FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Stochastik

FH Aachen – Studienstandort Aachen
28. November 2025

Prof. Dr. Florian Heinrichs

Lage- und Streumaße

Mittelwert

- Aus einer Umfrage kennen wir die Gehälter (in EUR) von $n = 21$ Personen

  3093, 5501, 4581, 3896, 5044, 4591, 3688,
  3977, 5487, 2812, 3031, 4084, 4134, 4796,
  3545, 2943, 3894, 3492, 4557, 4477, 3151

- Wie können wir die Daten zusammenfassen?
- Mit dem Mittelwert!
- Für reelle Zahlen $x_1, \ldots, x_n$ ist der Mittelwert definiert als

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Hier: $\bar{x}_n = 4037$
- Für Zufallsvariablen $X_1, \ldots, X_n$ ist $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

Lagemaß

- Der Mittelwert beschreibt die Lage einer “typischen” Beobachtung
- Es gibt unterschiedliche Parameter, die die Lage einer Stichprobe beschreiben
- Ein sinnvolles Lagemaß ℓ sollte *äquivariant* bzgl. Verschiebungen sein, d.h. für eine Stichprobe $x_1, \ldots, x_n$ und $a \in \mathbb{R}$

$$\ell(x_1 + a, \ldots, x_n + a) = \ell(x_1, \ldots, x_n) + a$$

- Der Mittelwert erfüllt diese Eigenschaft

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + a$$

Mittelwert

Bemerkung 21

Der Mittelwert minimiert den quadratischen Fehler

- Sei $x_1, \ldots, x_n$ eine Stichprobe
- $\bar{x}_n$ minimiert den Ausdruck $\sum_{i=1}^{n}(x_i - c)^2$, d.h.

$$\bar{x}_n = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n}(x_i - c)^2$$

- Beweis:
  - Definiere $f(c) = \sum_{i=1}^{n}(x_i - c)^2$
  - 1. Ableitung: $f'(c) = -2\sum_{i=1}^{n}(x_i - c)$
  - 2. Ableitung: $f''(c) = 2n$

$$f'(c) = 0 \iff \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} c \iff c = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$$

Mittelwert

Übung 78

Eine Studentin im ersten Semester misst die Anzahl der Stunden, die sie jede Woche für die Vorlesung Analysis I lernt. Nach dem Semester hat sie folgende Liste

$$0, 2, 3, 4, 3, 1, 2, 5, 1, 5, 4, 3.$$

Während der Klausurvorbereitung erhöht sich ihr Arbeitsaufwand auf $8, 12$ und 15 Stunden.

1. Berechnen Sie den Mittelwert über das Semester.
2. Berechnen Sie den Mittelwert unter Einbeziehung der Klausurvorbereitung.
3. Wie unterscheiden sich die Mittelwerte? Warum?

Mittelwert

Bemerkung 22

Wie hängt der Mittelwert mit dem Erwartungswert zusammen?

- Sei $x_1, \ldots, x_n$ eine Stichprobe
- Sei Y gleich verteilt auf der Stichprobe, d.h. $Y \sim \mathcal{U}_{\{x_1,\ldots,x_n\}}$
- Dann ist $\mathbb{E}[Y] = \bar{x}_n$
- Beweis:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{i=1}^{n} x_i \mathbb{P}(Y = x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n\end{aligned}$$

Ordnungsstatistik

Definition 51

Seien $X_1, \ldots, X_n$ Zufallsvariablen. Die Zufallsvariablen $X_{(1)}, \ldots, X_{(n)}$ heißen *Ordnungsstatistik* der Zufallsvariablen $X_1, \ldots, X_n$, falls für jedes $\omega \in \Omega$ gilt

$$X_{(1)}(\omega) \leq X_{(2)}(\omega) \leq \cdots \leq X_{(n)}(\omega)$$

und für jedes i ein k_i existiert, sodass $X_{(i)}(\omega) = X_{k_i}(\omega)$. $X_{(i)}$ ist die i-te *Ordnungsstatistik der Zufallsvariablen* $X_1, \ldots, X_n$.

Bemerkung 23

1. $X_{(1)} = \min\{X_1, \ldots, X_n\}$
2. $X_{(n)} = \max\{X_1, \ldots, X_n\}$
3. Allgemein: Eine *Statistik* ist eine messbare Abbildung

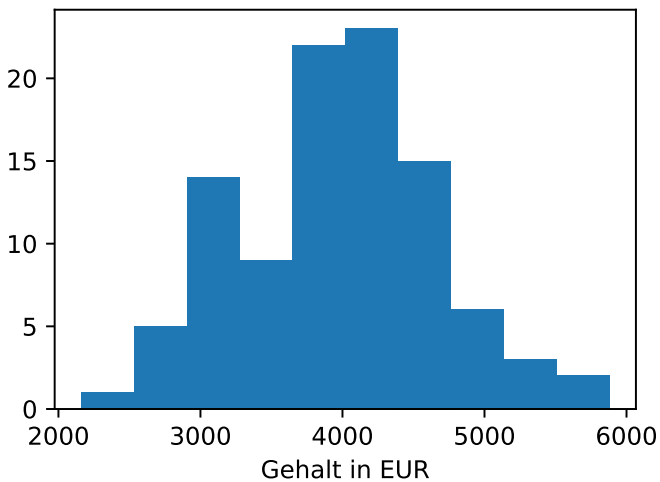
Ordnungsstatistik

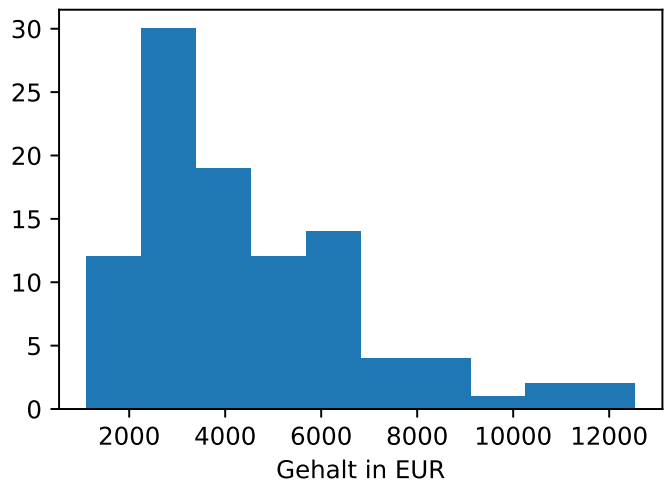
Bemerkung 24

- Für $\omega \in \Omega$ ist $X(\omega)$ eine **Realisierung** der Zufallsvariablen X
- Eine Stichprobe $x_1, \ldots, x_n$ ist eine Realisierung der Zufallsvariablen $X_1, \ldots, X_n$
- Die Ordnungsstatistik einer Stichprobe $x_1, \ldots, x_n$ ist $x_{(1)} \leq \cdots \leq x_{(n)}$
- Beispiel:
  - Sei eine Stichprobe $5, 1, 3, 4, 2$ gegeben
  - Die Ordnungsstatistik ist $x_{(1)} = 1, x_{(2)} = 2, \ldots, x_{(5)} = 5$

Median

- Fasst der Mittelwert die Daten hier gut zusammen?
- Nein: Der Mittelwert ist **nicht robust** gegenüber Ausreißern
- Besser: Median
- Median: Kennzahl, sodass 50% der Stichprobe jeweils kleiner und größer gleich dem Median sind
- Beispiel (rechts):
  - Mittelwert: 4528.37, Median: 3938.85





Median

Definition 52

Seien $X_1, \ldots, X_n$ Zufallsvariablen mit Ordnungsstatistik $X_{(1)}, \ldots, X_{(n)}$. Der *empirische Median* (auch *Stichprobenmedian* oder *Median der Stichprobe*) ist definiert als

$$med(X_1, \ldots, X_n) = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & \text{falls } n \text{ gerade} \end{cases}$$

> Der Median ist äquivariant bzgl. Verschiebungen

$$med(X_1 + a, \ldots, X_n + a) = med(X_1, \ldots, X_n) + a$$

> Der Median ist so gewählt, dass min. 50% der Stichprobe größer und kleiner sind

Median

Beispiel 132

Gegeben sei die Stichprobe $3.80, 4.33, 4.96, 3.71, 3.40$. Was ist der Median?

- $>$ Ordnungsstatistik: $3.40, 3.71, 3.80, 4.33, 4.96$
- $>$ Median: $n = 5$ ungerade und $\frac{n+1}{2} = 3$, also $med(X) = X_{(3)} = 3.80$

Median

Übung 79

Gegeben sei die Stichprobe $4.31, 4.82, 3.69, 4.02, 3.98, 3.57$. Bestimmen Sie die Ordnungsstatistik und den Median.

Übung 80

Gegeben seien die folgenden Verspätungen eines ICEs (in Minuten)

$$2, 5, 3, 185, 5, 4, 4, 3, 4, 7, 9, 175, 4, 6, 4, 9.$$

Bestimmen Sie den Mittelwert und Median der Verspätungen. Welche Größe beschreibt eine "typische" Verspätung besser?

Empirische Quantile

Wir wissen bereits dass der Median einer Verteilung ihrem 50%-Quantil entspricht. Das selbe gilt für eine Stichprobe:

Definition 53

Seien $X_1, \ldots, X_n$ Zufallsvariablen mit Ordnungsstatistik $X_{(1)}, \ldots, X_{(n)}$. Das *empirische p-Quantil* (auch *Stichprobenquantil* oder *p-Quantil* der Stichprobe) ist definiert als

$$q_p = \begin{cases} X_{(\lfloor n \cdot p + 1 \rfloor)} & \text{falls } n \cdot p \notin \mathbb{N} \\ \frac{1}{2}\left(X_{(n \cdot p)} + X_{(n \cdot p + 1)}\right) & \text{falls } n \cdot p \in \mathbb{N} \end{cases}$$

> Das empirische p-Quantil ist äquivariant bzgl. Verschiebungen

> Der Median entspricht dem 50%-Quantil:

$$med(X_1, \ldots, X_n) = q_{0.5}$$

Modalwert

Der häufigste Wert in einer Stichprobe heißt *Modus*. Genauer:

Definition 54

Sei $x_1, \ldots, x_n$ die Stichprobe eines diskreten Merkmals mit Klassen $c_1, \ldots, c_d$. Bezeichne mit h_i die Häufigkeit der Klasse c_i in der Stichprobe, d.h.

$$h_i = |\{x_j \in c_i : j = 1, \ldots, n\}|$$

für $i = 1, \ldots, d$. Dann ist die Klasse c_k ein *Modus* (auch: *Modalwert*) der Stichprobe, falls $h_k = \max_{i=1}^{d} h_i$.

> Der Modus ist äquivariant bzgl. Verschiebungen

> Der Modus kann auch für stetige Verteilungen mit Dichte f definiert werden: $x_m = \arg\max f(x)$

Streumaß

- Nach dem 1. Semester wurden 20 Studierende nach ihrer Punktzahl (von 100) gefragt

Analysis	41	47	46	48	50	49	50	47	42	45
	90	97	93	90	100	95	95	100	93	90
Lineare Algebra	62	64	62	79	73	62	72	72	84	73
	70	62	78	66	81	67	82	66	67	66

- Der Mittelwert beider Klausuren ist 70.4
- Der Median ist jeweils 70.0 bzw. 68.5
- Sind die Klausuren also etwa gleich ausgefallen?
- Nein! $\rightarrow$ Berücksichtige Streuung!

Streumaß

Ein sinnvolles Streumaß (auch Streuungsmaß) σ sollte *invariant* bzgl. Verschiebungen sein, d.h. $\sigma(x_1 + a, \ldots, x_n + a) = \sigma(x_1, \ldots, x_n)$.

Definition 55: (Strichprobenvarianz)

Seien $X_1, \ldots, X_n$ Zufallsvariablen mit Mittelwert $\bar{X}_n$. Die *empirische Varianz* (auch *Stichprobenvarianz* oder *Varianz der Stichprobe*) ist definiert als

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Die Wurzel der Varianz heißt *empirische Standardabweichung* $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$.

Die empirische Varianz s^2 (und damit auch s) ist invariant bzgl. Verschiebungen.

Empirische Varianz

Bemerkung 25

Wie hängt die empirische Varianz mit der Varianz einer Verteilung zusammen?

- Sei $x_1, \ldots, x_n$ eine Stichprobe
- Sei Y gleich verteilt auf der Stichprobe, d.h. $Y \sim \mathcal{U}_{\{x_1,\ldots,x_n\}}$
- Dann ist $\text{var}(Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$
- **Achtung:** $\text{var}(Y) \neq s^2$
- Es gilt $s^2 = \frac{n-1}{n} \text{var}(Y)$, also $s^2 \approx \text{var}(Y)$ für n groß
- Die Skalierung mit $\frac{1}{n-1}$ hat theoretische Vorteile

Empirische Varianz

Beispiel 133

Nach dem 1. Semester wurden 10 Studierende nach ihren Klausurergebnissen gefragt. Was ist die Varianz der Ergebnisse?

i	1	2	3	4	5	6	7	8	9	10
x_i (Analysis)	41	47	46	48	50	93	92	100	93	90
y_i (Lin. Algebra)	61	64	62	78	73	62	71	72	84	73
$x_i - 70$	-29	-23	-24	-22	-20	23	22	30	23	20
$(x_i - 70)^2$	841	529	576	484	400	529	484	900	529	400

- Es gilt $\bar{x}_{10} = \bar{y}_{10} = 70$.
- $\sum_{i=1}^{10}(x_i - 70)^2 = 5672$
- $s^2 = 630.22$

Empirische Varianz

Übung 81

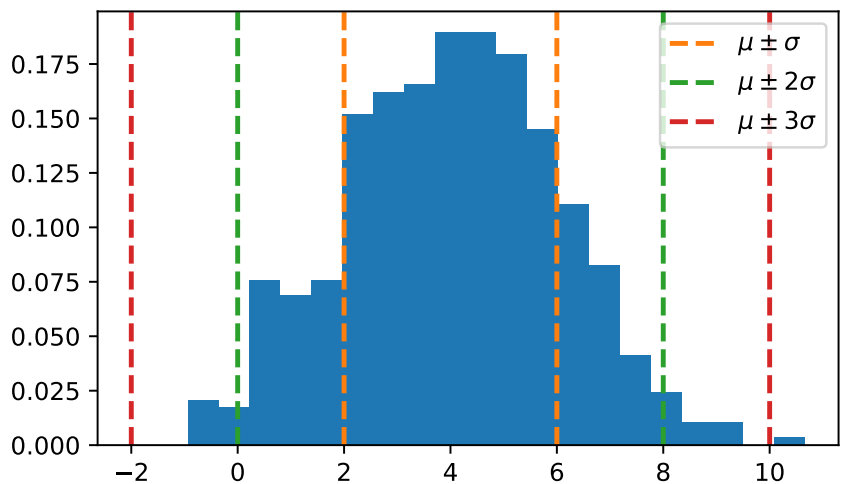
Nach dem 1. Semester wurden 10 Studierende nach ihren Klausurergebnissen gefragt. Berechnen Sie die Stichprobenvarianz der Ergebnisse zur linearen Algebra.

i	1	2	3	4	5	6	7	8	9	10
x_i (Analysis)	41	47	46	48	50	93	92	100	93	90
y_i (Lin. Algebra)	61	64	62	78	73	62	71	72	84	73
$y_i - 70$										
$(y_i - 70)^2$										

> Es gilt $\bar{x}_{10} = \bar{y}_{10} = 70$.
>
> $\sum_{i=1}^{10}(y_i - 70)^2 =$
>
> $s^2 =$

Empirische Varianz

- Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, dann gilt
  - $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6827$
  - $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
  - $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$
- Interpretation: Falls wir Realisierungen von $\mathcal{N}(\mu, \sigma^2)$ beobachten, liegen
  - ca. 68.27% der Beobachtungen innerhalb des Intervalls $[\mu - \sigma, \mu + \sigma]$
  - ca. 95.45% in $[\mu - 2\sigma, \mu + 2\sigma]$
  - ca. 99.73% in $[\mu - 3\sigma, \mu + 3\sigma]$



Andere Streumaße

Definition 56: (Mittlere Absolute Abweichung, Stichprobenspannweite)

Seien $X_1, \ldots, X_n$ Zufallsvariablen mit Mittelwert $\bar{X}_n$, Minimum $X_{(1)}$ und Maximum $X_{(n)}$.

1. Die *mittlere absolute Abweichung* (engl. *mean absolute deviation*) ist definiert als

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}_n|.$$

2. Die *Stichprobenspannweite* (engl. *range*) ist definiert als

$$r = X_{(n)} - X_{(1)} = \max\{X_1, \ldots, X_n\} - \min\{X_1, \ldots, X_n\}.$$

Beide Streumaße sind invariant bzgl. Verschiebungen.

Interquartilsabstand

- Aus einer Umfrage kennen wir die Gehälter (in EUR) von $n = 10$ Personen

$$4051, 4790, 10805, 3919, 3823, 5041, 5227, 4947, 4736, 4994$$

- Der Mittelwert beträgt 5233.30 EUR, entspricht dieser Wert einem "typischen" Einkommen?
- Der Median beträgt 4848.50 EUR und ist repräsentativer
- Der Median ist *robust* gegenüber Ausreißern, der Mittelwert nicht
- Die Standardabweichung beträgt 1917.76 EUR
  - Wie können wir diesen Wert interpretieren?
  - Ohne den hohen Wert (10805 EUR) beträgt die Standardabweichung 503.87 EUR
  - Wie können wir diesen Unterschied interpretieren?
- Die Standardabweichung ist nicht robust gegenüber Ausreißern
- Stattdessen: Nutze Interquartilsabstand (basierend auf Quartilen)

Interquartilsabstand

Definition 57: (Interquartilsabstand)

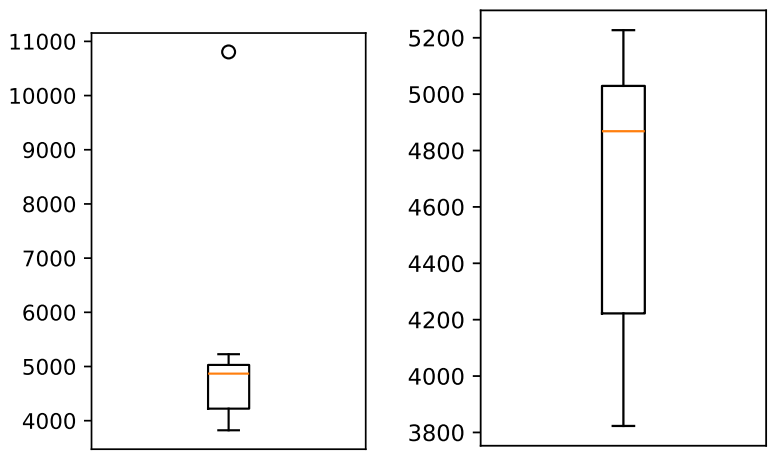
Seien $X_1, \ldots, X_n$ Zufallsvariablen mit 25%- und 75%-Quantilen $q_{0.25}$ und $q_{0.75}$. Der *Interquartilsabstand* ist definiert als

$$IQR = q_{0.75} - q_{0.25}.$$

> Aus einer Umfrage kennen wir die Gehälter (in EUR) von $n = 10$ Personen:
>
> $$4051, 4790, 10805, 3919, 3823, 5041, 5227, 4947, 4736, 4994$$
>
> > Der IQR für die gesamten Daten beträgt 990 EUR
> > Der IQR ohne Ausreißer beträgt 943 EUR
>
> Der Interquartilsabstand ist invariant bzgl. Verschiebungen
>
> Welcher Anteil der Stichprobe liegt zwischen $q_{0.25}$ und $q_{0.75}$?
> min. 50%

Box-Plot

- Stichprobe: $4051, 4790, 10805, 3919, 3823, 5041, 5227, 4947, 4736, 4994$
- Wie können wir die Daten visualisieren?
- Mit Hilfe eines Box-Plots
  - Untere Grenze der Box: $q_{0.25}$
  - Obere Grenze der Box: $q_{0.75}$
  - Markierung in der Mitte: $q_{0.5}$ (Median)
  - Unterer *Whisker* (seltener: *Fühler*): Kleinste Beobachtung in Stichprobe, die größer oder gleich $q_{0.25} - 1.5 \cdot IQR$ ist
  - Oberer *Whisker*: Größte Beobachtung in Stichprobe, die kleiner oder gleich $q_{0.75} + 1.5 \cdot IQR$ ist
  - Ausreißer: Punkte außerhalb von $[q_{0.25} - 1.5 \cdot IQR, q_{0.75} + 1.5 \cdot IQR]$



Woher kommt die 1.5? $\rightarrow$ Übliche Wahl (aber willkürlich)

Interquartilsabstand

Übung 82

Gegeben seien die Gehälter (in EUR) von $n = 10$ Personen:

$$4051, 4790, 10805, 3919, 3823, 5041, 5227, 4947, 4736, 4994.$$

Berechnen Sie den Interquartilsabstand $IQR = q_{0.75} - q_{0.25}$.

Beachten Sie, dass die empirischen Quartile für Zufallsvariablen $X_1, \ldots, X_n$ mit Ordnungsstatistik $X_{(1)}, \ldots, X_{(n)}$ definiert sind als

$$q_{0.25} = \begin{cases} X_{(\lfloor 0.25 \cdot n + 1 \rfloor)} & \text{falls } 0.25 \cdot n \notin \mathbb{N} \\ \frac{1}{2}\left(X_{(0.25 \cdot n)} + X_{(0.25 \cdot n + 1)}\right) & \text{falls } 0.25 \cdot n \in \mathbb{N} \end{cases}$$

$$q_{0.75} = \begin{cases} X_{(\lfloor 0.75 \cdot n + 1 \rfloor)} & \text{falls } 0.75 \cdot n \notin \mathbb{N} \\ \frac{1}{2}\left(X_{(0.75 \cdot n)} + X_{(0.75 \cdot n + 1)}\right) & \text{falls } 0.75 \cdot n \in \mathbb{N} \end{cases}$$

Variationskoeffizient

Beispiel 134: Kurse der Alphabet Inc. Aktie

Schlusskurse der Alphabet Inc. Aktie (Umrechnung: 1 USD $\hat{=}$ 1.1612 EUR) im Jahr 2025:

Datum	29.09.	30.09.	01.10.	02.10.	03.10.	06.10.	07.10.	08.10.
Kurs (USD)	244,05	243,10	244,90	245,69	245,35	250,43	245,76	244,62
Kurs (EUR)	210,17	209,35	210,90	211,58	211,29	215,66	211,64	210,66

- Empirische Varianz der Preise (in USD): $s^2_{USD} \approx 4.775$
- Empirische Varianz der Preise (in EUR): $s^2_{EUR} \approx 3.541$
- s^2_{USD} ist größer als s^2_{EUR}, ist der Aktienpreis in USD volatiler? $\rightarrow$ Nein! Die Preise haben eine andere Größenordnung!

Variationskoeffizient

Definition 58: (Variationskoeffizient)

Seien $X_1, \ldots, X_n$ Zufallsvariablen mit Mittelwert $\bar{X}_n$ und empirischer Varianz s^2. Der *(empirische) Variationskoeffizient* ist definiert als

$$V = \frac{s}{\bar{X}_n}.$$

Der Variationskoeffizient beschreibt die *relative* Streuung und bleibt unverändert, wenn jeder Wert mit einer Zahl $\alpha > 0$ multipliziert wird.

Beispiel 134: Kurse der Alphabet Inc. Aktie

- USD: $\bar{X}_n^{USD} \approx 245.49$, $s^2_{USD} \approx 4.775$
- EUR: $\bar{X}_n^{EUR} \approx 211.41$, $s^2_{EUR} \approx 3.541$
- Variationskoeffizient: $V_{USD} = 0.0089 = V_{EUR}$

Literatur I

Bundesministerium für Gesundheit (2024).
Infektionsradar.
https://infektionsradar.gesund.bund.de/de/covid/inzidenz.
Abgerufen: 2024-10-15.

Dehling, H. and Haupt, B. (2006).
Einführung in die Wahrscheinlichkeitstheorie und Statistik.
Springer-Verlag.

Literatur II

Dinnes, J., Sharma, P., Berhane, S., van Wyk, S., Nyaaba, N., Domen, J., Taylor, M., Cunningham, J., Davenport, C., Dittrich, S., Emperador, D., Hooft, L., Leeflang, M., McInnes, M., Spijker, R., Verbakel, J., Takwoingi, Y., Taylor-Phillips, S., Van den Bruel, A., and Deeks, J. (2022).
Rapid, point-of-care antigen tests for diagnosis of sars-cov-2 infection.
Cochrane Database of Systematic Reviews, (7).

Henze, N. et al. (1997).
Stochastik für Einsteiger, volume 4.
Springer.