Kapitel 3: Eindimensionale Häufigkeitsverteilungen

In Kapitel 2 wurden bereits Unterscheidungen hinsichtlich der Datenarten vorgenommen, auf die hier aufgebaut werden soll (vgl. Übersicht 3.1). Entscheidend auch für die in den folgenden Kapiteln behandelte Berechnung von Maßzahlen (gewogene und ungewogene Ansätze) ist danach die Unterscheidung in klassierte und unklassierte Daten und bei letzteren zwischen gruppierten Daten und Einzelbeobachtungen.

1. Unklassierte Daten

a) Häufigkeitsverteilung

Def. 3.1: Häufigkeiten

Seien $x_1,x_2,...x_m$ (gruppierte Daten) die m realisierbaren Ausprägungen eines diskreten Merkmals X, dann heißt die Anzahl der Beobachtungseinheiten mit der i-ten Ausprägung

(3.1) $n_i = n(x_i)$ absolute Häufigkeit (i = 1,2,...,m),

und mit $n = \Sigma n_i$ (Gesamthäufigkeit, Umfang der Beobachtungsgesamtheit) der Quotient

(3.2) $h_i = h(x_i) = \frac{n_i}{n}$ relative Häufigkeit

der i-ten Ausprägung des Merkmals X.

Bemerkungen und Folgerungen:

1. Offensichtlich ist n_i eine natürliche Zahl mit $n_i \geq 1$ und für die relativen Häufigkeiten gilt (bei nichthäufbaren Merkmalen):

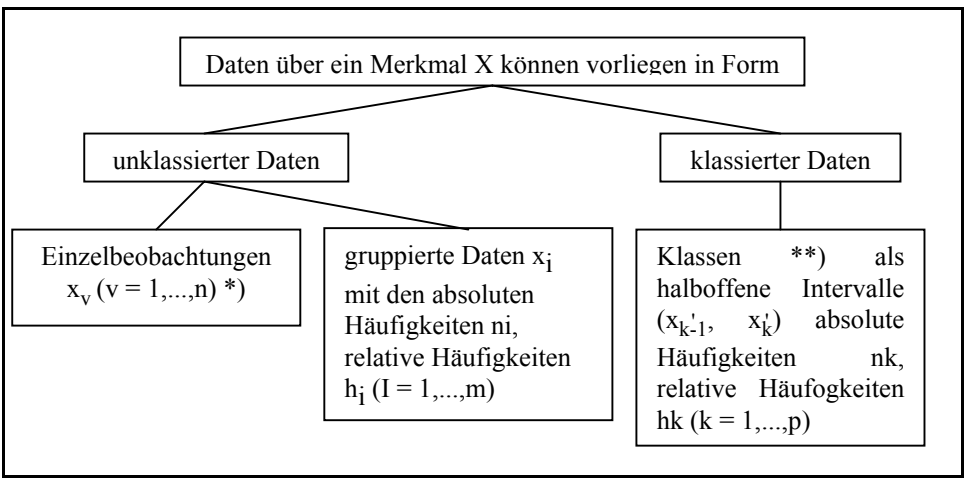
$0 \le h_i \le 1$ und (wegen $n = \Sigma n_i$) $\Sigma h_i = 1$.

Die mit 100 multiplizierten relativen Häufigkeiten heißen prozentuale Häufigkeiten.

2. Bei einem mindestens ordinalskalierten Merkmal sollten die Merkmalsausprägungen der Größe nach geordnet und numerisch codiert sein, so dass gilt $x_1 < x_2 < ... < x_m$.
   Eine solche Anordnung der Werte x_i ist erforderlich um kumulierte Häufigkeiten und die empirische Verteilungsfunktion (vgl. Def. 3.3) zu bestimmen. Eine Codierung (Zuordnung von Zahlen zu Merkmalsausprägungen) ist für die statistische Analyse nicht immer notwendig, erleichtert diese aber erheblich.

3. Bei den Merkmalswerten $x_1,x_2,...x_n$ der n Einheiten, die alle verschieden sind (Einzelbeobachtungen), ist $n_v = 1$ und $h_v = 1/n$ für alle Werte von v. Die Summe $\Sigma x_v = \Sigma x_i n_i = S$ heißt **Merkmalssumme**. Sie ist nur bei extensiven Merkmalen sinnvoll interpretierbar. Dichotome Merkmale werden zweckmäßig wie folgt codiert: $x_1 = 0$ und $x_2 = 1$, so dass die Merkmalssumme n_2 ist.

Übersicht 3.1: Daten



*) In späteren Abschnitten (insbes. im Kap. 8) wird gelegentlich auch x_i anstelle von x_v verwendet.

*) Es sei verabredet, dass x_k' die Obergrenze der k-ten Klasse (d.h. der k-ten der p aneinander grenzenden Größenklassen) ist, so dass x_{k-1}' die Obergrenze der (k-1)-ten Klasse und damit die Untergrenze der k-ten Klasse ist.

Def. 3.2: Häufigkeitsverteilung

Das m-Tupel $[(x_1,n_1),(x_2,n_2),...,(x_m,n_m)]$ heißt absolute Häufigkeitsverteilung und entsprechend ist $[(x_1,h_1),(x_2,h_2),...,(x_m,h_m)]$ die (relative) Häufigkeitsverteilung eines Merkmals X.

Sie ist eine Zuordnung von Häufigkeiten (n_i oder h_i) zu den Ausprägungen x_i (i = 1,2,...,m) des Merkmals X und zeigt, wie sich die n Einheiten über die möglichen Werte von X "verteilen". Häufigkeitsverteilungen können tabellarisch (Häufigkeitstabelle) oder graphisch dargestellt werden. Die Art der graphischen Darstellung hängt von der Skala des Merkmals X ab.

b) Tabellen und Graphiken

Für die praktische Anwendung der Statistik (weniger dagegen für die wissenschaftliche Beschäftigung mit Statistik) spielt die Gestaltung von Tabellen und "aussagefähigen" und eindrucksvollen Graphiken (meist mit einer speziell hierfür entwickelten Software) eine große Rolle. Es kann hier auf diese Gegenstände nur sehr kurz eingegangen werden.

Tabellen:

Für die Gestaltung von Tabellen gibt es Normen (DIN-Normblatt 55301). Eine Tabelle ist eine geordnete Zusammenstellung der Ergebnisse statistischer Erhebungen oder Berechnungen mit **Zeilen** (waagrecht) und **Spalten** (senkrecht), wobei in den so gebildeten **Tabellenfächern** i.d.R. Häufigkeiten eingetragen werden. Eine Tabelle hat stets eine Überschrift und eine Quellenangabe. Sie kann auch Fußnoten haben. Zeilen und Spalten können numeriert werden. Die **Überschrift** soll enthalten: Dargestellte Tatbestand, räumliche und zeitliche Abgrenzung der Erhebungsmasse. **Fußnoten** können Erklärungen zu Zahlen in einzelnen Tabellenfächern oder ergänzende Hinweise zu Textangaben enthalten.

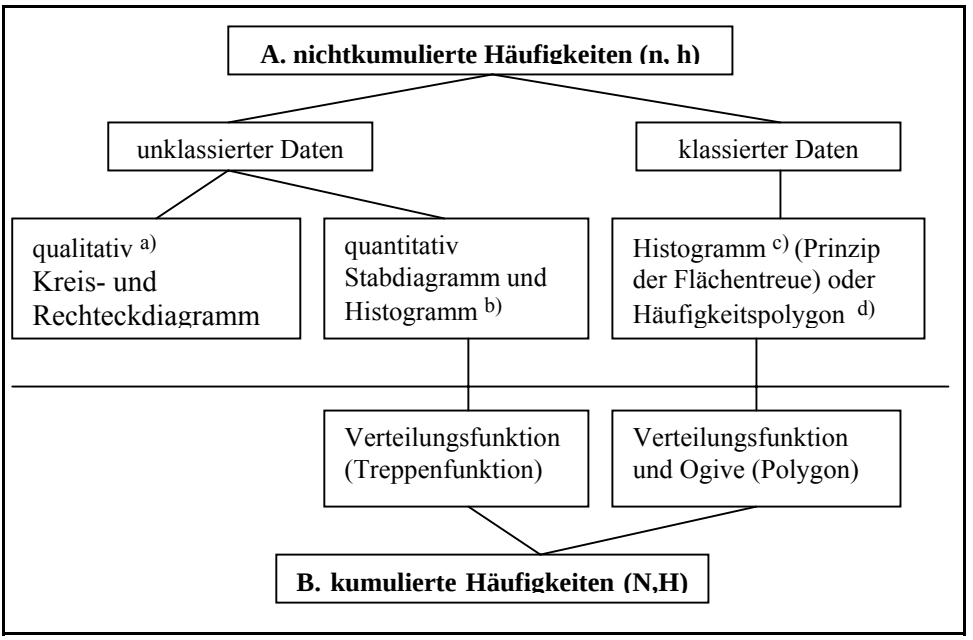
Graphiken:

Es gibt eine Fülle von Gestaltungsmöglichkeiten für graphische Darstellungen. Abgesehen von Piktogrammen (Bildgraphiken mit Verwendung anschaulicher Symbole) und Kartogrammen handelt es sich jedoch meist um Varianten der in Übers. 3.2 genannten Diagramme.

1. **Bei qualitativen Merkmalen** ist eine Reihenfolge der Merkmalsausprägungen nicht definiert, so dass die Häufigkeitsverteilung (bzw. in

diesem Fall, die **Struktur**) des Merkmals X am besten durch ein **Kreisdiagramm** (engl. pie chart) dargestellt wird (vgl. Beispiel 3.1 und Abb. 3.1 links). Eine Alternative ist das **Rechteckdiagramm** (z.T. auch Flächendiagramm genannt, Abb. 3.1 rechts).
Die Winkel a_i der Kreissegmente (und damit die Flächen der Kreissektoren) bzw. die Höhen der Rechtecke des Rechteckdiagramms sind proportional zu den Häufigkeiten. Damit verhalten sich auch die Flächen zueinander wie die relativen, bzw. absoluten Häufigkeiten und es gilt $a_i = 360°h_i$. Die Reihenfolge der Sektoren (Kreissegmente) kann beliebig gewählt werden (für X wird ja nur eine Nominalskala vorausgesetzt).

Übersicht 3.2: Graphische Darstellung von Häufigkeiten



a) kategorial, nominalskaliert;
b) in diesem Fall Stäbe, Säulen oder (nicht notwendig aneinander angrenzende) Blöcke gleicher Breite;
c) bei gleichen Breiten (äquidistante Klassen) ist die Höhe und die Fläche sowie bei ungleichen Breiten die Fläche der aneinander angrenzenden Blöcke proportional zur absoluten oder relativen Häufigkeit;
d) lineare Verbindung der Blockmitten (auch Kurvendiagramm genannt);
e) kumulierte Häufigkeiten (Summenhäufigkeiten) gem. Def. 3.3 (bei Resthäufigkeiten [Def. 3.4] erhält man jeweils fallende Treppenkurven).

Vergleicht man die Verteilungen mehrerer Massen A und B unterschiedlichen Umfangs (n_A,n_B) bezüglich des Merkmals X miteinander, so wird der Unterschied des Umfangs durch entsprechend unterschiedlich große Flächen der Kreise, bzw. Rechtecke zum Ausdruck gebracht. Da die Kreisfläche $F_A=r_A,^2p$ ist (und F_B entsprechend), muss für die Quadrate der Radien gelten $r_A^2 / r_B^2 = n_A / n_B$.

2. Bei **quantitativen diskreten Merkmalen** empfiehlt sich die Darstellung eines **Stabdiagramms** (andere Ausdrücke hierfür sind: **Balken-, Block-** oder **Säulendiagramm**, allgemein: **Histogramm**). Die Höhen der Stäbe bzw. Säulen sind proportional zu den relativen oder absoluten Häufigkeiten (Abb. 3.2). Bei ordinalskalierten Merkmalen sind ihre Abstände nicht eindeutig und bei qualitativen Merkmalen wäre auch ihre Reihenfolge nicht eindeutig.

Beispiel 3.1:

a) Im Jahr 1989 ergaben sich für die Bundesrepublik Deutschland folgende Anteile der einzelnen Wirtschaftsbereiche an der Gesamtzahl der Erwerbstätigen (Quelle: Gutachten SVR 1990):

Staat und Private Haushalte	19,8%
Dienstleistungsunternehmen	18,0%
Handel und Verkehr	18,7%
Warenproduzierendes Gewerbe	39,8%
Land- und Forstwirtschaft, Fischerei	3,7%

Veranschaulichen Sie die Struktur der Erwerbstätigen durch ein Kreis- und Rechteckdiagramm!

b) An einer Straßenkreuzung wurden an 128 Tagen die Unfallzahlen (Anzahl der Unfälle an einem Tag) gemessen.

Anzahl der Verkehrsunfälle (x_i)	0	1	2	3	4
Anzahl der Tage (n_i)	13	26	38	32	19

Stellen Sie die Häufigkeitsverteilung durch ein Stabdiagramms dar!

Lösung 3.1:

a) Zum Kreis- und Rechtecksdiagramm vgl. Abb. 3.1. Die Winkel des Kreisdiagramms sind:

Staat und Private Haushalte (Staat)	71° 17'
Dienstleistungsunternehmen (Dienst)	64° 48'
Handel und Verkehr (H+V)	67° 20'
Warenproduzierendes Gewerbe (Gew)	143° 17'
Land- und Forstwirtschaft, Fischerei (L+F)	13° 19'

b) Zum Stabdiagramm (oder Blockdiagramm) vgl. Abb. 3.2.

c) Summenhäufigkeiten

Def. 3.3: Summenhäufigkeit, Verteilungsfunktion

a) Die Summe N_i der absoluten Häufigkeiten n_j (j=1,2,...,i) aller Merkmalsausprägungen x_j eines mindestens ordinalskalierten Merkmals, die kleiner oder gleich x_i sind,

$$(3.3) \quad N_i = N(x_i) = n\,(X \le x_i) = \sum_{j=1}^{i} n_j$$

heißt absolute kumulierte Häufigkeit (**absolute Summenhäufigkeit**).

Abb. 3.1: Kreis- und Rechteckdiagramm

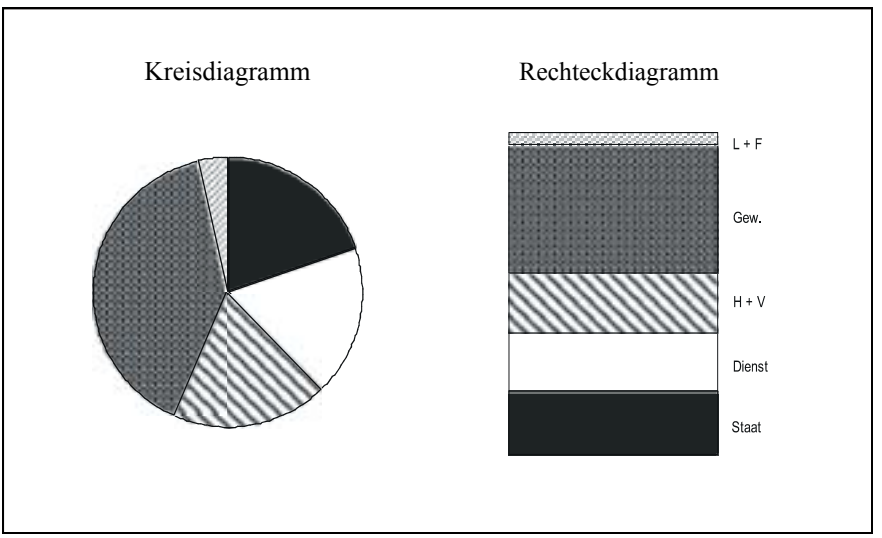
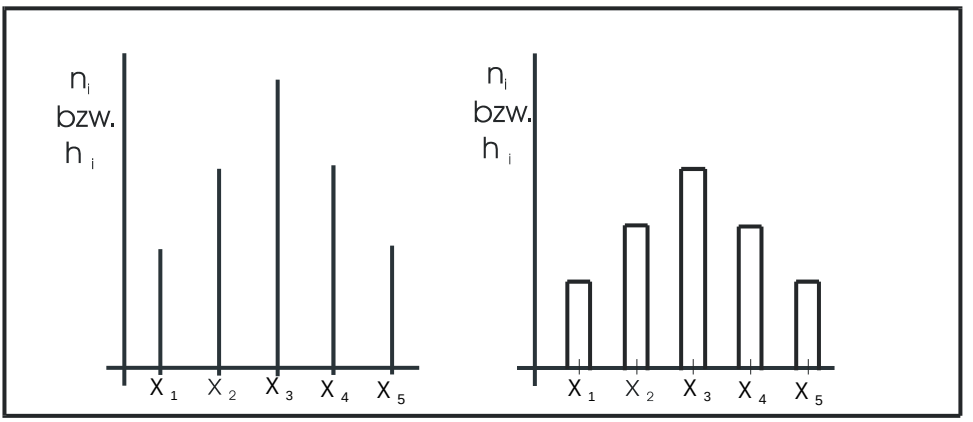


Abb. 3.2: Stabdiagramm



b) Entsprechend heißt

$$(3.4) \quad H_i = H(x_i) = h(X \le x_i) = \sum_{j=1}^{i} h_j = \frac{N_i}{n}$$

relative kumulierte Häufigkeit (relative Summenhäufigkeit).

c) Die Funktion

$$(3.5) \quad H(x) = \begin{cases} 0 & \text{für } x < x_1 \\ H_j & \text{für } x_j \le x < x_{j+1} (j = 1,2,...,m-1) \\ 1 & \text{für } x \mid x_m \end{cases}$$

der reellen Variable X heißt (empirische) **Verteilungsfunktion** oder (relative) Summenhäufigkeitskurve des diskreten Merkmals X.

d) Die Funktion x = G(H) ist die **inverse Verteilungsfunktion**.

Bemerkungen zu Def. 3.3:

1. Wie man leicht sieht, gilt: $N_1 = n_1$, $N_2 = n_1 + n_2$, $N_3 = n_1 + n_2 + n_3$ usw. und schließlich $N_m = \Sigma\, n_i = n$ (mit i=1,2,...,m). Ferner ist $H_m = 1$.
2. Im Fall von n verschiedenen **Einzel**beobachtungen $x_1,...,x_n$ gilt: $N_i = i$ und $H_i = i/n$.
3. Die empirische Verteilungsfunktion H(x) gibt die Summe der relativen Häufigkeiten aller Merkmalswerte an, die kleiner oder gleich x sind. Während die einzelnen Summenhäufigkeiten H_i dargestellt werden als Stäbe, bei denen die Häufigkeiten h_j "gestapelt" werden (stacked histogram) und zwischen den Stäben Lücken sind, hat

der Graph der Funktion H(x) die Gestalt einer Treppe. H_i ist geeignet für unklassierte Daten, H(x) auch für klassierte Daten.

4. H(x) ordnet jedem Wert x die bis dahin (also für $X \le x$) erreichte Summenhäufigkeit H zu. Mit der inversen Funktion G kann man für bestimmte Werte von H (etwa H = 1/2 oder H = 1/4) den Wert x bestimmen (was im Falle von 1/2 und 1/4 die Quartile Q_1 und $Q_2 = \tilde{x}_{0,5}$ [vgl. Kap. 4] sind).

5. Weniger gebräuchlich ist die Darstellung der absoluten Summenhäufigkeiten N_j als absolute Summenhäufigkeitskurve. Es gilt N(x) = nH(x) für jedes x.

Eigenschaften der empirischen Verteilungsfunktion H(x):

a) Aus $x < y$ folgt $H(x) < H(y)$ (H ist monoton nichtfallend).

b) $0 \le H(x) \le 1$ (mit $H(-\infty) = 0$ und $H(\infty) = 1$ wenn der Definitionsbereich nicht beschränkt ist).

c) H(x) ist für alle $x \in \mathbb{R}$ definiert und eine rechtsseitig stetige Funktion.

d) H(x) hat als Treppenfunktion Sprungstellen bei $x_1, x_2, ..., x_m$. Die Größe der Sprünge beträgt $h_i = H(x_i) - H(x_{i-1})$.

e) Unter bestimmten Voraussetzungen (X ein nichtnegatives Merkmal) ist die Fläche oberhalb von H(x) das arithmetische Mittel (vgl. Abb. 4.3).

Def. 3.4: Resthäufigkeit

Die Summe N_i^- der absoluten Häufigkeiten n_j (j = i+1, i+2, ..., m) aller Merkmalsausprägungen eines mindestens ordinalskalierten Merkmals, die größer als x_i sind,

$$(3.6) \quad N_i^- = N^-(x_i) = n(x > x_i) = \sum_{j=i+1}^{m} n_j = n - N_i$$

heißt **absolute Resthäufigkeit**. Entsprechend heißt

$$H_i^- = 1 - H_i \qquad \textbf{(relative) Resthäufigkeit}$$

und die analog zu Gl. 3.5 definierte Funktion

$$H^-(x) = 1 - H(x) \qquad \textbf{relative Resthäufigkeitsfunktion}.$$

Die Resthäufigkeiten spielen in statistischen Anwendungen eine weniger wichtige Rolle. Es ist leicht zu sehen, dass gilt:

1. Die in Abb. 4.3 schraffierte Fläche unter $\bar{H}(x)$ ist gleich der Fläche oberhalb von H(x) und damit gleich dem arithmetischen Mittel.
2. Ferner ist $N_i + \bar{N}_i = n$, $H_i + \bar{H}_i = 1$ und $H(x) + \bar{H}(x) = 1$.

Beispiel 3.2:

Stellen Sie anhand der Daten des Bsp. 3.1b die Summenhäufigkeiten H_i und die Resthäufigkeiten $\bar{H}_i$ graphisch dar.

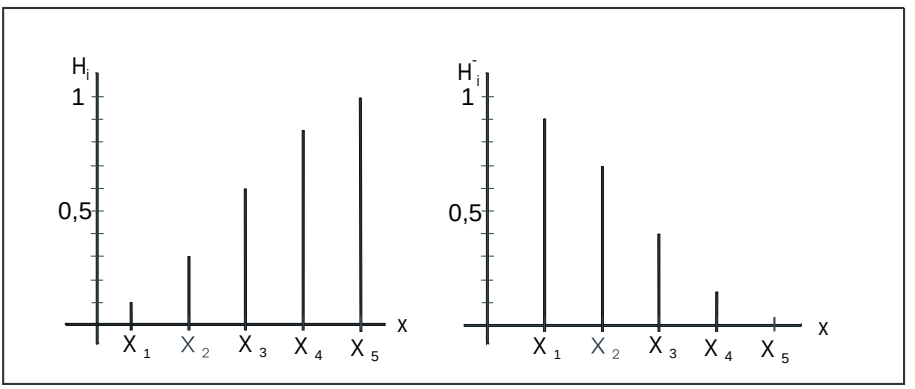
Lösung 3.2:

Zur Übersichtlichkeit wird folgende Arbeitstabelle aufgestellt:

Nr. i	Anzahl der Verkehrsunfälle (x_i)	Anzahl der Tage (n_i)	Anteil der Tage (h_i)*)	relative Summenhäufigkeit H_i	Resthäufigkeit $\bar{H}_i$
1	0	13	0,1	0,1	0,9
2	1	26	0,2	0,3	0,7
3	2	38	0,3	0,6	0,4
4	3	32	0,25	0,85	0,15
5	4	19	0,15	1,0	0

*) gerundet

Abb. 3.3: Summenhäufigkeiten H_i und Resthäufigkeiten $\bar{H}_i$ für das Beispiel 3.2



2. Klassierte Daten

a) Größenklassen

Notwendigkeit der Klassenbildung

Häufig enthält das Datenmaterial so viele unterschiedliche Merkmalsausprägungen, dass eine Darstellung sämtlicher Beobachtungen nach Art eines Stabdiagramms wenig aufschlußreich wäre (vgl. Abb. 3.4 wo die absoluten Häufigkeiten n_i jeweils nur 1 oder 2 betragen). In diesem Fall ist eine Klassenbildung (Klassierung) zu empfehlen um die Gestalt der Verteilung besser erkennbar zu machen.
Das gilt v.a. bei stetigen Merkmalen wie Gewicht, Körpergröße, Alter, Länge von Schrauben etc., die zumindest theoretisch beliebig genau gemessen werden können und bei quasistetigen Merkmalen wie Einkommen, Vermögen, Sparguthaben etc. Aber auch bei diskreten Merkmalen wie z.B. Punktzahlen in einer Klausur, IQ-Werte, Stückzahlen etc. kann eine unklassierte Verteilung sehr unübersichtlich sein, wie das Beispiel 3.3 (Abb. 3.4) zeigt.

Intervallabgrenzung

Für ein mindestens ordinalskaliertes Merkmal X lassen sich (beidseitig) offene oder geschlossene Intervalle wie folgt abgrenzen:
(a,b) soll bedeuten $a < x < b$ (offenes Intervall) und
[a,b] soll bedeuten $a \le x \le b$ (geschlossenes Intervall).
In diesen Fällen entstehen jedoch dann Unklarheiten, wenn x die (Grenz- oder Eck-) Werte x = a und x = b annimmt. Eine widerspruchsfreie Intervallabgrenzung ist jedoch möglich mit (a,b]:
$a < x \le b$ (oder mit [a,b), also wenn gilt $a \le x < b$).
Ist x_k' die Obergrenze der k-ten Größenklasse, dann ist x_{k-1}' die Untergrenze dieser Klasse. Damit lassen sich die Begriffe der Def. 3.5 definieren:

Def. 3.5: Klassierung

a) In einer klassierten Verteilung wird die Variable X in p Intervalle (**Klassen**) $(x_{k-1}', x_k']$ eingeteilt (linksseitig offene Intervalle) mit k = 1,2,..,p wobei x_k' die Obergrenze der k-ten Größenklasse ist (vgl. Bem. Nr. 4).

b) Die Differenz $b_k = x_k' - x_{k-1}'$ heißt **Klassenbreite**
und die Größe $m_k = ½(x_{k-1}' + x_k')$ heißt **Klassenmitte** der k-ten Klasse.

c) Die Anzahl n_k der betrachteten Einheiten, die in die k-te Klasse "fallen" [$n_k = n\ (x'_{k-1} < x \leq x'_k)$], ist die absolute **Klassenhäufigkeit** (der k-ten Klasse) und die Anteile $h_k = n_k/n$ sind die relativen Klassenhäufigkeiten.

d) Die Quotienten $h^*_k = h_k/b_k$ [k=1,2,...,p] sind **Häufigkeiten je Klassenbreite**. Mit $b_k \rightarrow 0$ wird X zu einer stetigen Variable und das Häufigkeitspolygon zu einem kontinuierlichen Kurvenzug (zur Dichtefunktion).

Bemerkungen zu Def. 3.5:

1. Geht man davon aus, dass sich die Merkmalsträger in den Klassenmitten konzentrieren, so können die Daten wie gruppierte Daten behandelt werden. Durch Klassierung wird ein stetiges Merkmal quasi zu einem diskreten (Diskretionierung). Klassenbildung ist eine Transformation, bei der Information (nämlich die Verteilung innerhalb der Klasse) verloren geht.

2. Die **Klassenmitten** m_k sind i.d.R. nicht identisch mit den wahren Klassenmittelwerten, es sei denn die Einheiten verteilen sich gleichmäßig innerhalb einer Klasse. m_k heißt auch "Präsumptivwert" (also Schätzwert des [wahren] **Klassenmittelwerts**).

3. Für die Entscheidung über Anzahl und Breite der Klassen lassen sich keine formalen Kriterien angeben. Hierbei sind auch Manipulationen möglich, d.h. das gleiche Datenmaterial kann optisch sehr unterschiedlich wirken (vgl. Abb. 3.5). Es sollte dabei berücksichtigt werden:

   - Zweck der Untersuchung
   - Meßgenauigkeit beim Merkmal X
   - Streuung der Merkmalswerte
   - Anzahl der Erhebungs- bzw. Darstellungseinheiten.

   Ungleiche Klassenbreiten empfehlen sich, wenn die Merkmalsausprägungen sehr unterschiedlich dicht liegen. Zu kleine Klassen lassen Meßfehler zu stark hervortreten, zu große Klassen verdecken wiederum Charakteristiken der Verteilung. Klassen sollten in jedem Fall so gebildet werden, dass keine leeren Klassen auftreten. Im allgemeinen wird man mit 5 bis 20 Klassen auskommen.

4. Häufig sind die erste und p-te Klasse nicht geschlossen. Mit solchen sog. offenen Randgruppen treten Schwierigkeiten bei der graphischen Darstellung und der Berechnung bestimmter Mittelwerte auf.

5. Üblicher als die obige Abgrenzung $(x'_k, x'_{k+1}]$ sind in der Praxis rechtsseitig offene Intervalle $[x'_k, x'_{k+1})$ "von ... bis unter ...". Die Abgrenzung $(x'_k, x'_{k+1}]$ wurde jedoch in Def. 3.5 gewählt um mit der Verteilungsfunktion (Def. 3.3) konsistent zu sein.

b) Graphische Darstellungen

Häufigkeitsverteilung bei einem klassierten Merkmal:

Graphisch wird eine klassierte Verteilung durch das Histogramm dargestellt. Es setzt sich aus Rechtecken über den Klassenbreiten b_k zusammen, deren Flächen proportional zu den Klassenhäufigkeiten h_k sind (**Prinzip der Flächentreue**). Daraus folgt, dass die Höhen der Rechtecke die Häufigkeitsdichten h_k^* sind. Häufigkeiten werden also zweidimensional (durch Flächen) repräsentiert, was für die Beurteilung schwieriger ist: man kann leicht Unterschiede in der Höhe von Blöcken feststellen aber nicht immer eindeutig die Fläche von Rechtecken vergleichen, es sei denn ein Rechteck ist in beiden Dimensionen (Höhe und Breite) größer als das andere. Gelegentlich verbindet man auch die Mitten der oberen Rechteckseiten eines Histogramms miteinander, wobei dieser Polygonzug auf der x-Achse im Wert $x_1' - \frac{1}{2}b_1$ beginnt und mit $x_p' + \frac{1}{2}b_p$ endet. Abb. 3.6 zeigt dieses **Häufigkeitspolygon** (dessen Gesamtfläche über der Abszisse gleich derjenigen des Histogramms ist) für ein fiktives Beispiel mit p = 5 Klassen.
Bei infinitesimal kleinen Klassenbreiten geht die Darstellung des Histogramms bzw. Häufigkeitspolygons über in eine **Dichtefunktion** eines stetigen Merkmals (die Dichte ist ein stetiger Kurvenzug). Sie ist nicht für die Deskriptive, sondern nur für die Induktive Statistik von Bedeutung.

Beispiel 3.3:

Die Messwerte für das Körpergewicht X von n = 25 Personen (in kg) seien: 63, 61, 70, 81, 72, 74, 69, 62, 75, 79, 77, 80, 86, 76, 78, 70, 80, 77, 73, 66, 85, 67, 83, 82, 71. Man stelle diese Daten dar als

- Stabdiagramm, also unklassiert (Abb. 3.4),
- klassierte Verteilung mit folgender Klasseneinteilung (Abb. 3.5):

a) einheitliche Klassenbreiten von jeweils 5 von 60 bis 90kg: beginnend mit "60 bis einschließlich 65" (also (60,65]), bis (85,90];
b) einheitliche Klassenbreiten von jeweils 10.

Lösung 3.3:

Das Stabdiagramm der Abb. 3.4 ist wenig sinnvoll, denn es ist kaum erkennbar, dass die Daten eine Verteilung darstellen. Bei den beiden Klasseneinteilungen (Abb. 3.5) zeigt sich, dass bei ein und demselben Datensatz eine unterschiedliche Gestalt der Verteilung möglich ist. Bei einer Klassierung mit einer einheitlichen Klassenbreite von $b_k = 5$ (Abb. 3.5a), also p = 6 Klassen treten die Charakteristiken der Verteilung recht gut hervor. Dagegen scheint eine Wahl von p = 3 Klassen zu grob zu sein. So kommt in Abb. 3.5b nicht zum Ausdruck, dass sich die Meßwerte in den Intervallen (70, 80] und (80, 90] in der oberen bzw. in der unteren Hälfte häufen.

Abb. 3.4: Daten des Beispiels 3.3 als Stabdiagramm

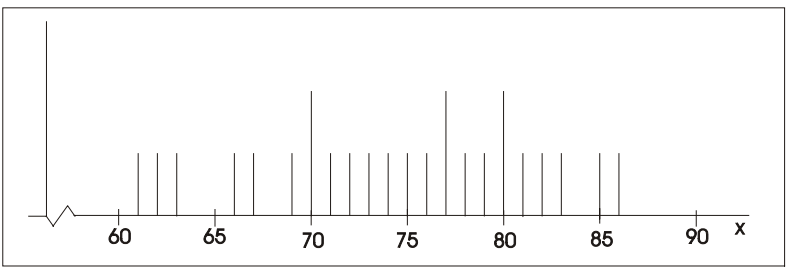
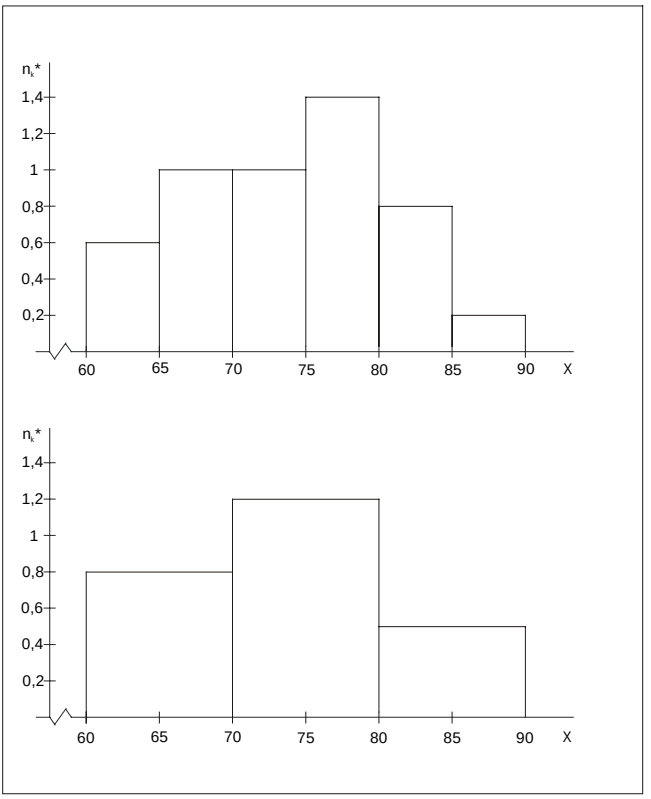


Abb. 3.5: Verschiedene Klasseneinteilungen für die Daten des Beispiels 3.3



Summenhäufigkeiten bei klassierter Verteilung:

Die Definitionen 3.2 und 3.3 für das diskrete Merkmal X gelten analog auch für ein klassiertes Merkmal X. Die Größen N_k bzw. H_k sind die absoluten bzw. relativen kumulierten Klassenhäufigkeiten. Letztere betragen $H_0 = 0$ vor der Untergrenze x_0' der ersten Klasse (denn $x_{1-1}' = x_0'$) und $H_p = 1$ nach der Obergrenze x_p' der letzten (p-ten) Klasse.

Analog zum Häufigkeitspolygon (Abb. 3.6) der Klassenhäufigkeiten kann man auch einen Polygonzug der Summenhäufigkeiten (kumulierten Klassenhäufigkeiten) bestimmen. Man nennt diese Kurve **Ogive** (Abb. 3.7 für das Beispiel der Abb. 3.6). Die Ogive[1] H(x) ist eine Näherung der exakten empirischen Verteilungsfunktion, die sich i.d.R. nicht angeben läßt, weil die Verteilung der Merkmalsträger (Einheiten)

[1] Die Ogive ist die lineare Verbindung der Treppenabsätze der Verteilungsfunktion H(x). Die Steigung ist dabei jeweils die Größe h_k^*. Um die Symbolik nicht zu kompliziert zu machen, soll die Ogive (oder "approximierende Verteilungsfunktion") auch H(x) genannt werden.

innerhalb der Klassen nicht bekannt ist. Sie ist unter der Annahme einer Gleichverteilung (Gleichhäufigkeit jeder Ausprägung in der Klasse) oder einer symmetrischen Verteilung innerhalb der Klassen konstruiert und deshalb eine stückweise lineare Funktion (Polygonzug) mit den genannten Eigenschaften von H(x).

Abb. 3.6: Histogramm und Häufigkeitspolygon p = 5 Klassen

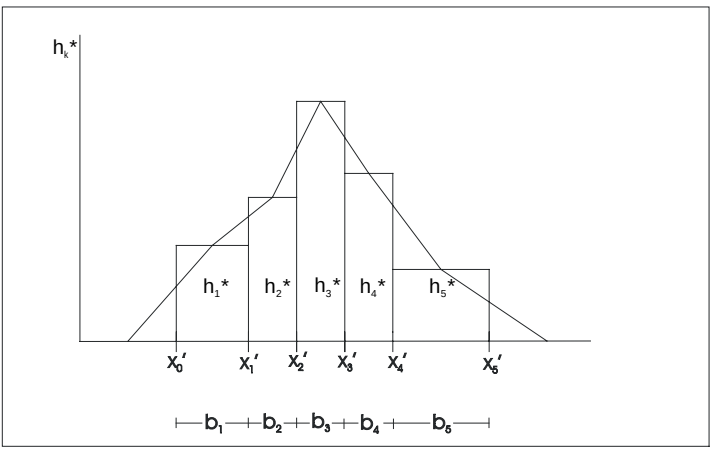
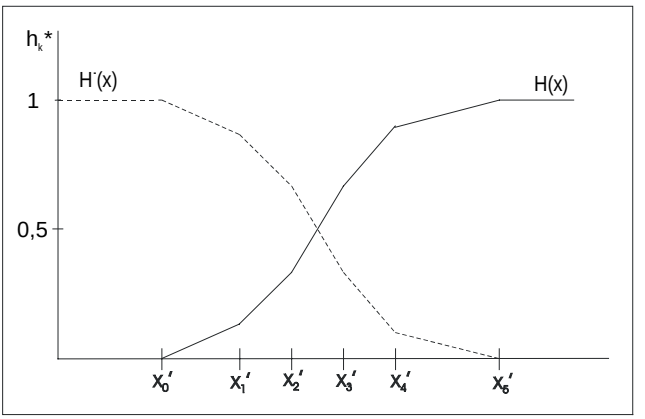


Abb. 3.7: Ogive H(x) und Resthäufigkeitsfunktion $H^-(x)$ für das Beispiel der Abb. 3.6



Beispiel 3.4:

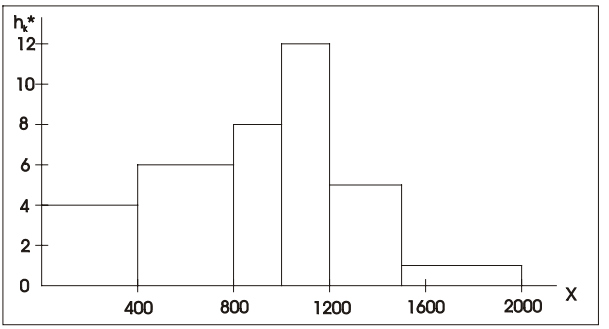
Gegeben sei die folgende Verteilung der Verdienste in einem Betrieb, für welche die Dichten h_K^*, sowie die Summen- und Resthäufigkeitskurve (also H(x) und $H^-(x)$) zu bestimmen und zu zeichnen sind:

Angaben		Lösung			
von...bis unter	h_k	Breite b_k	Dichte h_k^*	kumulierte Häufigk. H(x)	$H^-(x)$
0 - 400	0,16	400	4*)	0,16	0,84
400 - 800	0,24	400	6	0,40	0,60
800 - 1000	0,16	200	8	0,56	0,44
1000 - 1200	0,24	200	12	0,80	0,20
1200 - 1500	0,15	300	5	0,95	0,05
1500 - 2000	0,05	500	1	1	0

*) In dieser Spalte ist jeweils der 10000-fache Wert verzeichnet, d.h. die Angabe 4 in der ersten Zeile ist zu verstehen als $4 \cdot 10^{-4} = 0{,}16/400$. Die Dichte ist stets $h_k^* = h_k / b_k$. Es gilt $H^-(x) = 1 - H(x)$ und $H^-(x)$ ist 0,84 an der Stelle x = 400 und 1 bei x=0.

Abb. 3.8 zeigt die klassierte Verteilung mit den Höhen h_k^* und Abb. 3.9 die kumulierten Häufigkeiten H (Verteilungsfunktion) und Resthäufigkeiten H^- (gestrichelte Linie).

Abb. 3.8: Häufigkeitsverteilung der klassierten Verteilung des Beispiels 3.4



Man beachte, dass die Verteilungsfunktion (anders als bei einer diskreten Verteilung oder bei einer klassierten Verteilung mit gleich breiten Klassen) nicht einfach ein Aufeinanderstapeln der Stäbe bzw. Blöcke der Häufigkeitsverteilung sein kann. Denn das würde darauf hinauslaufen, die Höhen h_k^* und nicht (wie es richtig ist) die relativen Häufigkeiten h_k zu addieren.

Abb. 3.9: Verteilungsfunktion und Resthäufigkeitskurve (Beispiel 3.4)