3 Lage- und Streuungsmaße

Grafische Darstellungen geben einen allgemeinen Eindruck der Verteilung eines Merkmals, u.a. von

- Lage und Zentrum der Daten,
- Streuung der Daten um dieses Zentrum,
- Schiefe / Symmetrie und Unimodalität / Multimodalität der Daten.

Oft ist (zur weiteren Informationsverdichtung) die Beschreibung einer Verteilung durch *eine* bzw. wenige Maßzahlen gewünscht:

- Lagemaße sollen die *zentrale Tendenz* (das Zentrum) eines Merkmals beschreiben.
- Streuungsmaße beschreiben die *Variabilität* eines Merkmals.

3.1 Lagemaße

*Lage*maße beantworten Fragen über die *Lage* der Häufigkeitsverteilung, wie:

- Wo liegen die meisten Beobachtungen?
- Wo liegt der „Schwerpunkt“ einer Verteilung?
- Wo liegt die „Mitte“ der Beobachtungen?
- Was ist eine „typische“ Beobachtung?

Bemerkungen:

- Es gibt nicht das Lagemaß schlechthin. Die unterschiedlichen Lagemaße sind je nach Situation unterschiedlich geeignet.
- Die Eignung ist insbesondere abhängig von der Datensituation und dem Skalenniveau.

3.1.1 Arithmetisches Mittel

Definition 3.1.

Sei $x_1, \ldots, x_n$ die Urliste eines (mindestens) intervallskalierten Merkmals X. Dann heißt

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

das *arithmetische Mittel* der Beobachtungen $x_1, \ldots, x_n$.

Bemerkungen:

- Das arithmetische Mittel ist also das Lagemaß, das typischerweise als Mittelwert oder Durchschnitt bezeichnet wird.
- Das arithmetische Mittel muss nicht mit einer der beobachteten Ausprägungen zusammenfallen.

Beispiel: Anzahl von Statistikbüchern, die die Studierenden jeweils besitzen

Person i	Anzahl der Bücher x_i
1	0
2	2
3	1
4	2
5	2
6	3
7	0
8	12
9	1
10	2

$$\bar{x} \quad =$$

Alternative Berechnung basierend auf Häufigkeiten:

Hat das Merkmal X die Ausprägungen $a_1, \ldots, a_k$ und die (relative) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$, so gilt

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j h_j = \sum_{j=1}^{k} a_j f_j$$

Im Beispiel: Häufigkeitstabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12
$\vert\vert$												$\vert$

bzw.

Berechnung aus den Merkmalsausprägungen:

$$\bar{x} = \frac{1}{10} \cdot (0 + 2 + 1 + 2 + 2 + 3 + 0 + 12 + 1 + 2) =$$

Berechnung aus der Häufigkeitsverteilung:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{j=1}^{k} a_j \cdot h_j \\ &= \end{aligned}$$

Beispiel: Einfacher Tabellenmietspiegel

Nettomiete in Euro/qm (Fallzahlen)				
	Wohnfläche			
Baujahr	bis 50 qm	51 bis 80 qm	81 qm und mehr	
bis 1918	9.00 (45)	7.88 (164)	7.52 (200)	7.83 (409)
1919 bis 48	6.90 (42)	6.87 (94)	6.50 (52)	6.78 (188)
1949 bis 65	9.04 (129)	7.84 (237)	7.95 (70)	8.21 (436)
1966 bis 80	10.05 (173)	7.97 (313)	7.80 (156)	8.49 (642)
1981 bis 95	10.59 (45)	9.53 (162)	9.72 (63)	9.75 (270)
1996 bis 2001	10.60 (15)	10.28 (58)	9.69 (35)	10.14 (108)
	9.43 (449)	8.20 (1028)	7.93 (576)	8.39 (2053)

Beispiel: Augenfarbe

	h_j
0: grün	2
1: grau	2
2: rot	0
3: blau	6

$$\bar{x} =$$

Bemerkungen:

- Das arithmetische Mittel setzt zwingend ein intervallskaliertes Merkmal voraus.
- Auf einem niedrigeren Skalenniveau ist die Addition nicht erlaubt, und daher sind die entsprechenden Mittelwertsbildungen sinnlos und nicht interpretierbar (auch wenn sie das Software-Paket berechnet!).
- Einzige Ausnahme: Binäre Merkmale (mit nur zwei Ausprägungen), deren Ausprägungen als 0/1 kodiert werden. In diesem Fall kann das arithmetische Mittel als Anteil von Beobachtungen mit Ausprägung 1 interpretiert werden.

Transformationen

- Die Intervallskala erlaubt Transformationen der Form $aX + b$, die Verhältnisskala Transformationen der Form $a \cdot X$, wobei a und b feste Konstanten sind.

- Aus der Urliste $x_1, x_2, \ldots, x_n$ kann man eine Urliste, nämlich der transformierten Werte $y_1, y_2, \ldots, y_n$ mit $y_i = ax_i + b$, $i = 1, \ldots, n$ bestimmen.

Wie verändert sich das arithmetische Mittel bei diesen oder allgemeineren Transformationen?

Beispiele

- Linear affine Transformation $Y = a \cdot X + b$

X	jährliche Ausgaben von Studierenden 2010 in Euro
Y	jährliche Ausgaben von Studierenden 2010 in USD ohne Studiengebühren

- Nichtlineare Transformation
  Beispiel: 3 quadratische Zimmer mit den Seitenlängen 7, 4 und 10m.
  Sei X die Seitenlänge, dann ist

$$Y = g(X) = X^2 \quad \text{die Zimmerfläche,}$$

und es gilt

$$\bar{x} =$$

$$\bar{y} =$$

Satz 3.2. *Arithmetisches Mittel und lineare Transformationen.*

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) intervallskalierten Merkmals X. Betrachtet wird das (linear transformierte) Merkmal $Y = a \cdot X + b$ und die zugehörigen Ausprägungen $y_1, \ldots, y_n$.
Dann gilt:

$$\bar{y} = a \cdot \bar{x} + b.$$

Beweis: Von der Urliste $x_1 \ldots x_n$ von X zur Urliste $y_1 \ldots y_n$ von Y übergehen. Dabei gilt für jedes i: $y_i = a \cdot x_i + b$.

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a \cdot x_i + b) = \\
&= \\
&= \\
&= \\
&=
\end{aligned}$$

Bemerkungen:

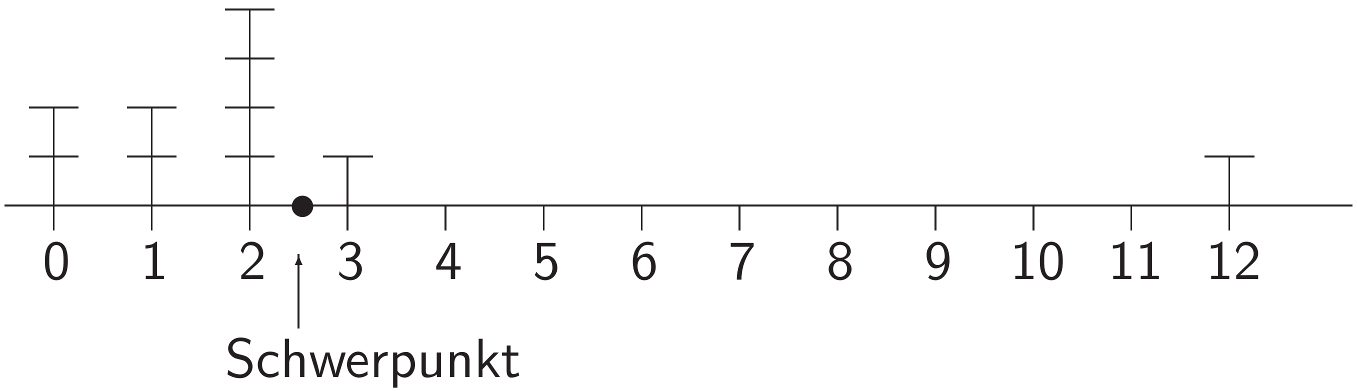
- Ist X verhältnisskaliert, so geht für $b \neq 0$ der natürliche Nullpunkt für Y verloren.
- Der Satz gilt im Allgemeinen nur, falls die Transformation von X auf Y linear ist. Z.B. ist bei $Y = X^2$ im Allgemeinen $\bar{y} \neq (\bar{x})^2$ (wie im Beispiel gezeigt).

Weitere Eigenschaften des arithmetischen Mittels:

- $\bar{x}$ ist derjenige Wert, den jede Beobachtungseinheit erhielte, würde man die Gesamtsumme der Merkmalsausprägungen gleichmäßig auf alle Einheiten verteilen.
- $\bar{x}$ ist der Schwerpunkt der $x_1, \ldots, x_n$, d.h. es gilt:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Vorstellung: Für jede Beobachtung i im Punkt x_i Gewicht mit 1kg hinlegen.



- Die Schwerpunktseigenschaft macht auch deutlich, dass extrem große und kleine Werte außerordentliche Hebelwirkung haben: lässt man die Beobachtung 12 im Beispiel weg, dann gilt:
$$\bar{x} = \frac{13}{9} = 1.44.$$
Das arithmetische Mittel ist sehr *ausreißeranfällig*, d.h. ein falsch gemessener Wert kann „den ganzen Mittelwert zerstören“oder ein tatsächlich extremer Wert ein völlig falsches Bild entstehen lassen.

- Befürchtet man Ausreißer, so weicht man gelegentlich auf das sogenannte *α-getrimmte Mittel* aus, bei dem man die $\alpha\%$ größten und kleinsten Werte (z.B. α=5) weglässt; meist verwendet man den sog. Median (s.u.).

Gruppierte Daten: Häufig hat man die Daten nur in gruppierter Form vorliegen. Wie lässt sich in diesem Fall ein sinnvoller Mittelwert definieren?

Typisches Beispiel: Einkommensverteilung

	Anzahl h'_l	
$0 \leq x < 750$	3	
$750 \leq x < 1250$	8	
$1250 \leq x < 1750$	6	
$1750 \leq x < 2250$	2	
$2250 \leq x < 3250$	1	
$\sum$	20	

Definition 3.3.

Sei X ein intervallskaliertes Merkmal, das in gruppierter Form mit k Klassen $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k]$ erhoben wurde. Mit $h'_l,\ l = 1, \ldots k$, als absoluter Häufigkeit der $l-$ten Klasse, f'_l als zugehöriger relativer Häufigkeit und $m_l := \frac{c_l + c_{l-1}}{2}$ als der jeweiligen Klassenmitte definiert man als *arithmetisches Mittel für gruppierte Daten*

$$\bar{x}_{grupp} := \frac{1}{n} \sum_{l=1}^{k} h'_l m_l = \sum_{l=1}^{k} f'_l m_l.$$

Im Beispiel:

Bemerkungen:

- Bei nach oben offener letzter Kategorie (Einkommen größer als 2250) wäre die Klassenmitte nicht definiert.
- Im Allgemeinen gilt $\bar{x} \neq \bar{x}_{grupp}$; nur in Extremfällen, z.B. wenn das Merkmal in jeder Gruppe gleichmäßig verteilt ist, erhält man die Gleichheit.
- $\bar{x}_{grupp}$ hängt von der Gruppenmitte und damit von der gewählten Gruppierung ab: Fasst man z.B. die ersten drei Gruppen und die letzten beiden jeweils zusammen, so erhält man

	h'_l	m_l
$0 \leq x < 1750$	17	
$1750 \leq x < 3250$	3	

und

$$\bar{x}_{grupp} = \frac{1}{n}\sum_{l=1}^{k} h'_l m_l$$

- Im Allgemeinen ist $\bar{x}_{grupp}$ nur eine grobe Approximation für den „echten“, d.h. auf ungruppierten Daten beruhenden, Mittelwert. Eigentlich kann man nur mit Sicherheit folgende Abschätzung geben: Jeder in der l-ten Gruppe verdient mindestens c_{l-1} und höchstens c_l. Damit ergibt sich als Abschätzung für das arithmetische Mittel

$$\frac{1}{n}\sum_{l=1}^{k} h_l c_{l-1} \leq \bar{x} \leq \frac{1}{n}\sum_{l=1}^{k} h_l c_l$$

  Diese Abschätzung ist oft relativ grob. Andererseits ist sie aber oft das Beste, was man ohne unüberprüfbare Zusatzannahmen aus den Daten herausholen kann.

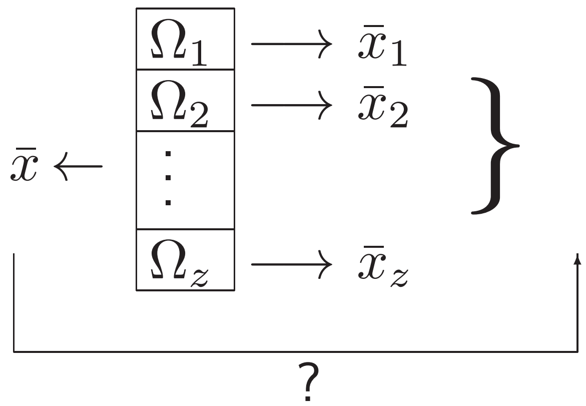
- Sind die ungruppierten Daten erhältlich, so ist $\bar{x}$ vorzuziehen, da jede Gruppierung Informationsverlust mit sich bringt.

- Andererseits sind gruppierte Daten leichter (und oft wahrheitsgetreuer) erhebbar.

Geschichtete Daten

Insbesondere bei Tertiäranalysen hat man häufig nicht die Urliste zur Verfügung, sondern nur Mittelwerte $\bar{x}_l$ in einzelnen Schichten $l = 1, \ldots, z$, in die die Grundgesamtheit zerlegt ist.

$X \colon \Omega \to \mathbf{R}$



Beachte: hier wird nicht das Merkmal sondern die Grundgesamtheit in Gruppen eingeteilt.

Beispiel:

$\bar{x}_l$	Durchschnittseinkommen in den einzelnen Bundesländern $(l = 1, \ldots, 16)$
$\bar{x}$	Durchschnittseinkommen in der BRD

Mittelwert der Mittelwerte $\frac{1}{z}\sum_{\ell=1}^{z}\overline{x}_\ell$?

3.1.2 Median & Quantile

- Wie lässt sich ein „Mittelwert“bei ordinalskalierten Merkmalen definieren?
- Das arithmetische Mittel besitzt die Schwerpunkteigenschaft

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

- Eine andere mögliche Schwerpunkteigenschaft: Rechts und links des Mittelwerts liegen jeweils (mindestens) 50% der Daten. Dies ergibt den Median.

Definition 3.4.

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) ordinalskalierten Merkmals X. Jede Zahl x_{med} mit

$$\frac{|\{i|x_i \leq x_{med}\}|}{n} \geq 0.5 \quad \text{und} \quad \frac{|\{i|x_i \geq x_{med}\}|}{n} \geq 0.5$$

heißt Median.

Beispiel: Klausurnoten

$$\underbrace{1{,}1{,}1, \ldots, 1}_{} \qquad \underbrace{2{,}2{,}2, \ldots, 2}_{} \qquad \underbrace{3{,}3{,}3, \ldots, 3}_{} \qquad \underbrace{4{,}4{,}4, \ldots, 4}_{} \qquad \underbrace{5{,}5{,}5, \ldots, 5}_{}$$

65 mal	96 mal	91 mal	78 mal	53 mal
17%	25,1%	23,8%	20,4%	13,8%

Verallgemeinerung: Quantile

Gegeben sei die Urliste
$x_1, \ldots, x_n$ eines (mindestens) ordinalskalierten Merkmals X und eine Zahl $0 < \alpha < 1$.
Jede Zahl x_α mit

$$\frac{|\{i|x_i \leq x_\alpha\}|}{n} \geq \alpha \quad \text{und} \quad \frac{|\{i|x_i \geq x_\alpha\}|}{n} \geq 1 - \alpha$$

heißt $\alpha \cdot 100\%$-*Quantil*.

Spezielle Quantile:

- Median: $x_{0.5} = x_{med}$.
- Quartile: $x_{0.25}$, $x_{0.75}$.
- Dezile: $x_{0.1}$, $x_{0.2}$, ... , $x_{0.8}$, $x_{0.9}$.

Beispiel Klausurnoten:

$x_{0.25} =$ $\qquad$ $x_{0.1} =$

Bemerkungen:

- Alternative Definition des Medians über die *geordnete* Urliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

$$x_{med} := \begin{cases} \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{für } n \text{ gerade} \\ x_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \end{cases}$$

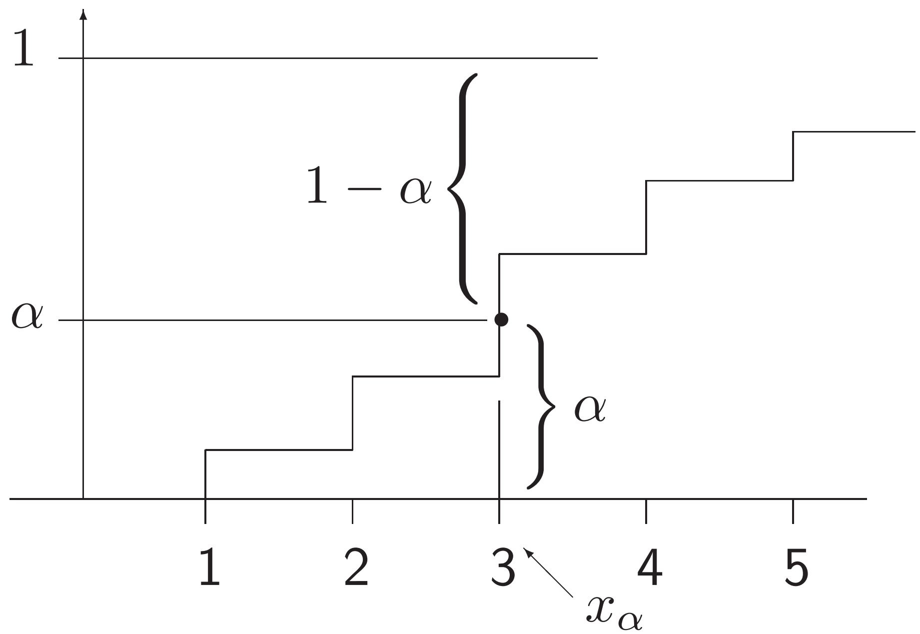
  Ähnliche Definitionen sind für andere Quantile möglich.

- Diese Definition ist insofern inkonsequent, als sie die bei ordinalen Daten evtl. nicht mögliche (bzw. nicht zulässige) Addition verwendet.

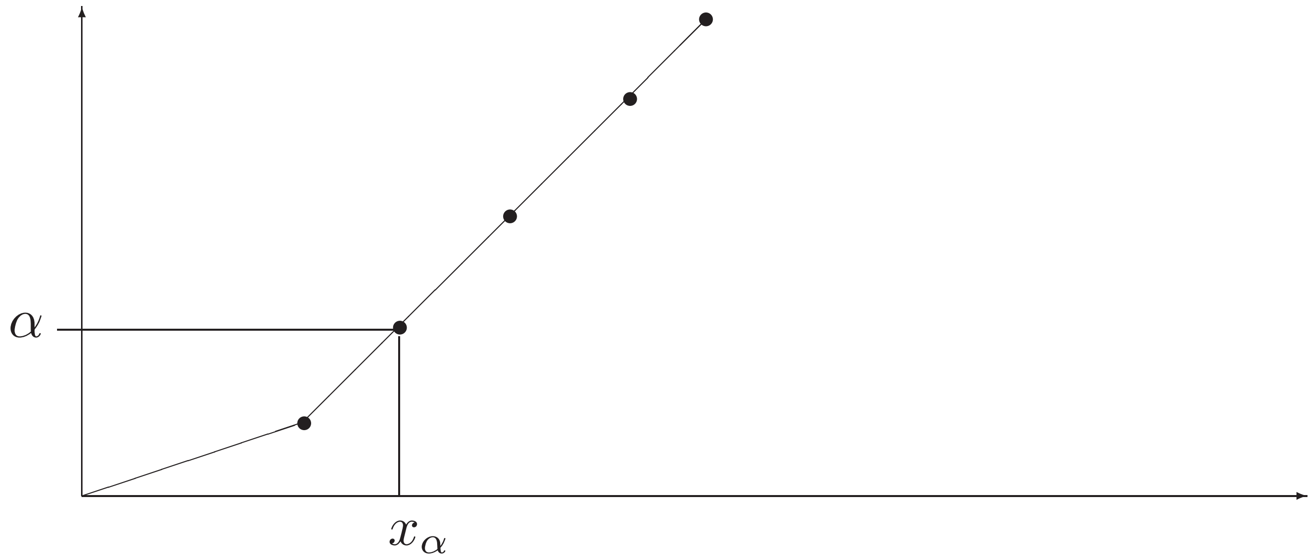
- Bei intervallskalierten Daten hingegen spricht vieles für diese Definition: In manchen Fällen können Quantile im Sinne der ursprünglichen Definition nicht eindeutig sein.

- In vielen praktisch relevanten Fällen sind beide Definitionen miteinander verträglich. Für n ungerade fallen sie stets zusammen, für n gerade stimmen sie überein, falls $x_{\left(\frac{n}{2}\right)} = x_{\left(\frac{n}{2}+1\right)}$.

- Quantile kann man einfach an der empirischen Verteilungsfunktion ablesen:



- Bei linearer Interpolation für gruppierte intervallskalierte Merkmale definiert man die Quantile analog über den Schnittpunkt mit der Verteilungsfunktion:



Transformationen: Wie ändert sich der Median bei Transformation der Daten?

Satz 3.5.

Sei $x_1, x_2, \ldots, x_n$ die Urliste eines (mindestens) ordinalskalierten Merkmals X, g eine streng monoton steigende Funktion und $y_1 = g(x_1), \ldots, y_n = g(x_n)$ die Urliste des Merkmals $Y = g(X)$. Dann gilt:

$$y_{med} = g(x_{med}).$$

$$\begin{tikzcd}
\text{Merkmal } X \arrow[r, "g \text{ streng monoton}"] \arrow[d] & \text{Merkmal } Y \arrow[d] \\
x_{med} \arrow[r, "g"] & y_{med}
\end{tikzcd}$$

Beispiel: Drei quadratische Zimmer

Für die Merkmale X (Seitenlänge) und $Y = f(X) = X^2$ (Fläche) galt ja mit den Daten

$$
\begin{array}{llll}
 & x_1 = 7, & x_2 = 4, & x_3 = 10 \\
\text{und} & y_1 = x_1^2 = 49 & y_2 = x_2^2 = 16 & y_3 = x_3^2 = 100
\end{array}
$$

Gegenbeispiel mit nicht monotoner Funktion: $g(X) = (X-6)^2$ ist nicht monoton, sondern u-förmig.

- Für das Merkmal $Z = g(X) = (X-6)^2$ ergeben sich die Merkmalsausprägungen $z_1 = 1, z_2 = 4$ und $z_3 = 16$ und damit der Median $z_{med} = 4$
- Für den transformierten Median gilt aber $g(x_{med}) = g(7) = 1$.

Der Median ist wegen seiner Invarianz gegenüber beliebigen streng monotonen Transformationen ein geeignetes Lagemaß auch in allen Situationen, in denen es trotz Intervallskala keine natürliche Maßeinheit gibt

- Bei vielen Messungen nicht klar, ob man auf einer linearen oder auf einer logarithmischen Skala messen soll.

$\Rightarrow$ Betrachtung von sogenannten Rangstatistiken, d.h. von Verfahren, die nicht den genauen Wert einer Beobachtung an sich verwenden, sondern nur den Rangplatz. („Verteilungsfreie Verfahren“)

3.1.3 Modus

- Gesucht: geeignetes Lagemaß bei auf Nominalskala gemessenen Daten
- Der exakte Wert der als Merkmalsausprägungen vergebenen Zahlen ist inhaltlich völlig bedeutungslos, d.h. (etwas formaler): beliebige eineindeutige Transformationen verändern die inhaltliche Aussage nicht (z.B. Parteienpräferenz: ob man die Partei alphabetisch durchnummeriert oder anhand ihrer Stimmenanteile bei der letzten Wahl ändert nichts).
- Als Lagemaß dient der *häufigste Wert*: genauer die Ausprägung a_j mit der größten Häufigkeit h_j.

Definition 3.6.

Sei $x_1, \ldots, x_n$ die Urliste eines nominalskalierten Merkmals mit den Ausprägungen $a_1, \ldots, a_k$ und der Häufigkeitsverteilung $h_1, \ldots, h_k$. Dann heißt a_{j^*} genau dann *Modus* x_{mod}, wenn $h_{j^*} \geq h_j$, für alle $j = 1, \ldots, k$.

Bemerkungen:

- Der Modus wird auch als Modalwert bezeichnet.
- Existieren mehrere Ausprägungen mit der gleichen größten Häufigkeit, so ist der Modus nicht eindeutig.
- Der Modus bleibt unter beliebigen eineindeutigen Transformationen erhalten: Betrachtet man das Merkmal X, eine eineindeutige Transformation g und das Merkmal $Y = g(X)$, so gilt

$$y_{mod} = g(x_{mod}).$$

$$\begin{tikzcd}
\text{Merkmal } X \arrow[r, "g \text{ eineindeutig}"] \arrow[d] & \text{Merkmal } Y \arrow[d] \\
\text{Modus } x_{mod} \arrow[r, "g"] & \text{Modus } y_{mod}
\end{tikzcd}$$

3.1.4 Vergleich der Lagemaße

- Bei intervallskalierten Daten kann man auch den Modus oder den Median angeben, am besten zusätzlich zum arithmetischen Mittel.
- Der Median geht nur auf die Ordnung der Beobachtungen und nicht auf die Abstände ein, der Modus gibt nur die am stärksten vertretene Ausprägung an.
- Anschaulich gesprochen ist der Median der mittlere Wert, was oft umgangssprachlich auch als Mittelwert bezeichnet wird. Immer genau angeben, welchen Mittelwert man verwendnet (Median, arithmetisches Mittel, ...); Vorsicht beim Lesen von Veröffentlichungen!
- Median und Modus sind unempfindlich gegenüber Ausreißern.

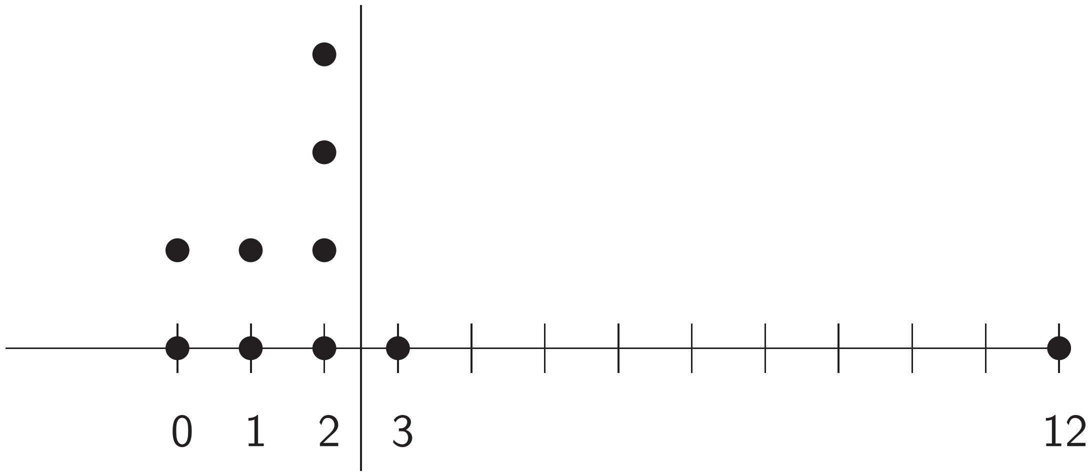
Beispiel: Einkommensverteilung

- Wird die größte Beobachtung verhundertfacht, so ändern sich Median und Modus nicht, das arithmetische Mittel reagiert dagegen stark.

- Generell ist bei der Betrachtung von Einkommen das arithmetische Mittel meist deutlich größer als der Median: Laut dem dritten Armuts- und Reichtumsbericht der Bundesregierung (2008) ist für das reale Bruttojahreseinkommen aus unselbstständiger Arbeit unter den Vollbeschäftigten für 2005
  - das arithmetische Mittel 33678
  - der Median 30157.

Beispiel: Statistikbücher. Häufigkeitsverteilung und graphische Veranschaulichung:

	Häufigkeiten
$a_1 = 0$	$h_1 = 2$
$a_2 = 1$	$h_2 = 2$
$a_3 = 2$	$h_3 = 4$
$a_4 = 3$	$h_4 = 1$
$a_5 = 12$	$h_5 = 1$

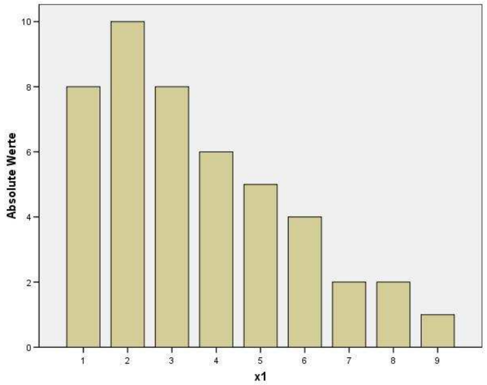


Mittelwert $\bar{x}$=2.5

$x_{0.25}$=1 Median x_{med} =2 $x_{0.75}$=2

Allgemeiner gilt: Die relative Lage von $\bar{x}, x_{med}, x_{mod}$ zueinander kann zur Charakterisierung von Verteilungen herangezogen werden:

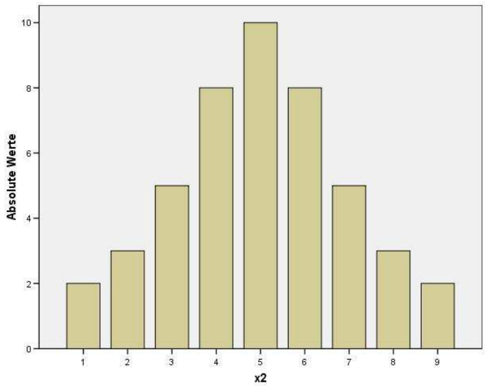
symmetrisch:	$\bar{x}$	$\approx$	x_{med}	$\approx$	x_{mod}
linkssteil:	$\bar{x}$	$>$	x_{med}	$>$	x_{mod}
rechtssteil:	$\bar{x}$	$<$	x_{med}	$<$	x_{mod}



$$\bar{x} = 3.57$$

$$x_{med} = 3$$

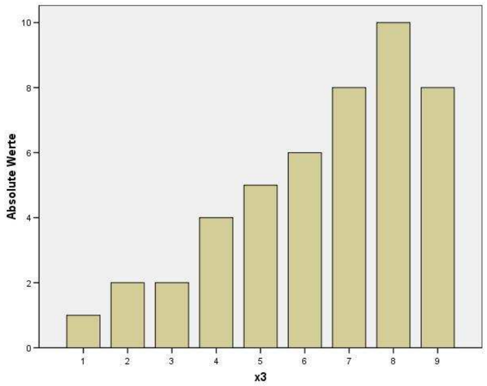
$$x_{mod} = 2$$



$$\bar{x} = 5$$

$$x_{med} = 5$$

$$x_{mod} = 5$$



$$\bar{x} = 6.43$$

$$x_{med} = 7$$

$$x_{mod} = 8$$