STATISTIK I – Übung 02
Median und Modus

1 Kurze Wiederholung

Median

Beim sogenannten Median handelt es sich ebenfalls um ein statistisches Lagemaß, welches – im Gegensatz zum arithmetischen Mittel – auch für ordinalskalierte Daten berechnet werden kann. Der Median ist als derjenige Wert definiert, der genau in der Mitte der geordneten Werte (die Ordnung von Werten setzt bekanntlich mindestens ordinalskalierte Daten voraus) einer Verteilung liegt. Da es bei einer ungeraden Anzahl von Werten tatsächlich genau einen „mittigen" Wert gibt, während bei einer geraden Anzahl von Werten zwei Werte in der Mitte der Verteilung liegen, existieren für die Berechnung des Median zwei unterschiedliche Formeln.

Bei einer ungeraden Anzahl von Werten wird der mittlere Wert der geordneten Verteilung gewählt:

$$x_{med} = x_{(\frac{n+1}{2})}$$

Bei einer geraden Anzahl von Werten wird das arithmetische Mittel der „mittigen" Werte gebildet:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

mit:

x_{med} = *Median*

n = Anzahl der Werte der Verteilung

x_n = *Wert an n-ter Stelle der geordneten Verteilung*

Robustheit des Median

Im Gegensatz zum arithmetischen Mittel ist der Median Ausreißern gegenüber äußerst robust. Dies zeigt sich am bereits bekannten Beispiel der Verteilungen [1; 2; 3; 4] und [1; 2; 3; 50]. Während das arithmetische Mittel der ersten Verteilung bei 2,5 liegt, liegt das arithmetische Mittel der zweiten Verteilung bei 14 – das arithmetische Mittel wird also durch den einzelnen, aus dem Rahmen fallenden Wert deutlich sichtbar nach oben verzerrt. Betrachtet man hingegen den Median, so ist festzustellen, dass dieser sowohl in der ersten als auch in der zweiten Verteilung bei 2,5 liegt – der Median wird durch den einzelnen Ausreißer also überhaupt nicht beeinträchtigt.

Der Grund hierfür liegt auf der Hand: Während in die Berechnung des arithmetischen Mittels sämtliche Werte der Verteilung mit exakt dem gleichen Gewicht eingehen (also auch sämtliche Ausreißer), werden für die Berechnung des Median in diesem Fall lediglich zwei Werte (bei einer ungeraden Anzahl von Werten sogar lediglich ein Wert) benötigt, die in der Mitte der geordneten Verteilung liegen und daher unmöglich Ausreißer sein können. (Der Sonderfall einer Verteilung mit zwei Werten ist per se unsinnig und wird an dieser Stelle nicht weiter betrachtet.) Im Datensatz eventuell vorhandene Ausreißer gehen daher nicht in die Berechnung des Median ein – und können diesen somit auch nicht beeinflussen. Das Gedankenexperiment zeigt: Auch wenn man in der zweiten Verteilung [1; 2; 3; 50] den Ausreißer auf 500, 5.000 oder 50.000 setzen würde, bliebe der Median stabil bei 2,5.

Perzentile

Im vorangegangenen Abschnitt haben wir die Definition des Median als den Wert kennengelernt, der exakt in der Mitte der geordneten Verteilung liegt. Diese Definition kann man abgewandelt auch wie folgt formulieren: 50% der Werte einer Verteilung sind entweder kleiner oder gleich dem Median, während die anderen 50% der Werte einer Verteilung entweder größer oder gleich dem Median sind. Diese Betrachtungsweise macht deutlich, dass es sich beim Median lediglich um den Sonderfall eines anderen Lagemaßes – des Perzentils – handelt, der den Datensatz genau an der Marke 50/50 teilt. Ebenso sind aber natürlich auch andere Perzentile vorstellbar, die den Datensatz etwa an der Marke 20/80, 45/55 oder 95/5 teilen. Der Median ist insofern lediglich das bekannteste – und meistberechnete – Perzentil. Gemeinsam mit den 25%-Perzentil sowie dem 75%-Perzentil, bildet der Median die sogenannten Quartile, die einen Datensatz exakt in vier gleich große Wertebereiche unterteilen:

- 25%-Perzentil (25% aller Werte liegen unterhalb dieses Wertes, 75% liegen oberhalb)
- 50%-Perzentil – Median (50% aller Werte liegen unter- bzw. oberhalb dieses Wertes)
- 75%-Perzentil (75% aller Werte liegen unterhalb dieses Wertes, 25% liegen oberhalb)

Anders formuliert handelt es sich bei dem 25%-Perzentil um den Median der Werte unterhalb des 50%-Perzentils, während das 75%-Perzentil den Median der Werte oberhalb des 50%-Perzentils darstellt. Die Differenz zwischen 75%-Perzentil und 25%-Perzentil spielt als sogenannter Interquartilsabstand (IQR = Inter Quartile Range) eine bedeutende Rolle bei der Konstruktion von Box-Plots und stellt zudem das einzige Streuungsmaß dar, das Ausreißern gegenüber robust ist.

Für die Berechnung beliebiger Perzentile existieren – analog zur Berechnung des Median – wiederum zwei Formeln. Ergibt die Multiplikation der gewünschten Perzentilgrenze p mit der Anzahl der Werte der Verteilung n (d.h. n * p), keinen ganzzahligen Wert, berechnet sich das Perzentil wie folgt:

$$x_p = x_{(k)}$$

Dabei ist k die nächste auf das Ergebnis der Multiplikation (n * p) folgende, ganze Zahl. Ergibt (n * p) dagegen einen ganzzahligen Wert (in diesem Falle k), berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

Modus

Der Modus – den wir der Kürze wegen an dieser Stelle noch an die Betrachtung des Medians anhängen wollen – ist das einzige Lagemaß, das auch für nominalskalierte Daten bestimmt werden kann. Er ist als der in den unklassierten Daten am häufigsten auftretende Wert definiert, bei gleichbreit klassierten Daten entspricht der Modus dagegen der Klassenmitte der Klasse, welche die meisten Werte auf sich vereinen kann. Der Modus lässt sich ohne Berechnung unmittelbar aus den Daten herauslesen – allerdings nur dann, wenn ein eindeutiges Maximum (d.h. eine unimodale Verteilung) vorliegt. Bei bi- oder multimodalen Verteilungen kann der Modus in der Regel (es sei denn, zwei Werte treten tatsächlich exakt gleich oft auf) zwar rechnerisch bestimmt, jedoch nicht mehr vernünftig interpretiert werden. Da keine scharfen Kriterien existieren, ist es der Anwenderin bzw. dem Anwender überlassen, wann eine Verteilung „gerade noch“ als unimodal gelten kann bzw. als bimodal gelten muss.

Auf eine Besonderheit bei der Verwendung von Software wie SPSS, PAST oder PSPP sei abschließend noch hingewiesen: Liegt ein bimodaler oder multimodaler Datensatz vor, wird zumeist nur der in der Häufigkeitstabelle zuoberst stehende, am häufigsten auftretende Wert als Modus ausgegeben. Dies ist dann selbstverständlich kein rechnerisch gültiger Modus – auch wenn SPSS das anders sieht.

2 Beispielrechnungen

Median, Perzentile und Modus

Für eine Gruppe von Studierenden liegt die folgende Altersverteilung vor:

Alter	Absolute Häufigkeit	Relative Häufigkeit	Kumulierte abs. Häufigkeit	Kumulierte rel. Häufigkeit
21 Jahre	5	0,25	5	0,25
22 Jahre	4	0,20	9	0,45
23 Jahre	3	0,15	12	0,60
24 Jahre	4	0,20	16	0,80
25 Jahre	4	0,20	20	1,00
Summe	20	1,00	20	1,00

Um den Median berechnen zu können, müssen die Werte der Verteilung zunächst in eine geordnete Reihenfolge gebracht werden:

21; 21; 21; 21; 21; 22; 22; 22; 22; 23; 23; 23; 24; 24; 24; 24; 25; 25; 25; 25

Da es sich um eine gerade Anzahl an Werten handelt, steht kein einzelner Wert direkt in der Mitte der geordneten Verteilung. Für die Bestimmung des Median wird in diesem Fall daher auf die zweite Medianformel zurückgegriffen:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

Es ist also das arithmetische Mittel des 10 (20/2) und des 11 (20/2+1) Wertes zu berechnen:

(23+23) / 2 = 23

Der Median liegt somit bei 23 Jahren.

Zusätzlich zum Median sollen nun noch das 25%-Perzentil sowie das 75%-Perzentil berechnet werden, um die Quartile vollständig angeben und den Interquartilsabstand bestimmen zu können. Hierzu wird wie folgt vorgegangen:

(0,25 * 20) = 5 -> ganzzahliger Wert -> k = 5

(0,75 * 20) = 15 -> ganzzahliger Wert -> k = 15

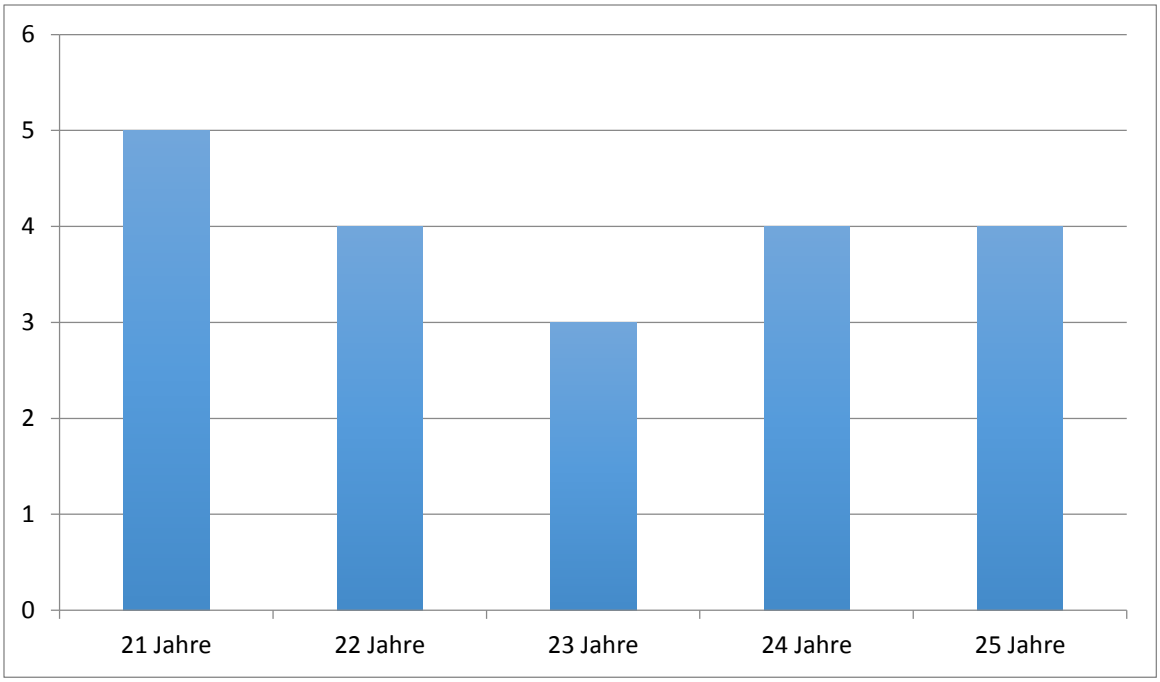
$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

$p_{0,25} = (x_5 + x_6) / 2 = (21+22) / 2 = 21{,}5$

$p_{0,75} = (x_{15} + x_{16}) / 2 = (24+24) / 2 = 24$

Die drei Quartile liegen demnach bei 21,5 Jahren (unteres Quartil), 23 Jahren (Median) und 24 Jahren (oberes Quartil), der Interquartilsabstand liegt bei 2,5 Jahren (24 – 21,5).

Da kein eindeutiges Maximum existiert, ist die Bestimmung des Modus in diesem Fall nicht angezeigt. Verwendet man zum Nachrechnen dieser Übungsaufgaben eine Statistiksoftware, wird diese eventuell 21 Jahre als Modus ausgeben. Dass dies kein sinnvolles Ergebnis ist, kann man sich leicht vor Augen führen, wenn man sich das Balkendiagramm für die Verteilung anzeigen lässt.



3 Übungsaufgaben (Lösungen folgen in der kommenden Woche)

Median und Perzentile

Im Rahmen eines Produkttests werden 20 Probandinnen und Probanden gebeten, den Geschmack eines neuartigen Joghurtprodukts auf einer Skala von 1 („hervorragend“) bis 5 („scheußlich“) zu bewerten. Der Test erbringt die folgenden Daten:

Bewertung	Anzahl an Probanden/innen
1 („hervorragend“)	5
2 („gut“)	6
3 („mittelmäßig“)	2
4 („ungenügend“)	1
5 („scheußlich“)	6

a) Bestimmen Sie den Median.

b) Bestimmen Sie den Interquartilsabstand.

Modus

Im Rahmen einer Qualitätsstichprobe werden 100 vom Band laufende Maschinenteile einer Genauigkeitskontrolle (Abweichung des Durchmessers von der zu erfüllenden Norm in mm) unterzogen. Die Stichprobenziehung erbringt die folgenden Daten.

Abweichung des Durchmessers von der Norm	Anzahl an Maschinenteilen in der Stichprobe
0 mm	17
[1 mm – 5 mm)	62
[5 mm – 9 mm)	12
[9 mm – 13 mm)	9

a) Bestimmen Sie den Modus.