Lernzettel - Statistik: Streuung und Datenauswertung

Klasse: Klasse 10

Zwei Klassen können beide den Notenschnitt 3,0 haben, und doch ist in der einen jeder im Mittelfeld, in der anderen gibt es lauter Einsen und Fünfen. Neben der Mitte musst du daher auch die Streuung betrachten.

Einstieg: Zwei Fußballteams, gleicher Schnitt

Team A schießt in sechs Spielen 2, 2, 3, 3, 2 und 3 Tore. Team B schießt 0, 5, 0, 5, 1 und 4 Tore. Beide Teams haben im Mittel 2,5 Tore pro Spiel. Trotzdem sind sie völlig verschieden: Team A ist verlässlich, Team B schwankt stark. Der Mittelwert allein verrät das nicht. Um zu beschreiben, wie stark Daten um die Mitte streuen, braucht man **Streuungsmaße** wie Spannweite, Quartilsabstand und Standardabweichung. In diesem Kapitel lernst du, sie zu berechnen, in Boxplots darzustellen und kritisch zu deuten.

Lage- und Streuungsmaße

Definition: Lagemaße

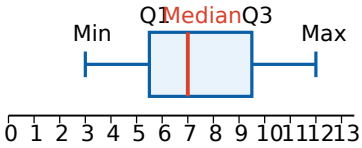
Der **Mittelwert** ist $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$. Der **Median** ist der mittlere Wert der geordneten Liste. Der **Modus** ist der häufigste Wert. Der Mittelwert reagiert stark auf Ausreißer, der Median kaum.

Definition: Spannweite und Quartile

Die **Spannweite** ist $R = x_{\max} - x_{\min}$. Das untere Quartil Q_1 ist der Median der unteren Hälfte, das obere Quartil Q_3 der Median der oberen Hälfte. Bei ungerader Anzahl wird der mittlere Wert nicht mitgezählt. Der **Quartilsabstand** ist $\mathrm{IQR} = Q_3 - Q_1$ und umfasst die mittleren 50 % der Daten.

Definition: Boxplot

Der **Boxplot** zeigt fünf Werte: Minimum, Q_1, Median, Q_3 und Maximum. Die Box reicht von Q_1 bis Q_3, im Inneren steht der Median, die Linien (Whisker) führen zu Minimum und Maximum.



Boxplot zu den Daten 3; 5; 6; 6; 7; 8; 9; 10; 12.

So gehst du vor: Fünf-Punkte-Zusammenfassung

1. Ordne die Daten der Größe nach.
2. Bestimme Minimum, Maximum und den Median.
3. Teile die Liste in zwei Hälften (bei ungerader Anzahl ohne den Median) und bestimme Q_1 und Q_3.
4. Berechne Spannweite und Quartilsabstand und zeichne den Boxplot.

Beispiel 1: Quartile

Daten: 3, 5, 6, 6, 7, 8, 9, 10, 12. Bestimme die Fünf-Punkte-Zusammenfassung.

$n = 9$, der Median ist der fünfte Wert: 7.

Untere Hälfte 3, 5, 6, 6: $Q_1 = \frac{5+6}{2} = 5{,}5$. Obere Hälfte 8, 9, 10, 12: $Q_3 = \frac{9+10}{2} = 9{,}5$.

$\mathrm{IQR} = 4$ und $R = 9$. Werte jenseits von $Q_1 - 1{,}5 \cdot \mathrm{IQR}$ und $Q_3 + 1{,}5 \cdot \mathrm{IQR}$ gelten als **Ausreißer**.

Achtung

Ordne die Daten immer zuerst, bevor du den Median bestimmst. Bei ungerader Anzahl nimmst du den Median nicht in beide Hälften auf, wenn du die Quartile berechnest.

Varianz und Standardabweichung

Definition: Varianz und Standardabweichung

Die **Varianz** ist die mittlere quadratische Abweichung vom Mittelwert:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

Die **Standardabweichung** ist $\sigma = \sqrt{\sigma^2}$ und hat dieselbe Einheit wie die Daten. Je größer σ, desto stärker streuen die Werte. (Bei Stichproben teilt man oft durch $n - 1$, hier rechnen wir mit n.)

So gehst du vor: Standardabweichung berechnen

1. Berechne den Mittelwert $\bar{x}$.
2. Bilde in einer Tabelle für jeden Wert die Abweichung und ihr Quadrat.
3. Addiere die Quadrate und teile durch n: Das ist die Varianz.
4. Ziehe die Wurzel: Das ist die Standardabweichung.

Beispiel 2: Standardabweichung

Daten: 2, 4, 4, 4, 5, 5, 7, 9.

$\bar{x} = \frac{40}{8} = 5$. Quadrierte Abweichungen: $9, 1, 1, 1, 0, 0, 4, 16$, Summe 32. Also $\sigma^2 = \frac{32}{8} = 4$ und $\sigma = 2$.

Beispiel 3: Gleicher Mittelwert

Klasse A hat die Noten 3, 3, 3, 3, 3, Klasse B die Noten 1, 2, 3, 4, 5. Vergleiche die Streuung.

Beide haben den Mittelwert 3. Bei A ist $\sigma = 0$. Bei B ist $\sigma^2 = \frac{4+1+0+1+4}{5} = 2$ und $\sigma = \sqrt{2} \approx 1{,}41$. Die Standardabweichung macht den Unterschied sichtbar.

Merke: Verschieben und Strecken der Daten

Addierst du zu jedem Wert dieselbe Zahl c, verschiebt sich der Mittelwert um c, aber σ bleibt gleich. Multiplizierst du jeden Wert mit $k > 0$, werden Mittelwert und σ mit k multipliziert.

Daten richtig auswerten

Merke: Welches Maß passt?

Bei Ausreißern oder schiefen Daten, zum Beispiel Gehältern, sind Median und Quartilsabstand besser. Bei symmetrischen Daten ohne Ausreißer passen Mittelwert und Standardabweichung.

Beispiel 4: Abgeschnittene Achse

Zwei Säulen haben die Höhen 98 und 102. Die y-Achse beginnt erst bei 96. Was fällt auf?

Die Säulen erscheinen 2 und 6 Einheiten hoch, die zweite wirkt dreimal so hoch. Tatsächlich ist $\frac{102}{98} \approx 1{,}04$: Der Unterschied beträgt nur 4 %.

Lagemaße im Vergleich

Der Mittelwert ist die Summe aller Werte geteilt durch ihre Anzahl. Der Median ist der Wert in der Mitte der geordneten Liste. Der Modus ist der häufigste Wert. Welches Maß sinnvoll

ist, hängt von den Daten ab. Der Mittelwert reagiert empfindlich auf **Ausreißer**, einzelne extrem große oder kleine Werte. Der Median bleibt dagegen stabil. Bei Gehältern, Immobilienpreisen oder Vermögen ist der Median meist aussagekräftiger als der Mittelwert. Beim Vergleichen zweier Verteilungen genügt jedoch keines der Lagemaße allein, denn Verteilungen mit gleichem Mittelwert können sehr verschieden aussehen.

Beispiel 6: Ausreißer verfälschen den Mittelwert

In einer kleinen Firma verdienen sieben Personen monatlich 2400 €, 2500 €, 2600 €, 2600 €, 2800 €, 3000 € und 9000 €. Berechne Mittelwert und Median und beurteile.

Die Summe beträgt 24 900 €, der Mittelwert also $\frac{24\,900}{7} \approx 3557$ €. Der Median ist der vierte Wert der geordneten Liste: 2600 €. Der Mittelwert liegt weit über dem, was die meisten verdienen, weil das Gehalt von 9000 € ihn nach oben zieht. Ohne diesen Wert ist der Mittelwert $\frac{15\,900}{6} = 2650$ €. Der Median beschreibt hier die typische Person besser.

Streuungsmaße

Die einfachste Kenngröße ist die Spannweite $R = x_{\max} - x_{\min}$. Sie ist leicht zu berechnen, wird aber von einem einzigen Ausreißer bestimmt. Robuster ist der Quartilsabstand $\mathrm{IQR} = Q_3 - Q_1$, die Breite der mittleren Hälfte der Daten. Die Standardabweichung s misst, wie weit die Werte im Durchschnitt vom Mittelwert entfernt sind. Sie ist die Wurzel aus der Varianz, dem Mittelwert der quadrierten Abweichungen: $s^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$. Das Quadrieren sorgt dafür, dass positive und negative Abweichungen sich nicht aufheben. Die Standardabweichung hat dieselbe Einheit wie die Daten und lässt sich direkt interpretieren.

Beispiel 7: Standardabweichung in einer Tabelle

Berechne die Standardabweichung der Daten 2, 4, 4, 4, 5, 5, 7, 9.

x_i	2	4	4	4	5	5	7	9
$x_i - \bar{x}$	-3	-1	-1	-1	0	0	2	4
$(x_i - \bar{x})^2$	9	1	1	1	0	0	4	16

Der Mittelwert ist $\bar{x} = \frac{40}{8} = 5$. Die Summe der quadrierten Abweichungen beträgt 32. Die Varianz ist $s^2 = \frac{32}{8} = 4$, die Standardabweichung $s = 2$. Die Werte weichen im Mittel um 2 vom Durchschnitt ab.

Beispiel 8: Mittelwert aus einer Häufigkeitstabelle

In einer Klasse mit 28 Schülern sind die Mathenoten: 3-mal die 1, 7-mal die 2, 9-mal die 3, 6-mal die 4, 2-mal die 5 und 1-mal die 6. Berechne Mittelwert, Median und Standardabweichung.

Der Mittelwert ist $\bar{x} = \frac{3\cdot 1 + 7\cdot 2 + 9\cdot 3 + 6\cdot 4 + 2\cdot 5 + 1\cdot 6}{28} = \frac{84}{28} = 3{,}0$. Für den Median zählst du: Die

Plätze 14 und 15 liegen in der Gruppe der Dreier (kumuliert 3, 10, 19), also ist der Median 3. Für die Varianz rechnest du $\sum f_i (x_i - \bar{x})^2 = 3 \cdot 4 + 7 \cdot 1 + 9 \cdot 0 + 6 \cdot 1 + 2 \cdot 4 + 1 \cdot 9 = 42$ und teilst durch 28: $s^2 = 1{,}5$, $s \approx 1{,}22$.

Achtung: Typische Fehler

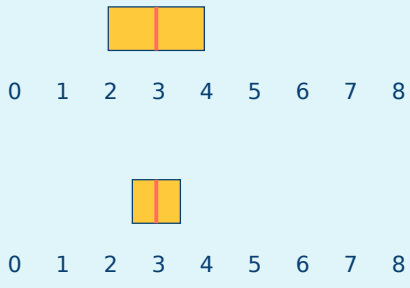
Erstens: Vor dem Bestimmen von Median und Quartilen müssen die Daten geordnet werden. Zweitens: Bei einer geraden Anzahl von Werten ist der Median der Mittelwert der beiden mittleren Werte, und die Quartile bestimmst du als Mediane der beiden Hälften. Drittens: Die Standardabweichung ist nicht der Mittelwert der Abweichungen. Viertens: Gleiche Mittelwerte bedeuten nicht gleiche Verteilungen, vergleiche immer auch ein Streuungsmaß. Fünftens: Einen Mittelwert von Mittelwerten darfst du nur bilden, wenn die Gruppen gleich groß sind.

Boxplots lesen und vergleichen

Ein Boxplot zeigt die Fünf-Punkte-Zusammenfassung Minimum, unteres Quartil, Median, oberes Quartil und Maximum. Die Box enthält die mittleren 50 % der Daten, die Linie in der Box ist der Median, die Antennen reichen bis zum kleinsten und größten Wert. Eine kurze Box bedeutet geringe Streuung, eine längere Antenne auf einer Seite eine schiefe Verteilung. Wenn die Boxen zweier Datensätze sich kaum überlappen, unterscheiden sich die Gruppen deutlich. Schreibe beim Vergleich immer einen Satz mit Zahlen: „Die Mediane sind gleich, aber der Quartilsabstand von Klasse A ist doppelt so groß.“

Beispiel 9: Zwei Klassen im Vergleich

Zwei Klassen schreiben denselben Test. Die Fünf-Punkte-Zusammenfassungen sind für A: 1; 2; 3; 4; 6 und für B: 2; 2,5; 3; 3,5; 4. Vergleiche.



Beide Klassen haben den Median 3. Die Spannweite ist bei A mit 5 mehr als doppelt so groß wie bei B mit 2, und der Quartilsabstand beträgt bei A 2 und bei B 1. Die Leistungen in Klasse B liegen enger beisammen, in Klasse A gibt es sowohl sehr gute als auch schwache Ergebnisse.

So gehst du vor: Einen Boxplot zeichnen

1. Ordne die Daten der Größe nach.
2. Bestimme den Median. Bei ungerader Anzahl ist es der mittlere Wert, bei gerader der Mittelwert der beiden mittleren.
3. Teile die Daten in die untere und obere Hälfte (bei ungerader Anzahl ohne den Median) und bestimme deren Mediane Q_1 und Q_3.
4. Zeichne eine Skala, die zu den Daten passt, und trage Minimum, Q_1, Median, Q_3 und Maximum ein.
5. Zeichne die Box von Q_1 bis Q_3, den Median als Strich und die Antennen zu Minimum und Maximum.

Ausreißer erkennen

Ein einzelner ungewöhnlich großer oder kleiner Wert kann den Mittelwert und die Spannweite stark verändern. Eine gebräuchliche Regel nennt einen Wert Ausreißer, wenn er mehr als das Eineinhalbfache des Quartilsabstands über dem oberen oder unter dem unteren Quartil liegt. Bei den Fahrzeiten des Busses ist das Quartil $Q_3 = 19{,}5$ Minuten und der Quartilsabstand 4 Minuten. Die Grenze liegt bei $19{,}5 + 1{,}5 \cdot 4 = 25{,}5$ Minuten, und 31 Minuten ist ein Ausreißer. Wichtig ist, dass du einen Ausreißer nicht einfach streichst. Prüfe, woher er kommt: Ein Messfehler wird korrigiert, aber eine echte Verspätung durch einen Unfall gehört zu den Daten und zeigt, dass der Bus manchmal sehr unpünktlich ist.

Daten richtig auswerten

Statistik besteht nicht nur aus Rechnen, sondern auch aus Fragen an die Daten. Wer wurde befragt? Eine Umfrage nur unter Gymnasiasten sagt wenig über alle Jugendlichen. Wie groß ist die Stichprobe? Mit fünf Befragten ist jedes Ergebnis zufällig. Wurde gerundet oder gruppiert? Wie sieht die Achse des Diagramms aus? Bei Kreis- und Säulendiagrammen solltest du auf die Beschriftung und die Skala achten, denn eine abgeschnittene Achse lässt kleine Unterschiede riesig erscheinen. Auch Durchschnittswerte können täuschen: Die Aussage „im Durchschnitt hat jeder Deutsche etwa ein Auto“ sagt nichts darüber, wie viele Menschen überhaupt keines besitzen. Gib deshalb zu jedem Mittelwert eine Streuung an.

Beispiel 10: Boxplot zeichnen

Zeichne den Boxplot für die Wartezeiten 4; 7; 8; 10; 11; 13; 17 (in Minuten).

Die Daten sind geordnet. Der Median ist der vierte Wert, also 10. Die untere Hälfte 4; 7; 8 hat den Median $Q_1 = 7$, die obere Hälfte 11; 13; 17 den Median $Q_3 = 13$. Das Minimum ist 4, das Maximum 17. Die Box reicht von 7 bis 13, und der Strich bei 10 liegt etwas links von der Mitte der Box. Der Quartilsabstand beträgt 6 Minuten, die Spannweite 13 Minuten. Die Hälfte aller Wartezeiten liegt zwischen 7 und 13 Minuten.

Hinweis: Faustregel für die Standardabweichung

Bei vielen natürlichen Daten, zum Beispiel Körpergrößen, liegen etwa zwei Drittel aller Werte höchstens eine Standardabweichung vom Mittelwert entfernt, und fast alle

Werte (etwa 95 %) höchstens zwei Standardabweichungen. Diese Faustregel gilt für glockenförmige Verteilungen. Bei einer Körpergröße mit dem Mittelwert 170 cm und der Standardabweichung 8 cm sind Größen zwischen 162 cm und 178 cm also normal, Größen über 186 cm schon ungewöhnlich.

Wusstest du? Tukey und der Boxplot

Der amerikanische Statistiker John Tukey entwickelte in den 1970er Jahren den Boxplot, weil er Daten schnell und ehrlich vergleichen wollte, ohne sich auf einen einzigen Wert zu verlassen. Er sagte, ein Bild sei oft mehr wert als tausend Tabellen. Heute findest du Boxplots in Wetterberichten, Studien und Berichten über Löhne. Achte aber immer auf die Achsen: Eine abgeschnittene Achse kann kleine Unterschiede riesig wirken lassen.

Merke: Das Wichtigste auf einen Blick

- Mittelwert $\bar{x} = \frac{\sum x_i}{n}$, Median (Mitte der geordneten Liste), Modus (häufigster Wert); der Median ist robust gegen Ausreißer.
- Spannweite $R = x_{\max} - x_{\min}$; Quartilsabstand $\mathrm{IQR} = Q_3 - Q_1$ umfasst die mittleren 50 % der Daten.
- Fünf-Punkte-Zusammenfassung: Minimum, Q_1, Median, Q_3, Maximum; Darstellung im Boxplot.
- Varianz $s^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$, Standardabweichung $s = \sqrt{s^2}$; gleiche Einheit wie die Daten.
- Bei Häufigkeitstabellen: $\bar{x} = \frac{\sum f_i x_i}{n}$ und $s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n}$.
- Addiert man eine Zahl zu allen Daten, verschiebt sich der Mittelwert, die Streuung bleibt; multipliziert man mit k, ändert sich s um den Faktor $|k|$.
- Gleiche Mittelwerte können zu sehr verschiedenen Verteilungen gehören: Streuung immer mit angeben.
- Diagramme kritisch lesen: abgeschnittene Achsen und ungleiche Skalen können täuschen.

Achtung

Diagramme können täuschen: abgeschnittene y-Achse, ungleiche Achsenabstände, 3D-Säulen. Verwechsle Prozentpunkte und Prozente nicht: Steigt ein Anteil von 40 % auf 44 %, sind das 4 Prozentpunkte, relativ aber 10 % mehr. Mittelwerte von Teilgruppen musst du nach den Gruppengrößen gewichten.

Beispiel 5: Median und Mittelwert

Fünf Gehälter in Euro: 2000, 2200, 2300, 2400, 12000. Vergleiche Mittelwert und Median.

Mittelwert: $\frac{20\,900}{5} = 4180$ €. Median: 2300 €. Der Ausreißer 12000 zieht den Mittelwert

stark nach oben, der Median bleibt typisch. Hier ist der Median das bessere Maß.

Ottos Tipp

Ordne die Daten immer zuerst der Größe nach. Danach findest du Median und Quartile fast von selbst. Bei der Standardabweichung hilft eine Tabelle mit den Spalten Wert, Abweichung und Abweichung zum Quadrat.