Statistik für Linguisten

Streuungsmaße

Andreas Blombach

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Kurze Zusammenfassung

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Frage

- Wenn blutrünstige Mutanten vom Mars im Durchschnitt sieben Augen haben, was wäre dann eure beste Schätzung dafür, wie viele Augen ein zufällig ausgewählter Marsmutant (nennen wir ihn Hubert) hat?

Zusammenfassung

- erster Überblick über Daten: graphische Darstellung
- Lagemaße (Maße der zentralen Tendenz) abhängig vom Skalenniveau
- bei mind. intervallskalierten Daten unbedingt Streuungsmaß angeben (besonders üblich: Standardabweichung)

Streuungsmaße

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Spannweite (range)

Spannweite der Daten (Minimal- bis Maximalwert)

- Stichprobe 1 (Schüler): MW 9,14
  - Alter: 7, 12, 9, 9, 8, 10, 9
  - Spannweite: 7-12
  - Differenz: 5
- Stichprobe 2 (Schüler): MW 9,43
  - Alter: 9, 9, 9, 9, 10, 10, 10
  - Spannweite: 9-10
  - Differenz: 1
- Stichprobe 3 (Schüler): MW 9,71
  - Alter: 6, 19, 19, 6, 6, 6, 6
  - Spannweite: 6-19
  - Differenz: 13

Quantile

- Aufsteigende Sortierung aller Werte einer Verteilung
- Quantile sind Schwellenwerte (und für sich nur Lage-, keine Streuungsmaße): z.B. 50%-Quantil: 50% aller Werte ≤ Schwellenwert

Beispielverteilung:

- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- 30%-Quantil: 3 (oder etwas größer, z.B. 3,7 – je nach Berechnungsart)
- 70%-Quantil: 7 (oder etwas größer, z.B. 7,3)

Interquartilsabstand

- Quartile: 25% (unteres Quartil), 50% (mittleres Quartil = Median) und 75 % (oberes Quartil)
- Interquartilsabstand (IQR): 0,75-Quartil – 0,25-Quartil; innerhalb dieses Bereichs liegen 50% der Werte
- Interquartilsabstand ist ein Streuungsmaß
- außerdem: Quintile, Dezile, Perzentile

Beispiel

- Stichprobe 1 (Schüler)
  - Alter: 7, 8, 9, 9, 9, 10, 12
  - Quartile: 8,5 (25%) – 9 (50%) – 9,5 (75%)
  - Interquartilsabstand: 1
- Stichprobe 2 (Schüler)
  - Alter: 9, 9, 9, 9, 10, 10, 10
  - Quartile: 9 – 9 – 10
  - Interquartilsabstand: 1
- Stichprobe 3 (Schüler)
  - Alter: 6, 6, 6, 6, 6, 19, 19
  - Quartile: 6 – 6 – 12,5
  - Interquartilsabstand: 6,5

Durchschnittliche Abweichung vom Mittelwert

- *Average absolute deviation*
- Durchschnitt der Differenz jedes einzelnen Datenpunkts zum Mittelwert

Durchschnittliche Abweichung

- Stichprobe 1 (Schüler)
  - Alter: 7, 12, 9, 9, 8, 10, 9
  - Mittelwert: 9,14
  - Mittlere Abweichung: 1,06
- Stichprobe 2 (Schüler)
  - Alter: 9, 9, 9, 9, 10, 10, 10
  - Mittelwert: 9,43
  - Mittlere Abweichung: 0,49
- Stichprobe 3 (Schüler)
  - Alter: 6, 19, 19, 6, 6, 6, 6
  - Mittelwert: 9,71
  - Mittlere Abweichung: 5,31

Formel

$$AD = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \bar{x} \right|$$

Es wird also von jedem einzelnen Wert der Mittelwert abgezogen, wobei negative Vorzeichen über den Betrag gelöscht werden, dann wird der Mittelwert aller Ergebnisse gebildet.
Die Betragsbildung ist notwendig, weil sich die Abweichungen vom Mittelwert sonst gegenseitig aufheben würden.

MAD

- Anstelle des Mittelwerts lässt sich hier auch der Median verwenden; dann erhält man den Median der Abweichungen vom Median (median absolute deviation, MAD).

$$MAD = \bar{x}_{median}\left(\left|x_i - \bar{x}_{median}\right|\right)$$

- Dieser Wert ist robuster gegenüber Ausreißern, also besonders hohen oder niedrigen Werten, als die üblicherweise verwendete Standardabweichung.
- Mit einer Anpassung kann er genutzt werden, um anhand einer Stichprobe die Standardabweichung in der Grundgesamtheit (robust) zu schätzen. Dazu wird der Wert mit einem Faktor multipliziert, dessen Wert davon abhängt, welche zugrundeliegende Datenverteilung angenommen wird – für normalverteilte Daten beträgt er 1,4826 (was auch der Standardwert von mad() in R ist).

Varianz

- Statt den absoluten Betrag zu bilden, kann man die Differenzen auch quadrieren, um einen positiven Wert zu erhalten. Damit lässt sich die **Varianz** berechnen, die Ausreißer stärker gewichtet als die durchschnittliche Abweichung vom Mittelwert.
- Formel (sehr ähnlich):

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

- Will man ausgehend von einer Stichprobe die Varianz in der Grundgesamtheit schätzen, wird allerdings nicht durch *n* geteilt, sondern durch *n – 1*. Man spricht dann von der **korrigierten** oder **unverzerrten Stichprobenvarianz s^2**.

Standardabweichung

- Zieht man die Wurzel aus der Varianz, erhält man die sog. **Standardabweichung**, ein sehr häufig verwendetes Streuungsmaß:

$$\sigma = \sqrt{\sigma^2}$$

- Warum die Wurzel? Weil die Einheit der Varianz etwas problematisch ist: Hätte man etwa die Varianz von Verkaufspreisen in Euro berechnet, würde die Varianz in „Quadrateuro“ gemessen.

Varianz und Standardabweichung

- korrigierte/unverzerrte Stichprobenvarianz:

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

- Standardabweichung:

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

Warum zur Hölle wird durch n – 1 geteilt statt durch n?

- Antwort 1: Weil uns diese Anpassung eine bessere Schätzung der Streuung in der Grundgesamtheit/Population erlaubt.
- Antwort 2: Das hat mit den sog. **Freiheitsgraden** zu tun … was leider nicht ganz so leicht zu verstehen ist.

Freiheitsgrade (degrees of freedom, df) (1)

- Nehmen wir an, ich möchte 3 Freunde besuchen, die ich selten sehe. Jedem davon möchte ich ein Mitbringsel schenken. Weil ich zu faul bin, mir ernsthaft Gedanken darüber zu machen, was ich wem mitbringen könnte, kaufe ich die erstbesten drei Dinge, die mir unterkommen: einen Roman, eine Flasche Wein und einen aufblasbaren Biber mit Sonnenbrille und Jetpack (der Laden ist ein wenig sonderbar).
- Nun besuche ich nacheinander meine drei Freunde und entscheide spontan, was jeder bekommt: Freund Nr. 1 erhält die Flasche Wein, bei Freund Nr. 2 entscheide ich mich für den Roman – und bei Freund Nr. 3 (der wirklich so heißt), habe ich keine Wahl mehr, sodass ich nur noch hoffen kann, dass er etwas für Biber übrig hat.
- Die Anzahl der Freiheitsgrade ist um 1 geringer als die Anzahl der Freunde.

Freiheitsgrade (2)

- Wenn ich fünf Werte habe und weiß, dass der Mittelwert 10 sein muss, dann kann ich die ersten Werte noch frei bestimmen, z.B.: 5, 8, 1, 12 – aber der fünfte Wert *muss* nun 24 sein, damit sich ein Mittelwert von 10 ergeben kann.
- Um zurück zur Varianz und zur Standardabweichung zu kommen: Wenn wir anhand unserer Stichprobe Varianz oder Standardabweichung der Grundgesamtheit schätzen wollen, brauchen wir in der Formel einen Mittelwert. Wir haben allerdings nur den der Stichprobe, den wir als Schätzung des Mittelwerts der Grundgesamtheit nehmen. Wenn wir das aber tun, ist dieser Wert festgelegt – und es können nicht mehr alle n Werte der Stichprobe frei variieren.
- Ergo: Freiheitsgrade = n – 1
- … und deshalb teilen wir durch n – 1.

Anmerkung zu den Symbolen

Oft werden für *Parameter* in der Grundgesamtheit/Population und *Statistiken* in der Stichprobe unterschiedliche Symbole verwendet.

	Mittelwert	Varianz	Standard-abweichung	Größe
Population	μ	σ^2	σ	N
Stichprobe	$\bar{x}$	s^2	s	n