Lagemasse und Streuung

Benjamin Schlegel
07. März 2016

Lagemasse sagen etwas über die Lage und das Zentrum der Daten aus, Streuungsmasse, wie die Daten um dieses Zentrum gestreut sind.

Lagemasse

Lagemasse sagen etwas über die Lage der Daten aus. Wo liegt das Zentrum? Welcher Wert kommt am häufigsten vor? Fragen, welche durch Lagemasse beantwortet werden können.

arithmetische Mittel

Das arithmetische Mittel (oft Durchschnitt genannt) ist die Summe der Werte durch die Anzahl der Werte geteilt.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n}$$

Das arithmetische Mittel ist empfindlich gegenüber Ausreissern. Ein sehr hoher oder sehr tiefer Wert kann alleine schon das arithmetische Mittel stark verändern. Bei Zufallszahlen wird das arithmetische Mittel oft Erwartungswert \(\mu\) genannt. Das arithmetische Mittel ist nur für numerische Werte definiert.

Beispiel: In einer Schulklasse haben die Kinder folgende Alter: 10, 10, 10, 11, 10, 11, 11, 12, 10, 10, 10, 10, 10, 11, 11, 11. Das Durchschnittsalter der Klasse ist 10.5. Der Lehrer ist kurz vor der Pensionierung und hat das Alter 64. Damit ist das Durchschnittsalter aller Personen im Klassenzimmer 13.64.

geometrische Mittel

Das geometrische Mittel ist die n-te Wurzel aus dem Produkt aller Werte. Es wird häufig bei Verhältnisse und Wachstumsraten verwendet und immer dann, wenn das Produkt sinnvoll ist und nicht die Summe von mehreren Werten.

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}$$

Ein Beispiel in der Politikwissenschaft eines geometrischen Mittels ist der Human Development Index (HDI). Der Wert ist das geometrische Mittel aus dem Lebenserwartungs-Index bei Geburt (LEI), dem Bildungs-Index (BI) und dem Einkommensindex (EI). Es wird folgende Formel verwendet:

$$HDI = \sqrt[3]{LEI \cdot BI \cdot EI}$$

Alle drei Indexe haben einen Wert zwischen 0 und 1. Um einen hohen Wert zu erzielen, müssen alle drei Werte hoch sein. Ein tiefer Wert verändert den Index viel stärker als es bei einem arithmetischen Mittel der Fall wäre. Der höchste HDI Wert hatte 2013 Norwegen mit 0.944, der tiefste Wert Niger mit 0.337 (Einige Länder lieferten keine Daten.).

Median

Der Median ist bei sortieren Werten der mittlere Wert. Er ist nicht nur für numerische Werte definiert, sondern auch für Werte auf der Ordinalskala. Der Median teil eine Stichprobe in zwei Hälften. Er ist robust gegen Ausreisser.

Zurück zum Beispiel mit der Schulklasse: Um den Median zu berechnen, werden die Werte sortiert aufgelistet: 10, 10, 10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 11, 12. Es sind 16 Werte. Damit ist der Median der 8. oder 9. Werte (resp. der Durchschnitt aus diesen beiden, sofern er sich, wie in diesem Beispiel, berechnen lässt.). Sowohl der 8. als auch der 9. Wert sind 10. Der Median ist damit 10. Nimmt man den Lehrer dazu, sind es 17 Werte. Der Median ist der 9. Wert. Dieser ist 10. Der Median hat sich in diesem Beispiel durch den Ausreisser (Lehrer) nicht verändert.

Modus

Der Modus ist der Wert, der am häufigen vorkommt. Er kann immer berechnet werden, als auch bei Variablen auf der Nominalskala. Der Modus ist aber nicht immer eindeutig. Eine Verteilung mit einem Modus ist unimodal (eingipflig), mit zwei Modi bimodal (zweigipflig) und mit mehreren Modi multimodal (mehrgipflig).

Beim Beispiel mit der Schulklasse lässt sich auch der Modus berechnen. Neun Schüler sind zehn Jahr alt, sechs Schüler elf Jahr alt und ein Schüler zwölf. Der Modus ist damit wie der Median 10. Der Lehrer ändert am Resultat nichts.

Quantile

Quantile teilen die Verteilung in n gleichgrosse Teile auf. Der Median ist das 1/2-Quantil, welches die Verteilung in zwei gleichgrosse Hälften aufteilt. Quartile sind die Werte, welche die Verteilung in vier gleichgrosse Teile aufteilen, Quintile in fünf gleichgrosse Teile.

Beim Beispiel mit der Schulklasse sind es 16 Werte. Das 1. Quartil (Q0.25) ist der 4. oder 5. Wert (=10), das 2. Quartil (Q0.5)ist gleich dem Median (=10) und das 3. Quartil (Q0.75) ist der 12. resp. 13. Wert (=11).

Streuung

Die Streuung sagt es über die Verteilung der Daten aus. Sind alle in der Nähe des Zentrum oder sind sie weiter weg? Sind die Daten gleichmässig verteilt oder gibt es viele Extremwerte? Sind die Daten einseitig oder symmetrisch verteilt? Fragen die mit Streuungsmassen beantwortet werden können.

Spannweite und Interquartilsabstand

Die Spannweite ist die Distanz zwischen dem kleinsten und dem grössten Wert. Die Spannweite ist nicht robust gegenüber Ausreissern, sondern hängt nur von den Extremwerten ab.

$$Spannweite = Max - Min$$

Beim Schulklassenbeispiel ist die Spannweite ohne Lehrer 2 und mit Lehrer 54.

Robuster ist der Interquartilsabstand. Er misst den Abstand zwischen dem 1. und 3. Quartil.

$$Interquartilsabstand = Q_{0.75} - Q_{0.25}$$

Bei der Schulklasse ist der Interquartilsabstand sowohl mit als auch ohne Lehrer 1 (11-10).

Varianz und Standardabweichung

Die Varianz ist die erwartete quadratische Abweichung einer Zufallsvariable von ihrem Erwartungswert.

$$\sigma^2 = E((X-\mu)^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Die Standardabweichung ist die Wurzel aus der Varianz.

Beim Beispiel mit der Schulklasse ist der Erwartungswert 10.5 (Durchschnitt). Damit kann die Varianz folgendermassen berechnet werden:

$$\frac{9\cdot(10-10.5)^2+6\cdot(11-10.5)^2+1\cdot(12-10.5)^2}{16} = 0.375$$

Die Varianz ist 0.375 und die Standardabweichung 0.612.

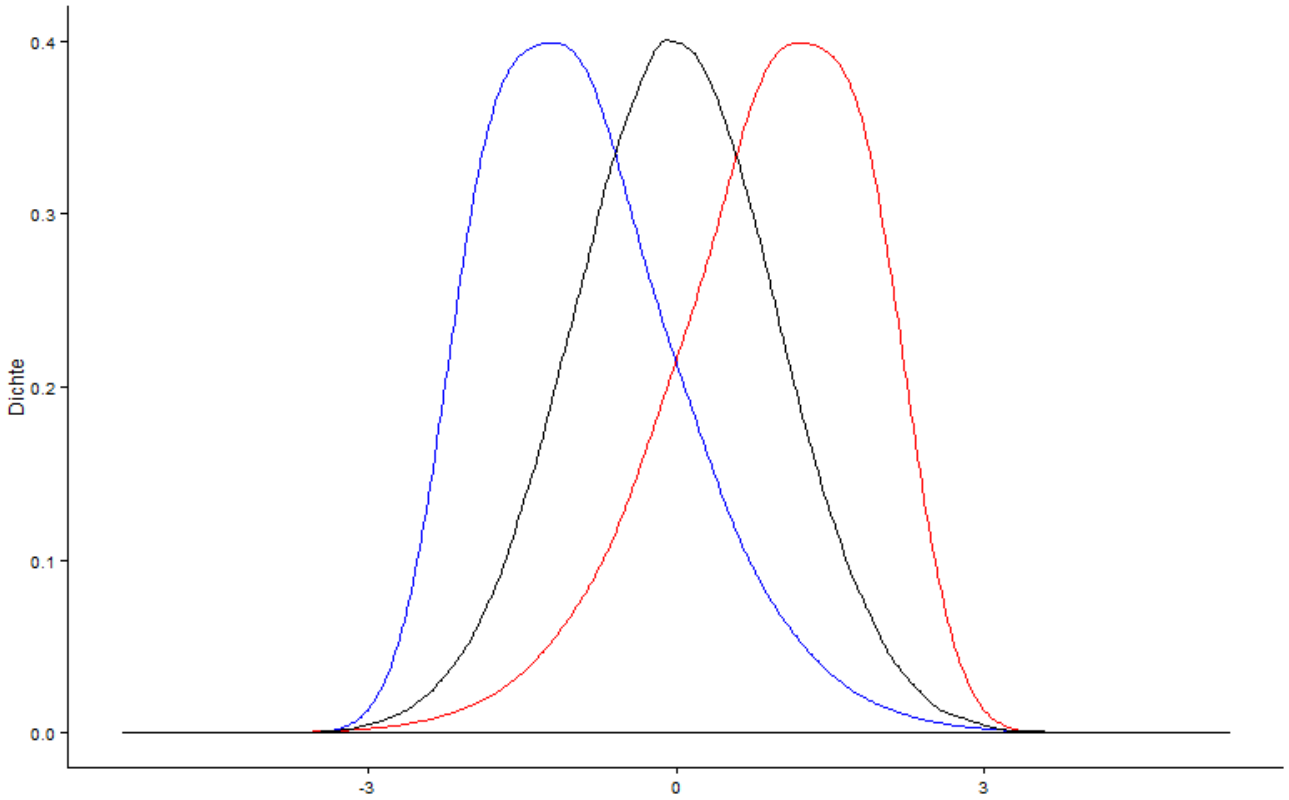
Will man die Stichprobenvarianz berechnen, so teilt man nicht durch n, sondern durch n-1.

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

Schiefe

Die Schiefe gibt an, inwiefern eine Verteilung auf eine Seite tendiert. Sie wird mit folgender Formel berechnet:

$$\nu = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^3$$



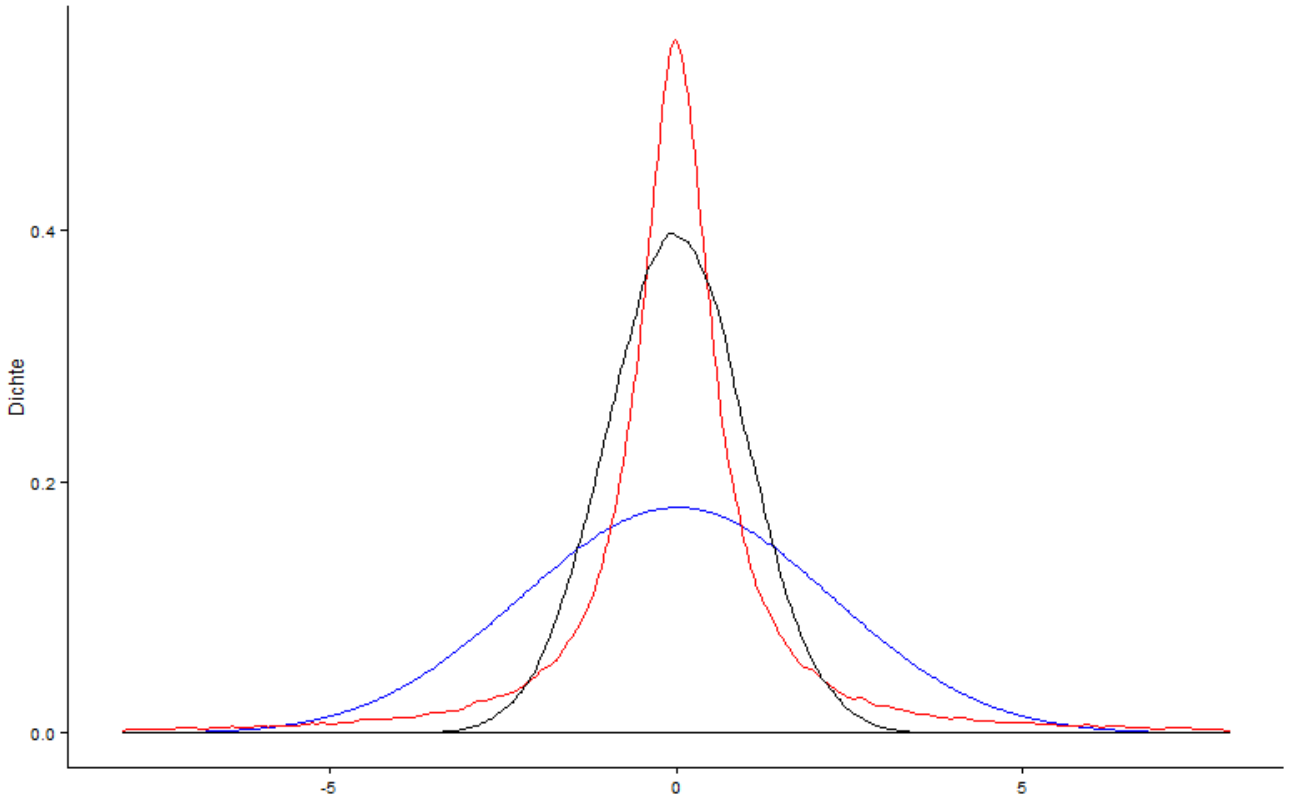
Neigt die Verteilung auf die linke Seite, spricht man von rechtsschief (blaue Kurve), neigt sie auf die rechte Seite, spricht man von linksschief (rote Kurve). Neigt die Kurve auf keine Seite, sie ist sie symmetrisch (schwarze Kurve).

Wölbung

Die Wölbung ist ein Mass für die Steilheit einer Verteilung. Eine geringe Wölbung bedeutet eine gleichmässige Streuung, bei einer hohen Wölbung besteht die Streuung aus mehr Extremwerten. Um die Werte besser vergleichen zu können, wird der Exzess berechnet, indem von der Wölbung 3 abgezogen wird. Ein Exzess von 0 bedeutet, dass die Verteilung normalgipflig (mesokurtisch) ist. Ein positiver Wert bedeutet, dass die Verteilung steilgipflig (leptokurtisch) ist. Bei einem negativen Exzess ist die Verteilung flachgipflig (platykurtisch).

Die Wölbung wird mit folgender Formel berechnet:

$$\omega = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s})^4$$



Die schwarze Linie ist die Normalverteilung. Sie ist mesokurtisch. Die blaue Linie ist platykurtisch. Die rote Linie ist leptokurtisch.

Variationsverhältnis

Bei nominal skalierten Variablen können die vorhergehenden Masse nicht berechnet werden. Immer berechnet werden kann jedoch das Variationsverältnis. Es gibt an, welcher Anteil der Werte nicht beim Modus liegen.

$$v = 1 - \frac{Anzahl\ Fälle\ beim\ Modus}{N}$$

Weiterführende Literatur

Weins, Cornelia (2010): Uni- und bivariate deskriptive Statistik. In: Christof Wolf und Henning Best (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlage. (Deutsch)