2: Maßzahlen

Statistische Verfahren in der Geographie

Till Straube <straube@geo.uni-frankfurt.de>

Institut für Humangeographie
Goethe-Universität Frankfurt

1 Lernziele dieser Sitzung

Sie können…

- die wichtigsten Lagemaße von Stichproben bestimmen.
- die wichtigsten Streumaße von Stichproben bestimmen.
- Boxplots interpretieren.

2 Einleitende Bemerkungen

Die im Folgenden besprochenen Maßzahlen (oder Kennzahlen, Parameter) verdichten (oder aggregieren) Häufigkeitsverteilungen einer Variable. Durch diese Parameter kann das Charakteristische einer Verteilung schnell erfasst und vergleichbar gemacht werden. Die Verdichtung auf Maßzahlen geht jedoch immer auch mit Informationsverlust einher.

Die Möglichkeit der Angabe statistischer Maßzahlen ist abhängig vom Skalenniveau der Daten, wie der Überblick in Tabelle 1 zeigt.

2.1 Beispielverteilung

Alle Berechnungen von Maßzahlen werden am folgenden Beispiel illustriert: Für die 14 Gemeinden im Landkreis Rothenberge wurde die jeweilige Anzahl an Gaststätten erhoben. Die Zählung ergab die Wertereihe in Tabelle 2.

3 Lagemaße

Lagemaße (auch Maße der Zentraltendenz, Lokalisationsparameter, Mittelwerte, engl. *measures of central tendency*) bezeichnen alle statistischen Maßzahlen, die eine Verteilung repräsentieren, indem sie die Lage der mittleren oder häufigsten Variablenwerte angeben.

Im Falle einer unimodalen, perfekt symmetrischen Verteilung (z.B. Glockenform) haben alle drei Lageparameter den gleichen Wert. Je weiter Verteilungen von dieser Form abweichen – durch Mehrgipfligkeit oder Asymmetrie – desto unpräziser ist die Beschreibung der Verteilung durch einen einzigen Parameter.

Tabelle 1: Die wichtigsten Maßzahlen

Parameter	Typ	Mindestes Skalenniveau	Formel
Modalwert	Lagemaß	nominal	Mo
Median	Lagemaß	ordinal	$Md = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{falls } n \text{ gerade} \\ x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \end{cases}$
Arithmetisches Mittel	Lagemaß	metrisch	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Spannweite	Streumaß	ordinal	$R = x_{(n)} - x_{(1)}$
Quartilsabstand	Streumaß	ordinal	$IQR = Q_3 - Q_1$
Varianz	Streumaß	metrisch	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Standardabweichung	Streumaß	metrisch	$s = \sqrt{s^2}$

Tabelle 2: Beispielverteilung

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
4	1	4	1	5	5	0	1	8	5	1	25	3	3

3.1 Median

Der Median (engl. *median*) einer Verteilung ist der Wert, der größer als genau 50% aller Werte ist.

Da dies eine Größer-kleiner-Relation der Werte voraussetzt, kann der Median nur für ordinale und metrische Skalenniveaus angegeben werden.

Im Folgenden wird die (einfachere) Bestimmung des Medians nach Bortz und Schuster (2010) verwendet. Benninghaus (2007) beschreibt ein anderes Verfahren, welches zu anderen Ergebnissen kommen kann.

Um den Median zu bestimmen, wird zunächst eine geordnete Liste angefertigt, indem die Werte aufsteigend sortiert werden. Diese sortierten Werte werden mit $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ bezeichnet (also mit Klammern). Für unsere Beispielverteilung ergibt sich Tabelle 3.

Tabelle 3: Sortierte Wertereihe

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
0	1	1	1	1	3	3	4	4	5	5	5	8	25

Bei einer ungeraden Stichprobengröße n teilt der $(\frac{n+1}{2})$-te Wert (also der Wert genau in der Mitte) die Stichprobe in zwei Hälften, weshalb gilt:

$$Md = x_{(\frac{n+1}{2})} \quad \text{falls } n \text{ ungerade.} \tag{1}$$

Bei geradem n entstehen zwei gleich große Hälften der Stichprobe: $x_{(1)}$ bis $x_{(\frac{n}{2})}$ einerseits, und $x_{(\frac{n}{2}+1)}$ bis $x_{(n)}$ andererseits. Der Durchschnitt zwischen $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$ teilt die Stichprobe in zwei Hälften. Es gilt:

$$Md = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} \quad \text{falls } n \text{ gerade.} \tag{2}$$

In unserem Beispiel ist $n = 14$ und damit gerade. Der Median errechnet also nach Formel 2 wie folgt:

$$\begin{aligned} Md &= \frac{x_{(7)} + x_{(8)}}{2} \\ &= \frac{3+4}{2} \\ &= 3{,}5 \end{aligned}$$

Softwarehinweis

In R gibt die Funktion `median()` den Median einer Verteilung aus.

3.2 Modalwert

Der Modalwert Mo (auch Modus, engl. *mode*) gibt den häufigsten Wert oder die häufigsten Werte einer Verteilung an.

Der Modalwert kann so auch (als einziger Mittelwert) für nominalskalierte Variablen angegeben werden.

Bei ordinalen und metrischen Skalenniveaus sind folgende Besonderheiten zu beachten:

- Wird der Modus einer Verteilung durch unmittelbar benachbarte Werte gebildet, wird er als Kombination (bei metrischen Variablen als arithmetisches Mittel) dieser Werte angegeben.
- Bei bimodalen (multimodalen) Verteilungen werden beide (alle) Modalwerte angegeben.

Hierzu müssen die Häufigkeiten der Werte bekannt sein, bzw. bestimmt werden (s. Tabelle 4).

Tabelle 4: Häufigkeiten der Beispielverteilung

Wert x_i	Häufigkeit f_i
0	1
1	4
3	2
4	2
5	3
8	1
25	1

Der Modalwert der Beispielverteilung beträgt 1, da der Wert 1 am häufigsten (viermal) vorkommt.

3.3 Arithmetisches Mittel

Das arithmetische Mittel (auch Mittelwert, Durchschnitt, engl. *mean*) ist das gebräuchlichste Lagemaß und Grundlage für viele statistische Verfahren.

Das arithmetische Mittel setzt ein metrisches Skalenniveau voraus.

Die Berechnung des arithmetischen Mittels einer Stichprobe erfolgt durch die Formel:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3}$$

Für unsere Beispielverteilung ergibt sich durch einsetzen in Formel 3:

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{14} x_i}{14} \\
&= \frac{4+1+4+1+5+5+0+1+8+5+1+25+3+3}{14} \\
&= \frac{63}{14} \\
&\approx 4{,}71
\end{aligned}$$

Softwarehinweis

Der Befehl für die Ermittlung des arithmetischen Mittels in R lautet `mean()`.

4 Streumaße

Streumaße (auch Streuungs-, Variabilitäts-, Dispersionswerte, engl. *measures of variability*) geben Auskunft darüber, wie heterogen die Werte einer Verteilung sind, d.h. wie breit sie gestreut sind. Während Lagemaße den typischen Wert einer Verteilung ermitteln, zeigen Streumaße, wie gut (oder eigentlich: wie schlecht) dieser typische Wert die Verteilung repräsentiert.

4.1 Spannweite

Die Spannweite (engl. *range*) gibt Auskunft darüber, wie groß der Wertebereich ist, der von einer Verteilung abgedeckt wird. Sie wird (für metrische Skalen) als die Differenz vom größten zum kleinsten Wert (also vom letzten zum ersten Wert einer geordneten Werteliste) angegeben:

$$R = x_{(n)} - x_{(1)} \tag{4}$$

Für unsere Beispielstichprobe ergibt sich (mit Blick auf Tabelle 3):

$$\begin{aligned} R &= x_{(14)} - x_{(1)} \\ &= 25 - 0 \\ &= 25 \end{aligned}$$

Softwarehinweis

In R gibt die Funktion `range()` die Werte für $x_{(1)}$ und $x_{(n)}$ aus.

4.2 Quartilsabstand

Der Quartilsabstand (engl. *interquartile range / IQR*) gibt die Größe des Wertebereichs der mittleren 50% einer Verteilung an.

Genau so wie der Median eine Messwertreihe in zwei gleich große Hälften „schneidet", schneiden die Quartile die Werte in Viertel. Dabei liegt der so genannte untere Angelpunkt Q_1 genau über 25% der Werte, Q_2 ist identisch mit dem Median und der obere Angelpunkt Q_3 liegt genau über 75% der Werte.

Der Angelpunkt Q_1 wird ermittelt, indem der Median für die unteren 50% (Q_3: die oberen 50%) der Werte bestimmt wird – also jener Werte, die theoretisch unterhalb des Medians der Gesamtverteilung liegen.

Dabei folgen wir Bortz und Schuster (2010) und nehmen im Fall eines ungeraden n den Median auf beiden Seiten hinzu.

Die Formel für den Quartilsabstand lautet:

$$IQR = Q_3 - Q_1 \tag{5}$$

Der Quartilsabstand ist Ausreißern gegenüber stabiler als die Spannweite, da extreme hohe oder niedrige Wert nicht in die Berechnung einfließen.

In unserem Beispiel (mit $n = 14$) ist die untere Hälfte der Verteilung:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
0	1	1	1	1	3	3

Q_1 ist der Median dieser Werte, also $x_{(4)} = 1$.

Die oberen 7 Werte lauten:

$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
4	4	5	5	5	8	25

Q_3 ist also $x_{(11)} = 5$.

Für den Quartilsabstand ergibt sich durch einsetzen in Formel 5:

$$\begin{aligned} IQR &= 5 - 1 \\ &= 4 \end{aligned}$$

Softwarehinweis

In R werden die Quartile üblicherweise mit `quantile()` und der Quartilsabstand mit `IQR()` bestimmt.

Achtung: Genau wie für den Median gibt es auch für die Ermittlung der Quartile bzw. des Quartilsabstands unterschiedliche Verfahren. Die Ergebnisse dieser R-Funktionen weichen hier deshalb meist leicht vom hier besprochenen Verfahren ab!

4.3 Varianz

Die Varianz einer Messwertreihe (engl. *variance*) kann verstanden werden als der durschnittliche quadrierte Abstand der Werte zum arithmetischen Mittel.

Die Formel lautet:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \tag{6}$$

Die Quadrierung der Differenz hat dabei einen doppelten Effekt: Zum einen bekommen auch negative Differenzen ein positives Vorzeichen, so dass sich positive und negative Differenzen nicht neutralisieren. Zum anderen werden hierdurch besonders große Abweichungen zum arithmetischen Mittel stärker gewichtet als dies ohne Quadrierung der Fall wäre.

Zudem fällt auf, dass im Gegensatz zur Formel für das arithmetische Mittel im Nenner $n - 1$ steht und nicht etwa n. Dies hat mit so genannten Freiheitsgraden zu tun, die wir allerdings erst in Sitzung 6 genauer kennenlernen.

Für unsere Beispielstichprobe wird die Berechnung für alle einzelnen $(x_i - \bar{x})^2$ schnell aufwendig und unübersichtich. Deshalb berechnen wir ihre Summe hier mit Hilfe einer Häufigkeitstabelle (s. Tabelle 5). Dabei werden alle distinkten Werte einzeln transformiert und in der letzten Spalte mit ihrer Häufigkeit multipliziert.

Tabelle 5: Häufigkeitstabelle zur Berechnung der Varianz

Wert	Häufigkeit	Abweichung vom Mittel	Quadrat der Abweichung	Produkt mit Häufigkeit
x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
0	1	-4,71	22,18	22,18
1	4	-3,71	13,76	55,04
3	2	-1,71	2,92	5,84
4	2	-0,71	0,50	1,00
5	3	0,29	0,08	0,24
8	1	3,29	10,82	10,82
25	1	20,29	411,68	411,68

Schließlich werden die Werte in Formel 6 eingesetzt:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{14} (x_i - \bar{x})^2}{14 - 1} \\ &\approx \frac{22{,}18 + 55{,}04 + 5{,}84 + 1 + 0{,}24 + 10{,}82 + 411{,}68}{13} \\ &= \frac{506{,}80}{13} \\ &\approx 38{,}98 \end{aligned}$$

Eine solche Tabelle lässt sich analog auch für die Berechnung von Summen größerer Messwertreihen für das arithmetische Mittel verwenden.

Zudem lässt dieses Verfahren sich auf klassierte Daten anwenden, wenn für x_i der Mittelwert der Klassen eingesetzt wird (womit allerdings Informations- und Präzisionsverlust einhergeht).

Softwarehinweis

In R lautet der Befehl für die Errechnung der Varianz `var()`.

4.4 Standardabweichung

Die Standardabweichung (engl. *standard deviation*) ist das gebräuchlichste Streumaß und spielt eine herausragende Rolle in den allermeisten statistischen Verfahren.

Die Standardabweichung einer Messwertreihe ist definiert als die Quadraturzel ihrer Varianz:

$$s = \sqrt{s^2} \tag{7}$$

Indem hier die Wurzel gezogen wird, wird in gewisser Weise die Quadrierung der Differenzen für die Varianz wieder „korrigiert“. Insbesondere wird die Quadrierung der Maßeinheit wieder aufgehoben – die Standardabweichung hat also die gleiche Einheit wie die Messreihe selbst.

In unserem Beispiel beträgt die Standardabweichung also:

$$s \approx \sqrt{38{,}98} \approx 6{,}24$$

Softwarehinweis

Die Standardabweichung wird in R mit der Funktion `sd()` berechnet.

5 Boxplot

Der Boxplot (auch Box-and-whisker-plot) kombiniert einige der gebräuchlichsten Maßzahlen in einer übersichtlichen Grafik (s. Abbildung 1).

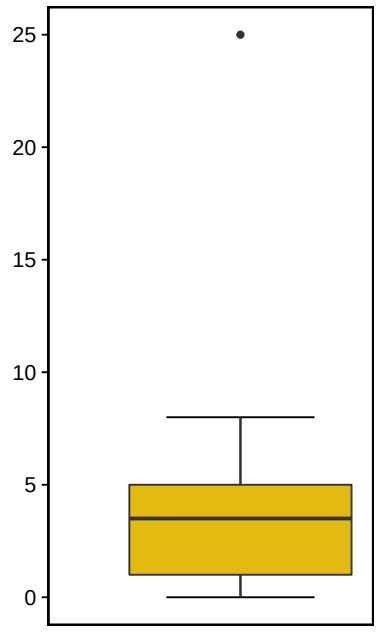

Abbildung 1: Boxplot der Beispielverteilung

Die Höhe der „Box“ definiert sich durch den Quartilsabstand, der mittlere Strich markiert den Median und die „Whisker“ markieren den Wertebereich insgesamt – wobei Ausreißer, deren Abstand zur Box mehr als das 1,5-Fache des Quartilsabstands beträgt, üblicherweise gar nicht oder (wie hier) gesondert mit Punkten markiert werden.

Softwarehinweis

In R lässt sich ein Boxplot mit dem Befehl `boxplot()` ausgeben.

6 Aufgaben

Die folgenden Aufgaben sind zur eigenständigen Überprüfung Ihrer Lernleistung gedacht (als Vor- oder Nachbereitung der Vorlesung, oder als Klausurübung) und nicht etwa als Hausaufgabe.

6.1 Aufgabe 1

Bei einer Befragung jedes 500. Studierenden im Matrikel einer privaten Hochschule wurden folgende Angaben zur Haushaltsgröße gemacht:

```
1 4 4 2 3 2 3 5 2 7 2 1 1
```

a) Welches Skalenniveau liegt vor? (Sitzung 1)
b) Berechnen Sie Modalwert,
c) Median und
d) arithmetisches Mittel der Stichprobe.
e) Berechnen Sie außerdem die Spannweite,
f) den Quartilsabstand,
g) die Varianz und
h) die Standardabweichung der Stichprobe.
i) Zeichnen Sie einen Boxplot der Stichprobenverteilung.

6.2 Aufgabe 2

In Australien betrug die durchnittliche Niederschlagsmenge in den 1970er und 80er Jahren:[1]

Jahr	Niederschlag (mm)
1970	384,52
1971	493,65
1972	364,65
1973	661,32
1974	785,27
1975	603,45
1976	527,75
1977	471,81
1978	525,65
1979	455,64
1980	433,01
1981	535,12
1982	421,36
1983	499,29
1984	555,21
1985	398,88
1986	391,96
1987	453,41
1988	459,84
1989	483,78

a) Welches Skalenniveau liegt vor? (Sitzung 1)
b) Legen Sie eine klassierte Häufigkeitstabelle an. Begründen Sie die Wahl der Klassen. (Sitzung 1)

[1] Auszug aus dem Datensatz `bomsoi` in Maindonald und Braun (2015)

c) Was ist der Modalwert der klassierten Verteilung?
d) Wie groß ist der Quartilsabstand?
e) Bestimmen Sie das arithmetische Mittel der klassierten Verteilung.
f) Berechnen Sie die Standardabweichung.
g) Zeichnen Sie einen Boxplot für die Verteilung.

7 Tipps zur Vertiefung

7.1 Lagemaße

- Kapitel 3.3.1 in Benninghaus (2007)
- Kapitel 2.1 in Bortz und Schuster (2010)
- Kapitel 4.2.1 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Arithmetisches, harmonisches und geometrisches Mittel
- YouTube-Kanal „Kurzes Tutorium Statistik“: Boxplots, Median, Quartile

7.2 Streumaße

- Kapitel 3.1.2 in Benninghaus (2007)
- Kapitel 2.2 in Bortz und Schuster (2010)
- Kapitel 4.2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Streumaße - Varianz, Standardabweichung, Variationskoeffizient und mehr!

7.3 Boxplot

- Kapitel 3.4 in Bortz und Schuster (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Boxplots, Median, Quartile

Quellen

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. 5. Aufl. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Benninghaus, Hans. 2007. *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler*. 11. Aufl. Wiesbaden: VS Verlag.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Maindonald, John H und W John Braun. 2015. DAAG. Data Analysis and Graphics Data and Functions. https://CRAN.R-project.org/package=DAAG.