2 Empirische Häufigkeitsverteilungen

2.1 Häufigkeit und Verteilung

In diesem Kapitel werden Sie mit den Grundelementen einer statistischen Datenauswertung bekannt gemacht; dazu zählen die Häufigkeitsverteilung einer Variable, die Berechnung von Häufigkeiten aus dem Datenmaterial und die graphische Darstellung beider erstgenannten Punkte.

Eine sozialwissenschaftliche Untersuchung verfolgt in der Regel den Zweck, Phänomene, die in der Realität auftreten, zu erklären. Um ein Phänomen erklären zu können, ist es nötig, Aussagen über Art und Zusammenhänge[12] des untersuchten Phänomens zu geben. Da sich Aussagen empirischer Untersuchungen grundsätzlich auf Häufigkeiten beziehen, kommt der *empirischen Häufigkeitsverteilung* bei der Erklärung eines Phänomens eine elementare Funktion zu. Denn selbst Aussagen über Zusammenhänge mehrerer Variablen gründen sich auf Häufigkeitsverteilungen; in diesem Fall auf die Häufigkeit des Auftretens von Ausprägungen mehrerer Variablen.

> **Kap.2 Definition 1:**
> Als *empirische Häufigkeitsverteilung* bezeichnet man die Gesamtheit aller absoluten oder relativen Häufigkeiten der in der Messung aufgetretenen Messwerte. Die absolute Häufigkeit gibt für jeden Messwert an, wie oft dieser Messwert in der gesamten Stichprobe anzutreffen ist. Die relative Häufigkeit spiegelt ebenfalls die Anzahl der in der Stichprobe aufgetretenen Messwerte wider, jedoch relativ zum Stichprobenumfang.

Wird pro Untersuchungseinheit nur ein Merkmal gemessen (z. B. das Geschlecht), so spricht man von einer *eindimensionalen* oder *monovariaten Häufigkeitsverteilung*. Werden pro Einheit zwei Merkmale gemessen (z. B. Geschlecht und Körpergröße), nennt man die Häufigkeitsverteilung *zweidimensional* oder *bivariat*. Bei der Messung von mehreren Merkmalen pro Untersuchungseinheit erhält man eine *polyvariate* oder *k-dimensionale Häufigkeitsverteilung* (*k* ist dann die Anzahl der gemessenen Merkmale).

---

[12] gemeint sind beispielsweise die Wechselwirkungen des untersuchten Systems mit der Umwelt oder interne Wirkmechanismen

2.1.1 Das Aufstellen einer Häufigkeitstabelle

Nachdem eine empirische Untersuchung durchgeführt wurde, bedürfen die erhobenen Daten in den meisten Fällen zur weiteren Verarbeitung und Analyse einer Nachbereitung. Dazu werden die Einzelergebnisse, z. B. die Fragebögen, zunächst in einer *Urliste* notiert. Die Daten werden dabei ungeordnet in der Reihenfolge ihres Auftretens festgehalten. Zuweilen werden die Daten bei diesem Arbeitsschritt bereits nach einem bestimmten Kriterium geordnet, beispielsweise dem Namen der Versuchspersonen nach.
Verdeutlichen wir uns die Urliste an einem Beispiel:

Beispiel 1:

Ziel ist es, das mathematische Leistungsvermögen einer 8. Klasse einer Dresdner Mittelschule – bestehend aus 20 Schülern – zu ermitteln. Dazu wurde ein mehrteiliger Test (Algebra, Geometrie, ...) durchgeführt. Die 6-stufige Bewertungsskala des Endergebnisses reicht von sehr gut =1 über befriedigend =3 bis sehr schlecht =6.

Die Auflistung der Einzelergebnisse in einer Urliste kann in verschiedener Form vollzogen werden. Etwa als Zahlenreihe: 2, 4, 4, 4, 3, 1, 1, 5, 3, 2, 1, 4, 5, 6, 5, 4, 1, 3, 2, 5

oder in Form einer Tabelle:

Schülername	Schülernummer	erreichte Note
Sabine Aal	1	2
Klaus Cedur	2	4
Holger Krause	3	4
...	...	...
Christa Löwenzahn	20	5

In den meisten Fällen wird bei der tabellarischen Urliste jedem Merkmalsträger, in diesem Beispiel dem Schüler, eine Nummer zugeordnet, welche in den folgenden Arbeitsschritten als Index einer Variable verwendet wird und der Unterscheidung der einzelnen Merkmalsausprägungen einer Variable dient. Benennt man in unserem Beispiel die Note mit der Variablen X, so kann man demgemäss für „Der Schüler mit der Nummer 20 hat die Note 5 erzielt“ auch kurz schreiben $x_{20} = 5$ (siehe dazu auch Abschnitt 1.5.2).

Besonders bei großem Stichprobenumfang empfiehlt es sich, die Urliste in Form eines Rechteckes aufzustellen. Man kann dabei die Nummer der Merkmalsträger

in Zeile und Spalte kodieren, indem man in jeder Zeile genau 10 Messergebnisse nebeneinander schreibt. Jedes Messergebnis kann durch Auswahl der 10-er Potenz seiner Nummer in der Zeile und durch Auswahl seiner 1-er Potenz in der Spalte erreicht werden. Aus der Kombination der Zeile und Spalte, sprich aus der Einer- und Zehnerstelle, ergibt sich die konkrete Ausprägung des gesuchten Merkmalsträgers. Beispielsweise gibt Zeile 2, Spalte 4 (ohne Berücksichtigung der Kopfzeile und –spalte) die Note des Schülers mit der Nummer 14: $x_{14} = 6$.

Laufende Nummer	1	2	3	4	5	6	7	8	9	0
1–10	2	4	4	4	3	1	1	5	3	2
11–20	1	4	5	6	5	4	1	3	2	5

Die Entscheidung darüber, welche Art der Urliste man verwendet, wird von der Zweckmäßigkeit der Darstellung geleitet. Wurden beispielsweise in der Untersuchung mehrere Merkmale untersucht, so verwendet man sehr häufig die tabellarische Darstellungsform, wenn aber nur der Durchschnittswert des einen gemessenen Messwertes von Interesse ist, so ist die Aufreihung der Messergebnisse ausreichend.

Die Urliste ist die Grundlage jeder weiteren Datenverarbeitung. Zugleich schafft sie einen ersten Überblick über die Ergebnisse der Stichprobe. So lassen sich zum Beispiel der niedrigste (x_{min}) und höchste Messwert (x_{max}) eines Merkmerkmals ablesen.

> In unserem Beispiel mit den Schulnoten ist der niedrigste Messwert $x_{min} = 1$, der höchste $x_{max} = 6$. Sehr gut sichtbar wird an diesem Beispiel, dass die Bezeichnungen „höchster“ und „niedrigster Messwert“ keine qualitativen Wertungen darstellen, sondern sich einzig auf die Lage der Messwerte auf der Messwertskala beziehen.

Urlisten sind vor allem bei größeren Datenmengen äußerst unübersichtlich. Um eine Aussage über die gesamte Stichprobe treffen zu können, sprich in welcher Art die Merkmalsausprägungen verteilt sind, muss nun der zweite Arbeitsschritt folgen: Die Urliste wird in eine *Häufigkeitstabelle* übertragen, in der festgehalten wird, wie oft eine bestimmte Merkmalsausprägung in der gesamten Messung aufgetreten ist. Dazu werden die Variablenwerte mit Hilfe einer *Strichliste* so zusammengefasst, dass sie ihrer Größe nach geordnet sind und dass gleiche Messwerte zusammenstehen.

In der ersten Spalte der Häufigkeitstabelle werden alle Merkmalsausprägungen der Größe nach geordnet eingetragen. Bei diskreten Merkmalen empfiehlt es sich alle möglichen Messwerte in die Tabelle mit aufzunehmen, selbst wenn die Häufigkeit ihres Auftretens gleich Null ist, da dadurch „Lücken“ in der Verteilung sichtbar werden.
Die zweite Spalte wird für die Strichliste verwendet.

In der dritten Spalte wird in absoluten Zahlen eingetragen, wie oft das jeweilige Merkmal (in unserem Beispiel 1 die Note im Test) aufgetreten ist.

Um zu prüfen, ob bei der Auszählung Messwerte doppelt oder gar nicht gezählt wurden, bildet man einfach die Summe aller Häufigkeiten. Sie muss den Stichprobenumfang *N* ergeben.

Für unser Beispiel mit den Noten des Mathematiktests sieht die Häufigkeitstabelle mit integrierter Strichliste wie folgt aus:

Messwert (Note)	Strichliste	Anzahl des Auftretens (absolute Häufigkeit f_i)
1	/ / / /	4
2	/ / /	3
3	/ / /	3
4	/ / / / /	5
5	/ / / /	4
6	/	1
		$\Sigma = 20$

2.1.2 Absolute, relative und prozentuale Häufigkeiten

Bisher war die Rede von Häufigkeiten oder absoluten Häufigkeiten. In der Statistik unterscheidet man grundsätzlich drei Darstellungsformen der Häufigkeit: die absolute, die relative und die prozentuale Häufigkeit. Später werden noch Klassenhäufigkeiten vorgestellt, die ebenfalls in diesen drei Formen auftreten können; sie beziehen sich jedoch nicht auf einen einzelnen Messwert, sondern auf Intervalle oder Klassen von Messwerten. Ein weiterer häufig anzutreffender Häufigkeitsbegriff ist die Summenhäufigkeit. Er schließt für jede Merkmalsausprägung auch alle kleineren Merkmalsausprägungen ein, wird also für eine Merkmalsausprägung durch die Summe der Häufigkeiten aller Merkmalsausprägungen bestimmt, die höchstens genauso groß sind. Auch die Summenhäufigkeit tritt in den Formen absolute, relative und prozentuale Summenhäufigkeit auf.

Die *absolute Häufigkeit* (h_i) gibt also an, wie oft eine bestimmte Merkmalsausprägung in der gesamten Stichprobe aufgetreten ist. Mit anderen Worten die *n* Einzelergebnisse $x_1, x_2, ..., x_n$ werden zusammengefasst, indem man nur noch die sich unterscheiden den *k* Merkmalsausprägungen mit der Häufigkeit ihres Auftretens angibt.

Kap.2 Definition 2:
Die *absolute Häufigkeit* h_i des Auftretens des i-ten Messwertes gibt an, wie oft diese Merkmalsausprägung in der Stichprobe aufgetreten ist. Die Summe der absoluten Häufigkeiten aller k Merkmalsausprägungen ergibt die Gesamtzahl N aller Einzelergebnisse:

$$\sum_{i=1}^{k} h_i = N \tag{2.1}$$

Die *relative Häufigkeit* (f_i) der i-ten Merkmalsausprägung setzt die absolute Häufigkeit dieser Ausprägung (h_i) in Relation zum Gesamtumfang der Stichprobe (N), d. h. die relative Häufigkeit resultiert aus der Division von absoluter Häufigkeit und Gesamtanzahl der Messungen. Diese Bezugnahme auf den Stichprobenumfang ist bei der Interpretation der Häufigkeiten sehr wichtig, denn es kann beispielsweise trotz gleicher absoluter Häufigkeiten ein großer Unterschied in den relativen Häufigkeiten bei zwei Stichproben mit N = 10 Messungen und N = 10 000 Messungen bestehen. (Wenn ein Messwert bei 10 Messungen 3 mal auftritt, kann man das als häufig bezeichnen; wenn ein Messwert jedoch bei 10 000 Messungen 3 mal auftritt, wird man dies als extrem selten bezeichnen. Genau diesen Umstand versucht man mit der Angabe der relativen Häufigkeit hervorzuheben.)

Kap.2 Definition 3:
Die *relative Häufigkeit* f_i des i-ten Messwertes in einer Stichprobe vom Umfang N ist definiert als:

$$f_i = \frac{h_i}{N}\,. \tag{2.2}$$

Relative Häufigkeiten besitzen folgende Eigenschaften:
Relative Häufigkeiten sind nichtnegative Zahlen, die immer größer oder gleich null und kleiner oder gleich eins sein müssen: $0 \le f_i \le 1$. Die relative Häufigkeit ist dann null, wenn die zugehörige absolute Häufigkeit $h_i = 0$ ist. Den Wert eins nimmt sie nur an, wenn alle Messwerte der Stichprobe den selben Wert besitzen.
Die Summe aller relativen Häufigkeiten einer Stichprobe muss immer gleich eins sein:

$$\sum_{i=1}^{k} f_i = 1\,, \tag{2.3}$$

wobei k gleich der Anzahl der möglichen Messwerte (Merkmalsausprägungen) ist.

Die relative Häufigkeit hat noch eine weitere wichtige Eigenschaft:
Besitzt man zwei Messreihen derselben Messung vom gleichen Stichprobenumfang, kann man die beiden Messreihen durch die Häufigkeitsverteilungen vergleichen und u.U. feststellen, dass sie sich wesentlich unterscheiden. Doch wie soll man die beiden Messreihen unterscheiden, wenn sich die Anzahl der Messwerte

wesentlich unterscheidet? Die absoluten Häufigkeiten werden sich stark voneinander unterscheiden! Durch die Bildung der relativen Häufigkeiten lässt sich dieses Problem lösen. Man kann also mit Hilfe der relativen Häufigkeit Stichproben mit verschiedenen Umfängen miteinander vergleichen.

Die relativen Häufigkeiten lassen sich problemlos in *prozentuale Häufigkeiten* überführen, indem sie mit hundert multipliziert werden. Die prozentuale Häufigkeit bezeichnet den prozentualen Anteil des jeweiligen Messwertes an der Stichprobe; beispielsweise wie viel Prozent der Wähler einer Landtagswahl eine bestimmte Partei gewählt haben.

Kap.2 Definition 4:
Die *prozentuale Häufigkeit* $\%f_i$ des i-ten Messwertes in einer Stichprobe vom Umfang N beträgt:

$$\% f_i = f_i \cdot 100 = \frac{h_i}{N} \cdot 100 \tag{2.4}$$

Die prozentuale Häufigkeit kann der Art ihrer Berechnung nach nur Zahlenwerte im Intervall von null bis einhundert annehmen.

Beispiel 2:

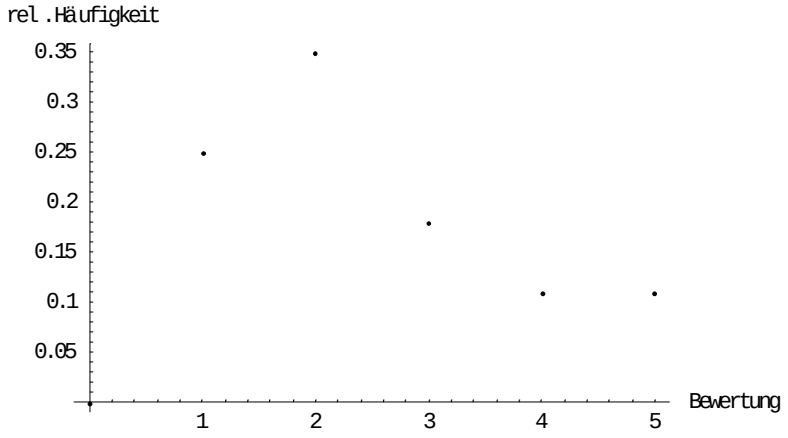
Die folgende Tabelle zeigt alle absoluten, relativen und prozentualen Häufigkeiten der Mathematiknoten aus Beispiel 1

Messwert (Note)	absolute Häufigkeit h_i	relative Häufigkeit $f_i = h_i / N$	prozentuale Häufigkeit $\%f_i = f_i * 100\%$
1	4	0,20	20 %
2	3	0,15	15 %
3	3	0,15	15 %
4	5	0,25	25 %
5	4	0,20	20 %
6	1	0,05	5 %
	N = Σ = 20	Σ = 1	Σ = 100 %

2.1.3 Die Häufigkeitsfunktion

Nach der Sichtung des Datenmaterials und der Aufstellung der Häufigkeitstabellen ist die grafische Darstellung der Häufigkeiten eine weitere Möglichkeit, einen Überblick über die ermittelten Daten zu erhalten. Das einfachste Mittel hierfür ist die Häufigkeitsfunktion. Sie beinhaltet die graphische Darstellung der relativen Häufigkeiten eines Merkmals in einem Funktionsgraphen. Auf diese Weise be-

schreibt sie, in welcher Art und Weise die gesamten Merkmalsausprägungen der Stichprobe über die Messwertskala verteilt sind.
Betrachtet man z. B. die Verteilungsfunktion einer Beliebtheitsumfrage mit fünf Antwortmöglichkeiten, so lässt sich auf den ersten Blick erkennen, ob die Masse der befragten Personen eher negativ oder positiv gegenüber der Fragestellung eingestellt sind. Ist das der Fall, so ergibt sich für diese Merkmalsausprägung(en) eine höhere Häufigkeit und der Graph der Häufigkeitsfunktion besitzt einen deutlich erkennbaren Gipfel. Je nach Lage drückt dieser dann eine positive, neutrale oder negative Einstellung der Mehrheit zum Erfragten aus.
Abbildung 2 zeigt eine Verteilungsfunktion, deren Hauptgewicht (der Gipfel der Funktion) leicht auf der Seite der negativen Beantwortung der Frage zu erkennen ist (linksgipflig). Dies bedeutet, dass die Mehrheit der Befragten den erfragten Sachverhalt eher als unbeliebt empfinden.



Kap. 2 Abbildung 1 - Häufigkeitsfunktion einer Beliebtheitsfrage

Häufigkeitsfunktionen können für alle Messniveaus gebildet werden, d.h. das zu beschreibende Merkmal kann nominal, ordinal oder metrisch gemessen sein[13]).

Kap.2 Definition 5:
Die *Häufigkeitsfunktion f(x)* zeigt, in welcher Art und Weise die relativen Häufigkeitswerte f_i über den Merkmalswerten x_i in der Stichprobe verteilt sind. Sie ist definiert als:

$$f(x) = \begin{cases} \textit{für} \quad x = x_i \\ \textit{für} \quad x \quad \textit{sonst} \end{cases} , \tag{2.5}$$

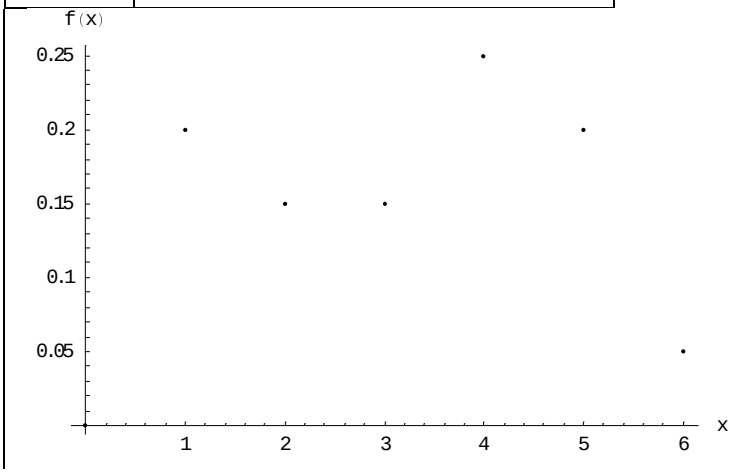
mit $-\infty < x < \infty$ und $0 \le f(x) \le 1$.

[13]) Schöffel, 1997: S. 4

Die Häufigkeitsfunktion *f(x)* wird in einem rechtwinkligen Koordinatensystem dargestellt; hierfür sind auf der Abszisse die Merkmalsausprägungen und auf der Ordinate die relativen (oder absoluten) Häufigkeiten abzutragen[14]).

Beispiel 3: Wertetafel von $f(x)$ für $x = x_i$:

$x = x_i$	1	2	3	4	5	6
$f(x)$	0,20	0,15	0,15	0,25	0,20	0,05



Kap.2 Abbildung 2 - Funktionsgraph der Häufigkeitsfunktion *f(x)* für die Schüler der 8. Klasse

2.1.4 Die Empirische Verteilungsfunktion

In vielen Untersuchungen spielt die Fragestellung eine große Rolle, wie viele Merkmalsträger Merkmalsausprägungen ober- oder unterhalb eines bestimmten Wertes aufweisen. So interessiert sich z.B. eine Hochschule mit einem Studiengang bei beschränkter Anzahl von Studienplätzen (numerus clausus) dafür, wie viele der Studienplatzanwärter einen Notendurchschnitt von z.B. 1.3 oder besser in ihrer Abiturausbildung erreicht haben. Um solchen Fragestellungen gerecht zu werden, ist die grafische Darstellung der Daten mit Hilfe eines Histogramms oder eines Balkendiagramms wenig nützlich.
Lässt sich eine sinnvolle Ordnung der Merkmalsausprägungen eines Merkmals angeben, sind die Daten also mindestens ordinalskaliert, kann die Häufigkeitsverteilung des Merkmals durch die *empirische Verteilungsfunktion* beschrieben werden.
Mit ihrer Hilfe lässt sich für jede beliebige Merkmalsausprägung x_i der Anteil aller Merkmalsausprägungen einer Stichprobe angeben, die diesen Messwert x_i nicht

[14]) Franz, 1991: S. 10

überschreiten. Die empirische Verteilungsfunktion kann folgendermaßen definiert werden:

Kap.2 Definition 6:
Eine Funktion, die für ein wenigstens ordinal skaliertes, zahlenmäßig erfasstes und geordnetes Merkmal *X* jeder reellen Zahl *x* den Anteil derjenigen Merkmalsträger einer statistischen Gesamtheit zuordnet, deren Merkmalswerte x_i diese Zahl nicht überschreiten, heißt empirische Verteilungsfunktion[15]).

Sind die Ausprägungen x_1, x_2,..., x_n des Merkmals *X* der Größe nach geordnet, so ist die empirische Verteilungsfunktion an der Stelle *x* gleich der *kumulierten Häufigkeit* (oder Summenhäufigkeit) aller Merkmalsausprägungen, die kleiner oder gleich *x* sind.

Kap.2 Definition 7:
Die *absolute Summenhäufigkeit* H_i gibt an, wieviele aller Messwerte kleiner oder gleich dem i-ten der möglichen Messwerte sind (i = 1, 2, 3,..., k). Mit Hilfe der absoluten Häufigkeiten lässt sich dies in folgender Form schreiben:

$$H_i = \sum_{j=1}^{i} h_j = h_1 + h_2 + ... + h_i \,. \tag{2.6}$$

Die *relative Summenhäufigkeit* F_i gibt an, welcher Anteil der Messwerte, bezogen auf den Stichprobenumfang *N*, kleiner oder gleich dem i-ten der möglichen Messwerte ist (i = 1, 2,..., k). Mit Hilfe der relativen Häufigkeiten lässt sich dies in folgender Form schreiben:

$$F_i = \sum_{j=1}^{i} f_j = \sum_{j=1}^{i} f\left(x_j\right) = f\left(x_1\right) + f\left(x_2\right) + ... + f\left(x_i\right). \tag{2.7}$$

Wenden wir die eben eingeführten Begriffe auf Beispiel 1, einen Mathematiktest, an, so erhalten wir folgende Häufigkeitstabelle:

Kap.2 Tabelle 1.

Note	h_i	f_i	H_i	F_i
1	4	0,20	4	0,20
2	3	0,15	7	0,35
3	3	0,15	10	0,50
4	5	0,25	15	0,75
5	4	0,20	19	0,95
6	1	0,05	20	1

[15]) Eckstein, 1998: S. 18

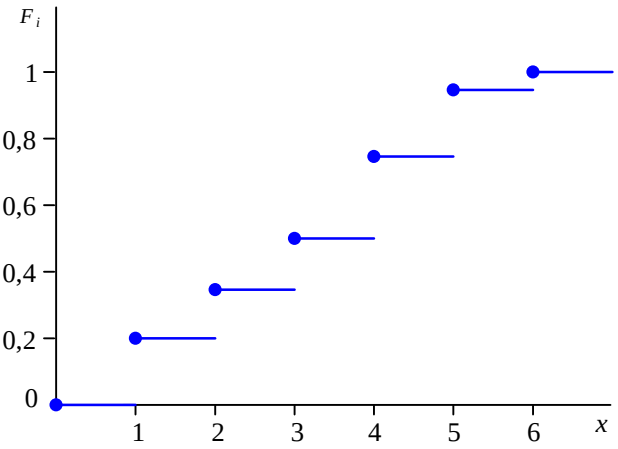
Zu beachten ist, dass der letzte Wert von H_i immer gleich dem Stichprobenumfang (in diesem Fall 20) und der letzte Wert von F_i immer gleich eins sein muss.

Kap.2 Definition 8:
Die empirische Verteilungsfunktion $F(x)$ ist monoton wachsend. Sie ist nur abschnittsweise definiert und ihre graphische Darstellung springt am rechten Rand jedes Abschnitts um die Häufigkeit des nächsten Messwertes nach oben. Im Abschnitt zwischen zwei aufeinanderfolgenden Messwerten ist $F(x)$ konstant.
Es gilt stets $0 \le F(x) \le 1$. Die empirische Verteilungsfunktion ist definiert als:

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_i = F(x_i) \quad \text{für} & x_i \le x < x_{i+1},\ i = 1,2,...,k-1 \\ 1 & x \ge x_k \end{cases} \qquad (2.8)$$

.

Die grafische Darstellung der empirischen Verteilungsfunktion (auch *Summenhäufigkeitskurve* genannt) nimmt die Form einer Treppe an und wird deshalb auch als *Treppenfunktion* bezeichnet.
Die Verteilungsfunktion des Beispiels 1 mit den Noten aus dem Mathematiktest besitzt folgendes Aussehen:



Kap.2 Abbildung 3 - Treppenfunktion zur Häufigkeitsverteilung der Schulnoten aus Beispiel 1

Über ein Merkmal lassen sich mit Hilfe der empirischen Verteilungsfunktion vier Aussagen formulieren:

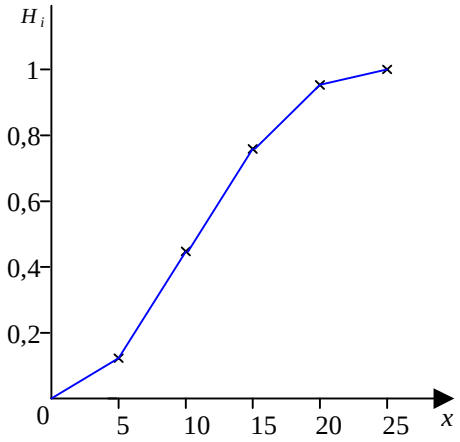
1. Wie viel Prozent der Messwerte einer Stichprobe kleiner oder gleich einem bestimmten Messwert sind,
2. Wie viel Prozent der Messwerte größer als ein bestimmter Messwerte sind,

3. Wie viel Prozent der Messwerte zwischen zwei Messwerten liegen und
4. Welcher Messwert die Stichprobe in zwei gleich große Hälften teilt.

> Für das Beispiel 1 lassen sich beispielsweise folgende Aussagen formulieren:
>
> 1. 75 % der Schüler haben eine Note erreicht, die kleiner oder gleich 4 ist; sprich die Noten 1, 2, 3 und 4.
> 2. 25 % der Schüler haben eine Note erhalten, die größer als 4 ist; sprich die Noten 5 und 6.
> 3. 60 % der Schüler haben eine Note zwischen 2 und 5 erreicht.
> 4. 50 % aller Noten sind kleiner bzw. gleich der Note 3.

Bei in Klassen eingeteilten Merkmalen wird innerhalb der Klasse eine Gleichverteilung der Merkmalsausprägungen unterstellt, da die Originalwerte nicht mehr bekannt sind. Es wird angenommen, dass sich die Messwerte gleichmäßig über das Klassenintervall verteilen. Die empirische Verteilungsfunktion kann deshalb im Bereich der Klasse als Diagonale dargestellt werden, wodurch man einen Polygonzug erhält.

Summenhäufigkeitspolygon der Daten aus Beispiel 3 (Bushaltestellen):



Kap.2 Abbildung 4 - Summenhäufigkeitspolygon der klassierten Daten zu den zurückgelegten Haltestellen aus Beispiel 3

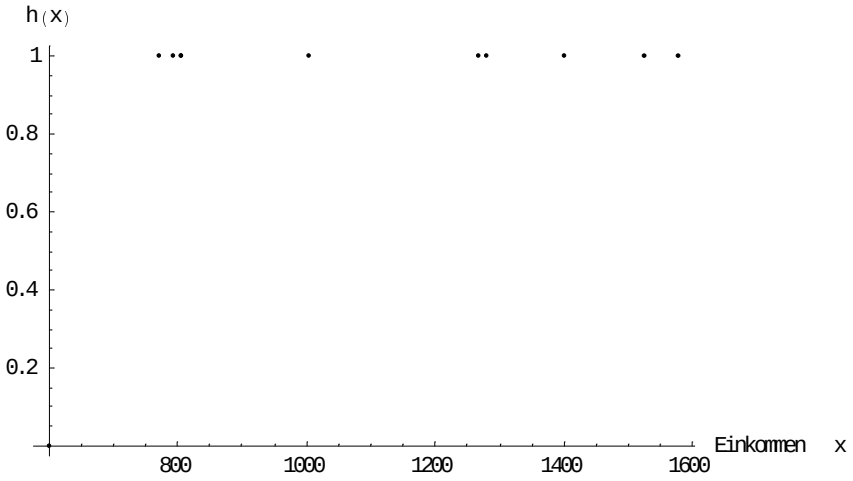
2.2 In Klassen eingeteilte Merkmale

Die Klassenbildung wird in diesem Kapitel eingeführt. Es wird um das Einteilen der Messwerte in Klassen gehen. Auch werden das Aufstellen der Klassenhäufigkeiten und offene Klassen behandelt. Der Leser sollte nach diesem Kapitel exakte Klassengrenzen erkennen können. Unter anderem anhand eines Beispiels wird der Informationsverlust durch Klasseneinteilung dem Leser verdeutlicht werden.

Bei Merkmalen, die auf dem Intervall- oder Verhältnisniveau gemessen werden, kommt es nicht selten vor, dass eine sehr große Vielzahl von unterschiedlichsten Merkmalsausprägungen gemessen werden. Im Extremfall ist die Anzahl k der gemessenen Merkmalsausprägungen gleich der Anzahl der Messungen n, so dass die absoluten Häufigkeiten h_i aller aufgetreten Merkmalsausprägungen gleich 1 sind.

Beispiel 4: Das Einkommen von 10 zufällig ausgewählten Studenten einer Seminargruppe:

Einkommen	770	794	804	806	1003	1266	1278	1399	1523	1576
Abs. Häufigkeit	1	1	1	1	1	1	1	1	1	1



Kap.2 Abbildung 5 - Häufigkeiten selten auftretender Merkmalsausprägungen aus Beispiel 4

Der Graph der Verteilungsfunktion zeigt, welche Probleme bei der Interpretation der Häufigkeiten jetzt auftreten. Die Häufigkeiten weisen kein lokales Maximum mehr auf, um die sich die Werte der Verteilungsfunktion häufen. Dies führt insbesondere dann zu Schwierigkeiten, wenn man Aussagen über die Verteilung der Daten von der Stichprobe auf die Grundgesamtheit, aus der die Stichprobe gezogen wurde, verallgemeinern möchte. Schließlich muss doch damit gerechnet werden, dass auch die Datenwerte, die bisher nicht aufgetreten sind, in der Grundgesamtheit anzutreffen sind.
Trotzdem lassen sich interessante Eigenschaften aus der Verteilungsfunktion erkennen. Man kann eine Teilmenge von relativ dicht zusammen liegenden Merkmalsausprägungen interpretieren als solche Teilmenge von Werten, die sich um einen bestimmten Wert häufen – nur mit dem Unterschied zu den Verteilungen bisher, dass sich dieser Umstand nicht in der relativen Häufigkeit selbst ausdrückt. Mit dieser Idee lassen sich Aussagen über die Verteilung der Stichprobe treffen, die auch auf die Grundgesamtheit verallgemeinern lassen, weil sie sich nicht mehr auf die relativen Häufigkeiten der Einzelwerte beziehen, sondern auf Gruppen von Einzelwerten. Damit werden die Aussagen allerdings auch weniger genau, da sie sich nicht mehr auf das Auftreten von Einzelwerten beziehen, sondern nur noch auf die Zugehörigkeit zu bestimmten Gruppen.
Dieser Ansatz soll nun der formalen Berechnung zugänglich gemacht werden.

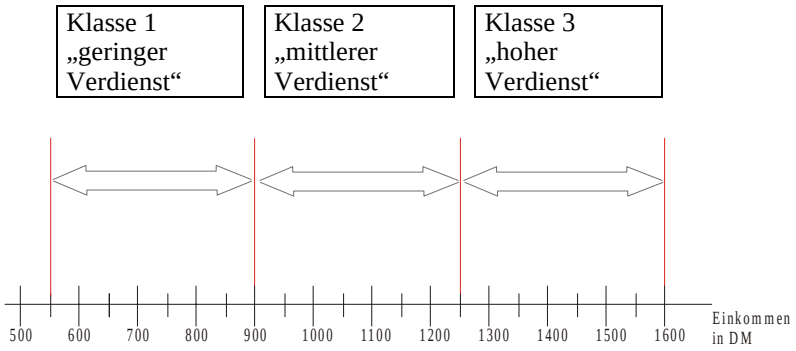
2.2.1 Das Einteilen der Messwerte in Klassen

Ein einfaches Mittel, um für solche wenig besetzten[16] Verteilungen eine aussagekräftige Verteilungsfunktion zu erhalten, ist das Einteilen der Merkmalsausprägungen in Klassen.
Unter einer Klassifizierung versteht man die vollständige Einteilung der Messwertskala in Intervalle, die sich nicht überschneiden. So wird jeder Messwert auf der Messwertskala genau einem Intervall zugeordnet, unabhängig davon, ob der Messwert in der Stichprobe gemessen wurde oder nicht. Die so entstandenen Intervalle der Messwertskala nennt man auch Klassen, die meist nach ihrer Lage auf der Messwertskala durchnummeriert werden.

Kap.2 Definition 9:
Eine vollständige Zerlegung einer Messwertskala in endlich viele paarweise verschiedener Klassen heißt Klassierung. Dabei heißt das k-te Teilintervall k-te Klasse. Jeder Messwert der Skala kann dabei genau einer Klasse zugeordnet werden. Zweiseitige Klassen besitzen eine untere x_{ku} und eine obere Klassengrenze x_{ko}. Die Klassenbreite b_k wird durch die Differenz der Klassengrenzen bestimmt.

[16]damit ist gemeint, dass die Mehrzahl der Messwerte nur selten vorkommen

Die Klassifizierung der Skala in unserem Beispiel könnte z.B. so erfolgen:



Kap.2 Abbildung 6 - Die Klassifizierung der Skala in einem Beispiel

Die Messwertskala wurde in drei Klassen bzw. Intervalle eingeteilt: 550 bis 900, 900 bis 1250 und Werte von 1250 bis 1600 DM. Diese Werte, welche die Klassen voneinander abgrenzen heißen Klassengrenzen. Für jede der hier gebildeten Klassen kann eine obere und eine untere Klassengrenze angegeben werden. Solche Klassen heißen auch zweiseitig begrenzte Klassen – oder kurz – zweiseitige Klassen. Weiterhin wurde jeder Klasse noch eine Nummer und – was nicht immer notwendig ist – eine verbale Bezeichnung zugeordnet.

Die Klassen in diesem Beispiel haben noch wichtige Eigenschaften:
Die Klassen sind gleich groß (alle Klassen erfassen ein Intervall von 350 DM), überschneiden sich nicht (es gibt keinen Punkt auf der Skala, der in zwei Klassen liegt) und alle gemessenen Werte liegen in einer jener Klassen.
Die beiden letzten Eigenschaften sind notwendig bei der Einteilung von Skalen in Klassen, während die Entscheidung über die Breite und damit auch die Anzahl der Klassen, in welche die Skala eingeteilt werden soll, vom Auswertenden selbst getroffen werden muss. Dafür gibt es keine fest vorgegebenen Regeln, aber einige Hinweise können trotzdem hilfreich sein:
Man sollte sich bei der Klasseneinteilung das Ziel der Klassierung vor Augen halten. Nach der Klassierung sollten in den meisten Klassen genügend viele Werte liegen, auf der anderen Seite muss die Klassenbreite so klein sein, dass auch noch genügend viele solcher Klassen gebildet werden können.

2.2.2 Aufstellen der Klassenhäufigkeiten

Nach der Bildung der Klassen können die zu jeder Klasse gehörenden absoluten Klassenhäufigkeiten berechnet werden. Dabei werden für jede Klasse die Anzahl der Merkmalsträger ermittelt, für welche die Merkmalsausprägung innerhalb der Klassengrenze liegt. Addiert man dies absoluten Klassenhäufigkeiten zusammen, muss sich (wie bei der absoluten Häufigkeit der Daten ohne Klassenbildung) die

Gesamtanzahl von gemessenen Merkmalsträgern ergeben. Ist diese Summe zu groß, liegt mindestens ein Merkmalsträger in zwei Klassen; ist die Summe zu klein, gibt es mindestens einen Merkmalsträger, der in keiner Klasse liegt.

Kap.2 Definition 10:
Die Klassenhäufigkeit h_j der Klasse j wird durch Addition der absoluten Häufigkeiten aller Messwerte, die innerhalb der Klasse j liegen, bestimmt:

$$h_{kl\ j} = \sum_{u_j \le x_i < o_j} h_i \qquad \text{für halbseitig offene Intervalle ,} \tag{2.9}$$

$$h_{kl\ j} = \sum_{u_j \le x_i \le o_j} h_i \qquad \text{für geschlossene Intervalle,} \tag{2.10}$$

mit u_j als untere und o_j als obere Klassengrenzen der Klasse j.

Aus den absoluten Klassenhäufigkeiten können die relativen und prozentualen Klassenhäufigkeiten wie gewohnt berechnet werden.
Halbseitig offene Intervalle bedeuten dabei, dass eine der Intervallgrenzen nicht zur Klasse gehört, bei geschlossenen Intervallen gehören die Intervallgrenzen zu dem Intervall dazu. Geschlossenen Intervalle verwendet man daher bei der Klassierung diskreter Merkmale, halboffene Intervalle bei stetigen Merkmalen.
Das Ergebnis der Einteilung ist eine interpretierbare Verteilungsfunktion. Die unüberschaubare Vielzahl unterschiedlichster Messwerte mit ihren sehr niedrigen Häufigkeiten wurde durch eine übersichtliche Anzahl von Klassen und den dazugehörigen großen Häufigkeiten ersetzt. Die Vielfalt der Ausprägungen wurde damit auf wenige wesentliche aber dennoch aussagekräftige Ausprägungen reduziert – natürlich nur insofern die Wahl der Klassengrenzen sinnvoll vollzogen wurde. Große Datenmengen werden auf diese Weise übersichtlicher, bei gleichzeitiger Verringerung des Arbeitsaufwandes für die weiteren Bearbeitungsschritte.
Eine Klasseneinteilung für unser Beispiel 4 (siehe S. 62) könnte folgendermaßen aussehen:

Einkommen	770	794	804	806	1003	1266	1278	1399	1523	1576
Klasse	1	1	1	1	2	3	3	3	3	3

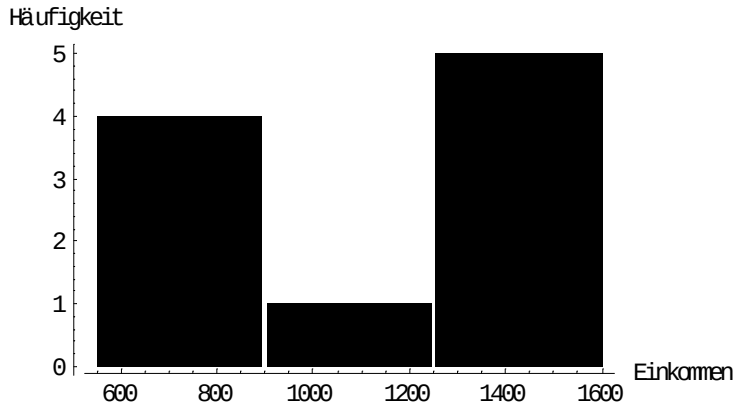
Damit ergeben sich folgende absolute Klassenhäufigkeiten:

Kap.2 Tabelle 2.

Klasse (k)	Klassenhäufigkeiten H_k
1	4
2	1
3	5

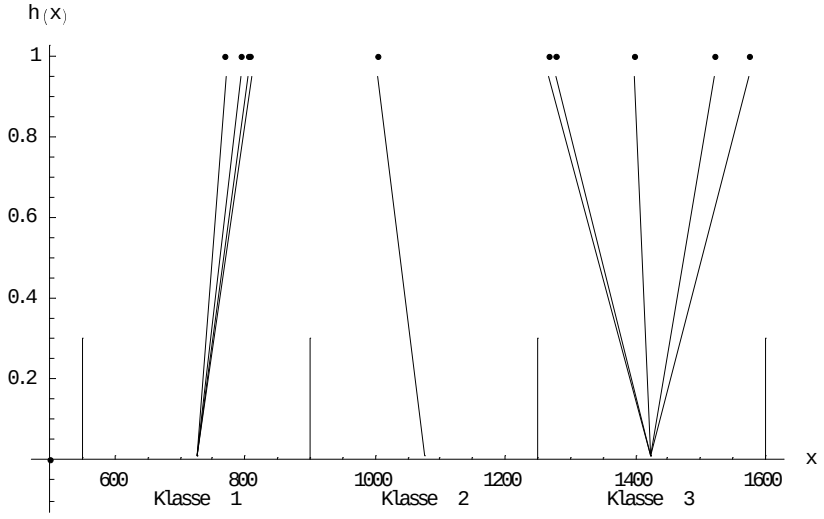
Diese Häufigkeiten sollen jetzt zur besseren Übersicht grafisch dargestellt werden. Dabei stellt sich die Frage, welchen Werten diese Klassenhäufigkeiten jetzt zugeordnet werden sollen. Bei den unklassierten Daten wurden die Häufigkeiten den

Messwerten selbst zugeordnet. Hier beziehen sich die Häufigkeiten aber auf Intervalle. Also kann man die Klassenhäufigkeiten den gesamten Intervallen zuordnen. Das Ergebnis ist dann ein Balkendiagramm.



Kap. 2 Abbildung 7

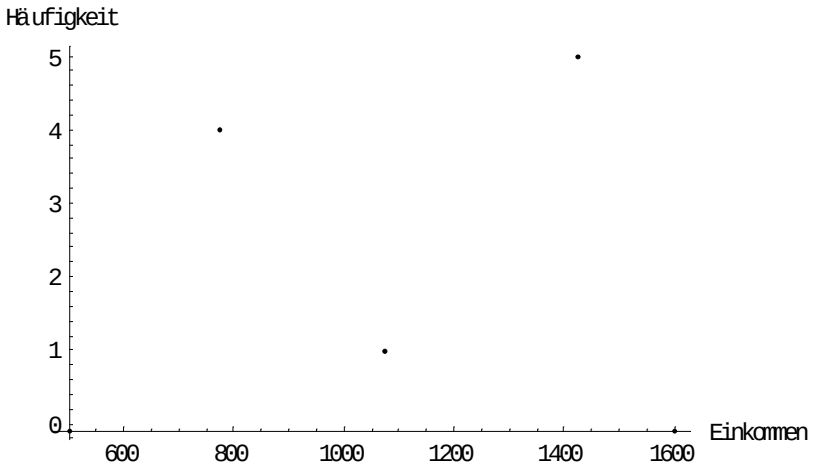
Dieses Vorgehen ist allerdings nicht üblich. Die Klassenhäufigkeiten werden - wie auch die gewöhnlichen Häufigkeiten - einem bestimmten Wert zugeordnet. Im Falle einer Klassierung werden die Klassenhäufigkeiten den Klassenmitten zugeordnet.



Kap.2 Abbildung 8

Dabei gehen alle Informationen darüber verloren, wie die Daten innerhalb der Klassengrenzen verteilt sind. An der Darstellung erkennt man, dass in der Klasse 1 die Originaldaten größer sind als die Klassenmitte, in Klasse 2 sind die Original-

daten kleiner als die Klassenmitte, in Klasse 3 etwa gleichverteilt. Nach der Zuordnung zu Klassen können keine Aussagen über die Anordnung der Daten in den Klassen gemacht werden. Deshalb wird angenommen, sie wären innerhalb der Klassengrenzen gleichverteilt.
Nach der Zuordnung der Daten zu den Klassenmitten ist die Darstellung der Häufigkeitsverteilung in einem Punktdiagramm möglich (allerdings erkennt man an der grafischen Darstellung nicht mehr, dass es sich um in Klassen eingeteilte Daten handelt!).



Kap.2 Abbildung 9

Die Verteilungsfunktion hat nach der Klasseneinteilung deutlich an Kontur gewonnen. Aus ihr ist nunmehr abzulesen, dass sich die gemessenen Einkommen der Stichprobe besonders in zwei Bereichen konzentrieren, und zwar in den Klassen „geringer Verdienst“ und „hoher Verdienst“.

Sozialwissenschaftler arbeiten jedoch sehr selten mit metrischen Daten, wesentlich häufiger werden ordinalskalierte Daten vorliegen, etwa zur Charakterisierung von Merkmalen wie Aggressivität, soziale Kompetenz, Zufriedenheit oder Geltungsbedürfnis. Bei deren Auswertung kommt es lediglich auf die Einhaltung einer Rangordnung an.

Beispiel 5:

Von Personalabteilungen werden häufig Persönlichkeitstests mit Bewerbern durchgeführt. Darin werden beispielsweise Merkmale wie „soziale Kompetenz“ getestet. Dieses Merkmal wird etwa aus Faktoren Frustrationstoleranz, Reizbarkeit, Integrationsvermögen, Menschenkenntnis gebildet. Auf einer neunstufigen Rangskala gemessen, könnte es wie folgt klassiert werden: die Ausprägungen 1 bis 3 stehen für „mangelhaft“, 4 bis 6 für „ausreichend“ und 7 bis 9 für „gut“.

> Denkbar wäre es auch, solcher Test für Studenten einzuführen, die den Lehrerberuf ergreifen wollen. Da Lehrer zwangläufig durch ihren Beruf viel Umgang mit Menschen haben, müssten Lehramtsanwärter in einem Test für „soziale Kompetenz“ mindestens die Klasse „ausreichend“ erreichen, um überhaupt für diesen Beruf als geeignet zu gelten.

Bisher haben wir die Einteilung von Merkmalen in Klassen unter dem Aspekt, eine Datenmenge auf eine interpretierbare Menge zu reduzieren betrachtet. Beleuchten wir die Nützlichkeit der Einteilung eines Merkmals in Klassen nochmals von einer anderen Seite.
In den Sozialwissenschaften beschäftigt man sich natürlich nicht nur mit der Auswertung von Fragebögen, in denen die Mehrzahl der gemessenen Merkmale ordinalskaliert und damit diskret sind, sondern auch mit stetigen Merkmalen etwa physiologischen Daten wie EEG-Signale, Puls, Hautleitfähigkeit oder die Frequenz der Stimme.
Die Wahrscheinlichkeit des Auftretens einer Merkmalsausprägung lässt sich bei diskreten Merkmalen leicht durch die Häufigkeiten beschreiben. Im Gegensatz dazu lassen sich Aussagen bei stetigen Merkmalen nicht direkt über die aufgetreten Elementarereignisse fällen. Man erinnere sich an das zur Einführung in diesen Abschnitt verwendete Beispiel 4. In den meisten Fällen beträgt die absolute Häufigkeit pro Merkmalsausprägung bei stetigen Merkmalen den Wert eins und jene der nicht aufgetretenen Null. Würde man diese Häufigkeiten auf die gleiche Weise interpretieren wie bei diskreten Merkmalen, so käme man zu dem Schluss, dass jene Messwerte, deren Häufigkeit Null beträgt, auch zukünftig nicht auftreten werden. Was sich bei nüchterner Betrachtung als Widerspruch zu der Stetigkeitsannahme herausstellt.
Das Problem, auf das man hierbei stößt, liegt in der Sache an sich. Bei unendlich genauer Messung können zwei stetige Merkmalsausprägungen noch durch die winzigsten Abweichungen voneinander unterschieden werden. Die Grundfrequenz der Stimme ist beispielsweise bei allen Menschen unterschiedlich. Wohl aber besteht eine hohe Wahrscheinlichkeit, dass die Grundfrequenz der Stimme eines Mannes im tiefen Frequenzbereich liegt. Um Aussagen über die Wahrscheinlichkeiten bei stetigen Merkmalen zu treffen, spricht man nicht von Wahrscheinlichkeiten des Auftretens einer einzelnen Merkmalsausprägung, sondern von der Wahrscheinlichkeit mit der sich eine Merkmalsausprägung in einem bestimmten Intervall befindet. Da Klassen Intervalle repräsentieren, ist eine Einteilung eines stetigen Merkmals in Klassen unabdingbar, um Aussagen über Wahrscheinlichkeiten fällen zu können.

Damit sind wir wieder bei unserem Ausgangspunkt angelangt. Die Einteilung von Merkmalen in Klassen ist nach den obigen Erläuterungen nicht nur dort nützlich, wo es gilt, eine Vielzahl von Ausprägungen auf ein überschaubares Maß zu reduzieren, sondern ist auch überall dort notwendig, wo es gilt, Fragen über die Wahrscheinlichkeiten bei stetigen Merkmalen zu erörtern.
Im Abschnitt über die graphische Darstellung eines Merkmals werden wir auf diesen Sachverhalt zurückkommen, da die Eigenschaft der Stetigkeit maßgebli-

ches Entscheidungskriterium über die Anwendbarkeit bestimmter Darstellungsformen ist.

2.2.3 Offene Klassen

In einigen Fällen ist die gleichzeitige Angabe von oberer und unterer Klassengrenze unmöglich oder unsinnig. Bei der Untersuchung des monatlichen Einkommens zum Beispiel ist es nicht möglich, eine geschlossene letzte Klasse zu bestimmen, weil das höchste mögliche Einkommen nicht vorhergesagt werden kann. In solchen Fällen benutzt man offene Klassen (manchmal auch halboffen genannt – nicht zu verwechseln mit halboffenen Intervallen!).
Bei solchen sogenannten *offenen Klassen* gibt man gewöhnlich nur eine Klassengrenze exakt an und lässt die andere Klassengrenze offen. Ein Beispiel für die Verwendung einer offenen Klasse ist die „Einkommen von 10.000 DM und mehr“. Diese Art der Klassendefinition zieht allerdings auch Probleme nach sich. Während bei offenen Klassen bei der Berechnung der absoluten und relativen Häufigkeiten keine Komplikationen auftreten, ist die Berechnung von Klassenmitten oder die Erstellung von Histogrammen nicht wie bisher möglich, da hier – wie für viele Parameterberechnungen – genau bestimmte Klassenbreiten benötigt werden. Für die graphische Darstellung kann man sich mit einem Trick aushelfen. Für Klassierungen mit sonst gleicher Klassenbreite kann man den halboffenen Klassen am Rand der Verteilung dieselbe Klassenbreite zuordnen wie den geschlossenen Klassen. Für die Berechnung von Verteilungsparametern aus klassifizierten Daten bleiben allerdings die Probleme bei halboffenen Klassen noch bestehen. Grundsätzlich sollten daher nur offene Klassen verwendet werden, wenn es sich nicht umgehen lässt.

2.2.4 Exakte Klassengrenzen

Das folgende Beispiel demonstriert, warum es notwendig ist, sich nochmals über die Klassengrenzen Gedanken zu machen:
Es wurde an 14 verschiedenen Tagen die Lufttemperatur gemessen. Dabei stand ein Thermometer mit °C- Einteilung zur Verfügung:

Tag	1	2	3	4	5	6	7	8	9	10	11	12	13	14
°C	8	9	11	12	14	11	11	14	15	18	20	19	19	22

Da nur ganzzahlige Werte auftraten, versuchen wir eine Klasseneinteilung der folgenden Art:

Kap.2 Tabelle 3.

Temperatur in °C	Absolute Klassenhäufigkeit
8-10	2
11-13	4

14-16	3
17-19	3
20-22	2

Bei dieser Klasseneinteilung scheinen im betrachteten Skalenbereich alle möglichen Messwerte einer Klasse zugeordnet zu sein. Trotzdem treten zwei Probleme auf:

- Zum einen werden in jeder Klasse drei Messwerte zusammengefasst. Die Klassenbreite beträgt aber bei jeder Klasse zwei.
- Zum anderen können beim Einsatz eines genaueren Messinstrumentes auch gebrochene Zahlen als Messwerte auftreten, etwa 13,6 °C. Diese Werte können dann nicht zugeordnet werden. Offensichtlich ist dies eine Folge der Messungenauigkeit des Erhebungsgerätes.

Auf dem ersten Blick scheint diese Messungenauigkeit eine Klassierung vorzunehmen, genau wie wir. Genauer betrachtet ergeben sich jedoch wesentliche Unterschiede. Bei der Messungenauigkeit eines Messgerätes wirkt ein interner Rundungsmechanismus. Wird z.B. die Luft mit einer Temperatur von 13,6 °C mit unserem Thermometer gemessen, das eine Messungenauigkeit von 1 °C hat, wird man eine Temperatur von 14 °C ablesen (weil das Thermometer alle Werten von 13,5 bis 14,5 den Wert 14 zuordnet). Würde das Thermometer aber eine Klassierung wie wir vornehmen, etwa wie in unserem Einkommensbeispiel, würde man 13 °C ablesen ($13 \le x < 14$).
Wünschenswert wäre doch nun für uns eine Form der Klasseneinteilung nach diesem natürlichen Prinzip, das heißt unter Benutzung der Rundungsmechanismen.
Dazu muss man die maximale Auflösung des Messinstrumentes kennen, das ist die kleinste Differenz zwischen Merkmalsausprägungen, die das Messinstrument noch unterscheiden kann. In unserem Beispiel zu den Lufttemperaturen wäre das gerade 1 °C, im Beispiel zu der Einkommensuntersuchung könnte der Wert auf 1 DM festgelegt werden. Weiterhin müssen die Klassengrenzen, die bisher festgelegt wurden, korrigiert werden. Wurden diese als beidseitig abgeschlossene Intervalle gebildet, wie im Temperaturbeispiel, müssen die Klassen auf beiden Seiten um den halben Betrag der maximalen Auflösung des Messinstrumentes erweitert werden. Wurden die Klassenintervalle halbseitig offen gebildet (wie im Beispiel der Einkommensverteilung), werden die Klassen um den halben Betrag der maximalen Auflösung verschoben. Die so entstandenen Klassengrenzen nennt man exakte Klassengrenzen.

Damit ergibt sich für unser Beispiel der Lufttemperaturen folgende Klassierung:

Kap.2 Tabelle 4.

Temperatur in °C	Exakte Klassengrenzen	Absolute Klassenhäufigkeit
8-10	7.5-10.5	2
11-13	10.5-13.5	4
14-16	13.5-16.5	3
17-19	16.5-19.5	3
20-22	19.5-22.5	2

Damit erfolgt die Zuordnung von gebrochenen Temperaturwerten zu den Klassen nach den üblichen Rundungsregeln. Die Klassenbreite entspricht jetzt der Anzahl der zu der Klasse gehörenden ganzzahligen Merkmalsausprägungen.
Im Beispiel der Einkommensverteilung ergibt sich folgende Klassierung:

Kap.2 Tabelle 5.

Klasse (k)	Klassengrenzen	Exakte Klassengrenzen	Klassenhäufigkeiten H_k
1	550-900	550.5-900.5	4
2	900-1250	900.5-1250.5	1
3	1250-1600	1250.5-1600.5	5

Bei der Klasseneinteilung wird folgendermassen vorgegangen:

1. Zuerst wird die Anzahl der Klassen (k) gewählt. Die Wahl sollte sich dabei an sachlogischen Gegebenheiten orientieren.
2. Nun erfolgt die Wahl der Klassenbreiten und Klassengrenzen. In der Regel bildet man äqudistante (gleichgroße) Klassen. Das erleichtert u.a. die Berechnung der Klassenbreiten (b_k). Sie beträgt in dem Fall: $b_k = (x_{max} - x_{min}) \div k$. Andernfalls müssen die einzelnen Klassenbreiten individuell festgelegt werden. Dabei sollte darauf geachtet werden, dass die Dispersionsspanne zwischen x_{min} und x_{max} vollständig durch die erstellten Klassen abgedeckt wird.
3. Die exakte untere Klassengrenze (x_{ku}) und die exakte obere Klassengrenze (x_{ko}) muss für jede Klasse festgelegt werden.
4. Im Folgenden werden die absoluten Häufigkeiten der Klassen ermittelt. Getreu dem Prinzip: „liegt der Messwert im Intervall der Klasse, so ist er ihr zugehörig“, werden dazu die Messwerte ausgezählt. Dies kann mit Hilfe einer Strichliste geschehen.

Die Berechnung der Häufigkeiten vollzieht sich bei klassierten Merkmalen analog der Berechnung von Häufigkeiten bei nichtklassierten Merkmalen.

Die Anzahl der Messwerte, die in eine Klasse fallen, heißt *absolute Klassenhäufigkeit*.

Die Division der absoluten Klassenhäufigkeiten mit der Gesamtanzahl der Messwerte ergibt die relative Häufigkeit der Klasse (oder: *relative Klassenhäufigkeit*) (siehe Definition 2).
Die absolute Summenhäufigkeit einer Klasse (oder: *absolute Klassensummenhäufigkeit*) ist die Summe ihrer absoluten Häufigkeit mit den absoluten Häufigkeiten ihrer Vorgängerklassen (siehe Definition 7).
Die relative Summenhäufigkeit einer Klasse (oder: *relative Klassensummenhäufigkeit*) ist die Summe ihrer relative Häufigkeit mit den relativen Häufigkeiten ihrer Vorgängerklassen (siehe Definition 8).
Die Häufigkeitsfunktion (siehe Definition 4) und Summenhäufigkeitsfunktion (siehe Definition 8) des in Klassen eingeteilten Merkmales werden auf die gleiche Weise, wie bei den nichtklassierten Merkmalen, erstellt.

Beispiel 6:

Im Rahmen einer umfangreichen Untersuchung zur Nutzung der Verkehrsmittel befragte eine sächsische Stadt unter anderem 100 Studenten, die täglich auf dem Weg von ihrer Wohnung zur Universität einen Bus benutzen, wie viele Stationen sie dabei zurücklegen. Das Befragungsergebnis ist in folgende Urliste abgebildet:

lfd. Nummer	1	2	3	4	5	6	7	8	9	0
1 - 10	1	11	3	15	10	19	5	7	15	10
11 - 20	17	2	1	20	4	11	12	8	21	17
21 - 30	14	11	22	10	14	1	8	14	9	12
31 - 40	13	12	3	6	2	16	17	1	9	12
41 - 50	7	20	7	11	17	8	11	19	12	10
51 - 60	18	13	18	17	8	2	13	10	9	15
61 - 70	6	7	11	7	6	11	19	16	11	23
71 - 80	18	10	13	10	18	9	15	9	9	14
81 - 90	14	10	20	7	16	14	6	24	17	25
91 - 100	8	13	6	14	8	11	9	7	11	6

Da die Datenmenge hier bereits recht groß ist, sollen die Werte in Klassen eingeteilt werden.

1. Aufgrund der Dispersionsspanne zwischen einer und 25 benutzten Stationen sollen fünf Klassen mit einer Breite von je fünf Einheiten (Stationen) gebildet werden. Die Messwertskala wird in fünf gleich große Intervalle unterteilt, d. h. $k = 5$.
2. Die Breite aller Klassen ergibt sich aus der Division Anzahl der möglichen Merkmalsausprägungen und der Anzahl der Klassen ($b = 25 / 5 = 5$).

3. Die erste Klasse besitzt demzufolge die Klassengrenzen $x_{1u} = 1$ und $x_{1o} = 5$, die zweite $x_{2u} = 6$ und $x_{2o} = 10$ usw.).
4. Mit Hilfe einer Strichliste werden die Messwerte x_i, also die Anzahl der zurückgelegten Stationen, den Klassen zugeordnet.

Klasse	Klassen Grenzen	Exakte Klassen-grenzen	Strich-liste	absolute Klassenhäu-figkeit (f_i)	relative Klassenhäu-figkeit (h_i)	absolute Summen-häufigkeit	relative Summen-häufigkeit
1	1 ≤ x ≤ 5	0.5 ≤ x < 5.5	///// ///// /	11	0,11	11	0,11
2	6 ≤ x ≤ 10	5.5 ≤ x < 10.5	///// ///// ///// ///// ///// ////	34	0,34	45	0,45
3	11≤ x ≤ 15	10.5 ≤ x < 15.5	///// ///// ///// ///// ///// ///// /	31	0,31	76	0,76
4	16≤ x ≤ 20	15.5 ≤ x < 20.5	///// ///// ///// ////	19	0,19	95	0,95
5	21≤ x ≤ 25	20.5 ≤ x < 25.5	/////	5	0,05	100	1,00
				Σ = 100	Σ = 1		

2.2.5 Repräsentation einer Klasse durch die Klassenmitte

In den meisten Fällen bleibt es jedoch nicht bei einer Betrachtung und Interpretation der Häufigkeitsverteilung. Will man darüber hinaus Aussagen treffen, so ist es notwendig, jeder Klasse neben ihren Klassengrenzen einen Zahlenwert zuzuweisen, der als repräsentativer Wert aller in diese Klasse befindlichen Werte (ungeachtet ihrer tatsächlichen Ausprägung) gehandelt werden kann. Alle Messwerte, die in eine Klasse fallen, gelten daher als gleich groß. In der Regel verwendet man für den repräsentativen Wert die *Klassenmitte*, das arithmetische Mittel der beiden exakten Klassengrenzen.
Man beachte, dass das arithmetische Mittel nur bei metrischen Merkmalen und nicht bei Merkmalen, die auf dem Nominal- bzw. Ordinalniveau gemessen wurden, berechnet werden kann. Im allgemeinen ist eine Klasseneinteilung bei Ordinalskalierten Merkmalen auch nicht notwendig, da sich bei ihnen die Anzahl der Merkmalsausprägungen meist auf 4 bis 7 Möglichkeiten beschränkt.

Kap.2 Definition 12:
Die *Klassenmitte* x_{km} der Klasse k wird bei metrischen Merkmalen folgendermassen berechnet:

$$x_{km} = \frac{x_{ku} + x_{ko}}{2} . \tag{2.11}$$