
Vorwort

Schriftlichkeit ist nichts anderes als eine Gedächtnishilfe, die nicht von sich allein Weisheit vermitteln kann.
Sokrates

Die Statistik ist ein enorm umfangreiches Wissensgebiet. Jede Wissenschaft entwickelt nicht nur eigene statistische Verfahren, sondern es sind auch viele Kenntnisse aus verschiedenen Wissensgebieten erforderlich, um diese Verfahren, ihre Anwendung und ihre Ergebnisse zu verstehen. So könnten bei einer Datenerhebung per Fragebogen oder Interview psychologische Kenntnisse benötigt werden. Für die Datenverarbeitung und Berechnung der statistischen Verfahren werden in der Regel mathematische und informatische Kenntnisse eingesetzt. Die Interpretation der Ergebnisse setzt wiederum sehr gute Fachkenntnisse des entsprechenden Wissensgebiets voraus. Von daher ist es schwierig, in einer Darstellung alle Aspekte gleichermaßen zu berücksichtigen. Ich habe mich bemüht, die Eigenschaften und Zusammenhänge statistischer Verfahren aufzuzeigen und ihre Anwendung durch zahlreiche Beispiele und Abbildungen verständlich zu machen.

Die Auswahl der statistischen Methoden von der Erhebung über beschreibende Techniken (Teil I) hin zur schließenden Analyse (Teil II) einschließlich einiger multivariater Verfahren (Teil III) erfolgte unter wirtschaftswissenschaftlicher Sicht. In einem Buch zur Einführung in die Statistik können außerdem nur die grundlegenden statistischen Verfahren berücksichtigt werden. Aber auch unter diesen ist nur eine Auswahl erfolgt. Dennoch sind über 500 Seiten zustandegekommen und es gäbe zu dieser Auswahl noch mehr zu schreiben!

Während der langen Zeit an der Arbeit an diesem Buch ist es mir immer wieder schwergefallen, bestimmte Themen und Aspekte nicht mehr aufzunehmen. Erst die Zeit wird zeigen, ob ich an den richtigen Stellen die Darstellung beendet habe.

Ich bedanke mich sehr herzlich bei Prof. Dr. Peter Naeve von der Universität Bielefeld für seine große Diskussionsbereitschaft und hilfreiche Kritik.

Bielefeld, im Mai 2004

Wolfgang Kohn

Inhaltsverzeichnis

1	Einleitung	1
----------	-------------------	----------

Teil I Deskriptive Statistik

2	Grundlagen	7
2.1	Einführung	7
2.2	Grundbegriffe der deskriptiven Statistik	8
2.3	Messbarkeitseigenschaften von Merkmalen	13
2.3.1	Nominalskala	13
2.3.2	Ordinalskala	13
2.3.3	Kardinalskala	14
2.4	Skalentransformation	15
2.5	Häufigkeitsfunktion	16
2.6	Klassierung von metrischen Merkmalswerten	18
2.7	Übungen	21
3	Datenerhebung und Erhebungsarten	23
3.1	Datenerhebung	23
3.1.1	Fragebogenerhebung	24
3.1.2	Interview	24
3.1.3	Gestaltung von Befragungen	25
3.2	Erhebungsarten	27
3.2.1	Zufallsstichproben	28
3.2.2	Nicht zufällige Stichproben	30
3.3	Datenschutz	32
4	Eindimensionale Datenanalyse	35
4.1	Einführung	36
4.2	Qualitative Merkmale	36
4.2.1	Modus	38

4.2.2	Informationsentropie	38
4.3	Komparative Merkmale	45
4.3.1	Empirische Verteilungsfunktion	45
4.3.2	Quantile	47
4.3.3	Boxplot	49
4.3.4	Summenhäufigkeitsentropie	51
4.4	Quantitative Merkmale	54
4.4.1	Stamm-Blatt Diagramm	54
4.4.2	Histogramm	57
4.4.3	Dichtespur	59
4.4.4	Arithmetisches Mittel	65
4.4.5	Harmonisches Mittel	67
4.4.6	Geometrisches Mittel	70
4.4.7	Spannweite	72
4.4.8	Median der absoluten Abweichung vom Median	73
4.4.9	Varianz und Standardabweichung	74
4.4.10	Variationskoeffizient	80
4.4.11	Relative Konzentrationsmessung	83
4.4.12	Absolute Konzentrationsmessung	92
4.5	Übungen	98
5	Zweidimensionale Datenanalyse	101
5.1	Zweidimensionale Daten und ihre Darstellung	101
5.2	Randverteilungen und bedingte Verteilungen	105
5.2.1	Randverteilung	105
5.2.2	Bedingte Verteilung	106
5.3	Zusammenhangsmaße qualitativer Merkmale	113
5.3.1	Quadratische Kontingenz	113
5.3.2	Informationsentropie bei zweidimensionalen Häufigkeitsverteilungen	116
5.3.3	Transinformation	118
5.3.4	Vergleich von C und T	119
5.4	Zusammenhangsmaße komparativer Merkmale	121
5.4.1	Kovarianz und Korrelationskoeffizient	121
5.4.2	Rangfolge und Rangzahlen	122
5.4.3	Rangkorrelationskoeffizient	123
5.5	Zusammenhangsmaße quantitativer Merkmale	126
5.5.1	Kovarianz	126
5.5.2	Korrelationskoeffizient	127
5.6	Interpretation von Korrelation	129
5.7	Simpson Paradoxon	131
5.8	Übungen	133

6	Lineare Regression	137
6.1	Einführung	137
6.2	Lineare Regressionsfunktion	138
6.3	Methode der Kleinsten Quadrate	141
6.3.1	Normalgleichungen	142
6.3.2	Kleinst-Quadrate Schätzung	143
6.3.3	Standardisierte Regressionskoeffizienten	146
6.4	Bestimmtheitsmaß	147
6.5	Spezielle Regressionsfunktionen	150
6.5.1	Trendfunktion	150
6.5.2	Linearisierung von Funktionen	150
6.5.3	Datentransformation	151
6.6	Lineares Modell	152
6.7	Prognose	153
6.8	Übungen	159
7	Verhältnis- und Indexzahlen	161
7.1	Einführung	161
7.2	Gliederungs-, Beziehungs- und Messzahlen	163
7.3	Umbasierung und Verkettung von Messzahlen	166
7.3.1	Umbasierung	166
7.3.2	Verkettung	167
7.4	Indexzahlen	167
7.4.1	Preisindex nach Laspeyres	169
7.4.2	Basiseffekt	172
7.4.3	Preisindex nach Paasche	173
7.4.4	Mengenindizes nach Laspeyres und Paasche	174
7.4.5	Umsatzindex	176
7.4.6	Deflationierung	176
7.4.7	Verkettung von Indexzahlen	177
7.4.8	Anforderungen an einen idealen Index	179
7.4.9	Preisindex nach Fischer	179
7.4.10	Kettenindex	180
7.5	Aktienindex DAX	181
7.6	Übungen	183

Teil II Schließende Statistik

8	Kombinatorik	187
8.1	Grundbegriffe	187
8.2	Permutation	189
8.2.1	Permutation ohne Wiederholung	189
8.2.2	Permutation mit Wiederholung	189
8.3	Variation	190

8.3.1	Variation ohne Wiederholung	191
8.3.2	Variation mit Wiederholung	192
8.4	Kombination	192
8.4.1	Kombination ohne Wiederholung	192
8.4.2	Kombination mit Wiederholung	193
8.5	Übungen	195
9	Grundzüge der Wahrscheinlichkeitsrechnung	197
9.1	Einführung	197
9.2	Zufallsexperiment	198
9.3	Ereignisoperationen	199
9.4	Wahrscheinlichkeitsbegriffe	204
9.4.1	Laplacescher Wahrscheinlichkeitsbegriff	205
9.4.2	Von Misesscher Wahrscheinlichkeitsbegriff	206
9.4.3	Subjektiver Wahrscheinlichkeitsbegriff	206
9.4.4	Axiomatische Definition der Wahrscheinlichkeit	207
9.4.5	Rechenregeln	209
9.5	Bedingte Wahrscheinlichkeiten	210
9.6	Satz von Bayes	216
9.7	Unabhängige Zufallsereignisse	220
9.8	Übungen	224
10	Zufallsvariablen und Wahrscheinlichkeitsfunktion	227
10.1	Zufallsvariablen	227
10.2	Wahrscheinlichkeitsfunktion	230
10.2.1	Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen	230
10.2.2	Wahrscheinlichkeitsfunktion einer stetigen Zufallsvariablen	232
10.3	Verteilungen von transformierten Zufallsvariablen	238
10.4	Erwartungswert	240
10.5	Modus und Quantil	244
10.6	Varianz	244
10.7	Erwartungswert und Varianz linear transformierter Zufallsvariablen	246
10.7.1	Erwartungswert linear transformierter Zufallsvariablen	246
10.7.2	Varianz linear transformierter Zufallsvariablen	247
10.8	Momente und momenterzeugende Funktion	249
10.8.1	Momente	249
10.8.2	Momenterzeugende Funktion	250
10.9	Ungleichung von Chebyschew	251
10.10	Übungen	254
11	Gemeinsame Verteilung von Zufallsvariablen	257
11.1	Einführung	257
11.2	Bedingte Verteilung	261
11.3	Bedingter Erwartungswert	262
11.4	Kovarianz	264

11.5	Übungen	268
12	Normalverteilung	271
12.1	Einführung	271
12.2	Standardnormalverteilung	276
12.3	Lognormalverteilung	281
12.4	Übungen	284
13	Bernoulli-verwandte Zufallsvariablen	285
13.1	Einführung	285
13.2	Bernoulliverteilung	285
13.3	Binomialverteilung	286
13.4	Hypergeometrische Verteilung	292
13.5	Geometrische Verteilung	297
13.6	Negative Binominalverteilung	300
13.7	Negative hypergeometrische Verteilung	303
13.8	Poissonverteilung	304
13.9	Exponentialverteilung	311
13.10	Übungen	314
14	Stichproben	315
14.1	Einführung	315
14.2	Identisch verteilte unabhängige Stichproben	316
14.3	Schwaches Gesetz der großen Zahlen	317
14.4	Starkes Gesetz der großen Zahlen	320
14.5	Hauptsatz der Statistik	321
14.6	Zentraler Grenzwertsatz	322
14.6.1	Approximation der Binomialverteilung durch die Normalverteilung	326
14.6.2	Approximation der hypergeometrischen Verteilung durch die Normalverteilung	329
14.6.3	Approximation der Poissonverteilung durch die Normalverteilung	329
14.7	Stichprobenverteilungen aus normalverteilten Grundgesamtheiten ..	330
14.7.1	χ^2 -Quadrat Verteilung	330
14.7.2	t -Verteilung	332
14.7.3	F -Verteilung	334
14.8	Hauptsatz der Stichprobentheorie	335
14.9	Stichproben aus bernoulli-, exponential- und poissonverteilten Grundgesamtheiten	338
14.9.1	Stichproben aus bernoulli- und binomialverteilten Grundgesamtheiten	338
14.9.2	Stichproben aus poissonverteilten Grundgesamtheiten	339
14.9.3	Stichproben aus exponentialverteilten Grundgesamtheiten ..	339
14.10	Übungen	341

15 Parameterschätzung	343
15.1 Einführung	343
15.2 Punktschätzung	344
15.2.1 Methode der Momente	345
15.2.2 Maximum-Likelihood Schätzung	346
15.2.3 Methode der Kleinsten Quadrate	349
15.3 Intervallschätzung	351
15.3.1 Konfidenzintervall für μ_X einer Normalverteilung bei bekanntem σ_X^2	352
15.3.2 Konfidenzintervall für μ_X einer Normalverteilung bei unbekanntem σ_X^2	353
15.3.3 Konfidenzintervall für Regressionskoeffizienten	354
15.3.4 Konfidenzintervall für eine ex post Prognose	357
15.3.5 Approximatives Konfidenzintervall für μ_X	358
15.3.6 Approximatives Konfidenzintervall für den Anteilswert θ	358
15.3.7 Konfidenzintervall für σ_X^2 bei normalverteilter Grundgesamtheit	358
15.4 Eigenschaften von Schätzstatistiken	359
15.4.1 Erwartungstreue	359
15.4.2 Mittlerer quadratischer Fehler	363
15.4.3 Konsistenz	367
15.4.4 Effizienz	369
15.5 Übungen	372
16 Statistische Tests	375
16.1 Einführung	375
16.2 Klassische Testtheorie	376
16.3 Parametertests bei normalverteilten Grundgesamtheiten	380
16.3.1 Gauss-Test	380
16.3.2 t -Test	383
16.3.3 Parametertest im linearen Regressionsmodell	384
16.4 Binomialtest	388
16.5 Testentscheidung	391
16.6 Gütefunktion	397
16.7 Operationscharakteristik	403
16.8 Test auf Gleichheit von zwei Mittelwerten	409
16.8.1 Vergleich zweier unabhängiger Stichproben	410
16.8.2 Vergleich von zwei verbundenen Stichproben	418
16.8.3 Unterschied zwischen verbundenen und unabhängigen Stichproben	420
16.9 Übungen	421

17 Statistische Tests für kategoriale Merkmale	423
17.1 Einführung	423
17.2 χ^2 -Anpassungstest	423
17.3 χ^2 -Homogenitätstest	433
17.4 χ^2 -Unabhängigkeitstest	439
17.5 Übungen	440

Teil III Einführung in die multivariaten Verfahren

18 Überblick über verschiedene multivariate Verfahren	445
18.1 Einführung	445
18.2 Asymmetrische Modelle	446
18.2.1 Regressionsanalyse	446
18.2.2 Varianzanalyse	447
18.2.3 Diskriminanzanalyse	447
18.3 Symmetrische Modelle	448
18.3.1 Kontingenzanalyse	448
18.3.2 Faktorenanalyse	449
18.3.3 Clusteranalyse	449
18.3.4 Multidimensionale Skalierung	450
18.3.5 Conjoint Analyse	451
19 Varianzanalyse	453
19.1 Einführung	453
19.2 Einfaktorielle Varianzanalyse	454
19.3 Zweifaktorielle Varianzanalyse	462
19.4 Andere Versuchspläne	474
19.5 Multivariate Varianzanalyse	475
20 Diskriminanzanalyse	481
20.1 Problemstellung der Diskriminanzanalyse	481
20.2 Klassische Diskriminanzanalyse nach Fisher	483
20.2.1 Klassifikationsregel	486
20.2.2 Spezialfall zwei Gruppen	487
20.2.3 Weitere Klassifikationsregeln	488
20.3 Überprüfung der Diskriminanzfunktion	489
20.4 Überprüfung der Merkmalsvariablen	490
20.5 Anwendungsbeispiel	491
21 Grundlagen der hierarchischen Clusteranalyse	503
21.1 Problemstellung der Clusteranalyse	504
21.2 Ähnlichkeits- und Distanzmaße	505
21.2.1 Nominalskalierte binäre Merkmale	506
21.2.2 Nominalskalierte mehrstufige Merkmale	510

21.2.3	Ordinalskalierte Merkmale	512
21.2.4	Quantitative Merkmale	513
21.2.5	Merkmale mit unterschiedlichem Skalenniveau	518
21.3	Hierarchische Clusteranalyse	519
21.4	Agglomerative Verfahren	519
21.4.1	Nearest Neighbour Verfahren	520
21.4.2	Furthest Neighbour Verfahren	523
21.4.3	Centroid Verfahren	525
21.4.4	Median Cluster Verfahren	527
21.4.5	Average Linkage Verfahren	528
21.4.6	Ward's Verfahren	533
21.4.7	Entropieanalyse	537
21.5	Fusionseigenschaften agglomerativer Verfahren	539
21.6	Divisive Verfahren	540
21.6.1	Ein polythetisches Verfahren	541
21.6.2	Ein monothetisches Verfahren	542
21.7	Probleme von Clusterverfahren	545
21.7.1	Definition von Clustern	545
21.7.2	Entscheidung über die Anzahl der Klassen	546
21.7.3	Beurteilung der Klassen	547
21.7.4	Anwendungsempfehlungen	549
21.8	Anmerkung	551
Lösungen zu den Übungen		553
A.1	Deskriptive Statistik	553
A.2	Induktive Statistik	570
Tabellen		595
B.1	Binomialverteilung	595
B.2	Poissonverteilung	600
B.3	Standardnormalverteilung	604
B.4	χ^2 -Verteilung	605
B.5	t -Verteilung	606
B.6	F -Verteilung	607
Literaturverzeichnis		611
Sachverzeichnis		617

Einleitung

Statistik hat ursprünglich mit der zahlenmäßigen Erfassung von Information zu tun, und zwar der eines Staats. Der Begriff leitet sich aus dem italienischen *statista* (Staatsmann) ab und wurde von dem Göttinger Staatenkundler Achenwall 1749 als neulateinisches Wort *statistica* im Zusammenhang mit der Verfassung eines Staates verwendet. Daraus resultiert auch, dass Volkswirtschaftslehre (ehemals auch Staatskunde genannt) und Statistik oft in einem Zusammenhang genannt werden. Daher spricht man auch von einer Statistik, wenn im Rahmen empirischer Fragestellungen Daten erhoben, dargestellt und analysiert werden (z. B. Volkszählung).

Die elementaren Methoden der beschreibenden oder deskriptiven Statistik reichen ebenso weit zurück, wie sich die Existenz von Erhebungsstatistiken nachweisen lässt. Etwa 3000 v. Chr. wurde in Ägypten der Verbrauch an Nahrungsmitteln beim Pyramidenbau festgehalten, etwa 2300 v. Chr. wurden in China die Bevölkerungs- und Besitzverhältnisse aufgelistet und etwa 400 v. Chr. wurde in Rom eine Volkszählung durchgeführt. Zu dieser historischen Herkunft der Statistik haben sich ab dem 17. Jahrhundert die Fragestellungen aus dem Glücksspiel gesellt. Viele Erkenntnisse sind aber erst in neuerer Zeit entwickelt worden. Heute wird dieser Bereich als induktive, mathematische Statistik oder Stochastik bezeichnet, wobei zwischen diesen Begriffen auch wieder inhaltliche Abgrenzungen existieren.

Allgemein lässt sich die Statistik als ein Instrument zur Informationsgewinnung und Informationsverdichtung beschreiben. Aus einer großen Zahl von Beobachtungen, wenn sie korrekt erhoben worden sind, können Informationen gewonnen werden, die – zumindest tendenziell – dann die Masse aus der sie stammen, beschreiben.

Beispiel 1.1. Ein Auto eines bestimmten Typs wird als sehr langlebig angesehen. Dies kann als Vorinformation oder Vorurteil gewertet werden. Woher kommt eine solche Aussage? Werden z. B. die Laufleistung der Motoren, das Durchschnittsalter von verkehrstauglichen Fahrzeugen, die Reparaturanfälligkeit erhoben (gemessen), so kann man aus diesen Daten, die im Einzelnen keine Übersicht geben und keine Aussage erkennen lassen, Kennzahlen (Maßzahlen) berechnen, die die Einzelinformationen verdichten und somit dann einen einfachen Vergleich zwischen verschiedenen Autotypen erlauben. Aus diesem Vergleich wird dann eine solche Aussage

hergeleitet und die Vorinformation überprüft. Jedoch kann aus einer solchen Aussage keine Vorhersage über die Haltbarkeit eines einzelnen, spezifischen Autos abgeleitet werden.

Wenn Sie die im Folgenden beschriebenen Verfahren unter diesem Aspekt betrachten, wird hoffentlich verständlich, dass es sich bei der Statistik nicht um eine Art anderer Mathematik handelt, sondern um ein Konzept, mit dem aus vielen Einzelinformationen eine Gesamtinformation abgeleitet wird, die zur Überprüfung einer Vorinformation (Vorurteil?), eines Erklärungsansatzes verwendet wird. Nun beeinflusst die Vorinformation aber häufig die Sicht der Dinge und damit wie und was untersucht wird, so dass zwischen Vorinformation und statistischer Untersuchung eine Art Wechselwirkung / Zusammenhang existieren kann. Bei der Statistik handelt es sich also um einen ähnlichen Vorgang wie bei einem (Landschafts-) Maler der ein Bild erstellt: Der Maler versucht eine Situation zu erfassen, auf Details zu verzichten und dennoch eine Abbildung der Realität in seinem Bild wiederzugeben. Dabei gehen natürlich seine Empfindungen, seine Interpretation, seine Sicht der Dinge (Vorurteil) in das Bild ein. Der Statistiker möchte ebenfalls ein Bild malen, jedoch misst er die Situation mit Zahlen und fasst diese in einer Statistik und/oder Abbildung zusammen. Dabei muss er wegen der Übersichtlichkeit ebenfalls auf Details verzichten, aber nicht auf die wesentlichen Dinge. Hierbei gehen, wie bei dem Maler, seine Erklärungsmodelle (Interpretationen) der Realität, seine Vorinformationen, seine Hypothesen in die Statistik ein. Dabei muss aber klar sein, dass die Daten – in Analogie zum Maler, das Wetter, die Wahl der Farben – „zufällig“ auch anders hätten ausfallen können.

In den Wirtschaftswissenschaften werden häufig neue Erkenntnisse durch Beobachtungen induziert. Auf der anderen Seite sollte jede theoretische Überlegung durch Beobachtungen an der Realität überprüft werden. Jeder Jahresbericht enthält Statistiken, jede Investition wird aufgrund von Absatzerwartungen – dies ist eine statistische Prognose, basierend auf einer statistischen Erhebung – mit entschieden. Die wenigen Beispiele zeigen, dass beinahe bei jeder Entscheidung, nicht nur in den Wirtschaftswissenschaften, Zahlen mitspielen, die statistisch ausgewertet und analysiert werden müssen. Aufgrund dieser Auswertungen und Analysen werden dann z. B. im betrieblichen Bereich (Management-) Entscheidungen vorbereitet und vielleicht sogar entschieden.

An dieser Stelle soll schon darauf hingewiesen werden, dass mit „Statistik“ nichts bewiesen werden kann. Mit der Statistik können theoretische Überlegungen überprüft, gestützt werden, jedoch die Richtigkeit im Sinne eines Beweises kann sie nicht liefern, da die Aussage immer an den untersuchten Datensatz gebunden ist. Wesentlich für die Interpretation statistischer Ergebnisse sind Kausalmodelle, die der Beobachter als Vorwissen für die Analyse der Daten miteinbringt (siehe Kapitel 5.7, 9.6).

Im vorliegenden Text werden immer wieder Herleitungen zu Formeln angegeben. Sie erklären die Entstehung der Formel und damit z. T. die aus ihr abgeleitete Interpretation. Die Zusammenhänge sind leider nicht immer auf den ersten Blick einsichtig. Daher ist es für ein Erfassen der Zusammenhänge m. E. unerlässlich, die

Beispiele nachzurechnen und die Übungen zu lösen. Diese Erkenntnis kann aber auch mit einer Statistik belegt werden: Handlungsorientiertes Lernen ist am effektivsten: Von dem, was wir mit den eigenen Händen tun, behalten wir 90 Prozent im Gedächtnis. Von der reinen Lektüre eines Buches werden nur 10 Prozent erinnert! (siehe Tabelle 1.1).

Tabelle 1.1. Lernerfolg

Der Mensch behält von dem . . .	
was er liest	10%
was er hört	20%
was er sieht	30%
was er sieht und hört	50%
was er spricht	70%
was er selbst ausführt	90%

[68, Seite 46]

Deskriptive Statistik

Grundlagen

Inhaltsverzeichnis

2.1	Einführung	7
2.2	Grundbegriffe der deskriptiven Statistik	8
2.3	Messbarkeitseigenschaften von Merkmalen	13
2.3.1	Nominalskala	13
2.3.2	Ordinalskala	13
2.3.3	Kardinalskala	14
2.4	Skalentransformation	15
2.5	Häufigkeitsfunktion	16
2.6	Klassierung von metrischen Merkmalswerten	18
2.7	Übungen	21

2.1 Einführung

In der deskriptiven Statistik werden Verfahren zur Beschreibung bestimmter Charakteristiken einer beobachteten Stichprobe erklärt. Es interessieren also die beobachteten Werte selbst, nicht die Grundgesamtheit, aus der sie stammen, es sei denn, die Grundgesamtheit selbst wird erhoben und untersucht. Interessiert die übergeordnete Grundgesamtheit, obwohl man nur eine Stichprobe erhoben hat, so begibt man sich in das Teilgebiet der induktiven (schließenden) Statistik. Sie bedient sich formaler Modelle. Innerhalb dieser ist es möglich, Rückschlüsse von einer zufällig ausgewählten Teilmenge auf die Gesamtheit, aus der sie entnommen wurde, zu ziehen.

Statistische Untersuchungen gliedern sich im Allgemeinen in vier Phasen, für die jeweils entsprechende statistische Methoden existieren.

1. Vorbereitung

- Zweck der Untersuchung bestimmen

2. Datenerhebung durch
 - Primärstatistik (Experiment, Beobachtung, Befragung)
 - Sekundärstatistik (aus Primärstatistiken erstellt)
3. Datenaufbereitung und -darstellung
 - grafisch
 - tabellarisch
4. Datenauswertung und -analyse
 - Berechnung von Maßzahlen
 - Darstellung von Verteilungsfunktionen

Bei empirischen Arbeiten benötigen die ersten beiden Phasen in der Regel häufig mehr als Zweidrittel des Zeitbudgets. Diese Phasen sind von größter Bedeutung, weil mit der Zweckbestimmung die interessierenden Daten und mit der Erhebung der Daten deren Skalenniveau (Informationsgehalt) und damit die Analysemethoden festgelegt werden. Die Datenanalyse selbst ist durch den Einsatz von Tabellenkalkulationsprogrammen und speziellen Statistikprogrammen leicht geworden.

Der Erfolg einer Untersuchung hängt von einer klaren Zweckbestimmung ab. Einstein schrieb, dass die Formulierung eines Problems häufig von größerer Bedeutung ist als die Lösung (vgl. [27, Seite 92]). Dies fängt bei der Bewilligung der Untersuchung an, setzt sich bei der Datenerhebung fort und endet bei der Präsentation der Ergebnisse. Niemand wird Zeit und Geld für eine statistische Untersuchung bewilligen, deren Ziel unklar ist. Befragte werden Antworten verweigern oder die Fragen unwissentlich oder wissentlich falsch beantworten, weil sie unklar sind oder wegen der schlechten Begründung nicht akzeptiert werden. Aufgrund einer unklaren Aufgabenstellung können irrelevante oder sogar falsche Daten erhoben werden. Dies alles wirkt sich natürlich fatal auf die Ergebnisse aus. Denn eine statistische Auswertung kann keine Fehler in den Daten beseitigen! Also legen Sie größte Sorgfalt auf die Erhebung der Daten, überprüfen Sie die Qualität der Daten und der Datenquellen kritisch!

2.2 Grundbegriffe der deskriptiven Statistik

Ein gegebener Vorgang soll aufgrund von Beobachtungen statistisch erfasst werden. Dies geht, wie in der Einleitung beschrieben, mit einer Reduktion der Komplexität einher. Bevor jedoch mit der Datenerhebung begonnen wird, müssen einige Festlegungen erfolgen. Die Grundbegriffe der Statistik dienen zur Bestimmung, was statistisch erfasst und was analysiert werden soll. Sie beschreiben den statistischen Messvorgang.

Definition 2.1. Die **statistische Einheit** (*Element*) ω_i ($i = 1, \dots, n$) ist das Einzelobjekt einer statistischen Untersuchung. Sie ist Träger der Information, für die man sich bei der Untersuchung interessiert. Die statistischen Einheiten heißen auch *Merkmalsträger*. Der Laufindex i kennzeichnet die statistische Einheit, n bezeichnet die Gesamtzahl der erfassten Elemente.

Beispiel 2.1. Bei einer statistischen Untersuchung der Einkommensverteilung in der Bundesrepublik Deutschland sind die „Einkommensbezieher“ die statistischen Untersuchungseinheiten.

Beispiel 2.2. Die Besucherdichte auf einer Messe kann durch Zählung der pro Zeiteinheit an festgelegten Beobachtungspunkten passierenden Besucher gemessen werden. Die statistischen Einheiten sind die „Beobachtungspunkte“, denn an diesen Einzelobjekten der Untersuchung werden die Informationen „Anzahl der pro Zeiteinheit passierenden Besucher“ erfasst.

Jede statistische Einheit muss im Hinblick auf das Untersuchungsziel durch

- sachliche
- räumliche
- zeitliche

Kriterien identifiziert bzw. abgegrenzt werden.

Beispiel 2.3. Für die in Beispiel 2.1 genannte Untersuchung der Einkommensverteilung sind die Identifikationskriterien

- räumlich: Gebiet der BRD
- sachlich: Einkommensbezieher
- zeitlich: Zeitraum (z. B. Jahr) der Untersuchung

Die Kriterien für die Abgrenzung von statistischen Einheiten ergeben sich meistens unmittelbar aus der jeweils zu behandelnden Fragestellung. Die Kriterien sind mit großer Sorgfalt auszuwählen, denn eine ungeeignete Abgrenzung der Einheiten macht die gesamte Untersuchung wertlos. Durch die Kriterien der Abgrenzung wird die statistische Grundgesamtheit bestimmt.

Definition 2.2. Die **statistische Masse oder Grundgesamtheit** Ω ist die Menge von statistischen Einheiten, die die vorgegebenen Abgrenzungskriterien erfüllen.

$$\Omega = \{\omega_i \mid \text{auf } \omega_i \text{ treffen die Abgrenzungskriterien zu}\} \quad (2.1)$$

Beispiel 2.4. Die Grundgesamtheit Ω in Beispiel 2.1 könnten alle Einkommensbezieher sein, die mehr als 10 000 € und weniger als 75 000 € pro Jahr verdienen.

$$\Omega = \{\omega_i \mid 10\,000 \leq \text{Einkommen}(\omega_i) \leq 75\,000\} \quad (2.2)$$

Bei der statistischen Grundgesamtheit kann es sich um so genannte Bestands- oder Ereignismassen handeln. Bei **Bestandsmassen** gehören die statistischen Einheiten nur für ein gewisses Zeitintervall zur Masse, d. h. sie treten zu einem bestimmten Zeitpunkt ein und zu einem späteren Zeitpunkt aus der Masse wieder aus. Bei einer **Ereignismasse** sind die statistischen Einheiten Ereignisse, die zu einem bestimmten Zeitpunkt auftreten und die statistische Masse nicht mehr verlassen.

Beispiel 2.5. Die Bestandsmasse „Einwohner einer Stadt“ oder „Auftragsbestand“ kann nur innerhalb eines bestimmten Zeitintervalls angegeben werden, da sie fortlaufend durch Ereignisse wie z. B. Geburten verändert wird. Die Angabe muss daher auf einen Zeitpunkt bezogen sein. Die Ereignismasse „Geburten in einem Jahr“ oder „Auftragseingang“ kann hingegen nur in einem Zeitraum z. B. in einem Jahr gemessen werden. Sie muss daher auf einen Zeitraum bezogen sein.

Für eine statistische Untersuchung wird in der Regel nur ein Teil der Grundgesamtheit Ω untersucht. Aus der Grundgesamtheit wird dann die so genannte Ergebnismenge \mathcal{X} gezogen, die auch als Stichprobe bezeichnet wird (siehe auch Kapitel 3.2).

Definition 2.3. Wird bei einer statistischen Untersuchung nur ein Teil der Grundgesamtheit Ω (interessierenden Masse) erfasst, dann heißt dieser Teil **Stichprobe** und wird \mathcal{X} bezeichnet.

Meistens interessiert man sich nicht für die statistischen Einheiten selbst, sondern für bestimmte Eigenschaften der Einheiten wie z. B. Alter, Geschlecht, Einkommen. Diese werden als Merkmale einer statistischen Einheit bezeichnet.

Definition 2.4. Eine bei einer statistischen Untersuchung interessierende Eigenschaft einer statistischen Einheit heißt **Merkmalsausprägung** $X(\omega_i)$ oder kurz X .

Da die Merkmale an den statistischen Einheiten erhoben werden, sind die Merkmale als eine Funktion (Abbildung) der statistischen Einheiten zu sehen.

Beispiel 2.6. In Beispiel 2.1 (Seite 9) ist das Einkommen das Merkmal X . Es wird an der statistischen Einheit ω_i „Einkommensbezieher“ erhoben.

Der bisher beschriebene statistische Messvorgang umfasst die Bestimmung der statistischen Einheit sowie des Merkmals. Jedoch muss auch bestimmt werden, wie das Merkmal erfasst werden soll. Dazu wird der Begriff der Merkmalsausprägung definiert.

Definition 2.5. Die interessierenden Werte (Kategorien), die ein Merkmal annehmen kann, heißen **Merkmalsausprägungen** und werden mit der Menge

$$\mathcal{A}_X = \{x_1, \dots, x_m\} \quad (2.3)$$

festgelegt. Mit m wird die Zahl der interessierenden Merkmalsausprägungen bezeichnet.

In der Regel ist $m \leq n$, d. h. die Anzahl der Merkmalsausprägungen m ist in der Regel kleiner als die Anzahl der Beobachtungen n . Davon zu unterscheiden ist, dass theoretisch weit mehr Merkmalsausprägungen existieren können.

Definition 2.6. Die Menge \mathcal{A} der möglichen Merkmalsausprägungen wird mit \mathcal{A} bezeichnet.

Es gilt: $\mathcal{A}_X \subseteq \mathcal{A}$. Die Zahl der beobachteten Merkmalsausprägungen kann wiederum kleiner sein als die Zahl der interessierenden Merkmalsausprägungen. Die Zahl der beobachteten Merkmalsausprägungen wird aber ebenfalls mit m bezeichnet, weil einerseits davon auszugehen ist, dass die interessierenden Merkmalsausprägungen auch beobachtet werden. Andererseits werden Merkmalsausprägungen die nicht beobachtet werden, auch nicht die Berechnungen beeinflussen, so dass die Unterscheidung nicht notwendig ist.

Die Festlegung der Merkmalsausprägungen sollte vor einer statistischen Untersuchung geschehen, um sicherzustellen, dass keine interessierenden Merkmalsausprägungen übersehen werden. Dies ist vor allem bei der Erhebung vom Daten wichtig.

Beispiel 2.7. Bei einer Untersuchung zum gewünschten Koalitionspartner mit der SPD könnten die interessierenden Merkmalsausprägungen sich nur auf $x_1 = \text{CDU}$, $x_2 = \text{Grüne}$ beschränken.

$$\mathcal{A}_X = \{\text{CDU}, \text{Grüne}\} \quad (2.4)$$

Die Menge \mathcal{A} der möglichen Merkmalsausprägungen umfasst alle zur Wahl zugelassenen Parteien. Bei einer Wahlprognose sollten hingegen alle antretenden Parteien als interessierende Merkmalsausprägungen in die Menge \mathcal{A}_X aufgenommen werden.

Bei dieser Art der Merkmalsausprägungen werden dann häufig die Merkmalsausprägungen bei der Weiterverarbeitung mit Zahlen kodiert (siehe Kapitel 2.4).

Beispiel 2.8. Bei dem Merkmal „Einkommenshöhe“ sind die möglichen Merkmalsausprägungen alle positiven reellen Zahlen: $\mathcal{A}_X = \mathbb{R}^+$. Hier ist $\mathcal{A}_X = \mathcal{A}$.

Man beobachtet mittels der statistischen Einheiten eine bestimmte Merkmalsausprägung. Durch die Beobachtung wird dem Merkmal die entsprechende Merkmalsausprägung zugeordnet. Dieser Vorgang ist der eigentliche statistische Messvorgang. Die beobachtete Merkmalsausprägung wird als Merkmalswert oder Beobachtungswert bezeichnet.

Definition 2.7. Eine bei einer statistischen Untersuchung an einer bestimmten statistischen Einheit festgestellte Merkmalsausprägung heißt **Merkmalswert** oder **Beobachtungswert** und wird mit x_i bezeichnet. Dem Merkmal X wird ein Element aus der Menge der Merkmalsausprägungen \mathcal{A}_X zugeordnet und zwar jenes, das beobachtet wurde.

$$X : \omega_i \mapsto X(\omega_i) = x_i \in \mathcal{A}_X \quad (2.5)$$

Die Schreibweise $x \mapsto f(x)$ bedeutet, dass die Funktion f (hier X) den Elementen x (hier ω_i) das Bild $f(x)$ (hier $X(\omega_i) = x_i$) zuordnet. Der statistische Messvorgang kann damit als folgende Abbildung beschrieben werden:

$$(\Omega, \mathcal{A}) \xrightarrow{X} (\mathcal{X}, \mathcal{A}_X) \quad (2.6)$$

Ein Teil der Grundgesamtheit Ω wird auf die Stichprobe \mathcal{X} abgebildet. Die Menge der theoretischen Merkmalsausprägungen \mathcal{A} – also aller möglichen Merkmalsausprägungen – wird dabei auf die Menge der Merkmalsausprägungen \mathcal{A}_X abgebildet, die die Menge der interessierenden Merkmalsausprägungen ist. Durch die Beobachtung, bezeichnet mit \xrightarrow{X} , wird den statistischen Einheiten ω_i das Bild $X(\omega_i) = x_i$ zugeordnet.

Beispiel 2.9. Es ist eine Person (statistische Einheit) nach der gewählten Partei (Merkmal) befragt worden. Die Antwort „SPD“ ist der Beobachtungswert, der aus der Menge der Merkmalsausprägungen \mathcal{A}_X dem Merkmal „gewählte Partei“ gegeben wird.

Der Unterschied zwischen den Merkmalsausprägungen und den Merkmalswerten ist, dass die Merkmalsausprägungen vor der Untersuchung festgelegt werden, während die Merkmalswerte (auch Beobachtungswerte genannt) die beobachteten Ausprägungen sind. Es kann also durchaus sein, dass eine Merkmalsausprägung definiert ist, die nicht beobachtet wird.

Hat man nun den Messvorgang abgeschlossen, so werden die Beobachtungswerte in einer Urliste notiert, mit der dann die statistische Analyse durchgeführt wird.

Definition 2.8. Die Aufzeichnung der Merkmalswerte für die statistischen Einheiten heißt **Urliste**. Sie enthält die **Rohdaten** einer statistischen Untersuchung.

Beispiel 2.10. Die Erhebung statistischer Daten läuft nun wie folgt ab: Von der statistischen Einheit ω_i „Student“ werden die Merkmale X „Alter“ und Y „Studienfach“ erhoben. Die möglichen Merkmalsausprägungen bei dem ersten Merkmal sind das Alter $\mathcal{A}_X \subset \mathbb{R}^+$ und bei dem zweiten Merkmal die möglichen (interessierenden) Studienfächer $\mathcal{A}_Y = \{\text{BWL, Mathematik, } \dots\}$. Die in der Urliste gemessenen Merkmalswerte für einen befragten Studenten ω_i könnten dann beispielsweise $x_i = 21$ Jahre und $y_i = \{\text{BWL}\}$ sein.

Häufig werden Merkmalsausprägungen, die selbst keine Zahlenwerte sind, mit Zahlen kodiert; z. B. könnte das Studienfach „BWL“ durch eine 2 kodiert sein: $y_i = 2$. Dadurch wird die Verarbeitung mittels der EDV ermöglicht (siehe Kapitel 2.4).

Zusammenfassend läuft also folgender Prozess von der Erhebung der statistischen Einheit bis zur Messung des entsprechenden Merkmalswertes ab, der die Merkmalsausprägung repräsentiert: An den statistischen Einheiten ω_i mit den Merkmalen X, Y, \dots und der festgelegten Menge der Merkmalsausprägungen $\mathcal{A}_X, \mathcal{A}_Y, \dots$ werden die Merkmalswerte x_i, y_i, \dots gemessen.

Die folgenden beiden Definitionen werden zur weiteren, genaueren Charakterisierung von Merkmalen benötigt.

Definition 2.9. Ein Merkmal heißt **häufbar**, wenn an derselben statistischen Einheit mehrere Ausprägungen des betreffenden Merkmals vorkommen können.

$$X(\omega_i) = \{x_i \mid x_i \in \mathcal{A}_X\} \quad (2.7)$$

Beispiel 2.11. An dem Merkmal „Unfallursache“ können z. B. „überhöhte Geschwindigkeit“ und „Trunkenheit am Steuer“ die Unfallursachen sein. Hingegen kann beim Merkmal „Geschlecht“ dieselbe statistische Einheit nicht beide mögliche Ausprägungen gleichzeitig auf sich vereinen.

Definition 2.10. *Kann ein Merkmal nur endlich viele Ausprägungen annehmen oder abzählbar unendlich viele Ausprägungen, so heißt es ein **diskretes Merkmal**. Kann ein Merkmal alle Werte eines Intervalls annehmen, heißt es ein **stetiges Merkmal**.*

Beispiel 2.12. Diskrete Merkmale sind z. B. Kinderzahl, Anzahl der Verkehrsunfälle und Einkommen. Es können nur einzelne Zahlenwerte auftreten, Zwischenwerte sind unmöglich.

Stetige Merkmale sind z. B. Körpergröße, Alter und Gewicht. Ein stetiges Merkmal kann wenigstens in einem Intervall der reellen Zahlen jeden beliebigen Wert aus dem Intervall annehmen.

2.3 Messbarkeitseigenschaften von Merkmalen

Merkmalsausprägungen werden in qualitative (klassifikatorische), komparative (intensitätsmäßige) und quantitative (metrische) Merkmale unterteilt. Die Ausprägungen qualitativer Merkmale unterscheiden sich durch ihre Art, die komparativer durch ihre intensitätsmäßige Ausprägung und die quantitativer Merkmale durch ihre Größe.

Um die Ausprägungen eines Merkmals zu messen, muss eine Skala festgelegt werden, die alle möglichen Ausprägungen des Merkmals beinhaltet. Die Skala mit dem niedrigsten Niveau ist die so genannte Nominalskala. Auf ihr werden qualitative Merkmale erfasst. Ein etwas höheres Messniveau hat die Ordinalskala. Auf ihr werden komparative Merkmale gemessen, die intensitätsmäßig unterschieden werden können. Die Skala mit dem höchsten Messniveau ist die metrische Skala. Auf ihr werden quantitative Merkmale gemessen, die metrische Eigenschaften aufweisen.

2.3.1 Nominalskala

Definition 2.11. *Eine Skala, deren Skalenwerte nur nach dem Kriterium gleich oder verschieden geordnet werden können, heißt **Nominalskala**.*

Beispiel 2.13. Qualitative Merkmale sind z. B. Geschlecht, Beruf, Haarfarbe. Die Merkmalsausprägungen lassen sich nur nach ihrer Art unterscheiden. Eine Ordnung der Merkmalsausprägungen ist nicht möglich. Man kann nicht sagen, dass der Beruf Schlosser besser als der Beruf Bäcker ist oder umgekehrt.

2.3.2 Ordinalskala

Definition 2.12. *Eine Skala, deren Skalenwerte nicht nur nach dem Kriterium gleich oder verschieden, sondern außerdem in einer natürlichen Reihenfolge geordnet werden können, heißt **Ordinalskala**.*

Beispiel 2.14. Komparative Merkmale sind z. B. Zensuren, Güteklassen oder Grad einer Beschädigung. Bei Zensuren weiß man, dass die Note 1 besser als die Note 2 ist, aber der Abstand zwischen 1 und 2 lässt sich nicht bestimmen. Ebenso verhält es sich mit den anderen aufgeführten Beispielen. Die Ausprägungen unterliegen einer Rangfolge, die Abstände sind aber nicht interpretierbar.

Da qualitative und komparative Merkmale nur in Kategorien eingeteilt werden können, werden sie auch als **kategoriale Merkmale** bezeichnet. Die Kategorien, auch als Merkmalsausprägungen dann bezeichnet (siehe Definition 2.5) werden manchmal auch (diskrete) Klassen genannt.

2.3.3 Kardinalskala

Definition 2.13. *Eine Skala, deren Skalenwerte reelle Zahlen sind und die die Ordnungseigenschaften der reellen Zahlen besitzt, heißt **Kardinalskala** oder **metrische Skala**.*

Beispiel 2.15. Quantitative Merkmale sind z. B. Alter, Jahreszahlen, Einkommen oder Währungen. Diese Merkmalsausprägungen können nicht nur nach ihrer Größe unterschieden werden, sondern es ist auch der Abstand zwischen den Merkmalsausprägungen interpretierbar.

Bei der Kardinalskala bzw. metrischen Skala wird die Skala weiterhin dahingehend unterschieden, ob ein natürlicher Nullpunkt und eine natürliche Einheit existieren. Diese Unterscheidung ist von nachgeordneter Bedeutung.

Definition 2.14. *Eine metrische Skala, die keinen natürlichen Nullpunkt und keine natürliche Einheit besitzt, heißt **Intervallskala**.*

Beispiel 2.16. Kalender besitzen keinen natürlichen Nullpunkt und auch keine natürliche Einheit. Der Anfangspunkt der Zeitskala wird willkürlich auf ein Ereignis gesetzt, von dem aus die Jahre der Ära gezählt werden. Solche Zeitpunkte sind z. B. in der römischen Geschichte das (fiktive) Gründungsjahr Roms (753 v. Chr.), in der islamischen Geschichte das Jahr der Hedschra (622 n. Chr.) und in der abendländischen Geschichte die Geburt Christi. Die heute übliche Einteilung der Zeitskala wird nach dem gregorianischen Kalender (der auf Papst Gregor XIII (1582) zurückgeht) vorgenommen. Die Zeitabstände (Intervalle) auf der Skala können miteinander verglichen werden. Eine Quotientenbildung $1980/1990$ ergibt jedoch kein interpretierbares Ergebnis.

Definition 2.15. *Eine metrische Skala, die einen natürlichen Nullpunkt, aber keine natürliche Einheit besitzt, heißt **Verhältnisskala**.*

Beispiel 2.17. Bei Währungen existiert ein natürlicher Nullpunkt. Null Geldeinheiten sind überall in der Welt nichts. Jedoch sind 100 \$ in der Regel nicht gleich 100 €. Jede Währung wird auf einer anderen Skala gemessen. Das Verhältnis zweier Währungen ist interpretierbar und wird als Wechselkurs bezeichnet.

$$Y[\$] = \text{Wechselkurs} \left[\frac{\$}{\text{€}} \right] \times X[\text{€}] \quad (2.8)$$

Definition 2.16. Eine metrische Skala mit einem natürlichen Nullpunkt und einer natürlichen Einheit heißt **Absolutskala**.

Beispiel 2.18. Stückzahlen besitzen einen natürlichen Nullpunkt und eine natürliche Einheit.

2.4 Skalentransformation

Bei der praktischen statistischen Analyse ist man manchmal daran interessiert, alle erhobenen Daten auf einer Skala zu messen. Sind nun die Merkmale auf unterschiedlichen Skalen erfasst, so kann man durch eine Skalentransformation eine Skala mit höherem Messniveau durch eine mit niedrigerem ersetzen. Andersherum geht es nicht!

Daraus ergibt sich auch, dass Maßzahlen, die für Merkmale eines niedrigeren Messniveaus konstruiert sind, auf Merkmale eines höheren Messniveaus angewandt werden können. Dabei werden dann aber nicht alle Informationen, die die Merkmalsausprägungen enthalten, verwendet. Wird beispielsweise eine Maßzahl für komparative Merkmale auf metrische angewendet, so bleibt die Abstandsinformation unberücksichtigt. Dies ist sogar manchmal wünschenswert (siehe Beispiel 4.22 auf Seite 65 und Vergleich von Mittelwert und Median auf Seite 67). Maßzahlen für z. B. metrische (quantitative) Merkmale dürfen aber nicht auf komparative oder qualitative Merkmale angewandt werden.

Definition 2.17. Unter einer **Skalentransformation** versteht man die Übertragung der Skalenwerte in Werte einer anderen Skala, wobei die Ordnungseigenschaften der Skala erhalten bleiben müssen.

Beispiel 2.19. Wird bei einer Befragung von Familien die Kinderzahl ermittelt und ist allein die Unterscheidung der Familien nach der Kinderzahl von Interesse, so reichen die Ordnungskriterien der Nominalskala für die Analyse aus. Es wird eine Absolutskala in eine Nominalskala transformiert.

Definition 2.18. Mit **Kodierung** bezeichnet man die Transformation der Merkmalsausprägungen, ohne dass dabei das Messniveau der erhobenen Daten geändert wird.

Die Gefahr bei einer Kodierung von kategorialen Daten in eine Zahlenskala besteht darin, dass die Daten als rechenbare Größen aufgefasst werden.

Beispiel 2.20. Die Merkmalsausprägungen „männlich“ und „weiblich“ einer Nominalskala können durch die Werte „0“ und „1“ kodiert werden. Diese Zahlenwerte dürfen daher nicht durch Addition und Multiplikation oder andere Operationen bearbeitet werden.

Beispiel 2.21. Die Noten „sehr gut“ bis „mangelhaft“ werden im Allgemeinen mit den Zahlen 1 bis 5 kodiert. Obwohl Zahlen verwendet werden, handelt es sich weiterhin um ordinale Merkmalsausprägungen. Deshalb ist es unzulässig Notendurchschnitte zu berechnen.

Statistische Merkmale unterscheiden sich also durch ihren Skalentyp, dem Informationsgehalt der Merkmalswerte. Will man die erhobenen Daten mit Hilfe statistischer Methoden verdichten, d. h. durch Maßzahlen beschreiben, so ist dem Skalentyp des betrachteten Merkmals Rechnung zu tragen. Daraus ergibt sich auch ein Teil der Gliederung in dem vorliegenden Text.

2.5 Häufigkeitsfunktion

In der Urliste kommen häufig einige oder alle Merkmalswerte mehrmals vor. Um die mehrmals vorkommenden Merkmalswerte nicht mehrmals aufschreiben zu müssen, werden diese mit der Anzahl des Auftretens der Merkmalsausprägung gekennzeichnet.

Definition 2.19. Die Anzahl der Beobachtungswerte mit der Merkmalsausprägung x_j heißt **absolute Häufigkeit** $n(x_j)$. Die absolute Häufigkeit ist eine natürliche Zahl oder null.

$$n(x_j) = \left| \{ \omega_i \mid X(\omega_i) = x_j \} \right| \quad i = 1, \dots, n; j = 1, \dots, m \quad (2.9)$$

Bei kategorialen Daten (nominales oder ordinales Merkmalsniveau) werden häufig die Kategorien auch als Klassen bezeichnet. In diesem Fall spricht man dann auch von **diskreter Klassierung**, wenn die Merkmalsausprägungen in einer Tabelle mit der auftretenden Häufigkeit aufgelistet werden. Die Anzahl der Kategorien bzw. Klassen m ist bei der Analyse von kategorialen und metrischen Daten in der Regel kleiner als die Anzahl der Beobachtungen: $m \leq n$. Im Text wird die Kategorie bzw. Klasse mit dem Index $j = 1, \dots, m$ gekennzeichnet.

Beispiel 2.22. Die Stimmen, die auf die Merkmalsausprägung „Partei“ entfallen, sind die absoluten Häufigkeiten. Bei der Landtagswahl am 27. Februar 2000 in Schleswig-Holstein erzielten die Parteien die in Tabelle 2.1 angegebenen Stimmen.

Die absolute Häufigkeit, mit der die Merkmalsausprägung auftritt, kann in Beziehung zu der Gesamtzahl n der Beobachtungswerte gesetzt werden. Dann erhält man relative Häufigkeiten.

Definition 2.20. Der relative Anteil der absoluten Häufigkeiten einer Merkmalsausprägung x_j an der Gesamtzahl der Beobachtungswerte heißt **relative Häufigkeit**.

$$f(x_j) = \frac{n(x_j)}{n} \quad (2.10)$$

¹ Mit SSW wird die Süd-Schleswigsche-Wählerversammlung bezeichnet.

Tabelle 2.1. Landtagswahl Schleswig-Holstein vom 27. Februar 2000

Partei x_j	Stimmen $n(x_j)$
SPD	689 764
CDU	567 428
F.D.P.	78 603
Grüne	63 256
SSW ¹	37 129
sonstige	13 189
Σ	1 449 369

Quelle: Statistisches Landesamt Schleswig-Holstein

Für die relativen Häufigkeiten gilt:

$$0 \leq f(x_j) \leq 1 \quad (2.11)$$

Die relative Häufigkeit kann null oder größer als null sein. Negative Werte kann sie nicht annehmen, da auch keine negativen absoluten Häufigkeiten beobachtet werden können. Ferner gilt:

$$\sum_{j=1}^m f(x_j) = 1 \quad (2.12)$$

Die relative Häufigkeit kann nicht größer als eins werden, da sie als Relation zu der Gesamtzahl aller Beobachtungen formuliert ist. Es ist dabei darauf zu achten, dass man auch die richtige Gesamtheit n in Beziehung setzt.

Beispiel 2.23. Soll beispielsweise ermittelt werden, ob bestimmte Autotypen besonders von „Dränglern“ gefahren werden, so ist die Zahl der „Drängler“ eines bestimmten Autotyps zur Zahl der insgesamt beobachteten Zahl der Autos dieses Typs und nicht zur Gesamtzahl der beobachteten Autos in Relation zu setzen.

Für kategoriale und klassierte metrische Daten sind die absoluten und relativen Häufigkeiten auf Kategorien bzw. Klassen bezogen, d. h. $n(x_j)$ gibt die Anzahl der Beobachtungswerte bzw. $f(x_j)$ den Anteil in der j -ten Kategorie bzw. Klasse an.

Beispiel 2.24. In der Landtagswahl am 27. Februar 2000 in Schleswig-Holstein waren 2 134 954 Personen wahlberechtigt. Davon haben 1 484 128 Personen insgesamt gewählt, wovon aber nur 1 449 369 Stimmen gültig waren. 689 764 Stimmen entfielen davon auf die „SPD“, die hier mit der Merkmalsausprägung x_j bezeichnet wird.

Merkmalsausprägung x_j : „SPD“
absolute Häufigkeit: $n(x_j) = 689\,764$
relative Häufigkeit: $f(x_j) = 689\,764 / 1\,449\,369 = 0.476$

Das Gesamtergebnis der Wahl ist im Beispiel 2.22 (siehe Seite 16) wiedergegeben.

Die Verwendung von Prozenten ist immer wieder eine Quelle für Fehler. Änderungsraten werden häufig zur Beschreibung einer Entwicklung verwendet (siehe hierzu auch Kapitel 7). Hierbei werden sehr häufig Prozent und Prozentpunkt verwechselt. Mit **Prozent** wird ein relativer Anteil von etwas ausgedrückt. Um eine Entwicklung zu beschreiben, verwendet man die Veränderung der relativen Anteile (Prozente). Hierzu benutzt man meistens die absolute Änderung des relativen Anteils, den **Prozentpunkt**, weil der leicht auszurechnen ist. Aber auch die Verwendung der relativen Änderung der relativen Anteile ist zulässig.

Beispiel 2.25. Die Änderung des Stimmenanteils für eine Partei von 36% auf 45% bedeutet eine Zunahme um 9 Prozentpunkte ($45\% - 36\%$). Die relative Zunahme des Stimmenanteils beträgt aber 25% ($9\%/36\%$) und nicht 9%!

2.6 Klassierung von metrischen Merkmalswerten

Bei der Untersuchung von metrischen Merkmalswerten ist die Erfassung und Auszählung aller einzelnen Merkmalsausprägungen nicht sinnvoll oder nicht möglich, weil

- die Anzahl der Merkmalsausprägungen zu groß ist,
- das Merkmal stetig ist,
- die Übersichtlichkeit bei der Darstellung und Datenaufbereitung verloren geht.

In diesem Fall werden die Merkmalswerte klassiert. Man spricht dann von einer **stetigen Klassierung**. Es werden nicht alle möglichen Merkmalsausprägungen einzeln erfasst, sondern benachbarte Merkmalsausprägungen zu einer Klasse zusammengefasst. Damit ist zwangsläufig ein Verlust an Information verbunden, da sich die einzelnen Ausprägungen der in eine Klasse fallenden Merkmalswerte nachträglich nicht mehr feststellen lassen.

Werden bei diskreten Merkmalen gleiche Merkmalswerte zusammengefasst, so spricht man von einer diskreten Klassierung.

Definition 2.21. Eine Klasse mit stetigen Merkmalswerten wird durch zwei Grenzen bestimmt, die untere Klassengrenze (x_{j-1}^*) und die obere Klassengrenze (x_j^*). Diese Klasse wird als j -te Klasse bezeichnet. Um alle Beobachtungswerte eindeutig einer Klasse zuordnen zu können, müssen die Klassen nicht überlappend sein. Die Merkmale x_i mit $i = 1, \dots, n$ Merkmalswerten werden in $j = 1, \dots, m$ Klassen eingeteilt mit $m < n$.

Es ist Konvention, den Wert der oberen Klassengrenze einzuschließen und den Wert der unteren Klassengrenze auszuschließen. Dies wird auch mit dem halboffenen Intervall $(x_{j-1}^*, x_j^*]$ beschrieben. Die Wahl der Klassengrenzen beeinflusst die Häufigkeit in den Klassen und damit die Verteilung der klassierten Werte (siehe hierzu auch Beispiel 4.18, Seite 57).

Beispiel 2.26. Die Merkmalswerte $x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 2$ und $x_5 = 4$ ($i = 1, \dots, n = 5$) sollen in drei Klassen ($j = 1, \dots, m = 3$) mit den Klassengrenzen $x_0^* = 0, x_1^* = 1, x_2^* = 3, x_3^* = 6$ eingeteilt werden.

Dann ist $x_1^* = 1$ die Obergrenze der ersten Klasse und $x_0^* = 0$ die Untergrenze der ersten Klasse. Die Untergrenze der zweiten Klasse ist $x_1^* = 1$ und die Obergrenze der zweiten Klasse ist $x_2^* = 3$. Der Merkmalswert $x_1 = 1$ wird aufgrund der Klassenabgrenzung der ersten Klasse zugeordnet; der Merkmalswert $x_2 = 3$ liegt ebenfalls an der Klassengrenze und wird der zweiten Klassen zugeordnet. Es ergibt sich dann die Häufigkeitsverteilung in Tabelle 2.2.

Tabelle 2.2. Klasseneinteilung

	1. Klasse	2. Klasse	3. Klasse
	$(x_0^* = 0, x_1^* = 1] \quad (x_1^* = 1, x_2^* = 3] \quad (x_2^* = 3, x_3^* = 6]$		
x_j	0.5	2	4.5
$n(x_j)$	1	2	2

Wird nun die Klassengrenze um 0.5 reduziert, so ergibt sich folgende veränderte Verteilung der klassierten Werte, obwohl es sich um die gleiche Urliste handelt (siehe Tabelle 2.3).

Tabelle 2.3. Klasseneinteilung

	1. Klasse	2. Klasse	3. Klasse
	$(x_0^* = -0.5, x_1^* = 0.5] \quad (x_1^* = 0.5, x_2^* = 2.5] \quad (x_2^* = 2.5, x_3^* = 5.5]$		
x_j	0	1.5	4
$n(x_j)$	0	2	3

Definition 2.22. Sei x_i ein kardinal skaliertes Merkmal, so wird die Differenz zweier aufeinander folgender Klassengrenzen

$$\Delta_j = x_j^* - x_{j-1}^* \quad \text{mit } j = 1, \dots, m \quad (2.13)$$

als **Klassenbreite** Δ_j bezeichnet.

Nach Möglichkeit sollten alle Klassen gleich breit sein (äquidistant). Jedoch ist die Verwendung gleich breiter Klassen nicht immer sinnvoll. Wenn sehr viele Beobachtungswerte in einem kleinen Bereich der Merkmalsausprägungen liegen und ein geringer Rest in einem sehr weiten Bereich, sollte man im kleinen Bereich vieler Werte fein klassieren, während man im übrigen Bereich breite Klassen wählen sollte. Ferner sollte darauf geachtet werden, dass die Klassen gleichmäßig besetzt sind, denn viele Berechnungen mit klassierten Werten unterstellen eine Gleichverteilung

der Merkmalswerte innerhalb der Klassen (siehe z. B. auch Beispiel 4.24 auf Seite 66 und Beispiel 17.3, Seite 427).

Beispiel 2.27. Bei einer Einkommensverteilung ist der untere und mittlere Einkommensbereich wesentlich stärker besetzt als der hohe und man wird daher in dem Bereich niedrigerer Einkommen eine kleinere Klassenbreite wählen, um nicht zu viel Information zu verlieren. Im Bereich hoher Einkommen wählt man hingegen breitere Klassen, um nicht insgesamt zu viele Klassen zu bekommen, von denen viele nur gering oder gar nicht besetzt sind. Die Nettoeinkommen der Erwerbstätigen in Ost- und Westdeutschland in Tabelle 2.4 sind hier in $m = 8$ Klassen eingeteilt.

Tabelle 2.4. Nettoeinkommen der Erwerbstätigen 1998

Klasse	von	bis	West	Ost
j	in DM		in %	
1	unter 600		6.6	6.6
2	600	1 000	8.7	8.0
3	1 000	1 400	6.8	11.2
4	1 400	1 800	7.5	17.3
5	1 800	2 200	11.8	20.1
6	2 200	3 000	25.2	23.5
7	3 000	4 000	16.9	8.8
8	4 000	und mehr	16.6	4.5

(aus [55, Tabelle 19])

In der Regel möchte man die durch einen Wert repräsentieren. Dies geschieht bei stetigen Merkmalen meistens durch die Klassenmitte (siehe Tabelle 2.2 und Tabelle 2.3, Seite 19). Bei diskreten Klassen repräsentiert der Merkmalswert selbst die Klasse.

Definition 2.23. Sei x_i eine kardinal skaliertes Merkmal so wird die Mittelung zwischen der Klassenuntergrenze x_{j-1}^* und der Klassenobergrenze x_j^* als **Klassenmitte** x_j bezeichnet.

$$x_j = \frac{(x_j^* + x_{j-1}^*)}{2} \quad \text{mit } j = 1, \dots, m \tag{2.14}$$

Man geht davon aus, dass sich alle Beobachtungswerte einer Klasse gleichmäßig über die Klasse verteilen (**Gleichverteilung innerhalb der Klasse**), so dass die Klassenmitte Repräsentant der Klasse ist. Mit diesen Klassenrepräsentanten und den entsprechenden Häufigkeiten werden dann die Berechnungen durchgeführt.

Bei der Klassenbildung erfolgt eine Zusammenfassung von Merkmalsausprägungen. Damit ist zwangsläufig ein Verlust an Information verbunden, da sich die einzelnen Ausprägungen der in eine Klasse fallenden Merkmalswerte nachträglich

nicht mehr feststellen lassen. Verzerrungen können sich zusätzlich ergeben, wenn die Beobachtungswerte nicht gleichmäßig in der Klasse verteilt sind (siehe hierzu das Beispiel 4.24 auf Seite 66).

Eine weitere Schwierigkeit bei der Klassenbildung stellen die so genannten offenen **Randklassen** dar. Sie werden eingeführt, um die Zahl der Klassen zu begrenzen und dennoch alle Werte zu erfassen oder weil die Angabe des kleinsten bzw. größten Wertes nicht möglich ist.

Beispiel 2.28. Werden bei einer Befragung die Einkommen durch bestimmte Einkommensklassen erfragt, so ist das größte Einkommen nicht im Vorhinein bekannt. Ist man hingegen in Besitz der Einzelangaben und kennt den größten Wert, so möchte man beispielsweise aus Gründen der Diskretion den größten Wert nicht durch die oberste Klassengrenze mitteilen.

Bei offenen Randklassen ergibt sich die Schwierigkeit der Bestimmung des repräsentativen Wertes. Als Klassenmitte kann man einen geschätzten Wert oder den aus den ursprünglichen Merkmalswerten berechneten Mittelwert der Klasse, sofern möglich, verwenden.

Beispiel 2.29. Fortsetzung von Beispiel 2.27 (Seite 20): Die Berechnung der Klassenmitte für die offenen Randklassen aus den Ursprungswerten scheidet bei dem Beispiel aus, da die Ursprungswerte nicht bekannt sind. Die Klassenmitte der unteren Randklasse könnte hier z. B. durch den Abstand der zweiten Klasse zur Klassenmitte approximiert werden. In diesem Fall ergäbe dies die Klassenmitte: $x_0 = x_1^* - (x_2 - x_1^*) = 600 - (800 - 600) = 400$. Die Bestimmung der Klassenmitte für die obere Randklasse könnte analog erfolgen: $x_8 = x_7^* + (x_7^* - x_7) = 4\,000 + (4\,000 - 3\,500) = 4\,500$. Wenn einem der Klassenrepräsentant für die offenen Randklassen zu hoch bzw. zu niedrig erscheint, wäre es auch möglich, die zweite Klassenbreite auf die erste Klasse und die vorletzte Klassenbreite auf die letzte Klasse zu übertragen: $x_1 = x_1^* - \Delta_2 = 600 - 400 = 200$ bzw. $x_8 = x_7^* + \Delta_7 = 4\,000 + 1\,000 = 5\,000$. Auch jede andere Berechnung ist grundsätzlich möglich, solange sie durch Plausibilitätsüberlegungen gerechtfertigt ist.

2.7 Übungen

Übung 2.1. Geben Sie für folgende Fragestellungen die statistischen Massen und Einheiten an und grenzen Sie diese räumlich, sachlich und zeitlich ab:

- Wählerverhalten in einer Landtagswahl
- Durchschnittliche Studiendauer von Studenten an deutschen Hochschulen bis zum Abschluss

Übung 2.2. Geben Sie zu den folgenden statistischen Massen an, ob es sich um Bestands- oder Ereignismassen handelt:

- Eheschließungen in Bielefeld

- Wahlberechtigte Bundesbürger
- Zahl der Verkehrsunfälle 1998 in Deutschland
- Auftragseingang

Übung 2.3. Geben Sie zu folgenden Merkmalen mögliche Merkmalsausprägungen an:

- Haarfarbe
- Einkommen
- Klausurnote
- Schulabschluss
- Freizeitbeschäftigung

Welche der Merkmale sind häufbar?

Übung 2.4. Geben Sie zu den folgenden Merkmalen an, auf welcher Skala die Merkmalsausprägungen gemessen werden können: Semesterzahl, Temperatur, Klausurpunkte, Längen- und Breitengrade der Erde, Studienfach, Handelsklassen von Obst.

Übung 2.5. Geben Sie für die beiden Klasseneinteilungen im Beispiel 2.26 auf Seite 19 jeweils die Verteilung der relativen Häufigkeiten die an.

Übung 2.6. Fuhr vor einigen Jahren noch jeder zehnte Autofahrer zu schnell, so ist es heute nur jeder fünfte. Doch auch fünf Prozent sind zu viel. Welcher Fehler wird in dieser Mitteilung begangen?

Statistik

Datenanalyse und Wahrscheinlichkeitsrechnung

Kohn, W.

2005, XVI, 624 S., Softcover

ISBN: 978-3-540-21677-3