

M24 Statistik 1: Sommersemester 2024

# Vorlesung 03: Lage- und Streuungsmaße

Prof. Matthias Guggenmos

Health and Medical University Potsdam



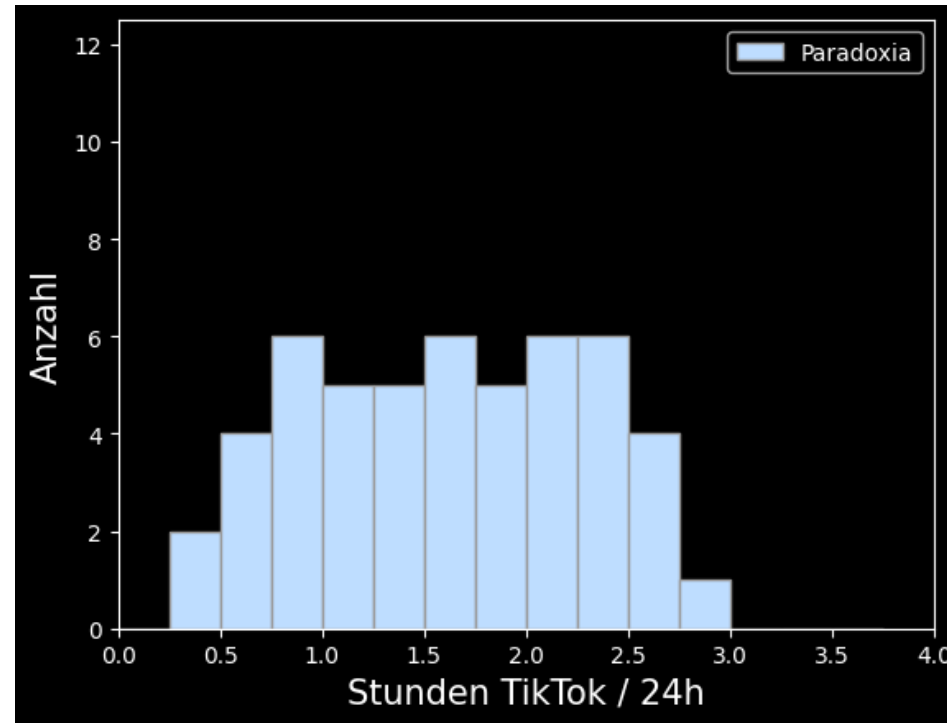


Die Daten der ersten Beobachtungsstudie zu Paradoxia sind frisch eingetroffen!

Table 1. Results.

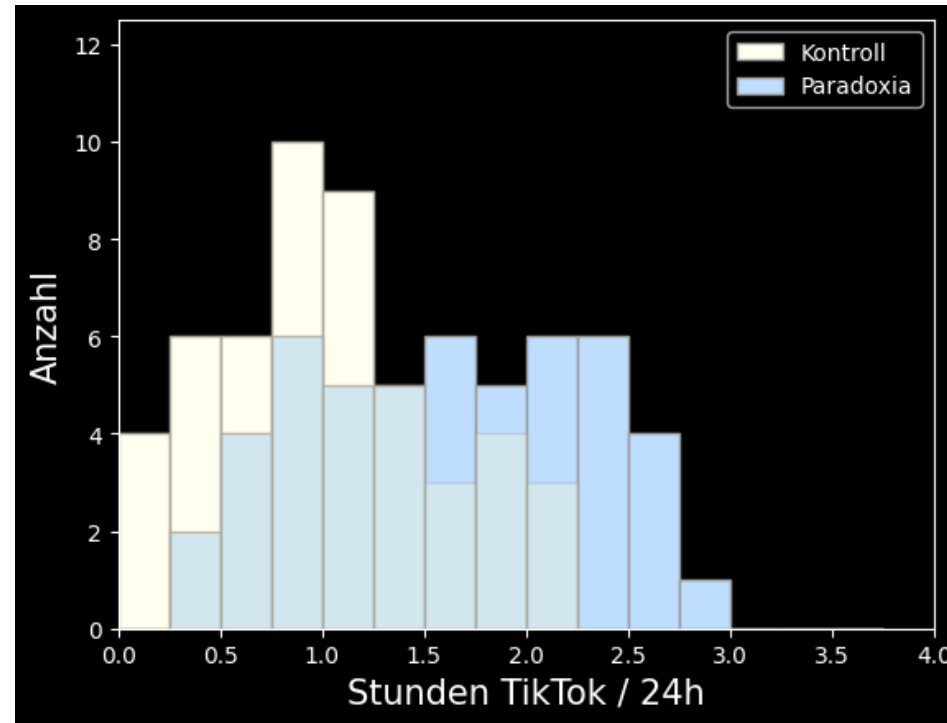
id	group	hours_tiktok_per_day	inflammation
1	control	1.14	0.24
2	control	2.24	0.19
...	...	...	...
50	control	1.14	0.13
51	paradoxia	1.57	0.03
52	paradoxia	2.59	0.21
...	...	...	...
100	paradoxia	0.52	0.12

Hier ist das Histogramm der TikTok-Zeiten von Paradoxikern:



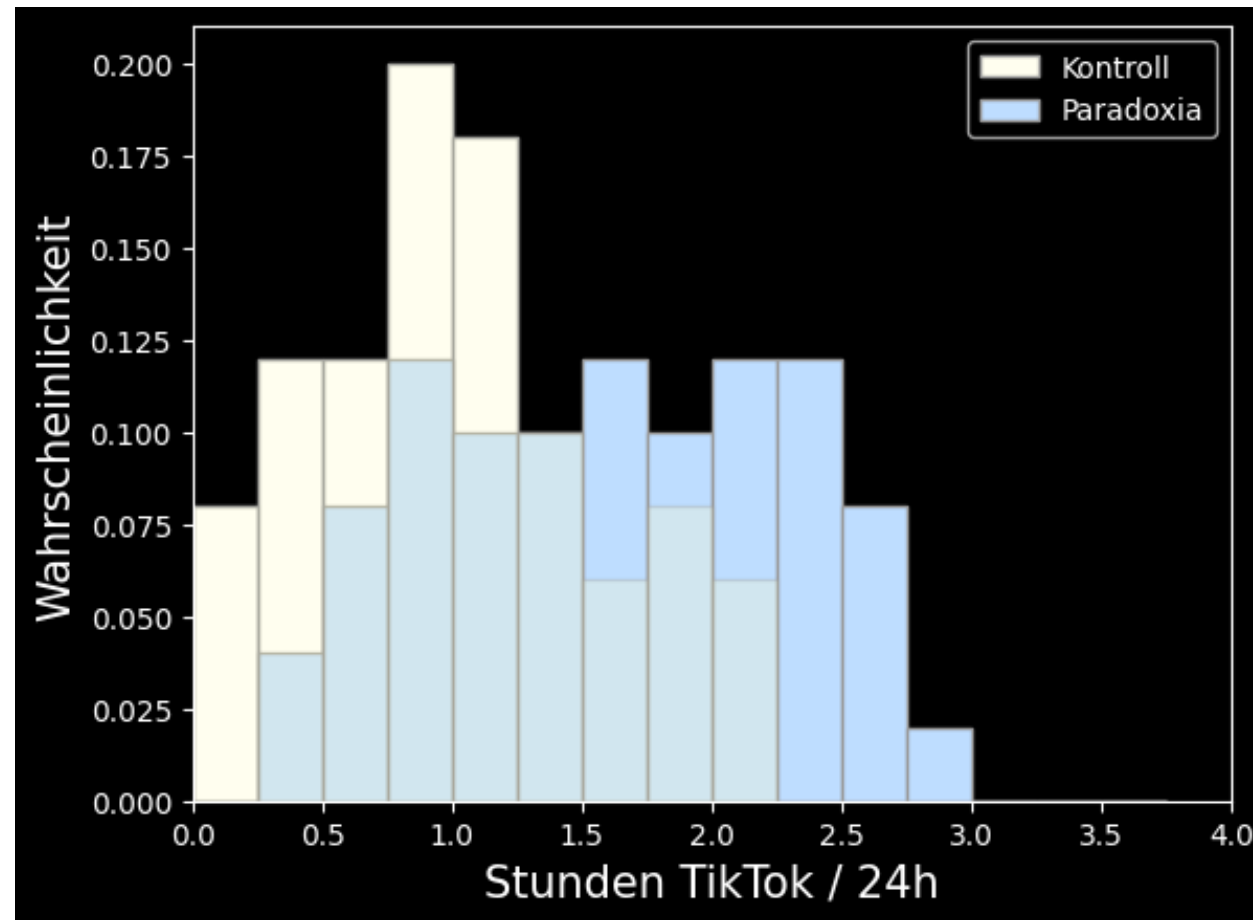
Überschlagen Sie: passt das Histogramm zur angegebenen Stichprobe von  $n=50$  Paradoxikern? Und handelt es sich um eine Abbildung relativer oder absoluter Häufigkeit?

Vergleich mit der Kontrollgruppe:



Wir können hier schon erahnen, dass die Studie tatsächlich Evidenz für einen erhöhte TikTok-Zeit bei Paradoxikern erbringt (Hypothese 1)!

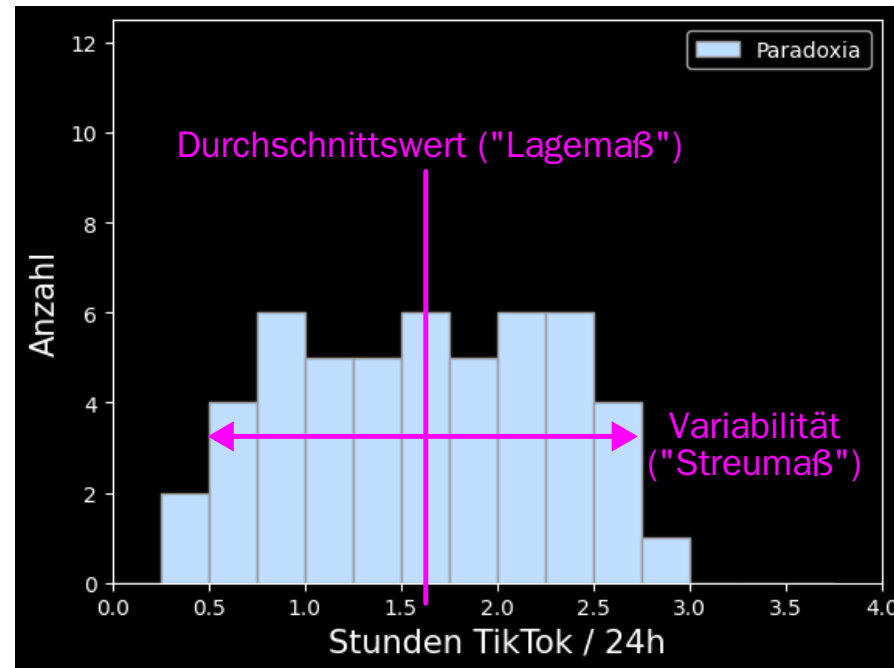
Erinnerung: statt der Anzahl (absolute Häufigkeit) kann auch die Wahrscheinlichkeit (relative Häufigkeit) dargestellt werden:



Jeder Wert in dieser Abbildung gibt also die Wahrscheinlichkeit an, dass ein Entzündungswert im Intervall des jeweiligen Balkens liegt.

Während sich die Balken eines Histogramms mit absoluter Häufigkeit (Anzahl) zur Stichprobengröße aufaddieren, addieren sie sich beim Histogramm mit relativer Häufigkeit (Wahrscheinlichkeit) zu 1.

Die ersten Daten sind also eingetroffen, und Sie machen sich nun an die Auswertung. Das führt zu Sie zum Thema der heutigen Vorlesung: **wie kann man Daten statistisch beschreiben?**



# Anwesenheitsliste



# Notation



# Drei Fälle bei der Berechnung statistischer Kennwerte

## Fall 1: Populationskennwerte

Vollständige Population liegt vor (z.B. alle US-Präsidenten)

⇒ Bestimmung exakter statistische Kennwerte für die Population

⇒ *Populationskennwerte* (z.B. Populationsstreuung)

(Hinweis: man spricht auch manchmal von *Populationsparametern*)

## Fall 2: Stichprobenkennwerte

Stichprobe liegt vor

⇒ Ziel ist Beschreibung der Stichprobe *ohne Inferenz auf die Population*

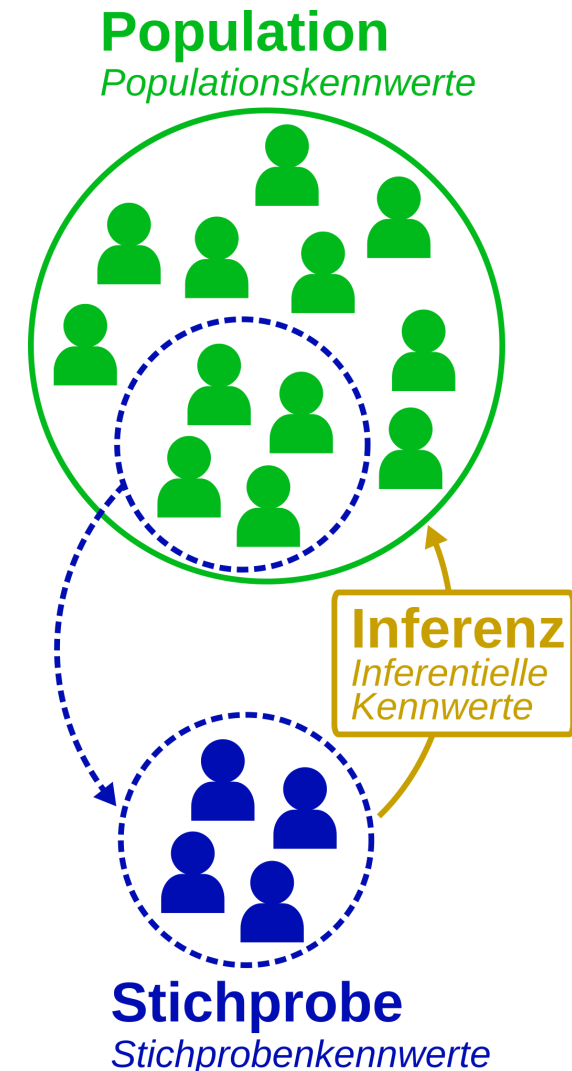
⇒ *Stichprobenkennwerte* (z.B. Stichprobenstreuung)

## Fall 3: Inferentielle Kennwerte

Stichprobe liegt vor

⇒ Ziel ist *Inferenz auf die Population*, d.h. die Populationskennwerte sollen auf Basis der Stichprobe geschätzt werden.

⇒ *Inferentielle Kennwerte* (z.B. inferentielle Streuung)



# Notation / Variablenbenennung

- Folgendes Schema hat sich in der Literatur etabliert:
  - **Populationskennwerte**: griechische Buchstaben
  - **Stichprobenkennwerte**: lateinische Buchstaben
  - **Inferentielle Kennwerte**: griechische Buchstaben mit Zirkumflex ( $\hat{\phantom{x}}$ ).

	<b>Populations- Kennwert</b>	<b>Stichproben- Kennwert</b>	<b>Inferentieller Kennwert</b>
Mittelwert	$\mu$	$\bar{x}, m, M$	$\hat{\mu}$
Standardabweichung	$\sigma$	$s$	$\hat{\sigma}$
Varianz	$\sigma^2$	$s^2, Var(X)$	$\hat{\sigma}^2, \hat{Var}(X)$
Kovarianz	$\sigma_{XY}$	$s_{XY}, Cov(X, Y)$	$\hat{\sigma}_{XY}, \hat{Cov}(X, Y)$
Korrelation	$\rho$	$r$	$\hat{\rho}$

- Grund für die aufwendige Unterscheidung in der Notation:
  - Kommunikation des Analyseziels in einer Formel
  - Formeln für Populations- und Stichprobenkennwerte sind identisch, die Formeln für inferentielle Kennwerte können jedoch verschieden sein!
  - (Ausnahme: beim Mittelwert gilt  $\hat{\mu} = \bar{x}$ )

# Hinweis

Um die Komplexität zu reduzieren, werden wir **in dieser Vorlesungsreihe konsequent Formeln für inferentielle Kennwerte behandeln** (zu erkennen am  $\wedge$ ). Dies ist der mit Abstand häufigste Fall in der Psychologie und de facto in allen empirischen Wissenschaften.



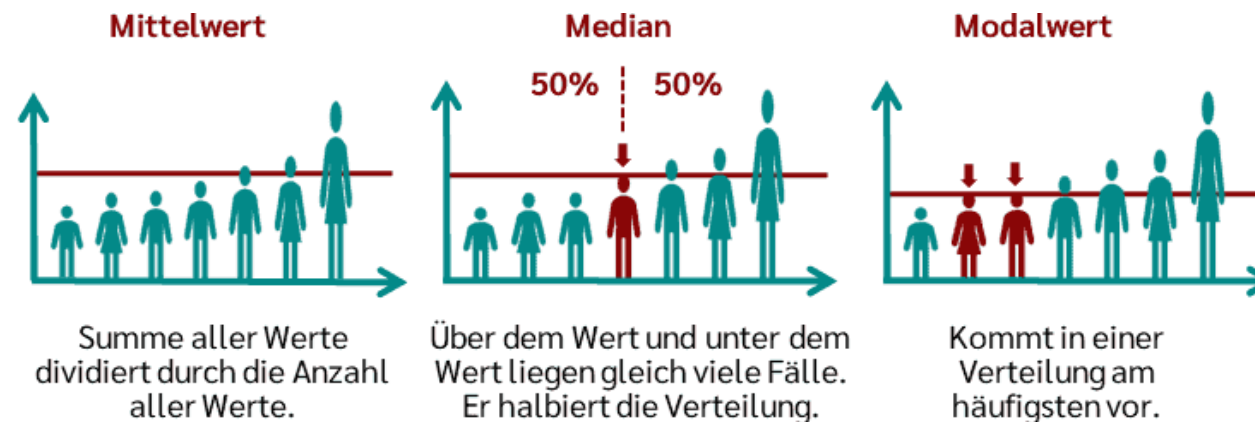
In der Literatur (Lehrbücher, Statistikportale, Wikipedia, ...) herrscht ein heilloses Durcheinander bezüglich der Verwendung der Begriffe Populationskennwert (z.B. Populationsvarianz) und Stichprobenkennwert (z.B. Stichprobenvarianz). Beide Begriffe werden zum Teil dafür verwendet, was hier als inferentieller Kennwert bezeichnet wird. Eine Ursache ist, dass sich für das Konzept des “inferentiellen Kennwerts” bislang kein etablierter Begriff herausgebildet hat. Den Begriff “inferentieller Kennwert” werden Sie daher bei Google nicht finden.

Die Lösung dieses Dilemmas in dieser Veranstaltung besteht eben in der Einführung des Begriffs **inferentieller Kennwert**. In dieser Logik beschreibt ein Populationskennwert die Population, ein Stichprobenkennwert beschreibt die Stichprobe, und ein inferentieller Kennwert schätzt die Populationskennwerte auf Basis der Stichprobe (“Inferenz auf die Population”).

# Lagemaße

# Lagemaße

- Der weithin bekannte **Durchschnittswert** oder **Mittelwert** ist ein Beispiel für ein **Lagemaß** von Verteilungen.
- Man spricht dabei auch von der **zentralen Tendenz** (engl. *central tendency*) einer Verteilung
- Die drei wichtigsten Lagemaße sind:
  - Mittelwert (gemeint ist das *arithmetische Mittel*)
  - Median
  - Modus (auch Modalwert)



Bildnachweis<sup>1</sup>

# Mittelwert

- Berechnung (verbal):
  1. Addiere alle  $n$  Stichprobenwerte
  2. Teile durch die Anzahl der Werte
- Berechnung (Formel):

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

- Notation: Strich über dem kleinen Letter der Zufallsvariable = Mittelwert der Zufallsvariable

## Beispiel

Folgende Beobachtungen der Zufallsvariable  $X$  "Punktzahl in der Abi-Matheprüfung" werden in einer Stichprobe von 7 Psychologiestudierenden gemacht:  $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14\}$

$$\bar{x} = \frac{1}{7} (13 + 7 + 15 + 8 + 4 + 9 + 14) = \frac{1}{7} \cdot 70 = 10$$

# Wann ist der Mittelwert sinnvoll?

- Der Mittelwert ist ein sinnvolles Lagemaß, wenn er nicht durch einzelne **Ausreißer** (extreme Werte) dominiert bzw. verzerrt wird.

$x$ (z.B. Punktzahlen im Abi)	$\bar{x}$	Histogramm	Mittelwert sinnvoll?
{13, 7, 15, 8, 4, 9, 14}	10		✓
{3, 1, 4, 2, 2, 6, 15}	4.7		✗
{13, 15, 11, 12, 9, 14, 1}	10.7		✗

# Median

- Gibt es dominante Ausreißer in den Daten, so ist häufig der **Median** das sinnvollere Lagemaß
- Berechnung:
  1. Alle  $n$  Stichprobenwerte der Größe nach aufreihen
  2. Der Wert, der genau in der Mitte liegt, ist der Median
  3. Bei einer geraden Anzahl von Werten bilden zwei Werte die Mitte – in diesem Fall ist der Median der Mittelwert dieser beiden Werte
  4. (Alternativ mit Formel:  $Tiefe_{\text{Median}} = \frac{n+1}{2}$ )
- Der Median wird mit einer Tilde ( $\sim$ ) über der Variablen bezeichnet:  $\tilde{x}, \tilde{y}, \dots$

## Beispiel

Folgende Beobachtungen der Zufallsvariable  $X$  "Punktzahl in der Abi-Matheprüfung" werden in einer Stichprobe von 7 Psychologiestudierenden gemacht:  $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14\}$

Sortierte Reihenfolge:  $\mathbf{x} = \{4, 7, 8, \mathbf{9}, 13, 14, 15\} \rightarrow \tilde{x} = 9$

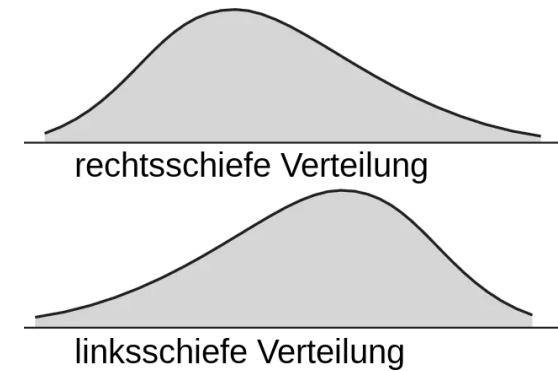
Wir fügen den Wert eines weiteren Studierenden hinzu:  $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14, 10\}$

Sortierte Reihenfolge:  $\mathbf{x} = \{4, 7, 8, \mathbf{9, 10}, 13, 14, 15\} \rightarrow \tilde{x} = \frac{9 + 10}{2} = 9.5$



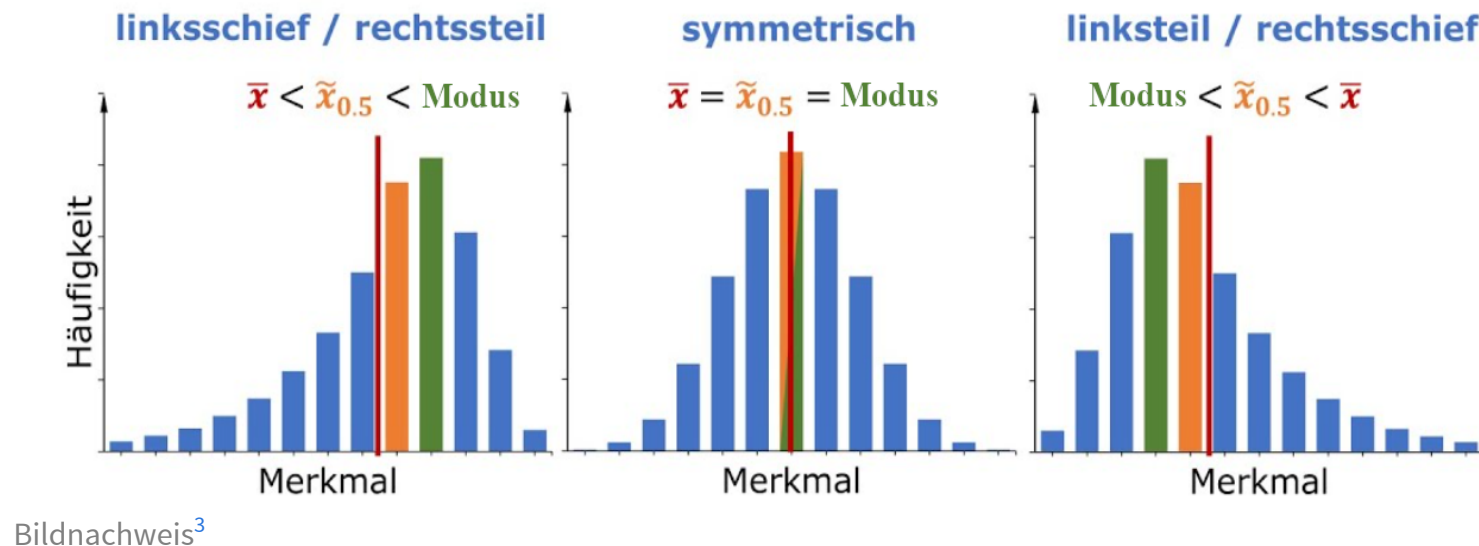
# Median

- Der **Median ist sinnvoll** bei:
  - Ausreißern in den Daten.
  - Schiefen Verteilungen der Daten (dazu kommen wir noch).
  
- Der **Median ist nicht sinnvoll**, wenn:
  - Auch der Mittelwert ein sinnvolles Lagemaß darstellt (der Mittelwert hat einige hilfreiche mathematische Eigenschaften<sup>2</sup>).

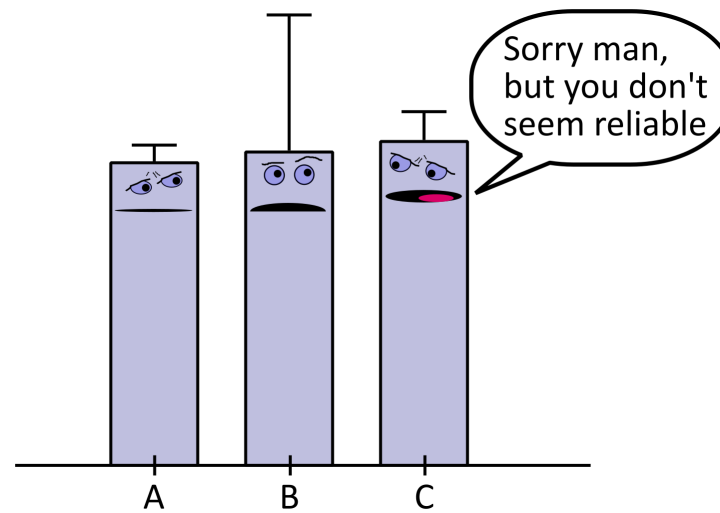


# Modus

- In manchen Fällen ist es interessant zu wissen, was der **häufigste Wert** in einem Datensatz ist – dies ist der **Modus**.
- Berechnung:
  1. Zähle die Häufigkeit aller vorkommenden Werte
  2. Der häufigste Wert ist der Modus
- Im Fall von **kategorialen Variablen** ist der Modus das einzig mögliche Lagemaß (Beispiel: aus welchem Bundesland kommen die meisten von Ihnen?)
- Ein weiterer sinnvoller Anwendungsfall können schiefe Verteilungen sein:

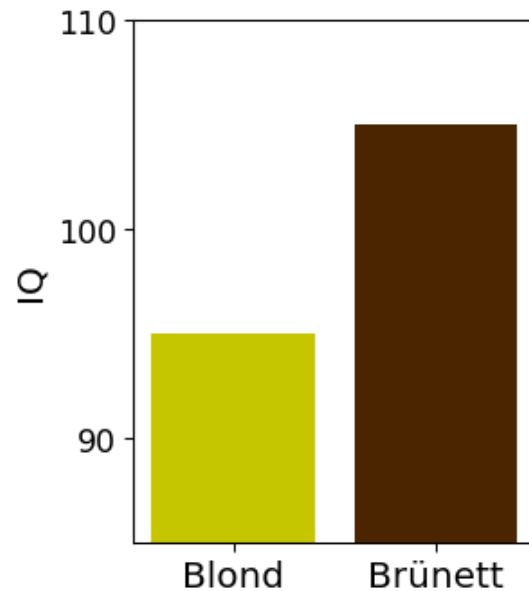


# Streuungsmaße

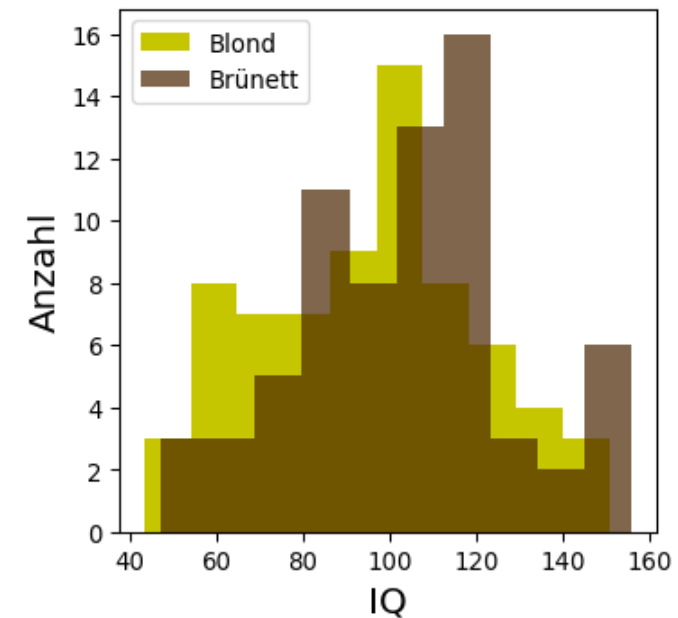


# Warum sind Streuungsmaße wichtig?

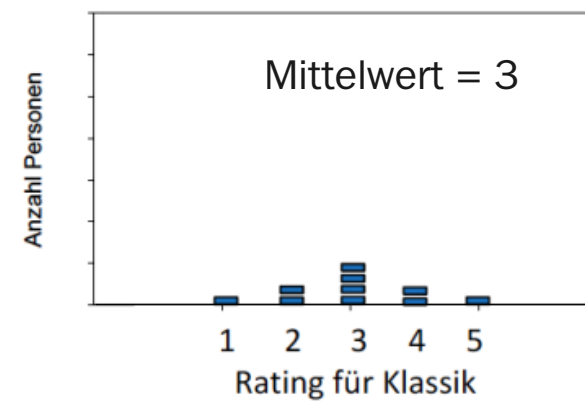
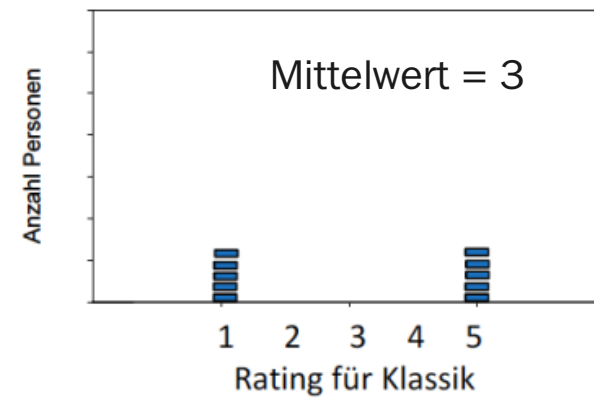
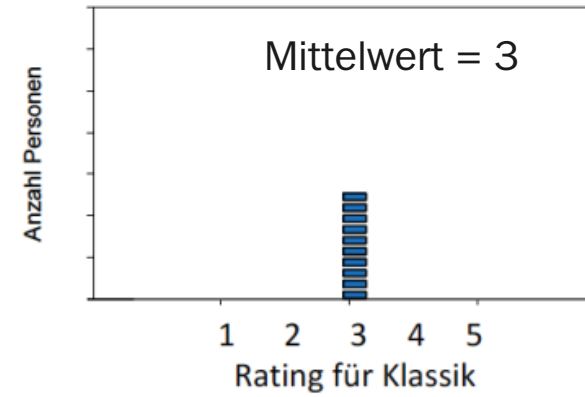
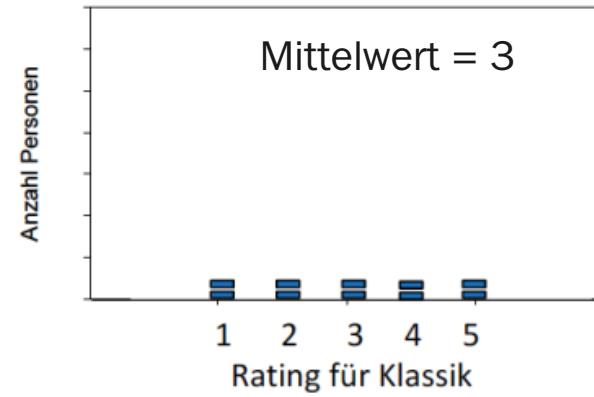
“In unserer Studie waren brünette Menschen im Schnitt 10 IQ-Punkte schlauer als blonde Menschen”



Behind the scenes:

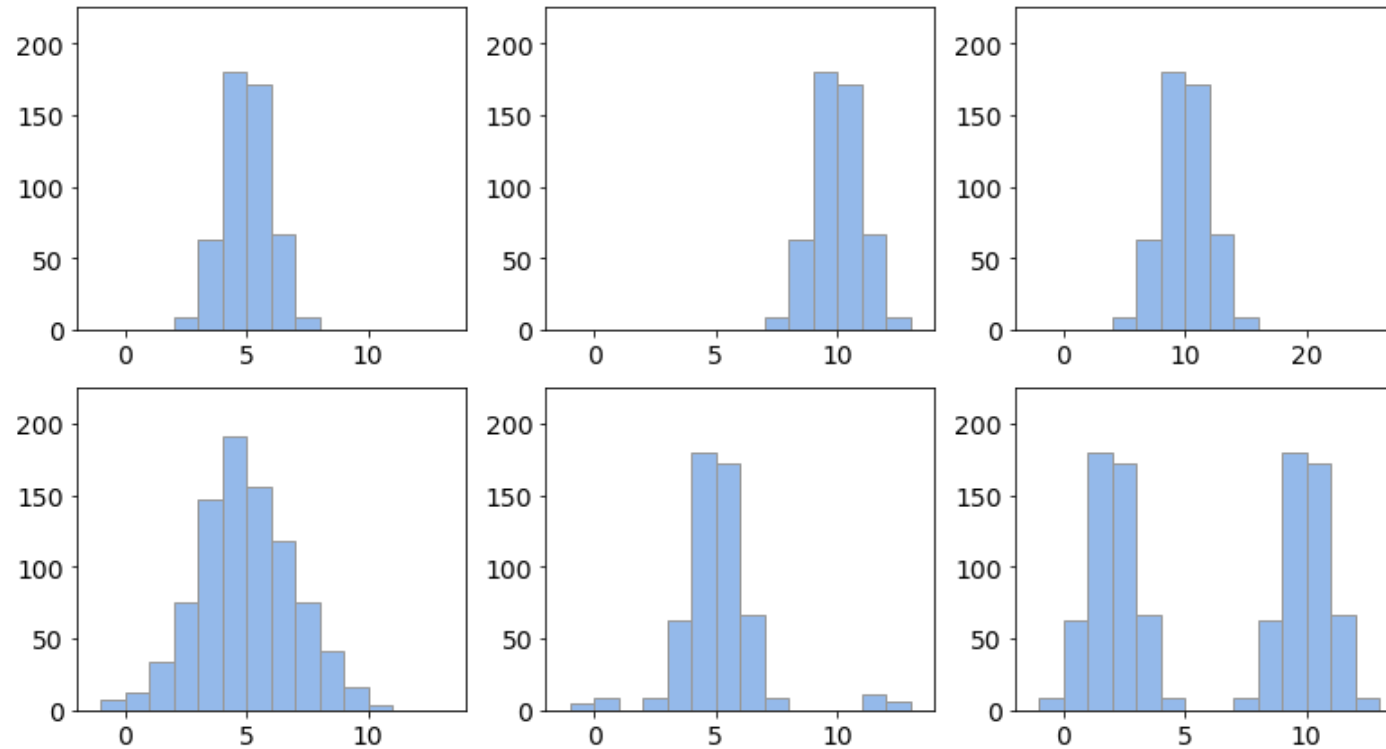


# Warum sind Streuungsmaße wichtig?



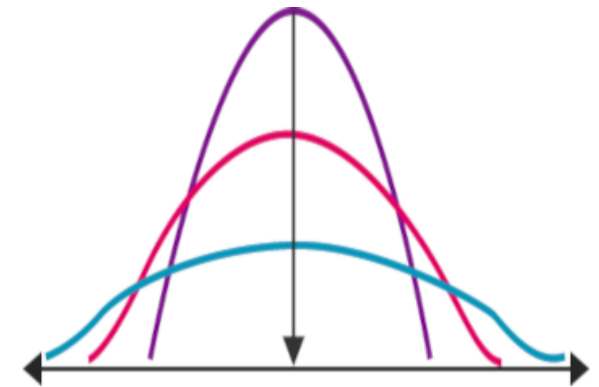


Wie schätzen Sie die Streuung / Variabilität folgender Verteilungen ein?



# Streuungsmaße

- **Streuungsmaße** geben die Variabilität von Daten an
- Streuungsmaße sind eine extrem wichtige Ergänzung zu Lagemaßen beim Bericht wissenschaftlicher Ergebnisse.
- Die Streuung von Daten kann ein Ausdruck *echter Variabilität* in der Stichprobe sein oder eine Folge der *Messungenauigkeit* (häufig beides).
- Je nach Skalenniveau und Zweck können verschiedene Streuungsmaße bestimmt werden:
  - Spannweite (Range)
  - Interquartilsabstand
  - Varianz
  - Standardabweichung

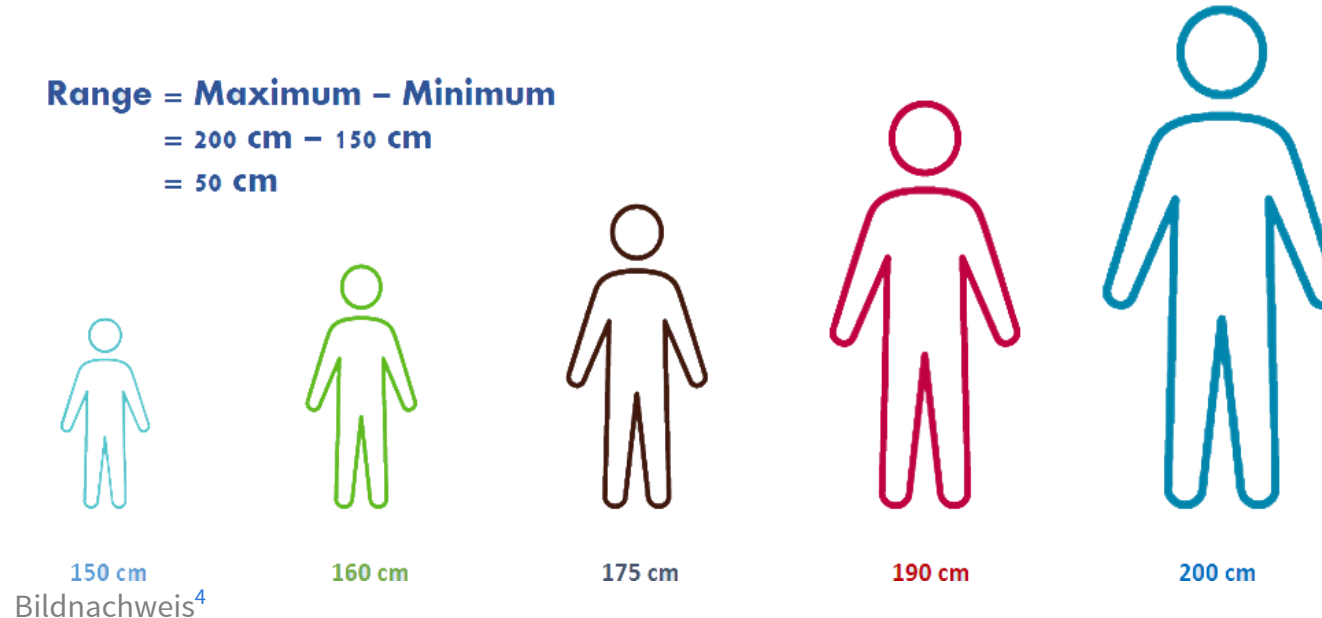


# Spannweite / Range

- Die **Spannweite** oder **Range** ist die Differenz zwischen dem kleinsten und dem größten Wert:

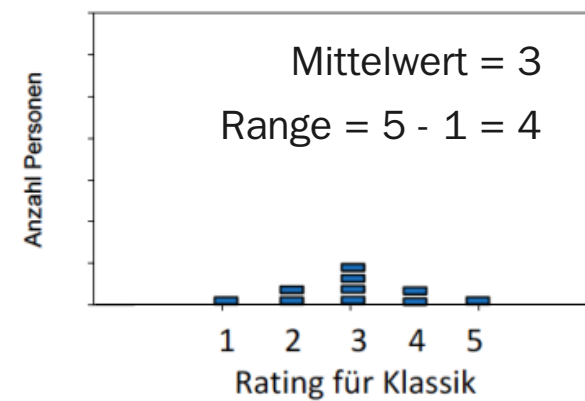
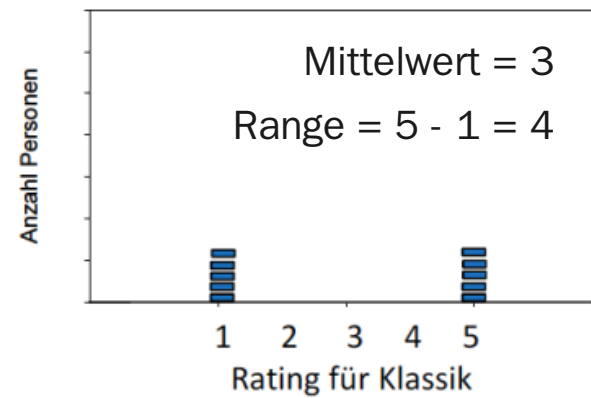
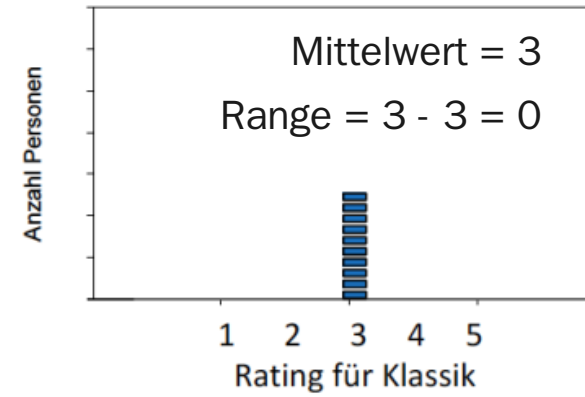
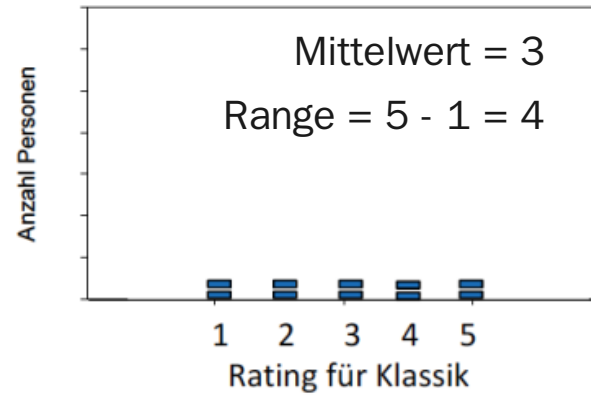
$$\text{Range} = x_{max} - x_{min}$$

- Einfachstes **Streuungsmaß** – sinnvoll, um dem Leser einen Eindruck der gesamten Spannweite von Daten zu geben.
- Allerdings kaum Aussagekraft über die tatsächliche Variabilität der Daten
  - Beispiel: die Spannweite von  $\mathbf{x} = \{1, 10, 10, 10, 10, 10, 10, 10, 10, 10, 101\}$  ist  $101 - 1 = 100$ , obwohl die Daten bis auf die zwei Ausreißer 1 und 101 keine Variabilität aufweisen.





# Range: Beispiele



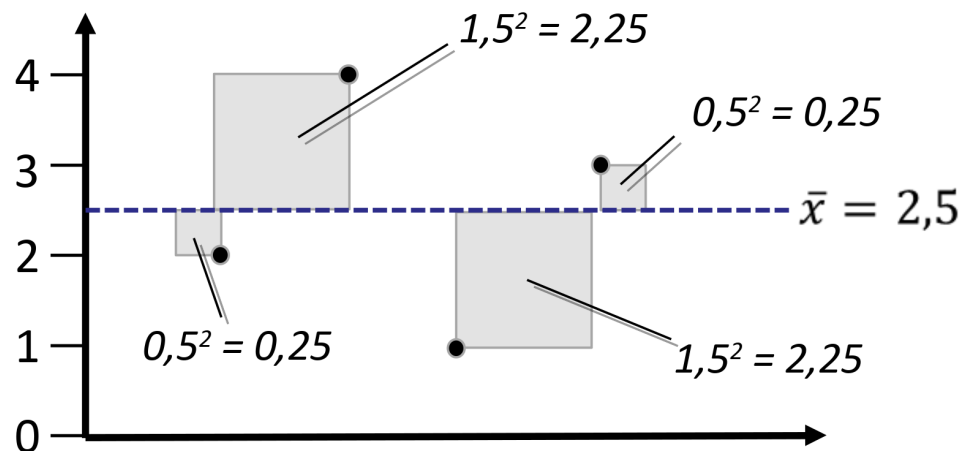
# Varianz

- Die Varianz ist definiert als **mittleres Abstandsquadrat aller Werte vom Mittelwert**:

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

(Die Varianz ist gleich der quadrierten Standardabweichung  $\sigma$  – letztere lernen wir noch kennen)

- Hinweis:** hierbei handelt es sich um die Formel für die **inferentielle Varianz**. Soll kein Rückschluss auf eine Population gezogen werden, ändert sich der Nenner von  $n - 1$  zu  $n$ .
- Beispiel:



$$\begin{aligned} \sigma^2 &= \frac{1}{4-1} (0,25 + 2,25 + 0,25 + 2,25) \\ &= \frac{1}{3} \cdot 5 = \frac{5}{3} \end{aligned}$$


# Besselkorrektur

- Streng genommen stellt die gezeigte Varianzformel nicht exakt das *mittlere* Abstandsquadrat aller Werte vom Mittelwert dar, da im Nenner durch  $n - 1$  und nicht durch  $n$  geteilt wird.
- Die Korrektur von  $n \longrightarrow n - 1$  heißt **Besselkorrektur** und tritt häufig auf, wenn **inferentielle Kennwerte** bestimmt werden sollen.
- Die Korrektur wird notwendig, weil die Schätzung der Populationsvarianz eigentlich erfordert, dass die Abstandsquadrate der  $x_i$  *um den wahren Populationsmittelwert  $\mu$*  berechnet werden.
- Das wahre  $\mu$  ist jedoch nicht bekannt, sondern lediglich die Schätzung  $\hat{\mu} = \bar{x}$  auf Basis der Stichprobe.
- Es stellt sich heraus, dass die Streuung der  $x_i$  um den Mittelwert  $\bar{x}$  der Stichprobe immer etwas kleiner ist, als es die Streuung der  $x_i$  um den wahren Wert  $\mu$  wäre.
- Aus diesem Grund muss die Varianzformel durch einen kleineren Wert ( $n - 1$  statt  $n$ ) geteilt werden, damit der Varianzschätzer “nach oben” korrigiert wird.

# Standardabweichung

- Ein Nachteil der Varianz ist, dass sie aufgrund der Abstandsquadrate in quadrierten Einheiten angegeben ist:

---


 Beispiel  $x = \{167 \text{ cm}, 181 \text{ cm}, 154 \text{ cm}, 192 \text{ cm}, 173 \text{ cm}\} \rightarrow \hat{V}ar(X) = 205.3 \text{ cm}^2$

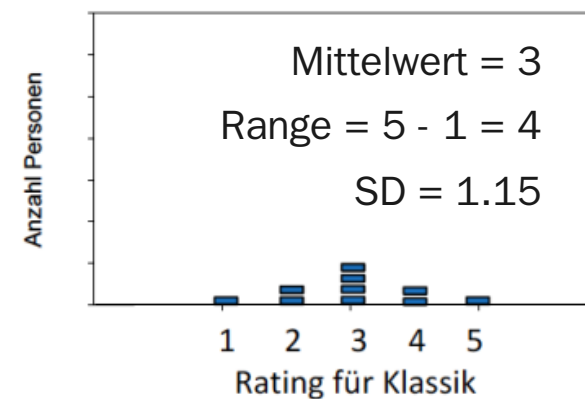
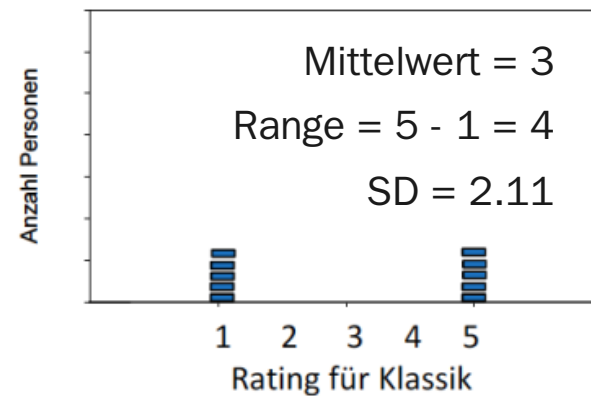
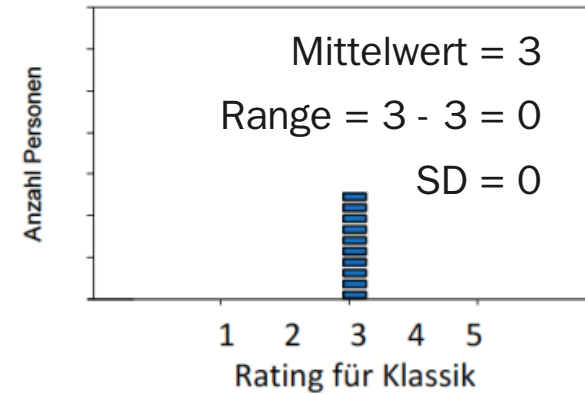
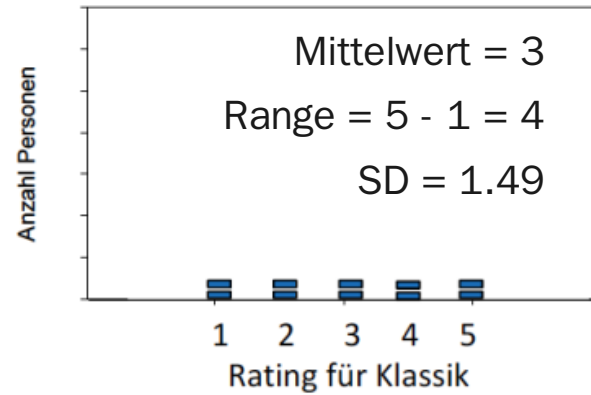
---

- Quadrierte Einheiten sind jedoch wenig intuitiv und schwer zu interpretieren.
- Aus diesem Grund wird häufig die Standardabweichung  $\hat{\sigma}$  angegeben, welche die Wurzel der Varianz darstellt:

$$\hat{\sigma} = \sqrt{\hat{V}ar(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Die Standardabweichung drückt die Streuung in den **Rohwerten der Skala** aus.
- Sie wird häufig auch mit *sd* oder *SD* (für *standard deviation*) abgekürzt.

# Standardabweichung: Klassik-Beispiele



# Interquartilsabstand (interquartile range = IQR)

- **Quartile** sind eine Spezialform von **Quantilen**
- **Quantile** teilen eine Verteilung von Daten in gleich große Abschnitte ein
  - “gleich groß” = jeder Abschnitt hat gleich viele Datenpunkte
- Beispiel *Dezile*: Einteilung der Verteilung in 10 gleich große Abschnitte
  - Hier: 20 Werte eingeteilt in 10 Abschnitte (Dezile) á 2 Werte

$1, 1, 1, 2, 2, 3, 3, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 8$   
 1.Dezil 2.Dezil ... 10.Dezil

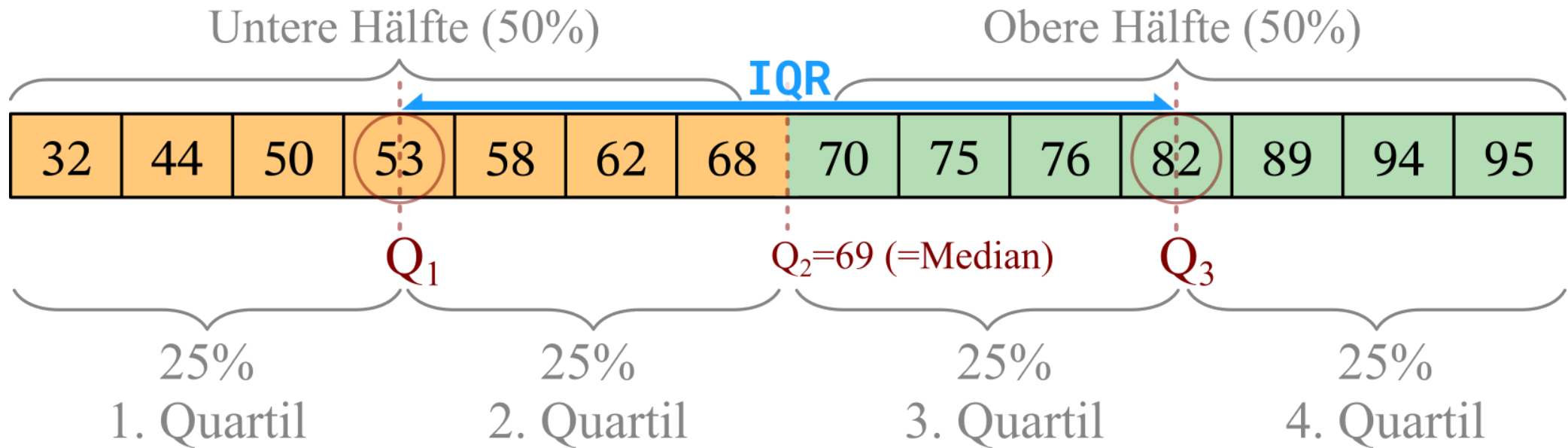
- Beispiel *Quintile*: Einteilung der Verteilung in 5 gleich große Abschnitte
  - Hier: 30 Werte eingeteilt in 5 Abschnitte (Quintile) á 6 Werte

$1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 9, 9$   
 1.Quartil 2.Quartil 3.Quartil 4.Quartil 5.Quartil

- Beispiel *Quartile*: Einteilung der Verteilung in 4 gleich große Abschnitte
  - Hier: 12 Werte in 4 Abschnitte (Quartile) á 3 Werte

$1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 6, 6$   
 1.Quartil 2.Quartil 3.Quartil 4.Quartil

# Interquartilsabstand (interquartile range = IQR)



- Um eine Reihe von Daten in 4 gleich große Quartile zu teilen, sind genau drei Quartilsgrenzen notwendig
- Diese Quartilsgrenzen werden mit  $Q_1$ ,  $Q_2$ ,  $Q_3$  (bezogen auf *Quartile*) bzw. mit  $Q_{25\%}$ ,  $Q_{50\%}$ ,  $Q_{75\%}$  (bezogen auf *Quantile*) bezeichnet
- Der Interquartilsabstand (IQR) ist die Differenz aus der 75%-*Quantilsgrenze* und der 25%-*Quantilsgrenze* bzw. die Differenz aus der 3. und der 1. *Quartilsgrenze*:

$$IQR = Q_{75\%} - Q_{25\%} = Q_3 - Q_1$$

# Interquartilsabstand (interquartile range = IQR)

- Berechnung des IQR:

1. Sortiere alle Werte von klein nach groß

2. Bestimme die Tiefe des Medians:  $Tiefe_{\text{Median}}$

3. Bestimme die Tiefe des Quartils:  $Tiefe_{\text{Quartil}} = \frac{Tiefe_{\text{Median}} + 1}{2} = \frac{\frac{n+1}{2} + 1}{2} = \frac{n+3}{4}$

4. Für das 25%-Quantil ( $Q_1$ ) geht man **von vorne** in die Datenreihe

5. Für das 75%-Quantil ( $Q_3$ ) geht man **von hinten** in die Datenreihe

Folgende 12 Werte werden beobachtet:  $x = \{1, 1, 2, 3, 3, 3, 4, 4, 6, 7, 7, 9\}$

In diesem Fall ist der "6,5"-te Wert der Median, da  $Tiefe_{\text{Median}} = \frac{n+1}{2} = \frac{12+1}{2} = 6.5$

Die Tiefe des Quartils ist damit  $Tiefe_{\text{Quartil}} = \frac{Tiefe_{\text{Median}} + 1}{2} = \frac{6.5+1}{3} = 3.75$

  
Beispiel

Der "3.75"-te Wert von vorne befindet sich auf 75% der Strecke zwischen dem 3. Wert (2) und dem 4. Wert (3),

also  $Q_1 = 2.75$ ; der "3.75"-te Wert von hinten befindet sich auf 75% der Strecke zwischen der 7 und der 6,

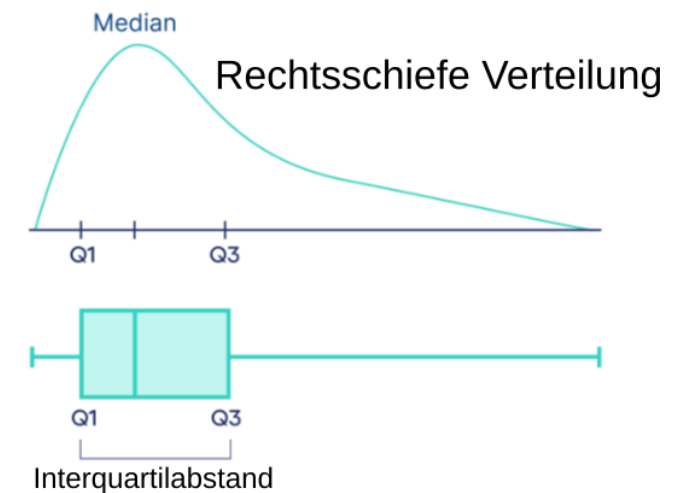
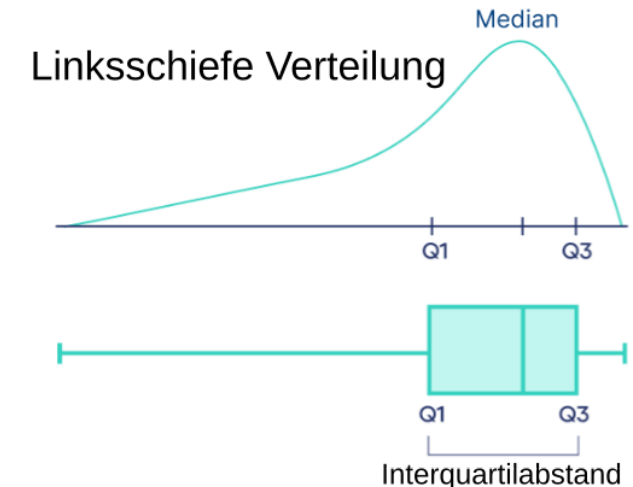
also  $Q_3 = 6.25$ .

$IQR = Q_3 - Q_1 = 6.25 - 2.75 = 3.5$



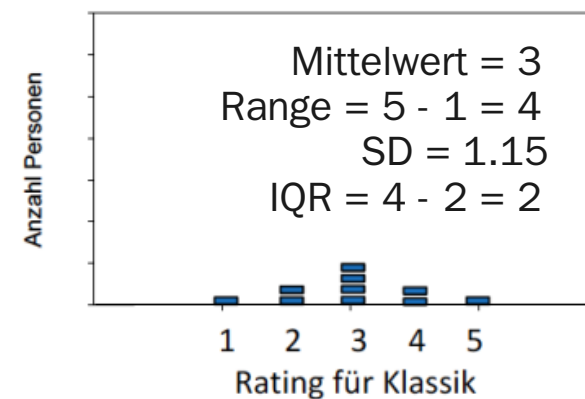
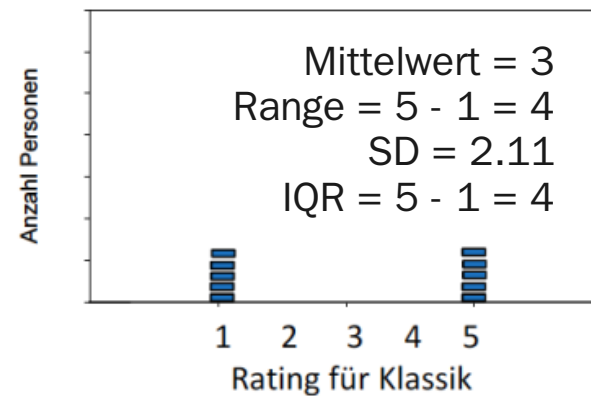
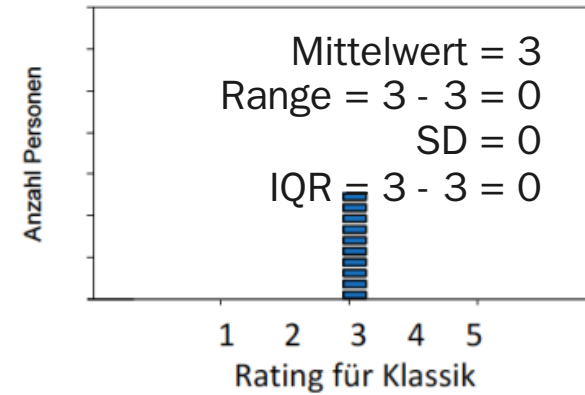
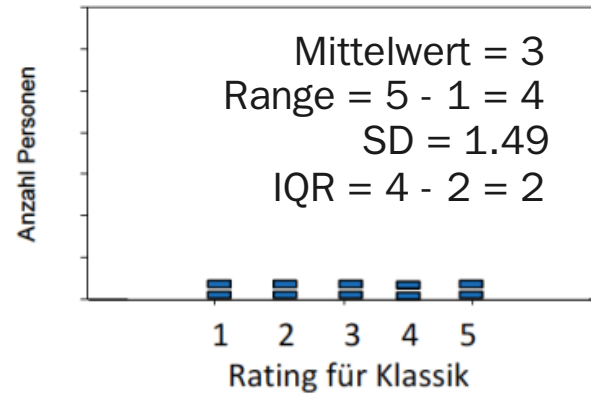
# Wann ist der Interquartilsabstand ein sinnvolles Streuungsmaß?

- Die klassische Standardabweichung kann leicht durch einen einzelnen oder wenige extreme Datenpunkte verzerrt werden (insbesondere durch die Abstandsquadrierung).
- Der Interquartilsabstand ist **robuster gegen Ausreißer**, da er auf einem ähnlichen Prinzip wie der Median basiert: statt einem arithmetischen Mittel werden die Werte auf tatsächlichen Datenpunkten *in der Mitte der Verteilung*.
- Wie dem Median ist es daher dem Interquartilsabstand “egal” ob etwa der höchste Wert 10 oder 10 Trillionen ist.
- Auch bei **stark schiefen Verteilungen** bietet sich der Interquartilsabstand an, da der von der Standardabweichung verwendete Bezugspunkt “Mittelwert” in diesem Fall problematisch ist.



Vereinfacht gesagt verhält sich der **Interquartilsabstand** zur **Varianz**, wie der **Median** zum **Mittelwert**.

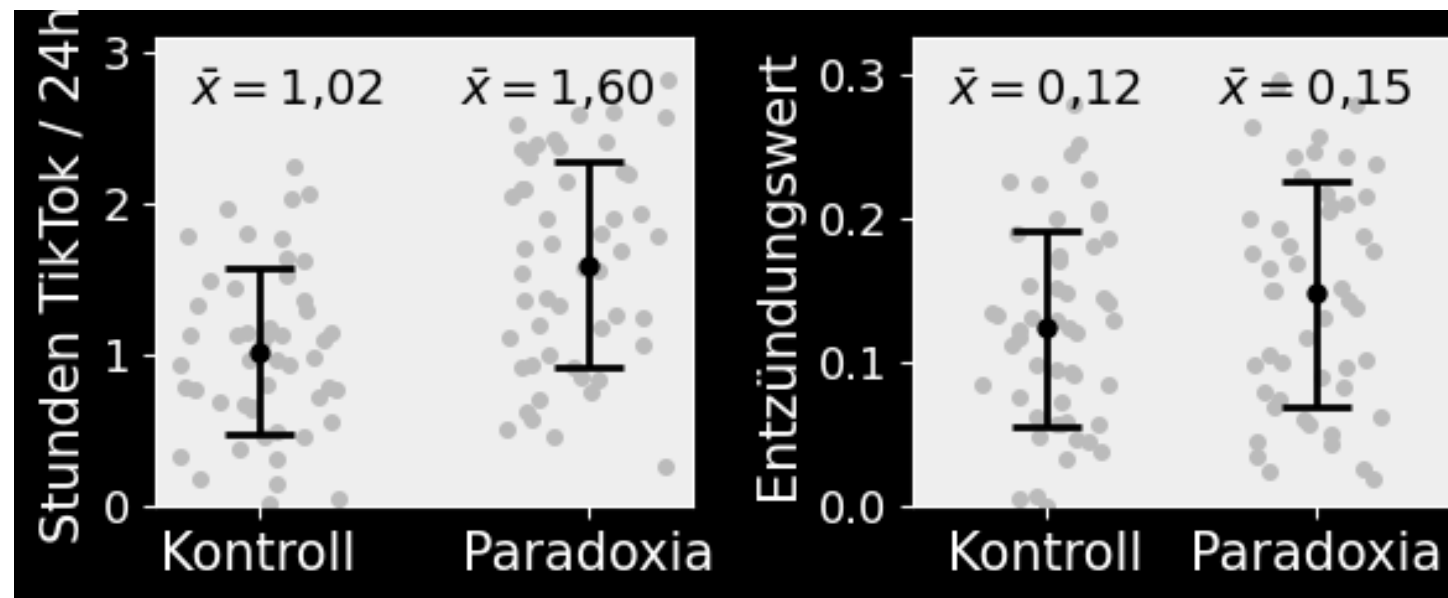
# Interquartilsabstand: Klassik-Beispiele

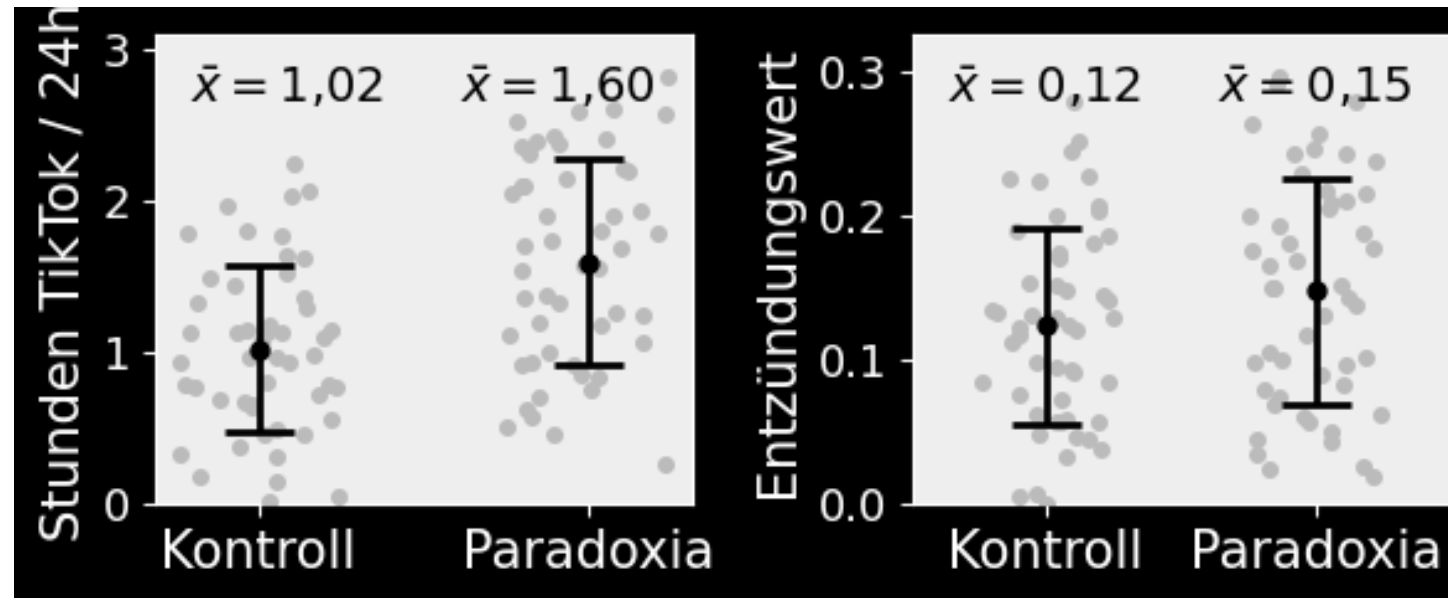


# Die Streuungsmaße im Vergleich

<b>Spannbreite (Range)</b>	<ul style="list-style-type: none"><li>▪ Gibt die Ausdehnung des gesamten Wertebereiches an</li><li>▪ Auf kein bestimmtes Lagemaß bezogen</li><li>▪ Geringer statistischer Nutzen, manchmal interessante Zusatzinfo</li><li>▪ Maximal abhängig von Ausreißern</li></ul>
<b>Varianz</b>	<ul style="list-style-type: none"><li>▪ Auf den Mittelwert bezogen (“wie stark streuen die Daten um den Mittelwert?”)</li><li>▪ Relativ anfällig gegenüber Ausreißern</li><li>▪ Unnatürliche quadrierte Einheiten</li></ul>
<b>Standardabweichung</b>	<ul style="list-style-type: none"><li>▪ Wie Varianz, aber natürliche unquadrierte Einheiten</li></ul>
<b>Interquartilsabstand</b>	<ul style="list-style-type: none"><li>▪ Auf kein bestimmtes Lagemaß bezogen</li><li>▪ Jedoch ähnliches Prinzip wie der Median und häufig im Zusammenhang mit diesem angegeben</li><li>▪ Robust gegenüber Ausreißern &amp; sinnvoll bei schiefen Verteilungen</li></ul>

Die Daten in Ihrer Beobachtungsstudie weisen keine größeren Ausreißer auf und Sie entscheiden sich für den Mittelwert als Lagemaß und die Standardabweichung als Streuungsmaß. Für eine erste Kommunikation mit den anderen Task Forces erstellen Sie folgende Abbildung:





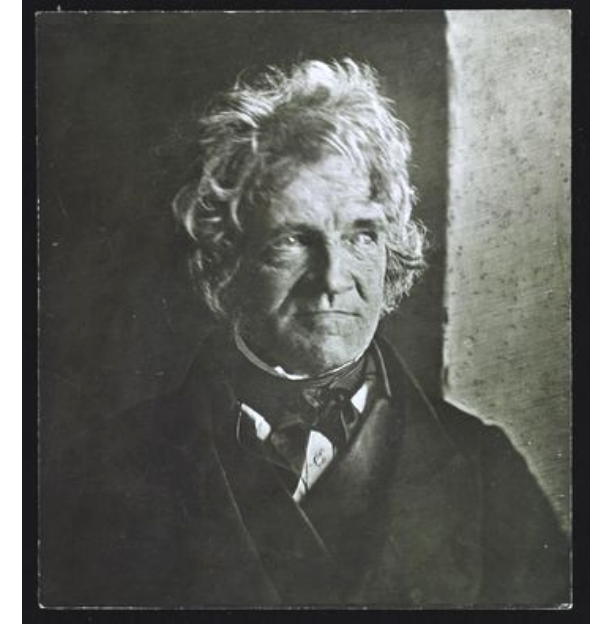
Aus den Werten und der Abbildung wird ersichtlich: tatsächlich sind auch die Entzündungswerte bei Paradoxikern erhöht! Auf Basis der Mittelwerte finden Sie also Evidenz für *beide Hypothesen*!

# Besselkorrektur

- Der deutsche Naturwissenschaftler Friedrich Wilhelm Bessel zeigte, dass für eine Schätzung  $\hat{\sigma}^2$  der Populationsstreuung auf Basis der Stichprobenstreuung  $s^2$ , der Faktor  $\frac{1}{n}$  in der Varianzformel durch  $\frac{1}{n-1}$  ersetzt werden muss:

Schätzung der Populationsvarianz:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$



Friedrich Wilhelm Bessel (1748-1846)<sup>5</sup>

- Die Größe  $\hat{\sigma}^2$  ist die **Schätzung der Populationsvarianz auf Basis der Stichprobe** und wird in dieser Vorlesungsreihe als **inferentielle Varianz** bezeichnet.
- Intuition: die Varianz wird im Fall von  $\frac{1}{n-1}$  *durch weniger geteilt* als beim ursprünglichen Faktor  $\frac{1}{n}$ , d.h. die resultierende Varianz wird *zu einem größeren Wert hin* korrigiert.
- Die “ $n - 1$ ”-Korrektur für Stichproben wird als **Besselkorrektur** bezeichnet.
- Ab ca.  $n = 30$  spielt die Besselkorrektur kaum eine Rolle mehr ( $\frac{1}{30} = 0.033$  vs.  $\frac{1}{29} = 0.034$ )



# Herleitung des Faktors $\frac{1}{n-1}$

- Der Ausgangspunkt der Besselkorrektur die Beobachtung, dass die (unkorrigierte) Varianz  $s^2$  der Stichprobe die Varianz  $\sigma^2$  der Population systematisch unterschätzt.
- Der Grund ist liegt im Stichprobenmittelwert  $\bar{x}$ : er ist keine perfekte Schätzung für den wahren Mittelwert  $\mu$  ist — er streut um das “wahre”  $\mu$ .
- Diese Streuung werden wir noch kennenlernen — es ist der Standardfehler  $\hat{s}e = \frac{\hat{\sigma}}{\sqrt{n}}$ . Als Varianz formuliert:

$$\hat{s}e^2 = \frac{\hat{\sigma}^2}{n}$$

- Für die Schätzung der Populationsvarianz  $\hat{\sigma}^2$  muss nun die *zusätzliche* Variabilität  $\hat{s}e^2$  aufgrund der Unsicherheit des Mittelwertes auf die Stichprobenvarianz  $s^2$  addiert werden:

$$\hat{\sigma}^2 = s^2 + \hat{s}e^2 = s^2 + \frac{\hat{\sigma}^2}{n} \quad \longrightarrow \quad \hat{\sigma}^2 - \frac{\hat{\sigma}^2}{n} = s^2$$

- Linke Seite umformen:

$$\hat{\sigma}^2 - \frac{\hat{\sigma}^2}{n} = \hat{\sigma}^2 \left( 1 - \frac{1}{n} \right) = \hat{\sigma}^2 \frac{n-1}{n} \stackrel{!}{=} s^2 \quad \longrightarrow \quad \hat{\sigma}^2 = \frac{n}{n-1} s^2$$



# Herleitung des Faktors $\frac{1}{n-1}$

Zwischenergebnis:  $\hat{\sigma}^2 = \frac{n}{n-1} s^2$

- Die unkorrigierte Varianz  $s^2$  lautet:

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

- Einsetzen:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{n}{n-1} \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

- Nun haben wir unseren Faktor  $\frac{1}{n-1}$ .
- Für die korrigierte Standardabweichung der Stichprobe wird häufig analog angenommen:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

- Diese analoge Korrektur für die Standardabweichung ist nicht 100% gültig<sup>6</sup> – für alle praktischen Zwecke reicht sie aber aus.





# Varianz: warum werden werden nicht einfach die Absolutwerte der Differenzen genommen?

Prinzipiell wäre auch eine Formel für die Varianz mit Absolutabständen denkbar:

$$\hat{Var}_{\text{absolut}}(X) = \frac{1}{n-1} \sum_{i=1}^n |x_i - \bar{x}|$$

Über die Gründe, warum sich  $\hat{Var}_{\text{absolut}}$  nicht durchgesetzt hat, streitet sich die Fachwelt. Neben historischen Gründen, gibt es aber einige Eigenschaften, die die Präferenz für Abstandsquadrate zumindest nachvollziehbar machen:

- Quadrierte Abstände gewichten Punkte, die weiter vom Mittelwert entfernt sind, höher. Dies entspricht einer “Bestrafung” von Ausreißern und kann ein wünschenswertes Verhalten sein.
- Analogie zur euklidischen Distanz (z.B. Satz des Pythagoras:  $a = \sqrt{b^2 + c^2}$ )
- Fortgeschritten: die Varianz mit Abstandsquadraten ist für alle  $x$  differenzierbar (hingegen ist  $\hat{Var}_{\text{absolut}}$  bei  $x = 0$  nicht differenzierbar)



# Fußnoten

1. <https://datatab.de/tutorial/mittelwert-median-modus>
- 2.

Beispielsweise ist der Mittelwert der Mittelwerte von zwei Gruppen mit je  $n$  Datenpunkten gleich dem Mittelwert aller  $2 \cdot n$  Datenpunkte. Beim Median ist dies nicht gegeben. Auch beziehen sich viele statistische Standardtests auf den Mittelwert und nicht den Median.

3. <https://youtu.be/inJ4OvU0zMA>
4. <https://www.shiksha.com/online-courses/articles/measures-of-dispersion-range-iqr-variance-standard-deviation/>
5. <http://www.digiporta.net/index.php?id=658493715>
6. [https://en.wikipedia.org/wiki/Unbiased\\_estimation\\_of\\_standard\\_deviation](https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation)
7. [https://web.archive.org/web/20221024193801/https://www4.hcmut.edu.vn/~ndlong/TK/mat/04\\_standard\\_deviation\\_vs\\_absolute\\_deviation.pdf](https://web.archive.org/web/20221024193801/https://www4.hcmut.edu.vn/~ndlong/TK/mat/04_standard_deviation_vs_absolute_deviation.pdf)