

Markus Tausendpfund

Quantitative Datenanalyse. Eine Einführung mit R

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

Vorwort

Die vorliegende Lerneinheit behandelt die sozialwissenschaftliche Datenanalyse, R und insbesondere die sozialwissenschaftliche Datenanalyse mit R. Dabei werden Kenntnisse vermittelt, um einfache Analysen selbstständig mit der Software R durchführen zu können.

Der Text entspricht dabei weniger einem klassischen Lehrbuch, sondern eher einem Begleitkurs für die Auseinandersetzung mit dem Programm R bzw. RStudio. Die Lerneinheit soll das Interesse an sozialwissenschaftlichen Fragestellungen wecken sowie die Möglichkeiten und Grenzen der quantitativen Datenanalyse aufzeigen.

Lehrmaterial, das in erster Linie zum Selbststudium angelegt ist, profitiert insbesondere durch Rückmeldungen der Leserinnen und Leser. Deshalb möchte ich mich herzlich bei allen Personen bedanken, die mich auf Fehler und Verbesserungsmöglichkeiten hingewiesen haben.

Aktuelle Ergänzungen und mögliche Korrekturen zu dieser Lerneinheit finden Sie in der Moodle-Lernumgebung des Moduls M1 „Quantitative Methoden der Sozialwissenschaften“ im BA-Studiengang „Politikwissenschaft, Verwaltungswissenschaft und Soziologie“. Dort werden auch Übungsaufgaben und Videos veröffentlicht, die die Auseinandersetzung mit den Inhalten dieser Lerneinheit fördern sollen.

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

Die vorliegende Lerneinheit ist kein „Endprodukt“. Die regelmäßige Aktualisierung stellt eine Daueraufgabe dar. Deshalb freue ich mich sehr über alle Hinweise und Anregungen zur weiteren Verbesserung der Lerneinheit (E-Mail: Markus.Tausendpfund@fernuni-hagen.de).

Hagen, im Dezember 2025

Markus Tausendpfund

Inhaltsverzeichnis

Abbildungsverzeichnis	VIII
Tabellenverzeichnis	IX
1 Einführung	10
1.1 Sozialwissenschaftlicher Forschungsprozess	10
1.2 Quantitative Datenanalyse	12
1.3 Warum R?	13
1.4 Struktur der Lerneinheit	15
2 R und RStudio kennenlernen	16
2.1 Installation	16
2.1.1 R	16
2.1.2 RStudio	17
2.1.3 Pakete	21
2.1.4 Aktualisierungen	21
2.1.5 posit Cloud: eine Alternative zur lokalen R-Installation	21
2.2 Ein erster Überblick	22
2.2.1 Console	22
2.2.2 Skripte	24
2.2.3 Befehle	25
2.2.4 Objekte	26
2.2.5 Vektoren	27
2.2.6 Tabellen	29
2.2.7 Erste Analysen	30
2.2.8 Fehlende Werte	32
2.2.9 Hilfe	33
3 Arbeiten mit R	34
3.1 Pakete installieren und laden	34
3.2 Daten laden	36
3.3 Daten importieren	37
3.3.1 Excel	38
3.3.2 SPSS	39
3.3.3 CSV	39
3.4 Objekttypen	41
3.5 Datenstrukturen	44

3.6	Saubere Skripte erstellen.....	45
3.7	Projekte in R.....	46
4	Datenmanagement und Datenaufbereitung.....	47
4.1	Daten und Pakete.....	47
4.2	Beispieldatensatz kennenlernen.....	47
4.3	Pipe-Operator.....	48
4.4	Datenmanagement mit dplyr.....	49
4.4.1	Variablen auswählen.....	50
4.4.2	Variablen umbenennen.....	50
4.4.3	Variablen filtern.....	51
4.4.4	Variablen erstellen.....	52
4.4.5	Einfache Berechnungen.....	52
4.4.6	Weitere Optionen.....	53
4.5	Datenaufbereitung mit sjmisc.....	53
4.5.1	Variablenwerte ändern.....	53
4.5.2	Variablen zusammenfassen.....	56
4.5.3	Werte zählen.....	58
4.5.4	Weitere Optionen.....	58
4.6	Labels konvertieren.....	59
5	Univariate Datenanalyse.....	61
5.1	Daten und Pakete.....	61
5.2	Häufigkeitstabellen.....	61
5.3	Lagemaße.....	63
5.4	Streuungsmaße.....	64
5.5	Formmaße.....	65
5.6	Kompakte Übersichten.....	67
6	Bivariate Datenanalyse.....	68
6.1	Daten und Pakete.....	68
6.2	Univariate Statistiken nach Gruppen.....	68
6.3	Kreuztabellen.....	69
6.4	Zusammenhangsmaße.....	72
6.4.1	Nominalskalierte Merkmale.....	73
6.4.2	Ordinalskalierte Merkmale.....	76
6.4.3	Metrische Merkmale.....	78

7	Grafiken.....	81
7.1	Einführung.....	81
7.2	Ausgewählte Diagramme.....	83
7.2.1	Säulen- und Balkendiagramm	83
7.2.2	Kreisdiagramm	85
7.2.3	Histogramm	86
7.2.4	Boxplot	87
7.2.5	Streudiagramm.....	89
7.2.6	Liniendiagramm.....	91
7.3	Weitere Pakete.....	92
8	Multivariate Datenanalyse	93
8.1	Einführung.....	93
8.2	Lineare Regression.....	95
8.2.1	Das Grundmodell	95
8.2.2	Daten und Pakete.....	103
8.2.3	Lineare Regression mit R.....	103
8.2.4	Interpretation der Ergebnisse	106
8.2.5	Weitere Möglichkeiten.....	109
8.2.6	Anwendungsvoraussetzungen	115
8.2.7	Praktische Hinweise	116
8.3	Logistische Regression	118
8.3.1	Das Grundmodell	118
8.3.2	Daten und Pakete.....	124
8.3.3	Logistische Regression mit R.....	125
8.3.4	Interpretation der Ergebnisse	126
8.3.5	Weitere Möglichkeiten.....	129
8.3.6	Praktische Hinweise	134
9	Inferenzstatistik	136
9.1	Daten und Pakete.....	136
9.2	Konfidenzintervalle bei Mittelwerten	137
9.3	Konfidenzintervalle bei Regressionen.....	138
9.4	Mittelwertvergleiche (t-Test)	141
10	Arbeiten mit ausgewählten Sekundärdaten	147
10.1	European Social Survey.....	147

10.2 Allgemeine Bevölkerungsumfrage der Sozialwissenschaften	149
Pakete im Überblick.....	152
Beispieldatensatz im Überblick	154
Literaturverzeichnis.....	158

Abbildungsverzeichnis

Abbildung 1: Phasen eines quantitativen Forschungsprojekts	11
Abbildung 2: The Comprehensive R Archive Network (CRAN)	16
Abbildung 3: Startbildschirm von R	17
Abbildung 4: RStudio mit drei Fenstern	18
Abbildung 5: RStudio mit vier Fenstern	19
Abbildung 6: Global Options bei RStudio	20
Abbildung 7: posit Cloud	22
Abbildung 8: Streudiagramm des Einkommens in Abhängigkeit des Alters	32
Abbildung 9: Installierte Pakete	35
Abbildung 10: „Import Dataset“-Funktion in RStudio	38
Abbildung 11: Excel-Datensatz importieren	38
Abbildung 12: SPSS-Datensatz importieren	39
Abbildung 13: Normalverteilung	66
Abbildung 14: Säulen- und Balkendiagramm	84
Abbildung 15: Kreisdiagramm	85
Abbildung 16: Histogramm	86
Abbildung 17: Elemente eines Boxplots	87
Abbildung 18: Boxplots	88
Abbildung 19: Plotsymbole bei R	90
Abbildung 20: Streudiagramm	90
Abbildung 21: Liniendiagramm	91
Abbildung 22: Auswahl von regressionsanalytischen Verfahren	93
Abbildung 23: Streudiagramm von Alter und Einkommen	96
Abbildung 24: Streudiagramm von Einkommen und Alter mit Regressionsgerade	97
Abbildung 25: Empirischer und geschätzter Wert einer linearen Regression	98
Abbildung 26: Grafische Darstellung einer multiplen Regression	102
Abbildung 27: Lineare Regressionsfunktion	120
Abbildung 28: Logistische Regressionsfunktion	121
Abbildung 29: Logistische Regression (Beispieldaten)	124
Abbildung 30: Grafische Darstellung der logistischen Regression I	133
Abbildung 31: Grafische Darstellung der logistischen Regression II	134
Abbildung 32: Koeffizientenplot der Determinanten der Lebenszufriedenheit	141
Abbildung 33: Aussage über die Grundgesamtheit auf Basis einer Zufallsstichprobe	142

Tabellenverzeichnis

Tabelle 1: SPSS, Stata und R im Überblick.....	14
Tabelle 2: Mathematische Funktionen in R	23
Tabelle 3: Logische Abfragen in R	23
Tabelle 4: Installieren und Laden von Paketen.....	35
Tabelle 5: Beispieldatensatz (peanuts_r)	36
Tabelle 6: Ausgewählte Funktionen zum Testen und Konvertieren von Objekten	43
Tabelle 7: Objekttypen in den Datensätzen.....	43
Tabelle 8: Ausgewählte Funktionen von dplyr.....	49
Tabelle 9: Auswahl von Filter-Möglichkeiten.....	51
Tabelle 10: Ausgewählte Möglichkeiten des rec-Arguments.....	55
Tabelle 11: Interpretation von Schiefe und Wölbung	67
Tabelle 12: Kreuztabelle zwischen Wahlbeteiligung und Bildung.....	71
Tabelle 13: Zusammenhangsmaße bei der bivariaten Datenanalyse	73
Tabelle 14: Arbeitstabelle für die Berechnung von Chi-Quadrat	74
Tabelle 15: Interpretation von Cramer's V	76
Tabelle 16: Interpretation von Spearman's Rho.....	78
Tabelle 17: Interpretation des Korrelationskoeffizienten nach Pearson.....	80
Tabelle 18: Argumente bei Grafikbefehlen	81
Tabelle 19: Pakete zur Erstellung von Grafiken	92
Tabelle 20: Beispieldaten für Alter und Einkommen	95
Tabelle 21: Bivariate Regression (Beispieldaten)	100
Tabelle 22: Beispieldaten für Alter, Einkommen und Berufserfahrung.....	101
Tabelle 23: Multiple Regression (Beispieldaten).....	102
Tabelle 24: Bestimmungsfaktoren der Lebenszufriedenheit I	104
Tabelle 25: Bestimmungsfaktoren der Lebenszufriedenheit II	105
Tabelle 26: Bestimmungsfaktoren der Lebenszufriedenheit III.....	106
Tabelle 27: Bestimmungsfaktoren der Lebenszufriedenheit IV	110
Tabelle 28: Bestimmungsfaktoren der Lebenszufriedenheit V (standardisierte Koeffizienten)..	111
Tabelle 29: Bestimmungsfaktoren der Lebenszufriedenheit V (mit Faktoren).....	113
Tabelle 30: Bestimmungsfaktoren der Lebenszufriedenheit VI (Regressionstabelle mit sjPlot)..	114
Tabelle 31: Informationen zu einer linearen Regression.....	115
Tabelle 32: Beispieldaten für Rauchen und Alter in Jahren	119
Tabelle 33: Bestimmungsfaktoren des Tabakkonsums.....	123
Tabelle 34: Bestimmungsfaktoren der Wahlbeteiligung	126
Tabelle 35: Bedeutung der Asteriske	128
Tabelle 36: Logistische Regressionskoeffizienten und Odds Ratio im Vergleich.....	130
Tabelle 37: Determinanten der Lebenszufriedenheit (OLS-Regression)	140
Tabelle 38: t-Test der Lebenszufriedenheit in Abhängigkeit des Geschlechts.....	144
Tabelle 39: t-Test der Demokratiezufriedenheit in Abhängigkeit des politischen Interesses	145
Tabelle 40: Argumente beim t-Test.....	146
Tabelle 41: Pakete im Überblick	152
Tabelle 42: Beispieldatensatz im Überblick	154

1 Einführung

Vorschau



Die quantitative Datenanalyse ist die Phase im sozialwissenschaftlichen Forschungsprozess, in der die theoretisch entwickelten Hypothesen empirisch geprüft werden. In diesem Kapitel werden die einzelnen Phasen des Forschungsprozesses knapp skizziert, die Bedeutung der Methodenkompetenz für die Auseinandersetzung mit empirischen Studien dargelegt und Statistikprogramme vorgestellt. Die Einführung schließt mit einem Ausblick auf die weiteren Kapitel dieser Lerneinheit und verweist auf ergänzende Materialien in der Moodle-Lernumgebung.

1.1 Sozialwissenschaftlicher Forschungsprozess

In einem quantitativen Forschungsprojekt lassen sich idealtypisch mehrere Phasen unterscheiden (siehe Abbildung 1). Nach der Entscheidung für ein Forschungsthema und der Entwicklung einer Forschungsfrage (1) müssen zunächst die zentralen Konzepte der Forschungsfrage identifiziert und theoretisch geklärt werden (2). Auf dieser Grundlage können Hypothesen formuliert (3) und Operationalisierungen der Konzepte (4) entwickelt werden (ausführlicher Tausendpfund 2018).

Bedeutung des Forschungsstands

Diese Phasen eines Forschungsprojekts erfolgen in intensiver Auseinandersetzung mit der existierenden Fachliteratur. Nur wer den Forschungsstand zu seinem Forschungsthema kennt, kann eine gehaltvolle 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 existierender Operationalisierungen notwendig.

Bei einer Primäranalyse werden neue Daten erhoben, um die Forschungsfrage zu untersuchen. Bei einer Sekundäranalyse werden existierende Daten genutzt, um die Forschungsfrage zu bearbeiten (5). Falls für die Bearbeitung einer Forschungsfrage bereits geeignetes Datenmaterial existiert, dann können die Phasen Forschungsdesign (6), Auswahlverfahren (7) und Datenerhebung (8) „übersprungen“ werden.

Die Datenaufbereitung und -analyse stellt eine Phase in einem sozialwissenschaftlichen Forschungsprojekt dar (9). In dieser Phase werden die theoretisch formulierten Hypothesen empirisch geprüft. Mittlerweile existieren zahlreiche Verfahren der Datenanalyse (für einen Überblick siehe z.B. Wolf und Best 2010a; Backhaus et al. 2021); alle Verfahren setzen jedoch eine vorherige intensive Auseinandersetzung mit dem jeweiligen Forschungsstand voraus. Mit anderen Worten: Die Datenanalyse kann die vorherige Auseinandersetzung mit dem Forschungsstand nicht ersetzen.

In Publikationen (10) werden die Forschungsergebnisse der Öffentlichkeit zugänglich gemacht.