STATISTIK I – Übung 04
Spannweite und IQR

1 Kurze Wiederholung

Was sind Dispersionsparameter?

Die sogenannten Dispersionsparameter oder statistischen Streuungsmaße geben Auskunft darüber, wie die Werte einer Verteilung um deren Zentrum (gekennzeichnet durch ein statistisches Lagemaß) streuen, d.h. ob sie tendenziell eher dicht am Zentrum oder eher weit davon entfernt liegen. Warum die Kenntnis der Streuung unter anderem für den Vergleich zweier Verteilungen von großer Bedeutung ist, soll das folgende fiktive Beispiel der Einkommensverteilung in zwei Abteilungen eines Unternehmens mit je sechs Mitarbeiterinnen und Mitarbeitern verdeutlichen.

Mitarbeiter Abt. A	Einkommen	Mitarbeiter Abt. B	Einkommen
MA 1	2.500,00 Euro	MA 1	4.130,00 Euro
MA 2	2.550,00 Euro	MA 2	1.060,00 Euro
MA 3	2.480,00 Euro	MA 3	1.110,00 Euro
MA 4	2.630,00 Euro	MA 4	5.020,00 Euro
MA 5	3.000,00 Euro	MA 5	4.000,00 Euro
MA 6	2.210,00 Euro	MA 6	1.250,00 Euro
Summe	15.370,00 Euro	Summe	16.570,00 Euro
Arithmetisches Mittel	2.561,67 Euro	Arithmetisches Mittel	2.761,67 Euro

Wie man sieht, liegen die arithmetischen Mittel beider Verteilungen mit 2.561 Euro in Abteilung A und 2.761 Euro in Abteilung B gerade einmal 200 Euro voneinander entfernt – dennoch wäre es mit Blick auf die Daten falsch, würde man aus dieser geringen Differenz den Schluss ziehen, dass die Mitarbeiterinnen und Mitarbeiter in beiden Abteilungen etwa gleich viel verdienen. Würde man in Abteilung A anheuern, könnte man in der Tat davon ausgehen, um 2.561 Euro zu verdienen. In Abteilung B liegt dagegen kein Mitarbeiter mit seinem Einkommen auch nur in der Nähe des Durchschnittswertes – hier verdient man entweder deutlich mehr oder deutlich weniger als 2.761 Euro.

Das Beispiel zeigt, dass es bei der Betrachtung von Verteilungen eben nicht nur darauf ankommt, das Zentrum der Verteilung zu kennen – vielmehr muss man auch eine Vorstellung davon haben, wie sich die Werte um dieses Zentrum herum verteilen. Auskunft hierüber geben die Dispersionsparameter, von denen wir im Rahmen der Vorlesung vier – Spannweite, Interquartilsabstand, Varianz und Standardabweichung – kennenlernen werden. In dieser Übungseinheit werden wir – neben der Theorie – zunächst die „einfacheren" Parameter Spannweite und Interquartilsabstand abarbeiten, während wir uns in der kommenden Übungseinheit der Varianz sowie der Standardabweichung zuwenden.

Dispersionsparameter und Skalenniveaus

Ähnlich wie schon bei den statistischen Lagemaßen, hängt auch die Wahl geeigneter Dispersionsparameter vom Skalenniveau der betrachteten Verteilung ab, wobei in diesem Fall nur zwei Skalenniveaus – Ordinalskala und Kardinalskala – von Bedeutung sind. Da nominalskalierte Werte nicht quantifiziert werden können, können sie auch nicht um das Zentrum einer Verteilung (in diesem Fall also um den Modus) streuen, da das Vorhandensein einer Streuung natürlich voraussetzt, dass sich die Abstände zwischen den Werten einer Verteilung und ihrem Zentrum ermitteln lassen. Da dies bei nominalskalierten Werten nicht möglich ist, existiert kein Maß für deren Streuung. Man kann sich dies leicht vor Augen führen, indem man mit dem Geschlecht ein typisches nominalskaliertes Merkmal betrachtet. Befragt man 40 Personen, von denen 30 männlich sind, ist „männlich" zwar der Modus – aber wie sollte in diesem Fall eine „Streuung" der Werte um dieses Zentrum aussehen?

Die Berechnung des Interquartilsabstands verlangt mindestens ordinalskalierte Werte, da – wie wir nachfolgend noch sehen werden – das obere und untere Quartil in diesen Parameter einfließen, deren Berechnung bekanntlich ebenfalls ordinaskalierte Werte voraussetzt. Da in die Berechnung der Varianz sowie der sich aus dieser ergebenden Standardabweichung das arithmetische Mittel einfließt, können diese beiden Parameter analog nur für metrisch skalierte Werte berechnet werden. Auch die Berechnung der Spannweite ist – aus Gründen, die wir nachfolgend noch betrachten werden – nur für metrisch skalierte Werte sinnvoll. Wie bei den statistischen Lagemaßen gilt auch für die Dispersionsparameter, dass sie abwärts-, aber nicht aufwärtskompatibel sind, d.h. der Interquartilsabstand lässt sich auch für metrisch skalierte Werte berechnen, während die Berechnung von Varianz, Standardabweichung und Spannweite für ordinalskalierte Werte nicht möglich ist.

Spannweite

Bei der Spannweite handelt es sich um die Differenz zwischen dem kleinsten und dem größten Wert im Datensatz. Da die Bildung von sinnvollen Differenzen voraussetzt, dass diese quantifizierbar sind, d.h. dass mit den Abständen zwischen den Werten gerechnet werden kann, kann die Spannweite nur für metrische Daten berechnet werden – auch wenn es vielleicht kontraintuitiv erscheint, dass etwa eine Spannweite von 5 Notenpunkten bei einer Klassenarbeit keine sinnvolle Größe sein soll.

$$d_s = x_{\max} - x_{\min}$$

Intuitiv dürfte dagegen die Erkenntnis sein, dass die Spannweite kein robustes Maß für die Streuung ist, da sie von Ausreißern – soweit im Datensatz vorhanden – extrem beeinflusst wird. Tritt an einem Ende der Verteilung mindestens ein Ausreißer auf, findet dieser in jedem Fall Eingang in die Berechnung der Spannweite – gibt es an beiden Enden der Verteilung mindestens einen Ausreißer, wird die Spannweite sogar ausschließlich durch diese bestimmt. Das nachfolgende Beispiel verdeutlicht, wie extrem die Spannweite durch Ausreißer beeinflusst werden kann.

Verteilung A: [100; 120; 170; 280; 290; 300]	Spannweite: $d_s = 300 - 100 = 200$
Verteilung B: [5; 120; 170; 280; 290; 10000]	Spannweite $d_s = 10000 - 5 = 9995$

Aber auch wenn keinerlei Ausreißer im Datensatz vorhanden sind, gibt die Spannweite aufgrund der Tatsache, dass lediglich zwei Werte in ihre Berechnung einfließen, nur ein eher unvollständiges Bild der Streuung wieder, wie anhand eines zweiten Beispiels demonstriert werden soll.

Verteilung A: [100; 100; 110; 115; 118; 300]	Spannweite: $d_s = 300 - 100 = 200$
Verteilung B: [100; 140; 185; 240; 285; 300]	Spannweite: $d_s = 300 - 100 = 200$

Liegen metrisch skalierte Werte vor, ist die Berechnung von Varianz und Standardabweichung also in jedem Fall zu bevorzugen. In der Praxis werden der größte und der kleinste Wert allerdings noch für die sogenannte Fünf-Werte-Zusammenfassung benötigt, auf die später eingegangen werden soll.

Interquartilsabstand (IQR)

Der Interquartilsabstand (nachfolgend als IQR – Inter Quartile Range – abgekürzt) ist als die Differenz zwischen dem oberen und dem unteren Quartil definiert, die wir bereits im Rahmen der statistischen Lagemaße kennengelernt haben.

$$IQR = x_{0,75} - x_{0,25}$$

Damit der IQR bestimmt werden kann, muss also die Berechnung der Quartile möglich sein, was wiederum – wie eingangs bereits angesprochen – mindestens ordinalskalierte Werte voraussetzt. Der IQR kann aber auch für metrisch skalierte Werte kalkuliert werden (Abwärtskompatibilität).

Im Gegensatz zur Spannweite sowie auch zur Varianz und Standardabweichung, die wir im Rahmen der nächsten Übungseinheit besprechen werden, wird der IQR durch Ausreißer nicht beeinflusst und ist somit der einzige in dieser Vorlesung besprochene robuste Dispersionsparameter. Dies lässt sich anhand des weiter oben genutzten Beispiels zur Robustheit der Spannweite demonstrieren:

Verteilung A: [100; 120; 170; 280; 290; 300] IQR = 290 – 120 = 170
Verteilung B: [5; 120; 170; 280; 290; 10000] IQR = 290 – 120 = 170

Der IQR spielt eine zentrale Rolle bei der Konstruktion von Box-Plots (Höhe der Box) – einer der zwei grafischen Darstellungsformen (zusammen mit dem Stem-and-Leaf-Plot) mit Relevanz für die Klausur. Wie ein Box-Plot konstruiert wird, wird Gegenstand einer zukünftigen Übungseinheit sein.

Fünf-Werte-Zusammenfassung

Bei der sogenannten Fünf-Werte-Zusammenfassung handelt es sich um eine hochkomprimierte Darstellung der Lage sowie der Streuung einer Verteilung, bestehend aus den drei Quartilen (oberes und unteres Quartil sowie Median) sowie dem größten und dem kleinsten Wert der Verteilung.

$$\left[x_{\min}; x_{0,25}; x_{med}; x_{0,75}; x_{\max}\right]$$

2 Beispielrechnungen

Im Rahmen einer Befragung machten 30 Probanden Angaben zu ihrem Körpergewicht.

Nr.	Körpergewicht	Nr.	Körpergewicht
1	61,2 kg	16	62,2 kg
2	72,8 kg	17	79,5 kg
3	62,3 kg	18	61,0 kg
4	85,7 kg	19	55,5 kg
5	91,8 kg	20	98,9 kg
6	64,9 kg	21	94,4 kg
7	74,3 kg	22	85,3 kg
8	95,2 kg	23	93,0 kg
9	87,3 kg	24	72,8 kg
10	82,4 kg	25	88,2 kg
11	78,4 kg	26	87,4 kg
12	91,2 kg	27	68,6 kg
13	89,3 kg	28	90,5 kg
14	76,3 kg	29	85,0 kg
15	68,2 kg	30	71,9 kg

Berechnung der Spannweite

Die Spannweite berechnet sich als Differenz zwischen dem größten und dem kleinsten Wert.

$d_s = x_{max} – x_{min} = 98,9 – 55,5 = \underline{43,4}$

Die Spannweite beträgt 43,4 kg.

Berechnung des Interquartilsabstands

Für die IQR-Berechnung müssen die Werte zunächst in eine geordnete Reihe gebracht werden.

Werte 1-10: 55,5; 61,0; 61,2; 62,2; 62,3; 64,9; 68,2; 68,6; 71,9; 72,8
Werte 11-20: 72,8; 74,3; 76,3; 78,4; 79,5; 82,4; 85,0; 85,3; 85,7; 87,3
Werte 21-30: 87,4; 88,2; 89,3; 90,5; 91,2; 91,8; 93,0; 94,4; 95,2; 98,9

Für die Berechnung beliebiger Perzentile (die Quartile sind bekanntlich drei Perzentilwerte) existieren – analog zur Berechnung des Median – zwei bereits in einer der Übungseinheiten zu den statistischen Lagemaßen vorgestellte 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 gesuchte 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 gesuchte Perzentil wie folgt:

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

In diesem Fall werden das obere (p = 0,75) und das untere (p = 0,25) Quartil benötigt, d.h.:

(30 * 0,25) = 7,5 -> kein ganzzahliger Wert -> k = 8 -> Der 8. Wert im Datensatz lautet 68,6
(30 * 0,75) = 22,5 -> kein ganzzahliger Wert -> k = 23 -> Der 23. Wert im Datensatz lautet 89,3

IQR = 89,3 – 68,6 = <u>20,7</u>

Der Interquartilsabstand beträgt 20,7 kg.

<u>Angabe der Fünf-Werte-Zusammenfassung</u>

Vier der fünf für die Fünf-Werte-Zusammenfassung benötigten Werte (größter Wert, kleinster Wert, oberes Quartil und unteres Quartil) sind aus den vorangegangenen Rechnungen bereits bekannt – es wird demnach nur noch der Median gesucht. Da eine gerade Anzahl von Werten (n = 30) vorliegt, wird für die Berechnung des Median auf folgende Formel zurückgegriffen:

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

$$x_{med} = \frac{1}{2}(x_{(\frac{30}{2})} + x_{(\frac{30}{2}+1)})$$

$$x_{med} = \frac{1}{2}(x_{15} + x_{16})$$

$$x_{med} = \frac{1}{2}(79,5 + 82,4)$$

$$x_{med} = 80,95$$

Wie bei jeder Perzentilberechnung kann natürlich auch analog zu oben vorgegangen werden:

(30 * 0,50) = 15 -> ganzzahliger Wert -> k = 15; k+1 = 16 -> ½* (79,5 + 82,4) = 80,95

Die Fünf-Werte-Zusammenfassung lautet somit:

[55,50 kg; 68,60 kg; 80,95 kg; 89,30 kg; 98,90 kg]

3 Übungsaufgaben (Lösungen folgen in der kommenden Woche)

Die gleichen 30 Probanden machten außerdem Angaben zu ihrem Alter.

Nr.	Alter	Nr.	Alter
1	17 Jahre	16	33 Jahre
2	36 Jahre	17	22 Jahre
3	41 Jahre	18	62 Jahre
4	34 Jahre	19	44 Jahre
5	32 Jahre	20	41 Jahre
6	54 Jahre	21	56 Jahre
7	45 Jahre	22	62 Jahre
8	22 Jahre	23	23 Jahre
9	71 Jahre	24	86 Jahre
10	14 Jahre	25	41 Jahre
11	86 Jahre	26	65 Jahre
12	44 Jahre	27	53 Jahre
13	34 Jahre	28	35 Jahre
14	21 Jahre	29	21 Jahre
15	54 Jahre	30	19 Jahre

a) Bestimmen Sie die Spannweite.

b) Bestimmen Sie den Interquartilsabstand.

c) Bestimmen Sie die Fünf-Werte-Zusammenfassung.