2 Häufigkeitsverteilungen

Ziel: Darstellung bzw. Beschreibung (Exploration) $einer$ Variablen.

Ausgangssituation

An n Einheiten $\omega_1, \ldots, \omega_n$ sei das Merkmal X beobachtet worden.

$\Rightarrow x_1 = X(\omega_1), \ldots, x_n = X(\omega_n)$

Also $x_i = X(\omega_i)$, d.h. x_i ist der Wert der i-ten Einheit

Bezeichnungen:

- $x_1, \ldots, x_n$: Urliste oder Rohdaten
- n : Stichprobenumfang
- Die verschiedenen Merkmalsausprägungen werden mit $a_1, \ldots, a_k$ bezeichnet.

Bemerkungen:

- Werden mehr Beobachtungen erhoben, so ändert sich n, aber i.A. k nicht.

- Meist bezeichnet $a_1, \ldots, a_k$ die beobachteten verschiedenen Merkmalsausprägungen, manchmal aber auch die prinzipiell möglichen Merkmalsausprägungen.

- Für mindestens ordinalskalierte Merkmale seien die Ausprägungen geordnet, d.h.

$$a_1 < a_2 < \ldots < a_k.$$

Beispiel: Häufigkeitsverteilung der Schichtzugehörigkeit einer Gesamtheit Ω von acht Personen $\Omega = \{\omega_1, \ldots, \omega_8\}$.

Tabelle:

2.1 Häufigkeiten

Absolute Häufigkeiten der Merkmalsausprägungen:

Für jedes a_j, $j = 1, \ldots, k$, bezeichnen h_j und $h(a_j)$ die absolute Häufigkeit der Ausprägung a_j, d.h. die Anzahl der x_i aus $x_1, \ldots, x_n$ mit $x_i = a_j$.

Formal:

$$h_j := h(a_j) := |\{\omega \in \Omega \mid X(\omega) = a_j\}|.$$

$|M|$ bezeichnet die Mächtigkeit der Menge M

$:=$ bedeutet „wird definiert als“.

$h_1, h_2, \ldots, h_k$ (als Ganzes) nennt man die absolute Häufigkeitsverteilung.

Es gilt

$$\sum_{j=1}^{k} h_j = n.$$

Erste Darstellung von Häufigkeiten anhand einer Strichliste

Relative Häufigkeiten der Merkmalsausprägungen:

Für jedes a_j, $j = 1, \ldots, k$, bezeichnen f_j und $f(a_j)$ die relative Häufigkeit der Ausprägung a_j, also

$$f_j := f(a_j) := \frac{h_j}{n}.$$

$f_1, f_2, \ldots, f_k$ nennt man die relative Häufigkeitsverteilung.

Es gilt

$$\sum_{j=1}^{k} f_j = 1.$$

Häufigkeitstabelle:

Allgemeine Form:

j	a_j	h_j	f_j
1	a_1	h_1	f_1
2	a_2	h_2	f_2
3	a_3	h_3	f_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	a_k	h_k	f_k
$\sum$		n	1

Im Beispiel:

Gruppierte Häufigkeitsverteilungen

- Insbesondere bei stetigen oder quasi-stetigen Merkmalen ist es häufig zweckmäßig, die Merkmalsausprägungen zu klassieren / zu gruppieren.

  $\Rightarrow$ gruppierte Häufigkeitsverteilung.

- Die gruppierte Häufigkeitsverteilung enthält nur die Häufigkeiten der Ausprägungen in den einzelnen Gruppen, die einzelnen a_i entsprechen in diesem Fall Intervallen.

- Festlegung der Klassen

  – nicht überlappend

  – Überdeckung aller Datenwerte

- Achtung: Die Gruppierung bedeutet einen Informationsverlust

Beispiel Mietspiegel: Merkmal = Nettomieten

Urliste für n=26 Wohnungen, bereits der Größe nach geordnet:

127	172	194	217	226	228	238	248	272	337	347	349	349
373	375	378	383	394	426	443	466	467	533	539	560	676

Klasse j	h_j	f_j
$100 < \ldots \leq 200$		
$200 < \ldots \leq 300$		
$300 < \ldots \leq 400$		
$400 < \ldots \leq 500$		
$500 < \ldots \leq 600$		
$600 < \ldots \leq 700$		
$\sum$		

2.2 Grafische Darstellung

Stabdiagramm: Trage über $a_1, \ldots, a_k$ jeweils einen zur x-Achse senkrecht stehenden Stab mit Höhe $h_1, \ldots, h_k$ (oder $f_1, \ldots, f_k$) ab.

Horizontal: Ausprägungen der Variablen $(a_1, a_2, \ldots, a_k)$

Vertikal: absolute / relative Häufigkeiten ($h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$)

Vorausgesetztes Skalenniveau: mindestens Nominalskala

Säulendiagramm: Ersetze die Stäbe durch Rechtecke (Säulen) gleicher Breite.

Balkendiagramm: Säulendiagramm mit vertauschten Achsen

Streifendiagramm: „gestapeltes“ Säulendiagramm

Kreisdiagramm („Tortendiagramm“)

- Der Kreis wird in Sektoren unterteilt, denen jeweils eine Ausprägung (oder Klasse) zugeordnet wird. Der jeweilige Winkel ist proportional zur Häufigkeit.

- Damit ist die **Fläche** proportional zur Häufigkeit: Prinzip der **Flächentreue**.

- Die Länge des Kreisbogens ist proportional zum Winkel und damit auch zur Häufigkeit.

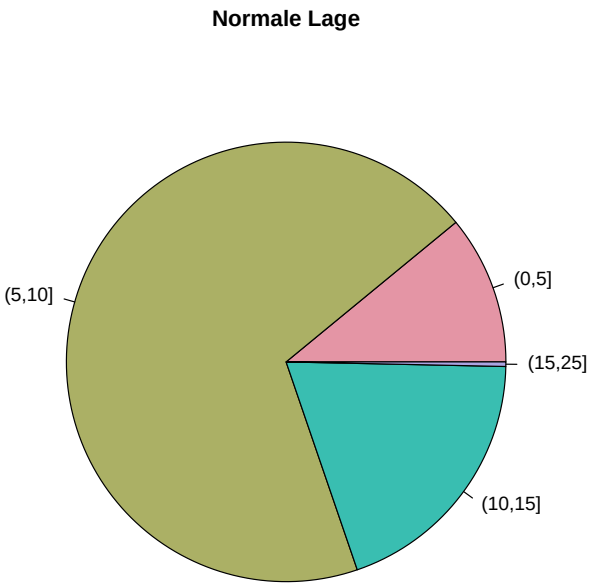
  Achtung: Für Stab-, Säulen- und Balkendiagramm gilt das Prinzip der Längentreue, d.h. die **Länge** der Stäbe / Säulen / Balken ist proportional zur Häufigkeit.

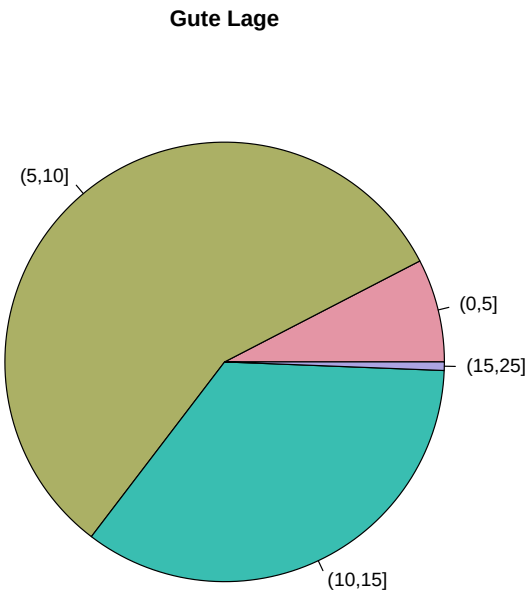
Berechnung

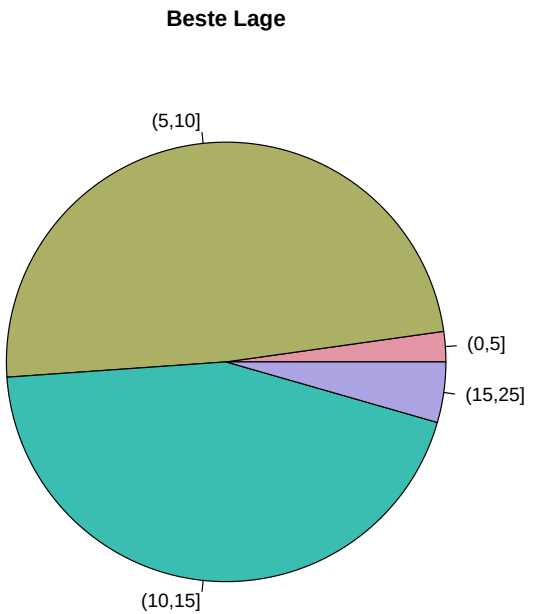
Winkel des Kreissektors j = relative Häufigkeit $\times 360°$

Häufigkeit	Winkel
$f_1 = \frac{1}{8}$	$\frac{360°}{8} = 45°$
$f_2 =$	
$f_3 =$	

Beispiel: Münchner Mietspiegel







Bemerkungen:

- Stab- / Balken- / Säulen- / Streifendiagramme können auch gut zum zeitlichen Vergleich von Häufigkeitsverteilungen eingesetzt werden.
- Für ordinalskalierte Merkmale lässt sich mit Stab- / Balken- / Säulen- / Streifendiagrammen auch die Ordnung der Kategorien darstellen.
- Tortendiagramme in $3D-$ Abbildungen vermitteln oft verzerrten Eindruck.
- Alle bisherigen Grafiken sind nur sinnvoll für kleine Kategorienzahlen k.
- Klasseneinteilungen bei großem k sind oft problematisch
  - Viele Klassen: Diagramm wird unübersichtlich
  - Wenige Klassen: starker Informationsverlust und eventuell starke Abhängigkeit von der konkreten Wahl der Klassen.

Stamm-Blatt-Diagramm

Semigrafisches Verfahren in Analogie zu Strichlisten, gut geeignet für mittelgroßes k.

Erklärung anhand des Mietspiegelbeispiels; Nettomieten:

127	172	194	217	226	228	238	248	272	337	347	349	349
373	375	378	383	394	426	443	466	467	533	539	560	676

Grundidee:

1. Gib einen groben Eindruck von dem Bereich, in dem Ausprägungen liegen (Stamm)

2. Veranschauliche Häufigkeiten in Klassen und bewahre zugleich Wissen über die detaillierte Lage der Ausprägungen (von jedem Punkt auf dem Stamm abzweigende Blätter)

Stamm: führende Ziffern

Blatt: nächste Ziffer (evtl. gerundet)

Vorgehen:

1. Suche den kleinsten und größten Wert der Urliste und zerlege den Wertebereich in Intervalle der Breite 10^q (d.h. die Intervallgrenzen sind ganzzahlige Vielfache von 10 oder 100 oder 1000 ...), wobei q geeignet zu wählen ist (Nettomieten: $q = 2$)

2. Runde die Daten auf die führenden q Stellen.

130	170	190	220	230	230	240	250	270	340	350	350	350
370	380	380	380	390	430	440	470	470	530	540	560	680

3. Bestimme den Stamm aus den jeweils führenden Ziffern

4. Bestimme die Blätter aus der jeweils zweiten Ziffer:

3	7	9	2	3	3	4	5	7	4	5	5	5
7	8	8	8	9	3	4	7	7	3	4	6	8

5. Trage für jeden Wert des Stamms die zugehörigen Blätter rechts von einer vertikalen Linie der Größe nach geordnet ab

Stamm-Blatt-Diagramm für den Mietspiegel:

127	172	194	217	226	228	238	248	272	337	347	349	349
373	375	378	383	394	426	443	466	467	533	539	560	676

Vorteile:

- Einfache Gruppierung ohne viel Informationsverlust: die Darstellung enthält bis auf Rundungen alle Werte der Urliste.
- Ermöglicht guten Einblick in die Datenstruktur für explorative Analysen, z.B. auch Erkennen von Ausreißern.

Nachteile:

- Wird bei großen Datensätzen schnell unübersichtlich.
- Lässt sich oft nicht mehr gut auf Papier präsentieren.

2.3 Histogramm

Zur Motivation: Die Häufigkeit einer Klasse und damit die Höhe im Säulendiagramm hängt stark von der Klassenbreite ab. Durch Zusammenfassen von Klassen erhält man ein ganz anderes Bild.

Nummer	Klasse	Anzahl
1	0 - 10	10
2	10 - 20	20
3	20 - 30	30
4	30 - 50	20
5	50 - 100	20

Ziel: Sinnvolle Häufigkeitsdarstellung für metrische Merkmale (ohne manuelle Kategorisierung).

- Gegeben: Urliste $x_1, \ldots, x_n$ eines (mindestens) intervallskalierten Merkmals.
- Wähle $c_0 \leq \min_{i=1,\ldots,n}(x_i)$ und $c_k \geq \max_{i=1,\ldots,n}(x_i)$
- Bilde Klasseneinteilung $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k]$.
- Für jede Klasse $[c_{j-1}, c_j)$, $j = 1, \ldots, k$ sei

$$d_j = c_j - c_{j-1}$$

die Breite des j-ten Intervalls und h_j bzw. f_j die absolute bzw. relative Häufigkeit in der j-ten Klasse.

- Zeichne über jedem Intervall ein Rechteck der Breite d_j so, dass die **Fläche** proportional zu f_j und h_j ist.

Achtung: Das Histogramm ist **flächentreu**, nicht längentreu!

Es gilt

$$\text{Fläche} = \text{Breite} \cdot \text{Höhe}$$

und damit

$$\text{Höhe} = \text{Fläche} \,/\, \text{Breite}.$$

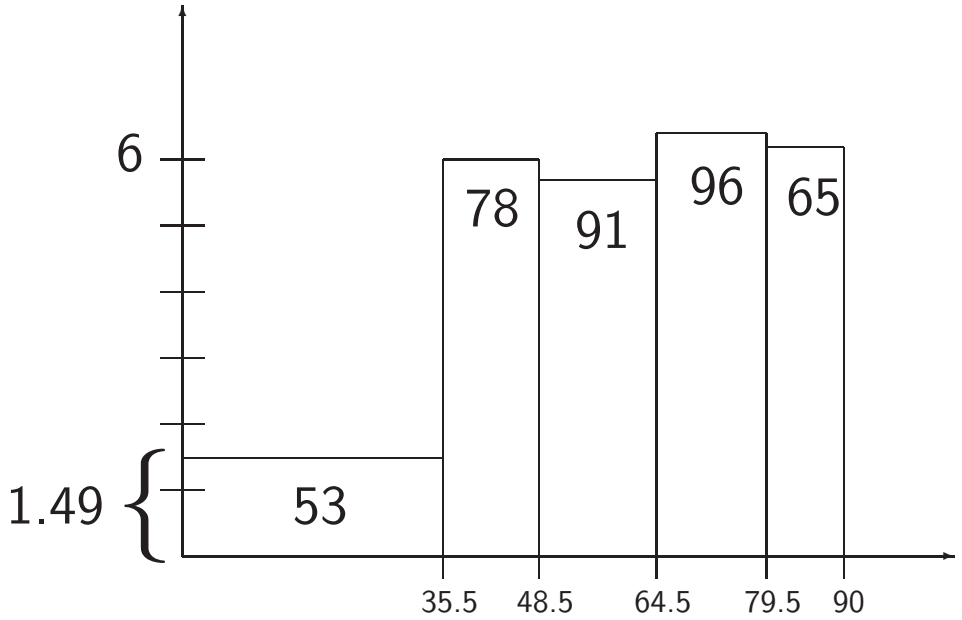
Also ist die Höhe der Rechtecke proportional zu

$$\frac{f_j}{d_j} \quad \text{bzw.} \quad \frac{h_j}{d_j}$$

(und nicht zu f_j bzw. h_j.)

Beispiel: Punkteverteilung in der Klausur

Klassen	h_j Hfgkt.	d_j Breite	h_j/d_j Höhe
$[0, 35.5)$	53	35.5	1.49
$[35.5, 48.5)$	78		
$[48.5, 64.5)$	91		
$[64.5, 79.5)$	96		
$[79.5, 90]$	65		
	383		

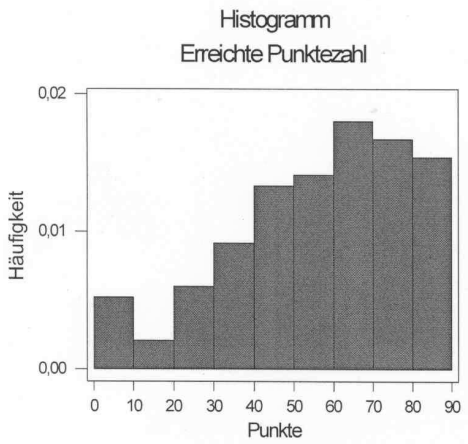

6
1.49
53
78
91
96
65
35.5
48.5
64.5
79.5
90

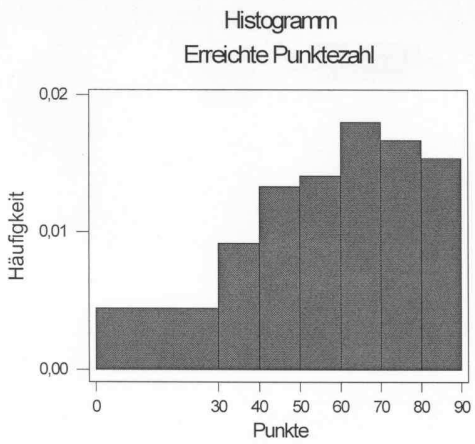
Vorteile des Histogramms:

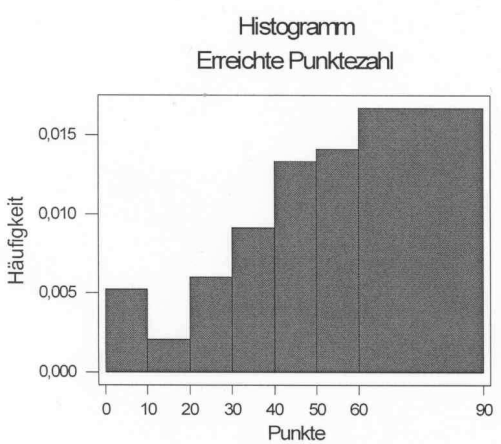
- Information der metrischen Skala (Differenzen) voll ausgenutzt
- etwas weniger empfindlich gegenüber Klasseneinteilung, da sich Häufigkeiten in der Fläche widerspiegeln

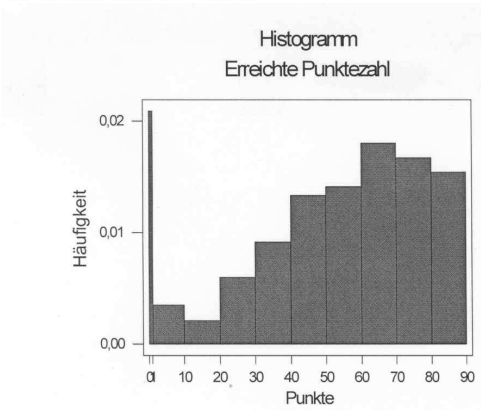
Klasseneinteilung:

- Faustregel: $k = \sqrt{n}$ oder $k = \sqrt{2n}$
- Wenn mögliche natürliche Klasseneinteilung nutzen (z.B: Notenstufen)
- Möglichst gleich große Klassenbreiten wählen.

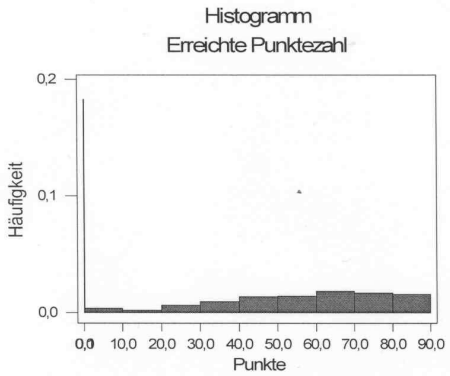








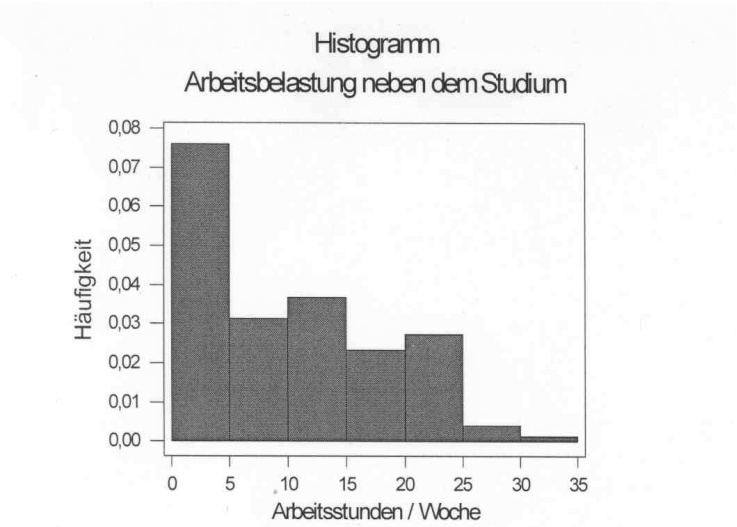
Breite des ersten Balkens 0.5

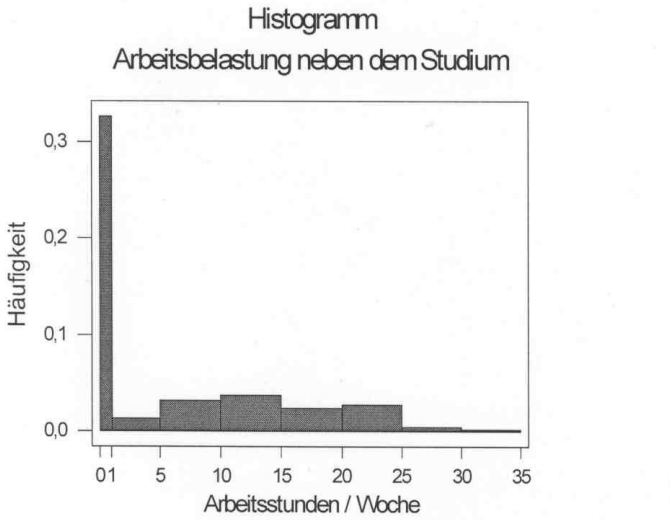


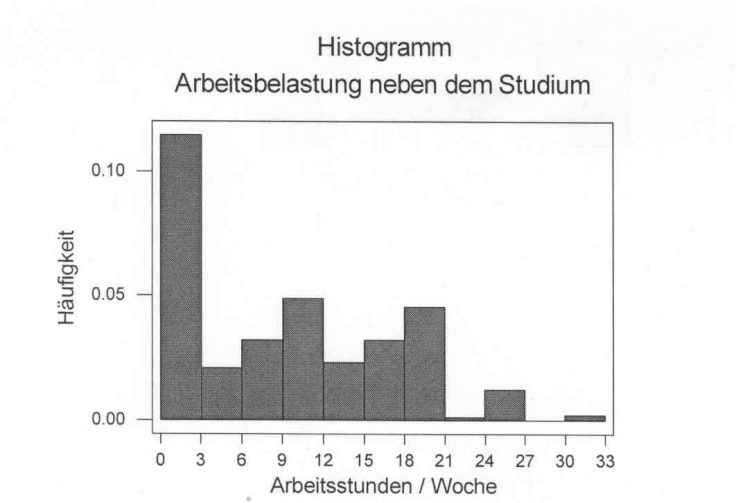
Breite des ersten Balkens 0.1

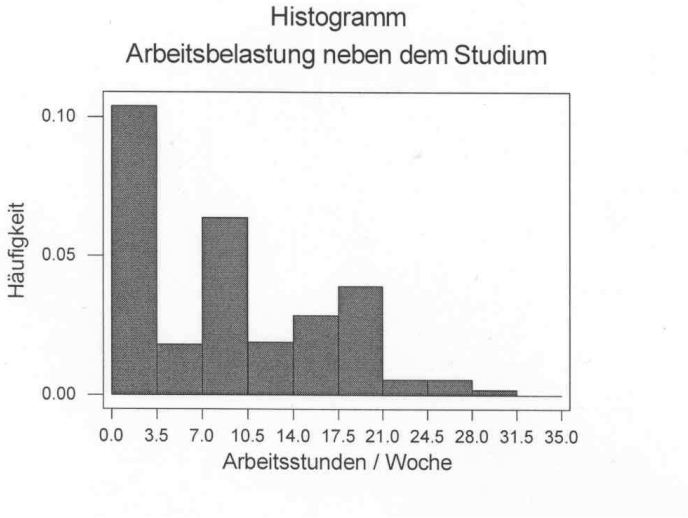
Die Ordinatenbezeichnung „Häufigkeiten“ ist in vielen Softwarepaketen standard, aber eigentlich sehr irreführend !

Beispiel: Arbeitsbelastung neben dem Studium









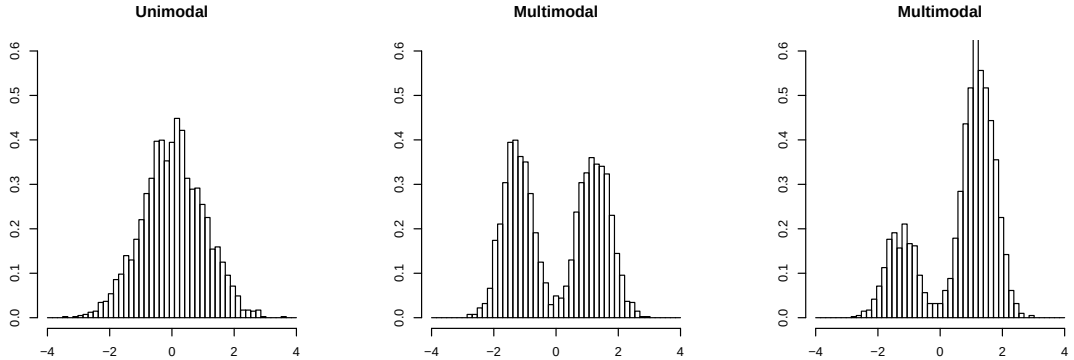
- Ein Wert bildet eine inhaltliche Kategorie für sich, z.B. der Wert 0 bei Arbeitsbelastung neben Studium:
  - die „natürliche Breite“ des zugehörigen Intervalls ist gleich 0 und damit die Höhe gleich unendlich$\Rightarrow$ beliebige Peaks produzierbar, die alle anderen Ausprägungen optisch verschwinden lassen.
  - Mögliche Lösung: Wert aus dem Histogramm nehmen und auf zwei Grafiken aufteilen: Arbeit ja/nein, Verteilung der Arbeitsstunden bei den Arbeitenden (Vorsicht bei der Interpretation!)

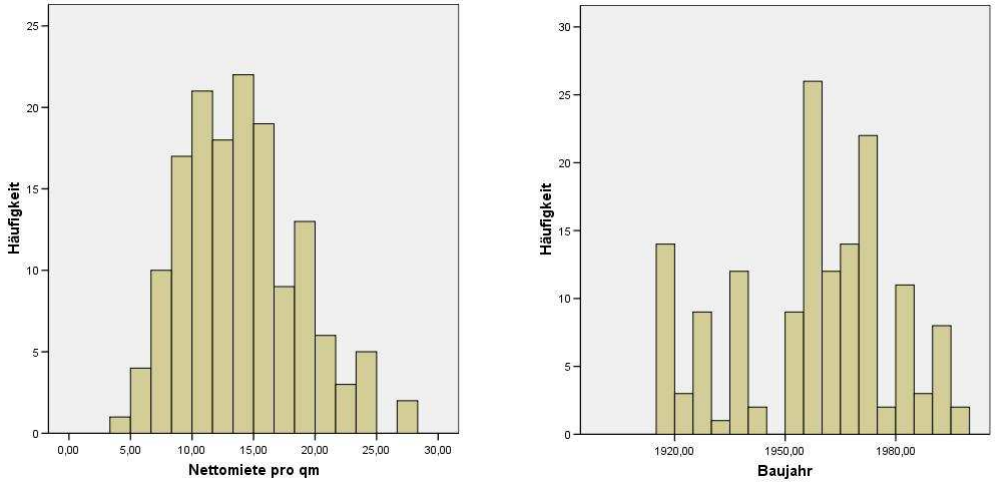
- Implizite Rundung auf „Jubiläumszahlen“ (attractive numbers, Heaping)
  z.B. 5, 10, 20 . . . ; $\frac{1}{2}$ Jahr, 1 Jahr . . . ; 16, 18, 25, 30 Monate
  Zur Vermeidung von Artefakten: Jubiläumszahlen nicht als Intervallgrenzen verwenden!

Typen von Häufigkeitsverteilungen

Histogramme eignen sich gut zur Beurteilung der Form von Häufigkeitsverteilungen

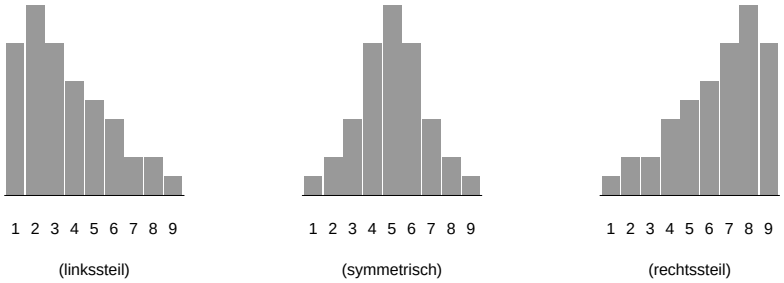
- **Unimodale und multimodale Verteilungen**
  Modus= Gipfel einer Verteilung



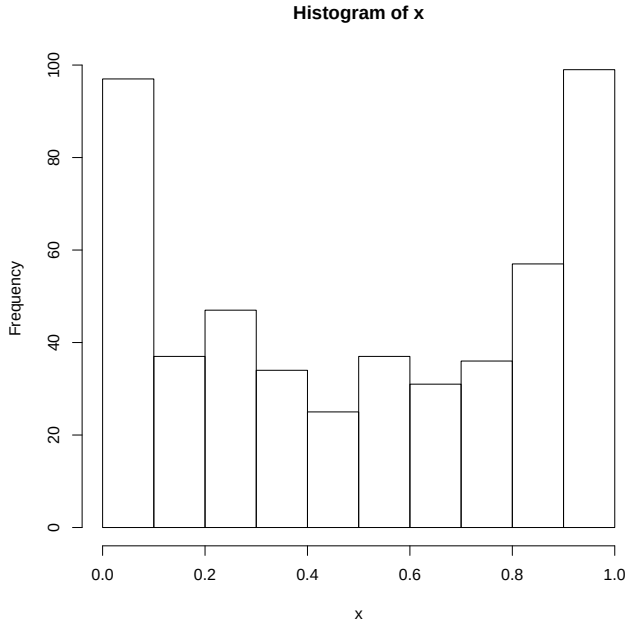


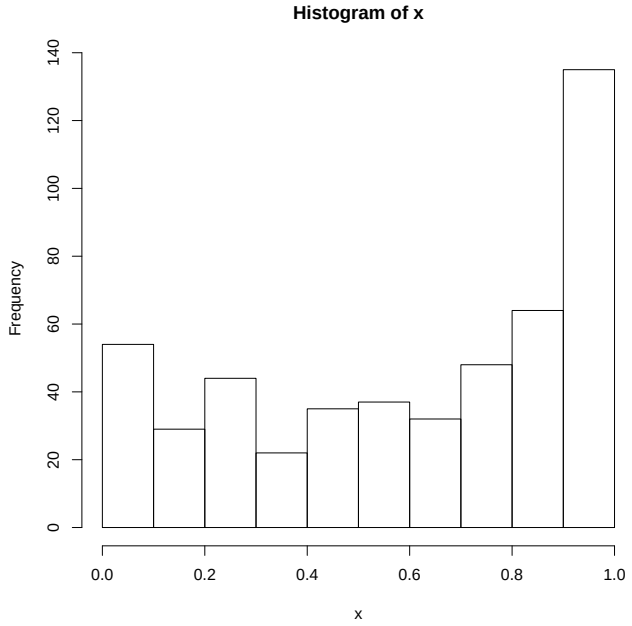
- **Symmetrie und Schiefe**

symmetrisch:	Rechte und linke Hälfte der Verteilung sind annähernd zueinander spiegelbildlich.
linkssteil (rechtsschief):	Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab.
rechtssteil (linksschief):	Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab.



- Andere typische Verteilungsformen:
  - U-förmig,
  - J-förmig.





2.4 Kumulierte Häufigkeiten und empirische Verteilungsfunktion

Oft sind kumulierte Häufigkeiten von Interesse, also eine Antwort auf die Frage „Wieviel Prozent der Daten über-/unterschreiten einen bestimmten Wert?“

- Wieviel Prozent der Studenten arbeiten bis zu 8 Stunden pro Woche neben dem Studium?
- Wieviel Prozent der Studenten arbeiten mehr als 8 Stunden pro Woche neben dem Studium?
- Wieviel Prozent der Studenten haben mindestens 35.5 Punkte, also die Klausur bestanden?

Voraussetzung: Mindestens ordinalskaliertes Merkmal.

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) ordinalskalierten Merkmals mit der Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$.

Dann heißt

$$\begin{aligned} H(x) &:= \text{Anzahl der Werte } x_i \text{ mit } x_i \leq x \\ &= \sum_{j:a_j \leq x} h(a_j) = \sum_{j:a_j \leq x} h_j \end{aligned}$$

absolute kumulierte Häufigkeitsverteilung und

$$\begin{aligned} F(x) &:= \text{Anteil der Werte } x_i \text{ mit } x_i \leq x \\ &= H(x)/n \\ &= \sum_{j:a_j \leq x} f(a_j) = \frac{1}{n} \sum_{j:a_j \leq x} h(a_j) \end{aligned}$$

relative kumulierte Häufigkeitsverteilung bzw. **empirische Verteilungsfunktion**.

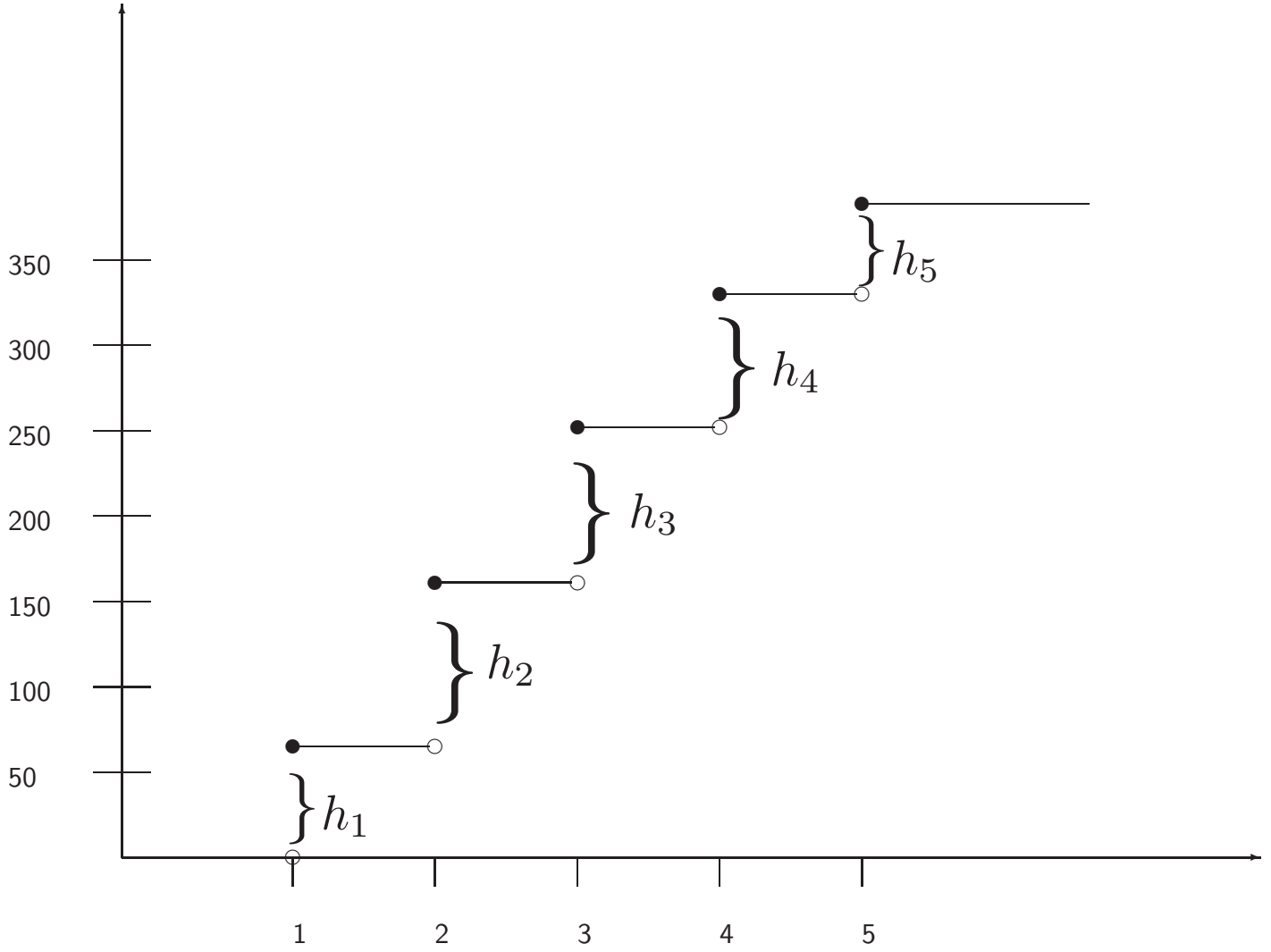
Die Schreibweise $H(x) := \sum_{j:a_j \leq x} h(a_j)$ ist eine Abkürzung für

$$H(x) := \sum_{j \in J_x} h(a_j) \text{ mit } J_x := \{j | a_j \leq x\},$$

d.h. für jedes x wird die Summe über alle j mit der Eigenschaft betrachtet, dass die zugehörigen Werte a_j kleiner gleich x sind (analog für $F(x)$).

Beispiel: Klausurnoten (zur Vereinfachung $a_j = j$)

Note: a_j	$h(a_j)$	$H(a_j)$	$f(a_j)$	$F(a_j)$
$a_1 = 1$	65	65	0.17	0.17
$a_2 = 2$	96	161	0.25	0.42
$a_3 = 3$	91		0.24	
$a_4 = 4$	78		0.20	
$a_5 = 5$	53		0.14	
	383		1.00	



Bemerkungen:

- $F(x)$ sieht genauso aus! Einfach Maßstab auf der y-Achse durch 383 teilen!

- Man kann aus $H(x)$ und $F(x)$ die Häufigkeitsverteilungen $h_1, \ldots, h_k$ und $f_1, \ldots, f_k$ reproduzieren, z.B. ist

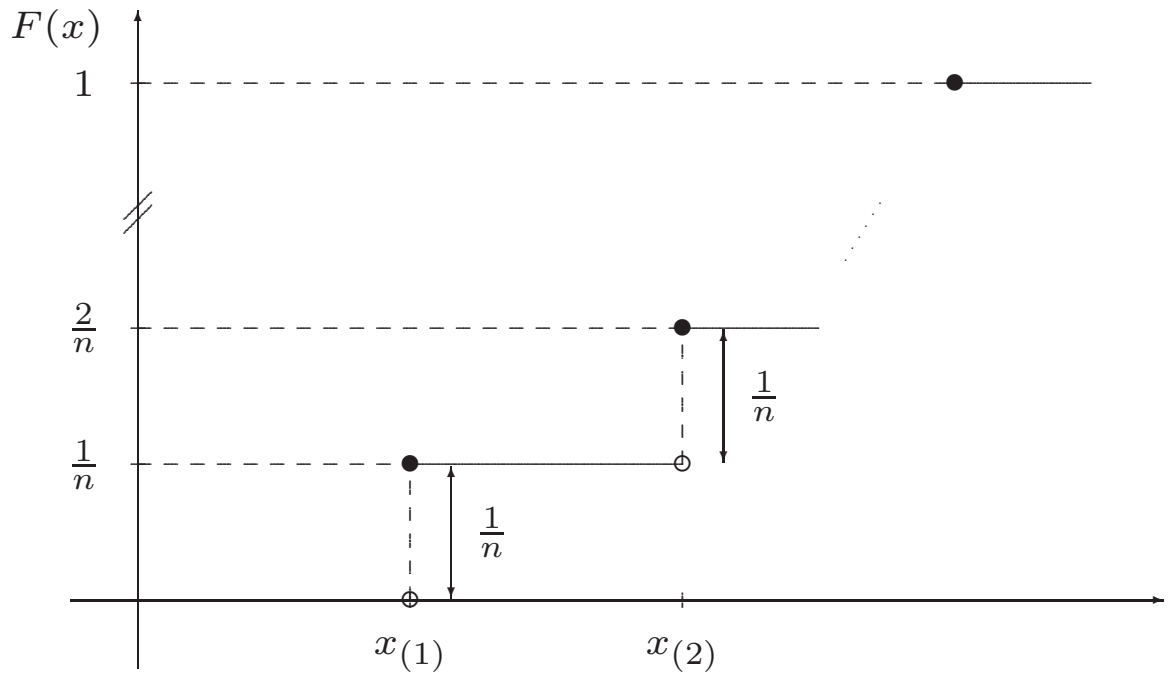
$$h(a_j) = H(a_j) - H(a_{j-1})$$

die Häufigkeit von a_j, beide Darstellungen enthalten also die volle Information über die Häufigkeitsverteilung.

- Bei rein ordinalen Merkmalen ist die Skaleneinteilung auf der Abszisse (x-Achse) völlig willkürlich

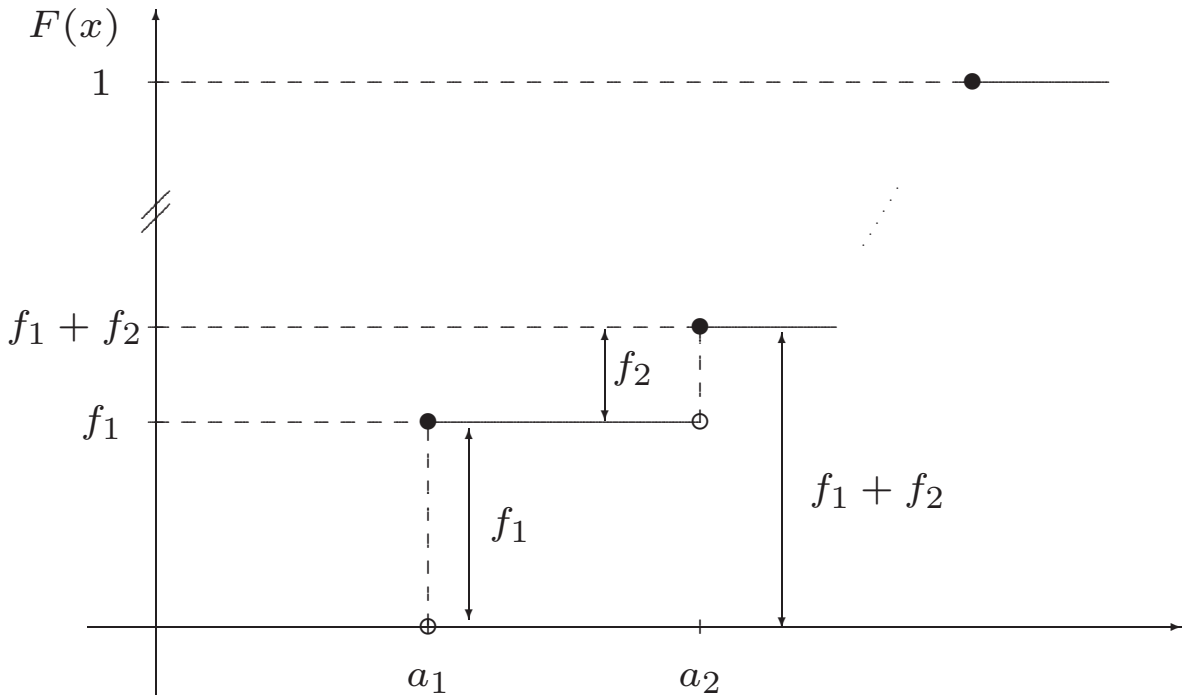
- Bei intervallskalierten Merkmalen ist diese Willkürlichkeit nicht mehr vorhanden $\Rightarrow$ kumulierte Häufigkeitsverteilungen werden fast nur bei intervallskalierten Merkmalen betrachtet.

- Empirische Verteilungsfunktion wenn alle Beobachtungen verschieden sind:



Die empirische Verteilungsfunktion ist eine Treppenfunktion mit Sprüngen an den Stellen $x_1, \ldots, x_n$. Die Höhe aller Sprünge ist $1/n$.

- Empirische Verteilungsfunktion bei gegebenen Häufigkeiten:



Die empirische Verteilungsfunktion ist eine Treppenfunktion mit Sprüngen an den Stellen $a_1 < \cdots < a_k$. Die Höhe des Sprunges an der Stelle a_i ist f_i.

Kumulierte Häufigkeiten bei gruppierten Merkmalen

Beispiel: Punkteverteilung in den Klassen

Klassen	Häufigkeiten	kumuliert
$[0, 35.5)$	53	53
$[35.5, 48.5)$	78	131
$[48.5, 64.5)$	91	
$[64.5, 79.5)$	96	
$[79.5, 90)$	65	

Kumulierte Häufigkeiten bei gruppierten Merkmalen

- Werte der kumulierten Häufigkeitsverteilungen an den zu den Intervallgrenzen gehörenden Punkten sind klar

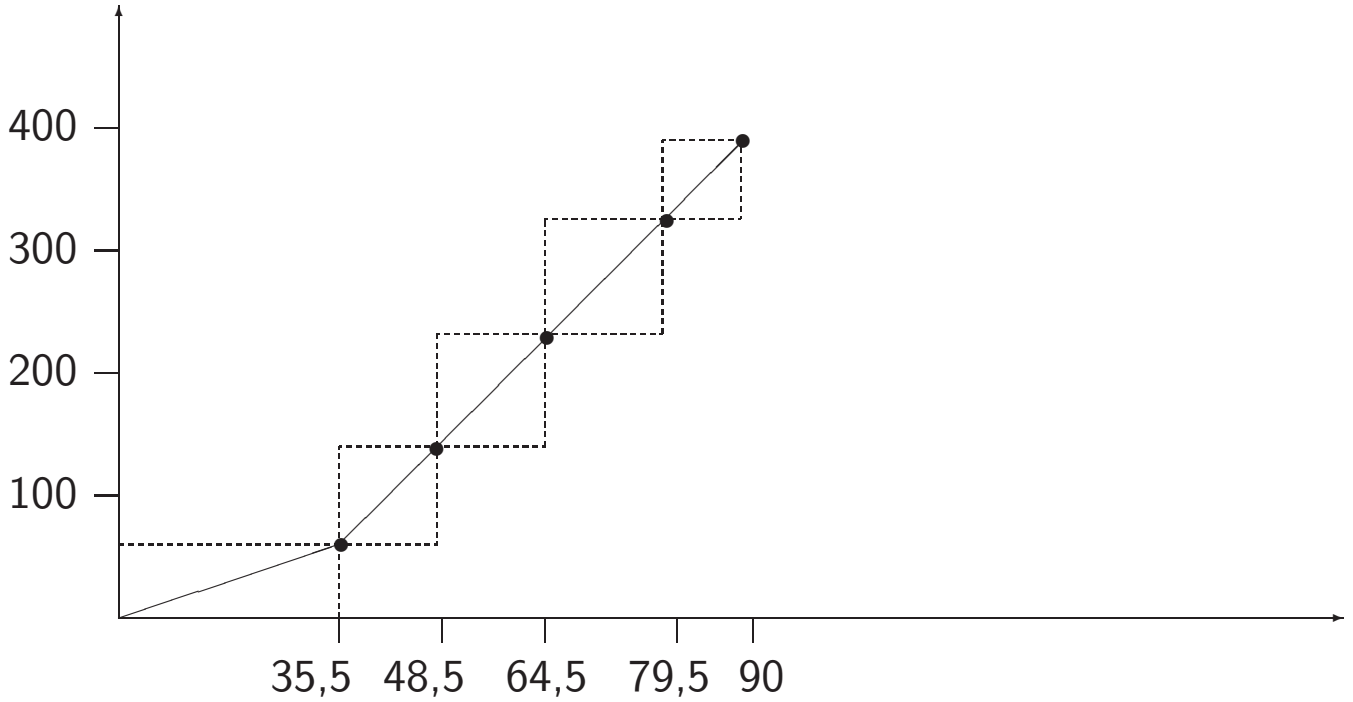
- Wie definiert man $H(x)$ und $F(x)$ zwischen diesen Punkten, was also ist z.B. $H(40)$?

  H ist nicht mehr notwendigerweise konstant zwischen den Klassengrenzen:

  – „40“ ist eine Ausprägung, die in den unklassierten Daten vorkam. Dass 40 ein möglicher Wert ist, soll auch in der Grafik zum Ausdruck kommen.

  – $H(40)$ ist aber aus den klassierten Daten nicht mehr rekonstruierbar. Eigentlich weiß man nur, dass $H(40)$ einen Wert in dem entsprechenden Rechteck annehmen kann.

$\Rightarrow$ **lineare Interpolation** : der unbekannte Verlauf zwischen $zwei\ Punkten$ wird durch eine Gerade durch diese Punkte angenähert.



Die Steigungen der „Gradenstücke“ sind (im Allgemeinen) unterschiedlich!

Allgemeine Formulierung:

- k Klassen $[c_0, c_1), \ldots, [c_{j-1}, c_j), \ldots, [c_{k-1}, c_k]$
- h_j Häufigkeit in j-ter Klasse.
- Verwende bei einem x aus der Klasse $[c_{j-1}, c_j)$ als Approximation für $H(x)$ den aus der linearen Interpolation gewonnenen Punkt
- Geradengleichung:

$$H(x) \quad \approx \quad H(c_{j-1}) \quad + \quad \frac{h_j}{(c_j - c_{j-1})} \quad \cdot \quad (x - c_{j-1})$$

$$H(40) \quad \approx$$