Statistik

Lehr- und Handbuch der angewandten Statistik

von

o. Prof. Dr. Joachim Hartung

Thomas-L.-Saaty-Preisträger

Dr. Bärbel Elpelt

Fakultät Statistik der
Technischen Universität Dortmund

Prof. Dr. Karl-Heinz Klösener

Fachhochschule Trier

15., überarbeitete und wesentlich erweiterte Auflage

Oldenbourg Verlag München

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© 2009 Oldenbourg Wissenschaftsverlag GmbH
Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
oldenbourg.de

Das Werk einschließlich aller Abbildungen ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Lektorat: Wirtschafts- und Sozialwissenschaften, wiso@oldenbourg.de
Herstellung: Anna Grosser
Coverentwurf: Kochan & Partner, München
Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Druckhaus „Thomas Müntzer“ GmbH, Bad Langensalza

ISBN 978-3-486-59028-9

Kapitelverzeichnis

Einführung und Grundlagen 1

KAPITEL I: Aufbereitung und Darstellung von Datenmaterial — Deskriptive Statistik 15

KAPITEL II: Grundbegriffe der Wahrscheinlichkeitsrechnung 91

KAPITEL III: Statistische Schlußweisen 123

KAPITEL IV: Spezielle Verteilungen und statistische Schlüsse über Kenngrößen von Verteilungen mittels einer Meßreihe (Stichprobe) 143

KAPITEL V: Aspekte der Datengewinnung — Stichprobentheorie, Meßfehler, Ausreißertests, Datentransformationen, Versuchsplanung, Klinische Versuche, Skalierung 269

KAPITEL VI: Qualitätskontrolle 381

KAPITEL VII: Analyse diskreten Datenmaterials in Form von Kontingenztafeln 407

KAPITEL VIII: Vergleich zweier Meßreihen (Stichproben) 505

KAPITEL IX: Die Korrelation von Merkmalen 545

KAPITEL X: Regressionsanalyse 569

KAPITEL XI: Varianzanalyse 609

KAPITEL XII: Zeitreihenanalyse 637

KAPITEL XIII: Analyse von Lebensdauern und Zuverlässigkeit von Systemen 745

KAPITEL XIV: Explorative Datenanalyse (EDA) und Robuste Verfahren 825

KAPITEL XV: Meta-Analyse zur Kombination von Studien, Experimenten und Prognosen 887

ANHANG 1023

Inhaltsverzeichnis

Vorworte XXVII

Ein kurzer Überblick XXXI

Einführung und Grundlagen **1**

1. Was ist Statistik 1
2. Das Experiment 3
3. Die Erhebung 5
4. Zur Statistik und ihren philosophischen Voraussetzungen 7
5. Zur Geschichte der Statistik 10

Kapitel I: Aufbereitung und Darstellung von Datenmaterial — Deskriptive Statistik **15**

1. Grundlegende Begriffe und Überblick 15

1.1. Untersuchungseinheiten, Merkmale und Merkmalsausprägungen 15
1.2. Charakterisierung von Merkmalen 16
1.3. Grundgesamtheit und Stichprobe 18
1.4. Überblick über die Methoden der deskriptiven Statistik 19

2. Der Häufigkeitsbegriff 20

2.1. Absolute und relative Häufigkeiten 20
2.2. Die graphische Darstellung von Häufigkeiten 21
2.3. Die empirische Verteilungsfunktion 23

3. Der Häufigkeitsbegriff bei Klassenbildung 24

3.1. Die Klassenbildung 26
3.2. Absolute und relative Häufigkeiten bei Klassenbildung 26
3.3. Die graphische Darstellung von Häufigkeiten bei Klassenbildung 27
3.4. Die empirische Verteilungsfunktion bei Klassenbildung 28

4. Lagemaße von Häufigkeitsverteilungen 31

4.1. Das arithmetische Mittel 31
4.2. Der Median und das α-Quantil 32
4.2.1. Der Median einer Beobachtungsreihe 32
4.2.2. Das α-Quantil einer Beobachtungsreihe 34
4.3. Der Modalwert 35
4.4. Das geometrische Mittel und das harmonische Mittel 35
4.5. Einige Bemerkungen zu den Lagemaßen 37

5. Streuungsmaße von Häufigkeitsverteilungen 40

5.1. Die Spannweite 40
5.2. Der Quartilsabstand 41
5.3. Die mittlere absolute Abweichung vom Median 42
5.4. Varianz, Standardabweichung und Variationskoeffizient 43
5.4.1. Die Varianz 44
5.4.2. Die Standardabweichung 46
5.4.3. Der Variationskoeffizient 47

5.5. Die Schiefe und der Exzeß 47
5.5.1. Die Schiefe einer Häufigkeitsverteilung 47
5.5.2. Der Exzeß einer Häufigkeitsverteilung 49
6. Konzentrationsmaße für Häufigkeitsverteilungen 50
6.1. Die Lorenzkurve 50
6.2. Das Lorenzsche Konzentrationsmaß; der Gini-Koeffizient 52
7. Verhältniszahlen 55
7.1. Gliederungszahlen 55
7.2. Beziehungszahlen 56
7.2.1. Verursachungszahlen 56
7.2.2. Entsprechungszahlen 56
7.3. Indexzahlen 57
7.3.1. Meßzahlen 57
7.3.2. Standardisierung von Meßzahlen, Sterbeziffern 60
7.3.3. Zusammengesetzte Indexzahlen 62
A. Der Wertindex 62
B. Preisindizes nach Laspeyres und nach Paasche 63
C. Ein Beispiel 63
D. Mengenindizes nach Laspeyres und nach Paasche 65
E. Preisbereinigung; Deflationierung 65
F. Preis- und Mengenindizes als gewogene Mittel von Meßzahlen: Subindizes 66
7.3.4. Vergleich von Preisindizes nach Laspeyres und nach Paasche; der Fishersche Idealindex; der Preisindex nach Lowe 70
8. Die empirische Ausfallrate 70
9. Darstellung zweidimensionalen Zahlenmaterials und deskriptive Korrelationsrechnung 72
9.1. Die Kontingenztafel 72
9.2. Der Korrelationskoeffizient nach Bravais-Pearson 73
9.3. Der Fechnersche Korrelationskoeffizient 78
9.4. Der Spearmansche Rangkorrelationskoeffizient 79
9.5. Der Kendallsche Rangkorrelationskoeffizient 81
9.6. Der Yulesche Assoziationskoeffizient für die Vierfeldertafel 82
10. Praktische Berechnung einiger Kenngrößen 83
10.1. Berechnung des arithmetischen Mittels und der Standardabweichung 83
10.2. Berechnung der mittleren absoluten Abweichung vom Median 87

Kapitel II: Grundbegriffe der Wahrscheinlichkeitsrechnung 91

1. Ereignisse und Zufallsexperimente 91
2. Wahrscheinlichkeiten 93
3. Kombinatorik und Beispiele für die Berechnung von Laplace-Wahrscheinlichkeiten 96
3.1. Permutationen 96
3.2. Kombinationen 96
3.3. Beispiele zur Berechnung von Laplace-Wahrscheinlichkeiten 97

4. Bedingte Wahrscheinlichkeiten und Unabhängigkeit 98

5. Die Bayessche Formel 102

6. Zufallsvariable und Verteilungen 103

7. Unabhängigkeit und Funktionen von Zufallsvariablen 108

7.1. Unabhängigkeit von Zufallsvariablen 108
7.2. Funktionen von Zufallsvariablen 108
A. Lineare Transformation; Normalverteilung 109
B. Summe; Faltung; Binomialverteilung 109
C. Maximum 111
D. Minimum 111

8. Kenngrößen von Zufallsvariablen 112

8.1. Lageparameter 112
A. Der Erwartungswert 112
B. Der Median und andere Quantile 114
C. Der Modalwert 116
8.2. Streuungsparameter 116
A. Die Varianz und die Standardabweichung; standardisierte Zufallsvariable; Tschebyscheffsche Ungleichung 116
B. Der Variationskoeffizient 117
C. Der Quartilsabstand 118
8.3. Momente von Zufallsvariablen; Schiefe; Exzeß 118
8.4. Kovarianz und Korrelation von Zufallsvariablen 119

9. Grenzwertsätze 121

Kapitel III: Statistische Schlußweisen 123

1. Schätzen von Parametern 124
A. Momentenmethode 126
B. Maximum-Likelihood-Methode 126
C. Methode der kleinsten Quadrate 128

2. Konfidenzintervalle 129

3. Prognose- und Toleranzintervalle 132

4. Statistische Tests 133

5. Beurteilungskriterien für statistische Tests 137

6. Arten von Hypothesen und allgemeine Bemerkungen 138

7. Nichtparametrische (verteilungsfreie) Verfahren 139

8. Zufällige Auswahl, Randomisation 141

9. Notation von Zufallsvariablen 142

Kapitel IV: Spezielle Verteilungen und Statistische Schlüsse über Kenngrößen von Verteilungen mittels einer Meßreihe (Stichprobe) 143

1. Die Normalverteilung und daraus abgeleitete Verteilungen 143
1.1. Die Normalverteilung und ihre Bedeutung 143
1.2. Einige in enger Beziehung zur Normalverteilung stehende Verteilungen 148
1.2.1. Aus der Normalverteilung abgeleitete Verteilungen 148
A. Die gestutzte Normalverteilung 148
B. Die Lognormalverteilung 151
1.2.2. Prüfverteilungen 152
A. Die χ^2-Verteilung 152
B. Die t-Verteilung 154
C. Die F-Verteilung 156
1.3. Punktschätzungen und Konfidenz-, Prognose- und Toleranzintervalle bei normalverteilter Grundgesamtheit 157
1.3.1. Schätzen der Parameter μ und σ^2 157
1.3.2. Konfidenzintervalle für μ, σ^2 und σ 160
1.3.3. Prognose- und Toleranzintervalle 163
1.4. Bestimmung von benötigten Stichprobenumfängen bei Intervallschätzungen 166
1.4.1. Einhaltung absoluter Genauigkeiten 166
1.4.2. Einhaltung prozentualer Genauigkeiten 173
A. Das Variationszahlverfahren 173
B. Das Streuzahlverfahren 175
1.5. Testen von Parameter-Hypothesen und Bestimmung des benötigten Stichprobenumfangs 178
1.5.1. Testen von Hypothesen über die Parameter einer normalverteilten Grundgesamtheit 178
A. Hypothesen über den Mittelwert μ 178
B. Hypothesen über die Varianz σ^2 179
1.5.2. Bestimmung des Stichprobenumfangs n beim Testen von Hypothesen über den Erwartungswert μ einer normalverteilten Grundgesamtheit bei vorgegebenem Fehler 1. Art α und Fehler 2. Art β 181
1.6. Anpassungstests an die Normalverteilung 182
A. Der χ^2-Anpassungstest 182
B. Der Kolmogoroff-Smirnov-Anpassungstest 183
C. Ein Beispiel 186
C1. χ^2-Anpassungstest 186
C2. Kolmogoroff-Smirnov-Anpassungstest 187
1.7. Weitere Verfahren zum Testen von Normalverteilungshypothesen 189
1.7.1. Test auf Schiefe und Exzeß 189
1.7.2. Überprüfung der Normalverteilungsannahme mit Hilfe von Wahrscheinlichkeitspapier 190

2. Die Gleichverteilung und die Dreiecksverteilung 192
2.1. Die stetige Gleichverteilung 192
2.1.1. Die eindimensionale Gleichverteilung und ihre Anwendung in der Computersimulation 192
2.1.2. Die zweidimensionale Gleichverteilung 194
2.2. Die Dreiecksverteilung 195

2.3. Punkt- und Intervallschätzungen für die Gleichverteilung 197
2.4. Der χ^2-Anpassungstest für die Gleichverteilung 198
3. Einige diskrete Verteilungen 199
3.1. Die Binomialverteilung 199
3.1.1. Punkt- und Intervallschätzung des Parameters p 202
3.1.2. Testen von Hypothesen über den Parameter p 205
3.1.3. Bestimmung des Stichprobenumfangs n beim Testen von Hypothesen über den Parameter p einer Binomialverteilung bei vorgegebenen Fehlern 1. und 2. Art 206
3.2. Die hypergeometrische Verteilung 207
3.2.1. Punktschätzungen für die Hypergeometrische Verteilung 208
3.3. Die Multinomialverteilung 209
3.3.1. Konfidenzbereich für die Multinomialverteilung 211
3.4. Die Poissonverteilung 212
3.4.1. Punkt- und Intervallschätzung für den Parameter λ einer Po(λ)-Verteilung 214
3.4.2. Test über den Parameter λ einer Po(λ)-Verteilung 214
3.4.3. Der χ^2-Anpassungstest für die Poissonverteilung 216
4. Einige Lebensdauerverteilungen 218
4.1. Die Exponentialverteilung 219
4.1.1. Punkt- und Intervallschätzung für den Parameter einer Ex(λ)-Verteilung 220
4.1.2. Tests von Hypothesen über den Parameter einer Ex(λ)-Verteilung 223
4.1.3. Der χ^2-Anpassungstest für die Exponentialverteilung 225
4.1.4. Der Kolmogoroff-Smirnov-Anpassungstest für die Exponentialverteilung 226
4.2. Die Weibullverteilung 230
4.2.1. Schätzen der Parameter α und β der Weibullverteilung 232
4.3. Die IDB-Verteilung (Hjorth-Verteilung) 232
4.3.1. Schätzen der Parameter α, β und γ der IDB-Verteilung 234
4.4. Die Erlang-n-Verteilung 234
5. Nichtparametrische Test- und Schätzmethoden im Ein-Stichproben-Fall 235
5.1. Konfidenzintervalle und Tests für Quantile 235
5.1.1. Ein Konfidenzintervall für Quantile 236
5.1.2. Tests für Quantile 237
5.2. Nichtparametrische Toleranzintervalle 238
5.3. Konfidenzstreifen für eine unbekannte Verteilungsfunktion 240
5.4. Nichtparametrische Einstichproben-Lokationsvergleiche und Tests auf Trend 242
5.4.1. Der Zeichentest 242
5.4.2. Der Vorzeichenrangtest nach Wilcoxon 243
5.4.3. Tests auf Trend 247
A. Der Test von Cox und Stuart 247
B. Der Test nach Mann 249
6. Sequentielle Quotiententests 251
6.1. Der sequentielle Quotiententest für die Binomialverteilung 252
6.2. Sequentieller Quotiententest für den Erwartungswert einer Normalverteilung 259
6.3. Sequentieller Quotiententest für eine Exponentialverteilung 261

Kapitel V: Aspekte der Datengewinnung — Stichprobentheorie, Meßfehler, Ausreißertests, Datentransformationen, Versuchsplanung, Klinische Versuche, Skalierung 269

1. Abriß der klassischen Stichprobentheorie am Beispiel der Inventur auf Stichprobenbasis 269
1.1. Die Stichprobe 270
1.2. Überlegungen und Vorgehensweisen bei Stichprobenerhebungen 271
1.3. Verteilungsannahmen bei Stichprobenerhebungen 272
1.4. Die einfache Zufallsauswahl 273
1.5. Geschichtete Zufallsauswahl 278
1.5.1. Die optimale Aufteilung (Neyman-Tschuprow-Aufteilung) 282
1.5.2. Die proportionale Aufteilung 285
1.5.3. Aufteilung nach Auswahl der Stichprobe 286
1.5.4. Genauigkeitsvergleiche 287
1.6. Klumpenstichprobenverfahren 288
1.6.1. Einstufige Auswahlverfahren 289
1.6.2. Mehrstufige Auswahlverfahren 291
2. Weitere Verfahren der Stichprobentheorie 292
2.1. Ziehen mit und ohne Zurücklegen 292
2.2. Schätzen von Anteilen 293
2.3. Die systematische Stichprobe 295
2.4. Stichproben mit ungleichen Auswahlwahrscheinlichkeiten 297
2.5. Die Formel von Horwitz-Thompson 299
2.6. Verhältnis-, Differenzen- und Regressionsschätzung, gebundene und freie Hochrechnung 300
2.6.1. Die Verhältnisschätzung 300
2.6.2. Die Differenzenschätzung 303
2.6.3. Die Regressionsschätzung 303
2.7. Zweiphasige Problemstellungen 304
3. Probleme bei der praktischen Durchführung einer Erhebung 305
3.1. Die Abgrenzung der Grundgesamtheit 305
3.2. Endliche und unendliche sowie fiktive Grundgesamtheiten 306
3.3. Auswahltechniken und Erhebungsprobleme 307
3.4. Probleme im Zusammenhang mit Befragungen 309
3.4.1. Fragestellung und Fragebogen 309
3.4.2. Typen von Befragungen 310
3.4.3. Das Problem der Nichtbeantwortung 311
3.5. Vergleich zwischen den Schichten 313
3.6. Stichprobenverfahren in der Marktforschung 314
3.6.1. Marktforschung — Zielsetzungen und Problemstellungen 314
3.6.2. Beurteilungsstichproben in der Marktforschung 316
A. Typische Auswahl 316
B. Auswahl nach dem Konzentrationsprinzip 317
C. Quotenauswahl 318
3.7. Die Bedeutung der Stichprobenverfahren 320
4. Theorie der Meßfehler, Ausreißertests, Datentransformationen 320
4.1. Der Meßfehler bei der Datengewinnung 321
4.2. Das Gaußsche Fehlerfortpflanzungsgesetz 326

4.3. Kontrolle und Erfassung von Meßfehlern ... 332
4.3.1. Kontrolle und Ermittlung systematischer Fehler ... 332
4.3.2. Die Verwendung von Kontrollkarten ... 335
4.3.3. Ringversuche: Inter- und Intralaboratorielle Vergleiche ... 337
4.3.4. Präzision, Spezifität, Richtigkeit und Sensibilität von Meß- und Analyseverfahren ... 341
4.4. Das Ausreißerproblem ... 343
A. Der David-Hartley-Pearson-Test ... 344
B. Der Grubbs-Test ... 345
C. Dixon's r-Statistiken ... 346
D. Test auf ein Ausreißerpaar ... 347
4.5. Transformationen ... 349
A. Die reziproke Transformation ... 349
B. Die Wurzel-Transformation ... 349
C. Die Logarithmische Transformation ... 351
D. Die Box-Cox-Transformation ... 352
E. Die Arcus-Sinus-Transformation ... 352
F. Die Fishersche Z-Transformation ... 354

5. Allgemeine Aspekte der Planung von Versuchen ... 354

6. Anlage von klinischen Versuchen ... 363

6.1. Ethische Probleme bei klinischen Versuchen ... 366
6.2. Auswahl und Zuordnung von Versuchspersonen ... 367
6.2.1. Die retrospektive Zuordnung ... 368
6.2.2. Zuordnung auf freiwilliger Basis ... 369
6.2.3. Pseudaleatorische und aleatorische Zuordnung ... 370
6.2.4. Einige weitere Zuordnungsverfahren ... 370
6.3. Die Vergleichbarkeit der Versuchsergebnisse ... 370
6.4. Auto- und Heterosuggestion, Blindversuche ... 371
6.5. Sequentielle Studien ... 372
6.6. Ein Beispiel ... 373

7. Skalierung von Merkmalsausprägungen und Testergebnissen ... 374

Kapitel VI: Qualitätskontrolle ... 381

1. Stichprobenpläne in der Eingangs- und Endkontrolle ... 381

1.1. Einfache Stichprobenpläne für qualitative Merkmale ... 383
A. Vorgabe zweier Punkte der Operationscharakteristik ... 384
B. Vorgabe des Indifferenzpunktes und der Steilheit ... 387
1.2. Mehrfache und sequentielle Stichprobenpläne für qualitative Merkmale ... 389
A. Doppelte Stichprobenpläne ... 390
B. Sequentielle Stichprobenpläne ... 392
1.3. Stichprobenpläne für quantitative Merkmale ... 395

2. Laufende Kontrolle der Produktion (Kontrollkarten) ... 401

2.1. Laufende Kontrolle bei quantitativen Merkmalen ... 401
2.2. Laufende Kontrolle bei qualitativen Merkmalen ... 404

3. Kontinuierliche Stichprobenpläne ... 406

Kapitel VII: Analyse diskreten Datenmaterials in Form von Kontingenztafeln 407

1. Die 2 x 2-Felder-Tafel 411

1.1. Hypothesen für die 2 x 2-Felder-Tafel 412
A. Die Unabhängigkeitshypothese 412
B. Die Homogenitätshypothese 412
C. Beziehungen zwischen den Hypothesen 412
1.2. Tests auf Unabhängigkeit in der 2 x 2-Tafel 413
1.2.1. Der χ^2-Test 413
1.2.2. Exakte Tests 414
A. Ein exakter Test, der auf größere Kontingenztafeln übertragbar ist ..414
B. Der exakte Test von Fisher 416
1.2.3. Einseitige Hypothesen in 2 x 2-Tafeln 416
1.3. Tests auf Homogenität in der 2 x 2-Tafel 418
1.3.1. Die Durchführung der Tests 418
1.3.2. Der erforderliche Stichprobenumfang 419
A. Die Formel mittels Approximation der Gütefunktion des χ^2-Tests ...419
B. Die Arcus-Sinus-Formel 420
C. Die Formel nach Casagrande/Pike/Smith 420
D. Exakte Stichprobenumfänge 421
1.4. Tests auf Symmetrie in der 2 x 2-Tafel 422
1.4.1. Der McNemar-Test 423
1.4.2. Cochran's Q 423

2. Loglineare Modelle und Tests für r x s-Tafeln 425

2.1. Das loglineare Modell für die r x s-Tafel 425
2.1.1. Entwicklung des Modells am Beispiel der 2 x 2-Tafel 425
A. Aufstellung des Modells 425
B. Schätzen der Parameter 427
C. Die approximative Varianz der Schätzungen 428
D. Die Interpretation der Parameter 428
2.1.2. Das Modell für die allgemeine r x s-Tafel 429
2.2. Hypothesen und Tests in r x s-Tafeln 432
2.2.1. Einige Hypothesen für r x s-Tafeln 433
A. Die Unabhängigkeits- bzw. Homogenitätshypothese 433
B. Die bedingte Gleichverteilungshypothese 434
C. Die totale Gleichverteilungshypothese 434
2.2.2. Einige Testverfahren für r x s-Tafeln 435
A. χ^2- und Likelihood-Quotienten-Test 435
B. Die Statistiken τ_A und τ_B 439
C. Der Test auf Symmetrie nach Bowker 440

3. Assoziationsmaße für 2 x 2 und r x s-Tafeln 442

3.1. Assoziationsmaße in der 2 x 2-Kontingenztafel 442
3.1.1. Assoziationsmaße, die in Beziehung zum cross-product ratio q stehen ..442
A. Die Eigenschaften des cross-product ratio 442
B. Der Q-Koeffizient von Yule 443
C. Der Verbundenheitskoeffizient von Yule 444

D. Punkt-und Intervallschätzungen für die Yuleschen Assoziationsmaße ... 444
E. Eigenschaften der Yuleschen Assoziationsmaße ... 446
3.1.2. Assoziationsmaße, die in Beziehung zum Korrelationskoeffizienten Q stehen ... 446
A. Der Korrelationskoeffizient Q (Phikoeffizient) und seine Eigenschaften ... 446
B. Der Pearsonsche Kontingenzkoeffizient ... 449
3.2. Assoziationsmaße in allgemeinen 2-dimensionalen Kontingenztafeln ... 450
3.2.1. Assoziationsmaße, die in Beziehung zur χ^2-Statistik stehen ... 451
A. Der Pearsonsche Kontingenzkoeffizient für die r x s-Tafel ... 451
B. Der korrigierte Pearsonsche Kontingenzkoeffizient ... 451
C. Das Assoziationsmaß von Tschuprow ... 451
D. Das Assoziationsmaß von Cramer ... 452
E. Schätzung der Varianzen der Assoziationsmaße ... 452
F. Ein Beispiel ... 452
3.2.2. Die λ- und die τ-Maße ... 455
A. Die λ-Maße λ_A, λ_B und λ ... 456
B. Die τ-Maße τ_A, τ_B und τ ... 459
4. Loglineare Modelle und Tests für mehrdimensionale Kontingenztafeln ... 464
4.1. Die Parameter des saturierten Modells ... 465
A. Schätzen der Parameter des saturierten Modells ... 466
B. Varianz- und Intervallschätzungen für die Parameter des saturierten Modells ... 471
4.2. Testen von Hypothesen über die Parameter des saturierten Modells ... 477
4.2.1. Ein iteratives Verfahren zur Schätzung erwarteter Häufigkeiten unter einer Hypothese ... 477
4.2.2. Die Bestimmung der Freiheitsgrade ... 485
4.2.3. Die Partitionierung der Teststatistiken ... 488
5. Verteilungsannahmen, Logit-Modell und Adjustieren bei Kontingenztafeln ... 492
5.1. Kontingenztafeln und Verteilungen ... 492
5.1.1. Verteilungsannahmen bei Kontingenztafeln ... 492
5.1.2. Vergleich der Parameter mehrerer diskreter Verteilungen ... 495
A. Vergleich der Parameter von s Poissonverteilungen ... 495
B. Vergleich der Parameter verschiedener Binomialverteilungen ... 496
C. Vergleich der Parameter mehrerer Multinomialverteilungen ... 498
5.2. Das Logit-Modell bei Kontingenztafeln ... 498
5.3. Adjustieren von Kontingenztafeln ... 501

Kapitel VIII: Vergleich zweier Meßreihen (Stichproben) ... 505

1. Vergleich zweier unabhängiger Meßreihen ... 505
1.1. Lokationsvergleiche bei normalverteilter Grundgesamtheit ... 505
1.1.1. Tests und Konfidenzintervalle bei bekannten Varianzen σ_1^2 und σ_2^2 der Grundgesamtheiten ... 505
1.1.2. Tests und Konfidenzintervalle bei unbekannten aber gleichen Varianzen σ_1^2 und σ_2^2 der beiden Grundgesamtheiten ... 508

1.1.3. Tests und Konfidenzintervalle bei unbekannten und ungleichen Varianzen σ_1^2 und σ_2^2 der beiden Grundgesamtheiten 510
1.1.4. Bestimmung von Stichprobenumfängen bei Tests und Konfidenzintervallen 511
1.2. Verteilungsfreie Lokationsvergleiche 513
1.2.1. Der Wilcoxon-Rangsummentest, der U-Test von Mann-Whitney 513
1.2.2. Der Kolmogoroff-Smirnov-Test 520
1.3. Dispersionsvergleiche bei normalverteilten Grundgesamtheiten — Tests und Konfidenzintervalle 524
1.4. Verteilungsfreie Dispersionsvergleiche 526
1.4.1. Der Test von Ansari-Bradley-Freund, der Siegel-Tukey-Test 526
1.4.2. Der Test von Moses 529
1.5. Test auf Trend 531
2. Vergleich zweier abhängiger Meßreihen 533
2.1. Lokationsvergleiche bei normalverteilten Grundgesamtheiten — Tests, Konfidenzintervalle und Bestimmung der Stichprobenumfänge 534
A. Die Varianz σ_d^2 ist bekannt 534
B. Die Varianz σ_d^2 ist unbekannt 536
2.2. Dispersionsvergleiche bei normalverteilten Grundgesamtheiten 538
2.3. Verteilungsfreie Lokationsvergleiche 539

Kapitel IX: Die Korrelation von Merkmalen 545

1. Die Korrelation zweier normalverteilter Merkmale 546
2. Die Rangkorrelation zweier Merkmale 553
2.1. Der Spearmansche Rangkorrelationskoeffizient 553
2.2. Der Kendallsche Rangkorrelationskoeffizient 559
3. Die partielle Korrelation 561
3.1. Die partielle Korrelation zwischen normalverteilten Merkmalen 561
3.2. Der partielle Rangkorrelationskoeffizient nach Kendall 563
4. Die bi-partielle Korrelation 563
5. Die multiple Korrelation 564
6. Ein Test auf Unabhängigkeit von p Meßreihen 567

Kapitel X: Regressionsanalyse 569

1. Lineare Regression 573
1.1. Die Methode der kleinsten Quadrate 574
1.2. Schätzen der Fehlervarianz σ^2 578
1.3. Zusammenhang zwischen Regressions- und Korrelationsrechnung; das Bestimmtheitsmaß 578
1.4. Konfidenzintervalle und Testen von Hypothesen über die unbekannten Parameter α, β und σ^2 580
1.5. Konfidenz- und Prognosestreifen 582

1.6. Regression durch einen vorgegebenen Punkt, Regression ohne Absolutglied (eigentlich-lineare Regression) ... 584
2. Residualanalyse ... 585
3. Transformationen auf Linearität ... 587
4. Nichtlineare Regression und Schätzen des Maximums (Minimums) einer quadratischen Regressionsfunktion ... 589
5. Multiple Regression ... 595
6. Regression bei Fehlern in den Variablen ... 601
A. Die Varianz σ_1^2 ist bekannt ... 601
B. Die Varianz σ_2^2 ist bekannt ... 602
C. Der Quotient der Varianzen σ_1^2 und σ_2^2 ist bekannt ... 603
D. Das Berkson-Modell ... 604
7. Regressionsgerade nach Wald ... 605

Kapitel XI: Varianzanalyse ... 609

1. Vergleich von p unabhängigen Meßreihen (Stichproben) — einfache Varianzanalyse, vollständig randomisierter Versuchsplan ... 610
1.1. Testen auf signifikante Lokationsunterschiede ... 611
A. Der F-Test (normalverteilte Grundgesamtheit) ... 611
B. Der Test von Kruskal und Wallis ... 613
1.2. Simultane Vergleiche von p Mittelwerten ... 614
A. Die Tests von Scheffé und Tukey ... 616
A1. Scheffe-Test ... 616
A2. Tukey-Test ... 616
B. Der Test von Steel und Dwass ... 616
1.3. Dispersionsvergleiche ... 617
A. Der Bartlett-Test ... 617
B. Der Levene-Test ... 617
C. Scheffe's χ^2-Test ... 617
1.4. Modellbetrachtung ... 619
2. Das einfache Blockexperiment ... 619
2.1. Verfahren bei normalverteilter Grundgesamtheit ... 620
2.2. Verteilungsfreie Verfahren ... 622
3. Zweifache Varianzanalyse mit mehreren Beobachtungen pro Faktorstufenkombination (pro Zelle) ... 624
3.1. Modell mit Wechselwirkungen zwischen den Faktoren A und B ... 625
3.2. Modell ohne Wechselwirkungen zwischen den Faktoren ... 627
4. Die Komponenten der Streuung — Modelle der Varianzanalyse mit zufälligen Effekten (Modell II) ... 629
4.1. Die einfach hierarchische Klassifikation ... 630
4.2. Ein nicht-klassisches Varianzanalysemodell aus der Geodäsie ... 634

Kapitel XII: Zeitreihenanalyse 637

1. Deskriptive Methoden der Zeitreihenanalyse 640

1.1. Die Komponenten einer Zeitreihe 640
1.2. Nichtlineare Trendmodelle — Trendschätzung mittels nicht-linearer Regression 642
1.2.1. Die logistische Funktion 642
1.2.2. Die Mitscherlich-Funktion und die Gompertz-Kurve 648
1.2.3. Die allometrische Funktion 654
1.3. Trend- und Saison-Schätzung bzw. -Elimination durch Glättung bzw. Filterung 660
1.3.1. Gleitende Durchschnitte 660
1.3.2. Polynome und Splines 666
1.3.3. Die Differenzenmethode 668
1.3.4. Exponentielles Glätten 672
1.3.5. Lineare Filter 673
1.4. Autokovarianzen, Autokorrelationen und partielle Autokorrelationen 675

2. Selbsterklärende Zeitreihenmodelle und die Methode von Box und Jenkins 678

2.1. Autoregressive Prozesse (AR-Prozesse) 678
2.2. Moving average Prozesse (MA-Prozesse) 681
2.3. Gemischte Prozesse (ARMA-Prozesse) 682
2.4. Instationäre stochastische Prozesse, ARIMA- und SARIMA-Prozesse, Box-Cox-Transformation 684
2.5. Die Methode von Box und Jenkins 686
2.5.1. Modellidentifikation 686
2.5.2. Schätzen der Modellparameter. 688
2.5.3. Modellüberprüfung 690
2.5.4. Prognose und Prognosegüte 691
2.6. Ein Beispiel zur Methode von Box und Jenkins 694

3. Die Spektralanalyse 699

3.1. Komplexe Zahlen 700
3.2. Spektrum und Spektraldichte 701
3.3. Spektraldichten gefilterter Prozesse 703
3.4. Schätzen der Spektraldichte 709
3.4.1. Das Periodogramm und das Stichprobenspektrum 709
3.4.2. Geglättete Spektraldichteschätzungen, Spektral- und Lag-Fenster 711
3.4.3. Ein Beispiel zum Stichprobenspektrum 715
3.5. Die harmonische Analyse einer Zeitreihe 717
3.6. Das Berliner Verfahren zur Saisonbereinigung 722

4. Analyse des Zusammenhangs zweier Zeitreihen 727

4.1. Analyse im Zeitbereich 728
4.2. Analyse im Frequenzbereich — Kreuzspektralanalyse 728
4.2.1. Das Kreuzspektrum 729
4.2.2. Schätzung kreuzspektraler Größen 734

5. Gemischte Regressions-Zeitreihen-Modelle 735

5.1. Regressionsmodelle mit korrelierten Fehlern ... 736
5.1.1. Allgemeine Vorgehensweise ... 736
5.1.2. Regressionsmodelle mit AR(1)-Fehler-Prozeß ... 737
A. Die Cochrane-Orcutt-Methode ... 738
B. Der Durbin-Watson-Test ... 740
C. Prognose und Prognosegüte ... 741
D. Ein Beispiel ... 741
5.2. Autoregressive Regressionsmodelle ... 744
6. Tests auf Stationarität einer Zeitreihe [Dickey-Fuller Unit Root Test] und eines kreuzkorrelierten Zeitreihenpanels ... 744

Kapitel XIII: Analyse von Lebensdauern und Zuverlässigkeit von Systemen ... 745

1. Die Zuverlässigkeit von Komponenten und Systemen ... 746
1.1. Die Zuverlässigkeit elementarer Systeme ... 746
1.2. Zuverlässigkeitsschaltbilder ... 748
1.3. Darstellung monotoner Systeme mittels minimaler Pfade und minimaler Schnitte ... 750
1.4. Systemstrukturfunktion und Systemzuverlässigkeitsfunktion ... 751
1.4.1. Die Strukturfunktion monotoner Systeme ... 751
A. Bestimmung einer Strukturfunktion mittels disjunktiver Normalform ... 752
B. Bestimmung einer Strukturfunktion mittels minimaler Pfad- oder Schnittmengen ... 754
C. Gewinnung des Zuverlässigkeitsschaltbildes zu einer Strukturfunktion ... 755
1.4.2. Die Zuverlässigkeitsfunktion monotoner Systeme ... 756
1.5. Stochastische Assoziiertheit von Systemkomponenten ... 759
1.6. Klassifizierung der Zuverlässigkeitsfunktionen monotoner Systeme mit Komponenten gleicher Zuverlässigkeit ... 759
1.7. Methoden zur Erhöhung der Zuverlässigkeit — Redundanz bei Komponenten und Systemen, Systeme mit heißer und kalter Reserve ... 760
1.8. Systeme mit mehr als zwei Zuständen (Multi-State-Systems) ... 763
1.8.1. Die Bestimmung des Systemzustandes mittels minimaler Pfad- oder Schnittmengen ... 763
1.8.2. Kritische Pfadvektoren bei Multi-State-Systemen ... 764
1.9. Die Fehlerbaumanalyse ... 765
1.10. Systembetrachtungen bei mehrphasigen Missionen ... 769
1.10.1. Mehrphasige Missionen ... 769
1.10.2. Phasenzuverlässigkeiten und Missionszuverlässigkeit ... 770
1.10.3. Die Transformation mehrphasiger Missionen ... 772
2. Klassen von Lebensdauerverteilungen ... 774
2.1. IFR- und DFR-Verteilungen ... 776
2.1.1. Die Verteilungsklassen IFR und DFR ... 776
2.1.2. IFR- und DFR-Tests ... 777
A. Der Proschan-Pyke-Test ... 777
B. Der cttot-Test nach Epstein ... 778

2.2. NBU- und NWU-Verteilungen 779
2.2.1. Die Verteilungsklassen NBU und NWU 779
2.2.2. NBU- und NWU-Tests: Der Hollander-Proschan-Test 780
2.3. IFRA- und DFRA-Verteilungen 782
2.3.1. Die Verteilungsklassen IFRA und DFRA 782
2.3.2. IFRA- und DFRA-Tests: Der cttot-Test 784
2.4. NBUE- und NWUE-Verteilungen 784
2.4.1. Die Verteilungsklassen NBUE und NWUE 784
2.4.2. NBUE- und NWUE-Tests: Der Hollander-Proschan-Test 784
2.5. Beziehungen zwischen den Verteilungsklassen 786
2.6. Zensierte Lebensdauerprüfungen 787
3. Punkt- und Intervallschätzungen für die Parameter einiger spezieller Lebensdauerverteilungen 788
3.1. Das Modell exponentialverteilter Lebensdauern 788
3.1.1. Punkt- und Intervallschätzungen bei fest vorgegebener Beobachtungsdauer T_0 788
A. Experimente mit Ersetzung ausgefallener Objekte 788
B. Experimente ohne Ersetzung ausgefallener Objekte 789
3.1.2. Punkt- und Intervallschätzungen bei fest vorgegebener Zahl von Ausfällen 789
A. Experimente mit Ersetzung ausgefallener Objekte 789
B. Experimente ohne Ersetzung ausgefallener Objekte 790
3.2. Das Modell Weibull-verteilter Lebensdauern 791
3.2.1. Schätzen der Parameter α und β bei fest vorgegebener Anzahl von Ausfällen ohne Ersetzung 791
3.2.2. Schätzen der Parameter α und β bei unzensierten Lebensdauerprüfungen 793
3.3. Das Modell der Lognormalverteilung 793
3.4. Das Modell der Hjorth-Verteilung (IDB-Verteilung) 794
4. Zur Problematik zeitraffender Zuverlässigkeitsprüfungen 794
4.1. Die Extrapolation der zeitraffenden Prüfung an Normalbedingungen 796
4.1.1. Das Eyring-Modell 796
4.1.2. Das Arrhenius-Modell 796
4.1.3. Das verallgemeinerte Eyring-Modell 799
4.2. Screening-Tests, Burn-Ins 803
4.3. Labor- und Einsatzbedingungen 803
4.3.1. Der Einfluß von Strahlungen auf den Alterungsprozeß 803
4.3.2. Wertungsfaktoren für im Labor ermittelte Ausfallraten 804
5. Wartungs- und Erneuerungsüberlegungen 806
5.1. Wartung und Wartbarkeit, Erneuerung 806
A. Die Wartbarkeit von Systemen 806
B. Die Wartung von Systemen 806
C. Die Erneuerung von Systemen 807
5.2. Erneuerungsstrategien, Schranken der Erneuerungsfunktion 807
A. Altersabhängige Erneuerungsstrategie und Gruppenerneuerungsstrategie ..807
B. Der Erneuerungsprozeß und die Erneuerungsfunktion 807
C. Schranken der Erneuerungsfunktion 808

5.3. Die Zuverlässigkeit von Straßenverkehrssignalanlagen — Ein Beispiel 808
5.3.1. Straßenverkehrssignalanlagen-Systeme 809
A. Das System ohne Reserve 809
B. Das System mit heißer Reserve 810
C. Systeme mit kalter Reserve 811
C1. Das System mit kalter Macro-Reserve 811
C2. Das System mit kalter Micro-Reserve 812
5.3.2. Vergleich der Straßenverkehrssignalanlagen-Systeme 812
5.3.3. Die Wirtschaftlichkeit der Systeme 815
A. Inspektionszeiträume und Mindestzuverlässigkeit der Systeme 817
B. Systemkosten bei Mindestzuverlässigkeit der Systeme 818

6. Verfügbarkeit von Systemen und Instandhaltungsstrategien 820

6.1. Die Verfügbarkeit von Systemen 820
A. Momentane Verfügbarkeit und Dauerverfügbarkeit 820
B. Punkt- und Intervallschätzungen, Testen von Hypothesen über die Dauerverfügbarkeit 820
6.2. Methoden zur Erhöhung der Verfügbarkeit 822
A. Die Redundanzplanung 822
B. Die vorbeugende Instandsetzung 823
B1. Bereitschafts- und Präventivstrategien 823
B2. Periodische, sequentielle und optionale Strategien 823

7. Vergleichende Lebensdauerprüfungen: Log-Rank Test, Wilcoxon-Gehan Test, Tarone-Ware Tests und Cox Proportional Hazards Regression Model 824

8. Schätzen der [Über-] Lebensdauerverteilung: Kaplan-Meier Survival Analysis 824

Kapitel XIV: Explorative Datenanalyse (EDA) und Robuste Verfahren 825

1. Verfahren für einzelne Merkmale in der EDA 827

1.1. Empirische Kenngrößen 827
1.2. Empirische Kenngrößen bei gruppierten Daten 831
1.3. Datentransformationen 832
1.4. Box-Plots 835
1.5. Stamm-und-Blätter-Darstellungen 838
1.6. Histogramm und empirische Verteilungsfunktion 839
1.7. Empirische Dichten 840
1.8. Wurzeldiagramme 844
1.9. Q-Q-Plots zur Überprüfung von Verteilungsannahmen 847

2. Verfahren für zwei Merkmale in der EDA 849

2.1. Glätten zweidimensionaler Punktescharen 850
2.2. Explorative Regressionsgeraden 852
2.3. Linearisieren zweidimensionaler Punktescharen 854

3. Verfahren für mehrdimensionale Daten in der EDA 857

3.1. Der Scatter-Plot 858
3.2. Weitere explorative Verfahren für mehrdimensionale Daten 860

4. Robuste Schätzungen 861
4.1. Charakterisierung von Robustheitseigenschaften 862
4.1.1. Die Sensitivitätskurve 862
4.1.2. Die Einflußkurve 863
4.1.3. Der Bruchpunkt. 864
4.2. Robuste Skalenschätzer 864
4.2.1. Der Median der absoluten Abweichungen vom Median 865
4.2.2. Der Quartilsabstand 866
4.3. M-Schätzer für die Lokation 866
4.3.1. Huber-k-Schätzer 869
4.3.2. Andrews' wave und Tukey's biweight 872
4.3.3. Die Berechnung von M-Schätzern 874
4.3.4. M-Schätzer für einige typische Beispiele 878
4.4. L-Schätzer für die Lokation 880
4.4.1. Das α-getrimmte Mittel 880
4.4.2. Das α-winsorisierte Mittel 881
4.4.3. Das α-Gastwirth-Cohen-Mittel 882
4.5. R-Schätzer für die Lokation 883

Kapitel XV: Meta-Analyse zur Kombination von Studien, Experimenten und Prognosen 887

1. Kombination von Wissen — Combining Information 887
1.1. Astronomie 887
1.2. Evidence Based Medicine 888
1.3. Verhaltensforschung — Kombination unabhängiger p-Werte und Tests 891
1.4. Landwirtschaft — Genetik — Heterogenitätsmaße 892
1.5. Gesamtwirtschaftliche Vorhersagen 893
1.6. Betriebswirtschaftliche Prognosen 894
1.7. Stationarität eines kreuz-korrelierten Zeitreihenpanels — Kombination abhängiger p-Werte 895
1.8. Umwelt — Physik — Technik 895
1.9. Pareto-Optimale Portfolioselektion im Wertpapiermanagement 897
1.10. Normalverteilungs-Annahmen 899
2. Fixed Effects Model (FE) der Meta-Analyse 899
2.1. Darstellung des Modells und Gauß-Markov-Schätzer 900
2.2. Kombinierte Schätzung des Gesamteffektes 900
2.3. Konfidenzintervalle für den Gesamteffekt 901
2.4. Homogenitätstest von Cochran 901
2.5. Beispiel: Klinische Prüfungen zu Multipler Sklerose MS [Zweiarmige Studien] 902
2.6. Korrelationen 903
2.6.1. Kombination von Korrelationskoeffizienten 904
2.6.2. Beispiel: Korrelation zwischen Rückenspeckdicke und Fleisch-Fett-Verhältnis im Schlachtkörper von Schweinen 904
3. Random Effects Model (RE) der Meta-Analyse 906
3.1. Schätzung des Heterogenitätsparameters 906
3.2. Gesamtergebnis und Konfidenzintervalle 907

3.3. Beispiel: MS-Studien im RE-Model ... 908
3.4. Beispiel: Landwirtschaft — Besamungsquote von Bullen [Einarmige Studien] ... 909
3.5. Studien-Cluster [Hartung-Knapp] ... 910
3.6. Beispiel: Fortsetzung Landwirtschaft ... 912
4. Diskrete Beobachtungen ... 913
4.1. Risiko Differenz RD ... 914
4.2. Odds Ratio OR ... 915
4.3. Relatives Risiko RR und Noninferiority ... 915
4.4. Beispiel: Klinische Prüfungen zu Funktioneller Dyspepsie [Zweiarmige Studien] ... 917
4.5. Odds Ratio Schätzer von Mantel-Haenzsel und Peto oder Yusuf-Peto ... 918
4.6. Poissonverteilte Daten ... 920
4.7. Beispiel Umwelt — Technik: Asbest-Faser Zählung [Einarmige Studien] ... 921
5. Überlebenszeiten und Time-to-Event-Daten ... 924
5.1. Tests auf Gleichheit zweier Survival-Funktionen ... 925
5.1.1. Der Log-Rank Test ... 926
5.1.2. Der Wilcoxon-Gehan Test und Tarone-Ware Tests ... 928
5.1.3. Beispiel: Progressionsfreies Überleben bei Tumorerkrankung ... 928
5.2. Kombination von vergleichenden Überlebenszeitstudien ... 930
5.2.1. Kombination der Hazard Ratios ... 931
5.2.2. Beispiel Progressionsfreies Überleben PFS [3 Studien]: Kombiniertes Konfidenzintervall für den Hazard Ratio ... 931
5.3. Schätzen der Überlebenswahrscheinlichkeiten und Konfidenzbereiche ... 932
5.3.1. Kaplan-Meier Schätzer der Survival-Funktion und Bestimmung eines Konfidenzstreifens ... 933
5.3.2. Schätzen der medianen Überlebenszeit und Konfidenzintervall ... 933
5.3.3. Beispiel Progressionsfreies Überleben PFS: Bestimmung der Kaplan-Meier PFS-Kurven und der medianen Überlebenszeiten ... 934
5.4. Kombination von Überlebenszeiten ... 935
5.4.1. Kombination der Kaplan-Meier Überlebenskurven ... 936
5.4.2. Kombination der medianen Überlebenszeiten ... 937
5.4.3. Kombination der Konfidenzintervalle für Überlebens-wahrscheinlichkeiten zu markanten Zeitpunkten ... 938
5.5. Beispiel Progressionsfreies Überleben PFS: Kombination der Kaplan-Meier Überlebenskurven, der medianen Überlebenszeiten und der 6-Monats-Konfidenzintervalle ... 938
5.6. Cox's Proportional Hazards Regression Model: Vergleich mehrerer Gruppen, Kovariablen und Kombination der Effekte ... 940
6. Generalized Common Mean (GCM) Problem und Varianzschätzung bei Unkenntnis der Einzelvarianzen ... 941
6.1. Das GCM-Problem ... 942
6.2. Direkter Varianzschätzer von Hartung-Böckenhoff-Knapp ... 942
7. Wirtschafts-Prognosen mit Scoringsystem ... 943
7.1. Entwicklung eines Scoringsystems ... 943
7.1.1. Bekannte Referenzwerte ... 943
7.1.2. Referenzbezug Gruppenmittel ... 944

7.2. Kombinierte Prognosen mit Scoringpunkten 944
7.3. Beispiel: Prognosen des BIP von sieben Wirtschaftsinstituten 945
8. Combining Forecasts mit geschätzten Zuverlässigkeiten und Korrelationen 947
8.1. Beispiel: Kombinierte Prognose in der Investitionsberatung 947
8.2. Beispiel: Investitionsberatung im RE-Model 948
8.3. Kombination von Prognosen bei geschätzten Korrelationen aber homogenen Prognosen [EFE-Model] 949
8.4. Beispiel: Investitionsberatung im EFE-Model und „Borrowing of Strength" 950
8.5. Kombination von Prognosen bei geschätzten Korrelationen und heterogenen Prognosen [ERE-Model] 951
8.6. Beispiel: Investitionsberatung im ERE-Model 951
8.7. Ein stabiler Hartung-Varianzschätzer für die kombinierte Prognose im ERE-Model 952
8.8. Vergleich der vier Prognose-Modelle im Beispiel zur Investitionsberatung 953
8.8.1. Vergleich der kombinierten Prognosen 953
8.8.2. Vergleich der kombinierten Konfidenzintervalle 953
9. Kombination unabhängiger p-Werte oder unabhängiger Tests 954
9.1. Inverse Normal- und Stouffer-Methode, Probits 954
9.2. Inverse χ^2- und Fisher-Methode, Fisher-Probits 955
9.3. Heterogenität: Hartung-Knapp-Methode 956
9.4. Beispiel Verhaltensforschung: Konformität bei Männern und Frauen 956
10. Combining Forecasts bei unbekannten Korrelationen und Kombination abhängiger p-Werte oder Tests 958
10.1. Strategie zu Combined Forecasting 959
10.2. Hartung-Methode zur Kombination von abhängigen p-Werten oder Tests 959
10.3. Hartung-Methode für Fisher-Probits 960
10.4. Beispiel: Investitionsprognosen mit unbekannter Korrelation zwischen Betrieb und Berater 960
11. Kombination von kreuz-korrelierten Zeitreihen zu einem Panel 962
11.1. Dickey-Fuller Unit Root Test auf Stationarität einer Zeitreihe 962
11.2. Prüfung eines Panels kreuz-korrelierter Zeitreihen auf Stationarität 964
11.2.1. Kombination der Dickey-Fuller p-Werte nach Hartung 964
11.2.2. Kombination der Dickey-Fuller p-Werte nach Hartung mit den Fisher-Probits 964
11.3. Interpretation des kombinierten Tests auf Stationarität 965
12. Delphi Methode für Vorhersagen und zur Gewinnung Subjektiver Wahrscheinlichkeiten 965
12.1. Anwendungsbereiche der Delphi Methode 965
12.2. Kombination der Einzelprognosen oder subjektiven Risikobeurteilungen mittels der Delphi Methode 966
12.3. Beispiel Investitionsberatung: Cash Flow Vorhersagen ohne Zuverlässigkeitsangaben 967
12.4. Iterative Delphi Methode 967
13. Heterogenitätsmaß für Meta-Analysen: Verallgemeinerter Intraklass-Korrelationskoeffizient 968
13.1. Hartung-Konfidenzintervalle für die Heterogenitätsmaße 969
13.2. Beispiel Landwirtschaft — Genetik: Erblichkeit der Besamungsfähigkeit 970

13.3. Ein weiterer expliziter Hartung-Schätzer für den Heterogenitätsparameter 972
13.4. Median Unverzerrte Hartung-Schätzer für die Heterogenitätsmaße 973
13.5. Zusammenfassendes Beispiel: Zweiarmige Studien zu Multipler Sklerose MS 974
14. Normalverteilte Daten: Exakte Verfahren und spezielle Effektmaße 976
14.1. Effektmaß — Differenz der Mittelwerte 978
14.1.1. Exakte Punkt- und Intervallschätzer im Fixed Effects Model [Hartung-Knapp-Verfahren] 978
14.1.2. Beispiel: Zusammenfassung zweier Noninferiority Studien zu Asthma Bronchiale [Differenz der Mittelwerte] 979
14.2. Effektmaß — Quotient der Mittelwerte [Hartung-Verfahren] 981
14.2.1. Exakte Methoden im Fixed Effects Model für Quotienten 981
14.2.2. Beispiel Asthma Bronchiale: Zusammenfassung zweier Noninferiority Studien [Quotient der Mittelwerte] 982
14.2.3. Random Effects Model für „Quotient der Mittelwerte" 982
14.2.4. Beispiel MS-Studien: RE-Model für „Quotient der Mittelwerte" 984
14.2.5. Alternatives FE-Model für Quotienten mit Beispiel Asthma Bronchiale 986
14.3. Effektmaß — Standardisierte Mittlere Differenzen 991
14.3.1. Exakte Methoden im FE-Model für Standardisierte Mittlere Differenzen [Hartung-Verfahren] 991
14.3.2. Beispiel: Zusammenfassung zweier kleiner Superiority Studien zu Acne Papulopustulosa [Standardisierte Mittlere Differenzen] — Klinische Relevanz 993
14.3.3. Bisherige Methoden im FE- und RE-Model für Standardisierte Mittlere Differenzen 994
14.3.4. Beispiel: Verbesserung der Belastbarkeit bei Angina Pectoris [Standardisierte Mittlere Differenzen] — Klinische Relevanz 995
15. Konfidenzintervalle und implizite Schätzer für den Heterogenitätsparameter ... 996
15.1. Hartung-Knapp-Konfidenzintervall für den Heterogenitätsparameter 996
15.2. Beispiel: Fortsetzung MS-Studien 997
15.3. Weitere Konfidenzintervalle und Tests für den Heterogenitätsparameter 998
15.4. Implizit definierte Schätzer für den Heterogenitätsparameter 999
15.4.1. Mandel-Paule Schätzer 999
15.4.2. Median Unverzerrte Schätzer nach Hartung 1000
15.4.3. Pseudo Maximum Likelihood Schätzer nach Hartung 1000
15.4.4. Beispiel: Fortsetzung MS-Studien 1000
15.4.5. Maximum Likelihood Schätzer 1000
15.4.6. Restringierte Maximum Likelihood Schätzer 1001
15.4.7. Beispiel: Fortsetzung MS-Studien 1001
15.5. Zusammenfassendes Beispiel: Fortsetzung Landwirtschaft — Genetik 1002
16. Portfolioselektion im Wertpapiermanagement 1004
16.1. Rendite und Volatilität, mit Beispiel 1004
16.2. Pareto-Optimale Portfolioselektion 1005
16.3. Die Rendite-Volatilitäts-Projektionsebene, mit Beispiel 1007
16.4. Rendite-Schwelle und Volatilitäts-Schranke, mit Beispiel 1008
16.5. Korrelation der Wertpapierkurse, mit Beispiel 1010

17. Weitere Aspekte der Meta-Analyse 1013
17.1. Ergänzende Tests auf Homogenität 1013
17.2. „Partial Weighting" nach Yates-Cochran bei Dominanz einer Studie oder Prognose und Verallgemeinertes RE-Model 1013
17.3. Meta-Regression, mit Beispiel „Tuberkulose Studien" 1013
17.4. Prospektive Meta-Analyse 1014
17.4.1. Adaptive und „Self-designing" Versuchsanlagen zur Prospektiven Meta-Analyse 1016
17.4.2. Beispiel: Adaptive Placebo-kontrollierte prospektive Nichtunterlegenheits-Studien zu Asthma Bronchiale 1016
17.5. Eigenschaften des Gauß-Markov-Schätzers und der Varianzschätzer sowie verwandter Schätzer 1018

Anhang **1023**

1. Tabellenverzeichnis 1023
1.1. Kritische Werte, Quantile 1023
1.2. Weitere allgemeine Tabellen 1024
2. Tabellenanhang 1025
3. Griechisches Alphabet 1043
4. Symbolverzeichnis 1044
5. Literaturverzeichnis 1048
6. Literaturhinweise zu den einzelnen Kapiteln 1075
7. Sach- und Namensregister 1077
8. Weiterführende und ergänzende Hinweise 1144

Vorwort zur 15. Auflage

In der nun schon 15. Auflage, überarbeitet und stark erweitert, dieses Lehr- und Handbuches der angewandten Statistik konnten wir modernsten Entwicklungen und vielfach geäußerten Leserwünschen nachkommen: Das Buch enthält jetzt das **umfangreiche neue Kapitel XV :**

Meta-Analyse zur Kombination
von
Studien, Experimenten und Prognosen

mit zahlreichen detailliert durchgerechneten realen Beispielen, dokumentiert in 39 Tabellen und illustriert in 36 Abbildungen, die das jeweilige Geschehen visuell verdeutlichen und somit helfen, das Verständnis zu festigen.

Die **Meta-Analyse** gehört zu unseren aktuellen Arbeitsgebieten und so konnten wir neueste Forschungsergebnisse direkt in ihrer relevanten praktischen Bedeutung berücksichtigen. Die Meta-Analyse ist eine moderne Disziplin und von rasant zunehmender Bedeutung. Mit dem Anstieg verfügbarer Informationen gewinnt die Meta-Analyse an Wichtigkeit. **Combining Information** oder **Kombination von Wissen** bzw. **Combining Forecasts** — wenn es sich um die Kombination von Prognosen oder Vorhersagen handelt — sind weitere Bezeichnungen für diese Disziplin; speziell die Kombination von Wertpapieren zu einem Portfolio wird als **Portfolioselektion** bezeichnet.

Wir behandeln in Kapitel XV konkrete und wichtige Anwendungen u. a. in den Bereichen Gesamtwirtschaft, Betriebswirtschaft, Wertpapiermanagement, Ernährungswirtschaft, Landwirtschaft, Genetik, Verhaltensforschung, Umwelt, Technik, Pharmazie und Medizin, wobei hier speziell auch Überlebenszeit- oder Time-to-Event-Studien mit zensierten Daten analysiert und zu einem relevanten Gesamtergebnis kombiniert werden. Dank der **Evidenz Basierten Medizin** [**EBM**] erlangt in diesem Bereich die Meta-Analyse außer der medizinischen auch eine zunehmende wirtschaftliche Bedeutung.

Wir sind davon überzeugt, dass das Buch in der nun vorliegenden überarbeiteten und wesentlich erweiterten Form, auch durch das aktualisierte und stark ausgedehnte Literaturverzeichnis, in entscheidend stärkerem Maße sowohl Lernenden als auch praktisch Arbeitenden in ihrem Bedürfnis nach aktuellen und relevanten statistischen Methoden gerecht wird.

Die Entscheidung, die Meta-Analyse neu in das Buch aufzunehmen, wurde beeinflusst zum einen durch an uns herangetragene Anregungen seitens der Leser und zum anderen auch durch die Verleihung des

Thomas L. Saaty Prize
vom
American Journal of
Mathematical and Management Sciences

im Jahre 2006 für „outstanding" anwendungsorientierte Forschungsleistungen auf dem Gebiet der Meta-Analyse an den erstgenannten Autor dieses Buches.

An dieser Stelle möchten wir **Stefan Tewes** aus Iserlohn für die komplette Erstellung des Typoskripts und **Jan Edgar Hartung** aus Königswinter für seine Assistenz bei numerischen Berechnungen sowie für die komplette Erstellung der 36 Abbildungen für das neue Kapitel XV danken. J.E. Hartung haben wir auch zu danken für anregende Diskussionen und sein sorgfältiges Korrekturlesen.

Wir danken **Prof. Dr. Peter Pflaumer** für intensive Gespräche, welche die Ausrichtung des neuen Kapitels XV wesentlich beeinflusst haben. Ferner danken wir den Lesern, für deren Anregungen wir auch künftig ein offenes Ohr haben werden. Unser Dank gilt natürlich auch dem Oldenbourg-Verlag, Lektorat Wirtschafts- und Sozialwissenschaften, Steuern, Recht, und da insbesondere Dr. Jürgen Schechler sowie Rainer Berger, die uns nach Kräften bei der Neugestaltung des Buches unterstützten.

Joachim Hartung
Bärbel Elpelt
Karl-Heinz Klösener

Vorwort zur 7. Auflage

Die unvermindert starke Nachfrage hat nunmehr das Erscheinen der siebten Auflage dieses Lehr- und Handbuches der angewandten Statistik zur Folge. Nachdem in Vorauflagen eine vollständige Überarbeitung und starke Erweiterung vorgenommen wurde, konnte — nach einer nochmaligen kritischen Durchsicht — der Text hier weitgehend unverändert bleiben.

Um den an uns herangetragenen Leserwünschen gerecht zu werden, sind mittlerweile

Elpelt / Hartung: Grundkurs Statistik

sowie

Hartung / Heine: Statistik-Übungen — Deskriptive Statistik

bzw.

Hartung / Heine: Statistik-Übungen — Induktive Statistik

im selben Verlag erschienen. Diese drei Bücher sowie das nun in 3. Auflage erschienene Buch Multivariate Statistik (Hartung / Elpelt) werden am Schluß dieses Buches in ihrer Konzeption vorgestellt.

Vorwort zur 4. Auflage

In der nun vorliegenden, überarbeiteten und stark erweiterten vierten Auflage dieses Lehr- und Handbuches der angewandten Statistik haben wir moderne Entwicklungen und vielfach geäußerte Leserwünsche berücksichtigt: Das Buch enthält stark erweiterte bzw. neue Kapitel über

Zeitreihenanalyse,
Robuste Verfahren,
Explorative Datenanalyse (EDA).

Zudem konnten wir die bestehenden Kapitel vollständig überarbeiten, wobei auch Anregungen der Leser einbezogen wurden. Wir sind davon überzeugt, daß das Buch in der nun vorliegenden Form, nicht zuletzt auch durch das wesentlich erweiterte Literaturverzeichnis, in noch stärkerem Maße sowohl Lernenden als auch praktisch Arbeitenden gleichermaßen gerecht wird.

An dieser Stelle möchten wir cand. stat. Rolf Dietl, cand. stat. Thomas Goddemeier und cand. stat. Manfred Jutzi danken, die uns bei der technischen Herstellung der neuen Kapitel behilflich waren. Weiterhin gilt unser Dank den aufmerksamen Lesern, für deren Anmerkungen wir auch weiterhin dankbar sind, sowie dem R. Oldenbourg-Verlag und insbesondere Herrn Martin Weigert, die uns bei der Überarbeitung freie Hand ließen.

Aus dem Vorwort zur 3. Auflage

Der von Lesern mehrfach an uns herangetragene Wunsch einer stärkeren Berücksichtigung multivariater Verfahren ließ sich wegen der ohnehin schon sehr hohen Seitenzahl nicht realisieren. Jedoch ist bereits

Hartung / Elpelt: MULTIVARIATE STATISTIK
Lehr- und Handbuch der angewandten Statistik

im selben Verlag erschienen.

Aus dem Vorwort zur 1. Auflage

Statistische Methoden werden überall dort in zunehmendem Maße angewandt, wo durch Experiment oder Erhebung gewonnene Daten ausgewertet werden müssen. Solche Bereiche sind die Wirtschaftswissenschaften, die Ingenieurwissenschaften, die Medizin, die Landwirtschaft, die Physik, die Biologie, die Chemie, die Pharmazie, die Geodäsie, die Raumplanung, die Soziologie, die Psychologie, die Pädagogik, die Linguistik, die Archäologie, die Geologie, die Geographie, die Sportwissenschaft, die Umweltforschung usw.

Dieses Buch eignet sich nun zum einen als Handbuch für alle diejenigen, die Statistik bzw. statistische Methoden benötigen, zum anderen dient es dem Studenten in allen Anwendungsgebieten als vorlesungsbegleitendes, vertiefendes Lehrbuch. Nicht zuletzt wegen der Vielzahl vollständig durchgerechneter Zahlenbeispiele eignet es sich aber auch in hervorragender Weise zum Selbststudium.

Um allen angesprochenen Personenkreisen gerecht zu werden, haben wir die benötigten mathematischen Vorkenntnisse auf ein Minimum reduziert, das durch reine Schulmathematik abgedeckt wird, und zudem haben wir nicht nur solche Methoden aufgenommen, die in Grundvorlesungen über Statistik behandelt werden.

Dem Spezialisten wird ein schnelles Nachschlagen, dem Studenten und dem Praktiker wird die Wiederholung, Vertiefung und Erweiterung seiner Kenntnisse sowie ein fruchtbares Selbststudium ermöglicht.

An dieser Stelle möchten wir cand. stat. Johanna Behrens, cand. stat. Jürgen Bochynek, cand. stat. Johannes Ecke, cand. stat. Bärbel Heine, Dipl.-Stat. Manfred Korter, cand. stat. Lothar Kreienbrock, Dipl.-Math. Heinz Krüger, cand. stat. Heike Küntzel, Dipl.-Stat. Harald Kunitz, cand. stat. Gerd Mettendorf, cand. stat. Rüdiger Ostermann, Frau Bärbel Skopp, cand. math. Bärbel Tjakraatmadja, cand. stat. Elisabeth Vieth, Dipl.-Math. Bernard Voet, cand. stat. Peter Woboril und Dipl.-Stat. Bernd Wolter danken, die zum Teil bei der technischen Herstellung des Manuskriptes und zum Teil auch beim Sammeln des Materials und der Beispiele — für das endgültige Manuskript oder eines seiner verschiedenen Vorläufer — behilflich waren. Weiterhin gilt unser Dank Herrn Prof. Dr. Franz Hering für Ratschläge und Diskussionen, sowie dem R. Oldenbourg-Verlag, insbesondere Herrn Martin Weigert, für die gute Zusammenarbeit und weitestgehende Freiheit bei der Gestaltung des Buches.

Ein kurzer Überblick

An dieser Stelle soll ein kurzer Überblick über den Inhalt dieses Buches gegeben werden; einen ausführlicheren Einblick bieten das detaillierte Inhaltsverzeichnis und nicht zuletzt auch die Einleitungen zu den einzelnen Kapiteln und Abschnitten.

Die beschreibende, deskriptive Statistik dient der Aufbereitung und tabellarischen bzw. graphischen Darstellung von Datenmaterial. Als recht elementares Gebiet der Statistik und Grundlage vieler komplizierterer Methoden wird sie in **Kapitel I** ausführlich behandelt. Hier werden unter anderem empirische Kenngrößen von Häufigkeitsverteilungen, Konzentrationsmaße, Indexzahlen sowie empirische Zusammenhangsmaße vorgestellt.

Ausgangspunkt der induktiven Statistik ist die in **Kapitel II** in einem kurzen Abriß dargestellte Wahrscheinlichkeitsrechnung.

Die Schlußweisen der induktiven Statistik, basierend auf Punkt- und Intervallschätzungen sowie statistischen Tests, beinhaltet das sich anschließende **Kapitel III**.

Auch das **Kapitel IV** kann in gewisser Weise noch als einführend angesehen werden. Hier werden die wichtigsten statistischen Verteilungen und ihre Eigenschaften dargestellt. Zugleich werden aber in diesem Kapitel viele spezielle Methoden bzw. Verfahren betrachtet; zu den verschiedenen Verteilungen werden direkt entsprechende Punktschätzer, Intervallschätzer und Tests über die Parameter sowie Anpassungstests vorgestellt. Außerdem beinhaltet das Kapitel IV nichtparametrische (verteilungsfreie) Test- und Schätzmethoden im Einstichprobenproblem sowie einen ausführlichen Abschnitt über sequentielle Tests.

Das **Kapitel V** befaßt sich dann mit Problemen der Datengewinnung. Neben der Behandlung klassischer Methoden der Stichprobentheorie wird auch auf spezielle Stichprobenprobleme, z.B. in der Marktforschung, eingegangen. In einem Abschnitt über Meßfehler wird die Berücksichtigung, Erfassung und Kontrolle derselben angesprochen; hier sind nicht zuletzt auch die vorgestellten Ausreißertests und Datentransformationen von Interesse. Außerdem enthält das Kapitel V ausführliche Hinweise zur Planung und Auswertung von Versuchen, auch speziell von klinischen Versuchen, sowie Methoden der Skalierung von Merkmalsausprägungen.

Kapitel VI ist dann der in allen Bereichen der Fabrikation wichtigen Qualitätskontrolle auf Stichprobenbasis gewidmet. Neben einfachen, mehrfachen und sequentiellen Stichprobenplänen für quantitative und qualitative Merkmale sowie kontinuierlichen Stichprobenplänen wird auch auf die laufende Kontrolle der Produktion mittels Kontrollkarten eingegangen.

Vor allem bei der Auswertung nominalen Datenmaterials ist die Kontingenztafelanalyse von hervorragender Wichtigkeit. Diese wird in **Kapitel VII** ausführlich dargestellt. Neben klassischen Methoden der Analyse von zweidimensionalen Kontingenztafeln ist hier die Vielzahl vorgestellter Assoziationsmaße, die Analyse mittels loglinearem Modell — auch in mehrdimensionalen Kontingenztafeln — und nicht zuletzt das Logit-Modell, das Adjustieren von Kontingenztafeln und die Behandlung von Tests zum Vergleich der Parameter mehrerer diskreter Verteilungen hervorzuheben.

Kapitel VIII beschäftigt sich dann mit dem Vergleich zweier Meßreihen bzw. Stichproben. Behandelt werden hier Tests und Konfidenzintervalle für Lokations- und Dispersionsparameter einschließlich verteilungsfreier Verfahren.

Die Untersuchung der Stärke eines Zusammenhanges zwischen verschiedenen Merkmalen erfolgt mittels der in **Kapitel IX** dargestellten Korrelationsanalyse; für nominal skalierte Merkmale wird dies jedoch schon in Kapitel VII behandelt. Für einfache, partielle, bi-partielle und multiple Korrelationskoeffizienten werden neben Schätzern auch entsprechende Testverfahren betrachtet. Weiterhin findet man in Kapitel IX einen Test auf Unabhängigkeit mehrerer Meßreihen.

Das Kapitel X ist der Regressionsanalyse gewidmet, die der funktionalen Spezifikation des Zusammenhanges verschiedener Merkmale dient; hier werden verschiedene Probleme der linearen und nichtlinearen (insbesondere quadratischen) Regression sowie der multiplen Regression angesprochen. Ebenfalls wird Wert gelegt auf die Regression bei Fehlern in den Variablen.

Modelle der Varianzanalyse zur Beurteilung gemeinsam wirkender Einflußfaktoren werden in **Kapitel XI** behandelt; speziell finden sich hier auch verschiedene Methoden zum Vergleich von mehreren unabhängigen Meßreihen bzw. Stichproben.

Mit den Methoden der Zeitreihenanalyse beschäftigen wir uns im **Kapitel XII**. Neben klassischen Verfahren der Trend- und Saison-Schätzung bzw. -Elimination werden hier Zeitreihen als Realisationen stationärer und instationärer Prozesse (Box-Jenkins-Methode) behandelt. Weiterhin enthält dieses Kapitel die wesentlichen Aspekte der Spektral- und Kreuzspektralanalyse, also der Analyse von Zeitreihen im Frequenzbereich, sowie einen Abschnitt über gemischte Regressions-Zeitreihen-Modelle. Tests auf Stationarität einer Zeitreihe [Dickey-Fuller] und eines [kreuzkorrelierten] Zeitreihenpanels sind wegen des methodischen Zusammenhangs in den Abschnitt 11 von Kapitel XV verlagert.

Das **Kapitel XIII** beschäftigt sich mit der Zuverlässigkeitsanalyse von Systemen (z. B. technischer oder biologischer Natur) und der Analyse von Lebensdauern. Neben klassischen Systembetrachtungen werden auch Multi-State-Systeme, Fehlerbaumanalyse und die Analyse mehrphasiger Missionen miteinbezogen. Weiterhin betrachten wir hier verschiedene Klassen von Lebensdauerverteilungen, speziell Tests über die Zugehörigkeit zu diesen Klassen, und Parameterschätzungen bei den wichtigsten Lebensdauerverteilungen. Besondere Aufmerksamkeit verdienen auch die Ausführungen zu beschleunigten Lebensdauerprüfungen (Analyse im Eyring- oder Arrhenius-Modell), Wartungs- und Instandhaltungsstrategien sowie Überlegungen zur Verfügbarkeit von Systemen. Vergleichende Lebensdauerprüfungen [Log-Rank Test, Wilcoxon-Gehan Test, Tarone-Ware Tests], Cox's Proportional Hazards Regression Model und Kaplan-Meier Survival Analysis sind methodisch bedingt in Abschnitt 5 von Kapitel XV dargestellt.

In **Kapitel XIV** werden die Methoden der Explorativen Datenanalyse (EDA) dargestellt, die i. w. eine Ergänzung zur deskriptiven Statistik in Kapitel I sind, aber auf Terminologien späterer Kapitel zurückgreifen. Außerdem behandeln wir die grundsätzlichen Vorgehensweisen im Bereich der Robusten Statistik: M-, L- und R-Schätzer sowie Kenngrößen der Robustheit werden vorgestellt.

Die in **Kapitel XV** dargestellte **Meta-Analyse** verarbeitet unterschiedliche Ergebnisse verschiedener Studien, Experimente, Erhebungen oder Prognosen bzw. Vorher-

sagen und kombiniert sie auf geeignete Weise zu einem Gesamtergebnis. Das Methodenspektrum für diese aktuelle und zunehmend wichtiger werdende **Kombination von Wissen** oder **Combining Information** wird vorgestellt und ausführlich für eine Reihe von Anwendungsbereichen an Hand realer Beispiele demonstriert, in 39 Tabellen dokumentiert und mit 36 Abbildungen graphisch veranschaulicht. Dies geschieht für qualitative, quantitative oder Lebensdauer- bzw. Time-to-Event-Daten mit Zensierungen in den beteiligten Studien. Insbesondere wird auch die Kombination wirtschaftlicher Vorhersagen real diskutiert [**Combining Forecasts**]. Etwa im Bereich Umwelt — Technik werden widersprüchliche Ergebnisse konkurrierender Meß-Labore zur Einschätzung der Asbestbelastung eines Objekts zu einem Gesamtergebnis kombiniert mit dem Anspruch, eine gerechte Beurteilung zu erreichen. Im Bereich der **Evidenz Basierten Medizin** [**EBM**] hängt sowohl die medizinische als auch die wirtschaftliche Zukunft eines Medikamentes entscheidend von der richtigen Meta-Analyse ab, d. h. von der adäquaten Kombination der Resultate aus aufwendig durchgeführten Klinischen Prüfungen oder Studien. Im Wertpapiermanagement umfaßt die Meta-Analyse etwa die **Portfolioselektion**, in der ein Portfolio aus Wertpapieren nach den Kriterien Rendite und Risiko wohl abgestimmt zusammenzustellen ist.

Natürlich konnten in diesem Buch nicht alle statistischen Methoden bis in die letzte Einzelheit berücksichtigt werden, da dies den Rahmen des ohnehin schon umfangreichen Buches gesprengt hätte. Wir haben aber die wichtigsten, mit elementaren Kenntnissen der Mathematik zu bewältigenden und für die Anwendungen relevanten Methoden erfaßt. Die Auswahl der Beispiele zu den einzelnen Verfahren, die aus verschiedensten Anwendungsbereichen entlehnt sind, streicht die weitgestreuten und somit für jeden Leser interessanten Anwendungsmöglichkeiten der vorgestellten Methoden und Verfahren heraus. Für den speziell an einem Gebiet weitergehend interessierten Leser haben wir zudem ausführliche Hinweise auf Literatur angegeben.

Diejenigen, die insbesondere auch an den Verfahren der sogenannten Multivariaten Statistik (z.B. Skalierungsverfahren, Faktorenanalyse, Clusteranalyse, multivariate Regressions- und Varianzanalyse), welche hier nur angerissen werden können, interessiert sind, seien hingewiesen auf das im selben Verlag erschienene Buch

Hartung / Elpelt: MULTIVARIATE STATISTIK
Lehr- und Handbuch der angewandten Statistik ,

in dem derartige Verfahren auf über 800 Seiten dargestellt werden.

Für die erstmals mit Statistik konfrontierten Leser möchten wir abschließend noch einige Hinweise zum Arbeiten mit diesem Buch geben. Es ist keinesfalls notwendig, das Buch fortlaufend Seite für Seite durchzuarbeiten. Vielmehr sollte man zunächst die ersten Teile der Abschnitte 1 bis 5 von Kapitel I und die Kapitel II und III lesen. Hier sind vor allem die Begriffe Zufallsvariable, Verteilungsfunktion, Dichte, (Mittel- oder) Erwartungswert, Varianz, Standardabweichung, Punktschätzer, Intervallschätzer, Konfidenzintervall und statistischer Test wichtig für das weitere Verständnis. Auch die in Kapitel IV angegebenen grundlegenden Eigenschaften von Verteilungen, insbesondere von Normal-, Binomial- und Exponentialverteilung, sind für den Anfänger wichtig. Beim weiteren Studium des Buches ist die Einhaltung einer bestimmten Reihenfolge nicht mehr nötig; alle Kapitel können weitgehend unabhängig voneinander gelesen und verstanden werden.

Einführung und Grundlagen

1. Was ist Statistik

Das Wort Statistik hat wohl jeder schon oft gehört und benutzt: Man spricht von Verkaufsstatistiken, Statistik bei Wahlen, Unfallstatistik, Bevölkerungsstatistik usw. In diesem Zusammenhang assoziiert man mit dem Begriff Statistik zunächst einmal Zahlenkolonnen, Tabellen, in denen große Datenmengen mehr oder weniger übersichtlich zusammengestellt sind. In eben diesen Rahmen paßt wohl die Herkunft des Wortes Statistik: Abgeleitet vom neulateinischen Begriff „status", der etwa „Staat" oder auch „Zustand" bedeutet, verstand man im achtzehnten und auch im neunzehnten Jahrhundert unter Statistik vorwiegend die „Lehre von der Zustandsbeschreibung des Staates", also das Sammeln von Daten über Heer, Gewerbe und Bevölkerung (Tauf-, Heirats-, Sterberegister).

Eine andere Assoziation zum Begriff Statistik, die einem sofort in den Sinn kommt, ist der Mittelwert, der Durchschnittswert. So spricht man oft vom mittleren Einkommen, von der durchschnittlichen Zuschauerzahl, vom durchschnittlichen Benzinverbrauch, von der mittleren Temperatur oder etwa vom Durchschnittsalter. Jemand, der schon einmal den Prospekt eines Reiseveranstalters studiert hat, wird auf Angaben von Regentagen, Sonnentagen usw. pro Jahr in bestimmten Gebieten gestoßen sein. Obwohl hier nicht explizit vermerkt ist, daß diese Angaben nur Durchschnittswerte der vergangenen Jahre sind, handelt es sich natürlich auch hier um Mittelwerte. Was ein Mittelwert eigentlich ist, welche Aussagekraft er hat, kommt in folgendem Gedicht, das wir Herrn **P. H. List**, Professor für pharmazeutische Technologie in Marburg, verdanken, vgl. auch Krafft (1977), treffend zum Ausdruck.

Ein Mensch, der von Statistik hört,
denkt dabei nur an Mittelwert.
Er glaubt nicht dran und ist dagegen,
ein Beispiel soll es gleich belegen:

Ein Jäger auf der Entenjagd
hat einen ersten Schuß gewagt.
Der Schuß, zu hastig aus dem Rohr,
lag eine gute Handbreit vor.

Der zweite Schuß mit lautem Krach
lag eine gute Handbreit nach.
Der Jäger spricht ganz unbeschwert
voll Glauben an den Mittelwert:
Statistisch ist die Ente tot.

Doch wär' er klug und nähme Schrot
– dies sei gesagt, ihn zu bekehren –
er würde seine Chancen mehren:
Der Schuß geht ab, die Ente stürzt,
weil Streuung ihr das Leben kürzt.

Wie dem Jäger in diesem Gedicht geht es vielen: Man sagt Mittelwert und interpretiert diese Kenngröße als den wahren, den wirklichen Wert. Weiß man etwa, daß eine Sorte von Fernsehgeräten eine mittlere Lebensdauer von sieben Jahren hat, so heißt dies aber noch lange nicht, daß ein bestimmtes Gerät, das man kauft, wirklich

sieben Jahre lang hält. Wenn man Pech hat, kann es schon nach wenigen Tagen kaputtgehen, wenn man Glück hat, vielleicht auch erst nach fünfzehn Jahren. Im Einzelfall ist also der Mittelwert nicht identisch mit dem wahren Wert. Daraus darf man aber nicht schließen, daß die Statistik ungenaue oder gar falsche Angaben hervorbringt. Man muß sich aber vor einer falschen Interpretation von Kenngrößen wie dem Mittelwert hüten: Er ist nicht gleichzusetzen mit einem wahren Wert; er charakterisiert zwar die Verhältnisse, die Charakterisierung ist aber nicht vollständig, denn er beschreibt nur ein Zentrum, einen Durchschnitt aller vorkommenden Werte. Genauso verhält es sich mit allen anderen statistischen Kenngrößen: Die Angabe des minimal und des maximal möglichen oder vorkommenden Wertes, der Index der Verbraucherpreise, der Reallohnindex usw. geben eine unvollständige, aber durchaus sinnvolle und hilfreiche Vorstellung von den wirklichen Verhältnissen. Durch die Angabe von statistischen Kenngrößen hofft man also, das Wesentliche oder Typische der wirklichen Gegebenheiten zu erfassen und erhebt nicht, wie oft fälschlich angenommen, den Anspruch auf Gültigkeit im Einzelfall. Bei der Erfassung des Wesentlichen ist nicht nur die Angabe von Kenngrößen, sondern vielmehr auch die übersichtliche tabellarische oder auch graphische Darstellung der wahren Gegebenheiten oft sehr hilfreich.

Die bisher angesprochene Charakterisierung und Darstellung der realen Verhältnisse ist der wohl jedem bekannte und geläufige Teil der Statistik. Er ist jedoch nur ein kleiner Teil, den man auch die beschreibende oder deskriptive Statistik nennt. Wesentlicher ist die schließende, die induktive Statistik. Aufgrund vorliegender Daten werden mit Hilfe der Methoden der induktiven Statistik Schlüsse gezogen und Aussagen gemacht:

Wie wird sich etwa der Absatz eines Produkts oder die Zahl der Arbeitslosen in Zukunft entwickeln? Kann man aus der bisherigen Entwicklung Prognosen gewinnen? Kann man aus dem Wahlverhalten einiger Personen auf das Verhalten aller Wähler schließen (Wahlprognose)? In diesen Beispielen kann man die Prognosen, die Schlüsse aus den bekannten Gegebenheiten auf unbekannte Gegebenheiten im nachhinein überprüfen. Es stellt sich aber die Frage, ob man nicht schon vorher Aussagen über die Zuverlässigkeit solcher Prognosen oder Schlußfolgerungen machen kann, insbesondere dann, wenn eine Überprüfung nicht vorgesehen oder gar nicht möglich ist. Befragt man etwa einige Personen in der Bundesrepublik Deutschland über ihren Gesundheitszustand, um auf den Gesundheitszustand der Gesamtbevölkerung zu schließen, so kann man die Richtigkeit, die Genauigkeit der Schlußfolgerung kaum überprüfen: Es wäre viel zu aufwendig im nachhinein den Schluß zu überprüfen, indem man alle interessierenden Personen medizinisch untersucht. Es stellen sich dann die Fragen, wie man die zu befragenden Personen auswählt (Frage nach dem Stichprobenplan), wieviele Personen man auswählt (Frage nach dem Stichprobenumfang) usw., um mit vertretbarem Aufwand eine möglichst hohe Genauigkeit zu erreichen.

Natürlich darf man nicht darüber hinwegsehen, daß Stichprobenergebnisse und auch einzelne Meßergebnisse stets Ungenauigkeiten beinhalten. Dies ist auch in einer so exakten Wissenschaft wie z. B. der Physik der Fall: Hier besagt etwa die Heisenbergsche Unschärferelation, daß es unmöglich ist, Ort und Geschwindigkeit eines Teilchens gleichzeitig exakt zu messen.

Sicherlich werden aber ganz genaue, ganz exakte Angaben auch gar nicht benötigt. Die Fehler müssen jedoch in angemessenen, vernünftigen Größen gehalten werden: Bei der Bestimmung der Entfernung zwischen zwei Städten ist ein Meßfeh-

ler von einem Meter i. a. sicherlich irrelevant. Dagegen wird man bei der Messung der Dicke einer Metallfolie einen Fehler von 0,5 mm schon als zu groß bezeichnen; hier wäre die mit einer Mikrometerschraube erreichbare Genauigkeit von 0,01 mm als angemessen zu betrachten.

Das benötigte Instrumentarium zur Abschätzung von Art und Größe von Meßfehlern, Fehlern bei Schlüssen von Stichproben auf Gesamtheiten usw. wird von der Statistik bereitgestellt. Sie ist überall dort einzusetzen, wo der Zufall eine Rolle spielt oder wo Situationen so komplex sind, daß sie sich einer deterministischen Beschreibung entziehen. Dies soll noch an einem einfachen Beispiel erläutert werden: Wirft man eine Münze, so sagt man, es ist zufällig, ob auf der Oberseite „Kopf" oder „Zahl" erscheint. In Wirklichkeit aber ist der Münzwurf deterministisch: Ob Kopf oder Zahl fällt, ist durch physikalische Gesetzmäßigkeiten (Abwurfwinkel, Fallgesetze usw.) eindeutig bestimmt. Die Zahl der Faktoren, die das Ergebnis eines Münzwurfs beeinflussen, ist jedoch so groß, daß man vom Zufall spricht.

Die induktive Statistik könnte man also als ein Modell des Zufalls verstehen. Sie versucht, „den Zufall in den Griff zu bekommen", „Sicherheit über Unsicherheit zu gewinnen".

2. Das Experiment

Jede Beobachtung in der Natur oder in der Umgebung zieht bestimmte Verhaltensweisen nach sich. Hat man beispielsweise einmal auf eine heiße Herdplatte gefaßt und sich dabei die Finger verbrannt, so wird man das aufgrund dieser schlechten Erfahrung nicht noch einmal tun. Wie hier, zieht einmalige schlechte Erfahrung schon eine bestimmte Verhaltensweise nach sich, was auch im Sprichwort „Gebranntes Kind scheut das Feuer" zum Ausdruck kommt.

Andere Beobachtungen werden erst zur Erfahrung, wenn man einen bestimmten Vorgang vielmals beobachtet hat. So wundert sich niemand darüber, daß ein Gegenstand, den man fallen läßt, nach unten und nicht nach oben fällt. Aber auch, wenn im Herbst die Blätter von den Bäumen fallen, erstaunt es nicht, wenn ein Blatt „nach oben fällt". Ganz selbstverständlich schließt man, daß es durch einen Windstoß nach oben getragen wird. Niemand wird deshalb Zweifel an der Tatsache hegen, daß Gegenstände ohne Einwirkung von außen stets nach unten fallen; dies ist einfach eine Erfahrung.

Das Experiment ist nun eine Formalisierung von Beobachtungen in der Natur: Aus Beobachtungen in der Natur (Erfahrungen) wird ein Prinzip abgeleitet, d. h. eine Hypothese wird aufgestellt und dann experimentell überprüft. Eine der Voraussetzungen zum Betreiben der Naturwissenschaften, der Wissenschaften des Messens und Vergleichens, zu denen auch die Statistik gehört, ist das Experimentieren.

Naturgesetze z. B. basieren auf Beobachtungen aus denen ein Prinzip abgeleitet und experimentell überprüft wird. Sie gelten dann als bewiesen (durch das Experiment), solange kein Widerspruch auftritt bzw. kein besseres Prinzip (Modell) vorhanden ist. Dies sei an einem Beispiel ausführlich dargestellt (vgl. H. Seitz (1953): Das Weltbild der Physik, in: Die moderne Volkshochschule, Schwab, Stuttgart).

Aufgrund der subjektiven Erfahrung vieler Personen hat sich der objektive Befund herauskristallisiert, daß alle Körper unabhängig von Beobachtungsort und

-zeit senkrecht nach unten fallen, wenn man sie losläßt oder ihnen die tragende Unterlage entzieht. Diese Aussage an sich ist aber noch nicht befriedigend: Man möchte auch wissen, wie die Körper fallen. Daher werden Experimente durchgeführt, in denen der Vorgang des Falls immer wieder unter den verschiedensten Bedingungen wiederholt wird: Man läßt gleich schwere, aber verschieden große, gleich große, aber verschieden schwere Körper immer wieder aus derselben Höhe fallen und stellt dabei fest, daß alle Körper – ganz gleich wie groß oder wie schwer sie sind – gleich schnell fallen, sofern kein Luftwiderstand den Fall stört. Zum Gesetz des freien Falls kommt man experimentell schließlich dadurch, daß man auch die Fallhöhe variiert. Nachdem man experimentell erforscht hat, wie ein Körper fällt, stellt sich natürlich die Frage, warum er so und nicht anders fällt. Diese Frage ist ungleich schwerer zu ergründen, da hier ein induktiver Schluß von der Wirkung auf die Ursache notwendig ist. Dieser Schluß ist nie eindeutig, so wie es der Schluß von der Ursache auf die Wirkung ist: Tritt etwa eine Warenverknappung auf, so werden die Preise natürlich steigen; sind dagegen die Preise gestiegen, so heißt das noch lange nicht, daß eine Verknappung eingetreten ist.

Zur Beantwortung der Frage „Warum ist etwas so und nicht anders?" stellt man eine (Arbeits-)Hypothese auf und vergleicht die Erfahrungen mit den Annahmen in der Hypothese. Zur Erklärung der Fallgesetze etwa wurde die Hypothese aufgestellt, daß die Erde eine Anziehungskraft auf die fallenden Körper ausübt. Diese Anziehungskraft nennt man dann Schwerkraft. Als Isaac Newton (1642–1727) nachweisen konnte, daß dieselbe Kraft, die einen Körper nach unten fallen läßt, auch den Mond zu seiner Bahn um die Erde zwingt, wurde die Vorstellung von der Schwerkraft auf das ganze Universum erweitert: Aus dem Fallgesetz entstand das Gravitationsgesetz. Aufgrund des Gravitationsgesetzes gelang es, die Existenz bisher unbekannter Planeten (z. B. Pluto) vorauszusagen, was zur Folge hatte, daß nicht mehr an der Richtigkeit der Hypothese über die Gravitationskraft gezweifelt wurde. Die Hypothese wurde zur Theorie. Eine Theorie ist aber nie eine unumstößliche Wahrheit, sie ist lediglich eine zweckmäßige Darstellung des Zusammenhangs aller bisherigen Erfahrungen und behält ihre Gültigkeit nur solange keine Widersprüche zur Erfahrung entstehen.

Das statistische Experiment oder das Zufallsexperiment ist nun dort besonders wichtig, wo Objekte (Untersuchungseinheiten) zur Verfügung stehen, an denen Effekte von Behandlungen meßbar, aber nicht exakt vorhersagbar sind, d. h. von Versuch zu Versuch verschieden sind. Anhand der gemessenen Effekte möchte man dann Schlüsse über die unbekannten, wahren Effekte ziehen, die Ergebnisse zukünftiger Experimente (Versuche) vorhersagen (bei Einräumung von spezifizierten Fehlerwahrscheinlichkeiten), d. h. man möchte „Sicherheit über Unsicherheit" gewinnen. Besonders wichtig ist das statistische Experiment also überall dort, wo Unsicherheiten (Zufälligkeiten) auftreten. Um nun wirklich „Sicherheit über Unsicherheit" gewinnen zu können, muß das Experiment genau geplant werden: Die Versuchsergebnisse müssen ja in möglichst hohem Maße vergleichbar sein, die Versuchsbedingungen müssen möglichst konstant gehalten werden, das Experiment muß kontrollierbar und jeder Zeit wiederholbar sein. Die Planung eines Experiments muß natürlich stets vor der Durchführung erfolgen und muß sich an den praktischen Gegebenheiten orientieren. (Natürlich sind die Planungsmöglichkeiten durch äußere Umstände eingeschränkt.)

Eine besonders große Rolle spielt das statistische Experiment naturgemäß dort, wo die Unsicherheit am größten ist. Dies ist in Bereichen wie z. B. der Medizin, der

Psychologie, der Pädagogik, der Biologie, der Landwirtschaft und den Ingenieurwissenschaften der Fall. Die Ergebnisse eines Experiments in solchen Bereichen, die O. Kempthorne die "noisy sciences" nennt, sind nie exakt: Störungen (Rauschen, noise) machen exaktes Messen unmöglich. Typisch ist hier, daß ein Objekt nach einmaliger Versuchsdurchführung nicht wieder in den ursprünglichen Zustand zurückversetzt werden kann. Hat man etwa die Lebensdauer eines Transistors festgestellt, so ist dieser dann defekt, und man kann nicht noch einmal seine Lebensdauer bestimmen. Hat man den Erfolg der Behandlung eines Krebspatienten mit einem bestimmten Medikament festgestellt, so kann man im selben Experiment das Medikament diesem Patienten i. a. nicht noch einmal verabreichen, um den Behandlungserfolg zu überprüfen, denn der Ausgangszustand des Patienten ist dann ein anderer. Eine Versuchswiederholung ist also nur mit einem anderen Objekt möglich. Dann stellt sich aber die Frage, ob sich die Objekte genügend ähnlich sind, d. h. die Frage nach der Konstanz der Versuchsbedingungen.

Am ehesten kontrollierbar sind wohl Experimente im physikalischen und chemischen Bereich. Aber auch hier treten Unsicherheiten auf; man kann nicht von absoluten Wahrheiten sprechen. Dies wurde schon bei den Ausführungen zum Fallgesetz und zum Gravitationsgesetz deutlich: Die Theorie hat nur so lange Gültigkeit, bis sie durch Erfahrung widerlegt wird. Die Unsicherheit bei solchen gut planbaren, kontrollierbaren Experimenten, die zwar relativ gering ist, kommt auch in einem Brief Maxwells an Lord Rayleigh zum Ausdruck; Maxwell schreibt dort: „Das zweite Gesetz der Thermodynamik hat denselben Wahrheitsgehalt wie die Aussage, daß man, nachdem man einen Eimer voll Wasser ins Meer geschüttet hat, nicht denselben Eimer Wasser wieder dem Meer entnehmen kann." Maxwell hält also die Gültigkeit des zweiten Gesetzes der Thermodynamik für sehr wahrscheinlich, erkennt jedoch, daß es nicht völlig gesichert ist.

Im wesentlichen sind also in allen Bereichen mehr oder weniger große Unsicherheiten vorhanden; ein „Rauschen" kann quasi niemals völlig ausgeschlossen werden. Die Statistik (das statistische Experiment) dient nun der Extraktion der vom Rauschen überlagerten Information, d. h. man möchte diejenige Information aus dem Experiment ziehen, die auf andere Umstände, andere Objekte usw. übertragbar ist; man vergleiche hierzu auch die Ausführungen von Krafft (1977).

3. Die Erhebung

Wir haben gesehen, daß sich Störeinflüsse auch in sogenannten exakten Wissenschaften nicht gänzlich ausschließen lassen; auch ein noch so gut kontrollierbares Experiment birgt gewisse Unsicherheiten. In wesentlich stärkerem Umfang treten unkontrollierbare Störungen z. B. im sozial- und wirtschaftswissenschaftlichen Bereich auf. Die Preisentwicklung etwa wird von saisonalen Einflüssen überlagert; oft aber rühren höhere Preise auch daher, daß Qualitätsverbesserungen mit ihnen einhergehen oder aus Angst Panikkäufe (Hamsterkäufe) getätigt werden. Hat man die Preisentwicklung eines Produktes beobachtet, so lassen sich diese Beobachtungen nicht in einem kontrollierten Experiment überprüfen: Die Zeit kann nicht zurückgedreht werden, Störeinflüsse können nicht ausgeschaltet werden. Wie will man etwa Saisonschwankungen ausschalten und beobachten, wie sich das Preisniveau entwickelt hätte, wenn keine saisonalen Effekte vorhanden gewesen wären?

Man ist hier gezwungen, seine Schlußfolgerungen direkt aus den einmal erhobenen (beobachteten) Daten zu ziehen. Dabei kann man natürlich Vergleichsdaten

(etwa aus anderen Perioden) zu Hilfe nehmen. Ist etwa die momentane Preisentwicklung „wesentlich" von der in anderen Jahren verschieden, so wird man dann sicherlich annehmen können, daß nicht nur saisonale Schwankungen das Preisniveau beeinflußt haben, denn diese dürften in allen Jahren etwa gleich sein.

Eine (nichtkontrollierbare) Erhebung, ein Sammeln von nicht im nachhinein überprüfbaren Daten kann also nicht dem experimentellen Nachweis einer Hypothese, sondern nur der Erfassung eines Sachverhaltes dienen.

Auch eine Befragung birgt ein gehöriges Maß an Unsicherheit. Befragt man etwa einige Personen nach ihrem Einkommen und will daraus auf die Einkommensstruktur z. B. in der Bundesrepublik Deutschland schließen, so macht man unter Umständen einen großen Fehler. Befragt man etwa nur Frauen, so ist ein Schluß auf die Gesamtheit sicher mit einem enormen Fehler behaftet, denn es ist bekannt, daß die Einkommensstruktur bei Frauen anders ist als bei Männern.

Auch eine Befragung ist natürlich i. a. kein Experiment zum Nachweis einer Hypothese. Man kann nicht die Einkommensstruktur beobachten und dann eine aus der Beobachtung gewonnene Hypothese experimentell überprüfen: Da die Zeit fortschreitet, kommt es ständig zu Verschiebungen in der Einkommensstruktur (z. B. durch Inflationsausgleiche oder einseitige Einkommenskürzungen im Öffentlichen Dienst).

Die Feststellung der momentanen Einkommensstruktur in der Bundesrepublik Deutschland, die natürlich nur eine Erhebung (und kein Experiment) ist, läßt sich aber planen: Man kann z. B. im vorhinein überlegen, welche Personen man befragt und wieviele Personen man befragt. Solche Fragen der Planung von Erhebungen werden von der Stichprobentheorie beantwortet. Man muß überlegen, auf welche Gesamtheit ein Erhebungsergebnis übertragen werden soll. Will man wissen, was der durchschnittliche momentane Verdienst aller Personen in der Bundesrepublik Deutschland ist, oder will man den Durchschnittsverdienst aller Berufstätigen ermitteln, d. h. welche Grundgesamtheit sollen die befragten Personen repräsentieren? Wie genau möchte man den Durchschnittsverdienst ermitteln? Je genauer das Ergebnis sein soll, desto größer muß in der Regel die Zahl der befragten Personen sein. Die geplante Erhebung unterscheidet sich also von der nichtgeplanten dadurch, daß ihr ein Stichprobenplan zugrundeliegt.

Man könnte nun natürlich einwenden, daß doch stets erreicht werden kann, daß bei einer Befragung kein Fehler beim Schluß auf eine Grundgesamtheit vorhanden ist. Theoretisch braucht man ja nur die gesamte Grundgesamtheit zu erfassen, d. h. etwa alle in Frage kommenden Personen zu befragen. In einem solchen Fall spricht man dann nicht mehr von einer Stichproben- oder Teilerhebung, sondern von einer Voll- oder Totalerhebung. Meistens sind aber Vollerhebungen viel zu teuer und zu aufwendig. Auch sind sie nur theoretisch fehlerfrei: Bei den meisten Erhebungen, die Personenbefragungen als Grundlage haben, kommen psychologische Effekte ins Spiel. Befragt man etwa alle Bewohner der Bundesrepublik Deutschland danach, ob sie arbeitslos sind oder nicht und ob sie „schwarzarbeiten", um den Prozentsatz der „arbeitslosen Schwarzarbeiter" festzustellen, so werden einige gar keine Antwort, andere eine falsche Antwort geben, da ihnen die Wahrheit unangenehm ist; wer gibt schon gerne zu, daß er arbeitslos ist und schwarzarbeitet? Praktisch enthalten also auch Totalerhebungen Fehlerquellen. Manchmal ist es sogar so, daß der Fehler bei einer Stichprobenerhebung geringer ist als bei einer Vollerhebung: Befragt man nur wenige Personen, so kann die Befragung durch ein Interview sehr intensiv durchgeführt werden. Will man dagegen eine Totalerhebung

durchführen, so ist man wohl auf das Versenden von Fragebögen angewiesen. Die Antworten sind dann oft unklar; ein Nachfragen ist nicht möglich. Auch werden viele Fragebögen erst gar nicht zurückgeschickt.

Bisher sind wir davon ausgegangen, daß eine Erhebung durchgeführt wird, um eine bestimmte Frage zu beantworten, einen Sachverhalt zu erfassen, z. B.: Wie entwickelt sich das Preisniveau? Wieviel Prozent aller Personen sind arbeitslos? Wie hoch ist der mittlere Verdienst in der Bundesrepublik Deutschland? Eine solche Erhebung nennt man auch eine primärstatistische Erhebung, denn die Erhebung wird eigens zur Beantwortung einer konkreten Frage durchgeführt.

Oft werden aber Daten sowieso erhoben. Beispielsweise wird, ohne daß eine bestimmte Frage verfolgt wird, in der Flensburger Verkehrskartei jede Neuzulassung von Personenkraftwagen registriert. Will man also die Anzahl der durchschnittlich pro Monat neu zugelassenen PKW berechnen, so braucht man nicht eigens zu diesem Zweck eine Erhebung durchzuführen; man kann die existierende Kartei zu Hilfe nehmen. Aber auch die Bestimmung der durchschnittlichen Neuzulassungen ist eine Erhebung, nämlich eine sogenannte sekundärstatistische Erhebung. Bei einer sekundärstatistischen Erhebung werden also die Daten nicht extra zwecks Beantwortung einer Fragestellung erhoben; vielmehr werden ohnehin vorhandene Originaldaten zur Klärung der Fragen herangezogen. Ist das Material einer Erhebung nicht nur zu einem anderen Zwecke erhoben worden, sondern handelt es sich auch noch um verfälschtes Material, d. h. um bereits verformte, manipulierte Daten, so spricht man schließlich von einer tertiärstatistischen Erhebung. Ist man bei primärstatistischen Erhebungen oft noch in der Lage, die Erhebung der zu beantwortenden Frage entsprechend optimal zu planen, d. h. einen möglichst gut angemessenen Stichprobenplan aufzustellen, so ist dies bei sekundär- und erst recht bei tertiärstatistischen Erhebungen natürlich nicht mehr möglich. Schlüsse, die aus nichtprimärstatistischen Erhebungen gezogen werden, sind daher mit einem noch größeren Fehler behaftet als bei primärstatistischen Erhebungen; die Aussagekraft von Schlußfolgerungen – ihr Wahrheitsgehalt – ist also mitunter nicht sehr groß. Dies gilt natürlich für tertiärstatistische Erhebungen in hohem Maße insbesondere dann, wenn man nicht weiß, in welcher Weise Verformungen und Manipulationen der Originaldaten vorgenommen wurden.

4. Zur Statistik und ihren philosophischen Voraussetzungen

Die Statistik ist als methodisches Instrumentarium an allen empirisch arbeitenden Wissenschaften bzw. Wissenschaftsbereichen orientiert. Philosophische Grundlagen/Voraussetzungen der Statistik sind daher mit den Grundlagen dieser Wissenschaften gegeben. Einige Hinweise auf solche philosophischen Voraussetzungen können dies aufzeigen. Charakteristisch für alle statistischen Methoden ist der Verzicht auf eine Vielzahl von Kenntnissen über die Untersuchungseinheiten einer Erhebung bzw. über die Objekte, die in einem Experiment berücksichtigt werden. Stellt man in einer Erhebung z. B. die mittlere Größe der Nutzfläche aller landwirtschaftlichen Betriebe in der Bundesrepublik Deutschland fest, so verzichtet man bewußt auf die Registrierung des Viehbestandes, der Region, in der der Betrieb liegt, usw. Bestimmt man in einem Experiment die mittlere Gewichtsabnahme in einem Zeitraum bei Verabreichung einer bestimmten Diät, so wird bewußt z. B. das Alter, das Anfangsgewicht und der Gesundheitszustand der Objekte (Personen, Tiere) außer acht gelassen.

Man sagt sich von vielen Kenntnissen los und begnügt sich mit einigen wenigen: „An die Stelle der tatsächlich existierenden Einheit, des tatsächlichen Geschehens, tritt der statistische Schatten, in dem nur Teile der Wirklichkeit registriert sind" (vgl. O. Anderson (1965): Probleme der statistischen Methodenlehre in den Sozialwissenschaften, Physica, Würzburg).

Real gesehen sind alle Objekte zufallsbedingt verschieden; durch den Verzicht auf Kenntnisse über das einzelne Objekt verliert es seine Individualität, an deren Stelle sein statistischer Schatten tritt: Objekte mit identischen Schatten sind statistisch gleich. Alle landwirtschaftlichen Betriebe mit gleich großer Nutzfläche sind in dieser Hinsicht gleich, obwohl sie in völlig verschiedenen Regionen liegen können, völlig unterschiedliche Viehbestände haben können usw. Alle Objekte, bei denen durch Verabreichung einer Diät die gleiche Gewichtsabnahme erreicht wird, werden als gleich angesehen, obwohl sie natürlich individuelle Unterschiede wie z. B. unterschiedliches Alter aufweisen.

Entscheidend für die Aussagekraft der Statistik (der Erhebung, des statistischen Experiments) ist es, inwieweit alle irgendwie relevanten Umstände im statistischen Schatten berücksichtigt werden können. Wesentlich für die Beurteilung der Ertragssteigerung durch Verwendung eines Düngemittels ist die Berücksichtigung der Bodengüte, der Witterungsbedingungen, der Bewässerungsmöglichkeiten, des Saatgutes usw. Sicherlich können nicht alle irgendwie relevanten Umstände im Schatten berücksichtigt werden, so daß Aussagen aufgrund von Erhebungen oder Experimenten nie den Anspruch auf absolute Wahrheit erheben können; wohl aber kann der Wahrheitsgehalt einer Aussage sehr hoch sein.

Andersons Charakterisierung des Wesens der Statistik (Ersetzen des Individuums durch seinen Schatten, bewußter Verzicht auf Kenntnisse über das Individuum) ist der Philosophie Immanuel Kants (1724–1804), der „kritisch" lehrt, daß man nicht ein „Ding an sich", sondern nur „Erscheinungen" von ihm beobachten kann, entnommen; demzufolge kann kein Anspruch auf absolute Gültigkeit erhoben werden: Nichts ist eben absolut wahr. Daher versucht man durch Beobachten, Erheben und Experimentieren im „Raum der Erscheinungen" Aussagen zu machen. Die Gewinnung von Erkenntnissen unter den Bedingungen von „Raum und Zeit" ist heute Grundlage aller modernen Wissenschaften. Daß dies nicht immer so war und diese Auffassung nicht immer vertreten wurde, wird bei Durchsicht der Geschichte der Philosophie deutlich.

Descartes (1596–1650) formulierte den bekannten Satz „Ich denke, also bin ich" („cogito, ergo sum"), d.h. nur das eigene Denken (Existenz) ist absolut sicher. Dieser Satz ist der Kern einer Philosophie, die sehr richtig bemerkt, daß absolute Wahrheit nicht durch Beobachtung, sondern nur durch logisches Denken deduktiv gewonnen werden kann, und die es deshalb generell ablehnt, sich auf empirische Beobachtungen als Grundlage zu stützen, um zu neuen Erkenntnissen zu gelangen. In der Antike wurde diese Philosophie von Plato (427–347 v. Chr.) vertreten. Aber auch damals gab es im Ansatz schon die heute vertretene Ansicht, daß man aus Beobachtungen sehr wohl induktiv zu neuen Erkenntnissen gelangen kann; dies kommt in der Philosophie des Aristoteles (384–322 v. Chr.) zum Ausdruck. Er akzeptiert das Vorhandensein von Variabilität, glaubt aber auch an die Immanenz des Parameters (des wahren Wertes, der nicht von Störungen überlagert ist) im Individuum. Die Unterschiede zwischen den Mitgliedern einer Gesamtheit sind nur vom Zufall bedingt, entstehen nur durch „Unfälle". Dieser philosophische Ansatz des Aristoteles wurde in der Folgezeit aber gelegentlich übersehen. Zwischenzeitlich

ging das philosophische Interesse auch weg von der Natur und der Wissenschaft und konzentrierte sich mehr auf die Theologie mit Ethik und Moral sowie Ästhetik. Unter dem Einfluß eines Philosophen wie z.B. Augustinus (354–430), der als höchste Erkenntnisstufe die göttliche Eingebung und als die niedrigste aller Erkenntnisstufen die Gewinnung von Erkenntnissen aus Beobachtungen und Experimenten betrachtet, kann sich keine philosophische Grundlage entwickeln, auf der Naturwissenschaften – und damit auch die Statistik – aufbauen können.

Die platonische Philosophie wird neben der aristotelischen auch weiter vertreten. So bekannte Philosophen wie Descartes und Leibniz (1646–1716) vertreten in der Konsequenz die Ansicht, daß Experimentieren kein wahres Wissen bringt. Descartes etwa vertritt den Standpunkt, daß die Axiome der Mathematik deduktiv, also nicht durch Beobachtungen (Erfahrungen), entstanden sind.

Die Lehren des Galileo Galilei (1564–1642) bereiten so etwas wie eine Wende im Denken vor. Er beschäftigte sich mit der Berechnung von Chancen bei Glücksspielen, wobei er natürlich Beobachtungen zugrundelegte. Dies findet etwa ein Jahrhundert später seinen Niederschlag u.a. in der Philosophie von Hume (1711–1776) und Kant. Hume vertritt einen extremen empiristischen Standpunkt. Er vertritt die philosophische Ansicht, daß man Dinge, die man nicht experimentell erarbeiten kann, auch nicht weiß. Nach Hume sind also nur Experimente geeignet, das Wissen zu vermehren; man kann nur induktiv, also durch den Schluß von der Wirkung auf die Ursache, vom Teil auf das Ganze, zu neuen Erkenntnissen gelangen. Der „kritische" Standpunkt Kants zeigt dagegen positive Möglichkeiten und Grenzen und kommt damit dem heutigen Wissenschaftsverständnis wohl am nächsten. Deshalb kann man die Kantsche Philosophie, die teilweise bei Aristoteles anknüpft, wohl als philosophische Grundlage der Naturwissenschaften, wie sie heute betrieben werden, betrachten. Kant versucht die „Begriffe" und die „empirische Anschauung" zu vereinen: Durch Konstruktion im „Gemüt" werden die vielfachen sensorischen Informationen, die das Gehirn empfängt, synthetisiert. Kant stellt fest, daß man nicht das „Ding an sich" (die absolute Realität), sondern nur Erscheinungen von ihm beobachtet, und daß man daher die Natur durch „Schemata" (Modelle) nicht vollständig erklären kann. Ein „Schema" (versinnlichter Begriff oder Hypothese) kann nach Kant sehr wohl dabei helfen, die empirische Wirklichkeit anhand von Beobachtungen „analytisch" und „synthetisch" zu erfassen.

Wie wichtig für die heutige Zeit die Wende im Denken durch die Werke Galileo Galileis und die daran anschließende philosophische Besinnung auf Aristoteles war, läßt sich vielleicht ermessen, wenn man weiß, daß zuvor praktisch nur die Gesetzmäßigkeiten der Sternenbewegung bekannt waren.

Faßt man die philosophischen Grundlagen der modernen Wissenschaften, und damit der Statistik noch einmal zusammen, so ergibt sich:

Das Experiment ist ein Verifikationsinstrumentarium, d. h. es dient zur Überprüfung von Hypothesen, die aufgrund von Beobachtungen aufgestellt werden. Die Erhebung hingegen ist nur ein Erkenntnisinstrumentarium: Aufgrund von Beobachtungen versucht man, reale Gesetzmäßigkeiten, reale Zustände zu erfassen. Ganz gleich, ob Experimente oder Erhebungen durchgeführt werden, sieht sich die Wissenschaft heute im Kantschen Verständnis.

Man muß Variabilität und Veränderung als wirklich und nichtillusorisch akzeptieren; man kann Teile (Einheiten, Objekte) nicht einzeln sehen, denn jedes Teil gibt nur ein unvollständiges Bild des Ganzen, zu dem es gehört; es gibt keinen speziellen

Grund für Variabilität, denn sie ist zufällig; Wissen und Erkenntnis können aus Beobachtungen abstrahiert werden; der induktive Prozeß, der es erlaubt, von Teilen auf das Ganze zu schließen, muß akzeptiert werden. (Deduktiv kann man nur vom Ganzen auf Teile schließen.)

5. Zur Geschichte der Statistik

Wir wollen hier einen kurzen Überblick über die wichtigsten Stationen der Entwicklung der Statistik geben, der natürlich keinesfalls den Anspruch auf Vollständigkeit erheben kann.

Die Grundlage aller statistischen Methoden ist die Wahrscheinlichkeitstheorie, die ihren Ursprung wohl in der Berechnung von Chancen bei Glücksspielen hat.

Schon sehr früh wird in einem Gedicht („de vetula"), das vermutlich Richard de Fournival (1200–1250) zuzuschreiben ist, die Anzahl der möglichen Wege mit Berücksichtigung der Reihenfolge angegeben, in denen drei Würfel fallen können, ohne sich dabei um irgendwelche Wahrscheinlichkeiten zu kümmern. Erst im 16. Jahrhundert legt der italienische Mathematiker, Physiker und Philosoph Geronimo Cardano (1501–1576) in seinem ca. 1526 erschienenen Buch „Liber de Ludo Aleae" empirisch fest, daß beim Würfeln mit einem „idealen" Würfel als Ergebnis alle Seiten gleichwahrscheinlich sind. Er verallgemeinert dieses Ergebnis fälschlicherweise auf einen aus dem Sprungbein von Schafen gewonnenen vierseitigen Würfel: Hier trifft die Gleichwahrscheinlichkeit wegen der unregelmäßigen Form nicht zu.

Exakte Berechnungen der Wahrscheinlichkeiten beim Würfeln mit 3 Würfeln finden sich dann in dem Werk „Sopra le Scoperte dei Dadi" des Italieners Galileo Galilei (1564–1642), das erstmals im Jahre 1718 unter dem Titel „Considerazione sopra il Giuco dei Dadi" veröffentlicht worden ist.

Die ersten wirklich mathematisch orientierten Betrachtungen zur Berechnung von Chancen bei Glücksspielen werden durch folgende Anfrage aus dem Jahre 1654 des Chevalier de Méré beim französischen Mathematiker Blaise Pascal (1623–1662) angeregt:

„Stimmt es, daß von den beiden folgenden Wetten

I. Ein Würfel wird viermal geworfen; gesetzt wird darauf, daß mindestens eine Sechs auftritt;

II. Zwei Würfel werden (gleichzeitig) vierundzwanzigmal geworfen; gesetzt wird darauf, daß mindestens ein Sechserpasch auftritt;

man bei der ersten Wette öfter gewinnt als bei der zweiten?"

Aus dieser Anfrage resultierte die Korrespondenz zwischen Blaise Pascal und Pierre de Fermat (1601–1665), die de Mérés Vermutung aufgrund häufigen Spielens bestätigte: Im ersten Fall gewinnt man in 51,77% aller Fälle, im zweiten nur in 49,14% aller Fälle. Die Ideen der französischen Mathematiker wie etwa Pascal und Fermat wurden von Christian Huygens (1629–1695) in „de Ratiociniis in Alea Ludo", einem Anhang zu seinem 1657 erschienenen „Exercitationes Mathematicae", herauskristallisiert und erweitert. Im 18. Jahrhundert untersuchte dann der englische Geistliche Thomas Bayes (1702–1761), welche Aussage man bei Kenntnis der Ausgänge eines Spiels über die zugrundeliegenden Wahrscheinlichkeiten ma-

chen kann. Schließlich erscheint im Jahre 1812 die „Théorie Analytique des Probabilités" von Pierre Simon Marquis de Laplace (1749–1827), in der er einen Überblick über den Stand der Wahrscheinlichkeitsrechnung gibt. Auch hier beschäftigt man sich fast ausschließlich noch mit solchen Problemen, wo nur endlich viele Ereignisse mit gleicher Wahrscheinlichkeit auftreten, also mit Würfel-, Karten- und Glücksspielen.

Die Wahrscheinlichkeitstheorie, wie sie heute betrieben wird, erlebte ihren Aufschwung erst mit ihrer Axiomatisierung durch A. N. Kolmogoroff im Jahre 1933.

Die eigentliche Statistik – bzw. was man heute unter Statistik versteht – entwickelte sich erst vom Beginn dieses Jahrhunderts an sehr stürmisch. Einige Ansätze und Grundlagen gehen jedoch schon auf frühere Zeiten zurück. So wurde im 16. Jahrhundert vom Astronom Tychonis Brahe Dani (1546–1601) schon das arithmetische Mittel verwandt. Er beobachtete vom Februar 1582 bis April 1588 in unregelmäßigen Abständen den Stand von α Arietis. Dabei faßte er stets zwei Beobachtungen zusammen und bildete deren arithmetisches Mittel, um so Parallaxen und Refraktionen auszugleichen. Die so erhaltenen Ergebnisse waren erstaunlich exakt und wurden auch von Kepler (1571–1630) verwandt. Die Hypothese, daß Mädchen- und Knabengeburten gleichverteilt sind, wurde vom schottischen Arzt und Satiriker John Arbuthnot (1667–1735) durch statistisches Zahlenmaterial widerlegt. Im Jahre 1713 wird das Buch „Ars conjectandi" von Jakob Bernoulli (1654–1705) veröffentlicht. Dort taucht der Begriff „Stochastisch" auf, die Binomialverteilung und die Gesetze der großen Zahlen werden wohl erstmals erwähnt.

Von Abraham de Moivre (1667–1754) wird dann der Zusammenhang zwischen Binomialverteilung und Normalverteilung erkannt und 1718 in seinem Werk "The Doctrine of Chances" veröffentlicht. Ein frühes Beispiel für das Vorkommen der Normalverteilung gab der belgische Astronom und Physiker Adolphe Quetelet (1796–1874) mit der Messung des Brustumfangs von 5738 schottischen Soldaten. Der Name Normalverteilung wird allerdings erst von Francis Galton (1822–1911) um 1880 verwandt; heute nennt man sie zu Ehren von Carl Friedrich Gauß (1777–1855) manchmal auch Gaußverteilung. Die Nützlichkeit der Vertafelung der Verteilungsfunktion der Standardnormalverteilung wird schon im Jahre 1783 von Pierre Simon Marquis de Laplace erkannt. Zu Beginn des 19. Jahrhunderts erscheinen dann kurz hintereinander 2 Arbeiten über die Methode der kleinsten Quadrate und zwar (in dieser Reihenfolge) von Adrien Marie Legendre (1752–1833) und Carl Friedrich Gauß; letzterer baute darauf seine Theorie der Beobachtungsfehler auf. Die Methode der kleinsten Quadrate ist z. B. die Grundlage der modernen Regressionsrechnung, die schon in ersten Ansätzen (Minimierung der Abstandssumme) im Jahre 1755 bei Boscovich zu finden ist. Die u. a. für Gauß' Theorie der Beobachtungsfehler wichtige χ^2-Verteilung taucht erstmals im Jahre 1863 in der Arbeit „Über die Gesetzmäßigkeit in der Verteilung bei Beobachtungsreihen" von Ernst Abbe (1840–1905) auf. Unabhängig davon wurde sie von dem Mathematiker, Astronomen und Geodäten Friedrich Robert Helmert (1843–1917) im Jahre 1876 entdeckt, geriet aber dann zunächst in Vergessenheit. Dann beginnt eine recht stürmische Phase in der Entwicklung der modernen Statistik.

1888 wird von Sir Francis Galton der Korrelationskoeffizient, der heute nach Karl Pearson benannt wird, angegeben. Karl Pearson (1857–1936) entdeckt im Jahre 1900 die χ^2-Verteilung wieder, entwickelt den χ^2-Anpassungstest, treibt die Korrelations- und Regressionsrechnung voran, vertafelt die vollständige Betafunktion, entwickelt die Gamma- und die Betaverteilung usw. Im Jahre 1901 gründet er

bereits die Zeitschrift Biometrika, die das wichtigste Organ der angelsächsischen Statistiker war. In den Jahren 1900, 1901 und 1903 erscheinen die Arbeiten von G. U. Yule (1871–1951) über Assoziationskoeffizienten in Vierfeldertafeln. Wesentliche Fortschritte auf dem Gebiet der induktiven Statistik werden dann dadurch erzielt, daß der Mathematiker und Bierbrauer William Sealy Gosset (1876–1937) die t-Verteilung entdeckte und dies im Jahre 1908 – unter dem Pseudonym Student – veröffentlichte. Die Entwicklung der Methoden der Varianzanalyse von Sir Ronald Aylmer Fisher (1890–1962), der erstmals im Jahre 1935 mit Erscheinen seines Buches "The Design of Experiments" eine ausführliche Darstellung von Varianzanalyse und Versuchsplanung veröffentlichte, wurde erst durch die Ableitung der F-Verteilung durch George W. Snedecor möglich; den Namen F-Verteilung erhielt diese Verteilung zu Ehren Fishers. Am Aufschwung der statistischen Testtheorie seit Beginn des Zweiten Weltkrieges waren vor allem Jerzy Neyman und Egon S. Pearson maßgeblich beteiligt. Schließlich wurden im Jahre 1950 die Schlußweisen der Statistik von Abraham Wald (1902–1950) in seinem Buch "Statistical Decision Functions" einer entscheidungstheoretischen Betrachtungsweise zugänglich gemacht.

Der Ursprung des Wortes „Statistik" selbst liegt wohl im Rahmen der bisher historisch kaum angesprochenen beschreibenden Statistik, genauer in der Beschreibung des Staates (seines Heeres, Gewerbes, seiner Bevölkerung usw.). Auf die geschichtliche Entwicklung in diesem Bereich wollen wir nun noch kurz eingehen.

Schon in der Bibel nimmt die Volkszählung eine wichtige Stellung ein: Im vierten Buch Moses, das ja „Numeri" heißt, wird von der „Musterung des Volkes" gesprochen. Diese Volkszählung wurde von Gott angeordnet; dagegen lastete auf der von David (2. Buch Samuel) eigenmächtig durchgeführten Zählung ein Fluch; David wurde dafür sogar von Gott bestraft. Noch im Mittelalter zeigte diese biblische Warnung ihre Wirkung: Die Erhebungen beschränken sich in dieser Zeit quasi auf Vermögenserhebungen. Karl der Große etwa ordnete die Bestandsaufnahme all seiner Güter, Domänen usw. an, und 250 Jahre später wurde ein Buch, das man als Grundkataster des Okzidents bezeichnen könnte, nämlich das Doomsday Book, im Auftrag von Wilhelm dem Eroberer angelegt.

In den Kirchenbüchern Englands beginnt dann Anfang des 16. Jahrhunderts die systematische Aufzeichnung von Geburts- und Todesfällen. Dies geschieht auf Anweisung des Lordkanzlers Thomas Cromwell (eines entfernten Verwandten von Oliver Cromwell), der (vielleicht auch aus zusätzlichen Gründen) des Hochverrats beschuldigt und hingerichtet wird. Im Jahre 1693 erscheint dann ein Werk des britischen Astronomen Edmond Halley (1656–1742) über Sterbewahrscheinlichkeiten (anhand der Kirchenbücher der Stadt Breslau) und daraus resultierenden Prämien für Lebensversicherungen. Im Jahre 1741, mit Erscheinen seines Werkes „Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen", verhilft der preußische Feldprediger Johann Peter Süßmilch (1707–1767) der Bevölkerungsstatistik im deutschsprachigen Raum zum Durchbruch.

Die z. B. von Halley und Süßmilch vertretene Art der Bevölkerungsstatistik, der Staatskunde, ist auch unter dem Namen der politischen Arithmetik bekannt. In krassem Gegensatz hierzu steht eine andere, gleichzeitig betriebene Art der Staatskunde: die Staatsbeschreibung. Eines der ersten Werke dieser Richtung ist Francesco Sansovisos 1562 erschienenes „Del governo et amministratione di diversi regai". Im Jahre 1656 erscheint „Teutscher Fürstenstaat" von Veit Ludwig von

Seckendorf. Dieses Werk ist ein Vorläufer der Staatsbeschreibungen Gottfried Achenwalls (1719–1772), die dieser selbst 1748 als „Statistik" bezeichnet. Vermutlich bezieht er das Wort Statistik auf das neulateinische „status", das soviel wie Zustand oder auch Staat heißt; Statistik ist im Sinne Achenwalls also die Zustandsbeschreibung des Staates. Auf diese Prägung des Wortes „statistique" von Achenwall verweisen auch die französischen Wissenschaftler des 18. Jahrhunderts ausdrücklich.

Verstand Achenwall unter der Statistik noch die Vollerhebung zur Beschreibung des Staates, so ging man in der Folgezeit immer mehr zu den heute üblichen Stichprobenerhebungen über.

Kapitel I:
Aufbereitung und Darstellung von Datenmaterial – Deskriptive Statistik

In vielen Bereichen wie z. B. den Wirtschaftswissenschaften, den Naturwissenschaften, der Medizin oder der Psychologie wird man häufig mit der Aufgabe konfrontiert, umfangreiches und unübersichtliches Datenmaterial zu beschreiben und eventuell zu analysieren.

Die **deskriptive Statistik** stellt nun zum einen **deskriptive** also **beschreibende Methoden** bereit, d. h. sie ist das direkte Instrumentarium zur Beschreibung von Datenmaterial; zum anderen ist jede deskriptive Methode aber auch eine Vorstufe der Analyse von Datenmaterial, eine Vorstufe der **schließenden Statistik**, die wir in den weiteren Kapiteln dieses Buches noch behandeln werden: Einerseits müssen Hypothesen aufgestellt werden, die aus den Resultaten der Deskription abgeleitet werden, andererseits müssen deskriptive Kenngrößen zur Hypothesenprüfung bestimmt werden.

Bevor wir uns mit den Methoden der deskriptiven Statistik konkret beschäftigen, soll nun im 1. Abschnitt ein Überblick über diese Methoden gegeben werden. Weiterhin werden in Abschnitt 1 einige der grundlegenden Begriffe, die uns nicht nur in diesem Kapitel, sondern auch in allen folgenden Kapiteln begegnen, erläutert.

1. Grundlegende Begriffe und Überblick

Die deskriptive Statistik dient der Beschreibung, der Strukturierung und der Verdeutlichung unübersichtlichen und umfangreichen Datenmaterials: Wo kommt dieses Material her, um was für Daten handelt es sich? Diese und ähnliche Fragen müssen zunächst einmal untersucht werden.

Daten kann man z. B. durch Befragung von Personen oder auch durch Messungen gewinnen.

Beispiel:

(a) Indem man eine Reihe von Personen nach ihrem Alter, ihrem Familienstand, ihrer Kinderzahl usw. fragt, erhält man eine Vielzahl von Daten.
(b) Befragt man Besitzer von landwirtschaftlichen Betrieben nach der Größe der ihnen zur Verfügung stehenden landwirtschaftlichen Nutzflächen, nach der Güte dieser Nutzflächen, nach dem Viehbestand ihres Betriebes und ähnlichem, so gewinnt man ebenfalls eine Vielzahl verschiedener Daten.
(c) Mißt man zu verschiedenen Zeitpunkten Temperatur, Luftdruck, Niederschlagsmengen, Luftfeuchtigkeit usw., so stehen umfangreiche Datenmengen, die Auskunft über das Klima geben sollen, zur Verfügung.
(d) Eine Vielzahl von Daten erhält man etwa auch dann, wenn mittels verschiedener Tests Musikalität, sprachliche Begabung, mathematische Begabung usw. von Personen ermittelt werden.

1.1. Untersuchungseinheiten, Merkmale und Merkmalsausprägungen

Die befragten Personen bzw. die Objekte, an denen Messungen vorgenommen wer-

den, nennt man auch **Untersuchungseinheiten**; die Größen, auf die sich Fragen oder Messungen beziehen, heißen **Merkmale**.

Beispiel:

(e) Im Beispiel (a) sind die Personen die Untersuchungseinheiten und etwa Alter und Familienstand die beobachteten Merkmale. Im Beispiel (b) ist der Besitzer eines landwirtschaftlichen Betriebes bzw. der Betrieb selbst eine Untersuchungseinheit. Die interessierenden Merkmale sind hier z. B. Größe und Güte der landwirtschaftlichen Nutzfläche.

(f) Im Beispiel (c) ist die Untersuchungseinheit das „Wetter" zu einem bestimmten Zeitpunkt, und die beobachteten Merkmale sind Klimafaktoren wie z. B. die Temperatur. Im Beispiel (d) schließlich sind die Personen wiederum Untersuchungseinheiten, an denen Merkmale wie Musikalität beobachtet und geeignet gemessen werden.

Bei der Beobachtung von Merkmalen an Untersuchungseinheiten stellt man die **Merkmalsausprägung** (den **Merkmalswert**) jedes Merkmals an jeder Untersuchungseinheit fest; die Merkmalsausprägungen insgesamt sind dann identisch mit dem gewonnenen Datenmaterial.

Beispiel:

(g) Im Beispiel (a) sind etwa die Ausprägungen des Merkmals „Alter" verschiedene Zahlen von Jahren, wie z. B. 20 Jahre oder 75 Jahre, und die möglichen Ausprägungen des Merkmals „Familienstand" sind: ledig, verheiratet, geschieden, verwitwet.

1.2. Charakterisierung von Merkmalen

Wie wir in Beispiel (g) schon sehen, können Merkmalsausprägungen sehr unterschiedlich sein: Die Ausprägungen sind nicht immer Zahlen, sie können z. B. auch Zustände sein. (Aus diesem Grund sprechen wir auch von Daten- und nicht von Zahlenmaterial.) Ausgehend von solchen Unterschieden möglicher Ausprägungen von Merkmalen werden Merkmale in Klassen eingeteilt. Eine Art der Klassifizierung ist die Unterscheidung **quantitativer** und **qualitativer Merkmale**: Die Ausprägungen quantitativer Merkmale unterscheiden sich durch ihre Größe, die Ausprägungen qualitativer Merkmale durch ihre Art.

Beispiel:

(h) Quantitative Merkmale sind z. B. Alter, Gewicht, Körpergröße und Einkommen von Personen, Wasser- und Lufttemperaturen, Lebensdauern von Maschinen oder Viehbestände von landwirtschaftlichen Betrieben. Dagegen sind Merkmale wie Geschlecht, Beruf, Rasse, Haarfarbe, Fabrikat qualitativ. Auch Schulnoten, Güteklassen und z. B. Härtegrade von Wasser sind qualitative Merkmale, die mitunter jedoch auch quantitativ aufgefaßt werden.

Um die Ausprägung eines Merkmals messen oder erfragen zu können, muß man natürlich zunächst eine **Skala** (diese ist oft schon vorgegeben) festlegen, die alle möglichen Ausprägungen eines Merkmals beinhaltet. Auch die Art, das **Niveau** von solchen Skalen kann zur Klassifizierung von Merkmalen herangezogen werden.

Die Skala mit dem niedrigsten Niveau ist die sogenannte **Nominalskala**: Die Werte einer Nominalskala unterliegen keiner Rangfolge und sind nicht vergleich-

bar. Merkmale, deren Ausprägungen nur einer solchen Skala genügen, nennt man auch **nominale Merkmale**.

Beispiel:

(i) Die Ausprägungen von Merkmalen wie Farbe, Beruf oder Rasse sind nicht vergleichbar und unterliegen keiner Rangfolge: Man kann z. B. nicht sagen, daß Schlosser besser als Bäcker sind oder umgekehrt. Es handelt sich hier also um nominale Merkmale.

Ein etwas höheres Niveau hat schon die **Ordinalskala**, deren Werte sich in ihrer Intensität unterscheiden und nach der Stärke der Intensität ordnen lassen. Ausprägungen **ordinaler Merkmale** unterliegen also einer bestimmten Rangfolge; Abstände zwischen verschiedenen Ausprägungen (Werten der Skala) lassen sich jedoch nicht interpretieren.

Beispiel:

(j) Merkmale mit Ausprägungen wie gut, mittel und schlecht, also etwa Zensuren oder Härtegrade, sind ordinale Merkmale: die Ausprägungen unterliegen einer Rangfolge, Abstände sind nicht interpretierbar. Bei Schulnoten wird man etwa sagen können 1 (sehr gut) ist besser als 3 (befriedigend); der Abstand der Noten $3 - 1 = 2$ läßt sich jedoch nicht interpretieren, da 1 und 3 willkürlich gewählte Zahlen, die gewissen Leistungen zugeordnet werden, sind. Man könnte die Leistung „befriedigend" etwa auch mit 6 und die Leistung „sehr gut" mit 10 beziffern; der Abstand wäre dann $10 - 6 = 4$.

Die Skala mit dem höchsten Niveau ist die **metrische Skala**, die dadurch gekennzeichnet ist, daß ihre Werte nicht nur einer Reihenfolge unterliegen, sondern die Abstände zwischen den Werten auch interpretierbar sind. **Metrische** oder **metrisch skalierte Merkmale** sind natürlich gerade solche, deren Ausprägungsskala metrisch ist.

Beispiel:

(k) Lebensdauern, Gewichte, Größen, Stromstärken u. ä. sind natürlich metrische Merkmale. Man kann etwa nicht nur sagen, daß eine Stromstärke von 10 Ampère größer ist als eine von 1 Ampère; vielmehr läßt sich auch der Abstand der Stromstärken interpretieren: 10 Ampère sind $10 - 1 = 9$ Ampère mehr als 1 Ampère.

Wir wollen nun noch eine weitere Art der Klassifizierung von Merkmalen kennenlernen, nämlich die Einteilung in **stetige** und **diskrete Merkmale**. Ein diskretes Merkmal kann nur endlich viele oder höchstens abzählbar unendlich viele Ausprägungen besitzen; dagegen nennen wir ein Merkmal stetig, wenn es als Ausprägung jeden beliebigen Wert in einem Bereich haben kann.

Beispiel:

(l) Ein diskretes Merkmal ist etwa das „Geschlecht" von Personen. Es hat nur zwei, also endlich viele Ausprägungen, nämlich männlich und weiblich. Auch die „Anzahl der Einwohner einer Stadt" ist ein diskretes Merkmal, dessen Ausprägungen prinzipiell alle positiven ganzen Zahlen (also 1,2,3, ...) sind; dies sind abzählbar unendlich viele Ausprägungen. Das Merkmal „Temperatur" dagegen ist stetig, denn es kann beliebige Werte in einem Bereich als Ausprägungen besitzen. Auch das Merkmal „Lebensdauer" ist stetig, denn die

Lebensdauer einer Untersuchungseinheit kann eine beliebige positive Zahl sein.

Es sei noch erwähnt, daß häufig stetige Merkmale nur diskret beobachtet werden, was z. B. daran liegen kann, daß ein Meßverfahren nicht beliebig genau ist: Mißt man etwa Temperaturen, so werden die Meßergebnisse zumeist nur in vollen Gradzahlen angegeben.

1.3. Grundgesamtheit und Stichprobe

Bisher sind wir davon ausgegangen, daß an Untersuchungseinheiten Merkmale beobachtet, erhoben werden. Was sind nun aber die Untersuchungseinheiten einer **Erhebung**? Wie wählt man diese aus? Für eine reine Beschreibung von Datenmaterial sind solche Fragen zweitrangig, denn Aussagen werden hier nur über die konkret beobachteten Untersuchungseinheiten gemacht. Sieht man jedoch die deskriptive Statistik als Vorstufe der schließenden Statistik, so sind diese Fragen grundlegend.

Beispiel:

(m) Befragt man 1000 Personen danach, ob sie arbeitslos sind oder nicht, so erlaubt die rein deskriptive Statistik nur Aussagen der Form „Von den 1000 Personen sind 50 Personen oder 5% arbeitslos". Möchte man jedoch ausgehend von einer solchen Aussage schließen können, wieviel Prozent aller Personen (z. B. in der Bundesrepublik Deutschland) arbeitslos sind, so muß im vorhinein überlegt werden, welche Personen grundsätzlich befragt werden sollen und wie man aus allen möglichen Untersuchungseinheiten 1000 Personen für die Erhebung auswählt.

Die Frage nach den möglichen Untersuchungseinheiten erfordert eine **Abgrenzung der Grundgesamtheit** aller Untersuchungseinheiten.

Beispiel:

(m) Interessiert man sich für den prozentualen Anteil der Arbeitslosen an den arbeitsfähigen Personen zwischen 17 und 65 Jahren, die in der Bundesrepublik Deutschland leben, oder für den Prozentsatz der Arbeitslosen an der Gesamtbevölkerung? Im ersten Fall sind alle in der Bundesrepublik Deutschland lebenden, arbeitsfähigen Personen zwischen 17 und 65 Jahren mögliche Untersuchungseinheiten (sie bilden die Grundgesamtheit), im zweiten Fall besteht die Grundgesamtheit aus allen in der Bundesrepublik Deutschland lebenden Personen.

Welche Personen nun wirklich in die Erhebung einbezogen werden, ist ein Problem der Stichprobentheorie. Diese wird erst im Kapitel V in den Abschnitten 1 bis 3 ausführlich behandelt und ist hier auch nicht so wesentlich.

Im Beispiel (m) haben wir postuliert, daß nur 1000 Personen befragt werden. Prinzipiell könnte man auch alle zur Grundgesamtheit gehörigen Personen befragen. Schöpft man eine Grundgesamtheit voll aus, so spricht man von einer **Vollerhebung** oder **Totalerhebung**; tut man dies nicht, so führt man eine **Stichprobenerhebung** oder **Teilerhebung** durch.

Beispiel:

(n) Die bekanntesten und grundlegendsten Erhebungen zur Bestandsaufnahme der Bevölkerung sind die Volkszählungen. Dies sind Totalerhebungen mit

Haushalten als Untersuchungseinheiten; sie sollen in der Bundesrepublik Deutschland in der Regel alle zehn Jahre durchgeführt werden. Der lange Abstand zwischen den Volkszählungen wird durch die 1957 eingeführte „Repräsentativerhebung der Bevölkerung und des Erwerbslebens" (**Mikrozensus**) überbrückt. Der Mikrozensus ist natürlich eine Teilerhebung.

Die Gründe für die Durchführung von Stichprobenerhebungen anstelle von Totalerhebungen sind vielfältig: Stichprobenerhebungen sind billiger und manchmal auch genauer als Totalerhebungen, da sie im Detail sorgfältiger durchgeführt werden können. Auch ist eine Vollerhebung mitunter sinnlos und zu langwierig: Will man etwa die Lebensdauern einer bestimmten Serie von Kühlschränken analysieren, so läge ein Ergebnis der Totalerhebung ja erst dann vor, wenn alle Kühlschränke defekt sind; eine Qualitätsbeurteilung ist für eventuelle Käufer dann uninteressant. Auch wäre die Dauer eines entsprechenden Versuchs viel zu lang.

1.4. Überblick über die Methoden der deskriptiven Statistik

Um das Datenmaterial einer Erhebung übersichtlich zu gestalten, bietet sich zunächst eine **tabellarische** und **graphische Darstellung**, die in den Abschnitten 2 und 3 behandelt wird, an. Bei metrischen und ordinalen Merkmalen empfiehlt es sich etwa, die beobachteten Merkmalsausprägungen der Größe nach zu ordnen. Weiterhin bietet es sich an, zu jeder Ausprägung eines Merkmals die Anzahl der Untersuchungseinheiten, bei denen sie beobachtet worden ist, zu berechnen und zum Beispiel durch unterschiedlich hohe Säulen graphisch darzustellen.

Oft ist es auch nützlich, große Datenmengen durch wenige charakteristische Größen (**Kenngrößen**) grob zu beschreiben. Zum einen kann man Zentren – sogenannte **Lageparameter** –, vergleiche auch Abschnitt 4, der beobachteten Ausprägungen eines Merkmals (z. B. die am häufigsten beobachtete Ausprägung) bestimmen, zum anderen können Maßzahlen der Variation der Ausprägungen – sogenannte **Streuungsparameter** –, vergleiche Abschnitt 5, berechnet werden (z. B. die Größe des Bereichs, in dem alle Ausprägungen liegen).

In Abschnitt 6 beschäftigen wir uns mit der graphischen Darstellung und der Bestimmung der **Konzentration** von Merkmalsausprägungen. Zwei Fragen, die mittels dieser Betrachtungen behandelt werden können, sind etwa die folgenden: Wieviel Prozent des gesamten Nettoeinkommens aller Beschäftigten entfallen auf wieviel Prozent der Beschäftigten? Entfällt vielleicht der größte Teil des Einkommens auf eine nur sehr kleine Zahl von Beschäftigten?

Im Beispiel (m) haben wir uns mit dem prozentualen Anteil der Arbeitslosen an der Gesamtbevölkerung beschäftigt. Hier wird die Zahl der Arbeitslosen zur Gesamtbevölkerung ins Verhältnis gesetzt. Solche sogenannten **Verhältniszahlen**, zu denen auch die **Indizes** gehören, werden in Abschnitt 7 ausführlich behandelt.

In Abschnitt 8 wird eine spezielle Kenngröße bei der Erhebung von Lebensdauern, nämlich die **empirische Ausfallrate**, vorgestellt.

Erhebt man an den Untersuchungseinheiten einer Erhebung mehrere Merkmale zugleich (z. B. Alter, Körpergröße und Gewicht von Personen), so kann man die einzelnen Merkmale (**eindimensionale Merkmale**) auch zu einem **mehrdimensionalen Merkmal** zusammenfassen. Auf die Darstellung und Beschreibung der Zusammenhänge bei speziell zwei Merkmalen wird in Abschnitt 9 eingegangen.

Schließlich werden in Abschnitt 10 noch einige Möglichkeiten zur Vereinfachung der Berechnung von Kenngrößen angegeben.

2. Der Häufigkeitsbegriff

Bei einer Erhebung wird nun an n **Untersuchungseinheiten** ein Merkmal X beobachtet, d.h. an jeder Einheit wird die Ausprägung dieses Merkmals festgestellt: Sind $a_1, \ldots, a_k$ die möglichen **Ausprägungen des Merkmals** X, so wird also der i-ten Untersuchungseinheit ($i = 1, \ldots, n$) seine Ausprägung a_j als **Merkmalswert** x_i zugeordnet:

$$x_i = a_j(i).$$

Beispiel: Die Arbeiter (Untersuchungseinheiten) einer Fabrik werden gefragt, ob sie mit öffentlichen Verkehrsmitteln (a_1), privaten Verkehrsmitteln (a_2) oder zu Fuß (a_3) zu ihrer Arbeitsstelle gelangen. Kommt etwa der sechste befragte Arbeiter zu Fuß, so ist der zugehörige Merkmalswert gerade

$$x_6 = a_3(6) = \text{„zu Fuß“}.$$

Die insgesamt n Merkmalswerte $x_1, \ldots, x_n$ heißen auch **Beobachtungsreihe** oder **Urliste**.

Beispiel: Bei n = 100 Schulkindern einer Jahrgangsstufe sollen die Zähne auf Karies untersucht werden: Bei jedem Kind wird die Anzahl der kariösen Zähne registriert, wobei sich die Urliste $x_1, \ldots, x_{100}$ aus **Tab. 1** ergibt.

Tab. 1: Urliste der Anzahl kariöser Zähne bei 100 Schulkindern

1,0,0,3,1,5,1,2,2,0,1,0,5,2,1,0,1,0,0,4,0,1,1,3,0,
1,1,1,3,1,0,1,4,2,0,3,1,1,7,2,0,2,1,3,0,0,0,0,6,1,
1,2,1,0,1,0,3,0,1,3,0,5,2,1,0,2,4,0,1,1,3,0,1,2,1,
1,1,1,2,2,0,3,0,1,0,1,0,0,0,5,0,4,1,2,2,7,1,3,1,5

Wie man am Beispiel erkennt, ist das Material einer Urliste oft sehr unübersichtlich, so daß man – zumal häufig nur interessiert, welche Ausprägung eines Merkmals wie oft auftritt – zu einer anderen Darstellungsform übergehen sollte.

2.1. Absolute und relative Häufigkeiten

Beobachtet man an n Untersuchungseinheiten ein Merkmal X, das in k Ausprägungen $a_1, \ldots, a_k$ vorkommt, so ist

$$H_n(a_j) = \text{„Anzahl der Fälle, in denen } a_j \text{ auftritt“ für } j = 1, \ldots, k$$

die **absolute Häufigkeit** der Ausprägung a_j; natürlich ist

$$H_n(a_1) + H_n(a_2) + \ldots + H_n(a_k) = \sum_{j=1}^{k} H_n(a_j) = n.$$

Die absoluten Häufigkeiten hängen stark vom Umfang n der Erhebung ab, so daß sie sich nicht zum Vergleich von Erhebungen unterschiedlichen Umfangs eignen. Daher bietet es sich an, eine Größe zu suchen, die unabhängig vom Umfang n ist. Eine solche ist die **relative Häufigkeit**

$$h_n(a_j) = \frac{1}{n} \cdot H_n(a_j) \quad \text{für} \quad j = 1, \ldots, k,$$

die den prozentualen Anteil der Untersuchungseinheiten angibt, die die Ausprä-

gung a_j tragen: $h_n(a_j)$ nimmt unabhängig von n stets Werte zwischen Null und Eins an, und es gilt

$$h_n(a_1) + \ldots + h_n(a_k) = \frac{1}{n} \cdot \sum_{j=1}^{k} H_n(a_j) = \frac{n}{n} = 1 .$$

Wie wir im Beispiel noch sehen werden, gewinnt man die absoluten und relativen Häufigkeiten der Ausprägungen eines Merkmals aus der Urliste am besten durch Anfertigen einer **Strichliste**.

Beispiel: Zur Urliste aus Tab. 1 sollen nun die absoluten und relativen Häufigkeiten der einzelnen Ausprägungen des beobachteten Merkmals „Anzahl kariöser Zähne" mittels einer Strichliste bestimmt werden, vgl. **Tab. 2.**

Tab. 2: Absolute und relative Häufigkeiten der verschiedenen Anzahlen kariöser Zähne von 100 Schulkindern, vgl. auch Tab. 1

Anzahl der kariösen Zähne a_j	Strichliste	$H_n(a_j)$	$h_n(a_j)$
0	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	30	0,30
1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||||	34	0,34
2	𝍸 𝍸 ||||	14	0,14
3	𝍸 𝍸	10	0,10
4	||||	4	0,04
5	𝍸	5	0,05
6	|	1	0,01
7	||	2	0,02
Gesamt		100	1,00

Die Folge der Zahlen $H_n(a_1), \ldots, H_n(a_k)$ nennt man auch die **absolute Häufigkeitsverteilung** und die Folge $h_n(a_1), \ldots, h_n(a_k)$ auch die **relative Häufigkeitsverteilung** des beobachteten Merkmals. Dazu sei noch angemerkt, daß es sich bei metrischen und ordinalen Merkmalen empfiehlt, die Ausprägungen $a_1, \ldots, a_k$ der Größe nach zu ordnen, wie dies im Beispiel bereits geschehen ist.

2.2. Die graphische Darstellung von Häufigkeiten

Eine Folge von Zahlen, wie dies die Häufigkeitsverteilungen sind, läßt sich besonders gut übersehen, wenn man sie graphisch darstellt. Hierzu eignen sich zunächst Stabdiagramme, Häufigkeitspolygone oder Histogramme.

Beim **Stabdiagramm** wird über jeder Ausprägung auf der Abszisse die zugehörige Häufigkeit als senkrechte Strecke abgetragen; es eignet sich besonders dann, wenn das beobachtete Merkmal diskret ist. In **Abb. 1** sind die relativen Häufigkeiten des Beispiels „Anzahl kariöser Zähne", vgl. auch Tab. 2, in Form eines Stabdiagramms dargestellt.

Das **Häufigkeitspolygon** erhält man, wenn man die Endpunkte der Strecken des Stabdiagramms miteinander verbindet, wie dies für die relativen Häufigkeiten unseres Beispiels in **Abb. 2** geschehen ist. Besonders geeignet ist ein Häufigkeitspolygon zur Darstellung zeitlicher Verläufe.

Beispiel: Drei Jahre lang wird an jedem Monatsende die Anzahl der Arbeitslosen registriert. Die insgesamt 36 Anzahlen werden dann durch ein Häufigkeitspolygon

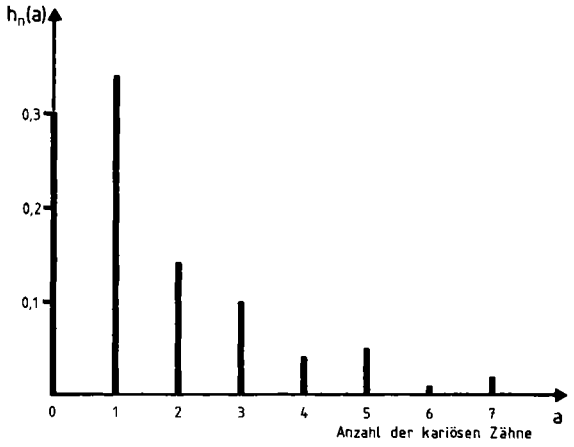

Abb. 1: Stabdiagramm

Abb. 2: Häufigkeitspolygon

beschrieben, wenn auf der Abzisse (= „Zeitachse") über jedem Monat die Anzahl der Arbeitslosen abgetragen wird und diese Anzahlen verbunden werden.

Ein **Histogramm** erhält man, wenn man die beobachteten Häufigkeiten als aneinanderstoßende Rechtecke, deren **Flächen proportional den Häufigkeiten** sind, darstellt. Diese graphische Darstellungsform, die in **Abb. 3** für unser Beispiel „Anzahl kariöser Zähne" veranschaulicht wird, eignet sich vor allem dann, wenn das beobachtete Merkmal stetig ist.

Will man speziell eine Struktur darstellen, so bedient man sich sogenannter **Flächendiagramme**, bei denen die Häufigkeiten durch Flächen repräsentiert werden.

Beispiel: In **Abb. 4** ist ein spezielles Flächendiagramm, nämlich ein **Kreisdiagramm** gewählt worden, um die Anzahlen der Beschäftigten in verschiedenen Wirtschafts-

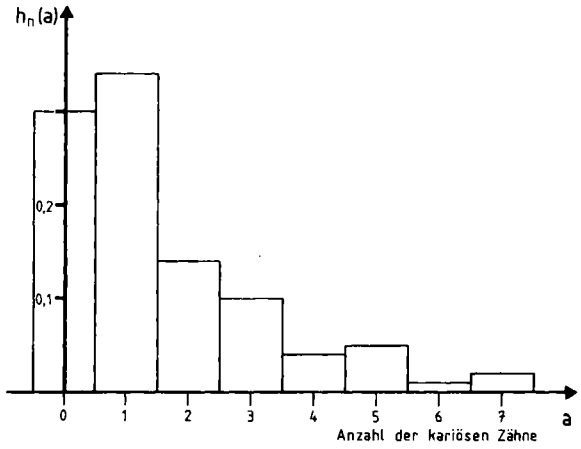

Abb. 3: Histogramm

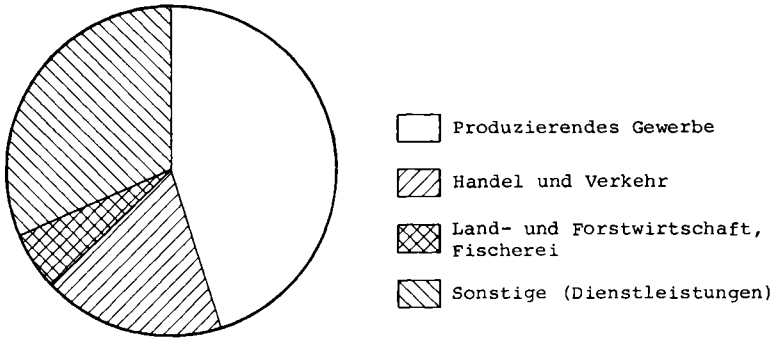

Abb. 4: Kreisdiagramm; Erwerbstätige nach Wirtschaftsbereichen 1979

zweigen darzustellen: Der Kreis wurde in Flächen zerlegt, die proportional zu den beobachteten Anzahlen sind.

2.3. Die empirische Verteilungsfunktion

Hat man bei ordinalen oder metrischen Merkmalen die möglichen Ausprägungen $a_1, \ldots, a_k$ der Größe nach geordnet, so bietet sich eine weitere Möglichkeit zur Beschreibung einer Beobachtungsreihe an. Statt wie bisher die Häufigkeiten einer jeden Ausprägung zu betrachten, bildet man sogenannte Summenhäufigkeiten:

Die **absolute Summenhäufigkeit** der j-ten Ausprägung a_j ist die Zahl der Untersuchungseinheiten, bei denen eine der Ausprägungen $a_1, a_2, \ldots, a_j$ beobachtet wurde, d.h. die absolute Summenhäufigkeit von a_j ist gegeben als

$$H_n(a_1) + H_n(a_2) + \ldots + H_n(a_j) = \sum_{i=1}^{j} H_n(a_i) \quad \text{für} \quad j = 1, \ldots, k.$$

Entsprechend versteht man unter der **relativen Summenhäufigkeit** von a_j die Größe

$$h_n(a_1) + \ldots + h_n(a_j) = \frac{1}{n} \cdot \sum_{i=1}^{j} H_n(a_i),$$

also die Summe der relativen Häufigkeiten von $a_1, \ldots, a_j$.

Beispiel: Im Beispiel der kariösen Zähne bei Schulkindern, vgl. Tab. 2, gibt die absolute Summenhäufigkeit von a_j an, wieviele der 100 untersuchten Kinder höchstens a_j kariöse Zähne haben, und die relative Summenhäufigkeit von a_j ist der Anteil der Kinder mit höchstens a_j kariösen Zähnen.

Durch die Folge der relativen Summenhäufigkeiten wird nun die sogenannte **empirische Verteilungsfunktion** oder **Summenhäufigkeitsfunktion** $S_n(x)$ des beobachteten Merkmals X bestimmt:

Für alle Zahlen x mit $a_j \leqq x < a_{j+1}$ ist

$$S_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{j} H_n(a_i) = \sum_{i=1}^{j} h_n(a_i).$$

Die empirische Verteilungsfunktion ist also auf jedem Intervall $[a_j, a_{j+1})$ konstant und springt bei a_{j+1} um den Wert $h_n(a_{j+1})$ nach oben. Die erste Sprungstelle liegt bei der kleinsten, die letzte bei der größten beobachteten Merkmalsausprägung. Natürlich ist

$$S_n(x) = 0 \quad \text{für} \quad x < a_1 \quad \text{und} \quad S_n(x) = 1 \quad \text{für} \quad x \geqq a_k.$$

Beispiel: Ausgehend von Tab. 2 wollen wir die empirische Verteilungsfunktion der Anzahl kariöser Zähne bei 100 Schulkindern bestimmen. In **Tab. 3** sind die relativen Summenhäufigkeiten angegeben, und in **Abb. 5** ist die zugehörige empirische Verteilungsfunktion graphisch dargestellt.

Tab. 3: Relative Summenhäufigkeiten im Beispiel, vgl. auch Tab. 2

Anzahl der kariösen Zähne a_j	relative Häufigkeit $h_n(a_j)$	relative Summenhäufigkeit $\sum_{i=1}^{j} h_n(a_i)$
0	0,30	0,30
1	0,34	0,64
2	0,14	0,78
3	0,10	0,88
4	0,04	0,92
5	0,05	0,97
6	0,01	0,98
7	0,02	1,00

3. Der Häufigkeitsbegriff bei Klassenbildung

Die Häufigkeitsverteilung wird dann unübersichtlich, wenn ein Merkmal in sehr vielen Ausprägungen auftritt. In solchen Fällen betrachtet man nicht mehr die einzelnen Ausprägungen, sondern faßt verschiedene Ausprägungen in Klassen zusammen. Dies ist natürlich insbesondere dann der Fall, wenn das beobachtete Merkmal stetig ist. Man geht dann zumeist von vornherein von Klassen aus, denn

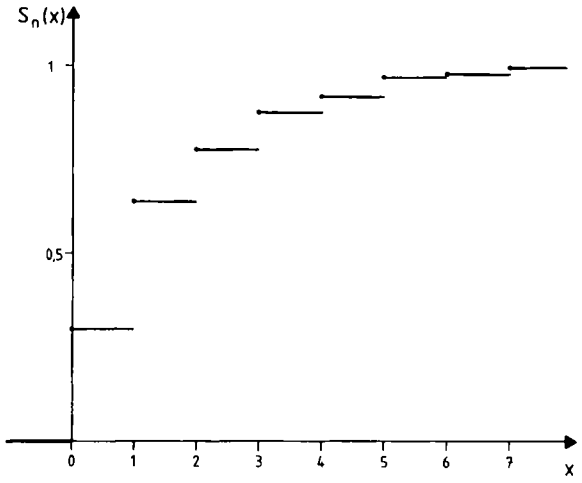

Abb. 5: Empirische Verteilungsfunktion im Beispiel

- selbst wenn man formal keine Klassen bildet, kann eine Merkmalsausprägung in Wirklichkeit einer Klasse entsprechen, weil man nicht beliebig genau messen kann;

Beispiel: An n Hühnereiern wird jeweils der größte Abstand x_i zwischen zwei Schalenpunkten gemessen. Wählt man als Maßeinheit „Millimeter", so entspricht etwa ein Meßwert $x_i = 57$ mm einem wahren Abstand, der im Intervall [56,5 mm, 57,5 mm) liegt. In Wirklichkeit betrachtet man also Klassen; der Meßwert x_i entsteht also durch Rundung oder weil nicht ganz exakt gemessen werden kann.

- bei stetigen Merkmalen kann eine Ausprägung jede beliebige Zahl in einem bestimmten Bereich sein, so daß es selten vorkommt, daß mehrere gleiche Ausprägungen beobachtet werden. Häufig kommt es jedoch vor, daß einige Ausprägungen dicht beieinander liegen. Es bietet sich dann an, diese in einer Klasse zusammenzufassen.

Beispiel: In **Tab. 4** sind die Lebensdauern von 87 Kühlaggregaten (der Größe nach geordnet) angegeben. Es kommt kaum einmal vor, daß mehrere Lebensdauern gleich sind; jedoch sind die Unterschiede oft relativ gering.

Tab. 4: Lebensdauern von 87 Kühlaggregaten (der Größe nach geordnet) in Jahren

0,05	0,06	0,06	0,08	0,11	0,13	0,15	0,16	0,20
0,22	0,24	0,25	0,25	0,28	0,31	0,34	0,37	0,42
0,43	0,47	0,51	0,51	0,53	0,59	0,60	0,61	0,63
0,68	0,75	0,76	0,76	0,79	0,87	0,88	0,88	0,92
0,99	1,12	1,16	1,18	1,22	1,27	1,35	1,38	1,39
1,42	1,45	1,49	1,53	1,69	1,74	1,81	1,83	1,87
1,92	1,93	2,07	2,09	2,15	2,22	2,24	2,36	2,39
2,41	2,47	2,49	2,53	2,64	2,69	2,83	2,90	3,21
3,25	3,49	3,61	3,80	3,88	4,37	4,58	4,62	5,29
5,68	6,02	6,23	6,71	7,82	9,93			

3.1. Die Klassenbildung

Wie geht man nun konkret bei der Klassenbildung vor? Wir betrachten hier nur den Fall ordinaler oder metrischer Merkmale, denn die Ausprägungen nominaler Merkmale sind meist nicht so zahlreich.

Zunächst zerlegt man das Intervall, in dem alle Beobachtungswerte liegen, in mehrere Teilintervalle, die sogenannten **Klassen** oder **Gruppen**.

Eine solche Klasse ist dann eindeutig bestimmt durch **Klassenmitte** und **Klassenbreite** oder durch **untere** und **obere Klassengrenze**. Natürlich muß im vorhinein festgelegt werden, zu welcher Klasse die Klassengrenzen gehören:
Zu jeder Klasse gehört die untere Klassengrenze bzw.
zu jeder Klasse gehört die obere Klassengrenze bzw.
man zählt einen Wert, der auf die Klassengrenze fällt, je zur Hälfte zu den beiden benachbarten Klassen.

Um Schwierigkeiten aus dem Weg zu gehen, sollte man, wenn möglich, die Klassengrenzen so wählen, daß kein Beobachtungswert dort liegt.

Obwohl sich rechnerische Vereinfachungen ergeben, wenn man als Klassenmitten möglichst einfache Zahlen und alle Klassen gleich breit wählt, legt man aus sachlichen Gründen oft verschieden breite Klassen fest.

Beispiel: Sind die beobachteten Merkmalsausprägungen die in einer Klausur erreichten Punktzahlen, so bietet es sich an, die Klasseneinteilung durch zugehörige Noten festzulegen, obwohl die zu den Noten gehörigen Punktintervalle nicht alle gleich groß sind.

3.2. Absolute und relative Häufigkeiten bei Klassenbildung

Genauso, wie wir in Abschnitt 2 absolute und relative Häufigkeiten von Merkmalsausprägungen bestimmt haben, können wir Klassenhäufigkeiten bestimmen.

Ist p die Anzahl der gebildeten Klassen, so ist die **absolute Klassenhäufigkeit** der i-ten Klasse K_i für $i = 1, \ldots, p$ gerade

$$H_i = \text{„Anzahl der Beobachtungswerte in der Klasse } K_i\text{“},$$

und die **relative Klassenhäufigkeit** dieser Klasse ist

$$h_i = \frac{1}{n} H_i .$$

Beispiel: (a) In Abschnitt 2 haben wir als ein Beispiel die Anzahl kariöser Zähne bei 100 Schulkindern betrachtet. Die Merkmalsausprägungen waren 0,1, ..., 7 kariöse Zähne, und die zugehörigen absoluten und relativen Häufigkeiten sind in Tab. 2 angegeben. Wir wollen nun in diesem Beispiel zur Klassenbildung übergehen:
geringer Kariesbefall $\hat{=}$ 0, 1 oder 2 kariöse Zähne,
mittlerer Kariesbefall $\hat{=}$ 3 oder 4 kariöse Zähne,
starker Kariesbefall $\hat{=}$ 5, 6 oder 7 kariöse Zähne.
Da es für die graphische Darstellung, wie wir noch sehen werden, günstig ist, daß die Klassen aneinanderstoßen, setzen wir $K_1 = [-0{,}5;\ 2{,}5)$, $K_2 = [2{,}5;\ 4{,}5)$ und $K_3 = [4{,}5;\ 7{,}5)$, obwohl etwa die Ausprägung 2,5 kariöse Zähne gar nicht auftreten kann.

Ausgehend von dieser Klasseneinteilung in 3 Klassen sind in **Tab. 5** die absoluten und relativen Klassenhäufigkeiten angegeben.

Tab. 5: Absolute und relative Klassenhäufigkeiten im Beispiel „Anzahl kariöser Zähne"; vgl. auch Tab. 2

Klasse K_i	kariöse Zähne	absolute Klassenhäufigkeit	relative Klassenhäufigkeit
$K_1 = [-0{,}5;\ 2{,}5)$	geringer Kariesbefall	78	0,78
$K_2 = [2{,}5;\ 4{,}5)$	mittlerer Kariesbefall	14	0,14
$K_3 = [4{,}5;\ 7{,}5)$	starker Kariesbefall	8	0,08
Gesamt		100	1

(b) Auch im Beispiel der Lebensdauern von 87 Kühlaggregaten, vgl. Tab. 4, wollen wir eine Klassenbildung vornehmen. Wir bilden 10 Klassen der Breite 1 (= 1 Jahr), d.h. es ist

K_1 = {Lebensdauer von weniger als einem Jahr} = [0,00; 1,00),
K_2 = {Lebensdauer von mindestens einem und weniger als zwei Jahren} = [1,00; 2,00), ...,
K_{10} = {Lebensdauer von mindestens neun und weniger als zehn Jahren} = [9,00; 10,00).

Die zugehörigen absoluten und relativen Klassenhäufigkeiten sind in **Tab. 6** angegeben.

Tab. 6: Klasseneinteilung für die Lebensdauern der 87 Kühlaggregate, vgl. auch Tab. 4

Klasse K_i	absolute Klassenhäufigkeiten H_i	relative Klassenhäufigkeiten h_i (gerundet)
$K_1 = [0;1)$	37	0,425
$K_2 = [1;2)$	19	0,218
$K_3 = [2;3)$	15	0,172
$K_4 = [3;4)$	6	0,069
$K_5 = [4;5)$	3	0,034
$K_6 = [5;6)$	2	0,022
$K_7 = [6;7)$	3	0,034
$K_8 = [7;8)$	1	0,011
$K_9 = [8;9)$	0	0,000
$K_{10} = [9;10)$	1	0,011

3.3. Die graphische Darstellung von Häufigkeiten bei Klassenbildung

Will man Häufigkeiten von in Klassen zusammengefaßten Beobachtungswerten graphisch darstellen, so bedient man sich zumeist eines **Histogramms**. Hier wird es wichtig, daß die Klassen aneinanderstoßend gewählt wurden, denn nur dann lassen sich die Häufigkeiten sinnvoll durch aneinanderstoßende Rechtecke darstellen. Die Flächen dieser Rechtecke beschreiben die Häufigkeit einer Klasse, d.h. nicht die Höhe des Rechtecks über einer Klasse ist proportional zur Klassenhäufigkeit, sondern die Fläche selbst ist es. Somit ist das Produkt von Rechteckhöhe und Klassenbreite einer Klasse proportional zur beobachteten Klassenhäufigkeit.

Wir wollen uns im folgenden auf die Darstellung relativer Klassenhäufigkeiten beschränken. Setzen wir die Fläche des Rechtecks über der i-ten Klasse K_i gleich der relativen Klassenhäufigkeit h_i dieser Klasse, so ist es recht einfach, die Höhe r_i des Rechtecks zu bestimmen, wenn wir die Klassenbreite mit b_i bezeichnen:

$$h_i = b_i \cdot r_i, \quad \text{d.h.} \quad r_i = \frac{h_i}{b_i}.$$

Beispiel: (a) In Tab. 5 sind die relativen und absoluten Klassenhäufigkeiten für das Beispiel „Anzahl kariöser Zähne" bei 3 Klassen angegeben. Wir wollen nun die relativen Klassenhäufigkeiten graphisch in Form eines Histogramms darstellen. Dazu müssen wir zunächst die Klassenbreiten angeben und daraus die Höhen r_i der Rechtecke über den Klassen bestimmen, wie dies in **Tab. 7** erfolgt ist. Nun können wir das Histogramm erstellen, vgl. **Abb. 6**.

Tab. 7 Berechnungen für das Histogramm in Abb. 6

Klasse K_i	relative Klassenhäufigkeit h_i	Klassenbreite b_i	Rechtecks-höhe r_i
K_1	0,78	3	0,26
K_2	0,14	2	0,07
K_3	0,08	3	0,027

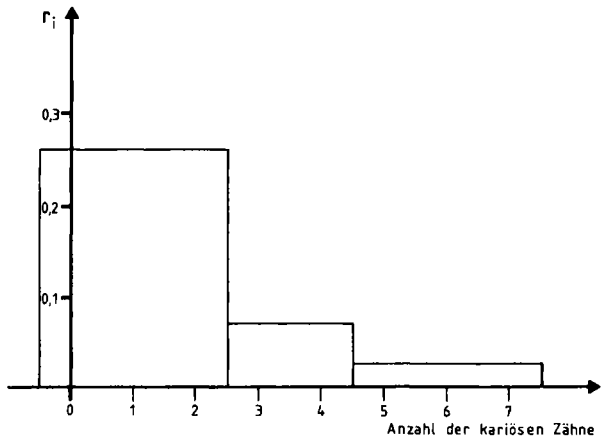

Abb. 6: Histogramm zum Beispiel

(b) Auch zum Beispiel der 87 Kühlaggregate wollen wir ein Histogramm ausgehend von Tab. 6 erstellen. Da hier die Klassenbreiten alle 1 Jahr sind, stimmen die relativen Klassenhäufigkeiten h_i jeweils mit der Höhe r_i der Rechtecke über den Klassen überein, so daß sich das Histogramm aus **Abb. 7** ergibt.

3.4. Die empirische Verteilungsfunktion bei Klassenbildung

Hauptsächlich dann, wenn die einzelnen Beobachtungswerte $x_1, \ldots, x_n$ nicht zur Verfügung stehen und nur noch gewisse Klassenhäufigkeiten bekannt sind, wird

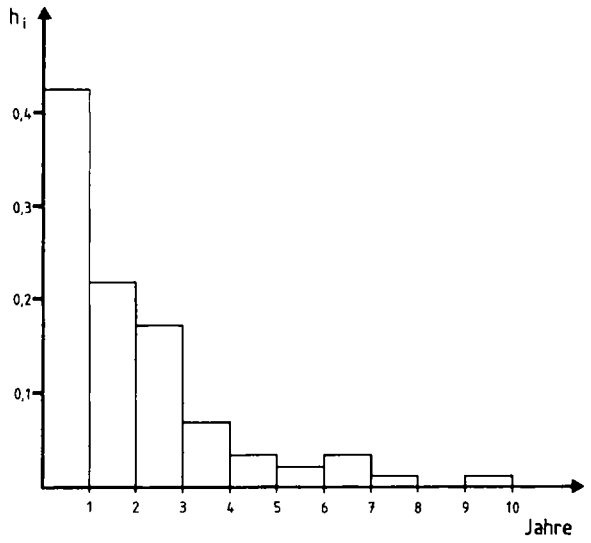

Abb. 7: Histogramm der Lebensdauer von 87 Kühlaggregaten bei Klassenbreite 1

man die **empirische Verteilungsfunktion** des beobachteten Merkmals X aus diesen Klassenhäufigkeiten berechnen. Hierbei gibt es verschiedene Vorgehensweisen, von denen zwei hier vorgestellt werden sollen, man vgl. aber auch Kap. XIV, 1.6.

Insbesondere im Fall symmetrisch um die Klassenmitten verteilter Beobachtungswerte bietet es sich an, den Wert der empirischen Verteilungsfunktion $S_p(x)$ (bei Bildung von p Klassen) an der Stelle x als die Summe der relativen Häufigkeiten aller Klassen, deren Mitten kleiner oder gleich x sind, festzulegen. Bezeichnet also m_i für $i = 1, \ldots, p$ die Klassenmitte der i-ten Klasse K_i und ist h_i die relative Klassenhäufigkeit dieser Klasse, so kann man

$$S_p(x) = \sum_{j=1}^{i} h_j \quad \text{für } m_i \leqq x < m_{i+1}$$

definieren. Die Sprungstellen der empirischen Verteilungsfunktion sind dann identisch mit den Klassenmitten; für $x < m_1$ ist $S_p(x) = 0$ und für $x \geqq m_p$ ist $S_p(x)$ natürlich gerade Eins.

Eine andere Möglichkeit der Definition der empirischen Verteilungsfunktion ist die folgende: Der Wert $S_p(x)$ ist für jede Zahl x gleich der Summe der relativen Klassenhäufigkeiten aller Klassen, deren Endpunkte kleiner oder gleich x sind. Ist also y_i für $i = 1, \ldots, p$ der Endpunkt der Klasse K_i, die mit Häufigkeit h_i auftritt, so ist

$$S_p(x) = \sum_{j=1}^{i} h_j \quad \text{für } y_i \leqq x < y_{i+1}.$$

Bei dieser Definition der empirischen Verteilungsfunktion fallen die Sprungstellen also mit den Klassenendpunkten zusammen und es ist $S_p(x) = 0$ für $x < y_1$ sowie $S_p(x) = 1$ für $x \geqq y_p$.

Beispiel: In **Abb. 8** und **Abb. 9** sind die empirischen Verteilungsfunktionen bei

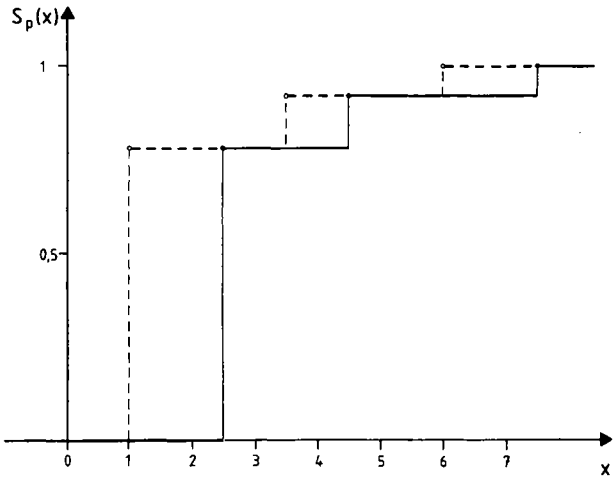

Abb. 8: Empirische Verteilungsfunktion der Anzahl kariöser Zähne bei Berechnung durch das Klassenende y_i (durchgehende Linie) und die Klassenmitte m_i (gestrichelte Linie)

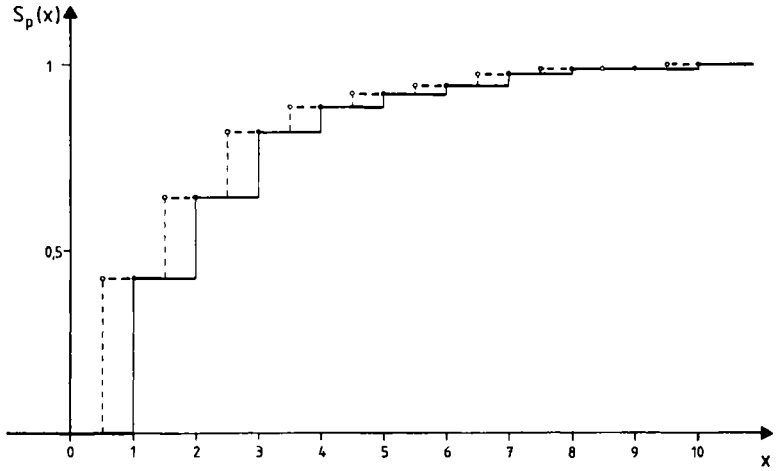

Abb. 9: Empirische Verteilungsfunktion der Lebensdauer von Kühlaggregaten bei Berechnung durch das Klassenende y_i (durchgehende Linie) und die Klassenmitte m_i (gestrichelte Linie)

Tab. 8: Berechnung der empirischen Verteilungsfunktion bei Klassenbildung im Beispiel „Anzahl kariöser Zähne"

Klasse K_i	Klassenmitte m_i	Klassenend-punkt y_i	relative Klassenhäufigkeit h_i	$S_p(x)$ für $m_i \leqq x < m_{i+1}$ bzw. für $y_i \leqq x < y_{i+1}$
K_1	1,0	2,5	0,78	0,78
K_2	3,5	4,5	0,14	0,92
K_3	6,0	7,5	0,08	1,00

Tab. 9: Berechnung der empirischen Verteilungsfunktion bei Klassenbildung im Beispiel „Lebensdauer von Kühlaggregaten"

Klasse K_i	Klassenmitte m_i	Klassenendpunkt y_i	relative Klassenhäufigkeit h_i	$S_p(x)$ für $m_i \leqq x < m_{i+1}$ bzw. für $y_i \leqq x < y_{i+1}$
K_1	0,5	1	0,425	0,425
K_2	1,5	2	0,218	0,644
K_3	2,5	3	0,172	0,816
K_4	3,5	4	0,069	0,885
K_5	4,5	5	0,034	0,920
K_6	5,5	6	0,022	0,943
K_7	6,5	7	0,034	0,977
K_8	7,5	8	0,011	0,989
K_9	8,5	9	0,000	0,989
K_{10}	9,5	10	0,011	1,000

Klassenbildung für die Beispiele „Anzahl kariöser Zähne", vgl. auch Tab. 5, und „Lebensdauer von Kühlaggregaten", vgl. Tab. 6, für beide Berechnungsmöglichkeiten dargestellt. Die nötigen Berechnungen findet man in **Tab. 8** bzw. **Tab. 9**.

4. Lagemaße von Häufigkeitsverteilungen

Oft läßt sich eine Beobachtungsreihe durch wenige charakteristische Größen kennzeichnen. Ein solches Charakteristikum ist z.B. eine Maßzahl, die in geeigneter Weise ein „Zentrum" der Beobachtungswerte $x_1, \ldots, x_n$ angibt. Einige solcher Maßzahlen, die man **Lagemaße** nennt, sollen hier vorgestellt werden.

4.1. Das arithmetische Mittel

Sind $x_1, \ldots, x_n$ die beobachteten Merkmalswerte eines Merkmals X mit Ausprägungen $a_1, \ldots, a_k$, so ist das **arithmetische Mittel (Mittelwert, durchschnittlicher Wert)**

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot \sum_{j=1}^{k} a_j \cdot H_n(a_j)$$

ein sinnvolles Lagemaß vor allem dann, wenn das Merkmal X metrisch skaliert ist.

Sind die Beobachtungswerte in p Klassen eingeteilt, so ist natürlich

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot H_i \quad \text{mit } \bar{x}_i = \frac{1}{H_i} \cdot \sum_{x_i \in K_i} x_i$$

das arithmetische Mittel aller Beobachtungswerte; $\bar{x}_i$ ist gerade das arithmetische Mittel der Beobachtungswerte in der i-ten Klasse.

Mitunter kommt es vor, daß bei gruppierten Daten die Klassenmittel nicht bekannt sind. Dann berechnet man als arithmetisches Mittel, ausgehend von den Klassenmitten $m_1, \ldots, m_p$,

$$\bar{x}_M = \frac{1}{n} \cdot \sum_{i=1}^{p} m_i \cdot H_i ;$$

natürlich gilt $\bar{x} = \bar{x}_M$ nur dann, wenn die Beobachtungswerte jeder Klasse symmetrisch um die Klassenmitte verteilt liegen.

Beispiel: Im Beispiel „Anzahl kariöser Zähne bei 100 Schulkindern" aus Abschnitt 2, vgl. Tab. 2, ergibt sich als arithmetisches Mittel der ungruppierten Daten

$$\bar{x} = \tfrac{1}{100}(0 \cdot 30 + 1 \cdot 34 + 2 \cdot 14 + 3 \cdot 10 + 4 \cdot 4 + 5 \cdot 5 + 6 \cdot 1 + 7 \cdot 2) = \\ = 1{,}53,$$

d.h.: Im Durchschnitt haben die 100 Kinder 1,53 kariöse Zähne. Gehen wir zu den gruppierten Daten aus Tab. 5 über, so ergibt sich natürlich ebenfalls

$$\bar{x} = 1{,}53,$$

denn es ist

$$\bar{x}_1 = \tfrac{1}{78}(0 \cdot 30 + 1 \cdot 34 + 2 \cdot 14) = \tfrac{31}{39} = 0{,}79,$$
$$\bar{x}_2 = \tfrac{1}{14}(3 \cdot 10 + 4 \cdot 4) = \tfrac{23}{7} = 3{,}29 \quad \text{und}$$
$$\bar{x}_3 = \tfrac{1}{8}(5 \cdot 5 + 6 \cdot 1 + 7 \cdot 2) = \tfrac{45}{8} = 5{,}625, \quad \text{so daß sich}$$
$$\bar{x} = \tfrac{1}{100}(\tfrac{31}{39} \cdot 78 + \tfrac{23}{7} \cdot 14 + \tfrac{45}{8} \cdot 8) = 1{,}53$$

ergibt.

Gehen wir einmal davon aus, daß wir die Klassenmittel nicht mehr kennen, und berechnen

$$\bar{x}_M = \tfrac{1}{100}(1 \cdot 78 + 3{,}5 \cdot 14 + 6 \cdot 8) = 1{,}75.$$

Der wirkliche Durchschnittswert $\bar{x}$ wird hier also überschätzt, da alle Klassenmittel durch die Klassenmitten überschätzt werden.

Gibt man sich allgemein Zahlen (**Gewichte**) $w_1, \ldots, w_n$ mit $0 \leqq w_i \leqq 1$ und $\sum_{i=1}^{n} w_i = 1$ vor, so erhält man ein sogenanntes **gewogenes arithmetisches Mittel**

$$\bar{x}^w = \sum_{i=1}^{n} w_i x_i .$$

Mit $w_1 = \ldots = w_n = 1/n$ ergibt sich wieder das gewöhnliche arithmetische Mittel $\bar{x}$.

4.2. Der Median und das α-Quantil

4.2.1. Der Median einer Beobachtungsreihe

Der **Median** oder **Zentralwert** ist dadurch charakterisiert, daß jeweils mindestens 50% der Beobachtungen $x_1, \ldots, x_n$ einen Wert größer oder gleich bzw. kleiner oder gleich dem Median annehmen. Sind also $x_{(1)} \leqq \ldots \leqq x_{(n)}$ die der Größe nach geordneten Beobachtungswerte (sind mehrere Werte gleich, so spielt deren Reihenfolge keine Rolle), so läßt sich der Median $\tilde{x}_{0,5}$ als

$$\tilde{x}_{0,5} = \begin{cases} x_{((n+1)/2)} & , \text{ falls n ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{((n+2)/2)}), & \text{falls n gerade} \end{cases}$$

angeben. Bei einer ungeraden Anzahl von Merkmalsausprägungen ist der Median also der Wert in der Mitte der geordneten Reihe, bei einer geraden Anzahl ist jede Zahl, die zwischen den beiden mittleren Werten liegt, ein Median; wir aber wählen hier stets das arithmetische Mittel der beiden Werte als speziellen Median. Der Median hat gegenüber dem durchschnittlichen Wert $\bar{x}$ als Lagemaß den Vorteil,

daß er auch bei ordinal skalierten Merkmalen verwandt werden kann und daß er weniger empfindlich gegenüber Ausreißern in der Beobachtungsreihe ist; d. h. Werte, die weit von allen übrigen entfernt liegen, beeinflussen den Wert des Medians kaum.

Beispiel: (a) Im Beispiel „Lebensdauern von 87 Kühlaggregaten" aus Abschnitt 3 wollen wir den Median $\tilde{x}_{0,5}$ der Beobachtungsreihe bestimmen; in Tab. 4 ist die geordnete Reihe aller 87 Beobachtungswerte zu finden. Da $n = 87$ ungerade ist, entspricht der Median gerade dem $(n+1)/2 = 88/2 = 44$ten Wert in der Meßreihe:

$$\tilde{x}_{0,5} = x_{(44)} = 1{,}38.$$

(b) Der Median der Beobachtungsreihe $x_{(1)} = 3$, $x_{(2)} = 5$, $x_{(3)} = 6$, $x_{(4)} = 9$ ist $\tilde{x}_{0,5} = (5+6)/2 = 5{,}5$.

Auch wenn gruppierte Daten vorliegen, läßt sich natürlich der Median bestimmen: Sind $K_1, \ldots, K_p$ die verschiedenen Klassen, die mit den relativen Häufigkeiten $h_1, \ldots, h_p$ auftreten, so bestimmt man zunächst die Klasse K_i, die den Median enthält. Für sie muß gelten

$$\sum_{j=1}^{i-1} h_j < 0{,}5 \quad \text{und} \quad \sum_{j=1}^{i} h_j \geqq 0{,}5, \quad \text{wobei} \sum_{j=1}^{0} h_j = 0 \text{ gesetzt wird.}$$

Hat man nun die Klasse K_i gefunden, so ermittelt man ihre untere Klassengrenze u_i und ihre Breite b_i. Der Median ist dann gerade

$$\tilde{x}_{0,5} = u_i + \frac{0{,}5 - \sum_{j=1}^{i-1} h_j}{h_i} \cdot b_i ;$$

graphisch bedeutet dies, daß der Median die gesamte Histogrammfläche in zwei gleichgroße Teile zerlegt, vgl. **Abb. 10**.

Beispiel: Im Beispiel der Kühlaggregate aus Abschnitt 3 soll nun der Median aus den gruppierten Daten bei Klassenbreite 1 bestimmt werden, vgl. Tab. 6. Der Median liegt hier in der Klasse K_2, denn es ist

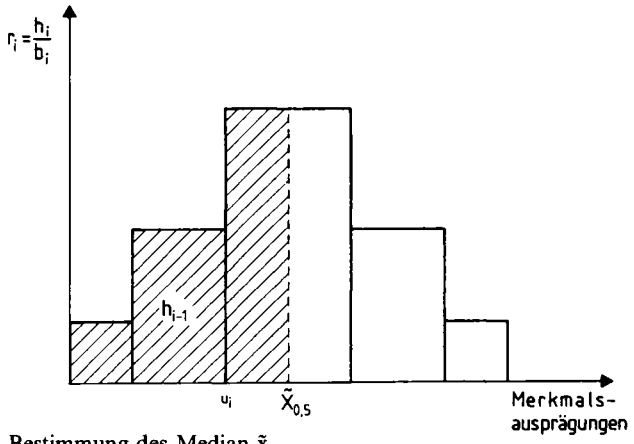

Abb. 10: Bestimmung des Median $\tilde{x}_{0,5}$

$$h_1 \simeq 0{,}425 < 0{,}5 \quad \text{und} \quad h_1 + h_2 \simeq 0{,}644 > 0{,}5.$$

Weiter ist

$$u_2 = 1 \quad \text{bzw.} \quad b_2 = 1$$

die untere Grenze bzw. die Breite der Klasse K_2, so daß sich als Median der Wert

$$\tilde{x}_{0,5} = u_2 + \frac{0{,}5 - h_1}{h_2} \cdot b_2 \simeq 1 + \frac{0{,}5 - 0{,}425}{0{,}218} \cdot 1 = 1 + \frac{0{,}075}{0{,}218} \simeq 1{,}344$$

ergibt.

4.2.2. Das α-Quantil einer Beobachtungsreihe

Ist $x_{(1)} \leqq \ldots \leqq x_{(n)}$ die geordnete Beobachtungsreihe zur Urliste $x_1, \ldots, x_n$, so bezeichnet man als α-**Quantil** $\tilde{x}_\alpha (0 < \alpha < 1)$ dieser Reihe gerade den Wert

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & , \text{ falls } n \cdot \alpha \text{ keine ganze Zahl ist} \\ & \quad (k \text{ ist dann die auf } n \cdot \alpha \text{ folgende ganze Zahl}) \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & \text{falls } n \cdot \alpha \text{ eine ganze Zahl ist} \\ & \quad (\text{Es ist dann } k = n \cdot \alpha) \end{cases}$$

Das α-Quantil $\tilde{x}_\alpha$ wird für ganze Zahlen $n \cdot \alpha$ so gewählt, daß $n \cdot \alpha$ Beobachtungswerte kleiner oder gleich und $n(1-\alpha)$ Werte größer oder gleich $\tilde{x}_\alpha$ sind. Ist $n \cdot \alpha$ keine ganze Zahl, so soll für das α-Quantil gelten, daß $k-1$ Werte kleiner und $n-k$ Werte größer als $\tilde{x}_\alpha$ sind, wobei k die auf $n \cdot \alpha$ folgende ganze Zahl ist.

Ist $n \cdot \alpha$ eine ganze Zahl, so würde diese Bedingung prinzipiell durch jede Zahl zwischen $x_{(n \cdot \alpha)}$ und $x_{(n \cdot \alpha + 1)}$ erfüllt; wir haben hier jedoch als spezielles α-Quantil den Durchschnitt dieser beiden Werte gewählt.

Wie man sieht ist der Median $\tilde{x}_{0,5}$ ein spezielles α-Quantil, nämlich gerade das 0,5-Quantil der Beobachtungsreihe. Weitere Quantile, die einen speziellen Namen tragen, sind das 0,25-Quantil und das 0,75-Quantil: Sie heißen **unteres** bzw. **oberes Quartil**.

Beispiel: Wie wollen das 0,1-Quantil der Lebensdauern von 87 Kühlaggregaten, vgl. Tab. 4 in Abschnitt 3, bestimmen. Es ist

$$n \cdot \alpha = 87 \cdot 0{,}1 = 8{,}7,$$

so daß, da 8,7 keine ganze Zahl ist, der neunte Wert $x_{(9)}$ in der Reihe der geordneten Beobachtungswerte dem 0,1-Quantil entspricht:

$$\tilde{x}_\alpha = \tilde{x}_{0,1} = x_{(9)} = 0{,}20.$$

Falls die Beobachtungsdaten nur gruppiert vorliegen, kann man bei der Bestimmung eines beliebigen α-Quantils analog zur Bestimmung des Medians vorgehen: Das α-Quantil liegt in der Klasse K_i, falls gilt

$$\sum_{j=1}^{i-1} h_j < \alpha \quad \text{und} \quad \sum_{j=1}^{i} h_j \geqq \alpha.$$

Ist dann u_i die untere Grenze dieser Klasse und b_i die Klassenbreite, so ist

$$\tilde{x}_\alpha = u_i + \frac{\alpha - \sum_{j=1}^{i-1} h_j}{h_i} \cdot b_i.$$

Durch diese Vorschrift wird $\tilde{x}_\alpha$ gerade so gewählt, daß die Fläche unterhalb des Histogramms der gruppierten Daten links von $\tilde{x}_\alpha$ die Größe α hat. (Die Gesamtfläche unterhalb des Histogramms hat gerade die Größe Eins.)

Beispiel: Zu den Daten aus Tab. 6 wollen wir das obere Quartil, d.h. das 0,75-Quantil, bestimmen. Dieses liegt in der Klasse K_3, deren untere Grenze bzw. Breite gerade

$$u_3 = 2 \quad \text{und} \quad b_3 = 1$$

sind, denn es ist

$$h_1 + h_2 \simeq 0{,}644 < 0{,}75 \quad \text{und} \quad h_1 + h_2 + h_3 \simeq 0{,}816 > 0{,}75.$$

Damit ergibt sich

$$\tilde{x}_\alpha = \tilde{x}_{0,75} = u_3 + \frac{\alpha - (h_1 + h_2)}{h_3} \cdot b_3 \simeq 2 + \frac{0{,}75 - 0{,}644}{0{,}172} \cdot 1 \simeq 2{,}616,$$

d.h. 75% der Kühlaggregate haben weniger als 2,616 Jahre gelebt.

4.3. Der Modalwert

Bei nominalskalierten Merkmalen wie Geschlecht oder Beruf kann man weder das arithmetische Mittel noch den Median als Lagemaß verwenden. Ein Lagemaß, das auch für diese Merkmale definiert ist, ist dagegen der **Modus (Modalwert, häufigster Wert)**. Er gibt die Ausprägung an, die die größte Häufigkeit in der Beobachtungsreihe besitzt.

Für den Modus x_{mod} gilt also: $h_n(x_{mod}) \geqq h_n(a_j)$ für alle Merkmalsausprägungen $a_1, \ldots, a_k$. Falls mehrere Ausprägungen diese Bedingung erfüllen, der Modus also nicht eindeutig ist, ist es nicht sinnvoll, ihn als Lagemaß zu verwenden.

Bei in Klassen eingeteilten Werten definiert man den Modus als Klassenmitte der am dichtesten besetzten Klasse, d.h. der Klasse K_i mit größter Histogrammhöhe r_i. Im Falle gleichbreiter Klassen bedeutet dies, daß der Modus die Klassenmitte der Klasse mit größter Klassenhäufigkeit h_i ist.

Beispiel: Im Beispiel mit den Kühlaggregaten besitzt, falls man Klassen der Breite Eins bildet, vgl. Tab. 6, die 1. Klasse K_1 die größte Klassenhäufigkeit. Daher ist $x_{mod} = 0{,}5$.

4.4. Das geometrische Mittel und das harmonische Mittel

Wenn die Merkmalsausprägungen relative Änderungen sind(z.B. Lohnerhöhung, Belegschaftsverminderung in %), ist es nicht sinnvoll, das arithmetische Mittel als Maß für einen Durchschnittswert zu verwenden, weil die Gesamtänderung nicht durch eine Summe, sondern durch ein Produkt beschrieben wird.

Bezeichnet $x_1, \ldots, x_n$ wiederum die Beobachtungsreihe und ist $x_i \geqq 0$ für $i = 1, \ldots, n$, so benutzt man als Lagemaß in solchen Fällen besser das **geometrische Mittel**

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}.$$

Beispiel: Vier Jahre lang wurden die Produktionssteigerungen eines Betriebes pro Jahr gemessen. Es ergaben sich die Werte aus **Tab. 10**.

Tab. 10: Produktionssteigerungen eines Betriebes in vier Jahren

Jahr	1. Jahr	2. Jahr	3. Jahr	4. Jahr
Produktionssteigerung	2%	11%	4%	7%

Falls die Produktion im Jahr 0 (Anfangsproduktion) a beträgt, gilt für die nächsten Jahre:

Produktion im 1. Jahr: $a \cdot 1{,}02$,
Produktion im 2. Jahr: $a \cdot 1{,}02 \cdot 1{,}11$,
Produktion im 3. Jahr: $a \cdot 1{,}02 \cdot 1{,}11 \cdot 1{,}04$,
Produktion im 4. Jahr: $a \cdot 1{,}02 \cdot 1{,}11 \cdot 1{,}04 \cdot 1{,}07$.

Die Gesamtsteigerung der Produktion ist also durch das Produkt $1{,}02 \cdot 1{,}11 \cdot 1{,}04 \cdot 1{,}07$ bestimmt. Es ist deshalb nicht sinnvoll, die Summe der Einzeländerungen zu bilden und durch die Anzahl der Beobachtungen zu teilen, um die durchschnittliche Steigerung zu erhalten. Statt dessen bildet man das geometrische Mittel:

$$\bar{x}_g = \sqrt[4]{1{,}02 \cdot 1{,}11 \cdot 1{,}04 \cdot 1{,}07} = \sqrt[4]{1{,}26} = 1{,}06.$$

Die durchschnittliche Produktionssteigerung beträgt also ungefähr 6%. Dies bedeutet, daß, falls in jedem Jahr eine Produktionssteigerung von 6% stattgefunden hätte, im 4. Jahr diese konstante Steigerung zur selben Gesamtsteigerung wie wirklich vorhanden geführt hätte.

Das geometrische Mittel wird oft auch dort als Lagemaß verwandt, wo eine logarithmische Skala sinnvoll ist, denn wenn man den Logarithmus des geometrischen Mittels bildet, erhält man wieder ein arithmetisches Mittel:

$$\ln \bar{x}_g = \ln \sqrt[n]{x_1 \dots x_n} = \frac{1}{n} (\ln x_1 + \dots + \ln x_n).$$

Der Wert des geometrischen Mittels ist nie größer als der des entsprechenden arithmetischen Mittels. Vielmehr gilt

$$\bar{x}_g \leqq \bar{x},$$

wobei die Gleichheit genau dann zutrifft, wenn

$$x_1 = x_2 = \dots = x_n$$

ist.

Gibt man sich Gewichte $w_1, \dots, w_n$ mit $0 \leqq w_i \leqq 1$ und $\sum_{i=1}^{n} w_i = 1$ vor, so erhält man ein **gewogenes geometrisches Mittel**

$$\bar{x}_g^w = x_1^{w1} \cdot x_2^{w2} \cdot \dots \cdot x_n^{wn}.$$

Mit $w_1 = \dots = w_n = 1/n$ ergibt sich wieder das gewöhnliche geometrische Mittel $\bar{x}_g$.

Hat man nur positive oder negative Beobachtungswerte $x_1, \dots, x_n$, so ist das **harmonische Mittel** $\bar{x}_h$ gegeben durch

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

Genauso wie beim arithmetischen Mittel und beim geometrischen Mittel können natürlich auch beim harmonischen Mittel die Beobachtungen $x_1, \ldots, x_n$ mit Zahlen $w_1, \ldots, w_n$, $0 \leqq w_i \leqq 1$ und $\sum_{i=1}^{n} w_i = 1$, gewichtet werden. Man erhält dann ein

gewogenes harmonisches Mittel

$$\bar{x}_h^w = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Für $w_1 = \ldots = w_n = 1/n$ ergibt sich natürlich gerade wieder das gewöhnliche harmonische Mittel $\bar{x}_h$.

Beispiel: Hat man etwa Beziehungen $U = P \cdot M$ und $u_i = x_i \cdot m_i$ und ist $\sum u_i = U$, $\sum m_i = M$, so ergibt sich hieraus

$$P = \frac{U}{M} = \frac{\sum u_i}{\sum m_i} = \frac{\sum u_i}{\sum \frac{u_i}{x_i}} = \frac{1}{\sum \frac{w_i}{x_i}} \quad \text{mit} \quad w_i = u_i / \sum u_i .$$

P ergibt sich somit als mit w_i gewogenes harmonisches Mittel der x_i.

Obige Größen können z. B. sein:

U = Gesamtumsatz
M = Gesamtmenge
P = durchschnittlicher Preis pro Mengeneinheit
u_i = Einzelumsatz des i-ten Gutes
x_i = Einzelpreis pro Mengeneinheit des i-ten Gutes
m_i = umgesetzte Menge des i-ten Gutes.

Der durchschnittliche Preis P ergibt sich als mit den Umsatzanteilen w_i gewogenes harmonisches Mittel der Einzelpreise x_i.

Analog berechnet man beispielsweise auch Durchschnittsgeschwindigkeiten aufgrund der Beziehung

Weg = Geschwindigkeit · Zeit

oder durchschnittliche Stückzahlen wegen der Beziehung

Gesamtstückzahl = (durchschnittliche Stückzahl/Zeiteinheit) · Zeit.

Hat man nur positive Beobachtungswerte $x_1, \ldots, x_n$, so gilt stets die Beziehung

$$\bar{x}_h \leqq \bar{x}_g \leqq \bar{x} ;$$

bei identischen Beobachtungswerten $x_1 = \ldots = x_n$ sind die Mittel gleich.

4.5. Einige Bemerkungen zu den Lagemaßen

Bei mehrgipfligen und insbesondere bei U-förmigen Häufigkeitsverteilungen sind – im Gegensatz zu eingipfligen Verteilungen –, vgl. **Abb. 11, 12, 13,** die Lagemaße oft nicht charakteristisch für die Häufigkeitsverteilung.

Beispiel: Die Lebensdauer der Menschen im Mittelalter hat eine U-förmige Häufigkeitsverteilung: Die Säuglingssterblichkeit war sehr hoch; hatte man jedoch die ersten Jahre überlebt, so war die Chance, ein recht hohes Alter zu erreichen, groß.

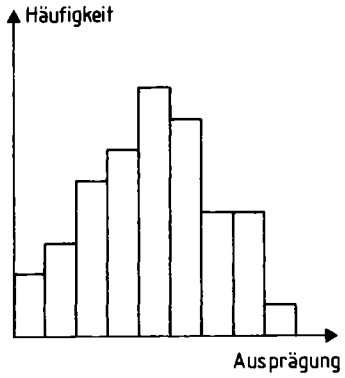

Abb. 11: Eingipflige Verteilung

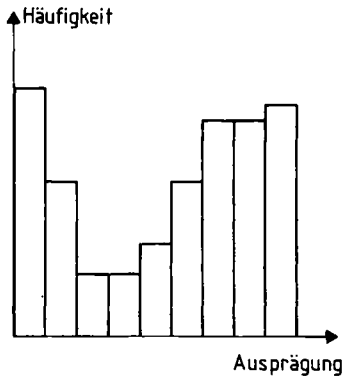

Abb. 12: U-förmige Verteilung

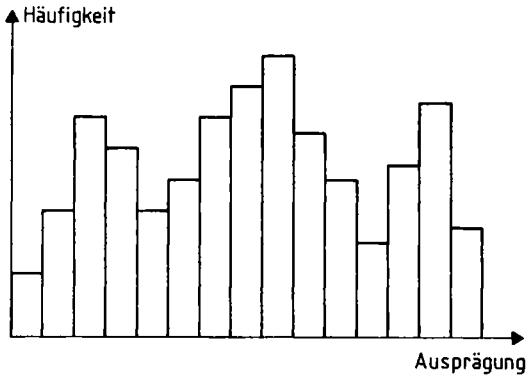

Abb. 13: Mehrgipflige Verteilung

Der Satz „Im Mittelalter wurden die Menschen durchschnittlich 25 Jahre alt" gibt daher eine falsche Vorstellung von der damaligen Lebenserwartung.

Bevor man also zur Charakterisierung einer nicht eingipfligen Häufigkeitsverteilung ein Lagemaß bestimmt, muß man überprüfen, ob dieses sinnvoll zu interpretieren ist.

Abschließend wollen wir noch die eingipfligen Häufigkeitsverteilungen näher charakterisieren. Bezeichnet $\bar{x}$ das arithmetische Mittel, $\tilde{x}_{0,5}$ den Median und x_{mod} den Modus einer eingipfligen Häufigkeitsverteilung, so heißt diese

rechtsschief oder **linkssteil**, falls $\bar{x} > \tilde{x}_{0,5} > x_{mod}$,
linksschief oder **rechtssteil**, falls $\bar{x} < \tilde{x}_{0,5} < x_{mod}$ und
symmetrisch , falls $\bar{x} = \tilde{x}_{0,5} = x_{mod}$,

vgl. auch **Abb. 14, 15** und **16**.

Beispiel: Die Verteilung der Anzahl kariöser Zähne bei 100 Schulkindern ist zwar mehrgipflig (Gipfel bei 1, 5 und 7 kariösen Zähnen), vgl. Abb. 3, aber nur der Gipfel

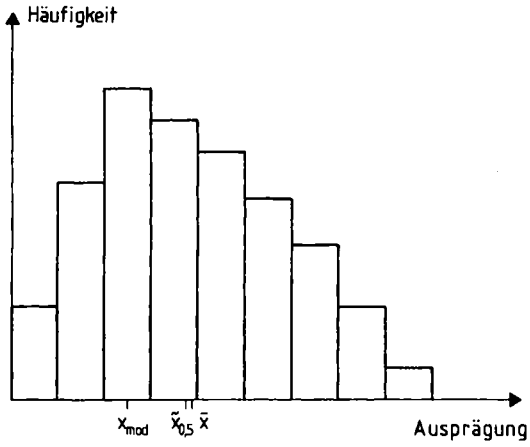

Abb. 14: Rechtsschiefe Verteilung

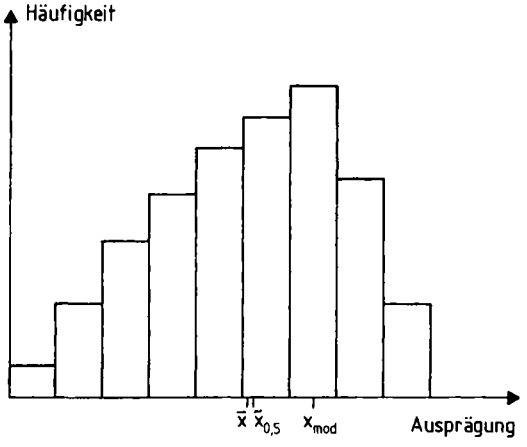

Abb. 15: Linksschiefe Verteilung

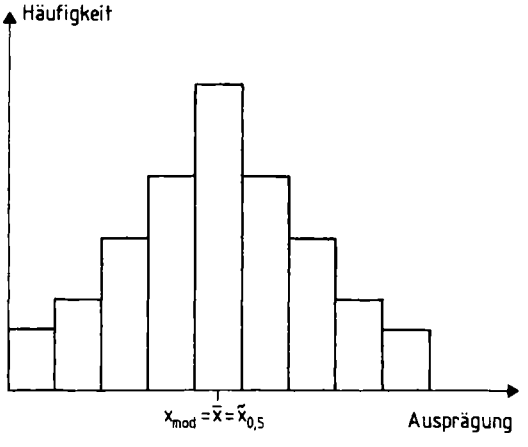

Abb. 16: Symmetrische Verteilung

bei Eins ist ausgeprägt, so daß man die Verteilung als annähernd eingipflig bezeichnen kann. Weiter ist

$$\bar{x} = 1{,}53 > \tilde{x}_{0,5} = 1 = x_{mod} = 1,$$

was auf eine rechtsschiefe Verteilung hindeutet. Bildet man zunächst Klassen wie in Abschnitt 3, Tab. 5 geschehen, so ist die Verteilung eingipflig, vgl. Abb. 6, und wegen

$$\bar{x} = 1{,}53 > \tilde{x}_{0,5} = 1{,}423 > x_{mod} = 1$$

rechtsschief.

5. Streuungsmaße von Häufigkeitsverteilungen

Lagemaße von Häufigkeitsverteilungen, wie wir sie in Abschnitt 4 kennengelernt haben, geben allein wenig Auskunft über eine Häufigkeitsverteilung. Sie beschreiben zwar ein Zentrum dieser Verteilung, liefern aber keinen Anhaltspunkt dafür, wie weit ein konkreter Merkmalswert von einem solchen Zentrum abweichen kann.

Beispiel: (a) Die Aussage, „Ein Beschäftigter der Firma Schulze verdient durchschnittlich 1347,– DM netto", gibt keine Informationen darüber, wieviel der einzelne Beschäftigte wirklich verdient. Es könnte sein, daß alle etwa diesen durchschnittlichen Verdienst haben oder daß einige sehr viel weniger, andere dafür aber sehr viel mehr verdienen.

(b) Beträgt der durchschnittliche Preis für ein Kilogramm Schweinefleisch in einer bestimmten Gegend 17,– DM, so heißt dies nicht unbedingt, daß der Preis zwischen 16,– und 18,– DM liegt. Es könnte auch sein, daß man in einigen Geschäften nur 12,– DM, in anderen dagegen etwa 20,– DM bezahlt.

Maße, die die Abweichung von einem Zentrum einer Häufigkeitsverteilung beschreiben, sind nun die sogenannten **Streuungsmaße** oder **Dispersionsmaße**. Kennt man ein Lagemaß und ein Streuungsmaß einer Häufigkeitsverteilung, so hat man schon eine recht präzise Vorstellung von derselben. Einige solche Streuungsmaße für metrisch skalierte Merkmale wollen wir im folgenden vorstellen.

5.1. Die Spannweite

Der **Streubereich** einer Häufigkeitsverteilung ist derjenige Wertebereich, in dem alle Merkmalswerte einer Beobachtungsreihe liegen. Er ist somit bestimmt durch die Angabe des kleinsten und des größten beobachteten Merkmalswertes in einer Urliste. Ist also $x_{(1)} \leqq \ldots \leqq x_{(n)}$ die geordnete Beobachtungsreihe, so ist das Intervall $[x_{(1)}; x_{(n)}]$ gerade der Streubereich der zugehörigen Häufigkeitsverteilung.

Beispiel: In $n = 40$ Geschäften wird der Preis für ein bestimmtes Farbfernsehgerät ermittelt. Das billigste Angebot lag bei $x_{(1)} = 1374$,– DM und das teuerste bei $x_{(n)} = 1829$,– DM. Diese beiden Zahlen geben den Streubereich der Häufigkeitsverteilung der Preise für das Gerät an.

Die Breite eines Streubereichs nennt man auch die **Spannweite (Range)** R einer Häufigkeitsverteilung:

$$R = x_{(n)} - x_{(1)}.$$

Beispiel: Die Spannweite für den Preis eines Fernsehgerätes war gerade $R = x_{(n)} - x_{(1)} = 1829 - 1374 = 455$.

Die Spannweiten verschiedener Beobachtungsreihen kann man nur dann vergleichen, wenn die Beobachtungsreihen die gleiche Länge n haben, denn natürlich kann die Verlängerung einer Reihe nur zur Folge haben, daß sich die Spannweite der Häufigkeitsverteilung entweder nicht verändert oder daß sie sich vergrößert. Eine Verkleinerung der Spannweite und damit des Streubereichs ist nicht möglich.

5.2. Der Quartilsabstand

Betrachten wir die Spannweite als Streuungsmaß einmal genauer, so fällt auf, daß sie lediglich vom kleinsten und vom größten Merkmalswert einer Beobachtungsreihe abhängt. Es kann nun vorkommen, daß z. B. der kleinste Wert sehr stark von allen anderen Werten abweicht, also ein sogenannter Ausreißer ist. Dann beeinflußt er den Wert R der Spannweite extrem stark.

Beispiel: Wäre im Beispiel aus Abschnitt 5.1 das zweitbilligste Angebot für den Farbfernseher etwa $x_{(2)} = 1509{,}-$ DM, so würde die Spannweite $R = 455$ eine schlechte Vorstellung von der Preisschwankung geben. Denn hätte man das Billigangebot $x_{(1)}$ nicht beobachtet, so wäre die Spannweite nur noch

$$1829 - 1509 = 320,$$

mithin sehr viel geringer gewesen.

Ein Streuungsmaß, das nicht so abhängig von Extremwerten und somit i. a. günstiger ist, ist der **Quartilsabstand**. Er gibt gerade die Größe des Bereichs zwischen oberem und unterem Quartil einer Beobachtungsreihe an; d. h. der Quartilsabstand ist

$$\tilde{x}_{0,75} - \tilde{x}_{0,25},$$

vgl. auch Abschnitt 4.2.2.

Zwischen unterem und oberem Quartil liegen mindestens 50% aller Merkmalswerte in einer Beobachtungsreihe, so daß der Quartilsabstand die Größe des Bereichs liefert, in dem etwa die Hälfte aller beobachteten Ausprägungen liegen.

Beispiel: n = 30 Glühbirnen werden einer Lebensdauerprüfung unterzogen; dabei stellt man die Lebensdauern x_i für $i = 1, \ldots, 30$ aus **Tab. 11** fest.

Tab. 11: Lebensdauer x_i von 30 Glühbirnen in Stunden

i	x_i	i	x_i	i	x_i
1	375,3	11	772,5	21	918,3
2	392,5	12	799,4	22	935,4
3	467,9	13	799,6	23	951,1
4	503,1	14	803,9	24	964,9
5	591,2	15	808,7	25	968,8
6	657,8	16	810,5	26	1006,5
7	738,5	17	812,1	27	1014,7
8	749,6	18	848,6	28	1189,0
9	752,0	19	867,0	29	1215,6
10	765,8	20	904,3	30	1407,2

Um den Quartilsabstand dieser Häufigkeitsverteilung bestimmen zu können, müssen zunächst die Quartile $\tilde{x}_{0,25}$ und $\tilde{x}_{0,75}$ berechnet werden, vgl. Abschnitt 4.2.2. Es ist

$$n \cdot 0{,}25 = 30 \cdot 0{,}25 = 7{,}5$$

und somit

$$\tilde{x}_{0,25} = x_{(8)} = x_8 = 749{,}6.$$

Weiterhin ergibt sich mit

$$n \cdot 0{,}75 = 30 \cdot 0{,}75 = 22{,}5$$

nun

$$\tilde{x}_{0,75} = x_{(23)} = x_{23} = 951{,}1.$$

Damit ist der Quartilsabstand der Häufigkeitsverteilung gerade

$$\tilde{x}_{0,75} - \tilde{x}_{0,25} = 951{,}1 - 749{,}6 = 201{,}5.$$

5.3. Die mittlere absolute Abweichung vom Median

Als Streuungsmaße lassen sich auch solche Größen verwenden, die eine durchschnittliche Abweichung der Merkmalswerte in einer Beobachtungsreihe von einem festen Wert messen. Eine Möglichkeit besteht etwa darin, die mittlere betragliche Abweichung von einem festen Wert c zu betrachten:

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - c| = \frac{1}{n} \sum_{j=1}^{k} |a_j - c| \cdot H_n(a_j);$$

hier bezeichnet x_i den Merkmalswert, der an der i-ten Beobachtungseinheit gemessen wurde und a_j die j-te Ausprägung des beobachteten Merkmals, die mit absoluter Häufigkeit $H_n(a_j)$ aufgetreten ist.

Es stellt sich nun die Frage, wie man den Wert c am besten wählt. Eine ausgezeichnete Stellung hat hier der Median $c = \tilde{x}_{0,5}$, denn es gilt die **Minimumeigenschaft des Medians**:

$$\sum_{i=1}^{n} |x_i - \tilde{x}_{0,5}| \leqq \sum_{i=1}^{n} |x_i - c| \quad \text{für alle Zahlen c.}$$

Die Größe

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}_{0,5}|$$

nennt man auch die **mittlere absolute Abweichung vom Median.**

Beispiel: Für das Beispiel der Glühbirnen aus Abschnitt 5.2, vgl. auch Tab. 11, wollen wir die mittlere absolute Abweichung vom Median berechnen. Der Median der Beobachtungsreihe selbst ist gerade das arithmetische Mittel der Werte

$$x_{(15)} = x_{15} = 808{,}7 \quad \text{und} \quad x_{(16)} = x_{16} = 810{,}5,$$

d.h. es ist

$$\tilde{x}_{0,5} = \frac{1}{2} \cdot (808{,}7 + 810{,}5) = \frac{1619{,}2}{2} = 809{,}6,$$

so daß sich die mittlere absolute Abweichung vom Median zu

$$d = \frac{1}{30} \sum_{i=1}^{30} |x_i - 809{,}6| =$$

$$= \frac{1}{30} \cdot (|375{,}3 - 809{,}6| + |392{,}5 - 809{,}6| + \ldots + |1407{,}2 - 809{,}6|) =$$

$$= \frac{1}{30} (434{,}3 + 417{,}1 + \ldots + 597{,}6) = 161{,}207$$

berechnet.

Auch wenn nur gruppierte Daten vorliegen, verwendet man die mittlere absolute Abweichung vom Median als Maß der Streuung einer Beobachtungsreihe.

Hat man p Klassen $K_1, \ldots, K_p$ gebildet, die mit absoluten Häufigkeiten $H_1, \ldots, H_p$ auftreten, so ist die mittlere absolute Abweichung vom Median gerade

$$\frac{1}{n} \sum_{j=1}^{p} |\bar{x}_j - \tilde{x}_{0,5}| \cdot H_j,$$

falls die Klassenmittel

$$\bar{x}_j = \frac{1}{H_j} \sum_{x_i \in K_j} x_i \quad \text{für } j = 1, \ldots, p$$

bekannt sind, und

$$\frac{1}{n} \sum_{j=1}^{p} |m_j - \tilde{x}_{0,5}| \cdot H_j$$

falls nur die Klassenmitten $m_1, \ldots, m_p$ bekannt sind; anstelle der Klassenmittel kann man auch die Klassenmediane verwenden.

Beispiel: Im Beispiel aus Abschnitt 3 „Lebensdauern von 87 Kühlaggregaten", vgl. auch Tab. 6, haben wir zehn Klassen der Breite 1 gebildet. Im Abschnitt 4.2.1 ist bereits der Median der so gruppierten Daten berechnet worden:

$$\tilde{x}_{0,5} = 1{,}344.$$

Nehmen wir einmal an, daß wir nur noch die Klassenmitten $m_1 = 0{,}5$, $m_2 = 1{,}5, \ldots, m_{10} = 9{,}5$ kennen, so ergibt sich als mittlere absolute Abweichung vom Median der Wert

$$\frac{1}{n} \sum_{j=1}^{p} |m_j - \tilde{x}_{0,5}| \cdot H_j = \frac{1}{87} \sum_{j=1}^{10} |m_j - 1{,}344| \cdot H_j$$

$$= \frac{1}{87} (|0{,}5 - 1{,}344| \cdot 37 + |1{,}5 - 1{,}344| \cdot 19 + \ldots + |9{,}5 - 1{,}344| \cdot 1)$$

$$= 112{,}028/87 = 1{,}288.$$

5.4. Varianz, Standardabweichung und Variationskoeffizient

Haben wir im vorigen Abschnitt als Bezugsgröße den Median einer Häufigkeitsverteilung gewählt, so beziehen sich die hier vorgestellten Maßzahlen auf das arithmetische Mittel einer Beobachtungsreihe.

Ein weiterer grundlegender Unterschied zwischen den hier betrachteten Maßen für die Streuung und der mittleren absoluten Abweichung vom Median besteht darin, daß nun als Abstandsmaß keine betragliche Differenz, sondern vielmehr ein quadratischer Abstand gewählt wird.

5.4.1. Die Varianz

Bezeichnen wiederum $x_1, \ldots, x_n$ die Merkmalswerte in der Urliste und $a_1, \ldots, a_k$ die Ausprägungen des betrachteten Merkmals, so ist die **Varianz** der Beobachtungsreihe, die wohl das am häufigsten verwandte Streuungsmaß ist, gerade

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

$$= \frac{1}{n-1} \sum_{j=1}^{k} (a_j - \bar{x})^2 H_n(a_j) = \frac{1}{n-1} \left(\sum_{j=1}^{k} a_j^2 H_n(a_j) - n\bar{x}^2 \right).$$

Bei Verwendung dieses quadratischen Abstandsmaßes wählt man speziell das arithmetische Mittel $\bar{x}$ als Bezugsgröße, da es – ähnlich wie der Median bei der absoluten Differenz – die **Minimumeigenschaft des arithmetischen Mittels** besitzt:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \leqq \sum_{i=1}^{n} (x_i - c)^2 \quad \text{für jede beliebige Zahl c.}$$

Beispiel: Wir wollen die Varianz s^2 der Lebensdauer von $n = 30$ Glühbirnen, vgl. Tab. 11, berechnen. Mit

$$\bar{x} = \frac{1}{30} (375{,}3 + \ldots + 1407{,}2) = 24791{,}8/30 = 826{,}39$$

ergibt sich

$$s^2 = \frac{1}{29} \sum_{i=1}^{30} (x_i - 826{,}39)^2 = 52025{,}361$$

als Varianz der Beobachtungsreihe.

Liegen die Beobachtungen nur in gruppierter Form vor, so läßt sich die Varianz mit Hilfe der Klassenmittel oder der Klassenmitten berechnen.

Ist p die Anzahl der Klassen, die mit absoluten Häufigkeiten $H_1, \ldots, H_p$ auftreten, und ist

$$\bar{x}_j = \frac{1}{H_j} \sum_{x_i \in K_j} x_i \quad \text{für } j = 1, \ldots, p$$

das Klassenmittel, so berechnet sich die Varianz der gruppierten Daten zu

$$s_0^2 = \frac{1}{n-1} \sum_{j=1}^{p} (\bar{x}_j - \bar{x})^2 H_j.$$

Tab. 11a: Arbeitstabelle zur Berechnung der Varianz der gruppierten Lebensdauern von 30 Glühbirnen, vgl. auch Tab. 11

Klasse K_j	H_j	$\bar{x}_j$	m_j
$K_1 = [300; 700)$	6	497,967	500
$K_2 = [700; 800)$	7	768,200	750
$K_3 = [800; 900)$	6	825,133	850
$K_4 = [900; 1000)$	6	940,467	950
$K_5 = [1000; 1500)$	5	1166,600	1250

Beispiel: Im Beispiel der Glühbirnen bilden wir 5 Klassen wie in **Tab. 11a**, wo auch die absoluten Häufigkeiten H_j und die Klassenmittel (sowie die Klassenmitten) angegeben sind, und berechnen die Varianz s_0^2 der gruppierten Daten.

Es ergibt sich hier

$$s_0^2 = \frac{1}{29} \sum_{j=1}^{5} (\bar{x}_j - 826{,}39)^2 \cdot H_j =$$
$$= \frac{1}{29} ((497{,}967 - 826{,}39)^2 \cdot 6 + \ldots + (1166{,}600 - 826{,}39)^2 \cdot 5) =$$
$$= 1327677{,}6/29 = 45781{,}986$$

als Varianz der gruppierten Daten.

In diesem Beispiel sehen wir, daß die Varianz der Beobachtungsreihe durch die Varianz der gruppierten Daten unterschätzt wird. Das gilt auch allgemein, denn bei der Berechnung aus gruppierten Daten wird die Varianz innerhalb der Klassen vernachlässigt: Es ist

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

und weiter ist

$$\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{p} \sum_{x_i \in K_j} x_i^2 \quad \text{sowie}$$
$$-n\bar{x}^2 = n\bar{x}^2 - 2n\bar{x}^2 = \bar{x}^2 \sum_{j=1}^{p} H_j - 2\bar{x} \sum_{j=1}^{p} \bar{x}_j H_j,$$

so daß gilt

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{p} \left[\sum_{x_i \in K_j} x_i^2 + \bar{x}^2 H_j - 2\bar{x}\bar{x}_j H_j \right]$$
$$= \frac{1}{n-1} \sum_{j=1}^{p} \left[\sum_{x_i \in K_j} x_i^2 - \bar{x}_j^2 H_j + \bar{x}_j^2 H_j + \bar{x}^2 H_j - 2\bar{x}\bar{x}_j H_j \right]$$
$$= \frac{1}{n-1} \sum_{j=1}^{p} \left[\sum_{x_i \in K_j} (x_i - \bar{x}_j)^2 + (\bar{x}_j - \bar{x})^2 H_j \right]$$
$$= \frac{1}{n-1} \sum_{j=1}^{p} \sum_{x_i \in K_j} (x_i - \bar{x}_j)^2 + s_0^2$$
$$\geqq s_0^2$$

Verwendet man nun zur Berechnung der Varianz der gruppierten Daten die Klassenmitten, also

$$s_*^2 = \frac{1}{n-1} \sum_{j=1}^{p} (m_j - \bar{x})^2 H_j,$$

so wird man die Varianz der Beobachtungsreihe i. a. überschätzen, falls die Zahl der Klassen klein im Verhältnis zur Zahl der Beobachtungswerte ist. Das liegt daran, daß dann die Verteilung in den Klassen meist unsymmetrisch ist. vgl. **Abb. 17**. Je symmetrischer die Verteilung in den Klassen ist, umso kleiner ist auch s_*^2 im Verhältnis zu s_0^2. Liegt völlige Symmetrie vor, so ist $s_*^2 = s_0^2 \leqq s^2$.

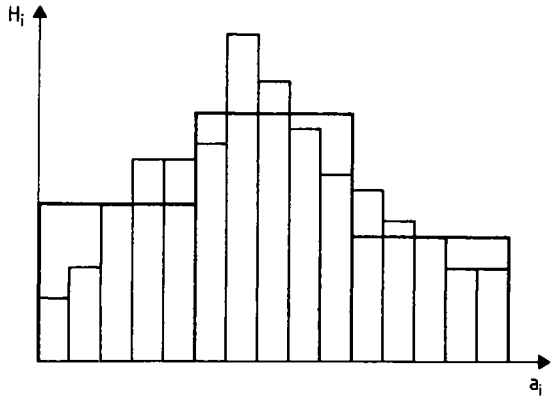

Abb. 17: Eine in den Klassen unsymmetrische Häufigkeitsverteilung

Beispiel: Bei Verwendung der Klassenmitten ergibt sich als Varianz der gruppierten Lebensdauern der Glühbirnen, vgl. Tab. 11a, der Wert

$$s_*^2 = \frac{1}{29} \cdot \sum_{j=1}^{5} (m_j - 826{,}39)^2 \cdot H_j$$
$$= \frac{1}{29} \cdot ((500 - 826{,}39)^2 \cdot 6 + \ldots + (1250 - 826{,}39)^2 \cdot 5)$$
$$= 1672279/29 = 57664{,}792.$$

Wir sehen also, daß die Varianz s^2 hier tatsächlich überschätzt wird, da die Verteilungen der Beobachtungswerte um die Klassenmitten unsymmetrisch sind.

Haben alle Klassen die gleiche Breite $b = b_1 = \ldots = b_p$, so läßt sich der Wert s_*^2 korrigieren. Hierzu verwendet man die **Sheppardsche Korrektur** $b^2/12$:

$$s_{**}^2 = s_*^2 - \frac{b^2}{12}$$

und hofft, daß man so eine gute Näherung für die Varianz s^2 der ungruppierten Beobachtungswerte gefunden hat.

5.4.2. Die Standardabweichung

Die **Standardabweichung**

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

ist gerade die Wurzel aus der Varianz s^2 einer Beobachtungsreihe. Sie hat gegenüber der Varianz den Vorteil, daß sie die gleiche Dimension wie die Beobachtungswerte und damit wie $\bar{x}$ besitzt.

Wird etwa $\bar{x}$ in Stunden gemessen, so hat auch s die Dimension „Stunden". Dagegen ist die Maßeinheit der Varianz (Stunde)^2.

Beispiel: Die Standardabweichung in der Beobachtungsreihe der Lebensdauern von 30 Glühbirnen, vgl. Tab. 11, ergibt sich wegen

$$s^2 = 52025{,}361$$

gerade zu

$$s = \sqrt{s^2} = \sqrt{52025{,}361} = 228{,}091.$$

5.4.3. Der Variationskoeffizient

Varianz und Standardabweichung benutzen zwar als Bezugspunkt das arithmetische Mittel $\bar{x}$, d.h. sie messen einen mittleren Abstand von $\bar{x}$, werden jedoch nicht ins Verhältnis zu $\bar{x}$ gesetzt. Ist aber der Mittelwert einer Beobachtungsreihe z.B. 10000, so wird man sagen, daß eine Varianz von 10 recht klein ist, dagegen ist diese Varianz ziemlich groß zu nennen, wenn $\bar{x} = 1$ ist. Ein von $\bar{x}$ bereinigtes Streuungsmaß, das also ohne direkte Nennung von $\bar{x}$ interpretiert werden kann, ist nun der **Variationskoeffizient**

$$v = \frac{s}{\bar{x}}.$$

Der Variationskoeffizient mißt also das Verhältnis von Standardabweichung und arithmetischem Mittel. Er ist nur für solche Beobachtungsreihen ein sinnvolles Maß, die ausschließlich positive Merkmalswerte beinhalten und eignet sich natürlich insbesondere zum Vergleich der Streuungen verschiedener Meßreihen.

Beispiel: Der Variationskoeffizient der Lebensdauer von 30 Glühbirnen, vgl. Tab. 11, ergibt sich zu

$$v = \frac{s}{\bar{x}} = \frac{228{,}091}{826{,}39} = 0{,}276.$$

5.5. Die Schiefe und der Exzeß

Obwohl Schiefe und Exzeß keine direkten Streuungsparameter von Häufigkeitsverteilungen sind, wollen wir sie in diesem Zusammenhang erläutern. Beide Maßzahlen können nur dort sinnvoll verwandt werden, wo eingipflige Häufigkeitsverteilungen vorliegen.

5.5.1. Die Schiefe einer Häufigkeitsverteilung

Im Abschnitt 4.5 haben wir schon die Begriffe rechtsschiefe, linksschiefe und symmetrische eingipflige Häufigkeitsverteilung kennengelernt. Die Schiefe ist nun eine Maßzahl, die uns nicht nur sagt, in welcher Richtung eine Häufigkeitsverteilung schief ist, sondern vielmehr auch die Größenordnung dieser Schiefe angibt.

Sind $x_1, \ldots, x_n$ die Merkmalswerte in einer Beobachtungsreihe mit Mittelwert $\bar{x}$, so ist die Schiefe der zugehörigen Häufigkeitsverteilung gerade

$$g_1 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^3}}.$$

Ist $g_1 = 0$, so heißt dies gerade, daß die Häufigkeitsverteilung symmetrisch ist. Je stärker negativ g_1 ist, desto linksschiefer ist die Verteilung, und sie ist umso rechtsschiefer, je stärker positiv der Wert g_1 ist.

Natürlich läßt sich die Schiefe einer Häufigkeitsverteilung auch aus gruppiertem Datenmaterial berechnen:

Bei Verwendung der Klassenmittel $\bar{x}_1, \ldots, \bar{x}_p$ ist

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^3 \cdot H_i}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2 \cdot H_i\right]^3}}$$

und bei Verwendung der Klassenmitten $m_1, \ldots, m_p$ ergibt sich

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{p}(m_i - \bar{x})^3 \cdot H_i}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{p}(m_i - \bar{x})^2 \cdot H_i\right]^3}}.$$

Beispiel: Die Schiefe der Häufigkeitsverteilung der Lebensdauer von 30 Glühbirnen ergibt sich aus gruppierten Daten, vgl. Tab. 11a, zu

$$g_1 = \left(\frac{1}{30}\sum_{i=1}^{5}(\bar{x}_i - 826{,}39)^3 H_i\right) \Big/ \sqrt{\left(\frac{1}{30}\sum_{i=1}^{5}(\bar{x}_i - 826{,}39)^3 H_i\right)^3}$$

$$= -\frac{8133122{,}1}{30} \Big/ \sqrt{\left(\frac{1327677{,}6}{30}\right)^3} = -\frac{271104{,}07}{9310158{,}2} = -0{,}029,$$

wenn man die Klassenmittel, und zu

$$g_1 = \left(\frac{1}{30}\sum_{i=1}^{5}(m_i - 826{,}39)^3 H_i\right) \Big/ \sqrt{\left(\frac{1}{30}\sum_{i=1}^{5}(m_i - 826{,}39)^2 H_i\right)^3}$$

$$= \frac{179742320}{30} \Big/ \sqrt{\left(\frac{1672279}{30}\right)^3} = \frac{5991410{,}7}{13160768} = 0{,}455,$$

wenn man die Klassenmitten verwendet.

Bei Verwendung der Klassenmittel ergibt sich eine schwach linksschiefe und bei Verwendung der Klassenmitten eine rechtsschiefe Häufigkeitsverteilung.

Hier müßte man also wirklich von den ungruppierten Daten ausgehen; für diese ergibt sich

$$g_1 = \frac{52076124}{30} \Big/ \sqrt{\left(\frac{1508735{,}5}{30}\right)^3} = \frac{1735870{,}8}{11278147} = 0{,}154,$$

also eine leichte Rechtsschiefe.

Wir sehen, daß in diesem Beispiel der Wert g_1 der ungruppierten Daten durch Klassenbildung einmal unter- und einmal überschätzt wurde.

5.5.2. Der Exzeß einer Häufigkeitsverteilung

Eine andere traditionelle Maßzahl für eingipflige Häufigkeitsverteilungen ist der **Exzeß (Kurtosis, Wölbung)**

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3,$$

der angibt, ob (bei gleicher Varianz) das absolute Maximum der Häufigkeitsverteilung größer als bei der Dichte der Normalverteilung, vgl. Abschnitt 1 in Kapitel IV, ist. Der theoretische Exzeß einer Normalverteilung ist gerade 0. Ist $g_2 > 0$, so ist das absolute Maximum der Häufigkeitsverteilung größer als das der zugehörigen Normalverteilung; ist $g_2 < 0$, so ist es kleiner.

Bei gruppierten Daten ergibt sich

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^4 \cdot H_i}{\left[\frac{1}{n}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2 \cdot H_i\right]^2} - 3$$

bei Verwendung der Klassenmittel und

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{p}(m_i - \bar{x})^4 \cdot H_i}{\left[\frac{1}{n}\sum_{i=1}^{p}(m_i - \bar{x})^2 \cdot H_i\right]^2} - 3$$

bei Verwendung der Klassenmitten.

Beispiel: Für die Häufigkeitsverteilung der ungruppierten Lebensdauern von $n = 30$ Glühbirnen, vgl. Tab. 11, ergibt sich ein Exzeß von

$$g_2 = \frac{\frac{1}{30}\sum_{i=1}^{30}(x_i - 826{,}39)^4}{\left(\frac{1}{30}\sum_{i=1}^{30}(x_i - 826{,}39)^2\right)^2} - 3 = \frac{2{,}6592080 \cdot 10^{11}/30}{(1508735{,}5/30)^2} - 3$$
$$= 3{,}504 - 3 = 0{,}504.$$

Das absolute Maximum dieser Häufigkeitsverteilung ist also größer als das der zugehörigen Normalverteilung.

Sind Schiefe und Exzeß einer Häufigkeitsverteilung wesentlich von 0 verschieden, so ist das ein Hinweis darauf, daß die zugrundeliegende theoretische Verteilung der Gesamtheit, aus der die Untersuchungseinheiten stammen, vgl. Kap. II, wesentlich von einer Normalverteilung abweicht.

6. Konzentrationsmaße für Häufigkeitsverteilungen

Beobachtet man an n Untersuchungseinheiten ein metrisch skaliertes Merkmal X mit nur positiven Ausprägungen, so ist es oftmals von Interesse zu untersuchen, wie sich die Summe der Merkmalswerte $x_1, \ldots, x_n$ in der Beobachtungsreihe auf die Untersuchungseinheiten aufteilt. Ist diese Summe vielleicht auf wenige Einheiten konzentriert? Mit solchen und ähnlichen Fragen wollen wir uns in diesem Abschnitt beschäftigen.

Beispiel: In einer Region hat man an n Untersuchungseinheiten, nämlich den landwirtschaftlichen Betrieben, die Größe der landwirtschaftlichen Nutzflächen $x_1, \ldots, x_n$ beobachtet. Wie teilt sich die gesamte Nutzfläche der Region auf die einzelnen Betriebe auf?

Ist sie auf einige Betriebe konzentriert, d. h. besitzen wenige Betriebe einen Großteil der Nutzfläche?

Wir wollen hier davon ausgehen, daß die n beobachteten Merkmalswerte durch q verschiedene Merkmalsausprägungen $a_1 < \ldots < a_q$, die mit absoluten Häufigkeiten $H_1, \ldots, H_q$ und relativen Häufigkeiten $h_1, \ldots, h_q$ auftreten, bestimmt werden; $a_1, \ldots, a_q$ können hier auch die Klassenmitten oder Klassenmittel von gruppierten Merkmalswerten sein. Dann ist natürlich der Gesamtbetrag der Merkmalswerte in der Beobachtungsreihe gerade

$$\sum_{j=1}^{n} x_j = \sum_{i=1}^{q} a_i H_i .$$

6.1. Die Lorenzkurve

Eine graphische Darstellung der Konzentration der Merkmalswerte liefert die sogenannte **Lorenzkurve**, die man, wie im folgenden beschrieben, erhält. Auf der Abszisse eines Koordinatenkreuzes werden die Größen

$$k_i = \sum_{j=1}^{i} \frac{H_j}{n} = \sum_{j=1}^{i} h_j \quad \text{für } i = 1, \ldots, q$$

abgetragen; es werden also die nach der Größe der Merkmalsausprägungen bzw. Gruppen geordneten relativen Häufigkeiten aufsummiert (im Beispiel: Es wird erst der Prozentsatz der Betriebe mit der kleinsten Fläche bestimmt, dann wird der Prozentsatz der Betriebe mit zweitkleinster Fläche bestimmt und zum Prozentsatz der Betriebe mit der kleinsten Fläche addiert, usw.). Auf der Ordinate wird

$$l_i = \frac{\sum_{j=1}^{i} a_j H_j}{\sum_{j=1}^{q} a_j H_j} \quad \text{für } i = 1, \ldots, q$$

abgetragen; die Ausprägungen werden also der Größe nach aufsummiert und auf die Summe aller Ausprägungen bezogen (im Beispiel: Die Flächenanteile der Betriebe bzgl. der Gesamtfläche werden der Flächengröße nach aufsummiert).

Ein Punkt $(k_i; l_i)$ der Lorenzkurve gibt dann an, daß auf $k_i \cdot 100\%$ der Untersuchungseinheiten $l_i \cdot 100\%$ des Gesamtbetrages aller Merkmalsausprägungen entfal-

len (im Beispiel: Auf $k_i \cdot 100\%$ der Betriebe entfallen $l_i \cdot 100\%$ der gesamten Nutzfläche).

Durch die Verbindung der Punkte (k_i, l_i) mit $(k_{i+1}; l_{i+1})$ für $i = 0, \ldots, k-1$ entsteht die Lorenzkurve. Dabei wird $k_0 = l_0 = 0$ gesetzt, und es gilt stets $k_q = l_q = 1$.

Beispiel: In **Tab. 12** ist die Aufteilung der landwirtschaftlich genutzten Fläche einer Region in Hektar (ha) abzulesen.

Tab. 12: Landwirtschaftlich genutzte Fläche einer Region

Flächengr. von ... bis ... ha	Anzahl der landw. Betriebe	Flächengr. pro Gruppe in ha	Prozentsatz der landw. Betriebe	Flächenanteile	aufsumm. Betriebsanteile	aufsumm. Flächenanteile
	H_i	$a_i \cdot H_i$	h_i	$\frac{a_i \cdot H_i}{\sum_i a_i \cdot H_i}$	k_i	l_i
bis 5 ha	21	63	0,42	0,063	0,42	0,063
5 bis 10 ha	9	72	0,18	0,072	0,60	0,135
10 bis 20 ha	9	135	0,18	0,135	0,78	0,270
20 bis 50 ha	8	280	0,16	0,280	0,94	0,550
> 50 ha	3	450	0,06	0,450	1,00	1,000
Gesamt	n = 50	1000	1,00	1,000		

Die zugehörige Lorenzkurve ist in **Abb. 18** zu sehen.

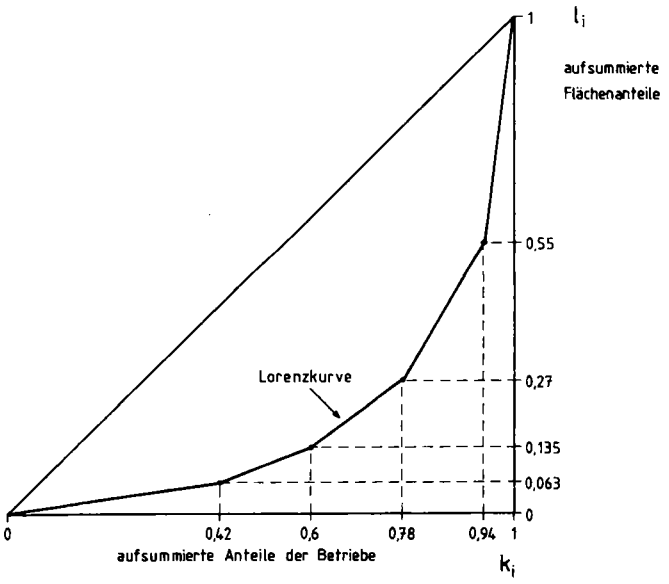

Abb. 18: Lorenzkurve im Beispiel

Der Punkt $(k_1; l_1) = (0{,}42; 0{,}063)$ auf der Lorenzkurve besagt z. B., daß auf 42% der landwirtschaftlichen Betriebe nur 6,3% der Fläche entfällt.

Es können zwei Extremfälle der Lorenzkurve auftreten:

- Es liegt keine Konzentration vor, d. h. alle Untersuchungseinheiten haben den gleichen Anteil am Gesamtbetrag (im Beispiel: Alle Betriebe besitzen gleich viel Fläche). In diesem Fall entfallen auf $z \cdot 100\%$ der Untersuchungseinheiten immer $z \cdot 100\%$ des Gesamtbetrages ($0 \leqq z \leqq 1$). Die Lorenzkurve ist also die Diagonale.
- Der Gesamtbetrag konzentriert sich fast vollständig auf eine Untersuchungseinheit, während der Anteil der anderen Untersuchungseinheiten verschwindend gering ist (im Beispiel: Ein landwirtschaftlicher Betrieb besitzt fast die ganze Fläche). Die Lorenzkurve nähert sich dann bei immer weiter fortschreitender Konzentration auf eine Untersuchungseinheit immer mehr der Abzisse und steigt immer näher am rechten Eckpunkt (1; 0) fast senkrecht zum Eckpunkt (1; 1) auf.

6.2. Das Lorenzsche Konzentrationsmaß; der Gini-Koeffizient

Mit Hilfe der Lorenzkurve möchte man eine Maßzahl für die Konzentration finden, die bei Nichtkonzentration den Wert 0 und bei vollständiger Konzentration den Wert 1 annimmt. Die Fläche zwischen Diagonale und Lorenzkurve wird mit zunehmender Konzentration immer größer.

Das **Lorenzsche Konzentrationsmaß (LKM)**, das man auch **Gini-Koeffizient** nennt, wird nun als das 2-fache der Fläche F zwischen Diagonale und Lorenzkurve definiert:

$$\text{LKM} = 2F$$

Die Fläche F kann durch die Summe der Trapezflächen F_i abzüglich der Fläche des oberen Dreiecks berechnet werden, vgl. auch **Abb. 19**:

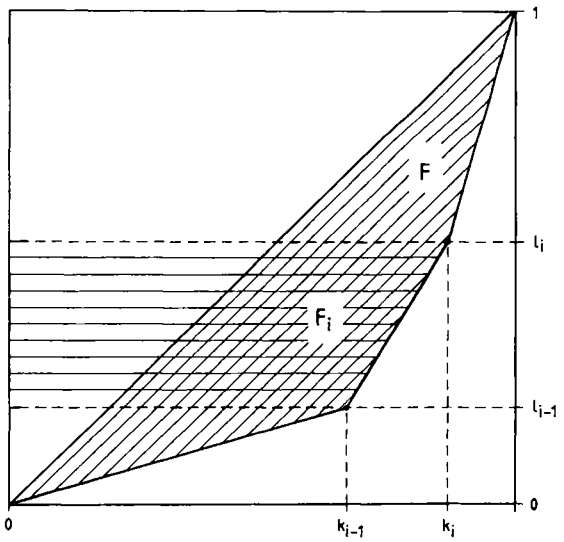

Abb. 19: Berechnung des Lorenzschen Konzentrationsmaßes

$$F = \sum_{i=1}^{q} F_i - 0{,}5 \quad \text{mit}$$

$$F_i = \frac{k_{i-1} + k_i}{2}(l_i - l_{i-1}) = \frac{k_{i-1} + k_i}{2} \frac{1}{\sum_{j=1}^{q} a_j \cdot H_j} \left(\sum_{j=1}^{i} a_j \cdot H_j - \sum_{j=1}^{i-1} a_j \cdot H_j \right)$$

$$= \frac{k_{i-1} + k_i}{2} \frac{a_i \cdot H_i}{\sum_{j=1}^{q} a_j \cdot H_j} \quad \text{für } i = 1, \ldots, q.$$

Damit berechnet sich das Lorenzsche Konzentrationsmaß zu

$$LKM = \sum_{i=1}^{q} (k_{i-1} + k_i) \cdot \frac{a_i H_i}{\sum_{j=1}^{q} a_j H_j} - 1,$$

wobei gilt:

$$k_i = \sum_{j=1}^{i} \frac{H_j}{n} \quad \text{für } i = 1, \ldots, q \quad \text{und} \quad k_0 = 0.$$

Für das Lorenzsche Konzentrationsmaß gilt stets

$$0 \leqq LKM \leqq \frac{n-1}{n};$$

es ist

$$LKM = 0,$$

falls alle Merkmalswerte in der Beobachtungsreihe gleich sind, und

$$LKM = \frac{n-1}{n},$$

falls eine einzige Untersuchungseinheit die gesamte Merkmalssumme trägt.

Damit man ein Konzentrationsmaß hat, das auch wirklich den Wert 1 annehmen kann (bei vollständiger Konzentration) normiert man den Gini-Koeffizienten oft wie folgt:

$$LKM_{nor} = \frac{n}{n-1} LKM.$$

Beispiel: Für die Aufteilung der landwirtschaftlichen Nutzflächen einer Region, vgl. Tab. 12, wollen wir das LKM berechnen. Aus **Tab. 13** ergibt sich

$$LKM = \sum_{i=1}^{5} 2F_i - 1 = 1{,}6408 - 1 = 0{,}6408 \quad \text{bzw.}$$

$$LKM_{nor} = \frac{n}{n-1} LKM = \frac{50}{49} \cdot 0{,}6408 = 0{,}6539.$$

Das Lorenzsche Konzentrationsmaß kann trotz unterschiedlicher Konzentrationsstrukturen gleich sein.

Tab. 13: Landwirtschaftlich genutzte Flächen, Arbeitstabelle zum Beispiel

Flächengröße von ... bis ... ha	Flächenanteile $\frac{a_i H_i}{\sum_{j=1}^{q} a_j H_j}$	aufsummierte Betriebsanteile k_i	$2 F_i$
bis 5 ha	0,063	0,42	0,02646
5 bis 10 ha	0,072	0,60	0,07344
10 bis 20 ha	0,135	0,78	0,18630
20 bis 50 ha	0,280	0,94	0,48160
> 50 ha	0,450	1,00	0,87300
Gesamt	1,000		1,64080

Beispiel: In **Abb. 20** ergibt sich für **Südafrika:** LKM = 0,63 und für die **Bundesrepublik Deutschland**: LKM = 0,64. Das LKM ist also in beiden Ländern fast gleich. Die Umsätze verteilen sich aber in Südafrika anders auf die Betriebe als in der Bundesrepublik Deutschland. Denn, während in Südafrika

weniger als 5% des Umsatzes auf fast 50% der Betriebe,
ungefähr 15% des Umsatzes auf ca. 20% der Betriebe,
ungefähr 25% des Umsatzes auf ca. 20% der Betriebe und
ungefähr 55% des Umsatzes auf ca. 10% der Betriebe entfallen,

konzentrieren sich in der Bundesrepublik Deutschland

ungefähr 10% des Umsatzes auf ca. 50% der Betriebe,
ungefähr 25% des Umsatzes auf ca. 40% der Betriebe und
ungefähr 65% des Umsatzes auf ca. 10% der Betriebe.

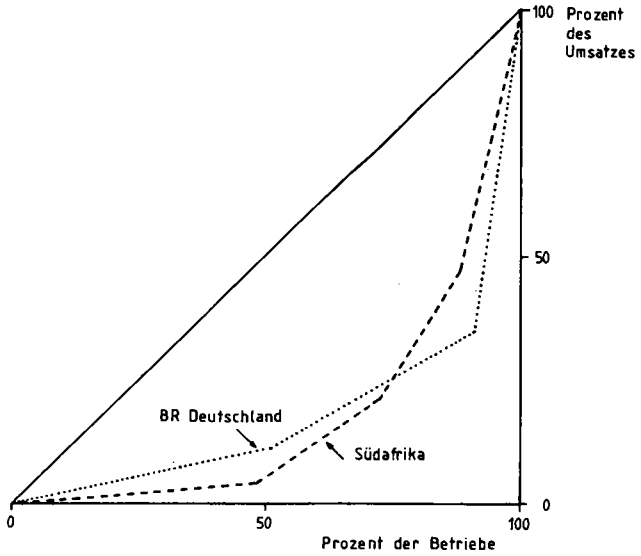

Abb. 20: Lorenzkurve der Umsatzkonzentration bei Ackerschleppern, Südafrika und Bundesrepublik Deutschland 1977

In Südafrika entfallen also sowohl auf Betriebe mit sehr geringem Anteil am Umsatz als auch auf Betriebe mit einem sehr hohen Anteil am Umsatz weniger Prozent des Umsatzes als in der Bundesrepublik Deutschland. Dabei ist der Anteil der Betriebe mit sehr geringem bzw. sehr hohem Umsatzanteil in beiden Ländern ungefähr gleich.

7. Verhältniszahlen

In der deskriptiven Statistik betrachtet man auch Zahlen, die einen bestimmten Sachverhalt quantitativ kennzeichnen. So beschreiben z. B. die Lagemaße ein „Zentrum" einer Häufigkeitsverteilung, die Streuungsmaße eine „Streuung" der Merkmalsausprägungen und die Merkmalsausprägungen selbst eine Eigenschaft der Untersuchungseinheit. Diese werden als **statistische Maßzahlen** bezeichnet.

Bildet man den Quotienten zweier Maßzahlen, so erhält man eine wichtige Klasse von Maßzahlen: die sogenannten **Verhältniszahlen**.

Diese lassen sich, wie in der **Abb. 21** gezeigt, untergliedern.

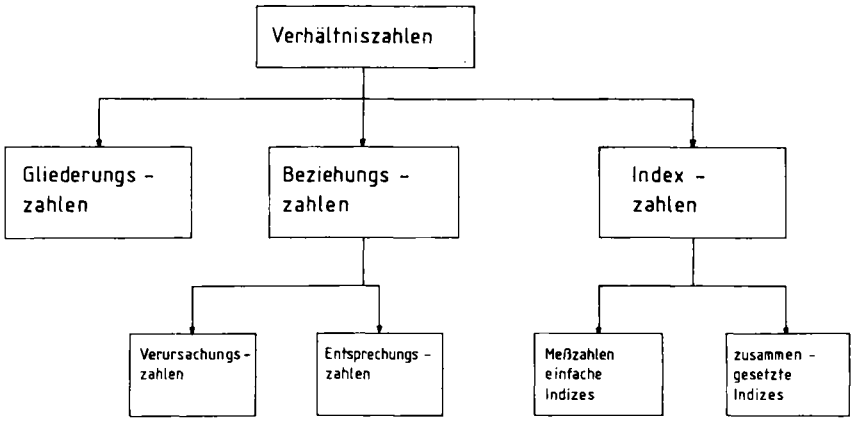

Abb. 21: Gliederung von Verhältniszahlen

7.1. Gliederungszahlen

Bezieht man eine Teilgröße auf eine ihr übergeordnete Gesamtgröße, so erhält man eine Gliederungszahl.

Besteht **z. B.** in einem Betrieb eine Tagesproduktion aus 1500 Maschinenteilen, wobei 300 Teile eine Qualitätskontrolle nicht überstehen, dann sind 20% der Tagesproduktion Ausschuß; dieser Ausschußanteil ist eine Gliederungszahl.

Weitere Beispiele findet man im Abschnitt 2: die relativen Häufigkeiten. Dort werden absolute Häufigkeiten eines Beobachtungswertes dividiert durch die Summe aller absoluten Häufigkeiten.

Man kann Gliederungszahlen in Anteilen, Prozenten oder in Promillen angeben.

7.2. Beziehungszahlen

Den Quotienten zweier sachlich sinnvoll in Verbindung stehender Maßzahlen, z. B. die Anzahl der Lebendgeborenen in einem Jahr bezogen auf die durchschnittliche Anzahl der Frauen in der Bundesrepublik Deutschland, bezeichnet man als **Beziehungszahl.**

Man unterscheidet zwischen **Verursachungs-** und **Entsprechungszahlen.**

7.2.1. Verursachungszahlen

Bei Verursachungszahlen bezieht man **Bewegungsmassen** auf die zugehörigen **Bestandsmassen.**

Dabei versteht man unter Bestandsmassen solche Massen, deren Elemente eine bestimmte Dauer haben; es wird zu einem festen Zeitpunkt der Bestand ermittelt. Dagegen sind die Elemente einer Bewegungsmasse Ereignisse, also einem festen Zeitpunkt zugeordnet. Eine zeitliche Abgrenzung erfolgt durch Festlegung eines Zeitraumes. Betrachten wir z. B. den Lagerbestand einer Brauerei am 15.4.1979 um 8^{00} Uhr von 500 hl Bier: Werden an diesem Tag 200 hl Bier abgesetzt, so sind dies 40% des Bestandes; bei einer Tagesproduktion von 150 hl wird das Lager um 30% des alten Bestandes aufgefüllt.

Beispiel: Betrachten wir als weiteres Beispiel für eine Bestandsmasse die Wohnbevölkerung einer Stadt zu einem bestimmten Zeitpunkt. Die zugehörigen Bewegungsmassen sind Geburten, Todesfälle, Zu- und Fortzüge. Die sogenannte **„Geburtenziffer"**

$$G = \frac{\text{Zahl der Lebendgeborenen}}{\text{Einwohner}} \cdot 1000$$

sagt, wie viele Geburten auf 1000 Einwohner einer Stadt entfallen, vgl. **Tab. 14.**

Tab. 14: Geburten in zwei Städten

	Wohnbevölk.	Geburten	Geburten-ziffer G	Frauen zwischen 15–45 Jahren	Fruchtbarkeits-rate F
Stadt A	178503	1794	10,05	34478	52,03
Stadt B	288495	3054	10,59	62728	48,69

Es ist zu beachten, daß der Aussagegehalt von Beziehungszahlen von der Wahl der Bezugsgröße abhängt.

Beispiel: Die Geburtenziffer für die Stadt B ist im obigen Beispiel größer als für die Stadt A, vgl. Tab. 14. Wenn man jedoch die Bestandsmasse von **sterilen Teilmassen** befreit und die **„Fruchtbarkeitsrate"**

$$F = \frac{\text{Zahl der Lebendgeborenen}}{\text{Frauen im gebärfähigen Alter (15–45 Jahre)}} \cdot 1000$$

für jede Stadt berechnet, so kehrt sich die Relation gerade um.

7.2.2. Entsprechungszahlen

Alle Beziehungszahlen, bei denen man Ereignisse nicht auf ihren Bestand beziehen kann, werden als **Entsprechungszahlen** bezeichnet.

Beispiel: Betrachten wir als Beispiel das Schüler-Lehrer-Verhältnis an Grund- und Hauptschulen in Nordrhein-Westfalen 1978:

$$\frac{\text{Zahl der Schüler}}{\text{Zahl der Lehrer}} = \frac{1581506}{66779} \simeq 23{,}7.$$

Es entfielen also auf einen Lehrer ungefähr 24 Schüler. Man beachte, daß diese Zahl i.a. nicht der durchschnittlichen Klassengröße entspricht.

Einige der Beziehungszahlen haben einen natürlichen Einzug in unseren Sprachgebrauch gefunden. Begriffe wie Hektarertrag oder Bevölkerungsdichte sind den meisten bekannt.

Doch bei der Interpretation von Beziehungszahlen ist, wie man bei der Konstruktion einer Fruchtbarkeitsrate in Abschnitt 7.2.1 gesehen hat, Vorsicht geboten.

7.3. Indexzahlen

Werden nicht mehr nur auch sachlich sinnvoll in Verbindung stehende Maßzahlen, sondern zwei Maßzahlen der gleichen Art in Beziehung gesetzt, so spricht man von **Indexzahlen.**

Man bezeichnet diese als **Meßzahlen (einfache Indizes)**, wenn die zugehörigen Maßzahlen einen realen Sachverhalt beschreiben.

Ist eine der Maßzahlen eine Zahl, die einen fiktiven Zustand beschreibt (z.B. das Ausgabeverhalten einer Person, wenn sie das Dreifache ihres Gehaltes verdienen würde), so erhält man eine **zusammengesetzte Indexzahl** (kurz **Indexzahl**).

7.3.1. Meßzahlen

In vielen Fällen hat man es nicht nur mit zwei Maßzahlen, sondern mit einer Reihe von Maßzahlen zu tun, welche man in Beziehung setzen möchte. Wir betrachten den Fall einer zeitlichen Reihe genauer:

Seien x_t Maßzahlen zu den Zeitpunkten (**Perioden**) t, wobei x_0 eine Maßzahl zum **Basiszeitpunkt** (Periode) 0 ist, dann heißt:

Tab. 15: Umsatzzahlen der beiden Firmen

	Maier & Co			Müller KG		
t Quartal ab 1.1.1976	Umsatz x_t in 1000 DM	Meßzahlen x_t/x_0	x_t/x_4	Umsatz y_t in 1000 DM	Meßzahlen y_t/y_0	y_t/y_4
0	1240	1,00	1,08	960	1,00	1,07
1	1210	0,98	1,05	980	1,02	1,09
2	1180	0,95	1,03	980	1,02	1,09
3	1190	0,96	1,03	920	0,96	1,03
4	1150	0,93	1,00	900	0,94	1,00
5	1080	0,87	0,94	860	0,90	0,96
6	1120	0,90	0,97	880	0,92	0,98
7	1130	0,91	0,98	920	0,96	1,02

$$I_{0t} = \frac{x_t}{x_0} \quad \text{für } t = 0, 1, 2, \ldots$$

eine **Zeitreihe einfacher Indizes** (Meßzahlen).

Beispiel: Untersuchen wie z.B. die vierteljährlichen Umsatzzahlen der Firma Maier & Co und der Firma Müller KG über einen Zeitraum von 2 Jahren, vgl. **Tab. 15.**

Man erhält Meßzahlen I_{0t} zur Basisperiode 0 als: $I_{0t} = \frac{x_t}{x_0}$. Jede beliebige Periode $k = 0, 1, \ldots, 7$ ist als Basisperiode wählbar. In diesem Beispiel wurde $I_{4t} = \frac{x_t}{x_4}$ berechnet.

In der graphischen Darstellung der Umsätze, vgl. **Abb. 22**, bedeutet ein paralleler Verlauf der Kurven, daß die Umsätze in diesem Quartal um den gleichen absoluten Betrag gestiegen oder gesunken sind. Dagegen bedeuten parallel verlaufende Strekken bei der graphischen Darstellung der Indizes, vgl. **Abb. 23, 24,** daß im Verhältnis zum Basisquartal die Umsätze in beiden Firmen sich gleich verhalten.

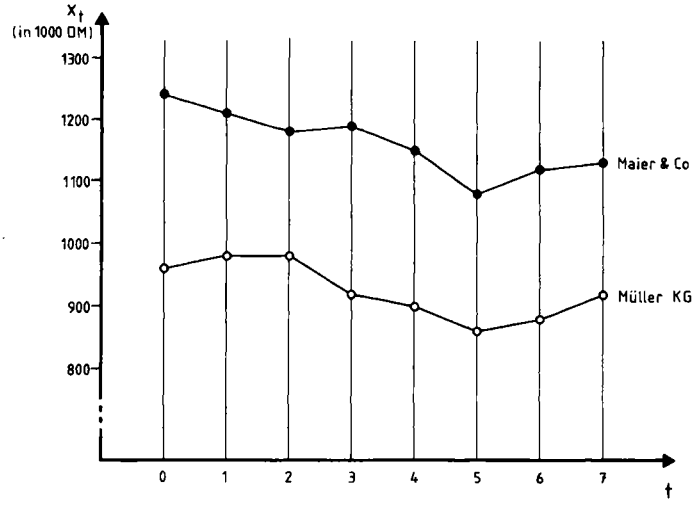

Abb. 22: Entwicklung des Umsatzes der beiden Firmen

Es ist üblich, die Meßzahlen mit 100 multipliziert zu betrachten. So besagt z.B. für Maier & Co

$$I_{05} \cdot 100 = \frac{x_5}{x_0} \cdot 100 = 87,$$

daß 87% des Umsatzes im **Basisquartal** im Quartal 5 umgesetzt werden, oder anders ausgedrückt, daß eine Minderung des Umsatzes um 13% vorliegt.

Vergleicht man z.B. die Meßzahlen

$$I_{05} \cdot 100 = 87 \quad \text{und} \quad I_{06} \cdot 100 = 90,$$

so sagt man, daß der Umsatz um 3 **Prozentpunkte** gestiegen ist.

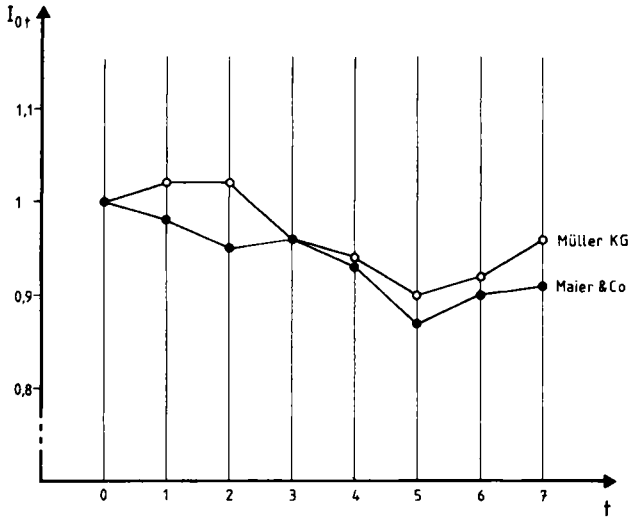

Abb. 23: Umsatzmeßzahlen mit Basisquartal 0

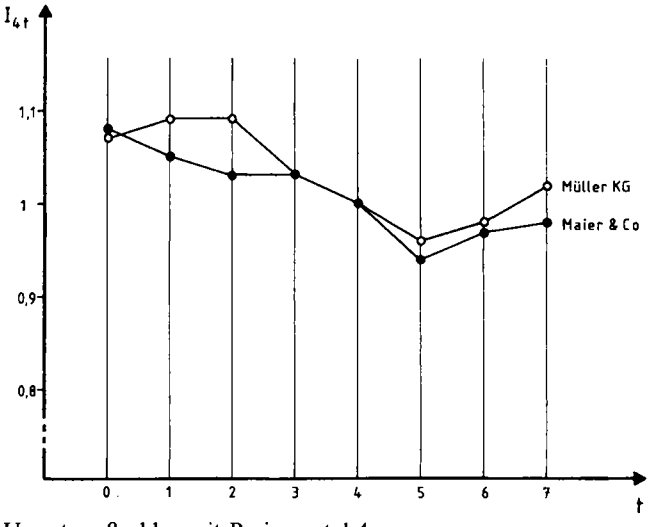

Abb. 24: Umsatzmeßzahlen mit Basisquartal 4

Im allgemeinen empfiehlt es sich nicht, lange zurückliegende Zeitperioden als Basisperiode zu wählen, und ebensowenig, Perioden mit außergewöhnlichen Zahlen zu wählen. Deshalb wird von Zeit zu Zeit eine **Umbasierung** der Zeitreihe notwendig. Dazu braucht man nicht die Ursprungsdaten zu kennen, da man die Indizes I_{kt} zur Basisperiode k wie folgt aus den Indizes zur Basisperiode 0 erhält:

$$I_{kt} = \frac{x_t}{x_k} = \frac{x_t/x_0}{x_k/x_0} = \frac{I_{0t}}{I_{0k}}.$$

Beispiel: In unserem Beispiel gilt für die Müller KG:

$$0{,}98 \simeq I_{46} = \frac{y_6}{y_4} = \frac{880}{900} \quad \text{und}$$

$$\frac{I_{06}}{I_{04}} = \frac{880/960}{900/960} \simeq \frac{0{,}92}{0{,}90} \simeq 0{,}98.$$

Direkt aus der Umbasierung läßt sich ihre inverse Operation, die **Verkettung** ableiten:

$$I_{0t} = I_{0k} \cdot I_{kt}.$$

Diese wird dann verwendet, wenn man für eine weitere Maßzahl x_t die Zeitreihe der Meßzahlen zur Basisperiode 0 ergänzen möchte, aber den Wert x_0 nicht mehr kennt; dabei muß man aber mit der Meßzahl I_{0k} auch die Maßzahl x_k kennen, um I_{kt} zu berechnen.

Beispiel: Im Beispiel sei für die Müller KG für $t = 8$ der Umsatz gleich 910 (in 1000 DM) gewesen. Dann ist

$$I_{08} = \frac{910}{960} \simeq 0{,}95 \quad \text{oder} \quad I_{08} = I_{04} \cdot I_{48} = 0{,}94 \cdot \frac{910}{900} \simeq 0{,}95.$$

Analog wie Zeitreihen von Meßzahlen lassen sich auch Reihen von Meßzahlen des örtlichen und sachlichen Verlaufs behandeln.

So ist z. B. der durchschnittliche Kartoffelpreis in verschiedenen Großstädten eine örtliche Reihe oder die Anzahl der verschiedenen Haustiere in 1000 Haushalten eine sachliche Reihe.

7.3.2. Standardisierung von Meßzahlen, Sterbeziffern

Beim Vergleich von gleichartigen Meßzahlen kann es zu Problemen bei der Interpretation kommen, wenn unterschiedliche Strukturen nicht berücksichtigt werden. So hat man z. B. bei der Konstruktion einer Fruchtbarkeitsrate, vgl. Abschnitt 7.2.1, versucht, diese Unterschiede weitgehend zu eliminieren, indem man die Bestandsmasse von sterilen Teilmassen befreit. Hier soll eine andere Methode, die der **Standardisierung**, anhand eines **Beispiels** erläutert werden.

Ziel ist es, für ein bestimmtes Jahr die Sterblichkeit der Wohnbevölkerung zweier Städte A und B zu vergleichen. Die Bevölkerung wird für jede Stadt zuerst in verschiedene Altersklassen – von 0 bis unter 1 Jahr, von 1 bis unter 10 Jahren usw. – eingeteilt, nun werden die Klassenhäufigkeiten H_{Ai} bzw. H_{Bi} und die Anzahl der Gestorbenen G_{Ai} bzw. G_{Bi} in den einzelnen Klassen bestimmt, vgl. auch **Tab. 16**.

Ein Vergleich der beiden globalen (mittleren) **Sterbeziffern** q_A und q_B – ermittelt aus der Gesamtzahl der Gestorbenen und der Wohnbevölkerung einer Stadt –

$$q_A = \frac{\sum_{i=1}^{k} G_{Ai}}{H_A} \quad \text{mit } H_A = \sum_{i=1}^{k} H_{Ai},$$

wobei k die Zahl der Altersklassen ist (q_B analog), scheint zu ergeben, daß die Sterblichkeit in der Stadt B um 4,4% höher liegt, als in der Stadt A:

Tab. 16: Wohnbevölkerung in zwei Städten A und B, Gestorbene, Sterbeziffern und Arbeitstabelle zur Berechnung von standardisierten Meßzahlen

	Stadt A			Stadt B			Arbeitswerte			
Alter von ... bis unter ... Jahren	Wohnbevölk. H_{Ai}	Gestorbene G_{Ai}	Sterbeziffern $q_{Ai}(\cdot 1000)$	Wohnbevölk. H_{Bi}	Gestorbene G_{Bi}	Sterbeziffern $q_{Bi}(\cdot 1000)$	p_{Ai}	p_{Bi}	$p_{Ai}q_{Bi}$	$p_{Bi}q_{Ai}$
0– 1	4812	84	17,46	731	13	17,78	0,0097	0,0095	0,172	0,166
1–10	59371	27	0,45	9024	4	0,44	0,1198	0,1177	0,053	0,053
10–20	84210	31	0,37	12800	4	0,31	0,1700	0,1669	0,053	0,062
20–30	85915	50	0,58	11818	7	0,59	0,1734	0,1541	0,102	0,089
30–40	72134	78	1,08	10963	11	1,00	0,1456	0,1430	0,146	0,154
40–50	49183	136	2,77	8027	21	2,62	0,0993	0,1047	0,260	0,290
50–60	56291	343	6,09	9286	55	5,92	0,1136	0,1211	0,673	0,737
60–70	42311	675	15,96	7040	110	15,63	0,0854	0,0918	1,335	1,464
70–80	33625	1677	49,87	5704	278	48,74	0,0679	0,0744	3,309	3,710
80–90	6715	955	142,22	1133	154	135,92	0,0136	0,0148	1,849	2,105
90 und mehr	890	277	311,24	150	43	286,67	0,0018	0,0020	0,516	0,622
insgesamt	495457	4333		76676	700		1	1	8,47	9,45

$$I_{AB} = \frac{q_B}{q_A} = \frac{\sum_{i=1}^{k} G_{Bi}}{\sum_{i=1}^{k} G_{Ai}} \cdot \frac{H_A}{H_B} = \frac{700}{4333} \cdot \frac{495457}{76676} \simeq 1{,}044.$$

Betrachtet man die altersspezifischen Sterbeziffern q_{Ai} (bzw. q_{Bi}), definiert durch $q_{Ai} = \frac{G_{Ai}}{H_{Ai}}\left(\text{bzw. } q_{Bi} = \frac{G_{Bi}}{H_{Bi}}\right)$, so zeigt sich, daß die globalen Sterbeziffern nicht nur von den altersspezifischen Sterbeziffern abhängen, sondern auch vom Altersaufbau einer Stadt:

$$q_A = \frac{\sum_{i=1}^{k} G_{Ai}}{H_A} = \frac{\sum_{i=1}^{k} H_{Ai} q_{Ai}}{H_A} = \sum_{i=1}^{k} p_{Ai} q_{Ai} \quad \text{mit } p_{Ai} = \frac{H_{Ai}}{H_A}.$$

Um beim Vergleich der Sterblichkeit den Einfluß der Altersstruktur auszuschalten, muß man **standardisierte Sterbeziffern** verwenden, d.h. für beide Städte den gleichen **„Standardaltersaufbau"** voraussetzen. So ist

$$q_B^{st} = \sum_{i=1}^{k} p_{Ai} q_{Bi}$$

die standardisierte Sterbeziffer der Stadt B, berechnet mit dem Altersaufbau $p_{A1}, \ldots, p_{Ak}$ der Stadt A, und entsprechend ist

$$q_A^{st} = \sum_{i=1}^{k} p_{Bi} q_{Ai} \quad \text{mit} \quad p_{Bi} = \frac{H_{Bi}}{H_B}$$

die standardisierte Sterbeziffer für die Stadt A mit dem Altersaufbau $p_{B1}, \ldots, p_{Bk}$ der Stadt B.

Für den **Vergleich der Sterbeziffern** bieten sich die Meßzahlen

$$I_{AB} = \frac{q_B^{st}}{q_A} = \frac{\sum_{i=1}^{k} p_{Ai} q_{Bi}}{q_A} \simeq \frac{8{,}47}{8{,}75} = 0{,}968 \quad \text{bzw.}$$

$$I_{AB} = \frac{q_B}{q_A^{st}} = \frac{q_B}{\sum_{i=1}^{k} p_{Bi} q_{Ai}} \simeq \frac{9{,}13}{9{,}45} \simeq 0{,}966$$

an.

Hätte die Stadt B die gleiche Altersstruktur wie die Stadt A, dann hätte die Stadt B 3,2% weniger Gestorbene als die Stadt A. Oder im zweiten Fall wären es 3,4% weniger Gestorbene, falls in beiden Städten der Altersaufbau der Stadt B zugrundegelegt würde.

Bei den oben berechneten Indexzahlen handelt es sich bereits um **zusammengesetzte Indexzahlen,** da eine der beiden Maßzahlen einen nichtrealen Zustand beschreibt, in diesem Fall eine fiktive Sterbeziffer unter der Annahme einer anderen Altersstruktur.

Weitere Beispiele zur Standardisierung (Mengenindex, Preisindex) werden im nächsten Abschnitt folgen.

7.3.3. Zusammengesetzte Indexzahlen

Hat man Reihen gleichartiger Daten, so möchte man den Zusammenhang dieser Reihen oft durch eine Zahl beschreiben. Im Abschnitt 7.3.2 haben wir eine Methode kennengelernt, mit der man unterschiedliche Strukturen der Reihen ausschalten kann: die Standardisierung. Ein Vergleich einer standardisierten Meßzahl mit einer anderen Meßzahl führt uns zum Begriff der **zusammengesetzten Indexzahlen**, kurz **Indizes**. Am **Beispiel** eines **Warenkorbes** mit n verschiedenen Waren werden einige Indizes vorgestellt. Kennt man für die Waren bestimmte Mengen $q_{t1}, \ldots, q_{tn}$ und bestimmte Preise $p_{t1}, \ldots, p_{tn}$ zu einer bestimmten Periode t (für $t = 0$ **Basisperiode** oder für $t > 0$ **Berichtsperiode** genannt), so kann man z. B. für einen Warenkorb, bestehend aus den Getränken Milch, Limonade, Bier und Wein, den durchschnittlichen Verbrauch eines 4-Personenhaushaltes mit den durchschnittlichen Preisen pro Liter oder auch die Absätze einer Getränkefirma mit deren Getränkepreisen vergleichen. Der **Wert eines Warenkorbes** $q_{t1}, \ldots, q_{tn}$ in einer Periode t wird durch

$$\sum_{i=1}^{n} p_{ti} q_{ti}$$

bestimmt. Dies entspricht dann im Beispiel dem tatsächlichen Aufwand eines 4-Personenhaushaltes für diese Getränke.

A. Der Wertindex

Vergleicht man den Wert eines Warenkorbes zur Berichtsperiode t mit dem zur Basisperiode 0, so erhält man einen **Wertindex** (eine Meßzahl)

$$W_{0t} = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{0i} q_{0i}}.$$

Mit Hilfe des Wertindex W_{0t} lassen sich z. B. Aussagen darüber machen, wie sich der tatsächliche Aufwand eines 4-Personenhaushaltes für Getränke im Berichtsmonat t in Bezug zum Basismonat 0 verändert hat.

B. Preisindizes nach Laspeyres und nach Paasche

Möchte man jedoch Aussagen über Preisentwicklungen machen, so muß man für verschiedene Perioden das gleiche Mengenschema verwenden. Dies führt uns zu den **Preisindizes**

$$P^P_{0,t} = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{0i} q_{ti}} \qquad \textbf{nach Paasche} \quad \text{und}$$

$$P^L_{0,t} = \frac{\sum_{i=1}^{n} p_{ti} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \qquad \textbf{nach Laspeyres}.$$

Beim Preisindex nach Paasche vergleicht man den Wert einer **Gütermengenkombination (Warenkorb)** $q_{t1}, \ldots, q_{tn}$ zur jeweiligen Berichtsperiode t mit dem Wert, den dieser unter der Preissituation zur Basisperiode gehabt hätte. Dagegen wird beim Preisindex nach Laspeyres die Gütermengenkombination $q_{01}, \ldots, q_{0n}$ der Basisperiode 0 für alle Berichtsperioden zugrundegelegt und ihr fiktiver Wert zur Berichtsperiode t mit ihrem Wert zur Basisperiode verglichen.

C. Ein Beispiel

Betrachten wir als Beispiel einen Preisindex für den durchschnittlichen Verbrauch an Fleisch aller privaten Haushalte in einer Gemeinde. Die durchschnittlich verbrauchten Mengen an Fleisch und die durchschnittlichen Fleischpreise wurden jeweils im Basismomat 0 und im Berichtsmonat 12 erfaßt, vgl. **Tab. 17.** Die Ergebnisse, die die Indizes liefern, werden in Prozent „auf der Basis 0" angegeben. Der Wertindex

$$100 \cdot W_{0\,12} = 100 \cdot \frac{\sum_{i=1}^{7} p_{12i} q_{12i}}{\sum_{i=1}^{7} p_{0i} q_{0i}} = 100 \cdot \frac{62750{,}5}{62012{,}5} \simeq 101{,}19$$

sagt uns, daß der tatsächliche Aufwand für Fleisch für die privaten Haushalte vom Basismonat bis zum Berichtsmonat um 1,19% gestiegen ist. Dabei ist nicht berücksichtigt, daß der durchschnittliche Verbrauch im Berichtsmonat um 205 g geringer ist.

Für die beiden Preisindizes erhält man:

$$100 \cdot P^P_{0,t} = \frac{\sum_{i=1}^{7} p_{12i} q_{12i}}{\sum_{i=1}^{7} p_{0i} q_{12i}} = 100 \cdot \frac{62750{,}5}{59962{,}5} \simeq 104{,}65$$

Tab. 17: Durchschnittlicher Verbrauch an Fleisch aller privaten Haushalte und die durchschnittlichen Fleischpreise in einer Gemeinde im Basismonat und im Berichtsmonat; Arbeitstabelle

Fleisch	p_{0i} Preis im Basismonat (DM/kg)	q_{0i} verbrauchte Mengen (g)	p_{12i} Preis im Berichtsmonat (DM/kg)	q_{12i} verbrauchte Mengen (g)	$p_{0i}q_{0i}$	$p_{12i}q_{12i}$	$p_{0i}q_{12i}$	$p_{12i}q_{0i}$
Rindfleisch zum Kochen	5,75	1450	6,20	1410	8337,5	8742	8107,5	8990
Rindfleisch zum Schmoren und Braten	12,25	1260	13,55	1200	15435	16260	14700	17073
Rindfleisch, Lendenfilet	24,10	50	24,80	50	1205	1240	1205	1240
Roastbeef	17,80	210	18,10	215	3738	3891,5	3827	3801
Schweinebauchfleisch	4,55	2450	4,60	2470	11147,5	11362	11238,5	11270
Schweinebraten/Schulter	10,25	1310	10,40	1230	13427,5	12792	12607,5	13624
Schweinefleisch, Kotelett ohne Filet	8,90	980	9,10	930	8722	8463	8277	8918
insgesamt		7710		7505	62012,5	62750,5	59962,5	64916

und

$$100 \cdot P^L_{0,t} = \frac{\sum_{i=1}^{7} p_{1\,2i}\, q_{0i}}{\sum_{i=1}^{7} p_{0i}\, q_{0i}} = 100 \cdot \frac{64916}{62012{,}5} \simeq 104{,}68.$$

Nimmt man für beide Monate den durchschnittlichen Verbrauch an Fleisch im Berichtsmonat als **Mengenschema (Warenkorb)** an, so sind die Preise in diesem Zeitraum um 4,65% gestiegen; oder: Für die im Basismonat verbrauchten Mengen an Fleisch muß man im Berichtsmonat 4,68% mehr Geld ausgeben.

D. Mengenindizes nach Laspeyres und nach Paasche

Ist man nun daran interessiert, wie sich die verbrauchten Mengen unabhängig von der Preisentwicklung verändert haben, so kann man einen **Mengenindex** betrachten. Standardisiert man nach den Preisen zur Berichtsperiode, so erhält man einen

Mengenindex nach Paasche:

$$Q^P_{0,t} = \frac{\sum_{i=1}^{n} p_{ti}\, q_{ti}}{\sum_{i=1}^{n} p_{ti}\, q_{0i}}$$

und entsprechend nach den Preisen zur Basisperiode einen

Mengenindex nach Laspeyres:

$$Q^L_{0,t} = \frac{\sum_{i=1}^{n} p_{0i}\, q_{ti}}{\sum_{i=1}^{n} p_{0i}\, q_{0i}}.$$

Beispiel: Für unser Beispiel, vgl. auch Abschnitt C, sagt

$$100 \cdot Q^P_{0,t} = 100 \cdot \frac{62750{,}5}{64916} \simeq 96{,}66,$$

daß der Verbrauch an Fleisch, gewichtet mit den Preisen im Berichtsmonat, um 3,34% gesunken ist.

Entsprechend besagt

$$100 \cdot Q^L_{0,t} = 100 \cdot \frac{59962{,}5}{62012{,}5} \simeq 96{,}69,$$

daß der Verbrauch an Fleisch, gewichtet mit den Preisen zum Basismonat, um 3,31% gesunken ist.

E. Preisbereinigung; Deflationierung

Die Umbasierung bzw. Verkettung, wie wir sie bei den Meßzahlen kennengelernt haben, läßt sich nicht auf zusammengesetzte Indizes übertragen, denn im allgemeinen ist $Q_{0,r} \cdot Q_{r,t}$ ungleich Q_{0t} bzw. $P_{0,r} \cdot P_{r,t}$ ungleich $P_{0,t}$.

Es gilt aber

$$Q^L_{0,t} = \frac{\sum_i p_{0i} q_{ti}}{\sum_i p_{0i} q_{0i}} \cdot \frac{\sum_i p_{ti} q_{ti}}{\sum_i p_{ti} q_{ti}} = \frac{W_{0t}}{P^P_{0,t}}$$

und

$$Q^P_{0,t} = \frac{\sum_i p_{ti} q_{ti}}{\sum_i p_{ti} q_{0i}} \cdot \frac{\sum_i p_{0i} q_{0i}}{\sum_i p_{0i} q_{0i}} = \frac{W_{0t}}{P^L_{0,t}}.$$

Die Berechnung eines Mengenindex Q_{0t} aus einem Wertindex W_{0t} nennt man **Preisbereinigung** oder **Deflationierung**.

F. Preis- und Mengenindizes als gewogene Mittel von Meßzahlen: Subindizes

Eine Zahl, die das Preisverhältnis einer Ware zwischen Berichtsperiode und Basisperiode zum Ausdruck bringt, heißt **Preismeßzahl** und ist definiert durch $\frac{p_{ti}}{p_{0i}}$.

Die Preisindizes lassen sich dann als gewogenes Mittel von Preismeßzahlen darstellen: Den Preisindex nach Laspeyres erhält man als gewogenes arithmetisches Mittel:

$$P^L_{0,t} = \frac{\sum_i p_{ti} q_{0i}}{\sum_i p_{0i} q_{0i}} = \sum_i \frac{p_{ti} q_{0i}}{\sum_j p_{0j} q_{0j}} \cdot \frac{p_{0i} q_{0i}}{p_{0i} q_{0i}} = \sum_i g_{0i} \frac{p_{ti}}{p_{0i}}$$

$$\text{mit } g_{0i} = \frac{p_{0i} q_{0i}}{\sum_j p_{0j} q_{0j}}$$

(die Gewichte g_{0i} entsprechen den Ausgabenanteilen zur Basisperiode), den Preisindex nach Paasche erhält man als gewogenes harmonisches Mittel von Preismeßzahlen:

$$\frac{1}{P^P_{0,t}} = \frac{\sum_i p_{0i} q_{ti}}{\sum_j p_{tj} q_{tj}} = \sum_i \frac{p_{0i} q_{ti}}{\sum_j p_{tj} q_{tj}} \cdot \frac{p_{ti} q_{ti}}{p_{ti} q_{ti}} = \sum_i g_{ti} \frac{1}{p_{ti}/p_{0i}}$$

$$\text{mit } g_{ti} = \frac{p_{ti} q_{ti}}{\sum_j p_{tj} q_{tj}}.$$

(Dabei fungieren die Ausgabenanteile zur jeweiligen Berichtsperiode als Gewichte.)

Durch Vertauschen der Symbole p und q gelangt man zu entsprechenden Darstellungen für Mengenindizes. Der Grund dafür, daß in der amtlichen Statistik in der Regel die Wertgewichtsmethode zur Berechnung von Indizes verwendet wird, liegt in der Aufteilung von Warengütern in Untergruppen. So wird z. B. für den **Preisindex für die Lebenshaltung** das Güterverzeichnis für den privaten Verbrauch in 9 Hauptgruppen aufgeteilt; vgl. **Tab. 18.**

Für den Preisindex nach Laspeyres (Mengenindex analog) gilt folgende Aufspaltung in zwei **Subindizes** bei einem Warenkorb mit den Waren $i = 1, \ldots, n$ unterteilt in zwei Warengruppen I und II mit den Waren $i = 1, \ldots, k$ und $i = k+1, \ldots, n$:

Tab. 18: Hauptgruppen des Güterverzeichnisses des Lebenshaltungsindex

1	Nahrungs- und Genußmittel
2	Kleidung, Schuhe
3	Wohnungsmieten u.ä.
4	Elektrizität, Gas, Brennstoffe u.ä.
5	Waren und Dienstleistungen für die Haushaltsführung
6	Verkehrszwecke, Nachrichtenübermittlung
7	Körper- und Gesundheitspflege
8	Bildungs- und Unterhaltungszwecke
9	Persönliche Ausstattung und sonstiges

$$P_{0,t}^{L} = \sum_{i=1}^{n} g_{0i} \frac{p_{ti}}{p_{0i}} = \sum_{i=1}^{k} g_{0i} \frac{p_{ti}}{p_{0i}} + \sum_{i=k+1}^{n} g_{0i} \frac{p_{ti}}{p_{0i}}$$

$$= g_0^{I} \sum_{i=1}^{k} \frac{g_{0i}}{g_0^{I}} \frac{p_{ti}}{p_{0i}} + g_0^{II} \sum_{i=k+1}^{n} \frac{g_{0i}}{g_0^{II}} \frac{p_{ti}}{p_{0i}}$$

mit den Gewichten $g_0^{I} = \dfrac{\sum_{i=1}^{k} p_{0i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}}$ und $g_0^{II} = \dfrac{\sum_{i=k+1}^{n} p_{0i} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}}$,

welche den Ausgabenanteilen der Warengruppen zur Basisperiode entsprechen.

Da $$g_{0i}^{I} = \frac{g_{0i}}{g_0^{I}} = \frac{p_{0i} q_{0i}}{\sum_{j=1}^{n} p_{0j} q_{0j}} \cdot \frac{\sum_{j=1}^{n} p_{0j} q_{0j}}{\sum_{j=1}^{k} p_{0j} q_{0j}} = \frac{p_{0i} q_{0i}}{\sum_{j=1}^{k} p_{0j} q_{0j}}$$

und $$g_{0i}^{II} = \frac{g_{0i}}{g_0^{II}} = \frac{p_{0i} q_{0i}}{\sum_{j=k+1}^{n} p_{0j} q_{0j}}$$

den Ausgabenanteilen einer Ware i an der jeweiligen Warengruppe zur Basisperiode entsprechen, erhält man die Preisindizes für diese Warengruppe; es ist also

$$P_{0,t}^{L} = g_0^{I} P_{0,t}^{L(I)} + g_0^{II} P_{0,t}^{L(II)}$$

mit den Subindizes

$$P_{0,t}^{L(I)} = \sum_{i=1}^{k} g_{0i}^{I} \frac{p_{ti}}{p_{0i}} \quad \text{und}$$

$$P_{0,t}^{L(II)} = \sum_{i=k+1}^{n} g_{0i}^{II} \frac{p_{ti}}{p_{0i}}.$$

Analog läßt sich der Preisindex nach Paasche als gewichtetes harmonisches Mittel von Subindizes darstellen:

$$\frac{1}{P^{P}_{0,t}} = \sum_{i=1}^{n} g_{ti} \frac{1}{p_{ti}/p_{0i}} = \sum_{i=1}^{k} g_{ti} \frac{1}{p_{ti}/p_{0i}} + \sum_{i=k+1}^{n} g_{ti} \frac{1}{p_{ti}/p_{0i}}$$

$$= g^{I}_{t} \frac{1}{P^{P(I)}_{0,t}} + g^{II}_{t} \frac{1}{P^{P(II)}_{0,t}}$$

mit den Gewichten $g^{I}_{t} = \dfrac{\sum_{i=1}^{k} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{ti} q_{ti}}$ und $g^{II}_{t} = \dfrac{\sum_{i=k+1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{ti} q_{ti}}$

und den Subindizes

$$\frac{1}{P^{P(I)}_{0,t}} = \sum_{i=1}^{k} \frac{g_{ti}}{g^{I}_{t}} \frac{1}{p_{ti}/p_{0i}} = \sum_{i=1}^{k} \frac{p_{ti} q_{ti}}{\sum_{j=1}^{k} p_{tj} q_{tj}} \frac{1}{p_{ti}/p_{0i}},$$

$$\frac{1}{P^{P(II)}_{0,t}} = \sum_{i=k+1}^{n} \frac{p_{ti} q_{ti}}{\sum_{j=k+1}^{n} p_{tj} q_{tj}} \frac{1}{p_{ti}/p_{0i}}$$

Beispiel: Sehen wir uns die Berechnung von Subindizes in unserem Beispiel für die Unterwarengruppen Rind- und Schweinefleisch an, und beschränken wir uns auf die Preisindizes, vgl. **Tab. 19**.

Tab. 19: Berechnung der Subindizes mit der Gewichtungsmethode, mit Gewichten

$g^{I}_{ti} = \dfrac{p_{ti} q_{ti}}{\sum_{i=1}^{4} p_{ti} q_{ti}}$ und $g^{II}_{ti} = \dfrac{p_{ti} q_{ti}}{\sum_{i=5}^{7} p_{ti} q_{ti}}$ für die Monate $t = 0$ und $t = 12$

	Preismeßzahlen	Gewichte			
Untergruppe Rindfleisch	$\frac{p_{12i}}{p_{0i}} \cdot 100$	g^{I}_{0i}	g^{I}_{12i}	$g^{I}_{0i} \frac{p_{12i}}{p_{0i}} \cdot 100$	$g^{I}_{12i} \frac{1}{p_{12i}/p_{0i}}$
zum Kochen	107,83	0,290	0,290	31,270	0,2690
zum Schmoren und Braten	110,61	0,538	0,540	59,509	0,4882
Lendenfilet	102,90	0,042	0,041	4,322	0,0398
Roastbeef	101,69	0,130	0,129	13,219	0,1269
gesamt		1	1	108,320	0,9239
Untergruppe Schweinefleisch	$\frac{p_{12i}}{p_{0i}} \cdot 100$	g^{II}_{0i}	g^{II}_{12i}	$g^{II}_{0i} \frac{p_{12i}}{p_{0i}} \cdot 100$	$g^{II}_{12i} \frac{1}{p_{12i}/p_{0i}}$
Bauchfleisch	101,10	0,335	0,348	33,868	0,3442
Braten/Schulter	101,46	0,403	0,392	40,890	0,3863
Kotelett (o. Filet)	102,25	0,262	0,260	26,790	0,2543
gesamt		1	1	101,548	0,9848

Zu jeder Ware werden die Preismeßzahlen berechnet, welche die prozentualen Veränderungen gegenüber dem Basismonat angeben. Für Rindfleisch zum Kochen z. B. betrug der Preis im Berichtsmonat 107,83% des Preises im Basismonat, anders ausgedrückt ist der Preis um 7,83% gestiegen. Die in der Tabelle 19 stehenden Gewichte $g_{0i}^{I(II)}$ bzw. $g_{12i}^{I(II)}$ geben an, welche Ausgaben für die i-te Ware im Verhältnis zu den Gesamtausgaben einer Unterwarengruppe, also Rindfleisch (I) oder Schweinefleisch (II), von allen privaten Haushalten der Gemeinde im Durchschnitt getätigt wurden (in beiden Monaten $t = 0$ und $t = 12$). Für die Preisindizes nach Laspeyres in den beiden Gruppen erhält man:

$$100 \cdot P_{0,12}^{L(I)} = 108{,}32 \quad \text{bzw.} \quad 100 \cdot P_{0,12}^{L(II)} = 101{,}55.$$

Also ist der Warenkorb des Basismonats für die Warengruppe Rindfleisch um 8,32% und für die Warengruppe Schweinefleisch nur um 1,55% gestiegen.

Mit den Ausgabenanteilen der Warengruppen an den Gesamtausgaben (zum Basismonat) als Gewichte

$$g_0^I = \frac{\sum_{i=1}^{4} p_{0i} q_{0i}}{\sum_{i=1}^{7} p_{0i} q_{0i}} = \frac{28715{,}5}{62012{,}5} \simeq 0{,}4631 \quad \text{und}$$

$$g_0^{II} = \frac{\sum_{i=5}^{7} p_{0i} q_{0i}}{\sum_{i=1}^{7} p_{0i} q_{0i}} = \frac{33297}{62012{,}5} \simeq 0{,}5369,$$

erhält man den gemeinsamen Preisindex nach Laspeyres:

$$100 \cdot P_{0,12}^{L} = 100 \cdot (g_0^I P_{0,12}^{L(I)} + g_0^{II} P_{0,12}^{L(II)}) = 104{,}7.$$

Die Preisindizes nach Paasche für die beiden Warengruppen berechnet man aus den Gewichten g_{12i}^I bzw. g_{12i}^{II} und den Kehrwerten der Preismeßzahlen:

$$100 \cdot P_{0,12}^{P(I)} = \frac{100}{\sum_{i=1}^{4} g_{12i}^I \frac{1}{p_{12i}/p_{0i}}} = \frac{100}{0{,}9239} \simeq 108{,}24,$$

$$100 \cdot P_{0,12}^{P(II)} = \frac{100}{\sum_{i=5}^{7} g_{12i}^{II} \frac{1}{p_{12i}/p_{0i}}} = \frac{100}{0{,}9848} \simeq 101{,}54.$$

Den Warenkorb des Berichtsmonats zugrundelegend, ist der Preis für Rindfleisch um 8,24% und für Schweinefleisch um 1,54% gestiegen. Für den gemeinsamen Preisindex nach Paasche gilt:

$$100 \cdot P_{0,12}^{P} = \frac{100}{g_{12}^I \frac{1}{P_{0,12}^{P(I)}} + g_{12}^{II} \frac{1}{P_{0,12}^{P(II)}}}$$

$$= \frac{100}{0{,}4802 \cdot 0{,}9239 + 0{,}5198 \cdot 0{,}9848} \simeq 104{,}65$$

mit den Gewichten

$$g_{12}^{I} = \frac{\sum_{i=1}^{4} p_{12i} q_{12i}}{\sum_{i=1}^{7} p_{12i} q_{12i}} = \frac{30133{,}5}{62750{,}5} \simeq 0{,}4802 \quad \text{und}$$

$$g_{12}^{II} = \frac{\sum_{i=5}^{7} p_{12i} q_{12i}}{\sum_{i=1}^{7} p_{12i} q_{12i}} = \frac{32617}{62750{,}5} \simeq 0{,}5198.$$

7.3.4. Vergleich von Preisindizes nach Laspeyres und nach Paasche; der Fishersche Idealindex; der Preisindex nach Lowe

Für den Preisindex nach Laspeyres spricht die Tatsache, daß der Warenkorb nur für die Basisperiode bestimmt werden muß und nicht für die weiteren Berichtsperioden wie beim Preisindex nach Paasche. Dabei ist nicht nur der Kostengesichtspunkt gemeint, sondern auch die Frage nach der Aktualität, da die umfangreichen Berechnungen eines Warenkorbes sich oft über einen längeren Zeitraum erstrecken. Sind diese Arbeiten einmal abgeschlossen, so läßt sich der Preisindex nach Laspeyres bei fortlaufender Preisnotierung schnell berechnen. Sein Nachteil besteht darin, daß die Ausgabenstruktur der Basisperiode veraltet, so daß nach einer bestimmten Zeit (ca. 4–8 Jahre) der Warenkorb neu festgesetzt werden muß. Der Preisindex nach Paasche paßt sich dagegen laufend der neuen Verbrauchssituation an. In den meisten Fällen wird daher der Preisindex nach Laspeyres benutzt und zu Kontrollzwecken der nach Paasche. Ergeben sich größere Abweichungen, so muß der Warenkorb neu festgelegt werden.

Häufig wird zur Überprüfung der Indizes auch noch der **Fishersche Idealindex** (nach Irving Fisher) berechnet, der sich gerade als geometrisches Mittel der Indizes nach Paasche und Laspeyres ergibt:

$$I_F = \sqrt{I_P \cdot I_L}.$$

Der Fishersche Idealindex nimmt also stets einen Wert an, der zwischen den Werten der Indizes von Paasche und Laspeyres liegt.

Manchmal wird nicht der Verbrauch der Basisperiode zur Berechnung der Indizes verwendet (z. B. wegen Saisonschwankungen), sondern ein anderer „geeigneter" Verbrauch, z. B. beim Preisindex der Lebenshaltung der Durchschnittsverbrauch eines Jahres. Dies führt zum **Preisindex nach Lowe**, auf den wir hier nicht weiter eingehen wollen.

8. Die empirische Ausfallrate

Beschreiben die Daten aus einer Erhebung die Zeit bis zum Tod von Untersuchungsobjekten, so kann die **empirische Ausfallrate** zur Beschreibung der Daten herangezogen werden. Man unterteilt die Zeitachse in Intervalle $[0; t_1)$, $[t_1; t_2)$, ... und berechnet die Anzahlen $H_1, H_2, \ldots$ der Versuchsobjekte mit Lebensdauern im Intervall $[0; t_1)$, $[t_1; t_2)$, ... Bei insgesamt n betrachteten Lebensdauern ist die empirische Ausfallrate definiert als:

$$r_{emp}(t) = \frac{H_i}{(n - \sum_{j=1}^{i-1} H_j)(t_i - t_{i-1})} \quad \text{für } t \in [t_{i-1}; t),\ t_0 = 0.$$

Sie gibt – bezogen auf die Klassenbreite – den Anteil der im Intervall $[t_{i-1}; t_i)$ ausfallenden bzw. sterbenden Objekte H_i bezogen auf die Anzahl der zum Zeitpunkt t_{i-1} noch funktionierenden bzw. lebenden Objekte an. Sie hängt recht stark von der gewählten Klasseneinteilung ab, wie man im nachfolgenden Beispiel sieht.

Ist die empirische Ausfallrate ziemlich konstant, so heißt dies, daß die Objekte der Erhebung mit fortschreitender Zeit weder „altern" noch sich „verjüngen". Steigt die empirische Ausfallrate mit wachsender Zeit t, so „altern" die Objekte bzw. zeigen Verschleißerscheinungen, fällt sie, so „verjüngen" sich die Objekte mit fortschreitender Zeit, zeigen also mit zunehmender Zeit immer weniger Neigung zum Ausfallen; vgl. auch Abschnitt 4 in Kap. IV und Kap. XIII.

Beispiel: Wir wollen die empirische Ausfallrate $r_{emp}(t)$ am Beispiel der 87 Kühlaggregate, vgl. Tab. 4, berechnen und dafür zwei verschiedene Klasseneinteilungen wählen. Bei der Einteilung in Klassen der Breite 1 Jahr ergibt sich z. B.:

$$r_{emp}(t) = \frac{37}{(87-0)(1-0)} = 0{,}425 \quad \text{für } t \in [0;1),$$

$$r_{emp}(t) = \frac{19}{(87-37)(2-1)} = 0{,}380 \quad \text{für } t \in [1;2),$$

$$r_{emp}(t) = \frac{15}{(87-37-19)(3-2)} = 0{,}484 \quad \text{für } t \in [2;3), \quad \text{usw.}$$

Diese empirische Ausfallrate ist in **Abb. 25** abgetragen. Als zweites wählen wir die Einteilung in Klassen der Breite 0,5 Jahre und erhalten durch

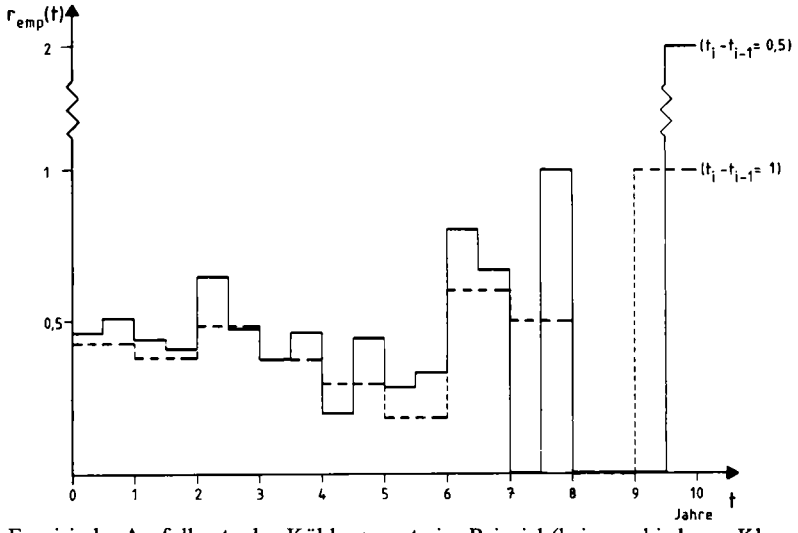

Abb. 25: Empirische Ausfallsrate der Kühlaggregate im Beispiel (bei verschiedenen Klasseneinteilungen)

$$r_{emp}(t) = \frac{20}{(87-0)(0{,}5-0)} = 0{,}460 \quad \text{für } t \in [0;0{,}5),$$

$$r_{emp}(t) = \frac{17}{(87-20)(1-0{,}5)} = 0{,}508 \quad \text{für } t \in [0{;}5, 1),$$

$$r_{emp}(t) = \frac{11}{(87-20-17)(1{,}5-1)} = 0{,}440 \quad \text{für } t \in [1;1{,}5); \quad \text{usw.}$$

die andere in Abb. 25 dargestellte empirische Ausfallrate.

Der Zeichnung kann man entnehmen, daß 0,45 eine grobe konstante Approximation der empirischen Ausfallrate darstellt. Wir können für die in der Zeit beobachteten Kühlaggregate schließen, daß sie im Schnitt keine wesentlichen „Verschleißerscheinungen" zeigten. Die relativen Ausfallhäufigkeiten zeigen keine gleichmäßig steigende oder fallende Tendenz.

9. Darstellung zweidimensionalen Zahlenmaterials und deskriptive Korrelationsrechnung

In Abschnitt 1 dieses Kapitels wurde der Begriff des mehrdimensionalen Merkmals eingeführt. Wir wollen uns hier auf zweidimensionale Merkmale beschränken und vor allem untersuchen, wie man die Abhängigkeit bzw. den Zusammenhang zwischen zwei Merkmalen messen kann.

Dabei können folgende Fragestellungen auftreten:

- Wie stark ist der Zusammenhang bzw. die Abhängigkeit?
- Läßt sich der Zusammenhang in einer bestimmten Form darstellen?

Wir wollen uns hier mit der ersten Fragestellung befassen, die unter dem Begriff **Korrelationsrechnung,** vgl. auch Kap. IX, bekannt ist. Die zweite Fragestellung wird in der **Regressionsrechnung** (vgl. Kap. X) behandelt.

Im folgenden betrachten wir n Untersuchungseinheiten, an denen die Merkmale X und Y beobachtet werden. Die verschiedenen Ausprägungen von X werden mit $a_1, \ldots, a_l$, die von Y mit $b_1, \ldots, b_m$ bezeichnet. Somit erhält man das zweidimensionale Merkmal (X, Y), dessen Ausprägungen (a_j, b_k) mit absoluter Häufigkeit H_{jk} und mit relativer Häufigkeit $h_{jk} = \frac{1}{n} H_{jk}$ auftreten.

9.1. Die Kontingenztafel

Die Häufigkeitsverteilung von (X, Y) wird meistens in Form einer **Kontingenztafel** dargestellt, vgl. **Tab. 20** und **Tab. 21**.

$H_{j.} = \sum_{k=1}^{m} H_{jk}$ nennt man **absolute Randhäufigkeit** und

$h_{j.} = \sum_{k=1}^{m} h_{jk}$ **relative Randhäufigkeit** von a_j für $j = 1, \ldots, l$.

Durch die Randhäufigkeiten erhält man die Häufigkeitsverteilung des Merkmals X, die dann **Randverteilung** genannt wird.

Tab.20 Kontingenztafel der absoluten Häufigkeiten H_{ij}

	b_1	b_2	...	b_m	$\sum$
a_1	H_{11}	H_{12}	...	H_{1m}	$H_{1.}$
a_2	H_{21}	H_{22}	...	H_{2m}	$H_{2.}$
⋮	⋮	⋮		⋮	⋮
a_l	H_{l1}	H_{l2}	...	H_{lm}	$H_{l.}$
$\sum$	$H_{.1}$	$H_{.2}$	...	$H_{.m}$	n

Tab.21: Kontingenztafel der relativen Häufigkeiten h_{ij}

	b_1	b_2	...	b_m	$\sum$
a_1	h_{11}	h_{12}	...	h_{1m}	$h_{1.}$
a_2	h_{21}	h_{22}	...	h_{2m}	$h_{2.}$
⋮	⋮	⋮		⋮	⋮
a_l	h_{l1}	h_{l2}	...	h_{lm}	$h_{l.}$
$\sum$	$h_{.1}$	$h_{.2}$	...	$h_{.m}$	n

Entsprechend liefern die Randhäufigkeiten von b_k, die durch

$$H_{.k} = \sum_{j=1}^{l} H_{jk} \quad \text{bzw.} \quad h_{.k} = \sum_{j=1}^{l} h_{jk} \quad \text{für } k = 1, \ldots, m$$

gegeben sind, die Randverteilung von Y.

Es gilt:

$$h_{j.} = \frac{1}{n} H_{j.}, \quad h_{.k} = \frac{1}{n} H_{.k},$$

$$\sum_{j=1}^{l} \sum_{k=1}^{m} H_{jk} = \sum_{j=1}^{l} H_{j.} = \sum_{k=1}^{m} H_{.k} = n \quad \text{und}$$

$$\sum_{j=1}^{l} \sum_{k=1}^{m} h_{jk} = \sum_{j=1}^{l} h_{j.} = \sum_{k=1}^{m} h_{.k} = 1.$$

Auf die Analyse von Kontigenztafeln wird in Kap. VII eingegangen.

9.2. Der Korrelationskoeffizient nach Bravais-Pearson

Wir betrachten nun die Ausprägungen eines metrisch skalierten Merkmals (X, Y). Die verschiedenen Ausprägungen (a_j, b_k) treten mit Häufigkeiten H_{jk} für $j = 1, \ldots, l$, $k = 1, \ldots, m$ auf. Ein Maß für den Zusammenhang zwischen X und Y liefert dann der **Korrelationskoeffizient (nach Bravais-Pearson)** r_{XY}, der gegeben ist durch

$$r_{XY} = \frac{\sum_{j=1}^{l} \sum_{k=1}^{m} (a_j - \bar{a})(b_k - \bar{b}) H_{jk}}{\sqrt{\left(\sum_{j=1}^{l} (a_j - \bar{a})^2 H_{j.}\right)\left(\sum_{k=1}^{m} (b_k - \bar{b})^2 H_{.k}\right)}}$$

$$= \frac{\sum_{j=1}^{l}\sum_{k=1}^{m} a_j b_k H_{jk} - n \cdot \bar{a} \cdot \bar{b}}{\sqrt{\left(\sum_{j=1}^{l} a_j^2 H_{j.} - n\bar{a}^2\right)\left(\sum_{k=1}^{m} b_k^2 H_{.k} - n\bar{b}^2\right)}}$$

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

wenn (x_i, y_i) die Beobachtung an der i-ten Untersuchungseinheit bezeichnet.

Der Zähler von r_{XY} ist im wesentlichen die empirische **Kovarianz** zwischen X und Y, die als

$$s_{XY} = \frac{1}{n-1}\sum_{j=1}^{l}\sum_{k=1}^{m}(a_j - \bar{a})(b_k - \bar{b})H_{jk}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

definiert ist und die man auch als Maß des Zusammenhangs benutzen kann. Wenn man s_{XY} durch die Standardabweichungen s_X und s_Y von X bzw. Y dividiert, erhält man dann als ein normiertes Zusammenhangsmaß den Korrelationskoeffizienten r_{XY}, der nur Werte zwischen -1 und $+1$ annehmen kann.

Um zu erkennen, wie sich der Korrelationskoeffizient r_{XY} bei verschiedenen Konstellationen von Ausprägungen des Merkmals (X, Y) verhält, werden im folgenden Beispiel jeweils für gleichbleibende Ausprägungen von X die Ausprägungen von Y so variiert, daß s_Y konstant bleibt.

Beispiel: In **Tab. 22** sind für je 8 Untersuchungseinheiten die Beobachtungsdaten (x_i, y_i) der Merkmale X und Y, wobei die Ausprägungen von X gleichbleiben und die von Y variieren, die Korrelationskoeffizienten nach Bravais-Pearson berechnet.

Tab. 22: Korrelationskoeffizient bei verschiedenen Konstellationen von Ausprägungen

x_i	y_i	y_i	y_i	y_i	y_i	y_i	y_i	y_i
0	1	2	4	1	2	3	8	8
2	2	1	3	8	7	8	6	7
4	3	4	2	2	5	5	4	6
6	4	3	1	7	3	2	2	5
8	5	6	8	3	8	7	7	4
10	6	5	7	6	4	4	3	3
12	7	8	6	4	1	1	5	2
14	8	7	5	5	6	6	1	1
$r_{XY} =$	1,00	0,905	0,52	0,19	0,00	−0,14	−0,62	−1,00

Um eine Vorstellung von der Größe des Korrelationskoeffizienten r_{XY} zu bekommen, haben wir in den **Abb. 26 bis 33** diese Kombinationen von Merkmalsausprägungen graphisch dargestellt.

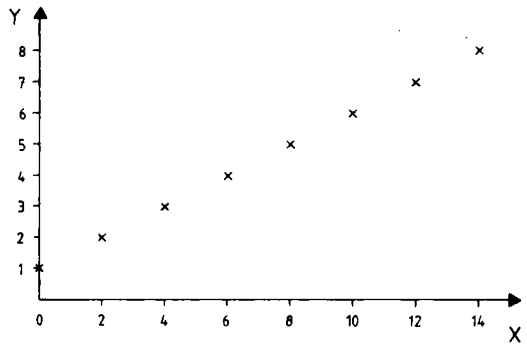

Abb. 26: $r_{XY} = 1{,}00$

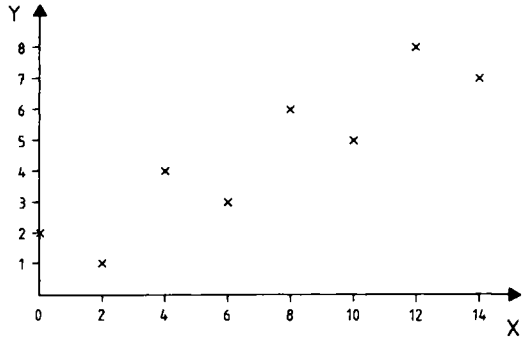

Abb. 27: $r_{XY} = 0{,}905$

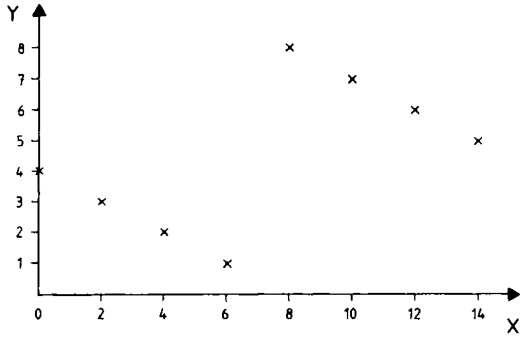

Abb. 28: $r_{XY} = 0{,}52$

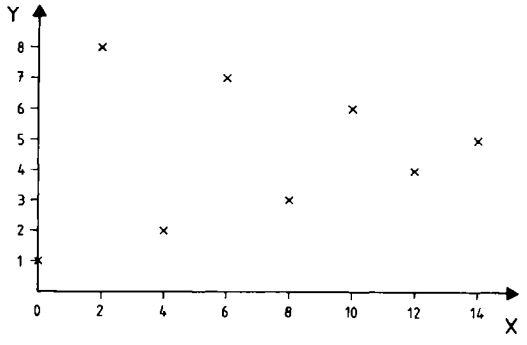

Abb. 29: $r_{XY} = 0{,}19$

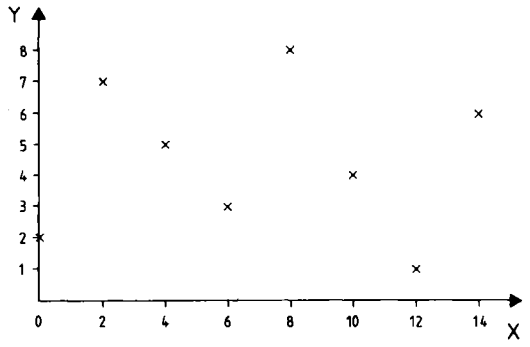

Abb. 30: $r_{XY} = 0{,}00$

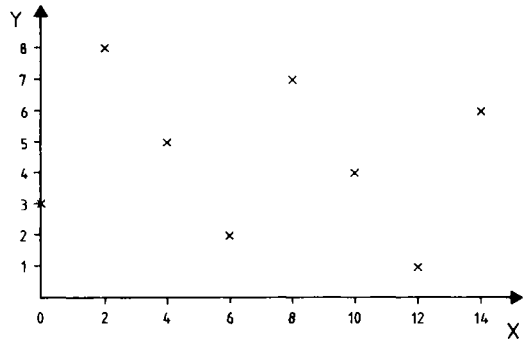

Abb. 31: $r_{XY} = -0{,}14$

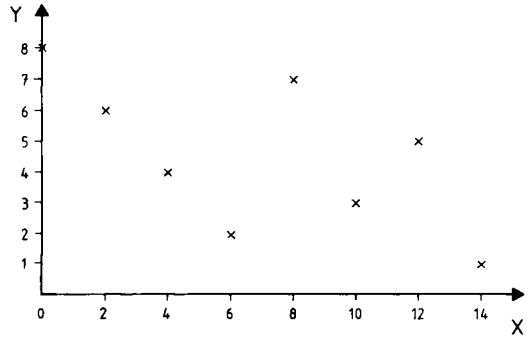

Abb. 32: $r_{XY} = -0{,}62$

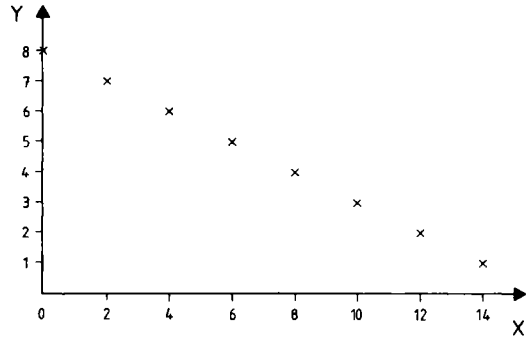

Abb. 33: $r_{XY} = -1{,}00$

Tab. 23: Studienanfänger im Jahr 1977 und Studenten insgesamt im Jahr 1977 an 10 Universitäten der Bundesrepublik Deutschland (vgl. Grund- und Strukturdaten 1979; der Bundesminister für Bildung und Wissenschaft)

	Univ. Bochum	Univ. Bremen	Univ. Clausthal	Univ. Dortmund	Univ. Frankfurt	Univ. Freiburg	Univ. Hamburg	Univ. Heidelberg	Univ. Kiel	Univ. Münster
Studienanfänger	3970	732	499	1300	3463	2643	3630	3294	1931	5048
Studenten insgesamt	24273	5883	2847	5358	23442	17076	28360	19812	12379	31433

Man sieht am Beispiel, daß r_{XY} nahe 1 liegt, wenn der Zusammenhang zwischen X und Y annähernd positiv linear ist und nahe -1 liegt, wenn eine negativ lineare Abhängigkeit zu erkennen ist. Je „verstreuter" die Ausprägungen (x_i, y_i) in der Ebene liegen, desto näher liegt r_{XY} bei Null.

Beispiel: Aus den Daten der **Tab. 23** wollen wir die empirische Korrelation zwischen der Anzahl der Studienanfänger X und der Gesamtzahl von Studenten an einer Universität bestimmen. Es ergibt sich

$$r_{XY} = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{10} (x_i - \bar{x})^2\right)\left(\sum_{i=1}^{10} (y_i - \bar{y})^2\right)}}$$

$$= \frac{\sum_{i=1}^{10} (x_i - 2651)(y_i - 17086{,}3)}{\sqrt{\left(\sum_{i=1}^{10} (x_i - 2651)^2\right)\left(\sum_{i=1}^{10} (y_i - 17086{,}3)^2\right)}}$$

$$= \frac{133195081}{\sqrt{20173924 \cdot 920380528}} = \frac{133195081}{136263300} = 0{,}9775.$$

Es besteht also ein starker linearer Zusammenhang zwischen der Anzahl der Studienanfänger und der Anzahl aller Studenten an einer Universität.

Schon an dieser Stelle soll auf die Gefahr der **Scheinkorrelation** hingewiesen werden. Scheinkorrelation liegt dann vor, wenn die beiden Merkmale X und Y beide von einem dritten Merkmal Z abhängen: Die Gefahr eines Waldbrandes und schlechter Kornertrag zum Beispiel hängen von der Stärke der Sonnenbestrahlung ab. Ebenso sollte man darauf achten, daß die zu untersuchenden Größen in einem sachlogischen Zusammenhang stehen, damit keine **Nonsenskorrelation** entsteht, wie es z. B. die Korrelation zwischen der Anzahl der Störche und der Anzahl der Geburten in einem Land oder die Korrelation zwischen Apfelsinen – Einfuhrmenge und Anzahl der Krebserkrankungen ist. Gewarnt sei auch noch davor, durch geschickte Bastelei eine optimale Korrelation zu erhalten. Dies kann dadurch geschehen, daß man einige „unerwünschte" Werte streicht, Werte einer Variablen einer zeitlichen Verschiebung unterwirft, bei einer Variablen zu logarithmischen Werten

übergeht etc. Ebenfalls sei noch einmal darauf hingewiesen, daß r_{XY} nur einen linearen Zusammenhang zwischen X und Y mißt und somit r_{XY} Null sein kann, obwohl ein starker nichtlinearer Zusammenhang vorliegt, vgl. **Abb. 34**.

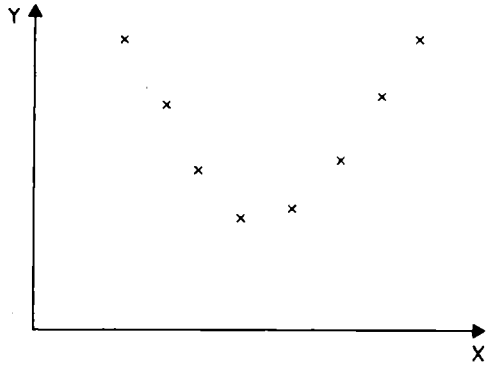

Abb. 34: Nichtlinearer Zusammenhang mit Korrelation $r_{XY} = 0$

9.3. Der Fechnersche Korrelationskoeffizient

Ein anderer – sehr leicht zu berechnender – Korrelationskoeffizient für metrisch skalierte Merkmalsausprägungen zweier Merkmale X und Y ist der **Fechnersche Korrelationskoeffizient** r_F. Dieser basiert nicht auf den ursprünglichen Merkmalsausprägungspaaren $(x_1, y_1), \ldots, (x_n, y_n)$, sondern auf den Vorzeichen innerhalb der transformierten Paare $(x_1 - \bar{x}, y_1 - \bar{y}), \ldots, (x_n - \bar{x}, y_n - \bar{y})$.

Wir bestimmen die Summe

$$V = \sum_{i=1}^{n} v_i, \quad \text{wobei } v_i = \begin{cases} 1, & \text{falls } x_i - \bar{x} \text{ und } y_i - \bar{y} \text{ das gleiche Vorzeichen haben oder beide Null sind} \\ 1/2, & \text{falls genau einer der Werte } x_i - \bar{x} \text{ bzw. } y_i - \bar{y} \text{ gleich Null ist} \\ 0, & \text{sonst} \end{cases}$$

und daraus den Fechnerschen Korrelationskoeffizienten

$$r_F = \frac{2V - n}{n},$$

der stets Werte zwischen -1 und $+1$ annimmt. Allerdings nimmt er z. B. den Wert $+1$ nicht nur bei positiv linearem Zusammenhang an, sondern schon dann, wenn für alle Paare (x_i, y_i) gilt:

$$(x_i < \bar{x} \text{ und } y_i < \bar{y}) \quad \text{oder} \quad (x_i > \bar{x} \text{ und } y_i > \bar{y}).$$

Beispiel: Bei $n = 15$ Hennen wurden die Merkmale $X \triangleq$ Körpergewicht und $Y \triangleq$ monatliche Legeleistung bestimmt. Als Maß für die Abhängigkeit zwischen diesen beiden Merkmalen wollen wir den Fechnerschen Korrelationskoeffizienten r_F bestimmen. Die festgestellten Merkmalsausprägungen sind in **Tab. 24** zusammengestellt, die gleichzeitig auch unsere Arbeitstabelle ist.

Damit ergibt sich

Tab. 24: Legeleistung und Gewicht von 15 Hennen

Henne i	Körpergewicht x_i	Legeleistung y_i	$x_i - \bar{x} = x_i - 1847$	$y_i - \bar{y} = y_i - 21$	v_i
1	1763	19	−84	−2	1
2	1890	24	+ 43	+3	1
3	1872	23	+ 25	+2	1
4	1938	26	+ 91	+5	1
5	1791	22	− 56	+1	0
6	1854	18	+ 7	−3	0
7	1960	21	+113	+0	1/2
8	1723	20	−124	−1	1
9	1898	21	+ 51	+0	1/2
10	1834	20	− 13	−1	1
11	1946	24	+99	+3	1
12	1755	19	− 92	−2	1
13	1846	21	+−1	+0	1/2
14	1752	17	− 95	−4	1
15	1884	20	+ 37	−1	0

$$r_F = \frac{2 \cdot 10{,}5 - 15}{15} = \frac{7}{15} = 0{,}4.$$

Der Fechnersche Korrelationskoeffizient ist also recht groß, da eindeutig eine positive Korrelation zwischen den beiden Merkmalen vorhanden ist.

Zum Vergleich wollen wir noch den Korrelationskoeffizienten nach Bravais-Pearson berechnen:

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 \cdot s_Y^2}} = \frac{120{,}28492}{\sqrt{5755{,}0666 \cdot 6}} = \frac{120{,}28492}{185{,}82357} = 0{,}648.$$

Der Fechnersche Korrelationskoeffizient, der die Größe der Abstände ja nicht berücksichtigt, hat in diesem Fall sogar einen kleineren Wert als r_{XY}.

9.4. Der Spearmansche Rangkorrelationskoeffizient

Bei ordinal skalierten Merkmalen kann man den Korrelationskoeffizienten nach Bravais-Pearson nicht verwenden. Um ein Zusammenhangsmaß für ordinal skalierte Merkmale zu erhalten, nutzt man aus, daß die Ausprägungen solcher Merkmale eine eindeutige Rangfolge haben. Deshalb muß der Begriff der **Rangzahl** eingeführt werden. Dazu betrachtet man die Ausprägungen $z_1, \ldots, z_n$ eines Merkmals Z und ordnet sie der Größe nach an (hier von der größten zur kleinsten). Die geordnete Reihe bezeichnet man mit $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Wenn alle $z_{(1)}, \ldots, z_{(n)}$ verschieden sind, numeriert man sie von 1 bis n durch und definiert die Rangzahl von $z_{(i)}$ als

$$R(z_{(i)}) = i \quad \text{für} \quad i = 1, \ldots, n.$$

Tritt eine Ausprägung mehrmals auf, so ordnet man diesen gleichen Ausprägungen als Rang das arithmetische Mittel der Ränge zu, die sie einnehmen: Man erhält z. B. für die Beobachtungsreihe $z_{(1)} = 8, z_{(2)} = 5, z_{(3)} = 5, z_{(4)} = 2$ die Ränge $R(z_{(1)}) = 1$, $R(z_{(2)}) = R(z_{(3)}) = 2{,}5$, $R(z_{(4)}) = 4$.

Für zweidimensionale ordinal skalierte Merkmale (X, Y) mit an den n Untersu-

chungseinheiten beobachteten Ausprägungen (x_i, y_i) für $i = 1, \ldots, n$ kann man dann den **Spearmanschen Rangkorrelationskoeffizienten** r_s berechnen, der dem gewöhnlichen Korrelationskoeffizienten (nach Bravais-Pearson) r_{XY} der Rangzahlen entspricht.

Es gilt also:

$$r_S = \frac{\sum_{i=1}^{n} R(x_i)\,R(y_i) - n\,\overline{R(x)}\;\overline{R(y)}}{\sqrt{\left(\sum_{i=1}^{n} R(x_i)^2 - n\,\overline{R(x)}^2\right)\left(\sum_{i=1}^{n} R(y_i)^2 - n\,\overline{R(y)}^2\right)}}.$$

Wenn man r_S umformt, erhält man mit $d_i = R(x_i) - R(y_i)$ die einfachere Formel

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \text{ falls alle } x_i \text{ und } y_i \text{ verschieden sind.}$$

Beispiel: Ein Professor hat in einem Semester Daten gesammelt und dabei festgestellt, daß die beste Klausur vom Studenten mit den meisten Übungsaufgabenpunkten und die schlechteste Klausur vom Studenten mit den wenigsten Übungsaufgabenpunkten geschrieben wurde, vgl. **Tab. 25**. Er vermutet deshalb einen Zusammenhang zwischen Punkten in der Klausur X und in den Übungsaufgaben Y.

Tab. 25: Klausur- und Übungspunkte von 10 Studenten

Student	A	B	C	D	E	F	G	H	I	J
Klausurpunkte	76	44	32	53	25	58	26	59	29	65
Rang der Klausur	1	6	7	5	10	4	9	3	8	2
Übungsaufgabenpunkte	122	67	68	101	42	59	118	79	83	89
Rang der Übungsaufgaben	1	8	7	3	10	9	2	6	5	4
d_i	0	−2	0	2	0	−5	7	−3	3	−2
d_i^2	0	4	0	4	0	25	49	9	9	4

Es ist

$$r_S = 1 - \frac{6 \cdot 104}{10(100-1)} = 0{,}37$$

dann ein Maß für den Zusammenhang zwischen Übungsaufgabenpunkten und Klausurergebnis.

Der Spearmansche Korrelationskoeffizient nimmt nicht nur bei streng linearem Zusammenhang zwischen X und Y den Wert 1 an, sondern auch dann, wenn die Beobachtungen monoton wachsend sind, d.h. wenn für alle (x_i, y_i), (x_j, y_j) gilt:

$$\text{mit } x_i < x_j \text{ ist auch } y_i < y_j,$$

da er ja nur auf den Rangzahlen, nicht aber auf tatsächlich interpretierbaren Zahlenabständen beruht.

9.5. Der Kendallsche Rangkorrelationskoeffizient

Ein anderes Maß für die Korrelation, das man mit Hilfe der Rangzahlen ermittelt, ist **Kendalls τ**:

Wir ordnen zunächst den n Ausprägungen der Merkmale X und Y jeweils wie vorne beschrieben Rangzahlen zu. Dann ordnen wie die Paare $(x_1, y_1), \ldots, (x_n, y_n)$ so, daß die Rangzahlen des Merkmals X der Größe nach geordnet sind und zwar von der kleinsten zur größten, wodurch auch eine eindeutige Reihenfolge der Rangzahlen von Y festgelegt ist. Dann bestimmen wir für jede Rangzahl $R(y_i)$ die Anzahl der Rangzahlen von Y, die kleiner oder gleich $R(y_i)$ sind und in der Anordnung rechts von $R(y_i)$ stehen. Für $R(y_i)$ bezeichnen wir diese Anzahl mit Q_i. Dann ergibt sich Kendall's τ zu

$$\tau = 1 - \frac{4 \sum_{i=1}^{n} Q_i}{n(n-1)};$$

τ liegt wie r_{XY}, r_F und r_S zwischen -1 und 1.

Diese Vorgehensweise soll nun an einem Beispiel erläutert werden.

Beispiel: Es liegen die gleichen Daten wie beim letzten Beispiel vor, vgl. Tab. 25.

Wir ordnen zunächst die Ausprägungen von X (Klausurpunkte) der Größe nach und vergeben die Ränge. Das gleiche machen wir mit den Ausprägungen von Y (Übungspunkte) und ordnen dann die Merkmalspaare nach den Rangzahlen von X. Nun berechnet man die Q_i, vgl. **Tab. 26**.

Tab. 26: Arbeitstabelle zur Berechnung von Kendall's τ im Beispiel

Klausurpunkte (geordn.)	$x_{(1)} = 76$	$x_{(2)} = 65$	$x_{(3)} = 59$	$x_{(4)} = 58$	$x_{(5)} = 53$	$x_{(6)} = 44$	$x_{(7)} = 32$	$x_{(8)} = 29$	$x_{(9)} = 26$	$x_{(10)} = 25$
Rangzahl der Klau.-Punkte	1	2	3	4	5	6	7	8	9	10
zugehörige Übungspunkte	122 $= y_{(1)}$	89 $= y_{(4)}$	79 $= y_{(6)}$	59 $= y_{(9)}$	101 $= y_{(3)}$	67 $= y_{(8)}$	68 $= y_{(7)}$	83 $= y_{(5)}$	118 $= y_{(2)}$	42 $= y_{(10)}$
Rangzahl der Üb.-Punkte	1	4	6	9	3	8	7	5	2	10
$Q_1, \ldots, Q_{10}$	0	2	3	5	1	3	2	1	0	0

In Tab. 26 ist z. B. $Q_4 = 5$, weil rechts von der Rangzahl 9 von Y fünf Rangzahlen von Y stehen, die kleiner als 9 sind, nämlich 3, 8, 7, 5 und 2.

Es ist also nach Tab. 26 gerade

$$\tau = 1 - \frac{4 \sum_{i=1}^{10} Q_i}{n(n-1)} = 1 - \frac{4 \cdot 17}{10 \cdot 9} = 1 - \frac{68}{90} = 0{,}244,$$

d. h. die Abhängigkeit zwischen Klausurpunkten und Übungspunkten ist nicht sehr groß.

9.6. Der Yulesche Assoziationskoeffizient für die Vierfeldertafel

Es fehlt nun noch der Fall, daß (X, Y) nominal skaliert ist. Um ein Maß für den Zusammenhang von X und Y zu finden, benötigt man die Häufigkeitsverteilung von (X, Y).

Wir wollen uns hier darauf beschränken, daß X und Y jeweils genau 2 Ausprägungen a_1, a_2 bzw. b_1, b_2 haben können. Die Kontingenztafel nennt man in diesem Fall **Vierfeldertafel**, sie hat die Form der **Tab. 27**.

Tab. 27: Vierfeldertafel

X \ Y	b_1	b_2	$\sum$
a_1	H_{11}	H_{12}	$H_{1.}$
a_2	H_{21}	H_{22}	$H_{2.}$
$\sum$	$H_{.1}$	$H_{.2}$	n

Als Maß für die Abhängigkeit von X und Y kann man dann den **Yuleschen Assoziationskoeffizienten** verwenden, der durch

$$A_{XY} = \frac{H_{11}H_{22} - H_{12}H_{21}}{H_{11}H_{22} + H_{12}H_{21}} = \frac{h_{11}h_{22} - h_{12}h_{21}}{h_{11}h_{22} + h_{12}h_{21}}$$

definiert ist. Es ist $-1 \leqq A_{XY} \leqq 1$, und falls ein $H_{ij} = 0$ ist, so gilt: $|A_{XY}| = 1$.

Das Vorzeichen des Assoziationskoeffizienten ist nur in Verbindung mit der Vierfeldertafel interpretierbar.

Beispiel: 100 Personen (50 Frauen, 50 Männer) im Alter von 40–50 Jahren wurden befragt, ob sie Abitur haben oder nicht.

Dann kann man die Vierfeldertafeln aus **Tab. 28** und **29** verwenden:

Es ist

$$A_{XY} = \frac{7 \cdot 32 - 43 \cdot 18}{7 \cdot 32 + 43 \cdot 18} = -0{,}551 \quad \text{bei der Tafel aus Tab. 28 und}$$

$$A_{XY} = \frac{43 \cdot 18 - 7 \cdot 32}{43 \cdot 18 + 7 \cdot 32} = +0{,}551 \quad \text{bei der Tafel aus Tab. 29.}$$

Tab. 28: 1. Vierfeldertafel im Beispiel

	Frauen	Männer	$\sum$
Abitur	7	18	25
kein Abitur	43	32	75
$\sum$	50	50	100

Tab. 29: 2. Vierfeldertafel im Beispiel

	Männer	Frauen	$\sum$
Abitur	18	7	25
kein Abitur	32	43	75
$\sum$	50	50	100

Den ersten Assoziationskoeffizienten könnte man dann so interpretieren, daß es einen negativen Zusammenhang zwischen Frauen und Abitur gibt, und den zweiten so, daß es einen positiven Zusammenhang zwischen Männern und Abitur gibt.

Beispiel: Bei psychologischen Tests führt man oft denselben Test zweimal durch (**Retestmethode**), um den „Grad der Genauigkeit" zu bestimmen, mit dem ein Persönlichkeits- oder Verhaltensmerkmal „gemessen" wird. Wenn man die Versuchspersonen bzgl. eines nominal skalierten Merkmals mit Ausprägungen A und B untersucht, kann man mit Hilfe des Assoziationskoeffizienten ein Maß für diese Genauigkeit erhalten.

Im Idealfall, also wenn beide Tests die gleichen Ergebnisse liefern, ergibt sich die Vierfeldertafel aus **Tab. 30** und somit $A_{XY} = 1$.

Tab. 30: Vierfeldertafel der Retest-Methode im Idealfall

Test 2 \ Test 1	A	B
A	H_{11}	0
B	0	H_{22}

Auch wenn nur im Feld (A, B) eine Null steht, ergibt sich $A_{XY} = 1$. Der Yulesche Assoziationskoeffizient ist deshalb in diesem Fall kein gutes Maß für den Zusammenhang.

Weitere Zusammenhangsmaße für nominal skalierte Merkmale findet man in Kap. VII.

10. Praktische Berechnung einiger Kenngrößen

10.1. Berechnung des arithmetischen Mittels und der Standardabweichung

Bei der Berechnung des arithmetischen Mittels und der Standardabweichung gibt es verschiedene Möglichkeiten, sich die Rechnung zu erleichtern.

Bestehen die einzelnen Werte aus wenigen Ziffern oder steht ein Rechner zur Verfügung, so wird man das arithmetische Mittel in üblicher Weise aus

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

und die Standardabweichung aus

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

berechnen. Hat man $\bar{x}$ und s schon so aus n Elementen bestimmt und stellt dann fest, daß noch ein weiterer (n + 1)-ter Wert berücksichtigt werden muß, so sind folgende Regeln nützlich:

$$\bar{x}_{n+1} = \frac{1}{n+1}(x_{n+1} + n \cdot \bar{x}),$$

$$s_{n+1} = \sqrt{(n+1) \cdot (\bar{x}_{n+1} - \bar{x})^2 + \frac{n-1}{n} \cdot s^2}$$

$$= \sqrt{\frac{1}{n+1}(x_{n+1} - \bar{x})^2 + \frac{n-1}{n} \cdot s^2}.$$

Bei vielstelligen Einzelwerten kann man sich die Arbeit erleichtern, wenn man zunächst einen im Grunde genommen beliebigen Wert f von den Einzelwerten subtrahiert. Dieser Wert sollte möglichst so gewählt werden, daß alle Einzelwerte nach Subtraktion von f klein sind. Es ist dann

$$\bar{x} = f + \frac{1}{n}\sum_{i=1}^{n}(x_i - f) \quad \text{und}$$

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - f)^2 - n \cdot (\bar{x} - f)^2\right)}.$$

Beispiel: Wir wollen arithmetisches Mittel und Standardabweichung der in **Tab. 31** zusammengestellten Einzelwerte berechnen. Da alle Zahlen in der Größenordnung 15000 liegen, subtrahieren wir zunächst f = 15000 von den Einzelwerten.

Tab. 31: Werte zum Beispiel: $\bar{x}$ und s bei vielstelligen Einzelwerten

i	x_i	$x_i - f$
1	15146	146
2	15128	128
3	15033	33
4	15048	48
5	15117	117
6	15098	98
7	15057	57
8	15133	133
9	15029	29
10	15110	110

Es ist nun

$$\bar{x} = f + \frac{1}{n}\sum_{i=1}^{n}(x_i - f) = 15000 + \frac{1}{10}\sum_{i=1}^{10}(x_i - f) = 15000 + \frac{899}{10}$$
$$= 15089{,}9$$

und

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - f)^2 - n \cdot (\bar{x} - f)^2\right)} = \sqrt{\frac{1}{9}(98265 - 80820{,}1)}$$
$$= \sqrt{1938{,}3222}$$
$$= 44{,}026.$$

Sind die Einzelwerte so beschaffen, daß viele Stellen hinter dem Dezimalpunkt stehen, so kann man zunächst mit einer geeigneten Potenz von 10 (etwa $10^1 = 10$, $10^2 = 100$ oder $10^3 = 1000$) multiplizieren, denn ist

$$y_i = x_i \cdot 10^k, \quad \text{so ist}$$
$$\bar{x} = \bar{y} \cdot 10^{-k} \quad \text{und} \quad s = s_y \cdot 10^{-k}.$$

Möchte man gar von den Einzelwerten x_i zuerst eine Zahl subtrahieren und dann diese neue Zahl noch durch eine andere Zahl dividieren, möchte man sie also einer linearen Transformation

$$y_i = \frac{1}{k_1}(x_i - k_2)$$

unterziehen, um eine möglichst gute Größenordnung zum Rechnen zu erhalten, so kann man dies tun und dann zunächst $\bar{y}$ und s_y in gewohnter Weise berechnen. Für die ursprünglichen Werte x_i gilt dann

$$\bar{x} = k_1 \cdot \bar{y} + k_2 \quad \text{und} \quad s = k_1 \cdot s_y.$$

Ist die Anzahl der zu berücksichtigenden Einzelwerte sehr groß (etwa $n > 25$), so ist es zweckmäßig, die Daten in Klassen einzuteilen. Diese Klassen sollten nun möglichst gleich groß sein. Sind sie es nicht, so können wir $\bar{x}$ und s nach untenstehenden Formeln approximativ berechnen; vgl. Abschnitt 4.1 und 5.4.1:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{p} m_j \cdot H_j,$$
$$s = \sqrt{\frac{1}{n-1}\left(\sum_{j=1}^{p} m_j^2 \cdot H_j - n \cdot \bar{x}^2\right)}.$$

Dabei bezeichnet p die Anzahl der Klassen, m_j die Klassenmitte der j-ten Klasse und H_j die Anzahl der Einzelwerte in der j-ten Klasse.

Beispiel: Wir wollen arithmetisches Mittel $\bar{x}$ und Standardabweichung s der gruppierten Einzelwerte aus **Tab. 32** bestimmen.

Tab. 32: Gruppierte Einzelwerte zur Berechnung von $\bar{x}$ und s

j	Klassengrenzen	m_j	H_j
1	[0;1)	0,5	6
2	[1;3)	2,0	11
3	[3;4)	3,5	13
4	[4;7)	5,5	8
5	[7;8)	7,5	7
6	[8;10)	9,0	10