3 Lage- und Streuungsmaße

- Häufigkeitsverteilungen geben die „quantitative Information“ von Merkmalen vollständig wieder. Oft will man charakteristische Aspekte von Verteilungen durch eine einzelne Zahl charakterisieren (Informationsverdichtung/-reduktion, einfacher Vergleich von Verteilungen).

- 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.

- Jetzt Quantifizierung / Charakterisierung

- Im Folgenden zunächst: Maßzahlen zur Beschreibung von Lage und Streuung durch $eine$ Zahl.

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

3.1 Arithmetisches Mittel und Varianz

3.1.1 Arithmetisches Mittel: Grundlegendes

Beachte: 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.

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 $x_1, \ldots, x_n$.

Bem. 3.2.

- 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.

Bsp. 3.3. *Anzahl von Statistikbüchern, die Studierende besitzen (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:

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

Weitergehend kann man Daten charakterisieren, indem man spezifische Subgruppen bildet und die artithmetischen Mittel in diesen Subgruppen tabellarisch gegenüberstellt.

Bsp. 3.4. *Einfacher Tabellenmietspiegel*

durchschnittliche 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)

Bsp. 3.5. *Augenfarbe*

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

$$\bar{x} =$$

Bem. 3.6.

- 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.

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 1 kg hinlegen.

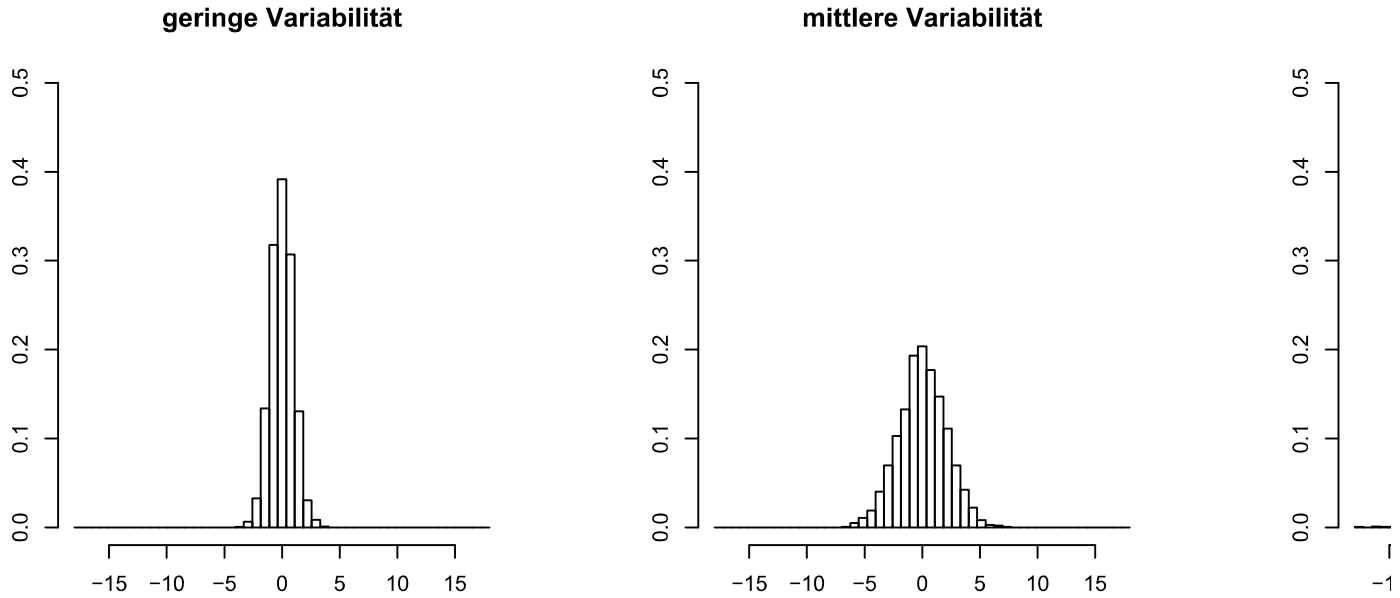
Die Schwerpunktseigenschaft macht auch deutlich: außerordentliche Hebelwirkung extrem großer und kleiner Werte: (lässt man die Beobachtung 12 im Beispiel weg, dann gilt: $\bar{x} = \frac{13}{9} = 1.44$).
Man muss den Effekt kennen, ob er gewünscht ist, oder nicht, ist eine inhaltliche Frage.

- Insbesondere ist damit das arithmetische Mittel sehr $ausreißeranfällig$, 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 dann oft den Median (s.u.).

3.1.2 Varianz und Standardabweichung: Grundlegendes

Eine Verteilung ist durch die Angabe von einem oder mehreren Lagemaßen nur unzureichend beschrieben.

Bsp. 3.7. *Häufigkeitsverteilungen mit gleicher zentraler Tendenz*



Streuungsmaße beantworten Fragen wie

- Wie groß ist die durchschnittliche Abweichung vom Mittelwert?
- Über welchen Bereich erstrecken sich die Beobachtungen?
- Wie stark schwanken die Beobachtungen?

Bem. 3.8.

Von Streuung im eigentlichen Sinne kann man nur bei mindestens intervallskalierten Daten sprechen, da nur dort Abstände interpretierbar sind. (Es gibt verschiedene Versuche, ein analoges Konzept für ordinal skalierte Daten zu definieren, aber bisher hat sich keine dieser Definitionen durchgesetzt.)

Varianz: Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißen

$$\tilde{s}_X^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *(empirische) Varianz* oder *Stichprobenvarianz* und

$$\tilde{s}_X := \sqrt{\tilde{s}_X^2}$$

die *(empirische) Streuung, Stichprobenstreuung* oder *(empirische) Standardabweichung von* X.

Bem. 3.9.

- Vorsicht: Der Begriff Streuung wird in einem doppelten Sinne gebraucht: Allgemein als Phänomen generell („wir suchen nach Maßzahlen zur Beschreibung der Streuung der Daten“), andererseits als eine bestimmte Maßzahl für das Problem.
- Die Varianz misst die durchschnittliche quadratische Abweichung vom Mittelwert.
- Durch das Quadrieren tragen negative und positive Abweichungen vom Mittelwert gleichermaßen zur Varianz bei.
- Die Varianz besitzt im Vergleich zum Merkmal X die quadrierte Einheit. Sie ist daher unanschaulicher zu interpretieren als die Standardabweichung, besitzt aber andererseits viele mathematische Vorzüge, wie sich später zeigen wird (siehe etwa Satz 3.21).
  Die Standardabweichung dagegen wird in der gleichen Einheit gemessen wie X.

- Sind die Ausprägungen $a_1, \ldots, a_k$ mit (relativer) Häufigkeitsverteilung $h_1, \ldots, h$ bzw. $f_1, \ldots, f_k$ gegeben, so gilt

$$\begin{aligned} \tilde{s}_X^2 &= \frac{1}{n} \sum_{j=1}^{k} h_j (a_j - \bar{x})^2 = \\ &= \sum_{j=1}^{k} f_j (a_j - \bar{x})^2. \end{aligned}$$

- Ist aus dem Kontext klar ersichtlich welches Merkmal betrachtet wird, so lässt man das X in der Notation auch häufig weg, schreibt also einfach $\tilde{s}^2$ und $\tilde{s}$.

Bsp. 3.10. *Statistikbücher*

Ausprägungen	h_j
0	2
1	2
2	4
3	1
12	1
$\sum$	10

Berechnung der Varianz über die ursprüngliche Formel:

$$\tilde{s}^2 \quad =$$

Berechnung über die Häufigkeitsverteilung:

$$\tilde{s}_X^2 \quad =$$

Standardabweichung:

$$\tilde{s} =$$

Verschiebungssatz: Es gilt

$$\begin{aligned}
\tilde{s}_X^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{x^2} - (\bar{x})^2, \\
&= \left(\frac{1}{n}\sum_{j=1}^{k} (a_j^2) \cdot h_j\right) - \left(\frac{1}{n}\sum_{j=1}^{k} a_j \cdot h_j\right)^2 \\
&= \sum_{j=1}^{k} (a_j^2) \cdot f_j - \left(\sum_{j=1}^{k} a_j \cdot f_j\right)^2
\end{aligned}$$

Achtung (sehr häufige Fehlerquelle):

Der Verschiebungssatz ist sehr bequem zum Berechnen der Varianz, es können aber beim Verwenden von Taschenrechnern bei sehr großen Ausprägungen starke Rundungsfehler auftreten, die das Ergebnis eventuell verfälschen. Für Aufgaben von Klausurlänge ist es aber meist geschickt, den Verschiebungssatz zu verwenden!

Bsp. 3.11. *Statistikbücher*

Berechne die empirische Varianz mit Hilfe des Verschiebungssatzes.

$$\tilde{s}_X^2 =$$

$$\tilde{s}_X =$$

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

Korrigierte empirische Varianz:

Neben der empirischen Varianz existiert noch eine alternative Definition der Varianz, die sog. *korrigierte (empirische) Varianz*:

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

$$s_X^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *korrigierte empirische Varianz* oder *korrigierte Stichprobenvarianz von* X.

- Der Sinn des Vorfaktors $\frac{1}{n-1}$, also der Begriff „korrigierte (empirische) Varianz“ wird erst in Statistik II deutlich: s_X^2 hat inferenz-theoretisch schönere Eigenschaften als $\tilde{s}_X^2$.
- Für großen Stichprobenumfang n nähern sich s_X^2 und $\tilde{s}_X^2$ an, weil dann $n - 1 \approx n$.

3.1.3 Arithmetisches Mittel und Varianz unter (linearen) Transformationen

Die Intervallskala erlaubt lineare Transformationen der Form $aX + b$, 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 $y_i = ax + b$, $i = 1, \ldots, n$. Wie verändert sich das arithmetische Mittel bei diesen oder allgemeineren Transformationen?

Bsp. 3.12.

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

- Nichtlineare Transformation

Satz 3.13. [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 für reelle Konstanten a, b das (linear transformierte) Merkmal $Y = a \cdot X + b$ und die zugehörigen Ausprägungen $y_1, \ldots, y$. Dann gilt für das arithmetische Mittel $\bar{y}$ von Y:

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

Beweis:

Bem. 3.14.

- 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).

Varianz unter Transformationen: Wie ändert sich die Varianz bei (linearer) Transformation eines Merkmals?

Satz 3.15.

Sei $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X und $y_1, \ldots, y$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$. Dann gilt

Bem. 3.16.

- Eine spezielle Transformation, die sogenannte $Standardisierung$, ist der Übergang zum Merkmal Z („Z-Score“) mit

$$z_i := \frac{x_i - \bar{x}}{\tilde{s}_X}.$$

  Z besitzt arithmetisches Mittel 0 und (empirische) Varianz 1. Man erzeugt damit in gewisser Weise eine natürlich Skala.

  Begründung:

3.1.4 Das arithmetische Mittel bei gruppierten Daten

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

Bsp. 3.17. *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 3.18.

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 $h'_\ell,\ \ell = 1, \ldots k$, als absoluter Häufigkeit der $\ell-$ten Klasse, f'_ℓ als zugehöriger relativer Häufigkeit und $m_\ell := \frac{c_\ell + c_{\ell-}}{2}$ als der jeweiligen Klassenmitte definiert man als *arithmetisches Mittel für gruppierte Daten*

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

Im Beispiel:

Bem. 3.19.

- 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. Eventuell entsteht durch die Gruppierung ein deutlicher Informationsverlust. Die Ungenauigkeit wird aber durch Angabe eines vermeintlich präzisen Wertes verdeckt.

- $\bar{x}_{gruppe}$ 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

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

- Im Allgemeinen ist $\bar{x}_{gruppe}$ natürlich nur eine grobe Approximation an den „echten“, d.h. auf ungruppierten Daten beruhenden, Mittelwert. Deshalb ist extreme Vorsicht bei „knappen“ Vergleichen zweier Gesamtheiten geboten.

* Eigentlich kann man nur mit Sicherheit folgende Abschätzung geben: Jede Einheit in der ℓ-ten Gruppe hat eine Ausprägung von mindestens $c_{\ell-1}$ und höchstens c_ℓ. Damit ergibt sich als Abschätzung für das arithmetische Mittel

$$\bar{x}_{unten} := \frac{1}{n}\sum_{\ell=1}^{k} h'_\ell c_{\ell-1} \leq \bar{x} \leq \frac{1}{n}\sum_{\ell=1}^{k} h'_\ell c_\ell =: \bar{x}_{oben}$$

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

Ein gesicherter Vergleich zweier Gesamtheiten ist dann und nur dann möglich, wenn $\bar{x}_{unten}$ einer Gesamtheit kleiner ist als $\bar{x}_{oben}$ einer anderen Gesamtheit.

- Sind die ungruppierten Daten auch erhältlich, so ist $\bar{x}$ vorzuziehen, da jede Gruppierung Informationsverlust mit sich bringt.
- Denoch werden, wie bereits diskutiert, oft nur gruppierte Daten erhoben, da sie leichter (und oft wahrheitsgetreuer) ermittelbar sind.

3.1.5 Arithmetisches Mittel und Varianz unter geschichteten Daten

Insbesondere bei Tertiäranalysen hat man häufig nicht die Urliste zur Verfügung, sondern nur Mittelwerte $\bar{x}^{(\ell)}$ in einzelnen Schichten $\ell = 1, \ldots, z$, in die die Grundgesamtheit zerlegt ist. (Man denke ferner an geschichtete Stichproben, wie sie in Kapitel 1 erwähnt wurden.)

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

Schicht	$1, \ldots, \ell, \ldots, z$	
Besetzungszahlen	$n^{(1)}, \ldots, n^{(\ell)}, \ldots, n^{(z)};$	$\sum_{\ell=1}^{z} n^{(\ell)} = n$
Mittelwerte	$\bar{x}^{(1)}, \ldots, \bar{x}^{(\ell)}, \ldots, \bar{x}^{(z)}$	
Varianzen	$\tilde{s}^{2^{(1)}}, \ldots, \tilde{s}^{2^{(\ell)}}, \ldots, \tilde{s}^{2^{(z)}}$	

Satz 3.20. *Betrachtet werde ein (mindestens) intervallskaliertes Merkmal X mit Urliste $x_1, \ldots, x_n$, die schichtweise zusammengefasst sei. Sei für $\ell = 1, \ldots, z$ jeweils $x_1^{(\ell)}, \ldots, x_{n^{(\ell)}}^{(\ell)}$ die Urliste in Schicht ℓ und ist $\bar{x}^{(\ell)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(\ell)}} x_i^{(\ell)}$ der Mittelwert in der ℓ-ten Schicht, $\ell = 1, \ldots, z$, so gilt für das arithmetische Mittel $\bar{x}$ von X:*

$$\bar{x} = \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \bar{x}^{(\ell)}$$

Beweis:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \left(\sum_{\ell=1}^{z} \sum_{i=1}^{n^{(\ell)}} x_i^{(\ell)} \right) \\ &= \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \bar{x}^{(\ell)} \end{aligned}$$

Varianzzerlegung / Streuungszerlegung bei geschichteten Daten:

Satz 3.21.

Für das arithmetische Mittel gilt

$$\bar{x} = \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \bar{x}^{(\ell)}.$$

Beobachtet werde die Situation von Satz 3.20. Mit $\tilde{s}^{2^{(1)}}, \ldots, \tilde{s}^{2^{(\ell)}}, \ldots, \tilde{s}^{2^{(z)}}$ als empirische Varianzen der Merkmalsausprägungen der zur jeweiligen Schicht gehörenden Einheiten definiere man

$$\tilde{s}^2_{innerhalb} := \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \tilde{s}^{2^{(\ell)}}$$

sowie

$$\tilde{s}^2_{zwischen} := \frac{1}{n} \sum_{\ell=1}^{z} n^{(l)} (\bar{x}^{(\ell)} - \bar{x})^2.$$

Dann gilt für die empirische Varianz S^2 des Merkmals X:

Bem. 3.22.

- Im Detail gilt also mit den Urlisten $\{x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_n^{(\ell)}\}$ in Schicht $\ell, \ell = 1, \ldots, z$,

$$\frac{1}{n}\sum_{\ell=1}^{z}(\sum_{i=1}^{n^{(\ell)}}(x_i^{(\ell)} - \bar{x})^2) = \underbrace{\frac{1}{n}\sum_{\ell=1}^{z}\sum_{i=1}^{n^{(\ell)}}(x_i^{(\ell)} - \bar{x}^{(\ell)})^2} + \underbrace{\frac{1}{n}\sum_{\ell=1}^{z} n^{(\ell)}(\bar{x}^{(\ell)} - \bar{x})^2}.$$

- Diese Zerlegungsmöglichkeit gilt *nur für Varianzen*, nicht aber für andere Streuungsmaße. Letztendlich ist sie der Grund für die Beliebtheit der Varianz – trotz anderer Unannehmlichkeiten. Deshalb sollte man eigentlich eher von der Varianzzerlegung als von der Streuungszerlegung sprechen.

- Die Streuungszerlegung entspricht der sogenannten Varianzanalyse, die gegen Ende der Vorlesungen zu Statistik I und Statistik II nochmals aus einer anderen Perspektive aufgegriffen wird. Im Prinzip lässt sich mit ihr untersuchen, wie stark sich die Mittelwerte in verschiedenen Schichten (hier auch in der Literatur als „Gruppen“ bezeichnet) unterscheiden und damit, wie stark der Einfluss der “schichtbildenden“ Variable auf das untersuchte Merkmal ist.

- Bei vielen Verfahren werden verallgemeinerte Formen einer Streuungszerlegung betrachtet; dies ist ein ganz grundlegendes Prinzip in der Statistik: Ein Modell ist umso besser, je mehr es an der Gesamtstreuung “erklärt“.

- Man kann die Wichtigkeit (Erklärungskraft) der schichtbildenden Variable durch folgende Überlegung bewerten: je größer $\tilde{s}^2_{zwischen}$ im Vergleich zu $\tilde{s}^2$ bzw. $\tilde{s}^2_{inn}$ ist, desto „mehr Variation“ wird durch die Schichtungsvariable „erklärt“, desto größer sind also die Mittelwertsunterschiede.

- Interpretation anhand des Beispiels mit den Einkommen der einzelnen Bundesländer.

3.2 Median & Quantile

3.2.1 Median

- 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 3.23.

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.

Bsp. 3.24. *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%

3.2.2 Quantile

Definition 3.25. *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*

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

heißt $\alpha \cdot 100\%$*-Quantil.*

Spezielle Quantile:

- Median: $x_{0.5} = x_{med}$.
- Quartile: $x_{0.25}$ ("unteres Quantil"), $x_{0.75}$ ("oberes Quantil").
- Dezile: $x_{0.1}$, $x_{0.2}$, . . . , $x_{0.8}$, $x_{0.9}$.

Bsp. 3.26. *Klausurnoten*

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

Bem. 3.27.

Alternative Definition des Medians: z.B. Fahrmeir et al., 2010 definieren den Median direkt über die $geordnete$ Urliste, also mit $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

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

Eine ähnliche Definition ist auch für andere Quantile möglich.

Allgemeine Regel zur Bestimmung des α-Quantils:

$\alpha \cdot n$ ist ganzzahlig	$x_\alpha \in [x_{(\alpha \cdot n)}; x_{(\alpha \cdot n+1)}]$
	bzw. zum Zeichnen: $x_\alpha = \frac{1}{2}(x_{(\alpha \cdot n)} + x_{(\alpha \cdot n+1)})$
$\alpha \cdot n$ ist nicht ganzzahlig	$x_\alpha = x_{(\lceil \alpha \cdot n \rceil)}$
	($\lceil x \rceil$ heißt kleinste ganze Zahl größer oder gleich x)

* 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. Hier wird sie vor allem bei der Zeichnung von Boxplots (s. Kapitel 3.7) verwendet.

* Andererseits können in gewissen Grenzfällen Quantile im Sinne der ursprünglichen Definition 3.25 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 $x_{\left(\frac{n}{2}\right)} = x_{\left(\frac{n}{2}+1\right)}$

* Für die Anwendung ist ganz wichtig: 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:
  Man beachte aber, dass es sich bei der Interpolation um eine Approximation handelt und damit auch der so ermittelte Median nur eine Approximation darstellt.

3.2.3 Verhalten unter Transformationen:

Wie ändert sich der Median bei Transformation der Daten?

Satz 3.28.

Sei $x_1, x_2, \ldots, x_n$ die Urliste eines (mindestens) ordinalskalierten Merkmals X und g eine monotone Funktion.
i) Ist x_{med} ein Median von X, so gilt mit $y_1 = g(x_1), \ldots, y_n = g(x_n)$ als Urliste des Merkmals $Y = g(X)$:

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

ist ein Median von Y.

ii) Fordert man zusätzlich, dass $g(\cdot)$ monoton steigend ist, so gilt die entsprechende Aussage für beliebige Quantile.

Bsp. 3.29.

- Geldnutzen: $Y = \ln X$ (implizit hier angenommen $X \geq 1$)

- Drei quadratische Zimmer

Korollar:
Bei gruppierten Daten gilt für alle $\alpha \in (0,1)$ und alle $\alpha-$Quantile x_α: Die Gruppe, in der x_α liegt, ist ein $\alpha-$Quantil für das gruppierte Merkmal X_{grupp}.

Bem. 3.30.

Man beachte, dass in obiger Situation geschichteter Gesamtheiten (vgl. Satz 3.20) eine korrekte Bestimmung des Gesamtmedians aus den Medianen in den einzelnen Schichten im Allgemeinen nicht möglich ist.

Bsp. 3.31. ***Ausführliches Beispiel*** *(nicht triviale Transformationen)*

Seien unter der unproblematischen Annahme. dass alle Größen ≥ 1 seien,

- X das steuerpflichtige Haushaltseinkommen
- V das steuerpflichtige Haushaltseinkommen umgerechnet in US-$,
- Y das Haushaltsnettoeinkommen,

- $W = \ln(Y)$ der 'Geldnutzen des Haushaltsnettoeinkommens' („Bernoullitransformation“, abnehmender Grenznutzen des Wirkens auf Lebensstile), und

- Z das Nettoäquivalenzeinkommen nach der OECD Skala

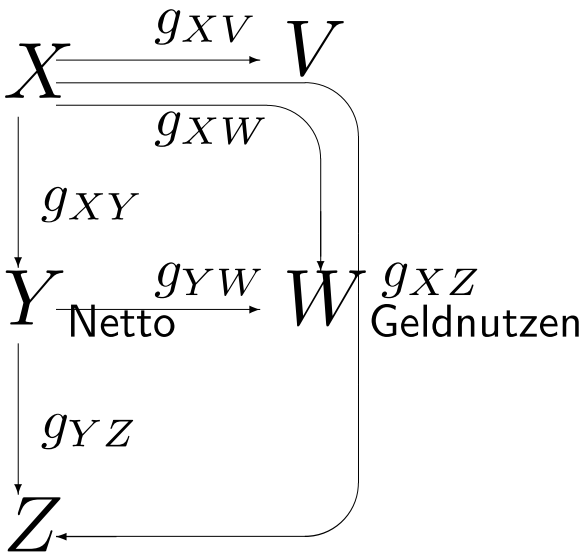
$$Z = \frac{Y}{korr}$$

mit

$$korr = 1 + 0.5 \cdot \text{Anzahl anderer HH Mitglieder mit Alter} \geq 14$$

$$+0.3 \cdot \text{Anzahl anderer HH Mitglieder mit Alter} < 14$$

Betrachtet werden folgende Transformationen unter den Variablen:



- Welche dieser Transformationen sind linear, welche streng monoton?
  Was lässt sich über die Beziehungen zwischen den arithmetischen Mitteln der Merkmale sagen, was über ihre Mediane und Quantile?

3.3 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 jede Ausprägung a_j mit der größten Häufigkeit h_j.

Definition 3.32.

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

$$h_{j^*} \geq h_j$$

Bem. 3.33.

- 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

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

3.4 Ein kurzer Vergleich der Lagemaße und einige weitere Bemerkungen

- Bei intervallskalierten Daten darf man auch den Modus oder den Median anwenden, man verschenkt (bei alleiniger Verwendung) aber meist viel 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, und wird deshalb oft umgangssprachlich auch als Mittelwert bezeichnet. Vorsicht bei nicht statistischen Veröffentlichungen! (Etwa Nachrichtenmeldungen im Rundfunk zum Armutsbericht.)
- Im Gegensatz zum arithmetischen Mittel sind Median und Modus unempfindlich gegenüber Ausreißern. Wird die größte Beobachtung verhundertfacht, so ändern sich Median und Modus nicht, das arithmetische Mittel reagiert dagegen stark.

Bsp. 3.34. *Statistikbücher*

Häufigkeitsverteilung und zur graphischen Veranschaulichung ein maßstabtreues „Pseudostabdiagramm“:

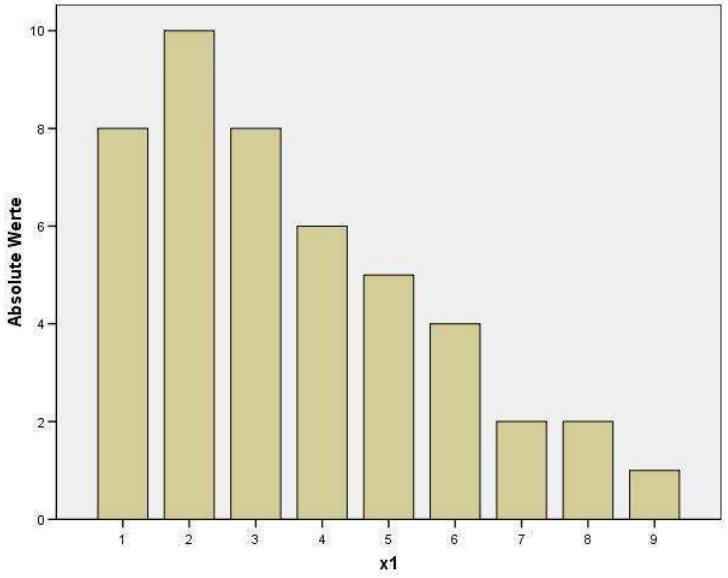
	Häufigkeiten
$a_1 = 0$	$h_1 = 2$
$a_2 = 1$	$h_2 = 2$
$a_3 = 2$	$h_3 = 4$
$a_4 = 3$	$h_4 = 1$
$a_5 = 12$	$h_5 = 1$

Bsp. 3.35. *Einkommensverteilung*

Generell ist bei der Betrachtung von Einkommen das arithmetische Mittel meist deutlich größer als der Median.

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

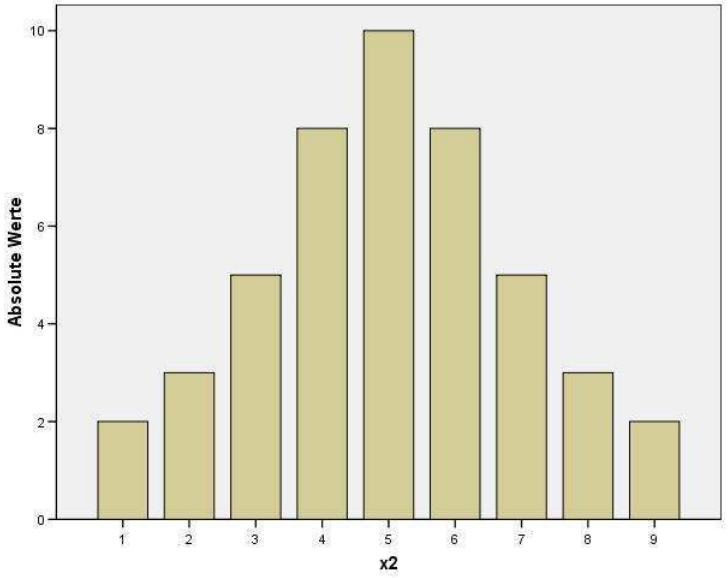
symmetrisch:	$\bar{x}$	$\approx$	x_{med}	$\approx$	x_{mod}
linkssteil:	$\bar{x}$	$>$	x_{med}	$>$	x_{mod}
rechtssteil:	$\bar{x}$	$<$	x_{med}	$<$	x_{mod}



$$\bar{x} = 3.57$$

$$x_{med} = 3$$

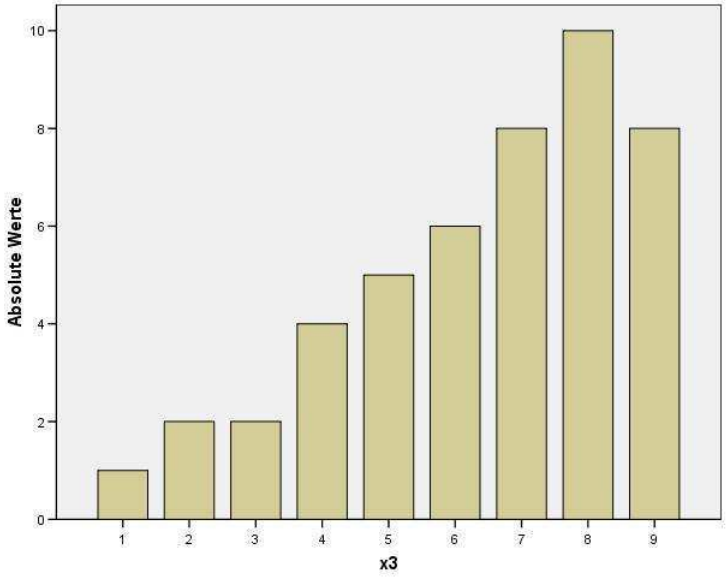
$$x_{mod} = 2$$



$$\bar{x} = 5$$

$$x_{med} = 5$$

$$x_{mod} = 5$$



$$\bar{x} = 6.43$$

$$x_{med} = 7$$

$$x_{mod} = 8$$

Exkurs: Lagemaße als Lösung eines Optimierungsproblems Alternative Möglichkeit, Lagemaße zu begründen, die später in der Regressionsanalyse verallgemeinert wird. (Typische statistische Sicht: Verfahren als in einem gewissen Sinn optimale Datenbeschreibung.)

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	$(x - a^*)^2$	so ergibt sich für a^* das arithmetische Mittel $\bar{x}$
linear durch den Absolutbetrag	$\lvert x - a^* \rvert$	so ergibt sich für a^* der Median x_{med}

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

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

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

Den Modus erhält man durch eine Grenzwertbetrachtung über die sogenannte Toleranzverlustfunktion. Mit

$$\mathbb{I}_{\{x \neq a\}} := \begin{cases} 1 & x_i \neq a \\ 0 & \text{sonst,} \end{cases}$$

ist für alle a

$$\sum_{i=1}^{n}\mathbb{I}_{\{x_i \neq x_{mod}\}} \leq \sum_{i=1}^{n}\mathbb{I}_{\{x_i \neq a\}}.$$

3.5 Geometrisches und harmonisches Mittel

3.5.1 Das geometrische Mittel

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

Definition 3.36.

Sei $\Omega = \{0, \ldots, n\}$ eine Menge von Zeitpunkten und $b_0, b_1, \cdots, b_n$ die Urliste eines Merkmals B mit $b_i := B(i) > 0$ für alle $i = 0, \cdots, n$ (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

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

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

Bsp. 3.37. *Wirtschaftwachstum gemessen zu drei Zeitpunkten*

Geometrisches Mittel der Wachstumsfaktoren:

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

Bem. 3.38.

- Es gilt

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

d.h. $\bar{x}_{geom}$ ist tatsächlich ein durchschnittlicher Wachstumsfaktor, also derjenige Wert, mit Hilfe dessen sich b_n aus 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:

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

  Da typischerweise $\bar{x}_{geom} \neq \bar{x}$, würde im Allgemeinen also die Angabe von $\bar{x}$ erhöhte Wachstumsraten vortäuschen.

3.5.2 Harmonisches Mittel

Bsp. 3.39.

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.

Definition 3.40.

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 $x_1, \ldots, x_n$.

3.6 Weitere Streuungsmaße

3.6.1 Variationskoeffizient:

Definition 3.41.

Ist $\bar{x} > 0$, so heißt die Größe

$$v_X := \frac{\tilde{s}_X}{\bar{x}}$$

Variationskoeffizient des Merkmals X.

Bem. 3.42.

- Gemessen wird hier die Streuung relativ zum Mittelwert. Insbesondere ist v_X dimensionslos.
- Der Variationskoeffizient erlaubt damit beispielsweise auch den Vergleich der Streuung von Preisen, die in verschiedenen Währungen gemessen wurden.

3.6.2 Inter-Quartils-Abstand:

Sind $x_{0.25}$ und $x_{0.75}$ das obere und das untere Quartil eines intervallskalierten Merkmals, so heißt

$$d_{QX} := x_{0.75} - x_{0.25}$$

der *Interquartilsabstand*.

3.6.3 Median-Absolute-Deviation:

Der Median der Werte $|x_i - x_{med}|$, $i = 1, \ldots, n$ heißt *Median-Absolute-Deviation* von X (MAD_X).

3.6.4 Spannweite:

Die Größe

$$R_X := x_{(n)} - x_{(1)}$$

heißt *Spannweite* von X.

Bem. 3.43.

- Alle betrachteten Streuungsmaße sind nur für (mindestens) intervallskalierte Merkmalse sinnvoll definiert, da sie auf Abständen (typischerweise dem Abstand der Beobachtungen zu einem Lagemaß) beruhen.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind die gebräuchlichsten Streuungsmaße.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind sehr empfindlich gegenüber Ausreißern! Das Gleiche gilt für die Spannweite R. Die Kennzahlen MAD und d_Q hingegen entstammen der sogenannten robusten Statistik, die sich um ausreißerresistente Methoden bemüht.
- Gilt $x_1 = x_2 = \ldots = x_n$, so weisen alle Streungsmaße den Wert 0 auf. Mit Ausnahme von d_Q gilt auch die Umkehrung: Sind die Steuungsmaße (außer eben d_Q) $= 0$, so sind alle Werte der Urliste gleich.

- Nochmals der Hinweis: Eine häufige Ursache für Verwirrung und Missverständnisse liegt in der Tatsache, dass der Begriff „Streuung“ in der Statistik in einem doppelten Sinn gebraucht wird:
  - in einem allgemeinen Sinn: Streuung als Phänomen („Die Daten streuen stark“).
  - in einem speziellen Sinn: als $eine$ Maßzahl für dieses Phänomen.

Bsp. 3.44. *Statistikbücher*

Man berechne den Variationskoeffizienten, den Interquartilsabstand und die Spannweite.

Ausprägungen	h_j
0	2
1	2
2	4
3	1
12	1
$\sum$	10

3.7 Box-Plot

Ziele:

- einfache Darstellung der Häufigkeitsverteilung von (mindestens) intervallskalierten Merkmal und ihrer Kennzahlen
- Identifikation von potentiellen Ausreißern
  $\Rightarrow$ nicht ausreißeranfällige Meßzahlen verwenden.

Idee:

0) zeichne einen Zahlenstrahl, der die Datenpunkte umfasst

i) markiere den Median (hier als eindeutig vorausgesetzt)

ii) symbolisiere Lage der „mittleren Werte“ durch eine Box
(optisch „passende Höhe“, die Ordinate ist ohne inhaltliche Bedeutung)

iii) wie weit reichen „weitere nicht atypische“ Werte?

iv) identifiziere potentielle Ausreißer: atypische (ungewöhnlich große, ungewöhnlich kleine) Werte, die genauerer Untersuchung bedürfen

zu ii) wähle die mittleren 50%: Die Box hat also Länge $dQ = x_{0.75} - x_{0.25}$

zu iii) als „nicht atypisch“ gelten alle Werte, die nicht weiter als $1.5dQ$ von der Box entfernt sind

Vorgehen: bestimme:

- $x_{0.25}, \quad x_{0.50}, \quad x_{0.75}$, in der Definitionsweise, die eindeutige Werte ergibt.
- Interquartilsabstand: $d_{QX} = x_{0.75} - x_{0.25}$
- Zäune z_u, z_o, die am kleinsten bzw. größten Datenpunkt im Bereich $x_{0.25} - 1.5 \cdot d \quad ;$ $x_{0.75} + 1.5 \cdot d_{QX}$ liegen.
- Ausserhalb der Zäune werden $alle$ Punkte eingezeichnet; sie sind ausreißerverdächtig.

- Vorsicht bei der Anwendung von Software! Vor allem außerhalb der Box sind auch andere Darstellungen üblich (z.B. Zäune immer bis $x_{(1)}$ und $x_{(n)}$).
  Toutenburg und Heumann (2009) beispielsweise unterscheidet zwischen Ausreißern ($1.5 \cdot d_{QX}$ bis $3 \cdot d_{QX}$ von Rändern der Box entfernt) und Extremwerten (mehr als $3 \cdot d_{QX}$ vom Rand entfernt).

- Oft wird der Median durch einen dicken Punkt ausgedrückt. Der Box-Plot gibt einen kompakten Überblick über die Form der Verteilung (Zentrale Tendenz, Variabilität, Schiefe, extreme Werte).

- Manchmal gibt es auch reduzierte Dartellungen; z. B. wird nur die Box gezeichnet.

Box-Plots können auch zum graphischen Vergleich von Verteilungen verwendet werden: