Frank Klinker

Statistische Grundlagen

Teil 2: Lagemaße und Streumaße

1 Lagemaße von Datenreihen

Ein **Lagemaß** einer Datenreihe soll einen Eindruck vermitteln, wo sich die Werte der Datenreihe konzentrieren oder um welchen speziellen Wert die weiteren Werte der Datenreihe sich – mehr oder weniger gleichmäßig – verteilen. Daher spricht man statt Lagemaß auch von Mittelwert.

Es gibt verschiedene Lagemaße und welches man gerade verwendet, liegt daran, welche Eigenschaft der Verteilung der Datenwerte man betonen möchte.

1.1 Median

Der **Median** Q_2 gibt die faktische Mitte der Datenreihe an. Das heißt links und rechts von diesem speziellen Wert liegen jeweils die Hälfte aller Werte der Datenreihe.

Um den **Median** zu bestimmen, sortiert man die Werte inklusive mehrfacher Werte der Datenreihe der Größe nach:

$$x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n \,.$$

- Ist der Umfang n der Datenreihe ungerade, dann ist der Median der Wert genau in der Mitte.
- Ist der Umfang n der Datenreihe gerade, dann gibt es in der Mitte zwei Werte der Datenreihe und der Median liegt genau in der Mitte dazwischen.[1]

Rechnerisch erhält man den Median durch die Formel

$$Q_2 = \begin{cases} x_{\frac{n+1}{2}} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{falls } n \text{ gerade} \end{cases}$$

Beispiel 1.

1. Die (sortierte) Datenreihe $\{12, 12, 12, 18, 18, 23, 27, 28, 28, 29, 33, 33\}$ hat den Umfang $n = 12$. Damit liegt der Median in der Mitte zwischen den zwei mittleren Werten der Datenreihe:
$$Q_2 = \frac{1}{2}(23 + 27) = 25$$

Adresse: Eduard-Spranger-Berufskolleg, 59067 Hamm
E-Mail: mail@frank-klinker.de
Version: 3. Oktober 2024

und ist selber nicht in der Datenreihe enthalten.

2. Die (sortierte) Datenreihe $\{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 19\}$ hat den Umfang $n = 13$. Damit ist der Median als Wert in der Mitte selbst in der Datenreihe enthalten:
$$Q_2 = 8\,.$$

3. Ersetzt man im vorigen Beispiel den Wert 19 durch den wesentlich größeren Wert 190, dann ändert sich der Median nicht.

1.2 Arithmetisches Mittel

Der Median hat zwar alle Werte einer Datenreihe verwendet, die Größe der Werte spielte dabei aber kaum eine Rolle.

Im Gegensatz dazu, verwendet das **arithmetische Mittel** auch die Größe aller Werte einer Datenreihe.

Das arithmetische Mittel ist auch der Wert, den man im Alltag verwendet, z. B. wenn man ermitteln möchte wie viel jeder Artikel eines Einkaufs im "Schnitt" gekostet hat, oder wie großdie Durchschnittsnote in einer Klausur ist.

> Das **arithmetische Mittel** $\bar{x}$ einer Datenreihe $\{x_1, x_2, \ldots, x_n\}$ mit dem Umfang n ist die Summe aller Datenwerte dividiert durch den Umfang:
> $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}\,.$$

Fakten 2. Eigenschaften des arithmetischen Mittels

1. $\bar{x}$ ist nach unten durch den kleinsten Wert der Datenreihe beschränkt und nach oben durch den größten.
2. Versieht man die die einzelnen Datenwerte mit ihren absoluten und relativen Häufigkeiten, dann ergibt sich
$$\bar{x} = \frac{H_1 \cdot x_1 + H_2 \cdot x_2 + \ldots + H_k \cdot x_k}{n}$$
und
$$\bar{x} = h_1 \cdot x_1 + h_2 \cdot x_2 + \ldots + h_k \cdot x_k\,.$$
3. Summiert man alle Differenzen $(x_i - \bar{x})$ auf, dann erhält man als Summe Null:
$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0$$
4. Summiert man alle Quadrate der Differenzen zu einer Zahl z, also $(x_i - z)^2$, dann erhält man die Funktion
$$f(z) = \frac{(x_1 - z)^2 + (x_2 - z)^2 + \ldots + (x_n - z)^2}{n}\,.$$
Diese Funktion besitzt ein Minimum an der Stelle $z = \bar{x}$, denn $f'(\bar{x}) = 0$ und $f''(\bar{x}) > 0$.

Beispiel 3.

1. Die Datenreihe $\{12, 12, 12, 18, 18, 23, 27, 28, 28, 29, 33, 33\}$ hat den Umfang $n = 12$ und es ist
$$\bar{x} = \frac{3 \cdot 12 + 2 \cdot 18 + 1 \cdot 23 + 1 \cdot 27 + 2 \cdot 28 + 1 \cdot 29 + 2 \cdot 33}{12} = \frac{273}{12} \approx 22{,}8\,.$$
2. Die Datenreihe $\{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 19\}$ hat den Umfang $n = 13$ und es ist
$$\bar{x} = \frac{3 \cdot 2 + 2 \cdot 4 + 2 \cdot 8 + 1 \cdot 13 + 2 \cdot 15 + 2 \cdot 17 + 1 \cdot 19}{13} = \frac{126}{13} \approx 9{,}7\,.$$
3. Ersetzt man im letzten Beispiel die 19 durch 190, so ergibt sich
$$\bar{x} = \frac{3 \cdot 2 + 2 \cdot 4 + 2 \cdot 8 + 1 \cdot 13 + 2 \cdot 15 + 2 \cdot 17 + 1 \cdot 190}{13} = \frac{297}{13} \approx 22{,}8\,.$$

1.3 Vergleich Median vs. arithmetisches Mittel

Wir sehen uns noch einmal zwei Datenreihen aus den vorigen Beispielen an, die Datenreihe A mit
$$A = \{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 19\}$$
und die Datenreihe B mit
$$B = \{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 190\}\,.$$
B unterscheidet sich von A durch den letzten Wert. Dabei kann es sich z. B. um einen Übertragungsfehler bei der Weitergabe der Daten handeln.

- Wie wir gesehen haben ist der Median bei a und B gleich, nämlich
$$Q_{2,A} = Q_{2,B} = 8$$
was sich aus der Tatsache ergibt, dass sich die faktische Mitte der Datenmenge durch den Tausch des maximalen Wertes nicht ändert. Man kann daher sagen:

Der Median sieht Ausreißer der Datenreihe nicht.

- Auf das arithmetische Mittel hingegen wirkt sich die Änderung von A zu B sehr stark aus, nämlich
$$\bar{x}_A \approx 9{,}7 \quad \text{und} \quad \bar{x}_B \approx 22{,}8\,.$$
Das liegt daran, dass alle Zahlenwerte der Datenreihe gleichwertig behandelt werden. Man kann daher sagen:

Das arithmetische Mittel wird von Ausreißern beeinflusst.

Vergleicht man beide Median und arithmetisches Mittel einer Datenreihe, dann gibt das Aufschluss darüber in welcher Hälfte der Datenreihe Seite sich Ausreißer befinden können:

- Ist $Q_2 < \bar{x}$, dann befinden sich eventuelle Ausreißer in der oberen Hälfte. Sicher ist jedoch, dass sich mehr Datenpunkte unterhalb des Mittelwertes befinden, als darüber.
- Ist $Q_2 > \bar{x}$, dann befinden sich eventuelle Ausreißer in der unteren Hälfte. Sicher ist jedoch, dass sich mehr Datenpunkte oberhalb des Mittelwertes befinden, als darunter.

2 Streumaße von Datenreihen

Das gewählte Lagemaß allein ist oft nicht aussagekräftig genug, um eine Datenreihe zu beschreiben: man möchte auch wissen, wie sich die Werte um diesen mittleren Werte herum verteilen.

Dazu definiert man sogenannte Streumaße einer Datenreihe.

Das einfachste Streumaß ist sicher die Gesamtausdehnung, also die Differenz zwischen größtem und kleinsten Datenwert.

Im Folgenden lernen wir zwei Streumaße kennen, die zu unseren vorher besprochenen Lagemaßen passen.

2.1 Quartile, Interquartilabstand und der BoxPlot

Hat man als Lagemaß den Median gewählt, dann kann man, statt wie beim Median die sortierte Datenreihe in zwei Teile aufzuteilen, die Datenreihe auch in mehr Teile zerlegen. Eine spezielle Wahl ergibt sich bei der Zerlegung in vier Teile.

2.1.1 Oberes und unteres Quartil, Interquartilabstand

Den oberen Wert des unteren Viertels und den unteren Wert des oberen Viertels einer Datenreihe bezeichnet man als **Quartil**.

Kennt man den Median, dann hat man die Datenreihe bereits in zwei Hälften zerlegt. Dabei ist zu beachten, dass der der Median weder zur unteren noch zur oberen Hälfte gezählt wird, wenn der Umfang n der Datenreihe ungerade ist:[2]

> Das **untere Quartil** Q_1 ist der Median der unteren Hälfte der Datenreihe.
>
> Das **obere Quartil** Q_3 ist der Median der oberen Hälfte der Datenreihe.

Beispiel 4.

[2] Der Begriff Quartil ist im Gegensatz zum Median in der Literatur nicht eindeutig festgelegt. Wendet man insbesondere die in Wikipedia:Empirisches Quantil beschriebene Formel an, dann gibt es leichte Abweichungen von der hier gewählten Definition.

- Die (sortierte) Datenreihe $\{12, 12, 12, 18, 18, 23, 27, 28, 28, 29, 30, 33, 33\}$ hat den Umfang $n = 12$. Die Zerlegung in zwei Hälften gibt $\{12, 12, 12, 18, 18, 23\} \cup \{27, 28, 28, 29, 33, 33\}$.

  Deren Mediane geben die beiden Quartile:

$$Q_1 = \frac{1}{2}(12 + 18) = 15\,, \quad Q_3 = \frac{1}{2}(28 + 29) = 28{,}5\,.$$

- Die (sortierte) Datenreihe $\{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 19\}$ hat den Umfang $n = 13$. Bei der Zerlegung in zwei Hälften nimmt man den Median heraus und erhält $\{2, 2, 2, 4, 4, 8\} \cup \{8\} \cup \{13, 15, 15, 17, 17, 19\}$.

  Deren Mediane geben die beiden Quartile:

$$Q_1 = \frac{1}{2}(2 + 4) = 3\,, \quad Q_3 = \frac{1}{2}(15 + 17) = 16\,.$$

- Die (sortierte) Datenreihe $\{1, 1, 1, 2, 8, 8, 13, 17, 18, 18, 19, 20, 23, 25, 33\}$ hat den Umfang $n = 14$. Die Zerlegung in obere und untere Hälfte gibt $\{1, 1, 1, 2, 8, 8, 13\} \cup \{17, 18, 19, 20, 23, 25, 33\}$.

  Deren Mediane geben die beiden Quartile:

$$Q_1 = 2\,, \quad Q_3 = 20\,.$$

- Die (sortierte) Datenreihe $\{1, 1, 2, 2, 7, 7, 713, 17, 18, 18, 19, 20, 25, 33\}$ hat den Umfang $n = 15$. Bei der Zerlegung in zwei Hälften nimmt man den Median heraus und erhält $\{1, 1, 2, 2, 7, 7, 7\} \cup \{13\} \cup \{17, 18, 18, 19, 20, 25, 33\}$.

  Deren Mediane geben die beiden Quartile:

$$Q_1 = 2\,, \quad Q_3 = 19\,.$$

Als **Interquantilabstand** (IQR) bezeichnet man den Abstand von oberem und unterem Quartil, also

$$IQR = Q_3 - Q_1$$

Das ist der Abstand in dem sich die Hälfte der Datenwerte symmetrisch um den Median befinden: ein Viertel unterhalb und ein Viertel oberhalb des Medians.

2.1.2 Der Boxplot

Median, unteres und oberes Quartil und damit auch den IQR kann man gut mit Hilfe eines **Boxplots** graphisch darstellen. Die namensgebende Box umfasst den Datenbereich zwischen dem unteren und den oberen Quartil.

Ergänzt wird dieser dann um die so genannten **Antennen**, die einen weiteren Einblick in die Verteilung der Werte einer Datenreihe gewähren. Diese Antennen verbinden den Rand der Box jeweils mit einem Datenwert oberhalb und unterhalb der Box.

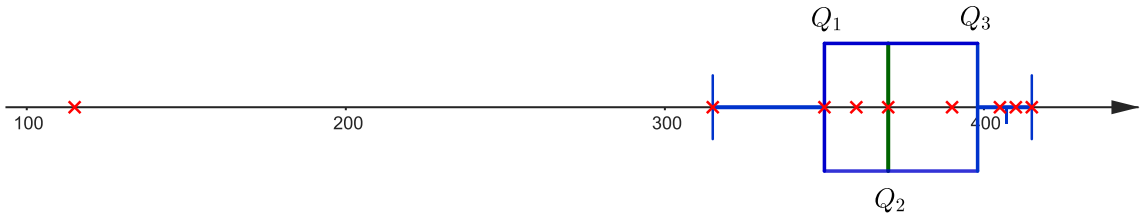
Die Länge der Antennen ist nicht festgelegt und wird an die gewünschte Aussagekraft des Boxplots in Bezug zur Datenreihe gewählt. Die maximal mögliche Länge der Antennen ist durch den größten und kleinsten Wert der Datenreihe begrenzt. Eine übliche Antennenlänge ist durch maximal dem 1,5-fache des IQR gegeben und diese Wahl wird hier in den Beispielen genutzt.

Die Werte der Datenreihe, die von den Antennen nicht mehr erreicht werde, nennt man **Ausreißer**.

Beispiel 5.

Datenwerte:	115	315	350	360	370	390	405	410	415
abs. Häufigkeit:	1	2	3	1	2	5	4	1	2

Wegen $n = 21$ ist der Median hier $Q_2 = x_{11} = 390$. Das untere Quartil ist $Q_1 = \frac{1}{2}(x_5 + x_6) = 350$ und das obere Quartil $Q_3 = \frac{1}{2}(x_{16} + x_{17}) \approx 398$. Damit liefert der Interquartilabstand $IQR = Q_3 - Q_1 = 48$ die Länge der Box. Die Antennen haben eine maximale Länge von $1{,}5 \cdot IQR = 72$:



Die obere Antenne reicht bis zum Maximalwert der Datenreihe, die untere Antenne endet jedoch schon vor dem Minimalwert. Deshalb kann der Wert 115 der Datenreihe als Ausreißer interpretiert werden.

Bemerkung 6. Ist im vorigen Beispiel der Ausreißer 115 ein Weitergabefehler und ersetzt man ihn z. B. 315 (der besser zur Datenreihe passt), dann ändert sich der Boxplot nicht.

Der Boxplot ist ebenso wie die Quartile ebenso robust gegenüber Ausreißern wie der Median.

2.2 Varianz und Standardabweichung

Wir haben im Punkt 4 von Fakt 2 gesehen, dass das arithmetische Mittel das Minimum einer speziellen Abstandsfunktion ist. Dieses Minimum heißt die **Varianz** der Datenreihe.

Die **Varianz** s^2 einer Datenreihe $\{x_1, x_2, \ldots, x_n\}$ mit Umfang n ist gegeben durch

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}.$$

Dabei ist $\bar{x}$ das arithmetische Mittel der Datenreihe.

Ein mit Blick auf die Werte der Datenreihe anschaulicheres Streumaß ist die **Standardabweichung** s mit

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

Im Gegensatz zur Varianz hat die Standardabweichung die gleiche Dimension wie die einzelnen Datenwerte. Sind z. B. die Datenwerte x_i Längenangaben mit der Einheit m, dann gilt das auch für die Standardabweichung, während die Varianz die Einheit m^2 besitzt.

Fakten 7. Eigenschaften der Varianz

1. Berechnet man die einzelnen Quadrate in der Summe dann folgt

$$s^2 = \frac{x_1^2 + x_2^2 + \ldots x_n^2}{n} - \bar{x}^2\,.$$

2. Versieht man die die einzelnen Datenwerte mit ihren absoluten und relativen Häufigkeiten, dann ergibt sich

$$s^2 = \frac{H_1(x_1 - \bar{x})^2 + H_2(x_2 - \bar{x})^2 + \ldots + H_k(x_k - \bar{x})^2}{n}$$

und

$$s^2 = h_1(x_1 - \bar{x})^2 + h_2(x_2 - \bar{x})^2 + \ldots + h_k(x_k - \bar{x})^2\,.$$

3. s^2 ist der minimale Wert der Funktion $g(z) = \frac{1}{n}\sum_{i=1}^{n}(x_i - z)^2$, nämlich an der Stelle $z = \bar{x}$.

Beispiel 8.

1. Die Datenreihe $\{12, 12, 12, 18, 18, 23, 27, 28, 28, 29, 33, 33\}$ hat den Umfang $n = 12$ und es ist $\bar{x} = 22{,}8$

$$s^2 = \frac{3\cdot(12-22{,}8)^2+2\cdot(18-22{,}8)^2+1\cdot(23-22{,}8)^2+1\cdot(27-22{,}8)^2+2\cdot(28-22{,}8)^2+1\cdot(29-22{,}8)^2+2\cdot(33-22{,}8)^2}{12}$$
$$\approx 59{,}5$$

und

$$s \approx 7{,}7\,.$$

2. Die Datenreihe $\{2, 2, 2, 4, 4, 8, 8, 13, 15, 15, 17, 17, 19\}$ hat den Umfang $n = 13$ und es ist $\bar{x} = 9{,}7$

$$s^2 = \frac{3\cdot(2-9{,}7)^2+2\cdot(4-9{,}7)^2+2\cdot(8-9{,}7)^2+1\cdot(13-9{,}7)^2+2\cdot(15-9{,}7)^2+2\cdot(17-9{,}7)^2+1\cdot(19-9{,}7)^2}{13}$$
$$\approx 39{,}1$$

und

$$s \approx 6{,}3\,.$$

3. Ersetzt man im letzten Beispiel die 19 durch 190, so ist $\bar{x} = 22{,}8$ ergibt sich

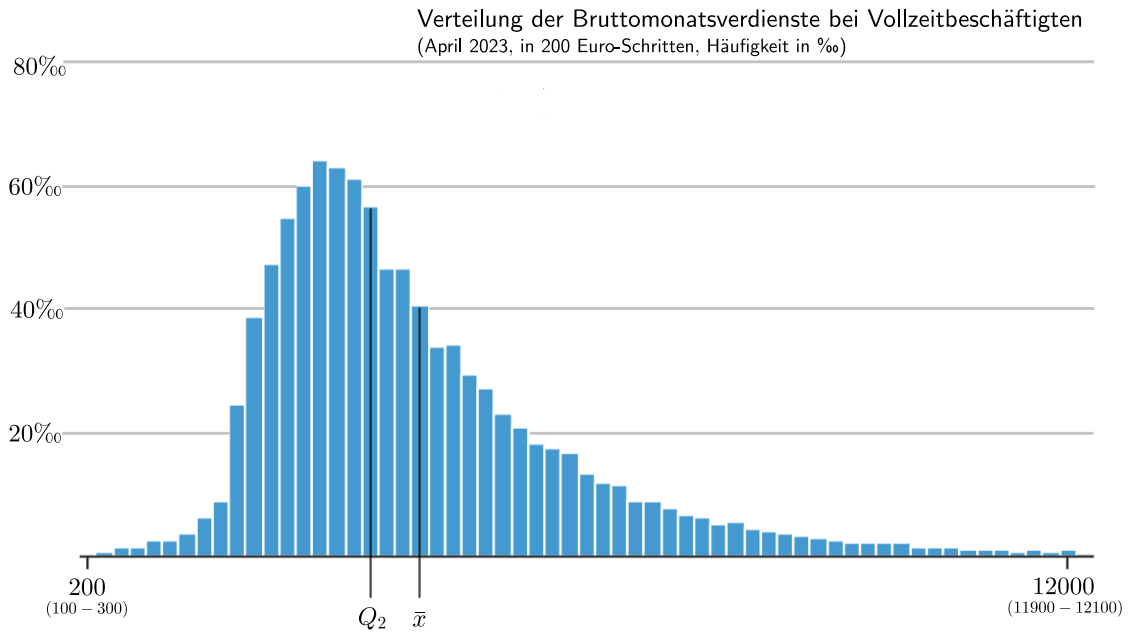
$$s^2 = \frac{3\cdot(2-22{,}8)^2+2\cdot(4-22{,}8)^2+2\cdot(8-22{,}8)^2+1\cdot(13-22{,}8)^2+2\cdot(15-22{,}8)^2+2\cdot(17-22{,}8)^2+1\cdot(190-22{,}8)^2}{13}$$
$$\approx 2360{,}3$$

und

$$s \approx 48{,}6\,.$$

Bemerkung 9. An den letzten beiden Beispielen sieht man wieder, wie stark sich Ausreißer auswirken: der Summand $\frac{(190-22{,}8)^2}{13} \approx 2150{,}4$ trägt mit großem Abstand am meisten zur Gesamtsumme in s^2 bei.

3 Beispiel: Einkommensverteilung in Deutschland, April 2023



Die Daten[3] sind in der folgenden Tabelle sortiert und die Häufigkeit ist in ‰

[3] Originalgrafik und Originaldaten: Statistisches Bundesamt (Destatis). Beides wurde leicht an die hier verwendete Notation angepasst.

angegeben. Rundungsbedingt ist der Umfang $n = 997$ (statt 1000).

x_i	200	400	600	800	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800	3000
H_i	0	1	2	1	3	3	4	6	9	25	39	48	55	60	64

x_i	3200	3400	3600	3800	4000	4200	4400	4600	4800	5000	5200	5400	5600	5800	6000
H_i	63	61	57	47	47	41	34	35	31	27	23	21	19	18	17

x_i	6200	6400	6600	6800	7000	7200	7400	7600	7800	8000	8200	8400	8600	8800	9000
H_i	14	13	13	9	9	8	7	7	5	6	5	5	4	4	3

x_i	9200	9400	9600	9800	10000	10200	10400	10600	10800	11000	11200	11400	11600	11800	12000
H_i	3	2	2	2	2	2	2	2	1	1	1	1	1	1	1

Weil der Umfang ungerade ist, ist der Median durch

$$Q_2 = x_{\frac{997+1}{2}} = x_{499} = 3600\,€$$

gegeben. Das arithmetische Mittel ist

$$\bar{x} = \frac{1}{997}\sum_{i=1}^{60} H_i x_i \approx 4200\ €$$

und liegt damit deutlich über dem Median.

Damit verdienen mehr als die Hälfte aller Arbeitnehmerinnen und Arbeitnehmer weniger als der Durchschnitt (mit dem in der Regel das arithmetische Mittel gemeint ist).

Unteres und oberes Quartil sind gegeben durch

$$Q_1 = 2800 \quad \text{und} \quad Q_3 = 5000\,,$$

sodass der Interquartilabstand $IQR = 2200$ ist. Wir wählen die maximale Antennenlängen wieder als $1{,}5 \cdot 2200 = 3300$.

Am Boxplot sieht man nun, dass die Hälfte aller Vollzeitbeschäftigten zwischen 2800 € und 5000 € verdienen. Bei unserer Wahl der Antennenlängen gibt es keine Ausreißer nach unten, aber alle Personen, die mehr als 8300 € verdienen sind Ausreißer nach oben; zum betrachteten Zeitpunkt sind das 4% aller Personen aus der beschriebenen Gruppe.