3 Lage- und Streuungsmaße

- Grafische Darstellungen geben einen allgemeinen Eindruck der Verteilung eines Merkmals:
  - Lage und Zentrum der Daten,
  - Streuung der Daten um dieses Zentrum,
  - Schiefe / Symmetrie und Unimodalität / Multimodalität der Daten.
- Im Folgenden: Maßzahlen zur Beschreibung von Lage und Streuung durch $eine$ Zahl.
- 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 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 ein Student besitzt (fiktiv).

Person	Anzahl
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

bzw.

Alte Berechnung:

$$\bar{x} =$$

Neue Berechnung:

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

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 niedrigerem Skalenniveau ist die Addition nicht erlaubt, und daher sind die entsprechenden Mittelwertsbildungen sinnlos und nicht interpretierbar (auch wenn sie ein Software-Paket ohne zu zögern ausspuckt).
- Einzige Ausnahme: Binäre Merkmale (mit nur zwei Ausprägungen), deren Ausprägungen als 0/1 (nur so!) kodiert werden. In diesem Fall kann das arithmetische Mittel als Anteil von Beobachtungen mit Ausprägung 1 interpretiert werden.

Transformationen: Die Intervallskala erlaubt lineare Transformationen der Form $a+bX$, die Ratioskala Transformationen der Form $b \cdot X$, wobei a und b feste Konstanten sind, so dass man aus der Urliste $x_1, x_2, \ldots, x_n$ eine neue Urliste $y_1, y_2, \ldots, y_n$ erhält, mit $y_i = ax_i + b$, $i = 1, \ldots, n$. Wie verändert sich das arithmetische Mittel bei diesen oder allgemeineren Transformationen?

Beispiele:

- Lineare Transformation $Y = a \cdot X + b$

- Nichtlineare Transformation

Satz 3.2. [Arithmetisches Mittel und lineare Transformationen.]

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) intervallskalierten Merkmals X mit arithmetischem Mittel $\bar{x}$. Betrachtet wird das (linear transformierte) Merkmal $Y = a \cdot X + b$ und die zugehörigen Ausprägungen $y_1, \ldots, y_n$.
Dann gilt für das arithmetische Mittel $\bar{y}$ von Y:

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

Beweis:

Bemerkungen:

- Vorsicht: 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: außerordentliche Hebelwirkung extrem großer und kleiner Werte: (lässt man die Beobachtung 12 im Beispiel weg, dann gilt: $\bar{x} = \frac{13}{9} = 1.44$. Insbesondere ist damit das arithmetische Mittel sehr $ausrei\text{-}ßeranfällig$, d.h. ein falsch gemessener Wert kann „den ganzen Mittelwert zerstören“.

- 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. Alternativ verwendet man oft den Median (s.u.).

Gruppierte Daten: Häufig hat man die Daten nur in gruppierter Form vorliegen.
Ferner: Anonymisierung
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 $\bar{x}_{grupp}$ natürlich nur eine grobe Approximation an 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

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

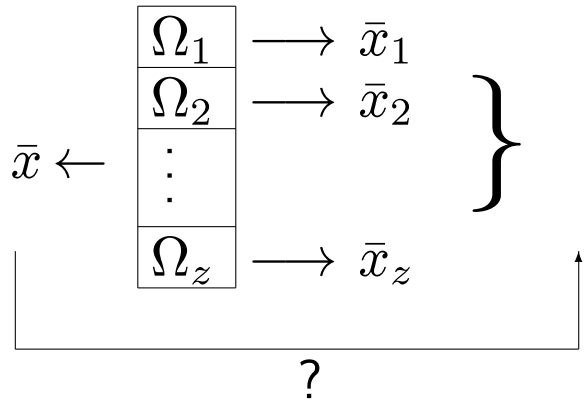
  Diese Abschätzung ist oft relativ grob. Andererseits ist sie aber oft das Beste, was man ohne unüberprüfbare Zusatzannahmen aus den Daten herausholen kann.

- 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:

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 „mittleren Wertes“ $x_{0.5}$ liegen jeweils mit dem Wert selbst (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

$\underbrace{1,1,1,\ldots,1}$	$\underbrace{2,2,2,\ldots,2}$	$\underbrace{3,3,3,\ldots,3}$	$\underbrace{4,4,4,\ldots,4}$	$\underbrace{5,5,5,\ldots,5}$
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
  (z.B. Fahrmeir et al., 2010)
  $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

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

  Ähnlich für andere Quantile möglich.

- Diese Definition ist insofern inkonsequent, als sie auf die bei ordinalen Daten streng genommen nicht zulässige Addiditionen rekurriert. Bei intervallskalierten Daten hingegen spricht vieles für diese Definition.

- Andererseits können in gewissen Grenzfällen Quantile im Sinne der ursprünglichen Definition nicht eindeutig sein:

- Beide Definitionen sind letztlich in vielen praktisch relevanten Fällen miteinander verträglich. Für n ungerade fallen sie stets zusammen, für n gerade stimmen sie überein, falls $x_{\left(\frac{n}{2}\right)} = x_{\left(\frac{n}{2}+1\right)}$
- Man kann Quantile einfach an der empirischen Verteilungsfunktion ablesen:

- Bei linearer Interpolation für gruppierte intervallskalierte Merkmalen definiert man die Quartile 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 mit Median x_{med}, und g eine streng monoton steigende Funktion. Mit $y_1 = g(x_1), \ldots, y_n = g(x_n)$ als Urliste des Merkmals $Y = g(X)$ gilt für den Median y_{med} von Y:

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

Beispiel: Drei quadratische Zimmer

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 $h_1, \ldots, h_k$, so heißt a_{j^*} $Modus$ x_{mod} genau dann, 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 unter beliebigen eineindeutigen Transformationen erhalten: Betrachtet man das Merkmal X, eine eineindeutige Transformation g und das Merkmal $Y = g(X)$, so gilt

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

3.1.4 Vergleich der Lagemaße

- Bei intervallskalierten Daten darf man auch den Modus oder den Median anwenden, man verschenkt (bei alleiniger Verwendung) aber eventuell Information.
- Der Median geht nur auf die Ordnung der Beobachtungen und nicht auf die Abstände ein, der Modus gibt nur die am stärksten vertretende Ausprägung an.
- Anschaulich gesprochen ist der Median der mittlere Wert, was oft umgangssprachlich auch als Mittelwert bezeichnet wird. Vorsicht bei nicht statistischen 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.

Beispiel: Statistikbücher. Häufigkeitsverteilung und zur graphischen Veranschaulichung ein maßstabtreues „Pseudostabdiagramm“:

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

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}

$$\bar{x} = 3.57$$

$$x_{med} = 3$$

$$x_{mod} = 2$$

$$\bar{x} = 5$$

$$x_{med} = 5$$

$$x_{mod} = 5$$

$$\bar{x} = 6.43$$

$$x_{med} = 7$$

$$x_{mod} = 8$$

Exkurs: Lagemaße als Lösung eines Optimierungsproblems

Alternative Möglichkeit, Lagemaße zu begründen, die später in der Regressionsanalyse verallgemeinert wird.

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines intervallskalierten Merkmals X, die zu einer Zahl a^* zusammengefasst werden soll. Man könnte sagen, das beste a^* ist dasjenige, das so gewählt wird, dass der Gesamtabstand zwischen a^* und den Daten minimal wird. Misst man den Abstand

quadratisch	so ergibt sich für a^*
linear durch den Absolutbetrag	so ergibt sich für a^*

Für alle anderen $a \in \mathbb{R}$ gilt:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2,$$

$$\sum_{i=1}^{n} \mid x_i - x_{med} \mid \leq \sum_{i=1}^{n} |x_i - a|.$$

3.1.5 Geometrisches Mittel

Es gibt Fälle, bei denen das arithmetische Mittel selbst bei intervallskalierten Merkmalen nicht angemessen ist, zum Beispiel für Wachstumsraten oder Geschwindigkeiten.

Sei $\Omega = \{0, \ldots, n\}$ eine Menge von Zeitpunkten und $B(i) =: b_i$ ein zum Zeitpunkt i erhobenes Merkmal, z.B. das Bruttosozialprodukt.

Für $i = 1, \ldots, n$ heißt

$$x_i = \frac{b_i}{b_{i-1}}$$

der i-te *Wachstumsfaktor* und

$$r_i = \frac{b_i - b_{i-1}}{b_{i-1}} = x_i - 1$$

die i-te *Wachstumsrate*.

Dann bezeichnet man

$$\bar{x}_{geom} := \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$

als das *geometrische Mittel der Wachstumsfaktoren* $x_1, \ldots, x_n$.

Beispiel: Wirtschaftwachstum gemessen zu drei Zeitpunkten.

Geometrisches Mittel der Wachstumsfaktoren:

$$\bar{x}_{geom} =$$

Bemerkungen:

- Es gilt

$$b_n = b_0 \cdot (\bar{x}_{geom})^n$$

d.h. $\bar{x}_{geom}$ ist tatsächlich ein durchschnittlicher Wachstumsfaktor, also derjenige Wert, der sich aus b_n und b_0 ergäbe, wenn zu allen Zeitpunkten konstantes Wachstum geherrscht hätte. Im Beispiel gilt in der Tat:

- Das geometrische Mittel kann auch zur Prognose (unter der Stabilitätsannahme, dass das durchschnittliches Wachstum gleich bleibt) verwendet werden:

$$b_{n+q} = b_n \cdot (\bar{x}_{geom})^q, \quad q \in \mathbb{N}.$$

- Logarithmieren liefert:

$$\ln \bar{x}_{geom} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i.$$

  Das geometrische Mittel ist also ein arithmetisches Mittel auf der logarithmierten Skala.

- Man kann zeigen:

$$\bar{x}_{geom} \leq \bar{x}$$

  i.A. würde also die Angabe von $\bar{x}$ erhöhte Wachstumsraten vortäuschen.

3.1.6 Harmonisches Mittel

Beispiel: Die Entfernung von A nach B sei 99 km. Herr K. humpelt von A nach B mit konstant 1 km/h und fährt zurück mit konstant 99 km/h. Wie groß ist seine Durchschnittsgeschwindigkeit?

Naive Lösung: 50 km/h.

Allgemein: Sei $x_1, \ldots, x_n$ mit $x_i \neq 0$ für alle i die Urliste eines verhältnisskalierten Merkmals X. Dann heißt

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

das *harmonische* Mittel der $x_1, \ldots, x_n$.