Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Arithmetisches Mittel

Für eine vollständig gegebene Urliste

$$x_1, \ldots, x_n$$

berechnet man

$$\begin{aligned} \bar{x} &= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i \end{aligned}$$

Arithmetisches Mittel

Bei diskreten Merkmalen kann man dies umformen zu

$$\begin{aligned}\bar{x} &= \frac{1}{n}(h_1 \cdot a_1 + h_2 \cdot a_2 + \ldots + h_m \cdot a_m) \\ &= \frac{1}{n}\sum_{j=1}^{m} h_j \cdot a_j \\ &= \sum_{j=1}^{m} f_j \cdot a_j\end{aligned}$$

Beispiel: Noten einer Klassenarbeit

$$\bar{x} = \frac{1}{25}(3 \cdot 1 + 5 \cdot 2 + 8 \cdot 3 + 6 \cdot 4 + 2 \cdot 5 + 1 \cdot 6) = 3.08$$

Arithmetisches Mittel

Berechnung in R mit Funktion `mean()`

- Beispiel: Noten einer Klassenarbeit

```
x <- c(3,5,4,3,2,3,4,6,1,2,3,3,4,5,2,1,3,4,2,4,3,1,2,3,4)
mean(x)
[1] 3.08
```

Arithmetisches Mittel

Bei gruppierten Daten berechnet man approximativ

$$\bar{x} \approx \sum_{j=1}^{m} f_j \cdot x'_j$$

dabei ist x'_j die Klassenmitte der Klasse K_j, d. h.

$$x'_j = \frac{u_{j-1} + u_j}{2}$$

für $j = 1, \ldots, m$

Nulleigenschaft bzw. Schwerpunkteigenschaft

Die Abweichungen der Datenpunkte vom arithmetischen Mittel summieren sich zu Null (d. h. sie heben sich auf)

$$
\begin{aligned}
(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) &= \sum_{i=1}^{n} x_i - n \cdot \bar{x} \\
&= n \cdot \left[\frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x} \right] \\
&= n \cdot [\bar{x} - \bar{x}] \\
&= 0
\end{aligned}
$$

Minimumeigenschaft

Die Summe der quadrierten Abweichungen

$$\sum_{i=1}^{n}(x_i - a)^2$$

der Datenpunkte x_i von a ist minimal für $a = \bar{x}$

Diese Eigenschaft werden wir in Statistik 2 nachweisen, im Rahmen der Besprechung von Verfahren zur Parameterschätzung

Weitere Eigenschaften

Das arithmetische Mittel $\bar{x}$ hat viele günstige Eigenschaften (z. B. gegenüber dem Median $\tilde{x}$), die erst im Rahmen der Theorie der statistischen Parameterschätzung klar werden

Ein wesentlicher Nachteil von $\bar{x}$ gegenüber $\tilde{x}$ ist die Tatsache, dass das arithmetische Mittel empfindlich gegenüber Ausreißerwerten in den Daten ist

- Gelegentlich wird daher ein *getrimmtes arithmetisches Mittel* berechnet, bei dem ein gewisser Prozentsatz (z. B. 10%) der extremen Datenpunkte ausgeschlossen wird

Geometrisches Mittel

Definition (Voraussetzung: $x_i > 0$ für alle $i = 1, \ldots, n$)

$$\begin{aligned} GM &= \sqrt[n]{\prod_{i=1}^{n} x_i} \\ &= \sqrt[n]{x_1 \cdot \ldots \cdot x_n} \\ &= (x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}} \end{aligned}$$

Beziehung zu arithmetischer Mittelung

$$\log(GM) = \frac{1}{n} \sum_{i=1}^{n} \log(x_i)$$

Geometrisches Mittel

Berechnung in R

- Es existiert keine explizite Funktion zur Berechnung des geometrischen Mittels, wir können aber definieren

```
geometric_mean <- function(x) {
  exp(mean(log(x)))
}
```

Damit kann man dann berechnen

```
x <- c(2.2, 2.5, 3.0, 2.8, 3.2)
geometric_mean(x)
[1] 2.716
```

Geometrisches Mittel

Kommt in der Psychologie zum Einsatz bei

- Psychophysischen Fragestellungen: Mittelung einzelner Urteile bei Experimenten zur Größenschätzung (magnitude estimation, S. S. Stevens)
- multiplikativem Zusammenhang aufeinander folgender Merkmalsausprägungen (z. B. exponentielles Wachstum, konstante Lernrate, Inflationsrate) zur Bestimmung der durchschnittlichen Veränderung

Die Daten müssen mindestens verhältnisskaliert sein

Harmonisches Mittel

Definition (Voraussetzung: $x_i > 0$ für alle $i = 1, \ldots, n$)

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Kommt in der Psychologie zum Einsatz bei

- Psychophysischen Fragestellungen: Mittelung der Exponenten in der Stevensschen psychophysischen Potenzfunktion (Rule, 1969)
- Berechnung des Durchschnitts aus Indexzahlen (wie z. B. Kilometer pro Stunde, Preis pro Liter, Einwohner pro Quadratkilometer) bei konstanter Zählervariable (Kilometer, Preis, Einwohnerzahl)

Bedeutsamkeit des Vergleichs von arithmetischen Mittelwerten

$\varphi(a)$ sei die mittels einer Skala eingeschätzte Stärke von Hustensymptomen (z. B. Ratingskala von 1 bis 5)

Aussage: Die mittlere Symptomstärke in der Patientengruppe $a_1, \ldots, a_n$ ist größer als in der Patientengruppe $b_1, \ldots, b_m$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(a_i) > \frac{1}{m}\sum_{j=1}^{m}\varphi(b_j)$$

Unter welchen Voraussetzungen ist diese Aussage bedeutsam?

Bedeutsamkeit des Vergleichs von arithmetischen Mittelwerten

Die Aussage ist bedeutsam genau dann, wenn für alle zulässigen Transformationen f der Ratings φ gilt

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(a_i) > \frac{1}{m}\sum_{j=1}^{m}\varphi(b_j) \quad \text{gdw.} \quad \frac{1}{n}\sum_{i=1}^{n} f[\varphi(a_i)] > \frac{1}{m}\sum_{j=1}^{m} f[\varphi(b_j)]$$

Die Aussage ist bedeutsam, wenn φ

- aus einer Verhältnisskala stammt
- aus einer Intervallskala stammt

Die Aussage ist *nicht* bedeutsam, wenn φ

- aus einer Ordinalskala stammt

Bedeutsamkeit des Vergleichs von arithmetischen Mittelwerten

Wir betrachten die Situation, dass jeder von n Beobachtern mittels einer Skala die Stärke der Hustensymptome von Patienten einschätzt

$\varphi_i(a)$ bezeichne die Einschätzung des Patienten a durch den Beobachter i

Aussage: Die mittlere Einschätzung der Stärke der Symptome des Patienten a ist größer als die des Patienten b

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_i(a) > \frac{1}{n}\sum_{i=1}^{n}\varphi_i(b)$$

Ist diese Aussage bedeutsam?

Bedeutsamkeit des Vergleichs von arithmetischen Mittelwerten

Die Aussage ist bedeutsam genau dann, wenn für alle zulässigen Transformationen f_i der Skalen φ_i gilt

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_i(a) > \frac{1}{n}\sum_{i=1}^{n}\varphi_i(b) \quad \text{gdw.} \quad \frac{1}{n}\sum_{i=1}^{n}f_i[\varphi_i(a)] > \frac{1}{n}\sum_{i=1}^{n}f_i[\varphi_i(b)]$$

Auch wenn die φ_i allesamt aus Verhältnisskalen stammen, können die zulässigen Transformationen unabhängig voneinander gewählt werden (verschiedene Konstanten!)

- Die Aussage ist *nicht* bedeutsam

Bedeutsamkeit des Vergleichs von geometrischen Mittelwerten

Wenn die φ_i aus voneinander unabhängigen Verhältnisskalen stammen, dann ist ein Vergleich von geometrischen Mittelwerten bedeutsam

$$\sqrt[n]{\prod_{i=1}^{n} \varphi_i(a)} > \sqrt[n]{\prod_{i=1}^{n} \varphi_i(b)}$$

gdw.

$$\sqrt[n]{\prod_{i=1}^{n} f_i[\varphi_i(a)]} > \sqrt[n]{\prod_{i=1}^{n} f_i[\varphi_i(b)]}$$

Streuungsmaße (Dispersionsmaße)

Spannweite (Range)

$$R = x_{(n)} - x_{(1)}, \text{ also } R = x_{max} - x_{min}$$

Quartilabstand

$$Q_{dif} = Q_3 - Q_1 = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

Mittlere absolute Abweichung

- Ein naheliegendes Maß ist

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

- Wird wegen ungünstiger mathematischer Eigenschaften kaum verwendet

Varianz

Deutlich bessere Eigenschaften besitzt die empirische Varianz

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

Hinweis

- In vielen Lehrbüchern wird bei der Definition der Varianz die Summe der quadrierten Abweichungen durch $(n-1)$ geteilt, statt durch n
- Der Grund hierfür wird erst im Rahmen der Theorie der statistischen Parameterschätzung deutlich werden
- Die Kenngröße bei Division durch $n-1$ wird als *Stichprobenvarianz* bezeichnet und spielt in Statistik 2 eine zentrale Rolle

Varianz

Da die empirische Varianz kein Maß in der Einheit der Beobachtungswerte ist, sondern in deren Quadrat, betrachtet man die Quadratwurzel

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

Die Kenngröße $\tilde{s}$ wird *(empirische) Standardabweichung* genannt

Varianz

Beispiel: Höchsttemperatur (°C) an fünf Tagen

i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$\lvert x_i - \tilde{x}\rvert$	x_i^2
1	16	6	36	7	256
2	4	-6	36	5	16
3	13	3	9	4	169
4	9	-1	1	0	81
5	8	-2	4	1	64
$\sum$	50	0	86	17	586

$$\bar{x} = 10, \quad \tilde{x} = 9, \quad MAD = \frac{17}{5} = 3.4$$

$$\tilde{s}^2 = \frac{86}{5} = 17.2, \quad \tilde{s} = \sqrt{17.2} \approx 4.15$$

Varianz

Vereinfachte Berechnung der Varianz

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

Für das vorige Beispiel errechnet man so

$$\tilde{s}^2 = \frac{1}{5} \cdot 586 - 10^2$$
$$= 17.2$$

Varianz

Beweis der Rechenvereinfachung

$$
\begin{aligned}
\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2 \cdot x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \bar{x} + \frac{1}{n} \cdot n \cdot \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}
$$

Varianz und Standardabweichung: Berechnung in R

Die Funktionen `var()` und `sd()` geben die Stichprobenvarianz (d. h. Division durch $n-1$) und die zugehörige Standardabweichung (Quadratwurzel) an

Zur Berechnung der empirischen Varianz und Standardabweichung der Werte eines Vektors `x` mittels dieser Funktionen müssen diese resultierenden Werte korrigiert werden

```
var_tilde <- (n - 1)/n * var(x)
  s_tilde <- sqrt((n - 1)/n) * sd(x)
```

Alternativ dazu kann man direkt berechnen

```
var_tilde <- mean(x^2) - mean(x)^2  # mean((x - mean(x))^2)
  s_tilde <- sqrt(var_tilde)
```