STOCHASTIK

Beschreibende Statistik

Hartmut Meyer
https://mathemeyer.com

Inhalt

Skalen

Für jede Statistik werden zunächst Daten in einer Stichprobe erhoben, diese Rohwerte erfüllen unterschiedliche Anforderungen an die Auswertung :

- **Nominalskala**

  Unterscheidbare Klassen ohne eine Rangfolge.

  Beispiel : Lieblingsfarbe „rot“, „blau“, „gelb“, „andere“

- **Rang- oder Ordinalskala**

  Unterscheidbare Klassen, die so aneinandergereiht sind, dass sich eine durchgehend auf- bzw. absteigende Rangordnung ergibt.

  Beispiel : Schulnoten von „1“ bis „6“

- **Intervallskala**

  Unterscheidbare Klassen mit auf- bzw. absteigender Rangordnung mit gleichen Abständen.

  Beispiel : Temperatur nach Grad Celsius

- **Rational- oder absolute Skala**

  Unterscheidbare Klassen mit auf- bzw. absteigender Rangordnung mit gleichen Abständen und variablengerechtem Nullpunkt.

  Beispiel : Körpergröße in cm

Mittelwerte

Der Zentralwert / Der Median

Der **Zentralwert** (Median) der Merkmalswerte einer *„rangierten"* Stichprobe ist für

- ungeraden Stichprobenumfang der in der Mitte stehende Wert,
- geraden Stichprobenumfang das arithmetische Mittel der beiden in der Mitte stehenden Werte.

Der Median wird immer dann verwendet, wenn

(1) die Merkmalswerte nicht mindestens „intervallskaliert" sind,
d.h. die Abstände nicht gleich groß sind,

Beispiel: *Mathenoten einer Schülergruppe*

erhobene Rohwerte : 1 / 3 / 4 / 3 / 3 / 5 / 3

Mittelwert : $1 \;\; 3 \;\; 3 \;\; \boxed{3} \;\; 3 \;\; 4 \;\; 5 \quad \Rightarrow \quad \overline{x} = 3$

(2) extreme Werte (Ausreißer) die Stichprobe einseitig beeinflussen

Beispiel: *Alter einer Schülergruppe mit Lehrer*

erhobene Rohwerte : 14 /13 / 13 / 13 / 14 / 48

Mittelwert : $13 \;\; 13 \;\; \underbrace{13 \;\; 14}_{\boxed{13{,}5}} \;\; 14 \;\; 48 \quad \Rightarrow \quad \overline{x} = 13{,}5 \; Jahre$

(das arithmetische Mittel wäre 19,1 Jahre !)

Das arithmetische Mittel

Sind n Merkmalswerte $x_1, x_2, \dots, x_n$ einer Stichprobe gegeben, so nennt man den Term

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

das arithmetische Mittel der Merkmalswerte.

Die natürliche Zahl n wird als **Umfang der Stichprobe** bezeichnet.

Beispiel: *Weitsprungleistung einer Schülergruppe*

gemessene Rohwerte (in m) : 3,25 / 2,93 / 4,02 / 3,58 / 3,88

Mittelwert : $\bar{x} = \frac{3,25+2,93+4,02+3,58+3,88}{5} = 3,532 \approx 3,53\ m$

Das arithmetische Mittel ist (strenggenommen) nur bei Intervall- bzw. Rationalskalen anwendbar.

Wenn nicht etwas anderes angegeben ist, ist mit dem Mittelwert $\overline{x}$ immer das arithmetische Mittel gemeint.

Das gewichtete arithmetische Mittel

Das gewichtete Mittel (auch *gewogenes* Mittel genannt) wird verwendet, wenn

(1) für k verschiedene Stichproben der gleichen Grundgesamtheit mit unterschiedlichen Stichprobenumfängen n_k ein gemeinsamer Mittelwert $\overline{x}$ aus den einzelnen Mittelwerten $\overline{x}_k$ berechnet werden soll:

$$\overline{x} = \frac{n_1 \cdot \overline{x}_1 + \cdots + n_k \cdot \overline{x}_k}{n_1 + \cdots + n_k} = \frac{\sum_{i=1}^{k}(n_i \cdot \overline{x}_i)}{\sum_{i=1}^{k} n_i}$$

Beispiel: *Körpergröße (in m) zweier Gruppen*

Gruppe 1: *Rohwerte: 1,72 / 1,91 / 1,82 / 1,76 / 1,79* $n_1 = 5$
Mittelwert : $\overline{x_1} = \frac{1{,}72+1{,}91+1{,}82+1{,}76+1{,}79}{5} = 1{,}80m$

Gruppe 2: *Rohwerte: 1,64 / 1,86 / 1,60* $n_2 = 3$
Mittelwert : $\overline{x_2} = \frac{1{,}64+1{,}86+1{,}60}{3} = 1{,}70m$

Gewichtetes Mittel: $\overline{x} = \frac{5\cdot 1{,}80+3\cdot 1{,}70}{5+3} = 1{,}7625$

(2) innerhalb einer Stichprobe (mit dem Umfang n) die Merkmale x_i mit unterschiedlichen Faktoren w_i gewichtet werden sollen:

$$\overline{x} = w_1 \cdot x_1 + \cdots + w_n \cdot x_n = \sum_{i=1}^{n}(w_i \cdot x_i)$$

Beispiel: *Ermittlung einer Gesamtpunktzahl in der Oberstufe*

schriftliche Note: 5 Pkte / Mitarbeitsnote: 10 Pkte
Gewichtung: schriftlich 40% / Mitarbeit 60%

Mittelwert : $\bar{x} = 0{,}4 \cdot 5P. + 0{,}6 \cdot 10P. = 8\ Punkte$

Das praktische Mittel

Sind die Merkmalswerte nicht genau bekannt, sondern nur in *Klassen* erhoben worden, muss der praktische Mittelwert berechnet werden :

Man bestimmt hierbei zunächst die sogenannten *Klassenmitten* und berechnet anschließend das arithmetische Mittel dieser Werte.

Beispiel: *Zeitaufwand für Hausaufgaben*

erhobene Rohwerte:

Zeitaufwand in min/Tag	*Klassen-mitte*	*Besetzungs-zahl*
[0 - 10]	*5*	*2*
]10 - 20]	*15*	*12*
]20 - 30]	*25*	*9*
]30 - 40]	*35*	*2*
]40 - 50]	*45*	*1*
]50 - 60]	*55*	*1*

Mittelwert: $\overline{x} = \frac{2 \cdot 5 + 12 \cdot 15 + 9 \cdot 25 + 2 \cdot 35 + 1 \cdot 45 + 1 \cdot 55}{27} = 21{,}\overline{6}\ min$

Streuung und Streuungsmaße

Wenn lediglich der Mittelwert einer Stichprobe bzw. einer Datenreihe vorliegt, ist damit noch keine Aussage über die Heterogenität der Einzelwerte gemacht.

Daten können unterschiedlich stark um den Mittelwert streuen. Das Ausmaß dieser Streuung kann durch unterschiedliche Methoden bestimmt werden.

Die verschiedenen Berechnungsmethoden unterscheiden sich prinzipiell durch ihre Beeinflussbarkeit beziehungsweise Empfindlichkeit gegenüber Ausreißern.

Zur Veranschaulichung betrachten wir ein Beispiel:

Experiment:

Zwei Schülergruppen (jeweils $n = 10$) messen mit einer Stoppuhr die „Fallzeit einer Kugel für 5m“

Messwerte:

Nr. der Messung	1. Schülergruppe (Fallzeit in s)	2. Schülergruppe (Fallzeit in s)
1	1,02	1,06
2	0,85	1,03
3	1,11	0,96
4	0,96	1,06
5	0,97	1,02
6	0,89	0,98
7	1,07	0,94
8	1,03	1,04
9	1,15	1,05
10	1,05	0,96

Das arithmetische Mittel beider Gruppen ist identisch: $\overline{x}_1 = \overline{x}_1 = 1{,}01\ s$

Die Spannweite einer Stichprobe

Die Differenz zwischen dem größtem x_{max} und dem kleinsten Wert x_{min} einer Stichprobe wird als Spannweite (oder *Variationsweite*) s* bezeichnet.

$$s^* = x_{max} - x_{min}$$

In unserem Beispiel ergibt sich:

Messwerte:

Nr. der Messung	1. Schülergruppe (Fallzeit in s)	2. Schülergruppe (Fallzeit in s)
1	1,02	1,06
2	0,85	1,03
3	1,11	0,96
4	0,96	1,06
5	0,97	1,02
6	0,89	0,98
7	1,07	0,94
8	1,03	1,04
9	1,15	1,05
10	1,05	0,96

$$s^*{}_1 = 1{,}15 - 0{,}85 = 0{,}3\ s$$

und

$$s^*{}_2 = 1{,}06 - 0{,}94 = 0{,}12\ s$$

*Die unterschiedliche Spannweite sagt aus, dass die Merkmalswerte der ersten Gruppe stärker **streuen** als bei der zweiten Gruppe.*

Die mittlere absolute Abweichung

Das arithmetische Mittel der Beträge der Abweichungen der Merkmalswerte einer Stichprobe vom Mittelwert $\overline{x}$ wird als mittlere absolute Abweichung s_m bezeichnet.

$$s_m = \frac{|x_1 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \overline{x}|$$

Beispiel: *für unser Beispiel gilt:*

Nr. der Messung	1. Schülergruppe Differenzen $\lvert(x_i - \overline{x})\rvert$	2. Schülergruppe Differenzen $\lvert(x_i - \overline{x})\rvert$
1	0,01	0,05
2	0,16	0,02
3	0,10	0,05
4	0,05	0,05
5	0,04	0,01
6	0,12	0,03
7	0,06	0,07
8	0,02	0,03
9	0,14	0,04
10	0,04	0,05
Summe :	0,74	0,40

$$s_{m_1} = \frac{0{,}74}{10} = 0{,}074\ s$$

und

$$s_{m_2} = \frac{0{,}40}{10} = 0{,}040\ s$$

Durchschnittlich weichen die Einzelwerte bei Gruppe 1 also um 0,074s und bei Gruppe 2 nur um 0,04s vom Mittelwert ab.

Die Streuung bei Gruppe 1 ist fast doppelt so groß wie bei Gruppe 2.

Die Varianz bzw. die mittlere quadratische Abweichung

Als mittlere quadratische Abweichung oder Varianz bezeichnet man das arithmetische Mittel der Abweichungsquadrate s^2 :

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ \dots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

In unserem Beispiel ergibt sich:

Messwerte:

Nr. der Messung	1. Schülergruppe $(x_i - \bar{x})^2$	2. Schülergruppe $(x_i - \bar{x})^2$
1	0,0001	0,0025
2	0,0256	0,0004
3	0,0100	0,0025
4	0,0025	0,0025
5	0,0016	0,0001
6	0,0144	0,0009
7	0,0036	0,0049
8	0,0004	0,0009
9	0,0196	0,0016
10	0,0016	0,0025
Summe :	0,0794	0,0188

$${s_1}^2 \approx \frac{0{,}0794}{10} \approx 0{,}008\ s^2$$

und

$${s_2}^2 \approx \frac{0{,}0188}{10} \approx 0{,}002\ s^2$$

Die Abweichungen als Maß dafür, wie heterogen die Messwerte sind, werden bei der mittleren quadratischen Abweichung s^2 stark gewichtet, nämlich quadriert.
Von Nachteil ist allerdings, dass die Dimension (hier „*Quadratsekunde*“) nicht mit der Dimension der Messwerte übereinstimmt. Dieser Nachteil wird geheilt durch die Standardabweichung s.

Die Standardabweichung

Die Standardabweichung s ist die Quadratwurzel aus der Varianz s^2:

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

In unserem Beispiel ergibt sich:

$$s_1 \approx \sqrt{0{,}008} \approx 0{,}089\ s$$

und

$$s_2 \approx \sqrt{0{,}002} \approx 0{,}045\ s$$