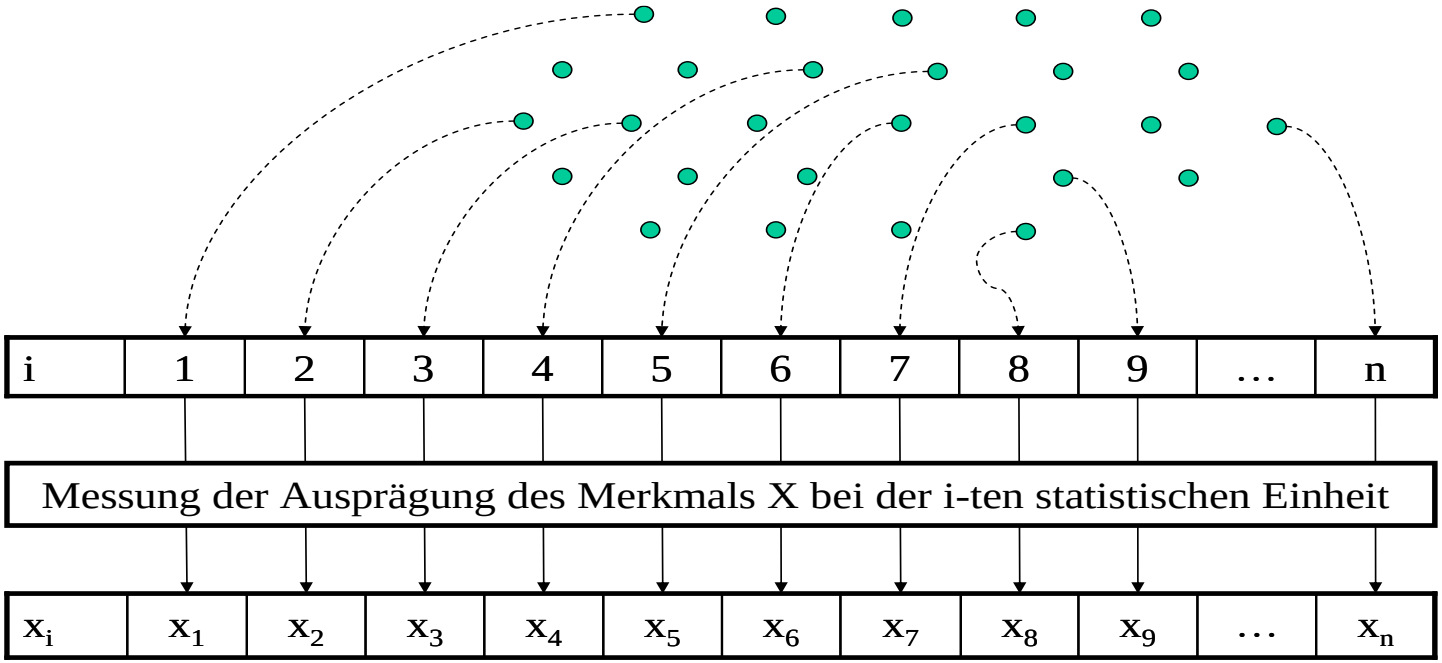
3. Univariate Häufigkeitsverteilungen

3.1 Häufigkeitsverteilungen bei unklassierten Merkmalen

Daten eines diskreten Merkmals X mit nicht zu vielen Ausprägungen werden bei insgesamt n statistischen Einheiten erhoben (i = 1, 2, …, n)

Die Merkmalswerte (Einzelwerte, Beobachtungswerte) x_i sind die bei den statistischen Einheiten gemessenen Merkmalsausprägungen

Merkmalswerte: $x_1, x_2, \ldots, x_n$



Beispiel: Kinderzahl

Zehn Studierende werden nach ihrer Kinderzahl gefragt:

- Die erste befragte Person ist kinderlos, d.h. $x_1 = 0$,
- die zweite Person hat ein Kind, d.h. $x_2 = 1$ usw.
- Insgesamt liegen folgende Beobachtungswerte bzw. Merkmalswerte vor:

$x_1 = 0$	$x_6 = 0$
$x_2 = 1$	$x_7 = 0$
$x_3 = 1$	$x_8 = 1$
$x_4 = 2$	$x_9 = 1$
$x_5 = 0$	$x_{10} = 0$

- Kinderzahl ist ein diskretes Merkmal mit nicht zu vielen Ausprägungen
- Es ist daher ein nicht zu klassierendes Merkmal
- Bei unklassierten Merkmalen werden die Merkmalsausprägungen ausgezählt

Die **Merkmalsausprägungen** werden mit x_j^* (j = 1, 2, …, m) bezeichnet. Es liegen also m Merkmalsausprägungen vor: $x_1^* < x_2^* < \cdots < x_m^*$ (mit m ≤ n).

- Bei nominalskalierten Merkmalen ist die Reihenfolge der Merkmalsausprägungen beliebig
- Ordinalskalierte Merkmale und metrisch skalierte Merkmale werden in aufsteigender Reihenfolge geordnet.

Beispiel (Fortsetzung): **Kinderzahl**

- Aus den Merkmalswerten $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, $x_4 = 2$, $x_5 = 0$, $x_6 = 0$, $x_7 = 0$, $x_8 = 1$, $x_9 = 1$, $x_{10} = 0$ lassen sich die Merkmalsausprägungen ableiten.
- Da das Merkmal „Kinderzahl“ metrisch skaliert ist, werden die Merkmalsausprägungen in aufsteigender Reihenfolge geordnet:
  - Begonnen wird mit der geringsten Kinderzahl, also den kinderlosen Studierenden.
  - Bei dieser Merkmalsausprägung setzen wir j gleich eins: $x_1^* = 0$
  - Anschließend folgt die nächst größere Kinderzahl: $x_2^* = 1$
  - Die größte gemessene Kinderzahl ist gleich zwei: $x_3^* = 2$

Absolute Häufigkeiten

Die Anzahl der statistischen Einheiten, bei denen die j-te Merkmalsausprägung gemessen wird, ist die absolute Häufigkeit n_j .

Summiert man die absoluten Häufigkeiten n_j auf, dann erhält man die Anzahl der statistischen Einheiten, bei denen das Merkmal erhoben wurde:

$n_1 + n_2 + \ldots + n_m = n$ oder $\sum_{j=1}^{m} n_j = n$

Summenzeichen Σ (großes Sigma): „Summe aller n_j von j = 1 bis m“

Beispiel (Fortsetzung): **Kinderzahl**

j	x_j^* (Kinderzahl)	Strichliste	n_j (Anzahl Studierende)
1	$x_1^* = 0$	IIIII	$n_1 = 5$
2	$x_2^* = 1$	IIII	$n_2 = 4$
3	$x_3^* = 2$	I	$n_3 = 1$

Anzahl der statistischen Einheiten durch Summierung der absoluten Häufigkeiten:

$$n = \sum_{j=1}^{3} n_j = n_1 + n_2 + n_3 = 5 + 4 + 1 = 10$$

Relative Häufigkeiten

Die relativen Häufigkeiten $h_j = n_j / n$ (j = 1, 2, …, m) geben den Anteil der statistischen Einheiten an, bei denen die j-te Merkmalsausprägung gemessen wird. Sie lassen sich auch prozentual interpretieren.

Eigenschaften:

- $0 \leq h_j \leq 1$
- $\sum_{j=1}^{m} h_j = h_1 + h_2 + \ldots + h_m = 1$

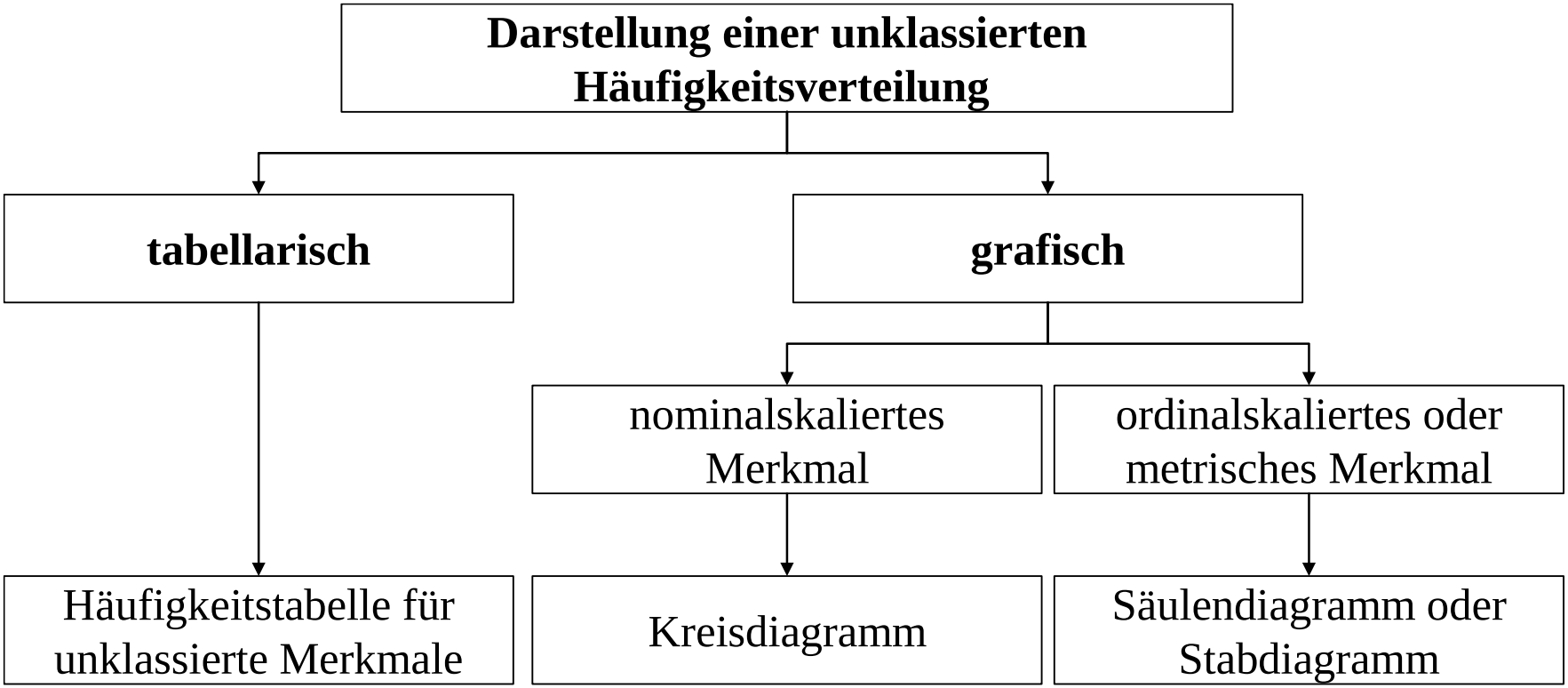
Beispiel (Fortsetzung): **Kinderzahl**

j	x_j^* (Kinderzahl)	n_j (Anzahl Studierende)	h_j (Anteil Studierende)
1	$x_1^* = 0$	$n_1 = 5$	$h_1 = \frac{n_1}{n} = \frac{5}{10} = 0{,}5$
2	$x_2^* = 1$	$n_2 = 4$	$h_2 = \frac{n_2}{n} = \frac{4}{10} = 0{,}4$
3	$x_3^* = 2$	$n_3 = 1$	$h_3 = \frac{n_3}{n} = \frac{1}{10} = 0{,}1$

Es gilt: $\sum_{j=1}^{3} h_j = h_1 + h_2 + h_3 = 0{,}5 + 0{,}4 + 0{,}1 = 1$

Darstellung unklassierter Häufigkeitsverteilungen

Unter einer unklassierten Häufigkeitsverteilung versteht man die Zuordnung von absoluten Häufigkeiten n_j oder relativen Häufigkeiten h_j zu den Merkmalsausprägungen.



Häufigkeitstabelle bei unklassierten Häufigkeiten

j	x_j^*	n_j	h_j
1	x_1^*	n_1	h_1
2	x_2^*	n_2	h_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m^*	n_m	h_m
$\sum$		n	1

Merkmalsausprägungen: x_j^* ($j = 1, 2, \ldots, m$)

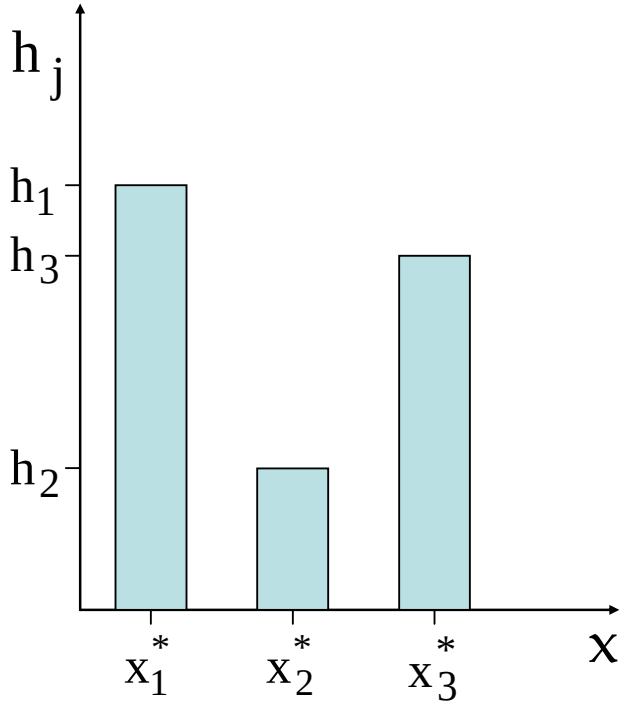
absolute Häufigkeiten: n_j

relative Häufigkeiten: h_j

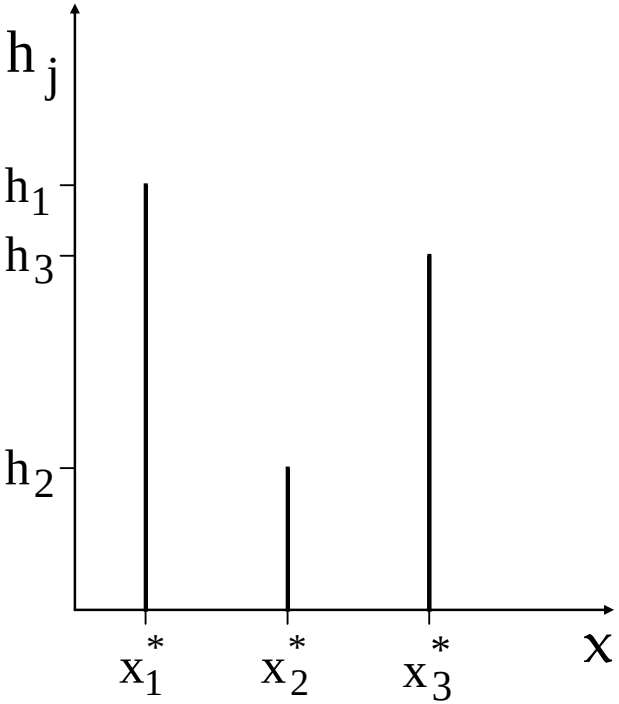
Grafische Darstellung bei unklassierten Häufigkeiten

Ordinal- und metrisch skalierte Merkmale lassen sich mit einem Säulen- oder Stabdiagrammen grafisch darstellen:

- Abszisse (x-Achse): Merkmalsausprägungen
- Ordinate (y-Achse): Relative (absolute) Häufigkeiten



Säulendiagramm



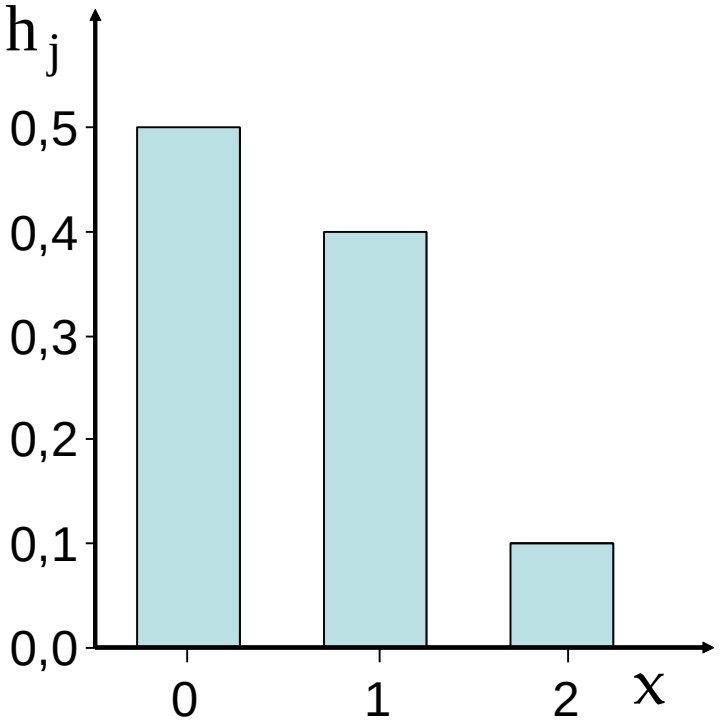
Stabdiagramm

Beispiel (Fortsetzung): **Kinderzahl**

Das Merkmal „Kinderzahl“ ist metrisch skaliert, deshalb ist ein Säulendiagramm eine geeignete grafische Darstellung

Tabellarische Darstellung	Grafische Darstellung

j	x_j^*	n_j	h_j
1	0	5	0,5
2	1	4	0,4
3	2	1	0,1
$\sum$		n=10	1



Prinzip der Flächenproportionalität

Die Häufigkeiten müssen proportional zu den zugehörigen Flächen im Diagramm sein.

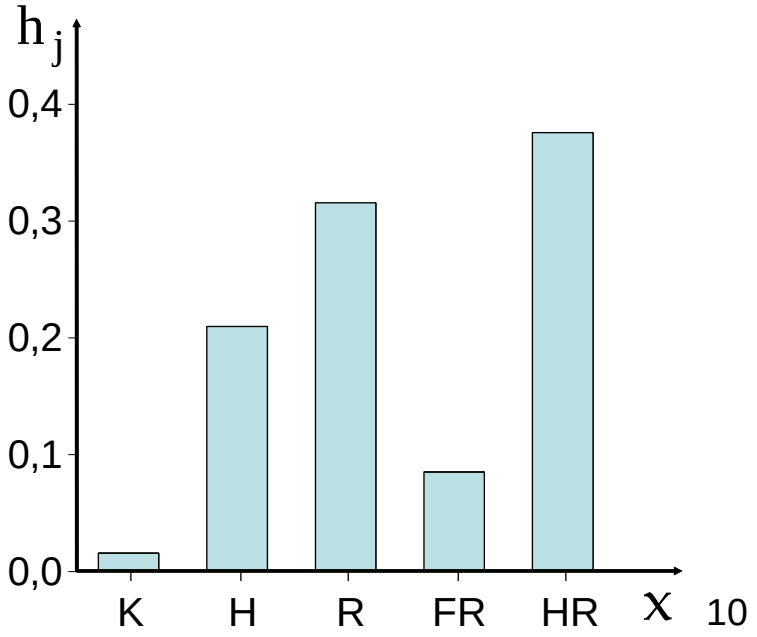
Beispiel: Schulabschluss

Es werden 2.000 Personen nach ihrem höchsten Schulabschluss befragt. 30 haben keinen Schulabschluss (K), 420 einen Hauptschulabschluss (H), 630 einen Realschulabschluss (R), 170 die Fachhochschulreife (FR) und die übrigen Befragten die Hochschulreife (HR). Das Merkmal Schulabschluss soll tabellarisch und grafisch dargestellt werden. Die Anzahl der Personen mit Hochschulreife wird als Differenz zwischen n = 2.000 und der Summe von Personen mit einem anderen Abschluss berechnet: $n_5 = 2.000 – 30 – 420 – 630 – 170 = 750$. Aufgrund des ordinalen Skalenniveaus bietet sich die Verwendung eines Säulendiagramms an.

Tabellarische Darstellung

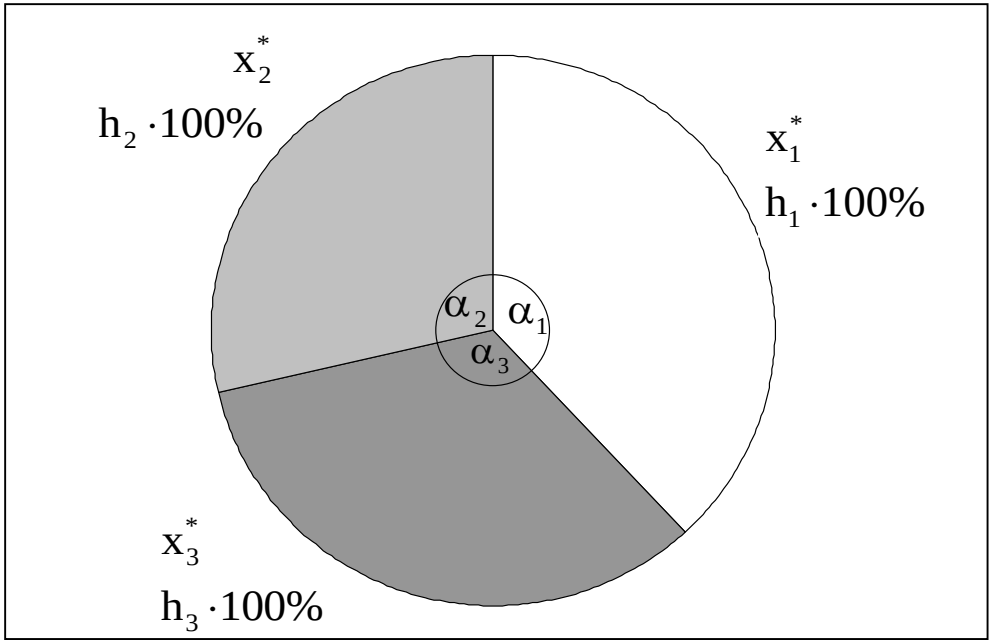
j	x_j^*	n_j	h_j
1	K	30	30/2000 = 0,015
2	H	420	420/2000 = 0,210
3	R	630	630/2000 = 0,315
4	FR	170	170/2000 = 0,085
5	HR	750	750/2000 = 0,375

Grafische Darstellung



Grafische Darstellung bei nominalskalierten Merkmalen

Bei nominalskalierten Merkmalen ist die Reihenfolge der Merkmalsausprägungen beliebig → **Kreisdiagramm** (statt Säulendiagramm → suggeriert Reihenfolge)



Flächenproportionalität:

$n_j : n = F_j : F$

F_j: Flächeninhalt des j-ten Kreissektors

F: Kreisfläche

Flächenproportionalität: Beim Kreisdiagramm soll die Fläche der Kreissektoren proportional zu den Häufigkeiten sein. Entsprechend dieser Forderung ist der Winkel des j-ten Kreissektors zu bestimmen. Man erhält ihn, indem die relative Häufigkeit der j-ten Merkmalsausprägung mit 360 multipliziert wird:

$$\alpha_j = h_j \cdot 360°$$

Kreisdiagramm: Herleitung der Formel

Man setzt die Berechnungsformel für den Flächeninhalt des Kreises

$$F = r^2 \cdot \pi$$

und des j-ten Kreissektors

$$F_j = r^2 \cdot \pi \cdot \frac{\alpha_j}{360°}$$

in die Bedingung für Flächenproportionalität $F_j / F = n_j / n$ ein:

$$\frac{r^2 \cdot \pi \cdot \frac{\alpha_j}{360°}}{r^2 \cdot \pi} = \frac{n_j}{n}$$

Nach Kürzen mit $r^2 \cdot \pi$ erhält man:

$\alpha_j / 360° = n_j / n$

woraus sich unter Verwendung der Definition der relativen Häufigkeit die Beziehung ergibt.

Beispiel: Stellung im Beruf

Im Mikrozensus des Statistischen Bundesamtes wird das Merkmal „Stellung im Beruf“ der Erwerbstätigen erhoben. Aufgrund der Stichprobenergebnisse ergibt sich nach der Hochrechnung folgende Struktur für die Anzahl der Erwerbstätigen (in 1000) in den einzelnen Berufen:

Stellung im Beruf	Selbst-ständige (S)	Mithelfende Familienan-gehörige (MF)	Beamte (B)	Angestell-te (AN)	Arbeiter (AR)	Erwerbs-tätige insges.
Erwerbs-tätige	2422	639	2370	11516	10419	27366

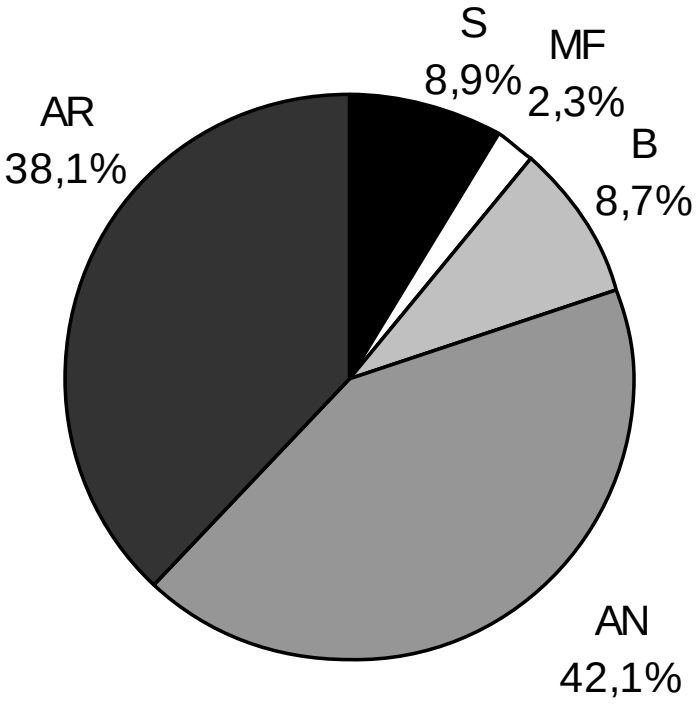
- Das Merkmal „Stellung im Beruf“ wird auf einer Nominalskala gemessen, so dass die Häufigkeitsverteilung adäquat in Form eines Kreisdiagramms dargestellt wird.
- In welcher Reihenfolge die Merkmalsausprägungen in der Tabelle angegeben bzw. die Kreissektoren gezeichnet werden, ist beliebig. In der Häufigkeitstabelle weisen wir zusätzlich die Winkel für die Kreissektoren α_j aus.

Tabellarische Darstellung				Grafische Darstellung

j	x_j^*	n_j	h_j	α_j
1	S	2.422	0,089	32,0°
2	MF	639	0,023	8,3°
3	B	2.370	0,087	31,3°
4	AN	11.516	0,421	151,6°
5	AR	10.419	0,381	137,2°
∑		27.366	1	360°

z.B. $h_1 = n_1 / n = 2422 / 27366 = 0{,}089$

$\alpha_1 = h_1 \cdot 360° = 0{,}089 \cdot 360° = 32{,}0°$



3.2 Häufigkeitsverteilungen bei klassierten Merkmalen

Klassierung (Klassenbildung):

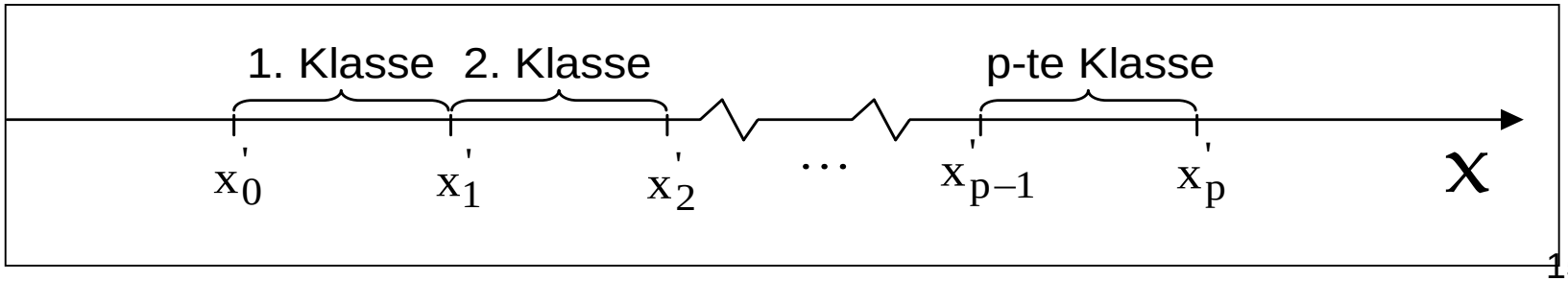
- Bei stetigen und quasi-stetigen Merkmalen
- Bei diskreten Merkmalen mit vielen Merkmalsausprägungen
- Voraussetzung: Ordinal oder metrisch skaliertes Merkmal X

Klassengrenzen

Die Klassengrenzen werden als x'_k bezeichnet. Der Index k läuft von 1 bis p, weil es p Klassen gibt.

Bei der Abgrenzung der Klassen ist zu beachten, dass mit Ausnahme der Randklassen die Obergrenze einer Klasse der Untergrenze der nächsten Klasse entspricht:

x'_k : Obergrenze der k-ten Klasse und Untergrenze der (k+1)ten Klasse.



Sofern die Daten selbst klassiert werden, wird ein Wert, der auf eine Klassengrenze fällt, der unteren Klasse zugeordnet. Wir definieren also **linksseitig offene** Klassen bzw. **rechtseitig geschlossene Klassen**:

- Verbal: „von über ... bis zu …“
- Formal: $(x'_{k-1}; x'_k]$

Absolute Klassenhäufigkeit

Die Anzahl der statistischen Einheiten, die der k-ten Klasse zugeordnet werden, heißt absolute Klassenhäufigkeit n_k

Es gilt: $n_1 + n_2 + \ldots + n_p = \sum_{k=1}^{p} n_k = n$

Relative Klassenhäufigkeit

Die relative Klassenhäufigkeit $h_k = n_k / n$ gibt den Anteil der statistischen Einheiten an, die in der k-ten Klasse liegen.

Es gilt: $h_1 + h_2 + \ldots + h_p = \sum_{k=1}^{p} h_k = 1$

Während die absoluten und relativen Häufigkeiten einer unklassierten Verteilung punktweise definiert sind, bezieht sich die Klassenhäufigkeit stets auf ein Intervall (Klasse).

Beispiel: **Rendite**

Von zehn Aktien wurde die Rendite (Verzinsung) erhoben (in %):
$x_1 = 0{,}6$, $x_2 = 2$, $x_3 = 1$, $x_4 = 2{,}1$, $x_5 = 2{,}4$, $x_6 = 3$, $x_7 = 1{,}6$, $x_8 = 2{,}2$, $x_9 = 3{,}5$, $x_{10} = 1{,}2$

Diese Werte sollen mit den Klassengrenzen 0,5; 1,5; 2; 2,5; 3,5 unter Angabe der absoluten und relativen Klassenhäufigkeiten klassifiziert werden.

k (Klasse)	von über … bis zu …	Strichliste	n_k (Anzahl)	h_k (Anteil)
1	0,5 – 1,5	III	$n_1 = 3$	$h_1 = \frac{3}{10} = 0{,}3$
2	1,5 – 2	II	$n_2 = 2$	$h_2 = \frac{2}{10} = 0{,}2$
3	2 – 2,5	III	$n_3 = 3$	$h_3 = \frac{3}{10} = 0{,}3$
4	2,5 – 3,5	II	$n_4 = 2$	$h_4 = \frac{2}{10} = 0{,}2$

Darstellung klassierter Häufigkeitsverteilungen

Unter einer klassierten Häufigkeitsverteilung versteht man die Zuordnung von relativen Klassenhäufigkeiten h_k oder absoluten Klassenhäufigkeiten n_k zu den Klassen.

Darstellung einer klassierten Häufigkeitsverteilung

tabellarisch	**grafisch**
Häufigkeitstabelle für klassierte Merkmale	Histogramm

Häufigkeitstabelle bei klassierter Häufigkeitsverteilung

k	von über … bis zu …	n_k	h_k	b_k	d_k
1	$x'_0 - x'_1$	n_1	h_1	b_1	d_1
2	$x'_1 - x'_2$	n_2	h_2	b_2	d_2
⋮	⋮	⋮	⋮	⋮	⋮
p	$x'_{p-1} - x'_p$	n_p	h_p	b_p	d_p
∑		n	1		

Klassenbreite: $b_k = x'_k - x'_{k-1}$

Die (relative) Häufigkeitsdichte d_k gibt an, wie „dicht“ die statistischen Einheiten in einer Klasse liegen:

$$d_k = \frac{h_k}{b_k}$$

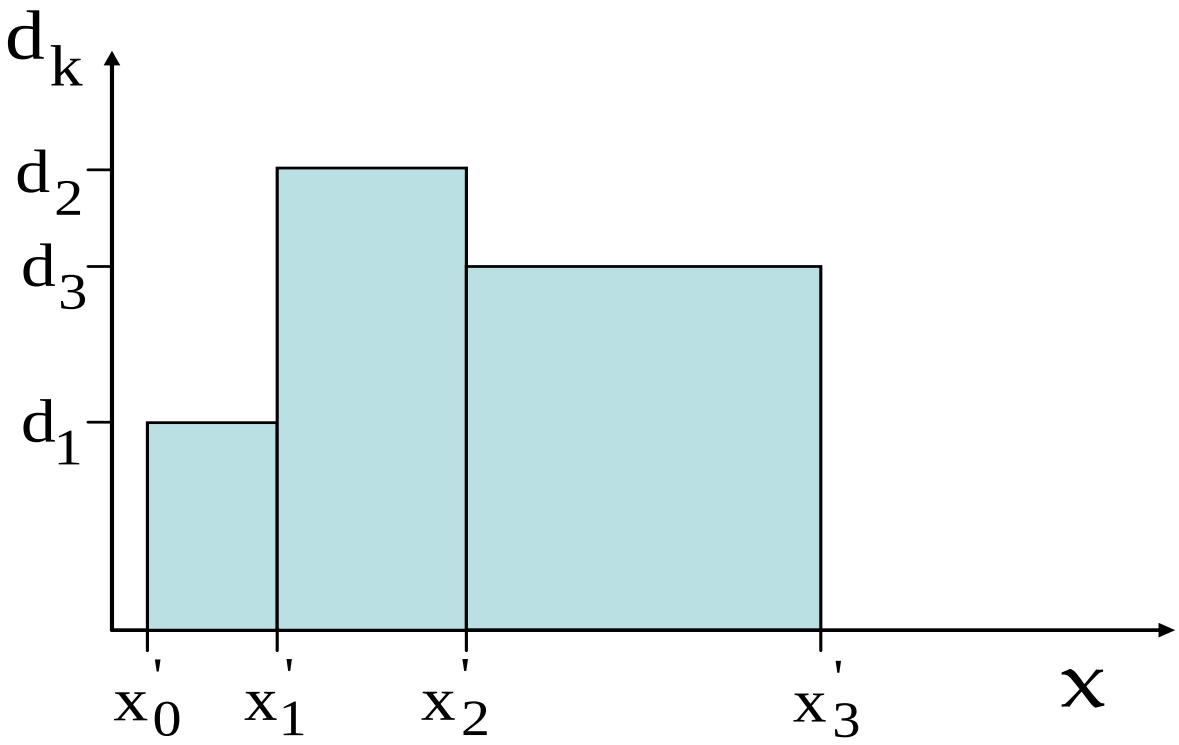
Beispiel (Fortsetzung): Rendite

Für das Merkmal „Rendite“ ergibt sich folgende tabellarische Darstellung:

k	von über… bis zu …	n_k	h_k	b_k	d_k
1	0,5 – 1,5	3	0,3	1,5 – 0,5 = 1	$\frac{h_1}{b_1} = \frac{0,3}{1} = 0,3$
2	1,5 – 2	2	0,2	2 -1,5 = 0,5	$\frac{h_2}{b_2} = \frac{0,2}{0,5} = 0,4$
3	2 – 2,5	3	0,3	2,5 – 2 = 0,5	$\frac{h_3}{b_3} = \frac{0,3}{0,5} = 0,6$
4	2,5 – 3,5	2	0,2	3,5 – 2,5 = 1	$\frac{h_4}{b_4} = \frac{0,2}{1} = 0,2$
∑		n = 10	1		

Histogramm

- Ein Koordinatensystem, in dem die Merkmalswerte auf der Abzisse und die Häufigkeitsdichten auf der Ordinate abgetragen werden, heißt Histogramm.
- Die Flächenproportionalität bei einem Histogramm verlangt, dass die relativen Häufigkeiten h_k proportional zu den Flächeninhalten sind, die sich als Produkt aus der Breite (b_k) und Höhe (d_k) ergeben. Bis auf einen Proportionalitätsfaktor muss daher die Beziehung $h_k = b_k \cdot d_k$ erfüllt sein.



Beispiel (Fortsetzung): Rendite

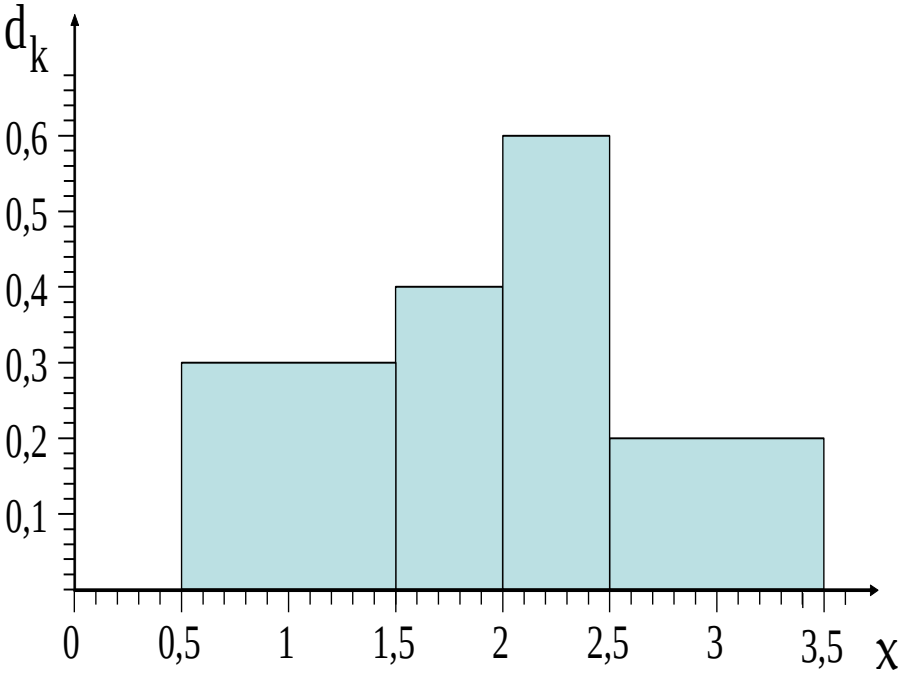
Tabellarische Darstellung

k	von über… bis zu …	n_k	h_k	b_k	d_k
1	0,5 – 1,5	3	0,3	1	0,3
2	1,5 – 2	2	0,2	0,5	0,4
3	2 – 2,5	3	0,3	0,5	0,6
4	2,5 – 3,5	2	0,2	1	0,2
∑		n = 10	1		

z.B. $d_1 = h_1 / b_1 = 0{,}3 / 1 = 0{,}3$

$d_2 = h_2 / b_2 = 0{,}2 / 0{,}5 = 0{,}4$

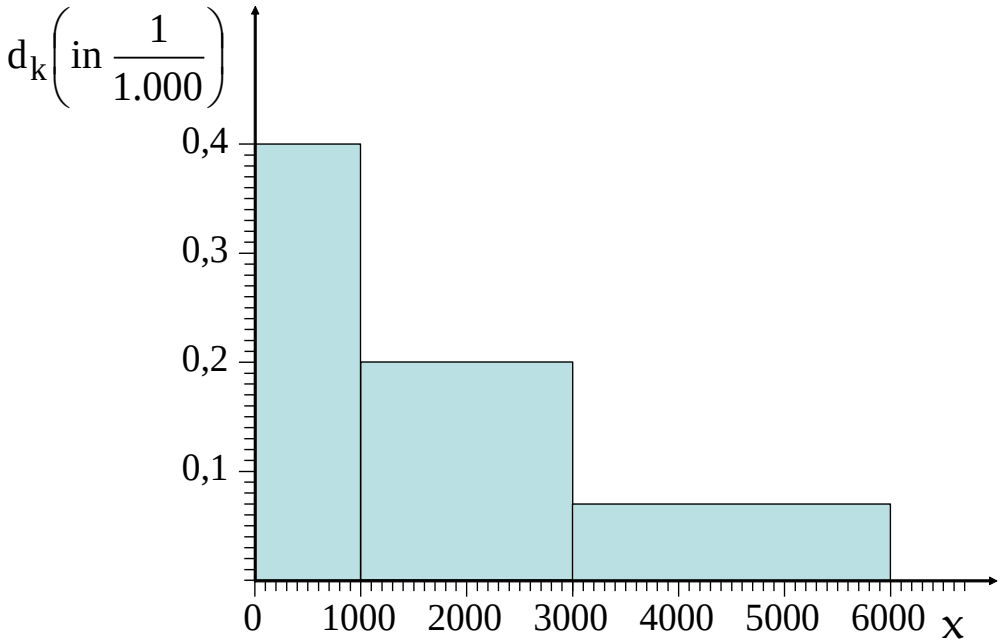
Histogramm



Beispiel: Einkommen

1.000 Personen werden nach ihrem Einkommen befragt. 400 Personen verdienen bis zu 1.000€, 400 von über 1.000€ bis zu 3.000€ und 200 von über 3.000€ bis zu 6.000€. Wie lässt sich (bei einem Proportionalitätsfaktor von 1/1.000) die Einkommensverteilung tabellarisch und grafisch darstellen?

k	von über… bis zu…	n_k	h_k	b_k	d_k	d_k in 1/1000
1	0–1000	400	0,4	1000	0,0004	0,4
2	1000–3000	400	0,4	2000	0,0002	0,2
3	3000–6000	200	0,2	3000	0,00007	0,07
∑		n = 1000	1			



Für Klasse 1:

$$h_1 = \frac{n_1}{n} = \frac{400}{1000} = 0{,}4$$

$$b_1 = x_1' - x_0' = 1000 - 0 = 1000$$

$$d_1\left(\text{in } \frac{1}{1000}\right) = \frac{h_1}{b_1}1000 = \frac{0{,}4}{1000}1000 = 0{,}4$$

Klassierung von Daten

- Sofern die Daten in unklassierter Form erhoben werden, ist zur Darstellung der Häufigkeitsverteilung vor allem bei stetigen und quasi-stetigen Merkmalen eine Klassierung vorzunehmen.
- Die Klassenbildung sollte stets auf sachlichen Erwägungen des jeweiligen Anwendungsbereichs basieren.
- Zusätzlich sind statistische Aspekte zu berücksichtigen:
  - Keine leeren Klassen
  - Zu kleinen Klassen betonen Messfehler
  - Zu große Klassen verdecken wichtige Eigenschaften der Verteilung
  - In der Regel sind 5 bis 20 Klassen ausreichend
- Bei kleinen Beobachtungsumfängen (n<100) gilt die Faustregel: $p \approx \sqrt{n}$

Beispiel: Aktien

Bei 59 Top-Aktien auf dem US-amerikanischen Aktienmarkt sind im Beobachtungszeitraum folgende Werte der Renditen (in %) zu verzeichnen:

3,3	4,4	5,4	5,8	3,0	4,2	6,2	9,0	3,6
2,2	4,2	3,9	7,0	6,9	1,8	4,2	3,2	1,6
5,4	5,1	6,9	4,4	3,4	3,2	3,9	6,0	5,4
2,7	5,7	3,0	1,5	3,1	0,1	4,9	8,3	1,9
3,0	4,1	5,1	2,6	1,9	4,1	0,4	3,9	4,3
2,4	2,7	4,9	3,8	2,6	0,2	8,4	1,9	8,7
3,7	3,2	3,2	4,5	6,7				

$\sqrt{n}$ Faustregel: Klassierung der Daten mit 7 oder 8 Klassen ($\sqrt{59}$ =7,681)

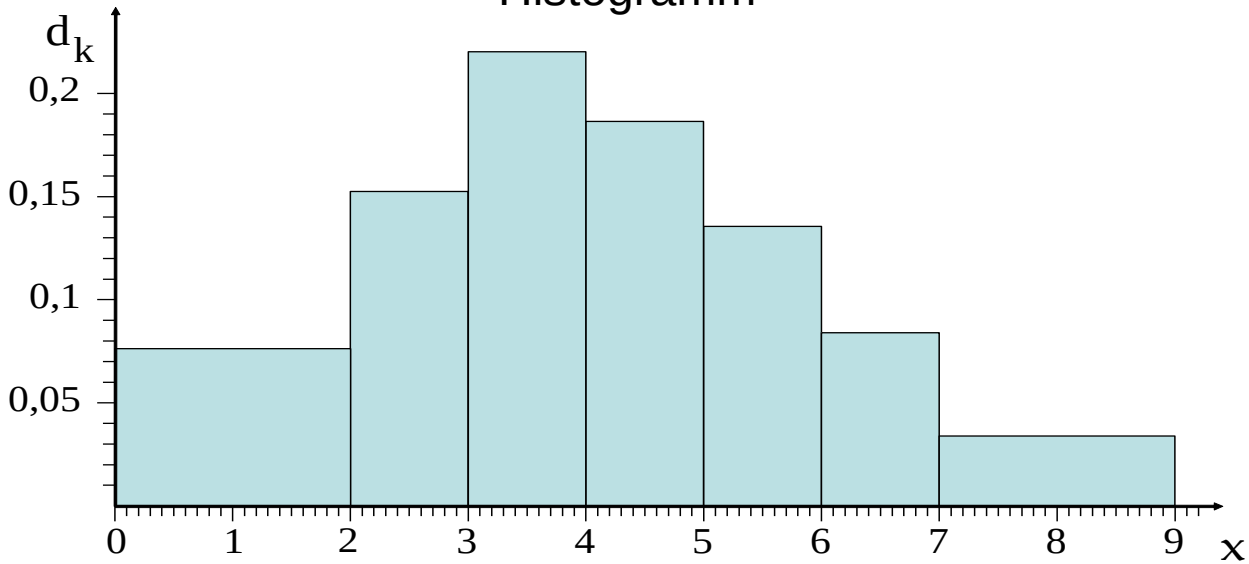
Verteilung der Rendite

```
                          xxxx      xx
   x x x x      xxxx   xxxx     xxxx    xxxx      xx
  x x x x x   xxxxx  xxxxx   xxxxx   xxxx     xxx      x x x x
|---------------|---------|---------|---------|---------|---------|--------------|--------->
0               2         3         4         5         6         7              9         x
```

Äußere Daten: Dünne Besetzung (Zwei-Prozentpunkt-Klassen)
Mittlere Daten: Dichtere Besetzung (Ein-Prozentpunkt-Klassen)

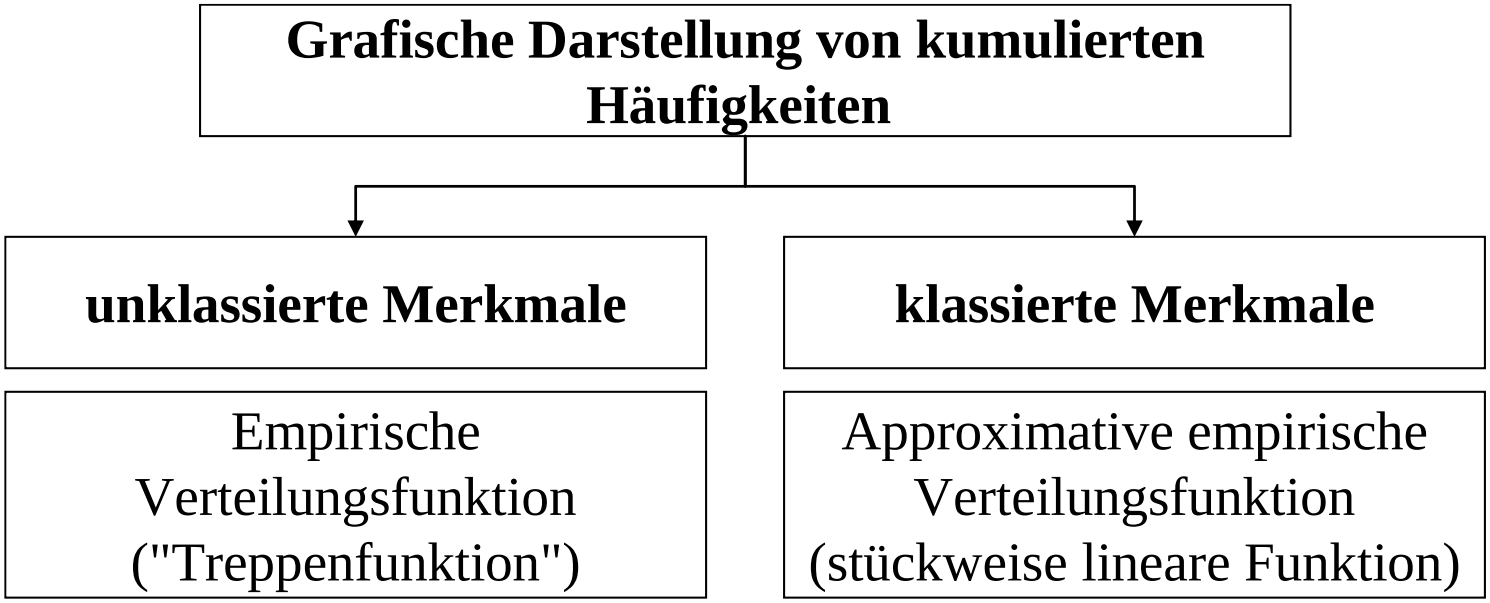
k	von über… bis zu…	n_k	h_k	b_k	d_k
1	0 - 2	9	0,153	2	0,077
2	2 – 3	9	0,153	1	0,153
3	3 – 4	13	0,220	1	0,220
4	4 – 5	11	0,186	1	0,186
5	5 – 6	8	0,136	1	0,136
6	6 – 7	5	0,085	1	0,085
7	7 – 9	4	0,068	2	0,034
∑		n = 59	1		

Histogramm



3.3 Empirische Verteilungsfunktion

- Empirische Verteilungsfunktion: Summenhäufigkeitsfunktion
- Voraussetzung: Ordinalskalierte oder metrisch skalierte Merkmale (Kumulierung von Häufigkeiten zulässig)



Empirische Verteilungsfunktion bei unklassierter Häufigkeitsverteilung

Die **absolute kumulierte Häufigkeit** N_j gibt an, wie viele statistische Einheiten sich bis zur j-ten Merkmalsausprägung kumuliert (= aufsummiert) haben.

- Absolute kumulierte Häufigkeiten (Aufaddieren der absoluten Häufigkeiten):
  $N_j = n_1 + n_2 + \ldots + n_j = N_{j-1} + n_j$
- Insgesamt liegen m Merkmalsausprägungen vor. Somit entspricht N_m der Gesamtzahl der statistischen Einheiten n:
  $N_m = n_1 + n_2 + \ldots + n_m = n$

Die **relative kumulierte Häufigkeit** H_j weist aus, welcher Anteil der statistischen Einheiten sich bis zur j-ten Merkmalsausprägung kumuliert (= aufsummiert) hat.

- Relative kumulierte Häufigkeiten (Aufaddieren der relativen Häufigkeiten):
  $H_j = h_1 + h_2 + \ldots + h_j = H_{j-1} + h_j$
- Insgesamt liegen m Merkmalsausprägungen vor. Somit ist H_m bei der m-ten Ausprägung gleich eins:
  $H_m = h_1 + h_2 + \ldots + h_m = 1$

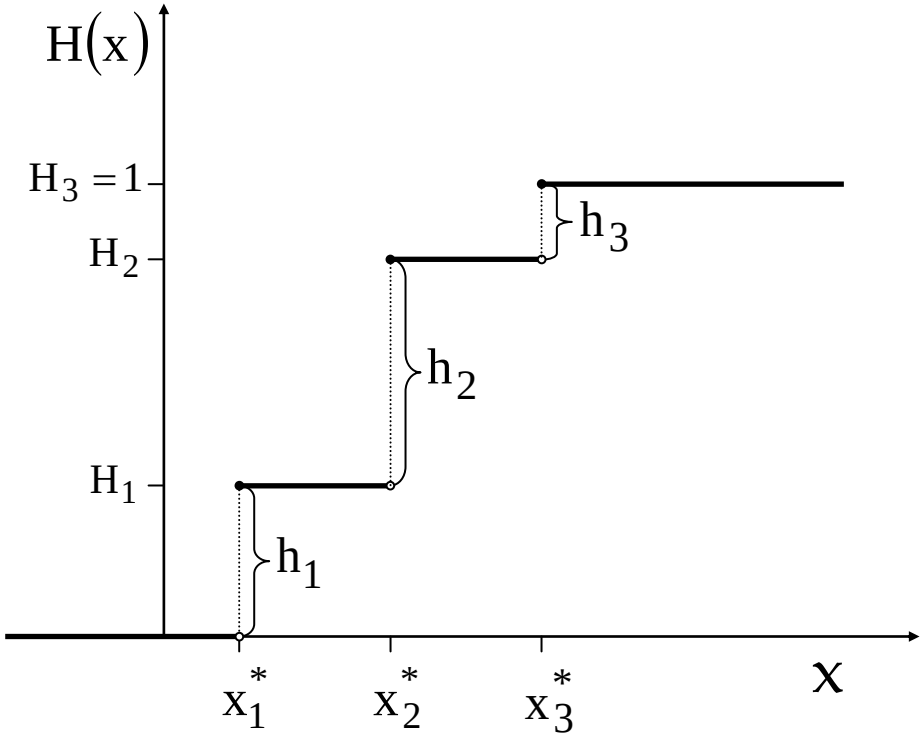
Beispiel (Fortsetzung): Kinderzahl

Für das Merkmal „Kinderzahl“ werden die relativen und absoluten kumulierten Häufigkeiten berechnet:

j	x_j^*	n_j	h_j	N_j	H_j
1	0	5	0,5	5	0,5
2	1	4	0,4	5 + 4 = 9	0,5 + 0,4 = 0,9
3	2	1	0,1	9 + 1 = 10	0,9 + 0,1 = 1
$\sum$		n = 10	1		

Definition der empirische Verteilungsfunktion

- Die empirische Verteilungsfunktion wird zur Darstellung der kumulierten Häufigkeiten verwendet. Im Unterschied zu den kumulierten Häufigkeiten ist sie nicht nur an den Stellen $x_1^*, x_2^*, \ldots, x_m^*$ sondern für alle reellen Zahlen x definiert.
- Die empirische Verteilungsfunktion gibt für jedes x den Anteil der statistischen Einheiten an, der sich bis zu x kumuliert (= aufsummiert) hat, d.h. die einen Merkmalswert haben, der kleiner oder gleich x ist.

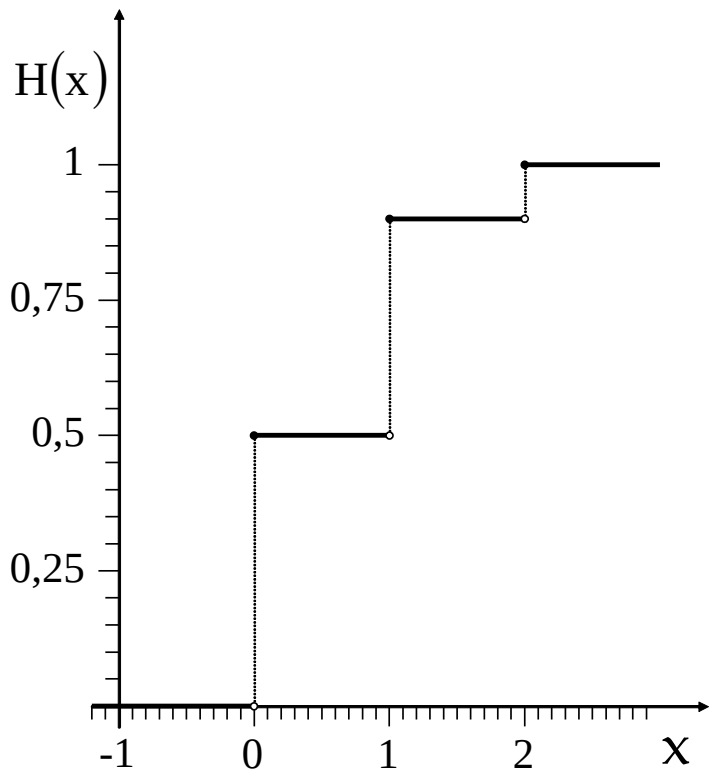


$$H(x) = \begin{cases} 0 \text{ für } x < x_1^* \\ H_j \text{ für } x_j^* \le x < x_{j+1}^* \text{ mit } j = 1,2,\ldots,m-1 \\ 1 \text{ für } x \ge x_m^* \end{cases}$$

Beispiel (Fortsetzung): Kinderzahl

Tabellarische Darstellung	Grafische Darstellung

j	x_j^*	h_j	H_j
1	0	0,5	0,5
2	1	0,4	0,9
3	2	0,1	1



Empirische Verteilungsfunktion bei klassierter Häufigkeitsverteilung

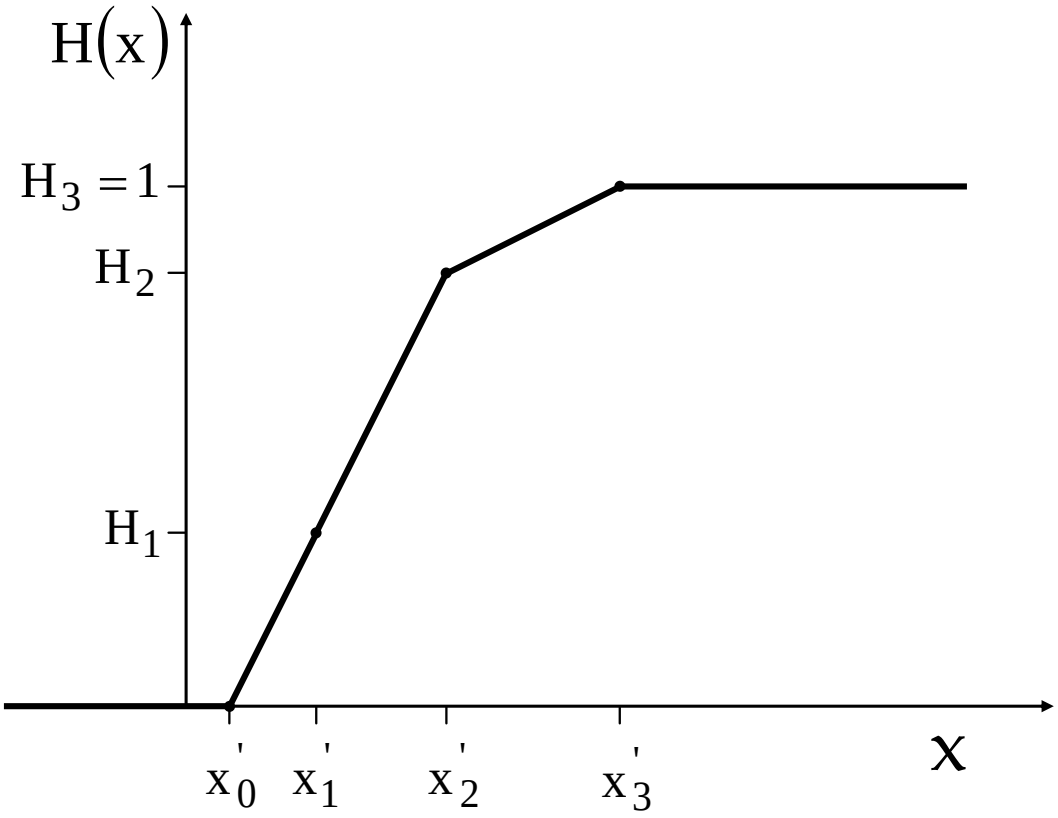
Die **absolute kumulierte Klassenhäufigkeit** N_k gibt an, wie viele statistische Einheiten sich bis zum Ende der k-ten Klasse kumuliert (= aufsummiert) haben.

- Absolute kumulierte Häufigkeiten (Aufaddieren der absoluten Häufigkeiten):
  $N_k = n_1 + n_2 + \ldots + n_k = N_{k-1} + n_k$
- Insgesamt liegen p Klassen vor. Somit entspricht N_p der Gesamtzahl der statistischen Einheiten n:
  $N_p = n_1 + n_2 + \ldots + n_p = n$

Die **relative kumulierte Klassenhäufigkeit** H_k weist aus, welcher Anteil der statistischen Einheiten sich bis zum Ende der k-ten Klasse kumuliert hat.

- Relative kumulierte Klassenhäufigkeiten (Aufaddieren der relativen Klassenhäufigkeiten):
  $H_k = h_1 + h_2 + \ldots + h_k = H_{k-1} + h_k$ bzw. $H_k = \frac{N_k}{n}$
- Insgesamt liegen p Klassen vor. Somit ist H_p bei der p-ten Klasse gleich eins:
  $H_p = h_1 + h_2 + \ldots + h_p = 1$

Approximative empirische Verteilungsfunktion

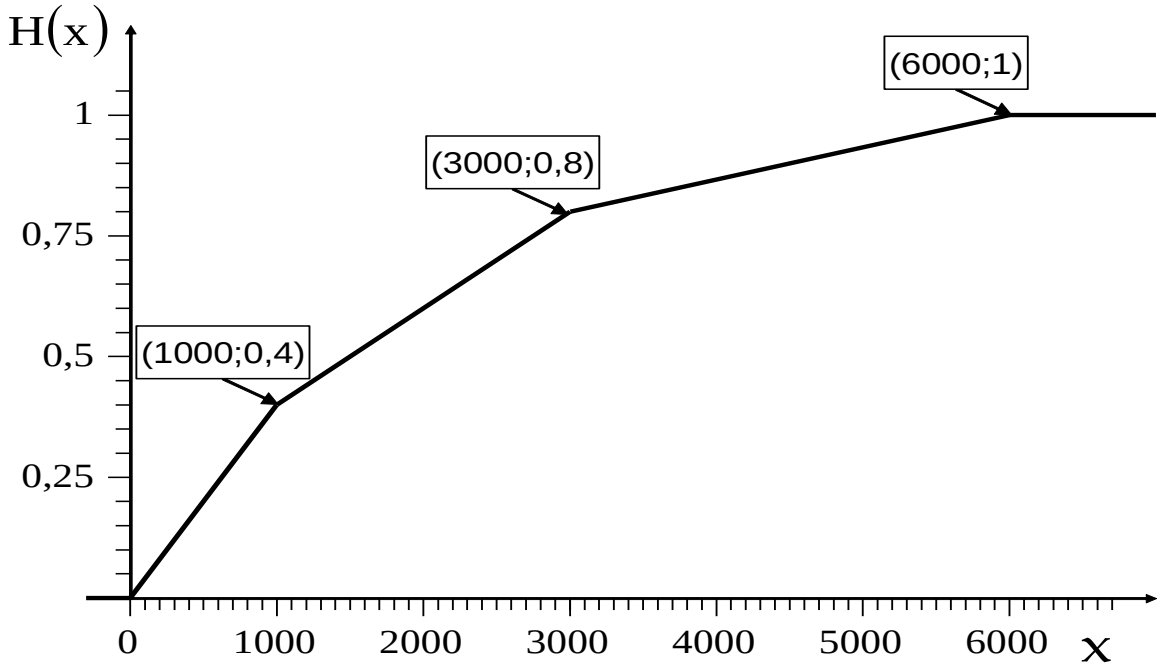


Annahme: Gleichverteilung der Merkmalswerte innerhalb der Klassen

Beispiel (Fortsetzung): Einkommen

Für das Merkmal „Einkommen“ erhalten wir folgende absoluten und relativen kumulierten Häufigkeiten

k	von über…bis zu …	n_k	h_k	N_k	H_k
1	0–1000	400	0,4	400	0,4
2	1000–3000	400	0,4	400+400=800	0,4+0,4=0,8
3	3000–6000	200	0,2	800+200=1000	0,8+0,2=1,0
∑		n=1000	1		

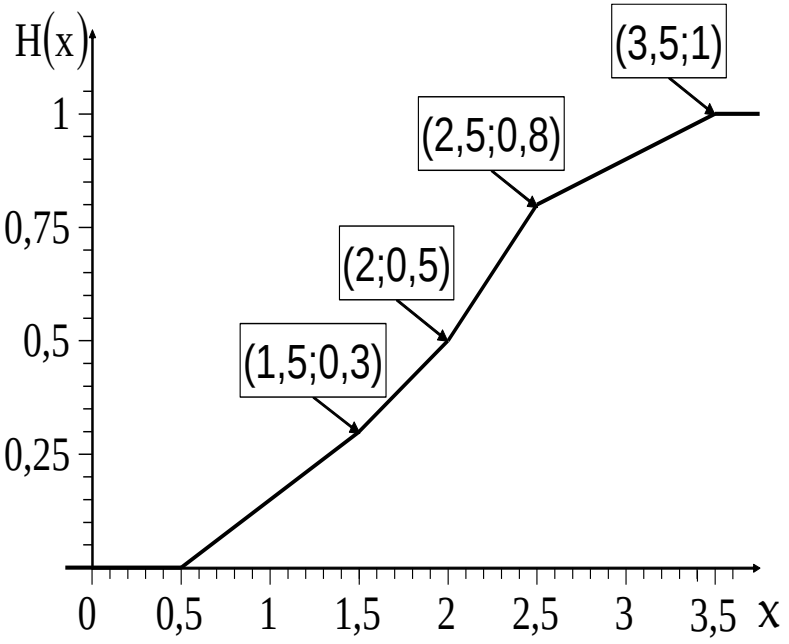


Beispiel (Fortsetzung): Rendite

Die kumulierten Klassenhäufigkeiten der Rendite werden ebenfalls mit einer approximativen empirischen Verteilungsfunktion grafisch dargestellt:

Kumulierte relative Häufigkeiten	Grafische Darstellung

k	von über bis…zu	h_k	H_k
1	0,5 – 1,5	0,3	0,3
2	1,5 – 2	0,2	0,5
3	2 – 2,5	0,3	0,8
4	2,5 – 3,5	0,2	1



Definition der approximativen empirischen Verteilungsfunktion

$$H(x) = \begin{cases} 0 \text{ für } x \le x'_0 \\ H_{k-1} + \dfrac{x - x'_{k-1}}{b_k} \cdot h_k \quad \text{für } x'_{k-1} < x \le x'_k \text{ mit } k = 1,2,\ldots,p \\ 1 \text{ für } x > x'_p \end{cases}$$

Proportionalitätsfaktor: $\dfrac{x - x'_{k-1}}{b_k}$

Mit der approximativen empirischen Verteilungsfunktion kann für x-Werte innerhalb der Klassen der Anteil der statistischen Einheiten geschätzt werden, deren Merkmalswerte kleiner oder gleich x sind.

Beispiel (Fortsetzung): Einkommen

Wie groß ist der geschätzte Anteil von Befragten mit einem Einkommen von bis zu 500€, von bis zu 4000€ und zwischen 500€ und 4000€?

k	von über… bis zu …	n_k	h_k	H_k
1	0 – 1000	400	0,4	0,4
2	1000 – 3000	400	0,4	0,4 + 0,4 = 0,8
3	3000 – 6000	200	0,2	0,8 + 0,2 = 1,0
∑		n = 1000	1	

x = 500 fällt in die erste Klasse (k = 1):

$$H(500) = H_0 + \frac{500 - x_0'}{b_1} \cdot h_1 = 0 + \frac{500 - 0}{1000} \cdot 0{,}4 = 0{,}2 \left[\hat{=}\, 20\,\%\right]$$

x = 4000 fällt in die dritte Klasse (k = 3):

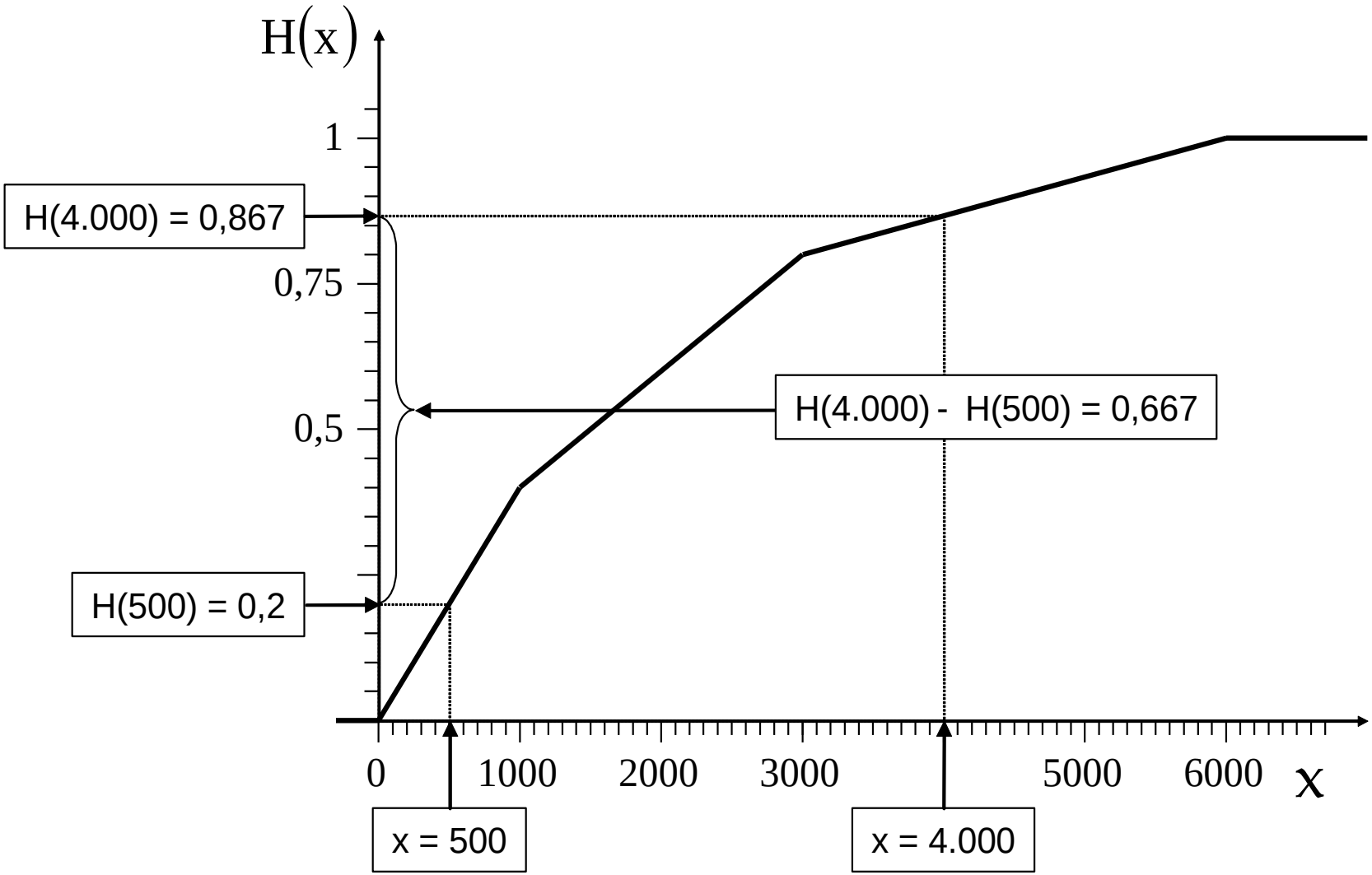
$$H(4000) = H_2 + \frac{4000 - x_2'}{b_3} \cdot h_3 = 0{,}8 + \frac{4000 - 3000}{3000} \cdot 0{,}2$$

$$= 0{,}867 \left[\hat{=}\, 86{,}7\,\%\right]$$

Zwischen 500 und 4000:

$$H(4000) - H(500) = 0{,}867 - 0{,}2 = 0{,}667 \left[\hat{=}\, 66{,}7\,\%\right]$$

Grafische Veranschaulichung der berechneten Anteile



Resthäufigkeitsfunktion

Mit der Resthäufigkeitsfunktion R(x) lässt sich für x-Werte innerhalb der Klassen der Anteil der statistischen Einheiten schätzen, deren Merkmalswerte größer als x sind.

- Die Resthäufigkeitsfunktion berechnet sich durch:
  R(x) = 1 – H (x)

- **Beispiel:**
  Der geschätzte Anteil der Befragten mit einem Einkommen von mehr als 4000€ beträgt:

$$R(4000) = 1 - H(4000) = 1 - 0{,}867 = 0{,}133 \left[\hat{=} 13{,}3\,\%\right]$$