Kapitel 1

Vorwort

In dem vorliegenden Werk zur Angewandten Statistik werden statistisch-mathematische Methoden aus der deskriptiven und induktiven, also schließenden Statistik vorgestellt. Das Buch wendet sich vorwiegend an Studierende technischer Fächer zur Unterstützung ihrer Grundausbildung in Statistik. Dabei wurde eine möglichst kurze sowie übersichtliche Darstellung angestrebt und besonderer Wert auf erläuternde Beispiele und zahlreiche Übungsaufgaben gelegt.

Häufig untersucht die Statistik Massenerscheinungen und Ergebnisse von Entscheidungen, die von sehr vielen Personen getroffen werden oder die sich im Laufe der Zeit wiederholen. Im Bereich der Wirtschafts- und Sozialwissenschaften sind dies beispielsweise Resultate von Konsum- und Investitionsentscheidungen, Entscheidungen über die Neueinstellung von Personen oder die Entlassung von Beschäftigten, häufig auch Ergebnisse zu Meinungen über politische Fragen. Mit Hilfe der Statistik wird versucht, diese Erscheinungen und ihre Gesetzmäßigkeiten zu erfassen. Das ist jedoch bei Einzelerscheinungen nicht möglich. Die Zahl möglicher Einflussfaktoren auf Entscheidungen oder Verhaltensweisen ist groß, ihre Wirkungsstärke oft ungewiss, so dass bei einzelnen Erscheinungen die Zufallseinflüsse überwiegen können. Die Untersuchung einer großen Zahl von Beobachtungen dient hingegen der Analyse von Gesetzmäßigkeiten.

Die statistischen Methoden sind, im Sinne einer zusammenfassenden Charakterisierung, Hilfsmittel zur Reduktion von Daten. Diese vollzieht sich oft über die Bildung von statistischen Verteilungen. So wird die Statistik auch als „Lehre von Verteilungen" bezeichnet. Beispielsweise kann man das Gewicht jedes Teilnehmers bei einer Reihenuntersuchung bestimmen, die Daten in Gewichtsklassen zusammenfassen und Maßzahlen wie Durchschnittswerte oder Varianzen berechnen.

Im Kapitel zur *deskriptiven Statistik* steht hier nicht die Datenerhebung, sondern die Auswertung der Daten im Vordergrund. „Deskriptiv" heißt *beschreibend*, im Gegensatz zu „explanatorisch" oder *erklärend*. Deskriptive statistische Methoden sollen empirische Vorgänge quantitativ beschreiben, nicht aber erklären. Statistisch erfasste Sachverhalte werden mit Hilfe statistischer Methoden analysiert. Die Ergebnisse gelten dann für die betrachtete Gesamtheit. Eine Verallgemeinerung dieser Ergebnisse ist nicht Ziel der deskriptiven Statistik. Zudem findet die Wahrscheinlichkeitstheorie im Rahmen der deskriptiven Statistik keine Anwendung.

Auch wenn Sachverhalte als Daten numerisch charakterisiert werden und eine möglichst genaue zahlenmäßige Beschreibung dieser Sachverhalte das Ziel einer ernsthaften Beschäftigung mit statistischen Fragen sein soll, so muss man sich jedoch davor hüten, der Beschreibung einen beliebig hohen Genauigkeitsgrad zuzuschreiben. In jeder statistischen Erhebung, beispielsweise auch in einer sorgfältig geplanten und vorbereiteten Volkszählung, gibt es in jedem Stadium der

Erhebung, sei es in Planung, Durchführung, Aufbereitung oder in der tabellarischen Darstellung, eine große Zahl von Fehlerquellen. Selbst wenn diese klein gehalten werden, so muss bei Teilerhebungen mittels Stichproben der so genannte Stichprobenfehler berücksichtigt werden.

Komplementär zur deskriptiven Statistik hat sich die *induktive* oder *inferentielle Statistik* entwickelt. Diese wird, nach dem Abschnitt zur Wahrscheinlichkeitstheorie, im dritten Kapitel des Buches vorgestellt. Der induktiven Statistik fällt die Aufgabe zu, eine Verbindung zwischen Theorie und Empirie herzustellen. Dazu greift sie auf Methoden der Wahrscheinlichkeitstheorie zurück. Vorgänge werden modellhaft, in Form von Verteilungen von Gesamtheiten, oft mit unbekanntem Parametern, wie beispielsweise dem Mittelwert, formuliert. Die vorliegenden Daten lassen sich dabei jeweils als Stichprobe aus einer Gesamtheit auffassen und wahrscheinlichkeitstheoretische Verfahren nutzen, um Rückschlüsse von den Stichproben auf die unbekannten Parameter der Gesamtheit zu ziehen. Die Analyse vollzieht sich dabei im Rahmen stochastischer Methoden, ohne sich bei der Modellbildung auf anderen Fachdisziplinen eigene Erklärungen zu beziehen.

Begleitend zu den einzelnen Theorieabschnitten finden sich, jeweils am Ende der drei beschriebenen Kapitel, Übungsaufgaben und Lösungen, die zu einem vertieften Verständnis der behandelten Inhalte beitragen sollen. Dabei wird, ähnlich wie beispielsweise in [4, 10, 12, 13], Wert auf eine vollständige und ausführliche Darstellung der Lösungen gelegt.

Kapitel 2

Deskriptive Statistik

Im Rahmen einer Erhebung wird in der Regel eine große Zahl von Einzeldaten, die bestimmte Objekte hinsichtlich ausgewählter Merkmale charakterisieren, gewonnen. Aufgabe der Statistik ist es, diese Daten so darzustellen und aufzubereiten, dass die in der Menge der Einzeldaten enthaltene statistische Information mit statistischen Methoden herausgefiltert und analysiert werden kann. Um diese Verfahren kennenzulernen, soll zunächst ein Blick auf die betreffenden Objekte und Merkmale geworfen werden.

Die Objekte, deren Merkmale in einer gegebenen Fragestellung von Bedeutung sind und im Rahmen einer empirischen Untersuchung beobachtet oder erfragt werden sollen, heißen *statistische Einheiten*. Statistische Einheiten können materielle oder immaterielle Objekte, Lebewesen oder Institutionen sein. Die statistische Einheit ist Träger der Information, die erhoben werden soll. Die Gesamtheit der für die Untersuchung relevanten Einheiten wird als *Grundgesamtheit der Untersuchungseinheiten*, einfach als *Grundgesamtheit*, oder auch als *statistische Masse* bezeichnet.

Die Bezeichnung *Untersuchungseinheit* beschreibt, dass die Abgrenzung dieser Einheiten dem Untersuchungszweck entsprechen soll. Jedoch ist das in der Praxis schwer zu verwirklichen. Soll beispielsweise die Anzahl der Beschäftigten bei Lastkraftwagen-Herstellern erhoben werden, so kann man die Gesamtheit der Betriebe, die Lkw produzieren, als Untersuchungsgesamtheit ansehen. Allerdings stellen Betriebe neben Lastkraftwagen oft auch andere Produkte her, beispielsweise Pkw oder Maschinenbauerzeugnisse. In der amtlichen Statistik wird nur derjenige Betrieb als Lkw-Produzent ausgewiesen, dessen Haupterwerbszweig die Lkw-Herstellung ist. Wird die Gesamtheit dieser Betriebe als Untersuchungsgesamtheit definiert und deren Beschäftigung erfasst, so ist das aus folgenden Gründen fehlerhaft:

(i) Der Betrieb mit Lkw-Herstellung als Hauptaktivität ist eine ungenaue Untersuchungseinheit. Soweit dort andere Produkte hergestellt werden, dient auch ein Teil der Beschäftigung der Herstellung dieser Nebenprodukte. Folglich müsste man die Aktivität des Betriebes auf Produktgruppen aufteilen. Innerhalb einer Kostenträgerrechnung wären „fachliche Betriebsteile" zu bilden, um dann auch die genaue Beschäftigung für die Lkw-Herstellung ausweisen zu können. Demnach wäre der *fachliche Betriebsteil* die geeignete statistische Einheit.

(ii) Die Betriebe mit Lkw-Herstellung als Hauptaktivität bilden für die vorliegende Fragestellung eine unvollständige Gesamtheit, gibt es doch auch Betriebe mit Lkw-Herstellung als Nebenaktivität. Weiterhin kann die so definierte statistische Masse unvollständig sein, weil in der amtlichen Statistik nur Betriebe mit mindestens 20 Beschäftigten berücksichtigt werden.

Von dem Begriff der *Untersuchungseinheit* sind die Begriffe der *Darstellungseinheit* und *Erhebungseinheit* zu unterscheiden. Als Darstellungseinheit wird diejenige Einheit, für die Ergebnisse in Veröffentlichungen vorliegen, bezeichnet. Unter Erhebungseinheit versteht man diejenige Einheit, für welche Informationen erhoben werden. Dies ist in Deutschland häufig das Unternehmen, das für alle seine Betriebe die geforderten Daten an das Statistische Bundesamt meldet.

Das Interesse der Statistik gilt jedoch nicht den statistischen Einheiten selbst, sondern ihren *Merkmalen*. In diesem Sinne sind die statistischen Einheiten *Merkmalsträger*. Unter einem Merkmal wird eine Eigenschaft einer statistischen Einheit verstanden. Das kann beispielsweise das Geschlecht, der Familienstand oder die Körpergröße jeweils als Eigenschaft von Personen sein. Ein Merkmal lässt sich zudem als Zuordnung, die Elementen der Untersuchungsgesamtheit betrachtete Eigenschaftssymbole zuweist, beschreiben. Jeder Wert eines Merkmals, das in der Untersuchung auftreten kann, wird *Merkmalsausprägung* genannt; der Wert des Merkmals, der tatsächlich aufgetreten ist, heißt *Beobachtungs-*, *Mess-* oder *Ursprungswert*. Wie üblich verwenden wir die Bezeichnungen Merkmal und beobachtete Merkmalsausprägung demnach einerseits als Werte, andererseits auch als Zuordnungen.

Ein Merkmal kann in Form einer Qualität, wie dem Familienstand oder dem Geschlecht, oder aber in Form einer Quantität, wie dem Alter oder dem Gewicht, auftreten. Entsprechend lässt sich von *qualitativen* bzw. *quantitativen Merkmalen* sprechen. Weiterhin unterscheidet man zwischen *diskreten* und *stetigen Merkmalen*. Diskrete Merkmale liegen vor, falls diese jeweils endlich viele oder abzählbar unendlich viele Ausprägungen aufweisen, stetige Merkmale hingegen, falls sie jeweils überabzählbar viele Ausprägungen besitzen. Weitere Beispiele für Merkmale aus der eigenen Lebenswelt und dem, hier nicht beschriebenen Bereich der Pflege und Gesundheit findet der interessierte Leser u. a. in [20, 21, 16]. Ein zusätzliches Unterscheidungskriterium von Merkmalen bildet der Bezug auf einen Zeitpunkt oder auf einen Zeitraum. Dann wird im ersten Fall von *Bestandsmerkmalen*, im zweiten von *Bewegungsmerkmalen* gesprochen. Auch eine Unterteilung in *häufbare* und *nicht häufbare* Merkmale ist üblich. Merkmale werden *extensiv* genannt, falls die Summe von Merkmalen eine sinnvolle Größe bildet. Eine sehr wichtige Einteilung von Merkmalen stellt die Unterscheidung der Beobachtungswerte nach ihrem Skalenniveau dar. Einen Überblick zu Skalen bietet der folgende Abschnitt.

2.1 Skalen

Ein Merkmal beschreibt eine Zuordnung von Untersuchungseinheiten zu Eigenschaftswerten dieser Einheiten. Handelt es sich bei den Werten dieser Zuordnung um Zahlen, so muss diesen Werten ein bestimmtes Messverfahren zugrunde liegen. Die Erörterung der messtheoretischen Fragen und Skalenprobleme beschäftigt sich speziell mit diesen Messverfahren und den Eigenschaften der Merkmalswerte. Unter einer (Mess-)Skala verstehen wir ein Messverfahren, mit dessen Hilfe Objekten Werte bezüglich der untersuchten Merkmale zugeordnet werden. Nach

den möglichen Relationen zwischen den Werten lassen sich, wie im Folgenden dargestellt, verschiedene Skalentypen unterscheiden:

(i) **Nominalskala**

Bei Nominalskalen wird lediglich die Gleichheit oder Ungleichheit von Beobachtungen innerhalb eines Merkmals festgestellt. Dabei wird das Merkmal durch verschiedene Begriffe oder Namen (lateinisch: *nomen*) untergliedert.

Beispiel: Das Merkmal „Familienstand" wird durch Zuweisung eines Begriffs, wie ledig, verheiratet, geschieden, verwitwet, charakterisiert.

Relationen zwischen Merkmalswerten: Gleichheit oder Ungleichheit

Zulässige Skalentransformationen: bijektive Funktionen

(ii) **Ordinalskala**

Liegt eine Ordinalskala vor, so kann, über Nominalskalen hinaus, auch die Ordnung (lateinisch: *ordo*) innerhalb eines Merkmals unterschieden werden.

Beispiel: Prüfungsleistungen lassen sich auf einer Ordinalskala anordnen.

Relationen zwischen Merkmalswerten: Gleichheit, größer (gleich), kleiner (gleich)

Zulässige Skalentransformationen: streng monoton steigende Funktionen

(iii) **Intervallskala**

Unter einer Intervallskala wird eine Skala verstanden, die bei gleichen Differenzen der Quantitäten gleiche Differenzen der Messwerte impliziert.

Beispiel: Die Celsius- und die Fahrenheit-Skalen zur Temperaturmessung

Relationen zwischen Merkmalswerten zusätzlich zu (ii): Quotienten von Differenzen von Merkmalswerten

Zulässige Skalentransformationen: affin-lineare Transformationen mit

$$x \mapsto y = a + bx\,,\ a, b \in \mathbb{R}\,,\ b > 0$$

(iv) **Verhältnisskala**

Eine Verhältnisskala verfügt, neben den Eigenschaften einer Intervallskala, über einen absoluten Nullpunkt, und weist demnach keine negativen Werte auf.

Beispiel: Die Kelvin-Skala zur Temperaturmessung

Relationen zwischen Merkmalswerten zusätzlich zu (iii): Quotienten von Merkmalswerten

Zulässige Skalentransformationen: lineare Transfomationen mit

$$x \mapsto y = bx\,,\ b \in \mathbb{R}_+^*$$

Zusammenfassend spricht man bei intervall- oder verhältnisskalierten Merkmalen von *metrisch skalierten Merkmalen*. Die Unterscheidung der Merkmale der Objekte in Hinblick auf Gleichheit oder Ungleichheit, wie dies mit Hilfe einer Nominalskala vorgenommen wird, bildet folglich das niedrigste Informations- und Messniveau, jedoch auch das einzige Skalenniveau, das bei qualitativen Merkmalen zur Verfügung steht. Für höherwertige Skalen werden quantitative Merkmale benötigt. Auf Grundlage dieser beschäftigen wir uns nun mit dem Begriff der Häufigkeit und der Definition von Häufigkeitsverteilungen.

2.2 Der Begriff der Häufigkeit

Im vorangegangenen Abschnitt wurde der Begriff des statistischen Merkmals definiert. Die Beobachtungswerte des Merkmals sind nun Ausgangspunkt der Definition der Häufigkeit und Häufigkeitsverteilung dieses Merkmals. Hierzu gehen wir vorerst nochmals auf den Merkmalsbegriff ein.

Seien $i = 1, ..., n$ die Untersuchungseinheiten und $x_1, ..., x_n$ die Beobachtungswerte eines Merkmals X. Die beobachteten Merkmalsausprägungen $a_1, ..., a_m$ beschreiben dann eine Zuordnung von Untersuchungseinheiten zu Beobachtungswerten. Dabei existiert zu jedem $i \in \{1, ..., n\}$ ein $l \in \{1, ..., m\}$ mit

$$x_i = a_l(i)\,.$$

Beispiele

- Bei einer Befragung der Studierenden wird untersucht, welche Verkehrsmittel vorwiegend für den Weg zur Hochschule genutzt werden. Dabei kennzeichne
  - a_1: die vorwiegende Nutzung öffentlicher Verkehrsmittel,
  - a_2: die vorwiegende Nutzung eines privaten KFZ,
  - a_3: weitere Möglichkeiten.

  Gibt beispielsweise der fünfte Studierende an, er nutze vorwiegend öffentliche Verkehrsmittel, so kann

  $$x_5 = a_1(5)$$

  geschrieben werden.

- Eine weitere Befragung hat zum Ziel, die Länge des täglichen Anfahrtsweges zur Hochschule zu bestimmen. Dabei ergibt sich, bei einer Stichprobe mit 10 befragten Personen, für den Anfahrtsweg in km

  $$(x_1, ..., x_{10}) = (2, 10, 5, 24, 3, 7, 15, 17, 6, 15)\,.$$

  Somit gilt beispielsweise

  $$x_5 = 3\,km\,.$$

Beschreiben wir den Anfahrtsweg durch das Merkmal seiner Länge mit den Ausprägungen

- a_1: Weglänge kleiner oder gleich $10\,km$,
- a_2: Weglänge größer $10\,km$,

so können wir

$$x_5 = a_1(5)$$

notieren.

Einer Merkmalsausprägung lässt sich ein Gewicht zuordnen. Das Gewicht der Merkmalsausprägung soll im Maße ihres Einflusses auf die betrachtete Untersuchung steigen. Abhängig von der Art der Untersuchung werden dabei verschiedene Größen verwendet. Doch bilden die absolute und die relative Häufigkeit die bekanntesten unter den Gewichtsmaßen. Als *absolute Häufigkeit* $H(a_l)$ der Ausprägung a_l wird die Anzahl der Fälle, in denen a_l auftritt, bezeichnet. Es gilt

$$\sum_{l=1}^{m} H(a_l) = n\,.$$

Der Wert

$$h(a_l) = \frac{H(a_l)}{n}$$

wird als *relative Häufigkeit* bezeichnet.

Auf Grundlage der inzwischen geklärten Begriffe, können wir uns nun der Definition der Häufigkeitsverteilung zuwenden.

2.3 Eindimensionale Häufigkeitsverteilungen

Die Gesamtheit aller Aussagen über ein Merkmal, zusammen mit den zugehörigen Gewichten, heißt die empirische *statistische Verteilung des betrachteten Merkmals*. Werden dabei Häufigkeiten als Gewichte verwendet, so spricht man von einer *Häufigkeitsverteilung* des Merkmals. Für die praktische Untersuchung ist es unerlässlich, Häufigkeitsverteilungen in einer Art nutzen zu können, die das Gewicht einer Aussage schnell erkennen lässt. Alle Aussagen mit ihren Gewichten aufzulisten, wäre jedoch sehr umständlich. Daher werden andere Darstellungsformen, nämlich die der Graphen von Häufigkeits-, Häufigkeitsdichte- und Häufigkeitssummenfunktionen, gewählt.

2.3.1 Darstellung von Häufigkeitsverteilungen

Wir stellen Häufigkeitsverteilungen durch

(i) Häufigkeitsfunktionen bzw. Häufigkeitsdichtefunktionen und

(ii) Häufigkeitssummenfunktionen

dar. Gehen wir von Punktklassen aus, so wird die Menge $\{H(a_l) \in \mathbb{N} \,|\, l = 1, .., m\}$ *absolute Häufigkeitsverteilung* und die Menge $\{h(a_l) \in \mathbb{R} \,|\, l = 1, .., m\}$ *relative Häufigkeitsverteilung* des beobachteten Merkmals genannt. Mit $M = \{a_1, ..., a_m\}$ können diese Häufigkeitsverteilungen als Häufigkeitsfunktionen

$$H : M \to \mathbb{N}\,,\ a_l \mapsto H(a_l)$$

bzw.

$$h : M \to \mathbb{R}\,,\ a_l \mapsto h(a_l) := \frac{H(a_l)}{n}$$

geschrieben werden. Bevor wir jedoch diese Funktionsgraphen darstellen, ist es sinnvoll, über die Aufbereitungsform der Messwerte und die Merkmalsausprägungen nachzudenken. So lassen sich die Daten folgendermaßen nach der *Aufbereitungsform der Messwerte* A und der *Anzahl der den Messwerten zugrundeliegenden Merkmalsausprägungen* N einteilen:

A	$N = 1$	$N > 1$
Ursprungswerte	Punktmesswerte	Mehrpunktmesswerte
Klassen von Messwerten	Punktklassen	Mehrpunktklassen

Ursprungswerte sind

- *Punktmesswerte*, wenn sie aus einer Zahl oder einem Symbol bestehen, wie beispielsweise bei der Anzahl der Kinder einer Familie,
- *Mehrpunktmesswerte*, wenn sie aus mehreren Zahlen oder Symbolen bestehen, wie beispielsweise bei geeigneten Einkommensintervallen.

Klassen von Messwerten heißen

- *Punktklassen*, wenn jede Klasse genau eine Ausprägung besitzt,
- *Mehrpunktklassen*, wenn sie mehr als eine Ausprägung besitzen.

Punktmesswerte können zu Punktklassen oder zu Mehrpunktklassen zusammengefasst werden. In Mehrpunktklassen zusammengefasste Daten werden auch als *klassiertes Material* bezeichnet. Bei klassiertem Material sind folgende Zuordnungen der Beobachtungswerte innerhalb der Klassen üblich:

(i) Abbildung der Werte auf den Klassenmittelwert (womit aus Mehrpunktklassen Punktklassen entstehen),

(ii) Gleichverteilung der Werte innerhalb der Klassen.

Häufigkeitsverteilungen in Form von Punktklassen

(i) *Häufigkeitsfunktionen*

Werden die Ursprungswerte $x_1, ..., x_n$ zu $a_1, ..., a_k$ Beobachtungswerten zusammengefasst, so entstehen k Punktklassen. Für diese Werte lassen sich die absoluten und die relativen Häufigkeiten folgendermaßen bestimmen:

l	a_l	$H(a_l)$	$h(a_l)$
1	a_1	n_1	$\frac{n_1}{n}$
2	a_2	n_2	$\frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	a_k	n_k	$\frac{n_k}{n}$
	sonstige Werte	0	0
	Summe	$\sum_{l=1}^{k} n_l = n$	$\sum_{l=1}^{k} \frac{n_l}{n} = 1$

Beispiele

– *Einkommensverteilung*
Hierbei betrachten wir das Einkommen von 10 Personen in Tsd. €:

i	1	2	3	4	5	6	7	8	9	10
x_i	1,0	1,5	1,5	2,5	2,5	2,5	3,0	5,0	5,0	6,5

Bei den weiteren Darstellungen, die sich auf diese Tabelle beziehen, wird auf die Nennung der Einheit „Tsd. €" verzichtet.

Folglich erhalten wir für die Gesamtheit unterschiedlicher Werte,

$$\{1{,}0;\ 1{,}5;\ 2{,}5;\ 3{,}0;\ 5{,}0;\ 6{,}5\},$$

die Häufigkeitstabelle

l	a_l	$H_n(a_l)$	$h_n(a_l)$
1	1,0	1	0,1
2	1,5	2	0,2
3	2,5	3	0,3
4	3,0	1	0,1
5	5,0	2	0,2
6	6,5	1	0,1

und beispielsweise diese graphische Darstellung:

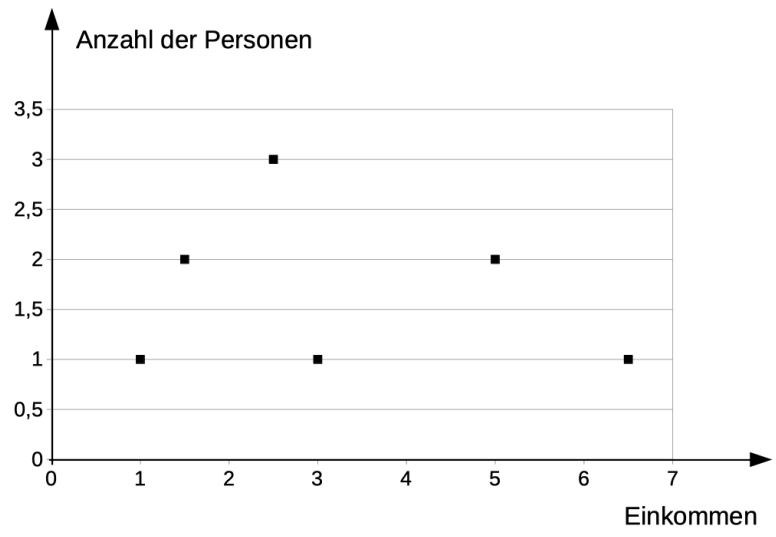

Abbildung 2.3.1 Einkommensverteilung

Mit Hilfe der Häufigkeitsfunktion $H : M \to \mathbb{N}$ lässt sich auch das Gewicht einer Menge bestimmen. So erhalten wir hier beispielsweise für $\{x \in \mathbb{R} \,|x \leq 2{,}5\}$ das Gewicht

$$H(x \leq 2,5) = H(\{1{,}0;\ 1{,}5;\ 2{,}5)\} = 6\,.$$

Die zugehörige relative Häufigkeit beträgt $0{,}6$.

– *Familienstand*

Wir betrachten den Familienstand von 10 Personen:

i	1	2	3	4	5	6	7	8	9	10
x_i	l	vh	l	g	vw	vw	l	l	l	vh

Die Ursprungswerte sind dabei Elemente der Menge

$$\{\text{ledig, verheiratet, verwitwet, geschieden}\} = \{l, vh, vw, g\}\,.$$