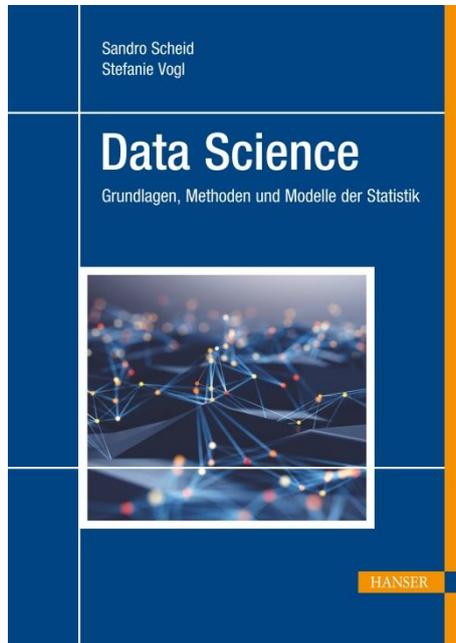


HANSER



Leseprobe

zu

Data Science

von Sandro Scheid and Stefanie Vogl

Print-ISBN: 978-3-446-46663-0

E-Book-ISBN: 978-3-446-47001-9

Weitere Informationen und Bestellungen unter

<https://www.hanser-kundencenter.de/fachbuch/artikel/9783446466630>

sowie im Buchhandel

© Carl Hanser Verlag, München

Inhalt

I	Deskriptive Statistik	13
1	Grundlagen	15
1.1	Aufgaben der deskriptiven Statistik	15
1.2	Grundgesamtheit und Stichprobe	16
1.3	Merkmale und Skalenniveaus	17
2	Betrachtung eines Merkmals	22
2.1	Häufigkeitsverteilungen bei diskreten Merkmalen	22
2.1.1	Absolute und relative Häufigkeitsverteilung	22
2.1.2	Graphische Darstellung	25
2.2	Häufigkeitsverteilungen bei stetigen Merkmalen	26
2.2.1	Prinzip der Klassenbildung	26
2.2.2	Histogramm	28
2.3	Statistische Maßzahlen	30
2.3.1	Lagemaße	30
2.3.2	Streuungsmaße	38
2.3.3	Formmaße	43
2.3.4	Box-Plots	44
3	Betrachtung zweier Merkmale	48
3.1	Kontingenztabelle	48
3.2	Bedingte Häufigkeiten	54
3.3	Kontingenzkoeffizient	59
3.3.1	Pearsons χ^2 -Statistik	59
3.3.2	Kontingenzmaß nach Cramer	62
3.3.3	Kontingenzkoeffizient nach Pearson	63
3.4	Streudiagramme	65
3.5	Korrelationsanalyse	66
3.5.1	Kovarianz	66
3.5.2	Korrelationskoeffizient nach Bravais-Pearson	71
3.5.3	Korrelationskoeffizient nach Spearman	75

3.6	Regressionsanalyse	80
3.6.1	Schätzung der Regressionskoeffizienten	81
3.6.2	Prognose	85
3.6.3	Güte der Anpassung	86

II Wahrscheinlichkeitstheorie 93

4 Das Rechnen mit Wahrscheinlichkeiten 95

4.1	Zufallsvorgänge und deren Beschreibung	95
4.2	Die Verknüpfung von Ereignissen	98
4.3	Die Axiome von Kolmogoroff	102
4.4	Die Laplace-Wahrscheinlichkeit	103
4.5	Statistische und subjektive Wahrscheinlichkeit	104
4.6	Rechenregeln für Wahrscheinlichkeiten	107
4.7	Bedingte Wahrscheinlichkeiten und Unabhängigkeit von Ereignissen	109
4.8	Totale Wahrscheinlichkeit	112
4.9	Der Satz von Bayes	114

5 Diskrete Zufallsvariable 118

5.1	Bedeutung und Definition einer diskreten Zufallsvariablen	118
5.2	Verteilung einer diskreten Zufallsvariablen	120
5.2.1	Wahrscheinlichkeitsfunktion	120
5.2.2	Verteilungsfunktion	125
5.3	Unabhängigkeit von diskreten Zufallsvariablen	129
5.3.1	Gemeinsame Verteilung von unabhängigen Zufallsvariablen.....	129
5.3.2	Rechnen mit Zufallsvariablen.....	131
5.4	Parameter von diskreten Zufallsvariablen	133
5.4.1	Erwartungswert	133
5.4.2	Varianz	138
5.5	Spezielle diskrete Verteilungen	143
5.5.1	Die Binomialverteilung	143
5.5.2	Die Poisson-Verteilung.....	147
5.5.3	Die hypergeometrische Verteilung	151

6 Stetige Zufallsvariable 154

6.1	Definition und Verteilung	154
6.2	Unabhängigkeit von stetigen Zufallsvariablen	160
6.3	Parameter von stetigen Zufallsvariablen	161

6.3.1	Erwartungswert stetiger Zufallsvariablen	161
6.3.2	Varianz stetiger Zufallsvariablen	163
6.3.3	Quantile stetiger Verteilungen	165
6.4	Die Normalverteilung	166
6.5	Sätze der Wahrscheinlichkeitsrechnung	173
6.5.1	Ungleichung von Tschebyscheff	173
6.5.2	Gesetz der großen Zahlen	175
6.5.3	Zentraler Grenzwertsatz	176
6.6	Prüfverteilungen	177
7	Gleichzeitige Betrachtung mehrerer Zufallsvariablen	181
7.1	Kovarianz – Korrelation	181
7.2	Lineare Funktionen von Zufallsvektoren	184
7.3	Die multivariate Normalverteilung	186
III	Induktive Statistik	189
8	Punktschätzung von Parametern	191
8.1	Der Begriff der Punktschätzung	192
8.2	Kriterien zur Güte einer Schätzung	196
8.2.1	Eigenschaften von Schätzfunktionen	196
8.2.2	Vergleich von Schätzfunktionen	201
8.2.3	Asymptotische Gütekriterien	202
8.3	Spezielle Schätzfunktionen	205
8.3.1	Schätzen von Anteilswerten	205
8.3.2	Schätzen von Mittelwerten	206
8.3.3	Schätzen der Varianz	207
8.3.4	ML-Schätzung	208
9	Intervallschätzung	210
9.1	Bedeutung des Konfidenzintervalls	210
9.2	Konfidenzintervalle für den Erwartungswert	212
9.2.1	Konfidenzintervall für μ bei bekanntem σ^2	212
9.2.2	Konfidenzintervall für μ bei unbekanntem σ^2	214
9.2.3	Approximatives Konfidenzintervall für μ	216
9.3	Konfidenzintervalle für eine Wahrscheinlichkeit	217

10	Das Prinzip eines statistischen Tests	220
10.1	Der Binomial-Test und Gaußtest	220
10.1.1	Binomial-Test	220
10.1.2	Gaußtest	225
10.2	Fehlentscheidungen	230
10.3	Gütefunktion	231
11	Spezielle Testverfahren	235
11.1	t -Tests (Lagetests)	235
11.1.1	Einfacher t -Test	235
11.1.2	Doppelter t -Test	237
11.1.3	t -Test für verbundene Stichproben	239
11.2	Einfaktorielle Varianzanalyse	242
11.3	Unabhängigkeitstest	246
11.4	Weiterführende Themen	248
11.4.1	Testen von Anteilswerten	248
11.4.1.1	Testen eines Anteilswerts mit Software	248
11.4.1.2	Testen auf Gleichheit von Anteilswerten mit Software.....	249
11.4.2	Umsetzung von Signifikanztests in gängiger Software	249
11.4.3	Tests zum Prüfen von Annahmen	250
11.4.3.1	Levene-Test	250
11.4.3.2	Kolmogorov-Smirnov-Test	250
11.4.4	Nichtparametrische Tests zur Lage.....	251
11.4.4.1	Mann-Whitney-U-Test	251
11.4.4.2	Wilcoxon-Vorzeichen-Rang-Test.....	252
IV	Angewandte Methoden der Datenanalyse	253
12	Regressionsanalyse	255
12.1	Lineare Einfachregression.....	255
12.1.1	Modellannahmen.....	256
12.1.2	Stochastische Eigenschaften der KQ-Schätzer	258
12.1.3	Konfidenzintervalle und Tests	260
12.2	Multiple lineare Regression	263
12.2.1	Das multiple lineare Regressionsmodell	263
12.2.2	Schätzen der Modellparameter	266
12.2.3	Streuungszerlegung und Bestimmtheitsmaß	269
12.2.4	Stochastische Eigenschaften der KQ-Schätzer	271
12.2.5	Konfidenzintervalle und Tests	272

13	Das Logit-Modell	276
	13.1 Formulierung des logistischen Regressionsmodells	276
	13.2 Schätzung des Modells	278
	13.3 Asymptotische Konfidenzintervalle und asymptotisches Testen einzelner Koeffizienten	280
	13.4 Asymptotisches Testen linearer Hypothesen	284
	13.5 Regressionsmodelle in der Anwendung	285
	13.5.1 Kategoriale Einflussgrößen	285
	13.5.2 Interaktion zweier Dummy-Variablen	288
	13.5.3 Modellierung nicht monotoner Einflüsse metrischer Größen	289
	13.5.4 Modellierung eines Polynoms in einer metrischen Einflussgröße	289
	13.5.5 Stückweise konstante Funktion	292
	13.5.6 Stückweise lineare Funktion	294
	13.5.7 Kubischer Spline mit Stützstellen	295
14	Diskriminanzanalyse	297
	14.1 Quadratische Diskriminanzanalyse	297
	14.2 Klassische Diskriminanzanalyse	300
	14.3 Bayesianische Diskriminanzanalyse	302
	14.4 Lineare Diskriminanzanalyse nach R. A. Fisher	305
	14.4.1 Aufgabenstellung	305
	14.4.2 Lösung des Maximierungsproblems	306
	14.4.3 Verallgemeinertes Eigenwertproblem	308
	14.4.4 Zusammenfassung und Illustration an einem Beispiel	308
	14.4.5 Güte der Diskriminanzfunktion	313
15	Clusteranalyse	316
	15.1 Hierarchische Verfahren	316
	15.2 Dendrogramm	322
	15.3 Partitionierende Verfahren	325
16	Neuronale Netze	330
	16.1 Grundidee Neuronaler Netze	330
	16.2 Mathematische Neuronen	330
	16.3 Das 3-layer Perzeptron	333
	16.4 Lernen mit dem Gradientenverfahren	336
	16.5 Modellierung mit Hilfe Neuronaler Netze – Regression	343
	16.6 Modellierung mit Hilfe Neuronaler Netze – Klassifikation	347

Literatur	353
Sachwortverzeichnis	355

Vorwort

In diesem Buch, das aus Vorlesungen für Studierende der Betriebswirtschaftslehre entstanden ist, werden wichtige Methoden der Datenanalyse vorgestellt. Statistische Verfahren werden stets dann benötigt und eingesetzt, wenn im Rahmen empirischer Fragestellungen Daten erhoben, dargestellt und analysiert werden sollen. In allen empirischen Wissenschaften – wir nennen beispielhaft die Wirtschaftswissenschaften – hat die Statistik daher eine große praktische Bedeutung. Zum Studium dieser Wissenschaften gehört deshalb auch eine intensive Beschäftigung mit Statistik.

Dieses Buch ist ein idealer Begleiter zu den Statistikvorlesungen des Bachelor-Studiums an Hochschulen und Universitäten und für das Nacharbeiten statistischer Themen im Masterstudium. Für Praktiker bietet es die Gelegenheit, sich im Selbststudium mit statistischen Fragestellungen zu befassen.

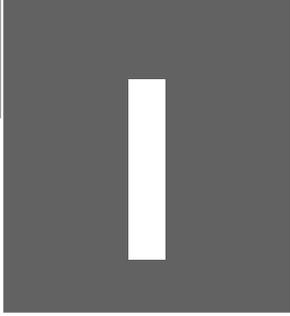
Das Buch setzt keine besonderen mathematischen Kenntnisse voraus. Der Leser benötigt die üblichen Grundkenntnisse der Elementarmathematik, wie sie an Fachoberschulen und Gymnasien unterrichtet wird. Die darüber hinausgehenden Anforderungen beschränken sich im Wesentlichen auf das Rechnen mit dem Summenzeichen.

Das Buch behandelt im ersten Teil (Kapitel 1 bis 3) die deskriptive Statistik, mit der die Datenanalyse beginnen sollte. Im zweiten Teil (Kapitel 4 bis 7) wird die Wahrscheinlichkeitsrechnung behandelt, die als Grundlage für die weiteren Kapitel benötigt wird. Teil drei behandelt die klassischen Themen der induktiven Statistik (Kapitel 8 bis 11). Danach werden im vierten Teil (Kapitel 12 bis 16) verschiedene weiterführende Methoden der Datenanalyse behandelt.

Zu beachten ist, dass die Rechnungen zu den komplexeren Beispielen teilweise mit Hilfe von Software durchgeführt wurden. Hier sind, um die Kumulation von Rundungsfehlern zu vermeiden, Zwischenergebnisse gerundet angegeben. Intern wurde aber mit einer höheren Genauigkeit weitergerechnet. Endergebnisse können sich also beim Rechnen mit gerundeten Zwischenergebnissen leicht unterscheiden.

An dieser Stelle wollen wir unseren Dank aussprechen. Unser Dank gilt zunächst den Studierenden der Hochschule München, die durch zahlreiche Fragen und Anmerkungen zur Gestaltung des Lehrtextes beigetragen haben. Weiter danken wir ganz herzlich Frau Julia Pelger, die durch ihre Korrekturvorschläge eine unersetzliche Arbeit geleistet hat, und Herrn Oliver Hien, durch dessen konstruktive Anregungen und Kommentare der Text weiter wesentlich verbessert werden konnte.

Unser ganz besonderer Dank gilt Herrn Frank Katzenmayer und Frau Christina Kubiak, Hanser Verlag, für die exzellente Betreuung des Buchprojekts.



Deskriptive Statistik

1

Grundlagen

Die deskriptive Statistik stellt Methoden bereit, um Untersuchungsgegenstände, die aus einer Vielzahl von einzelnen Objekten bestehen, zu beschreiben. Hinter dieser abstrakten Beschreibung stehen vielfältige unterschiedliche Anwendungen. Beispielsweise soll das Wahlverhalten der deutschen Bürger beschrieben werden oder die Nebenwirkungen bei der Einnahme eines Medikamentes oder das Interesse der Konsumenten an einem neuen Produkt. Die drei genannten Beispiele haben gemeinsam, dass das konkrete Interesse auf eine Gesamtheit bezogen ist. Dies sind die wahlberechtigten Personen in Deutschland, die Patienten, die das Medikament einnehmen oder der Kundenkreis des neuen Produkts. Diese Gesamtheiten sollen durch die statistische Analyse quantitativ beschrieben werden. Dazu werden an den Einheiten – in den genannten Beispielen sind dies Personen – Untersuchungsmerkmale betrachtet. Eine deskriptive Auswertung sollte das Untersuchungsziel und auch die Gesamtheit genauer beschreiben. Wichtige Begriffe, die dabei nützlich sind, werden in diesem Kapitel behandelt.

■ 1.1 Aufgaben der deskriptiven Statistik

Die deskriptive Statistik beschäftigt sich mit der Aufbereitung und Komprimierung von Daten, die in den verschiedensten Bereichen anfallen oder erhoben werden. Mithilfe der deskriptiven Statistik können Sachgebiete quantitativ beschrieben werden. Dabei werden viele Daten im Hinblick auf die Eigenschaften, die von Interesse sind, zusammengefasst. Zum Aufgabengebiet der deskriptiven Statistik zählen im Einzelnen:

- das Zusammenfassen und Ordnen der Daten in Tabellen,
- das Erstellen von Diagrammen und
- das Berechnen charakteristischer Kenngrößen oder Maßzahlen.

Beispiel 1.1

- Der Absatz der verschiedenen Modelle eines Kfz-Herstellers soll graphisch dargestellt werden.
- Die Einkommensstruktur der Bundesbürger soll komprimiert dargestellt werden. Von Interesse ist hier der Anteil der Bevölkerung an der Gesamtbevölkerung, der zu einer Einkommensklasse gehört, wobei das Einkommen in verschiedene Größenklassen aufgeteilt ist.
- Ein Kraftfahrzeugversicherer möchte die eingegangenen Schadensfälle verschiedenen Merkmalen zuordnen. So ordnet er sie dem Kraftfahrzeugtyp und dem Alter der Versicherungsnehmer zu.



Deskriptive Statistiken werden in unterschiedlichsten Bereichen benötigt. So stellen etwa die staatlichen Statistikämter unter anderem Statistiken der Bevölkerungsstruktur, der Erwerbstätigkeit, des produzierenden Gewerbes, der Sozialleistungen für Handel und Verkehr oder für das Gesundheitswesen zur Verfügung. Im Bereich Betriebsstatistik erfassen Unternehmen etwa Auftragseingänge, Produktion, Erzeugerpreise sowie Lohn- und Materialkosten. In der Forschung werden mittels Statistik Forschungsgegenstände wie z.B. die Verträglichkeit von Medikamenten, das Wahlverhalten, die Belastungsfähigkeit unterschiedlicher Materialien, das Kaufverhalten der Geschlechter oder das Armutsrisiko von Patchworkfamilien beschrieben.

■ 1.2 Grundgesamtheit und Stichprobe

Ausgangspunkt einer statistischen Fragestellung ist ein Untersuchungsgegenstand, der sich in der Regel auf viele einzelne Objekte bezieht. Dabei sollte der Untersuchungsgegenstand präzise formuliert und außerdem sachlich, räumlich und zeitlich abgegrenzt sein, damit eindeutig feststeht, auf welche konkreten Objekte er sich bezieht. Für die Vielzahl an möglichen Untersuchungsgegenständen sollen nun exemplarisch drei Fälle illustriert werden.

Beispiel 1.2

Eine Firma überlegt, ein neues Produkt einzuführen. Sie möchte wissen, wie es auf dem Markt ankommen würde. Um das herauszufinden, führt das Unternehmen eine Umfrage durch.

Untersuchungsgegenstand ist also die Akzeptanz eines neuen Produkts auf dem Markt. Dazu muss der potenzielle Absatzmarkt festgelegt werden. Welche geographische Größe ist relevant, ist der Markt national, europaweit oder global definiert? Sollen nur Ballungszentren bedient werden oder die ganze Fläche? Welches Geschlecht und welche Altersgruppen kommen infrage? Welche Einkommensklassen? Welche Vertriebswege sollen genutzt werden?



Beispiel 1.3

Ein Kfz-Versicherer möchte die Schadensfälle des abgelaufenen Geschäftsjahres nach Schadenshöhe und geografischer Häufigkeit charakterisieren.

Untersuchungsgegenstand ist die Gesamtheit der im abgelaufenen Geschäftsjahr eingegangenen Schadensfälle. Somit müssen Beginn und Ende des Geschäftsjahres durch konkrete Zeitpunkte markiert werden. Sollen sämtliche Versicherungstarife analysiert werden oder gilt das Interesse nur einer Auswahl?



Wie an den Beispielen deutlich wird, bezieht sich der Untersuchungsgegenstand stets auf eine Menge von Objekten. Diese werden in der Statistik als Untersuchungseinheiten bezeichnet. Die Menge aller Untersuchungseinheiten bildet dann die Grundgesamtheit.

Untersuchungseinheit

Untersuchungseinheiten sind die Objekte, auf die sich die statistische Analyse bezieht.

Grundgesamtheit

Die Menge aller Untersuchungseinheiten wird als Grundgesamtheit oder Population bezeichnet.

Meist werden in der Praxis nicht alle Untersuchungseinheiten, auf die sich eine Fragestellung bezieht, analysiert. Dies wäre aus organisatorischen und zeitlichen Gründen oft viel zu aufwendig oder sogar vollkommen unmöglich. Häufig ist das auch gar nicht notwendig. Die moderne Statistik stellt nämlich Methoden zur Verfügung, die es ermöglichen, basierend auf einer relativ kleinen Auswahl von Untersuchungseinheiten allgemein gültige Aussagen bezüglich einer weitaus größeren Grundgesamtheit herzuleiten.

Stichprobe

Eine Stichprobe bezeichnet eine Auswahl an Untersuchungseinheiten einer Grundgesamtheit.

Die Vorgehensweise, Ergebnisse einer Teilgesamtheit auf eine übergeordnete Gesamtheit zu verallgemeinern, ist Aufgabe der induktiven Statistik. Die deskriptive Statistik dient ausschließlich nur zur Beschreibung einer vollständig bekannten Grundgesamtheit oder zur Beschreibung einer Stichprobe. Interpretationen, die darüber hinausgehen, sind nicht Gegenstand der Untersuchungen und bleiben der induktiven Statistik vorbehalten.

■ 1.3 Merkmale und Skalenniveaus

Bei einer Untersuchung werden an den Untersuchungseinheiten Merkmale betrachtet, die dem jeweiligen Interesse entsprechen. Merkmale werden mit Großbuchstaben X, Y, Z, \dots notiert. Allgemein sind an einer statistischen Einheit in der Regel mehrere Merkmale beobachtbar.

Beispiel 1.4

Sind die Untersuchungseinheiten Personen, so könnten folgende Merkmale von Interesse sein:

X = das Monatseinkommen,

Y = die bei der letzten Bundestagswahl gewählte Partei,

Z = das Geschlecht.



Beispiel 1.5

Sind die Untersuchungseinheiten Schadensereignisse, die bei einer Versicherung innerhalb eines Zeitraums eingehen, so könnten folgende Merkmale von Interesse sein:

X = die Schadenshöhe,

Y = der Zeitpunkt des Schadens,

Z = die Versicherungshöhe des Versicherungsnehmers.



Neben dem Begriff Merkmal wird häufig auch der Begriff Variable verwendet. Beide Begriffe werden synonym angewandt. Der Begriff Variable deutet schon darauf hin, dass die Variable oder das Merkmal für die verschiedenen Untersuchungseinheiten verschiedene Werte annehmen kann. Die Werte, die die Merkmale annehmen, werden als Merkmalsausprägungen bezeichnet. Die möglichen Ausprägungen eines Merkmals bilden den Merkmals- oder Zustandsraum. Die folgende Übersicht enthält nochmals eine Zusammenfassung der beschriebenen Begriffe.

Merkmal

Unter einem Merkmal versteht man bestimmte Aspekte oder Eigenschaften einer Untersuchungseinheit.

Merkmalsausprägung

Den konkreten Wert, den eine Untersuchungseinheit für ein Merkmal annimmt, bezeichnet man als Merkmalsausprägung.

Merkmalsraum

Die Menge der möglichen Ausprägungen eines Merkmals wird mit Merkmals- oder Zustandsraum bezeichnet.

In der Literatur ist es üblich, folgende Begriffe synonym zu verwenden:

Untersuchungseinheit = Merkmalsträger = Objekt = Fall (Case),

Merkmal = Variable = Attribut.

Beispiel 1.6

Im *Beispiel 1.4* wurden Personen als Einheiten betrachtet. Die genannten Merkmale könnten die Werte

X = Monatseinkommen: 1 700 €, 2 100 €, 1 900 €, ...

Y = gewählte Partei: CDU, SPD, Grüne, FDP, ...

Z = Geschlecht: männlich, weiblich
annehmen.



Das Beispiel illustriert, dass die Art der Ausprägungen unterschiedlich sein kann. Im Folgenden sollen die verschiedenen möglichen Arten von Merkmalsausprägungen beschrieben werden. Je nach Art der Merkmalsausprägungen werden sich später die möglichen statistischen Methoden, die zur Auswertung der Daten herangezogen werden können, unterscheiden. Eine grobe Einteilung der Merkmalsausprägungen ergibt sich durch die Unterscheidung in zahlenmäßige und nicht zahlenmäßige Werte. Die verwendeten Begriffe sind:

Qualitative Merkmale

Die Ausprägungen unterscheiden sich in ihrer Art. Sie können nicht durch eine Größenordnung beschrieben werden, aber in verschiedene Kategorien eingeteilt werden.

Quantitative Merkmale

Die Ausprägungen lassen sich durch Zahlen beschreiben. Sie besitzen eine Ausprägung, bei der die Größe interpretiert werden kann.

Oft werden die Kategorien von qualitativen Merkmalen im praktischen Gebrauch mit Zahlen belegt. So stehen z. B. Steuer- oder Kundennummern für verschiedene Personen oder Steuerklassen für unterschiedliche Personengruppen. Bankleitzahlen stehen für Banken, Postleitzahlen für Orte. Wichtig für das Verständnis des Begriffs des qualitativen Merkmals ist hierbei, dass die jeweiligen Zahlen in diesen Fällen nur der Unterscheidung dienen. Eine Anordnung der Merkmale im Sinne einer Größe ist hier nicht sinnvoll.

Beispiel 1.7

- Qualitative Merkmale: Geschlecht, Familienstand, Konfession, Steuerklasse, Farbe eines Autos, Rechtsform eines Unternehmens.
- Quantitative Merkmale: Körpergröße, Alter, Einkommen, Umsatz, Temperatur.

Quantitative Merkmale lassen sich weiter unterteilen, je nachdem, ob die Merkmalsausprägungen nur diskrete oder kontinuierliche Werte annehmen können.

Diskrete Merkmale

Die Merkmalsausprägungen nehmen nur bestimmte, separate Zahlenwerte an. Es werden nur endlich viele oder abzählbar unendlich viele Werte angenommen.

Stetige Merkmale

Es können Werte aus einem reellen Zahlenintervall angenommen werden. Die Werte sind kontinuierlich. Man stellt sich vor, dass sie auf beliebig viele Nachkommastellen messbar sind.

Bemerkung:

- Alle qualitativen Merkmale sind trivialerweise diskret. Quantitative Merkmale sind dann diskret, wenn die Merkmalsausprägungen durch einen Zählvorgang ermittelt werden können.
- Merkmale, die sich sehr fein unterteilen lassen, aber im Prinzip diskrete Merkmale sind, werden auch quasi-stetige Merkmale genannt und zu den stetigen Merkmalen gezählt.

Beispiel 1.8

- Diskrete Merkmale: Anzahl der Semester eines Studierenden, die Anzahl der Angestellten in einem Betrieb, Anzahl der Fehltage eines Arbeitnehmers.
 - Stetige Merkmale: Körpergröße, Temperatur, Zeit, Benzinverbrauch.
 - Quasi-stetige Merkmale: Einkommen eines Haushalts, Umsatz einer Firma, Quadratmeterpreis einer Wohnung.
-

Je nach Art des betrachteten Merkmals können die Merkmalsausprägungen anhand verschiedener Skalen gemessen werden. Im Folgenden sind die verschiedenen Skalenniveaus, geordnet nach dem Informationsgehalt, beginnend beim niedrigsten, beschrieben.

Skalenniveaus

- **Nominales Skalenniveau**
Bei einem nominalen Skalenniveau lassen sich die verschiedenen Ausprägungen des Merkmals lediglich unterscheiden. Es gibt keine natürliche Anordnung der Merkmalsausprägungen. Man spricht von einem nominalen Merkmal.
- **Ordinales Skalenniveau**
Es gibt eine Rangfolge bzw. Ordnung innerhalb der Ausprägungen des Merkmals. Der Abstand zwischen den Ausprägungen ist aber nicht sinnvoll interpretierbar. Man spricht von einem ordinalen Merkmal.
- **Metrisches Skalenniveau**
Die Ausprägungen des Merkmals lassen sich der Größe nach anordnen. Zudem sind die Abstände der Ausprägungen interpretierbar. Man spricht von einem quantitativen bzw. metrischen Merkmal.

Beispiel 1.9 Skalen

- Beispiele für Merkmale, die auf einer nominalen Skala gemessen werden, sind: Geschlecht, Familienstand, Konfession, Augenfarbe, Rechtsform eines Unternehmens, Branche eines Unternehmens.
- Beispiele für Merkmale, die auf einer ordinalen Skala gemessen werden, sind: Leistungsbeurteilung, Bewertung bei einem Schönheitswettbewerb, Zufriedenheit mit einem Produkt mit beispielsweise den Ausprägungen: sehr zufrieden, zufrieden, unzufrieden, sehr unzufrieden.

- Beispiele für Merkmale, die auf einer metrischen Skala gemessen werden, sind: Körpergröße, Wartezeit auf den Bus, Gewinn eines Unternehmens, Einkommen eines Arbeitnehmers.

Für die metrischen Merkmale sind weitere Unterteilungen möglich.

Arten metrischer Skalen

- **Intervallskala**
Nur die Differenzen zwischen Ausprägungen können interpretiert werden.
- **Verhältnisskala**
Das Merkmal besitzt einen absoluten Nullpunkt. Als Folge davon ergibt eine Aussage wie „die eine Ausprägung ist doppelt so hoch wie die andere“ Sinn. Allgemein sind die Verhältnisse der verschiedenen Ausprägungen interpretierbar.

Beispiel 1.10 Metrische Skalen

- Ein Merkmal, das auf einer Intervallskala gemessen wird, ist die Temperatur. Abstände sind hier interpretierbar. Eine Aussage wie „heute ist es 3 Grad wärmer als gestern“ ist sinnvoll. Bei null Grad Celsius handelt es sich nicht um einen natürlichen Nullpunkt. Angenommen heute sind 6°C und gestern 3°C gemessen worden, so ist es nicht sinnvoll zu sagen, dass es heute doppelt so warm ist wie gestern.
- Ein Beispiel für ein Merkmal, das auf einer Verhältnisskala gemessen wird, ist das Einkommen eines Arbeitnehmers. Hier bietet es sich an, Verhältnisse zu bilden. Aussagen wie „Peter verdient doppelt so viel wie Dieter“ oder „Petra verdient 15% mehr als Susanne“ sind zulässig.

Sachwortverzeichnis

- χ^2 -Koeffizient, 59
 - Wertebereich, 60
- χ^2 -Verteilung, 177

- abhängige Variable, 277
- agglomeratives Verfahren, 316, 319
- Aktivierungsfunktion, 331
 - Sigmoid-Funktion, 338
 - Softmax-Aktivierung, 349
 - Sprungfunktion, 331
 - Tangens Hyperbolicus, 338
- Alternativhypothese, 222
- arithmetisches Mittel, 35
- Average Linkage Verfahren, 322

- Batch-Size, 340
- Bayes
 - Satz von, 114, 116
- bayesianische Diskriminanzanalyse, 303, 304
- Bernoulli-Experiment, 123
- Bernoulli-Variable, 123
- Bernoulli-Verteilung, 123
 - Erwartungswert, 135
 - Varianz, 141
- Bestimmtheitsmaß, 87
 - angepasstes-, 270
 - multiples-, 269
- Bias, 197
- Bias-Neuron, 334
- Binomialverteilung, 147
 - Additivität, 147
 - Erwartungswert, 146
 - Varianz, 146
 - Wahrscheinlichkeitsfunktion, 144
- Box-Plot, 45
 - modifizierter, 45

- Clusteranalyse, 316
- Complete Linkage Verfahren, 322

- Datentransformation
 - Min-Max Skalierung, 344
 - Standardisierung, 344
- Dendrogramm, 322, 324
- Dichte, 208
 - Normalverteilung, 166
- Dichtefunktion, 154
- Differenz, 99
- Diskriminanzanalyse, 297
- Diskriminanzfunktion
 - Güte, 313
- divisives Verfahren, 316
- Dummy-Kodierung, 285
- Dummy-Variable, 285, 292
- Durchschnitt, 99

- Early Stopping, 342
- effizient
 - asymptotisch, 280
- Einflussgrößen, 276, 277
 - kategoriale, 285
 - nicht monotone, 289
- Eintrittswahrscheinlichkeit, 276, 277
- Elementarereignis, 96
- Ereignis, 97
 - disjunktes, 100
 - sicheres, 98
 - unabhängiges, 111
 - unmögliches, 98
 - unvereinbares, 100
- Ereignisalgebra, 101
- Ergebnisraum, 96
- erwartungstreu
 - asymptotisch, 280
- Erwartungstreue, 196, 199
 - asymptotische, 204
 - beste Schätzung (UMVU), 202
- Erwartungswert
 - Bernoulli-Verteilung, 135
 - diskrete Zufallsvariable, 134

- geometrische Verteilung, 138
- Gleichverteilung, 162
- Rechenregeln, 135, 163
- stetige Zufallsvariable, 162
- Erwartungswertvektor, 183
- F-Test
- Overall-, 274
- F-Verteilung, 179
- Fehlerfunktion
- absoluter Fehler, 337
- Log-Loss, 349
- quadratische Fehlerfunktion, 336
- Fisher-Informationsmatrix, 280
- furthest neighbour, 322
- Generalisierungsmenge, 342
- geometrische Verteilung
- Erwartungswert, 138
- Varianz, 142
- Gesetz der großen Zahlen, 105, 175
- Gleichverteilung
- diskrete, 123
- stetige, 157
- Gradientenabstieg, 337
- Grundgesamtheit, 17
- Gütefunktion, 231
- Häufigkeiten
- absolute, 23, 27
- bedingte, 56
- gemeinsame, 50
- relative, 23, 27
- Häufigkeitsinterpretation, 211
- Hidden-Schicht, 334
- Histogramm, 29
- Hit-Rate, 351
- hypergeometrische Verteilung, 151
- Erwartungswert, 153
- Varianz, 153
- Hypothesen, 282
- Input-Schicht, 334
- Interaktion, 288, 289
- Intervallschätzung, 210
- Intervallskala, 21
- Irrtumswahrscheinlichkeit, 210
- k-means, 325, 328
- kanonischer Korrelationskoeffizient, 315
- kategoriale Einflussgrößen, 286
- Klasse, 27
- Klassenbreite, 27
- Klassengrenzen, 27
- klassische Diskriminanzanalyse, 302
- Kleinste-Quadrate-Schätzer, 83
- multiple Regression-, 267
- Koeffizienten, 276
- Komplement, 99
- Konfidenzgrenze, 210
- Konfidenzintervall, 281
- μ bei bekanntem σ^2 , 213
- μ bei unbekanntem σ^2 , 215
- Anteilswert, 218
- approximatives, 217
- Logit-Modell, 281
- zweiseitiges, 210
- Konfidenzniveau, 210
- Konfusionsmatrix, 351
- Konsistenz
- quadratisches Mittel, 205
- schwache, 204
- Kontingenzkoeffizient
- korrigierter nach Pearson, 63
- nach Pearson, 63
- Kontingenzmaß
- nach Cramer, 62
- Kontingenztabelle, 51, 52
- Korrelation, 181
- Korrelationskoeffizient
- nach Bravais-Pearson, 71
- nach Spearman, 76
- Kovarianz, 67, 181
- Kovarianzmatrix, 183, 271
- Schätzung-, 271
- KQ-Methode, 81
- KQ-Schätzer, 83
- erwartungstreu-, 271
- Erwartungswert-, 259
- Konfidenzintervalle-, 261, 272
- konsistent-, 260, 271
- Standardfehler-, 259
- Varianz-, 259
- Verteilung-, 260
- Kreisdiagramm, 25

- kritischer Bereich, 282, 285
kubischer Spline, 295
- Lernepoche, 340
Lernverfahren
– Batch Learning, 340
– Pattern by Pattern Learning, 339
Likelihood-Funktion, 208
Likelihood-Quotienten-Test, 284
lineare Diskriminanzanalyse, 305
lineare Diskriminanzfunktion
– Bestimmung, 308
– Zuordnung, 306
linearer Prädiktor, 276
log-Likelihood, 209
logistisches Regressionsmodell, 276, 277
Logit-Modell, 276, 277
– asymptotische Eigenschaften, 280
– Informationsmatrix, 280
– log-Likelihood, 279
– Scorefunktion, 279
– Test, 281
– Test einzelner Koeffizienten, 282
– Testen linearer Hypothesen, 284
- Maximum-Likelihood-Schätzung, 208, 209
Mean Square Error (MSE), 198, 199, 201
Median, 33
– bei Urliste, 33
Merkmal, 18
– diskretes, 19
– kategoriales-, 31
– qualitatives, 19
– quantitatives, 19
– quasi-stetiges, 20
– stetiges, 19
Merkmalsausprägung, 18
Merkmalsraum, 18
Meta-Parameter, 340
Metrik, 321
Modus, 31
Multiplikationssatz, 111
– für unabhängige Ereignisse, 112
- Neuron, 332
Neuronales Netz
– 3-Layer Perzeptron, 333
– Feedforward Netz, 343
– Rekurrentes Netz, 343
– tiefes Neuronales Netz, 342
Nominalskala, 20
Normalgleichungen, 82
– multiple Regression-, 267
normalverteilt
– asymptotisch, 280
Normalverteilung, 166
– Erwartungswert, 167
– lineare Transformation, 168
– multivariat, 186
– Quantile, 171
– Standardisierung, 170
– Symmetrie, 168
– Varianz, 167
Nullhypothese, 222
- One-Hot-Kodierung, 344, 348
Ordinalskala, 20
Output-Schicht, 334
Overfitting, 341
- Parameter, 193
Parameterraum, 193
partitionierende Verfahren, 325
Perzentile, 34
Poisson
– Grenzwertsatz, 148
Poisson-Verteilung
– Additionssatz, 150
– Erwartungswert, 150
– Varianz, 150
– Wahrscheinlichkeitsfunktion, 148
Polynom, 289, 290
Potenzmenge, 101
Prognose, 85
- quadratische Diskriminanzanalyse, 299
Quantile, 34, 165
Quartile, 34
Quartilsabstand, 39
- Randhäufigkeiten
– absolute, 50
Randverteilung
– absolute, 51
– relativ, 52

- Rang, 75
- Realisierungen, 119
- Rechenregeln
 - stetige Zufallsvariable, 159
- Referenzkategorie, 285
- Regression
 - lineare Einfach-, 80
- Regressionsmodell
 - lineares einfaches-, 256
 - multiples-, 265
- Residualplot, 90
- Restriktion, 285
- Root Mean Squared Error, 346

- Satz von Bernoulli, 176
- Satz von Gosset, 179
- Säulendiagramm, 25
- Scatterplot, 345
- Schätzfolge, 203
- Schätzfunktion, 195
- Schätzung, 195
 - Anteilswerte, 206
 - Mittelwerte, 206
 - Varianz, 207
- Schiefe, 43
- Scoretest, 284
- Signifikanz, 287, 289
- Signifikanztest, 224
- Single Linkage Verfahren, 322
- Skalenniveau
 - metrisches, 20
 - nominales, 20
 - ordinales, 20
- Spannweite, 38
- Spline, 296
- Standardabweichung
 - diskrete Zufallsvariable, 139
 - empirische, 39
- Standardfehler, 200
- Standardnormalverteilung, 170
- Stichprobe, 17, 194
- Stichprobenvarianz, 178
 - Verteilung, 178
- Strecken zug, 294
- Streudiagramm, 65
- Streuungszerlegung, 87

- stückweise konstante Funktion, 292
- stückweise lineare Funktion, 294

- t-Verteilung, 178
- Teilmodell, 284
- Test
 - Ablehnbereich-, 262, 273
 - Ablehnungsbereich, 224
 - Annahmehbereich, 224
 - Binomial, 220
 - doppelter t -Test, 238
 - einfacher t -Test, 236
 - einseitiger, 226
 - Fehlentscheidung, 231
 - Gauß, 225, 230
 - Gleichheit mehrerer Erwartungswerte, 244
 - Gütefunktion, 231
 - Interpretation Ergebnisse, 225
 - kritischer Bereich, 224, 262, 273
 - Modellannahmen, 222
 - parametrischer, 222
 - Prüfgröße, 223
 - Unabhängigkeitstest χ^2 , 247
 - verbundene Stichprobe, 240, 241
 - zweiseitiger, 226
- Testen linearer Hypothesen, 284
- Teststatistik, 282, 284
- Training, 336
- Trainingsmenge, 342

- Überwachtes Lernen, 330
- Unabhängigkeit
 - diskrete Zufallsvariable, 129, 130
 - empirische, 58
 - stetige Zufallsvariable, 160
- Ungleichung von Tschebycheff, 174
- Untersuchungseinheit, 16

- Validierungsmenge, 342
- Varianz
 - Bernoulli-Verteilung, 141
 - diskrete Zufallsvariable, 139
 - empirische, 39
 - externe, 42
 - geometrische Verteilung, 142
 - Gleichverteilung, 164
 - interne, 42

- Rechenregeln, 142, 164
- stetiger Zufallsvariablen, 163
- Verschiebesatz, 41, 141
- Varianz der Störvariablen
 - Einfachregression-, 258
 - multiple Regression-, 267
- Variationskoeffizient, 43
- Venn-Diagramm
 - Differenz, 100
 - Durchschnitt, 99
 - Komplement, 100
 - Vereinigung, 99
- Vereinigung, 98
- Verhältnisskala, 21
- Verschiebungssatz, 181
- Verteilung
 - geometrische, 124
- Verteilungsannahme
 - parametrische, 193
- Verteilungsfunktion
 - Bernoulli-Variable, 126
 - Eigenschaften, 125
 - einer diskreten Zufallsvariablen, 125
 - geometrische Verteilung, 127
 - Normalverteilung, 167
- Verteilungsfunktion
 - stetige Zufallsvariable, 158
- Wahrscheinlichkeit
 - a-posteriori, 117
 - a-priori, 117
 - bedingte, 109
 - nach Laplace, 104
 - statistische, 105
 - subjektive, 106
 - totale, 114
- Wahrscheinlichkeitsfunktion, 120, 208
- Wahrscheinlichkeitsrechnung
 - Additionssatz, 107
 - Axiome, 102
 - Multiplikationssatz, 111
 - Rechenregeln, 107
- Wald-Test, 284
- Wechselwirkung, 288
- Wilks-Lambda, 315
- Wölbung, 44
- zentraler Grenzwertsatz, 176
- Zentroid Distanz, 322
- Zentroid Verfahren, 322
- Zerlegung
 - disjunkte, 112
- Zielvariable, 277
- Zufallsgröße, 194
- Zufallsvariable
 - Dichtefunktion, 154
 - diskrete, 119
 - Erwartungswert, 162
 - stetige, 156, 157
- Zufallsvariablen
 - multivariate, 181
- Zufallsvektor, 183
 - lineare Funktionen, 184
- Zufallsvorgang, 95