Gebräuchliche Bezeichnungen:	
• *Merkmal*: interessierendes Kennzeichen, Charakteristikum bzw. interessierende Größe; ein Merkmal ist dabei *qualitativ*: nicht durch Messen oder Zählen erfassbar *quantitativ*: durch Messen oder Zählen erfassbar	**a)** Wohnort; Haarfarbe **b)** Länge; Druckfehleranzahl
• *Merkmalsträger*: Objekte (Elemente), welche die in Frage kommenden Merkmale aufweisen	**a)** Schüler **b)** Bolzen; Buchseiten
• *Merkmalsausprägung* (*Beobachtungs-* bzw. *Merkmalswert*) x_i: eine konkrete Beschreibung oder Realisierung bei einem qualitativen bzw. ein konkretes Mess- oder Zählergebnis bei einem quantitativen Merkmal Bemerkung: Schulnoten, Güteklassen usw. sind trotz der Zahlenangaben *keine quantitativen sondern qualitative* Merkmale (die Zahlen dienen der Reihung)!	**a)** Linz; blond **b)** 56 mm; 2 Fehler
• *diskretes* quantitatives Merkmal: nur endlich viele, meist ganzzahlige Werte sind möglich (Zählergebnisse) *stetiges* quantitatives Merkmal: beliebige reelle Zahlen eines Intervalls können Beobachtungswerte sein (Messergebnisse)	Fehleranzahl; defekt – fehlerfrei Längen; Zeitdauer; Zugfestigkeit
• *Grundgesamtheit*: die Menge aller (vor allem: sehr vieler) Merkmalsträger mit dem interessierenden Merkmal	**a)** alle Schüler Oberösterreichs **b)** alle Bolzen einer gewissen Art; eines bestimmten Romans
• *Umfang N der Grundgesamtheit*: Anzahl der Elemente der Grundgesamtheit	
• *Stichprobe*: eine *zufällige* Auswahl von Merkmalsträgern aus der Grundgesamtheit *zufällig*: gleiche Auswahlchancen für alle Objekte (oft problematisch, die Chancen-Ungleichheit zu erkennen)	**a)** alle Volksschüler **b)** eine Bolzen-Tagesproduktion; alle Seiten des 3. Kapitels
• *Stichprobenumfang n*: Anzahl der Elemente der Stichprobe	
• *Urliste*: bezüglich der Merkmalsausprägungen noch ungeordnete Datenerhebung	**a)** Linz, Zwettl, Linz, Steyr ... **b)** 59 mm, 56 mm, 56 mm, 57 mm, ...
• *Strichliste*: bezüglich der Beobachtungsergebnisse geordnete Datenliste; jedes Auftreten einer bestimmten Ausprägung wird durch einen Strich registriert (Blockbildung ist dabei empfehlenswert)	**b)** 56 mm 𝍸 𝍸 || 57 mm 𝍸 𝍸 58 mm 𝍸 𝍸 𝍸 |
• *Absolute Häufigkeit* (bzw. *Besetzungszahl*) n_i: die Anzahl der Objekte mit der Merkmalsausprägung x_i	**b)** $x_i = 56\,\text{mm} \rightarrow n_i = 12$
• *Relative Häufigkeit* $h(x_i)$ bzw. h_i: der Anteil der Objekte mit der Ausprägung x_i am Gesamtumfang der Daten (wird oft in Prozent angegeben) Bemerkung: relative Häufigkeiten sind nie negativ und nie größer als eins: $0 \le h_i \le 1$	$\boxed{h(x_i) = h_i := \frac{n_i}{n}} \cdot 100\,\%$ $0 \le n_i \le n \quad \rightarrow \quad 0 \le \frac{n_i}{n} \le 1$
• *Absolute Summenhäufigkeit* (*kumulierte* Häufigkeit) N_k: Summe der Besetzungszahlen von n_1 bis n_k	$\boxed{N_k := \sum_{i=1}^{k} n_i}$
• *Relative Summenhäufigkeit* (*relative kumulierte* Häufigkeit) $H(x_k)$ bzw. H_k: Summe der relativen Häufigkeiten von h_1 bis h_k Bemerkung: Die relative Summenhäufigkeit ist eine von null bis eins monoton steigende Funktion.	$\boxed{H(x_k) = H_k := \sum_{i=1}^{k} h_i} \cdot 100\,\%$ $H_{k+1} = H_k + h_{k+1} \ge H_k$

Bei sehr vielen verschiedenen, möglichen oder tatsächlichen Ausprägungen eines quantitativen (stetigen) Merkmals wird das Intervall reeller Zahlen, in dem alle Messwerte liegen, zweckmäßigerweise in Teilintervalle (so genannte *Klassen*) unterteilt, die Daten werden *klassiert*. Dazu einige

Bemerkungen:

- Da bei stetigen Merkmalen gemessen wird, ist von vornherein eine Klassierung durch die Messgenauigkeit gegeben. Diese ist hier aber nicht gemeint.
- Im Allgemeinen erfolgt eine Klassierung ab einem Stichprobenumfang von $n \geq 50$, darunter ist eine Klasseneinteilung unüblich. Dabei ist es gleichgültig, ob dieser Umfang tatsächlich vorliegt oder ob dieser Umfang möglich ist (Bestimmung des Geburtsgewichts von Säuglingen; dabei könnten die Ergebnisse beispielsweise zwischen 800 g und 4500 g liegen, bei einer Genauigkeit von 1 g sind daher sehr viele Ausprägungen möglich).
- Klassenbildung schafft Übersichtlichkeit, gleichzeitig gehen aber Informationen verloren. Man wird im Einzelnen die Vor- und Nachteile gegeneinander abwägen.
- Die Anzahl der Klassen richtet sich nach dem (möglichen) Stichprobenumfang. In der Praxis üblich sind 7 bis 20 Klassen, als Faustregel gilt:

$$\text{Anzahl der Klassen} \approx \sqrt{n}$$

- Die Klassen sollten nach Möglichkeit gleich breit sein.

Empfohlene Vorgangsweise bei der Klassenbildung:

- Die Unterteilung muss so erfolgen, dass *jeder Messwert eindeutig einer Klasse zugeordnet* werden kann. Als Klassengrenzen wählt man daher zweckmäßigerweise Rundungsgrenzen der Messwerte, da diese „um eine Stelle genauer" als die Messergebnisse selbst sind.

  Messwerte sind Zahlen begrenzter Genauigkeit: 3,2 repräsentiert das Intervall [3,15; 3,25[
  Rundungsgrenzen
  Hinweis: In der Praxis wird auch „nichtmathematisch" gerundet (z.B. Zeitangabe bei Digitaluhren)!

- Die *Klassenweite* w errechnet man aus den Rundungsgrenzen ($_{o(u)}x_i$ ... obere (untere) Rundungsgrenze von x_i) des größten bzw. kleinsten Messwerts nach den Formeln

$$w = \frac{{}_{o}x_{max} - {}_{u}x_{min}}{\sqrt{n}} \quad \text{für} \quad 50 \leq n \leq 400$$

$$w = \frac{{}_{o}x_{max} - {}_{u}x_{min}}{20} \quad \text{für} \quad n > 400$$

w auf die Genauigkeit der Messwerte runden

  Beispiel:

  $x_{max} = 531\,\text{g}$, $x_{min} = 422\,\text{g}$, $n = 60$

  $\rightarrow \quad w = \dfrac{531{,}5\,\text{g} - 421{,}5\,\text{g}}{\sqrt{60}} \approx 14\,\text{g}$

- *Anzahl k der Klassen*: Durch Runden von $\dfrac{{}_{o}x_{max} - {}_{u}x_{min}}{w}$ auf die *nächstgrößere* ganze Zahl erhält man die Anzahl der Klassen.

  $\rightarrow \quad \dfrac{531{,}5\,\text{g} - 421{,}5\,\text{g}}{14\,\text{g}} \approx 7{,}9 \approx 8 = k$

- *Klassengrenzen*: Der linke Rand der ersten und der rechte Rand der letzten Klasse werden so gewählt,
  - dass die Klassenränder Rundungsgrenzen sind und
  - dass das von ihnen gebildete Intervall der „Länge" $k \cdot w$ das Messintervall möglichst symmetrisch umfasst.

  $k \cdot w = 8 \cdot 14\,\text{g} = 112\,\text{g}$

  Rand: $\dfrac{112\,\text{g} - (531\,\text{g} - 422\,\text{g})}{2} = 1{,}5\,\text{g}$

  $\rightarrow \begin{cases} \text{1. Klasse:} & [420{,}5\,\text{g}\,;\,434{,}5\,\text{g}[\\ \text{8. Klasse:} & [518{,}5\,\text{g}\,;\,532{,}5\,\text{g}[\end{cases}$

- Als *Repräsentanten* der Klassen werden die Klassenmitten gewählt. Sie dienen – als Stellvertreter aller Messwerte innerhalb einer Klasse – zur näherungsweisen Berechnung der Stichprobenkennwerte (*praktische* Kennwerte).

  427,5 g ist der Repräsentant der ersten Klasse (ist aber selbst gar kein Messwert).

Bemerkungen:

- Die Anzahl der Objekte einer Stichprobe, die innerhalb einer Klasse liegen, ergibt die absolute Häufigkeit dieser Klasse und damit das „Gewicht“ ihres Repräsentanten.
- Die Repräsentanten müssen selbst nicht unbedingt Messwerte sein.

Die Strichliste (zum vorigen Beispiel) hat z. B. folgendes Aussehen:

422 g | , 425 g ||| , 426 g | ,

432 g | , 435 g || , ...

Dann hat die 1. Klasse (und ihr Repräsentant) die absolute Häufigkeit $n_1 = 6$

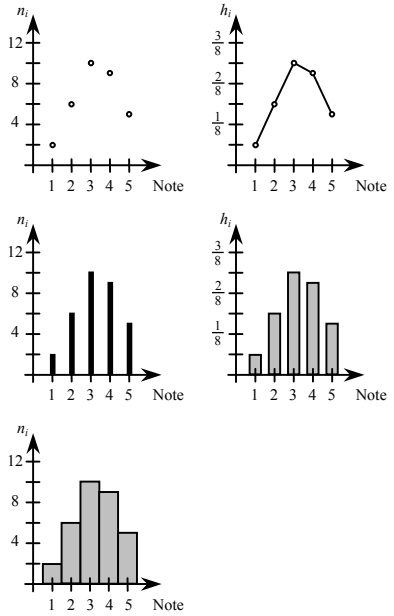
Graphische Darstellungen: Die Besetzungszahlen alleine vermitteln ein recht unanschauliches Bild von der *Häufigkeitsverteilung* einer Stichprobe. Besser geeignet dafür sind graphische Darstellungen.

- *Koordinatensystem*: Es kann nur bei einem Merkmal verwendet werden, das entweder quantitativ ist oder dessen Ausprägungen gereiht werden können. Dabei trägt man die absoluten oder relativen Häufigkeiten in Abhängigkeit von den Merkmalsausprägungen bzw. den Klassenmitten auf, und zwar als
  - Punkte (unüblich): *Punktdiagramm*; verbindet man die Punkte miteinander, so erhält man ein *Liniendiagramm* (*Häufigkeitspolygon*).
    Achtung: Letzteres täuscht bei diskreten Merkmalen Stetigkeit vor!
  - senkrechte Striche oder Balken, deren Längen den Häufigkeiten entsprechen: *Stab-* oder *Balkendiagramm*.
    Achtung: Bei Klassenbildung geht der Klasseneindruck verloren, stetige Merkmale erscheinen diskret!
  - aneinander gereihte Rechtecke, deren Flächen (bei gleichen Breiten der Rechtecke auch die Höhen) den Häufigkeiten entsprechen: *Histogramm* (*Staffelbild*); gut geeignet für klassierte Daten, die Rechtecksbreiten entsprechen dabei den Klassenweiten.
    Bemerkung: Die Gesamtfläche entspricht dem Stichprobenumfang n bei absoluten bzw. 1 bei relativen Häufigkeiten.

Beispiel:

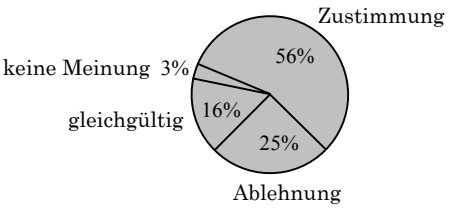
Notenverteilung einer Schularbeit:

Note	1	2	3	4	5
Anzahl	2	6	10	9	5



- *Kreisdiagramm*: Diese Darstellungsform wird vorzugsweise für qualitative Merkmale verwendet. Die Kreisfläche wird entsprechend den Häufigkeiten in Sektoren unterteilt (neben den Flächen entsprechen auch die Winkel den Häufigkeiten).

Ergebnis einer Meinungsumfrage:



- *Bildhafte Graphiken*, *Piktogramme*: Anhand von Bildern und Symbolen werden die Häufigkeiten qualitativer Merkmale meist durch die Größe (Länge, Fläche) und die Anzahl der verwendeten Objekte dargestellt.

Zustimmung:  56%

Ablehnung:  25%

gleichgültig:  16%

keine Meinung:  3%

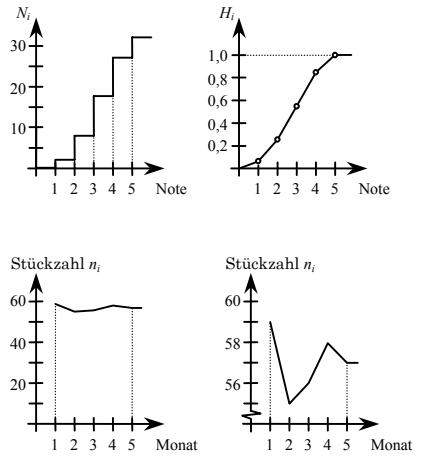
Bemerkungen:

- Neben der Darstellung von absoluter bzw. relativer Häufigkeit ist auch die Veranschaulichung der Summenhäufigkeit gebräuchlich: *Summenhäufigkeitsverteilung* (üblicherweise als Liniendiagramm oder Histogramm ausgeführt); es dient für Fragestellungen der Art „wann wurden so und so viel Prozent erreicht“, „wie viele Stücke liegen unterhalb der Toleranzgrenze“ usw.

  Bemerkung: Bei klassierten Daten erfolgen die Sprünge an den oberen Klassengrenzen.

- Die Aussagekraft einer graphischen Darstellung hängt stark vom gewählten Maßstab und der Lage der Koordinatenachsen ab! Es ist also Vorsicht geboten, sonst entsteht ein falscher Eindruck.

Beispiel: Notenverteilung



Um gleichartige Stichproben rasch miteinander vergleichen zu können (z. B. wöchentliche Umfrageergebnisse oder Proben einer laufenden Produktion), bedient man sich einiger Kenngrößen:

- *Lagekennwerte* (spiegeln die Häufigkeiten und die Größenordnung der Merkmalsausprägungen wider); wichtig ist dabei die Idee der *Mittelwerte*: Diese, nur für quantitative Merkmale erklärten Größen, stellen sozusagen ein „Kondensat“ der Messwerte dar und bilden bei entsprechender Anzahl einen für die Stichprobe charakteristischen Ersatz der (unterschiedlichen) Merkmalsausprägungen.
- *Streuungskennwerte* (beschreiben die Abweichungen der Merkmalsausprägungen voneinander)

Die wichtigsten Kennwerte der Lage: (k ... Anzahl der unterscheidbaren Merkmalsausprägungen)	Beispiel: Absolventenanzahl der E- und M-Klassen der letzten 6 Jahre	
a) Der (*arithmetische*) *Mittelwert* $\bar{x}$ (*arithmetisches Mittel*) wird gebildet, indem man die Summe aller Merkmalsausprägungen x_i durch den Stichprobenumfang n teilt: $\boxed{\bar{x} := \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{1}{n}\cdot\sum_{i=1}^{k} n_i\cdot x_i = \sum_{i=1}^{k} h_i\cdot x_i}$ Der Betrag des Mittelwerts kann (sollte) eine Dezimalstelle mehr als die Messwerte haben.	16, 17, 20, 15, 18, 18, 17, 18, 14, 17, 18, 13, 19, 18, 14 $\bar{E} = \frac{252}{15} = 16{,}8$	22, 23, 23, 22, 18, 24, 19, 21, 27, 22, 24, 24 $\bar{M} \approx 22{,}4$
b) Der *Median* $\tilde{x}$ (*Zentralwert*) ist der mittlere Wert bzw. das arithmetische Mittel der beiden mittleren Werte einer der Größe nach geordneten Stichprobe mit dem Umfang n: $\tilde{x} := x_{\frac{n+1}{2}}$ für n ungerade $\tilde{x} := \frac{1}{2}\cdot(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$ für n gerade	13, 14, 14, 15, 16, 17, 17, **17**, 18, 18, 18, 18, 18, 19, 20 $\tilde{E} = 17$	18, 19, 21, 22, 22, **22**, **23**, 23, 24, 24, 24, 27 $\tilde{M} = \frac{22+23}{2} = 22{,}5$
c) Gibt es genau eine häufigste Ausprägung – gleichgültig, ob ein qualitatives oder quantitatives Merkmal vorliegt – so bezeichnet man diese als *Modus* $\hat{x}$ (*Modalwert*) der Stichprobe.	$\hat{E} = 18$	$\hat{M}$ existiert nicht, weil die Zahlen 22 und 24 gleich häufig sind.

Bemerkungen:

- „Schwerpunkteigenschaft“ des arithmetischen Mittelwerts: Die Summe aller Differenzen zwischen den Merkmalswerten und dem Mittelwert ist null:

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

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

- Vorteilhafte Ermittlung des Mittelwerts (wenn „händisch“ gerechnet werden muss): Durch die Annahme eines vorläufigen Mittelwerts x_v (meist der Modalwert) lassen sich große Zahlen vermeiden; für den Mittelwert gilt dann:

  $$\bar{x} = x_v + \tfrac{1}{n} \cdot \sum_{i=1}^{n}(x_i - x_v)$$

  $$x_v + \tfrac{1}{n}\sum_{i=1}^{n}(x_i - x_v) = x_v + \underbrace{\tfrac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}} - \underbrace{\tfrac{1}{n}\sum_{i=1}^{n} x_v}_{x_v}$$

  z. B. $x_i = 8; 11; 11; 15; 15; 15; 15; 17; 19$

  $\rightarrow \quad \bar{x} = 15 + \tfrac{1}{9} \cdot (-7 + 2 \cdot (-4) + 2 + 4) = 14$

- Der Mittelwert wird durch *Ausreißer* stark beeinflusst. In diesen Fällen wird die Lage durch den Median meist besser beschrieben.

  Ausreißer: einzelner Wert, der stark von den anderen Werten abweicht.

  z. B. $x_i = 1; 21; 21; 22; 22; 22; 23; 23; 24$

  $\rightarrow \quad \bar{x} = \tfrac{1}{9} \cdot (1 + 2 \cdot 21 + ... + 24) \approx 19{,}9$

  statt $\bar{x} \approx 22{,}3$ Median: $\tilde{x} = 22$

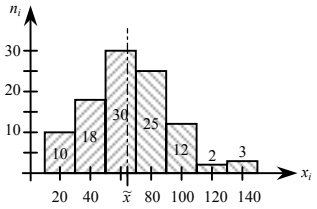
- Die Anwendung der Mittelwertbildung auf qualitative Merkmale, die gereiht werden können (wie Schulnoten oder Güteklassen), ist eigentlich nicht zulässig: Die Zuordnung von verbalen Merkmalsbeschreibungen und Zahlen ist willkürlich, die Wahl der Zahlenskala beeinflusst aber die Mittelwertbildung entscheidend. Darüber hinaus muss der Mittelwert wieder rückinterpretiert werden.

  Die Beurteilungen „Nicht genügend – Gut – Gut“ würden bei der üblichen Notenskala ein „Befriedigend“ als Mischnote ergeben, nach der Skala: „Sehr gut“(10) – „Gut“(6) – „Befriedigend“(5) – „Genügend“(4) – „Nicht genügend“(0) aber nur ein „Genügend“.

- Bei klassierten Daten erfolgt die Bestimmung der Lagekennwerte sehr oft näherungsweise:

  - Für die Mittelwertberechnung werden die Repräsentanten genommen,
  - als Median dient der Wert, der die Gesamtfläche des Histogramms für die Häufigkeiten halbiert (der Median teilt ja jede (geordnete) Stichprobe in zwei gleich große Teilmengen) und
  - als Modalwert nimmt man meistens den Repräsentanten jener Klasse, welche die größte Häufigkeit aufweist.

  Beispiel:

  $$\bar{x} \approx \frac{10 \cdot 20 + 18 \cdot 40 + ... + 3 \cdot 140}{100} = 65{,}8$$

  $$\tilde{x} \approx 50 + 20 \cdot \frac{50 - (10 + 18)}{30} \approx 64{,}7$$

  $$\hat{x} \approx 60$$

- Das arithmetische Mittel ist aber nicht immer zur Charakterisierung der Lage geeignet, wie nebenstehendes Beispiel jährlicher Umsatzsteigerungen zeigt (Umsatzsteigerungen werden immer auf das jeweils vergangene Jahr bezogen). Deshalb kommen auch andere Mittelwerte, wenn auch nicht sehr oft, zur Anwendung.

  Beispiel: Der Umsatz erhöhte sich in drei aufeinander folgenden Jahren um 10%, 20% bzw. 30%, insgesamt um 71,6% ($1{,}1 \cdot 1{,}2 \cdot 1{,}3 - 1 = 0{,}716$).
  Würde man mit einer mittleren Steigerung von 20% rechnen, so ergäbe sich ein Zuwachs von 72,8%.

Weitere Mittelwerte:
(n ... Stichprobenumfang, k ... Anzahl der unterscheidbaren Merkmalsausprägungen)

a) Den *gewichteten Mittelwert* (*gewogenes arithmetisches Mittel*) erhält man, wenn die Merkmalsausprägungen x_i durch eine *Gewichtsfunktion* $g(x_i)$ verschieden stark bewertet werden:

$$\boxed{\bar{x} := \frac{\sum_{i=1}^{n} g(x_i)\cdot x_i}{\sum_{i=1}^{n} g(x_i)} = \frac{\sum_{i=1}^{k} h_i \cdot g(x_i)\cdot x_i}{\sum_{i=1}^{k} h_i \cdot g(x_i)}}$$

(Anmerkung: Die absoluten bzw. relativen Häufigkeiten stellen selbst bereits eine Gewichtung dar.)

Beispiel: Beim Testen von Bildschirmen werden Bildqualität, Technik, Ausstattung, Handhabung und Bedienungsanleitung mit jeweils zehn Punkten bewertet. Daraus wird ein gewichteter Mittelwert gebildet (für Bildschirme ist z. B. die Bildqualität von größerer Bedeutung als die Bedienungsanleitung).

b) Das *geometrische Mittel* $\bar{x}^g$ von positiven Merkmalswerten x_i ist die n-te Wurzel aus deren Produkt:

$$\boxed{\bar{x}^g := \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = x_1^{h_1} \cdot x_2^{h_2} \cdot \ldots \cdot x_k^{h_k} = e^{\overline{\ln x_i}}} \qquad x_i > 0$$

Anwendung: Werden Größen (z. B. Jahresumsatz, Einwohnerzahl) fortlaufend aufeinander bezogen (relative Größen), so ist als Mittelwert dieser Verhältniszahlen das geometrische statt des arithmetischen Mittels zu nehmen:

$$x_1 = \frac{a_1}{a_0};\ x_2 = \frac{a_2}{a_1};\ x_3 = \frac{a_3}{a_2} \ \rightarrow\ x_1 \cdot x_2 \cdot x_3 = \frac{\cancel{a_1}}{a_0} \cdot \frac{\cancel{a_2}}{\cancel{a_1}} \cdot \frac{a_3}{\cancel{a_2}} = \frac{a_3}{a_0} = (\bar{x}^g)^3$$

Beispiel: Die Produktion sank im ersten Quartal um 20%, stieg aber in den nächsten Quartalen um 60%, 20% und 35%, bezogen jeweils auf das vorige Quartal. Welche mittlere Steigerung pro Quartal ergibt dies?

$x_i = 0{,}80\ ;\ 1{,}60\ ;\ 1{,}20\ ;\ 1{,}35$

$\rightarrow\ \bar{x}^g = \sqrt[4]{0{,}80 \cdot 1{,}60 \cdot 1{,}20 \cdot 1{,}35} = 1{,}20$

$\rightarrow$ mittlere Steigerung: 20%

c) Der Kehrwert des *harmonischen Mittels* $\bar{x}^h$ der von null verschiedenen Merkmalswerte x_i ist das arithmetische Mittel der Merkmalskehrwerte:

$$\boxed{\frac{1}{\bar{x}^h} := \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{k} \frac{h_i}{x_i}} \qquad x_i \neq 0$$

Anwendung: Diese Mittelwertbildung ist für Größen sinnvoll, die sich aus dem Verhältnis einer konstanten zu einer veränderlichen Größe ergeben:

$$x_1 = \frac{c}{a_1};\ x_2 = \frac{c}{a_2};\ x_3 = \frac{c}{a_3} \ \rightarrow\ \bar{x}^h = \frac{c}{\bar{a}} = \frac{3c}{a_1 + a_2 + a_3} = \frac{3\cancel{c}}{\frac{\cancel{c}^1}{x_1} + \frac{\cancel{c}^1}{x_2} + \frac{\cancel{c}^1}{x_3}}$$

Beispiel: Ein Autofahrer tankte fünf Tankstellen hintereinander nur um 20,- €. Der Treibstoff kostete je Liter 0,800 € / 0,875 € / 0,700 € / 1,000€ und noch einmal 0,700 €. Wie hoch waren die Durchschnittskosten pro Liter Treibstoff?

$$\bar{x}^h = \frac{5}{\frac{1}{0{,}8} + \frac{1}{0{,}875} + \frac{2}{0{,}7} + \frac{1}{1}}\ € = 0{,}800\ €$$

d) Das *quadratische Mittel* $\bar{x}^q$ ist die Quadratwurzel aus dem arithmetischen Mittelwert der Quadrate der Merkmalsausprägungen x_i:

$$\boxed{\bar{x}^q := \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2} = \sqrt{\sum_{i=1}^{k} h_i \cdot x_i^2}}$$

Dieses Mittel benötigt man z. B. in der Thermodynamik (Abhängigkeit der Geschwindigkeit eines ausströmenden Gases von den Molekülgeschwindigkeiten, ...).

Bemerkung: Für positive Stichprobenwerte gilt:

$$x_{\min} \le \bar{x}^h \le \bar{x}^g \le \bar{x} \le \bar{x}^q \le x_{\max}$$

kleinster bzw. größter Merkmalswert

Beispiel: 16, 17, 20, 15, 18, 18, 17, 18, 14, 17, 18, 13, 19, 18, 14

$x_{\min} = 13$; $\bar{x}^h \approx 16{,}6$; $\bar{x}^g \approx 16{,}7$

$\bar{x} = 16{,}8$; $\bar{x}^q \approx 16{,}9$; $x_{\max} = 20$

Die wichtigsten Kennwerte der Streuung von Messwerten:
(n ... Stichprobenumfang, k ... Anzahl der unterscheidbaren Merkmalsausprägungen)

a) Die *Spannweite* R ist die Differenz zwischen der größten und der kleinsten Merkmalsausprägung:

$$\boxed{R := x_{\max} - x_{\min}}$$

Beispiel: Vergleich zweier Stichproben mit dem gleichen Mittelwert

$x_i = 3; 5; 3; 4; 4; 8$ | $y_i = 6; 1; 3; 3; 5; 9$

$R = 8 - 3 = 5$ | $R = 8$

b) Die *mittlere lineare Abweichung* $\bar{s}$ ist der Mittelwert der Differenzbeträge aller Merkmalswerte x_i vom Mittelwert $\bar{x}$:

$$\boxed{\bar{s} := \tfrac{1}{n} \cdot \sum_{i=1}^{n} |\, x_i - \bar{x} \,| = \sum_{i=1}^{k} h_i \cdot |\, x_i - \bar{x} \,|}$$

Bemerkung: Die Beträge sind wegen der Schwerpunkteigenschaft des arithmetischen Mittels (Seite 5 oben) nötig.

$\bar{x} = \frac{27}{6} = 4{,}5$ | $\bar{y} = 4{,}5$

$\bar{s}_x = \frac{8}{6} \approx 1{,}3$ | $\bar{s}_y \approx 2{,}2$

c) Die (*empirische*) *Varianz* s^2 (mittlere quadratische Abweichung) ist der Mittelwert der Abweichungsquadrate aller Merkmalswerte x_i vom Mittelwert $\bar{x}$:

$$\boxed{s^2 := \tfrac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{k} h_i \cdot (x_i - \bar{x})^2}$$

Der Grund für die Verwendung der Quadrate an Stelle der Beträge liegt in der bequemeren Handhabung bei Abschätzungen und Beweisen.

$s_x^2 = \frac{17{,}5}{6} \approx 2{,}9$ | $s_y^2 \approx 6{,}6$

Soll die Varianz der Stichprobe gleichzeitig als *Schätzwert* für die Varianz der Grundgesamtheit dienen, so benutzt man besser

$$\boxed{s^2 := \tfrac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \tfrac{n}{n-1} \cdot \sum_{i=1}^{k} h_i \cdot (x_i - \bar{x})^2}$$

Der Grund für die Division durch den um eins kleineren Stichprobenumfang liegt in der Forderung nach Erwartungstreue eines Schätzwerts.

$s_x^2 = \frac{17{,}5}{5} = 3{,}5$ | $s_y^2 = 7{,}9$

d) Die (*empirische*) *Standardabweichung* s ist die Wurzel aus der Varianz (und hat damit die gleiche Einheit wie die Merkmalswerte und der Mittelwert; sie ist deshalb mit diesen vergleichbar):

$$\boxed{s := \sqrt{s^2}}$$

Für die Grundgesamtheit ist daher die Standardabweichung das quadratische Mittel der Abweichungen aller Merkmalswerte vom arithmetischen Mittelwert.

$s_x = \sqrt{\frac{17{,}5}{6}} \approx 1{,}7$ | $s_y \approx 2{,}6$ bzw.

bzw. $s_x \approx 1{,}9$ | $s_y \approx 2{,}8$

e) Der *Variationskoeffizient* V (*relative Streuung*) ist der Quotient aus der Standardabweichung und dem Mittelwert:

$$\boxed{V := \frac{s}{\bar{x}}}$$

Der Variationskoeffizient relativiert die Streuung: Stichproben mit gleicher Streuung aber unterschiedlichen Mittelwerten können hinsichtlich des Streuverhaltens der Merkmalswerte besser miteinander verglichen werden.

$V_x \approx \frac{1{,}7}{4{,}5} \approx 0{,}38$ | $V_y \approx 0{,}57$ bzw.

bzw. $V_x \approx 0{,}42$ | $V_y \approx 0{,}62$

Bemerkung: Die Varianz und die Standardabweichung sind die in der Praxis am häufigsten verwendeten Streumaße.

Einige wichtige Eigenschaften der Varianz bzw. der Standardabweichung:
(n ... Stichprobenumfang)

a) Die Varianz ist die Differenz des Mittelwerts der quadrierten Daten und dem Quadrat des Mittelwerts.

$$\boxed{s^2 = \tfrac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2 = (\bar{x}^q)^2 - \bar{x}^2}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 = \\ = \sum_{i=1}^{n} x_i^2 - 2n \cdot \bar{x}^2 + n \cdot \bar{x}^2$$

b) Minimaleigenschaft der Varianz: Die Varianz ist die kleinste quadratische Abweichung der Stichprobenwerte, d. h. es gilt für alle Werte $a \neq \bar{x}$:

$$\boxed{s^2 < \tfrac{1}{n} \cdot \sum_{i=1}^{n}(x_i - a)^2}$$

$$\sum_{i=1}^{n}(x_i - (\bar{x} \pm \delta))^2 = \sum_{i=1}^{n}((x_i - \bar{x}) \mp \delta)^2 = \\ \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{n \cdot s^2} \mp 2\delta \cdot \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{0} + \underbrace{\sum_{i=1}^{n}\delta^2}_{>0}$$

c) Ungleichung von ČEBYŠEV

Für alle $k \in \mathbb{R}$ gilt: Im Intervall $]\bar{x} - k \cdot s\,;\bar{x} + k \cdot s[$ liegen mindestens $n \cdot (1 - \frac{1}{k^2})$ Merkmalswerte.

Bemerkung: Diese Ungleichung erlaubt es, eine Abschätzung bezüglich der Streuung und damit der *Messungenauigkeit* für eine bestimmte Anzahl von Messwerten (z. B. 90% der Stichprobe) zu machen. Die Schätzung fällt im Allgemeinen recht „pessimistisch" aus (der Messfehler wird viel größer vermutet als er tatsächlich ist), bei bestimmten Häufigkeitsverteilungen (z. B. Normalverteilung) lassen sich die Messungenauigkeiten enger eingrenzen.

Sei n^* die Anzahl aller Messwerte x_i^*, die in $]\bar{x} - k \cdot s\,;\bar{x} + k \cdot s[$ liegen und $n - n^*$ die Anzahl der restlichen Werte x_j^{**}; dann gilt:

$$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{n \cdot s^2} = \underbrace{\sum_{i=1}^{n^*}(x_i^* - \bar{x})^2}_{\geq 0} + \underbrace{\sum_{j=1}^{n-n^*}(x_j^{**} - \bar{x})^2}_{\geq (k \cdot s)^2}$$

$$\rightarrow \quad n \cdot s^2 \geq (n - n^*) \cdot k^2 \cdot s^2$$