Deskriptive Statistik und explorative Datenanalyse

2

Zusammenfassung

Die statistische Datenanalyse ist heute eine Kernaufgabe im aktuariellen Umfeld. Die Arbeit mit zum Teil sehr großen Datenmengen und der Einsatz spezieller Software zur Datenanalyse sind im beruflichen Alltag eines Aktuars zu Grundkompetenzen geworden. Mittels deskriptiver und explorativer Verfahren werden Datensätze systematisch untersucht, durch Kennzahlen beschrieben und durch grafische Darstellungen charakterisiert. Die Methoden der deskriptiven Statistik und der explorativen Datenanalyse stehen oft am Beginn von weiterführenden, induktiven Verfahren, wie z. B. der statistischen Modellbildung. Deskriptive und explorative Verfahren der Statistik sind in der Regel der erste Schritt, um einen Datensatz zu beschreiben und inhaltlich kennenzulernen. Diese Methoden werden aber auch unterstützend innerhalb von induktiven statistischen Verfahren verwendet. Am Ende einer statistischen Modellbildung steht z. B. in der Regel die Überprüfung der Modellvoraussetzungen und die Beurteilung der Modellgüte, wobei oft wieder deskriptive und explorative Verfahren zum Einsatz kommen. Ein wichtiger Grund für die heute weit verbreitete Anwendung von deskriptiver Statistik und explorativer Datenanalyse sind sicher die damit einhergehenden, großen Entwicklungen in der Datenverarbeitung, in der Datenverfügbarkeit und bei statistischen Analysesoftwaresystemen.

2.1 Grundlagen

In diesem Abschnitt werden grundlegende Begriffe und Vorgehensweisen, die in der angewandten Statistik verwendet werden, vorgestellt. Die angewandte Statistik erweitert die mathematische Statistik vor allem im Hinblick auf die praktische Durchführung von statistischen Untersuchungen. Im Folgenden soll dem Leser der Grundwortschatz der angewandten Statistik nahegebracht werden. Der für die Statistik zentrale Begriff der

T. Becker et al., *Stochastische Risikomodellierung und statistische Methoden*,
Statistik und ihre Anwendungen, DOI 10.1007/978-3-662-49407-3_2

Stichprobe wird sowohl in seiner Bedeutung in der angewandten Datenanalyse als auch in der für die mathematische Statistik und Wahrscheinlichkeitstheorie typischen Definition eingeführt.

Zu dem Themenbereich Statistik (angewandte Statistik und mathematische Statistik) und Datenanalyse gibt es umfangreiche Literatur. Die Bandbreite der Literatur geht von Lehrbüchern mit eher theoretischem Hintergrund bis zu ganz pragmatischen Beschreibungen von praktischen Analysefällen. Letzere findet man oft im Kontext von Statistik-Softwarepaketen und können für die praktische Datenanalyse sehr hilfreich sein. Ausführliche Darstellungen zur deskriptiven Statistik und explorativen Datenanalyse findet man z. B., eher einführend, bei Fahrmeir et al. [3] und Pruscha [7]. Einen sehr ausführlichen Überblick über angewandte statistische Methoden geben z. B. Sachs und Hedderich [8] oder Hartung et al. [5].

2.1.1 Grundaufgaben der Statistik

Eine immer noch zeitgemäße Definition von Statistik geht auf Abraham Wald (1902–1950) zurück: **Statistik** ist eine Zusammenfassung von Methoden, die uns erlauben, vernünftige optimale Entscheidungen im Falle von Ungewissheit zu treffen. Die Grundlage jeder praktischen, statistischen Analyse sind Daten (man sagt auch Stichprobe, Messreihe etc.), aus denen Erkenntnisse über einen stochastischen Vorgang abgeleitet werden sollen.

Die **deskriptive Statistik** stellt Methoden bereit, mit denen grundlegende Eigenschaften eines Datensatzes beschrieben werden können. Dazu verwendet der Statistiker genormte Maßzahlen, z. B. das arithmetische Mittel für die zentrale Lage und die emprische Standardabweichung für die Streuung eines Datensatzes. Zusätzlich kommen die Daten charakterisierende, grafische Darstellungsformen, wie z. B. Histogramme, zum Einsatz. Die deskriptive Statistik legt ihren Fokus auf einen vorliegenden Datensatz und es werden keine Aussagen bzgl. Kennzahlen, Gesetzmäßigkeiten, Zusammenhänge etc. über den speziellen Datensatz hinaus postuliert.

Die **explorative Datenanalyse**, vgl. Tukey [10], geht über die reine Beschreibung von Daten hinaus, hin zu einer Suche von Auffälligkeiten in einem Datensatz. Die explorative Statistik trifft, wie auch die deskriptive Statistik, im Allgemeinen nur Aussagen zu einem vorliegenden Datensatz. Die Exploration der Daten gibt dem Anwender aber wichtige Impulse für die Formulierung von Hypothesen und Fragestellungen, die auch über den vorliegenden Datensatz hinaus interessieren. Innerhalb der explorativen Datenanalyse gibt es eine Vielzahl von grafischen Methoden. Ein bekanntes Beispiel für eine explorative Datenvisualisierung ist der Box-Whisker-Plot.

Oft sind explorative Verfahren, insbesondere bei großen Datensätzen, sehr rechenintensiv. Die weite Verbreitung der explorativen Verfahren und ihre vielfältige Weiterentwicklung in den letzten Jahren geht stark einher mit der sich parallel dazu schnell entwickelnden Computer- und Softwaretechnologie. So hat sich etwa die Visualisierung von Daten zu einem eigenen Gebiet der Statistik bzw. der Informatik entwickelt.

Neben der Deskription und Exploration von Daten gehört zu den Grundaufgaben der Statistik noch die **induktive Statistik**. In der induktiven Statistik werden, basierend auf Ergebnissen der Wahrscheinlichkeitstheorie und mathematischen Statistik, über den vorliegenden Datensatz hinaus probabilistisch-bewertbare Aussagen getroffen. Induktive Verfahren sind z. B. statistische Signifikanztests oder auch die statistische Modellbildung.

In einer fortgeschrittenen, statistischen Analyse werden meist alle drei Grundaufgaben der Statistik angewendet. Eine fundierte, **statistische Arbeitsweise** zeichnet sich durch den folgenden Ablauf einer Analyse aus:

- **1. Schritt:** Am Beginn jeder statistischen Untersuchung steht immer eine deskriptive und explorative Analyse der Daten. Der Anwender verschafft sich so einen Überblick über den Datensatz. In diesem Analyseschritt können fehlerhafte oder fehlende Daten entdeckt, entfernt oder auch ersetzt werden.
- **2. Schritt:** Explorative Verfahren zeigen mögliche Hypothesen und Modellierungsansätze für eine weiterführende Analyse.
- **3. Schritt:** Die formulierten Hypothesen werden mit den Methoden der induktiven Statistik überprüft. Der zu untersuchende Zufallsvorgang wird durch eine statistische Modellbildung beschrieben und analysiert. Es werden Aussagen über den speziellen, vorliegenden Datensatz hinaus getroffen.
- **4. Schritt:** Am Ende der Analysen steht oft nochmals eine Bewertung der wahrscheinlichkeitstheoretischen Voraussetzungen der verwendeten induktiven Methoden. So findet z. B. im Allgemeinen nach der Entwicklung eines Regressionsmodells die Überprüfung der Voraussetzungen des statistischen Modells, die für die induktiven Verfahren innerhalb der Modellbildung (z. B. statistische Signifikanztests) notwendig sind, statt. Dazu verwendet man dann oft wieder Verfahren der deskriptiven und explorativen Statistik.

Die Abgrenzung zwischen deskriptiven, explorativen und induktiven Verfahren ist in der Literatur nicht immer scharf vollzogen und so wird manchmal auch die explorative Statistik als ein Teil der deskriptiven Statistik betrachtet. Manche explorativen Analysen nähern sich zudem stark der induktiven Statistik an, indem die verwendeten Konzepte zum Teil auf einem erheblichen wahrscheinlichkeitstheoretischen Hintergrund basieren. Weiterhin beachte man, dass viele der Maßzahlen, die in der deskriptiven Statistik verwendet werden, innerhalb der induktiven Statistik als Punktschätzer für Verteilungsparameter Verwendung finden.

2.1.2 Grundgesamtheiten und Stichproben

Im Folgenden werden die für die angewandte Statistik zentralen Begriffe der Grundgesamtheit und der Stichprobe definiert. Die Festlegung bzw. klare Abgrenzung der Grundgesamtheit einer statistischen Untersuchung ist der erste Schritt bei einer Datenerhebung und die Grundlage für die spätere Bewertung der Untersuchungsergebnisse.

Wir werden die Begriffe Grundgesamtheit und Stichprobe zunächst aus dem Blickwinkel der angewandten Statistik definieren, der meist in der praktischen statistischen Arbeit vorliegt. Nachfolgend wird der Stichprobenbegriff in der Sichtweise der mathematischen Statistik ergänzt. Diese Betrachtung einer Stichprobe ist vor allem für das Verständnis von induktiven Verfahren grundlegend.

Definition 2.1 *Die Menge G aller möglichen (Untersuchungs-)Einheiten (man sagt auch Individuen oder Fälle), die einer statistischen Untersuchung zugrundeliegen und von Interesse sind, nennt man die* ***Grundgesamtheit*** *einer statistischen Untersuchung.*

Man unterscheidet prinzipiell zwei Fälle von Grundgesamtheiten. Zum einen den Fall einer endlichen Grundgesamtheit, die eine endliche Menge realer Objekte (Einheiten) darstellt. Bei Datenerhebungen, wie z. B. Umfragen, ist dieser Typ einer Grundgesamtheit gegeben. Zum anderen gibt es die Situation einer unendlichen Grundgesamtheit, die hypothetische Objekte (Einheiten) enthält. In diesem Fall wird der datengenerierende Prozess als sich wiederholende Realisationen von Zufallsvariablen betrachtet. Dieser Betrachtung folgt man im Allgemeinen innerhalb der induktiven Statistik.

Für Datenerhebungen ist eine klare Festlegung der für die Untersuchung relevanten, endlichen Grundgesamtheit notwendig. So muss z. B. für eine Erhebung unter den Kunden eines Unternehmens (d. h. die Grundgesamtheit sollen alle Kunden des Unternehmens sein) klar definiert werden, wen man als Kunde des Unternehmens betrachtet. Sind z. B. in einem Versicherungsunternehmen nur alle Versicherungsnehmer Kunden oder auch alle versicherten Personen?

Definition 2.2 *Jede endliche Teilmenge $S \subset G$, die aus einer Grundgesamtheit G ausgewählt wird, heißt* ***Stichprobe*** *von G. Die Mächtigkeit $|S| = n$, $n \in \mathbb{N}$, nennt man den* ***(Stichproben-)Umfang*** *von S. Man nennt eine Stichprobe vom Umfang n eine* ***einfache Zufallsstichprobe****, falls durch die Auswahlmethodik sichergestellt ist, dass die Wahrscheinlichkeit für alle $S \subset G$ mit $|S| = n$ als Stichprobe ausgewählt zu werden, identisch ist.*

Die zufällige Auswahl einer Stichprobe aus der Grundgesamtheit ist ein Grundprinzip der Statistik. Die Zufälligkeit der Stichprobe ermöglicht einen Rückschluss von den Gegebenheiten der Stichprobe auf die Gegebenheiten der Grundgesamtheit. Innerhalb der statistischen Versuchsplanung spricht man in diesem Zusammenhang von **Randomisierung**.

Bei der praktischen Durchführung von Zufallsauswahlen muß streng darauf geachtet werden, dass die Auswahl wirklich zufällig erfolgt. Bei einer nicht zufälligen Auswahlmethodik droht die Gefahr eines sogenannten **Stichproben-Bias**, einem methodischen Fehler in einer statistischen Untersuchung, der im weiteren Verlauf der Untersuchung in der Regel nicht mehr korrigiert werden kann.

In der angewandten Statistik sind Versuchsplanung und Datenerhebung wichtige Teilbereiche der statistischen Analysearbeit. In dem vorliegenden Text werden diese Themen nicht weiter vertieft und der Leser sei dazu auf ergänzende Literatur, wie z. B. einführend Fahrmeir et al. [3], Kapitel 1, verwiesen.

Es folgt die Definition des Stichproben-Begriffs, die in der mathematischen Statistik verwendet wird. Hier werden die Stichprobenwerte als Realisationen von Zufallsvariablen identifiziert. Damit ist eine Verbindung von der eher praxisorientierten reinen Datensicht mit einer wahrscheinlichkeitstheoretischen Betrachtungsweise gegeben.

Definition 2.3 *Jede Realisation*

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$$

eines Zufallsvektors

$$\mathbf{X} = (X_1, \ldots, X_n)^\top,$$

der auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ *definiert ist, heißt* ***Stichprobe*** *vom Umfang* n*. D. h. man betrachtet die Realisationen*

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$$

der Zufallsvariablen $X_1, \ldots, X_n$ *als Stichprobenwerte. Der Zufallsvektor* $\mathbf{X}$ *wird auch als Zufallsstichprobe bezeichnet. Die der Stichprobe zugrundeliegenden Zufallsvariablen* $X_1, \ldots, X_n$ *werden auch* ***Stichprobenvariablen*** *genannt.*

Entsprechend ist die Folge von Stichprobenwerten $\{x_i\}_{i \in \mathbb{N}}$ *als Realisation einer Folge von Stichprobenvariablen* $\{X_i\}_{i \in \mathbb{N}}$ *definiert.*

Man beachte, dass bei der Definition 2.3 die Stichprobe ein n-Tupel von reellen Zahlen bezeichnet und in der Definition 2.2 die Stichprobe eine Menge von Untersuchungseinheiten darstellt. Der Stichprobenbegriff in Definition 2.3 bezeichnet also die Werte der in einer Untersuchung betrachteten Messgröße, die an den ausgewählten Untersuchungseinheiten gemessen wurden. Die Zufallsvariablen $X_1, \ldots, X_n$ repräsentieren im Allgemeinen die immer gleiche Messgröße, die in der Untersuchung von Interesse ist und wiederholt n-mal gemessen wurde.

Es wird häufig der Fall betrachtet, dass die Zufallsvariablen $X_1, \ldots, X_n$ in dem Zufallsvektor $\mathbf{X}$ unabhängig und identisch wie eine Zufallsvariable X_0 verteilt sind. Man betrachtet also n unabhängige Versionen einer Zufallsvariablen X_0. Im Folgenden werden wir diesen wichtigen Spezialfall einer Stichprobe als **i. i. d. Stichprobenvariablen** X_i**,** $i \geq 1$**,** bezeichnen. Die Abkürzung i. i. d. steht hier für independent and identically distributed.

Eine Hauptaufgabe der induktiven Statistik ist es, auf Basis der wiederholten Realisationen von X_0 (d. h. auf Basis einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$) Aussagen über unbekannte Paramter der Verteilung von X_0 zu treffen.

Definition 2.4 *Eine **Stichprobe***

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top,$$

$n \in \mathbb{N}$, *heißt **unabhängig**, falls die zugrundeliegenden Zufallsvariablen, d. h. die Stichprobenvariablen*

$$X_1, \ldots, X_n$$

stochastisch unabhängig sind. Zwei Stichproben

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top \text{ und } \boldsymbol{y} = (y_1, \ldots, y_m)^\top,$$

$n, m \in \mathbb{N}$, *nennt man unabhängig, falls die zugehörigen Stichprobenvariablen*

$$X_1, \ldots, X_n, Y_1, \ldots, Y_m$$

stochastisch unabhängig sind. Ganz analog wird die Unabhängigkeit von $r > 2$ Stichproben definiert. Eine Stichproben-Folge $\{x_i\}_{i\in\mathbb{N}}$ nennt man unabhängig, falls die zugehörige Folge der Stichprobenvariablen $\{X_i\}_{i\in\mathbb{N}}$ unabhängig ist.

Die bisher betrachteten Stichproben beinhalten immer nur Werte einer Messgröße, man spricht daher auch von **univariaten Stichproben**. Werden mehrere, $p > 1$ Messgrößen an einer Untersuchungseinheit erhoben, gelangt man zu dem Begriff der multivariaten ($p-$variaten) Stichprobe.

Definition 2.5 *Man nennt die p-Tupel*

$$\left(x_{11}, \ldots, x_{1p}\right)^\top, \ldots, \left(x_{n1}, \ldots, x_{np}\right)^\top,$$

$p \in \mathbb{N}$, $p > 1$, ***p-variate Stichprobe** vom Umfang n, falls $\left(x_{i1}, \ldots, x_{ip}\right)^\top$ für jedes $1 \le i \le n$ die Realisation eines Zufallsvektors $\left(X_{i1}, \ldots, X_{ip}\right)^\top$ ist. Für $p = 2$ erhält man eine **bivariate Stichprobe***

$$(x_{11}, x_{12})^\top, (x_{21}, x_{22})^\top, \ldots, (x_{n1}, x_{n2})^\top.$$

Eine p-variate Stichprobe vom Umfang n entspricht einer Datensituation, in der bei n Untersuchungseinheiten an jeder Einheit jeweils p Messgrößen erfasst werden. In diesem Sinn repräsentiert die Zufallsvariable X_{ij}, $1 \le i \le n$, $1 \le j \le p$, die $j-$te Messgröße gemessen an der $i-$ten Einheit. Ein wichtiger Spezialfall ist hier die Situation, dass die Zufallsvariablen $X_{i1}, \ldots, X_{ip}$ für jedes $i \in \{1, \ldots, n\}$ stochastisch abhängig sind, während die Zufallsvariablen $X_{1j}, \ldots, X_{nj}$ für jedes $j \in \{1, \ldots, p\}$ stochastisch unabhängig sind.

Beispiel 2.6 Von 1000 Versicherungsnehmern ist jeweils das Alter a_i und die Schadensumme s_i, $i = 1\ldots,1000$, erfasst. Die Daten bilden eine bivariate Stichprobe $(a_1, s_1)^\top, \ldots, (a_{1000}, s_{1000})^\top$. Dabei sind Alter und Schadenhöhe im Allgemeinen nicht unabhängig. □

Als Realisationen von Zufallsvariablen sind Stichprobenwerte $x_i, i = 1\ldots,n$, zunächst immer reelle Zahlen. Für Messgrößen in einer statistischen Untersuchung mit anderen Messskalen, z. B. Klassenbezeichnungen, werden dann die Stichprobenwerte durch reelle Zahlen repräsentiert. So können z. B. Klassenbezeichnungen über die Kombination von dichotomen Stichprobenvariablen, d. h. Zufallsvariablen mit der Wertemenge {0,1}, dargestellt werden.

2.1.3 Merkmale und Skalenniveaus

In diesem Abschnitt wenden wir uns wieder stärker den Sprachregelungen in der angewandten Statistik zu. Die in einer statistischen Untersuchung betrachteten Messgrößen werden hinsichlich ihrer unterschiedlichen Werteskalen unterschieden.

Definition 2.7 *Die in einer statistischen Untersuchung interessierenden Messgrößen $X_1, \ldots, X_p$ werden **Merkmale** (oder auch **Variablen**) genannt. Die Untersuchungseinheiten, d. h. die Objekte, an denen man die Merkmale erfasst, nennt man **Merkmalsträger** (oder auch **statistische Einheiten**, **Individuen**, **Fälle**). Die Menge A aller in einer Stichprobe auftretenden Werte eines Merkmals nennt man **Ausprägungen**. Sei A_0 die Menge aller theoretisch möglichen Ausprägungen eines Merkmals. Ist A_0 endlich oder abzählbar, spricht man von einem **diskreten** Merkmal. Besitzt ein Merkmal eine überabzählbare Ausprägungsmenge A_0 (z. B. ein Intervall in $\mathbb{R}$), nennt man das Merkmal **stetig**.*

Beispiel 2.8 In einer Stichprobe von 200 Wohngebäuden wurden die Merkmale *Wohnfläche in Quadratmeter* und *Anzahl der Räume* erfasst. Die *Wohnfläche* ist ein stetiges Merkmal mit $A_0 = (0, \infty)$ und die *Raumanzahl* ist ein diskretes Merkmal mit $A_0 = \mathbb{N}$. □

Definition 2.9 (Statistische Skalenniveaus) *Ein Merkmal ist **nominalskaliert**, wenn seine möglichen Ausprägungen Klassen oder Kategorien sind, die keine Anordnung erlauben. Sind die möglichen Ausprägungen eines Merkmals anordbar, aber es können keine Abstände der Ausprägungen interpretiert werden, ist das Merkmal **ordinalskaliert**. Bei einem **intervallskalierten** Merkmal sind die möglichen Ausprägungen eine Teilmenge der reellen Zahlen und die Abstände der Ausprägungen sind somit interpretierbar. Die Intervallskala besitzt aber keinen absoluten, natürlichen Nullpunkt, daher sind Quotientenbildungen nicht sinnvoll interpretierbar. Ein Merkmal heißt **verhältnisskaliert**, falls über die Eigenschaften der Intervallskala hinaus noch ein absoluter, natürlicher Nullpunkt in der Skala existiert. Zusammenfassend spricht man bei der Intervall- und Verhältnisskala*

*auch von der **Kardinalskala** und kardinalskalierte Merkmale werden auch als **metrische** Merkmale bezeichnet.*

Die Bezeichnung Skalenniveaus bezieht sich bei den Skalentypen auf den Informationsgehalt der Skalierung und den möglichen Operationen, die die Skalierung erlaubt. So kann z. B. bei einer Stichprobe eines nominalskalierten Merkmals nur die Gleichheit bzw. Unterscheidung von Ausprägungen verwendet werden, während die Ordinalskala zusätzlich Reihenfolgen bzw. Rangbildungen erlaubt. Höhere Skalenniveaus können immer auf niedrigere Niveaus umgerechnet werden. So kann z. B. ein in der Kardinalskala gemessenes Merkmal immer auf eine Ordinal- oder Nominalskala transformiert werden (durch Klassenbildung), die Umkehrung gilt aber nicht. Statistische Verfahren setzen für ihre Anwendung immer ein bestimmtes minimales Skalenniveau voraus.

Beispiel 2.10

- Nominalskala: Geschlecht, Wohnort, Farbe, Beruf.
- Ordinalskala: Schulnoten, Kreditwürdigkeitsranking, Hotelkategorie.
- Intervallskala: Temperaturmessung in Grad Celsius, Kalenderdatum, Intelligenzquotient.
- Verhältnisskala: Alter, Schadenanzahl, Schadenhöhe. □

In manchen weiterführenden, statistischen Verfahren, wie z. B. bei Regressionsmodellen, werden die Merkmale eines Datensatzes nicht gleichwertig betrachtet, sondern den Merkmalen werden verschiedene Rollen zugeordnet. Die eigentlich interessierende Größe, für die man z. B. aus Prognosezwecken eine statistische Modellbildung durchführt, nennt man dann **Kriteriumsvariable** oder **abhängige Variable**, **Response**, **Zielfunktion**. Diejenigen Merkmale eines Datensatzes, die die Kriteriumsvariable funktional beeinflussen und nach einer Modellbildung beschreiben sollen, nennt man **Einflussgrößen** oder auch **unabhängige Variablen**. Metrische Einflussgrößen werden oft als **Kovariate** oder **Kovariablen** (z. B. in der Regressionsanalyse) bezeichnet, während man im Fall von nominalen Einflussgrößen von **Faktoren** (z. B. in der Varianzanalyse) spricht.

2.2 Häufigkeitsverteilungen

Im folgenden Abschnitt wird die Häufigkeitsverteilung einer Stichprobe betrachtet. Für Stichproben eines metrischen Merkmals sind das Histogramm, die empirische Verteilungsfunktion und die empirischen Quantile die grundlegenden Größen zur Darstellung und Analyse von Häufigkeitsverteilungen. Im Fall einer bivariaten Stichprobe nominaler Merkmale wird die Häufigkeitsverteilung in Kontingenztafeln zusammengefasst.

Sei

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top$$

eine Stichprobe eines Merkmals vom Umfang n und

$$A = \{a_1, \ldots, a_m\}$$

die Menge der Ausprägungen in der Stichprobe, d. h. die Menge aller unterschiedlichen Stichprobenwerte. Offensichtlich gilt stets $m \leq n$.

Definition 2.11 (Häufigkeitsverteilung) *Die Zahlenwerte*

$$h_i := h(a_i) := \sum_{j=1}^{n} 1_{\{a_i\}}(x_j),\ i = 1, \ldots, m,$$

nennt man ***absolute Häufigkeitsverteilung*** *der Stichprobe* $\boldsymbol{x}$.

Die Zahlenwerte

$$f_i := f(a_i) := \frac{h_i}{n},\ i = 1, \ldots, m,$$

nennt man ***relative Häufigkeitsverteilung*** *der Stichprobe* $\boldsymbol{x}$.

Ergänzend können die Häufigkeiten für zusätzliche, theoretisch mögliche Ausprägungswerte $b \in A_0$, die nicht in der Stichprobe explizit auftreten, als

$$h(b) = f(b) := 0$$

definiert werden. Man beachte, dass

$$\sum_{i=1}^{m} h_i = n \quad \text{und} \quad \sum_{i=1}^{m} f_i = 1\,.$$

Die Häufigkeitsverteilung einer Stichprobe kann mithilfe von Kreis-, Stab-, Säulen-, Balkendiagrammen oder auch Dotcharts grafisch dargestellt werden.

Beispiel 2.12 Gegeben sei eine Stichprobe vom Umfang $n = 10$

$$\boldsymbol{x} = (\mathrm{m}, \mathrm{m}, \mathrm{w}, \mathrm{m}, \mathrm{m}, \mathrm{w}, \mathrm{m}, \mathrm{w}, \mathrm{m}, \mathrm{m})^\top$$

des Merkmals Geschlecht, wobei die Codierung m für männlich und w für weiblich verwendet wurde. Man erhält die Häufigkeitsverteilungen

$$h(\mathrm{m}) = 7,\ h(\mathrm{w}) = 3 \text{ bzw. } f(\mathrm{m}) = \frac{7}{10},\ f(\mathrm{w}) = \frac{3}{10}.$$

In Abb. 2.1 ist die absolute Häufigkeitsverteilung grafisch dargestellt. □

Besteht eine Stichprobe aus Realisationen unabhängiger und identisch verteilter (kurz: i. i. d.) Zufallsvariablen X_i, $i \geq 1$, sind die relativen Häufigkeiten konsistente und erwartungstreue Schätzer für die entprechenden Wahrscheinlichkeiten.

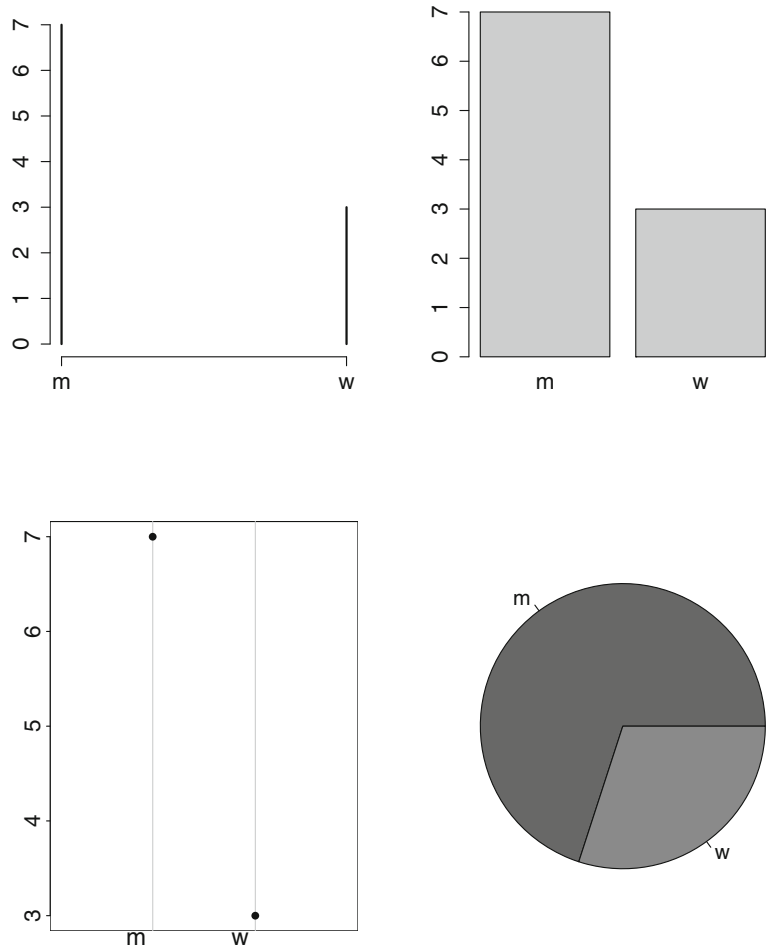

Abb. 2.1 Verschiedene grafische Darstellungen der absoluten Häufigkeitsverteilung aus Beispiel 2.12: Stabdiagramm, Säulendiagramm, Dotchart und Kreisdiagramm

Lemma 2.13 (Starkes Gesetz der großen Zahlen für relative Häufigkeiten) *Seien* X_i, $i \geq 1$, *i. i. d. Zufallsvariablen, dann gilt für alle* $a \in \mathbb{R}$ *und* $n \in \mathbb{N}$

$$E\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=a\}}\right) = P(X_i = a) \text{ (Erwartungstreue)}$$

und für alle $a \in \mathbb{R}$ *und* $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=a\}} \overset{f.s.}{\to} P(X_i = a) \text{ (starke Konsistenz).} \tag{2.1}$$

Man beachte, dass $\widehat{P}(X_i = a) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=a\}}$ einen Schätzer (Schätzfunktion) darstellt (d. h. $\widehat{P}(X_i = a)$ ist als Funktion der Stichprobenvariablen X_i, $i \geq 1$, selbst wieder eine Zufallsvariable), während die relative Häufigkeit $\frac{1}{n}\sum_{i=1}^{n} 1_{\{a\}}(x_i)$ als Zahlenwert (mit den Realisationen x_i der Zufallsvariablen X_i, $i = 1, \ldots, n$,) dann ein konkreter Schätzwert ist.

Beweis Da X_i, $i \geq 1$, i. i. d. Zufallsvariablen sind, folgt für alle $a \in \mathbb{R}$, dass auch die Zufallsvariablen $1_{\{X_i=a\}}$, $i \geq 1$, unabhängig und identisch verteilt sind.
Für alle $a \in \mathbb{R}$ und $i \geq 1$ gilt

$$E\left(\left|1_{\{X_i=a\}}\right|\right) \leq E(1) = 1 < \infty.$$

Mit den üblichen Rechenregeln des Erwartungswertes folgt, dass für alle $a \in \mathbb{R}$ und $n \in \mathbb{N}$

$$E\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=a\}}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(1_{\{X_i=a\}}\right) = \frac{1}{n} n E\left(1_{\{X_1=a\}}\right) = P\left(X_1 = a\right).$$

Nach dem starken Gesetz der großen Zahlen nach Komogorov, vgl. z. B. Pruscha [6], S. 343, folgt dann die Konsistenzeigenschaft (2.1). □

Allgemeiner als das Lemma 2.13 gilt das **Theorem von Bernoulli** (vgl. Fahrmeir et al. [3], S. 312), in dem die Konsistenzaussage (2.1) von $\{X_i = a\}$ auf beliebige Ereignisse $\{X_i \in A\}$, $A \subseteq \mathbb{R}$, erweitert wird.

2.2.1 Histogramm

Im Fall einer Stichprobe

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top$$

eines stetigen, metrischen Merkmals sind die Häufigkeitsverteilungen und ihre direkten grafischen Darstellungen, z. B. mittels eines Stabdiagramms, nicht sehr hilfreich, denn im Allgemeinen gilt hier

$$f_i \approx \frac{1}{n} \; \forall\, i = 1, \ldots, m.$$

D. h. die Stichprobenwerte sind fast alle verschieden. In dieser Situation klassifiziert man den Wertebereich der Stichprobe und bildet ein Histogramm.

Definition 2.14 (Histogramm) *Der Wertebereich*

$$W = [\min\{x_1, \ldots, x_n\}, \max\{x_1, \ldots, x_n\}]$$

einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ *reeller Zahlen sei in* $k \in \mathbb{N}$ *benachbarte, disjunkte Teilintervalle*

$$I_1 = [c_0, c_1), \ldots, I_k = [c_{k-1}, c_k]$$

mit $c_{i-1} < c_i$ *für* $i = 1, \ldots, k$ *und* $\bigcup_{i=1}^{k} I_i \supseteq W$ *aufgeteilt. Bezeichne für* $i = 1, \ldots, k$

$$h_i = \sum_{j=1}^{n} 1_{I_i}(x_j)$$

die absoluten Klassenhäufigkeiten der Teilintervalle. Das ***Histogramm der absoluten Klassenhäufigkeiten*** *der Stichprobe* $\boldsymbol{x}$ *besteht dann aus k Rechtecken über den Intervallen* I_i, $i = 1, \ldots, k$, *mit Rechtecksbreiten* $c_i - c_{i-1}$ *und geeignet gewählten Rechteckshöhen* H_i *mit der Eigenschaft, dass*

$$h_i = C \cdot H_i \cdot (c_i - c_{i-1}) \quad \textit{für alle } i = 1, \ldots, k,$$

wobei C *eine fest gewählte, positive reelle Zahl (Proportionalitätsfaktor) bezeichnet.*

In einem Histogramm werden demnach die Klassenhäufigkeiten proportional (mit Proportionalitätsfaktor C) zu den entsprechenden Rechtecksflächen dargestellt. Man spricht hier von dem **Prinzip der Flächentreue**.

Mithilfe eines Histogramms kann die Häufigkeitsverteilung einer Stichprobe unter anderem hinsichtlich Uni- oder Multimodalität und bzgl. Symmetrie bzw. Asymmetrie (Schiefe) untersucht werden.

Bemerkung 2.15

a) Das **Histogramm der relativen Klassenhäufigkeiten** wird ganz analog gebildet, indem man h_i durch die relative Klassenhäufigkeit $f_i := \frac{h_i}{n}$ ersetzt. Bei einem Histogramm der relativen Klassenhäufigkeiten mit Proportionalitätsfaktor $C = 1$ gilt, dass die Gesamtfläche aller Rechtecke identisch 1 ist.

b) Alternativ können die disjunkten Teilintervalle auch in der Form

$$I_1 = [c_0, c_1], I_2 = (c_1, c_2] \ldots, I_k = (c_{k-1}, c_k],$$

d. h. als rechts geschlossene und links offene Intervalle gebildet werden. Entscheidend ist, dass die Intervalleinteilung disjunkt ist und der gesamte Wertebereich der Stichprobe überdeckt wird.

c) Die Rechteckshöhen $H_i = \frac{h_i}{c_i - c_{i-1}}$ bzw. $H_i = \frac{f_i}{c_i - c_{i-1}}$ (mit $C = 1$) werden auch als **Häufigkeitsdichte** bezeichnet.
d) Für den Spezialfall, dass alle Teilintervalle I_i, $i = 1, \ldots, k$, identische Breite besitzen, können die Rechteckshöhen direkt als Klassenhäufigkeiten interpretiert werden. In der Anwendung wird oft diese äquidistante Intervalleinteilung aufgrund der einfacheren Interpretation verwendet.
e) In der Literatur zur angewandten Statistik (vgl. z. B. Fahrmeir et al. [3], S. 42) findet man verschiedene Regeln für die bei einem vorliegenden Stichprobenumfang n zu wählende Anzahl k von Teilintervallen, z. B. $k = \lfloor \sqrt{n} \rfloor$ oder $k = \lfloor 10 \log_{10} n \rfloor$, wobei $\lfloor x \rfloor$ den ganzzahligen Anteil von $x \in \mathbb{R}$ bezeichnet. Andere Empfehlungen für die Intervalleinteilung berücksichtigen auch die Streuung der Daten.
f) Sowohl der gewählten Anzahl k der Teilintervalle als auch der Wahl der Intervallgrenzen ist bei Histogrammen besondere Aufmerksamkeit zu widmen, da diese Festlegungen die resultierende Interpretation der Häufigkeitsverteilung stark beeinflusssen können.

In der Abb. 2.2 sind drei Histogramme der relativen Klassenhäufigkeiten mit unterschiedlichen Intervalleinteilungen einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_{100})^\top$ dargestellt. Die Stichprobe $\boldsymbol{x}$ besteht aus 100 auf dem Intervall [0,5] gleichverteilten Pseudo-Zufallszahlen.

Analog zu dem Beweis von Lemma 2.13 zeigt man das folgende Konsistenzergebnis für die Rechtecksflächen in einem Histogramm.

Korollar 2.16 (Starke Konsistenz der Histogramm-Rechtecke) *Im Fall von i. i. d. Stichprobenvariablen $X_1, X_2, \ldots$ gilt für die Schätzfunktionen*

$$F_i := \frac{1}{n} \sum_{j=1}^{n} 1_{\{X_j \in I_i\}},\ i = 1, \ldots, k,$$

$$F_i \xrightarrow{f.s.} P(c_{i-1} \leq X_j < c_i) = P(X_j \in I_i).$$

Die Schätzfunktionen F_i entsprechen den Flächen der Rechtecke in einem Histogramm der relativen Häufigkeiten mit Proportionalitätsfaktor $C = 1$ und der Intervalleinteilung $I_1 = [c_0, c_1), \ldots, I_k = [c_{k-1}, c_k]$.

Neben der rein deskriptiven Darstellung der Häufigkeitsverteilung einer Stichprobe können Histogramme auch zur Schätzung der unbekannten Dichte f der Stichprobenvariablen verwendet werden. Besteht eine Stichprobe $\boldsymbol{x}$ aus Realisationen der i. i. d. Zufallsvariablen X_i, $i \geq 1$, mit existierender (aber unbekannter) Wahrscheinlichkeitsdichte f, so stellt ein Histogramm der relativen Häufigkeiten (mit Proportionalitätsfaktor $C = 1$) einen einfachen, elementaren Schätzer $\widehat{f}$ für die Dichte f dar.

Wir gehen dazu von einer vorgegebenen, äquidistanten Intervalleinteilung

$$I_i := [x_0 + i \cdot h, x_0 + (i + 1) \cdot h),\ i \in \mathbb{Z},$$

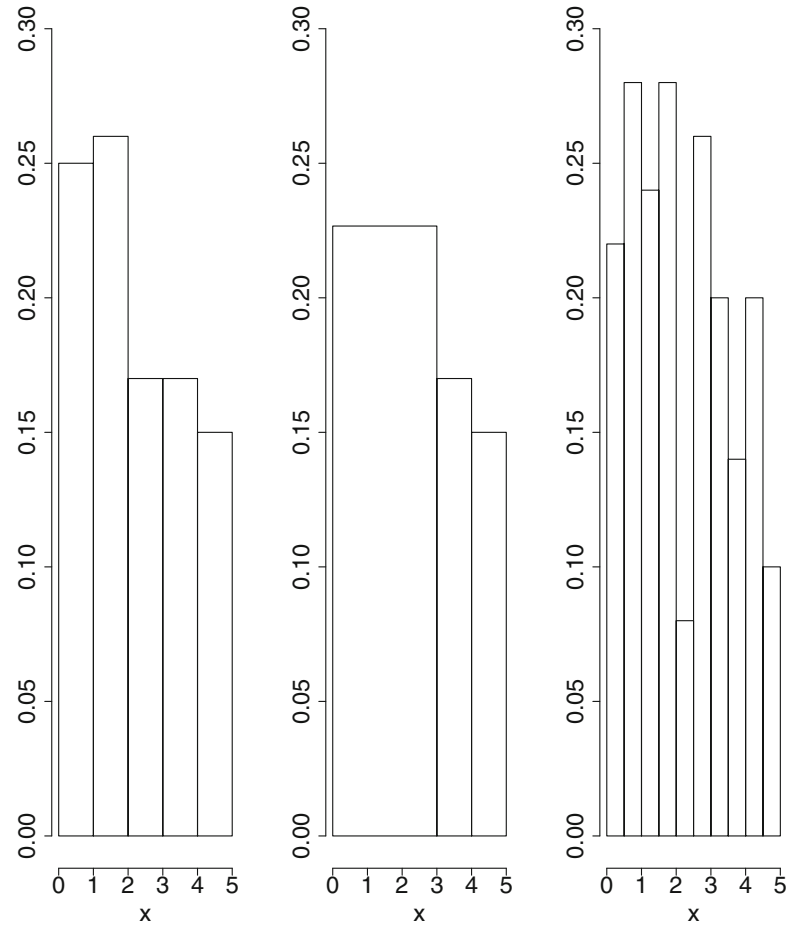

Abb. 2.2 Histogramme einer Stichprobe mit unterschiedlichen Intervalleinteilungen

mit Intervallbreite $h > 0$ und mit vorab festgelegtem $x_0 \in \mathbb{R}$ aus. Für alle $x \in \mathbb{R}$ definiert man dann als **Histogramm-Schätzer**

$$\widehat{f}_n(x) := \widehat{f}_{n,x_0,h}(x) := \frac{1}{nh} \sum_{j=1}^{n} 1_{\{X_j \in I(x)\}}, \tag{2.2}$$

wobei $I(x) = I_i$, falls $x \in I_i$.

Lemma 2.17 (Eigenschaften des Histogramm-Schätzers) *Der in (2.2) definierte Histogramm-Schätzer besitzt im Fall von i. i. d. Stichprobenvariablen X_i, $i \geq 1$, die*

Eigenschaften

$$\forall\, \omega \in \Omega, x \in \mathbb{R}, n \in \mathbb{N}: \qquad \widehat{f}_n(x) \geq 0$$

$$\forall\, \omega \in \Omega, n \in \mathbb{N}: \qquad \int_{-\infty}^{\infty} \widehat{f}_n(x)\, dx = 1$$

$$\forall\, x \in \mathbb{R} \text{ und } n \to \infty: \qquad \widehat{f}_n(x) \overset{f.s.}{\to} \frac{1}{h} \int_{I(x)} f(t)\, dt$$

Beweis Bezeichne $(\Omega, \mathcal{A}, P)$ den Wahrscheinlichkeitsraum, über dem die i. i. d. Stichprobenvariablen X_i, $i \geq 1$, mit der unbekannten Dichte f definiert sind.
Die erste Eigenschaft folgt sofort aus der Definition (2.2) des Histogramm-Schätzers.
Für den Beweis der zweiten Eigenschaft rechnet man für alle $\omega \in \Omega$

$$\int_{-\infty}^{\infty} \widehat{f}_n(x)(\omega)\, dx = \frac{1}{nh} \sum_{j=1}^{n} \int_{-\infty}^{\infty} 1_{\{X_j \in I(x)\}}(\omega)\, dx = \frac{1}{nh} \sum_{j=1}^{n} \int_{a_j(\omega)}^{b_j(\omega)} 1\, dx,$$

wobei $[a_j(\omega), b_j(\omega)) := I(X_j(\omega))$. Da $b_j(\omega) - a_j(\omega) = h$ für alle $j \geq 1$ und $\omega \in \Omega$ erhält man weiter

$$\frac{1}{nh} \sum_{j=1}^{n} \int_{a_j(\omega)}^{b_j(\omega)} 1\, dx = \frac{1}{nh} \sum_{j=1}^{n} h = 1.$$

Da für alle $x \in \mathbb{R}$

$$E\left(\frac{1}{h} 1_{\{X_i \in I(x)\}}\right) = \frac{1}{h} \cdot P\left(X_i \in I(x)\right) = \frac{1}{h} \int_{I(x)} f(t)\, dt,$$

folgt mit dem starken Gesetz der großen Zahlen nach Komogorov, vgl. z. B. Pruscha [6], S. 343, dass

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} 1_{\{X_j \in I(x)\}} \xrightarrow{f.s.} \frac{1}{h} \int_{I(x)} f(t)\, dt \quad \forall\, x \in \mathbb{R} \text{ und } n \to \infty,$$

d. h. die dritte Behauptung des Lemmas. □

In statistischen Analysen stellt sich oft die Frage, ob eine Verteilungsannahme für die Stichprobenvariablen gerechtfertigt ist. Eine einfache, deskriptive bzw. explorative Vorgehensweise ist nach den obigen Ergebnissen der Vergleich des Histogramms bzw. des

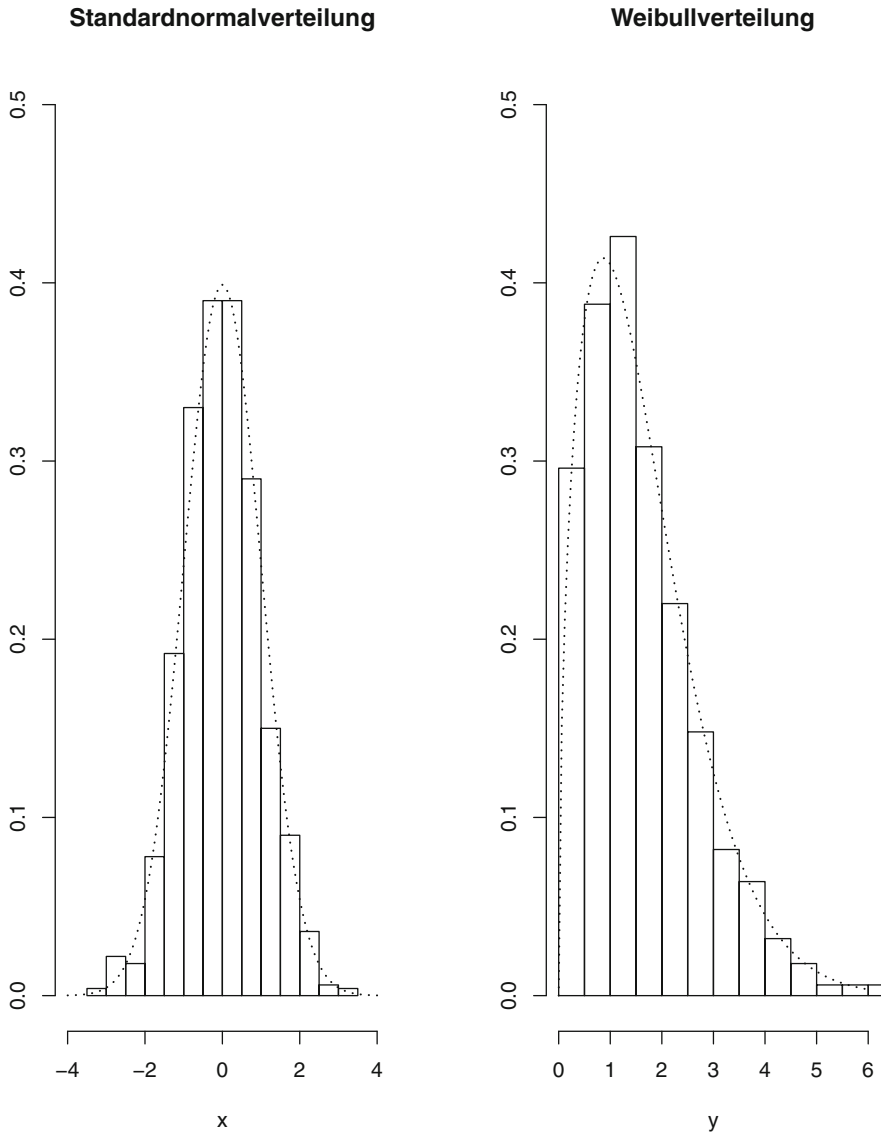

Abb. 2.3 Histogramme und theoretische Dichtefunktionen

Histogrammschätzers mit der zur Verteilungsannahme gehörigen, theoretischen Dichtefunktion. Nach Lemma 2.17 sollte sich bei genügend großem Stichprobenumfang und genügend klein gewählten Intervallbreiten der Histogrammschätzer der theoretischen Dichtefunktion annähern.

In der Abb. 2.3 sind das Histogramm einer i. i. d. Stichprobe $\boldsymbol{x}$ von 1000 standardnormalverteilten Pseudo-Zufallszahlen und das Histogramm einer i. i. d. Stichprobe $\boldsymbol{y}$ von

1000 Pseudo-Zufallszahlen, die nach einer Weibullverteilung mit Formparameter $\beta = \frac{3}{2}$ und Skalenparameter $\alpha = \frac{9}{5}$ verteilt sind, zusammen mit den entsprechenden theoretischen Dichten dargestellt. Die Annäherung der Histogramm-Schätzungen an die theoretischen Dichtefunktionen sind deutlich zu erkennen.

Der in (2.2) definierte Histogramm-Schätzer besitzt als Dichtekurven-Schätzer zwei wesentliche Nachteile. Der Schätzer hängt von der vorgegebenen Intervalleinteilung (über die Fixierungsgröße $x_0 \in \mathbb{R}$) ab und die resultierende Dichtekurvenschätzung führt zu einer unstetigen Funktion (Treppenfunktion). Die Abhängigkeit des Schätzers von x_0 kann durch die etwas modifizierte Definition

$$\widehat{f}_n(x) := \widehat{f}_{n,h}(x) := \frac{1}{2nh} \sum_{j=1}^{n} 1_{\{X_j \in [x-h, x+h[\}} \tag{2.3}$$

leicht vermieden werden. Ein Vergleich von (2.3) mit der für die zu schätzende Dichte f gültigen Darstellung

$$f(x) = \lim_{h \to 0} P\,(x - h \leq X_1 \leq x + h)$$

zeigt deutlich die Verwandtschaft von Schätzer und zu schätzender Dichte.

Der Nachteil der Unstetigkeit von Histogramm-Schätzern bleibt aber bestehen und so werden Histogramm-Schätzer auch als naive Dichteschätzer bezeichnet. Weiterentwicklungen von Dichteschätzern, die dann auch stetige Schätzfunktionen liefern, sind z. B. **Kerndichteschätzer**, vgl. Abschn. 2.4.5, oder **Orthogonalreihenschätzer**. Eine Einführung in die Theorie dieser Dichteschätzverfahren gibt z. B. Pruscha [6] in Kapitel VIII. Hier werden auch grundlegende Eigenschaften der Dichteschätzer, wie z. B. Konsistenz und Konvergenzordnung, dargestellt. In der praktischen Datenanalyse mit Softwareunterstützung werden Histogramme oft kombiniert mit Dichteschätzern wie z. B. Kerndichteschätzern verwendet.

2.2.2 Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion einer Stichprobe gibt für alle $x \in \mathbb{R}$ den relativen Anteil der Stichprobenwerte an, die kleiner oder gleich dem Wert x sind.

Definition 2.18 (Empirische Verteilungsfunktion) *Die empirische Verteilungsfunktion F_n einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ eines metrischen Merkmals ist definiert als die Funktion*

$$F_n : \mathbb{R} \to [0,1],\ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(x_i).$$

Die empirische Verteilungsfunktion $F_n(x), x \in \mathbb{R}$, einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit den Ausprägungen $a_1, \ldots, a_m$ ist eine rechtsseitig stetige, monoton wachsende Treppenfunktion mit den Sprungstellen $a_1, \ldots, a_m$ und den entsprechenden relativen Häufigkeiten $f(a_1), \ldots, f(a_m)$ als Sprunghöhen. Für $x < a_{\min} := \min\{a_1, \ldots, a_m\}$ ist $F_n(x) = 0$ und für $x \geq a_{\max} := \max\{a_1, \ldots, a_m\}$ ist $F_n(x) = 1$.

In Teilintervallen $I \subset [a_{\min}, a_{\max}]$ in denen viele Beobachtungen liegen, besitzt die empirische Verteilungsfunktion einen starken Anstieg. Verläuft der Graph der empirischen Verteilungsfunktion in Teilintervallen $J \subset [a_{\min}, a_{\max}]$ eher flach, sind dort nur wenige Stichprobenwerte vorhanden. Die Abb. 2.4 zeigt beispielhaft den Graph der empirischen Verteilungsfunktion einer Stichprobe $\boldsymbol{x}$.

Im Fall von i. i. d. Stichprobenvariablen $X_i, i \geq 1$, mit (unbekannter) Verteilungsfunktion F ist die empirische Verteilungsfunktion (jetzt betrachtet als Schätzfunktion)

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

ein erwartungstreuer, stark konsistenter Schätzer für F, d. h.

$$E\left(F_n(x)\right) = F(x) \quad \forall\, x \in \mathbb{R},$$

und

$$F_n(x) \xrightarrow{f.s.} F(x) \quad \forall\, x \in \mathbb{R} \text{ und } n \to \infty.$$

Es gilt sogar, dass F_n fast sicher gleichmäßig auf $\mathbb{R}$ gegen F konvergiert, d. h. dass

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{f.s.} 0 \text{ für } n \to \infty. \tag{2.4}$$

Das Konvergenzergebnis (2.4) ist als **Satz von Glivenko-Cantelli** bekannt. Einen Beweis des Satzes von Glivenko-Cantelli findet man z. B. bei Pruscha [6], S. 156. Man nennt diese grundlegende Beziehung zwischen der empirischen Verteilungsfunktion einer Stichprobe und der theoretischen, der Stichprobe zugrundeliegenden, aber in der Praxis meist unbekannten Verteilungsfunktion auch den **Hauptsatz der mathematischen Statistik**. Die gleichmäßige Konvergenz (2.4) impliziert als prinzipielle Methode zur Untersuchung einer Verteilungsannahme den Vergleich von empirischer und theoretischer Verteilungsfunktion.

In der induktiven Statistik wird die gleichmäßige Konvergenz (2.4) bei der Konstruktion von nichtparametrischen (verteilungsfreien) Signifikanztests, wie z. B. den **Kolmogorov-Smirnov-Test**, angewandt. Der Kolmogorov-Smirnov-Test verwendet die Teststatistik

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

und ermöglicht die induktive Beurteilung von Verteilungsannahmen, vgl. z. B. Pruscha [7], S. 25–26.

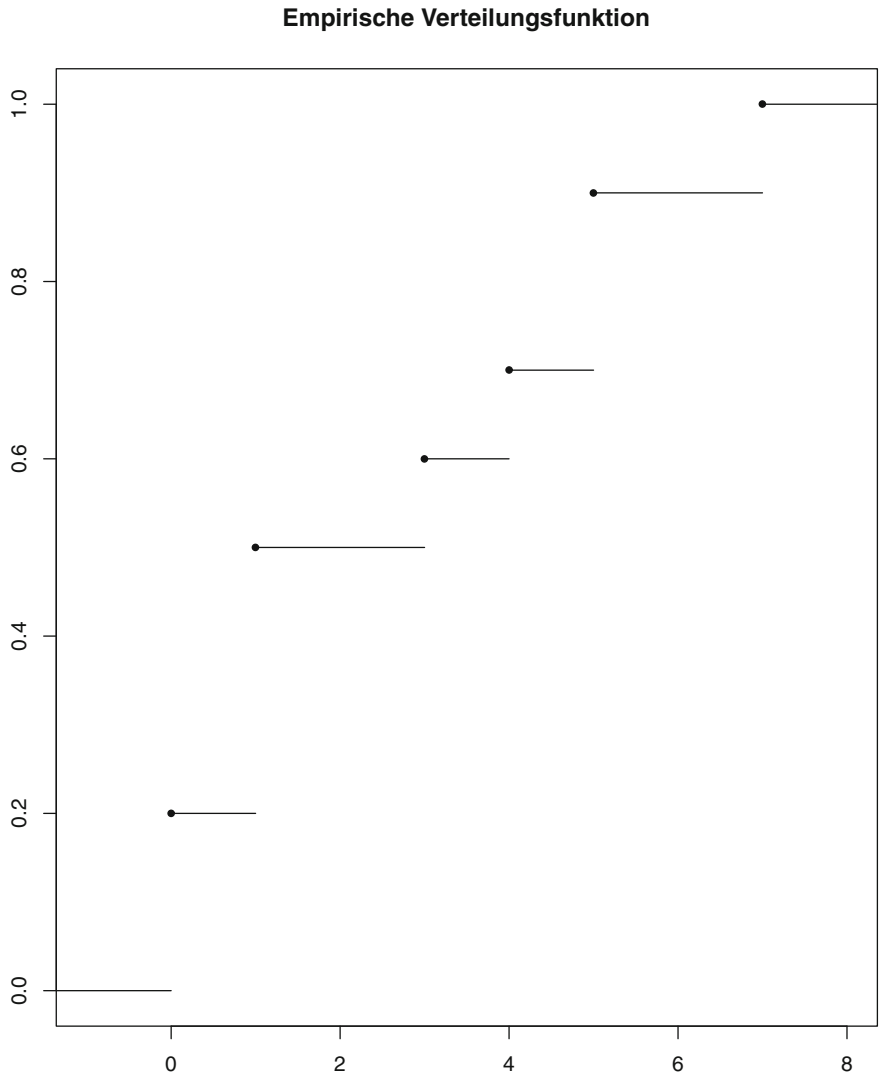

Abb. 2.4 Empirische Verteilungsfunktion F_{10} der Stichprobe $\boldsymbol{x} = (0, 0, 1, 1, 1, 3, 4, 5, 5, 7)^\top$

Mit dem Satz von Glivenko-Cantelli folgt auch eine Konvergenzaussage für die empirischen Quantilsfunktionen.

Definition 2.19 (Empirische Quantilsfunktion) *Die empirische Quantilsfunktion $F_n^{\leftarrow}$ einer empirischen Verteilungsfunktion F_n ist definiert als die verallgemeinerte Inverse von F_n, d. h. als*

$$F_n^{\leftarrow} : (0,1) \to \mathbb{R},\ F_n^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F_n(x) \geq p\}\,.$$

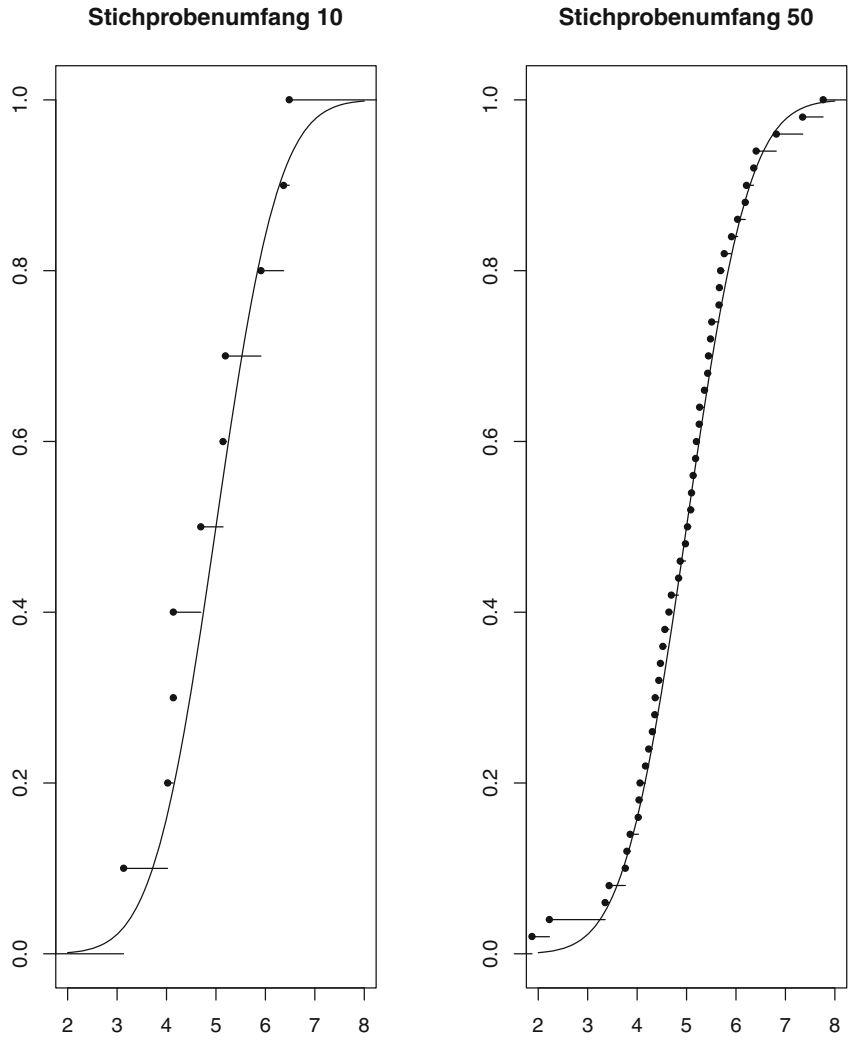

Abb. 2.5 Empirische Verteilungsfunktionen zu simulierten Stichproben mit unterschiedlichen Stichprobenumfängen und theoretische Verteilungsfunkion einer $\mathcal{N}(5,1)$-verteilten Zufallsvariablen

Lemma 2.20 *Für i. i. d. Zufallsvariablen $X_1, \ldots, X_n$ mit Verteilungsfunktion F gilt*

$$F_n^{\leftarrow}(p) \xrightarrow{f.s.} F^{\leftarrow}(p) \quad \text{für } n \to \infty$$

an jeder Stetigkeitsstelle $0 < p < 1$ von $F^{\leftarrow}$, wobei $F^{\leftarrow}$ die Quantilsfunktion (verallgemeinerte Inverse) von F bezeichnet.

Beweis Man wendet Satz 5.67 bei Witting und Müller-Funk [11], S. 71 f., und den Satz von Glivenko-Cantelli an. □

In der deskriptiven und explorativen Analyse verwendet man die empirische Verteilungsfunktion für die Bewertung von Verteilungsannahmen. In der Abb. 2.5 sind für zwei Stichproben $\mathcal{N}(5,1)$-verteilter Pseudo-Zufallszahlen (für diesen Begriff s. Kap. 5.1) mit den Umfängen $n = 10$ und $n = 50$ jeweils die Graphen der empirischen Verteilungfunk-

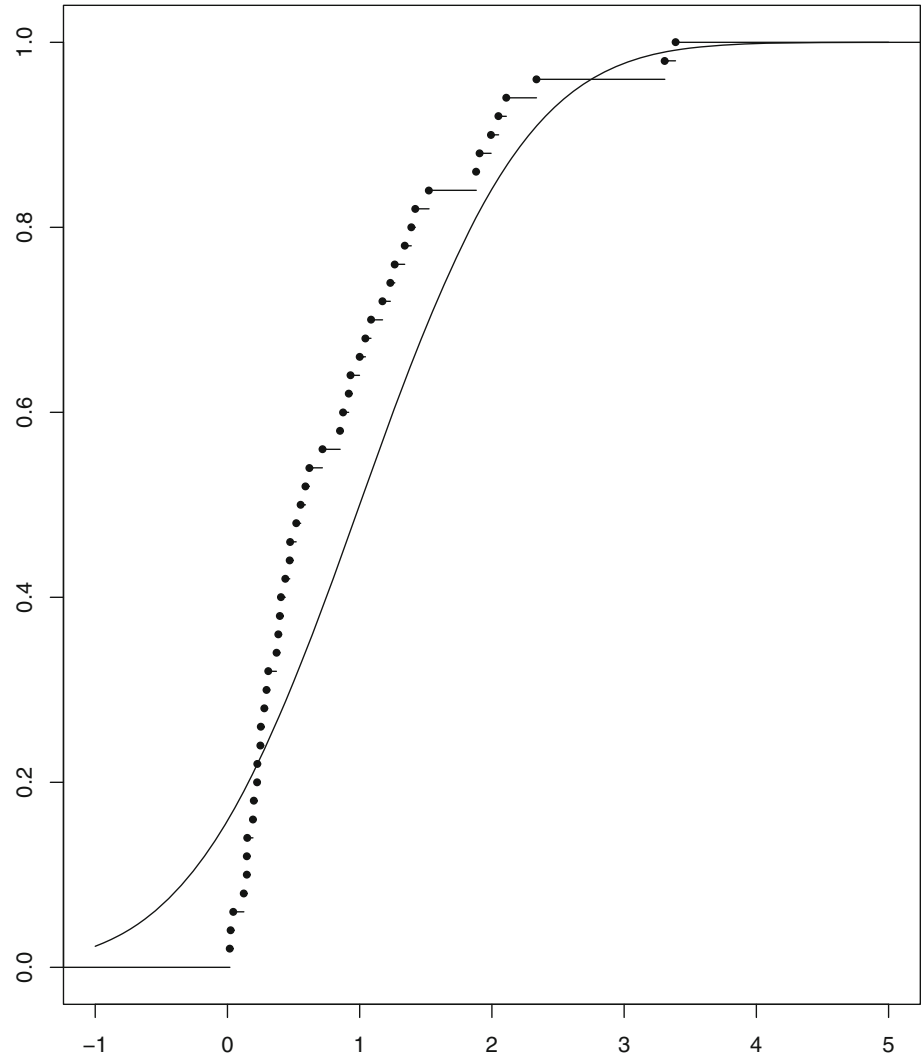

Abb. 2.6 Empirische Verteilungsfunktion einer Stichprobe von 50 $\Gamma(1,1)$-verteilten Pseudo-Zufallszahlen und theoretische Verteilungsfunktion einer $\mathcal{N}(1,1)$-verteilten Zufallsvariablen

tionen F_n zusammen mit der theoretischen Verteilungsfunktion F einer $\mathcal{N}(5,1)$-verteilten Zufallsvariablen dargestellt.

Die Abb. 2.6 zeigt die empirische Verteilungsfunktion einer simulierten Stichprobe von 50 $\Gamma(1,1)$-verteilten Zufallszahlen gemeinsam mit der theoretischen Verteilungsfunktion einer $\mathcal{N}(1,1)$-verteilten Zufallsvariablen. Man erkennt eine deutliche Abweichung der beiden Graphen.

2.2.3 Empirische Quantile

Im Folgenden sei $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ die Stichprobe eines metrischen oder ordinal skalierten Merkmals und

$$(x_{(1)}, \ldots, x_{(n)})^\top$$

bezeichne die zugehörige **geordnete Stichprobe**, d. h. die Stichprobenwerte $x_1, \ldots, x_n$ werden ihrer Größe nach geordnet (von klein nach groß). Es gilt also

$$x_{(i)} \leq x_{(i+1)} \quad (i = 1, \ldots, n-1)$$

Definition 2.21 (Empirische Quantile) *Für $p \in (0,1)$ ist das empirische p-Quantil x_p einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ definiert als*

$$x_p := \begin{cases} q \in \left[x_{(np)}, x_{(np+1)}\right], & \text{falls } np \in \mathbb{N} \\ x_{(\lfloor np \rfloor + 1)} & \text{, falls } np \notin \mathbb{N} \end{cases} \tag{2.5}$$

wobei $\lfloor np \rfloor$ die größte ganze Zahl bezeichnet, die kleiner oder gleich np ist.

Für den Fall, dass $np \in \mathbb{N}$, werden auch die modifizierten Definitionen

$$x_p := F_n^{\leftarrow}(p) = x_{(np)}, \tag{2.6}$$

oder, falls es sich um eine Stichprobe eines metrischen Merkmals handelt,

$$x_p := \frac{x_{(np)} + x_{(np+1)}}{2}, \tag{2.7}$$

verwendet.

Man beachte, dass die Definition (2.5) keine eindeutige Wertzuweisung liefert, sondern das p-Quantil als einen beliebigen Wert innerhalb eines ganzen **Quantilintervalls** festlegt. Die modifizierten Definitionen (2.6) und (2.7) formulieren eine eindeutige Zuweisung des

p-Quantils. In der Literatur und innerhalb statistischer Software finden sich auch noch weitere Definitionsmodifikationen.

Das empirische p-Quantil x_p (man sagt auch: $p \cdot 100\,\%$-Quantil) einer Stichprobe $\boldsymbol{x}$ besitzt die grundlegende Eigenschaft, dass mindestens ein Anteil von $p \cdot 100\,\%$ der Stichprobenwerte kleiner oder gleich als x_p ist und mindestens ein Anteil von $(1-p) \cdot 100\,\%$ der Stichprobenwerte größer oder gleich x_p ist.

Das empirische $\frac{1}{2}$-Quantil (50 %-Quantil) $x_{\frac{1}{2}}$ nennt man den **empirischen Median** der Stichprobe, er teilt die Stichprobe in zwei (etwa) gleich mächtige Mengen von Stichprobenwerte, die kleiner oder gleich dem empirischen Median bzw. größer oder gleich dem empirischen Median sind. Die häufig verwendeten speziellen Quantile $x_{\frac{1}{4}}$ und $x_{\frac{3}{4}}$ werden als **unteres** bzw. **oberes Quartil** bezeichnet. Die Spezialfälle $x_{\frac{k}{10}}$ für $k \in \mathbb{N}$, $k \leq 9$, werden **Dezile** genannt. Die Quantile $x_{\frac{k}{100}}$ für $k \in \mathbb{N}$, $k \leq 99$, bezeichnet man als **Perzentile**.

Für die empirische Verteilungsfunktion F_n einer Stichprobe $\boldsymbol{x}$ vom Umfang n und $p \in]0{,}1[$ gilt entweder

$$F_n^{-1}(\{p\}) = \emptyset \tag{2.8}$$

oder

$$F_n^{-1}(\{p\}) = [x_{(np)}, x_{(np+1)}), \tag{2.9}$$

wobei $F_n^{-1}(\{p\})$ das Urbild von p unter der empirischen Verteilungsfunktion bezeichnet. Im Fall (2.8) erhält man das empirische p-Quantil dann als

$$x_p = x_{(\lfloor np \rfloor + 1)}$$

und für den Fall (2.9) erhält man (nicht mehr eindeutig)

$$x_p \in [x_{(np)}, x_{(np+1)})$$

bzw. eindeutig z. B. $x_p = \frac{x_{(np)} + x_{(np+1)}}{2}$. Die empirischen Quantile können aus dem Graphen der empirischen Verteilungsfunktion dementsprechend abgelesen werden. In Abb. 2.7 ist für die bereits geordnete Stichprobe

$$\boldsymbol{x} = (0, 0, 1, 1, 1, 3, 4, 5, 5, 7)^\top$$

exemplarisch das Auffinden des Quantilintervalls [1,3] für den empirischen Median $x_{\frac{1}{2}}$ und des 80 %-Quantils $x_{0,8} = 5$ skizziert. Nach der modifizierten Definition (2.7) würde man als empirischen Median $x_{\frac{1}{2}} = \frac{1+3}{2} = 2$ setzen.

Bemerkung 2.22 Ersetzt man in der Definition der empirischen Quantile die Realisationen, d. h. die Stichprobenwerte $x_1, \ldots, x_n$, durch die zugrundeliegenden i. i. d. Stichprobenvariablen $X_1, \ldots, X_n$, so erhält man Schätzer $\widehat{\kappa}_p$ für die (theoretischen) Quantile κ_p,

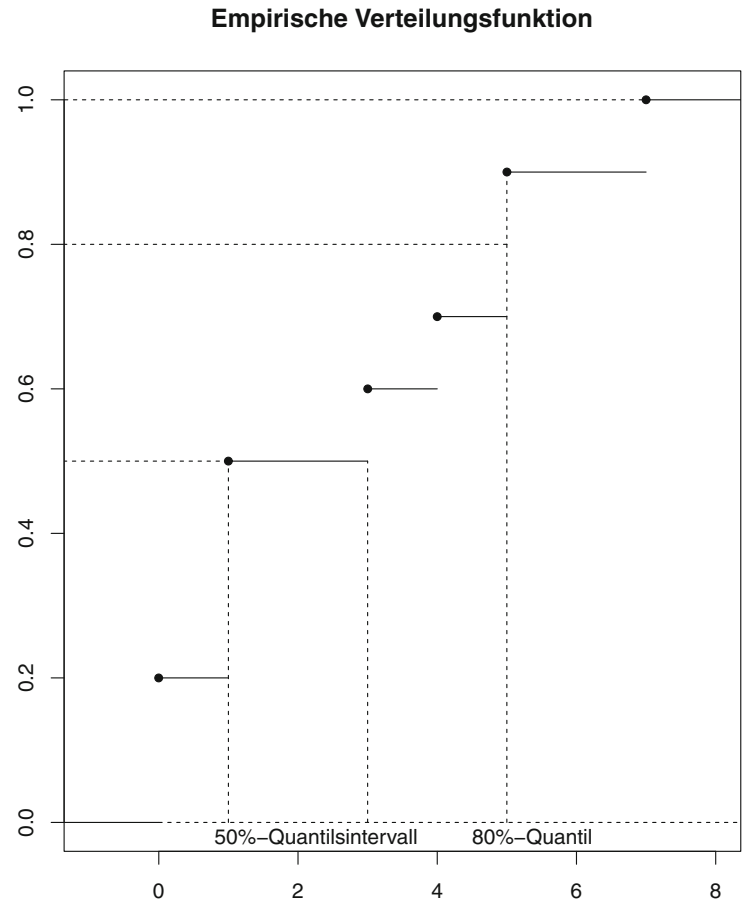

Abb. 2.7 Bestimmung von empirischen Quantilen aus dem Graphen der empirischen Verteilungsfunktkion der Stichprobe $(0, 0, 1, 1, 1, 3, 4, 5, 5, 7)^\top$

$0 < p < 1$, der zugrundeliegenden Verteilung mit Verteilungsfunktion F. Die dann in den Formeln auftretenden, geordneten Stichprobenvariablen $X_{(1)}, \ldots, X_{(n)}$ nennt man die **Ordnungsstatistik** der Zufallsstichprobe $X_1, \ldots, X_n$ und die Größen $X_{(i)}$, $i = 1, \ldots, n$, werden i-te **Ordnungsgrößen** genannt. Ist F stetig und sind die Quantile κ_p eindeutig bestimmt (z. B. falls F streng monoton ist), so sind die Schätzer $\widehat{\kappa}_p$ konsistent für κ_p, vgl. Witting und Müller-Funk [11], S. 71 f. und S. 575 f. D. h. bei großem Stichprobenumfang n erwartet man, dass

$$x_p \approx \kappa_p$$

für alle $0 < p < 1$. Man beachte, dass bei großem Stichprobenumfang n zudem die Approximation

$$x_p \approx x_{(np)}$$

gilt. Mithilfe der Ordnungsgrößen können auch Konfidenzintervalle für die Quantile κ_p, $0 < p < 1$, konstruiert werden, vgl. Pruscha [6], S. 49.

2.2.4 Kontingenztafeln

Die Häufigkeitsverteilung einer bivariaten Stichprobe

$$(x_1, y_1)^\top, \ldots, (x_n, y_n)^\top \tag{2.10}$$

zweier Merkmale X und Y vom Umfang n kann mithilfe einer Kontingenztafel (Kontingenztabelle) notiert werden. Für die praktische Anwendung ist der Spezialfall, dass beide Merkmale nominal skaliert sind, von besonderer Bedeutung. Der Begriff Kontingenz, also Zusammenhang, deutet bereits an, dass Fragestellungen bzgl. des Zusammenhangs der Merkmale oft im Mittelpunkt stehen.

Bezeichne

$$A = \{a_1, \ldots, a_k\}, k \leq n,$$

die Menge der Ausprägungen der Teilstichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ und

$$B = \{b_1, \ldots, b_m\}, m \leq n,$$

die Menge der Ausprägungen der Teilstichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Dann definiert man für $1 \leq i \leq k$, $1 \leq j \leq m$,

$$h_{ij} := h(a_i, b_j) := \sum_{1 \leq r \leq n} \sum_{1 \leq t \leq n} 1_{\{(a_i, b_j)\}} ((x_r, y_t)) .$$

h_{ij} bezeichnet also die absolute Häufigkeit der Merkmalskombination (a_i, b_j) in der bivariaten Stichprobe.

Als **Kontingenztafel der absoluten Häufigkeiten** bezeichnet man dann die $k \times m$ Matrix (bzw. die entsprechende Tabelle)

$$\mathbf{K} := \begin{pmatrix} h_{11} & \ldots & h_{1m} \\ h_{21} & \ldots & h_{2m} \\ \vdots & & \vdots \\ h_{k1} & \ldots & h_{km} \end{pmatrix}$$

oder auch $\mathbf{K}^\top$. Ganz analog ist mit den relativen Häufigkeiten $f_{ij} := \frac{h_{ij}}{n}$ anstelle der absoluten Häufigkeiten h_{ij} die **Kontingenztafel der relativen Häufigkeiten** definiert.

Die Kontingenztafel einer bivariaten Stichprobe wird genauer als **2-dimensionale Kontingenztafel** bezeichnet. Entsprechend erhält man für eine p-variate Stichprobe mit $p > 2$ dann eine **p-dimensionale Kontingenztafel**. Bei Pruscha [7], S. 181 ff., werden als Beispiele für mehrdimensionale Kontingenztafeln 3- und 4-dimensionale Kontingenztafeln erläutert.

Zusätzlich zu den Häufigkeiten h_{ij} bzw. f_{ij} sind die absoluten **Randhäufigkeiten**

$$h_{i.} := \sum_{j=1}^{m} h_{ij} \quad \text{und} \quad h_{.j} := \sum_{i=1}^{k} h_{ij}$$

für $i = 1, \ldots, k$ und $j = 1, \ldots, m$ und ganz analog die relativen Randhäufigkeiten von Interesse. Als Tabelle erhält man mit den Randhäufigkeiten eine Kontingenztafel der absoluten Häufigkeiten der Form

h_{11}	$, \ldots,$	h_{1m}	$, h_{1.}$
h_{21}	$, \ldots,$	h_{2m}	$, h_{2.}$
$\vdots$		$\vdots$	$\vdots$
h_{k1}	$\ldots,$	h_{km}	$h_{k.}$
$h_{.1}$	$\ldots,$	$h_{.m}$	n

Der Eintrag n rechts unten in der Kontingenzabelle entspricht der Summe der Zeilen- oder Spaltenhäufigkeiten, die sich jeweils zum Stichprobenumfang n addieren.

Um Hinweise auf einen eventuell vorliegenden Zusammenhang der beiden Merkmale X und Y zu gewinnen, bildet man die bedingten relativen Häufigkeitsverteilungen.

Definition 2.23 (Bedingte relative Häufigkeitsverteilung) *Die bedingte relative Häufigkeitsverteilung von Y gegeben die Bedingung $X = a_i$, $i = 1 \ldots, k$, ist durch die relativen Häufigkeiten*

$$f_Y(b_1|a_i) := \frac{h_{i1}}{h_{i.}}, \ldots, f_Y(b_m|a_i) := \frac{h_{im}}{h_{i.}} \tag{2.11}$$

definiert. Die bedingte relative Häufigkeitsverteilung von X gegeben die Bedingung $Y = b_j$, $j = 1, \ldots, m$, ist durch die relativen Häufigkeiten

$$f_X(a_1|b_j) := \frac{h_{1j}}{h_{.j}}, \ldots, f_X(a_k|b_j) := \frac{h_{kj}}{h_{.j}} \tag{2.12}$$

definiert. Man setzt dabei voraus, dass die in den Nennern auftretenden Randhäufigkeiten nicht identisch 0 sind.

Beispiel 2.24 Von 100.000 Versicherungsnehmern ist jeweils das Geschlecht Y (Ausprägungen: $b_1 :=$ weiblich und $b_2 :=$ männlich) und der berufliche Status X in den Ausprägungen:

$$a_1 := \text{ohne Beruf},\ a_2 := \text{angestellt},\ a_3 := \text{selbständig}$$

gegeben. Die bivariate Stichprobe sei in der folgenden Kontingenztabelle der absoluten Häufigkeiten zusammengefasst.

	ohne Beruf	angestellt	selbständig	$\sum$
weiblich	2400	28 910	12 460	43 770
männlich	2320	31 470	22 440	56 230
$\sum$	4720	60 380	34 900	100 000

Als bedingte relative Häufigkeitsverteilungen von X unter der Bedingung $Y = b_1$ bzw. unter der Bedingung $Y = b_2$ ergibt sich

$$\begin{aligned}
f_X(a_1|b_1) &= \frac{2400}{43\,770} \approx 0{,}055 \qquad \text{bzw. } f_X(a_1|b_2) = \frac{2320}{56\,230} \approx 0{,}041,\\
f_X(a_2|b_1) &= \frac{28\,910}{43\,770} \approx 0{,}660 \qquad \text{bzw. } f_X(a_2|b_2) = \frac{31\,410}{56\,230} \approx 0{,}559,\\
f_X(a_3|b_1) &= \frac{12\,460}{43\,770} \approx 0{,}285 \qquad \text{bzw. } f_X(a_3|b_2) = \frac{22\,440}{56\,230} \approx 0{,}398.
\end{aligned}$$

Aufgrund der Werte kann man einen potentiellen Zusammenhang zwischen dem beruflichen Status und dem Geschlecht vermuten, da z. B. der Selbständigen-Anteil unter den Frauen deutlich geringer ist, als bei den Männern. □

Besteht zwischen den beiden einer Kontingenztabelle zugrundeliegenden Merkmalen X und Y kein Zusammenhang, würde man erwarten, dass für die bedingten relativen Häufigkeiten gilt

$$f_Y(b_j|a_i) \approx f_Y(b_j|a_l) \approx \frac{h_{\cdot j}}{n} \quad \forall\, i, l = 1, \ldots, k \text{ und } j = 1, \ldots, m$$

und

$$f_X(a_i|b_j) \approx f_X(a_i|b_r) \approx \frac{h_{i\cdot}}{n} \quad \forall\, j, r = 1, \ldots, m \text{ und } i = 1, \ldots, k.$$

D. h. die bedingte realtive Häufigkeit einer Ausprägung hängt nicht von der Wahl der Ausprägung ab, bzgl. der man die Häufigkeit bedingt. Diese Überlegung führt zu der folgenden Definition.

Definition 2.25 *Eine bivariate Stichprobe* (x_i, y_i), $i = 1, \ldots, n$, *zweier mindestens nominal skalierter Merkmale* X *und* Y *sei in einer* $k \times m$*-Kontingenztabelle der absoluten Häufigkeiten mit den Randhäufigkeiten* $h_{i.}$ *und* $h_{.j}$, $i = 1, \ldots, k$, $j = 1, \ldots, m$, *zusammengefasst. Unter der Annahme, dass zwischen den Merkmalen* X *und* Y *kein Zusammenhang besteht, heißt*

$$\widetilde{h}_{ij} := \frac{h_{i.} h_{.j}}{n}$$

die ***erwartete Häufigkeit bei Unabhängigkeit*** *für die Merkmalskombination* (a_i, b_j), $i = 1, \ldots, k$, $j = 1, \ldots, m$.

2.3 Lage- und Streuungsmaße

In diesem Abschnitt werden die am häufigsten verwendeten Maßzahlen für die zentrale Lage und die Streuung einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ vorgestellt.

2.3.1 Lagemaße einer Stichprobe

Im Folgenden sei $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ eine Stichprobe eines Merkmals X und $A := \{a_1, \ldots, a_m\}$ die Menge aller Ausprägungen in der Stichprobe.

Definition 2.26 (Arithmetisches Mittel) *Ist* X *ein metrisches Merkmal, dann heißt*

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.13}$$

das ***arithmetische Mittel*** *(oder auch* ***empirischer Mittelwert****) der Stichprobe* $\boldsymbol{x}$.

Für eine Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit arithmetischem Mittel $\overline{x}$ gilt für das arithmetische Mittel $\overline{y}$ der linear transformierten Stichprobe

$$y_i := a x_i + b,\ i = 1, \ldots, n, \text{ mit } a, b \in \mathbb{R},$$

die Beziehung

$$\overline{y} = a\overline{x} + b.$$

Definition 2.27 (Modus) *Besitzt das Merkmal X mindestens nominales Skalenniveau, dann heißt*

$$x_{\text{Mod}} := \arg\max_{a \in A} \sum_{i=1}^{n} 1_{\{a\}}(x_i)$$

*der **Modus** (**Modalwert**) der Stichprobe $\boldsymbol{x}$.*

Der Modus x_{Mod} ist demnach jede Ausprägung der Stichprobe, die maximale Häufigkeit besitzt. Man beachte, dass der Modus nicht eindeutig bestimmt sein kann.

Ein weiteres wichtiges Lagemaß, dass nur ordinales Skalenniveau voraussetzt, ist der **empirische Median** $x_{\frac{1}{2}}$ einer Stichprobe, der bereits in Abschn. 2.2.3 als Spezialfall eines empirischen Quantils eingeführt wurde. Für ungeradzahligen Stichprobenumfang n gilt

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

und im Fall eines geradzahligen Stichprobenumfangs n erhält man das Medianintervall

$$x_{\frac{1}{2}} \in \left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right].$$

Im Fall einer metrischen Stichprobe kann der empirische Median auch eindeutig als

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

definiert werden.

Während der Median die zentrale Lage einer Stichprobe als empirisches 50 %-Quantil beschreibt, bildet das arithmetische Mittel den Schwerpunkt der Stichprobenwerte als Lagemaß der Stichprobe.

Satz 2.28 (Eigenschaften des arithmetischen Mittels) *Sei $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ eine Stichprobe eines metrischen Merkmals, dann gilt die Schwerpunktseigenschaft*

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0 \tag{2.14}$$

und die Minimierungseigenschaft

$$\arg\min_{z \in \mathbb{R}} \sum_{i=1}^{n} (x_i - z)^2 = \overline{x}. \tag{2.15}$$

Beweis Die Eigenschaft (2.14) folgt sofort aus der Defintion (2.13) und einfachem Nachrechnen. Zum Nachweis der Minimierungseigenschaft (2.15) bildet man die Ableitung

$$\frac{d}{dz}\sum_{i=1}^{n}(x_i - z)^2 = -2\sum_{i=1}^{n}(x_i - z)$$

und erhält dann über die Bedingung

$$-2\sum_{i=1}^{n}(x_i - z) = 0$$

die Behauptung. □

Der empirische Median einer Stichprobe minimiert die Betragsabstände zu den Beobachtungswerten.

Satz 2.29 (Minimierungseigenschaften des Medians) *Sei* $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ *eine Stichprobe eines metrischen Merkmals, dann gilt*

$$\arg\min_{z\in\mathbb{R}}\sum_{i=1}^{n}|z - x_i| = x_{\frac{1}{2}}.$$

Beweis Für alle $z \neq x_i$, $i = 1, \ldots, n$, besitzt die Funktion

$$h : \mathbb{R} \to \mathbb{R},\ h(z) := \sum_{i=1}^{n}|z - x_i|,$$

die Ableitung

$$\frac{\mathrm{d}}{\mathrm{d}z}h(z) = \sum_{i=1}^{n}\operatorname{sgn}(z - x_i).$$

Ist n ungeradzahlig, gilt für alle $z \neq x_i$, $i = 1, \ldots, n$,

$$\frac{\mathrm{d}}{\mathrm{d}z}h(z)\begin{cases} < 0, & \text{falls } z < x_{(\frac{n+1}{2})} \\ > 0, & \text{falls } z > x_{(\frac{n+1}{2})} \end{cases}$$

Da h stetig auf ganz $\mathbb{R}$ ist, folgt mit dem Mittelwertsatz der Differentialrechnung, dass h in $(-\infty, x_{(\frac{n+1}{2})}]$ streng monoton fallend und in $[x_{(\frac{n+1}{2})}, \infty)$ streng monoton wachsend ist, d. h. h besitzt an der Stelle $z = x_{(\frac{n+1}{2})}$ ein globales Minimum.

Ist n geradzahlig, gilt für alle $z \neq x_i, i = 1, \ldots, n$,

$$\frac{\mathrm{d}}{\mathrm{d}z} h(z) \begin{cases} < 0\,, \text{ falls } z < x_{\left(\frac{n}{2}\right)} \\ > 0\,, \text{ falls } z > x_{\left(\frac{n}{2}+1\right)} \\ = 0\,, \text{ falls } z \in [x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}] \end{cases}$$

Da h stetig auf ganz $\mathbb{R}$ ist folgt wieder mit dem Mittelwertsatz der Differentialrechnung, dass h in $(-\infty, x_{\left(\frac{n}{2}\right)}]$ streng monoton fallend und in $[x_{\left(\frac{n}{2}+1\right)}, \infty)$ streng monoton wachsend ist. D. h. alle $z \in [x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}]$ sind Stellen globaler Minima von h, insbesondere auch $z = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$. □

Der Modus x_{Mod} einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit der Ausprägungsmenge A besitzt die Minimierungseigenschaft

$$x_{\text{Mod}} = \arg\min_{z \in A} \sum_{i=1}^{n} \left(1 - 1_{\{z\}}(x_i)\right).$$

Bemerkung 2.30

a) Für die Berechnung des Modus wird nur nominales Skalenniveau vorausgesetzt, während der empirische Median erst bei mindestens ordinal skalierten Merkmalen verwendet werden kann. Das arithmetische Mittel setzt ein kardinales Skalenniveau in den Daten voraus.

b) Das arithmetische Mittel $\overline{x}$ reagiert sehr sensibel auf das Auftreten von extremen Werten innerhalb einer Stichprobe und kann durch Ausreißerwerte oder falsche Werte in einem Datensatz verzerrt werden. Der empirische Median $x_{\frac{1}{2}}$ hingegen verhält sich robust bzgl. extremer Werte.

c) Bei der Stichprobe eines kardinal skalierten Merkmals werden in der Anwendung oft alle drei Lagemaße, d. h. arithmetisches Mittel $\overline{x}$, empirischer Median $x_{\frac{1}{2}}$ und der Modus x_{Mod}, berechnet. Durch die Lage der drei Maßzahlen zueinander kann die Schiefe bzw. Symmetrie einer Stichprobenverteilung charakterisiert werden. Bei unimodalen Verteilungen gilt
symmetrische Verteilung: $x_{\text{Mod}} \approx x_{\frac{1}{2}} \approx \overline{x}$,
rechtsschiefe Verteilung: $x_{\text{Mod}} < x_{\frac{1}{2}} < \overline{x}$,
linksschiefe Verteilung: $x_{\text{Mod}} > x_{\frac{1}{2}} > \overline{x}$.
Eine große Abweichung von $\overline{x}$ und $x_{\frac{1}{2}}$ kann auch ein Hinweis auf das Vorliegen von extremen Stichprobenwerten, eventuellen Ausreißern oder auch von falschen Datenwerten sein.
Unterscheiden sich die Lagemaße stark, muss je nach Anwendung entschieden werden, welches Lagemaß mit seiner eigenen Interpretation der zentralen Lage einer

Stichprobe die geeignete Kennzahl für die Beschreibung der zentralen Lage der Stichprobe darstellt.

d) Bei unimodalen Häufigkeitsverteilungen verwendet man als deskriptive Maßzahlen für die Form (Schiefe und Wölbung) der Verteilung die **Schiefe** und den **Exzess** (**Kurtosis**), vgl. z. B. Hartung et al. [5], S. 47 - 49.

Im Fall von i. i. d. Stichprobenvariablen $X_1, \ldots, X_n$ mit existierendem Erwartungswert $\mu := E(X_1)$ und Varianz $\mathrm{Var}(X_1)$ ist das arithmetische Mittel

$$\widehat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

ein erwartungstreuer und konsistenter Schätzer für μ, vgl. z. B. Pruscha [6], S. 19.

2.3.2 Streuungsmaße einer Stichprobe

Streuungsmaße sind Kennzahlen einer Stichprobe, die die Schwankung bzw. Variabilität der Stichprobenwerte charakterisieren. Neben der empirischen Varianz, die die Streuung der Stichprobenwerte als quadratische Abweichung vom arithmetischen Mittelwert beschreibt, gibt es noch weitere Maßzahlen, die die Streuung auf andere Weise messen.

Definition 2.31 *Sei* $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ *eine Stichprobe eines metrischen Merkmals mit arithmetischem Mittel* $\overline{x}$, *Median* $x_{\frac{1}{2}}$, *den empirischen Quartilen* $x_{\frac{1}{4}}$ *bzw.* $x_{\frac{3}{4}}$, *minimalem Stichprobenwert* $x_{(1)}$ *und maximalem Stichprobenwert* $x_{(n)}$, *dann heißt*

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \text{ die } \textbf{\textit{empirische Varianz}},$$

$$\delta := \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}| \text{ die } \textbf{\textit{mittlere absolute Abweichung vom Mittelwert}},$$

$$\delta_{\mathrm{Med}} := \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{\frac{1}{2}}| \text{ die } \textbf{\textit{mittlere absolute Abweichung vom Median}},$$

$$MAD := \text{Median der Stichprobe } \left(|x_1 - x_{\frac{1}{2}}|, \ldots, |x_n - x_{\frac{1}{2}}|\right)^\top \text{ die } \textbf{\textit{Median-Deviation}},$$

$$R := x_{(n)} - x_{(1)} \text{ die } \textbf{\textit{Spannweite (range)}},$$

$$IQD := x_{\frac{3}{4}} - x_{\frac{1}{4}} \text{ die } \textbf{\textit{Inter-Quartil-Distanz}}$$

der Stichprobe $\boldsymbol{x}$.

In rein deskriptiven Anwendungen (speziell bei der Betrachtung von Grundgesamtheiten) wird die empirische Varianz manchmal auch in der modifizierten Form

$$\tilde{s}^2 := \frac{n-1}{n} s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

verwendet.

Die Quadratwurzel der empirischen Varianz

$$s := \sqrt{s^2} \text{ bzw. } \tilde{s} := \sqrt{\tilde{s}^2}$$

wird als **empirische Standardabweichung** bezeichnet.

Eine Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit empirischer Varianz $s^2 = 0$ bzw. empirischer Standardabweichung $s = 0$ besitzt minimale Streuung, d. h. alle Stichprobenwerte x_i, $i = 1, \ldots, n$, sind identisch.

Die Spannweite R einer Stichprobe ist als Differenz von maximaler und minimaler Ausprägung innerhalb der Stichprobe ein sehr anschauliches Streuungsmaß, allerdings ist sie sehr anfällig für den Einfluß extremer Stichprobenwerte. Die Inter-Quartil-Distanz *IQD* gibt den Abstand der in der geordneten Stichprobe zentral gelegenen 50 % der Stichprobenwerte an und verhält sich weit robuster gegenüber extremen Stichprobenwerten.

Die Streuungsmaße δ_{Med}, *MAD*, R und *IQD* können auch im Fall ordinal skalierter Daten verwendet werden.

Die besondere Rolle der empirischen Varianz s^2 in der induktiven Statistik zeigt der folgende Satz, vgl. z. B. Pruscha [6], S. 20.

Satz 2.32 (Eigenschaften des Varianz-Schätzers) *Im Fall von i. i. d. Stichprobenvariablen $X_1, \ldots, X_n$ mit existierendem Erwartungswert $\mu = E(X_1)$ und Varianz $\sigma^2 = Var(X_1) > 0$ ist der Varianz-Schätzer*

$$\widehat{\sigma}_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2,$$

wobei $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$, erwartungstreu und konsistent für σ^2.

Der entsprechend der modifizierten empirischen Varianz $\tilde{s}^2 := \frac{n-1}{n} s^2$ gebildete Varianz-Schätzer ist ebenfalls konsistent, aber nur asymptotisch erwartungstreu für σ^2, d. h.

$$\lim_{n \to \infty} E_{\sigma^2} \left(\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \right) = \sigma^2 \text{ für alle } \sigma^2 > 0.$$

Für die empirische Varianz s^2 einer Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit arithmetischen Mittel $\overline{x}$ gilt die **Verschiebungsformel**

$$(n-1)s^2 = \sum_{i=1}^{n} (x_i - c)^2 - n\,(\overline{x} - c)^2\,,$$

für beliebige $c \in \mathbb{R}$.

Für eine Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit empirischer Varianz s_x^2 gilt für die empirische Varianz s_y^2 der linear transformierten Stichprobe

$$y_i := ax_i + b,\ i = 1, \ldots, n,\ \text{mit}\ a, b \in \mathbb{R},$$

die Beziehung

$$s_y^2 = a^2 s_x^2.$$

Bei multiplikativen Maßstabsumrechnungen durch einen Faktor $a \in \mathbb{R}$ muss also beachtet werden, dass die empirische Varianz entsprechend maßstabsabhängig ist, während sich die empirische Varianz einer Stichprobe nicht verändert, wenn alle Stichprobenwerte nur um eine additive Konstante $b \in \mathbb{R}$ verschoben werden.

Für den Vergleich der Streuungen von Stichproben mit unterschiedlichen arithmetischen Mitteln verwendet man den **Variationskoeffizienten**. Der Variationskoeffizient ist ein relatives Streuungsmaß und ist invariant bzgl. Stichprobentransformationen, bei denen die Stichprobenwerte mit einem konstanten Faktor multipliziert werden.

Definition 2.33 (Variationskoeffizient) *Sei $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit $x_i > 0$, $i = 1, \ldots, n$, eine Stichprobe eines verhältnisskalierten Merkmals mit arithmetischem Mittel $\overline{x}$ und empirischer Varianz s^2, dann nennt man*

$$v := \frac{s}{\overline{x}}$$

den Variationskoeffizienten von $\boldsymbol{x}$.

Der Variationskoeffizient misst die empirische Standardabweichung s in Einheiten des artihmetischen Mittels $\overline{x}$. Oft wird der Variationskoeffizient auch als prozentuale Größe, z. B. $v = \frac{2}{10} = 20\,\%$ angegeben.

Bemerkung 2.34 Neben Lage-, Streuungsmaßen und Maßzahlen zur Schiefe und Wölbung einer Häufigkeitsverteilung werden in der Anwendung häufig **Konzentrationsmaße** verwendet. Konzentrationsmaße quantifizieren für eine Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ mit

$x_i > 0, i = 1, \ldots, n$, eines metrisch skalierten Merkmals, wie sich die Stichprobensumme $\sum_{i=1}^{n} x_i$ auf die n Untersuchungseinheiten aufteilt. Neben der grafischen Darstellung der relativen Konzentration mithilfe der **Lorenzkurve** verwendet man hier häufig als Maßzahl den (aus der Lorenzkurve abgeleiteten) **Gini-Koeffizienten** (**Gini Index**) , vgl. z. B. Hartung et al. [5], S. 50–55. Konzentrationsmaße beschreiben z. B. wie sich der Gesamtschadenbedarf in einem Kollektiv auf die einzelnen Versicherungsverträge aufteilt oder ob eine Kreditausfallsumme auf einzelne Kreditverträge konzentriert ist. Liegt eine gleichmäßige Aufteilung der Stichprobensumme auf die Untersuchungseinheiten vor, spricht man von einer Null-Konzentration.

2.4 Grafische und explorative Methoden

Neben der Beschreibung einer Stichprobe mittels Kennzahlen werden oft grafische Darstellungsformen verwendet. Mit speziellen Grafiken können nicht nur die Charakteristika von Stichproben visualisiert werden, sondern auch Hypothesen abgeleitet werden, die dann mit induktiven Verfahren weiter untersucht werden. Als Standardwerk für explorative Datenanalyse gilt Tukey [10]. Einen Überblick zu grafischen Verfahren in der statistischen Datenanalyse geben z. B. auch Chambers et al. [1].

2.4.1 Streudiagramm

Gegeben sei eine bivariate Stichprobe $(x_i, y_i)^\top$, $i = 1 \ldots, n$, zweier metrischer oder ordinaler Merkmale mit den Teilstichproben

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top \text{ und } \boldsymbol{y} = (y_1, \ldots, y_n)^\top.$$

Die Darstellung der Punkte (x_i, y_i) , $i = 1 \ldots, n$, in einem kartesischen Koordinatensystem nennt man **Streudiagramm** oder auch **Scatter-Plot** der bivariaten Stichprobe bzw. der Teilstichproben.

Der folgenden Abb. 2.8 liegt eine bivariate Stichprobe zugrunde, in der für verschiedene Versicherungsnehmer jeweils das Alter und die Schadensumme in einem bestimmten Zeitintervall erfasst sind. Ein Scatter-Plot gibt Hinweise auf den möglichen Zusammenhang zweier Merkmale bzw. der zugrundeliegenden Zufallsvariablen. Bei ordinal skalierten Merkmalen kann ein Streudiagramm nur einen monotonen Zusammenhang der Merkmale verdeutlichen, während für eine bivariate Stichprobe metrischer Merkmale auch ein funktionaler Zusammenhang der Merkmale erkannt werden kann. Von besonderem Interesse ist oft die Frage, ob ein linearer Zusammenhang besteht. In diesem Fall spricht man von **Korrelation** der Merkmale oder der Teilstichproben (bzw. der Stichprobenvariablen).

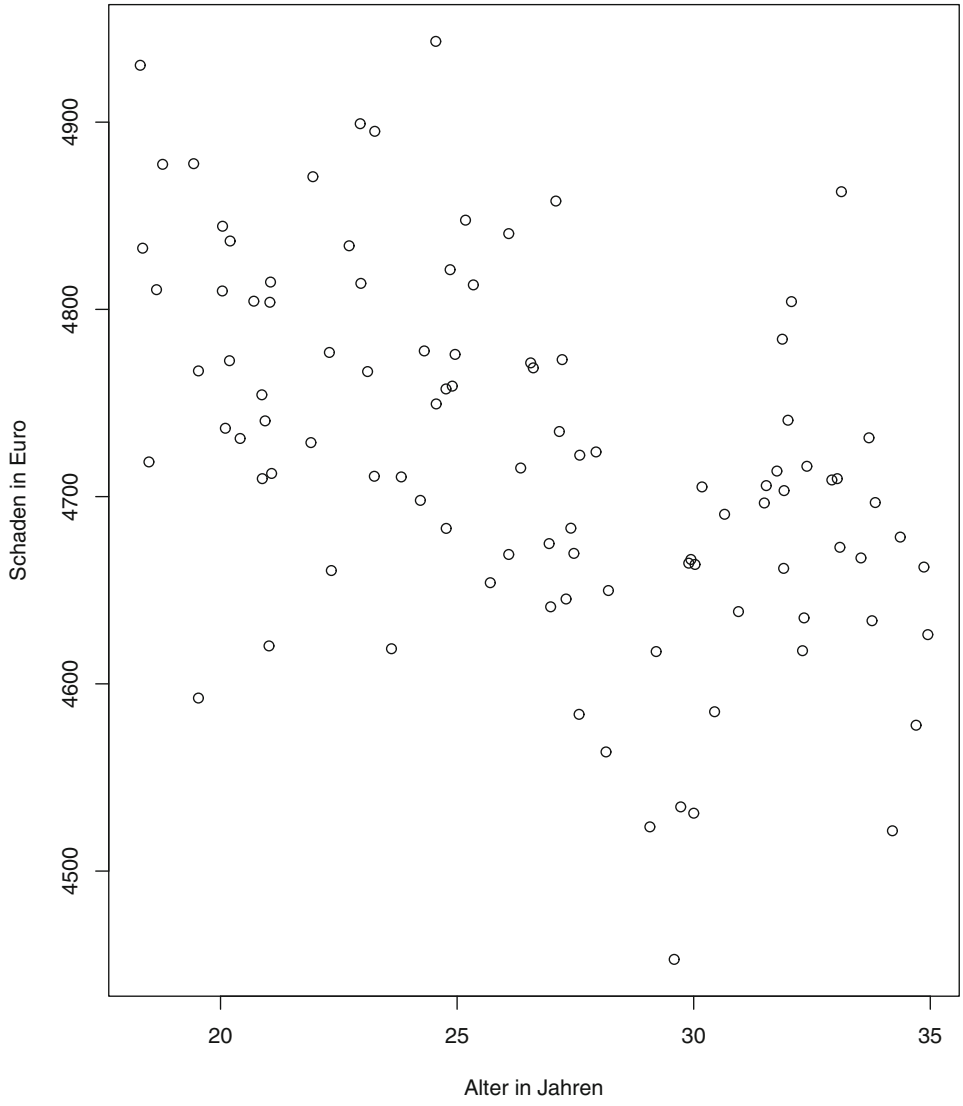

Abb. 2.8 Streudiagramm einer bivariaten Stichprobe, die das Alter und die Schadensumme von 100 Versicherungsnehmern beinhaltet

2.4.2 Box-Whisker-Plot

Ein **Box-Plot** oder **Box-Whisker-Plot** ist eine explorative Methode, um den Median $x_{\frac{1}{2}}$, das untere und obere Quartil ($x_{\frac{1}{4}}$ und $x_{\frac{3}{4}}$) und den Minimal- und Maximalwert ($x_{(1)}$ und $x_{(n)}$) einer Stichprobe $\boldsymbol{x}$ innerhalb einer Grafik darzustellen. Der Bereich zwischen den

Quartilen, d. h. der Bereich der mittleren 50 % der Daten, wird als Kasten (Box) visualisert. Der Median ist in einem Box-Whisker-Plot als eine Linie (manchmal auch als Kreis) dargestellt, die den Kasten zwischen den Quartilen in zwei Bereiche aufteilt. Die Form der Darstellung des Bereichs zwischen Minimal- und Maximalwert erinnert an einen Schnurrbart (engl.: whisker). In der Regel ist ein Box-Whisker-Plot mit einer Skala versehen, die die Zahlenwerte der dargestellten Größen erkennen lässt.

In modifizierten Formen des Box-Whisker-Plots werden anstelle des Minimalwerts $x_{(1)}$ und Maximalwerts $x_{(n)}$ der Stichprobe andere Grenzen für die Definition der Schnurrbartenden verwendet und Extremwerte, d. h. potentielle Ausreisserwerte, in der Grafik gesondert ausgewiesen. Eine oft verwendete Variante als Ersatz für $x_{(1)}$ und $x_{(n)}$ ist der kleinste Stichprobenwert größer als $x_{\frac{1}{4}} - c \cdot IQD$ und der größte Stichprobenwert kleiner als $x_{\frac{3}{4}} + c \cdot IQD$, mit $c = \frac{3}{2}$ oder auch $c = 3$. Das Intervall

$$\left[x_{\frac{1}{4}} - \frac{3}{2} \cdot IQD,\ x_{\frac{3}{4}} + \frac{3}{2} \cdot IQD \right]$$

stellt einen Bereich der unaufälligen Streuung dar. Die Stichprobenwerte, die außerhalb der Schurrbartenden liegen, werden im Box-Whisker-Plot als mögliche Ausreisser z. B. durch Kreise gekennzeichnet. Manchmal werden zur Definition des Bereichs der unaufälligen Streuung auch die Dezile $x_{\frac{1}{10}}$ und $x_{\frac{9}{10}}$ verwendet und dann die Stichprobenwerte, die nicht im Intervall $\left[x_{\frac{1}{10}}, x_{\frac{9}{10}} \right]$ liegen, gesondert gekennzeichnet.

Die Abb. 2.9 zeigt schematisch einen Box-Whisker-Plot, in dem der kleinste Stichprobenwert größer als $x_{\frac{1}{4}} - \frac{3}{2} \cdot IQD$ und der größte Stichprobenwert kleiner als $x_{\frac{3}{4}} + \frac{3}{2} \cdot IQD$ zur Festlegung der Schnurrbartenden verwendet werden und die so definierten Extremwerte mit Kreissymbolen gesondert gekennzeichnet sind.

Ein Box-Whisker-Plot eignet sich nicht nur zur Darstellung der Verteilung einer Stichprobe, sondern besonders für den Vergleich mehrere Stichprobenverteilungen hinsichtlich Lage und Streuung.

Man betrachtet z. B. eine bivariate Stichprobe $(x_i, y_i)^\top, i = 1, \ldots, n$, bestehend aus Werten eines metrischen Merkmals X (z. B. die Schadensumme eines Versicherungsnehmers) und eines nominalen Merkmals Y (z. B. der Beruf eines Versicherungsnehmers). Aufgeteilt nach den k Ausprägungen des nominalen Merkmals (man spricht hier auch von den Stufen eines Faktors) erhält man k Teilstichproben des metrischen Merkmals X. Nun ist die Fragestellung von Interesse, ob sich die Verteilungen der k Teilstichproben des metrischen Merkmals hinsichtlich Lage bzw. Streuung unterscheiden.

Die Abb. 2.10 zeigt Box-Whisker-Plots für die (Teil-)Stichproben $\boldsymbol{x}$, $\boldsymbol{y}$ und $\mathbf{z}$, die aus jeweils 100 erzeugten Pseudozufallszahlen bestehen. Für $\boldsymbol{x}$ wurde eine Standardnormalverteilung, für $\boldsymbol{y}$ eine $\mathcal{N}(0,5)$-Verteilung und für $\mathbf{z}$ eine $\mathcal{N}(5,1)$-Verteilung zur Erzeugung der Zufallszahlen verwendet. Box-Whisker-Plots wie in der Abb. 2.10 lassen dem Anwender Streuungs- und Lageunterschiede in den Teilstichproben vermuten. Mit Methoden der **Varianzanalyse**, vgl. z. B. Sachs und Hedderich [8], S. 577 ff., werden Hypothesen zu Lageunterschieden in den Teilstichproben dann in induktiver Weise weiter untersucht.

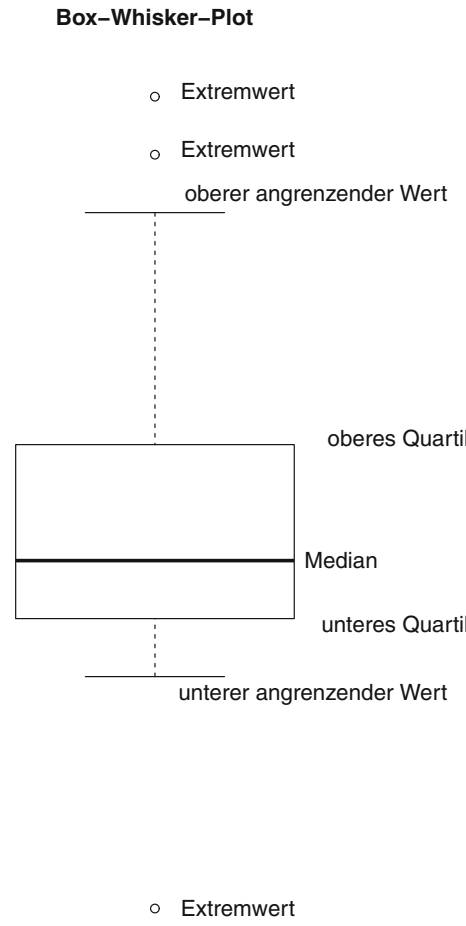

Abb. 2.9 Schematische Darstellung: Box-Whisker-Plot mit oberer angrenzender Wert := größter Stichprobenwert kleiner als $x_{\frac{3}{4}} + \frac{3}{2} \cdot IQD$, unterer angrenzender Wert := kleinster Stichprobenwert größer als $x_{\frac{1}{4}} - \frac{3}{2} \cdot IQD$

Durch Box-Whisker-Plots erhält man auch Hinweise auf die Schiefe einer Stichprobenverteilung. Dazu betrachtet man u. a. die Lage des Medians innerhalb der Box.

Alternativ können die Stichprobenverteilungen mehrerer Teilstichproben auch grafisch durch Diagramme, die die arithmetischen Mittel und z. B. die empirischen Standardabweichungen der Teilstichproben enthalten, dargestellt und verglichen werden. Dabei ist allerdings zu beachten, dass diese Lage- und Streuungsmaße nicht robust gegenüber extremen Werten in den Stichproben sind. Für den Fall von symmetrischen Stichprobenverteilungen (ohne extreme Werte) sind der empirische Median und das arithmetische Mittel für große Stichprobenumfänge mit hoher Wahrscheinlichkeit annähernd identisch. Grafiken, die die arithmetischen Mittel und Vielfache von den empirischen Standardabweichungen verwenden, orientieren sich an entsprechend konstruierten Konfidenzintervallen für die unbekannten Erwartungswerte der den Teilstichproben zugrundeliegenden Stichprobenvariablen, die z. B. im Fall von unabhängig und normalverteilten Stichprobenvariablen diese Form besitzen.

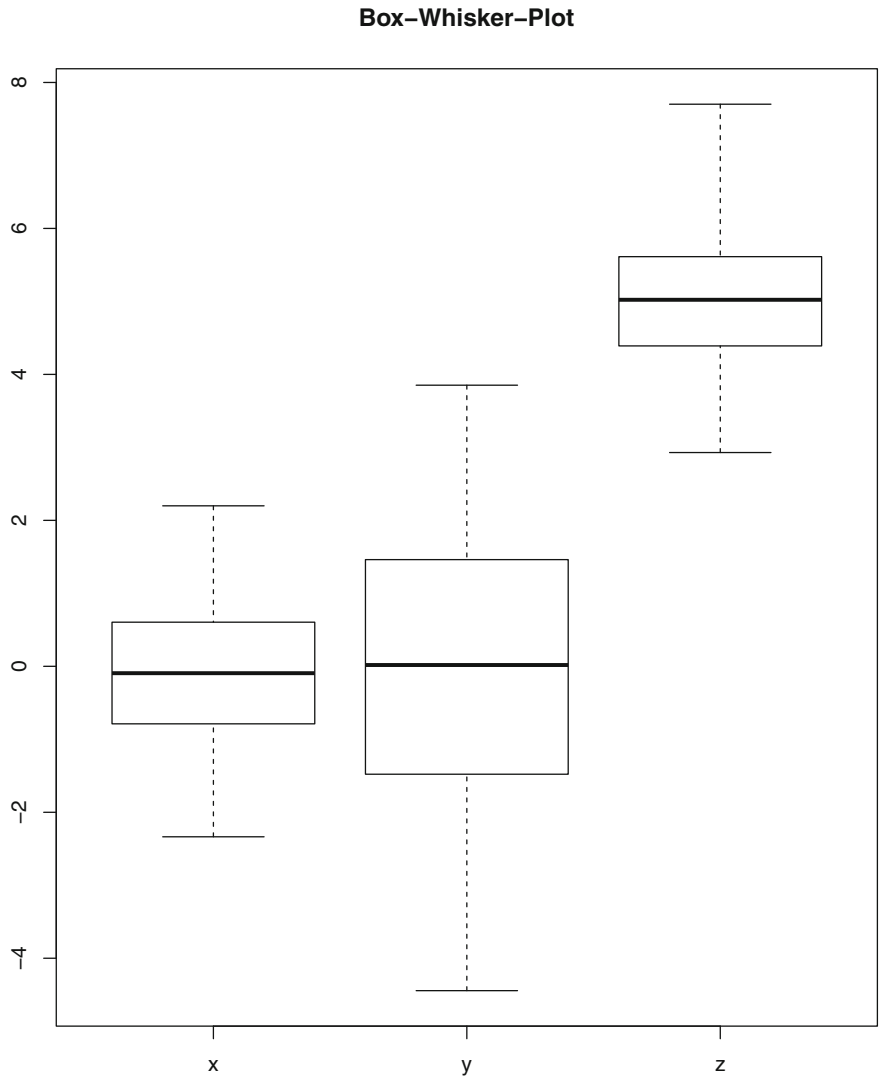

Abb. 2.10 Box-Whisker-Plot von drei Teilstichproben $\boldsymbol{x}$ (100 simulierte Realisationen einer $\mathcal{N}(0,1)$-verteilten Zufallsvariablen), $\boldsymbol{y}$ (100 simulierte Realisationen einer $\mathcal{N}(0,5)$-verteilten Zufallsvariabeln) und $\mathbf{z}$ (100 simulierte Realisationen einer $\mathcal{N}(5,1)$-verteilten Zufallsvariablen)

2.4.3 Mosaik-Plot

In einem **Mosaik-Plot** wird die Häufigkeitsverteilung einer p-variaten Stichprobe $(x_{i1}, \ldots, x_{ip})^\top$, $i = 1 \ldots, n$, von $p \geq 2$ nominalen Merkmalen $X_1, \ldots, X_p$ grafisch dargestellt. Für den Fall einer bivariaten Stichprobe (d. h. $p = 2$) bildet der Mosaik-Plot ei-

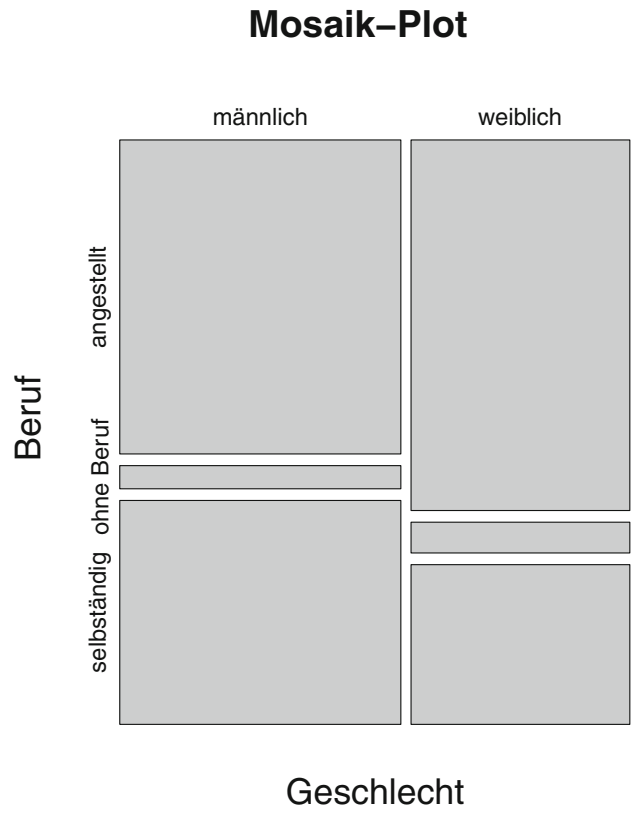

Abb. 2.11 Mosaik-Plot zu der bivariaten Stichprobe aus Beispiel 2.24

ne Visualisierung der entsprechenden 2-dimensionalen Kontingenztafel. Ein Mosaik-Plot gibt dem Anwender Hinweise, ob zwischen den betrachteten Merkmalen Zusammenhänge zu vermuten sind. Dazu betrachtet man wie in Beispiel 2.24 das Verhalten der bedingten Häufigkeiten.

Beispiel 2.35 Die Abb. 2.11 zeigt einen Mosaik-Plot zu der 2-dimensionalen Kontingenztafel aus Beispiel 2.24.

Die Ausprägungen des Merkmals *Geschlecht* sind am oberen Rand der Grafik angetragen und alle Daten werden nach den Ausprägungen *männlich* und *weiblich* in zwei Blöcke aufgeteilt. Die Aufteilung der Blöcke erfolgt dabei nach der Häufigkeit der Ausprägungen und führt daher hier zu unterschiedlichen Breiten der Teilblöcke. Man sieht, dass die Ausprägung *männlich* eine größere Häufigkeit besitzt, als die Ausprägung *weiblich*. Die Ausprägungen des Merkmals *Beruf* sind am linken Rand der Grafik angeordnet. In jedem der beiden durch das Merkmal *Geschlecht* bestimmten vertikalen Teilblöcke erfolgt eine weitere horizontale Unterteilung, die jeweils durch die entsprechenden bedingten Häufigkeiten definiert wird. Insgesamt erhält man eine Aufteilung in $2 \cdot 3 = 6$ Mosaik-Teile,

deren Flächen proportional zu den Häufigkeiten der Ausprägungskombinationen der beiden Merkmale sind. □

Mosaik-Plots können prinzipiell für beliebige p-variate Stichproben, $p \geq 2$, nominaler Merkmale erstellt werden. Allerdings werden die resultierenden Grafiken bei hoher Merkmalsanzahl p schnell unübersichtlich. Die Abb. 2.12 zeigt den Mosaik-Plot einer trivariaten Stichprobe mit $p = 3$ Merkmalen, der noch sehr gut interpretierbar ist.

Beispiel 2.36 In einer trivariaten Stichprobe $(x_i, y_i, z_i)^\top$, $i = 1, \ldots, 10.000$, seien für $n = 10.000$ Versicherungsnehmer die Merkmale $X :=$ *Geschlecht* (mit den Ausprägungen: *männlich* und *weiblich*), $Y :=$ *Berufsgruppe* (mit den Ausprägungen: A, B, C) und $Z :=$ *Schaden* (mit den Ausprägungen: *Ja* und *Nein*) in einer festgelegten Zeitperiode erfasst. Der zugehörige Mosaik-Plot ist in der Abb. 2.12 dargestellt. Als Erweiterung zum 2-dimensionalen Mosaik-Plot in Beispiel 2.35 wird nun noch die bedingte Häufigkeitsverteilung eines dritten Merkmals $Z =$ *Schaden* in die Grafik integriert. Das Merkmal *Schaden* wird zusätzlich an der oberen Seite der Grafik angeordnet und die, aus der Aufteilung nach den Häufigkeitsverteilungen der ersten beiden Merkmale X und Y resultierenden, Mosaik-Bereiche werden entsprechend der Häufigkeitsverteilung des dritten Merkmals Z jeweils in zwei Teilbereiche unterteilt. Man erkennt z. B., dass in der Gruppe der Männer in der Berufsgruppe A weniger Schaden-Fälle vorliegen, als in der Berufsgruppe B. □

Weitere Ausführungen zu Mosaik-Plots kann man z. B. bei Friendly [4] finden.

2.4.4 Quantile-Quantile-Plot

Ein **Quantile-Quantile-Plot** (kurz: Q-Q-Plot) ist eine grafische Methode zur Beurteilung von Verteilungsannahmen. Dazu werden in ein kartesisches Koordinatensystem die empirischen Quantile zweier Stichproben oder die empirischen Quantile einer Stichprobe und die theoretischen Quantile einer hypothetischen Verteilung gegeneinander angetragen. Eine sinnvolle Anwendung von Q-Q-Plots setzt Stichproben mit großen Stichprobenumfängen voraus.

Mit einem Q-Q-Plot kann für zwei Stichproben $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ und $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ metrischer Merkmale untersucht werden, ob die den beiden Stichproben zugrundeliegenden Stichprobenvariablen X_i, $i = 1, \ldots, n$, bzw. Y_j, $i = j, \ldots, m$, identisch verteilt sind.

Die zweite, wichtige Anwendung des Q-Q-Plots ist die Frage, ob die Stichprobenvariablen einer gegebenen Stichprobe eine spezielle, hypothetische Verteilung (z. B. eine Normalverteilung) besitzen. Man spricht in diesem Fall auch von einem **Wahrscheinlichkeits-Plot**.

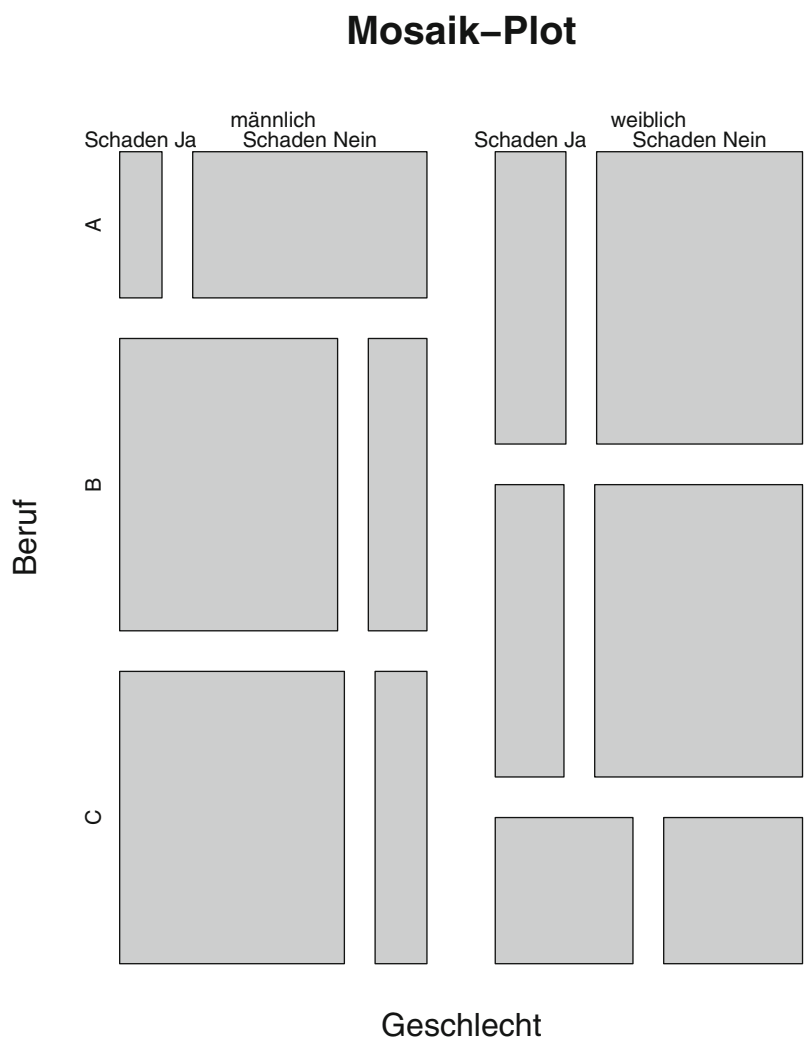

Abb. 2.12 Mosaik-Plot einer trivariaten Stichprobe

Wir betrachten zunächst den Fall zweier Stichproben

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top \text{ und } \boldsymbol{y} = (y_1, \ldots, y_m)^\top,\ n \leq m.$$

Wir nehmen an, dass die zugrundeliegenden Stichprobenvariablen X_i, $i = 1, \ldots, n$, und Y_j, $j = 1, \ldots, m$, alle identisch verteilt sind mit der stetigen, streng monotonen Verteilungsfunktion F.

Nach Lemma 2.20 bzw. Bemerkung 2.22 gilt dann für große Stichprobenumfänge n, m und alle $i = 1, \ldots, n-1$, dass

$$x_{\frac{i}{n}} \approx F^{-1}\left(\frac{i}{n}\right) = \kappa_{\frac{i}{n}} \text{ und auch } y_{\frac{i}{n}} \approx F^{-1}\left(\frac{i}{n}\right) = \kappa_{\frac{i}{n}},$$

wobei $x_{\frac{i}{n}}, y_{\frac{i}{n}}$ die empirischen $\frac{i}{n}$-Quantile der Stichprobe $\boldsymbol{x}$ bzw. $\boldsymbol{y}$ und $\kappa_{\frac{i}{n}}$ das theoretische $\frac{i}{n}$-Quantil der zugrundeliegenden Verteilung bezeichnet.

Für den Fall identisch verteilter Stichprobenvariablen erwartet man also, dass sich die empirischen Quantile der Stichprobe $\boldsymbol{x}$ und der Stichprobe $\boldsymbol{y}$ entsprechen. Im Q-Q-Plot werden die Punkte

$$(x_{\frac{i}{n}}, y_{\frac{i}{n}}),\ i = 1, \ldots, n-1,$$

oder z. B. auch die Punkte

$$(x_{(i)}, y_{(i)}),\ i = 1, \ldots, n,$$

eingetragen und sollten, falls die Stichprobenvariablen wirklich identisch verteilt sind, annähernd auf der Identitätsgeraden liegen.

Die Abb. 2.13 beinhaltet Q-Q-Plots für die Stichproben $\boldsymbol{x}$, bestehend aus 100 simulierten $\mathcal{N}(1,1)$-verteilten Zufallszahlen, $\boldsymbol{y}$, bestehend aus einer anderen, unabhängigen Simulation von 100 $\mathcal{N}(1,1)$-verteilten Pseudozufallszahlen und $\mathbf{z}$, bestehend aus 100 $\mathcal{E}(1)$-verteilten Pseudozufallszahlen.

Für den Fall, dass die Stichprobenvariablen Y_i eine lineare Transformation der Stichprobenvariablen X_i, $i = 1, \ldots, n$ sind, liegen die Punkte im Q-Q-Plot nicht mehr entlang der Identitätsgeraden, sondern sind um eine andere Sollgerade verteilt.

Lemma 2.37 (Lineare Transformation) *Seien X_i, $i = 1, \ldots, n$, i. i. d. Zufallsvariablen mit stetiger, streng monotoner Verteilungsfunktion F. Für die Zufallsvariablen $Y_1, \ldots, Y_n$ mit Verteilungsfunktion G gelte*

$$Y_i = a + bX_i,\ a \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\},\ \textit{für alle } i = 1, \ldots, n.$$

Dann gilt für die p-Quantile y_p, $0 < p < 1$, von G

$$y_p = a + bF^{-1}(p) = a + bx_p,$$

wobei x_p das p-Quantil von F bezeichnet.

Beweis Für alle $0 < p < 1$, $i = 1, \ldots, n$, und $a \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\}$ gilt

$$G(y_p) = P(Y_i \leq y_p) = P(a + bX_i \leq a + bx_p) = P(X_i \leq x_p) = F(x_p) = p. \quad \square$$

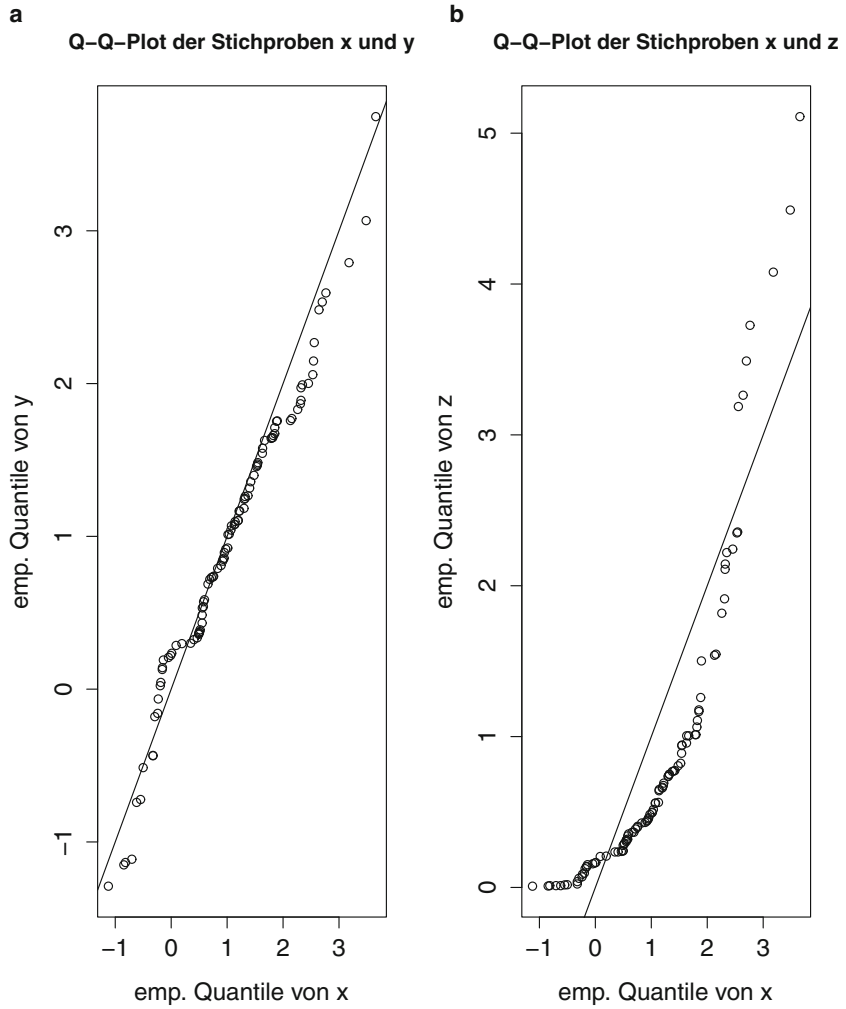

Abb. 2.13 **a** Q-Q-Plot der $\mathcal{N}(1,1)$-verteilten Stichproben $\boldsymbol{x}$ und $\boldsymbol{y}$. **b** Q-Q-Plot von $\boldsymbol{x}$ und der standardexponentialverteilten Stichprobe **z**. Im Fall identischer Verteilungen sollten die Punkte annähernd auf der eingezeichneten Identitätsgeraden liegen. Im Q-Q-Plot in (**b**) erkennt man eine deutliche Abweichung der Punkte von der Identitätsgeraden

Mit Lemma 2.37 folgert man für Stichprobenvariablen der Form

$$Y_i = a + bX_i,\ a \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\},\ \text{für alle } i = 1, \ldots, n,$$

dass hier die Punkte im Q-Q-Plot entsprechend entlang einer Sollgeraden mit Steigung b und Ordinatenabschnitt a ausgerichtet sind. Lageunterschiede werden demnach als Ver-

schiebung der Sollgeraden zur Identitätsgeraden angezeigt und Skalenunterschiede sind an der zu 1 verschiedenen Steigung der Sollgeraden zu erkennen.

Für den Fall, dass man für eine Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ eine hypothetische Verteilungsannahme mit stetiger, streng monotoner Verteilungsfunktion F der i.i.d Stichprobenvariablen $X_i, i = 1, \ldots, n$, überprüfen will, werden in einem Wahrscheinlichkeits-Plot die Punkte

$$\left(F^{-1}\left(\frac{i}{n}\right), x_{(i)}\right),\ i = 1, \ldots n-1,$$

oder auch

$$\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right),\ i = 1, \ldots n,$$

betrachtet. In der Praxis werden meist Punkte verwendet, die noch um eine Randkorrektur ergänzt sind, z. B.

$$\left(F^{-1}\left(\frac{i-\frac{1}{2}}{n}\right), x_{(i)}\right) \text{ für } n > 10 \text{ bzw. } \left(F^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right), x_{(i)}\right) \text{ für } n \leq 10.$$

Besitzen die Stichprobenvariablen X_i, $i = 1, \ldots, n$, die identische Verteilungsfunktion F, erwartet man wieder, dass die Punkte im Q-Q-Plot bei großem Stichprobenumfang n approximativ auf der Identitätsgeraden liegen.

In der Praxis bildet man Wahrscheinlichkeits-Plots oft mit einer standardisierten hypothetischen Verteilung der Stichprobenvariablen. So wird z. B. ein Q-Q-Plot zur Überprüfung einer Normalverteilungsannahme der Stichprobenvariablen oft mit der Standardnormalverteilung als hypothetische Verteilung gebildet. In vielen Anwendungen (z. B. bei der Überprüfung der Voraussetzungen für Signifikanztests) steht nur die Frage im Mittelpunkt, ob die Stichprobenvariablen als normalverteilt angenommen werden können, die Parameter der Normalverteilung sind hier nur von sekundärer Bedeutung.

Ist die wahre Verteilung der Stichprobenvariablen X_i, $i = 1, \ldots, n$, über eine lineare Transformation auf die hypothetische Verteilung zurückzuführen, so ergibt sich im Q-Q-Plot für genügend großen Stichprobenumfang approximativ ebenfalls ein linearer Trend der Punkte im Q-Q-Plot. Allerdings ist dann bei tatsächlichem Vorliegen der hypothetischen Verteilung im Allgemeinen nicht mehr die Identitätsgerade die Sollgerade, an der die Punkte ausgerichtet sind. Gilt für die Stichprobenvariablen

$$X_i = a + bY_i,\ a \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\},\ \text{für alle } i = 1, \ldots, n,$$

wobei die Zufallsvariablen Y_i die hypothetische Verteilung besitzen sollen, erwartet man mit Lemma 2.37, dass die Punkte des Q-Q-Plots bei großem Stichprobenumfang entlang einer Sollgeraden mit Steigung b und Ordinatenabschnitt a ausgerichtet sind.

In der praktischen Anwendung wird die Sollgerade in Q-Q-Plots aus den Punkten des Q-Q-Plots z. B. über eine einfache, lineare Regression oder, um den Einfluss von extremen Werten zu reduzieren, durch robuste Regressionsverfahren geschätzt.

Für den wichtigen Spezialfall, dass die hypothetische Verteilung die Standarnormalverteilung ist und die unabhängigen Stichprobenvariablen tatsächlich $\mathcal{N}(\mu, \sigma^2)$-verteilt sind, folgt wegen

$$X_i = \mu + \sigma Z_i,\ i = 1, \ldots, n,$$

wobei $Z_i, i = 1, \ldots, n$, unabhängige, standardnormalverteilte Zufallsvariablen bezeichnen, dass hier die Punkte eines Wahrscheinlichkeits-Plots entlang einer Sollgeraden mit Steigung $\sigma > 0$ und Ordinatenabschnitt $\mu \in \mathbb{R}$ liegen.

Man nennt in diesem Spezialfall den Q-Q-Plot auch **Normal Q-Q-Plot** oder **Normal-Wahrscheinlichkeits-Plot**.

In der Regel sind der Erwartungswert μ und die Standardabweichung σ unbekannt. Man könnte die Sollgerade durch Verwendung der üblichen Schätzwerte $\widehat{\mu} := \overline{x}$ (arithmetisches Mittel) für μ und $\widehat{\sigma} = s$ (empirische Standardabweichung) für σ approximieren oder mittels einfacher linearer Regression eine Schätzung der Sollgeraden bestimmen. In der Praxis verwendet man allerdings meist als robuste Schätzung der Sollgerade diejenige Gerade, die durch die unteren und oberen empirischen und theoretischen Quartile verläuft. D. h. die Gerade mit der Gleichung

$$y(x) = \frac{x_{\frac{3}{4}} + x_{\frac{1}{4}}}{2} + \frac{x_{\frac{3}{4}} - x_{\frac{1}{4}}}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)} \cdot x,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Das arithmetische Mittel des unteren und oberen emprischen Quartils

$$\frac{x_{\frac{3}{4}} + x_{\frac{1}{4}}}{2}$$

ist für symmetrische Verteilungen ein robuster Schätzwert für den Median, der im Fall der Normalverteilung mit dem Erwartungswert μ übereinstimmt. Der empirische Quartilsabstand $x_{\frac{3}{4}} - x_{\frac{1}{4}}$ ist nach Lemma 2.20 ein Schätzwert für den theoretischen Quartilsabstand $F_X^{-1}(\frac{3}{4}) - F_X^{-1}(\frac{1}{4})$, wobei F_X die Verteilungsfunktion der $\mathcal{N}(\mu, \sigma^2)-$verteilten Stichprobenvariablen $X_1, \ldots, X_n$ ist. Da weiter

$$\begin{aligned} F_X^{-1}\left(\frac{3}{4}\right) - F_X^{-1}\left(\frac{1}{4}\right) &= \left(\sigma\Phi^{-1}\left(\frac{3}{4}\right) + \mu\right) - \left(\sigma\Phi^{-1}\left(\frac{1}{4}\right) + \mu\right) \\ &= \sigma\left(\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)\right), \end{aligned}$$

gilt, dass die Steigung der approximierten Sollgeraden einen geeigneten Schätzwert für die Standardabweichung σ darstellt.

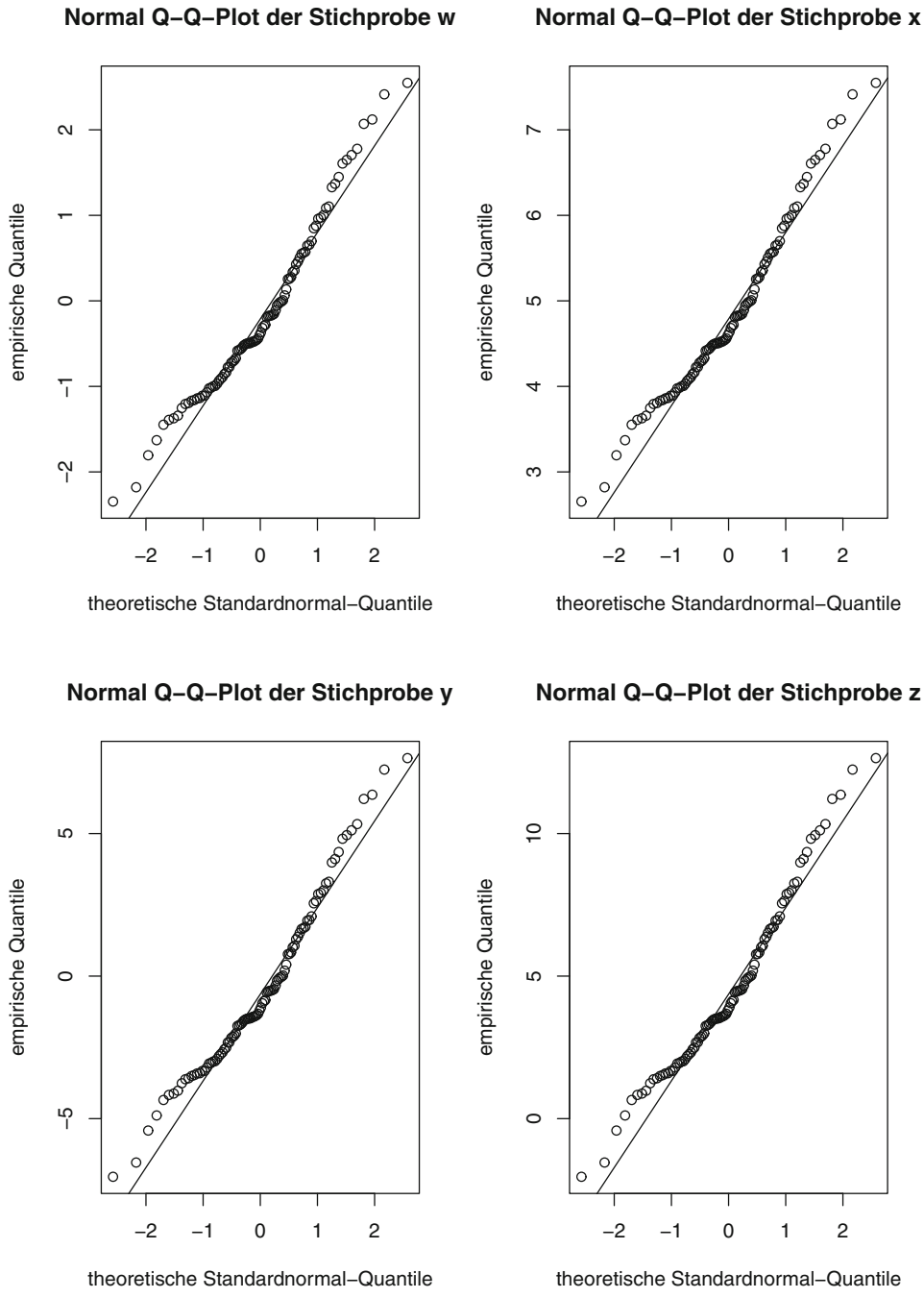

Abb. 2.14 Normal Q-Q-Plots der $\mathcal{N}(0,1)$-verteilten Stichproben $\boldsymbol{w}$, $\mathcal{N}(5,1)$-verteilten Stichprobe $\boldsymbol{x}$, $\mathcal{N}(0,9)$-verteilten Stichprobe $\boldsymbol{y}$ und der $\mathcal{N}(5,9)$-verteilten Stichprobe $\boldsymbol{z}$. Der Ordinatenabschnitt der eingezeichneten Sollgeraden ist ein Schätzwert für den Erwartungswert und die Geradensteigung ein Schätzwert für die Standardabweichung der Stichprobenverteilungen SRMfig2.14

In der Abb. 2.14 sind Normal Wahrscheinlichkeits-Plots (mit der Standardnormalverteilung als hypothetische Verteilung) für 4 simulierte Stichproben $\boldsymbol{w}$, $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ jeweils vom Umfang $n = 1000$ dargestellt. Als Verteilung bei der Erzeugung der Pseudo-Zufallszahlen wurde bei der Stichprobe $\boldsymbol{w}$ eine Standardnormalverteilung, bei $\boldsymbol{x}$ eine $\mathcal{N}(5,1)$-Verteilung, für $\boldsymbol{y}$ eine $\mathcal{N}(0,9)$-Verteilung und bei der Stichprobe $\boldsymbol{z}$ eine $\mathcal{N}(5,9)$-Verteilung verwendet.

In der folgenden Aufzählung sind die Anwendungsmöglichkeiten von Q-Q-Plots zur explorativen Analyse einer Stichprobe mit großem Stichprobenumfang zusammengefasst, vgl. Chambers [2], S. 90.

a) Verteilungsannahme: Stimmt die Verteilung der (linear transformierten) Stichprobenvariablen mit der hypothetischen Verteilung überein, zeigen die Punkte des Q-Q-Plots einen linearen Verlauf entlang einer Sollgeraden.
b) Lage- und Skalenunterschiede: Besitzen die Stichprobenvariablen nach einer linearen Transformation tatsächlich die hypothetische Verteilung, können mit dem Ordinatenabschnitt und der Steigung der Sollgeraden grafisch Lage- und Skalierungsparameter der Stichprobenverteilung geschätzt werden.
c) Ausreißer: Entsprechen die Punkte des Q-Q-Plots mehrheitlich einem approximativ linearen Verlauf, so können einzelne abweichende Punkte als potentielle Ausreißer identifiziert werden.
d) Unterschiede in Form und Schiefe: Ein systematisches Abweichen der Punkte im Q-Q-Plot von der Sollgeraden an den Rändern ist ein Hinweis auf Unterschiede an den Rändern der hypothetischen Verteilung und der tatsächlichen Stichprobenverteilung. Besitzt die Stichprobenverteilung im Vergleich zur hypothetischen Verteilung z. B. stärker (schwächer) besetzte Verteilungsränder, verlaufen die Punkte im Q-Q-Plot an den Rändern horizontal (vertikal) von der Sollgeraden weg, vgl. Abb. 2.15.

Mit Q-Q-Plots können Verteilungsannahmen insbesondere auch hinsichtlich ihrer Gültigkeit an den Rändern explorativ untersucht werden. Dies ist z. B. im Hinblick auf eine geeignete Modellierung von Schadenverteilungen mit möglichen Großschäden eine wichtige Anwendung. Bei der Interpretation von Q-Q-Plots an den Randbereichen sollte allerdings berücksichtigt werden, dass in Abhängigkeit von der vorliegenden Verteilung an den Rändern größere Abweichungen von der Sollgeraden, auch für den Fall, dass die Stichprobenvariablen (bzw. ihre lineare Transformation) wirklich die hypothetische Verteilung besitzen, vorliegen können. Dieses Verhalten kann z. B. durch wiederholte Simulationen von Normal Q-Q-Plots mit pseudo-normalverteilten Stichproben verdeutlicht werden, vgl. Thas [9], S. 56 ff.

Für die angemessene Interpretation eines Q-Q-Plots, vor allem auch bzgl. des Verhaltens an den Rändern, können Konfidenzintervalle, die die Sollgerade bzw. die Quantile der Stichprobenverteilung (das sind die Ordinatenkoordinaten der Punkte im Q-Q-Plot) zu Bereichsschätzern erweitern, sehr hilfreich sein.

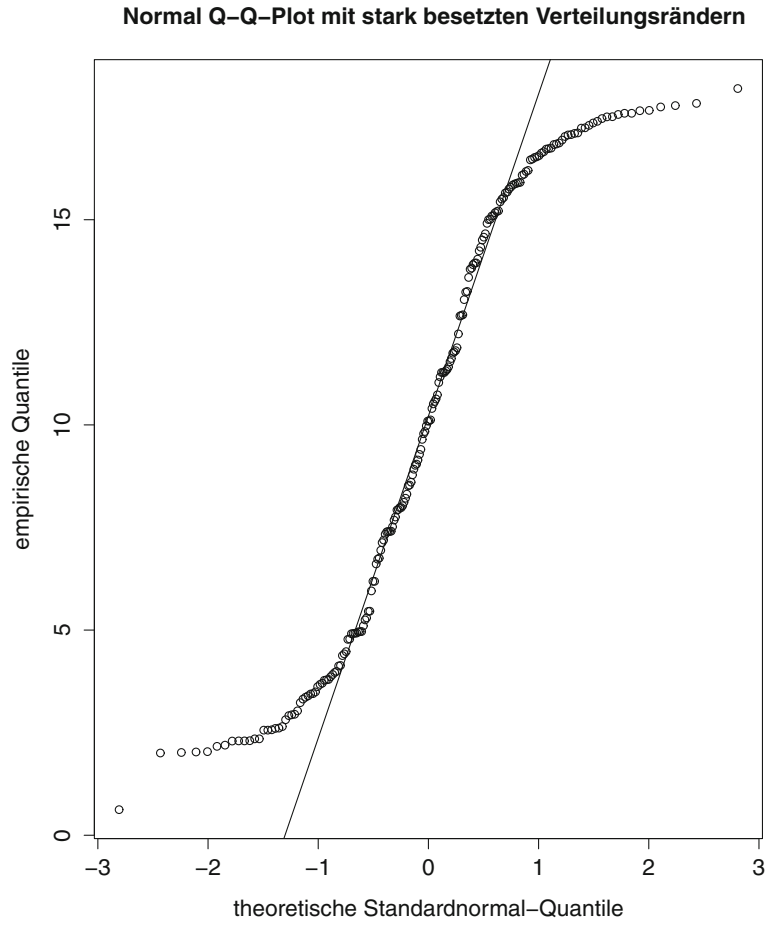

Abb. 2.15 Normal Q-Q-Plot einer simulierten Stichprobe $\boldsymbol{x}$ vom Umfang $n = 200$, die aus 100 $\mathcal{N}(10,2)$-verteilten Pseudo-Zufallszahlen und je 50 $\mathcal{U}[2,5]$- und $\mathcal{U}[15,18]$-verteilten Pseudo-Zufallszahlen erzeugt wurde. Man erkennt deutlich die horizontalen Abweichungen der Punkte an den Rändern, die mit den im Vergleich zur hypothetischen Standardnormalverteilung stärker besetzten Verteilungsrändern korrespondieren

In der Abb. 2.16 ist ein Normal Q-Q-Plot für eine Stichprobe vom Umfang $n = 100$ standardnormalverteilter Pseudo-Zufallszahlen dargestellt. Weiter beinhaltet die Abbildung einen Q-Q-Plot mit der Standardexponential-Verteilung als hypothetische Verteilung (kurz: **Exponential Q-Q-Plot**) für eine Stichprobe vom Umfang $n = 100$ standardexponential-verteilter Pseudo-Zufallszahlen. Beide Q-Q-Plots sind mit punktweisen Konfidenzintervallen zum Konfidenzniveau 99 % (jeweils verbunden zu einem Konfidenzband) ergänzt. Die variierenden Breiten der Konfidenzbänder über den Abszissenbereich zeigen die besonderen Randcharakteristika. Bei beiden Q-Q-Plots sind z. B.

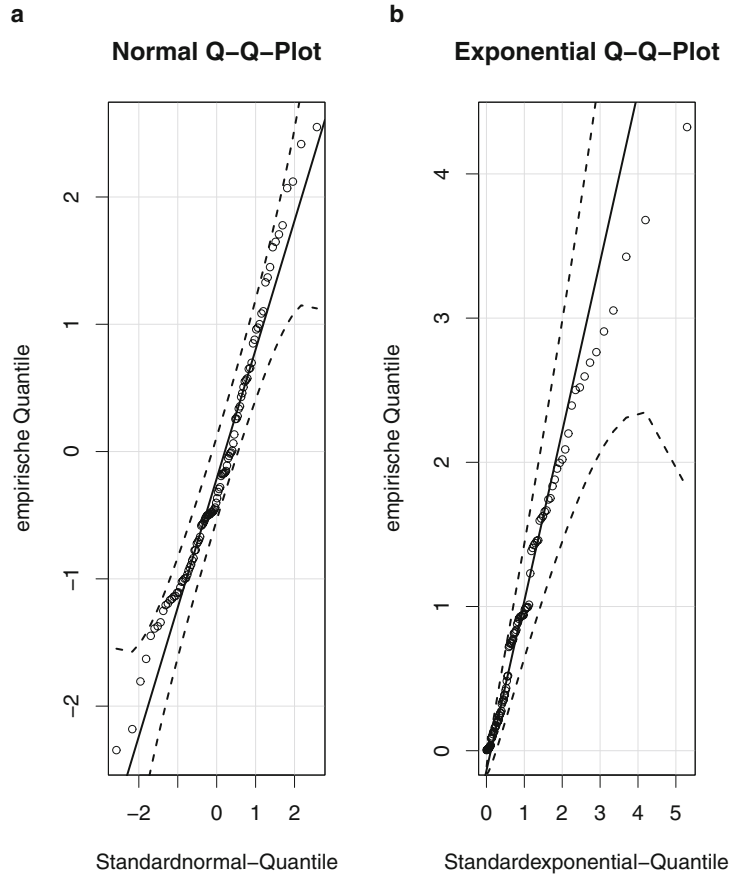

Abb. 2.16 **a** Normal Q-Q-Plot einer Stichprobe von $n = 100$ standardnormalverteilten Pseudo-Zufallszahlen. **b** Exponential Q-Q-Plot einer Stichprobe von $n = 100$ standardexponentialverteilten Pseudo-Zufallszahlen. Zusätzlich zu der Sollgeraden (*durchgezeichnete Linie*) sind die punktweisen Konfidenzintervalle verbunden als *gestrichelte Linie* markiert SRMfig2.16

am rechten Rand erst relativ große Abweichungen der Punkte von der Sollgeraden als signifikante Abweichungen von der Sollgeraden zu interpretieren.

2.4.5 Kerndichteschätzer

Wie bereits in Abschn. 2.2.1 erläutert, besitzen Histogramme als Schätzer für Wahrscheinlichkeitsdichten zwei große Nachteile. Zum einen ist das Histogramm abhängig von der gewählten Intervalleinteilung, zum anderen liefert ein Histogramm immer eine unstetige Funktion (Treppenfunktion) als Dichteschätzung. Viele für die Anwendung relevanten Wahrscheinlichkeitsdichten sind allerdings stetige Funktionen.

Ein **Kerndichteschätzer** ist eine Methode zur Schätzung einer Dichte f von i. i. d. Stichprobenvariablen $X_1, \ldots, X_n$, die als Schätzung eine stetige Funktion bereitstellt. Mit einem sogenannten **Kern** (oder auch als **Fenster** bezeichnet) $K : \mathbb{R} \to \mathbb{R}$, z. B. dem **Epanechnikov-Kern**

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{für } |x| \leq 1 \\ 0 & , \text{ sonst} \end{cases}$$

und einer zu wählenden **Bandbreite** (man sagt auch **Fensterbreite**) $h \in (0, \infty)$ ist der Kerndichteschätzer $\widehat{f}_{n,h}$ von f über die Abbildungsvorschrift

$$\widehat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

definiert. Konkrete Schätzwerte ergeben sich dann wieder, indem man die Stichprobenvariablen X_i durch die Stichprobenwerte x_i, $i = 1, \ldots, n$, ersetzt. Eine exakte Definition der geforderten Eigenschaften an eine Funktion $K : \mathbb{R} \to \mathbb{R}$, die einen Kern (bzw. ein Fenster) darstellt, gibt z. B. Pruscha [6], S. 302.

Kernschätzverfahren sind unabhängig von einer speziellen Intervalleinteilung, allerdings bestimmen der verwendete Kern K und die gewählte Bandbreite h die Form der Dichteschätzung. Vor allem die Wahl der Bandbreite beeinflusst stark den Grad der Glattheit der resultierenden Schätzfunktion. Kerndichteschätzer sind heute weit verbreitet in der statistischen Analysesoftware und werden oft kombiniert mit einem Histogramm verwendet.

Einen knappen Einblick zu Kerndichteschätzern findet man bei Fahrmeir et al. [3] S. 97–101. Bei Pruscha [6], S. 293–311, werden allgemein Dichteschätzer und im speziellen Kerndichteschätzer sehr ausführlich behandelt.

Die Abb. 2.17 zeigt für eine Stichprobe von 10.000 standardnormalverteilten Pseudo-Zufallszahlen das zugehörige Histogramm und die Approximation der Dichte durch einen Kerndichteschätzer.

2.5 Assoziationsmaße

In diesem Abschnitt werden bivariate Stichproben zweier Merkmale bzw. Stichprobenvariablen X und Y betrachtet. Die jetzt interessierende Fragestellung ist, ob die Stichprobe einen Zusammenhang (eine Assoziation) der Merkmale vermuten lässt. Für verschiedene Skalenniveaus der Merkmale werden unterschiedliche Zusammenhangsformen und Maßzahlen betrachtet. Im folgenden Abschnitt werden die Maßzahlen zur Beurteilung von Zusammenhangsstrukturen vorgestellt, die in der Anwendung sehr häufig zum Einsatz kommen.

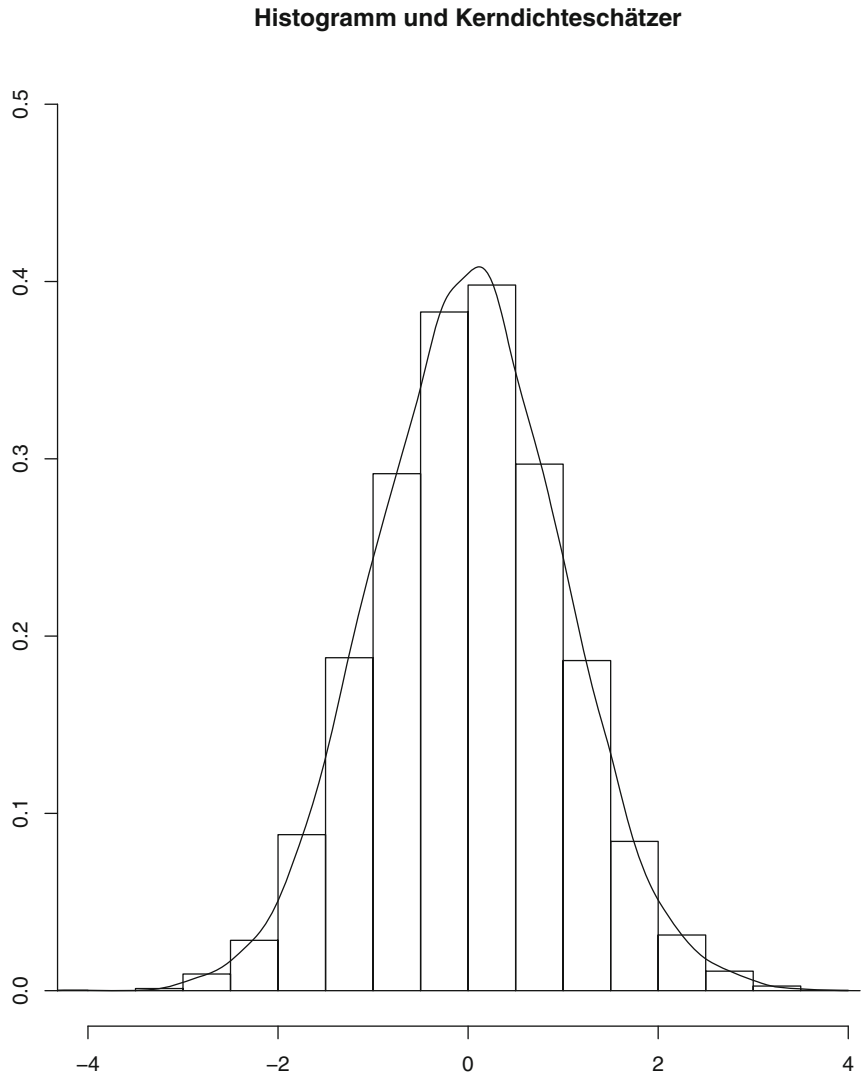

Abb. 2.17 Histogramm und Kerndichteschätzung einer Stichprobe von 10.000 standardnormalverteilten Pseudo-Zufallszahlen

2.5.1 Korrelationskoeffizienten

Zunächst wird der Zusammenhang zweier metrischer Merkmale betrachtet. In einem Streudiagramm kann eine bivariate Stichprobe metrischer Merkmale durch eine Punktwolke visualisiert werden. Die folgende Kennzahl ist ein Maß für die lineare Ausrichtung einer solchen Punktwolke.

Definition 2.38 (Empirischer Korrelationskoeffizient) *Sei* $(x_i, y_i)^\top$, $i = 1, \ldots, n$, *eine bivariate Stichprobe zweier kardinal skalierter Merkmale mit den arithmetischen Mitteln* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *und* $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ *der Teilstichproben* $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ *bzw.* $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. *Für die empirischen Varianzen* $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ *und* $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ *der Teilstichproben gelte* $s_x^2 s_y^2 \neq 0$. *Dann ist durch*

$$r_{x,y} := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.16}$$

der ***empirische Korrelationskoeffizient*** *(auch* ***Pearson-Korrelationskoeffizient*** *oder* ***gewöhnlicher Korrelationskoeffizient****) der bivariaten Stichprobe bzw. der beiden Teilstichproben definiert.*

Die Voraussetzung $s_x^2 s_y^2 \neq 0$ ist äquivalent dazu, dass weder die Stichprobenwerte $x_1, \ldots, x_n$ der Teilstichprobe $\boldsymbol{x}$, noch die Stichprobenwerte $y_1, \ldots, y_n$ der Teilstichprobe $\boldsymbol{y}$ alle identisch sind. Mit der **empirischen Kovarianz**

$$s_{x,y} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

und den empirischen Standardabweichungen der Teilstichproben $\boldsymbol{x}$ und $\boldsymbol{y}$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ und } s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

gilt die Darstellung

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}.$$

Der empirische Korrelationskoeffizient mit den i. i. d. Stichprobenvariablen X_i bzw. Y_i, $i = 1, \ldots, n$, anstelle der Stichprobenwerte x_i bzw. y_i, $i = 1, \ldots, n$, ist ein Schätzer für den theoretischen Korrelationskoeffizienten $\varrho(X_i, Y_i)$ der Stichprobenvariablen X_i und Y_i. Ebenso bildet die empirische Kovarianz eine Schätzfunktion für die Kovarianz $\mathrm{Cov}(X_i, Y_i)$. Entsprechend der Bedeutung des theoretischen Korrelationskoeffizienten können die Schätzwerte des empirischen Korrelationskoeffizienten interpretiert werden.

Rein deskriptiv kann die empirische Kovarianz als eine Maßzahl für die Ausrichtung der Punktwolke (x_i, y_i), $i = 1, \ldots, n$, im Streudiagramm um den gemeinsamen Schwerpunkt $(\overline{x}, \overline{y})$ verstanden werden. Wählt man für das Streudiagramm ein kartesisches Koordinatensystem mit dem Ursprung $(\overline{x}, \overline{y})$, so besitzt die Größe $(x_i - \overline{x})\,(y_i - \overline{y})$ (das ist das Produkt der vertikalen und horizontalen Abstände des Punktes (x_i, y_i) zum Schwerpunkt $(\overline{x}, \overline{y})$) in Abhängigkeit des Quadranten, in dem der Punkt (x_i, y_i) liegt, entweder ein positives oder ein negatives Vorzeichen. In der emprischen Kovarianz wird die Summe aller solcher Abweichungsprodukte gebildet. Liegt die Punktwolke z. B. gleichmäßig um den Schwerpunkt verteilt, so ergibt sich (aufgrund der gleichmäsig auftretenden positiven und negativen Summanden) eine empirische Kovarianz nahe 0. Der empirische Korrelationskoeffizient ist dann, ganz analog zum theoretischen Korrelationskoeffizienten, die um eine Normierung im Nenner ergänzte empirische Kovarianz. Eine ausführliche Darstellung der geometrischen Interpretation des empirischen Korrelationskoeffizienten findet man z. B. bei Fahrmeir et al. [3], S. 134–135.

Im folgenden Satz sind die grundlegenden Eigenschaften des empirischen Korrelationskoeffizienten zusammengefasst.

Satz 2.39 (Eigenschaften des empirischen Korrelationskoeffizienten) *Gegeben sei eine bivariaten Stichprobe* $(x_i, y_i)^\top$, $i = 1, \ldots, n$, *metrischer Merkmale. Für die empirischen Varianzen der Teilstichproben* $\boldsymbol{x}$ *und* $\boldsymbol{y}$ *gelte* $s_x^2 s_y^2 \neq 0$. *Dann gilt für den empirischen Korrelationskoeffizienten* $r_{x,y}$ *der bivariaten Stichprobe*

a) *Symmetrie:* $r_{x,y} = r_{y,x}$
b) *Maßstabsunabhängigkeit: Sei* $a, b, c, d \in \mathbb{R}$, $b, d \neq 0$, *dann gilt für die linear transformierte Stichprobe* (x_i^t, y_i^t) *mit* $x_i^t := a + bx_i$ *und* $y_i^t := c + dy_i$, $i = 1, \ldots, n$, *dass*

$$r_{x^t,y^t} = \frac{bd}{|b||d|} r_{x,y}.$$

Ist zusätzlich z. B. $b, d > 0$, *dann gilt* $r_{x^t,y^t} = r_{x,y}$.
c) *Wertebereich:* $-1 \leq r_{x,y} \leq 1$.
d) *Extremwerte:*

$$r_{x,y} = 1 \Leftrightarrow \exists a, b \in \mathbb{R},\, b > 0 : y_i = a + bx_i \quad \forall\, i = 1, \ldots, n.$$
$$r_{x,y} = -1 \Leftrightarrow \exists a, b \in \mathbb{R},\, b < 0 : y_i = a + bx_i \quad \forall\, i = 1, \ldots, n.$$

Beweis Die Aussagen a) und b) folgen sofort durch Nachrechnen aus der Definitionsgleichung (2.16).

Für die Beweise der Aussagen c) und d) verwendet man die **Ungleichung von Cauchy-Schwarz** für das Standardskalarprodukt $< \mathbf{a}, \mathbf{b} >$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$.

Bezeichne $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ und $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ die beiden Teilstichproben der bivariaten Stichprobe $(x_i, y_i)^\top$, $i = 1, \ldots, n$, mit den arithmetischen Mitteln $\overline{x}$, $\overline{y}$ und den empirischen Varianzen s_x^2, s_y^2.

Zu c): Mit den n-dimensionalen Vektoren

$$\boldsymbol{x}_0 := \boldsymbol{x} - \overline{x} \cdot \mathbf{1} = \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \text{ und } \boldsymbol{y}_0 := \boldsymbol{y} - \overline{y} \cdot \mathbf{1} = \begin{pmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix}$$

schreibt man

$$r_{x,y} = \frac{< \boldsymbol{x}_0, \boldsymbol{y}_0 >}{|\boldsymbol{x}_0||\boldsymbol{y}_0|}.$$

Man beachte, dass wegen $s_x^2 s_y^2 \neq 0$ schon $|\boldsymbol{x}_0||\boldsymbol{y}_0| \neq 0$ gilt. Damit erhält man mit der Ungleichung von Cauchy-Schwarz

$$\left|r_{x,y}\right| = \frac{|< \boldsymbol{x}_0, \boldsymbol{y}_0 >|}{|\boldsymbol{x}_0||\boldsymbol{y}_0|} \leq \frac{|\boldsymbol{x}_0||\boldsymbol{y}_0|}{|\boldsymbol{x}_0||\boldsymbol{y}_0|} = 1$$

und somit die Behauptung c)

$$-1 \leq r_{x,y} \leq 1.$$

Zu d): Wir zeigen zunächst, dass aus

$$y_i = a + bx_i \quad \forall\, i = 1, \ldots, n \text{ mit } a, b \in \mathbb{R},\ b \neq 0, \tag{2.17}$$

schon

$$r_{x,y} = \operatorname{sgn}(b) \tag{2.18}$$

folgt. Es gelte also (2.17), dann folgt $\overline{y} = a + b\overline{x}$. Damit erhält man, dass

$$\boldsymbol{y}_0 := \boldsymbol{y} - \overline{y} \cdot \mathbf{1} = \begin{pmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix} = \begin{pmatrix} a + bx_1 - \overline{y} \\ \vdots \\ a + bx_n - \overline{y} \end{pmatrix} = \begin{pmatrix} b(x_1 - \overline{x}) \\ \vdots \\ b(x_n - \overline{x}) \end{pmatrix} = b \cdot \boldsymbol{x}_0,$$

mit dem n-dimensionalen Vektor

$$\boldsymbol{x}_0 := \boldsymbol{x} - \overline{x} \cdot \mathbf{1} = \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix}.$$

Somit folgt

$$r_{x,y} = \frac{< \boldsymbol{x}_0, \boldsymbol{y}_0 >}{|\boldsymbol{x}_0||\boldsymbol{y}_0|} = \frac{b \cdot < \boldsymbol{x}_0, \boldsymbol{x}_0 >}{|\boldsymbol{x}_0| \cdot |b| \cdot |\boldsymbol{x}_0|} = \frac{b}{|b|} = \operatorname{sgn}(b),$$

d. h. (2.18).

Sei nun $|r_{x,y}| = 1$, dann gilt mit den n-dimensionalen Vektoren

$$\boldsymbol{x}_0 := \boldsymbol{x} - \overline{x} \cdot \mathbf{1} \text{ und } \boldsymbol{y}_0 := \boldsymbol{y} - \overline{y} \cdot \mathbf{1},$$

dass

$$|r_{x,y}| = \frac{|< \boldsymbol{x}_0, \boldsymbol{y}_0 >|}{|\boldsymbol{x}_0||\boldsymbol{y}_0|} = 1 \Leftrightarrow |< \boldsymbol{x}_0, \boldsymbol{y}_0 >| = |\boldsymbol{x}_0||\boldsymbol{y}_0|.$$

Nach der Ungleichung von Cauchy-Schwarz und unter Beachtung, dass $\boldsymbol{x}_0 \neq \mathbf{0}$ und $\boldsymbol{y}_0 \neq \mathbf{0}$ ist dies äquivalent dazu, dass ein $\lambda \in \mathbb{R} \setminus \{0\}$ existiert mit

$$\boldsymbol{y}_0 = \lambda \cdot \boldsymbol{x}_0.$$

Man erhält also, dass

$$\boldsymbol{y} = (\overline{y} - \lambda \overline{x}) \cdot \mathbf{1} + \lambda \cdot \boldsymbol{x},$$

wobei nach dem ersten Teil des Beweises zu d) weiter gilt

$$\operatorname{sgn}(\lambda) = \operatorname{sgn}\left(r_{x,y}\right).$$

□

Der empirische Korrelationskoeffizient ist eine Maßzahl für die Stärke und die Ausrichtung (positiv, d. h. gleichsinnig oder negativ, d. h. gegensinnig) des linearen Zusammenhangs der Teilstichproben $(x_1, \ldots, x_n)^\top$ und $(y_1, \ldots, y_n)^\top$. Je deutlicher die Punkte (x_i, y_i) im Streudiagramm auf einer Geraden mit positiver Steigung liegen, umso größer ist $r_{x,y}$. Umso mehr sich die Punkte einer Geraden mit negativer Steigung annähern, umso kleiner ist der Wert $r_{x,y}$. Die Stärke des linearen Zusammanhangs wird also durch $|r_{x,y}|$ beschrieben, während das Vorzeichen von $r_{x,y}$ die Ausrichtung des Zusammenhangs angibt.

Ein positiv (negativ) linearer Zusammenhang der Stichproben wird auch als **positive** (**negative**) **Korrelation** der Stichproben bezeichnet. Im Fall $r_{x,y} = 0$ liegt kein linearer Zusammenhang der Stichprobenwerte vor, man sagt auch die Teilstichproben sind **unkorreliert**.

In der Literatur finden sich verschiedene Vorschläge, ab welchem Wert von $|r_{x,y}|$ man von einer schwachen, mittleren oder starken Korrelation spricht. Fahrmeir et al. [3], S. 136,

schlagen z. B. vor, den Fall von $|r_{x,y}| < \frac{1}{2}$ als **schwache Korrelation** zu bezeichnen und ordnen dem Fall $\frac{1}{2} \leq |r_{x,y}| < \frac{4}{5}$ eine **mittlere Korrelation** zu. Für Stichproben mit $|r_{x,y}| \geq \frac{4}{5}$ spricht man dann von einer **starken Korrelation**.

In der Abb. 2.18 sind zu vier verschiedenen bivariaten Stichproben, jeweils mit Stichprobenumfang $n = 200$, die zugehörigen Streudiagramme und die empirischen Korrelationskoeffizienten (gerundet auf zwei Nachkommastellen) angegeben. Man beachte, dass die Stichprobe mit offensichtlich vorliegendem quadratischen Zusammenhang der Teilstichproben einen empirischen Korrelationskoeffizienten von nahe Null besitzt. Dies verdeutlicht, dass der empirische Korrelationskoeffizient nur den linearen Zusammenhang zweier Stichproben und nicht einen allgemeinen Zusammenhang misst.

In dem Fall, dass nicht beide der bivariaten Stichprobe $(x_i, y_i)^\top, i = 1, \ldots, n$, zugrundeliegenden Merkmale kardinal skaliert sind, kann der empirische Korrelationskoeffizient nicht sinnvoll verwendet werden. Sind allerdings beide Merkmale mindestens ordinal skaliert, kann man den empirischen Korrelationskoeffizient für die Stichprobe der zugeordneten Rangzahlen $(rg(x_i), rg(y_i))^\top$, $i = 1, \ldots, n$, berechnen. Die resultierende Maßzahl beschreibt dann die Stärke und Ausrichtung des monotonen Zusammenhangs der Teilstichproben.

Dabei ist für eine geordnete Stichprobe (eines mindestens ordinal skalierten Merkmals) $(x_{(1)}, \ldots, x_{(n)})^\top$ die **Rangzahl** (kurz: **Rang**) definiert als

$$rg(x_{(i)}) := i, \text{ falls kein } k \in \{1, \ldots, n\}, k \neq i, \text{ existiert mit } x_{(i)} = x_{(k)}.$$

Liegen in der Stichprobe **Bindungen** vor, d. h. es gibt $N > 1$ identische Stichprobenwerte

$$x_{(k)} = x_{(i)},\ k \in B \subset \{1, \ldots, n\}$$

wird $rg(x_{(i)})$ als **Durchschnittsrang**

$$rg(x_{(i)}) := \frac{1}{|B|} \sum_{k \in B} rg(x_{(k)}) = \frac{1}{N} \sum_{k \in B} k$$

definiert.

Beispiel 2.40 Für die bereits geordnete Stichprobe

$$\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)^\top = (-1, 1, 1, 1, 2, 5, 5)^\top$$

reeller Zahlen erhält man den Rang-Vektor

$$\begin{aligned}
&(rg(x_1), rg(x_2), rg(x_3), rg(x_4), rg(x_5), rg(x_6), rg(x_7))^\top \\
&= \left(rg(x_{(1)}), rg(x_{(2)}), rg(x_{(3)}), rg(x_{(4)}), rg(x_{(5)}), rg(x_{(6)}), rg(x_{(7)})\right)^\top \\
&= \left(1, 3, 3, 3, 5, \frac{13}{2}, \frac{13}{2}\right)^\top.
\end{aligned}$$

□

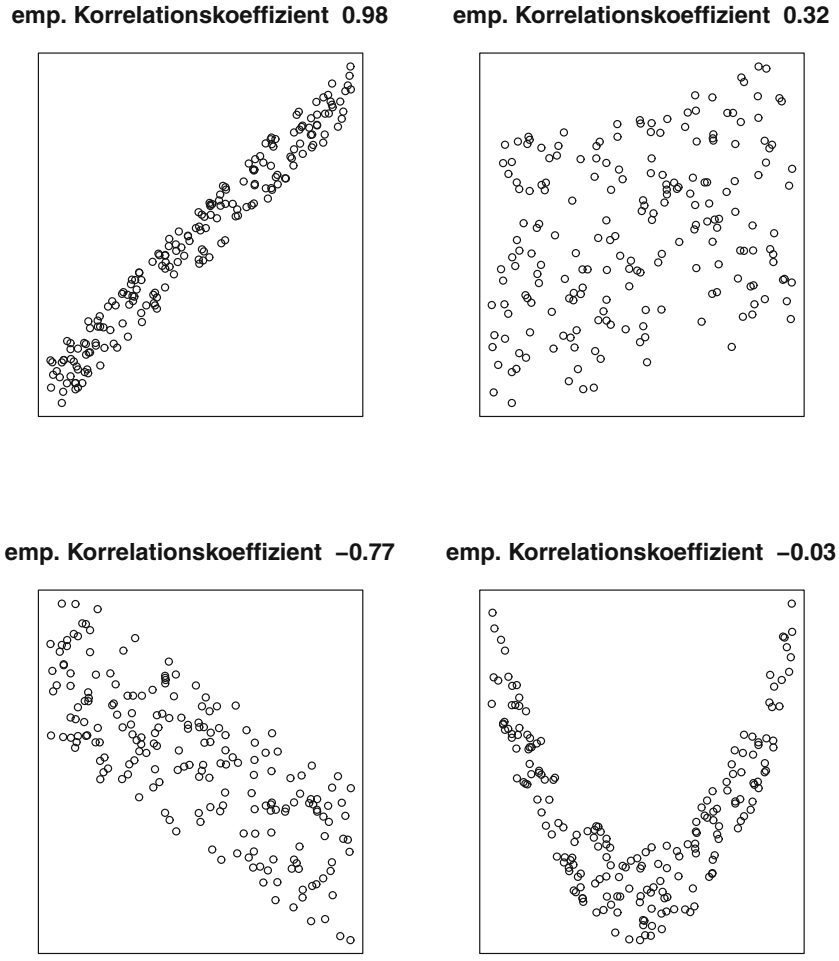

Abb. 2.18 Streudiagramme und empirische Korrelationskoeffizienten zu vier bivariaten Stichproben

Definition 2.41 (Rang-Korrelationskoeffizient nach Spearman) *Sei* $(x_i, y_i)^\top$, $i = 1, \ldots, n$, *eine bivariate Stichprobe zweier kardinal oder ordinal skalierter Merkmale. Weder die Teilstichprobenwerte* $x_1, \ldots, x_n$, *noch die Teilstichprobenwerte* $y_1, \ldots, y_n$ *seien alle identisch. Dann ist der (Spearman) Rang-Korrelationskoeffizient definiert als*

$$r_{x,y}^S := \frac{\sum_{i=1}^{n} \left(rg(x_i) - \overline{rg_x}\right)\left(rg(y_i) - \overline{rg_y}\right)}{\sqrt{\sum_{i=1}^{n} \left(rg(x_i) - \overline{rg_x}\right)^2} \sqrt{\sum_{i=1}^{n} \left(rg(y_i) - \overline{rg_y}\right)^2}},$$

mit $\overline{rg_x} := \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{n+1}{2}$ *und* $\overline{rg_y} := \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{n+1}{2}$.

Der Spearman Rang-Korrelationskoeffizient ist also identisch dem empirischen Korrelationskoeffizient der Rang-Stichprobe $(rg(x_i), rg(y_i))^\top$, $i = 1, \ldots, n$, wobei die Ränge $rg(x_i)$ und $rg(y_i)$, $i = 1, \ldots, n$, jeweils getrennt für jede der beiden Teilstichproben $\boldsymbol{x}$ und $\boldsymbol{y}$ gebildet werden.

Aufgrund der Definition 2.41 und mit Satz 2.39 ergeben sich folgende Eigenschaften für den Rang-Korrelationskoeffizienten.

Korollar 2.42 (Eigenschaften des Rang-Korrelationskoeffizienten) *Für den Rang-Korrelationskoeffizienten* $r^S_{x,y}$ *einer bivariaten Stichprobe* $(x_i, y_i)^\top$, $i = 1, \ldots, n$, *metrischer oder ordinaler Merkmale gilt*

a) Symmetrie: $r^S_{x,y} = r^S_{y,x}$
b) Maßstabsunabhängigkeit bei kardinal skalierten Merkmalen: Seien $a, b, c, d \in \mathbb{R}$, $b, d \neq 0$, *dann gilt für die linear transformierte Stichprobe* $(x_i^t, y_i^t)^\top$ *mit* $x_i^t := a + bx_i$ *und* $y_i^t := c + dy_i$, $i = 1, \ldots, n$

$$r^S_{x^t,y^t} = \frac{bd}{|b||d|} r^S_{x,y}$$

Allgemeiner gilt für jede streng monoton wachsende Transformation t_w *und jede streng monoton fallende Transformation* t_f *der Teilstichproben* $\boldsymbol{x}$ *und* $\boldsymbol{y}$, *dass*

$$\begin{aligned} r^S_{x,y} &= r^S_{t_f(x),t_f(y)} = r^S_{t_w(x),t_w(y)} \\ -r^S_{x,y} &= r^S_{t_f(x),t_w(y)} = r^S_{t_w(x),t_f(y)}, \end{aligned}$$

wobei z. B. $t_f(x)$ *die transformierte Stichprobe* $(t_f(x_1), \ldots, t_f(x_n))^\top$ *bezeichnet.*
c) Wertebereich: $-1 \leq r^S_{x,y} \leq 1$.
d) Extremwerte:

$$\begin{aligned} r^S_{x,y} = 1 &\Leftrightarrow rg(x_i) = rg(y_i) \quad \forall\, i = 1 \ldots, n. \\ r^S_{x,y} = -1 &\Leftrightarrow rg(x_i) + rg(y_i) = n + 1 \quad \forall\, i = 1 \ldots, n. \end{aligned}$$

Man beachte zu Korrolar 2.42 d), dass der Rang-Korrelationskoeffizient der Stichprobe $(x_i, y_i)^\top$, $i = 1, \ldots, n$, z. B. genau dann identisch 1 ist, falls der empirische Korrelationskoeffizient der Rang-Stichprobe $(rg(x_i), rg(y_i))^\top$ identisch 1 ist, d. h. falls die Punkte $(rg(x_i), rg(y_i))$, $i = 1, \ldots, n$, alle exakt auf einer Geraden mit positiver Steigung liegen. Der Rang-Korrelationskoeffizient besitzt demnach genau dann den Wert 1, falls zwischen den Teilstichproben ein eindeutig positiv monotoner Zusammenhang besteht, während der

Extremwert -1 genau dann angenommen wird, wenn die Teilstichproben sich in einem eindeutig negativ monotonen Zusammenhang befinden. Ganz analog zum empirischen Korrelationskoeffizienten kann die Größe $|r^S_{x,y}|$ als Maßzahl für die Stärke des monotonen Zusammenhangs verwendet werden.

Die Abb. 2.19 zeigt das Streudiagramm einer bivariaten Stichprobe mit eindeutig positiv monotonem Zusammenhang der Teilstichproben $\boldsymbol{x}$ und $\boldsymbol{y}$. Die Teilstichproben besitzen keinen strikt linearen Zusammenhang. Entsprechend erhält man für die resultierenden Korrelationskoeffizienten das Ergebnis

$$r_{x,y} < r^S_{x,y} = 1.$$

Bemerkung 2.43

a) Vor allem bei kleinen Stichprobenumfängen ist der empirische Korrelationskoeffizient sehr anfällig hinsichtlich Extremwerten in der Stichprobe. Der Rang-Korrelationskoeffizient stellt dagegen ein robustes Korrelationsmaß dar.

b) Ein weiterer bekannter Rangkorrelationskoeffizient für ordinal skalierte, bivariate Stichproben $(x_i, y_i)^\top$, $i = 1, \ldots, n$, ist der **Rangkorrelationskoeffizient nach Kendall** r_τ, vgl. etwa Sachs und Hedderich [8], S.67 - 68. Der Korrelationskoeffizient r_τ wird über so genannte **Inversionen** gebildet. Dazu werden die Stichprobenpaare $(x_i, y_i)^\top$ nach der Teilstichprobe $\boldsymbol{x}$ geordnet und die Rangpaare $(rg(x_i), rg(y_i))^\top$, $i = 1, \ldots, n$ betrachtet. Eine Inversion liegt vor, falls

$$rg(y_i) > rg(y_j) \text{ für } rg(x_i) < rg(x_j).$$

Der Rangkorrelationskoeffizient nach Kendall ist definiert als

$$r_\tau := 1 - \frac{4 \cdot A}{n(n-1)},$$

wobei A die Anzahl der vorliegenden Inversionen bezeichnet.

c) In der induktiven Statistik werden für die theoretischen Korrelationskoeffizienten sowohl Konfidenzintervalle, vgl. etwa Sachs und Hedderich [8], S. 297 ff., als auch Signifikanztests, vgl. z. B. Sachs und Hedderich [8], S. 544 ff. und S. 557 ff., verwendet. Die induktiven Verfahren basieren dabei jeweils auf den oben eingeführten, empirischen Korrelationskoeffizienten, welcher entsprechend als Schätzer interpretiert wird.

2.5.2 Empirischer χ^2-Koeffizient und Kontingenzkoeffizienten

Man betrachtet eine bivariate Stichprobe $(x_i, y_i)^\top$, $i = 1, \ldots, n$ vom Umfang n zweier diskreter Merkmale X, Y und die zugehörige 2-dimensionale $k \times m$ Kontingenztafel $\mathbf{K}$

der absoluten Häufigkeiten der Merkmalskombinationen. Wir setzen voraus, dass alle Randhäufigkeiten positiv sind. In der anschließenden Definition werden die Häufigkeitsbezeichnungen aus Abschn. 2.2.4 verwendet.

Definition 2.44 (Empirischer χ^2-Koeffizient)

$$\widehat{\chi}^2 := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{h_{i\cdot} h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot} h_{\cdot j}}{n}}. \tag{2.19}$$

Der empirische χ^2-Koeffizient ist die gewichtete Summe der Quadratabstände der tatsächlich vorliegenden Häufigkeiten h_{ij} zu den erwarteten Häufigkeiten bei Unabhängigkeit (vgl. Definition 2.25) über alle $k \cdot m$ Zellen der Kontingenztabelle. Die Gewichtung der Häufigkeitsabweichungen (Nenner in (2.19)) erfolgt je Zelle der Kontingenztabelle über die jeweilige erwartete Häufigkeiten bei Unabhängigkeit.

Aufgrund der Definition des empirischen χ^2-Koeffizienten folgt sofort, dass

$$0 \leq \widehat{\chi}^2 < \infty.$$

Kleine Werte von $\widehat{\chi}^2$ unterstützen die Hypothese, dass die zugrundeliegenden Merkmale keinen Zusammenhang aufweisen. Je größer der empirische χ^2-Koeffizient ausfällt, umso deutlicher liegt in der Stichprobe eine Abweichung von der empirischen Unabhängigkeit vor.

Die Werte des empirischen χ^2-Koeffizienten sind von der Dimension der Kontingenztafel (d. h. der Anzahl der unterschiedlichen Ausprägungen beider Merkmale) und vom Stichprobenumfang n abhängig. Daher ist ein reiner Zahlenwert des empirischen χ^2-Koeffizienten für die Bewertung der Stärke des Zusammenhangs zweier Merkmale nur schwer zu interpretieren. Ebenso sind Vergleiche der Zusammenhangstendenzen bei mehreren Kontingenztafeln mit unterschiedlichen Stichprobenumfängen oder Dimensionen der Tafeln alleine über die Grössenverhältnisse der χ^2-Koeffizienten nicht möglich.

Mithilfe des empirischen χ^2-Koeffizienten (dann auch **χ^2-Teststatistik** genannt) wird in der induktiven Statistik der asymptotische **χ^2-Unabhängigkeitstest** durchgeführt. Der χ^2-Unabhängigkeitstest wird z. B. bei Pruscha [7], S. 45–46, oder auch bei Fahrmeir et al. [3], S. 465–467, vorgestellt. Für einen geeignet großen Stichprobenumfang prüft der Signifikanztest die Unabhängigkeits-Nullhypothese

$$\begin{aligned} H_0 : P\left(X = a_i, Y = b_j\right) &= P\left(X = a_i\right) \cdot P\left(Y = b_i\right), \\ \forall\, i = 1, \ldots, k \text{ und } j &= 1, \ldots, m, \end{aligned}$$

wobei a_i, $i = 1, \ldots, k$, die Ausprägungen von X und b_j, $j = 1, \ldots, m$, die Ausprägungen von Y bezeichnen.

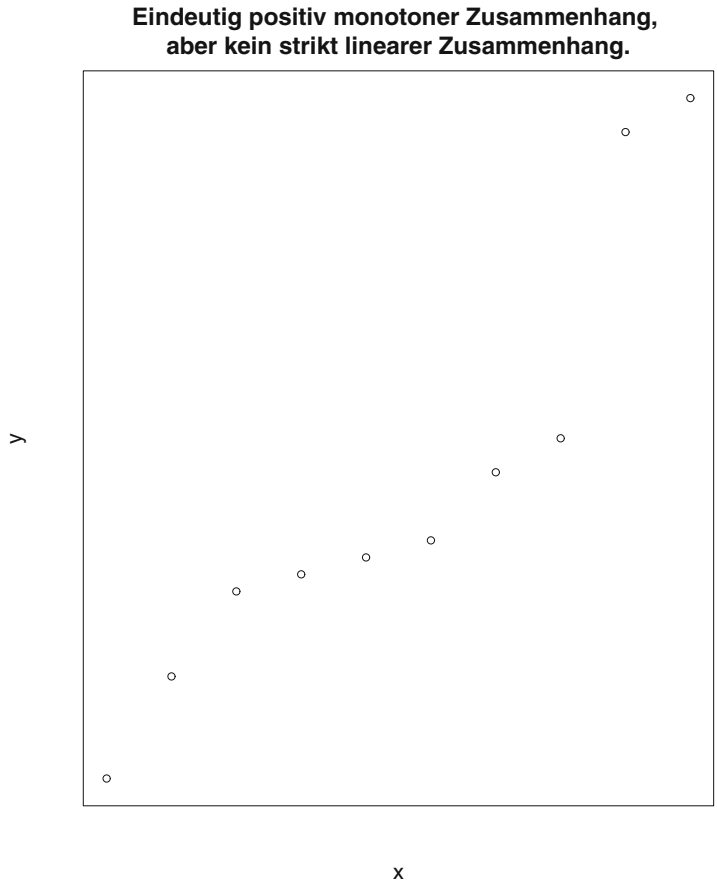

Abb. 2.19 Streudiagramm einer bivariaten Stichprobe (x_i, y_i), $i = 1, \ldots, 10$, reeller Zahlen mit Spearman Rangkorrelationskoeffizienten $r^S_{x,y} = 1$ und empirischen Korrelationskoeffizienten $r_{x,y} \approx \frac{9}{10}$

Für eine rein deskriptive bzw. explorative Bewertung des Grades der Abhängigkeit von X und Y (man sagt auch **Straffheit des Zusammenhangs**) verwendet man die folgenden Kontingenzkoeffizienten, die jeweils hinsichtlich der Interpretierbarkeit verbesserte Modifikationen des empirischen χ^2-Koeffizienten darstellen.

Definition 2.45 (Kontingenzkoeffizienten) *Für eine bivariate Stichprobe $(x_i, y_i)^\top$, $i = 1, \ldots, n$, zweier Merkmale X und Y vom Umfang n mit $k \times m$ Kontingenztafel $\boldsymbol{K}$ der absoluten Häufigkeiten und empirischen χ^2-Koeffizienten $\widehat{\chi}^2$ definiert man den **Kontingenzkoeffizienten nach Pearson***

$$K := \sqrt{\frac{\widehat{\chi}^2}{n + \widehat{\chi}^2}},$$

den ***korrigierten Kontingenzkoeffizienten nach Pearson***

$$K_{korr} := \sqrt{\frac{M}{M-1} \cdot \frac{\widehat{\chi}^2}{n + \widehat{\chi}^2}}$$

und den ***Kontingenzkoeffizienten nach Cramér***

$$V := \sqrt{\frac{\widehat{\chi}^2}{n \cdot (M-1)}},$$

wobei $M := \min\{k, m\}$ *das Minimum der Spalten- und Zeilenanzahl der zugrundeliegenden Kontingenztabelle* $\boldsymbol{K}$ *bezeichnet.*

Je größer die Kontingenzkoeffizienten sind, umso stärker ist der Zusammenhang der Merkmale in der Stichprobe ausgeprägt. Für den Wertebereich des Kontingenzkoeffizienten nach Pearson K gilt

$$0 \leq K \leq \sqrt{\frac{M-1}{M}},$$

daher folgt für den Wertebereich des korrigierten Kontingenzkoeffizienten nach Pearson K_{korr} aufgrund seiner Konstruktion

$$0 \leq K_{\text{korr}} \leq 1.$$

Der Kontingenzkoeffizient nach Cramér besitzt als Maximum den Wert 1.

Beispiel 2.46 Für die im Beispiel 2.24 betrachtete bivariate Stichprobe mit der gegebenen Kontingenztafel der absoluten Häufigkeiten rechnet man, dass

$$\widehat{\chi}^2 \approx 1433{,}5 \text{ und } K_{\text{korr}} = \sqrt{\frac{2}{2-1} \cdot \frac{\widehat{\chi}^2}{n + \widehat{\chi}^2}} \approx 0{,}17. \qquad \square$$

Für quadratische $k \times k$ Kontingenztafeln lässt sich die maximal straffe Zusammenhangsstruktur zweier Teilstichproben $\boldsymbol{x}$ und $\boldsymbol{y}$ bzw. zweier Merkmale X und Y sehr einfach charakterisieren. In diesem Fall besitzt die quadratische Kontingenztafel der absoluten Häufigkeiten in jeder Spalte und in jeder Zeile nur genau eine Zellen-Häufigkeit $h_{ij} \neq 0$. D. h. die Stichprobe besitzt die extreme Eigenschaft, dass durch die Ausprägung des einen Merkmals die Ausprägung des zweiten Merkmals schon eindeutig bestimmt ist. Der korrigierte Kontingenzkoeffizient nach Pearson K_{korr} ist genau dann identisch 1, falls

eine quadratische Kontingenztafel diese spezielle Form des maximal straffen Zusammenhangs besitzt.

Die Werte des korrigierten Kontingenzkoeffizienten nach Pearson K_{korr} sind unabhängig von der Zeilen- und Spaltenanzahl der zugrundeliegenden Kontingenztafel. Damit sind auch Kontingenztafeln mit unterschiedlichen Zeilen- bzw. Spaltenanzahlen über die entsprechenden, korrigierten Kontingenzkoeffizienten nach Pearson hinsichtlich der in den Stichproben vorliegenden Stärke des Zusammenhangs der Merkmale vergleichbar.

In der praktischen Anwendung kann man so etwa mehrere, nominale Merkmale mit unterschiedlich mächtigen Ausprägungsmengen hinsichtlich ihrer Zusammenhangsstärke bzgl. eines speziellen nominalen Ziel-Merkmals vergleichen. Wie der empirische χ^2-Koeffizient hängen die Kontingenzkoeffizienten allerdings weiterhin vom Stichprobenumfang ab. Daher ist bei einem Vergleich der Zusammenhangsstärke für unterschiedliche Kontingenztafeln auf Basis von Kontingenzkoeffizienten darauf zu achten, dass die den Kontingenztafeln zugrundeliegenden Stichproben ungefähr gleiche Umfänge besitzen.

Bemerkung 2.47 Sowohl Korrelationskoeffizienten, als auch Kontingenzkoeffizienten messen nur die Stärke einer Zusammenhangsstruktur in den Stichproben, sie geben aber keine Wirkungsrichtung in der Zusammenhangsstruktur (z. B. große Ausprägungen des einen Merkmals X führen zu großen Ausprägungen des anderen Merkmals Y) an. Weiter beweisen Assoziationsmaße alleine keine kausal-logischen Zusammenhänge zwischen Merkmalen, sondern interpretieren nur die datenstrukturellen Gegebenheiten. Für die praktische Anwendung ist in diesem Zusammenhang besonders auf die typischen Interpretationsfehler bei vorliegender **Scheinkorrelation** oder **verdeckter Korrelation** zu achten, vgl. ausführlicher bei Fahrmeir et al. [3], S. 145 ff..

Literatur

1. Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A.: Graphical Methods for Data Analysis. Wadsworth International Group, Belmont, California (1983)
2. Chambers, J. M.: Computional Methods for Data Analysis. Wiley, New York (1977)
3. Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G.: Statistik: der Weg zur Datenanalyse. Springer, Berlin (2003)
4. Friendly, M.: Mosaic displays for multi-way contingency tables. Journal of the American Statistical Association, **89**, 190–200 (1994)
5. Hartung, J., Elpelt, B., Klösener, K.-H.: Statistik: Lehr- und Handbuch der angewandten Statistik. Oldenbourg, München (2009)
6. Pruscha, H.: Vorlesungen über Mathematische Statistik. Teubner, Stuttgart (2000)
7. Pruscha, H.: Statistisches Methodenbuch: Verfahren, Fallstudien, Programmcodes. Springer, Berlin (2006)
8. Sachs, L., Hedderich, J.: Angewandte Statistik: Methodensammlung mit R. Springer, Berlin (2006)

9. Thas, O.: Comparing Distributions. Springer, New York (2010)
10. Tukey, J. W.: Exploratory Data Analysis. Addison-Weseley, Reading, Massachusetts (1977)
11. Witting, H. und Müller-Funk, U.: Mathematische Statistik II. Teubner, Stuttgart (1995)

http://www.springer.com/978-3-662-49406-6

Stochastische Risikomodellierung und statistische Methoden

Ein anwendungsorientiertes Lehrbuch für Aktuare

Becker, T.; Herrmann, R.; Sandor, V.; Schäfer, D.; Wellisch, U.

2016, XIV, 375 S. 65 Abb., Softcover
ISBN: 978-3-662-49406-6