinne der Hypothese schließen, dass je höher die Schichtzugehörigkeit ist, auch der eingeschätzte Gesundheitszustand desto besser ist und umgekehrt.

Einschränkungen und Alternativen

Im obigen Beispiel kam bei der Vergabe von Rängen gemäß den geordneten Merkmalsausprägungen jeder Rang nur einmal vor. In Fällen, in denen dies zutrifft, kann die oben vorgestellte vereinfachte Formel für Spearman's Rho angewendet werden. Insbesondere bei „echten“ Rangkorrelationen, wenn beispielsweise das Ranking von Ländern im Human Development Index mit dem Ranking auf einem Demokratieindex in Bezug gesetzt werden soll, funktioniert dies problemlos, da es sich um Rangplätze handelt, die nur einmal vergeben werden (Tausendpfund 2018b, S. 164). Bei größeren Stichproben, die auch mehr Befragte als Merkmalsausprägungen umfassen, kommt es stattdessen häufig vor, dass mindestens zwei oder mehr Fälle die gleiche Merkmalsausprägung aufweisen. Das heißt, hier werden Rangpositionen mit sogenannten Bindungen (engl. ties) vergeben. Um hier dennoch den Rangkorrelationskoeffizienten berechnen zu können, weisen Benninghaus (2007, S. 182) sowie Kuckartz et al. (2013, S. 218) auf ein pragmatisches Vorgehen hin: Den verbundenen Beobachtungseinheiten wird das arithmetische Mittel derjenigen Rangplätze zugewiesen, die ohne Bindungen vergeben worden wären. Um dies an einem konkreten Beispiel zu veranschaulichen: Hätten im obigen Beispiel sowohl der Befragte mit der ID 3 als auch der Befragte mit der ID 4 beim Gesundheitszustand „sehr gut“ angegeben, dann wäre beiden der Rangplatz 4,5 zugewiesen worden.

Übersteigt die Anzahl der verknüpften Ränge allerdings den Anteil von 20 % aller Ränge, muss Spearman's Rho mit der folgenden komplexeren Formel berechnet werden (Mittag 2017, S. 133):

$$r_{SP} = \frac{\sum_{i=1}^n (rg(x_i) - \overline{rg_x}) * (rg(y_i) - \overline{rg_y})}{\sqrt{\sum_{i=1}^n (rg(x_i) - \overline{rg_x})^2} * \sqrt{\sum_{i=1}^n (rg(y_i) - \overline{rg_y})^2}}$$

Ein alternatives Assoziationsmaß, welches zur Berechnung von Korrelationen zwischen zwei ordinalskalierten Merkmalen zur Verfügung steht, ist Kendall's Tau (Kühnel und Krebs 2007, S. 374-375). Dieses basiert auf dem Vergleich zwischen Paaren von Merkmalsausprägungen: Weisen beide Merkmalsausprägungen eines Paares die gleiche Rangreihenfolge auf, dann spricht man von einem konkordanten Paar. Sind die Merkmalsausprägungen gegenläufig in ihren Rangreihenfolgen, dann wird das Paar als diskordant bezeichnet. Als „verbunden“ wird ein Paar dann bezeichnet, wenn beide Beobachtungseinheiten auf einem Merkmal unterschiedliche Werte, auf dem anderen Merkmal aber identische Werte aufweisen (Ludwig-Mayerhofer et al. 2014, S. 226-228). Kendall's Tau-Assoziationsmaße berechnen nun alle möglichen Paarkombinationen. Überwiegen konkordante Paare, zeugt dies von einem positiven Zusammenhang, finden sich mehr diskordante Paare, weist dies auf einen negativen Zusammenhang hin, während überwiegend verbundene Paare auf keinen Zusammenhang hinweisen. Die errechneten Maßzahlen umfassen ebenfalls einen Wertebereich zwischen -1 und $+1$ und können hinsichtlich Richtung und Stärke des Zusammenhangs entsprechend interpretiert werden.

3.4 Zusammenhangsmaße für metrische Merkmale

Liegen zwei Merkmale vor, die ein metrisches Skalenniveau aufweisen, bietet sich zunächst eine grafische Anschauung der Merkmalskombinationen an, um mögliche Zusammenhänge in Augenschein nehmen zu können. Diese Art der Darstellung wird Streudiagramm, Punktwolkendiagramm oder Scatterplot genannt. An einem (fiktiven) Beispiel wird dies illustriert. Es wird überprüft, ob der IQ einer Person mit dem Abschneiden bei einem Test für räumliches Denken zusammenhängt. Dazu wird bei 8 Probanden zunächst der IQ ermittelt. Danach müssen die Teilnehmenden noch den Test absolvieren. Im Test für räumliches Denken konnten maximal 110 Punkte erreicht werden. Es wird vermutet, dass je höher der IQ einer Person ist, desto mehr Punkte erzielt sie auch im Test für räumliches Denken. Tabelle 34 zeigt die erzielten Werte.



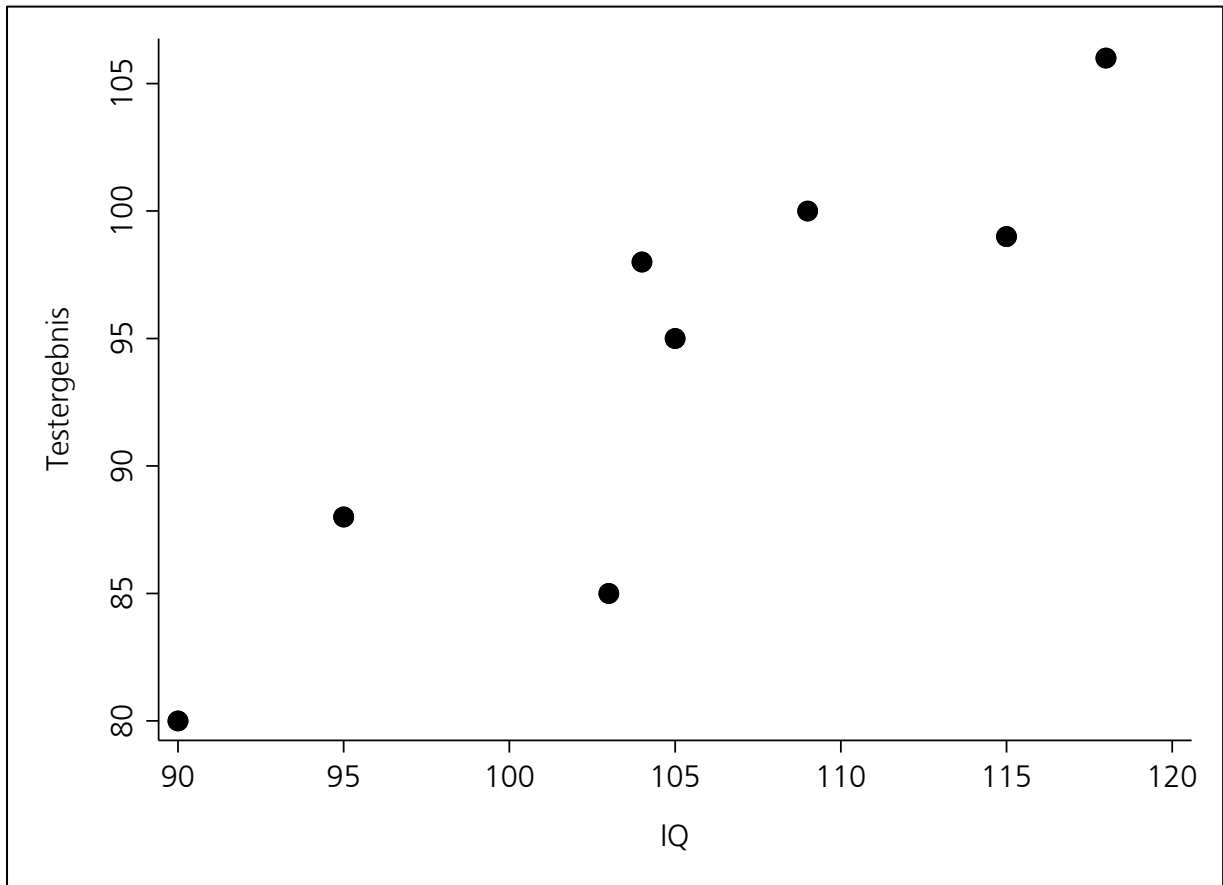
Tabelle 34: IQ und Testergebnis beim räumlichen Denken – Urliste

ID	IQ	Testergebnis
1	104	98
2	90	80
3	103	85
4	115	99
5	105	95
6	118	106
7	109	100
8	95	88

Quelle: Eigene Darstellung

Um zu erkennen, welche Art des Zusammenhangs zwischen zwei Merkmalen eines Datensatzes besteht, empfiehlt sich die Erstellung von Streudiagrammen. Anhand der Darstellung im Streudiagramm in Abbildung 13 können wir die Art des Zusammenhangs zwischen den beiden Merkmalen beschreiben. Höhere IQ-Werte gehen offensichtlich mit einer höheren Testpunktzahl einher.

Abbildung 13: IQ und Testergebnis beim räumlichen Denken – Streudiagramm



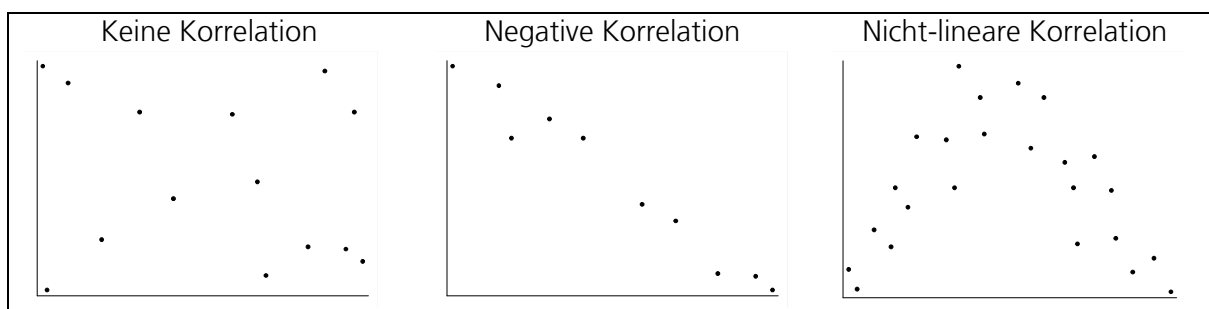
Quelle: Eigene Darstellung

Auf Basis unserer (fiktiven) Stichprobe lässt sich obige Vermutung, dass höhere IQ-Werte mit einem besseren Testergebnis einhergehen, bestätigen. Beide Variablen „korrelieren“ miteinander, das heißt sie stehen in einer Wechselbeziehung zueinander. Einen solchen „Je-mehr-desto-mehr“-Zusammenhang nennt man auch einen positiv linearen Zusammenhang bzw. eine positive lineare Korrelation. An dieser Stelle soll bereits darauf hingewiesen werden, dass eine Korrelation keine Aussage zur Kausalität trifft, sondern die Korrelation und die damit verbundenen Korrelationskoeffizienten nur messen, ob sich zwei Merkmale „im Gleichklang bewegen“ (Tausendpfund 2018b, S. 174-176).

Neben einer linearen Korrelation gibt es auch andere Arten des Zusammenhangs zwischen zwei (metrischen) Variablen, die in Abbildung 14 dargestellt werden. Das linke Streudiagramm in Abbildung 14 zeigt keine Beziehung zwischen zwei Variablen an; es liegt keine Korrelation vor. Die Datenpunkte ergeben kein Beziehungsmuster. Das mittlere Streudiagramm in Abbildung 14 zeigt eine negative lineare Korrelation. Je höher die Werte bei der einen Variable sind, desto niedriger

sind die Werte bei der anderen Variable bzw. umgekehrt. Als Beispiel könnte man sich den Zusammenhang zwischen dem Alter eines Autos in Jahren und seinem Wert in Euro vorstellen. Das heißt, je älter das Auto ist, desto geringer ist sein Wert (eine Ausnahme wären sicherlich Oldtimer). Wie auch bei der positiven linearen Korrelation in Abbildung 13 kann man sich eine Gerade vorstellen, die durch bzw. zwischen den einzelnen Datenpunkten durchgeht. Dass nicht alle Beziehungen zwischen zwei Merkmalen linear sind, zeigt das rechte Streudiagramm in Abbildung 14. Hier wird ein „umgekehrter u-förmiger“ Zusammenhang zwischen zwei Merkmalen dargestellt. Man könnte sich beispielsweise die Beziehung zwischen der Anspannung bei einer Klausur und der Leistungsfähigkeit vorstellen. Danach wäre die Leistung bei sehr geringer bzw. sehr hoher Anspannung schlechter als bei mittlerer Anspannung (Yerkes und Dodson 1908).

Abbildung 14: Weitere Arten des Zusammenhangs von zwei Merkmalen



Quelle: Eigene Darstellung

Werden sogenannte „Je-desto“-Hypothesen formuliert, werden lineare Korrelationen postuliert. Wie Abbildung 13 zeigt, hängen die beiden Merkmale linear positiv zusammen. Allerdings wird aus dem Streudiagramm nicht ersichtlich, wie stark der Zusammenhang genau ist. Dieser Zusammenhang lässt sich mit der sogenannten Bravais-Pearson-Produktkorrelation berechnen.

Dazu bietet sich als Zwischenschritt zunächst die Berechnung der Kovarianz an.

Kovarianz

Die Kovarianz ist ein Maß für den linearen Zusammenhang zweier Variablen x und y und zeigt die wechselseitige Varianz von zwei Merkmalen an. Diese ergibt sich, indem für jedes Wertepaar zunächst berechnet wird, wie weit der betreffende x -Wert bzw. der y -Wert vom jeweils zugehörigen arithmetischen Mittel entfernt ist (Kuckartz et al. 2013, S. 211). Das hat zur Folge, dass der Wert für die Kovarianz durch die Maßeinheiten der beiden Merkmale beeinflusst wird. Die Kovarianz (cov) gibt also die gemeinsame Streuung zweier Variablen x und y an. Sie stellt ein (unstandardisiertes) Maß zur Bestimmung eines linearen Zusammenhangs von zwei metrisch skalierten Merkmalen dar (Gehring und Weins 2009, S. 169-170). Analog zur Varianz eines einzelnen (metrischen) Merkmals, wird die (empirische) Kovarianz mit folgender Formel berechnet (Kuckartz et al. 2013, S. 211; Ludwig-Mayerhofer et al. 2014, S. 219)⁸:

⁸ Strenggenommen wird die empirische Kovarianz nur bei Daten aus einer Vollerhebung auf diese Weise berechnet. Weins (2010, S. 86) weist darauf hin, dass man, wenn die vorliegenden Daten keiner Vollerhebung, sondern einer Stichprobe entstammen, die (korrigierte) Kovarianz berechnen sollte, die im Zähler $n-1$ aufweist. Kuckartz et al. (2013, S. 211) gibt nur die korrigierte Kovarianzformel an.

$$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n}$$

Ein positives Vorzeichen des Kovarianzwerts gibt an, dass sich beide Variablen in dieselbe Richtung bewegen (das bedeutet, dass sich der Wert einer Variable erhöht, wenn der Wert der anderen Variable ansteigt). Ein negatives Vorzeichen sagt das Gegenteil über den Zusammenhang aus (das heißt, wenn der Wert einer Variable steigt, fällt der Wert der anderen). Ein Wert von Null oder nahe Null weist darauf hin, dass kein oder nur ein sehr geringer Zusammenhang besteht. Der Kovarianzwert steigt zwar mit der Stärke des Zusammenhangs, da der Wert selbst aber unstandardisiert bzw. abhängig von den Maßeinheiten der Merkmale ist, ist eine Interpretation der Stärke des Zusammenhangs schwierig.

Für das Beispiel IQ und Testergebnis wird in Tabelle 35 nun die Kovarianz händisch anhand der Formel berechnet. Dazu wird zunächst eine Arbeitstabelle erstellt.

Tabelle 35: Arbeitstabelle zur Berechnung der Kovarianz

ID	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	104	98	-0,88	4,12	-3,63
2	90	80	-14,88	-13,88	206,53
3	103	85	-1,88	-8,88	16,69
4	115	99	10,12	5,12	51,81
5	105	95	0,12	1,12	0,13
6	118	106	13,12	12,12	159,01
7	109	100	4,12	6,12	25,21
8	95	88	-9,88	-5,88	58,09
$\sum = 839$		$\sum = 751$	$\sum = 513,84$		
$\bar{x} = 104,88$		$\bar{y} = 93,88$			

Quelle: Eigene Darstellung

Die errechneten Zwischenergebnisse lassen sich nun in die Formel zur Errechnung der Kovarianz eintragen:

$$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n} = \frac{513,84}{8} = 64,23$$

Der Kovarianzwert beträgt 64,23. Das heißt, der vermutete positive Zusammenhang zwischen den beiden Variablen erhärtet sich.

Pearson's r

Als ein von den Maßeinheiten unabhängiges Maß für den Zusammenhang zwischen zwei metrischen Variablen eignet sich der sogenannte Korrelationskoeffizient.

fizient nach Bravais-Pearson, gemeinhin auch Pearson's r genannt. Diesen Korrelationskoeffizienten für zwei (mindestens) intervallskalierte Merkmale erhält man, indem man die Kovarianz durch das Produkt der Standardabweichungen der beiden Variablen x und y teilt:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Neben diesem „Umweg“ über die Kovarianz, lässt sich Pearson's r auch mit folgender Formel berechnen:

$$r = \frac{\overline{x * y} - (\bar{x} * \bar{y})}{\sqrt{\overline{x^2} - \bar{x}^2} * \sqrt{\overline{y^2} - \bar{y}^2}}$$

Der auf diese Weise errechnete Korrelationskoeffizient Pearson's r kann Werte zwischen -1 und $+1$ annehmen, wobei 0 keinen Zusammenhang anzeigt. Das Vorzeichen des Koeffizienten gibt die Art des Zusammenhangs an: Ist der Koeffizient positiv, so spricht man von einem gleichgerichteten oder gleichläufigen (bzw. positiven) Zusammenhang – sowohl die Werte von x als auch die Werte von y steigen oder fallen gleichzeitig. Ist das Vorzeichen negativ, so spricht man von einem gegengerichteten oder gegenläufigen (bzw. negativen) Zusammenhang – steigen die Werte von x , so fallen gleichzeitig die Werte von y und andersherum. Die Interpretation der Stärke des Zusammenhangs bezieht sich auf den Betrag des Koeffizienten. Sie kann somit für die entsprechenden positiven und negativen Werte auf gleiche Weise erfolgen. Tabelle 36 gibt Faustregeln an die Hand, nach denen die Beträge angemessen interpretiert werden können (Kühnel und Krebs 2007, S. 404-405; Kuckartz et al. 2013, S. 213):

Tabelle 36: Interpretation von Pearson's r

Korrelationskoeffizient ($ r $)	Interpretation
$\leq 0,05$	kein Zusammenhang
$> 0,05$ bis $\leq 0,20$	schwacher Zusammenhang
$> 0,20$ bis $\leq 0,50$	mittelstarker Zusammenhang
$> 0,50$ bis $\leq 0,70$	starker Zusammenhang
$> 0,70$	sehr starker Zusammenhang

Quelle: Eigene Darstellung

Für unser Beispiel mit IQ und Testergebnis lässt sich Pearson's r mit obiger Formel berechnen. Dazu empfiehlt es sich wiederum, zunächst eine Arbeitstabelle zu erstellen (siehe Tabelle 37).

Tabelle 37: Arbeitstabelle zur Berechnung von Pearson's r

ID	x_i	y_i	$x_i * y_i$	$(x_i)^2$	$(y_i)^2$
1	104	98	10192	10816	9604
2	90	80	7200	8100	6400
3	103	85	8755	10609	7225
4	115	99	11385	13225	9801
5	105	95	9975	11025	9025
6	118	106	12508	13924	11236
7	109	100	10900	11881	10000
8	95	88	8360	9025	7744
$\Sigma = 839$		$\Sigma = 751$	$\Sigma = 79275$	$\Sigma = 88605$	$\Sigma = 71035$
$\bar{x} = 104,88$		$\bar{y} = 93,88$	$\bar{x} * \bar{y} = 9909,38$	$\bar{x}^2 = 11075,63$	$\bar{y}^2 = 8879,38$

Quelle: Eigene Darstellung

Die Ergebnisse der Arbeitstabelle lassen sich nun wie folgt in die Formel für Pearson's r einsetzen:

$$\begin{aligned}
 r &= \frac{\bar{x} * \bar{y} - (\bar{x} * \bar{y})}{\sqrt{\bar{x}^2 - \bar{x}^2} * \sqrt{\bar{y}^2 - \bar{y}^2}} = \\
 &= \frac{9909,38 - (104,88 * 93,88)}{\sqrt{(11075,63 - 104,88^2)} * \sqrt{(8879,38 - 93,88^2)}} = \\
 &= \frac{9909,38 - 9846,13}{\sqrt{(11075,63 - 10999,81)} * \sqrt{(8879,38 - 8813,45)}} = \\
 &= \frac{63,25}{\sqrt{75,82} * \sqrt{65,93}} = \frac{63,25}{8,71 * 8,12} = \frac{63,25}{70,73} = 0,89
 \end{aligned}$$

Den für das obige Beispiel errechneten Zusammenhang würde man als sehr starken positiven Zusammenhang beschreiben.

Die in der sozialwissenschaftlichen Forschung betrachteten und relevanten Merkmale sind allerdings nur selten metrisch skaliert (Ausnahmen sind z.B. Alter in Lebensjahren und Einkommen). Deshalb werden in der Forschungspraxis häufig auch ordinalskalierte Merkmale, die mindestens fünf Merkmalsausprägungen aufweisen und deren (inhaltliche) Abstände zwischen den Ausprägungen gleich sind, als metrisch behandelt. Insbesondere bei der standardisierten Erfassung von latenten Merkmalen wie Einstellungen liegen sogenannte Likert-Skalen zugrunde, die als „pseudo-“ oder „quasi-metrisch“ behandelt werden (siehe auch Abschnitt 1.3).



Beispiel

An einem weiteren fiktiven Beispiel soll die Berechnung von Pearson's r Schritt für Schritt erläutert werden. Hierzu werden die beiden Merkmale „Einkommen“ und „Lebenszufriedenheit“ genutzt. Folgende Hypothese wird formuliert: Je höher das finanzielle Einkommen ist, desto zufriedener ist

die Person mit dem Leben. Um diese Hypothese beispielhaft mittels Pearson's r zu untersuchen, werden Einkommen und Lebenszufriedenheit mit zwei metrisch bzw. pseudometrisch skalierten Variablen untersucht. So wurden 13 (fiktive) Befragte gebeten, ihr monatliches Nettoeinkommen in Euro zu beziffern. Anschließend wurden sie mit einer 11-stufigen Skala (von 0 bis 10) nach ihrer allgemeinen Lebenszufriedenheit befragt, wie es beispielsweise auch in der ALLBUS- oder ESS-Befragung erfasst wird. In der ALLBUS lautet die allgemeine Frage: „Wie zufrieden sind Sie gegenwärtig – alles in allem – mit ihrem Leben?“ Als Antwort können die Personen einen Wert von 0 bis 10 wählen, bei der 0 „ganz und gar unzufrieden“ und 10 „ganz und gar zufrieden“ bedeutet. Mit den Zahlen dazwischen kann die Antwort abgestuft werden. Höhere Werte deuten also auf größere Lebenszufriedenheit hin. Tabelle 38 zeigt die Ergebnisse.

Tabelle 38: Nettoeinkommen und Lebenszufriedenheit – Urliste

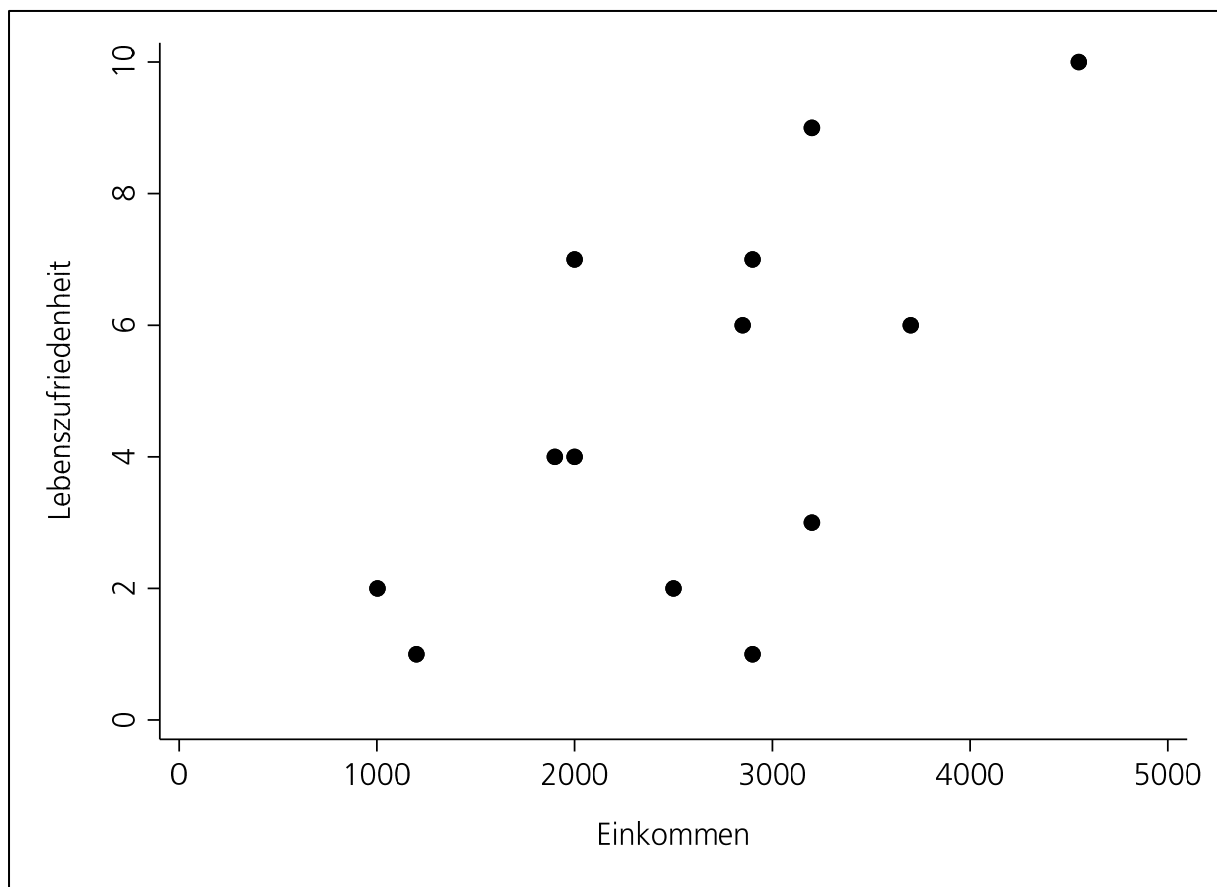
ID	Einkommen	Lebenszufriedenheit
1	2000	7
2	4550	10
3	1003	2
4	3200	9
5	2900	7
6	2850	6
7	1900	4
8	3700	6
9	2000	4
10	2500	2
11	3200	3
12	2900	1
13	1200	1

Quelle: Eigene Darstellung

Diese Datenpunkte lassen sich in einem Streudiagramm eintragen. Abbildung 15 deutet bereits einen positiven linearen Zusammenhang im Sinne der Hypothese an. Wie stark ist dieser Zusammenhang? Um Pearson's r auszurechnen, wird nun auf die Formel zurückgegriffen:

$$r = \frac{\overline{x * y} - (\bar{x} * \bar{y})}{\sqrt{\overline{x^2} - \bar{x}^2} * \sqrt{\overline{y^2} - \bar{y}^2}}$$

Abbildung 15: Nettoeinkommen und Lebenszufriedenheit – Streudiagramm



Daten: Eigene Darstellung

Für die händische Berechnung von Pearson's r bietet sich wieder eine Arbeitstabelle an (siehe Tabelle 39). In der ersten Spalte sind die Identifikationsnummern dokumentiert. In den Spalten daneben finden sich die Angaben für das Einkommen und die Lebenszufriedenheit. Die erste Berechnung findet sich in der vierten Spalte. Dort werden die jeweiligen Werte von x und y multipliziert. In der fünften und sechsten Spalte werden die Werte jeweils quadriert.

Tabelle 39: Arbeitstabelle zur Berechnung von Pearson's r

ID	x_i	y_i	$x_i * y_i$	$(x_i)^2$	$(y_i)^2$
1	2000	7	14000	4000000	49
2	4550	10	45500	20702500	100
3	1003	2	2006	1006009	4
4	3200	9	28800	10240000	81
5	2900	7	20300	8410000	49
6	2850	6	17100	8122500	36
7	1900	4	7600	3610000	16
8	3700	6	22200	13690000	36
9	2000	4	8000	4000000	16
10	2500	2	5000	6250000	4
11	3200	3	9600	10240000	9
12	2900	1	2900	8410000	1
13	1200	1	1200	1440000	1
$\Sigma = 33903$		$\Sigma = 62$	$\Sigma = 184206$	$\Sigma = 100121009$	$\Sigma = 402$
$\bar{x} = 2607,92$		$\bar{y} = 4,77$	$\bar{x} * \bar{y} = 14169,69$	$\overline{x^2} = 7701616,08$	$\overline{y^2} = 30,92$

Quelle: Eigene Darstellung

Die einzelnen Werte werden zunächst addiert und anschließend durch die Fallzahl dividiert, um die arithmetischen Mittel \bar{x} und \bar{y} zu erhalten. Tabelle 40 trägt diese sowie weitere Zwischenergebnisse ab.

Tabelle 40: Zwischenergebnisse zur Berechnung von Pearson's r

\bar{x}	$33903 \div 13 = 2607,92$
\bar{x}^2	$(33903 \div 13)^2 = 2607,92^2 = 6801246,73$
$\overline{x^2}$	$100121009 \div 13 = 7701616,08$
\bar{y}	$62 \div 13 = 4,77$
\bar{y}^2	$(62 \div 13)^2 = 4,77^2 = 22,75$
$\overline{y^2}$	$402 \div 13 = 30,92$
$\bar{x} * \bar{y}$	$184206 \div 13 = 14169,69$

Quelle: Eigene Darstellung

Die Zwischenergebnisse, die in Tabelle 40 eingetragen wurden, werden anschließend in die Formel eingesetzt: Pearson's r kann danach Schritt für Schritt ausgerechnet werden:

$$\begin{aligned}
 r &= \frac{\overline{x * y} - (\bar{x} * \bar{y})}{\sqrt{\overline{x^2} - \bar{x}^2} * \sqrt{\overline{y^2} - \bar{y}^2}} = \\
 &= \frac{14169,69 - (2607,92 * 4,77)}{\sqrt{7701616,08 - 6801246,73} * \sqrt{30,92 - 22,75}} = \\
 &= \frac{14169,69 - 12439,78}{\sqrt{900369,35} * \sqrt{8,17}} = \frac{1729,91}{948,88 * 2,86} = \frac{1729,91}{2713,80} = 0,64
 \end{aligned}$$

Pearson's r beträgt 0,64. Dies entspricht einem starken Zusammenhang. In der sozialwissenschaftlichen Praxis sind solche hohen Korrelationskoeffizienten allerdings eher selten, da erklärungsbedürftige Phänomene häufig von mehr als nur einem weiteren Merkmal abhängen.

3.5 Eta-Quadrat für metrische und nominale Merkmale

Es wurde besprochen, dass – wenn die beiden zu betrachtenden Merkmale unterschiedliche Skalenniveaus aufweisen – das Assoziationsmaß ausgewählt werden soll, welches zum niedrigsten Skalenniveau passt. Soll beispielsweise der Zusammenhang zwischen Familienstand und Nettoeinkommen berechnet werden, würde man nach dieser Logik Cramer's V nutzen, da die Variable Familienstand nominal skaliert ist und in diesem Beispiel das niedrigere Skalenniveau umfasst. Allerdings steht mit dem Eta-Koeffizienten auch ein Parameter zur Verfügung, mit dem sich spezifisch die Beziehung zwischen einer nominalen und einer metrischen Variable berechnen lässt (Benninghaus 2007, S. 228-250). Es wird dabei ermittelt, inwieweit bzw. wie „gut“ die Varianz einer (abhängigen) mindestens intervallskalierten Variable (z.B. Einkommen) durch eine unabhängige nominale Variable (z.B. Familienstand) erklärt wird. Eta-Quadrat wird auch genutzt, um Auskunft über die sogenannte Effektstärke der unabhängigen Variable zu erhalten (Kuckartz et al. 2013, S. 195).

Proportional Reduction of Error

Für die Berechnung von Eta-Quadrat muss die abhängige Variable mindestens Intervallskalenniveau haben, während die unabhängige Variable jedes andere, also auch Nominalskalenniveau, haben kann. Ein Beispiel wäre, wenn untersucht werden soll, ob der Familienstand einer Person das Einkommen beeinflusst. Inhaltlich möchte man hier wissen, ob bzw. wie stark der Familienstand mit dem Nettoeinkommen zusammenhängt. Statistisch lässt sich mit Eta-Quadrat angeben, wie viel Prozent der Gesamtvarianz der abhängigen Variable durch die unabhängige Variable aufgeklärt wird. Damit gehört der Eta-Koeffizient bzw. Eta-Quadrat zur Familie der sogenannten PRE-Maße (**P**roportional **R**eduction of **E**rror). Die relative Verbesserung der Vorhersage (bzw. die Verringerung der Fehlerquote) wird im PRE-Maß ausgedrückt. Für ein PRE-Maß benötigt man zwei Vorhersageregeln: Eine für den Fall, dass man keine Kenntnis über den Zusammenhang zwischen x (= uV) und y (= aV) hat sowie und eine für den Fall, dass man über entsprechende Kenntnisse verfügt.

PRE-Maße

Die Ausgangsfrage bei der Berechnung von PRE-Maßen lautet, wie gut die Werte einer abhängigen Variable durch die Werte einer unabhängigen Variable vorhergesagt werden können. Anders ausgedrückt versuchen PRE-Maße zu ermitteln, wie gut eine unabhängige Variable die abhängige Variable vorhersagen kann. Je nach Skalenniveau lassen sich verschiedene Maße berechnen. Das Vorgehen folgt dabei immer derselben schrittweisen Logik (Ludwig-Mayerhofer et al. 2014, S. 235). Erstens: Wie lautet die Prognose des Wertes der abhängigen Variable ohne Kenntnis der unabhängigen Variablen? (Vorhersagefehler E1). Hier stehen als Annäherungswerte die Mittelwerte der abhängigen Variable zur Verfügung (arithmetisches Mittel bei intervallskaliertem Variable und Modus bei nominaler Variable). Zweitens: Prognose des Wertes der abhängigen Variable mit Kenntnis der Verteilung der unabhängigen Variable (Vorhersagefehler E2). Drittens: Ermittlung des PRE-Maßes und Aussage, ob die Vorhersage durch die unabhängige Variable verbessert wurde. Subtrahiert man Fehler 2 (E2) von Fehler 1 (E1) und dividiert dies durch Fehler 1 (E1), erhält man das entsprechende PRE-Maß. Je nach PRE-Maß werden dabei die Fehler unterschiedlich berechnet. Die allgemeine PRE-Formel lautet:

$$\text{PRE} = (E1 - E2) / E1$$

Die Berechnung des PRE-Maßes Eta-Quadrat basiert auf sogenannten Quadratsummen. Die Quadratsumme stellt die Summe der quadrierten Abweichungen vom Mittelwert eines Merkmals dar (geteilt durch die Fallzahl ergibt sich daraus die empirische Varianz). Verschiedene Quadratsummen lassen sich unterscheiden:

- Die „Quadratsumme Gesamt“: Entspricht der Summe aller quadrierten Abweichungen vom Mittelwert der (abhängigen) Variable. Dies ist Vorhersagefehler E1.
- „Quadratsumme innerhalb“: Wird die unabhängige Gruppenvariable berücksichtigt, wird innerhalb der Gruppen berechnet, wie stark die Gruppenmittelwerte vom Gesamtmittelwert abweichen (also beispielsweise wird bei der Gruppenvariable „Familienstand“ geschaut, wie die Mittelwerte der Gruppen der Verheirateten, der Ledigen, der Geschiedenen etc. ausfallen und in Bezug zum Gesamtmittelwert gesetzt). Dies ist Vorhersagefehler E2.

Subtrahiert man von der „Quadratsumme Gesamt“ die „Quadratsumme innerhalb“, ergibt sich die „Quadratsumme zwischen“: Dies sind die quadrierten Abweichungen der unterschiedlichen Gruppenmittelwerte. Es wird also untersucht, wie stark die Vorhersagekraft gegenüber dem Mittelwert des abhängigen Merkmals verbessert wird, wenn die unabhängige Gruppenvariable (z.B. Familienstand oder Geschlecht) berücksichtigt wird. Ist der Unterschied zwischen den einzelnen Gruppen, die durch die unabhängige Variable konstituiert werden, substantiell und innerhalb der einzelnen Gruppen niedrig, deutet dies auf einen Zusammenhang hin (Ludwig-Mayerhofer et al. 2014, S. 235).

Um herauszufinden, wie sich die Gruppenzugehörigkeit auf das abhängige metrische Merkmal auswirkt, wird auf Basis obiger PRE-Formel Eta-Quadrat wie folgt berechnet:

$$\eta^2 = \frac{\text{Quadratsumme Gesamt} - \text{Quadratsumme innerhalb}}{\text{Quadratsumme Gesamt}} = \frac{\text{Quadratsumme Zwischen}}{\text{Quadratsumme Gesamt}}$$

Der Eta-Koeffizient kann Werte zwischen 0 und 1 annehmen. Je höher der Wert ist, desto besser trägt die untersuchte unabhängige Variable zur Erklärung der abhängigen Variable bei. Cohen (1988) unterscheidet bei der Interpretation von Eta zwischen folgenden drei Effekten (siehe auch Tabelle 41): kleine Effekte, mittlere und große Effekte (siehe auch Kuckartz et al. 2013, S. 195; Döring und Bortz 2016, S. 821). Damit lässt sich ein Eta-Koeffizient ab 0,14 dahingehend interpretieren, dass die beiden untersuchten Variablen sich in einem starken Zusammenhang befinden, das heißt, die unabhängige Variable einen großen Effekt auf die abhängige Variable ausübt, während ein niedriger Wert auf eine schwache Assoziation verweist. Multipliziert man den Wert von Eta-Quadrat mit 100 %, lässt sich eine Aussage darüber treffen, wie viel Prozent der Varianz der abhängigen Variable von der Zugehörigkeit zu einer bestimmten Gruppe vorhergesagt oder erklärt werden kann.

Tabelle 41: Interpretation von Eta-Quadrat

Eta-Quadrat	Interpretation
< 0,01	kein Effekt
0,01 bis < 0,06	kleiner Effekt
0,06 bis < 0,14	mittlerer Effekt
≥ 0,14	großer Effekt

Quelle: Eigene Darstellung



An einem fiktiven Beispiel wird die Berechnung und Interpretation von Eta-Quadrat verdeutlicht. Wissen Kinder ohne Migrationshintergrund mehr über das politische System als Kinder mit einem Migrationshintergrund, das heißt, hat ein Migrationshintergrund einen Effekt auf das politische Wissen von Kindern (Abendschön und Tausendpfund 2017)? Um diese Frage zu untersuchen, absolvieren zehn (fiktive) Kinder ein kleines Politikquiz, bei dem 16 Punkte erzielt werden können. Tabelle 42 informiert über den Migrationshintergrund und die erreichte Punktzahl der Kinder.

Tabelle 42: Migrationshintergrund und politisches Wissen

ID	Migrationshintergrund	Politisches Wissen (y_i)
1	1 (Nein)	15
2	2 (Ja)	10
3	2 (Ja)	2
4	2 (Ja)	9
5	1 (Nein)	7
6	1 (Nein)	16
7	1 (Nein)	14
8	2 (Ja)	6
9	2 (Ja)	4
10	1 (Nein)	12

Quelle: Eigene Darstellung

Zunächst wird der Gesamtmittelwert beim Test ermittelt. Dieser liegt bei 9,5 Punkten. Er stellt die „beste“ Vorhersage dar, wenn man keine unabhängige Variable berücksichtigt. Es zeigt sich allerdings beim Kind mit der ID 1, dass der Mittelwert hier eine um 5,5 Punkte zu niedrige Schätzung darstellt (vorletzte Spalte in Tabelle 43). Bei dem Kind mit der ID 9 liegt mit dem Mittelwert dagegen eine Überschätzung um 5,5 Punkte vor. Um Eta-Quadrat berechnen zu können, werden die jeweiligen Abweichungen vom Gesamtmittelwert quadriert (letzte Spalte in Tabelle 43). Insgesamt beträgt die Summe der quadrierten Abweichungen 204,5. Dies entspricht dem Vorhersagefehler E1.

Tabelle 43: Arbeitstabelle Migrationshintergrund und politisches Wissen

ID	Migrationshintergrund	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	1 (Nein)	15	5,5	30,25
2	2 (Ja)	10	0,5	0,25
3	2 (Ja)	2	-7,5	56,25
4	2 (Ja)	9	-0,5	0,25
5	1 (Nein)	7	-2,5	6,25
6	1 (Nein)	16	6,5	42,25
7	1 (Nein)	14	4,5	20,25
8	2 (Ja)	6	-3,5	12,25
9	2 (Ja)	4	-5,5	30,25
10	1 (Nein)	12	2,5	6,25
		$\Sigma = 95$		$\Sigma = 204,5$
		$\bar{y} = 9,5$		

Quelle: Eigene Darstellung

Tabelle 44 und Tabelle 45 berechnen nun die Gruppenmittelwerte nach Migrationshintergrund bzw. die Abweichungen von diesen, einmal für die Gruppe der Kinder mit (Tabelle 45) und einmal für die Gruppe der Kinder ohne Migrationshintergrund (Tabelle 44).

Tabelle 44: Arbeitstabelle Migrationshintergrund (Nein) und politisches Wissen

ID	Migrationshintergrund	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	1 (Nein)	15	2,2	4,84
5	1 (Nein)	7	-5,8	33,64
6	1 (Nein)	16	3,2	10,24
7	1 (Nein)	14	1,2	1,44
10	1 (Nein)	12	-0,8	0,64
		$\Sigma = 64$		$\Sigma = 50,8$
		$\bar{y} = 12,8$		

Quelle: Eigene Darstellung

Das arithmetische Mittel für Kinder ohne Migrationshintergrund liegt bei 12,8 Punkten, während es bei den Kindern mit Migrationshintergrund bei 6,2 Punkten liegt. Auch hier lassen sich für die einzelnen Kinder wieder die Abweichungen vom (Gruppen-)Mittelwert berechnen (vorletzte Spalte) und quadrieren (letzte Spalte). Bei den Kindern ohne Migrationshintergrund liegt die Abweichungssumme bei 50,8. Bei den Kindern mit Migrationshintergrund liegt die Abweichungssumme bei 44,8. Beide quadrierten Abweichungssummen ergeben E2, im vorliegenden Beispiel 95,6.

Tabelle 45: Arbeitstabelle Migrationshintergrund (Ja) und politisches Wissen

ID	Migrationshintergrund	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	2 (Ja)	10	3,8	14,44
3	2 (Ja)	2	-4,2	17,64
4	2 (Ja)	9	2,8	7,84
8	2 (Ja)	6	-0,2	0,04
9	2 (Ja)	4	-2,2	4,84
		$\Sigma = 31$		$\Sigma = 44,8$
		$\bar{y} = 6,2$		

Quelle: Eigene Darstellung

Auf Basis der PRE-Formel wird im nächsten Schritt Eta-Quadrat berechnet:

$$\eta^2 = \frac{E1 - E2}{E1} = \frac{204,5 - (50,8 + 44,8)}{204,5} = \frac{204,5 - 95,6}{204,5} = \frac{108,9}{204,5} = 0,53$$

Der Eta-Koeffizient ergibt einen Wert von 0,53.

Auf Basis der in Tabelle 41 vorgestellten Faustregeln der Interpretation deutet dies im vorliegenden Beispiel auf einen sehr großen Effekt des Migrationshintergrunds auf das politische Wissen hin.

53 Prozent der Varianz des politischen Wissens (gemessen durch das Politikquiz) lassen sich durch das Merkmal „Migrationshintergrund“ erklären.

3.6 Zusammenfassung

Was können bivariate Analysen leisten? Es wurde gezeigt, dass – je nach Skalierung der interessierenden Merkmale – auf verschiedene Weise Zusammenhänge zwischen zwei Variablen beschrieben und analysiert werden können.

Kreuztabellen und die damit verbundenen Maße (Chi-Quadrat, Phi und Cramer's V) geben Auskunft über Zusammenhänge zwischen zwei nominalskalierten Variablen bzw. einer nominal- und einer ordinalskalierten Variable. Der Rangkorrelationskoeffizient Spearman's Rho bietet sich an, wenn echte Rangkorrelationen zu berechnen sind bzw. wenn Assoziationen zwischen zwei ordinalskalierten Merkmalen berechnet werden sollen. Für zwei (pseudo-)metrisch skalierte Variablen wird Pearson's r genutzt. Eta-Quadrat eignet sich für bivariate Merkmalskombinationen, bei denen die uV nominal skaliert und die aV über ein metrisches Skalenniveau verfügt. Die kennengelernten Assoziationsmaße beschreiben die Stärke eines Zusammenhangs zweier Variablen. Verfügen beide interessierenden Merkmale über unterschiedliche Skalenniveaus, dann sollte man sich ein geeignetes Assoziationsmaß für das niedrigere Skalenniveau auswählen. Möchte man beispielsweise eine ordinalskalierte Bildungsvariable mit der nominalskalierten Familienstandsvariable in Beziehung setzen und mögliche Zusammenhänge untersuchen, sollte man zu Cramer's V als Assoziationsmaß für nominale Variablen greifen (Weins 2010, S. 74).

Was können bivariate Analysen nicht leisten? Art und Stärke des Zusammenhangs zweier Merkmale können durch die kennengelernten Assoziationsmaße ermittelt werden. Ob eine kausale Beziehung zwischen den Merkmalen besteht, lässt sich jedoch „technisch“ oder statistisch nicht bestimmen. Zum einen ist dies nicht möglich, weil mögliche und plausible (aber nicht erhobene) weitere Merkmale, sogenannte Drittvariablen, mit beiden Merkmalen zusammenhängen könnten. Somit könnte auch eine bestimmte Beziehung zwischen zwei Merkmalen nur vorgetäuscht sein – in solchen Fällen spricht man von einer Scheinkorrelation (Tausendpfund 2018b, S. 191). Ein klassisches Beispiel ist hier eine festgestellte hohe Korrelation zwischen der Anzahl von Störchen in einer Wohngegend und der Anzahl Neugeborener. Diese belegt nun nicht die alte „These“, dass Klapperstörche Babys bringen, vielmehr ist eine dritte Variable entscheidend, die mit beiden Merkmalen im Zusammenhang steht: Die Bebauungsart des Wohngebiets. In ländlichen Gegenden siedeln sich vorzugsweise sowohl Störche als auch junge Familien an.

Zum anderen ist es in vielen Fällen nicht möglich, die Richtung der Kausalität festzulegen. In den Fällen, in denen allerdings vor der Durchführung der bivariaten Analyse sozialwissenschaftliche Hypothesen formuliert werden, hat man in der Regel aufgrund von Literaturstudium und theoretischen Vorarbeiten eine Vorstellung von der Richtung des Zusammenhangs bzw. zugrundeliegenden kausalen Mechanismen. So wäre die dem obigen Beispiel zugrundeliegende Annahme, dass Menschen eine höhere Lebenszufriedenheit aufweisen, wenn sie mehr Geld verdienen. Allerdings lässt sich diese Kausalität statistisch nicht belegen. Das zeigt allerdings erneut die Wichtigkeit sozialwissenschaftlicher Theoriearbeit im Vorfeld auf.

4 Multivariate Datenanalyse

Simone Abendschön

Vorschau



In diesem Kapitel werden mit der linearen und der logistischen Regression zwei multivariate Analyseverfahren vorgestellt, die in den Sozialwissenschaften häufig verwendet werden, um Hypothesen über Einflusststrukturen zu überprüfen. Multivariate Analyseverfahren tragen dabei dem Umstand Rechnung, dass sich in der sozialen Wirklichkeit zu erklärende Phänomene so gut wie nie auf nur einen Faktor zurückführen lassen, sondern in der Regel verschiedenen Einflüssen unterworfen sind. Die lineare Regressionsanalyse lässt sich immer dann anwenden, wenn eine lineare Beziehung zwischen der abhängigen Variable und den unabhängigen Variablen besteht und die abhängige Variable ein metrisches Skalenniveau aufweist. Die logistische Regression kommt bei einer dichotomen abhängigen Variable zum Einsatz. Im vorliegenden Kapitel werden Grundlagen und Durchführung beider Verfahren anhand von Beispielen besprochen.

4.1 Einführung

Mit den im vorherigen Kapitel kennengelernten bivariaten Analyseverfahren lassen sich Beziehungen zwischen zwei Variablen untersuchen. Sie ermöglichen es, Zusammenhangshypothesen zwischen zwei Variablen zu testen. Da allerdings sozialwissenschaftliche Phänomene in aller Regel von mehreren Faktoren beeinflusst werden, stößt man mit bivariaten Verfahren schnell an die Grenzen der Erklärungskraft. Ein Beispiel: Ob ein Mensch zur Wahl geht, hängt nicht nur von seinem politischen Interesse, sondern unter anderem auch von seiner Bildung, seinem Einkommen und der Frage ab, ob er sich mit einer bestimmten politischen Partei identifiziert. Multivariate Analyseverfahren bieten die Möglichkeit, den Einfluss verschiedener erklärender Faktoren oder Variablen auf ein erklärungsbedürftiges Phänomen zu untersuchen. Das am häufigsten angewandte Verfahren ist hierbei die Regressionsanalyse.

Regression als Analyseverfahren

Das lateinische Wort „regressere“ lässt sich mit „zurückführen auf“ übersetzen – statistisch gesprochen bedeutet dies, dass die Ausprägungen einer zu erklärenden oder „abhängigen“ Variable auf die erklärenden oder „unabhängigen“ Variablen zurückgeführt werden sollen. Man möchte, um beim obigen Beispiel zu bleiben, die Wahlbeteiligung von Bürgern zurückführen auf ihre Bildung, ihr Einkommen, ihr politisches Interesse oder auch weitere relevante Faktoren. Neben der Erklärung von abhängigen Variablen lässt sich auf der anderen Seite eine Regressionsanalyse auch nutzen, um auf Basis der unabhängigen Variablen Vorhersagen für die abhängige Variable zu machen. Das heißt, die Regressionsanalyse kann auch für Prognosen eingesetzt werden. Um wieder auf obiges Beispiel zurückzukommen: Wenn wir das politische Interesse und den sozialen Status einer Person kennen, können wir auf Basis dieser Informationen „schätzen“, wie wahrscheinlich es ist, dass sie zur Wahlurne gehen wird.

Im Kontext der Regressionsanalyse werden häufig auch alternative Bezeichnungen für abhängige und unabhängige Variablen verwendet (siehe Tabelle 46). Anstelle von abhängiger Variable spricht man auch von erklärter Variable, Kriteriumsvariable, endogener Variable oder auch Regressand. Die unabhängige Variable wird je nach Fachkultur auch als erklärende Variable, Prädiktorvariable, exogene Variable oder auch Regressor bezeichnet.

Tabelle 46: Unterschiedliche Bezeichnungen für Variablen der Regressionsanalyse

Abhängige Variable (y)	Unabhängige Variable (x)
Erklärte Variable	Erklärende Variable
Kriteriums(variable)	Prädiktor(variable)
Endogene Variable	Exogene Variable
Regressand	Regressor

Quelle: Eigene Darstellung

Im Unterschied zu bivariaten Analyseverfahren untersucht die Regressionsanalyse auch Kausalbeziehungen, das heißt, sie bestimmt die Richtung und den Ursprung eines Einflusses und möchte dessen Stärke messen. Wir gehen also beispielsweise nicht nur davon aus, dass die Merkmale „politisches Interesse“ und „Wahlbeteiligung“ zusammenhängen, sondern dass das politische Interesse der Wahlberechtigten einen Einfluss darauf nimmt, ob er sie ins Wahllokal geht. Damit gehört die Regressionsanalyse zu den sogenannten „strukturprüfenden Verfahren“ (siehe auch Backhaus et al. 2018), mittels derer sich theoretisch entwickelte Hypothesen über die Beziehungsstruktur zwischen Variablen prüfen lassen. Zu beachten ist hierbei, dass die Regressionsanalyse aus statistischer Perspektive Kausalität selbst nicht nachweisen kann, sondern Korrelationen zwischen Variablen berechnet werden. Auch aus diesem Grund sind eine Theoriebildung und aus der Theorie abgeleitete Hypothesen als Vorbereitung einer Regressionsanalyse unerlässlich.

Die sozialwissenschaftlichen Fragestellungen, die mit Hilfe von Regressionsanalysen untersucht werden, können dabei sowohl auf der Individual- bzw. Mikroebene als auch auf der Aggregat- bzw. Makroebene angesiedelt sein. Das zeigt nochmals das Beispiel der Wahlbeteiligung, welche einerseits auf Mikro- und andererseits auf Makroebene untersucht werden kann. Auf der Mikroebene würde man fragen: Was beeinflusst die individuelle Wahlbeteiligung? Diese Frage könnte man mit Hilfe von Individualdaten untersuchen (z.B. GLES). Möchte man die Höhe der Wahlbeteiligung auf Länderebene erklären, ist die Fragestellung auf der Makroebene angesiedelt und man untersucht Makrofaktoren wie zum Beispiel die Existenz einer Wahlpflicht, das Wahlsystem und/oder das Bruttosozialprodukt eines Landes. In beiden Fällen wird man Querschnittsdaten nutzen, also Daten, die zu einem bestimmten Zeitpunkt zur Verfügung stehen. Für Längsschnitts- und Mehrebenenbetrachtungen stehen erweiterte Regressionsmodelle zur Verfügung.

Fragen auf Mikro- und Makroebene

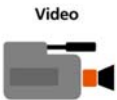
Ganz allgemein gesprochen lassen sich mit Regressionsanalysen zwei (verwandte) Fragen beantworten:

- 1) Welchen Einfluss üben einzelne (unabhängige) Variablen auf eine abhängige Variable aus? Dies beinhaltet auch Aussagen über die Stärke und die Richtung (positiv oder negativ) des Einflusses der einzelnen unabhängigen Variablen auf die abhängige Variable.
- 2) Wie gut erklären bestimmte (unabhängige) Variablen eine abhängige Variable insgesamt? Hier geht es um die Gesamtgüte des Regressionsmodells.

Wird mit Stichprobendaten, wie beispielweise mit der ALLBUS oder dem European Social Survey, gearbeitet, lassen sich innerhalb der Regressionsanalyse auch die statistische Signifikanz sowohl der einzelnen Einflussfaktoren als auch des Gesamtmodells berechnen.

Je nach Messniveau der zu erklärenden bzw. abhängigen Variable unterscheidet man hierbei zwischen der linearen Regressionsanalyse (bei metrischen bzw. pseudometrischen abhängigen Variablen) und der logistischen Regressionsanalyse (bei dichotomen abhängigen Variablen). In den folgenden Abschnitten werden diese beiden verschiedenen Analyseverfahren vorgestellt und erläutert.

4.2 Lineare Regression



Bei einer linearen Regression wird der lineare Zusammenhang zwischen mindestens einer unabhängigen Variable und einer abhängigen Variable untersucht. Sie geht dabei von einem linearen Zusammenhang der Variablen aus. Um das Prinzip, die Grundlagen, das generelle Vorgehen und die grundlegenden Begrifflichkeiten der linearen Regressionsanalyse zu verstehen, wird zunächst beispielhaft der bivariate Fall betrachtet (4.2.1), also eine unabhängige und eine abhängige Variable, um danach weitere unabhängige Variablen mit einzubeziehen (4.2.2).

Mit der linearen Regressionsanalyse werden aus statistischer Sicht verschiedene Ziele verfolgt. Dabei liegt das übergeordnete Ziel darin, eine Schätzgleichung zu ermitteln, die eine möglichst genaue Beschreibung der durchschnittlichen linearen Abhängigkeit einer Variable (aV) von mindestens einer anderen Variable (uV) darstellt. Es geht also darum, die beobachtete Varianz bzw. Streuung einer abhängigen Variable durch andere (theoretisch relevante) Variablen zu erklären.

Beispiel: Determinanten der Lebenszufriedenheit

Die Frage, was ein glückliches und zufriedenes Leben ausmacht und welche Faktoren ein solches befördern, steht seit einigen Jahren nicht nur im Mittelpunkt von Talkshows und populärwissenschaftlichen Publikationen, sondern auch der soziologischen und ökonomischen Forschung. Physischem Wohlbefinden und Befriedigung menschlicher Grundbedürfnisse sowie sozialer Wertschätzung kommt dabei, neben weiteren Faktoren wie der Persönlichkeitsstruktur oder Religiosität, eine wichtige Rolle zu (Frey 2017; Helliwell et al. 2018). Im vorliegenden Kapitel werden wir untersuchen, welche Rolle Einkommen, Gesundheitszustand, Bildung und Familienstand als mögliche Bestimmungsfaktoren der Lebenszufriedenheit spielen.

In Bezug auf unser Beispiel „Determinanten der Lebenszufriedenheit“ bedeutet das, dass wir die Streuung der beobachteten Daten durch das Einkommen (und später durch weitere Merkmale) erklären möchten. Dabei möchten wir auch ermitteln, wie groß der Einfluss der einzelnen Faktoren jeweils ist, und prüfen, wie genau alle Einflussvariablen zusammen, welche wir auch als unser „Regressionsmodell“ bezeichnen können, die abhängige Variable erklären. Wie gut können diese Einflussfaktoren einzeln und gemeinsam die Höhe der Lebenszufriedenheit bestimmen und damit auch vorhersagen?

4.2.1 Bivariate Regression

Wie in der kurzen Einführung im letzten Abschnitt bereits angedeutet, hängt es vom Skalenniveau der abhängigen Variable ab, welches Regressionsmodell eingesetzt werden kann. Die lineare Regression ist immer dann möglich, wenn die abhängige Variable zumindest intervallskaliert vorliegt, wie beispielsweise beim Einkommen in Euro.

In der sozialwissenschaftlichen Praxis hat man es aber eher selten mit „echten“ metrischen Variablen zu tun. Allerdings erhebt die sozialwissenschaftliche Forschung Einstellungen und Zustimmungen häufig mit sogenannten Likert-Skalen bzw. sogenannten „endpunktbenannten Skalen“ (Porst 2014). Damit stehen Messinstrumente zur Erfassung von Merkmalen, die als quasi- oder pseudometrisch gelten können, zur Verfügung (Tausendpfund 2018b, S. 122). Diese Skalen erheben – unter Vorbehalt – den Anspruch, dass sie sich als intervallskaliert behandeln lassen. Um diesen Anspruch zu erfüllen, sollten diese Skalen aber zum einen mindestens fünf Ausprägungen besitzen und zum anderen sollten die Abstände zwischen den Ausprägungen der Skala durch entsprechende Wertzuweisungen als gleich groß verstanden werden können (Urban und Mayerl 2018, S. 13-14).

Voraussetzung der linearen Regression

Am bivariaten linearen Regressionsmodell, also mit einer abhängigen und einer unabhängigen Variable, sollen nun zunächst die zentralen Begriffe vorgestellt und beispielhaft in die Regressionsanalyse eingeführt werden. Im Beispiel wird dazu als abhängige Variable die individuelle Lebenszufriedenheit und die Variable Einkommen als unabhängige Variable gewählt. Die Hypothese, die wir überprüfen möchten, lautet:

H1: Je höher das Einkommen ist, desto größer ist die Lebenszufriedenheit.

Wie kann die formulierte Hypothese begründet werden? Ein höheres Einkommen verbessert die Konsummöglichkeiten, fördert die Teilnahme am gesellschaftlichen Leben und wirkt sich dadurch günstig auf die Lebenszufriedenheit aus (Neller 2004). Dagegen müssen sich Menschen mit einem niedrigen Einkommen ständig ums Geld kümmern. Ein unerwarteter Vorfall, etwa eine defekte Waschmaschine, wird als belastend empfunden und kann die Lebenszufriedenheit senken (Frey 2017, S. 15).

Hypothese begründen

Die Variable Lebenszufriedenheit wird mit einer elfstufigen Skala erfasst. In der ALLBUS lautet die Frageformulierung wie folgt: „Und jetzt noch eine allgemeine Frage. Wie zufrieden sind Sie gegenwärtig – alles in allem – mit ihrem Leben?“ Die Befragten können dabei Werte von 0 bis 10 angeben, wobei der Wert 0 „ganz und gar unzufrieden“ und der Wert 10 „ganz und gar zufrieden“ bedeutet. Das Einkommen wird über das monatliche Nettoeinkommen der befragten Person (offene Abfrage) operationalisiert. Tabelle 47 zeigt die beobachteten Werte auf beiden Variablen für acht fiktive Befragte.

Tabelle 47: Bivariate lineare Regression mit Lebenszufriedenheit und Einkommen

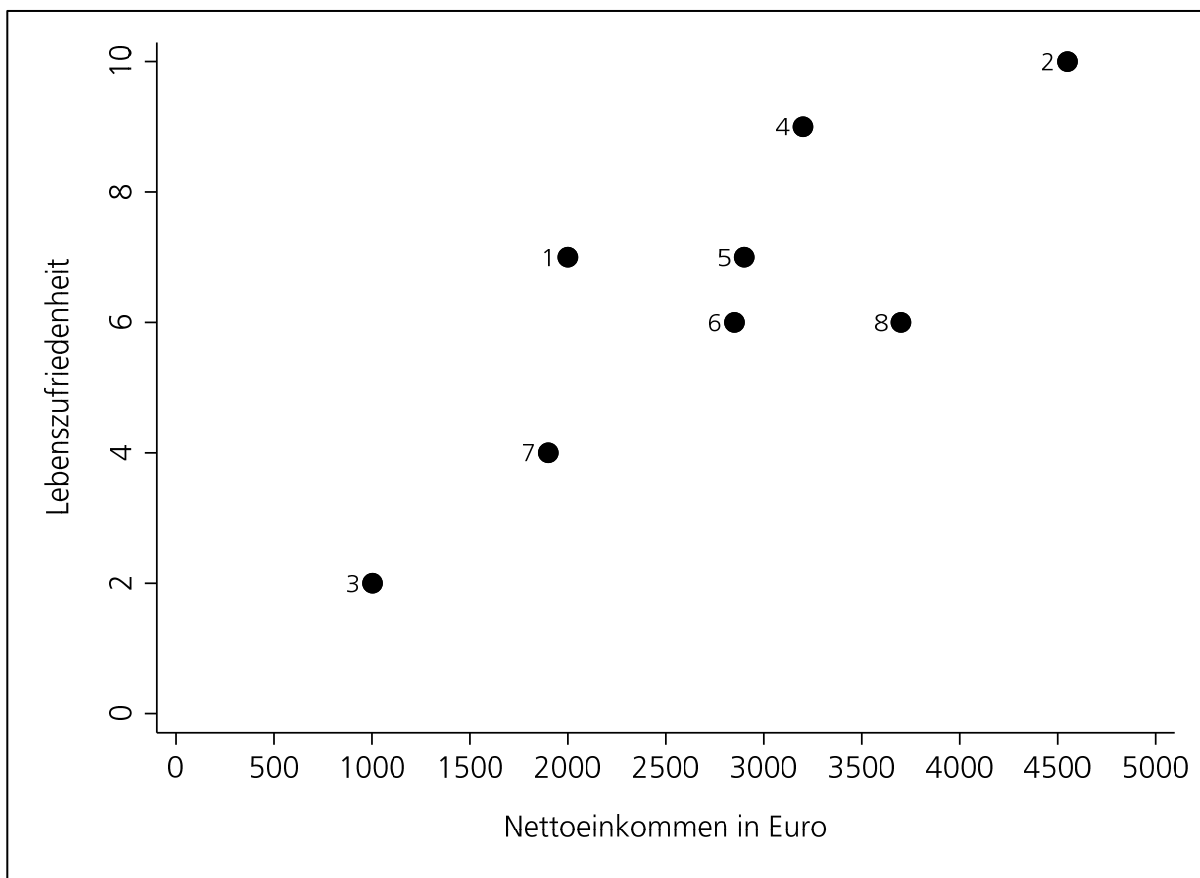
ID	Lebenszufriedenheit (Wert auf Skala von 0 bis 10)	Nettoeinkommen im Monat in Euro
1	7	2000
2	10	4550
3	2	1003
4	9	3200
5	7	2900
6	6	2850
7	4	1900
8	6	3700

Quelle: Eigene Darstellung

Diese beobachteten Werte lassen sich auch grafisch in einem Koordinatensystem darstellen. Dazu fertigen wir ein sogenanntes Streudiagramm oder Scatterplot an. Die Punktwolke in Abbildung 16 stellt eine solche bivariate Illustration dar. Sie zeigt die gemessenen Werte der acht fiktiven Befragten aus Tabelle 47 auf beiden Variablen als Punkte in einem Koordinatensystem. Mit Hilfe von Statistikprogrammen wie SPSS oder Microsoft Excel lassen sich solche Streudiagramme schnell erstellen.

Auf der y-Achse wird dabei die Lebenszufriedenheit abgetragen, auf der x-Achse das monatliche Nettoeinkommen in Euro. Um die Hypothese testen zu können, wird im (bivariaten) Regressionsmodell die abhängige Variable y auf die unabhängige Variable x – in unserem Beispiel die Lebenszufriedenheit auf das Einkommen – zurückgeführt. Die Form der Punktwolke bestätigt bereits die Richtung der in der Hypothese geäußerten Vermutung: Je höher das Einkommen ist, desto größer ist offensichtlich auch die geäußerte Lebenszufriedenheit der befragten Person.

Abbildung 16: Streudiagramm



Quelle: Eigene Darstellung

Lineare Beziehungen zwischen zwei Variablen werden mathematisch durch eine Gerade dargestellt. So wie sich in der sozialen Wirklichkeit aber keine perfekten Korrelationen zwischen zwei sozialen Phänomenen finden lassen, wird sich in der sozialwissenschaftlichen Praxis eine Variable nie vollständig auf eine andere zurückführen lassen, so dass immer ein unerklärter Rest übrig bleibt. In der Regel gibt es in der Sozialforschung also keinen perfekten Zusammenhang zwischen zwei Variablen – zu viele „Störvariablen“ existieren außerhalb des Labors im „wirklichen“ Leben.

Dennoch gehen wir davon aus, dass die einzelnen Ausprägungen unserer abhängigen Variable Y funktional von den jeweiligen Ausprägungen der unabhängigen Variablen X abhängen. Das heißt:

$$y = f(x)$$

Oder auf unser Beispiel bezogen:

$$\text{Lebenszufriedenheit} = f(\text{Einkommen})$$

Da die generelle Gleichung einer Geraden $f(x) = \alpha + \beta * x$ lautet, wobei α den Schnittpunkt der Geraden mit der y-Achse und β die Steigung der Geraden darstellt, können wir folgende Gleichung für unsere Regressionsgerade aufstellen:

$$y = \alpha + \beta * x$$

Verschiedene Störgrößen

Es wurde bereits angesprochen, dass verschiedene Störgrößen verhindern, dass alle Beobachtungen auf einer Regressionsgeraden liegen. Erstens sind dies systematische Fehler, die dadurch zustande kommen, dass bestimmte Variablen (noch) nicht im Regressionsmodell integriert sind. In unserem Beispiel könnte dies der Gesundheitszustand der befragten Personen sein, der – unabhängig von der finanziellen Situation – sicherlich einen Einfluss auf die Lebenszufriedenheit hat. Zweitens sind dies Beobachtungs- und Messfehler, die während der Datenerhebung auftreten können (z.B. Befragte machen falsche Angaben). Drittens gibt es unbekannte Störgrößen, die sich nicht näher spezifizieren lassen.

Die Aufgabe der linearen Regressionsanalyse besteht nun darin, die Regressionsgerade mit Hinblick auf die Datenpunkte bestmöglich zu schätzen. Trotz Störungen besteht nämlich die Möglichkeit, eine Regressionsgerade anzupassen, die die empirischen Beobachtungen so genau wie möglich erklärt. Die in Abbildung 17 eingezeichnete (vorweggenommene) lineare Regressionsgerade erklärt die Punktwolke in dieser Hinsicht und gibt dabei Auskunft über Richtung und Stärke des Einflusses von X (Einkommen) auf Y (Lebenszufriedenheit).

Daraus ergibt sich für die zu schätzende Regressionsgerade folgende Gleichung:

$$\hat{y} = \alpha + \beta * x$$

Dabei ist:

\hat{y} : geschätzter Wert der unabhängigen Variable y (das „Dach“ über y zeigt an, dass es sich um eine Schätzung handelt).

α : Schnitt mit der y-Achse (bei $x = 0$), auch Konstante oder englisch „intercept“ genannt.

β : mathematisch gesehen die Steigung der Regressionsgeraden zur Vorhersage von y. Im Kontext der Regressionsanalyse wird β als Regressionskoeffizient bezeichnet. Seine Ausprägung beantwortet die Frage nach der Stärke des Effekts von X auf Y. Das heißt, der Wert β bedeutet inhaltlich die durchschnittliche Veränderungsrate der y-Werte pro Zunahme einer Einheit von x-Werten.

X: Wert der unabhängigen Variable X für den der \hat{y} -Wert vorhergesagt wird.

Für die geschätzte Regressionsgerade des fiktiven Beispiels heißt das dann:

$$\text{Geschätzte Lebenszufriedenheit} = \alpha + \beta * \text{Einkommen}$$

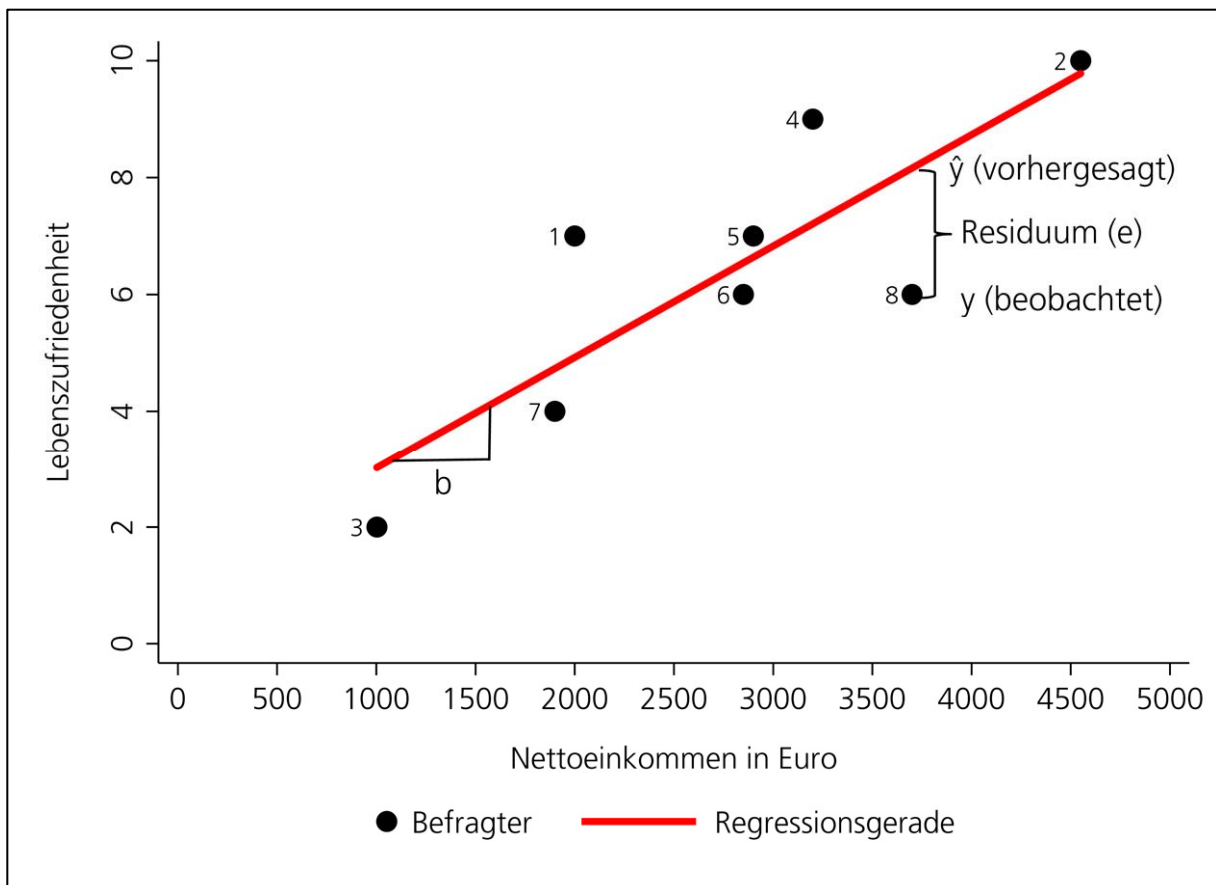
Abbildung 17 veranschaulicht das Verhältnis von tatsächlich beobachteten Werten und geschätzten Werten. Die Residuen sind Störungen, also die Abweichungen der Datenpunkte von der Regressionsgeraden. Unter Berücksichtigung der Residuen lässt sich für die Regressionsfunktion dann folgende Gleichung aufstellen:

$$y = \alpha + \beta \cdot x + e, \text{ wobei } e = y - \hat{y} \text{ entspricht}$$

(der Buchstabe e leitet sich vom englischen Wort „error“ ab)

Die Regressionsfunktion setzt sich damit zusammen aus der geschätzten Regressionsgeraden und den Residuen e. Für das verwendete Beispiel heißt das, dass sich die beobachtete Lebenszufriedenheit durch die mit der Regressionsgeraden geschätzten Lebenszufriedenheit und einem Rest, der sich durch das Einkommen nicht erklären lässt, ergibt.

Abbildung 17: Streudiagramm mit OLS-Regressionsgerade



Quelle: Eigene Darstellung

Doch wie erhält man diese Regressionsgerade? Durch die bestehende Punktwolke könnten durchaus verschiedene Geraden gezogen werden. Um die Varianz aber bestmöglich zu erklären, wird diejenige Gerade gesucht, bei der der Abstand aller beobachteten Punkte zur Geraden minimal ist. Damit eine Gleichung die beobachteten Daten bestmöglich beschreibt, müssen also die Fehler bzw. Residuen minimiert werden.

Ordinary Least Square-Verfahren

Dieser Abstand wird mathematisch durch Messung des vertikalen Abstands (bezeichnet als Residuen oder e) der beobachteten y -Werte von den vorhergesagten \hat{y} -Werten ermittelt. Um diesen Abstand von Vorzeichen unabhängig zu gestalten, wird er quadriert. Man nennt dieses Vorgehen auch Ordinary Least Square- oder OLS-Verfahren, im Deutschen mitunter auch Kleinste-Quadrat-Schätzung (KQ-Schätzung) oder „Methode der kleinsten Quadrate“. Die „beste“ Regressionsgerade erhält man also mit einem mathematischen Verfahren, das eine Konstante α und eine Steigung β schätzt sowie dabei die lineare Beziehung zwischen X und Y abbildet, indem die Quadrate der Residuen minimiert werden.

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \alpha + \beta * x_i)^2 = \text{Minimum}$$

Diese Gleichung lässt sich nun nach α und β ableiten (Gehring und Weins 2009, S. 181-184), sodass die Konstante sowie die Steigung der Regressionsgeraden berechnet werden können. In der Praxis übernimmt diese Aufgabe das Statistikprogramm. Für das fiktive Beispiel ergibt sich eine Regressionsgerade mit folgender Gleichung: $\hat{y} = 1,11 + 0,002x$ oder Lebenszufriedenheit = $1,11 + 0,002 \text{ Einkommen}$.

Interpretation der Koeffizienten

Wie bereits besprochen, gibt der Regressionskoeffizient β Auskunft über Stärke und Richtung des Einflusses. Sein Vorzeichen bestimmt, ob es sich um einen positiven oder negativen Zusammenhang handelt. Bezüglich der Stärke des Effekts ist festzuhalten, dass mit Zunahme der Variablen x (uV) um eine Einheit die Variable y (aV) um den Wert des Koeffizienten ansteigt. Dieser als „unstandardisiert“ bezeichnete Regressionskoeffizient berücksichtigt die Maßeinheiten der Variablen und hat den Vorteil, dass der Effekt leicht interpretierbar ist. In der Regel interessiert bei der Interpretation einer Regressionsanalyse vor allem der Regressionskoeffizient β , da er Informationen über die Stärke eines Einflusses gibt. Die Konstante stellt den Schnittpunkt der Regressionsgeraden mit der y -Achse dar und zeigt den Wert an, den die abhängige Variable annimmt, wenn die unabhängige Variable gleich 0 ist.

Im Falle des verwendeten Beispiels heißt das, dass jemand ohne Einkommen eine geschätzte Lebenszufriedenheit von 1,11 Skalenpunkten aufweisen würde. In der regressionsanalytischen Interpretationspraxis wird der Regressionskonstanten meist wenig Bedeutung beigemessen, da sie häufig außerhalb des interessierenden Wertebereichs liegt. In unserem Fall ergeben sich ein Regressionskoeffizient von 0,002 (β) und eine Konstante von 1,11 (α). Das heißt, dass jeder Euro Einkommen mehr im Monat einen Zuwachs von 0,002 auf der Lebenszufriedenheitsskala erbringt. Bei 1000 Euro Einkommen im Monat mehr entspricht das einer um zwei Skalenpunkte höheren Lebenszufriedenheit.

Bewertung des Regressionsmodells

Mit Hilfe der Methode der kleinsten Quadrate können also die Regressionsgerade, die den kleinsten Abstand zu den quadrierten Fehlern aufweist, sowie der entsprechende Regressionskoeffizient ermittelt werden. Wie gut erklärt nun aber die Regressionsgerade unsere beobachteten Datenpunkte insgesamt? Dies kann mit dem sogenannten „Bestimmtheitsmaß“ bzw. Determinationskoeffizienten R^2 geprüft werden. Dieses Maß gibt an, wie groß der Anteil der durch die Regressionsgerade erklärten Varianz der beobachteten Datenpunkte

ist. Mit dem Determinationskoeffizienten R^2 kann also regressionsanalytisch ein Parameter berechnet werden, der uns etwas über die Güte des Gesamtmodells verrät. Bei seiner Berechnung wird die Gesamtvarianz s_y^2 der abhängigen Variable y in zwei Teile zerlegt: erstens in die durch die Regressionsfunktion erklärte Varianz $s_{\hat{y}}^2$ und zweitens in die dadurch nicht erklärte „Restvarianz“. Daraus ergibt sich folgende Formel zur Berechnung:

$$R^2 = \frac{\text{Varianz der vorhergesagten Werte}}{\text{Varianz der beobachteten Werte}}$$

Das heißt:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Mit dem R^2 lässt sich dann bestimmen, wie viel Varianz durch das aufgestellte Regressionsmodell „erklärt“ werden kann, weshalb man auch von erklärter Varianz spricht. R^2 gehört damit zu den sogenannten PRE-Maßen (Proportional Reduction of Error), die Auskunft über die prozentuale Verringerung der Fehler durch ein Vorhersagemodell von y geben (siehe auch Abschnitt 3.5).

R^2 wird auch als „Prozentsatz der erklärten Varianz“ bezeichnet und lässt sich dadurch als ein Maß für die Güte der Anpassung der Regressionsfunktion an die beobachteten Daten verstehen. R^2 kann Werte zwischen 0 und 1 annehmen. Multipliziert man es mit 100, lässt sich die erklärte Varianz als Prozentwert ausdrücken. Ein Wert von 1 bedeutet folglich, dass 100 Prozent der Varianz aufgeklärt werden können. Damit würden im bivariaten Modell alle beobachteten Datenpunkte auf der prognostizierten Regressionsgerade liegen. Der Wert 0 sagt dagegen aus, dass die untersuchte unabhängige Variable nichts zur Erklärung der abhängigen Variable beiträgt. Je näher R^2 an 1 ist, desto besser erklärt das spezifizierte Modell die Streuung.

**Gütemaß
der Regression: R^2**

Rechnet man für unser Beispiel das Bestimmtheitsmaß aus, ergibt sich ein R^2 von 0,69, das heißt, 69 Prozent der Varianz unseres fiktiven Beispiels können durch die unabhängige Variable Einkommen erklärt werden. In der sozialwissenschaftlichen Praxis ist ein solch hohes R^2 selten, denn wie gesagt: Die soziale Wirklichkeit befindet sich jenseits von Laborbedingungen und damit sind die Modelle für Störungen anfällig. Ein niedriges R^2 kann auf einer fehlenden Integration wichtiger erklärender Merkmale beruhen, es kann aber auch einer fehlerhaften Operationalisierung von Variablen geschuldet sein (siehe zur Diskussion weiterer Fehlerquellen Urban und Mayerl 2018).

Sicherlich liegt ein Ziel sozialwissenschaftlicher Regressionsanalysen darin, die abhängige Variable möglichst gut erklären zu können. Allerdings sollte man davon absehen, das Bestimmtheitsmaß allein als Gütekriterium für eine durchgeführte Regressionsanalyse zu betrachten.

In diesem Abschnitt wurden Grundlagen der einfachen linearen Regression vorgestellt. Wird mit Stichprobendaten gearbeitet, können die Koeffizienten sowie das Gesamtmodell mit R^2 auch auf statistische Signifikanz getestet werden. So können wir herausfinden, ob die Ergebnisse, die wir auf Basis einer Zufallsstichprobe erzielt haben, mit einer gewissen Irrtumswahrscheinlichkeit auch auf die Grundgesamtheit zutreffen. Dieses Vorgehen werden wir im Kontext der multiplen linearen Regression im nächsten Abschnitt näher kennenlernen.

4.2.2 Multiple Regression

In sozialwissenschaftlichen Analysen können in der Regel viele Variablen einen Einfluss auf die abhängige Variable ausüben, sodass bivariate Modelle meist nicht ausreichen, um einen Sachverhalt zu erklären. Mit einer multiplen Regressionsanalyse können zusätzlich relevante (erklärende/unabhängige) Variablen integriert und ihr Einfluss berechnet bzw. geschätzt werden. Im Folgenden wird die multiple Regressionsanalyse als Erweiterung des zuvor beschriebenen einfachen bivariaten Regressionsmodells vorgestellt. Zusätzlich geht der Kapitelabschnitt auf die Interpretation der Regressionskoeffizienten, das Skalenniveau der unabhängigen Variablen sowie die Möglichkeit, Aussagen zur statistischen Signifikanz des Regressionsmodells vorzunehmen, ein. Zum Abschluss des Kapitels werden mit den sogenannten „BLUE-Annahmen“ Voraussetzungen vorgestellt, die aus statistischer Sicht gegeben sein müssen, um eine möglichst unverzerrte und effiziente lineare Regressionsanalyse durchzuführen.

Berücksichtigung mehrerer Variablen

Die multiple Regressionsanalyse ermöglicht es mit dem Einschluss mehrerer erklärender Variablen erstens, ein theoretisch basiertes und empirisch bestmögliches Modell zur Erklärung bzw. Vorhersage von y zu schätzen. Zweitens lässt sich die Stärke des Einflusses einzelner erklärender Variablen auf die abhängige Variable unabhängig voneinander untersuchen. Indem wir alle theoretisch interessierenden Variablen in die Regressionsanalyse aufnehmen, erhalten wir die Möglichkeit, gewissermaßen den „Netto-Effekt“ einer unabhängigen Variable, auf das zu erklärende Phänomen zu ermitteln. Wir halten die „konkurrierenden“ erklärenden Variablen in der Regressionsanalyse konstant, um die Auswirkung der Änderung einer Variable zu untersuchen. Dies nennt man auch für diese Variablen „kontrollieren“.

Prinzipiell gilt, dass das im bivariaten Beispiel kennengelernte Grundmodell erhalten bleibt, jetzt allerdings die Regressionsgeraden um die zusätzlichen unabhängigen Variablen erweitert werden müssen. Eine solche multiple Regressionsgleichung sieht formal folgendermaßen aus:

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k + e$$

Wir gehen zwar davon aus, dass die abhängige Variable durch den Einschluss weiterer theoretisch relevanter Variablen besser erklärt werden kann als durch einen einzelnen Prädiktor. Jedoch kann nicht erwartet werden, dass diese Faktoren das Phänomen vollständig erklären. Auch hier gibt es einen nicht erklärten Rest, der in der Regressionsgleichung durch die Residuen (e) ausgedrückt wird. Im multiplen Fall versuchen jetzt also mehrere Prädiktoren die Kriteriumsvariable zu erklären und schätzen gemeinsam das multiple Regressionsmodell. Damit wird auch die zweidimensionale Betrachtungsweise verlassen, da wir nicht mehr eine Regressionsgerade erhalten, sondern so viele Regressionskoeffizienten wie unabhängige Variablen in die Regressionsgleichung eingehen. Diese k -Variablen spannen einen mehrdimensionalen Raum auf und erstellen eine Regressions-(hyper)-Ebene:

$$\hat{y}_i = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k$$

Dadurch wird eine grafische Darstellung – je nach Anzahl der einbezogenen unabhängigen Variablen – schwierig. Drei Ziele lassen sich für die in der Sozialforschung praktizierte Durchführung multipler Regressionsanalysen formulieren:

1. die abhängige Variable (aV) mit den einbezogenen theoretisch relevanten Variablen (uVs) bestmöglich zu erklären
2. die Richtung und Stärke einzelner Einflussfaktoren zu ermitteln
3. herauszufinden, ob das anhand von Stichprobendaten vorhergesagte Regressionsmodell sowie die Einflüsse der einzelnen Prädiktoren auch auf die Grundgesamtheit übertragen werden kann, also statistisch signifikant ist.

Auch für die im letzten Abschnitt kennengelernte abhängige Variable „Lebenszufriedenheit“ können wir davon ausgehen, dass noch weitere Faktoren einen Effekt ausüben (Neller 2004; Weick 2012). Wir wollen an dieser Stelle drei exemplarische Merkmale berücksichtigen: den wahrgenommenen Gesundheitszustand einer Person, die Bildung sowie den Familienstand. Folgende Hypothesen werden formuliert:

H2: Je positiver die Einschätzung des eigenen Gesundheitszustands ist, desto größer ist die Lebenszufriedenheit.

H3: Je gebildeter eine Person ist, desto größer ist die Lebenszufriedenheit.

Eine (subjektiv) besser eingeschätzte Gesundheit sollte sich günstig auf die Lebenszufriedenheit auswirken. Schließlich können eine Krankheit oder sonstige Gebrechen die persönliche Lebensgestaltung einschränken und wünschenswerte Aktivitäten unmöglich machen. Eine ähnliche Argumentation bietet sich auch bei einem vermuteten Zusammenhang zwischen Bildung und Lebenszufriedenheit an. Eine höhere Bildung fördert die Teilhabe- und Gestaltungsmöglichkeiten in vielen Bereichen des privaten und öffentlichen Lebens (z.B. Mitgliedschaft in Vereinen). Dies sollte sich günstig auf die Lebenszufriedenheit auswirken.⁹ Wenn ich einen Partner bzw. eine Partnerin an meiner Seite habe, bin ich weniger einsam und steigere dadurch mein soziales Wohlempfinden. Deshalb erwarten wir für unsere Analyse:

H4: Verheiratete Personen sind zufriedener mit ihrem Leben als nicht-verheiratete Personen.

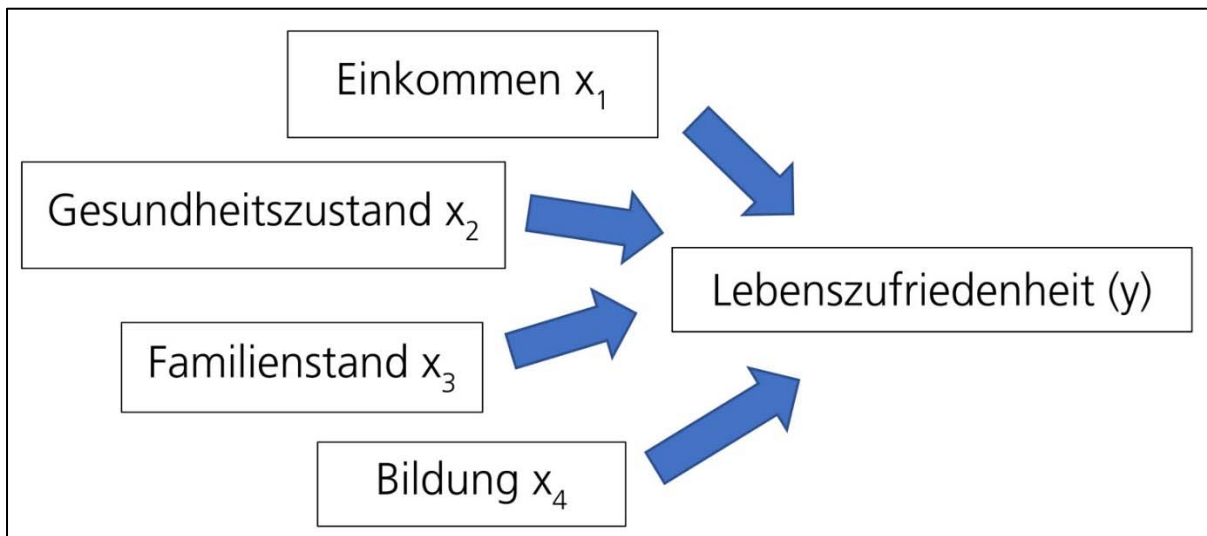
Familiensoziologische Analysen sprechen den Effekt auch nicht-verheirateten, aber zusammenlebenden Menschen zu. Allerdings zeigt sich in vielen Studien häufig, dass die institutionalisierte Ehe tatsächlich einen stärkeren Effekt ausübt (z.B. Lee und Ono 2012), auch wenn der dahinterstehende Mechanismus noch immer nicht ganz klar ist.

Gleichzeitig lässt sich am Merkmal Familienstand noch ein weiterer wichtiger Aspekt bei der Durchführung und Modellierung von Regressionsanalysen erörtern und erneut auf die Wichtigkeit von theoretischer Vorarbeit verweisen. Denn in welche Richtung geht nun die Kausalität? Ist der Umstand, dass jemand verheiratet ist, ausschlaggebend für die höhere Lebenszufriedenheit oder geht der „Pfeil“ gewissermaßen in die umgekehrte Richtung, also gehen lebenszufriedene Menschen eher eine Partnerschaft ein als Unzufriedene (bzw. finden sie womöglich eher einen Partner,

⁹ In einer Hausarbeit müssen die formulierten Hypothesen natürlich ausführlicher begründet und mit Forschungsliteratur verknüpft werden.

weil sie zufriedener sind). Diese Fragen lassen sich mit Querschnittsdaten nicht abschließend klären. Dazu greift die Soziologie dann auf sogenannte Ereignisdatenanalysen zurück, die auf Basis von Paneluntersuchungen durchgeführt werden können. Allerdings lässt sich auf Basis der oben ausgeführten theoretischen Überlegungen H4 durchaus vertreten. Eine theoretische und methodische Reflexion möglicher Limitationen der eigenen Annahmen und des eigenen Datenmaterials sollten dennoch erfolgen. Abbildung 18 illustriert die angenommenen Beziehungen zwischen den unabhängigen und der abhängigen Variablen schematisch.¹⁰

Abbildung 18: Schematische Darstellung der vermuteten multivariaten Einflussstruktur



Quelle: Eigene Darstellung

Übertragen in eine multiple Regressionsgleichung wird nun also davon ausgegangen, dass:

$$\text{Lebenszufriedenheit} = \alpha + \beta_1 * \text{Einkommen} + \beta_2 * \text{Gesundheitszustand} + \beta_3 * \text{Familienstand} + \beta_4 * \text{Bildung} + e$$

Mit Einschluss dieser vier unabhängigen Merkmale in das Regressionsmodell erhalten wir die bereinigten Effekte jeder einzelnen Variable unter Konstanthaltung oder „Kontrolle“ der jeweils anderen Variablen.

Die Berechnung der Regressionskoeffizienten im multiplen Regressionsmodell erfordert fortgeschrittene Mathematikkenntnisse (z. B. Matrixalgebra). Im Forschungsalltag wird die Berechnung der Regressionsparameter von Statistikprogrammen wie SPSS, Stata oder R übernommen. Auf Basis der Daten der ALLBUS werden diese Berechnungen ausgeführt und in Regressionstabellen, wie man sie auch in sozialwissenschaftlichen Publikationen findet, übertragen (siehe Tabelle 49 und Tabelle 50).

Dummy-Kodierung

Alle Merkmale, die wir für theoretisch relevant erachten und in den Hypothesen aufgenommen wurden, können mit der ALLBUS 2016 operationalisiert werden.

¹⁰ Wir können davon ausgehen, dass auch einige der unabhängigen Variablen untereinander korrelieren (z.B. Bildung und Einkommen). Abbildung 18 illustriert allerdings nur die angenommene Kausalstruktur zwischen den unabhängigen Variablen und der abhängigen Variable.

Allerdings wird deutlich, dass wir es nicht bei allen erklärenden Variablen mit metrisch skalierten Variablen zu tun haben. Das lineare Regressionsmodell setzt allerdings implizit voraus, dass nicht nur die abhängige, sondern auch die erklärenden Variablen metrisches Messniveau aufweisen, damit der lineare Zusammenhang abgebildet werden kann. Damit auch ordinal- oder nominalskalierte erklärende Variablen ohne Informationsverlust in ein Regressionsmodell integriert werden können, müssen sie entsprechend umgeformt werden. Hier kommt die sogenannte „Dummy-Kodierung“ zum Einsatz, bei der die kategorialen Variablen in verschiedene dichotome Variablen mit den Ausprägungen 0 („Merkmal liegt nicht vor“) und 1 („Merkmal liegt vor“) transformiert werden. Diese neu konstruierten dichotomen Variablen nennt man Dummy-Variablen (oder Design- bzw. Indikatorvariablen). Das heißt, dass man die betreffenden nicht metrisch vorliegenden Variablen so transformiert, dass sie ohne Informationsverlust in der Regressionsanalyse mitberücksichtigt werden können. Dies gilt in unserem Beispiel für die Variablen Familienstand sowie Bildung.

Der Familienstand ist in der ALLBUS 2016 folgendermaßen kodiert: 1) verheiratet, zusammenlebend, 2) verheiratet, getrennt lebend, 3) verwitwet, 4) geschieden, 5) ledig, 6) in eingetragener Lebenspartnerschaft, zusammenlebend, 7) in eingetragener Lebenspartnerschaft, getrennt lebend 8) Lebenspartner/in verstorben, 9) Lebenspartnerschaft aufgehoben.

Zunächst wurde die Variable umkodiert, da aus theoretischen Gründen eine Trennung zwischen hetero- und homosexuellen Partnerschaften nicht angezeigt ist.¹¹ Danach wurden aus den verbleibenden fünf Variablen die Dummy-Variablen gebildet. Bei insgesamt k Ausprägungen einer kategorialen Variable sind im Prinzip k Dummies denkbar (in unserem Fall handelt es sich um fünf Ausprägungen). Praktisch sind aber lediglich $(k - 1)$ Dummies nötig, um die k -Ausprägungen vollständig abzubilden. Im vorliegenden Fall wurden die Dummy-Variablen „verheiratet“, „getrennt“, „geschieden“ und „verwitwet“ gebildet. Tabelle 48 zeigt die vorgenommene Dummy-Kodierung für die verschiedenen Möglichkeiten.

Tabelle 48: Dummy-Kodierung für Familienstand

Dummy-Variable	verheiratet	getrennt	geschieden	verwitwet
geschieden	0	0	1	0
verheiratet	1	0	0	0
verwitwet	0	0	0	1
getrennt	0	1	0	0
ledig	0	0	0	0

Quelle: Eigene Darstellung

¹¹ Zudem sind die Fallzahlen bei den Kategorien zur eingetragenen Lebenspartnerschaft sehr niedrig, so dass eine Umkodierung auch aus empirischer Sicht sinnvoll ist. Die Umkodierung fasste die Werte 6 bis 9 mit ihren heterosexuellen Pendants zusammen, so dass fünf Kategorien bleiben: verheiratet, zusammenlebend, getrennt lebend, verwitwet, geschieden und ledig.

Die nicht dichotomisierte k-te Ausprägung (in unserem Beispiel „ledig“) erkennt man daran, dass sie auf allen anderen Dummies den Wert 0 aufweist. Dadurch wird sie zur sogenannten „Referenzkategorie“ im linearen Regressionsmodell, was bedeutet, dass die Regressionskoeffizienten der anderen Merkmale in Bezug auf dieses Merkmal interpretiert werden. So kann auch Hypothese 4, die dem Verheiratetsein einen positiven Effekt auf die Lebenszufriedenheit zuschreibt, untersucht werden. Hinsichtlich des gesamten Regressionsmodells bedeutet das:

$$\begin{aligned} \text{Lebenszufriedenheit} = & \alpha + \beta_1 * \text{Einkommen} + \beta_2 * \text{Gesundheitszustand} + \beta_3 * \text{verheiratet} \\ & + \beta_4 * \text{getrennt} + \beta_5 * \text{geschieden} + \beta_6 * \text{verwitwet} + \beta_7 * \text{Bildung} \\ & + e \end{aligned}$$

Weitere Operationalisierung

Zur weiteren Operationalisierung in unserem Beispiel: Dummy-Variablen müssen auch für die kategoriale Variable Bildung und ihre Ausprägungen erzeugt werden. Dabei werden zuvor die Ausprägungen der Variable „Allgemeiner Schulabschluss“ so rekodiert, dass die Kategorien „Hauptschulabschluss“, „Mittlere Reife“, „(Fach-)Hochschulreife“ und „Sonstiges“ übrig bleiben. Diese vier Merkmale werden in Dummy-Variablen mit der Referenzkategorie „Mittlere Reife“ transformiert.

Der wahrgenommene Gesundheitszustand der Befragten, der mit einer Fünf-Punkt-Skala (von 0 „sehr gut“ bis 5 „schlecht“) gemessen wurde, wird als pseudometrische Variable in die Regressionsanalyse aufgenommen. Allerdings wird die Skala auf zweierlei Weise transformiert: So weisen höhere Werte nun auch einen als besser wahrgenommenen Gesundheitszustand auf (0 „schlecht“, 5 „sehr gut“). Dies erleichtert zum einen die Interpretation des entsprechenden Regressionskoeffizienten, denn schließlich wird in H2 davon ausgegangen, dass die Lebenszufriedenheit desto höher ist, je besser der Gesundheitszustand eingeschätzt wird. Gleichzeitig wird der schlechteste Zustand mit dem Wert „0“ kodiert. Dass ein schlechter Gesundheitszustand nun mit dem Wert „0“ kodiert wird, hat zum anderen den Vorteil, dass auch die Regressionskonstante (also der Schnittpunkt der Geraden mit der y-Achse für $x = 0$) sinnvoll interpretiert werden kann. Zuletzt wird das monatliche Nettoeinkommen noch in die Variable wöchentliches Nettoeinkommen transformiert, damit Effekte leichter illustriert werden können.

Nachdem nun alle unabhängigen Variablen für die Analyse operationalisiert sind, wird zunächst auf Basis der ALLBUS das vom letzten Abschnitt bekannte bivariate Regressionsmodell mit der unabhängigen Variable Einkommen (wöchentlich) berechnet, um dann auch ermitteln zu können, inwieweit der Einbezug der weiteren unabhängigen Variablen eine Verbesserung des Modells insgesamt mit sich bringt. Das einfache bivariate Modell ergibt ein R^2 von 0,033 und einen (unstandardisierten) Regressionskoeffizienten β von 0,001. Das heißt, dass die Variable „Nettoeinkommen“ etwas mehr als drei Prozent der Varianz der beobachteten Lebenszufriedenheit erklären kann. Es bedeutet zudem, dass ein Euro Nettoeinkommen mehr pro Woche die Lebenszufriedenheit um 0,001 Skalenpunkte erhöht. Was ändert sich nun, wenn die drei weiteren unabhängigen Variablen in das Modell integriert werden? In Tabelle 49 sind die Ergebnisse dokumentiert.

Interpretation der Ergebnisse

Wie lassen sich diese Ergebnisse vor dem Hintergrund der aufgestellten Hypothesen interpretieren? Alle Vorzeichen weisen in die erwartete Richtung. Das Nettoeinkommen hat (wie schon im bivariaten Fall) einen positiven Effekt von 0,001. Der wahr-

genommene Gesundheitszustand zeigt ebenfalls den erwarteten positiven Effekt: Mit einer Verbesserung des subjektiven Gesundheitszustands um einen Skalenpunkt geht nach unserer Regressionsschätzung eine um 0,69 Skalenpunkte erhöhte Lebenszufriedenheit einher. Beim Familienstand und der Bildung zeigt sich, dass Verheiratete eine um 0,6 Skalenpunkte höhere Lebenszufriedenheit im Vergleich zu Ledigen aufweisen. Menschen mit (Fach-)Hochschulreife sind im Vergleich zu Personen mit mittlerer Reife um 0,18 Skalenpunkte zufriedener mit ihrem Leben. Die formulierten Hypothesen können im Hinblick auf die ALLBUS-Stichprobe bestätigt werden. Abschließend betrachten wir noch die Konstante, deren Koeffizient etwas mehr als 5 aufweist – wie lässt sich dies interpretieren? Der Wert entspricht der vorhergesagten Lebenszufriedenheit einer Person, die auf allen unabhängigen Variablen den Wert 0 aufweist, das heißt eine Person, die kein wöchentliches Einkommen hat, bei sich selbst einen schlechten Gesundheitszustand konstatiert, die ledig ist und deren höchster Schulabschluss die mittlere Reife ist. Wie bereits beim Beispiel des einfachen Regressionsmodells besprochen, wird die Regressionskonstante in aller Regel nicht interpretiert.

Tabelle 49: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 1)

	unstand. Koeffizienten	standard. Koeffizienten	Standard- fehler	t-Wert	p-Wert
Wöchentliches Einkommen	0,001	0,090	0,000	5,250	0,000
Gesundheitszustand	0,685	0,382	0,030	22,467	0,000
<i>Familienstand (Referenz: ledig)</i>					
verheiratet	0,600	0,167	0,070	8,585	0,000
geschieden	0,017	0,003	0,110	0,150	0,881
getrennt	-0,216	-0,016	0,220	-0,981	0,327
verwitwet	0,505	0,072	0,127	3,992	0,000
<i>Bildung (Referenz: Mittlere Reife)</i>					
Hauptschulabschluss	-0,029	-0,007	0,076	-0,385	0,700
(Fach-)Hochschulreife	0,177	0,048	0,070	2,525	0,012
Sonstiges	-0,050	-0,003	0,234	-0,212	0,832
Konstante	5,294		0,111	47,627	0,000
R ²	0,201				
Korrigiertes R ²	0,199				
N	3489				

Daten: ALLBUS 2016. Eigene Berechnungen

Die Güte des Gesamtmodells hat sich – ablesbar am R^2 – im Vergleich zum bivariaten Fall ebenfalls substanziell verbessert: Fast 20 Prozent, also ein Fünftel der Varianz, können durch die vier Variablen erklärt werden. Das entspricht einem Zuwachs von 17 Prozentpunkten im Vergleich zum bivariaten Modell mit dem Einkommen als einziger unabhängiger Variable. Da die Formel für das Bestimmtheitsmaß R^2 abhängig ist von der Anzahl der im Modell aufgenommenen Prädiktoren (je mehr Variablen in einem Regressionsmodell berücksichtigt werden, desto größer ist das R^2), wird bei multivariaten Regressionsanalysen meist das sogenannte korrigierte R^2 berichtet, um eine verlässliche Aussage über die globale Güte des Modells machen zu können. Da bei der Berechnung des korrigierten R^2 die Anzahl der unabhängigen Variablen berücksichtigt wird, fällt es immer etwas niedriger aus als das unkorrigierte R^2 .

Standardisierte Koeffizienten

Die Hypothesen können also auf Basis der unstandardisierten Koeffizienten mit der sozialen Wirklichkeit unserer Stichprobe abgeglichen werden. Will man allerdings zusätzlich die Einflussstärke verschiedener Variablen miteinander vergleichen, stößt man mit unstandardisierten Regressionskoeffizienten auf Schwierigkeiten, da diese ja mit Hinblick auf die ihnen zugrundeliegenden Maßeinheiten interpretiert werden müssen (z. B. Euro oder Skalenpunkte). Um Effektstärken der unterschiedlichen unabhängigen Variablen im multiplen Regressionsmodell dennoch miteinander vergleichen zu können, wird in der Forschungspraxis häufig zu sogenannten „standardisierten“ Koeffizienten gegriffen. Diese werden von Statistikprogrammen automatisch mit ausgegeben, doch wie kommen sie zustande? Sie lassen sich auf zwei verschiedene Arten berechnen. Zum einen, indem die unabhängigen Variablen vor Berechnung der Regressionsgleichung z-transformiert, also standardisiert, werden. Dabei werden sie so umgeformt, dass sie ein arithmetisches Mittel von 0 und eine Standardabweichung von 1 aufweisen. Wie alle z-transformierte oder standardisierte Variablen reichen die Werte der standardisierten Regressionskoeffizienten von -1 bis $+1$. Damit ermöglichen sie auch eine Vereinheitlichung der Regressionskoeffizienten. Standardisierte Koeffizienten erhält man zum anderen, wenn jeder der unstandardisierten Regressionskoeffizienten mit seiner Standardabweichung multipliziert und durch die Standardabweichung der abhängigen Variable geteilt wird. Damit sind nun alle Koeffizienten vereinheitlicht, wenn auch ihre Interpretation nun nicht mehr so anschaulich und leicht nachvollziehbar erfolgen kann, sondern Bezug auf die Standardisierung nehmen muss.

Ein standardisierter Regressionskoeffizient gibt nun Auskunft darüber, um wie viele Standardabweichungen sich die abhängige Variable verändert, wenn das unabhängige Merkmal um eine Standardabweichung ansteigt. Tabelle 49 zeigt in der zweiten Spalte die standardisierten Regressionskoeffizienten, auch Beta genannt. Für den Gesundheitszustand lautet die Interpretation: Verbessert sich der wahrgenommene Gesundheitszustand einer Person um eine Standardabweichung auf der Fünf-Punkt-Skala, dann steigt die Lebenszufriedenheit dieser Person um 0,38 Standardabweichungen. Die standardisierten Regressionskoeffizienten ermöglichen uns nun auch, die Einflüsse der vier untersuchten erklärenden Merkmale miteinander zu vergleichen. Es zeigt sich, dass der Gesundheitszustand den stärksten Effekt auf die Lebenszufriedenheit ausübt, gefolgt vom Familienstand „verheiratet“ (im Vergleich zu „ledig“). Es wird auch deutlich, dass das Nettoeinkommen den geringsten Effekt ausübt.

Die Verwendung standardisierter Koeffizienten ist aus statistischer Sicht nicht ganz unproblematisch (Wolf und Best 2015; Kohler und Kreuter 2017, S. 288; Urban und Mayerl 2018). Beachtet werden sollte bei der Verwendung der standardisierten Koeffizienten, dass sie nur innerhalb einer Stichprobe miteinander verglichen werden können. Würden wir eine ähnliche Regressionsanalyse auf Basis der Daten des European Social Surveys durchführen, sollten wir nicht unhinterfragt die standardisierten Regressionskoeffizienten der beiden Regressionsmodelle miteinander vergleichen. Die Überprüfung vorab formulierter Hypothesen kann sowohl mit standardisierten als auch mit unstandardisierten Regressionskoeffizienten vorgenommen werden. Richtung und Stärke eines möglichen Effekts unterscheiden sich nur in der konkreten Interpretation, nicht aber in der Substanz.

Wie ausgeführt, können die aufgestellten Hypothesen für die ALLBUS-Stichprobendaten bestätigt werden. Aber lassen sich die Ergebnisse auch auf die Grundgesamtheit übertragen? Da der ALLBUS eine Zufallsstichprobe der deutschen Wohnbevölkerung darstellt, könnten bei Vorliegen statistischer Signifikanz auch Aussagen über die gesamte deutsche Wohnbevölkerung getroffen werden. Sind also die Ergebnisse der Regressionsanalyse statistisch signifikant oder womöglich nur zufällig in unserer Stichprobe vorhanden? Dies ist aus Sicht der empirischen Sozialforschung eine zentrale Frage, da sie zum Ziel hat, Zusammenhänge zu verallgemeinern.

Statistische Signifikanz

Mittels inferenzstatistischer Verfahren kann bei einer Regressionsanalyse die statistische Signifikanz des Gesamtmodells sowie der einzelnen Regressionskoeffizienten geprüft werden. Die letzten drei Spalten von Tabelle 49 geben Parameter an, die sich mit der statistischen Signifikanz der Einflussfaktoren beschäftigen. Mit ihrer Hilfe können wir Aussagen darüber machen, ob das Gesamtmodell sowie die Effekte der unabhängigen Variablen auch auf die Grundgesamtheit übertragen werden können. Das heißt, dass bei Vorliegen statistischer Signifikanz die Passgenauigkeit des Regressionsmodells insgesamt sowie die Effektstärken der einzelnen unabhängigen Merkmale als Schätzer der „wahren“ (in der Grundgesamtheit vorliegenden) Zusammenhänge verstanden werden – selbstverständlich mit einer gewissen statistischen Irrtumswahrscheinlichkeit versehen. Hierzu werden zwei verschiedene Hypothesentests durchgeführt: das Gesamtmodell wird mit Hilfe der F-Statistik geprüft und t-Tests überprüfen, inwieweit die einzelnen Regressionskoeffizienten von 0 verschieden sind.

Bei der Berechnung der Prüfgröße F werden wie bei der Berechnung des Determinationskoeffizienten R^2 die erklärte und die nicht-erklärte Streuung des Modells sowie der Stichprobenumfang und die Anzahl der Einflussfaktoren berücksichtigt (Urban und Mayerl 2018, S. 125-129). Der F-Test prüft, inwieweit die ins Regressionsmodell aufgenommenen unabhängigen Variablen zusammengekommen einen statistisch signifikanten Beitrag zur Erklärung der abhängigen Variable leisten. Die Ergebnisse des F-Tests werden in sozialwissenschaftlichen Artikeln, in denen Regressionsanalysen angewandt werden, meist nicht berichtet. Allerdings sollte – wenn keine statistische Signifikanz vorliegt – überlegt werden, inwieweit das Regressionsmodell möglicherweise falsch spezifiziert ist: inwieweit üben die einbezogenen unabhängigen Variablen einen linearen Effekt auf die abhängige Variable aus oder wurde womöglich eine theoretisch relevante Variable vergessen? Hier müssen dann zunächst die theoretischen Annahmen

Berechnung der Prüfgröße F

überprüft werden. Empirisch kann bivariat geprüft werden, ob die entsprechenden unabhängigen Variablen tatsächlich mit der abhängigen Variablen in einem linearen Zusammenhang stehen.

Berechnung der t-Werte

Auch die einzelnen Regressionskoeffizienten können auf ihre statistische Signifikanz getestet werden. Da die theoretisch abgeleiteten Hypothesen den Einfluss eines bestimmten Merkmals vermuten, ist dies für die Interpretation der Ergebnisse meist relevanter als die Prüfung des Gesamtmodells. Hier wird nun zunächst der Standardfehler des (unstandardisierten) Regressionskoeffizienten inferenzstatistisch ermittelt und auf seiner Basis dann der sogenannte t-Wert berechnet. Mit Hilfe des Standardfehlers von β lässt sich ein Konfidenzintervall oder „Vertrauensbereich“ ermitteln, also ein Bereich, von dem wir auf Basis von inferenzstatistischen Annahmen davon ausgehen können, dass in 95 Prozent aller Schätzungen der „wahre“ Regressionskoeffizient nicht weiter als $\pm 1,96$ Standardfehler entfernt zu finden ist. Um dieses Konfidenzintervall zu berechnen, wird die Student-t-Verteilung (anstelle der Standardnormalverteilung) genutzt. Der berichtete t-Wert lässt also Rückschlüsse darauf zu, dass der berechnete Regressionskoeffizient tatsächlich auch auf die Grundgesamtheit übertragen werden kann. Mit Hilfe der t-Tabelle kann dann auch der p-Wert (letzte Spalte in Tabelle 49) angegeben werden. Am p-Wert für den zweiseitigen t-Test lässt sich ablesen, wie hoch die Irrtumswahrscheinlichkeit bei einer Übertragung von der Stichprobe auf die Grundgesamtheit ist.

Statistikprogramme berechnen bei regressionsanalytischen Verfahren automatisch die Standardfehler für die Konstante und die einzelnen (unstandardisierten) Regressionskoeffizienten. Der Standardfehler (engl. standard error) wird auch mit S. E. abgekürzt. Auf Basis der Werte des unstandardisierten Regressionskoeffizienten und des Standardfehlers können die t-Werte für die Regressionskoeffizienten berechnet und auf deren Basis dann entschieden werden, ob der Befund einer Stichprobe (mit einer gewissen Irrtumswahrscheinlichkeit) auf die Grundgesamtheit übertragen werden darf.

Der t-Wert berechnet sich dabei wie folgt:

$$t = \frac{\text{unstandardisierter Regressionskoeffizient}}{\text{Standardfehler des Regressionskoeffizienten}}$$

In den Sozialwissenschaften werden meist drei Signifikanzniveaus unterschieden (siehe ausführlich Abschnitt 5.4): p-Wert kleiner als 0,05 (Irrtumswahrscheinlichkeit liegt bei 5 Prozent, abgekürzt mit einem Sternchen *), p-Wert kleiner als 0,01 (Irrtumswahrscheinlichkeit liegt bei 1 Prozent, abgekürzt mit zwei Sternchen **) und p-Wert kleiner als 0,001 (Irrtumswahrscheinlichkeit liegt bei 0,1 Prozent, abgekürzt mit drei Sternchen ***). Auf Grundlage dieser Konvention spricht man dann bei Vorliegen der entsprechenden p-Werte von statistischer Signifikanz; Damit können die beobachteten Zusammenhänge und Einflüsse von der Stichprobe auf die Grundgesamtheit übertragen werden. Tabelle 50 zeigt die unterschiedlichen Signifikanzniveaus mit der Anzahl der Sternchen an. Tabellen dieser Art finden sich häufig in empirischen Artikeln. Befindet sich der Betrag des empirisch ermittelten t-Werts für den betreffenden Parameter über dem für das gewählte Signifikanzniveau kritischen t-Wert (bei $p < 0,05$ liegt dieser bei $\pm 1,96$), so sprechen wir von einem statistisch signifikanten Koeffizienten.

Für die in unserem Beispiel untersuchten Variablen „Einkommen“, „Gesundheitszustand“, „Fachhochschulreife“ und „Verheiratetsein“ sehen wir, dass die entsprechenden Regressionskoeffizienten statistisch signifikant sind. Das heißt, die Stärke der Einflussfaktoren ist (unter Berücksichtigung des Standardfehlers sowie der inferenzstatistischen Irrtumswahrscheinlichkeit) nicht durch Zufall zustande gekommen, sondern lässt sich auch in der Grundgesamtheit finden bzw. auf die Grundgesamtheit übertragen.

Tabelle 50: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 2)

	unstandardisierte Koeffizienten	standardisierte Koeffizienten (Beta)
Wöchentliches Einkommen	0,001***	0,090
Gesundheitszustand	0,685***	0,382
<i>Familienstand (Referenz: ledig)</i>		
Verheiratet	0,600***	0,167
Geschieden	0,017	0,003
Getrennt	−0,216	−0,016
Verwitwet	0,505***	0,072
<i>Bildung (Referenz Mittlere Reife)</i>		
Hauptschulabschluss	−0,029	−0,007
(Fach-)Hochschulreife	0,177*	0,048
Sonstiges	−0,050	−0,003
Konstante	5,294***	
R ²	0,201	
Korrigiertes R ²	0,199	
N	3489	

Anmerkungen: Multiple lineare Regression. Dargestellt sind unstandardisierte und standardisierte Regressionskoeffizienten. Signifikanzniveaus: * = $p < 0,05$; ** = $p < 0,01$; *** = $p < 0,001$. Daten: ALLBUS 2016. Eigene Berechnungen

Um tatsächlich eine möglichst unverzerrte und effiziente lineare Regressionsanalyse durchführen zu können, müssen verschiedene statistische Voraussetzungen erfüllt sein. Diese sind in der Literatur als sogenannte BLUE-Annahmen (*Best Linear Unbiased Estimator*) nach dem Gauss-Markov-Theorem bekannt. Sind sie gegeben, liefert die Methode der kleinsten Quadrate bestmögliche Vorhersagen. Einige beziehen sich auf die Vorhersage der Regressionsfunktion insgesamt, andere auf die Aussagen zur statistischen Signifikanz. Diese Annahmen sollte man in der eigenen Analysepraxis mittels einer sogenannten „Regressionsdiagnostik“ überprüfen (Urban und Mayerl 2018). Auf folgende Voraussetzungen möchten wir besonders hinweisen:

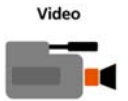
Voraussetzungen für die Regression

1. Möchte man eine lineare Regressionsbeziehung analysieren, sollte zwischen den einzelnen unabhängigen Variablen und der abhängigen Variable tatsächlich ein linearer Zusammenhang bestehen. Dies lässt sich am besten mit Streudiagrammen wie in Abbildung 16 kontrollieren. Sollte die Linearitätsannahme verletzt sein, gibt es verschiedene Möglichkeiten, damit umzugehen (siehe Urban und Mayerl 2018).
2. Damit die vorhergesagten Regressionsparameter auch auf die Grundgesamtheit übertragen werden können, muss es sich bei der Stichprobe um eine Zufallsstichprobe handeln.
3. Multikollinearität sollte vermieden werden. Multikollinearität bedeutet, dass einige der erklärenden Variablen untereinander sehr hoch korrelieren. Ist dies der Fall, wird eine unverzerrte Schätzung der Regressionsparameter beeinträchtigt. Zwar lassen sich Korrelationen zwischen unabhängigen Variablen nicht vollkommen vermeiden, diese sollten aber vorab getestet werden. Bei Korrelationen von $r > 0,9$ sollte eine der Variablen ausgeschlossen werden.
4. Die Residuen sollten zufällig auftreten und sich gegenseitig „ausgleichen“, also den Erwartungswert 0 haben. Sie sollten nicht miteinander und auch nicht mit der abhängigen Variablen Y korreliert sein.
5. Es sollte Homoskedastizität vorliegen, das heißt, die Varianz der Residuen sollte konstant sein. Ist diese Annahme verletzt, spricht man von Heteroskedastizität. Heteroskedastizität kann dazu führen, dass der Standardfehler der Regressionskoeffizienten nicht mehr korrekt berechnet werden kann und dadurch auch inferenzstatistische Aussagen zur Signifikanz erschwert werden.
6. Schließlich sollte auf eine korrekte Spezifikation des Modells geachtet werden, das heißt, alle theoretisch relevanten Faktoren sollten im Modell enthalten sein.

Diese Regressionsannahmen können (bis auf Annahme 6, die theoretisch angegangen werden muss) mittels regressionsdiagnostischer Verfahren überprüft werden. Das Verfahren der OLS-Regression ist robust genug, um leichtere Verletzungen dieser Annahmen kompensieren zu können. Sollten die Voraussetzungen allerdings stark verletzt sein, ist eine lineare OLS-Regression womöglich wenig geeignet, um die Einflussstruktur der Variablen zu beschreiben.

4.3 Logistische Regression

Im vorangegangenen Abschnitt wurde das Modell der linearen Regressionsanalyse vorgestellt. Um diese durchführen zu können, muss die abhängige Variable mindestens „pseudometrisch“ skaliert sein. Häufig ist das bei interessierenden sozialwissenschaftlichen Merkmalen allerdings nicht der Fall. Viele Phänomene, die man erklären möchte, sind binär: Geht jemand zur Wahl oder nicht, haben Bürgerinnen eine Präferenz für eine bestimmte Partei oder nicht, ist eine Person Raucher oder Nichtraucher? Aus Sicht der empirischen Sozialforschung möchte man zum Beispiel ermitteln, wer nun zur Gruppe der (Nicht-)Wählerinnen gehört und warum. Mit der (binären) logistischen Regressionsanalyse steht ein Analyseinstrument zur Verfügung, um solche dichotomen abhängigen Variablen untersuchen zu können.¹²



Mit Hilfe der logistischen Regressionsanalyse wird ein Modell aufgestellt, das auf Grundlage der Ausprägung(en) einer oder mehrerer unabhängiger Variablen die Wahrscheinlichkeit des Eintretens eines bestimmten Ereignisses vorhersagt. Da es bei einer binären abhängigen Variable nur zwei Möglichkeiten gibt, entweder tritt das Ereignis ein oder nicht (also beispielsweise jemand beteiligt sich an Wahlen oder nicht), liegt die Wahrscheinlichkeit für den Eintritt eines Ereignisses immer zwischen 0 und 1. Wie im linearen Fall, geht man auch im logistischen Regressionsmodell davon aus, dass eine Veränderung in den unabhängigen Variablen eine Veränderung der abhängigen Variable mit sich bringt. Diese Veränderung muss nun jedoch nicht mehr linear sein. Vielmehr werden die Wahrscheinlichkeiten dafür geschätzt, ob also zum Beispiel jemand zur Wahl geht oder nicht.

Auch bei der logistischen Regressionsanalyse kann eine Regressionsgleichung aufgestellt werden, um etwas über die Richtung und Stärke der Beziehungen der unabhängigen Variablen zur abhängigen Variable aussagen zu können. Da es sich bei der logistischen Regressionsanalyse um eine Form des „verallgemeinerten linearen Modells“ (*generalized linear model*) handelt (Fromm 2012), ähnelt die Gleichung – wie noch deutlich werden wird – einer linearen Gleichung. Bei solchen Modellen werden die Werte der abhängigen Variable nicht direkt durch eine lineare Gleichung geschätzt, sondern indirekt über eine Verknüpfungs- oder Linkfunktion zwischen Schätzparameter und linearer Gleichung ermittelt.

Die logistische Regressionsanalyse macht damit – wie im linearen Fall – eine Vorhersage für Y, wenn X bestimmte Werte annimmt. Allerdings wird bei der logistischen Regression nicht auf das Verfahren der Kleinsten Quadrate (Ordinary-Least-Squares OLS) zurückgegriffen, sondern auf das Maximum-Likelihood-Verfahren (Gautschi 2010). Dabei handelt es sich um ein iteratives Vorgehen, mit Hilfe dessen diejenige Kombination von Regressionskoeffizienten gesucht wird, mit der sich die empirischen Werte der abhängigen Variable möglichst gut reproduzieren lassen.

¹² Die sogenannte multinomiale logistische Regression ermöglicht die Untersuchung einer kategorialen abhängigen Variable mit mehr als zwei Ausprägungen. Diese ist jedoch nicht Bestandteil dieser Einführung (siehe für eine Einführung Kohler und Kreuter 2017, S. 397-401; Fromm 2012, S. 149-158).

In diesem Abschnitt wird die Vorgehensweise bei einer logistischen Regressionsanalyse näher erläutert und zwar am Beispiel folgender Fragestellung: Welche Faktoren beeinflussen, ob Bürgerinnen und Bürger zur Wahl gehen oder nicht?

Beispiel: Determinanten der Wahlbeteiligung

Sich bei freien und gleichen Wahlen mit der eigenen Stimmabgabe zu beteiligen, ist eine demokratische Errungenschaft, die als die zentrale sowie am einfachsten durchzuführende politische Partizipationsform gilt (z.B. Gabriel 2012; Schmitt 2014). Dennoch gab es in den letzten Jahrzehnten immer wieder Rückgänge in der Wahlbeteiligung bei einzelnen Wahlen zu verzeichnen, so dass die Untersuchung von Faktoren, die den Gang zur Wahlurne beeinflussen, sowohl aus demokratischen als auch aus wissenschaftlichen Gesichtspunkten relevant ist. Generell gilt, dass die Wahlbeteiligung bei nationalen Wahlen höher ist als bei Europawahlen, aber auch als bei Landtags- bzw. Kommunalwahlen. Dies lässt sich unter anderem durch den höheren gesamtpolitischen Stellenwert nationaler Wahlen erklären (z.B. Schmitt und Teperoglou 2017). Seit mehr als sieben Jahrzehnten erörtert und untersucht die sozialwissenschaftliche Wahlforschung eine Vielzahl von Einflussfaktoren bezüglich der Frage, ob und warum Bürgerinnen und Bürger zur Wahlurne gehen oder nicht. Dabei lassen sich unterschiedliche „Schulen“ unterscheiden (für einen Überblick siehe z.B. Blais 2006; Smets und van Ham 2013; Caballero 2014; Wass und Blais 2017).

Zunächst erfolgt wieder eine bivariate Betrachtung. Daran schließt sich eine theoretische Diskussion möglicher Einflussfaktoren sowie eine darauf aufbauende Hypothesenbildung an. Schließlich geht es im Beispiel darum, die Effekte der verschiedenen Einflussfaktoren bzw. unabhängigen Variablen auf die dichotome Variable „Wahlbeteiligung“ (mit den zwei Ausprägungen „geht wählen“ und „geht nicht wählen“) mittels binärer logistischer Regressionsanalyse zu untersuchen. Dabei wird – anders als bei der linearen Regressionsanalyse – kein konkreter Wert für die abhängige Variable ermittelt, sondern vielmehr die Wahrscheinlichkeit für das Eintreten des Ereignisses „Wahlbeteiligung“.

4.3.1 Bivariate Regression

Zunächst illustrieren wir wieder anhand eines fiktiven bivariaten Beispiels mit einer unabhängigen Variable die Logik des logistischen Regressionsmodells. Für die bivariate Anschauung wird zunächst als unabhängige Variable das Alter ausgewählt. Wählen gilt, wie auch andere Formen politischen Engagements, als ein sogenanntes „Lebenszyklusphänomen“ (Abendschön und Roßteutscher 2011, S. 63-64). In der Phase des Jugendalters gilt die politische Beteiligungsbereitschaft als vergleichsweise instabil, da hier erst einmal verschiedene Entwicklungsaufgaben bewältigt werden müssen. Mit Einstieg ins Erwachsenenleben mit seinen Verpflichtungen und seiner gesellschaftlichen Integration in Beruf und Familie steigt die politische Involvierung sowie die Beteiligung an Wahlen kontinuierlich an (Kleinhenz 1995, S. 27).¹³ Daher können wir an dieser Stelle folgende Hypothese formulieren:

¹³ Mit Eintritt ins Rentenalter erfolgt wiederum ein allmählicher Rückzug aus dem politischen Leben bedingt durch abnehmende Eingebundenheit in soziale Netzwerke sowie körperliche Beeinträchtigungen. Das

H1: Je älter eine Person ist, desto wahrscheinlicher ist ihre Wahlbeteiligung.

In Tabelle 51 sind das Alter in Jahren und die Wahlbeteiligung (Ja/Nein) von neun Befragten dokumentiert. Die Variable „Wahlbeteiligung“ wurde dabei so kodiert, dass Wähler den Wert 1 und Nichtwählerinnen den Wert 0 erhalten. Die Befragten mit den IDs 2, 4, 6 und 9 haben also angegeben, an der Wahl teilgenommen zu haben.

Tabelle 51: Bivariate logistische Regression mit Wahlbeteiligung und Alter

ID	Wahlbeteiligung	Alter in Jahren
1	0	20
2	1	45
3	0	18
4	1	32
5	0	21
6	1	63
7	0	42
8	0	27
9	1	51

Quelle: Eigene Darstellung

In Abbildung 19 sind die Angaben der neun Befragten in einem Streudiagramm eingetragen. Aufgrund der dichotomen Ausprägung der abhängigen Variable „Wahlbeteiligung“ sind die Werte der Personen nur auf zwei (horizontalen) Linien abgetragen, und zwar auf der unteren Linie ($y = 0$) die Nichtwähler und auf der oberen Linie ($y = 1$) die Wählerinnen.

Stellt man sich eine lineare Regressionsgerade vor, wird deutlich, dass die Daten durch diese nicht optimal repräsentiert werden können. Die Gerade würde auch Werte unter 0 bzw. über 1 beinhalten, die aber nicht sinnvoll interpretiert werden können. Schließlich kann niemand mit einer negativen bzw. mehr als hundertprozentigen Wahrscheinlichkeit zur Wahlurne gehen. Anstelle einer Regressionsgeraden wird deshalb eine logistische Funktion genutzt, die sich den Beobachtungspunkten anpasst und nur Werte zwischen 0 und 1 annehmen kann. Damit wird auch argumentiert, dass sich die Wahrscheinlichkeiten in den Extrembereichen 0 und 1 nicht linear sind, sondern sich allmählich annähern (Best und Wolf 2010, S. 829; Fromm 2012, S. 111). Zur Schätzung der Wahrscheinlichkeit des Eintretens eines bestimmten Ereignisses werden also nichtlineare – in unserem Fall logistische Funktionen – genutzt.

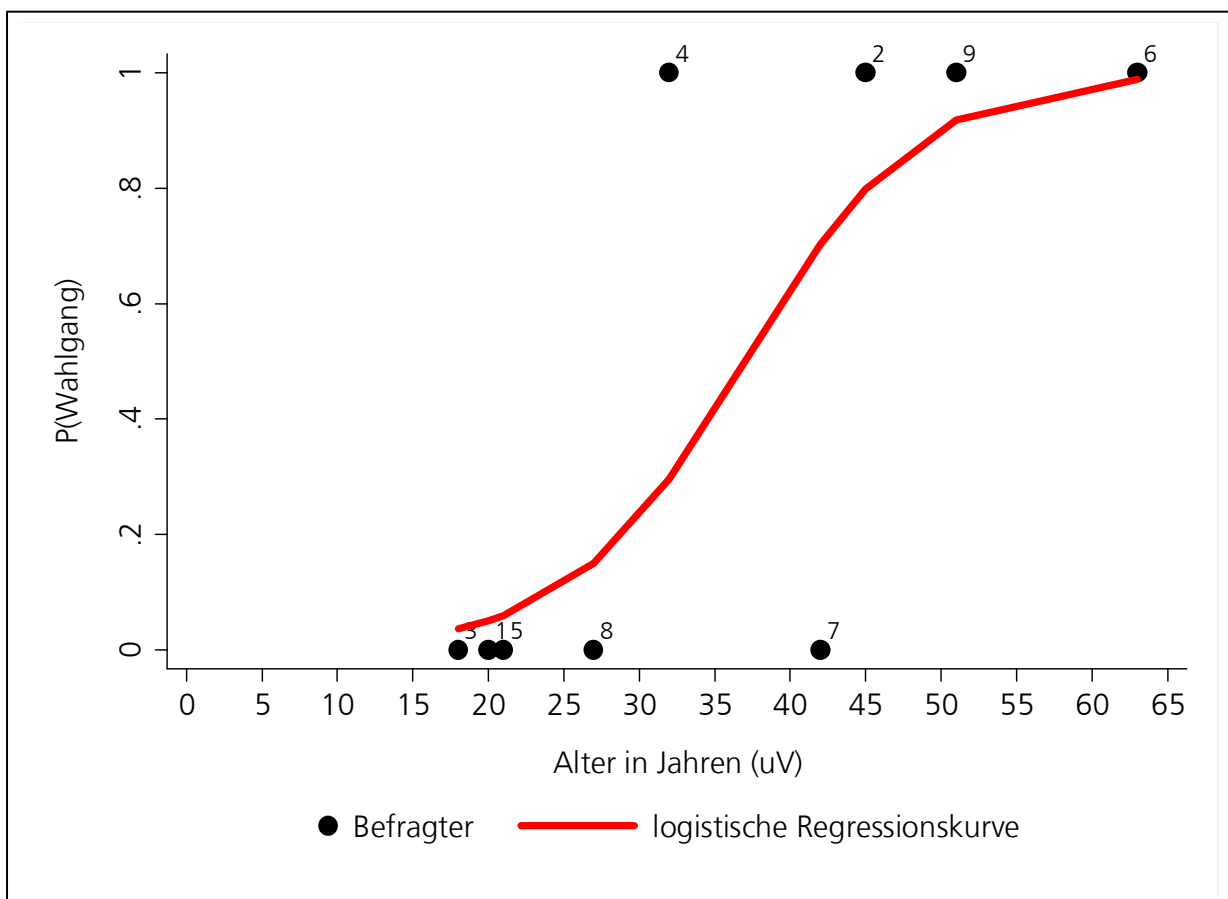


heißt, es handelt sich hier weniger um einen linearen als vielmehr um einen kurvilinearen Zusammenhang, der an dieser Stelle in den Regressionsmodellen allerdings nicht weiter berücksichtigt wird.

Schätzung der Regressionskurve

Wie wird nun die Regressionskurve geschätzt? Hierbei wird nicht das OLS-Verfahren eingesetzt, das wir im linearen Regressionsmodell kennengelernt haben und welches mit einer Reihe von Voraussetzungen verbunden ist (z. B. Homoskedastizität). Stattdessen wird das sogenannte Maximum-Likelihood-Schätzverfahren genutzt, um die zu den Daten bestmöglich passende Regressionskurve zu finden. Dabei handelt es sich um ein iteratives Vorgehen, bei dem schrittweise verschiedene Werte für die Parameter „ausprobiert“ werden, bis die bestmögliche Lösung gefunden ist. Das iterative Vorgehen bei der Maximum-Likelihood-Schätzung wird durch die Verwendung der logarithmierten Likelihood-Funktion (LL) erleichtert (ausführlich Best und Wolf 2010; Gautschi 2010; Kohler und Kreuter 2017, S. 363). In der Forschungspraxis übernimmt das Statistikprogramm die iterative Schätzung der Regressionskoeffizienten.

Abbildung 19: Streudiagramm mit Regressionskurve



Quelle: Eigene Darstellung

Wie aus Abbildung 19 ersichtlich, hat die allgemeine logistische Regressionsfunktion typischerweise einen s-förmigen Verlauf. Die Abbildung zeigt, dass die nichtlineare Regressionsfunktion immer innerhalb der Grenzen von 0 und 1 verläuft. Weiterhin ist sie symmetrisch um den Wendepunkt $P(y = 1) = 0,5$ und nähert sich zwar asymptotisch den Werten 0 und 1 an, überschreitet diese Grenzen aber nicht.

Zudem erlaubt dieser nichtlineare Verlauf häufig eine bessere Anpassung an die Daten, da eine Veränderung der unabhängigen Variable um eine Einheit (z.B. ein Anstieg des Alters um ein Jahr) nicht an allen Stellen der Funktion die gleiche Wirkung auf die abhängige Variable hat (z.B. Wähler

oder Nichtwählerin). Im Bereich sehr kleiner und sehr großer x-Werte (also bei sehr jungen bzw. sehr alten Befragten) sind die Veränderungen nur sehr gering; starke Effekte der unabhängigen Variable auf die abhängige Variable finden sich im mittleren Bereich der unabhängigen Variable (Fromm 2012, S. 111).

Die allgemeine Logit-Funktion für den bivariaten Fall lässt sich mathematisch wie folgt darstellen:

$$P(y = 1) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = \frac{e^{\text{Logit}}}{1 + e^{\text{Logit}}}$$

In der Gleichung ist neben der Euler'schen Zahl ($e \approx 2,718$) mit Logit die Regressionsgleichung gemeint. Die Regressionsgleichung lautet:

$$\text{Logit}(L_i) = \alpha + \beta * x$$

In ihrer Struktur ähnelt die logistische Regressionsgleichung bzw. Logit-Funktion der linearen Regressionsgleichung. Mit α wird wieder der Achsenabschnitt bezeichnet, mit β der Steigungskoeffizient.

Der s-förmige Verlauf ist nötig, um – wie oben dargestellt – prognostizierte Wahrscheinlichkeiten kleiner als 0 bzw. größer als 1 zu vermeiden. Diese Transformation der linearen Regressionsgleichung in eine Logit-Funktion bedeutet aber auch, dass die Schätzwerte für die Parameter auf andere Weise interpretiert werden müssen. Dabei ist angesichts der Kurve leicht ersichtlich, dass die Interpretation der Steigungskoeffizienten nicht für alle Werte der unabhängigen Variablen gleich ausfallen kann, wie dies im linearen Modell der Fall ist.

Wie wird nun die logistische Regressionskurve interpretiert? Die Regressionsgleichung oben zeigt an, dass auch im logistischen Modell als Parameter wieder der Schnittpunkt mit der y-Achse sowie ein (bzw. im multiplen Fall mehrere) Regressionskoeffizient(en) Auskunft über die Steigung der Kurve geben und im Modell geschätzt werden. Für unser bivariates Modell ergibt sich ein Schnittpunkt mit der y-Achse von $-6,40$ (d.h. für jemanden mit 0 Jahren beträgt die logarithmierte Chance, zur Wahl zu gehen, $-6,40$). Der Steigungskoeffizient von $0,17$ bedeutet, dass mit jedem Jahr, das eine Person älter wird, die logarithmierte Chance, dass sie zur Wahl geht, um $0,17$ steigt. Die Interpretation mittels logarithmierter Chancen ist nicht sehr intuitiv, wie wir auch im multiplen Fall noch sehen werden. Wichtig ist aber zur Überprüfung der Hypothese 1, dass der Steigungskoeffizient ein positives Vorzeichen hat. Die Hypothese kann bestätigt werden.

Interpretation der Ergebnisse

4.3.2 Multiple Regression

Die Durchführung einer logistischen Regressionsanalyse ist mittels Statistikprogrammen wie SPSS und Stata heutzutage vergleichsweise einfach. Allerdings sollte man wissen, wie die entsprechend berechneten Koeffizienten interpretiert werden. Für unser multiples logistisches Regressionsmodell werden exemplarisch drei weitere „klassische“ Faktoren bzw. unabhängige Variablen ausgewählt und ihr Einfluss auf die individuelle Wahlbeteiligung auf Basis des GLES-Datensatzes von 2017 untersucht. Sie betreffen soziodemografische sowie sozialpsychologische und einstellungsbezogene Faktoren der Stimmabgabe.

Hypothesen

In der Wahlforschung zeigt sich immer wieder, dass eine höhere Bildung mit einer höheren Bereitschaft, zur Wahl zu gehen, zusammenhängt. Schließlich geht ein formal höheres Bildungsniveau auch mit einer stärkeren Eingebundenheit in soziale Netzwerke und dem Erwerb sogenannter „civic skills“ einher, die politische Partizipation erleichtern (z.B. Verba et al. 1995; Smets und van Ham 2013, S. 348). Daher formulieren wir für unsere Untersuchung:

H2: Je höher das Bildungsniveau ist, desto wahrscheinlicher ist die Wahlbeteiligung.

Politisches Interesse kann die individuelle Wahlbeteiligung auf zweifache Weise beeinflussen (zum politischen Interesse siehe van Deth 1990, 2004, 2013). Interessiert man sich für das politische Geschehen, dann verfügt man über eine gesteigerte politische Aufmerksamkeit sowie politische Informationen, die es erleichtern, eine Wahlentscheidung zu treffen. Zweitens fühlt man sich dem Gemeinwohl stärker verbunden und man ist von der Wichtigkeit der eigenen Stimmabgabe überzeugt. Daher können wir vermuten, dass sich das politische Interesse einer Person positiv auf die Wahlbeteiligung auswirkt:

H3: Je stärker das politische Interesse ist, desto wahrscheinlicher ist die Wahlbeteiligung.

Eines der ersten Modelle zur Erklärung des (amerikanischen) Wahlverhaltens war das sogenannte Ann-Arbor- oder Michigan-Modell (Campbell et al. 1954). Dieses sozialpsychologische Erklärungsmodell der Wahlbeteiligung spricht insbesondere der Identifikation mit einer politischen Partei eine wichtige Rolle bei der Wahlentscheidung zu. Fühlt sich eine Person mit einer bestimmten Partei verbunden, dann möchte diese Person die Partei auch bei anstehenden Wahlen unterstützen und geht folglich eher zur Wahlurne als jemand, der über keine Parteiidentifikation verfügt (zum Konzept der Parteiidentifikation siehe Schoen und Weins 2014, S. 262-284; Green und Baltes 2017):

H4: Personen mit einer Parteiidentifikation beteiligen sich eher an Wahlen als Personen ohne Parteiidentifikation.

Seit mehr als 100 Jahren dürfen Frauen in Deutschland wählen. Trotzdem war bis vor einigen Jahren ein gender gap bezüglich der Stimmabgabe zu verzeichnen. Frauen gingen lange Zeit signifikant seltener zur Wahl als Männer (z.B. Norris 2004). Dies lag zum einen daran, dass Männer eher über die entsprechenden Partizipationsressourcen wie Bildung, Einkommen und soziale Netzwerke verfügten als Frauen. Frauen haben in Bezug auf diese strukturellen Variablen aufgeholt und machen zumindest in Deutschland meist genauso häufig von ihrem Wahlrecht Gebrauch wie Männer, so dass sich in jüngerer Zeit bezüglich des Wahlgangs sogar ein umgekehrter gender gap ausmachen lässt (Smets und van Ham 2013). In unserer Analyse geht das Geschlecht daher als Kontrollvariable ein, mit der keine spezifische Erwartung verknüpft ist.

Für eine binäre logistische Regressionsanalyse mit mehr als einer unabhängigen Variable können wir obige Gleichung für die multiple Logit-Funktion erweitern:

$$\text{Logit}(L_i) = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k$$

Mit den Logit-Koeffizienten und den Exp(B)-Koeffizienten (Odds Ratios) zeigt Tabelle 52 zwei verschiedene Möglichkeiten, wie in empirischen Publikationen die entsprechenden Koeffizienten von logistischen Regressionen dargestellt werden. Beide sind zulässig, unterscheiden sich aber in der Interpretation. Zudem werden in Tabelle 52 Informationen zur statistischen Signifikanz dokumentiert.

Die Logit-Koeffizienten (Steigungsparameter β) geben uns die logarithmierte Chance an, dass das Ereignis „Wahlbeteiligung“ eintritt. Dabei beschreibt die Konstante der logistischen Regression – wie bei der linearen Regression – Fälle, die bei allen unabhängigen Variablen den Wert 0 aufweisen (bzw. bei Dummy-Variablen die Referenzkategorie bezeichnen). Wie auch im linearen Fall schon angesprochen, handelt es sich hier in erster Linie um einen technischen Wert, den wir nicht überinterpretieren sollten. In unserem Beispiel beschreibt die Konstante auf Basis der Kodierungen der unabhängigen Variablen Männer im Alter von 0 Jahren mit mittlerer Bildung, keinem politischen Interesse und keiner Parteiidentifikation. Bei dieser Merkmalskonstellation liegt die vorhergesagte logarithmierte Chance, zur Wahl zu gehen, bei $-0,544$. Für die Prüfung der Hypothesen sind auch im logistischen Fall die (inhaltlichen) Regressionskoeffizienten wichtiger. Die Konstante lässt sich allerdings – wie noch gezeigt wird – nutzen, um bestimmte Merkmalskombinationen zu berechnen.

Logit-Koeffizienten

Tabelle 52: Bestimmungsfaktoren der Wahlbeteiligung

	Logit-Koeffizient	Odds Ratio/ Exp(B)-Koeffizienten	Standardfehler	p-Wert
Geschlecht (weiblich)	-0,138	0,871	0,163	0,397
Politisches Interesse	0,703	2,019	0,089	0,000
Parteiidentifikation (Ja)	0,617	1,853	0,165	0,000
Alter	0,021	1,021	0,005	0,000
Bildung (Referenz: Mittlere Reife)				
ohne Abschluss/Hauptschule	-0,779	0,459	0,076	0,000
(Fach-)Hochschulreife	0,741	2,098	0,213	0,000
Sonstige	0,963	2,619	0,234	0,371
Konstante	-0,544	0,581	0,288	0,059
Nagelkerkes R^2		0,205		
N		2032		

Daten: GLES 2017. Eigene Berechnungen

Der Steigungskoeffizient β gibt Hinweise auf die Lage der Logit-Funktion: Wie flach bzw. wie steil ist die Kurve für die jeweilige unabhängige Variable? Wie beim linearen Fall interpretieren wir die Regressionskoeffizienten für die unabhängigen Variablen im Hinblick auf eine Veränderung der abhängigen Variable, wenn die jeweils betrachtete unabhängige Variable um eine Einheit steigt. Allerdings sind die geschätzten Werte bei der logistischen Regression die logarithmierten Chancen

für das Eintreten des Ereignisses (im Beispiel: zur Wahl gehen). Mit Blick auf die Analyse auf Basis der GLES-Daten von 2017 lässt sich als Ergebnis festhalten: Pro Lebensjahr steigt die logarithmierte Wahrscheinlichkeit, einen Wahlzettel abzugeben, um 0,021. Für Personen mit hoher Bildung steigt die logarithmierte Chance, an der Bundestagswahl teilzunehmen, im Vergleich zu einer mittleren Bildung um 0,741.

! Diese Beispiele zeigen, dass die Interpretation der logarithmierten Chancen nicht sehr intuitiv ist. Deshalb werden bei einer logistischen Regression häufig nur die Vorzeichen der Regressionskoeffizienten interpretiert. Ein positives Vorzeichen bedeutet, dass die logarithmierte Chance (oder auch Wahrscheinlichkeit), zur Wahl zu gehen, steigt. Ein negatives Vorzeichen bedeutet eine sinkende (logarithmierte) Wahrscheinlichkeit für die Stimmabgabe bei der Wahl. Die Betragshöhe des Regressionskoeffizienten gibt einen Hinweis auf die Effektstärke, allerdings können (ohne mathematische Umrechnungen) keine Aussagen über das genaue Ausmaß der Veränderung der Wahrscheinlichkeit gemacht werden.

Mit Blick auf Tabelle 52 und vor dem Hintergrund der vorher vorgenommenen Hypothesenbildung können wir schließen: Personen mit niedriger Bildung gehen mit einer geringeren Wahrscheinlichkeit zur Bundestagswahl als Personen mit mittlerer Bildung. Dagegen beteiligen sich Personen mit höherer Bildung eher an Bundestagswahlen als Befragte mit mittlerer Bildung (H2). Das politische Interesse hat ebenfalls einen positiven Einfluss auf die Wahlbeteiligung. Je stärker sich eine Person für Politik interessiert, desto höher ist auch ihre (logarithmierte) Chance, einen Stimmzettel abzugeben (H3). Auch beteiligen sich Personen mit einer Parteiidentifikation eher an der Wahl als Befragte ohne Parteiidentifikation (H4). Schließlich kann auch auf Basis dieser Daten die erste Hypothese, die einen positiven Zusammenhang zwischen Alter und Wahlbeteiligung beschreibt, bestätigt werden (H1). Für das Geschlecht lassen sich dagegen keine statistisch signifikanten Ergebnisse vorweisen. Dies spiegelt wieder, was bereits diskutiert wurde: Bei der Wahlbeteiligung lässt sich in Deutschland kein gender gap feststellen.

Odds-Ratios und Exp(B)

Eine alternative Möglichkeit der Interpretation bieten die Koeffizienten in der Spalte „Exp(B)“. Diese Koeffizienten werden als Effektkoeffizienten (Fromm 2012; Urban und Mayerl 2018, S. 399-405) oder auch als Odds-Ratios (Schendera 2014, S. 161; Kohler und Kreuter 2017, S. 367-368) bezeichnet. Bei Odds-Ratios handelt es sich um Chancenverhältnisse. Ein Exp(B)-Koeffizient von 2,098 heißt aber nicht, dass hochgebildete Personen eine 2,1-mal höhere Wahrscheinlichkeit aufweisen, wählen zu gehen, sondern, dass das Vorliegen eines hohen Bildungsniveaus die Chance des Eintretens des Ereignisses $y = 1$ (Wahl) im Verhältnis zum Nichteintreten $y = 0$ (Nicht-Wahl) um den Faktor 2,098 erhöht. Es handelt sich also um ein Chancenverhältnis, das die vorhergesagte Veränderung in der abhängigen Variable angibt, wenn sich die unabhängige Variable um eine Einheit ändert. Ein Exp(B)-Koeffizient von 2,019 bedeutet, dass sich das Chancenverhältnis von Wähler zu Nichtwähler um den Faktor 2 zugunsten des Wahlgangs verändert, wenn eine Person ein um einen Skalenpunkt höheres politisches Interesse aufweist. Dagegen sinkt das Chancenverhältnis für den Wahlgang um 0,495, wenn Befragte über keinen oder nur über einen Hauptschulabschluss (anstelle der Mittleren Reife) verfügen.

Der Wertebereich der Exp(B)-Koeffizienten unterscheidet sich vom Wertebereich der Logit-Regressionkoeffizienten. Während der Wertebereich der oben besprochenen Logit-Regressionkoeffizienten zwischen $-\infty$ und $+\infty$ liegt, liegt der Wertebereich der Exp(B)-Koeffizienten zwischen 0

und $+\infty$. Ein $\text{Exp}(B)$ -Koeffizient kleiner als 1 deutet auf eine geringere Chance und ein $\text{Exp}(B)$ -Koeffizient größer als 1 auf eine höhere Chance hin. Bei einem $\text{Exp}(B)$ -Koeffizient von 1 bleibt die Chance gleich.

Mit den Effektkoeffizienten bzw. Odds-Ratios ändert sich selbstverständlich nur die Art der Interpretation – die substantiellen Ergebnisse bleiben gleich. Ein positiver Logit-Regressionskoeffizient wird einen Odds-Ratio größer als 1 haben, ein negativer Logit-Regressionskoeffizient wird einen Odds-Ratio kleiner als 1 haben (aber niemals kleiner als 0). Die Kenntnis der unterschiedlichen Wertebereiche ist wichtig, da in Veröffentlichungen entweder die Logit-Regressionskoeffizienten oder die $\text{Exp}(B)$ -Koeffizienten ausgewiesen werden. Bei den Logit-Regressionskoeffizienten deuten negative Werte auf eine geringere Chance hin, bei den $\text{Exp}(B)$ -Koeffizienten hingegen Werte zwischen 0 und 1. Bei den Logit-Regressionskoeffizienten deuten positive Werte auf eine höhere Chance hin, bei den $\text{Exp}(B)$ -Koeffizienten sind die Werte größer als 1. Die absoluten Beträge der Regressionskoeffizienten bzw. der Effektkoeffizienten sollten allerdings nicht interpretiert werden.

Für ausgewählte Kombinationen können aber die Wahrscheinlichkeiten berechnet werden. Dafür wird das Logit der Regressionsgleichung berechnet und in die Formel eingetragen¹⁴:

Wahrscheinlichkeit berechnen

$$P(y = 1) = \frac{e^{\text{Logit}}}{1 + e^{\text{Logit}}}$$

Für eine 40-jährige Person mit Parteiidentifikation und hoher Bildung in unserem Beispiel wird das Logit wie folgt berechnet: $-0,544$ (Konstante) + $0,617$ (Parteiidentifikation) + $(40 \cdot 0,021)$ (Alters-effekt) + $(0,741)$ (Bildungseffekt). Das Logit von $1,654$ wird in die Formel eingetragen und damit die Wahrscheinlichkeit berechnet:

$$P(y = 1) = \frac{e^{1,654}}{1 + e^{1,654}} = \frac{5,228}{1 + 5,228} = 0,84$$

Das heißt, die Wahrscheinlichkeit, dass eine 40-jährige Person mit hoher Bildung und Parteiidentifikation zur Wahl geht, liegt bei etwa 84 Prozent.

Wie Tabelle 52 auch zeigt, können wir bei der logistischen Regression ebenfalls wieder Aussagen darüber treffen, ob wir auf Basis der Stichprobe auch auf die Grundgesamtheit schließen können. Die letzte Spalte berichtet die sogenannten p-Werte. Ein p-Wert gibt die Wahrscheinlichkeit für ein Ergebnis unter der Bedingung an, dass die Nullhypothese zutrifft (Sedlmeier und Renkewitz 2013, S. 389). In den Sozialwissenschaften wird bei einem p-Wert kleiner als 0,05 die Nullhypothese vorläufig verworfen und die Alternativhypothese angenommen. Dabei werden in den Sozialwissenschaften meist drei Signifikanzniveaus unterschieden (siehe auch den Abschnitt zur linearen Regression): p-Wert kleiner als 0,05 (Irrtumswahrscheinlichkeit liegt bei 5 Prozent, abgekürzt mit einem Sternchen *), p-Wert kleiner als 0,01 (Irrtumswahrscheinlichkeit liegt bei 1 Prozent, abgekürzt mit zwei Sternchen **) und p-Wert kleiner als 0,001 (Irrtumswahrscheinlichkeit liegt bei 0,1 Prozent, abgekürzt mit drei Sternchen ***). Auf

Aussagen zur Signifikanz

¹⁴ Die Euler'sche Zahl (e) beträgt etwa 2,718.

Grundlage dieser Konvention spricht man dann bei Vorliegen der entsprechenden p-Werte von statistischer Signifikanz. Damit können (vorbehaltlich der Irrtumswahrscheinlichkeit) die beobachteten Zusammenhänge von der Stichprobe auf die Grundgesamtheit übertragen werden.

Der p-Wert bzw. die Signifikanz darf keineswegs als Indikator für die Effektstärke interpretiert werden. Je größer die Stichprobe ist, desto kleiner ist in der Regel der p-Wert. Ein p-Wert sagt nichts über die Bedeutsamkeit eines Befunds aus. Der p-Wert informiert lediglich darüber, ob ein in der Stichprobe gefundener Zusammenhang auf die Grundgesamtheit übertragen werden kann.

Modellgüte und Pseudo-R²

Im Abschnitt zur linearen Regression haben wir mit dem Determinationskoeffizienten R^2 ein Maß kennengelernt, welches uns anschaulich über die Güte des Gesamtmodells informiert und uns den Anteil erklärter Varianz an der Gesamtvarianz berichtet. Auch im logistischen Fall gibt es die Möglichkeit einer solchen Angabe. Analog zum R^2 sagen sogenannte Pseudo- R^2 etwas über die globale Güte des Modells aus. Gängige Pseudo- R^2 , die auch von den Statistikprogrammen ausgegeben werden, sind Cox und Snell sowie Nagelkerke. Nagelkerkes R^2 wird häufig interpretiert wie das Bestimmtheitsmaß in der linearen Regression, nämlich als Anteil der durch alle unabhängigen Variablen erklärten Varianz (Fromm 2012, S. 130). Diese Interpretation ist allerdings problematisch, da Pseudo- R^2 -Maßzahlen auf der Veränderung des Maximum-Likelihood-Werts basieren. Deshalb sollten Pseudo- R^2 -Maßzahlen nicht mit dem R^2 der linearen Regression verglichen werden (Urban und Mayerl 2018, S. 420). Grundsätzlich deutet ein höherer Wert auf ein passenderes Modell hin und kann als Indiz für die globale Güte des Modells interpretiert werden. Tabelle 52 berichtet ein Nagelkerkes R^2 von 0,205. Diesen Wert würden viele Sozialforscher als akzeptabel interpretieren.

Voraussetzungen

Wie wir gesehen haben, unterscheidet sich die logistische Regression vor allem in zwei Punkten von der linearen Regression. Erstens wird statt einer linearen Regressionsgeraden eine nichtlineare Regressionskurve geschätzt (die sogenannte Logit-Funktion). Zweitens wird bei der Schätzung der Regressionsparameter nicht auf das OLS-Verfahren, sondern auf das Maximum-Likelihood-Prinzip zurückgegriffen. In der Literatur wird vor allem auf zwei wesentliche Voraussetzungen für die Durchführung einer binären logistischen Regressionsanalyse hingewiesen: die Fallzahl und die Skalenniveaus der abhängigen und unabhängigen Variablen.

Für eine logistische Regression ist eine größere Fallzahl als bei einer linearen Regression erforderlich. Backhaus et al. (2018, S. 327) und Fromm (2012, S. 108-109) nennen als Minimum 50 Fälle. Dabei sollten für jede Ausprägung der abhängigen Variable mindestens 25 Fälle vorliegen. Nach Fromm (2012, S. 109) können aussagekräftige Ergebnisse sogar erst ab einer Fallzahl von 100 Beobachtungen erreicht werden. Die erforderliche Fallzahl steigt zudem mit der Anzahl der unabhängigen Variablen. Für jede weitere unabhängige Variable sollte die Fallzahl jeweils um zehn Beobachtungen steigen (Backhaus et al. 2018, S. 327).

Was die Messung der verwendeten Variablen anbelangt, muss die abhängige Variable binär kodiert sein und die unabhängigen Variablen müssen – wie im linearen Fall – ein metrisches oder binäres kategoriales Skalenniveau aufweisen. Letzteres kann durch eine Dummy-Kodierung – wie im linearen Fall gezeigt – erreicht werden.

5 Inferenzstatistik

Markus Tausendpfund

Die folgenden Abschnitte behandeln die Grundlagen der Inferenzstatistik. Die Inferenzstatistik beschäftigt sich mit der Frage, ob bzw. wie empirische Befunde aus Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden dürfen. In den Sozialwissenschaften sind Forscherinnen in der Regel mit Zufallsstichproben konfrontiert, so dass für ein Verständnis empirischer Forschungsergebnisse basale Kenntnisse der Inferenzstatistik erforderlich sind. In diesem Kapitel werden in einem ersten Schritt die mit Zufallsstichproben verbundenen Probleme dargestellt, ehe mit dem zentralen Grenzwertsatz und dem Standardfehler zwei zentrale Konzepte der Inferenzstatistik erläutert werden. Anschließend werden mit der Punktschätzung und der Intervallschätzung zwei Schätzungsarten vorgestellt, bevor das statistische Testen und die Durchführung eines t-Tests behandelt werden.

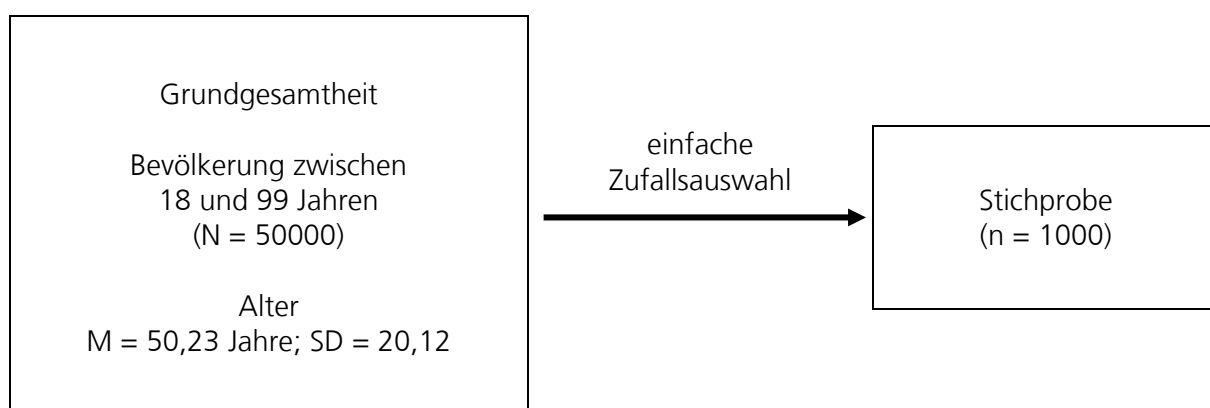


5.1 Was ist das Problem?

Warum müssen sich Sozialwissenschaftler überhaupt mit der Inferenzstatistik auseinandersetzen? Aus zeitlichen, finanziellen und forschungspraktischen Gründen dominieren in den Sozialwissenschaften Stichproben. Bei empirischen Untersuchungen werden (in der Regel) nicht alle Elemente der Grundgesamtheit untersucht, sondern es gerät nur eine zufällige Auswahl dieser Elemente in den Blick. Der ALLBUS basiert beispielsweise auf einer Zufallsstichprobe der Bevölkerung in Deutschland ab 18 Jahren (Wasmer et al. 2017). Eine zentrale Frage ist, ob bzw. wie die Ergebnisse einer Zufallsstichprobe auf die Grundgesamtheit übertragen werden können.

In Abbildung 20 wird der Zusammenhang zwischen Grundgesamtheit und Zufallsstichprobe illustriert. Unsere Beispiel-Grundgesamtheit besteht aus 50000 fiktiven Personen im Alter zwischen 18 und 99 Jahren. In der Regel sind unsere Informationen über die Grundgesamtheit begrenzt. An dieser Stelle unterscheidet sich unsere „fiktive“ Grundgesamtheit von einer „echten“ Grundgesamtheit. Wir wissen (aus einer fiktiven Volksbefragung), dass die Personen unserer Grundgesamtheit zwischen 18 und 99 Jahre alt sind. Das mittlere Alter der Grundgesamtheit liegt bei 50,23 Jahren, die Standardabweichung beträgt 20,12.

Abbildung 20: Grundgesamtheit und Stichprobe



Quelle: Eigene Darstellung

Aus dieser Grundgesamtheit wird eine Stichprobe von 1000 Personen zufällig gezogen. Zufällig bedeutet, dass jedes Element der Grundgesamtheit (also jede Person) die gleiche und von 0 verschiedene Chance hat, in die Stichprobe zu gelangen. Dies entspricht der klassischen Vorgehensweise bei einer (einfachen) Zufallsstichprobe. Wir unterstellen dabei, dass es keine Probleme mit Undercoverage und Overcoverage gibt. Auch verweigert keine Person die Teilnahme an der Befragung (Nonresponse).

Overcoverage und Undercoverage

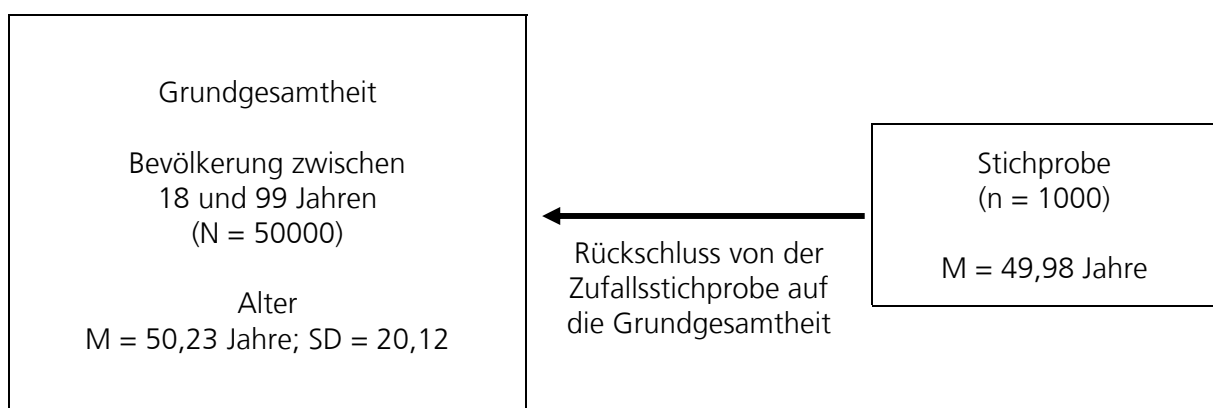
Mit Overcoverage ist die Menge an Untersuchungsobjekten gemeint, die in die Stichprobe gelangen kann, aber eigentlich gar nicht zur Grundgesamtheit gehört. In unserem Beispiel könnten dies Personen sein, die jünger als 18 Jahre sind. Mit Undercoverage sind Untersuchungsobjekte gemeint, die eigentlich eine Chance haben sollten, in die Stichprobe zu gelangen, aber faktisch nicht ausgewählt werden können. Bei einer Online-Erhebung sind das beispielsweise Personen, die keinen Internetanschluss haben.

Auf Grundlage unserer Stichprobe berechnen wir das mittlere Alter der 1000 Personen. Das mittlere Alter – das Durchschnittsalter – der Personen in unserer Stichprobe beträgt 49,98 Jahre.

Punktschätzung

Angenommen, wir haben keine Informationen über das mittlere Alter in der Grundgesamtheit. Dann könnten (und würden) wir das mittlere Alter der Grundgesamtheit auf Grundlage unserer Stichprobe schätzen. Wir würden also davon ausgehen, dass das mittlere Alter in der Grundgesamtheit ebenfalls 49,98 Jahre beträgt. Diese Vorgehensweise wird allgemein als Punktschätzung bezeichnet (siehe auch Abschnitt 5.3.1). Der nicht bekannte Parameter der Grundgesamtheit „Mittleres Alter“ wird auf Basis unserer Stichprobe geschätzt (Abbildung 21).

Abbildung 21: Rückschluss von der Stichprobe auf die Grundgesamtheit



Quelle: Eigene Darstellung

Das in Abbildung 21 skizzierte Vorgehen entspricht der üblichen Vorgehensweise in den Sozialwissenschaften. Wir nutzen Informationen unserer Stichprobe (z.B. arithmetisches Mittel und Standardabweichung), um Aussagen über die eigentlich interessierende Grundgesamtheit zu treffen. In der Regel sind wir überhaupt nicht an den Informationen der Stichprobe interessiert, sondern an Informationen über die Grundgesamtheit. Da wir aber meist keine Vollerhebungen durchführen können, nutzen wir die Informationen der Stichprobe, um etwas über unsere

Grundgesamtheit zu erfahren. Die Stichprobe ist nur ein „Hilfsmittel“, um Aussagen über die eigentlich interessierende Grundgesamtheit treffen zu können.

Unsere (fiktive) Grundgesamtheit (siehe Abbildung 20) entspricht an dieser Stelle allerdings nicht der Realität, da uns verlässliche Informationen über sie nicht zur Verfügung stehen. Wir wissen, dass das mittlere Alter unserer Grundgesamtheit 50,23 Jahre beträgt. Es gibt also eine Differenz zwischen dem mittleren Alter unserer Stichprobe (49,98) und dem mittleren Alter der Grundgesamtheit (50,23 Jahre). Diese Differenz beträgt 0,25 Jahre. Mit unserer Stichprobe unterschätzen wir das mittlere Alter um 0,25 Jahre. Diese Abweichung wird allgemein als Stichprobenfehler oder als Stichprobenschwankung bezeichnet.

Stichprobenfehler

Schauen wir uns diesen Stichprobenfehler einmal genauer an. Wir ziehen neun weitere Zufallsstichproben. Jede dieser Stichproben umfasst wieder 1000 Personen. In Tabelle 53 sind die Mittelwerte der einzelnen Stichproben und die Abweichung der Stichprobenmittelwerte vom Mittelwert in der Grundgesamtheit dokumentiert. Auf den ersten Blick zeigt sich, dass die einzelnen Mittelwerte der Stichproben vom Mittelwert der Grundgesamtheit abweichen. Mit anderen Worten: Bei der Abweichung der ersten Stichprobe handelt es sich nicht um ein singuläres Ereignis, sondern das Problem der Abweichung betrifft alle Stichproben.

Mit den Stichproben 1, 2, 3, 5, 6 und 8 werden die Mittelwerte der Grundgesamtheit (50,23) unterschätzt. Das durchschnittliche Alter der Befragten ist in den Stichproben geringer als das tatsächliche Alter in der Grundgesamtheit. Bei den Stichproben 4, 7, 9 und 10 findet sich das gegenteilige Muster. Diese Stichproben überschätzen das mittlere Alter der Grundgesamtheit.

Tabelle 53: Mittelwerte in Zufallsstichproben (Stichprobengröße jeweils 1000 Personen)

Stichprobe	Mittelwert der Stichprobe	Abweichung vom Mittelwert der Grundgesamtheit (50,23 Jahre)
1	49,98	-0,25
2	50,20	-0,03
3	50,10	-0,13
4	50,35	0,12
5	50,14	-0,09
6	49,67	-0,56
7	50,42	0,19
8	49,80	-0,43
9	50,52	0,29
10	50,54	0,31

Quelle: Eigene Darstellung

Die Angaben in Tabelle 53 zeigen aber auch, dass die Abweichung vom Mittelwert der Grundgesamtheit nicht in allen Stichproben gleich groß ausfällt. Bei Stichprobe 2 ist die Abweichung zum Mittelwert der Grundgesamtheit fast schon vernachlässigbar (−0,03), bei den Stichproben 6 (−0,56) und 8 (−0,43) sind die Abweichungen deutlich höher.

Stichprobenergebnisse variieren

In den Medien werden wir heute täglich mit Umfrageergebnissen konfrontiert. Ein bekanntes Beispiel ist die sogenannte Sonntagsfrage: „Welche Partei würden Sie wählen, wenn am kommenden Sonntag Bundestagswahl wäre?“ Die Ergebnisse variieren von Institut zu Institut (siehe für eine aktuelle Übersicht: www.wahlrecht.de/umfragen/index.htm). Die präsentierten Stichprobenergebnisse weisen immer eine Streuung auf.

! In der Forschungspraxis sind wir an dieser Stelle allerdings mit zwei zentralen Problemen konfrontiert: Erstens kennen wir den tatsächlichen Wert der Grundgesamtheit nicht. Das durchschnittliche Alter der Befragten, das durchschnittliche Einkommen oder auch Einstellungen (z.B. Demokratiezufriedenheit) und Verhaltensweisen (z.B. Wahlentscheidung) sind uns ebenfalls nicht bekannt. Wir nutzen ja gerade Stichproben wie ALLBUS oder ESS, um auf Basis der Stichproben etwas über unsere Grundgesamtheit zu erfahren. Zweitens ziehen wir in der Praxis nicht zehn (oder noch mehr) Stichproben, sondern nur eine Stichprobe. Es kann sich dabei um eine „gute“ Stichprobe (Stichprobe 2) oder um eine „schlechte“ Stichprobe (Stichprobe 6) handeln. Da wir allerdings den Mittelwert oder Anteilswert der Grundgesamtheit nicht kennen, können wir – auf den ersten Blick – nicht beurteilen, ob unsere Stichprobe eine präzise Schätzung ermöglicht oder nicht.

Deshalb wäre es wichtig, etwas mehr über Abweichung der Mittelwerte der Stichproben vom Mittelwert der Grundgesamtheit zu erfahren. Dazu berechnen wir einmal den Mittelwert unserer zehn Stichprobenmittelwerte:

$$\bar{\bar{x}} = \frac{49,98 + 50,20 + 50,10 + 50,35 + 50,14 + 49,67 + 50,42 + 49,80 + 50,52 + 50,54}{10} = 50,17$$

Der Mittelwert unserer zehn Stichprobenmittelwerte liegt bei 50,17 Jahren. Die Differenz zum Mittelwert der Grundgesamtheit beträgt 0,06 Jahre. Mit anderen Worten: Der Mittelwert der zehn Stichproben erlaubt in den meisten Fällen eine bessere Schätzung des Mittelwerts der Grundgesamtheit als eine einzelne Schätzung. Wie verändert sich der Mittelwert der Stichproben, wenn die Anzahl der Stichproben steigt?

In Tabelle 54 finden sich die Angaben für sechs Versuche. Für den ersten Versuch wurden zehn Stichproben gezogen und es wurde der Mittelwert dieser Stichproben berechnet. Für den zweiten Versuch wurden 20 Stichproben gezogen und es wurde wieder der Mittelwert dieser Stichproben berechnet. Dabei weicht der Mittelwert der 20 Stichproben etwas stärker ab als beim ersten Versuch. Deshalb erhöhen wir weiter die Anzahl der Stichproben. Beim dritten Versuch ziehen wir 50, beim vierten Versuch 100 und beim fünften Versuch sogar 500 Stichproben. Dabei berechnen wir jeweils den Mittelwert. Insgesamt ist die Abweichung der Mittelwerte der Stichproben nur noch geringfügig höher als der Mittelwert der Grundgesamtheit. Zum Schluss ziehen wir 1000 Stichproben mit jeweils 1000 Personen und berechnen erneut den Mittelwert. Dieser Mittelwert entspricht praktisch dem Mittelwert unserer Grundgesamtheit. Mit steigender Stichprobenzahl

bieten die Stichprobenmittelwerte also eine immer bessere Schätzung des Mittelwerts unserer Grundgesamtheit.

Tabelle 54: Mittelwerte von Zufallsstichproben (Stichprobengröße jeweils 1000 Personen)

Anzahl der Stichproben	Mittelwert der Stichproben	Abweichung vom Mittelwert der Grundgesamtheit (50,23 Jahre)	Min	Max
10	50,17	-0,06	49,67	50,54
20	50,35	0,12	48,74	51,54
50	50,20	-0,03	48,40	51,54
100	50,19	-0,04	48,88	52,14
500	50,27	0,04	47,98	52,03
1000	50,23	0,00	47,78	52,12

Quelle: Eigene Darstellung

In Tabelle 54 sind neben den Mittelwerten der Stichproben und deren Abweichung vom Mittelwert der Grundgesamtheit auch Minimum (Min) und Maximum (Max) eingetragen. Was bedeuten diese Werte? In der letzten Zeile wurden 1000 Stichproben gezogen. Jede dieser Stichproben umfasst 1000 Personen. Die 1000 Mittelwerte liegen zwischen 47,78 (Minimum) und 52,12 (Maximum). Zwar bietet der Mittelwert der 1000 Mittelwerte eine sehr gute Schätzung, aber ein Mittelwert dieser 1000 Stichproben kann immer noch sehr stark vom tatsächlichen Mittelwert der Grundgesamtheit abweichen. Und in der Forschungspraxis wird in der Regel nur eine Stichprobe gezogen. Im Extremfall könnte der Mittelwert der Grundgesamtheit also um 2,45 Jahre unterschätzt ($47,78 - 50,23 = -2,45$) oder aber um 1,89 Jahre überschätzt werden ($52,12 - 50,23 = 1,89$).

Das Beispiel sollte vier Herausforderungen illustrieren, mit denen Sozialwissenschaftler bei der Arbeit mit Zufallsstichproben konfrontiert sind:

- Erstens weichen die empirischen Ergebnisse einer Zufallsstichprobe immer (mehr oder weniger) vom tatsächlichen Wert in der Grundgesamtheit ab. Diese Abweichung wird – wie oben dargestellt – als Stichprobenfehler oder als Stichprobenschwankung bezeichnet.
- Zweitens können wir auf den ersten Blick nicht beurteilen, ob die Abweichung einer Stichprobe von der Grundgesamtheit als „hoch“ oder „niedrig“ bezeichnet werden kann. Schließlich kennen wir den tatsächlichen Wert der Grundgesamtheit nicht.
- Drittens bieten viele Stichproben – und deren Mittelwerte – eine bessere Schätzung des Mittelwerts der Grundgesamtheit als der Mittelwert einer Stichprobe.
- Viertens „streuen“ die Mittelwerte vieler Stichproben um den tatsächlichen Mittelwert der Grundgesamtheit. Dabei kann ein einzelner Mittelwert einer Stichprobe deutlich vom tatsächlichen Mittelwert in der Grundgesamtheit abweichen.

Die in diesem Abschnitt gemachten Beobachtungen verknüpfen wir im folgenden Abschnitt mit zwei zentralen Konzepten der Inferenzstatistik: dem zentralen Grenzwertsatz und dem Standardfehler.

5.2 Zentrale Konzepte der Inferenzstatistik

Das Ziel der Inferenzstatistik ist es, von den bekannten Kennwerten einer Stichprobe (z.B. Mittelwert) auf die unbekannten Parameter einer Grundgesamtheit zu schließen. Das Beispiel in Abschnitt 5.1 hat zwei Probleme illustriert: Erstens weichen Stichprobenwerte mehr oder weniger vom tatsächlichen Wert in der Grundgesamtheit ab. Zweitens streuen die Mittelwerte vieler Stichproben um den wahren Wert der Grundgesamtheit. Für den Rückschluss von der Stichprobe auf die Grundgesamtheit greift die Inferenzstatistik auf zwei zentrale Konzepte zurück: den zentralen Grenzwertsatz (Central Limit Theorem) und den Standardfehler.

5.2.1 Zentraler Grenzwertsatz und Normalverteilung

Der zentrale Grenzwertsatz ist eine der wichtigsten Aussagen der Statistik (Hedderich und Sachs 2012, S. 248-249; siehe auch Fischer 2011). Er beschreibt die Verteilung von Mittel- und Anteilswerten, die auf Basis von Zufallsstichproben berechnet werden. In einführenden Lehrbüchern wird der zentrale Grenzwertsatz wie folgt definiert:

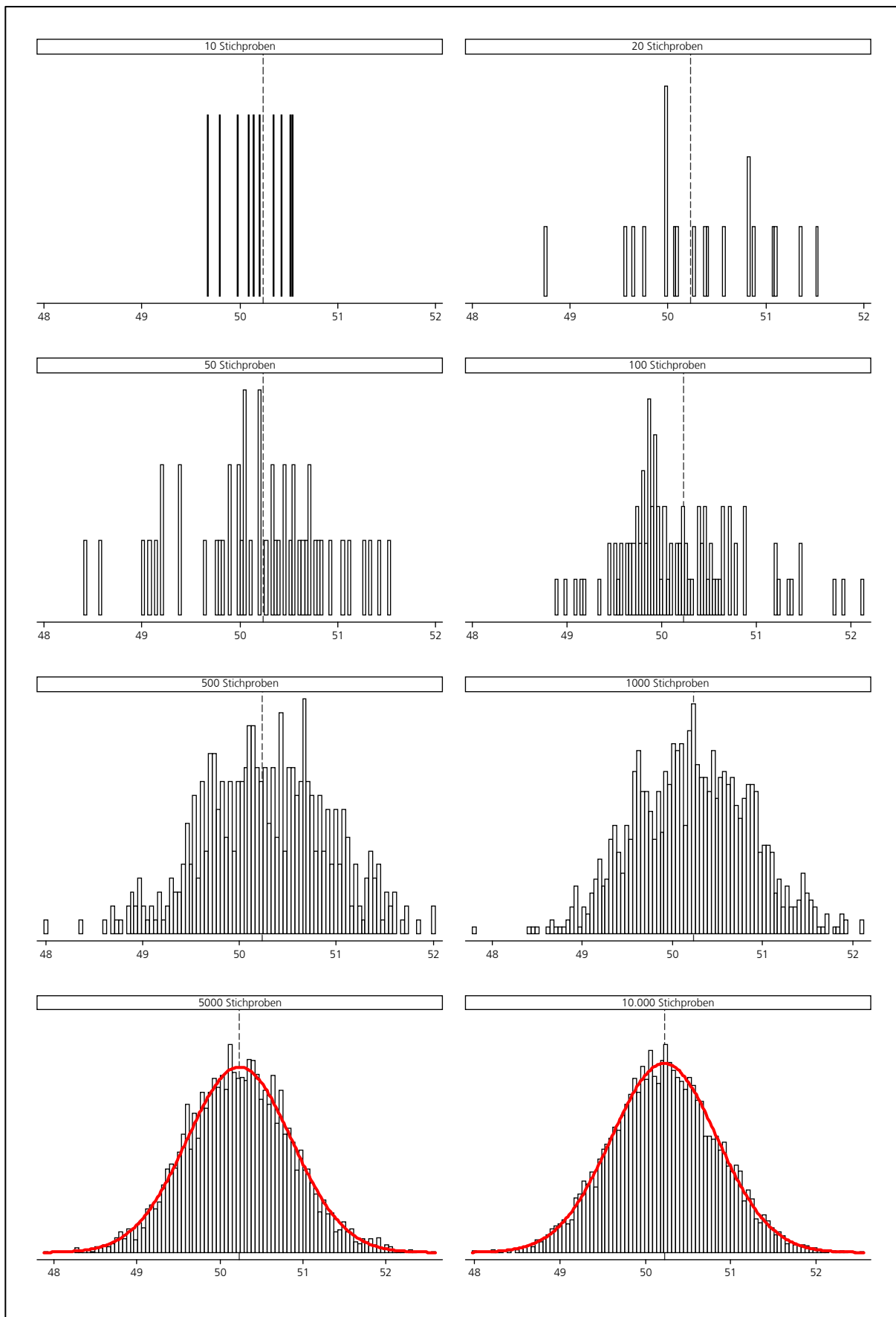
„Die Verteilung von Mittelwerten aus Stichproben des Umfangs n , die derselben Grundgesamtheit entnommen werden, geht mit wachsendem Stichprobenumfang in eine Normalverteilung über.“ (Bortz und Schuster 2010, S. 86)

„Die Mittelwerte von hinreichend großen Stichproben ($n \geq 30$) verteilen sich normal um μ , den Mittelwert der Grundgesamtheit. Diese Verteilung ist unabhängig von der Verteilung der Werte der Grundgesamtheit, d.h. diese müssen nicht normalverteilt sein.“ (Kuckartz et al. 2013, S. 140)

Ähnliche Formulierungen finden sich auch in anderen Lehrbüchern (Kühnel und Krebs 2007, S. 196-198; Gehring und Weins 2009, S. 248-249; Ludwig-Mayerhofer et al. 2014, S. 114-116; Tachtsoglou und König 2017, S. 280-281). Vereinfacht formuliert, macht der zentrale Grenzwertsatz eine Aussage über die Verteilung von Mittel- und Anteilswerten in Zufallsstichproben. Diese liegen – so der zentrale Grenzwertsatz – „normal“ um den tatsächlichen Wert der Grundgesamtheit, vorausgesetzt die jeweiligen Stichproben sind ausreichend groß (Rumsey 2010, S. 171). Ausreichend groß ist eine Stichprobe, wenn sie mindestens 30 Untersuchungsobjekte (z.B. Personen) umfasst (Gehring und Weins 2009, S. 249; Bortz und Schuster 2010, S. 87).

Wir wollen die praktische Bedeutung des zentralen Grenzwertsatzes an einem Beispiel illustrieren. In Abbildung 22 sind die Häufigkeitsverteilungen der Zufallsstichproben aus Tabelle 54 grafisch dargestellt. Oben links sind die Mittelwerte der zehn Zufallsstichproben dokumentiert. Dabei liegen die einzelnen Stichprobenmittelwerte jeweils links und rechts vom tatsächlichen Mittelwert der Grundgesamtheit.

Abbildung 22: Wiederholte Ziehung von Zufallsstichproben



Quelle: Eigene Darstellung

Mit steigender Stichprobenzahl (ab etwa 500 Stichproben) verteilen sich die Mittelwerte der einzelnen Stichproben gleichmäßig (normal) um den tatsächlichen Mittelwert in der Grundgesamtheit. In die beiden unteren Grafiken in Abbildung 22 ist neben der Häufigkeitsverteilung auch die Normalverteilung eingezeichnet. An dieser Stelle sind 5000 bzw. 10000 Stichproben nah genug an einer „unendlichen“ Anzahl von Stichproben, um die Aussage des zentralen Grenzwertsatzes zu illustrieren.

Stichprobenverteilung

Die Verteilung der einzelnen Mittelwerte (oder auch Anteilswerte) wird in der Statistik als Stichprobenverteilung (engl. sampling distribution) bezeichnet (Bortz und Schuster 2010, S. 82-83). Die Stichprobenverteilung beschreibt, wie sich aus Stichproben gewonnene Kennwerte (z.B. Mittel- und Anteilswerte) verteilen, wenn man die Stichprobenziehung unendlich oft wiederholen würde. Zwei Merkmale der Stichprobenverteilung sind besonders wichtig:

- Der Mittelwert der (unendlich) vielen Stichproben entspricht dem tatsächlichen Wert in der Grundgesamtheit. Dieser Mittelwert wird auch Erwartungswert genannt.
- Die Mittelwerte der einzelnen Stichproben streuen um diesen Erwartungswert. Dabei folgt die Streuung (bei einer Stichprobengröße von $n \geq 30$) der Normalverteilung.

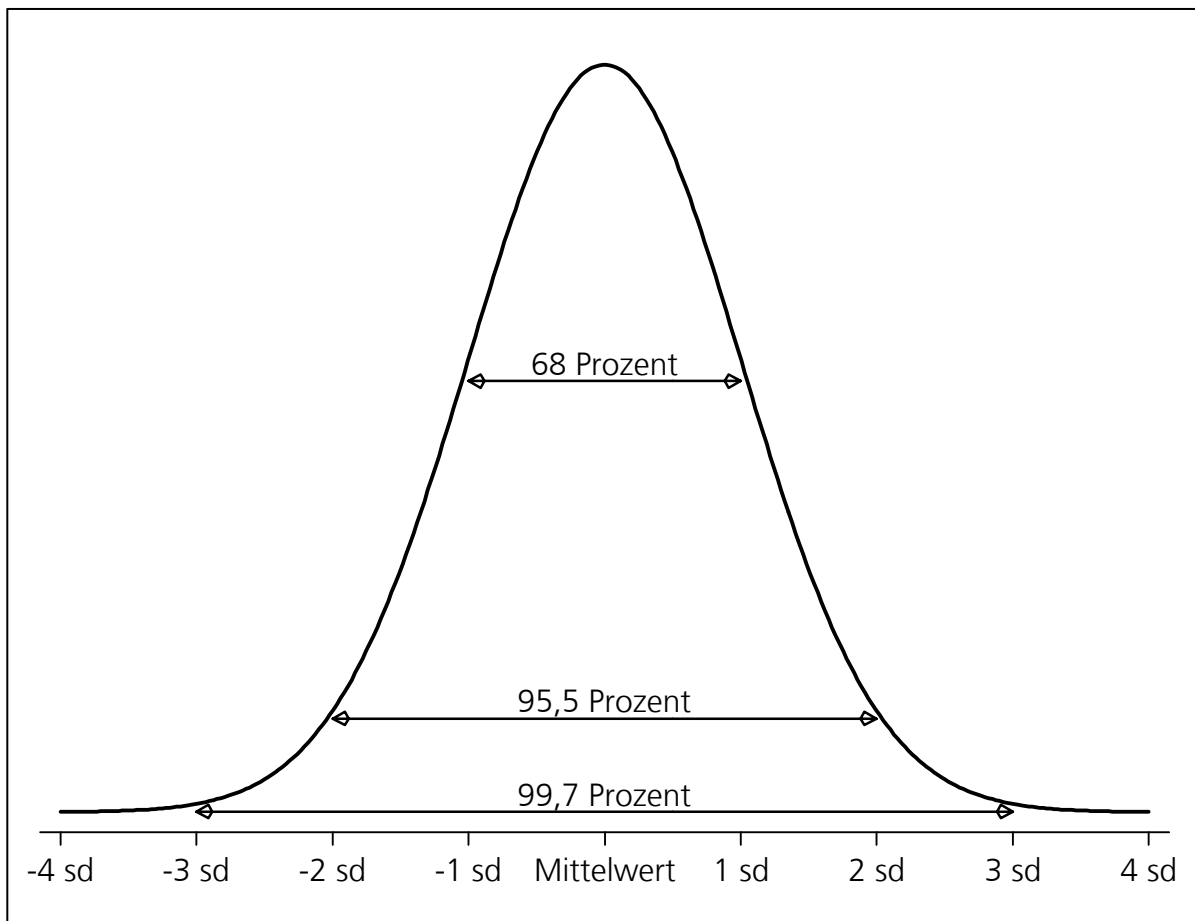
Beachten Sie: Die Stichprobenverteilung beschreibt die Verteilung der einzelnen Stichprobenmittelwerte, wenn man unendlich viele Stichproben ziehen würde. In der Forschungspraxis ziehen wir meist nur eine Stichprobe. Deshalb handelt es sich bei der Stichprobenverteilung um eine theoretische Verteilung. Wenn wir aber unendlich viele Stichproben ziehen würden, dann würden die einzelnen Stichprobenwerte normal um den tatsächlichen Wert der Grundgesamtheit streuen.

Normalverteilung

Aus dem zentralen Grenzwertsatz lässt sich ableiten, dass die Streuung der Stichprobenkennwerte (z.B. Mittel- oder Anteilswert) von (unendlich) vielen Stichproben der Normalverteilung folgt. Die typische Form der Normalverteilung – häufig auch Glockenkurve oder Gauß-Kurve genannt – ist in Abbildung 23 dargestellt. Jede Normalverteilung hat bestimmte Eigenschaften, die bei der Anwendung der Inferenzstatistik genutzt werden (Gehring und Weins 2009, S. 239; Bortz und Schuster 2010, S. 70-74; Kuckartz et al. 2013, S. 127-133):

- Die Normalverteilung ist symmetrisch.
- Mittelwert, Median und Modus sind identisch, liegen genau in der Mitte und teilen die Normalverteilung in zwei gleich große Hälften.
- Die Standardabweichung (SD) ist die typische (durchschnittliche) Entfernung aller Daten vom Mittelwert. Die durchschnittliche Streuung der Mittelwerte wird als Standardfehler bezeichnet.
- 68 Prozent aller Werte liegen in einem Bereich von \pm einer Standardabweichung.
- 95,5 Prozent aller Werte liegen in einem Bereich von \pm zwei Standardabweichungen.
- 99,7 Prozent aller Werte liegen in einem Bereich von \pm drei Standardabweichungen.

Abbildung 23: Normalverteilung



Quelle: Eigene Darstellung

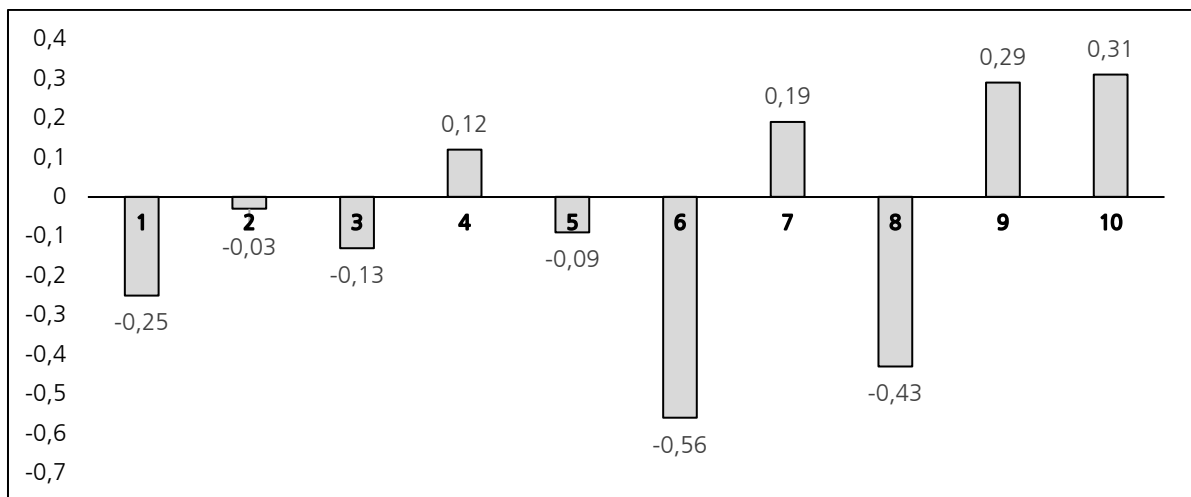
5.2.2 Standardfehler

Aus dem zentralen Grenzwertsatz lässt sich die Verteilung der Stichprobenwerte ableiten, wenn unendlich viele Stichproben gezogen werden. Die einzelnen Stichprobenwerte verteilen sich normal um den tatsächlichen Wert der Grundgesamtheit.

In Abbildung 24 sind die Abweichungen der ersten zehn Stichprobenmittelwerte vom Mittelwert unserer Grundgesamtheit dokumentiert (siehe Tabelle 53). Bei der ersten Stichprobe haben wir das durchschnittliche Alter der Grundgesamtheit beispielsweise um 0,25 Jahre unterschätzt, bei der siebten Stichprobe wurde das durchschnittliche Alter in der Grundgesamtheit um 0,19 Jahre überschätzt. Die einzelnen Abweichungen der Stichprobenwerte vom wahren Wert in der Grundgesamtheit sind offensichtlich nicht identisch, sondern variieren von Stichprobe zu Stichprobe.

Die einzelnen Stichprobenwerte streuen um den tatsächlichen Wert in der Grundgesamtheit. Bei einigen Stichproben ist die Abweichung „niedrig“, bei anderen Stichproben ist die Abweichung „hoch“. Die durchschnittliche Abweichung von (theoretisch) unendlich vielen Stichproben wird als Standardfehler bezeichnet.

Abbildung 24: Abweichungen einzelner Stichprobenmittelwerte vom wahren Mittelwert



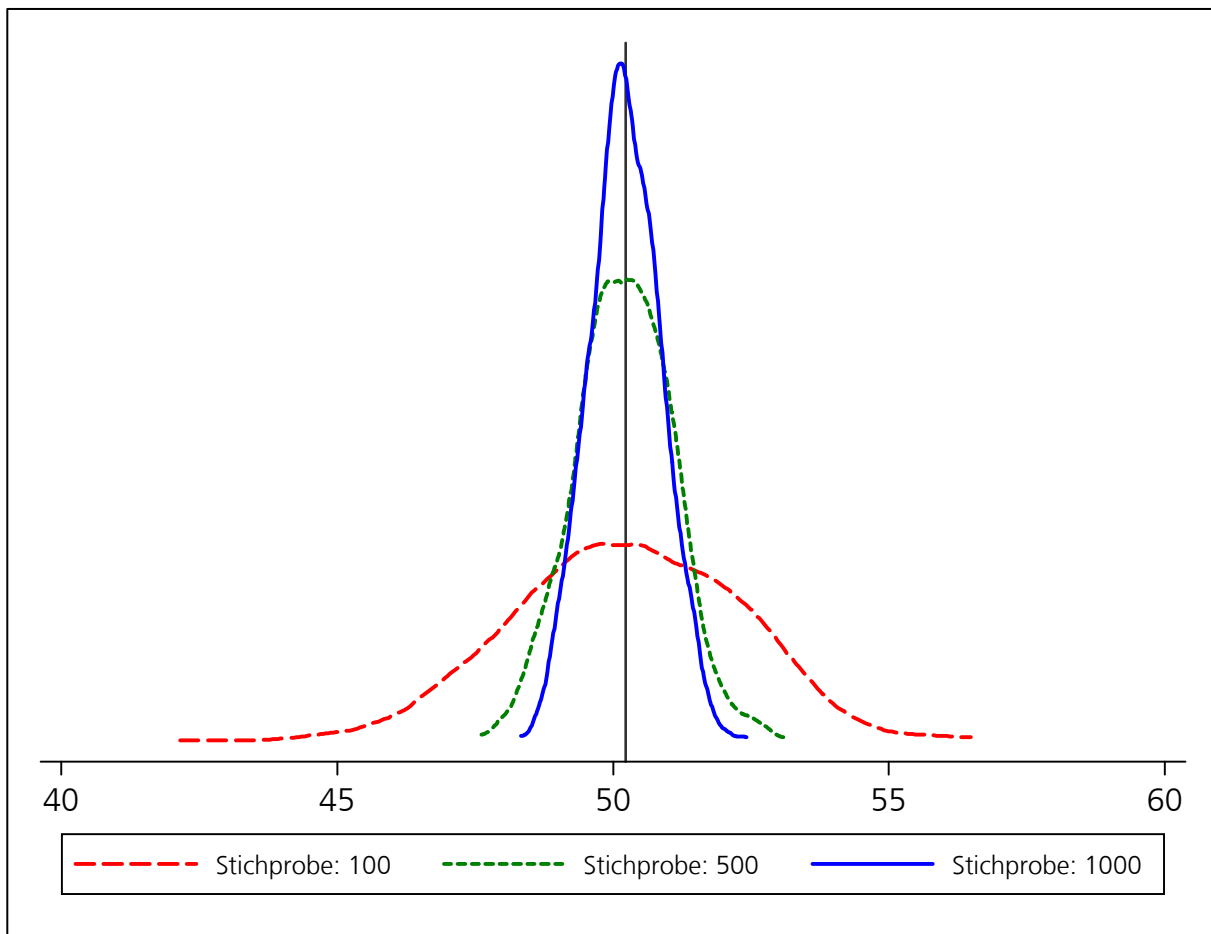
Quelle: Eigene Darstellung

In der Forschungspraxis steht uns allerdings immer nur eine Stichprobe zur Verfügung. Die zentrale Frage ist, ob unsere eine Stichprobe eine „gute“ (genaue) oder eine „schlechte“ (ungenau) Schätzung des wahren Wertes in der Grundgesamtheit bietet. Der Standardfehler informiert über die durchschnittliche Streuung der einzelnen Stichproben. Je größer der Standardfehler, desto unsicherer die Schätzung. Der Standardfehler ist damit ein Maß für die Genauigkeit einer Schätzung.

! Der Standardfehler – die durchschnittliche Abweichung vom tatsächlichen Wert in der Grundgesamtheit – ist von zwei Faktoren abhängig: Erstens von der Varianz des Merkmals in der Grundgesamtheit und zweitens von der Stichprobengröße (Gehring und Weins 2009, S. 245; Kuckartz et al. 2013, S. 141; Kühnel und Krebs 2007, S. 232-237). Beide Faktoren werden nachfolgend kurz erläutert:

- Der Standardfehler ist von der Varianz des Merkmals in der Grundgesamtheit abhängig. Warum? Je stärker ein Merkmal in der Grundgesamtheit streut, desto stärker werden die Mittelwerte des Merkmals in verschiedenen Stichproben variieren. Allerdings ist die Varianz des Merkmals in der Grundgesamtheit in der Regel nicht bekannt. Deshalb wird die Varianz des Merkmals in der Grundgesamtheit auf Basis der Varianz des Merkmals in der Stichprobe geschätzt.
- Der Standardfehler ist von der Stichprobengröße abhängig. Je größer die Stichprobe ist, desto kleiner ist der Standardfehler. Zur Illustration dieses Zusammenhangs ziehen wir erneut 1000 Stichproben aus unserer „fiktiven“ Grundgesamtheit. Allerdings variieren wir jeweils die Anzahl der Personen. Wir ziehen zunächst 1000 Stichproben mit jeweils 100 Personen, dann ziehen wir 1000 Stichproben mit jeweils 500 Personen und schließlich ziehen wir 1000 Stichproben mit jeweils 1000 Personen. In Abbildung 25 sind diese drei verschiedenen Stichprobenverteilungen dargestellt. In allen drei Fällen liegen die Stichprobenwerte zwar normal um den tatsächlichen Wert der Grundgesamtheit, aber die Breite der Verteilungen ist völlig unterschiedlich. Je größer die Anzahl der Personen ist, desto geringer fällt die durchschnittliche Streuung der Stichprobenwerte aus.

Abbildung 25: Stichprobenverteilungen bei unterschiedlicher Fallzahl



Quelle: Eigene Darstellung

Die (durchschnittliche) Streuung der Stichprobenmittelwerte oder auch Stichprobenanteile wird mit dem Standardfehler erfasst. Der Standardfehler beruht damit auf der gleichen Idee wie die Standardabweichung – beide informieren über die typische (durchschnittliche) Abweichung vom Mittelwert (Rumsey 2010, S. 171). Allerdings zielen die beiden Konzepte auf unterschiedliche Sachverhalte. In Tabelle 55 werden Standardfehler und Standardabweichung gegenübergestellt (Koschack 2008).

Standardfehler und Standardabweichung

Der Standardfehler ist ein statistisches Maß und erlaubt eine Aussage über die Genauigkeit einer Schätzung eines Merkmals auf Grundlage einer Zufallsstichprobe. Die Standardabweichung informiert über die durchschnittliche Abweichung eines Merkmals in der Grundgesamtheit bzw. Stichprobe. Während der Standardfehler stark von der Stichprobengröße abhängt (je größer die Stichprobe, desto geringer der Standardfehler), spielt die Fallzahl für die Standardabweichung eine vernachlässigbare Rolle.

Damit es in der Forschungspraxis nicht zu Missverständnissen darüber kommt, ob sich eine Aussage auf die durchschnittliche Abweichung in der Grundgesamtheit oder die Genauigkeit einer Schätzung bezieht, wird die Standardabweichung einer Stichprobenverteilung als Standardfehler bezeichnet. Der Standardfehler informiert darüber, wie unterschiedlich Stichprobenkennwerte (z.B. Mittelwerte) von Stichproben ausfallen können.

Tabelle 55: Vergleich zwischen Standardfehler und Standardabweichung

Standardfehler	Standardabweichung
– ist ein statistisches Maß	– ist ein beschreibendes Maß
– erlaubt eine Aussage über die Genauigkeit einer Schätzung in einer Zufallsstichprobe	– erlaubt eine Aussage über die Streuung eines Merkmals in der Grundgesamtheit bzw. Stichprobe
– ist von Fallzahl der Stichprobe abhängig	– ist nicht unmittelbar von der Fallzahl abhängig

Quelle: In Anlehnung an Koschack (2008, S. 259)

Standardfehler für Mittelwerte

Der Standardfehler (engl. standard error) wird mit dem griechischen Buchstaben σ (Sigma) dargestellt. Mit dem $\hat{\sigma}$ wird dokumentiert, dass es sich um eine Schätzung handelt. Um eine Schätzung handelt es sich, da die zur Berechnung des Standardfehlers benötigte Varianz des Merkmals in der Grundgesamtheit nicht bekannt ist. In der Regel stehen uns keine Informationen dazu zur Verfügung, wie stark das interessierende Merkmal (hier: Alter) in der Grundgesamtheit streut. Da die Varianz des Merkmals in der Grundgesamtheit nicht bekannt ist, wird die korrigierte Varianz (s^{2*}) bzw. korrigierte Standardabweichung (s^*) des Merkmals verwendet. Diese Information kann auf Grundlage der Stichprobe einfach berechnet werden. Neben der (korrigierten) Varianz bzw. der Standardabweichung ist für die Berechnung des Standardfehlers die Fallzahl der Stichprobe (n) erforderlich. Die Berechnung des Standardfehlers für Mittelwerte basiert auf folgender Formel:

$$\text{Standardfehler für Mittelwerte: } \hat{\sigma} = \sqrt{\frac{s^{2*}}{n}} = \frac{s^*}{\sqrt{n}}$$

Die korrigierte Varianz (s^{2*}) des Merkmals in der Stichprobe wird durch die Stichprobengröße (n) dividiert. Anschließend wird die Wurzel gezogen. Anstelle der korrigierten Varianz kann natürlich auch die korrigierte Standardabweichung verwendet werden. In diesem Fall wird die korrigierte Standardabweichung durch \sqrt{n} dividiert. Der Standardfehler informiert dann über die durchschnittliche Streuung der Stichprobenmittelwerte. Ein großer Standardfehler deutet auf eine große Unsicherheit hin, ein kleiner Standardfehler auf eine geringe Unsicherheit.



Wir wollen die Berechnung des Standardfehlers an einem Beispiel illustrieren. Bei der ersten Stichprobe beträgt der geschätzte Mittelwert des Alters 49,98 Jahre, die korrigierte Varianz des Alters liegt bei 423,99 Jahren. Die Stichprobengröße beträgt 1000 Personen. Für die Berechnung des Standardfehlers müssen diese Angaben in die Formel eingesetzt werden.

$$\text{Standardfehler für Mittelwerte: } \hat{\sigma} = \sqrt{\frac{423,99}{1000}} = \frac{20,59}{\sqrt{1000}} = 0,65$$

Der Standardfehler des Mittelwerts beträgt 0,65. Wie kann der Standardfehler nun interpretiert werden? Würden wir aus der Grundgesamtheit unendlich viele Stichproben mit jeweils 1000 Personen ziehen, dann würde der durchschnittliche Schätzwert 0,65 Jahre um den tatsächlichen Wert streuen. Je größer der Standardfehler wäre, desto stärker wäre die Unsicherheit der Schätzung.

Bei einer größeren Stichprobe wird der Standardfehler kleiner. Ersetzen wir 1000 Personen durch eine Stichprobengröße von 2000 Personen:

$$\text{Standardfehler für Mittelwerte: } \hat{\sigma} = \sqrt{\frac{423,99}{2000}} = \frac{20,59}{\sqrt{2000}} = 0,46$$

Bei einer Stichprobe von 2000 Personen beträgt der Standardfehler nur noch 0,46. Die Unsicherheit der Schätzung verringert sich. Allerdings müssten auch deutlich mehr Personen befragt werden.

Neben dem Mittelwert bzw. dem Standardfehler für einen Mittelwert sind Sozialwissenschaftlerinnen häufig auch an Anteilswerten, wie zum Beispiel dem Stimmenanteil einer Partei (Sonntagsfrage), interessiert. Auch für Anteilswerte kann der Standardfehler geschätzt werden. Mit p ist der Anteil des interessierenden Merkmals in der Stichprobe gemeint, mit n die Stichprobengröße. Auch hier wird der Standardfehler auf Grundlage von Stichprobendaten geschätzt. Dies wird mit dem $\hat{\sigma}$ dargestellt.

Standardfehler für Anteilswerte

$$\text{Standardfehler für Anteilswerte: } \hat{\sigma} = \sqrt{\frac{p * (1 - p)}{n}}$$

Die Berechnung des Standardfehlers wird an einem Beispiel illustriert. Bei einer (fiktiven) Umfrage wurde für eine Partei ein Anteilswert von 40 Prozent (0,4) ermittelt, insgesamt wurden 1000 Personen befragt. Der Anteilswert von 0,4 wird in die Formel eingetragen und mit der Gegenwahrscheinlichkeit von 0,6 multipliziert ($0,4 * 0,6 = 0,24$). Der Wert 0,24 wird durch die Fallzahl dividiert. Anschließend wird die Wurzel gezogen.

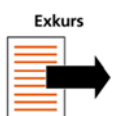


$$\text{Standardfehler für Anteilswerte: } \hat{\sigma} = \sqrt{\frac{0,4 * (1 - 0,4)}{1000}} = \sqrt{\frac{0,4 * 0,6}{1000}} = \sqrt{\frac{0,24}{1000}} = 0,015$$

Für unsere Partei liegt der Standardfehler bei 0,015. Auch hier deutet ein kleinerer Standardfehler auf eine präzisere Schätzung und ein größerer Standardfehler auf eine unpräzisere Schätzung hin.

Der Standardfehler ist ein Maß für die Unsicherheit einer Schätzung und erforderlich, um sogenannte Konfidenzintervalle zu berechnen (siehe Abschnitt 5.3.2). Bei Konfidenzintervallen handelt es sich um Bereiche, in denen der tatsächliche Wert der Grundgesamtheit mit einer bestimmten Sicherheit vermutet wird.

Die oben präsentierten Formeln für die Berechnung des Standardfehlers gelten für Stichproben, die *mit* Zurücklegen gezogen wurden (Kühnel und Krebs 2007, S. 232-237; Gehring und Weins 2009, S. 245-246; Braunecker 2016, S. 107-108). In der Forschungspraxis sind wir allerdings mit Stichproben *ohne* Zurücklegen konfrontiert. Ein Untersuchungsobjekt (z.B. eine Person), das einmal in die Stichprobe gelangt ist, wird bei einer sozialwissenschaftlichen Studie (z.B. ALLBUS oder ESS) nicht ein zweites Mal gezogen bzw. befragt.



Bei Stichproben *ohne* Zurücklegen muss für die Berechnung des Standardfehlers formal ein Korrekturfaktor (Endlichkeitsfaktor) berücksichtigt werden, der wie folgt definiert ist:

$$\text{Korrekturfaktor bei Stichproben ohne Zurücklegen: } \sqrt{\frac{N-n}{N-1}}$$

Dieser Korrekturfaktor (grau hinterlegt) berücksichtigt das Verhältnis zwischen Grundgesamtheit (N) und Stichprobe (n) und wird mit dem berechneten Standardfehler multipliziert.

$$\hat{\sigma} = \frac{s^*}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

Aus praktischer Sicht kann dieser Korrekturfaktor allerdings vernachlässigt werden, wenn es das Verhältnis zwischen Grundgesamtheit und Stichprobe erlaubt. Mit anderen Worten: Nur bei einer großen Stichprobe aus einer kleinen Grundgesamtheit beeinflusst der Korrekturfaktor das Ergebnis. In unserem Beispiel – 1000 Personen aus einer Grundgesamtheit von 50000 Personen – beträgt der Korrekturfaktor 0,99. Dieser Wert, multipliziert mit dem oben berechneten Standardfehler (0,65), verringert den Standardfehler gerade einmal um 0,01.

$$\hat{\sigma} = \frac{20,59}{\sqrt{1000}} * \sqrt{\frac{50.000-1000}{50.000-1}} = 0,65 * 0,99 = 0,64$$

Bestünde die Grundgesamtheit allerdings statt aus 50000 Personen lediglich aus 5000 Personen, hätte dies stärkere Auswirkungen auf den Standardfehler. Im konkreten Fall läge der Korrekturfaktor dann bei 0,89 und der Standardfehler bei 0,58.

$$\hat{\sigma} = \frac{20,59}{\sqrt{1000}} * \sqrt{\frac{5000-1000}{5000-1}} = 0,65 * 0,89 = 0,58$$

Allerdings werden in den Sozialwissenschaften in der Regel Stichproben aus sehr großen Grundgesamtheiten entnommen (z.B. wahlberechtigte Bevölkerung in Deutschland), so dass der Korrekturfaktor vernachlässigt werden kann. Nach Gehring und Weins (2009, S. 245) kann bei Stichproben *ohne* Zurücklegen der Korrekturfaktor vernachlässigt werden, wenn der Umfang der Grundgesamtheit (N) mindestens dem 20-Fachen des Umfangs der Stichprobe entspricht (ähnlich auch Braunecker 2016, S. 108). Diese Voraussetzung ist bei sozialwissenschaftlichen Umfragen praktisch immer erfüllt, so dass der Korrekturfaktor in der Forschungspraxis keine Bedeutung hat. Da das Verhältnis zwischen Grundgesamtheit und Stichprobe bei den folgenden Berechnungen immer größer als 20 sein wird, werden wir den Korrekturfaktor im Folgenden auch nicht weiter berücksichtigen.

5.3 Schätzungsarten

Die Schätzung der Populationsparameter – also der Merkmale der Grundgesamtheit – kann als Punktschätzung oder als Intervallschätzung vorgenommen werden. Bei einer Punktschätzung wird der Parameter der Grundgesamtheit (z.B. Mittel- oder Anteilswert) durch einen einzigen Wert der Stichprobe dargestellt. Bei einer Intervallschätzung wird ein Intervall (Bereich) angegeben, in dem der gesuchte Parameter der Grundgesamtheit vermutet wird (Gehring und Weins 2009, S. 254; Tachtsoglou und König 2017, S. 278).

5.3.1 Punktschätzung

Wird nur ein Schätzwert für den unbekannten Parameter der Grundgesamtheit (z.B. Mittel- oder Anteilswert) berechnet, dann handelt es sich um eine Punktschätzung (Hedderich und Sachs 2012, S. 295). Wir berechnen beispielsweise das arithmetische Mittel des Alters der Personen einer Zufallsstichprobe, um das durchschnittliche Alter der Grundgesamtheit zu schätzen. Bei der Sonntagsfrage („Wenn am nächsten Sonntag wirklich Bundestagswahl wäre...“) werden die Anteils- werte der Parteien in der Stichprobe berechnet, um Aussagen über die Parteianteile in der Grundgesamtheit treffen zu können.

Der Mittelwert und der Anteilswert einer Zufallsstichprobe sind die bekanntesten Schätzer, um Aussagen über die Grundgesamtheit zu machen. Vor der Darstellung der einzelnen Schätzer werden zunächst die Kriterien einer „guten“ Schätzung vorgestellt.

5.3.1.1 Kriterien einer „guten“ Schätzung

Allgemein gelten der Mittelwert bzw. der Anteilswert einer Zufallsstichprobe als „guter“ Schätzer für den Mittelwert bzw. Anteilswert in der Grundgesamtheit. Was sind aber die Kriterien für eine „gute“ Schätzung? In der „Theory of Statistical Estimation“ hat Fisher (1925a) Kriterien formuliert, die ein statistischer Kennwert erfüllen muss, um einen Populationsparameter der Grundgesamtheit bestmöglich schätzen zu können. Dies sind Erwartungstreue, Effizienz, Konsistenz und Suffizienz (Kühnel und Krebs 2007, S. 227-231; Bortz und Schuster 2010, S. 88-90; Ludwig-Mayerhofer et al. 2014, S. 120-124; Tachtsoglou und König 2017, S. 290-291).

Ein erwartungstreuer Schätzer ist ein unverzerrter Schätzer; er entspricht im Mittel dem Wert der Grundgesamtheit. Die Erwartungstreue bzw. Unverzerrtheit eines Schätzers lässt sich anhand des Beispiels aus Abschnitt 5.1 illustrieren. Auf Grundlage einer Stichprobe wird eine Aussage über das Durchschnittsalter der Grundgesamtheit gemacht (siehe Tabelle 53). Eine einzelne Stichprobe hat das Alter der Grundgesamtheit entweder überschätzt oder unterschätzt. Der Mittelwert vieler – im Prinzip unendlich vieler – Stichproben entsprach allerdings dem Mittelwert der Grundgesamtheit. Mit anderen Worten: Ein erwartungstreuer Schätzer wird im Mittel den Wert der Grundgesamtheit nicht systematisch unterschätzen bzw. überschätzen, sondern dem tatsächlichen Wert der Grundgesamtheit entsprechen.

Erwartungstreue

Bei einer Stichprobe wird ein erwartungstreuer Schätzer den „wahren“ Wert der Grundgesamtheit in der Regel über- bzw. unterschätzen. Ein erwartungstreuer Schätzer zeichnet sich aber dadurch aus, dass bei „unendlich“ vielen Stichproben der Mittelwert der einzelnen Schätzungen

dem tatsächlichen Wert der Grundgesamtheit entspricht. Die Abweichungen der einzelnen Stichproben gleichen sich aus und der tatsächliche Wert der Grundgesamtheit wird nicht systematisch unterschätzt bzw. überschätzt.

Effizienz Die zweite Eigenschaft eines „guten“ Schätzers ist die Effizienz. Damit ist die Präzision einer Schätzung gemeint. Der Standardfehler informiert über die durchschnittliche Abweichung des geschätzten Werts vom wahren Wert. Je kleiner der Standardfehler, also je geringer die durchschnittliche Abweichung des geschätzten Werts vom wahren Wert ist, desto effizienter ist der Schätzer.

Konsistenz Die Konsistenz beschreibt das Verhalten eines Schätzers bei Vergrößerung der Stichprobe (Gehring und Weins 2009, S. 257). Ein Schätzer wird als konsistent bezeichnet, wenn bei steigender Stichprobengröße die Differenz zwischen dem geschätzten Wert und dem wahren Wert der Grundgesamtheit geringer wird.

In Tabelle 56 wird die Konsistenz eines Schätzers an einem Beispiel illustriert. Wir nutzen unsere Grundgesamtheit von 50000 Personen; das mittlere Alter dieser Grundgesamtheit liegt bei 50,23 Jahren, die Standardabweichung beträgt 20,12 (siehe Abschnitt 5.1). Wir ziehen zunächst eine sehr kleine Stichprobe von zehn Personen; das mittlere Alter dieser Stichprobe liegt bei 64,10 Jahren und weicht stark vom Mittelwert der Grundgesamtheit ab (13,87 Jahre). Die zweite Stichprobe umfasst 50 Personen. Die Abweichung ist mit 4,23 Jahren noch sehr ausgeprägt, aber deutlich geringer als bei der ersten Stichprobe.

Bei der dritten Stichprobe ($n = 100$) beträgt die Abweichung nur noch 0,64 Jahre, bei der vierten Stichprobe ($n = 1000$) wird das mittlere Alter der Grundgesamtheit lediglich noch um 0,26 Jahre unterschätzt. Bei der Stichprobe mit 10000 Befragten liegt die Differenz zwischen der Stichprobe und der Grundgesamtheit nur noch bei 0,14 Jahren. Mit steigender Stichprobengröße hat sich die Differenz zwischen dem Mittelwert der Stichprobe und dem Wert der Grundgesamtheit verringert. Diese Eigenschaft eines Schätzers wird als Konsistenz bezeichnet.

Tabelle 56: Mittelwerte von Zufallsstichproben

Größe der Stichprobe	Mittelwert der Stichprobe	Standardabweichung der Stichprobe	Abweichung vom Mittelwert der Grundgesamtheit (50,23 Jahre)
10	64,10	18,39	13,87
50	54,46	22,54	4,23
100	50,87	20,69	0,64
1000	49,97	20,59	-0,26
10000	50,37	20,16	0,14

Quelle: Eigene Darstellung

Bortz und Schuster (2010, S. 90) bezeichnen einen Schätzer als suffizient oder erschöpfend, „wenn er alle in einer Stichprobe enthaltenen Informationen berücksichtigt“.

Suffizienz

Bei der Schätzung von Populationsmerkmalen sind Sozialwissenschaftlerinnen in der Regel an folgenden Informationen interessiert: dem Mittelwert bzw. dem Anteilswert eines Merkmals in der Grundgesamtheit sowie der Streuung des Merkmals. Im Folgenden werden die Schätzer für diese Informationen vorgestellt.

5.3.1.2 Schätzer für Mittelwerte

In der Statistik wird der wahre Mittelwert einer Grundgesamtheit häufig mit dem griechischen Buchstaben μ (lies: Mü) dargestellt (Gehring und Weins 2009, S. 254-255; Griffiths 2009, S. 445). Dieser wahre Mittelwert der Grundgesamt – also beispielsweise das durchschnittliche Alter der Grundgesamtheit – ist in der Regel nicht bekannt. Wir nutzen die Daten einer Zufallsstichprobe, um den Mittelwert der Grundgesamtheit zu schätzen. Der Punktschätzer für den Mittelwert wird mit einem \wedge -Symbol gekennzeichnet ($\hat{\mu}$). Mit dem \wedge wird angezeigt, dass es sich um eine Schätzung handelt. Mit den Daten einer Zufallsstichprobe lässt sich der Mittelwert der Grundgesamtheit schätzen.

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Wir schätzen den Mittelwert der Grundgesamtheit ($\hat{\mu}$) durch den Mittelwert der Stichprobe (\bar{x}). Dazu werden die Werte in der Stichprobe addiert und durch die Stichprobengröße dividiert. Der Mittelwert einer Stichprobe bietet die bestmögliche Schätzung für den Mittelwert einer Grundgesamtheit. Der Schätzwert $\hat{\mu}$ einer Stichprobe wird allerdings mehr oder weniger stark vom tatsächlichen Mittelwert der Grundgesamtheit μ abweichen.

5.3.1.3 Schätzer für Anteilswerte

Der Anteilswert eines Merkmals in der Grundgesamtheit (z.B. Anteil der SPD-Wähler in der Grundgesamtheit) wird mit dem griechischen Buchstaben θ (lies: Theta) gekennzeichnet. Auch dieser Wert ist in der Regel nicht bekannt und muss durch die Stichprobendaten geschätzt werden. Dies wird mit dem \wedge -Symbol gekennzeichnet ($\hat{\theta}$). Dabei dient der Anteil des Merkmals in der Stichprobe (z.B. Anteil der SPD-Wählerinnen in der Stichprobe) als Schätzer für den Anteil in der Grundgesamtheit:

$$\hat{\theta} = p$$

Dabei ist $\hat{\theta}$ die Schätzung des Anteils in der Grundgesamtheit; p ist der Anteil des Merkmals in der Stichprobe.

5.3.1.4 Schätzer für Varianzen

Statistisch lässt sich zeigen, dass der Mittelwert bzw. der Anteilwert einer Zufallsstichprobe die formulierten Kriterien für eine „gute“ Schätzung erfüllt (siehe ausführlich Bortz und Schuster 2010, S. 90-92). Der Mittelwert bzw. der Anteilwert einer Zufallsstichprobe bietet damit die bestmögliche Schätzung für den Mittelwert bzw. Anteilwert in einer Grundgesamtheit.

Die empirische Varianz (s^2) bzw. empirische Standardabweichung (s) ist allerdings nicht die bestmögliche Schätzung für die Varianz bzw. Standardabweichung eines Merkmals in der Grundgesamtheit (zur Erinnerung: Die Varianz bzw. Standardabweichung ist ein Maß für die Streuung der Daten). Die empirische Varianz (s^2) bzw. empirische Standardabweichung (s) wird mit folgender Formel berechnet:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Die empirische Varianz bzw. empirische Standardabweichung unterschätzt allerdings die Varianz bzw. Standardabweichung in der Grundgesamtheit. Je kleiner die Stichprobe, desto stärker wird die Varianz bzw. Standardabweichung in der Grundgesamtheit abweichen. Mit anderen Worten: Je größer die Stichprobe ausfällt, desto geringer ist die Unterschätzung.

Für die Schätzung der Varianz bzw. Standardabweichung eines Merkmals auf Basis einer Zufallsstichprobe wird die Formel leicht modifiziert. Statt durch $\frac{1}{n}$ wird durch $\frac{1}{n-1}$ dividiert. Es handelt sich um die korrigierte Varianz (s^{*2}) bzw. um die korrigierte Standardabweichung (s^*).

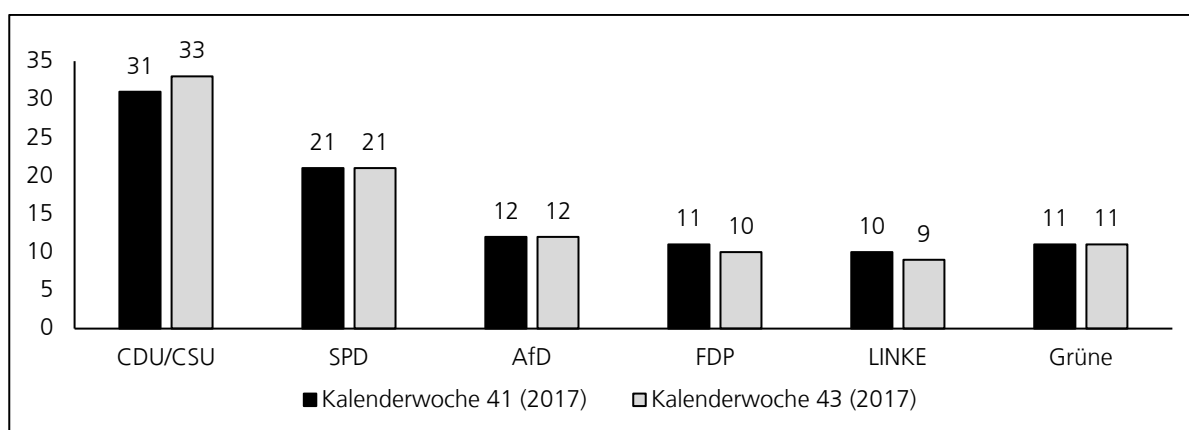
$$s^{*2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \qquad s^* = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Soll die Varianz bzw. Standardabweichung eines Merkmals (z.B. Alter) auf Grundlage einer Zufallsstichprobe geschätzt werden, dann bietet die korrigierte Varianz bzw. die korrigierte Standardabweichung die bestmögliche Schätzung (Ludwig-Mayerhofer et al. 2014, S. 75, 122). Bei der korrigierten Varianz bzw. Standardabweichung handelt es sich um einen erwartungstreuen (unverzerrten) Schätzer der Varianz bzw. Standardabweichung der Grundgesamtheit. Die empirische Varianz bzw. Standardabweichung unterschätzt dagegen die Varianz bzw. Standardabweichung eines Merkmals in der Grundgesamtheit systematisch.

5.3.2 Intervallschätzung

Bei der Punktschätzung wird die (unvermeidliche) Stichprobenstreuung nicht dokumentiert. Dies hat zur Folge, dass insbesondere Personen ohne Statistikkenntnisse die Genauigkeit von Stichprobenergebnissen häufig überschätzen. Ein typisches Beispiel sind Ergebnisse von (politischen) Meinungsumfragen in den Medien. Abbildung 26 dokumentiert das Ergebnis des Politbarometers der Forschungsgruppe Wahlen in den Kalenderwochen 41 und 43 des Jahres 2017 zur Projektion der Sonntagsfrage (Wahlentscheidung). Demnach hat die CDU/CSU in der Kalenderwoche 43 um zwei Prozentpunkte zugelegt, während FDP und LINKE jeweils einen Prozentpunkt verloren haben.

Abbildung 26: Ergebnisse des Politbarometers zu zwei Zeitpunkten (in Prozent)



Quelle: Eigene Darstellung

Veränderungen von ein bzw. zwei Prozentpunkten sind häufig schon Anlass für kontroverse Diskussionen über die Ursache dieser Veränderungen. Möglicherweise gibt es tatsächlich auch inhaltliche Gründe für die jeweilige Zunahme bzw. Abnahme, aber vermutlich ist die Veränderung in erster Linie eine Konsequenz des Stichprobenfehlers. Die Forschungsgruppe Wahlen informiert zwar über den Stichprobenfehler, aber in der öffentlichen Debatte und auch in der Berichterstattung der Medien wird dieser häufig nicht zur Kenntnis genommen.¹⁵

Bei der Intervallschätzung wird statt eines Wertes ein Bereich angegeben, in dem der gesuchte Parameter der Grundgesamtheit (z.B. Anteile von Parteien) vermutet wird. Ein solcher Bereich wird als Vertrauens- oder Konfidenzintervall bezeichnet (Gehring und Weins 2009, S. 260; Bortz und Schuster 2010, S. 93; Ludwig-Mayerhofer et al. 2014, S. 125-136). Ein Vertrauens- oder Konfidenzintervall wirkt damit auf den ersten Blick zwar ungenauer als eine Punktschätzung, bildet aber die Unsicherheit von Stichprobenergebnissen deutlich besser ab als eine Punktschätzung.

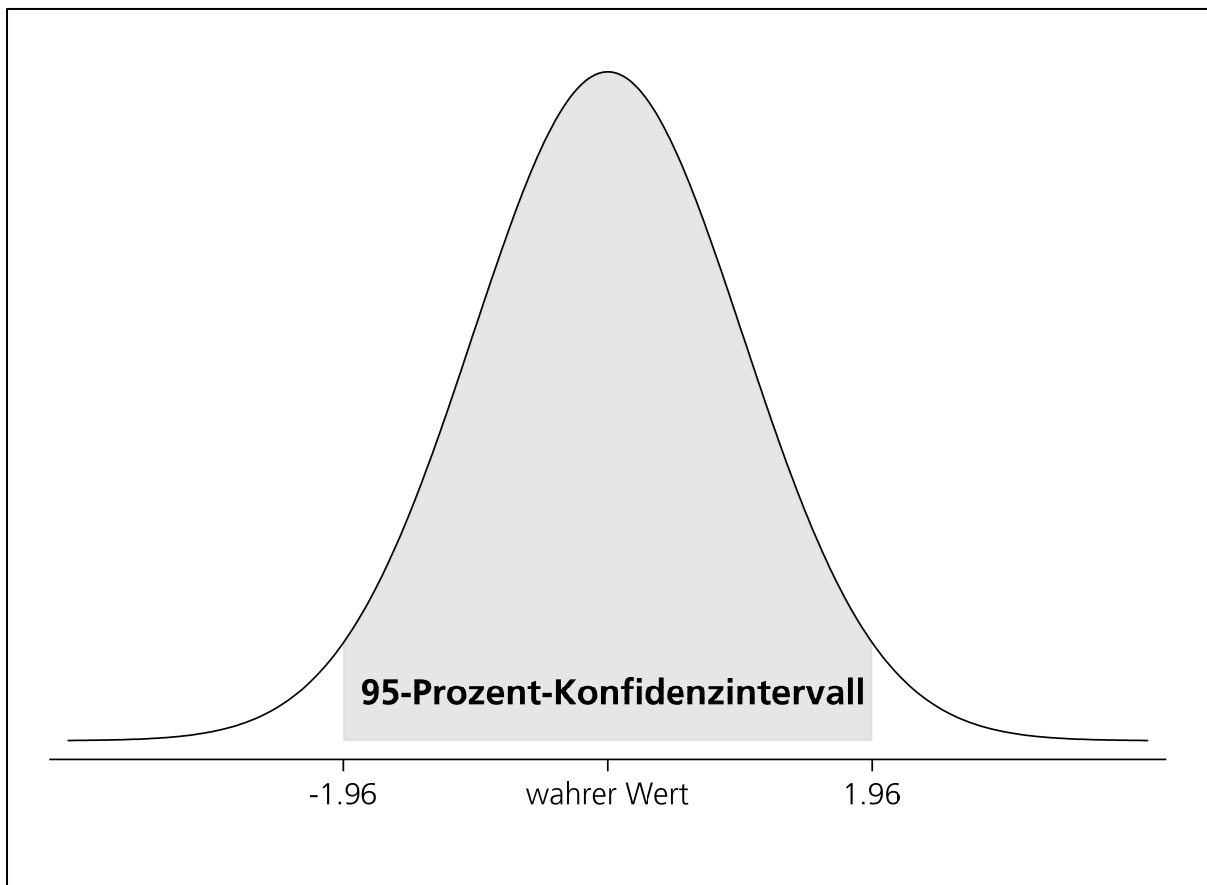
Konfidenzintervall

Ein Konfidenzintervall, auch Vertrauensintervall genannt, ist ein Bereich (Intervall), in dem der tatsächliche Wert der Grundgesamtheit mit großer Sicherheit (meist 95 oder 99 Prozent) vermutet wird.

Für die Berechnung eines Konfidenzintervalls greifen wir auf die zentralen Konzepte der Inferenzstatistik zurück. Auf Basis des zentralen Grenzwertsatzes wissen wir, dass hinreichend große Stichproben ($n > 30$) normal um den tatsächlichen Wert der Grundgesamtheit streuen. Die Streuung der (unendlich) vielen Stichproben wird dabei durch den Standardfehler erfasst. Abbildung 27 und Abbildung 28 zeigen zwei wichtige Konfidenzintervalle. Abbildung 27 illustriert das sogenannte 95-Prozent-Konfidenzintervall, Abbildung 28 das 99-Prozent-Konfidenzintervall.

¹⁵ In den entsprechenden Veröffentlichungen der Forschungsgruppe Wahlen findet sich folgende Information: „Der Fehlerbereich beträgt bei einem Anteilswert von 40 Prozent rund +/- drei Prozentpunkte und bei einem Anteilswert von 10 Prozent rund +/- zwei Prozentpunkte.“ Die Daten des Politbarometers werden bei GESIS archiviert und stehen für Sekundäranalysen kostenlos zur Verfügung (für weitere Informationen siehe www.gesis.org/wahlen/politbarometer).

Abbildung 27: 95-Prozent-Konfidenzintervall



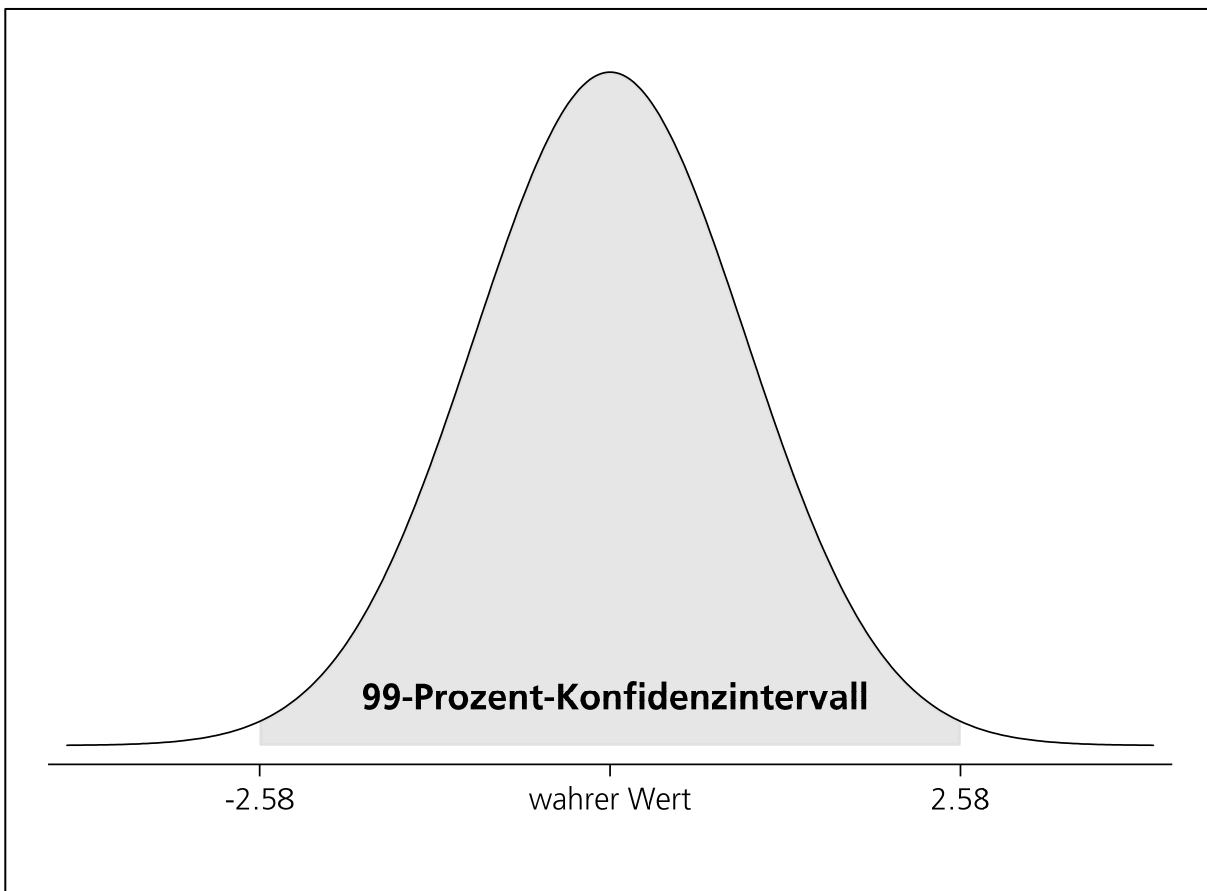
Quelle: Eigene Darstellung

Wie können diese beiden Konfidenzintervalle interpretiert werden? Stellen Sie sich vor, Sie ziehen unendlich viele Stichproben aus der Grundgesamtheit und berechnen für jede dieser Stichproben den Mittel- oder Anteilswert. Dann liegen 95 Prozent aller Stichprobenwerte in einem Bereich von $\pm 1,96$ Standardfehler um den tatsächlichen Wert der Grundgesamtheit. Die graue Fläche in Abbildung 27 markiert das 95-Prozent-Konfidenzintervall. Diese Fläche umfasst 95 Prozent der Werte, die beiden nicht schraffierten Flächen rechts und links davon beinhalten jeweils 2,5 Prozent der Stichprobenwerte.

Das 95-Prozent-Konfidenzintervall umfasst also 95 Prozent aller Stichprobenwerte. Falls Ihnen 95 Prozent zu wenig sind, bietet sich das 99-Prozent-Konfidenzintervall an (Abbildung 28). Beim 99-Prozent-Konfidenzintervall liegen 99 Prozent aller Stichprobenwerte in einem Bereich $\pm 2,58$ Standardfehler um den tatsächlichen Wert. Die graue Fläche in Abbildung 28 markiert das 99-Prozent-Konfidenzintervall. Diese Fläche umfasst 99 Prozent der Werte, die beiden nicht schraffierten Flächen links und rechts beinhalten jeweils 0,5 Prozent der Stichprobenwerte.

Selbstverständlich geht es auch noch genauer: das 99,9-Prozent-Konfidenzintervall beinhaltet 99,9 Prozent aller Stichprobenwerte in einem Bereich von $\pm 3,29$ Standardfehler um den tatsächlichen Wert. Mit anderen Worten: 99,9 Prozent aller Stichprobenwerte liegen dann in einem Bereich $\pm 3,29$ Standardfehler um den wahren Wert der Grundgesamtheit. In der (sozialwissenschaftlichen) Forschungspraxis sind aber 95- und 99-Prozent-Konfidenzintervall meist ausreichend.

Abbildung 28: 99-Prozent-Konfidenzintervall



Quelle: Eigene Darstellung

5.3.2.1 Konfidenzintervalle für Mittelwerte

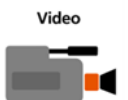
Für die Berechnung eines Konfidenzintervalls eines Mittelwerts sind zwei Angaben erforderlich: Erstens der Mittelwert der Stichprobe und zweitens der Standardfehler dieses Mittelwerts. Der Mittelwert einer Stichprobe (\bar{x}) ist die Summe der einzelnen Werte dividiert durch die Anzahl der Werte:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Der Standardfehler des Mittelwerts ($\hat{\sigma}$) wird mit folgender Formel geschätzt:

$$\hat{\sigma} = \frac{s^*}{\sqrt{n}}$$

Dabei ist s^* die korrigierte Standardabweichung und n die Stichprobengröße. Für die Berechnung des 95-Prozent-Konfidenzintervalls wird der Wert 1,96 mit dem Standardfehler multipliziert ($1,96 * \hat{\sigma}$). Dieser Wert addiert mit dem Mittelwert ($\bar{x} + 1,96 * \hat{\sigma}$) bildet die obere Grenze des 95-Prozent-Konfidenzintervalls. Die untere Grenze des 95-Prozent-Konfidenzintervalls ergibt sich nach der Subtraktion vom Mittelwert ($\bar{x} - 1,96 * \hat{\sigma}$).



$$\bar{x} \pm 1,96 * \hat{\sigma}$$

Für die Berechnung des 99-Prozent-Konfidenzintervalls wird statt dem Wert 1,96 der Wert 2,58 verwendet:

$$\bar{x} \pm 2,58 * \hat{\sigma}$$



Die Berechnung des Konfidenzintervalls wird an einem Beispiel illustriert. Wir berechnen das 95-Prozent-Konfidenzintervall des Alters in der ALLBUS 2014. Das mittlere Alter der Befragten in der ALLBUS-Stichprobe liegt bei 49,02 Jahren. Die korrigierte Standardabweichung des Mittelwerts ist 17,55; die Stichprobengröße ist 3467. Mit diesen Angaben wird der Standardfehler des Mittelwerts geschätzt:

$$\hat{\sigma} = \frac{s^*}{\sqrt{n}} = \frac{17,55}{\sqrt{3467}} = 0,298$$

Der Standardfehler des Mittelwerts liegt bei 0,298. Der Mittelwert (49,02) und der Standardfehler (0,298) werden in die Formel zur Berechnung des 95-Prozent-Konfidenzintervalls eingetragen:

$$\bar{x} \pm 1,96 * \hat{\sigma} = 49,02 \pm 1,96 * 0,298$$

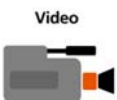
Die untere Grenze des Konfidenzintervalls liegt bei 48,44 Jahren; die obere Grenze des Konfidenzintervalls liegt bei 49,60 Jahren. Mit einer Sicherheit von 95 Prozent liegt das mittlere Alter in der Grundgesamtheit zwischen 48,44 und 49,60 Jahren. Häufig findet sich auch folgende Formulierung zur Beschreibung des Konfidenzintervalls: Bei einer Irrtumswahrscheinlichkeit von fünf Prozent ist zu vermuten, dass der Mittelwert der Grundgesamt zwischen 48,44 und 49,60 Jahren liegt.

Falls die Irrtumswahrscheinlichkeit von fünf Prozent zu gering ist, können wir eine Irrtumswahrscheinlichkeit von einem Prozent festlegen und das 99-Prozent-Konfidenzintervall berechnen. Es muss lediglich der Wert 1,96 durch 2,58 ersetzt werden:

$$\bar{x} \pm 2,58 * \hat{\sigma} = 49,02 \pm 2,58 * 0,298$$

Die untere Grenze des 99-Prozent-Konfidenzintervalls liegt bei 48,25 Jahren; die obere Grenze des Konfidenzintervalls liegt bei 49,79 Jahren. Mit einer Sicherheit von 99 Prozent liegt das mittlere Alter in der Grundgesamtheit zwischen 48,25 und 49,79 Jahren. Eine geringere Irrtumswahrscheinlichkeit führt damit zu einem breiteren Konfidenzintervall.

5.3.2.2 Konfidenzintervalle für Anteilswerte



Die Vorgehensweise für die Berechnung eines Konfidenzintervalls für Anteilswerte (z. B. Parteianteile) ist mit der Berechnung für Mittelwerte vergleichbar. Statt eines Mittelwerts einer Stichprobe benötigen wir den Anteilswert p (relative Häufigkeit). Die relative Häufigkeit eines Merkmals ist die absolute Häufigkeit dividiert durch die Fallzahl:

$$p = \frac{\text{absolute Häufigkeit}}{n}$$

Der Standardfehler des Anteilswerts ($\hat{\sigma}$) wird mit folgender Formel geschätzt (siehe auch Abschnitt 5.2.2):

$$\hat{\sigma} = \sqrt{\frac{p * (1 - p)}{n}}$$

Dabei ist p der Anteilswert und n die Stichprobengröße. Für die Berechnung des 95-Prozent-Konfidenzintervalls wird wieder der Wert 1,96 mit dem Standardfehler multipliziert ($1,96 * \hat{\sigma}$). Dieser Wert, addiert mit dem Anteilswert ($p + 1,96 * \hat{\sigma}$), bildet die obere Grenze des 95-Prozent-Konfidenzintervalls. Die untere Grenze des 95-Prozent-Konfidenzintervalls ergibt sich nach der Subtraktion vom Anteilswert ($p - 1,96 * \hat{\sigma}$).

$$p \pm 1,96 * \hat{\sigma}$$

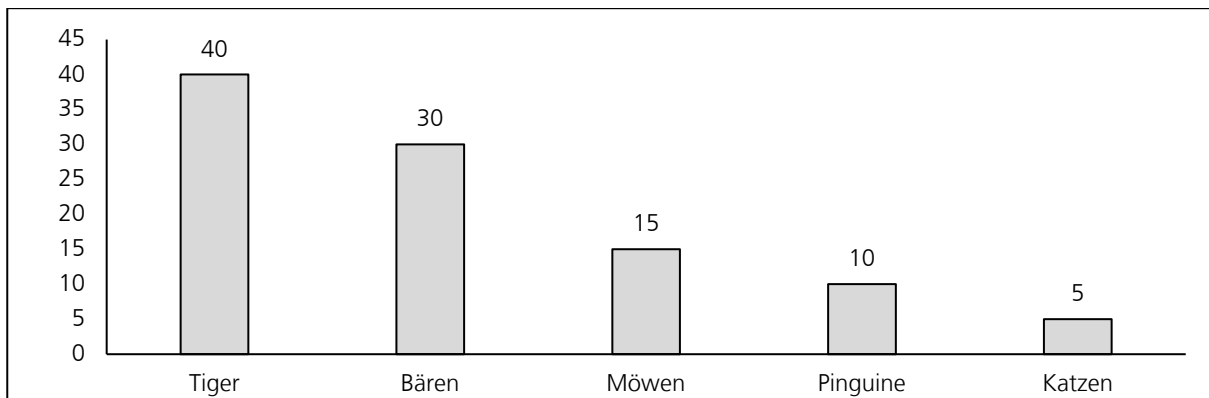
Für die Berechnung des 99-Prozent-Konfidenzintervalls wird statt des Werts 1,96 der Wert 2,58 verwendet:

$$p \pm 2,58 * \hat{\sigma}$$

Wir illustrieren die Berechnung eines Konfidenzintervalls an einem Beispiel. Bei einer Umfrage mit einer Zufallsstichprobe von 1000 Personen wurden die Personen gefragt, welche Partei sie wählen würden, wenn am nächsten Sonntag Wahl wäre. Abbildung 29 zeigt, dass 40 Prozent der Personen die Tigerpartei und 30 Prozent die Bärenpartei wählen würden. 15 Prozent der Personen votieren für die Möwenpartei, zehn Prozent für die Partei der Pinguine und fünf Prozent für die Katzenpartei.



Abbildung 29: Fiktive Befragung zur Wahlentscheidung von 1000 Personen (in Prozent)



Quelle: Eigene Darstellung

An dieser Stelle berechnen wir das 95-Prozent-Konfidenzintervall der Tigerpartei. Der Anteilswert der Tigerpartei liegt bei 0,4. Der Standardfehler des Anteilswerts wird mit folgender Formel geschätzt:

$$\hat{\sigma} = \sqrt{\frac{p * (1 - p)}{n}} = \sqrt{\frac{0,4 * (1 - 0,4)}{1000}} = \sqrt{\frac{0,4 * 0,6}{1000}} = \sqrt{0,00024} = 0,015$$

Der Standardfehler beträgt 0,015. Der Anteilswert (0,4) und der Standardfehler (0,015) werden in die Formel zur Berechnung des 95-Prozent-Konfidenzintervalls eingetragen:

$$p \pm 1,96 * \hat{\sigma} = 0,4 \pm (1,96 * 0,015) = 0,4 \pm 0,029$$

Die untere Grenze des 95-Prozent-Konfidenzintervalls liegt bei 0,371; die obere Grenze des 95-Prozent-Konfidenzintervalls liegt bei 0,429. Mit einer Sicherheit von 95 Prozent liegt der wahre Anteil der Tigerpartei in der Grundgesamtheit zwischen 37,1 und 42,9 Prozent.

Als zweites Beispiel berechnen wir das 99-Prozent-Konfidenzintervall der Katzenpartei. Der Anteilswert der Katzenpartei liegt in der Stichprobe bei 0,05. Der Standardfehler wird – wie oben – mit folgender Formel geschätzt:

$$\hat{\sigma} = \sqrt{\frac{p * (1 - p)}{n}} = \sqrt{\frac{0,05 * (1 - 0,05)}{1000}} = \sqrt{\frac{0,05 * 0,95}{1000}} = \sqrt{0,0000475} = 0,007$$

Der Standardfehler der Katzenpartei liegt bei 0,007. Der Anteilswert ($p = 0,05$) und der Standardfehler (0,007) werden in die Formel zur Berechnung des 99-Prozent-Konfidenzintervalls eingetragen:

$$p \pm 2,58 * \hat{\sigma} = 0,05 \pm (2,58 * 0,007) = 0,05 \pm 0,018$$

Die untere Grenze des 99-Prozent-Konfidenzintervalls liegt bei 0,032; die obere Grenze des Konfidenzintervalls bei 0,068. Mit einer Sicherheit von 99 Prozent liegt der wahre Anteil der Katzenpartei zwischen 3,2 und 6,8 Prozent.

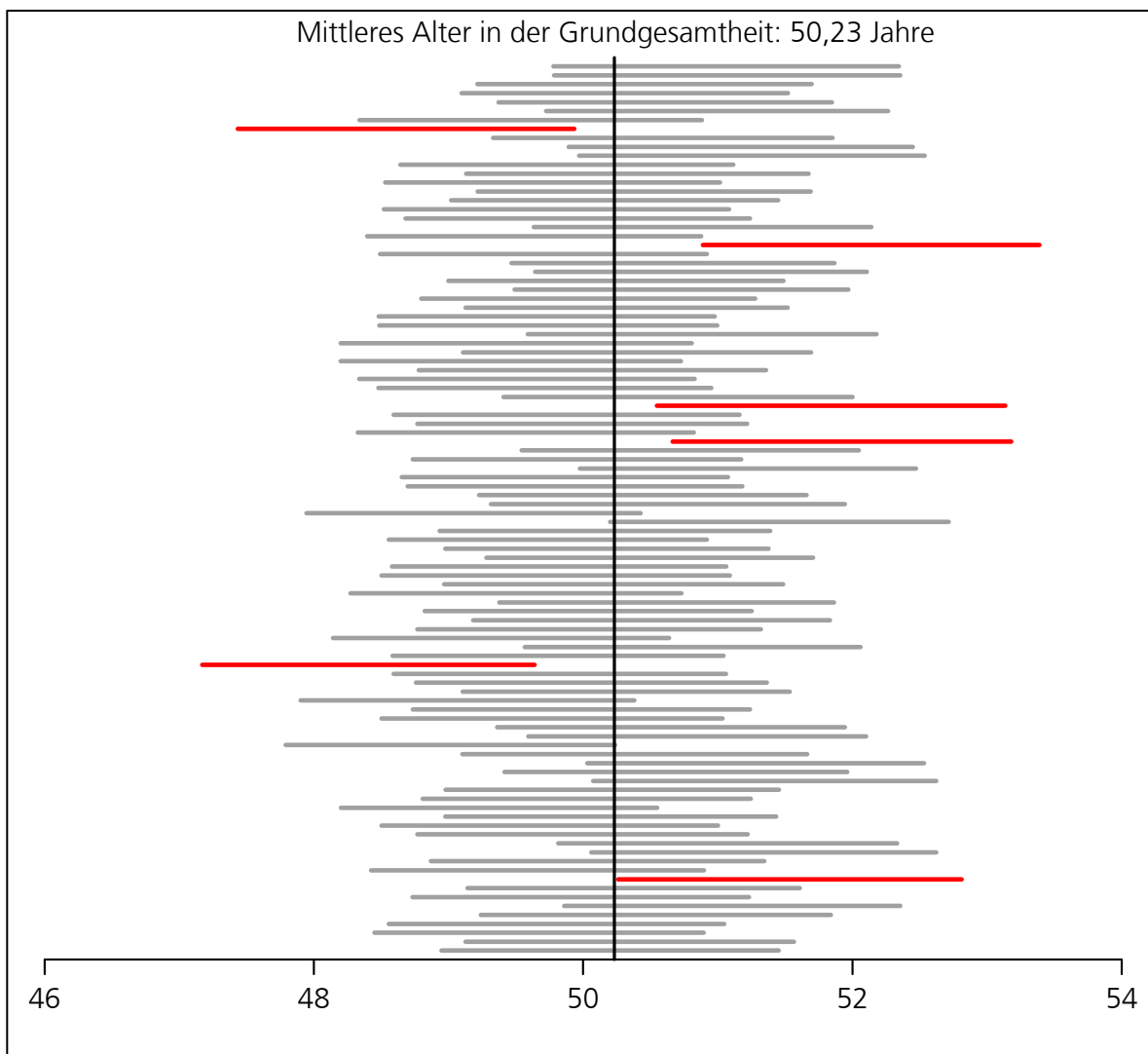
5.3.2.3 Interpretation von Konfidenzintervallen

In der Forschungspraxis werden Konfidenzintervalle gelegentlich falsch interpretiert. Ein typisches Beispiel ist folgender Satz:

„Mit einer Wahrscheinlichkeit von 95 Prozent liegt der tatsächliche Wert der Grundgesamtheit innerhalb des Konfidenzintervalls.“

Diese Interpretation ist formal nicht korrekt. Wir wollen diesen Sachverhalt in Anlehnung an ein Beispiel von Kohler und Kreuter (2017, S. 243-244; ähnlich auch Kühnel und Krebs 2007, S. 242-244) illustrieren. In Abbildung 30 markiert die schwarze senkrechte Linie den tatsächlichen Wert der Grundgesamtheit. Das mittlere Alter der 50000 Personen liegt bei 50,23 Jahren. Für die Abbildung haben wir 100 einfache Zufallsstichproben mit jeweils 1000 Personen aus unserer fiktiven Grundgesamtheit gezogen (siehe Abschnitt 5.1). Für jede dieser Stichproben wurde das 95-Prozent-Konfidenzintervall gebildet. In Abbildung 30 sind diese 100 95-Prozent-Konfidenzintervalle jeweils mit einer horizontalen Linie dargestellt. Die grauen Linien enthalten den tatsächlichen Wert der Grundgesamtheit, die roten Linien enthalten den wahren Wert der Grundgesamtheit nicht.

Abbildung 30: 95-Prozent-Konfidenzintervalle (Stichprobengröße jeweils 1000 Personen)



Quelle: Eigene Darstellung

Von 100 Konfidenzintervallen umfassen 94 Konfidenzintervalle den tatsächlichen Wert der Grundgesamtheit. Dies ist ein Anteil von 94 Prozent. Sechs Konfidenzintervalle (6 Prozent) enthalten den tatsächlichen Wert nicht. Werden statt 100 Konfidenzintervallen jetzt unendlich viele Konfidenzintervalle berechnet und in die Abbildung eingetragen, dann werden 95 Prozent dieser (unendlich vielen) Konfidenzintervalle den tatsächlichen Wert der Grundgesamtheit abdecken. Fünf Prozent (der unendlich vielen) Konfidenzintervalle werden den wahren Wert nicht berücksichtigen.

In der Forschungspraxis ziehen wir allerdings nicht unendlich viele Zufallsstichproben. Wir ziehen auch nicht 100 Zufallsstichproben. Wir ziehen lediglich eine (!) Zufallsstichprobe und berechnen das entsprechende Konfidenzintervall. Dieses *eine* Konfidenzintervall beinhaltet den tatsächlichen Wert der Grundgesamtheit oder es beinhaltet den wahren Wert der Grundgesamtheit nicht. Diese Frage können wir nicht beantworten, da wir den tatsächlichen Wert nicht kennen.

Auf den ersten Blick ist diese Feststellung nicht sonderlich zufriedenstellend, aber wir wissen, dass von unendlich vielen Konfidenzintervallen immerhin 95 Prozent den tatsächlichen Wert enthalten.

Diese Unsicherheit können (und sollten) wir auch verbal ausdrücken. Anstelle von „Wahrscheinlichkeit“ ist es bei der Beschreibung von Konfidenzintervallen formal korrekt, von Sicherheit zu sprechen. Folgende Aussage ist formal korrekt:

„Mit einer Sicherheit von 95 Prozent beinhaltet das Konfidenzintervall den tatsächlichen Wert der Grundgesamtheit.“

Alternativ bietet sich auch folgende Formulierung an:

„Bei einer Irrtumswahrscheinlichkeit von 5 Prozent beinhaltet das Konfidenzintervall den tatsächlichen Wert der Grundgesamtheit.“

I Dies klingt an dieser Stelle zwar sehr spitzfindig, aber mit Blick auf Abbildung 30 lässt sich die weit verbreitete Interpretation, dass es eine 95-prozentige Wahrscheinlichkeit gibt, dass der tatsächliche Wert innerhalb des Konfidenzintervalls liegt, als nicht korrekt entlarven. Ein bestimmtes Konfidenzintervall kann den tatsächlichen Wert enthalten oder nicht. Berechnen wir jedoch unendlich viele Konfidenzintervalle, dann werden 95 Prozent dieser Konfidenzintervalle den tatsächlichen Wert enthalten. Die Wahrscheinlichkeitsaussage bezieht sich nicht auf ein bestimmtes Konfidenzintervall, sondern auf (unendlich) viele Konfidenzintervalle. Da wir es meist aber nur mit einem Konfidenzintervall zu tun haben, können (und sollten) wir diese Wahrscheinlichkeitsaussage nicht auf das konkrete Konfidenzintervall anwenden.

5.3.3 Berechnung der benötigten Fallzahl

Die Logik von Konfidenzintervallen kann auch genutzt werden, um vor der Durchführung einer Erhebung den erforderlichen Stichprobenumfang zu ermitteln (Borg 2003, S. 186-189; Kühnel und Krebs 2007, S. 249-250; Gehring und Weins 2009, S. 268-270; Häder und Häder 2014, S. 286-289; Dillman et al. 2014, S. 77-83).

Für die Berechnung der erforderlichen Fallzahl müssen das Konfidenzniveau, der vermutete Anteilswert des Merkmals in der Stichprobe (p) und der akzeptierte Stichprobenfehler festgelegt werden. Bei einem 95-Prozent-Konfidenzniveau wird der Wert 1,96 eingetragen, bei einem 99-Prozent-Konfidenzniveau der Wert 2,58. Der vermutete Anteilswert eines Merkmals in der Grundgesamtheit (p) liegt zwischen 0 und 1. Als Stichprobenfehler werden häufig \pm fünf Prozentpunkte oder \pm drei Prozentpunkte akzeptiert. Bei großen Grundgesamtheiten (z.B. wahlberechtigte Bevölkerung) kann folgende Formel zur Berechnung der benötigten Fallzahl verwendet werden:

$$n = \left(\frac{\text{Konfidenzniveau} * \sqrt{p*(1-p)}}{\text{Stichprobenfehler}} \right)^2$$

Bei kleinen Grundgesamtheiten wird noch ein Korrekturfaktor (grau hinterlegt) hinzugefügt (Borg 2003, S. 183):

$$n = \left(\frac{\text{Konfidenzniveau} * \sqrt{p*(1-p)}}{\text{Stichprobenfehler}} * \sqrt{1 - \frac{n}{N}} \right)^2$$

Wir wollen die Berechnung der erforderlichen Stichprobengröße an einem Beispiel erläutern. Bei einem 95-Prozent-Konfidenzniveau (1,96) wird ein Stichprobenfehler von ± 3 Prozentpunkten (0,03) festgelegt. Als Anteil von p in der Grundgesamtheit vermuten wir 0,5. Dabei handelt es sich um den konservativsten Wert (ein kleineres p verringert die erforderliche Stichprobengröße). Unsere Grundgesamtheit ist die wahlberechtigte Bevölkerung in Deutschland, so dass wir auf den Korrekturfaktor verzichten können. Eingetragen ist die Formel:



$$n = \left(\frac{\text{Konfidenzniveau} * \sqrt{p*(1-p)}}{\text{Stichprobenfehler}} \right)^2 = \left(\frac{1,96 * \sqrt{0,5 * (1-0,5)}}{0,03} \right)^2 = \left(\frac{0,98}{0,03} \right)^2 = 1067,11$$

Die erforderliche Stichprobengröße liegt bei 1067 Personen. Wenn 1067 Personen aus der Grundgesamtheit (zufällig) befragt werden, dann kann der tatsächliche Anteil des Merkmals (z.B. Anteil der Wähler einer Partei) mit einer Irrtumswahrscheinlichkeit von 5 Prozent und einem Stichprobenfehler von ± 3 Prozentpunkten bestimmt werden.

In Tabelle 57 sind die erforderlichen Stichprobengrößen bei einer einfachen Zufallsstichprobe in Abhängigkeit der Größe der Grundgesamtheit, des Stichprobenfehlers sowie des vermuteten Anteilswerts ($p = 0,2$ bzw. $p = 0,5$) dokumentiert. Bei einer Grundgesamtheit von 1000 Millionen Personen ist eine Zufallsstichprobe von rund 1067 Personen ausreichend, um den Anteil eines Merkmals mit einer Irrtumswahrscheinlichkeit von 5 Prozent und einem Stichprobenfehler von ± 3 Prozentpunkten bestimmen zu können.

Tabelle 57: Erforderliche Stichprobengröße


Grundgesamtheit	Stichprobengröße für das 95-Prozent-Konfidenzintervall			
	Stichprobenfehler: ± 5 Prozent		Stichprobenfehler: ± 3 Prozent	
	$p = 0,5$	$p = 0,2$	$p = 0,5$	$p = 0,2$
1000	278	198	517	406
2000	322	219	696	509
4000	351	232	843	584
8000	367	239	942	629
10000	370	240	965	640
40000	381	244	1040	672
100000	384	246	1056	679
1000000	384	246	1066	683
1000000000	384	246	1067	683

Quelle: In Anlehnung an Dillman et al. (2014, S. 80). Die Tabelle zeigt die erforderliche Stichprobengröße bei einer einfachen Zufallsstichprobe und einem 95-Prozent-Konfidenzintervall in Abhängigkeit des Stichprobenfehlers (± 3 bzw. ± 5 Prozent) sowie des Anteils des Merkmals in der Grundgesamtheit ($p = 0,5$ bzw. $p = 0,2$).

In der Forschungspraxis werden sich die meisten Stichprobengrößen aber nicht am (absoluten) Minimum orientieren, da hierbei nur der Stichprobenfehler berücksichtigt wird. Neben dem Stichprobenfehler muss häufig auch dem Stichprobenverfahren sowie möglichen Ausfällen (Non-response) Rechnung getragen werden. Diese Aspekte führen uns zum nächsten Abschnitt.

5.3.4 Anwendungsprobleme in der Praxis

Die Grundlagen der Inferenzstatistik – zentraler Grenzwertsatz und Standardfehler sowie Punkt- und Intervallschätzung – gehören sicherlich nicht zu den anschaulichsten Inhalten der empirischen Sozialforschung. In der Forschungspraxis stellt die Berechnung von Standardfehler und Konfidenzintervallen meist keine größere Herausforderung dar. Im Gegenteil: Die meisten Statistikprogramme geben diese Informationen bei der Datenanalyse einfach mit aus. Häufig werden dann erforderliche Voraussetzungen für Standardfehler und Konfidenzintervalle nicht mehr hinterfragt und es wird den vom Statistikprogramm berechneten Werten „blind“ vertraut.

 Deshalb möchten wir abschließend auf drei zentrale Voraussetzungen für die korrekte Berechnung des Standardfehlers und damit auch für die Berechnung eines Konfidenzintervalls hinweisen:

- Einfache Zufallsstichprobe
- Overcoverage und Undercoverage
- Nonresponse

Einfache Zufallsstichprobe

In Abschnitt 5.1 haben wir auf Grundlage einer (fiktiven) Grundgesamtheit viele (einfache) Zufallsstichproben gezogen. Bei einer einfachen Zufallsstichprobe hat jedes Element (z.B. jede Person) der Grundgesamtheit die gleiche Chance, in diese Stichprobe zu gelangen. Auf dieser Grundlage haben wir Standardfehler und Konfidenzintervalle berechnet. Dieses Vorgehen und auch die Berechnungen sind völlig korrekt. Allerdings sind wir in der Forschungspraxis nur selten mit einfachen Zufallsstichproben konfrontiert. Meist handelt es sich um mehrstufige Zufallsverfahren. Beim ADM-Design lassen sich drei Auswahlsschritte unterscheiden. Im ersten Schritt werden (zufällig) die Sample Points ausgewählt, im zweiten Schritt werden per Random Route zufällig die Haushalte festgelegt und im dritten Schritt wird die konkrete Zielperson per Schwedenschlüssel oder Last-Birthday-Methode zufällig ausgewählt (Tausendpfund 2018b, S. 225-229). Aus praktischen Gründen müssen für die Stichprobenziehung mehrere Zufallsverfahren kombiniert werden.

Kohler und Kreuter (2017, S. 225-233) bezeichnen Zufallsstichproben, die nicht auf der einfachen Zufallsauswahl beruhen, als komplexe Stichproben. Bei komplexen Stichproben ist der Standardfehler aus verschiedenen Gründen größer als bei einfachen Stichproben. Deshalb werden auch die Konfidenzintervalle breiter. Bei den meisten Statistikprogrammen (z.B. SPSS) basiert die Berechnung von Standardfehler und Konfidenzintervallen allerdings auf der Annahme einer einfachen Zufallsstichprobe. Die Konsequenz: Die Statistikprogramme unterschätzen in der „Voreinstellung“ den Standardfehler und die Konfidenzintervalle sind zu eng. Mittlerweile gibt es eine Reihe von Techniken, um auch bei mehrstufigen Zufallsverfahren korrekte Standardfehler und angemessene Konfidenzintervalle berechnen zu können (z.B. Bootstrapping, Jackknife, Balanced

Repeated Replication). Die Darstellung dieser Verfahren, die meist auch in Statistikprogrammen implementiert sind, sprengt allerdings den Rahmen dieser Einführung. Deshalb sei an dieser Stelle auf Valliant et al. (2013), Heeringa et al. (2017) sowie Valliant und Dever (2018) verwiesen.

Während es bei komplexen Stichproben Möglichkeiten gibt, korrekte Standardfehler und Konfidenzintervalle zu schätzen, gilt dies nicht für willkürliche Stichproben (Convenience Sample). Mit anderen Worten: bei willkürlichen Stichproben können keine Standardfehler und Konfidenzintervalle berechnet werden. Das Statistikprogramm kann auch nicht prüfen, ob es sich bei den verwendeten Daten um eine (einfache) Zufallsstichprobe oder eine willkürliche Stichprobe handelt. Diese Aufgabe fällt in den Verantwortungsbereich des Anwenders.

Die Grundgesamtheit umfasst alle Elemente, über die Aussagen beabsichtigt sind. Die Auswahlgesamtheit beinhaltet alle Elemente, die eine Chance haben, in die Stichprobe zu gelangen. Im Idealfall ist die Auswahlgesamtheit mit der Grundgesamtheit deckungsgleich, in der Forschungspraxis ist aber ein sogenannter Abdeckungsfehler (coverage error) unvermeidlich. Dabei lassen sich allgemein Overcoverage und Undercoverage unterscheiden (siehe Kasten in Abschnitt 5.1). Der coverage error betrifft praktisch alle gängigen Befragungsvarianten – persönlich, telefonisch, schriftlich, online –, da es zumindest für eine Zufallsstichprobe der allgemeinen Bevölkerung in der Regel keinen vollständigen Auswahlrahmen für alle Elemente der Grundgesamtheit gibt. Bei einer telefonischen Befragung bleiben beispielsweise Personen ohne Telefon unberücksichtigt, bei einer Online-Erhebung werden Personen ohne Internetzugang systematisch ausgeschlossen (Fuchs 2010, S. 230; Harrison 2005). Bei der Berechnung von Standardfehler und Konfidenzintervallen werden mögliche Abdeckungsfehler nicht berücksichtigt. Abdeckungsfehler führen zu größeren Standardfehlern, zu unsicheren Schätzungen und gegebenenfalls auch zu verzerrten Ergebnissen.

Overcoverage und Undercoverage

Bei unserer (einfachen) Zufallsstichprobe haben wir 1000 (fiktive) Personen „gezogen“ und jeweils ihr Alter notiert. Alle Personen haben sich an der Erhebung beteiligt. Wir hatten keine Probleme, die „ausgewählten“ Personen zu finden und für die Teilnahme zu motivieren. In der Forschungspraxis werden sich in der Regel nicht alle (zufällig) ausgewählten Personen an der Erhebung beteiligen. Wir können dabei zwischen Unit- und Item-Nonresponse unterscheiden. Bei Unit-Nonresponse handelt es sich um einen vollständigen Ausfall. Es liegen keine Informationen für diese Person vor. Bei Item-Nonresponse hat sich die Person an der Befragung beteiligt, aber einzelne Fragen (z.B. zum Einkommen) nicht beantwortet.

Nonresponse

Die Konsequenzen von Unit-Nonresponse und Item-Nonresponse sind vergleichbar: die Standardfehler werden größer und die Konfidenzintervalle breiter. Selektive Ausfälle (z.B. insbesondere Personen mit hohem Einkommen verweigern die Teilnahme) können zudem zu verzerrten Schätzergebnissen führen. Deshalb existieren mittlerweile zahlreiche Verfahren, die versuchen, die Konsequenzen von Unit-Nonresponse und Item-Nonresponse zumindest zu verringern. Bei Unit-Nonresponse wird häufig auf Gewichtungsverfahren zurückgegriffen, bei Item-Nonresponse auf Verfahren der (multiplen) Imputation. Diese Techniken gehören einerseits zu den fortgeschrittenen Techniken der Datenanalyse, andererseits sind sie in der Literatur auch nicht völlig unumstritten. Deshalb soll dieses Problem (und die damit verbundenen vorgeschlagenen Lösungen) nicht weiter vertieft werden. Wichtig ist, dass Unit- und Item-Nonresponse in der Regel zu größeren Standardfehlern führen und damit die Unsicherheit der Schätzung erhöhen.

5.4 Statistisches Testen

Die Prüfung formulierter Hypothesen spielt in der quantitativen Sozialforschung eine herausragende Rolle (Ludwig-Mayerhofer et al. 2014, S. 136) und Signifikanztests gehören vermutlich zu den am weitesten verbreiteten statistischen Verfahren in den Sozialwissenschaften (Dubben und Beck-Bornholt 2006; Sedlmeier und Renkewitz 2013, S. 358). Während einerseits in praktisch allen quantitativen Studien die Ergebnisse statistischer Signifikanztests berichtet werden, spricht andererseits Walter Krämer (2006, S. 58) auch von einem „Signifikanztest-Ritual“.¹⁶

Was sind Hypothesen?

Bei sozialwissenschaftlichen Hypothesen handelt es sich um (begründete) Zusammenhänge zwischen zwei Merkmalen, die sich auf reale Sachverhalte beziehen, in Form eines Konditionalsatzes formuliert sind, über den Einzelfall hinaus reichen und falsifizierbar sind (Tausendpfund 2018b, S. 91-98). Typische Beispiele für sozialwissenschaftliche Hypothesen sind „Je höher die Bildung einer Person ist, desto größer ist ihre Lebenszufriedenheit“ oder „Je stärker das politische Interesse einer Person ist, desto wahrscheinlicher ist ihre Wahlbeteiligung“. Solche Hypothesen beziehen sich auf eine Grundgesamtheit.

Die Grundidee statistischer Testverfahren basiert auf Überlegungen von Ronald Fisher (1925b), die von Jerzy Neyman und Egon S. Pearson (1933) weiterentwickelt wurden (für eine Darstellung der frühen Ansätze siehe Sedlmeier und Renkewitz 2013, S. 361-386). Worum geht es – vereinfacht gesprochen – beim statistischen Testen?

Während sich eine Hypothese auf eine Grundgesamtheit bezieht, erfolgt die Überprüfung einer Hypothese auf Grundlage einer Stichprobe dieser Grundgesamtheit. Wir nutzen beispielsweise die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) oder den European Social Survey (ESS), um unsere Hypothesen zu testen. Solche Zufallsstichproben haben allerdings einen unvermeidbaren Stichprobenfehler. Wenn wir in einer Stichprobe einen Zusammenhang zwischen zwei Merkmalen finden (z.B. zwischen Bildung und Einkommen) oder sich zwei Gruppen unterscheiden (z.B. Lebenszufriedenheit zwischen Männern und Frauen), stellt sich die Frage, ob der in der Stichprobe nachgewiesene Zusammenhang (oder Unterschied) jetzt zufallsbedingt (also ein Resultat des Stichprobenfehlers) ist oder auch in der Grundgesamtheit existiert. Diese Frage wird auf Grundlage statistischer Testverfahren beantwortet.

Ein Signifikanztest prüft, wie wahrscheinlich der Befund unserer Stichprobe ist, wenn in der entsprechenden Grundgesamtheit *kein* Zusammenhang zwischen zwei Merkmalen besteht bzw. sich zwei Gruppen *nicht* unterscheiden. Ein empirisches Ergebnis wird als signifikant bezeichnet, wenn

¹⁶ Trotz der weiten Verbreitung von Signifikanztests sind diese in den Sozialwissenschaften durchaus umstritten (Morrison und Henkel 1970; Cohen 1994; Gill 1999; siehe auch Ludwig-Mayerhofer et al. 2014, S. 136). In einer jüngeren Veröffentlichung hat auch die *American Statistical Association* (Wasserstein und Lazar 2016) auf häufige Fehlinterpretationen von Signifikanztests hingewiesen. Seit 2018 verzichtet *Political Analysis* auf die Ausweisung von p-Werten, da sie häufig falsch interpretiert werden: „Furthermore, there is evidence that a large number of social scientists misunderstand p-values in general and consider them a key form of scientific reasoning“ (Gill 2018, S. 2).

wir uns sehr sicher sein können, dass der Zusammenhang von zwei Merkmalen bzw. der Unterschied zwischen zwei Gruppen in einer Stichprobe nicht zufallsbedingt ist. Mit anderen Worten: Der Zusammenhang bzw. Unterschied existiert (vermutlich) nicht nur in der Stichprobe, sondern auch in der Grundgesamtheit.

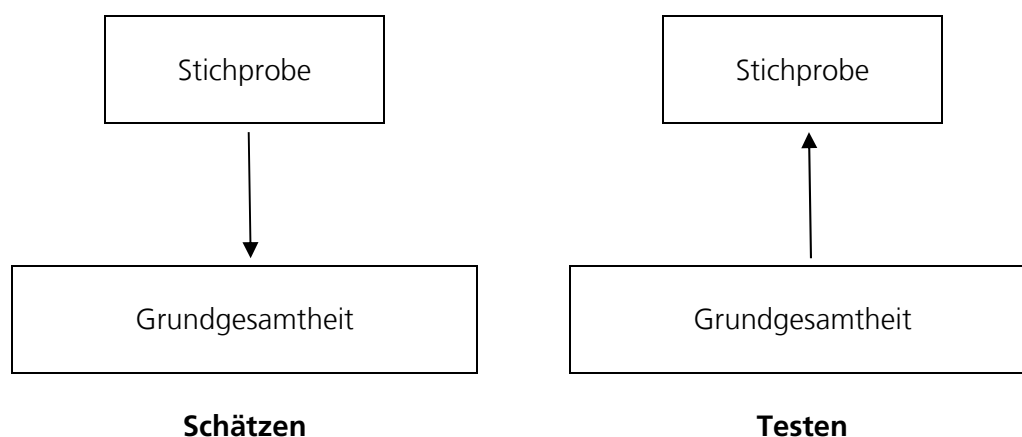
Wie sollten Signifikanztests *nicht* interpretiert werden

Signifikanztests werden häufig völlig falsch interpretiert. Ein Signifikanztest trifft keine Aussage über die Stärke eines Zusammenhangs oder die Bedeutung eines Befunds (Kuckartz et al. 2013, S. 153; Diaz-Bone 2018, S. 183). Auf Grundlage eines Signifikanztests entscheiden wir, ob ein Befund einer Stichprobe (z.B. Lebenszufriedenheit zwischen Männern und Frauen unterscheidet sich) auf die entsprechende Grundgesamtheit übertragen werden darf. Die Vorgehensweise ist dabei konservativ. Nur, wenn wir uns sehr sicher sind, dass der Befund nicht zufallsbedingt ist, handelt es sich um einen sogenannten signifikanten Befund.

Beim Schätzen und Testen gibt es einige Gemeinsamkeiten, aber auch wichtige Unterschiede (Ludwig-Mayerhofer et al. 2014, S. 137; Diaz-Bone 2018, S. 118-119). Beide Verfahren basieren jeweils auf einer Zufallsstichprobe. Wie Abbildung 31 illustriert, wird beim Schätzen eine Zufallsstichprobe genutzt, um von einer Maßzahl in der Stichprobe auf einen Parameter in der Grundgesamtheit zu schließen (z.B. vom mittleren Alter in der Stichprobe auf das mittlere Alter in der Grundgesamtheit). Beim Testen bezieht sich die zu überprüfende Hypothese auf die Grundgesamtheit. Wir nutzen die Stichprobe, um zu prüfen, ob die Hypothese korrekt ist oder nicht.

Schätzen und Testen im Vergleich

Abbildung 31: Schätzen und Testen im Vergleich



Quelle: Eigene Darstellung

Wir wollen die unterschiedlichen Logiken an zwei Beispielen illustrieren. Wir berechnen beim Schätzen für eine Zufallsstichprobe das mittlere Alter der Personen (z.B. 49,98 Jahre). Das mittlere Alter dieser Stichprobe ist unsere Schätzung für das mittlere Alter der entsprechenden Grundgesamtheit. Bei Kenntnis des Standardfehlers können wir einen Bereich angeben, in dem mit einer Sicherheit von 95 Prozent (alternativ: 99 Prozent) das mittlere Alter der Grundgesamtheit liegt.

Beim statistischen Testen ist eine Hypothese über die Grundgesamtheit der Ausgangspunkt, etwa: die Lebenszufriedenheit unterscheidet sich zwischen Männern und Frauen. Dieser Forschungshypothese wird die sogenannte Nullhypothese gegenübergestellt, die den Zusammenhang verneint (z.B. Die Lebenszufriedenheit unterscheidet sich nicht zwischen Männern und Frauen). Bei einem statistischen Test wird geprüft, ob die Nullhypothese mit den Daten der vorliegenden Stichprobe vereinbar ist. Falls nicht, wird die Nullhypothese verworfen und die Alternativhypothese (z.B. Die Lebenszufriedenheit unterscheidet sich zwischen Männern und Frauen) akzeptiert.

Vielfältige Testverfahren

In der Forschungspraxis gibt es nicht einen Signifikanztest, sondern sehr unterschiedliche Testverfahren (siehe für einen Überblick z.B. Kanji 2006; Bortz und Lienert 2008). Der bekannteste Signifikanztest ist der t-Test, der genutzt wird, um Mittelwerte zu vergleichen. Weitere wichtige Testverfahren sind der Chi-Quadrat-Test und der F-Test. Die allgemeine Vorgehensweise folgt bei allen Signifikanztests der gleichen Logik. Deshalb wird im nächsten Abschnitt zunächst diese allgemeine Vorgehensweise vorgestellt, ehe ausgewählte Testverfahren dargestellt werden.

5.4.1 Allgemeine Vorgehensweise bei einem Signifikanztest

Mit Ludwig-Mayerhofer et al. (2014, S. 139-141; ähnlich auch Kühnel und Krebs 2007, S. 264-272; Tachtsoglou und König 2017, S. 318; Diaz-Bone 2018, S. 182) lässt sich das allgemeine Vorgehen bei einem Signifikanztest in vier Schritten beschreiben. Erstens muss eine Forschungshypothese formuliert werden, die der Nullhypothese gegenübergestellt wird. Zweitens wird die geeignete Teststatistik ausgewählt. Drittens wird das Signifikanzniveau festgelegt. Viertens wird die Teststatistik der vorliegenden Stichprobe berechnet. Auf Grundlage der Teststatistik der vorliegenden Stichprobe wird eine Entscheidung für oder gegen die Nullhypothese getroffen. Vorausgesetzt wird, dass es sich bei der Stichprobe um eine Zufallsstichprobe handelt.

Hypothesen formulieren

Im ersten Schritt wird eine Forschungshypothese bzw. Alternativhypothese formuliert. Eine Forschungshypothese bzw. Alternativhypothese beinhaltet die eigentlich interessierende Aussage (z.B. Die Lebenszufriedenheit unterscheidet sich zwischen Männern und Frauen). Dieser Forschungshypothese wird die Nullhypothese gegenübergestellt, die den in der Forschungshypothese formulierten Zusammenhang verneint (z.B. Die Lebenszufriedenheit unterscheidet sich nicht zwischen Männern und Frauen). Die Forschungshypothese und die Nullhypothese bilden dabei ein Hypothesenpaar. Die Forschungshypothese wird meist mit H_1 abgekürzt und die Nullhypothese mit H_0 . In der Forschungspraxis wird in der Regel nur die Forschungshypothese explizit formuliert, die dann gegen die entsprechende Nullhypothese getestet wird.

Forschungshypothesen können ungerichtet oder gerichtet sein. Ungerichtete Hypothesen treffen keine Aussagen über die Richtung eines Unterschieds. Die Hypothesen „Die Lebenszufriedenheit unterscheidet sich zwischen Männern und Frauen“ und „Es besteht ein Zusammenhang zwischen Bildung und Einkommen“ sind Beispiele für ungerichtete Hypothesen. Es wird postuliert, dass sich zum Beispiel die Lebenszufriedenheit zwischen Männern und Frauen unterscheidet. Eine andere Bezeichnung für ungerichtete Hypothesen sind zweiseitige Hypothesen, weil der Unterschied bzw. der Zusammenhang in beide Richtungen gehen kann (z.B. Männer können eine höhere Lebenszufriedenheit haben oder Frauen können eine höhere Lebenszufriedenheit haben).

Bei einer gerichteten Hypothese wird eine Aussage über die Richtung eines vermuteten Unterschieds bzw. Zusammenhangs gemacht. „Frauen haben eine höhere Lebenszufriedenheit als Männer“ sowie „Je höher die Bildung einer Person ist, desto größer ist ihr Einkommen“ sind Beispiele für gerichtete Hypothesen. Eine gerichtete Hypothese wird auch als einseitige Hypothese bezeichnet.

Wie oben ausgeführt, lassen sich verschiedene Testverfahren unterscheiden. Der bekannteste Test, der auch in diesem Kurs im Mittelpunkt steht, ist der t-Test. Mit dem t-Test wird der Unterschied zwischen zwei arithmetischen Mittelwerten untersucht. Vereinfacht formuliert wird geprüft, ob der Unterschied zwischen zwei arithmetischen Mittelwerten in einer Stichprobe auf zufallsbedingte Abweichungen zurückzuführen ist oder ob sich diese beiden Werte (wahrscheinlich) auch in der entsprechenden Grundgesamtheit unterscheiden.

Teststatistik auswählen

Beim Chi-Quadrat-Test wird geprüft, ob die empirisch beobachteten Häufigkeiten von den bei statistischer Unabhängigkeit erwarteten Häufigkeiten abweichen (Ludwig-Mayerhofer et al. 2014, S. 212-216; Tachtsoglou und König 2017, S. 337-340). Die Berechnung des Zusammenhangsmaßes Cramer's V basiert auf Chi-Quadrat (siehe Kapitel 3.2). In der Regressionsanalyse wird häufig ein F-Test genutzt, um „die Abhängigkeit der Gesamtschätzung eines Regressionsmodells von zufälligen Verzerrungen zu überprüfen“ (Urban und Mayerl 2018, S. 144). Ein F-Test bezieht sich nicht auf einen einzelnen Regressionskoeffizienten, sondern auf das Regressionsmodell als Ganzes. Es wird geprüft, ob das Regressionsmodell zur Varianzaufklärung beitragen kann oder auf Zufallsfehler zurückzuführen ist (Kohler und Kreuter 2017, S. 281). In der Forschungspraxis spielt die Signifikanz der einzelnen Regressionskoeffizienten, die mit einem t-Test geprüft wird, allerdings eine wichtigere Rolle als die Prüfung des Gesamtmodells.

In unserer (fiktiven) Stichprobe unterscheidet sich die Lebenszufriedenheit zwischen Männern und Frauen. Die mittlere Lebenszufriedenheit der Frauen liegt bei 7,8 Punkten und die mittlere Lebenszufriedenheit der Männer bei 7,5 Punkten (die Lebenszufriedenheit wurde dabei auf einer Skala von 0 bis 10 erfasst). Eine Stichprobe kann die Grundgesamtheit allerdings nicht exakt abbilden. Deshalb könnte der Stichprobenbefund auch auf zufällige Abweichungen der beiden Stichprobenmittelwerte – also der mittleren Lebenszufriedenheit der Frauen und der Männer in der Stichprobe – von den Mittelwerten in der Grundgesamtheit zurückgehen. Die entscheidende Frage ist: Wie wahrscheinlich ist der Stichprobenbefund, wenn sich in der Grundgesamtheit die mittlere Lebenszufriedenheit zwischen Männern und Frauen nicht unterscheidet?

Signifikanzniveau festlegen

Die Festlegung dieser Wahrscheinlichkeit wird in der Sprache der empirischen Sozialforschung als Signifikanzniveau bezeichnet. In den Sozialwissenschaften ist das 5-Prozent-Signifikanzniveau üblich, gelegentlich wird auch das 1-Prozent- oder das 0,1-Prozent-Signifikanzniveau gewählt. Bei kleinen Stichproben arbeiten Forscherinnen auch mit dem 10-Prozent-Signifikanzniveau. Das 5-Prozent-Signifikanzniveau gilt allerdings meist als ungeschriebene Regel (siehe Cowles und Davis 1982 für den Ursprung dieser Konvention).

Mit dem Signifikanzniveau wird ein kritischer Schwellenwert festgelegt. Die Ausgangsfrage lautet: Wie wahrscheinlich ist der Stichprobenbefund (also der gefundene Unterschied bei der Lebenszu-

friedenheit zwischen Männern und Frauen), wenn in der Grundgesamtheit kein Unterschied existiert? Wenn diese Wahrscheinlichkeit kleiner als 5 Prozent ist (5-Prozent-Signifikanzniveau), dann wird die Nullhypothese abgelehnt und die Forschungs- bzw. Alternativhypothese (vorläufig) akzeptiert.

! Mit anderen Worten: Bei einem statistischen Test wird nicht die eigentlich interessierende Forschungshypothese getestet, sondern die Nullhypothese. Wenn die Wahrscheinlichkeit für die Bestätigung der Nullhypothese kleiner als 5 Prozent ist, dann wird die Nullhypothese verworfen. Die Vorgehensweise ist konservativ. Nur, wenn wir uns sehr sicher sind, dass die Nullhypothese nicht zutrifft, wird sie verworfen und die Forschungs- bzw. Alternativhypothese vorläufig akzeptiert.

Teststatistik berechnen

Schließlich wird die Teststatistik der vorliegenden Stichprobe ermittelt. Dabei wird eine empirische Prüfgröße (z.B. t-Wert) berechnet, die mit dem kritischen Wert der jeweiligen Teststatistik abgeglichen wird. Der kritische Wert der jeweiligen Teststatistik ist in Tabellen dokumentiert und kann direkt abgelesen werden (siehe exemplarisch Tabelle 60). Wenn der Betrag der berechneten Prüfgröße größer ist als der kritische Wert der Teststatistik, dann wird die Nullhypothese verworfen und die Alternativhypothese vorläufig akzeptiert.

In der Forschungspraxis berechnen Statistikprogramme wie SPSS die Signifikanztests. Neben der empirischen Prüfgröße (z.B. t-Wert) geben Statistikprogramme auch den p-Wert an. Der p-Wert gibt die Wahrscheinlichkeit für den empirischen Befund (oder für ein noch extremeres Ergebnis) an, wenn in der Grundgesamtheit die Nullhypothese gilt. Dabei werden allgemein drei zentrale p-Werte unterschieden:

- $p < 0,05$: Der p-Wert liegt unter 0,05. Die Wahrscheinlichkeit für den empirischen Befund, wenn die Nullhypothese tatsächlich gültig wäre, ist kleiner als 5 Prozent.
- $p < 0,01$: Der p-Wert liegt unter 0,01. Die Wahrscheinlichkeit für den empirischen Befund, wenn die Nullhypothese tatsächlich gültig wäre, ist kleiner als 1 Prozent.
- $p < 0,001$: Der p-Wert liegt unter 0,001. Die Wahrscheinlichkeit für den empirischen Befund, wenn die Nullhypothese tatsächlich gültig wäre, ist kleiner als 0,1 Prozent.

In der Forschungspraxis werden Befunde mit einem p-Wert, der kleiner als 0,05 ($p < 0,05$) ist, als signifikant bezeichnet. Die Nullhypothese wird abgelehnt und die Forschungshypothese (vorläufig) akzeptiert. Der in der Stichprobe gefundene Unterschied (oder Zusammenhang) existiert wahrscheinlich nicht nur in der Stichprobe, sondern auch in der entsprechenden Grundgesamtheit.

In Publikationen werden signifikante Befunde häufig mit Sternchen (Asterisk) gekennzeichnet. Die Sternchen informieren dabei über den jeweiligen p-Wert:

- *: Der p-Wert liegt unter 0,05, ist aber größer als oder gleich 0,01. Das Ergebnis ist auf dem 5-Prozent-Niveau signifikant.
- **: Der p-Wert liegt unter 0,01, ist aber größer als oder gleich 0,001. Das Ergebnis ist auf dem 1-Prozent-Niveau signifikant.
- ***: Der p-Wert liegt unter 0,001. Das Ergebnis ist auf dem 0,1-Prozent-Niveau signifikant.

Bei kleinen Stichproben wird gelegentlich auch das 10-Prozent-Signifikanzniveau ausgewiesen ($p < 0,10$). Welche Bedeutung die Sternchen im Einzelfall haben, muss stets aus der jeweiligen Ergebnistabelle hervorgehen. In der Tabellenlegende sollte sich beispielsweise folgender Hinweis finden: „Signifikanzniveaus: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$.“

Grundsätzlich wird die Nullhypothese also erst verworfen, wenn sich das empirische Ergebnis nur schlecht mit dieser vereinbaren lässt. Ein Signifikanzniveau von fünf Prozent bedeutet aber auch, dass wir bei rund fünf Prozent aller Stichproben empirische Ergebnisse erhalten, die unwahrscheinlich sind und der Nullhypothese widersprechen, aber grundsätzlich möglich sind. Mit anderen Worten: Das Ergebnis ist zwar unwahrscheinlich, aber dennoch könnte die Nullhypothese korrekt sein. Ein statistischer Test kann folglich keine korrekte Testentscheidung garantieren. In der Literatur werden meist zwei Fehlerarten unterschieden, die als Alpha-Fehler und Beta-Fehler bezeichnet und im nächsten Abschnitt behandelt werden.

5.4.2 Alpha- und Beta-Fehler

Bei der Forschungs- bzw. Alternativhypothese und der Nullhypothese handelt es sich um Aussagen über eine Grundgesamtheit. Auf Basis einer Stichprobe wird eine Entscheidung getroffen, ob die Nullhypothese verworfen und die Alternativhypothese (vorläufig) akzeptiert wird, oder ob die Nullhypothese beibehalten und die Alternativhypothese abgelehnt wird. Bei Stichproben sind allerdings Stichprobenfehler unvermeidlich. Aufgrund der durch Stichprobenfehler bedingten Abweichung einer Stichprobe von der entsprechenden Grundgesamtheit kann es beim Hypothesentest zu zwei Fehlentscheidungen kommen, die als Fehler 1. Art (α -Fehler) und Fehler 2. Art (β -Fehler) bezeichnet werden (Gehring und Weins 2009, S. 273-275; Bortz und Schuster 2010, S. 100; Ludwig-Mayerhofer et al. 2014, S. 167-172; Diaz-Bone 2018, S. 172-177). Bei einem statistischen Hypothesentest lassen sich vier Entscheidungen unterscheiden (siehe Tabelle 58).

Tabelle 58: Fehlerarten beim Hypothesentest

		in der Grundgesamtheit gilt die	
		H_0	H_1
Entscheidung aufgrund der Stichprobe	H_0	richtige Entscheidung	β -Fehler (Fehler 2. Art)
	H_1	α -Fehler (Fehler 1. Art)	richtige Entscheidung

Quelle: Eigene Darstellung

Ein statistischer Test führt entweder zur Ablehnung der Nullhypothese und damit zur Annahme der Forschungshypothese oder zur Beibehaltung der Nullhypothese und damit zur Ablehnung der Forschungshypothese. Jede der beiden Testentscheidungen kann richtig oder falsch sein.

Entscheidung 1: In der Grundgesamtheit besteht kein Zusammenhang zwischen zwei Merkmalen. Aufgrund der Teststatistik der Stichprobe entscheiden wir uns für die Nullhypothese und lehnen die Forschungs- bzw. Alternativhypothese ab. Wir haben die richtige Entscheidung getroffen.

Fehler 1. Art oder α -Fehler

Entscheidung 2: In der Grundgesamtheit besteht kein Zusammenhang zwischen zwei Merkmalen. Aufgrund der Teststatistik der Stichprobe entscheiden wir uns allerdings für die Forschungs- bzw. Alternativhypothese, die einen Zusammenhang zwischen den beiden Merkmalen postuliert. Wir lehnen also fälschlicherweise die Nullhypothese ab und akzeptieren die Forschungs- bzw. Alternativhypothese. Ein solcher Fehler wird als Fehler 1. Art oder auch als α -Fehler bezeichnet. Die Wahrscheinlichkeit für einen Fehler 1. Art entspricht dem gewählten Signifikanzniveau α (deshalb auch die Bezeichnung α -Fehler). Bei einem Signifikanzniveau von fünf Prozent werden wir bei fünf Prozent aller Stichproben einen solchen Fehler begehen. Das Ergebnis der Teststatistik spricht zwar sehr wahrscheinlich gegen die Nullhypothese, aber eben nur „sehr wahrscheinlich“. Mit anderen Worten: Bei einem Signifikanzniveau von fünf Prozent werden bei 100 Stichproben etwa fünf Stichproben ein Ergebnis liefern, das gegen die Nullhypothese spricht, auch wenn in der Grundgesamtheit die Nullhypothese gültig ist.

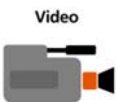
Entscheidung 3: In der Grundgesamtheit gilt die Forschungs- bzw. die Alternativhypothese. Es besteht ein Unterschied zwischen zwei Merkmalen. Auf Grundlage der Teststatistik einer Stichprobe entscheiden wir uns für die Forschungs- bzw. Alternativhypothese. Es handelt sich um die richtige Entscheidung.

Fehler 2. Art oder β -Fehler

Entscheidung 4: In der Grundgesamtheit gilt die Forschungs- bzw. die Alternativhypothese. Auf Basis der Teststatistik entscheiden wir uns allerdings gegen die Forschungs- bzw. Alternativhypothese. Wir akzeptieren also die Nullhypothese, obwohl in der Grundgesamtheit die Forschungs- bzw. Alternativhypothese gilt. Ein solcher Fehler wird „Fehler 2. Art“ oder auch β -Fehler genannt.

Alpha- und Beta-Fehler verhalten sich gegenläufig. Je kleiner die Wahrscheinlichkeit für einen α -Fehler ist, desto größer ist die Wahrscheinlichkeit für einen β -Fehler (Gehring und Weins 2009, S. 275). Die (mathematische) Berechnung für die Wahrscheinlichkeit eines β -Fehlers ist allerdings deutlich schwieriger als die Berechnung für die Wahrscheinlichkeit eines α -Fehlers. Für die Berechnung des β -Fehlers benötigen wir präzise Annahmen über die Stärke des vermuteten Zusammenhangs bzw. die Größe eines Unterschieds. In den meisten sozialwissenschaftlichen Hypothesen wird allerdings nur formuliert, dass es einen Zusammenhang bzw. Unterschied zwischen zwei Merkmalen gibt, nicht aber, wie stark dieser Zusammenhang bzw. wie groß dieser Unterschied ist. Die Wahrscheinlichkeit eines β -Fehlers ist also von der konkreten Hypothese abhängig, während die Wahrscheinlichkeit für einen α -Fehler konstant ist, da ein Zusammenhang bzw. Unterschied stets verneint wird (Nullhypothese). In der Forschungspraxis dominiert daher der α -Fehler, der „einen auf Schritt und Tritt begleitet“ (Kuckartz et al. 2013). Beispiele für die Berechnung des β -Fehlers finden sich bei Diaz-Bone (2018) und Ludwig-Mayerhofer et al. (2014, S. 167-172).

5.4.3 t-Test



Der t-Test ist eines der am häufigsten eingesetzten Testverfahren (Gehring und Weins 2009, S. 286-298; Sedlmeier und Renkewitz 2013, S. 397-415; Kuckartz et al. 2013, S. 159-184; Ludwig-Mayerhofer et al. 2014, S. 136-155). Mit einem t-Test wird geprüft, ob der Unterschied zwischen zwei arithmetischen Mitteln in einer Stichprobe auf zufällige Abweichungen zurückzuführen ist oder ob der Unterschied (wahrscheinlich) auch in entsprechenden Grundgesamtheit existiert.

Mit Daten der ALLBUS stellen Sie fest, dass sich die mittlere Lebenszufriedenheit zwischen Männern und Frauen unterscheidet. Frauen haben eine mittlere Lebenszufriedenheit von 7,6 und Männer haben eine mittlere Lebenszufriedenheit von 7,2 auf einer 11-Punkt-Skala (höhere Werte deuten auf eine größere Lebenszufriedenheit hin). Da es sich bei der ALLBUS um eine Stichprobe handelt, könnte der Unterschied zufallsbedingt sein. Ein t-Test bietet die Möglichkeit, zu prüfen, ob der Unterschied zwischen Männern und Frauen auf zufallsbedingte Abweichungen zurückzuführen ist oder ob der Unterschied von der Stichprobe auf die Grundgesamtheit übertragen werden kann. Ein statistisch signifikantes Ergebnis liegt vor, wenn sich der Unterschied zwischen Männern und Frauen nur schwierig mit zufallsbedingten Abweichungen erklären lässt und man daher mit sehr großer Sicherheit davon ausgehen kann, dass sich die Lebenszufriedenheit zwischen Männern und Frauen auch in der Grundgesamtheit unterscheidet.



Signifikanz und Bedeutsamkeit

Mit einem t-Test wird geprüft, ob ein Unterschied zwischen zwei arithmetischen Mitteln in einer Stichprobe wahrscheinlich auch in der entsprechenden Grundgesamtheit existiert. Ein empirischer Befund wird als „signifikant“ bezeichnet, wenn der Befund einer Zufallsstichprobe auf die Grundgesamtheit übertragen werden kann. Ein Signifikanztest macht keine Aussage darüber, ob man einen bedeutenden bzw. wichtigen Unterschied gefunden hat oder nicht (Kuckartz et al. 2013, S. 153-154; Ludwig-Mayerhofer et al. 2014, S. 172-173).

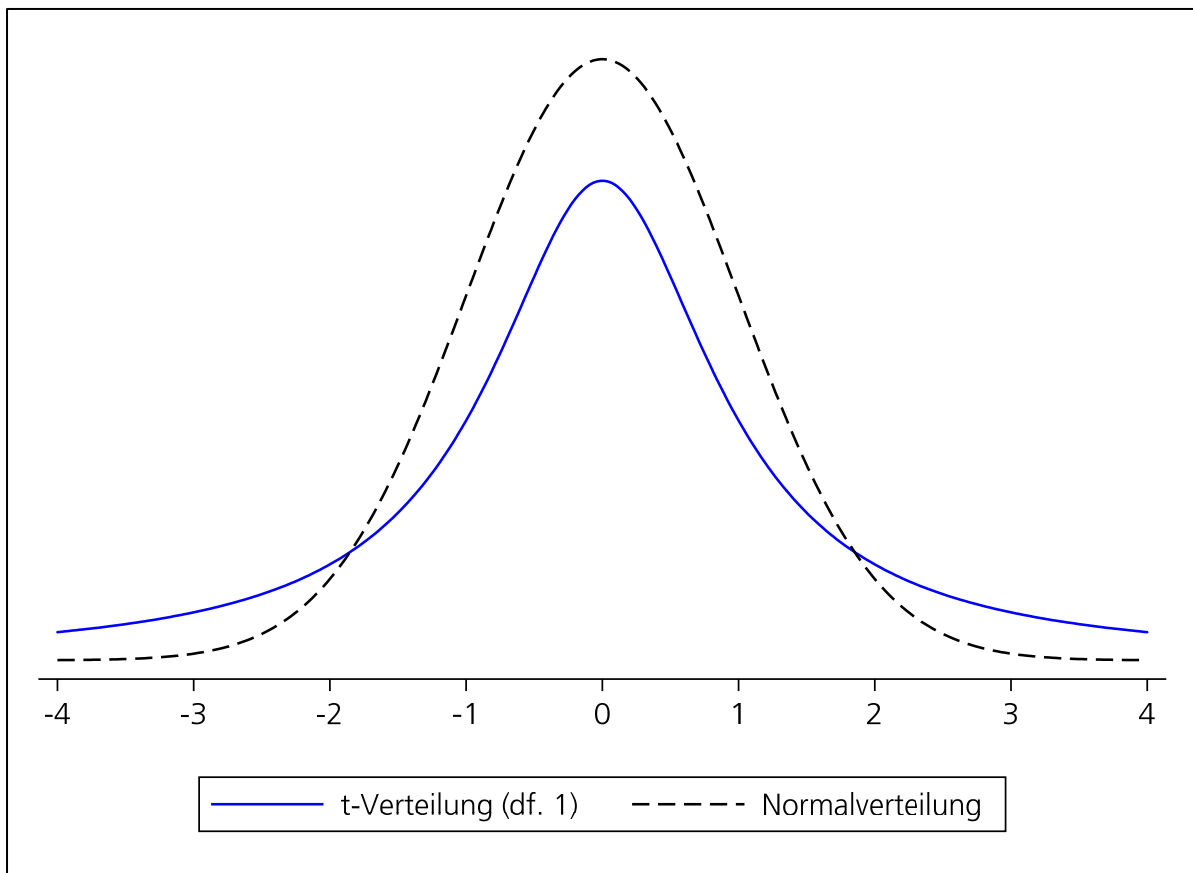
Der t-Test trägt seinen Namen nicht zufällig. Mit Hilfe des t-Tests wird die (empirische) Prüfgröße T berechnet. Diese Prüfgröße T wird mit dem sogenannten kritischen t-Wert der t-Verteilung abgeglichen. Ist der Betrag der Prüfgröße T größer als der kritische t-Wert, wird die Nullhypothese verworfen und die Forschungshypothese bzw. Alternativhypothese akzeptiert.

t-Verteilung

Die t-Verteilung wird häufig als „kleine Schwester“ (Kuckartz et al. 2013, S. 134) der Standardnormalverteilung bezeichnet. Die t-Verteilung wurde von William Sealy Gosset unter dem Pseudonym „Student“ (1908) entwickelt und ist – wie die Standardnormalverteilung – symmetrisch und eingipflig mit einem Mittelwert von 0. Während die Standardnormalverteilung bei großen Stichproben genutzt wird, kommt die t-Verteilung vor allem bei kleinen Stichproben zum Einsatz.

Abbildung 32 zeigt die t-Verteilung im Vergleich zur Normalverteilung. Die t-Verteilung ist flacher als die Normalverteilung, „denn gewissermaßen besteht der Preis, den man für die geringe Datenmenge zahlen muss, darin, dass die Werte stärker streuen und weniger Werte nahe am Mittelwert liegen“ (Kuckartz et al. 2013, S. 134).

Abbildung 32: t-Verteilung und Normalverteilung



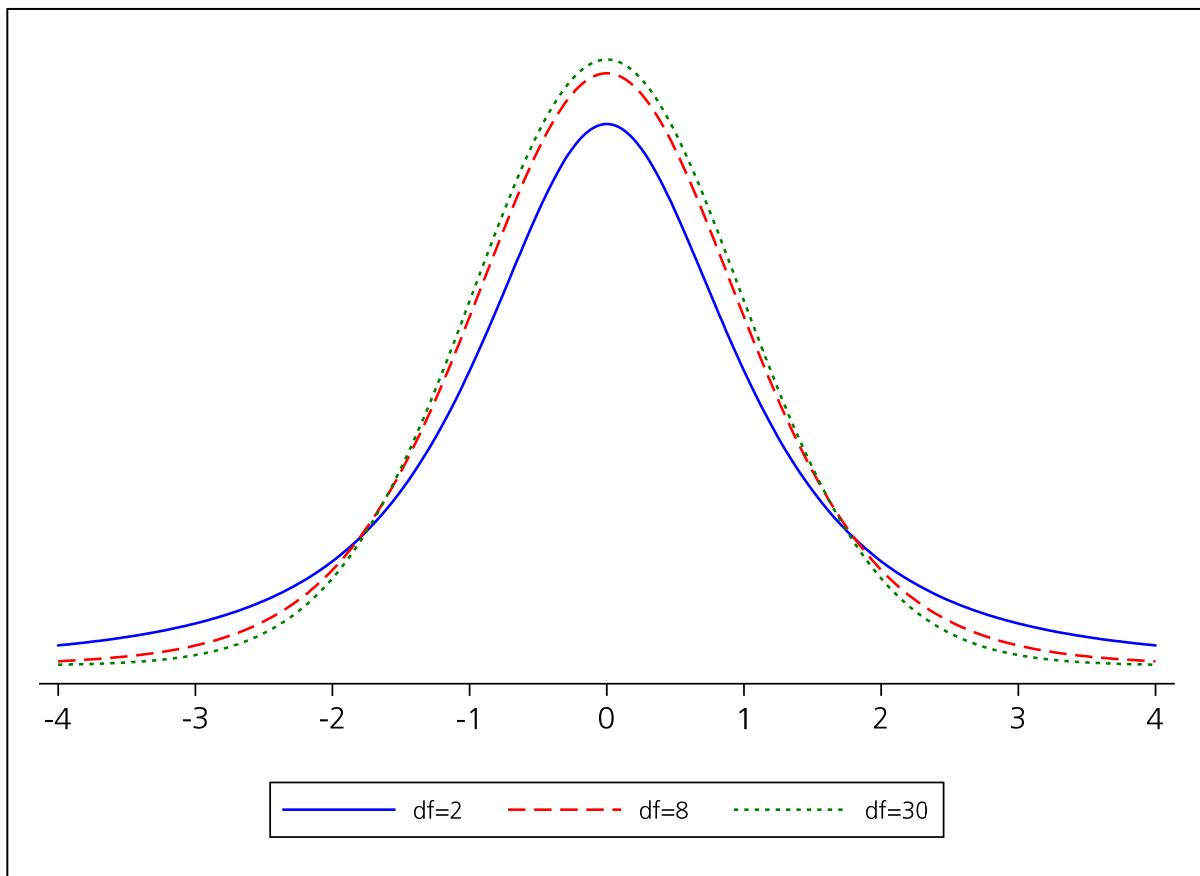
Quelle: Eigene Darstellung

Wie Abbildung 33 allerdings illustriert, gibt es nicht eine t-Verteilung, sondern viele t-Verteilungen. Je größer die Fallzahl ist, desto stärker nähert sich die t-Verteilung der Normalverteilung an. Bereits ab einer Fallzahl von 30 ist praktisch kein Unterschied mehr zwischen der t-Verteilung und der Normalverteilung zu erkennen. Bei einer Fallzahl von 120 sind die kritischen Werte der t-Verteilung mit der Normalverteilung nahezu identisch. Die kritischen t-Werte bei kleineren Fallzahlen sind Tabelle 60 zu entnehmen; das praktische Vorgehen wird weiter unten illustriert.

Was sind Freiheitsgrade?

Die Beschreibung einer Verteilung ist von sogenannten Freiheitsgraden abhängig. Was ist damit gemeint? Die Anzahl der Freiheitsgrade (englisch: degrees of freedom (df)) ist die Anzahl der Werte, die frei geändert werden können, ohne den interessierenden statistischen Parameter oder ein zur Berechnung des statistischen Parameters benötigtes Zwischenergebnis zu ändern. Ein Beispiel: Wir haben das Alter von drei Personen erfasst: 36, 49 und 65 Jahre. Das arithmetische Mittel ist 50 Jahre. Wir könnten jetzt das Alter von zwei anderen Personen erfassen (z. B. 40 und 80 Jahre), dann müsste die dritte Person allerdings 30 Jahre alt sein, um das arithmetische Mittel von 50 Jahren zu erhalten. Für die Berechnung des arithmetischen Mittels sind also nur zwei Werte (bzw. zwei Altersangaben) frei, der dritte Wert kann nicht geändert werden (Bortz und Schuster 2010, S. 121; Ludwig-Mayerhofer et al. 2014, S. 117-118).

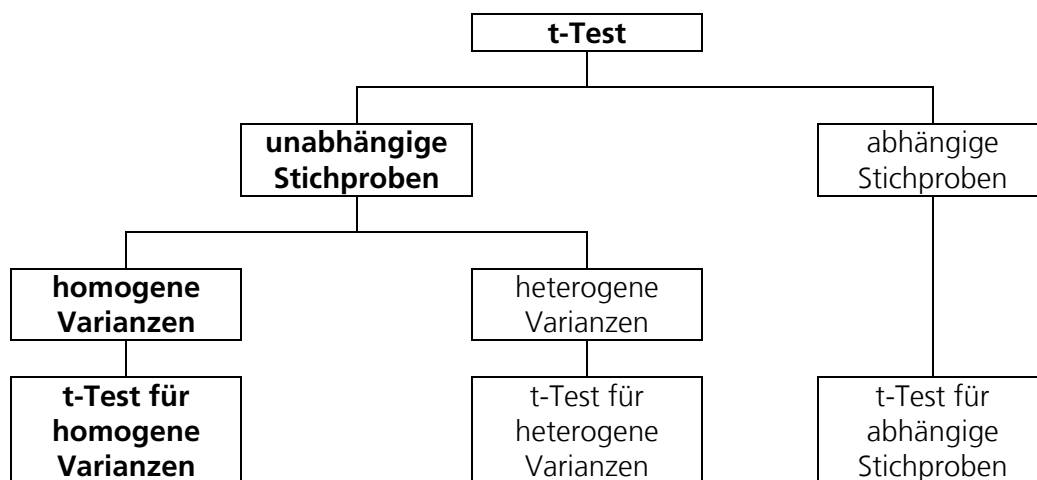
Abbildung 33: Verschiedene t-Verteilungen



Quelle: Eigene Darstellung

In Abbildung 34 werden verschiedene Varianten eines t-Tests unterschieden. Zunächst muss geklärt werden, ob es sich um abhängige oder unabhängige Stichproben handelt. Bei unabhängigen Stichproben muss zudem geprüft werden, ob sich die Varianzen des Merkmals (z.B. Lebenszufriedenheit) in den beiden Stichproben unterscheiden. In Abhängigkeit von diesen Entscheidungen wird ein bestimmter t-Test ausgewählt.

Abbildung 34: Varianten des t-Tests



Quelle: Eigene Darstellung

Art der Stichprobe

Bei der Art der Stichprobe ist zu klären, ob es sich um abhängige oder unabhängige Stichproben handelt. Ein typisches Beispiel für eine abhängige Stichprobe wäre eine wiederholte Befragung. Die Personen werden beispielsweise zu zwei Zeitpunkten zu ihrer Lebenszufriedenheit gefragt. Die erste Befragung fand dabei im Sommer und die zweite Befragung im Winter statt (Panelbefragung). Eine abhängige Stichprobe liegt aber auch vor, wenn beispielsweise Personen zufällig gezogen werden und dann jeweils zu diesen Personen noch eine weitere Person berücksichtigt wird, die in bestimmter Weise von der Person abhängt. Beispiele wären die Ehepartner oder die Kinder der zufällig gezogenen Personen (Ludwig-Mayerhofer et al. 2014, S. 152-153).

Eine unabhängige Stichprobe liegt vor, wenn die Personen der zwei Stichproben in keiner sich beeinflussenden Beziehung stehen. Eine Erhebung im Rahmen der ALLBUS oder des ESS sind typische Beispiele für unabhängige Stichproben. Die Personen sind jeweils unabhängig von anderen Personen ausgewählt.

Prüfung der Varianzhomogenität

Liegen zwei unabhängige Stichproben vor, dann muss zusätzlich die Varianzhomogenität des Merkmals geprüft werden. Die Varianzhomogenität bezieht sich auf die Streuung des Merkmals in den beiden Stichproben (z.B. die Streuung der Lebenszufriedenheit). Varianzhomogenität wäre gegeben, wenn sich die Varianz der Lebenszufriedenheit zwischen Männern und Frauen in der Grundgesamtheit nicht signifikant unterscheidet. Varianzheterogenität liegt vor, wenn sich die Streuung der Lebenszufriedenheit in beiden Stichproben signifikant unterscheidet. Für die Prüfung der Varianzhomogenität stehen wieder eigene Testverfahren zur Verfügung (z.B. der Bartlett-Test oder der Levene-Test). Diese Tests sind in den gängigen Statistikprogrammen implementiert. In SPSS werden bei der Berechnung eines t-Tests für unabhängige Stichproben automatisch auch die Resultate eines Levene-Tests ausgegeben, so dass die Prüfung der Varianzhomogenität nur ein (weiterer) Zwischenschritt bei einem t-Test ist.

An dieser Stelle wollen wir den Test auf Varianzhomogenität nicht vorstellen, sondern nur die Testentscheidung illustrieren. Bei einem Test auf Varianzhomogenität ist die Nullhypothese, dass es keinen signifikanten Unterschied zwischen den Varianzen der beiden Gruppen in der Grundgesamtheit gibt. Entsprechend lautet die Alternativhypothese, dass sich die Varianzen in den beiden Gruppen unterscheiden. Bei einem signifikanten Testergebnis ($p < 0,05$) muss folglich die Nullhypothese verworfen und die Alternativhypothese akzeptiert werden. Mit anderen Worten: Bei einem signifikanten Testergebnis ist davon auszugehen, dass sich die Varianzen unterscheiden und der t-Test für heterogene Varianzen angewendet werden muss.

5.4.3.1 t-Test für unabhängige Stichproben mit homogenen Varianzen

Bei einem t-Test wird die Prüfgröße t berechnet. Dieser auf Basis der Stichprobe berechnete empirische t -Wert wird mit dem kritischen t -Wert der t -Verteilung verglichen. Falls der Betrag des empirischen t -Werts größer ist als der (entsprechende) kritische t -Wert, wird die Nullhypothese abgelehnt und die Forschungshypothese bzw. Alternativhypothese akzeptiert. Ein größerer empirischer t -Wert deutet darauf hin, dass der Unterschied zwischen zwei arithmetischen Mitteln in der Stichprobe wahrscheinlich nicht zufallsbedingt ist, sondern auf tatsächliche Unterschiede in der Grundgesamtheit zurückzuführen ist.

Bei unabhängigen Stichproben mit homogenen Varianzen wird die Prüfgröße t mit folgender Formel bestimmt (Kuckartz et al. 2013, S. 162-163; Ludwig-Mayerhofer et al. 2014, S. 144):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1) * S_1^{*2} + (n_2 - 1) * S_2^{*2}}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Dabei sind \bar{X}_1 und \bar{X}_2 die arithmetischen Mittel der Variable X (z.B. Lebenszufriedenheit) in Gruppe 1 und Gruppe 2. Mit n_1 und n_2 sind die Fallzahlen der beiden Gruppen gemeint. S_1^{*2} und S_2^{*2} sind die korrigierten Varianzen der Variable in Gruppe 1 und Gruppe 2. Für die Berechnung der Prüfgröße t sind also die gruppenspezifischen Angaben zur Stichprobengröße, zum arithmetischen Mittel sowie zur korrigierten Varianz erforderlich.

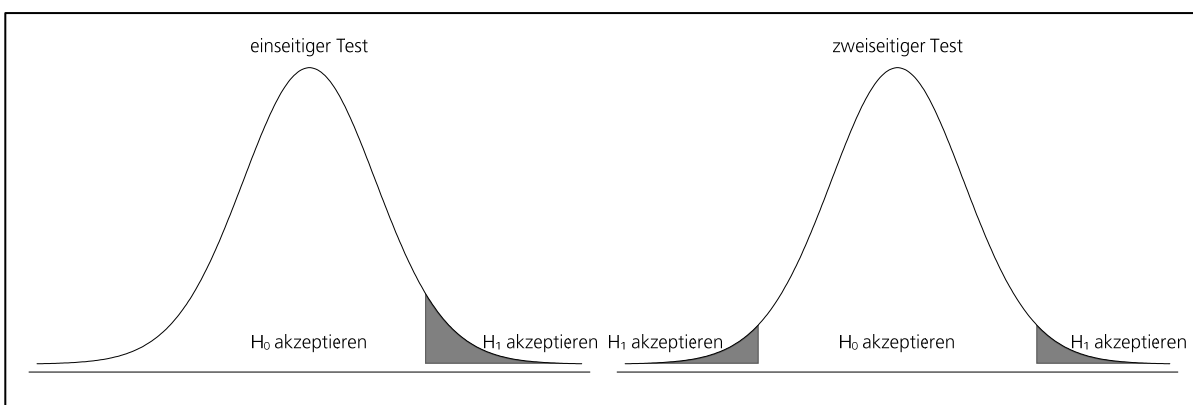
In Abhängigkeit von den formulierten Hypothesen können einseitige und zweiseitige t -Tests unterschieden werden (siehe Abbildung 35). Bei einem zweiseitigen Test können Werte, die auf Abweichungen sowohl nach oben als auch nach unten hindeuten, als Indiz gegen die Nullhypothese aufgefasst werden. Ein Beispiel: Die ungerichtete Hypothese „Die Lebenszufriedenheit zwischen Männern und Frauen unterscheidet sich“ wäre der Ausgangspunkt für einen zweiseitigen Test. Starke Abweichungen zugunsten der Frauen oder der Männer deuten darauf hin, dass die Nullhypothese nicht zutrifft.

Einseitiger und zweiseitiger t-Test

Bei einer gerichteten Hypothese wird ein einseitiger t -Test durchgeführt. Ausgangspunkt wäre folgende Hypothese: „Frauen haben eine höhere Lebenszufriedenheit als Männer.“ Nur starke Abweichungen zugunsten der Frauen sprechen gegen die Nullhypothese.

In Abbildung 35 sind die Logiken des einseitigen und des zweiseitigen t -Tests illustriert. Bei einem zweiseitigen t -Test (ungerichtete Hypothese) können starke Abweichungen in beide Richtungen auf die Ablehnung der Nullhypothese hindeuten (grau markierter Bereich). Bei einem einseitigen t -Test (gerichtete Hypothese) deuten nur starke Abweichungen in eine Richtung auf die Ablehnung der Nullhypothese hin.

Abbildung 35: Einseitiger und zweiseitiger t -Test



Quelle: Eigene Darstellung



Die Durchführung eines t-Test wird im Folgenden an einem Beispiel illustriert. Wir unterscheiden insgesamt vier Schritte: Erstens wird die Forschungs- bzw. Alternativhypothese formuliert, der die Nullhypothese gegenübergestellt wird. Zweitens wird die empirische Prüfgröße T berechnet. Drittens wird der kritische t-Wert abgelesen. Die Entscheidung über die Ablehnung der Nullhypothese bzw. der Annahme der Forschungshypothese fällt viertens auf Grundlage des Vergleichs der Prüfgröße T mit dem entsprechenden kritischen t-Wert.

Hypothesen formulieren

Vor der Berechnung der Prüfgröße T werden Forschungshypothese und Nullhypothese formuliert. Folgende Forschungshypothese wird formuliert.

H_1 : Die Lebenszufriedenheit zwischen Männern und Frauen unterscheidet sich.

In der Forschungshypothese wird zwar ein Unterschied zwischen Männern und Frauen postuliert, aber es wird keine Aussage über die Richtung getroffen. Gemäß der Hypothese können Frauen eine höhere Lebenszufriedenheit als Männer oder Männer eine höhere Lebenszufriedenheit als Frauen haben. Es handelt sich also um einen zweiseitigen Signifikanztest. Der Forschungshypothese wird die Nullhypothese gegenübergestellt, die keinen Zusammenhang zwischen den Merkmalen Geschlecht und Lebenszufriedenheit unterstellt. Sie lautet:

H_0 : Die Lebenszufriedenheit zwischen Männern und Frauen unterscheidet sich nicht.

Berechnung der Prüfgröße t

In Tabelle 59 sind die (fiktiven) Ergebnisse einer Befragung zur Lebenszufriedenheit dokumentiert. Die Zufallsstichprobe umfasst die Angaben von 42 Personen. Die mittlere Lebenszufriedenheit liegt bei den Frauen bei 7,5 Punkten, bei den Männern bei 6,5 Punkten. Die korrigierte Varianz bei den Frauen beträgt 1 und bei den Männern 1,2. Für unser Beispiel unterstellen wir, dass die Varianzhomogenität gewährleistet ist.

Tabelle 59: Lebenszufriedenheit von Frauen und Männern

	Frauen	Männer
Stichprobengröße	20	22
Arithmetisches Mittel (in Punkten)	7,5	6,5
Korrigierte Varianz	1,0	1,2

Quelle: Eigene Darstellung

Die Angaben der Stichprobe werden in die Formel zur Berechnung der Prüfgröße T eingetragen:

$$T = \frac{7,5 - 6,5}{\sqrt{\frac{(20-1) \cdot 1,0 + (22-1) \cdot 1,2}{20+22-2} \cdot \left(\frac{1}{20} + \frac{1}{22}\right)}} = \frac{1}{\sqrt{\frac{19+25,2}{40} \cdot \left(\frac{1}{20} + \frac{1}{22}\right)}} = \frac{1}{\sqrt{1,11 \cdot 0,1}} = \frac{1}{0,33} = 3,03$$

Die Prüfgröße beträgt: 3,03.

Berechnung des kritischen t-Werts

Für die Bestimmung des kritischen t-Werts sind drei Informationen erforderlich: Erstens die Anzahl der sogenannten Freiheitsgrade, zweitens die Festlegung des Signifikanzniveaus und drittens die Entscheidung, ob es sich um einen einseitigen oder zweiseitigen Signifikanztest handelt. Liegen diese Informationen vor, kann der kritische t-Wert aus Tabelle 60 abgelesen werden.

Für die Bestimmung der Anzahl der Freiheitsgrade wird die Fallzahl der beiden Gruppen addiert und anschließend um zwei verringert:

$$\text{Freiheitsgrade (df)} = n_1 + n_2 - 2$$

Mit n_1 und n_2 sind die Fallzahlen der Gruppen gemeint. Im Beispiel betragen die Fallzahlen 20 (n_1) und 22 (n_2). Die Gesamtfallzahl ist 42. Diese wird um 2 verringert (40). Die Anzahl der Freiheitsgrade ist 40.

Das Signifikanzniveau wird auf fünf Prozent festgelegt ($\alpha = 0,05$). Die Hypothese postuliert einen Unterschied bei der Lebenszufriedenheit zwischen Männern und Frauen, aber es wird keine Aussage über die Richtung getroffen. Es handelt sich also um eine ungerichtete Hypothese und damit um einen zweiseitigen Signifikanztest.

Mit der berechneten Anzahl an Freiheitsgraden (40), der Festlegung des Signifikanzniveaus ($\alpha = 0,05$) sowie der Art des Hypothesentests (zweiseitig) kann in Tabelle 60 der kritische t-Wert direkt abgelesen werden. Ausgehend von der Anzahl an Freiheitsgraden (40) muss der Wert der Tabelle in der Spalte „zweiseitig“ und „ $\alpha = 0,05$ “ abgelesen werden. Der kritische t-Wert liegt bei 2,021.

Im letzten Schritt wird der Betrag des empirischen t-Werts (|3,03|) mit dem kritischen t-Wert verglichen. Ist der Betrag des empirischen t-Werts größer als der kritische t-Wert, dann wird die Nullhypothese verworfen und die Alternativhypothese akzeptiert. Dies ist in unserem Beispiel der Fall. Die Wahrscheinlichkeit, dass bei gegebener Stichprobe die Nullhypothese korrekt ist, ist kleiner als 5 Prozent. Deshalb wird die Nullhypothese abgelehnt und die Alternativhypothese akzeptiert. Die Lebenszufriedenheit zwischen Männern und Frauen unterscheidet sich nicht nur in der Stichprobe, sondern (wahrscheinlich) auch in der Grundgesamtheit. Der Unterschied bei der Lebenszufriedenheit ist statistisch signifikant.

**Testentscheidung
treffen**

An dieser Stelle nochmals zur Erinnerung: Die statistische Signifikanz trifft keine Aussage darüber, wie groß oder bedeutend der Unterschied in der Grundgesamtheit ist. Signifikanztests sagen auch nichts über die Stärke eines Zusammenhangs aus. Auf Basis eines Signifikanztests können wir die Entscheidung treffen, ob ein empirischer Befund einer Zufallsstichprobe (z.B. Unterschiede bei der Lebenszufriedenheit) auf die entsprechende Grundgesamtheit übertragen werden kann (wir also die Nullhypothese verwerfen können). Keinesfalls darf auf Basis eines statistischen Tests eine Aussage über die Bedeutung des empirischen Befunds abgeleitet werden.



Tabelle 60: Kritische Werte der t-Verteilung

Freiheitsgrade	einseitig		zweiseitig	
	0,95 ($\alpha = 0,05$)	0,99 ($\alpha = 0,01$)	0,975 ($\alpha = 0,05$)	0,995 ($\alpha = 0,01$)
1	6,314	31,821	12,706	63,657
2	2,920	6,965	4,303	9,925
3	2,353	4,541	3,182	5,841
4	2,132	3,747	2,776	4,604
5	2,015	3,365	2,571	4,032
6	1,943	3,143	2,447	3,707
7	1,895	2,998	2,365	3,499
8	1,860	2,896	2,306	3,355
9	1,833	2,821	2,262	3,250
10	1,812	2,764	2,228	3,169
11	1,796	2,718	2,201	3,106
12	1,782	2,681	2,179	3,055
13	1,771	2,650	2,160	3,012
14	1,761	2,624	2,145	2,977
15	1,753	2,602	2,131	2,947
16	1,746	2,583	2,120	2,921
17	1,740	2,567	2,110	2,898
18	1,734	2,552	2,101	2,878
19	1,729	2,539	2,093	2,861
20	1,725	2,528	2,086	2,845
21	1,721	2,518	2,080	2,831
22	1,717	2,508	2,074	2,819
23	1,714	2,500	2,069	2,807
24	1,711	2,492	2,064	2,797
25	1,708	2,485	2,060	2,787
26	1,706	2,479	2,056	2,779
27	1,703	2,473	2,052	2,771
28	1,701	2,467	2,048	2,763
29	1,699	2,462	2,045	2,756
30	1,697	2,457	2,042	2,750
31	1,696	2,453	2,040	2,744
32	1,694	2,449	2,037	2,739
33	1,692	2,445	2,035	2,733
34	1,691	2,441	2,032	2,728
35	1,690	2,438	2,030	2,724
36	1,688	2,435	2,028	2,720
37	1,687	2,431	2,026	2,715
38	1,686	2,429	2,024	2,712
39	1,685	2,426	2,023	2,708
40	1,684	2,423	2,021	2,705
50	1,676	2,403	2,009	2,678
60	1,671	2,390	2,000	2,660
120	1,658	2,358	1,980	2,617
∞	1,645	2,326	1,960	2,576

Quelle: Eigene Darstellung

Wir wollen die Durchführung eines t-Tests an einem weiteren Beispiel illustrieren. Dabei wird exemplarisch eine einseitige Hypothese formuliert und das Signifikanzniveau auf 1 Prozent festgelegt ($\alpha = 0,01$). Inhaltlich wird geprüft, ob Westdeutsche eine höhere Lebenszufriedenheit haben als Ostdeutsche. Folgende Forschungshypothese wird formuliert:



H_1 : Westdeutsche haben eine höhere Lebenszufriedenheit als Ostdeutsche.

Dieser Forschungshypothese wird die Nullhypothese gegenübergestellt. Sie lautet:

H_0 : Die Lebenszufriedenheit zwischen West- und Ostdeutschen unterscheidet sich nicht.

In Tabelle 61 sind (fiktive) Angaben zur Lebenszufriedenheit von Ost- und Westdeutschen dokumentiert. In unserer Stichprobe haben Westdeutsche mit einem arithmetischen Mittel von 8,0 eine höhere Lebenszufriedenheit als Ostdeutsche, die ein arithmetisches Mittel von 6,5 aufweisen. Dieser Stichprobenbefund spricht für unsere Forschungshypothese. Mit dem t-Test wird geprüft, ob der Unterschied von 1,5 Punkten möglicherweise auf Zufallsfehler zurückzuführen ist oder ob von der Stichprobe auf die Grundgesamtheit geschlossen werden kann.

Tabelle 61: Lebenszufriedenheit von West- und Ostdeutschen

	Westdeutsche	Ostdeutsche
Stichprobengröße	31	31
Arithmetisches Mittel (in Punkten)	8,0	6,5
Korrigierte Varianz	1,0	1,2

Quelle: Eigene Darstellung

Die Angaben der Stichprobe werden in die Formel zur Berechnung der Prüfgröße t eingetragen:

$$T = \frac{8,0 - 6,5}{\sqrt{\frac{(31-1) \cdot 1,0 + (31-1) \cdot 1,2}{31+31-2} \cdot \left(\frac{1}{31} + \frac{1}{31}\right)}} = \frac{1,5}{\sqrt{\frac{30+36}{60} \cdot \left(\frac{1}{31} + \frac{1}{31}\right)}} = \frac{1,5}{\sqrt{1,1 \cdot 0,06}} = \frac{1,5}{0,26} = 5,77$$

Der empirische t-Wert beträgt 5,77. Dieser Wert muss im Folgenden mit dem kritischen t-Wert abgeglichen werden. Die Anzahl der Freiheitsgrade beträgt 60 ($31+31-2 = 60$). Es handelt sich um einen einseitigen Test mit einem Signifikanzniveau von einem Prozent ($\alpha = 0,01$). Der entsprechende kritische t-Wert beträgt 2,39.

Der Betrag des empirischen t-Werts ist mit 5,77 größer als der kritische t-Wert mit 2,39. Das Ergebnis ist auf dem 1-Prozent-Niveau signifikant. Auch in der Grundgesamtheit werden Westdeutsche eine größere Lebenszufriedenheit haben als Ostdeutsche.

In den beiden folgenden Abschnitten wird exemplarisch die Berechnung eines t-Test für unabhängige Stichproben mit heterogenen Varianzen und eines t-Tests für abhängige Stichproben illustriert. In der Forschungspraxis werden solche Berechnungen Statistikprogrammen überlassen, die die erforderlichen Berechnungen in der Regel schneller und fehlerfrei durchführen können als ein Mensch. Die folgenden Ausführungen sind daher als Exkurse zu verstehen.

5.4.3.2 t-Test für unabhängige Stichproben mit heterogenen Varianzen

Exkurs



Bei den bisherigen Beispielen haben wir (ungeprüft) vorausgesetzt, dass sich die korrigierten Varianzen der beiden Gruppen in der Grundgesamtheit nicht unterscheiden. Die jeweiligen korrigierten Varianzen waren sehr ähnlich und konnten als vergleichbar interpretiert werden. Bei der Arbeit mit einem Statistikprogramm (z. B. SPSS) wird die Annahme der Varianzhomogenität meist automatisch geprüft und die Ergebnisse zweier t-Tests werden ausgegeben: für homogene Varianzen und für heterogene Varianzen. In der Praxis und der Arbeit mit großen Stichproben unterscheiden sich die Ergebnisse dieser beiden t-Tests zudem nur geringfügig.

Zur Vollständigkeit wollen wir an dieser Stelle die Unterschiede bei der Durchführung eines t-Tests mit heterogenen Varianzen darstellen. Für die Berechnung der empirischen Prüfgröße T bei heterogenen Varianzen wird folgende Formel verwendet (Kuckartz et al. 2013, S. 164; Ludwig-Mayerhofer et al. 2014, S. 144):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}}}$$

Deutlich aufwändiger bei der Durchführung eines t-Tests mit heterogenen Varianzen ist die Berechnung der Anzahl der Freiheitsgrade. Während bei einem t-Test mit homogenen Varianzen die Fallzahl der beiden Stichproben addiert und um 2 verringert wird ($n_1 + n_2 - 2$), wird bei heterogenen Varianzen folgende Berechnung der Freiheitsgrade vorgeschlagen (Bortz und Schuster 2010, S. 123; Ludwig-Mayerhofer et al. 2014, S. 144).

$$\text{Freiheitsgrade (df)} = \frac{\left(\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{S_1^{*2}}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{S_2^{*2}}{n_2}\right)^2}$$

Beispiel



Wir illustrieren die Durchführung eines t-Test mit heterogenen Varianzen und prüfen, ob sich die Zufriedenheit mit der Demokratie, die auf einer Skala von 0 bis 10 erfasst wurde (höhere Werte deuten auf eine positivere Bewertung der Demokratie hin), zwischen Männern und Frauen unterscheidet. Dabei werden folgende Hypothesen formuliert:

H_1 : Die Zufriedenheit mit der Demokratie unterscheidet sich zwischen Männern und Frauen.

Dieser Forschungshypothese wird die Nullhypothese gegenübergestellt. Sie lautet:

H_0 : Die Zufriedenheit mit der Demokratie unterscheidet sich nicht zwischen Männern und Frauen.

Es handelt sich um eine ungerichtete Hypothese und folglich um einen zweiseitigen Hypothesentest. Das Signifikanzniveau wird auf 5 Prozent festgelegt ($\alpha = 0,05$). In Tabelle 62 sind die fiktiven Daten zur Zufriedenheit mit der Demokratie dokumentiert. Die mittlere Demokratiezufriedenheit bei den Frauen liegt bei 7,0 und einer korrigierten Varianz von 2,0. Bei den Männern liegt die durchschnittliche Demokratiezufriedenheit bei 6,8 und einer korrigierten Varianz von 1,0.

Tabelle 62: Zufriedenheit mit der Demokratie

	Frauen	Männer
Stichprobengröße	30	20
Arithmetisches Mittel (in Punkten)	7,0	6,8
Korrigierte Varianz	2,0	1,0

Quelle: Eigene Darstellung

Die Angaben der Stichprobe werden in die Formel zur Berechnung der Prüfgröße T eingetragen:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}} = \frac{7 - 6,80}{\sqrt{\frac{2}{30} + \frac{1}{20}}} = \frac{0,2}{\sqrt{\frac{2}{30} + \frac{1}{20}}} = \frac{0,2}{\sqrt{\frac{7}{60}}} = \frac{0,2}{0,34} = 0,59$$

Der empirische t-Wert beträgt 0,59. Der empirische t-Wert muss mit dem entsprechenden kritischen t-Wert der t-Verteilung verglichen werden. Dazu muss die Anzahl der Freiheitsgrade ermittelt werden.

$$df = \frac{\left(\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^{*2}}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^{*2}}{n_2}\right)^2} = \frac{\left(\frac{2}{30} + \frac{1}{20}\right)^2}{\frac{1}{30-1} \left(\frac{2}{30}\right)^2 + \frac{1}{20-1} \left(\frac{1}{20}\right)^2} = \frac{\left(\frac{7}{60}\right)^2}{\frac{1}{29} * \frac{1}{225} + \frac{1}{19} * \frac{1}{400}} = \frac{\frac{49}{3600}}{\frac{1}{6525} + \frac{1}{7600}} = 47,79$$

Die Berechnung der Anzahl an Freiheitsgraden ergibt einen Wert von 47,79. In Tabelle 60 ist der kritische t-Wert für 47,79 Freiheitsgrade nicht dokumentiert. Als Ersatz wird der kritische t-Wert bei 40 Freiheitsgraden herangezogen, der bei einem zweiseitigen Test und einem Signifikanzniveau von 5 Prozent bei 2,021 liegt.

Der Betrag des empirischen t-Werts (0,59) ist kleiner als der kritische t-Wert (2,021). Folglich wird die Nullhypothese akzeptiert, die keinen Unterschied in der Zufriedenheit mit der Demokratie zwischen Männern und Frauen postuliert. Das Ergebnis ist folglich nicht signifikant.

5.4.3.3 t-Test für abhängige Stichproben

Bei den bisher betrachteten t-Tests waren die Stichproben (z. B. Frauen und Männer) jeweils unabhängig und nicht miteinander verbunden. Bei experimentellen Designs oder auch bei Panelstudien sind die Stichproben allerdings häufig nicht unabhängig voneinander. Ein einfaches Beispiel ist die wiederholte Messung eines Merkmals (z. B. Einstellung zum Umweltschutz) an zwei Zeitpunkten (z. B. vor und nach einer Veranstaltung zum Umweltschutz). Mit einem t-Test für abhängige Stichproben kann untersucht werden, ob sich beispielsweise die Einstellung zum Umweltschutz geändert hat oder nicht.



Für die Berechnung des empirischen t-Werts bei abhängigen Stichproben wird folgende Formel verwendet (Kuckartz et al. 2013, S. 171; Ludwig-Mayerhofer et al. 2014, S. 153):

$$T = \frac{\bar{X}_D}{\frac{s_D^*}{\sqrt{n}}}$$

Dabei ist \bar{x}_D der Mittelwert der Differenzen aller Wertpaare (identisch mit der Differenz der beiden Stichprobenmittelwerte), s_D^* die korrigierte Standardabweichung der Differenzen und n die Anzahl der Wertpaare.

Die Berechnung der korrigierten Standardabweichung der Differenzen (s_D^*) erfolgt nach folgender Formel:

$$s_D^* = \sqrt{\frac{\sum_{i=1}^n (x_{Di} - \bar{x}_D)^2}{n - 1}}$$

Dabei ist x_{Di} die Differenz des Wertpaares, \bar{x}_D ist das arithmetische Mittel der Differenzen aller Wertpaare und n ist die Anzahl der Wertpaare.

Der empirische t-Wert wird wieder mit dem kritischen t-Wert der t-Verteilung verglichen. Die Anzahl der Freiheitsgrade beim t-Test für abhängige Stichproben beträgt $n-1$. Ist der Betrag des empirischen t-Werts größer als der kritische t-Wert, dann wird die Nullhypothese verworfen und die Alternativ- bzw. Forschungshypothese (vorläufig) akzeptiert.



Wir wollen die Durchführung eines t-Werts für abhängige Stichproben an einem Beispiel illustrieren. Aus Gründen der Übersichtlichkeit basiert das Beispiel auf einer sehr kleinen Datengrundlage. In Tabelle 63 finden sich (fiktive) Angaben von sechs Personen, die vor und nach einer Veranstaltung zu ihrer Einstellung zum Umweltschutz befragt wurden. Für jede der befragten Personen liegen also zwei Angaben vor, die ein Wertpaar bilden. Die Einstellung zum Umweltschutz wurde dabei auf einer Skala von 0 bis 10 erfasst, bei der höhere Werte auf eine positivere Haltung zum Umweltschutz hindeuten.

Die mittlere Einstellung der sechs Befragten vor der Veranstaltung liegt bei 5 und die mittlere Einstellung nach der Veranstaltung liegt bei 8. Die Differenz beträgt drei Punkte. Auf den ersten Blick hat die Teilnahme an der Veranstaltung die Einstellung zum Umweltschutz erhöht.

Tabelle 63: Beispieldaten für die Berechnung eines t-Tests bei abhängigen Stichproben

ID	Wert vor der Veranstaltung	Wert nach der Veranstaltung	Differenz
1	6	8	2
2	4	7	3
3	5	10	5
4	5	8	3
5	4	7	3
6	6	8	2
Mittelwert: 5		Mittelwert: 8	Mittelwert der Differenzen: 3

Quelle: Eigene Darstellung

Mit einem t-Test für abhängige Stichproben kann nun geklärt werden, ob die Mittelwertdifferenz von 3 Punkten möglicherweise auf zufällige Schwankungen oder aber auf den Besuch der Veranstaltung zurückgeführt werden kann. Folgende Hypothesen werden formuliert:

H_1 : Die durchschnittliche Haltung zum Umweltschutz vor und nach der Veranstaltung unterscheidet sich.

H_0 : Die durchschnittliche Haltung zum Umweltschutz vor und nach der Veranstaltung unterscheidet sich nicht.

Wir berechnen zunächst die korrigierte Standardabweichung der Differenzen:

$$s_D^* = \sqrt{\frac{(2-3)^2 + (3-3)^2 + (5-3)^2 + (3-3)^2 + (3-3)^2 + (2-3)^2}{6-1}} = \sqrt{\frac{1+0+4+0+0+1}{5}} = \sqrt{\frac{6}{5}} = 1,10$$

Das Zwischenergebnis kann anschließend in die Formel zur Berechnung der Prüfgröße T eingesetzt werden. Für unser Beispiel ergibt sich ein empirischer t-Wert von 6,71:

$$T = \frac{\bar{X}_D}{\frac{s_D^*}{\sqrt{n}}} = \frac{3}{\frac{1,1}{\sqrt{6}}} = \frac{3}{0,45} = 6,67$$

Der Betrag des empirischen t-Werts mit 6,67 wird abschließend mit dem kritischen t-Wert verglichen. Der kritische t-Wert für einen zweiseitigen Signifikanztest mit einer Irrtumswahrscheinlichkeit von 5 Prozent und fünf Freiheitsgraden liegt bei 2,571. Der Betrag des empirischen t-Werts ist größer als der kritische t-Wert. Die Nullhypothese wird abgelehnt und die Alternativ- bzw. Forschungshypothese wird vorläufig akzeptiert. Die Mittelwertdifferenz von drei Punkten ist vermutlich nicht zufällig entstanden, sondern auf die Veranstaltung zurückzuführen.

Das Beispiel zeigt, dass der Aufwand für die Berechnung des empirischen t-Werts – insbesondere durch die Berechnung der korrigierten Standardabweichung – mit steigender Fallzahl erheblich zunimmt. In der Forschungspraxis sind wir aber nicht mit kleinen, sondern mit großen Stichproben konfrontiert. Deshalb wird der empirische t-Wert (bzw. die Prüfgröße T) nicht händisch, sondern mit einem Statistikprogramm (z.B. SPSS) ermittelt.

5.4.3.4 Voraussetzungen für einen t-Test

Abschließend werden die wichtigsten Voraussetzungen für die Durchführung eines t-Tests erläutert:

- (einfache) Zufallsstichprobe
- (pseudo-)metrisches Skalenniveau der Variablen
- Merkmal in der Grundgesamtheit normalverteilt
- angemessene Fallzahl

Die Anwendung statistischer Testverfahren setzt (einfache) Zufallsstichproben

**(Einfache)
Zufallsstichprobe**

voraus. Die in diesem Kurs dargestellten und in den meisten Statistikprogrammen implementierten Testverfahren basieren auf der Annahme einer einfachen Zufallsstichprobe. Bei komplexen Stichproben sind ggf. Korrekturverfahren erforderlich. Bei willkürlichen Stichproben (Convenience Sample) können statistische Testverfahren nicht sinnvoll verwendet werden. Der Forscher bzw. die

Forscherin muss prüfen, ob es sich bei den verwendeten Daten um eine (einfache) Zufallsstichprobe handelt.

Skalenniveau der Variable

Beim t-Test werden die arithmetischen Mittel von zwei Gruppen verglichen. Die Berechnung des arithmetischen Mittels bzw. die Berechnung der korrigierten Varianz setzt ein metrisches Skalenniveau voraus. Die Variablen müssen also zumindest ein pseudometrisches Skalenniveau aufweisen, um das arithmetische Mittel und die korrigierte Varianz berechnen zu können.

Merkmal ist normalverteilt

Das untersuchte Merkmal (z.B. Lebenszufriedenheit) sollte in der Grundgesamtheit annähernd normalverteilt sein. Für die Überprüfung der Normalität einer Verteilung gibt es verschiedene Testverfahren, z.B. den Kolmogoroff-Smirnov-Test und den Lilliefors-Test (Bortz und Schuster 2010, S. 144-145). Auf die Verletzung dieser Voraussetzung reagiert der t-Test bei großen Stichproben allerdings relativ robust (Bortz und Schuster 2010, S. 122, 125).

Angemessene Fallzahl

Schließlich ist bei der Durchführung eines t-Tests auf eine angemessene Fallzahl zu achten. Zwar gewährleisten die Tests, dass auch bei kleiner Fallzahl das festgelegte Signifikanzniveau eingehalten wird, aber aus verschiedenen Gründen sind große Stichprobenumfänge wünschenswert. Bortz und Schuster (2010, S. 126) nennen als groben Orientierungspunkt eine Fallzahl von $n > 30$ je Stichprobe.

Bei kleineren Stichproben oder bei der Verletzung einzelner Voraussetzungen (z.B. Normalverteilung) existieren voraussetzungsärmere Verfahren, die als nicht-parametrische bzw. verteilungsfreie Verfahren bezeichnet werden. Für den Vergleich von zwei unabhängigen Stichproben bietet sich der U-Test von Mann-Whitney und für den Vergleich von zwei verbundenen Stichproben der Wilcoxon-Test an. Beide Testverfahren sind jeweils nicht an die Normalverteilungsvoraussetzung geknüpft (siehe ausführlich Bortz und Schuster 2010, S. 130-136). In der Soziologie und in der Politikwissenschaft sind wir allerdings meist mit großen Stichproben konfrontiert, so dass in der Regel auf den t-Test zurückgegriffen wird.

6 Literatur

- Abendschön, Simone, und Sigrid Roßteutscher. 2011. Jugend und Politik: Verliert die Demokratie ihren Nachwuchs? In *Der unbekannte Wähler? Mythen und Fakten über das Wahlverhalten der Deutschen*, Hrsg. Evelyn Bytzek und Sigrid Roßteutscher, 59-80. Frankfurt: Campus.
- Abendschön, Simone, und Markus Tausendpfund. 2017. Political Knowledge of Children and the Role of Sociostructural Factors. *American Behavioral Scientist* 61 (2): 204-221.
- Backhaus, Klaus, Bernd Erichson, Wulff Plinke, und Rolf Weiber. 2018. *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Berlin: Springer Gabler.
- Baur, Nina. 2011. Das Ordinalskalenproblem. In *Datenanalyse mit SPSS für Fortgeschrittene 1*, Hrsg. Leila Akremi, Nina Baur und Sabine Fromm, 211-221. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Becker, Rolf. 2017. *Lehrbuch der Bildungssoziologie*. Wiesbaden: Springer VS.
- Benninghaus, Hans. 2007. *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Best, Henning, und Christof Wolf. 2010. Logistische Regression. In *Handbuch der sozialwissenschaftlichen Datenanalyse*, Hrsg. Christof Wolf und Henning Best, 827-854. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blais, André. 2006. What affects voter turnout? *Annual Review of Political Science* 9: 111-125.
- Borg, Ingwer. 2003. *Führungsinstrument Mitarbeiterbefragung*. Göttingen: Hogrefe.
- Bortz, Jürgen, und Gustav A. Lienert. 2008. *Kurzgefasste Statistik für die Klinische Forschung. Leitfaden für die verteilungsfreie Analysen kleiner Stichproben*. Heidelberg: Springer.
- Bortz, Jürgen, und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Braunecker, Claus. 2016. *How to do Empirie, how to do SPSS. Eine Gebrauchsanleitung*. Stuttgart: UTB.
- Caballero, Claudio. 2014. Nichtwahl. In *Handbuch Wahlforschung*, Hrsg. Jürgen W. Falter und Harald Schoen, 437-488. Wiesbaden: Springer VS.
- Campbell, Angus, Gerald Gurin, und Warren E. Miller. 1954. *The Voter Decides*. Evanston/Illinois: Row, Peterson and Company.
- Cleveland, William S. 1994. *The Elements of Graphing Data*. New Jersey: Hobart Press.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.
- Cohen, Jacob. 1994. The Earth Is Round ($p < .05$). *American Psychologist* 49 (12): 997-1003.
- Cowles, Michael, und Caroline Davis. 1982. On the origins of the 05 level of statistical significance. *American Psychologist* 37 (5): 553-558.

- Degen, Horst. 2010. Graphische Datenexploration. In *Handbuch der sozialwissenschaftlichen Datenanalyse*, Hrsg. Christof Wolf und Henning Best, 91-116. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Diaz-Bone, Rainer. 2018. *Statistik für Soziologen*. Stuttgart: UTB.
- Dillman, Don A., Jolene D. Smyth, und Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys. The Tailored Design Method*. Hoboken: Wiley.
- Dollmann, Jörg. 2017. Ethnische Bildungsungleichheiten. In *Lehrbuch der Bildungssoziologie*, Hrsg. Rolf Becker, 487-510. Wiesbaden: Springer VS.
- Döring, Nicola, und Jürgen Bortz. 2016. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Dubben, Hans-Hermann, und Hans-Peter Beck-Bornholt. 2006. Die Bedeutung der statistischen Signifikanz. In *Methoden der Sozialforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie. Sonderheft 44/2004*, Hrsg. Andreas Diekmann, 61-74. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Faulbaum, Frank, Peter Prüfer, und Margrit Rexroth. 2009. *Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Fischer, Hans. 2011. *A History of the Central Limit Theorem. From Classical to Modern Probability Theory*. Springer: New York.
- Fisher, Ronald. 1925a. Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 22 (5): 700-725.
- Fisher, Ronald A. 1925b. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Frey, Bruno S. 2017. *Wirtschaftswissenschaftliche Glücksforschung. Kompakt – verständlich – anwendungsorientiert*. Wiesbaden: Springer Gabler.
- Fromm, Sabine. 2011. Neue Variablen berechnen. In *Datenanalyse mit SPSS für Fortgeschrittene 1*, Hrsg. Leila Akremi, Nina Baur und Sabine Fromm, 109-132. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Fromm, Sabine. 2012. *Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Fuchs, Marek. 2010. Herausforderungen der Umfrageforschung. In *Gesellschaftliche Entwicklungen im Spiegel der empirischen Sozialforschung*, Hrsg. Frank Faulbaum und Christof Wolf, 227-252. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gabriel, Oscar. 2012. Wahlen in der Demokratie. In *Wählerverhalten in der Demokratie. Eine Einführung*, Hrsg. Oscar Gabriel und Bettina Westle, 13-42. Stuttgart: UTB.
- Gautschi, Thomas. 2010. Maximum-Likelihood Schätztheorie. In *Handbuch der sozialwissenschaftlichen Datenanalyse*, Hrsg. Christof Wolf und Henning Best, 205-236. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Gehring, Uwe W., und Cornelia Weins. 2009. *Grundkurs Statistik für Politologen und Soziologen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52 (3): 647-674.
- Gill, Jeff. 2018. Comments from the New Editor. *Political Analysis* 26 (1): 1-2.
- Green, Donald P., und Susanne Baltes. 2017. Party Identification: Meaning and Measurement. In *The SAGE Handbook of Electoral Behaviour*, Hrsg. Kai Arzheimer, Jocelyn Evans und Michael S. Lewis-Beck, 287-312. London: Sage.
- Griffiths, Dawn. 2009. *Statistik von Kopf bis Fuß*. Köln: O'Reilly.
- Häder, Michael, und Sabine Häder. 2014. Stichprobenziehung in der quantitativen Sozialforschung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 283-297. Wiesbaden: Springer VS.
- Harrison, Chase H. 2005. Coverage Error. In *Polling America: An encyclopedia of public opinion*, Hrsg. Samuel J. Best und Benjamin Radcliff, 134-140. Westport: Greenwood.
- Hedderich, Jürgen, und Lothar Sachs. 2012. *Angewandte Statistik*. Heidelberg: Springer.
- Heeringa, Steven G., Brady T. West, und Patricia A. Berglund. 2017. *Applied Survey Data Analysis*. Boca Raton: CRC Press.
- Helliwell, John F., Richard Layard, und Jeffrey D. Sachs. 2018. *World Happiness Report 2018*. New York: Sustainable Development Solutions Network.
- Kanji, Gopal K. 2006. *100 Statistical Tests*. London: Sage.
- Kleinhenz, Thomas. 1995. *Die Nichtwähler. Ursachen der sinkenden Wahlbeteiligung in Deutschland*. Opladen: Westdeutscher Verlag.
- Kleining, Gerhard, und Harriett Moore. 1968. Soziale Selbsteinstufung (SSE). Ein Instrument zur Messung sozialer Schichten. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 20: 502-552.
- Kohler, Ulrich, und Frauke Kreuter. 2017. *Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. Berlin: de Gruyter.
- Koschack, Janka. 2008. Standardabweichung und Standardfehler: der kleine, aber feine Unterschied. *Zeitschrift für Allgemeinmedizin* 84 (6): 258-260.
- Kosfeld, Reinhold, Hans Friedrich Eckey, und Matthias Türck. 2016. *Deskriptive Statistik. Grundlagen – Methoden – Beispiele – Aufgaben*. Wiesbaden: Springer Gabler.
- Krämer, Walter. 2006. Statistik: Vom Geburtshelfer zum Bremser der Erkenntnis in den Sozialwissenschaften. In *Methoden der Sozialforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie. Sonderheft 44/2004.*, Hrsg. Andreas Diekmann, 51-60. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kromrey, Helmut, Jochen Roose, und Jörg Strübing. 2016. *Empirische Sozialforschung*. Stuttgart: UTB.

- Kuckartz, Udo, Stefan Rädiker, Thomas Ebert, und Julia Schehl. 2013. *Statistik. Eine verständliche Einführung*. Wiesbaden: Springer VS.
- Kühnel, Steffen-M., und Dagmar Krebs. 2007. *Statistik für die Sozialwissenschaften. Grundlagen. Methoden. Anwendungen*. Reinbek: Rowohlt.
- Lampert, Thomas, Matthias Richter, Sven Schneider, Jacob Spallek, und Nico Dragano. 2016. Soziale Ungleichheit und Gesundheit. Stand und Perspektiven der sozialepidemiologischen Forschung in Deutschland. *Bundesgesundheitsblatt* 59 (2): 152-165.
- Lee, Kristen Schultz, und Hiroshi Ono. 2012. Marriage, Cohabitation, and Happiness: A Cross-National Analysis of 27 Countries. *Journal of Marriage and Family* 74 (5): 953-972.
- Lück, Detlev, und Nina Baur. 2011. Wie kommen die Daten in den Datensatz? Arbeitsschritte vom Fragebogen zum fertigen Datensatz. In *Datenanalyse mit SPSS für Fortgeschrittene 1*, Hrsg. Leila Akremi, Nina Baur und Sabine Fromm, 22-58. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Lück, Detlev, und Uta Landrock. 2014. Datenaufbereitung und Datenbereinigung in der quantitativen Sozialforschung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 397-409. Wiesbaden: Springer VS.
- Ludwig-Mayerhofer, Wolfgang, Uta Liebeskind, und Ferdinand Geißler. 2014. *Statistik. Eine Einführung für Sozialwissenschaftler*. Weinheim: Beltz Juventa.
- Mittag, Hans-Joachim. 2017. *Statistik. Eine Einführung mit interaktiven Elementen*. Berlin: Springer Spektrum.
- Morrison, Denton E., und Ramon E. Henkel, Hrsg. 1970. *The significance test controversy*. Chicago: Aldine.
- Neller, Katja. 2004. Politik und Lebenszufriedenheit. In *Deutschland in Europa. Ergebnisse des European Social Survey 2002-2003*, Hrsg. Jan W. van Deth, 27-53. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Neyman, Jerzy, und Egon S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A* 231: 289-337.
- Noll, Heinz-Herbert. 1999. Subjektive Schichteinstufung. Aktuelle Befunde zu einer traditionellen Frage. In *Deutschland im Wandel. Sozialstrukturelle Analysen*, Hrsg. Wolfgang Glatzer und Ilona Ostner, 147-162. Opladen: Leske+Budrich.
- Norris, Pippa. 2004. *Electoral Engineering. Voting Rules and Political Behavior*. Cambridge: Cambridge University Press.
- Plümper, Thomas. 2012. *Effizient schreiben. Leitfaden zum Verfassen von Qualifizierungsarbeiten und wissenschaftlichen Texten*. München: Oldenbourg.
- Pollak, Reinhard. 2016. Schicht, soziale. In *Grundbegriffe der Soziologie*, Hrsg. Johannes Kopp und Anja Steinbach, 294-296. Wiesbaden: Springer VS.
- Porst, Rolf. 2014. *Fragebogen. Ein Arbeitsbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Richter, Matthias, und Klaus Hurrelmann. 2007. Warum die gesellschaftlichen Verhältnisse krank machen. *Aus Politik und Zeitgeschichte* (42): 3-10.
- Rumsey, Deborah. 2010. *Statistik für Dummies*. Weinheim: Wiley.
- Schendera, Christian FG. 2014. *Regressionsanalyse mit SPSS*. München: Oldenbourg.
- Schendera, Christian FG. 2015. *Deskriptive Statistik verstehen*. Konstanz: UVK.
- Scheuch, Erwin K., und Hans-Jürgen Daheim. 1961. Sozialprestige und soziale Schichtung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 5*: 65-103.
- Schmitt, Annette. 2014. Die Rolle von Wahlen in der Demokratie. In *Handbuch Wahlforschung*, Hrsg. Jürgen W. Falter und Harald Schoen, 3-35. Wiesbaden: Springer VS.
- Schmitt, Hermann, und Eftichia Teperoglou. 2017. The Study of Less Important Elections. In *The SAGE Handbook of Electoral Behaviour*, Hrsg. Kai Arzheimer, Jocelyn Evans und Michael S. Lewis-Beck, 56-79. London: Sage.
- Schnell, Rainer. 1994. *Graphisch gestützte Datenanalyse*. München: Oldenbourg.
- Schoen, Harald, und Cornelia Weins. 2014. Der sozialpsychologische Ansatz zur Erklärung von Wahlverhalten. In *Handbuch Wahlforschung*, Hrsg. Jürgen W. Falter und Harald Schoen, 241-329. Wiesbaden: Springer VS.
- Sedlmeier, Peter, und Frank Renkewitz. 2013. *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson.
- Smets, Kaat, und Carolien van Ham. 2013. The embarrassment of riches? A meta-analysis of individual-level research on voter turnout. *Electoral Studies* 32 (2): 344-359.
- Stanat, Petra. 2003. Schulleistungen von Jugendlichen mit Migrationshintergrund: Differenzierung deskriptiver Befunde aus PISA und PISA-E. In *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*, Hrsg. Deutsches PISA-Konsortium, 243-260. Opladen: Leske+Budrich.
- Stein, Petra. 2014. Forschungsdesigns für die quantitative Sozialforschung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 135-151. Wiesbaden: Springer VS.
- Stevens, S. S. 1946. On the Theory of Scales of Measurement. *Science* 103 (2684): 677-680.
- Student. 1908. The Probable Error of a Mean. *Biometrika* 6 (1): 1-25.
- Tachtsoglou, Sarantis, und Johannes König. 2017. *Statistik für Erziehungswissenschaftlerinnen und Erziehungswissenschaftler. Konzepte, Beispiele und Anwendungen in SPSS und R*. Wiesbaden: Springer VS.
- Tausendpfund, Markus. 2018a. *Quantitative Datenanalyse. Eine Einführung mit SPSS*. Hagen: FernUniversität in Hagen (Kurs: 33205).
- Tausendpfund, Markus. 2018b. *Quantitative Methoden in der Politikwissenschaft. Eine Einführung*. Wiesbaden: Springer VS.

- Tukey, John W. 1977. *Exploratory Data Analysis*. Reading: Addison Wesley.
- Urban, Dieter, und Jochen Mayerl. 2018. *Angewandte Regressionsanalyse: Theorie, Technik und Praxis*. Wiesbaden: Springer VS.
- Valliant, Richard, und Jill A. Dever. 2018. *Survey Weights. A Step-by-Step Guide to Calculation*. College Station: Stata Press.
- Valliant, Richard, Jill A. Dever, und Frauke Kreuter. 2013. *Practical Guide to Designing and Weighting Survey Samples*. New York: Springer.
- van Deth, Jan W. 1990. Interest in Politics. In *Continuities in Political Action. A Longitudinal Study of Political Orientations in Three Western Democracies*, Hrsg. M. Kent Jennings und Jan W. van Deth et al., 275-312. Berlin: de Gruyter.
- van Deth, Jan W. 2004. Politisches Interesse. In *Deutschland in Europa. Ergebnisse des European Social Survey 2002-2003*, Hrsg. Jan W. van Deth, 275-292. Wiesbaden: VS Verlag für Sozialwissenschaften.
- van Deth, Jan W. 2013. Politisches Interesse. In *Politik im Kontext: Ist alle Politik lokale Politik? Individuelle und kontextuelle Determinanten politischer Orientierungen*, Hrsg. Jan W. van Deth und Markus Tausendpfund, 271-296. Wiesbaden: Springer VS.
- Verba, Sidney, Kay Lehman Schlozman, und Henry E. Brady. 1995. *Voice and Equality. Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.
- Völkl, Kerstin, und Christoph Korb. 2018. *Deskriptive Statistik. Eine Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler*. Wiesbaden: Springer VS.
- Warwick, Paul V. 2002. Toward a Common Dimensionality in West European Policy Spaces. *Party Politics* 8 (1): 101-122.
- Wasmer, Martina, Michael Blohm, Jessica Walter, Regina Jutz, und Evi Scholz. 2017. *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) 2014. GESIS Papers 2017/20*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften.
- Wass, Hanna, und André Blais. 2017. Turnout. In *The SAGE Handbook of Electoral Behaviour*, Hrsg. Kai Arzheimer, Jocelyn Evans und Michael S. Lewis-Beck, 459-487. London: Sage.
- Wasserstein, Ronald L., und Nicole A. Lazar. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70 (2): 129-133.
- Weick, Stefan. 2012. Persönliches und soziales Wohlbefinden. In *Deutschlands Metamorphosen. Ergebnisse des European Social Survey 2002 bis 2008*, Hrsg. Silke I. Keil und Jan W. van Deth, 391-425. Baden-Baden: Nomos.
- Weins, Cornelia. 2010. Uni- und bivariate deskriptive Statistik. In *Handbuch der sozialwissenschaftlichen Datenanalyse*, Hrsg. Christof Wolf und Henning Best, 65-90. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Weiß, Christel. 2013. *Basiswissen Medizinische Statistik*. Heidelberg: Springer.

- Westle, Bettina. 2009. Die unpolitische Frau – ein Methodenartefakt der Umfrageforschung? In *Politik – Wissenschaft – Medien. Festschrift für Jürgen W. Falter zum 65. Geburtstag*, Hrsg. Hanna Kaspar, Harald Schoen, Siegfried Schumann und Jürgen R. Winkler, 179-201. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Westle, Bettina. 2018. Grundlagen der quantitativen Datenanalyse. In *Methoden der Politikwissenschaft*, Hrsg. Bettina Westle, 324-339. Baden-Baden: Nomos.
- Westle, Bettina, und Harald Schoen. 2002. Ein neues Argument in einer alten Diskussion: ‚Politikverdrossenheit‘ als Ursache des *gender gap* im politischen Interesse? In *Das Ende der politisierten Sozialstruktur?*, Hrsg. Frank Brettschneider, Jan van Deth und Edeltraud Roller, 215-244. Opladen: Leske+Budrich.
- Wolf, Christof, und Henning Best. 2015. Linear regression. In *The SAGE Handbook of Regression Analysis and Causal Inference*, Hrsg. Henning Best und Christof Wolf, 57-81. Los Angeles: SAGE.
- Yerkes, Robert M., und John D. Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology* 18 (5): 459-482.