

M24 Statistik 1: Wintersemester 2024 / 2025

Vorlesung 03: Lage- und Streuungsmaße

Prof. Matthias Guggenmos

Health and Medical University Potsdam



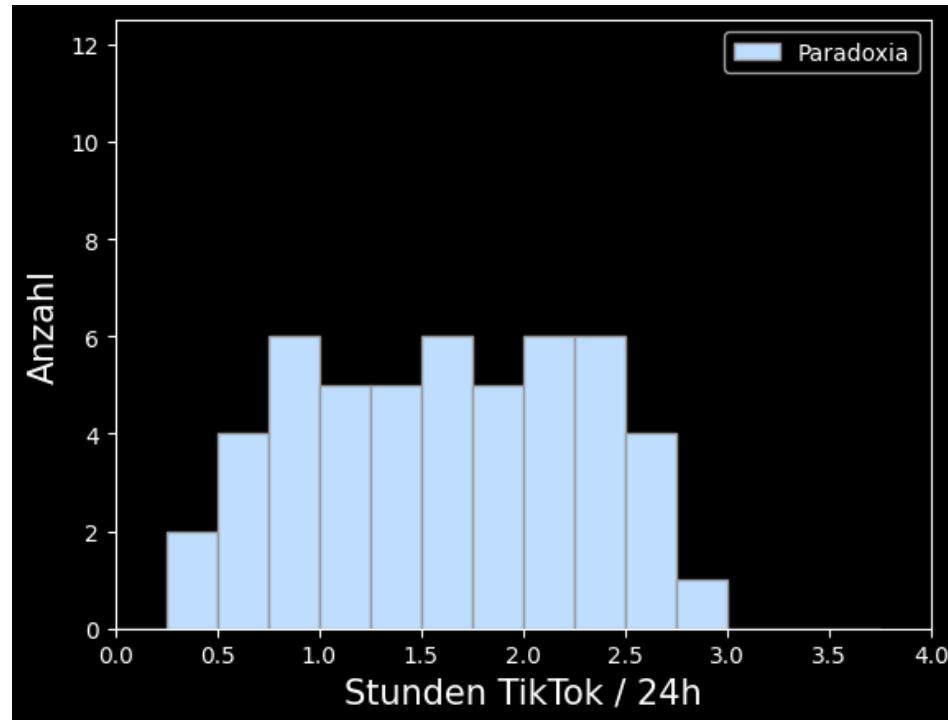


Die Daten der ersten Beobachtungsstudie zu Paradoxa sind frisch eingetroffen!

Table 1. Results.

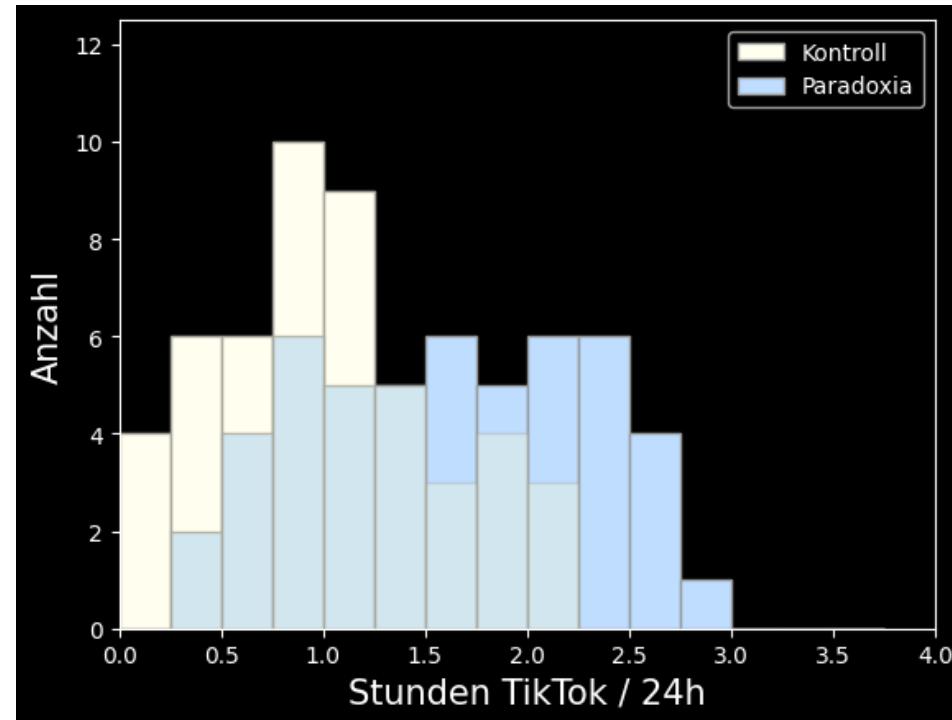
id	group	hours_tiktok_per_day	inflammation
1	control	1.14	0.24
2	control	2.24	0.19
...
50	control	1.14	0.13
51	paradoxa	1.57	0.03
52	paradoxa	2.59	0.21
...
100	paradoxa	0.52	0.12

Hier ist das Histogramm der TikTok-Zeiten von Paradoxikern:



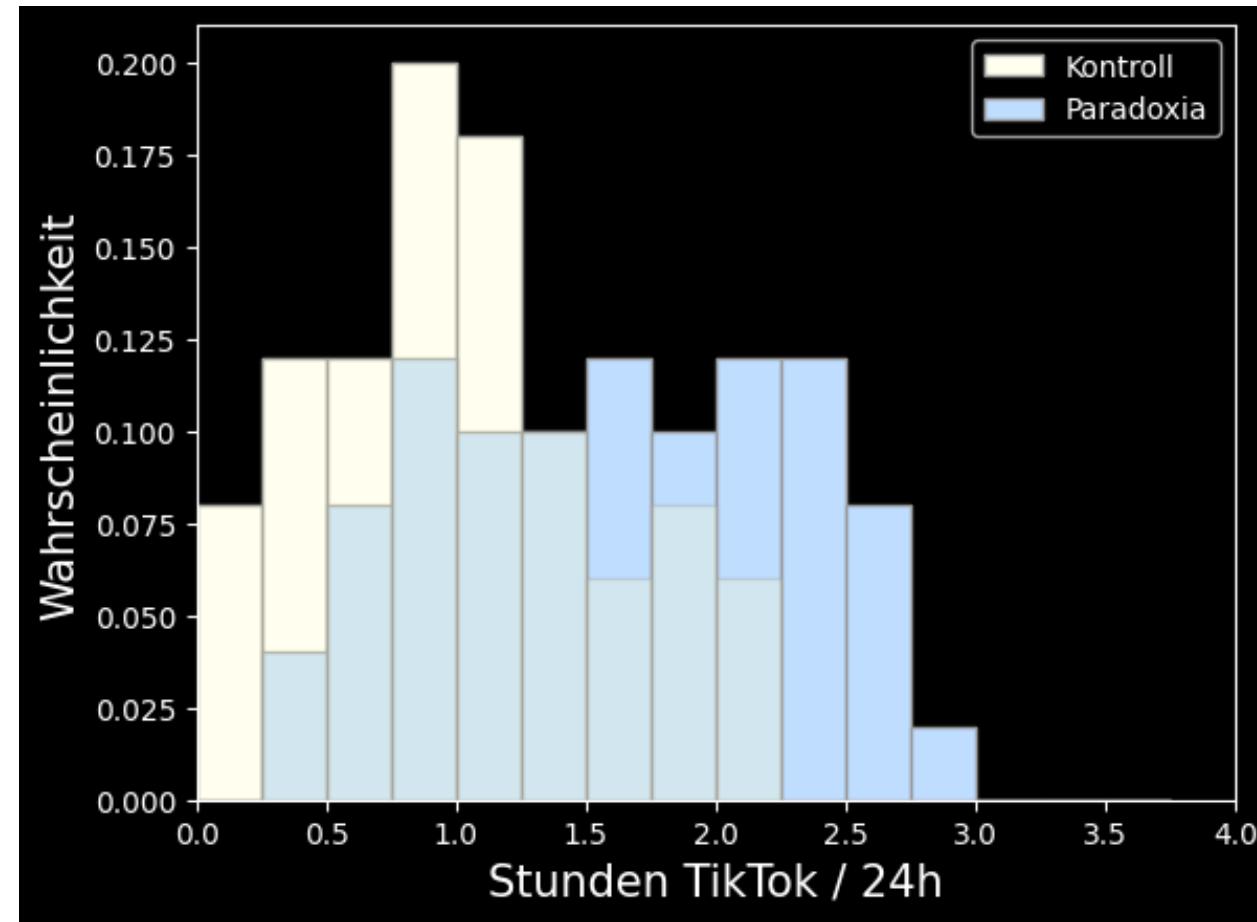
Überschlagen Sie: passt das Histogramm zur angegeben Stichprobe von $n=50$ Paradoxikern? Und handelt es sich um eine Abbildung relativer oder absoluter Häufigkeit?

Vergleich mit der Kontrollgruppe:



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

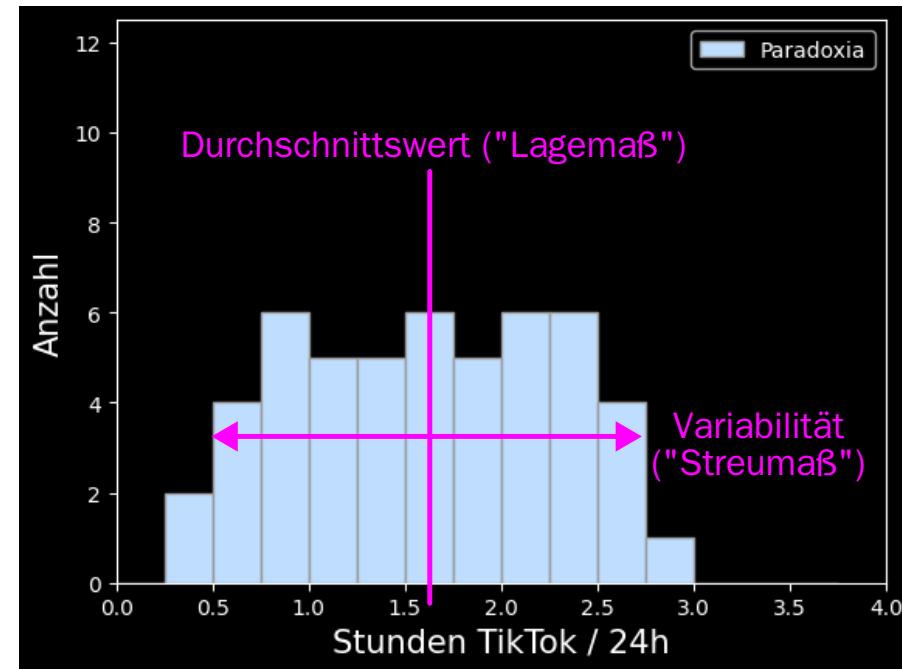
Erinnerung: statt der Anzahl (absolute Häufigkeit) kann auch die Wahrscheinlichkeit (relative Häufigkeit) dargestellt werden:



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

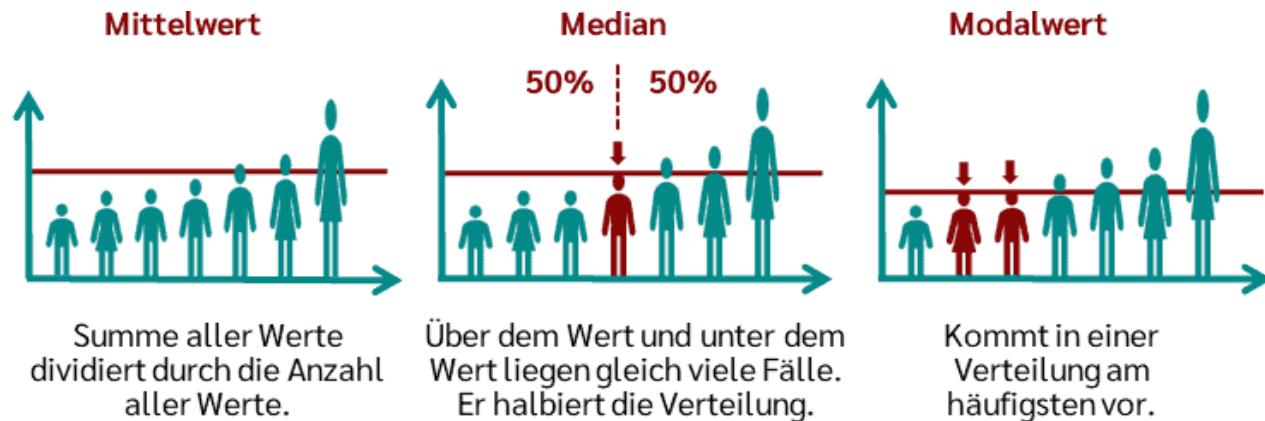


TraiNex

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)



Bildnachweis¹

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 (~) 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: $\mathbf{x} = \{13, 7, 15, 8, 4, 9, 14\}$

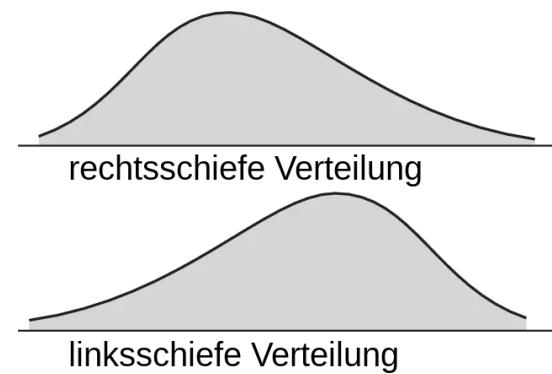
Sortierte Reihenfolge: $\mathbf{x} = \{4, 7, 8, \mathbf{9}, 13, 14, 15\} \rightarrow \tilde{x} = 9$

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

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

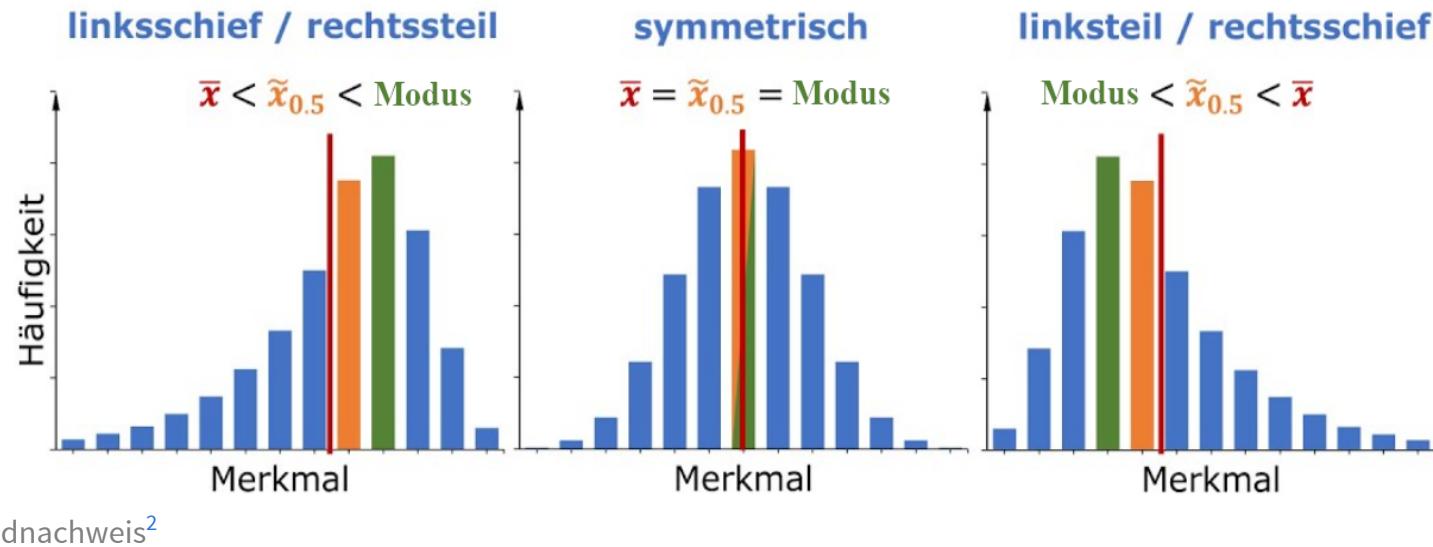
Median

- Der Median ist sinnvoll bei:
 - Ausreißern in den Daten.
 - Schiefen Verteilungen der Daten (dazu kommen wir noch).
- Der Median ist nicht sinnvoll, auch der Mittelwert ein sinnvolles Lagemaß darstellt, denn der Mittelwert hat einige hilfreiche mathematische Eigenschaften.
 - 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.

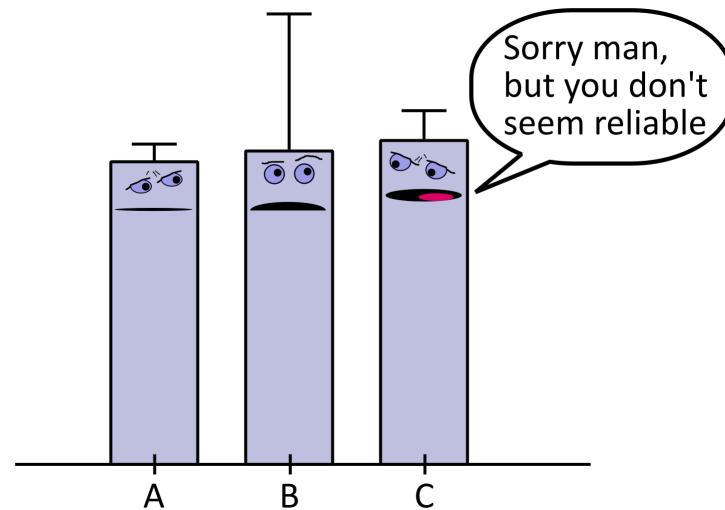


Modus

- 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 einzige mögliche Lagemaß (Beispiel: aus welchem Bundesland kommen die meisten von Ihnen?)
- Ein weiterer sinnvoller Anwendungsfall können schiefe Verteilungen sein:



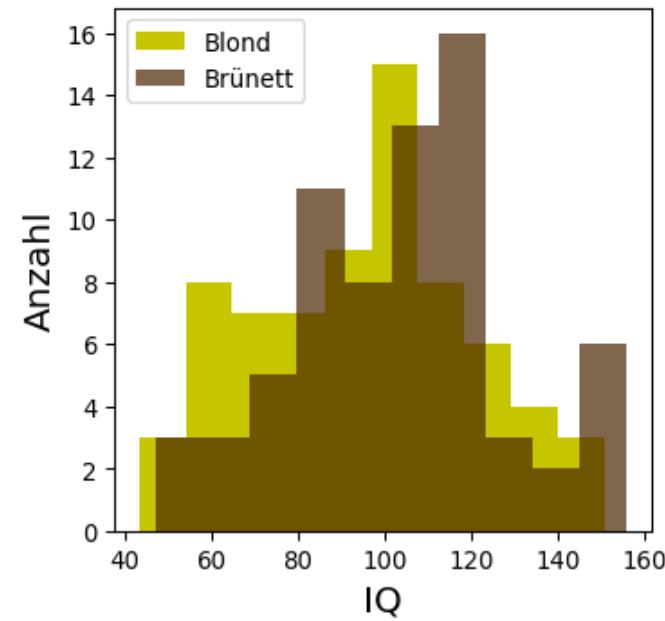
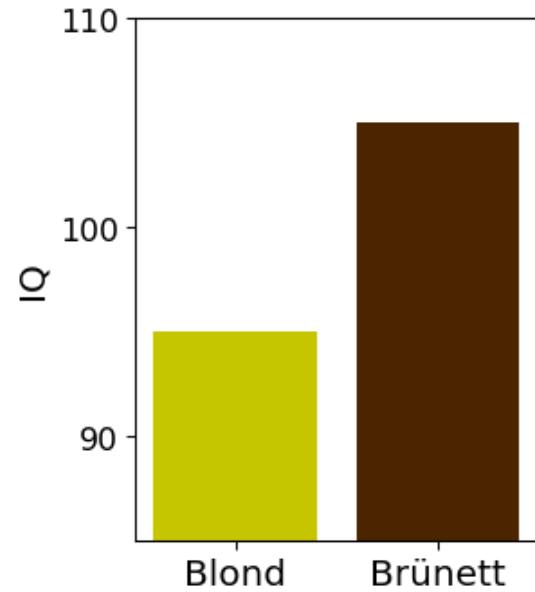
Streuungsmaße



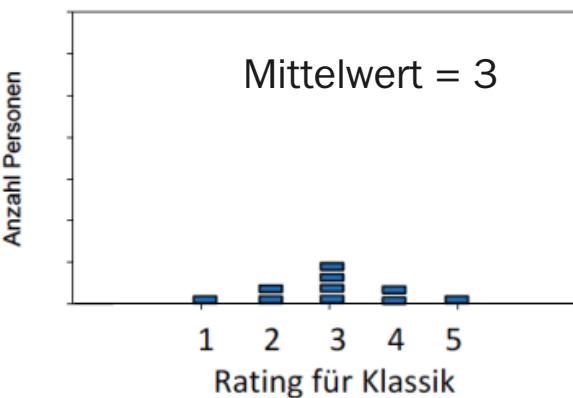
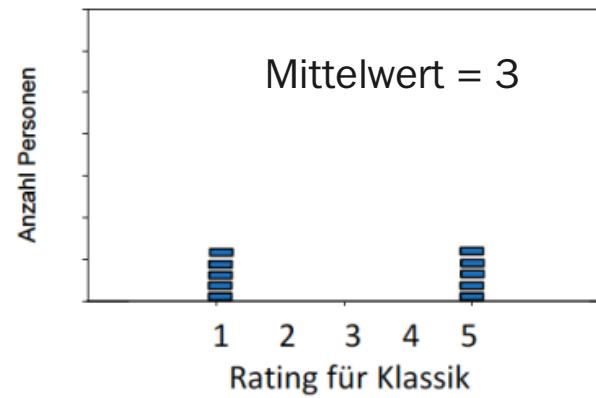
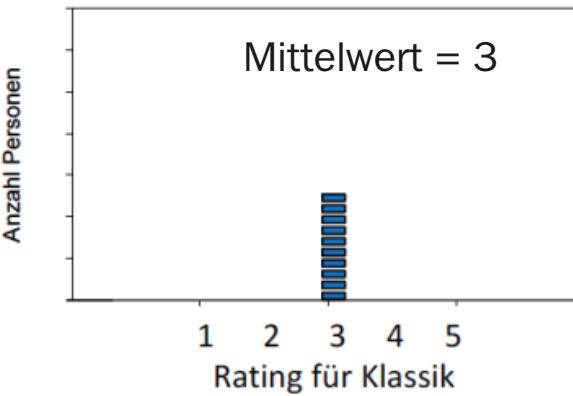
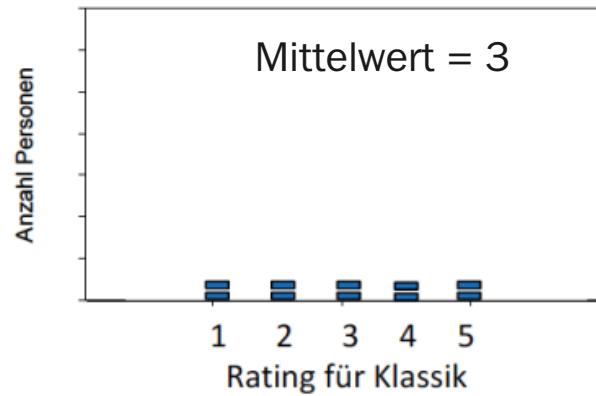
Warum sind Streuungsmaße wichtig?

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

Behind the scenes:

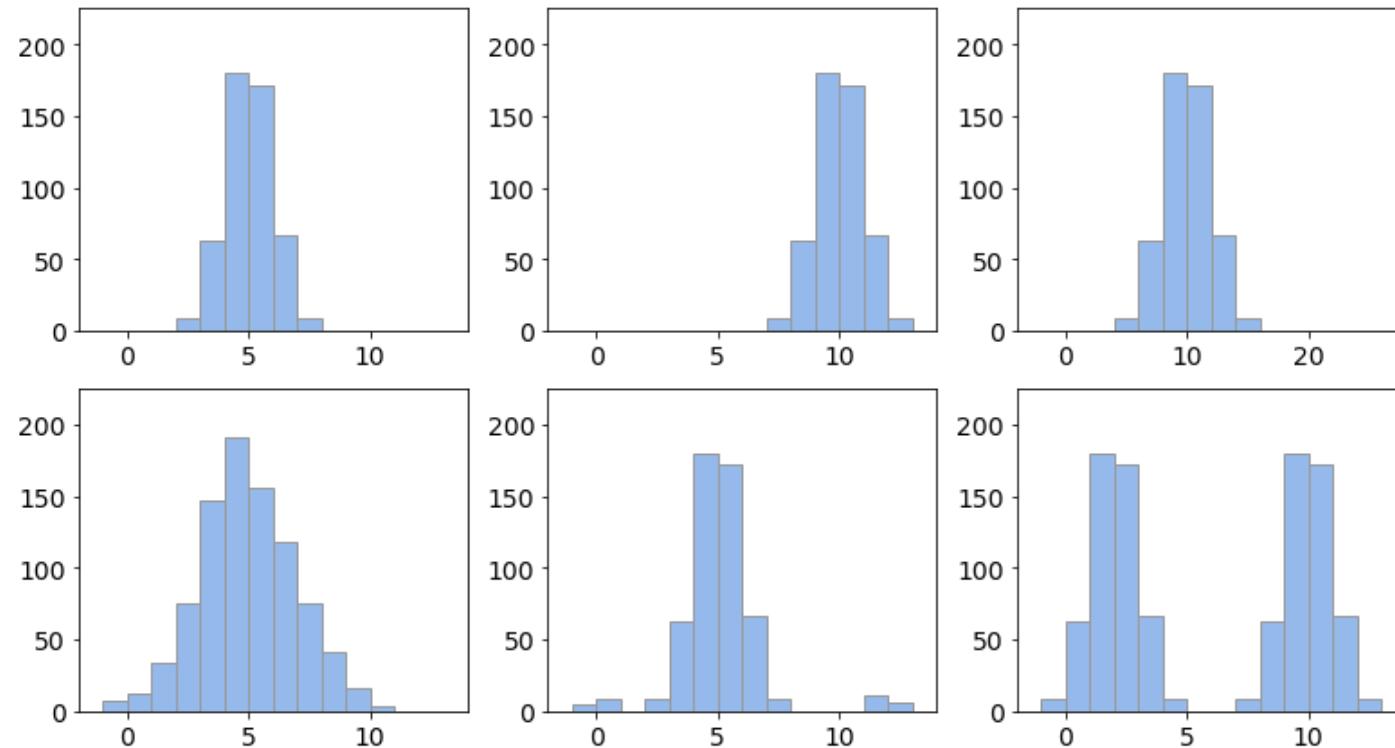


Warum sind Streuungsmaße wichtig?



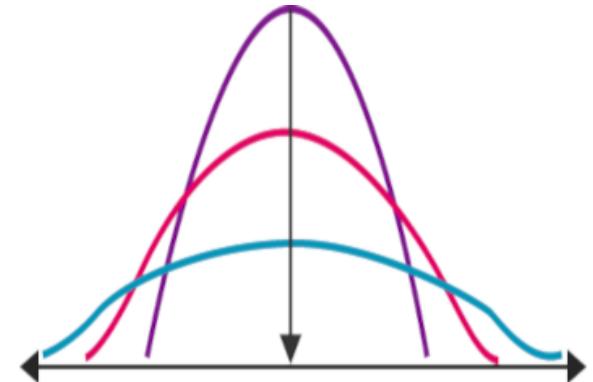


Wie schätzen Sie die Streuung / Variabilität folgender Verteilungen ein?



Streuungsmaße

- **Streuungsmaße** geben die Variabilität von Daten an
- Streuungsmaßes 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

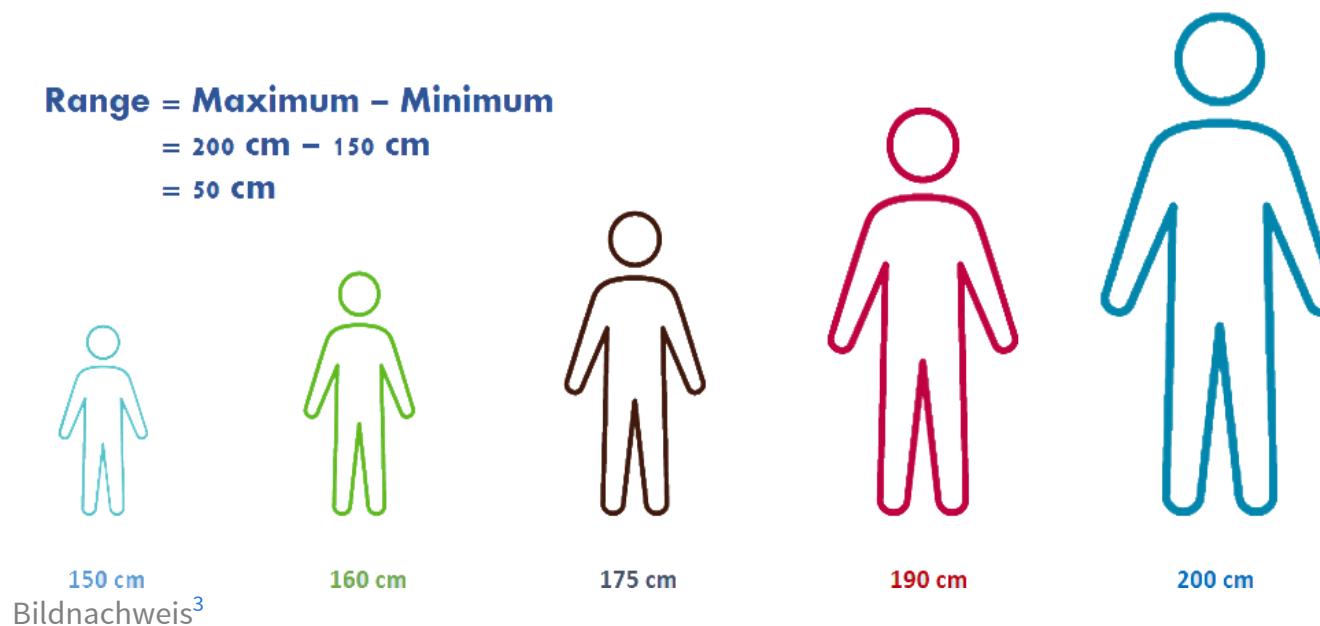


Spannweite / Range

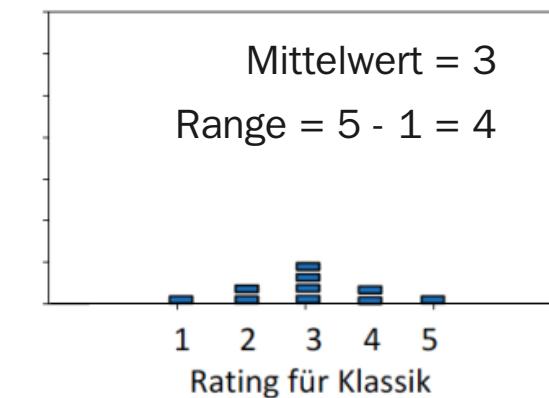
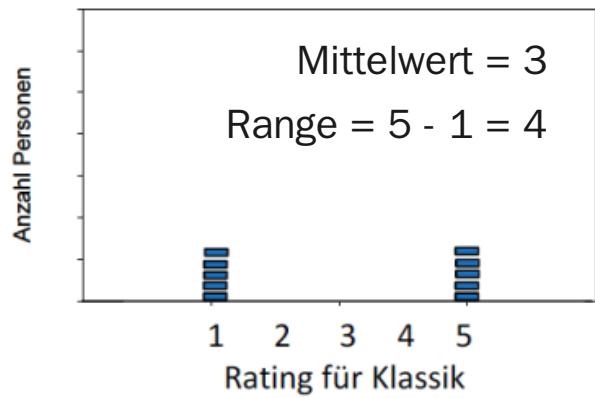
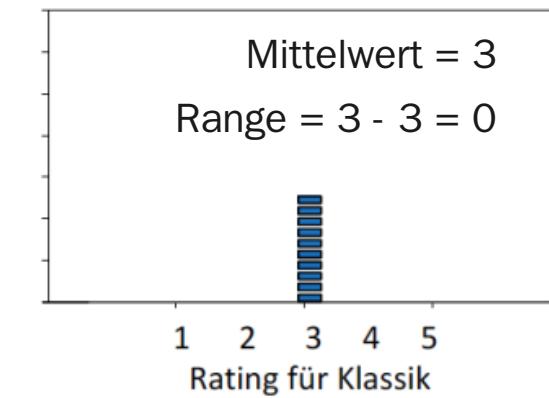
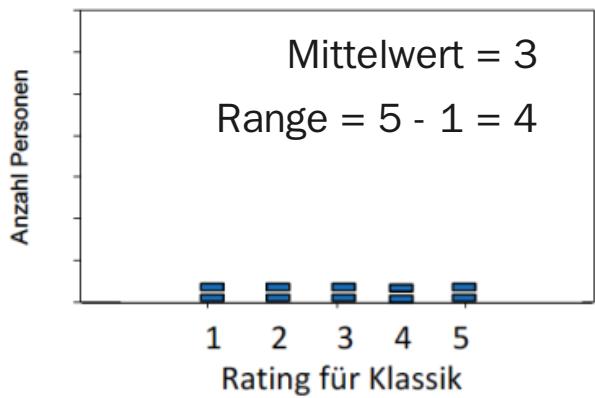
- Die **Spannweite** oder **Range** ist die Differenz zwischen dem kleinsten und dem größten Wert:

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

- Einfachstes **Streuungsmaß** – sinnvoll, um dem Leser einen Eindruck der gesamten Spannbreite von Daten zu geben.
- Allerdings kaum Aussagekraft über die tatsächliche Variabilität der Daten
 - Beispiel: die Spannbreite 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



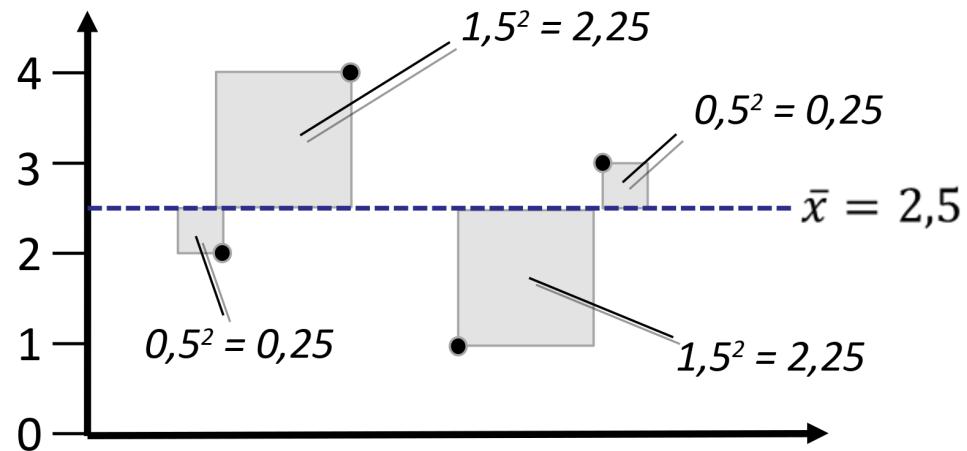
Varianz

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

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

(Die Varianz ist gleich der quadrierten Standardabweichung σ – 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 $n \rightarrow n - 1$ 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 μ berechnet werden.
- Das wahre μ 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 μ 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:



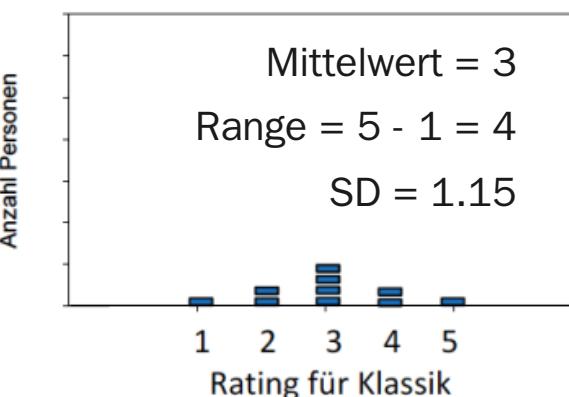
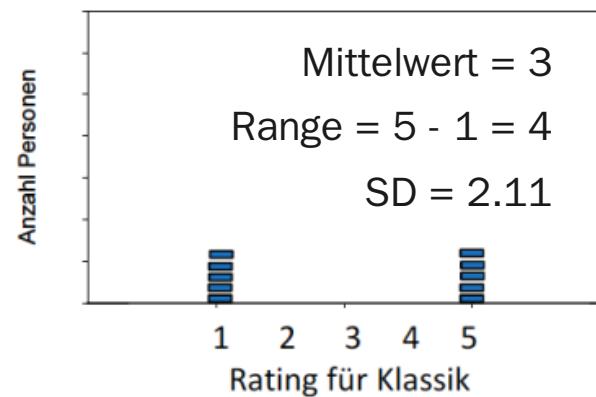
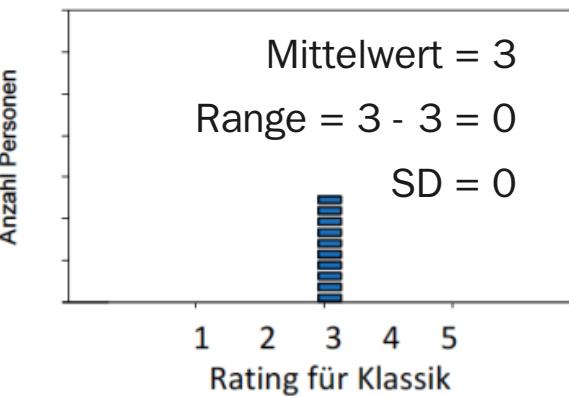
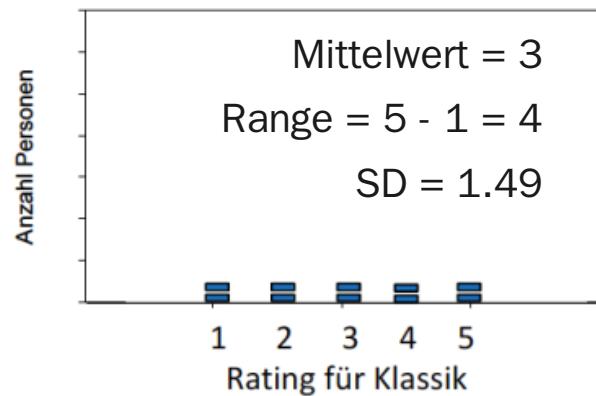
$$x = \{167 \text{ cm}, 181 \text{ cm}, 154 \text{ cm}, 192 \text{ cm}, 173 \text{ cm}\} \rightarrow \hat{Var}(X) = 205.3 \text{ cm}^2$$

- 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:

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

- 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 Datenwerte in gleich große Abschnitte ein (gleich groß = jeder Abschnitt hat gleich viele Datenpunkte)
 - Beispiel *Dezile*: Einteilung der Datenwerte in 10 gleich große Abschnitte
 - Hier: 20 Werte eingeteilt in 10 Abschnitte (Dezile) á 2 Werte

$\underbrace{1, 1,}_{\text{1.Dezil}} \underbrace{2, 2,}_{\text{2.Dezil}} 3, 4, 5, 5, 6, 6, 7, 7, 8, 9, 10, 10, 11, 11, \underbrace{12, 13}_{\text{10.Dezil}}$...

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

$$\underbrace{1, 1, 2,}_{\text{1.Quartil}} \underbrace{3, 3, 3,}_{\text{2.Quartil}} \underbrace{4, 4, 5,}_{\text{3.Quartil}} \underbrace{6, 6, 7}_{\text{4.Quartil}}$$

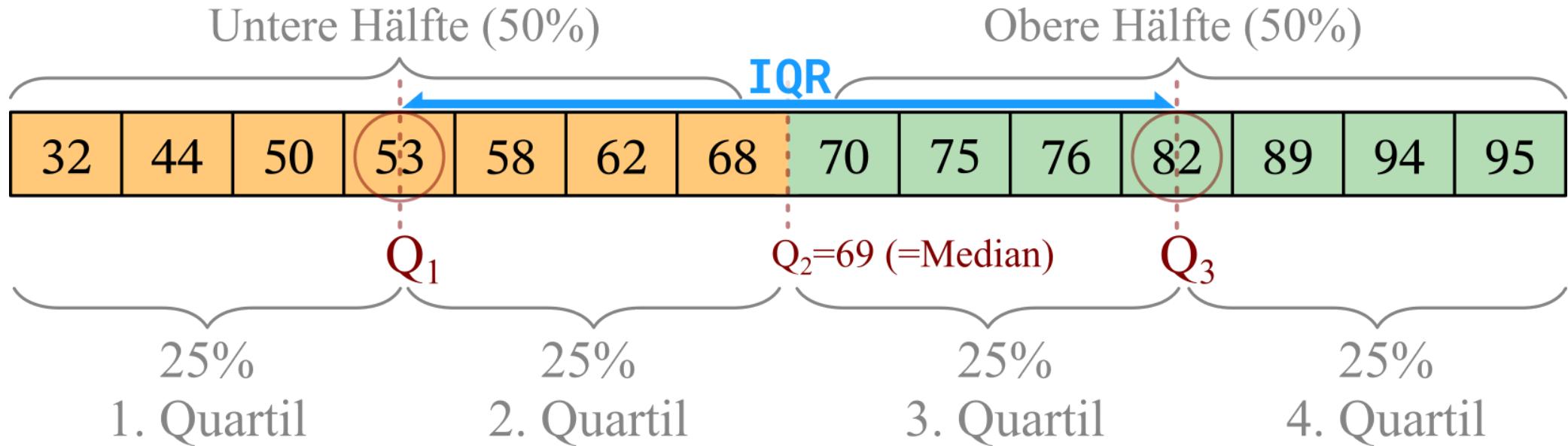
- Achtung: Zuordnung von Werten zu Quantilen nicht immer eindeutig

$$\underbrace{1, 1, 2,}_{\text{1.Quartil}} \underbrace{2, 3, 3,}_{\text{2.Quartil}} \underbrace{4, 4, 5,}_{\text{3.Quartil}} \underbrace{6, 6, 7}_{\text{4.Quartil}}$$

Sollte hier die rote 2 zum 1. Quartil oder 2. Quartil gehören? ⇒ unklar!

Glücklicherweise spielt dieses Problem für den Interquartilsabstand keine Rolle.

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),

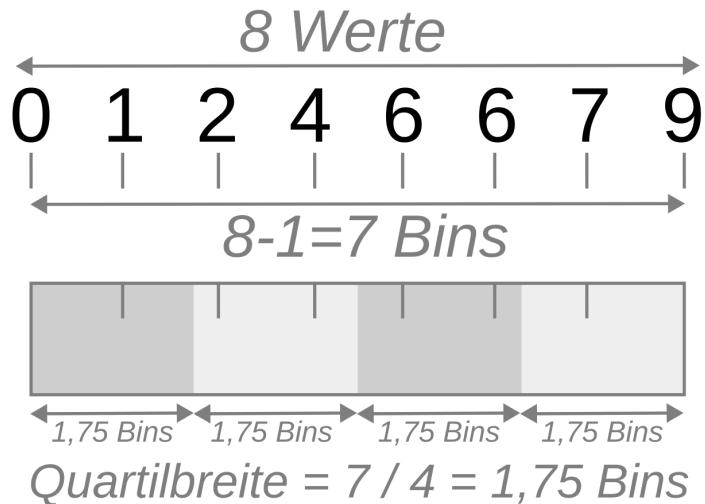
also $Q_1 = 2.75$; der “3.75”-te Wert von hinten befindet sich auf 75% der Strecke zwischen der 7 und der 6,

also $Q_3 = 6.25$.

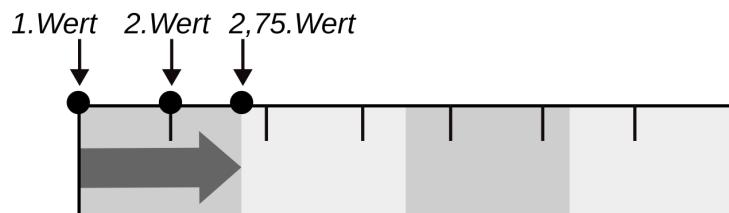
$$IQR = Q_3 - Q_1 = 6.25 - 2.75 = 3.5$$

Interquartilsabstand (interquartile range = IQR)

Beispiel

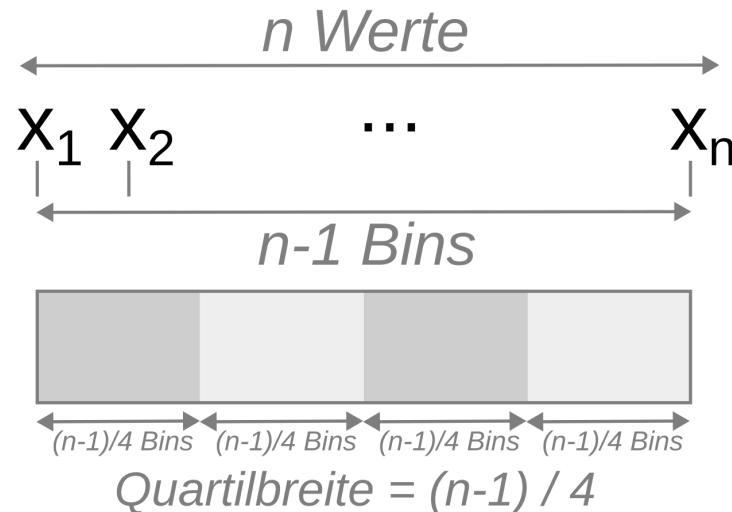


Wie tief muss man von links rein gehen für die erste Quartilsgrenze?

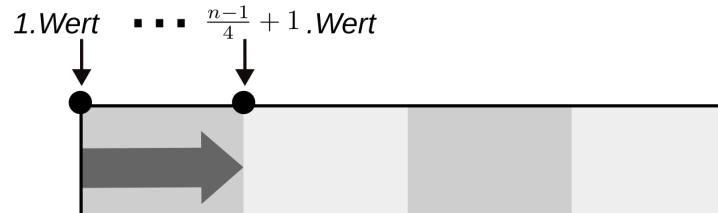


Man muss also 2,75 Werte von links in die Zahlenreihe hineingehen.

Allgemeiner Fall



Wie tief muss man von links rein gehen für die erste Quartilsgrenze?

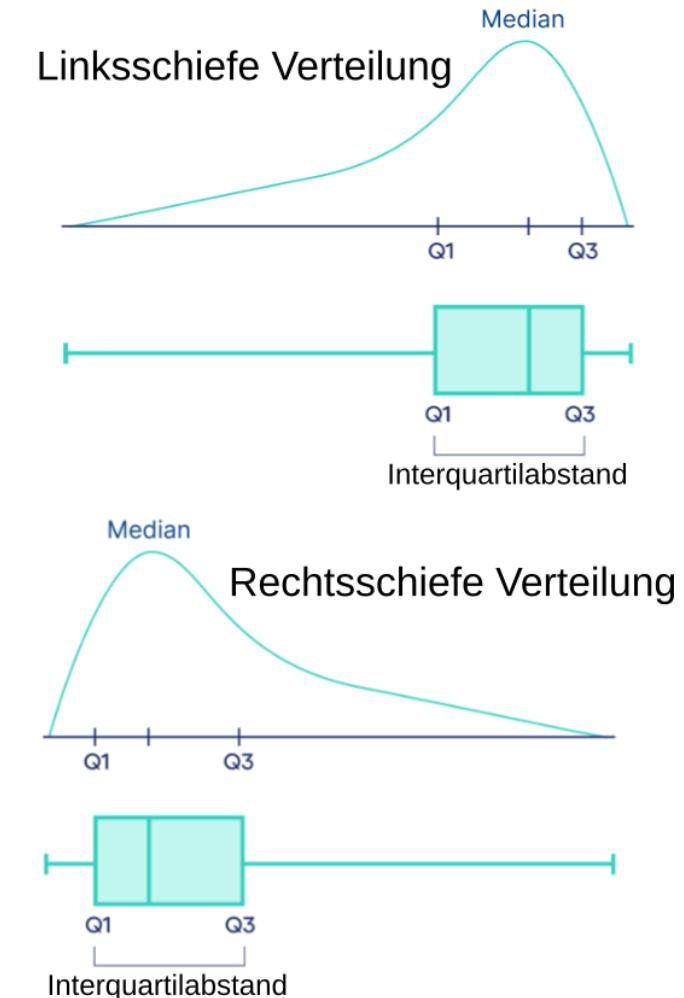


Man muss also $\frac{n-1}{4} + 1$ Werte von links in die Zahlenreihe hineingehen.

$$\frac{n-1}{4} + 1 = \frac{n-1}{4} + \frac{4}{4} = \frac{n-1+4}{4} = \frac{n+3}{4}$$

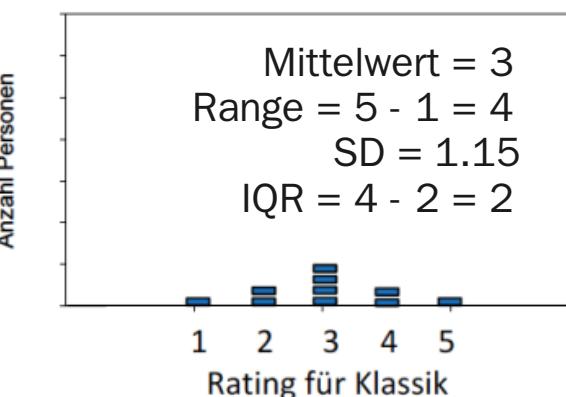
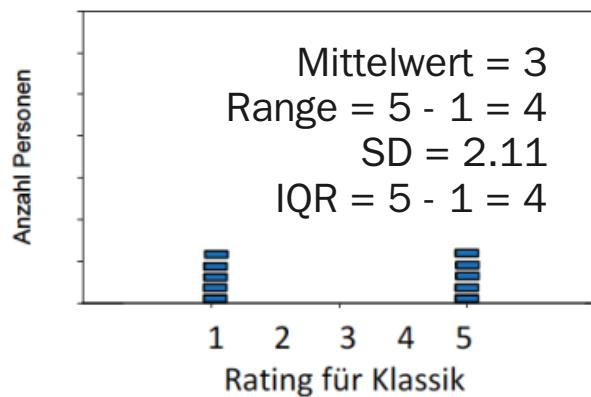
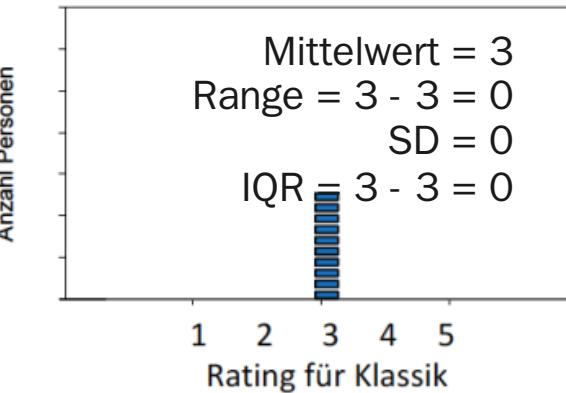
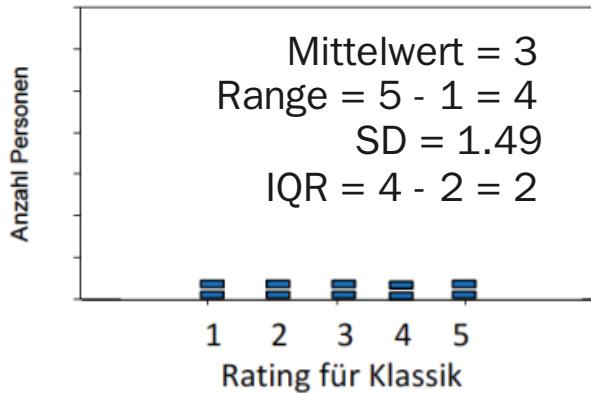
Wann ist der Interquartilsabstand ein sinnvolles Streuungsmaß?

- 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 basiert der Wert 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

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

Darstellung von Lage- und Streuungsmaßen in Text und Bild

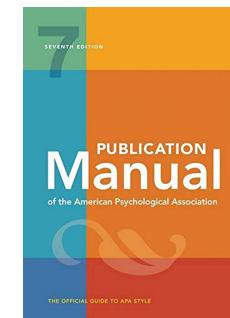
Angabe von Lage- und Streuungsmaßen in wissenschaftlichen Arbeiten

- Handelt es sich um einzelne Werte, können diese übersichtlich im Fließtext berichtet werden:

Respondents' mean rating for self-estimated musicality was 6.1 ($SD = 2.4$), and the mean rating for importance of music in their life was 8.2 ($SD = 1.8$). Thus, partici-

- Handelt es sich um eine größere Anzahl von Werten (z.B. bei mehreren Bedingungen) bietet sich eine Darstellung in Tabellenform an.
- Auch bei dieser Darstellung sollten Lage- und Streuungsmaße angegeben werden.
- In wissenschaftlichen Manuskripten werden Tabellen häufig nach den APA-Richtlinien (7. Edition, 2020) formatiert:

Musikstil	<i>M</i>	<i>SD</i>
Klassik	2,9	1,1
Rock	4,1	0,8
Rap	3,3	1,2



Siehe Link⁴ für Informationen zur Formatierung von Tabellen im APA-Stil

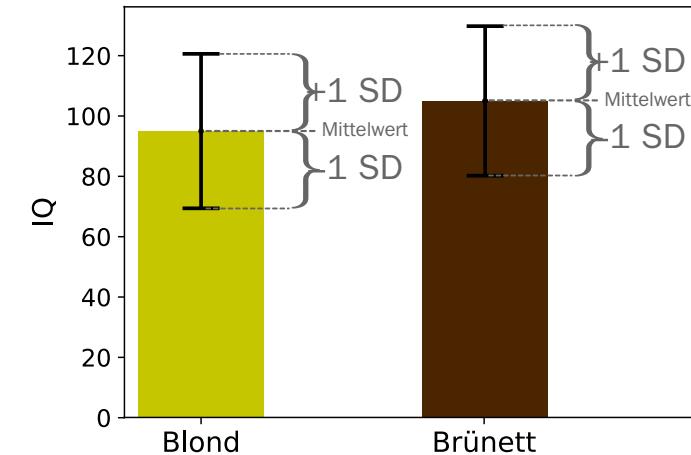
Beispieltabelle

TABLE 4 Mean ratings (and SDs) for the functions of music within the six dimensions of musical styles

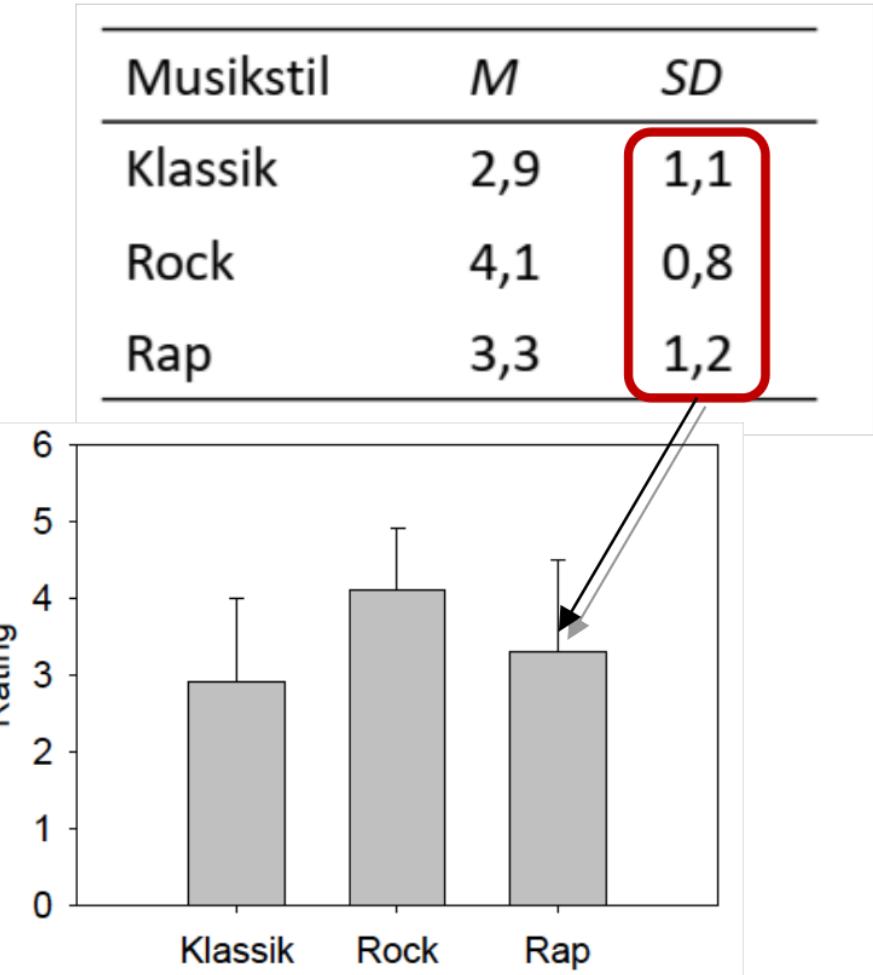
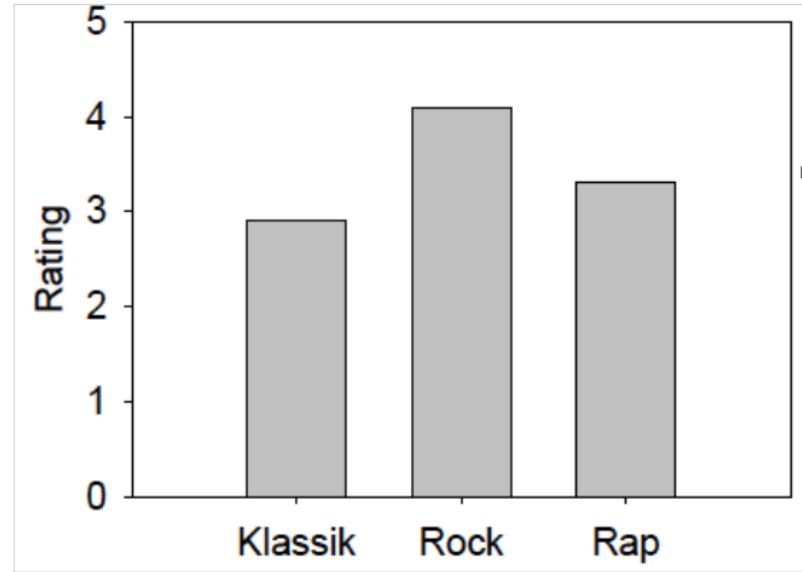
Statement	Preference dimensions (genres)						
	All styles (N = 503)	Sophisticated (n = 38)	Electronic (n = 30)	Rock (n = 218)	Rap (n = 37)	Pop (n = 88)	Beat, folk, & country music (n = 10)
Puts me in a good mood	8.2 (1.3)	8.3 (1.6)	8.4 (1.0)	8.3 (1.2)	8.0 (1.9)	7.8 (1.4)	8.7 (0.5)
Helps me chill and tune out	6.9 (2.3)	7.6 (2.1)	6.9 (2.6)	6.8 (2.4)	6.4 (2.9)	7.2 (1.7)	7.3 (2.2)
Energizes me	6.8 (2.0)	6.2 (2.3)	7.3 (1.8)	6.8 (1.9)	7.2 (2.0)	6.4 (2.1)	7.6 (1.8)
Lets me appreciate as art	6.5 (2.6)	8.1 (1.5)	6.4 (2.6)	6.7 (2.5)	5.7 (3.1)	5.7 (2.9)	6.3 (2.9)
Enables me to reminisce	6.4 (2.5)	5.2 (2.9)	6.1 (2.7)	6.8 (2.2)	6.0 (2.3)	6.6 (2.4)	7.8 (1.1)
Enables me to better understand my thoughts and feelings	6.0 (2.5)	5.8 (3.0)	5.5 (2.8)	6.2 (2.4)	5.0 (3.1)	5.9 (2.3)	6.5 (2.0)
Is what I listen to as background music	5.7 (2.8)	5.4 (2.8)	6.2 (2.3)	5.5 (2.8)	5.5 (2.9)	6.2 (2.7)	5.0 (3.1)
Is what I like to dance to	5.6 (3.2)	2.8 (2.9)	7.5 (2.8)	5.7 (3.0)	7.4 (2.3)	5.4 (3.0)	5.5 (3.2)
Expresses my identity	5.5 (2.7)	5.5 (2.3)	5.2 (2.4)	6.0 (2.6)	4.7 (2.9)	4.9 (2.7)	6.3 (2.8)
Expresses my values	5.5 (2.6)	5.7 (2.6)	4.1 (2.9)	6.0 (2.5)	5.2 (2.4)	5.1 (2.5)	6.3 (2.5)
Lets me forget my problems	5.5 (2.6)	5.8 (2.3)	6.0 (2.4)	5.3 (2.7)	4.5 (2.9)	5.1 (2.4)	6.6 (1.8)
Helps me feel close to others	5.4 (2.7)	5.0 (3.1)	5.2 (2.8)	5.6 (2.6)	5.1 (2.9)	5.3 (2.5)	5.3 (2.8)
Makes me feel ecstatic	4.9 (3.2)	3.8 (3.1)	6.5 (3.1)	5.3 (3.0)	4.3 (3.2)	3.6 (3.0)	6.8 (1.8)
Lets me experiment with different sides of my personality	4.8 (3.0)	3.6 (2.8)	4.7 (3.4)	5.1 (2.9)	4.8 (3.1)	4.4 (2.9)	4.6 (3.7)
Helps me meet people	4.4 (2.8)	3.6 (3.0)	5.1 (3.0)	4.7 (2.8)	4.3 (2.8)	3.8 (2.8)	4.0 (2.7)
Makes me identify with the artists	3.4 (2.9)	3.5 (3.3)	3.3 (3.2)	3.6 (2.8)	3.2 (2.9)	2.7 (2.5)	4.3 (3.4)
Gives me information	3.3 (2.8)	3.5 (3.1)	1.9 (2.3)	3.6 (2.7)	2.8 (2.8)	2.8 (2.4)	4.1 (3.0)

Balkendiagramm mit Fehlerbalken

- Der **Fehlerbalken** ist die klassische Wahl zur Darstellung der Streuung von Daten.
- Die Ausdehnung des Fehlerbalkens entspricht dem Lagemaß (z.B. Mittelwert) plus/minus dem Streuungsmaß (z.B. Standardabweichung).
- Der gesamte Fehlerbalken hat also die Ausdehnung $2 \cdot \text{Streuungsmaß}$
- Klassische Kombinationen von Lagemaß und Streuungsmaß sind:
 - Mittelwert \pm Standardabweichung
 - Mittelwert \pm Standardfehler (dazu kommen wir noch)
 - Mittelwert \pm Konfidenzintervall (dazu kommen wir noch)
 - Median \pm IQR
 - Median \pm Median-Abweichung (das behandeln wir nicht)

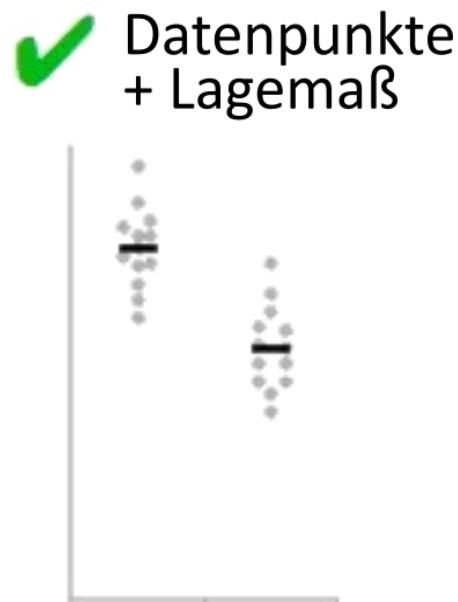
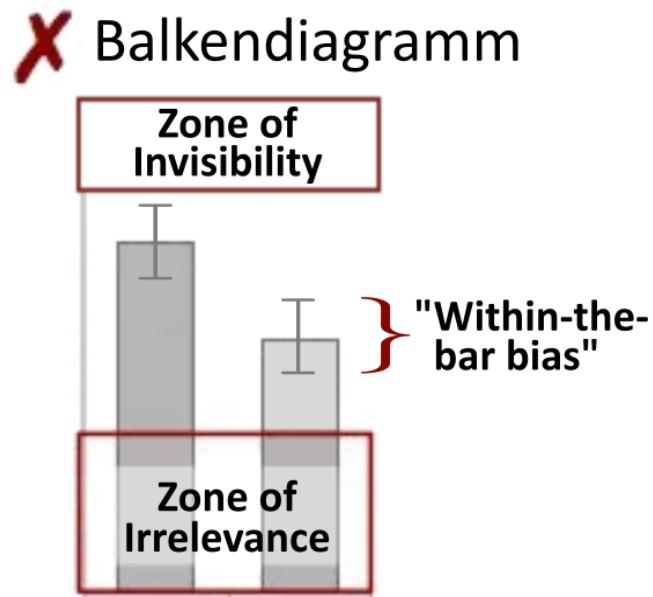


Beispiel



The case against bar plots

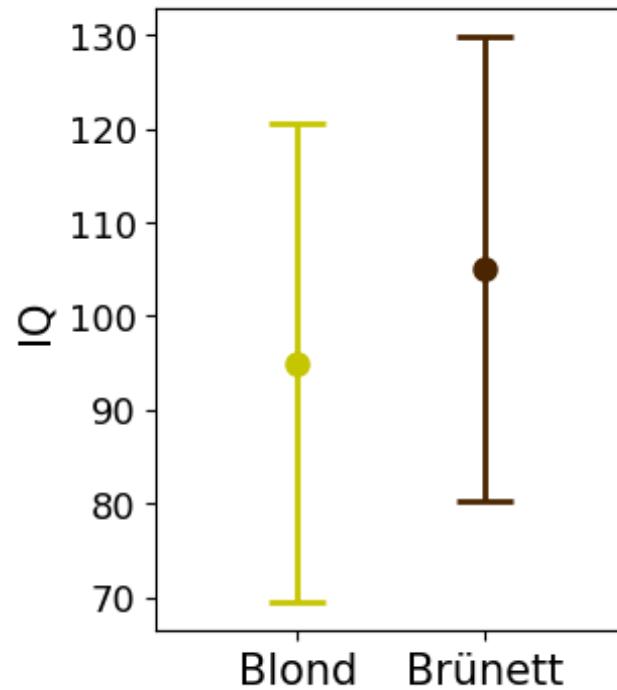
- Trotz ihrer hohen Verbreitung haben Balkendiagramme (Barplots) eine Reihe von Nachteilen:
 - Sie geben kaum Information über die spezifische Verteilung der Daten und mögliche Ausreißer
 - Die intransparente Darstellungsweise verdeckt häufig, dass 1) Daten unrealistisch sind oder 2) Ausreißer das Lagemaß verzerren oder 3) die Verteilung der Daten unpassend für das verwendete Lagemaß sind.
 - Sie legen den Fokus auf irrelevante Bereiche der Skala (siehe Abbildung unten)



Durch die tatsächliche Verteilung der Datenpunkte im rechten Plot wird klar, dass im Balkendiagramm ein vergleichsweise starker Fokus auf Bereiche gelegt wird, in den gar keine Daten enthalten sind ("Zone of Irrelevance"), und andererseits Extremwerte, insbesondere oberhalb des Fehlerbalkens, visuell völlig unrepräsentiert sind ("Zone of Invisibility"). Der Fehlerbalken ist außerdem leicht mit der Illusion verbunden, dass sich alle Datenpunkte innerhalb des angezeigten Bereiches befinden ("Within-the-bar bias").⁵.

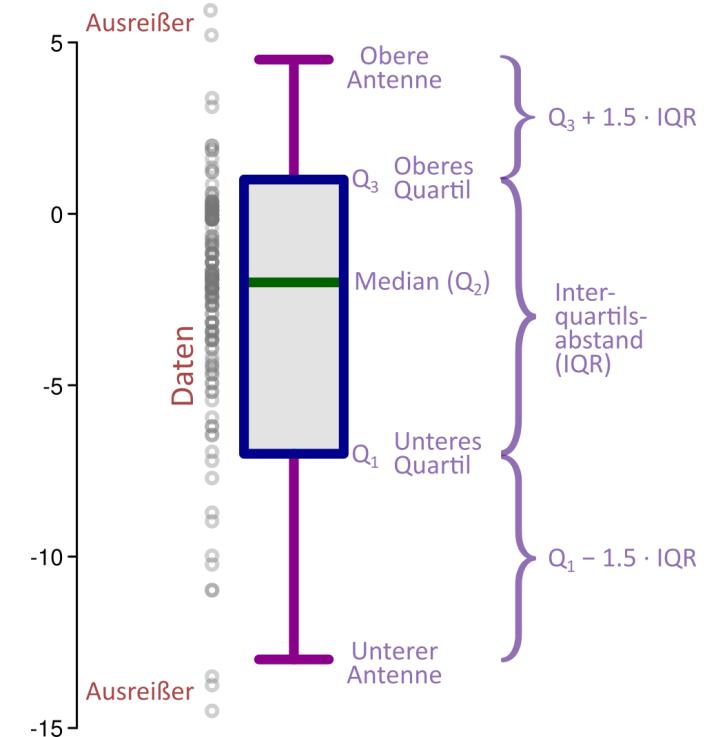
Einzelner Punkt statt Balken

- Statt eines Balkens kann der Mittelwert auch durch einen einzelnen Punkt repräsentiert werden – manche empfinden das als eleganter:



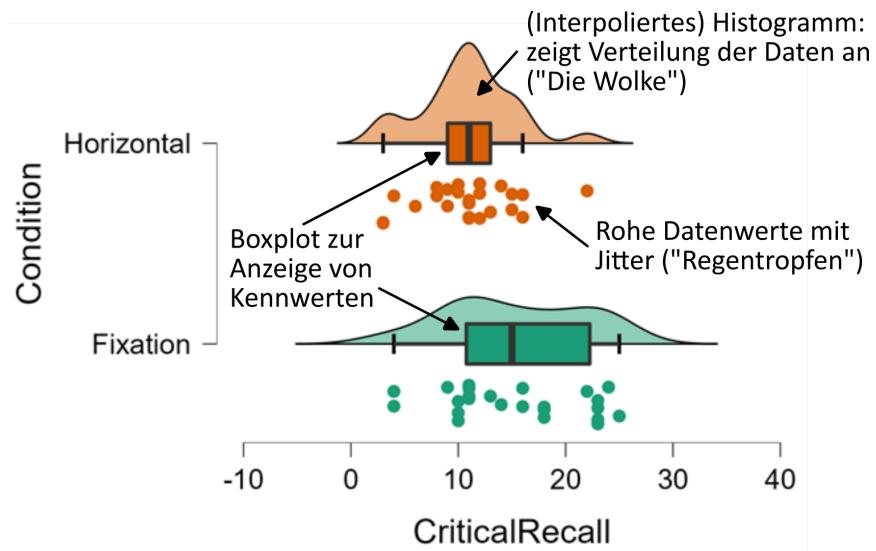
Der Box-Plot

- Eine bekannte Darstellungsform für den Median ist der **Boxplot**
- In einer gängigen Variante zeigt der Boxplot drei Informationen an:
 1. **Median (als einfache Linie)**
 2. **Box:** die “mittleren 50% der Daten” (25% über dem Median, 25% unter dem Median)
 3. **Antennen (“Whiskers”):** zeigen Grenzen der Ausreißer-Definition an (Ausreißer = alle Punkte unter- und überhalb der Antennen). Häufig ist hier das Kriterium, dass die Daten im Bereich $[Q_1 - 1.5 \cdot IQR; Q_3 + 1.5 \cdot IQR]$ liegen müssen.
- Der Boxplot gibt eine schnelle Übersicht über wesentliche Kennwerte eines Datensatzes
- Zu beachten ist, dass zahlreiche Varianten des Boxplots existieren
 - Bei einer weiteren häufigen Variante zeigen die Antennen das absolute Maximum und Minimum der Daten an (also *inklusive* möglicher Ausreißer)



Moderne Darstellungsformen

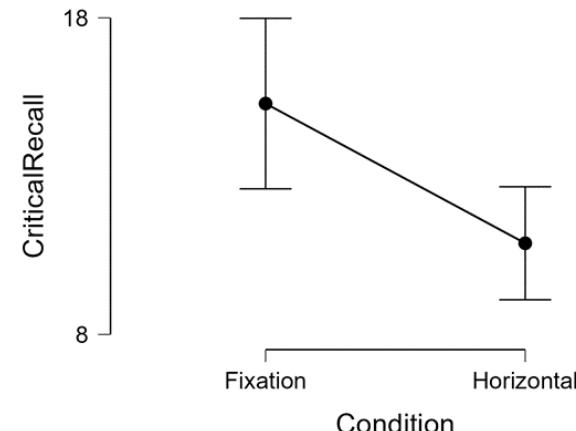
Raincloud-Plot



Bildnachweis⁶

- Moderne Statistik- und Plotsoftware ermöglicht heutzutage eine verbesserte und transparentere Darstellung von Daten:
 - Anzeige einzelner Datenpunkte, meist getrennt durch “Jitter” (d.h. leichte horizontale oder vertikale Versetzung mit zufälligen Abständen, um Überschneidung der Datenpunkte zu reduzieren)
 - Anzeige der Verteilung mittels (interpolierter) Histogramme (wichtige Information für die Auswahl geeigneter Lage- und Streumaße, aber auch statistischer Tests)
 - Zusätzliche Anzeige von Kennwerten

Klassische Darstellung mit Lage- und Streuungsmaß

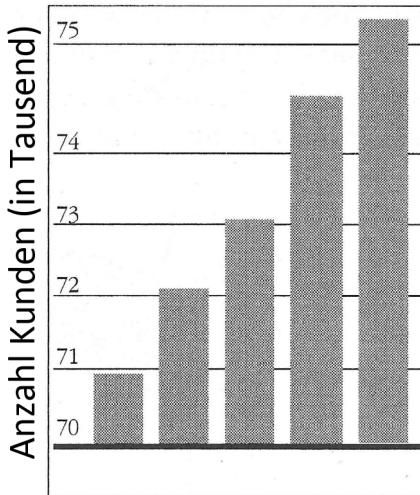


Bildnachweis⁷

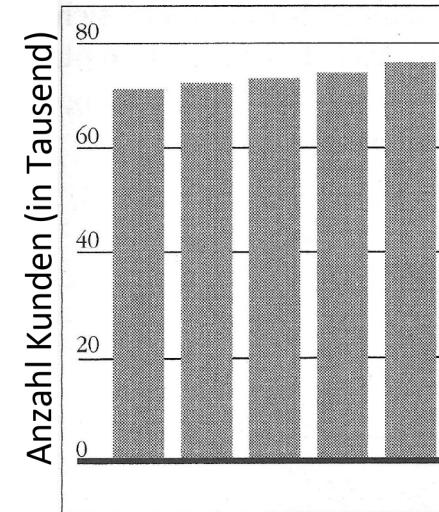
Abschneiden der y-Achse

- Das Abschneiden der y-Achse verzerrt häufig die Stärke von Effekten.

Kundenentwicklung wie von einer deutschen Bank dargestellt

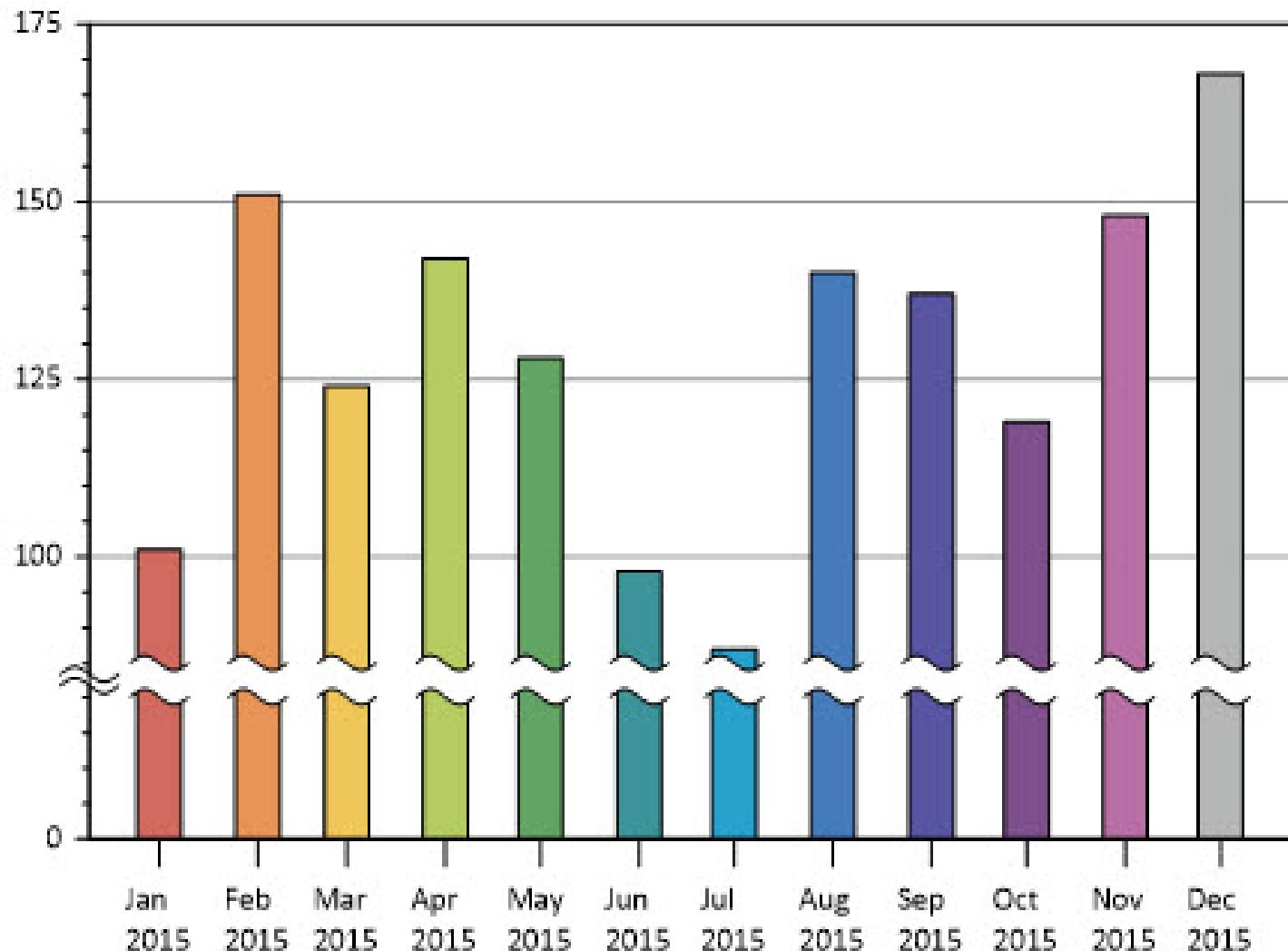


Kundenentwicklung mit Beginn der y-Achse bei 0



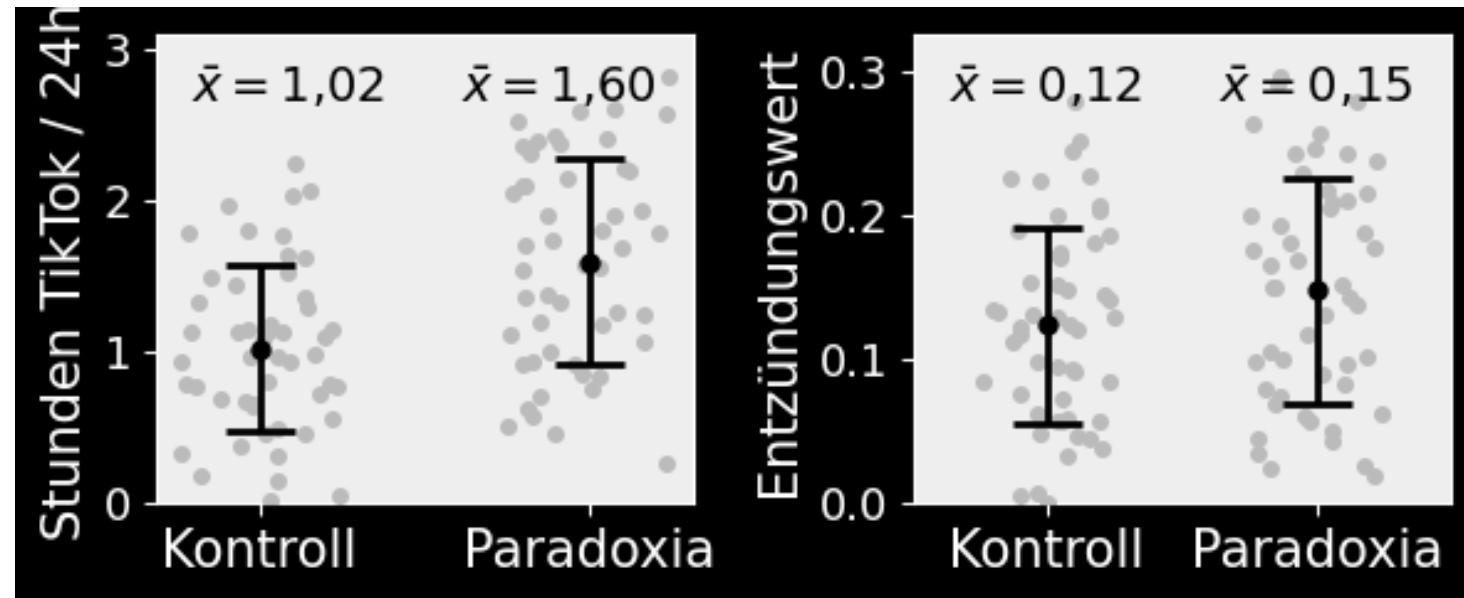
- Die grundsätzliche Empfehlung ist daher, die y-Achse bei 0 beginnen zu lassen.
- Es gibt aber Ausnahmen:
 - Vorliegen eines anderen natürlichen Referenzwertes (z.B. IQ-Wert 100, wenn alle Werte über 100 liegen).
 - Wären *tatsächlich vorhandene Unterschiede* zwischen Balken verschiedener Bedingungen überhaupt nicht mehr wahrnehmbar, kann ein Abschneiden der y-Achse sinnvoll sein (oder eine Logarithmus-Skala!).
 - In manchen Fällen können Messwerte niemals unter einen Mindestwert fallen. Beispielsweise sind motorische Reaktionszeiten physiologisch bedingt fast immer über 100ms – in diesem Fall ist der Bereich 0-100ms “Totraum” und kann sinnvollerweise weggelassen werden.
 - Im Idealfall wird das Abschneiden der y-Achse durch einen “Bruch” angezeigt (siehe nächste Folie).

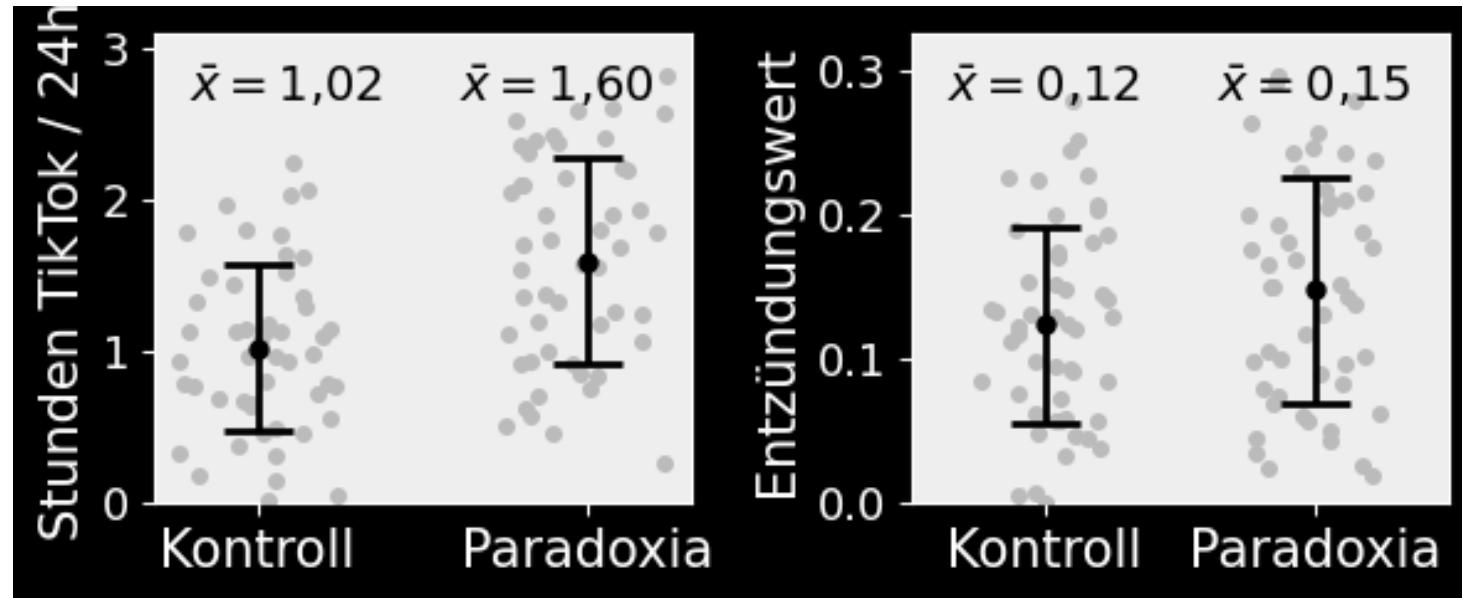
Beispiel für gebrochene y-Achse



Durch den Bruch der y-Achse wird betont, dass die Balken zur besseren Übersichtlichkeit "abgeschnitten" wurden. So werden fälschliche Wahrnehmungen und Interpretationen durch die abgeschnittene Achse eher vermieden. Bildnachweis⁸

Die Daten in Ihrer Beobachtungsstudie weisen keine größeren Ausreißer auf uns 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!*

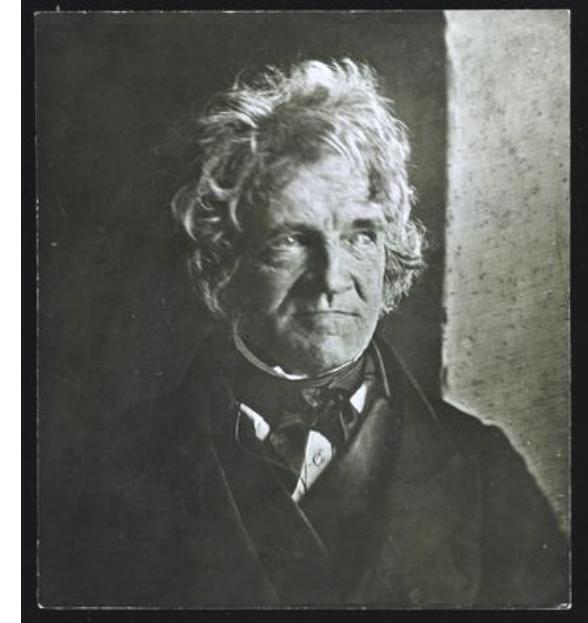
Bonuscontent

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)⁹

- 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 σ^2 der Population systematisch unterschätzt.
- Der Grund ist liegt im Stichprobenmittelwert \bar{x} : er ist keine perfekte Schätzung für den wahren Mittelwert μ ist – er streut um das “wahre” μ .
- Diese Streuung werden wir noch kennenlernen – es ist der Standardfehler $\hat{se} = \frac{\hat{\sigma}}{\sqrt{n}}$. Als Varianz formuliert:

$$\hat{se}^2 = \frac{\hat{\sigma}^2}{n}$$

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

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

- 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% gültig¹⁰ – 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 Abstandsquadrate ist für alle x differenzierbar
(hingegen ist $\hat{Var}_{\text{absolut}}$ bei $x = 0$ nicht differenzierbar)



Weitergehende Literatur ¹¹

Fußnoten

1. <https://datatab.de/tutorial/mittelwert-median-modus>
 2. <https://youtu.be/inJ4OvU0zMA>
 3. <https://www.shiksha.com/online-courses/articles/measures-of-dispersion-range-iqr-variance-standard-deviation/>
 4. <https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables>
 - 5.
- Weissgerber TL, Winham SJ, Heinzen EP, Milin-Lazovic JS, Garcia-Valencia O, Bukumiric Z, Savic MD, Garovic VD, Milic NM (2019) Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. *Circulation* 140:1506–1518.
6. <https://jasp-stats.org/2021/10/05/raincloud-plots-innovative-data-visualizations-in-jasp/>
 7. <https://jasp-stats.org/2021/10/05/raincloud-plots-innovative-data-visualizations-in-jasp/>
 8. <https://support.goldensoftware.com/hc/en-us/articles/360019762133-Using-Break-Axis-in-Grapher>
 9. <http://www.digiporta.net/index.php?id=658493715>
 10. https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation
 11. https://web.archive.org/web/20221024193801/https://www4.hcmut.edu.vn/~ndlong/TK/mat/04_standard_deviation_vs_absolute_deviation.pdf