Thema 04:

Univariate Deskriptivstatistik

QM1, SoSe 22

HOCHSCHULE
ANSBACH

Überblick

Die deskriptive Statistik ist da, das Leben zu vereinfachen

… hört sich erstmal unglaubwürdig an 🤓

Puh, so viele Zahlen. Ich check nix!

Vorbereitungszeit für die Klausur pro Student

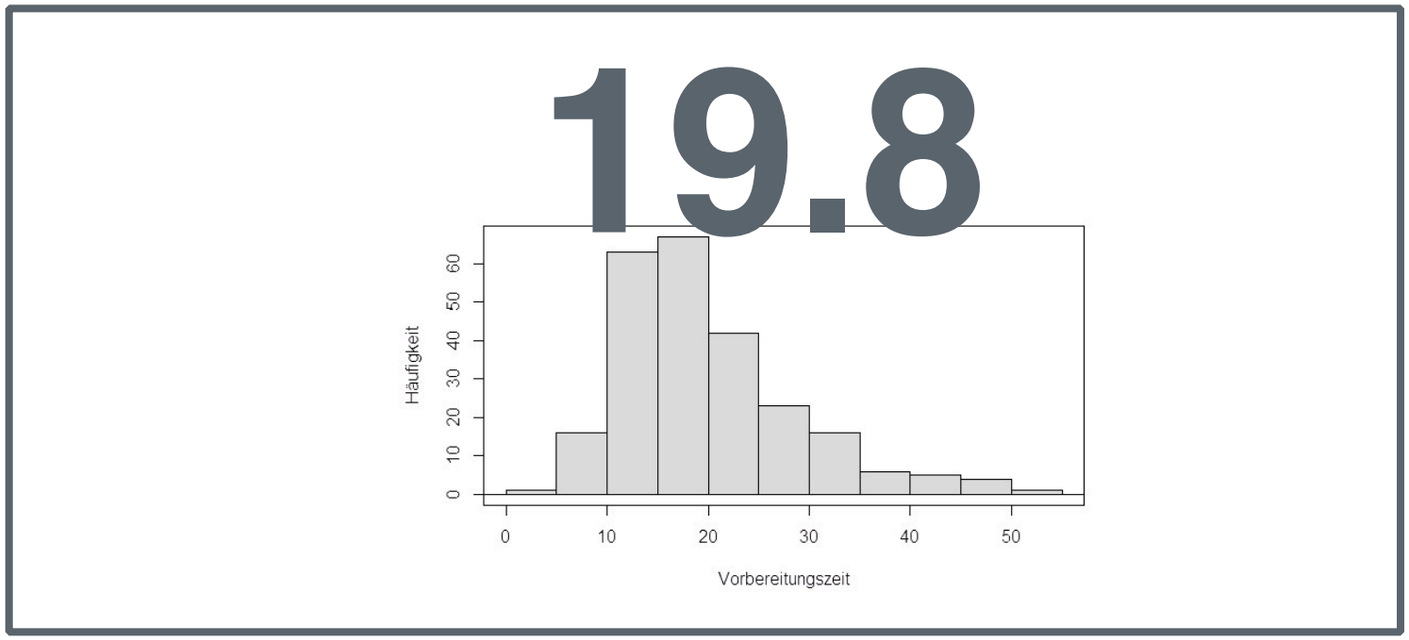
[1] 16.99 10.34 21.01 23.68 24.59 25.29 8.77 26.88 15.04 14.78 10.27 35.26 15.42 18.43 14.83 [16] 21.58 10.33 16.29 16.97 20.65 17.92 20.29 15.77 39.42 19.82 17.81 13.37 12.69 21.70 19.65 [31] 9.55 18.35 15.06 20.69 17.78 24.06 16.31 16.93 18.69 31.27 16.04 17.46 13.94 9.68 30.40 [46] 18.29 22.23 32.40 28.55 18.04 12.54 10.29 34.81 9.94 25.56 19.49 38.01 26.41 11.24 48.27 [61] 20.29 13.81 11.02 18.29 17.59 20.08 16.45 3.07 20.23 15.01 12.02 17.07 26.86 25.28 14.73 [76] 10.51 17.92 27.20 22.76 17.29 19.44 16.66 10.07 32.68 15.98 34.83 13.03 18.28 24.71 21.16 [91] 28.97 22.49 5.75 16.32 22.75 40.17 27.28 12.03 21.01 12.46 11.35 15.38 44.30 22.42 20.92 [106] 15.36 20.49 25.21 18.24 14.31 14.00 7.25 38.07 23.95 25.71 17.31 29.93 10.65 12.43 24.08 [121] 11.69 13.42 14.26 15.95 12.48 29.80 8.52 14.52 11.38 22.82 19.08 20.27 11.17 12.26 18.26 [136] 8.51 10.33 14.15 16.00 13.16 17.47 34.30 41.19 27.05 16.43 8.35 18.64 11.87 9.78 7.51 [151] 14.07 13.13 17.26 24.55 19.77 29.85 48.17 25.00 13.39 16.49 21.50 12.66 16.21 13.81 17.51 [166] 24.52 20.76 31.71 10.59 10.63 50.81 15.81 7.25 31.85 16.82 32.90 17.89 14.48 9.60 34.63 [181] 34.65 23.33 45.35 23.17 40.55 20.69 20.90 30.46 18.15 23.10 15.69 19.81 28.44 15.48 16.58 [196] 7.56 10.34 43.11 13.00 13.51 18.71 12.74 13.00 16.40 20.53 16.47 26.59 38.73 24.27 12.76 [211] 30.06 25.89 48.33 13.27 28.17 12.90 28.15 11.59 7.74 30.14 12.16 13.42 8.58 15.98 13.42 [226] 16.27 10.09 20.45 13.28 22.12 24.01 15.69 11.61 10.77 15.53 10.07 12.60 32.83 35.83 29.03 [241] 27.18 22.67 17.82 18.78

Quelle

Prof. Dr. I. Ch. Weiß-Ois

Die deskriptive Statistik fasst Datenmassen zusammen

Vorbereitungszeit für die Klausur **im Schnitt**



Vorb.zeit
16.99
10.34
21.01
23.68
…

Zusammenfassen

Aus Spalte wird Zahl

MW
19.80

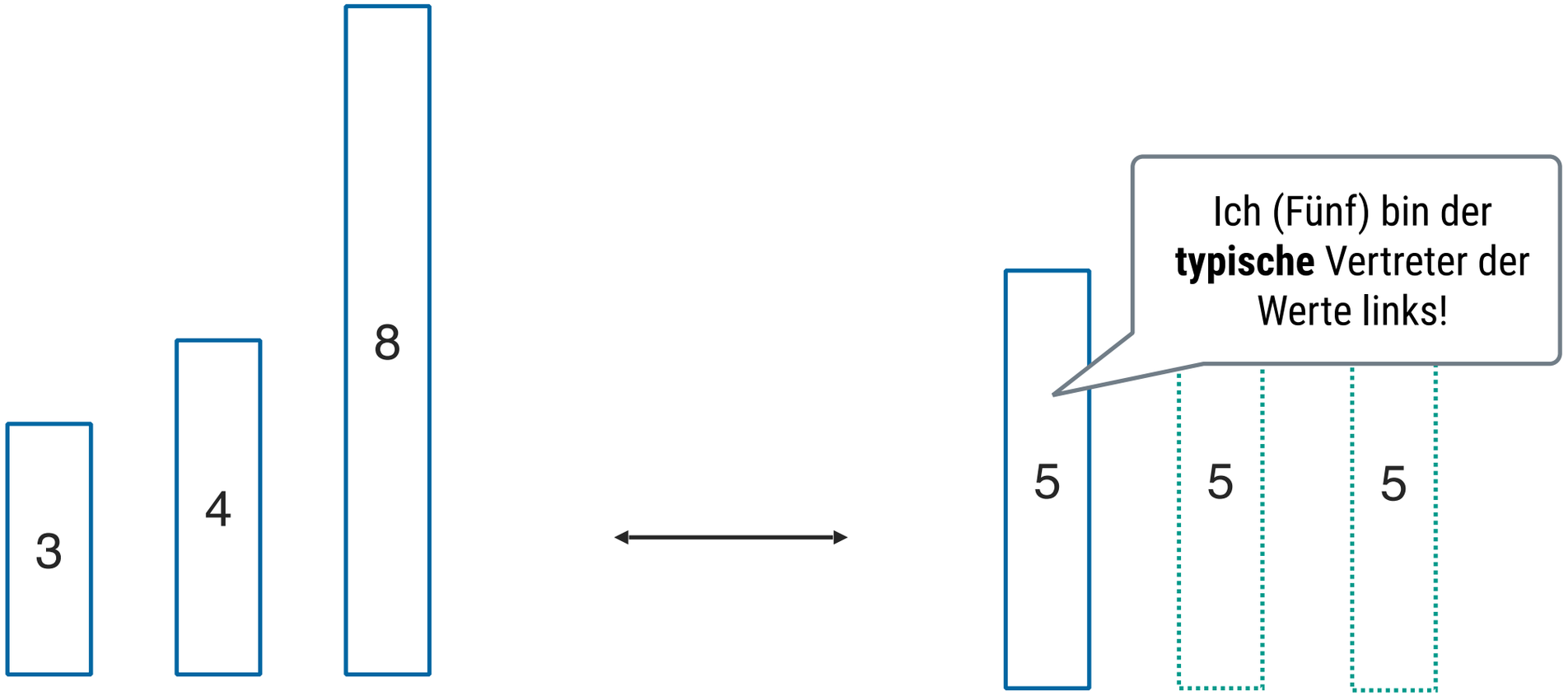
Ah! 20 Stunden lernen die Studis im Schnitt! Viel zu wenig natürlich!

Quelle

Prof. Dr. I. Ch. Weiß-Ois

Lagemaße

Ein Lagemaß sucht einen „typischen“ Vertreter



Ein Lagemaß gibt die Lage des typischen Werts in einer Reihe von Werte (Verteilung) an.
Entsprechend spricht man auch von der „zentralen Tendenz“ einer Verteilung.

Wenn ich alle Werte der Verteilung durch einen Wert ersetzen sollte, so dass jeder Wert dadurch „gut“ repräsentiert ist, welchen Wert würde ich wählen?
Diesen Wert bezeichnet man als Lagemaß.
Es gibt verschiedene Antworten auf diese Frage.

Das arithmetische Mittel ist ein Beispiel für ein Lagemaß

Synonym: Mittelwert, M, MW, aM, Durchschnitt, Mittel, $\bar{X}$ oder ∅(X)

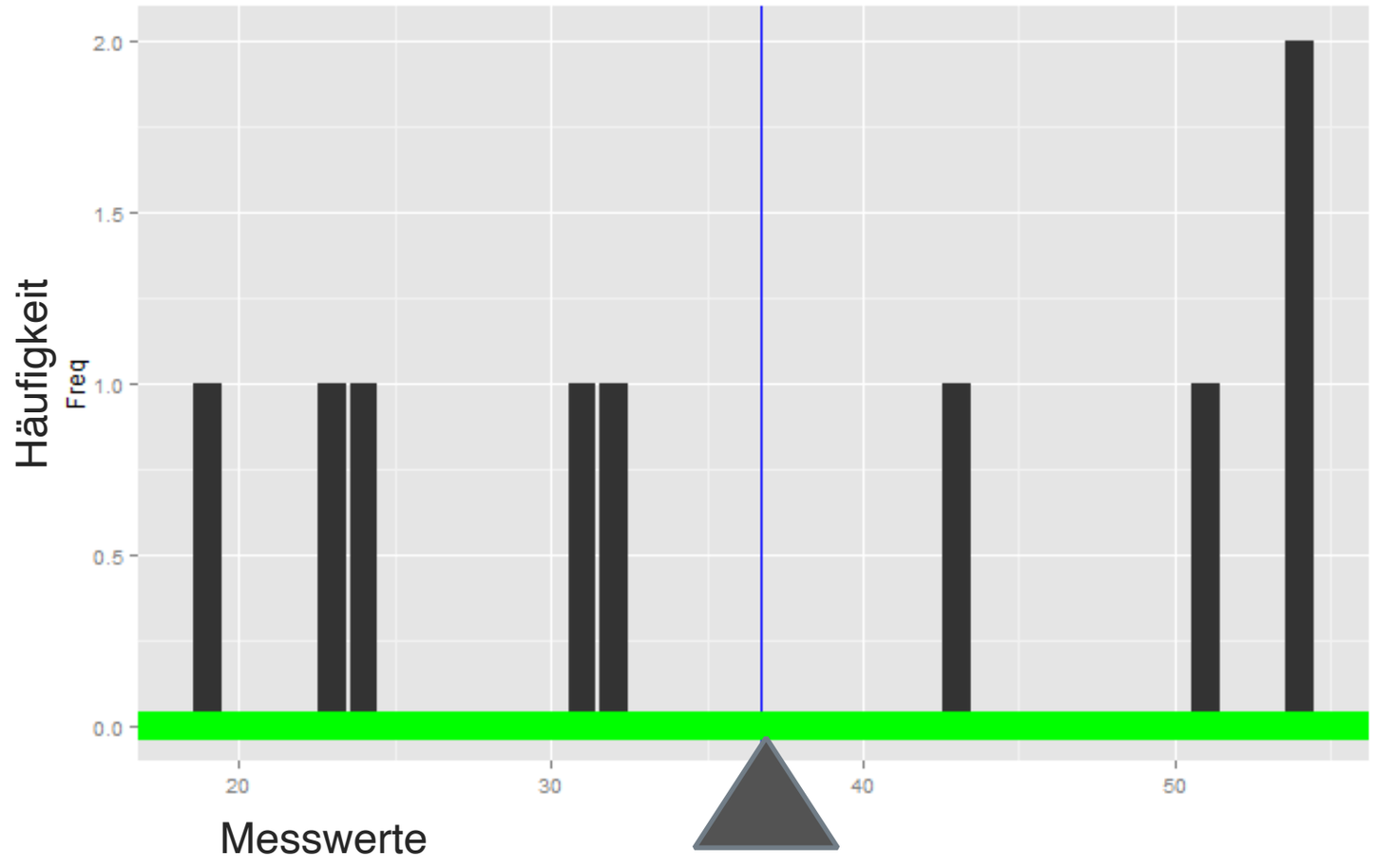
Wenn wir von Durchschnitt sprechen, meinen wir i.d.R. das arithmetische Mittel
Der Mittelwert berechnet sich als Summe aller Einzelwerte geteilt durch deren Anzahl *n*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

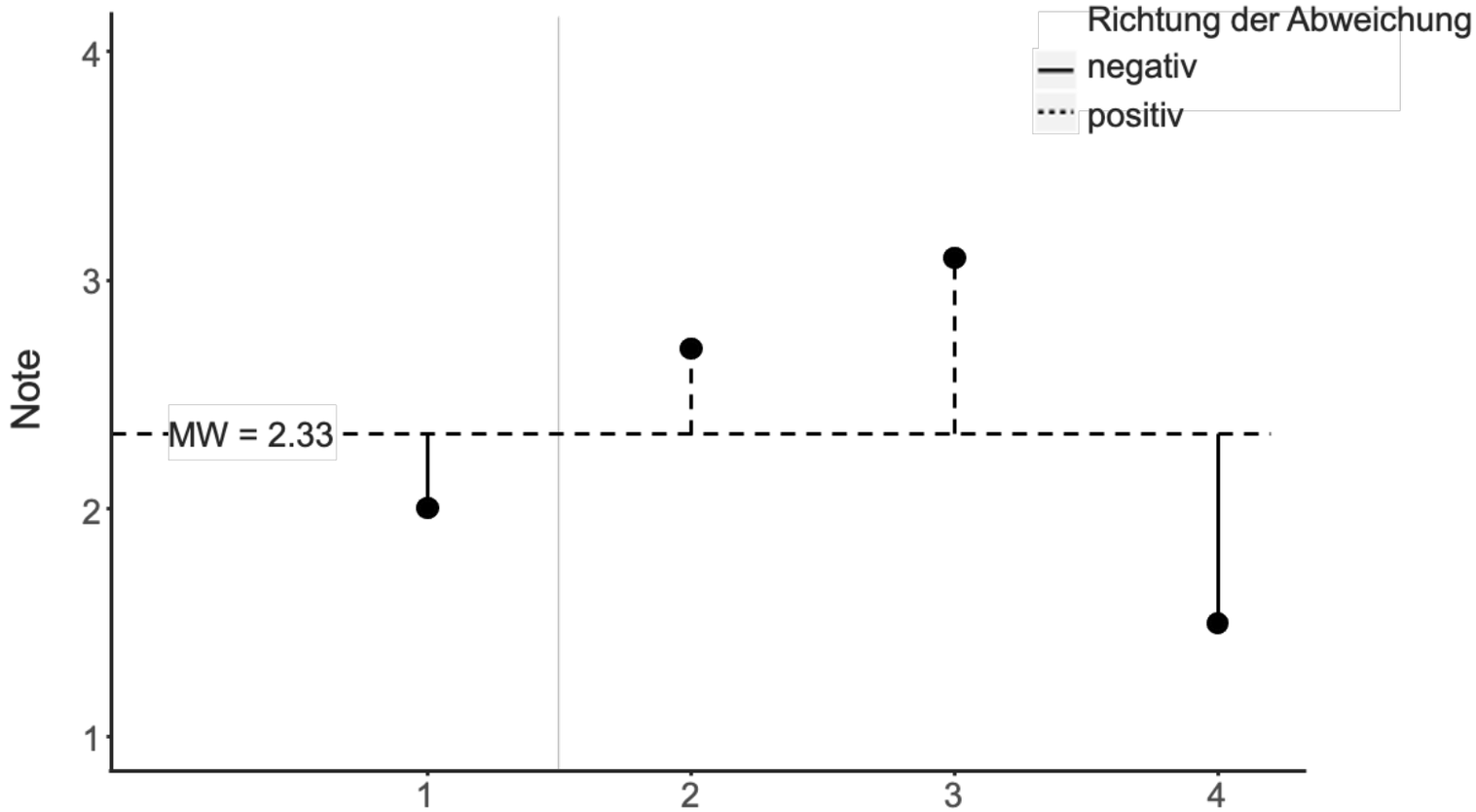
Das arithmetische Mittel als Waage oder Wippe

Der Mittelwert kann als der Wert einer Wippe veranschaulicht werden, an dem die Wippe im Schwerpunkt liegt.

Die Messwerte sind dabei wie Legosteine auf der Wippe aufgereiht.

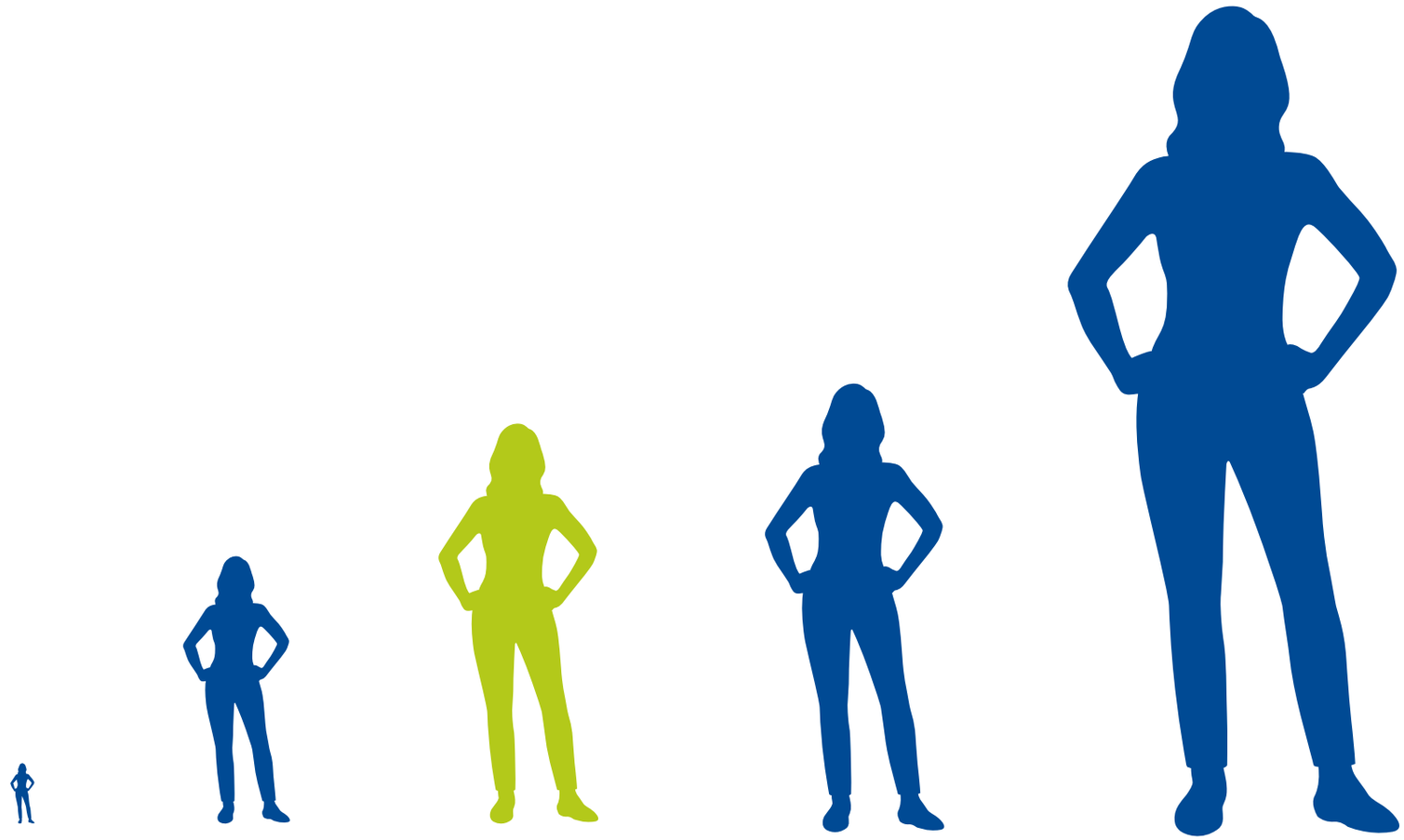


Die Abweichungen vom Mittelwert summieren sich zu Null auf



$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

So bastelt man sich einen Median



1. Sortiere die Messobjekte aufsteigend.
2. Finde das Messobjekt, zu dem es gleich viele Objekte mit größerem und kleineren Wert gibt.
3. Der Wert dieses Objekts ist der Median.

Median

- Der Median (Md, md) ist definiert als die Merkmalsausprägung, die bei (aufsteigend) sortierten Beobachtungen in der Mitte liegt.
  - Er beschreibt den mittleren Wert einer Verteilung (bei ungeradem n); der mittlere Wert einer Verteilung ist derjenige, zu dem es gleich viele kleinere und größere Werte gibt.
  - Bei geradem n werden die beiden mittleren Werte betrachtet und das arithmetische Mittel aus diesen beiden Werten gebildet: Bei der Messreihe 1, 2, 3, **4, 5**, 6, 8, 9 beträgt der Median 4.5.
  - Der Median kann ab ordinalskalierten Daten verwendet werden.

n ist gerade: md $= x_{(n+1)/2}$

n ist UNgerade: md $= \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$

Arithmetischer Mittelwert und Median

- Der **arithmetische Mittelwert** minimiert die Summe der quadratischen Abweichungen der Beobachtungen von einer Zahl c (Beweis):

$$\bar{x} = \arg\min_c \sum_i^n (c_i - c)^2$$

- Er ist der Durchschnitt in dem Sinne, dass alle Merkmalsträger den gleichen Anteil an der Merkmalssumme haben.
- Der **Median** minimiert die Summe der absoluten Abweichungen der Beobachtungen von einer Zahl c (Beweis):

$$\text{md} = \arg\min_c \sum_i^n |(c_i - c)|$$

- Der Median ist die Merkmalsausprägung eines (im Sinne des Merkmals) typischen, d. h. mittleren Merkmalsträgers.
- Der Median ist robust gegen Ausreißer, der arithmetische Mittelwert nicht. D. h. $\bar{X}$ kann stark durch einzelne extreme Werte verändert werden, MD nicht.

Hinweis:

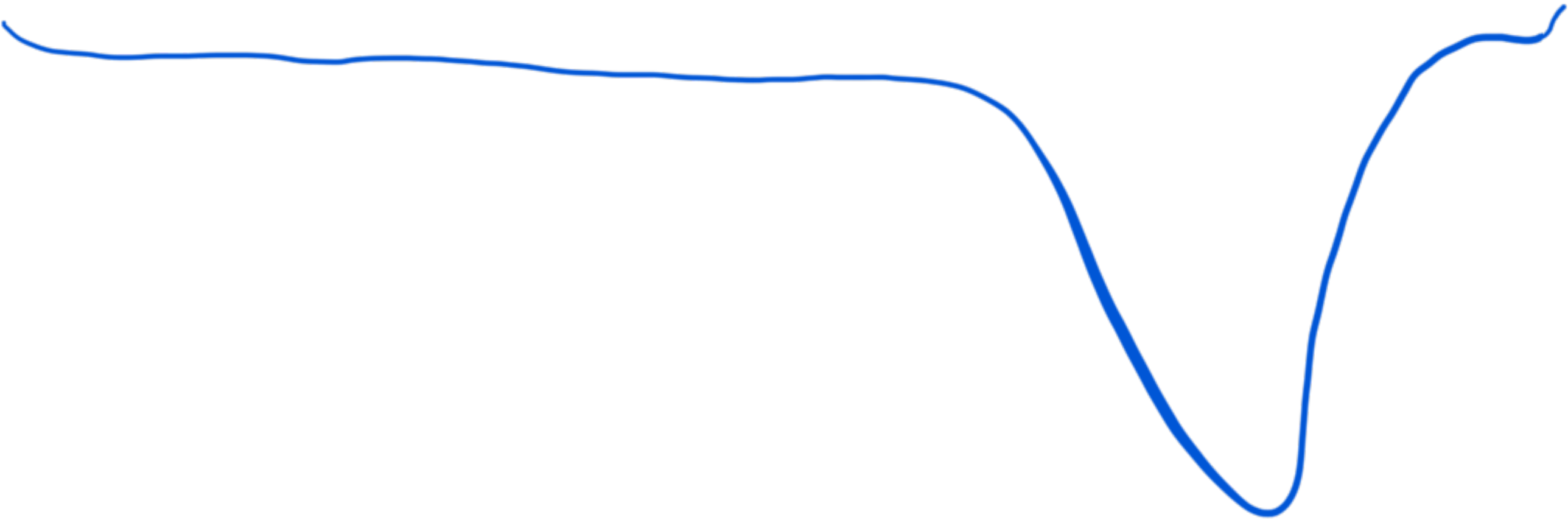
$$f(x) = x^2 + 1 \rightarrow \min_c f(x) = 1, \arg\min_c f(x) = 0$$

Streuungsmaße

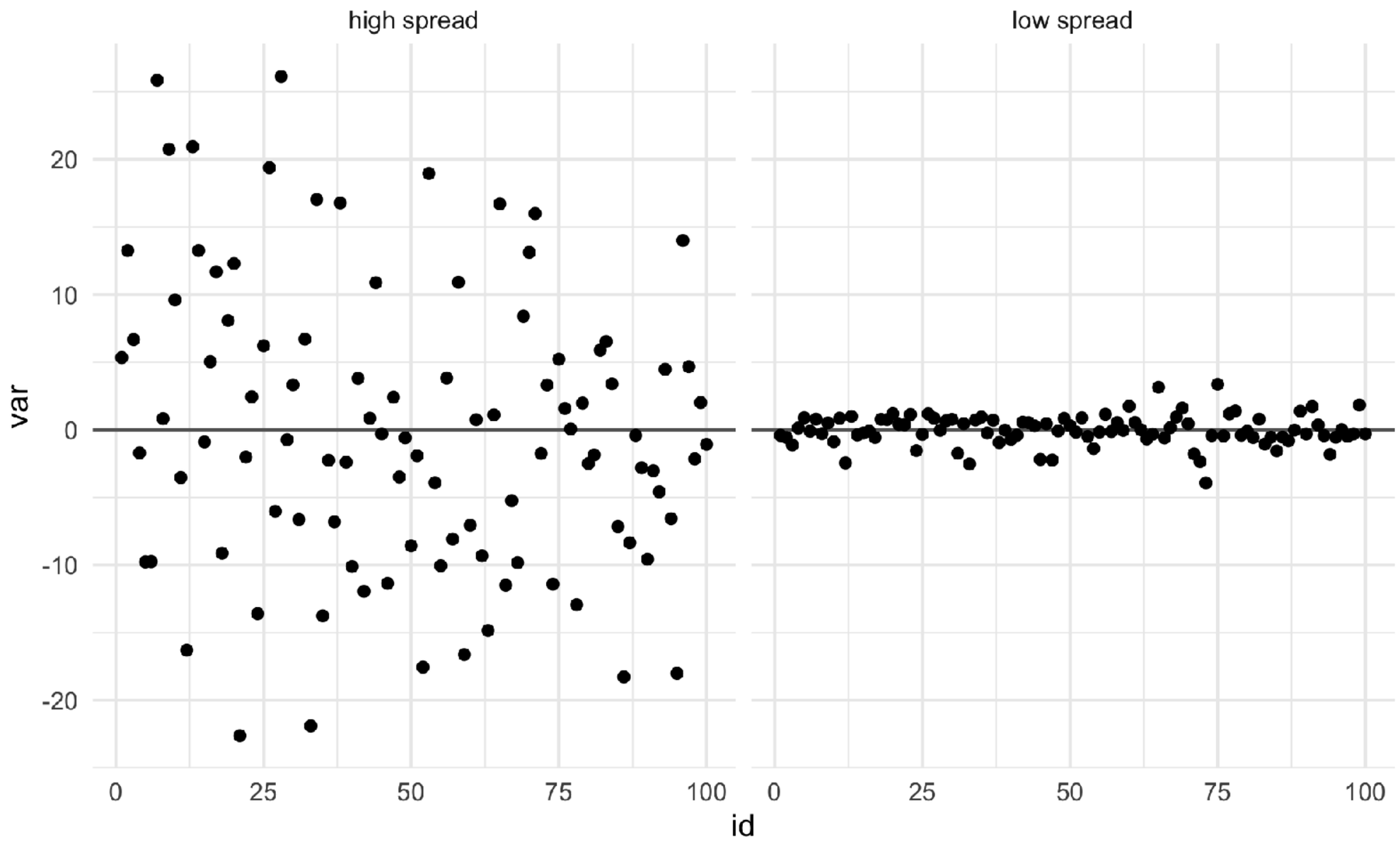
Streuung kann eine wichtige Information sein

„Ist der Fluss tief?“

„Im Schnitt ist er nur einen Meter tief.“

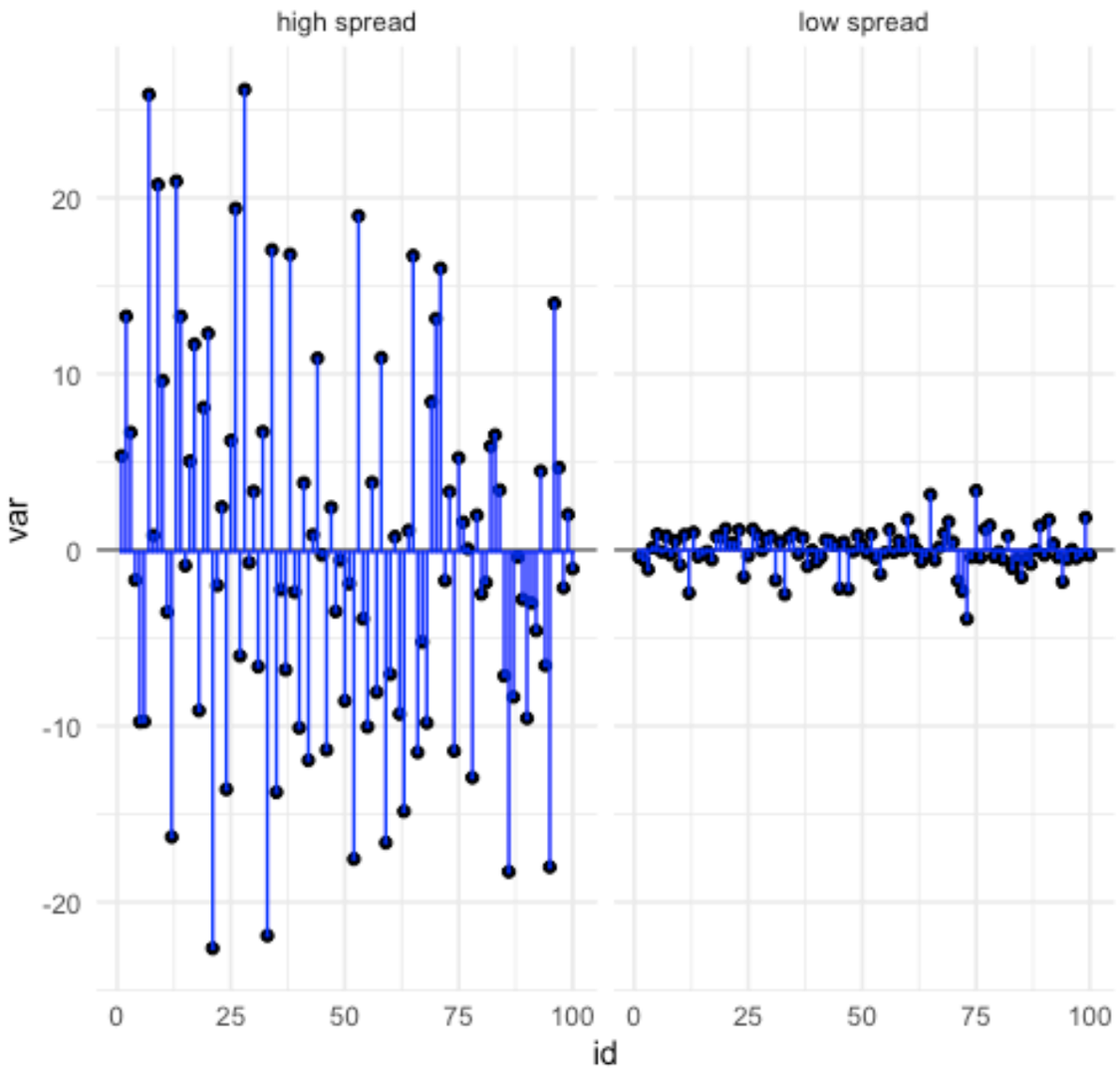


Viel Streuung vs. wenig Streuung



[Quelle]

Viel Streuung vs. wenig Streuung

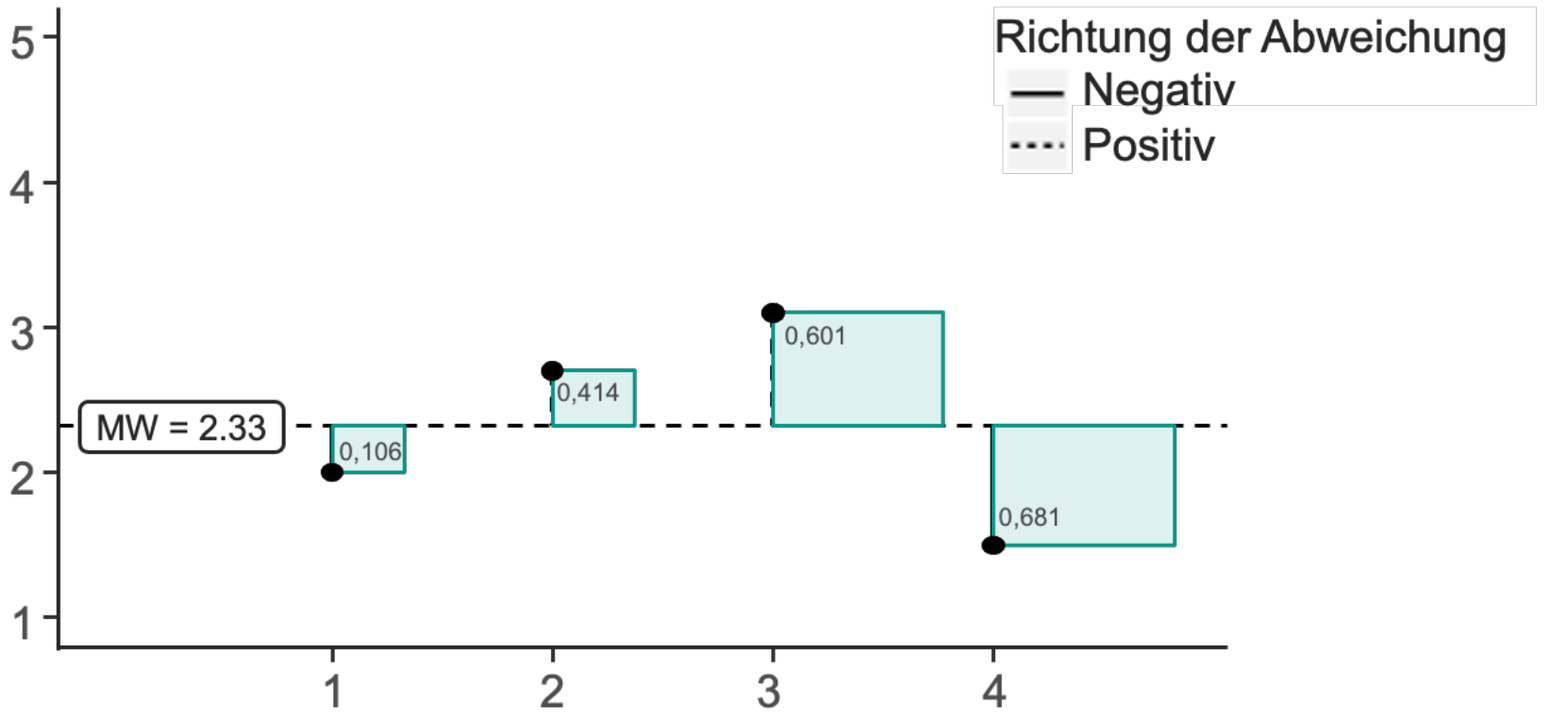


- Die „Balkenlänge“ (blaue vertikale Balken) *d* kann als Maß der Streuung verstanden werden.
- Je kürzer die blauen vertikalen Balken, desto geringer die Streuung.
- Je geringer die Streuung, desto ähnlicher sind sich die Messwerte.
- Genauer gesagt, desto näher sind die Messwerte an ihrem Mittelwert.
- Diesen Kennwert nennt man den mittleren Absolutabstand (MAA, mean absolute deviation, MAD, mad).
- Anschaulich gesprochen zeigt der MAA die mittlere Balkenlänge.
- Als Bezugswert für den MAD wird entweder Mittelwert oder Median gewählt.

$$\text{mad} = \frac{1}{n} \sum d_i$$

Quelle

Varianz als quadrierte Abweichungsbalken



Varianz und Standardabweichung

- Die Varianz (σ^2, s^2, V) ist ein Maß der Streuung.
- Damit gibt sie die Unterschiedlichkeit der Messwerte an.
- Die Varianz einer Stichprobe berechnet sich als der Mittelwert der quadrierten Abstände zum Mittelwert (d).
- Zieht man aus der Varianz die Wurzel, so erhält man die Standardabweichung (σ, s, SD, sd).
- Somit besitzt die Standardabweichung in etwa (!) die gleiche Größenordnung wie die Messwerte der Beobachtungsreihe. Die sd bleibt etwas größer als der MAD! (Achtung: i.d.R. gilt: sd ≠ MAD)

$$\sigma^2 = V = \frac{1}{n} \sum d_i^2$$

$$\sigma = \sqrt{V}$$

Veranschaulichung der Varianz

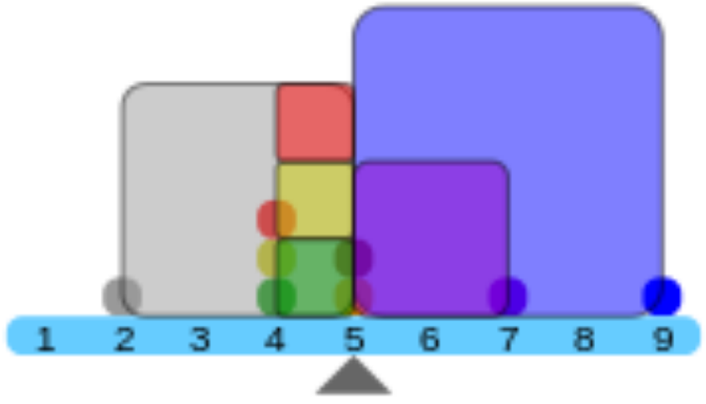
1.

Sieben Objekte liegen geordnet entsprechend ihrem Wert.

2.

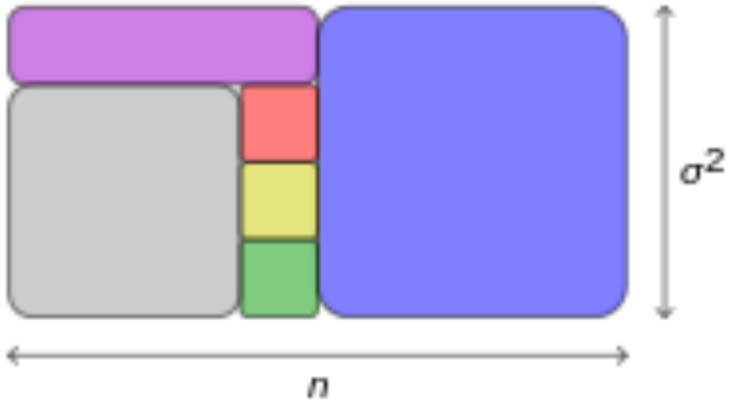
Der Schwerpunkt der Messwertreihe ist das arithmetische Mittel.

3.



Wir bilden ein Quadrat für jedes Objekt; die Kantenlänge jedes Quadrats ist gleich dem Abstand des Wertes des Objekts zum Schwerpunkt .

4.



Legt man die Quadrate so zu einem Rechteck zusammen, dass die eine Seitenlänge der Anzahl der Objekten (n) entspricht, so entspricht die andere Seitenlänge der Varianz (σ^2).

https://en.wikipedia.org/wiki/Standard_deviation

MAD gegenüber der Varianz bevorzugen?

Der „gesunde Menschenverstand“ würde spontan den mittleren Absolutabstand (MAA oder MAD) der Varianz (oder der Standardabweichung, SD) vorziehen. Das ist vernünftig, denn die MAA ist anschaulicher und damit nützlicher als die Varianz und die SD.

Warum sollte man überhaupt ein unanschauliches Maß wie die Varianz verwenden? Wenn es nur um deskriptive Statistik geht, braucht man die Varianz nicht unbedingt.

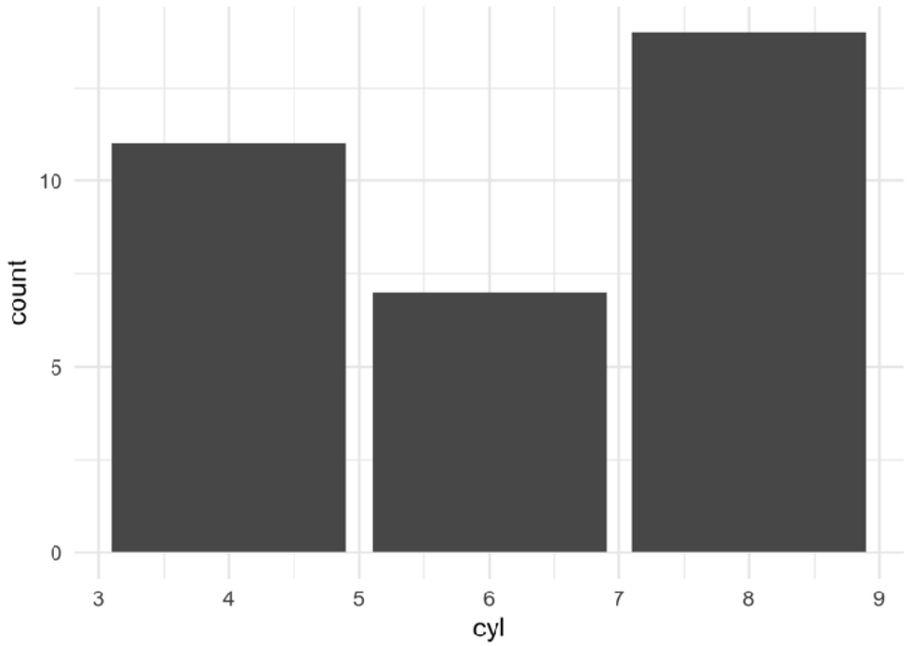
Gründe, warum Sie die Varianz (SD) kennen und nutzen sollten, sind:

1. Die Varianz wird häufig verwendet bzw. in Forschungsarbeiten berichtet, also müssen Sie die Varianz kennen
2. Die Varianz ist sensitiv zu Extremwerten, das kann eine nützliche Eigenschaft sein (kann aber auch unerwünscht sein)
3. Die Varianz (bzw. die SD) ist sehr nützlich zur Beschreibung der Normalverteilung
4. Die Varianz hat die erstaunliche Eigenschaft, dass die additiv ist. Für zwei unkorrelierte Variablen A und B gilt dann: Var(A+B) = Var(A) + Var(B). Andere Streuungsmaße haben diese praktische Eigenschaft nicht.

Verteilungen

Häufigkeitsverteilungen

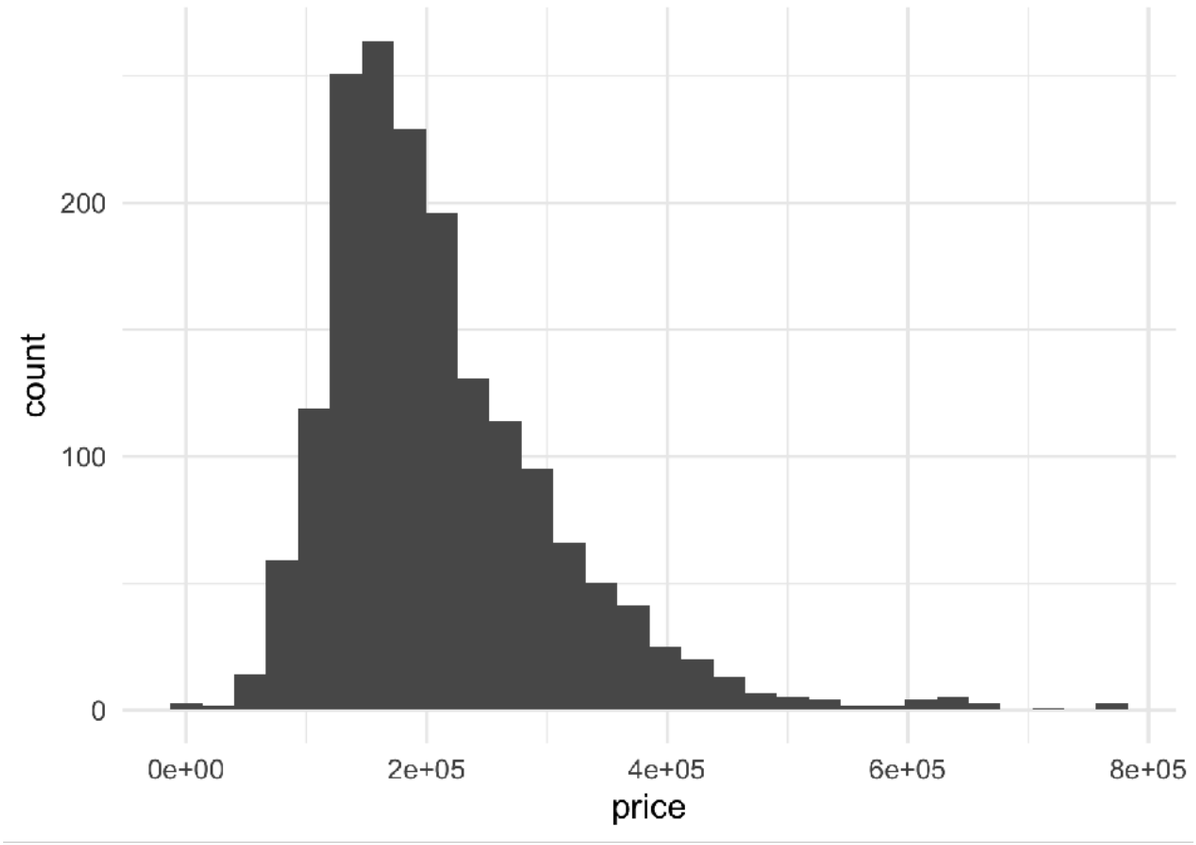
cyl	n
4	11
6	7
8	14



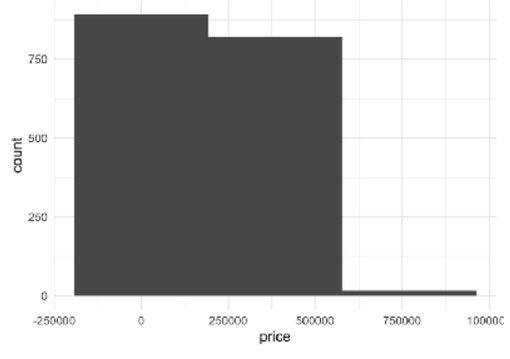
- Eine Häufigkeitsverteilung gibt an, wie häufig jeder der Ausprägungen (Stufen) einer Variablen *X* in einer Stichprobe ist.
- Beispiel:
  - Eine Stichprobe umfasse n=32 Autos.
  - Die Variable *cyl* (Zylinderzahl) hat 3 Ausprägungen: 4,6,8.
  - Jede dieser Ausprägungen findet sich mit einer bestimmten Häufigkeit in der Stichprobe (11, 7, 14).
- Ein Balkendiagramm (Säulendiagramm) eignet sich zur Darstellung einer Häufigkeitsverteilung, sofern die Variable nicht zu viele Ausprägungen hat.
- Der Modus (Modalwert) gibt die häufigste Ausprägung an, im Balkendiagramm entspricht der Modus der höchsten Säule.

Histogramm für Häufigkeitsverteilungen mit vielen Stufen

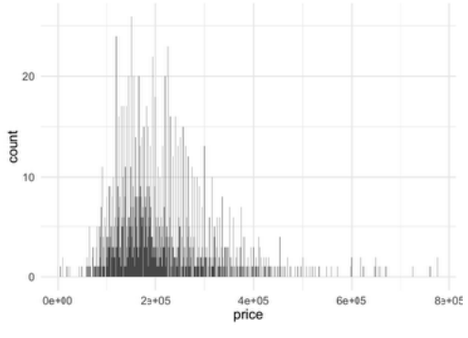
Häufigkeitsverteilung der Hauspreise im Saratoga County, New York, USA, 2006



- Histogramme eignen sich, um die Häufigkeitsverteilung einer Variablen mit vielen Ausprägungen darzustellen.
- Häufig werden Histogramme für stetige, metrische Variablen verwendet.
- Dabei stellt ein Balken einen Bereich von Ausprägungen (ein Intervall) dar.
- Bei gleich großer Intervallbreite ist die Höhe des Balkens proportional zur Anzahl der Werte in diesem Intervall.
- Für die Anzahl der *k* Balken gibt es keine feste Regel, aber die Balkenzahl sollte dem Erkenntnisziel zuträglich sein.

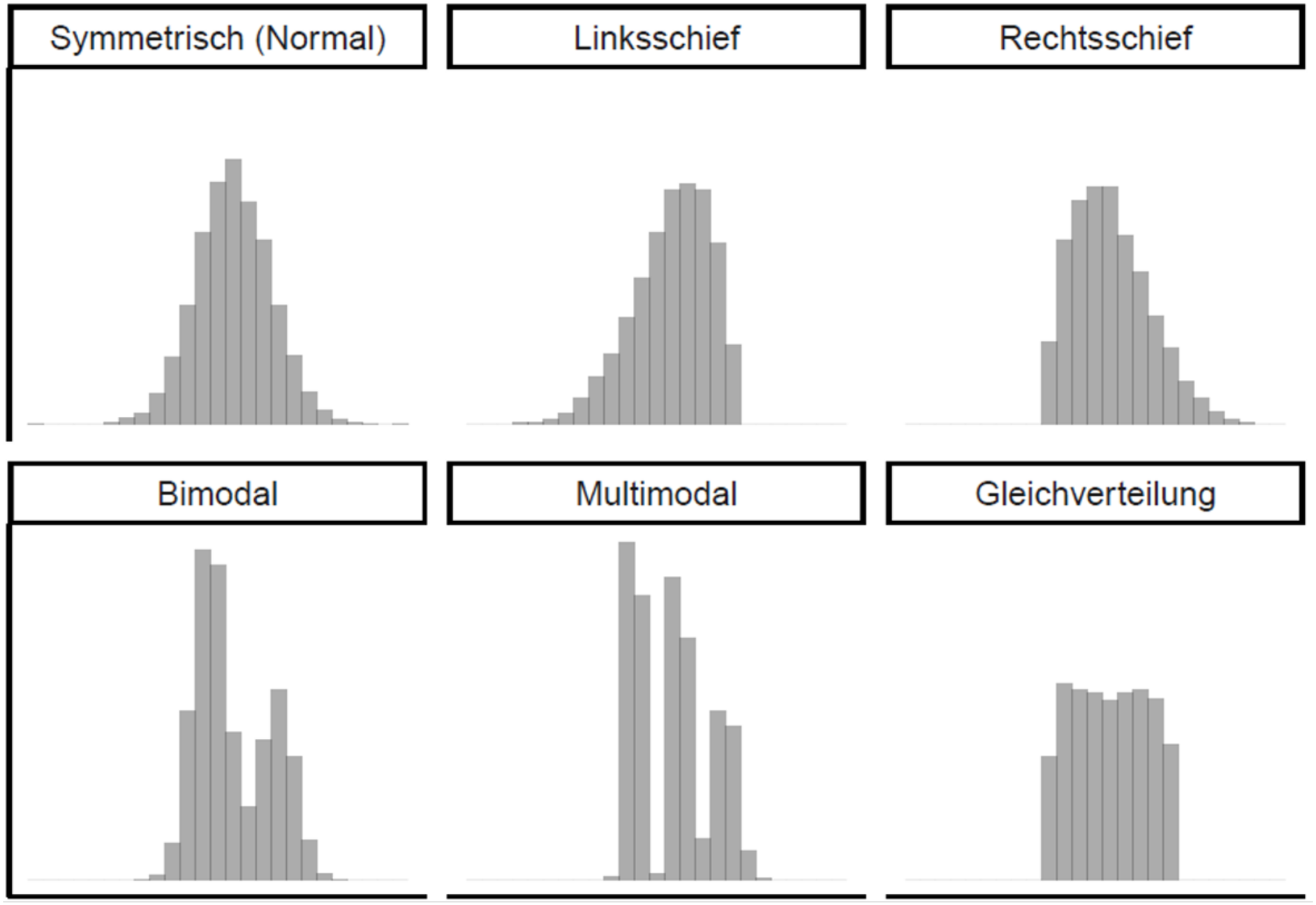


zu wenig Balken (k=3)



zu viele Balken (k=1000)

Verteilungsformen

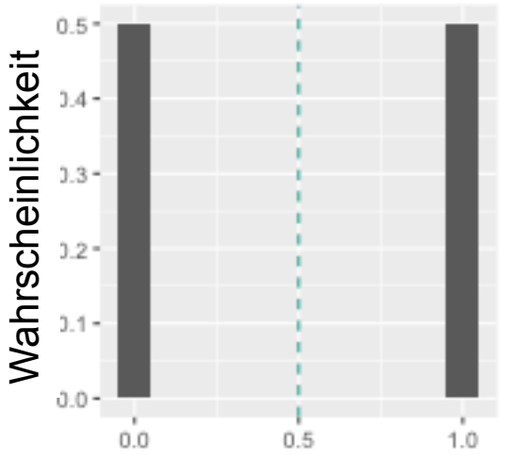


Wahrscheinlichkeits- vs. Häufigkeitsverteilung

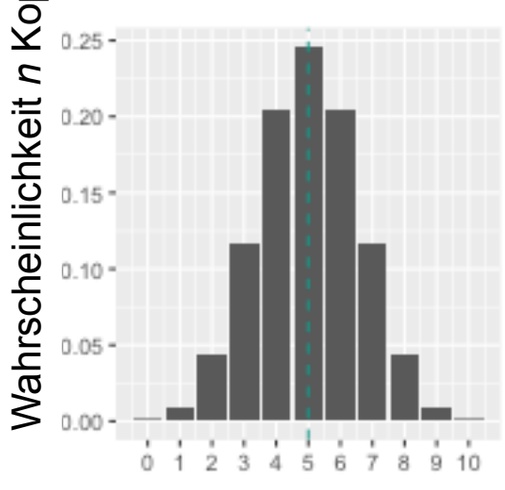
- Eine **Wahrscheinlichkeitsverteilung** ordnet jedem Wert einer Variablen eine Wahrscheinlichkeit zu:

Beispiele:

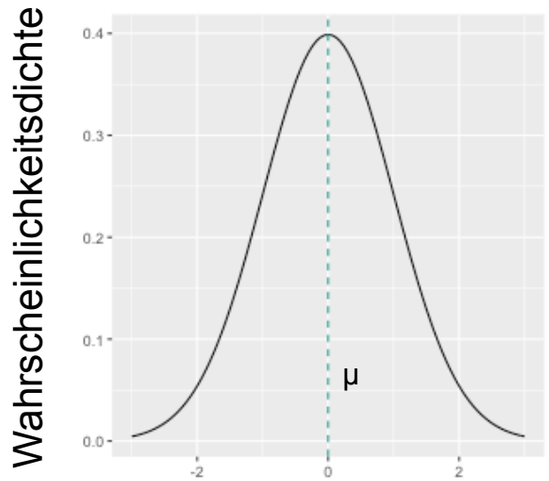
Münzwurf;
p(Kopf) = .5; p(Zahl) = .5



10 Münzwürfe



IQ in der Bevölkerung

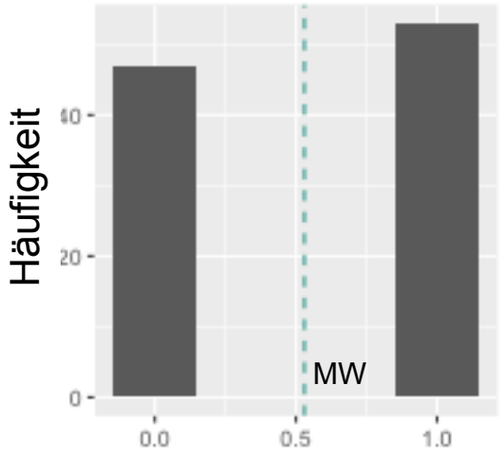


- Den "Mittelwert" einer Wahrscheinlichkeitsverteilung nennt man Erwartungswert (μ).

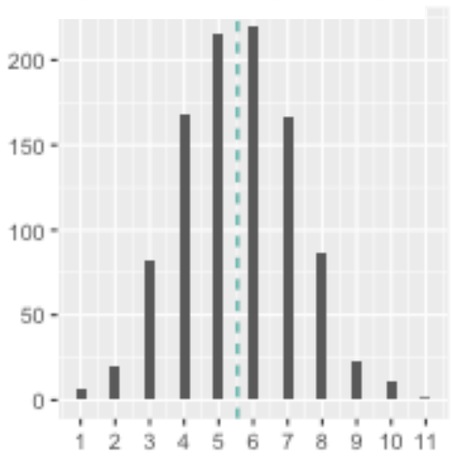
- Eine **Häufigkeitsverteilung** ordnet jedem Wert einer empirischen Variablen eine Häufigkeit zu:

Beispiele:

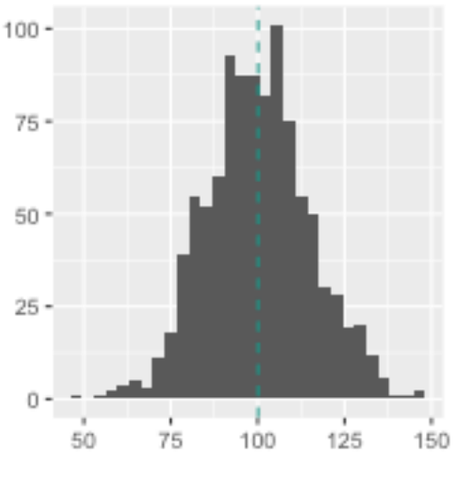
10 Münzwürfe*



10 Münzwürfe*,
1000 Mal wiederholt



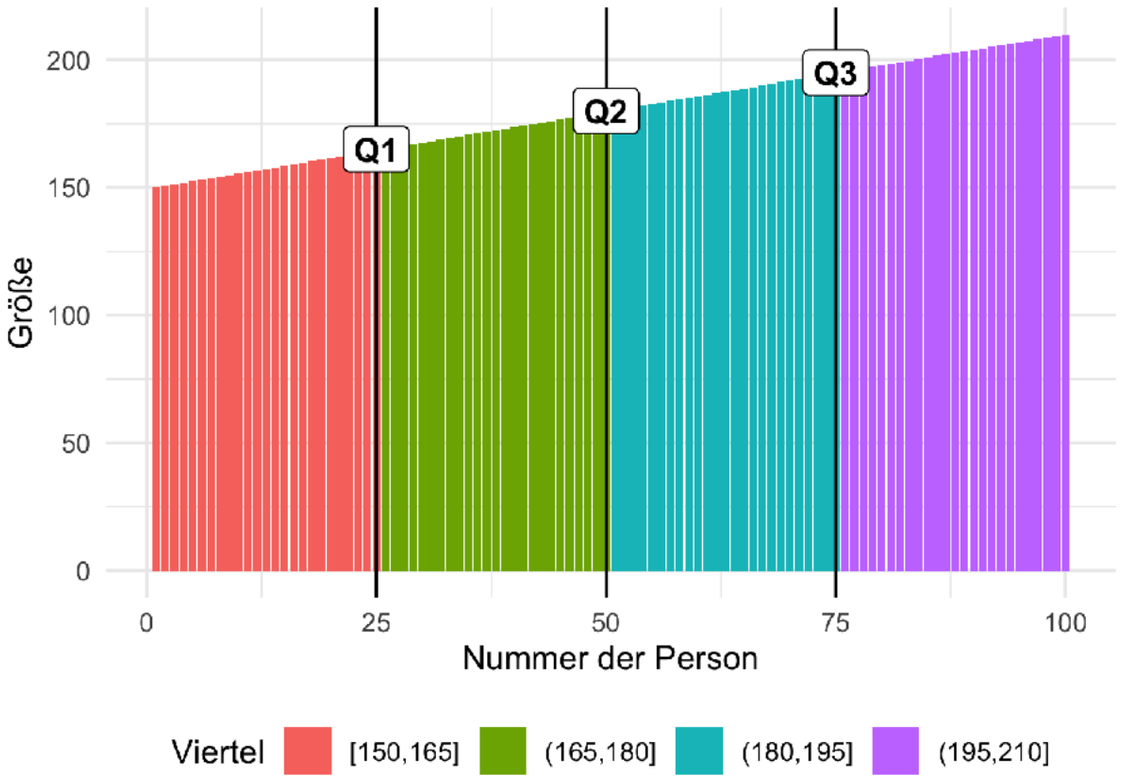
IQ von 1000 Personen



* Kopf ≙ 1; Zahl ≙ 0

Hundert Studentis der Größe nach sortiert

Etwa hundert Studentis stellen sich der Größe nach sortiert auf.



Quelle: Sauer, 2019, S. 106

Ein kauziger Statistik-Prof läuft die Reihe ab, er ruft:

- „1. Quartil!“ bei Person 25, (25% kleiner, Größe 1.65m)
- „2. Quartil!“ bei Person 50, (50% kleiner, Größe 1.80m)
- „3. Quartil!“ bei Person 75, (75% kleiner, Größe 1.95m)

- Das *1. Quartil* kennzeichnet denjenigen Wert der Körpergröße der Studentis, für den gilt, dass ein Viertel der Studentis kleiner (und drei Viertel größer sind).

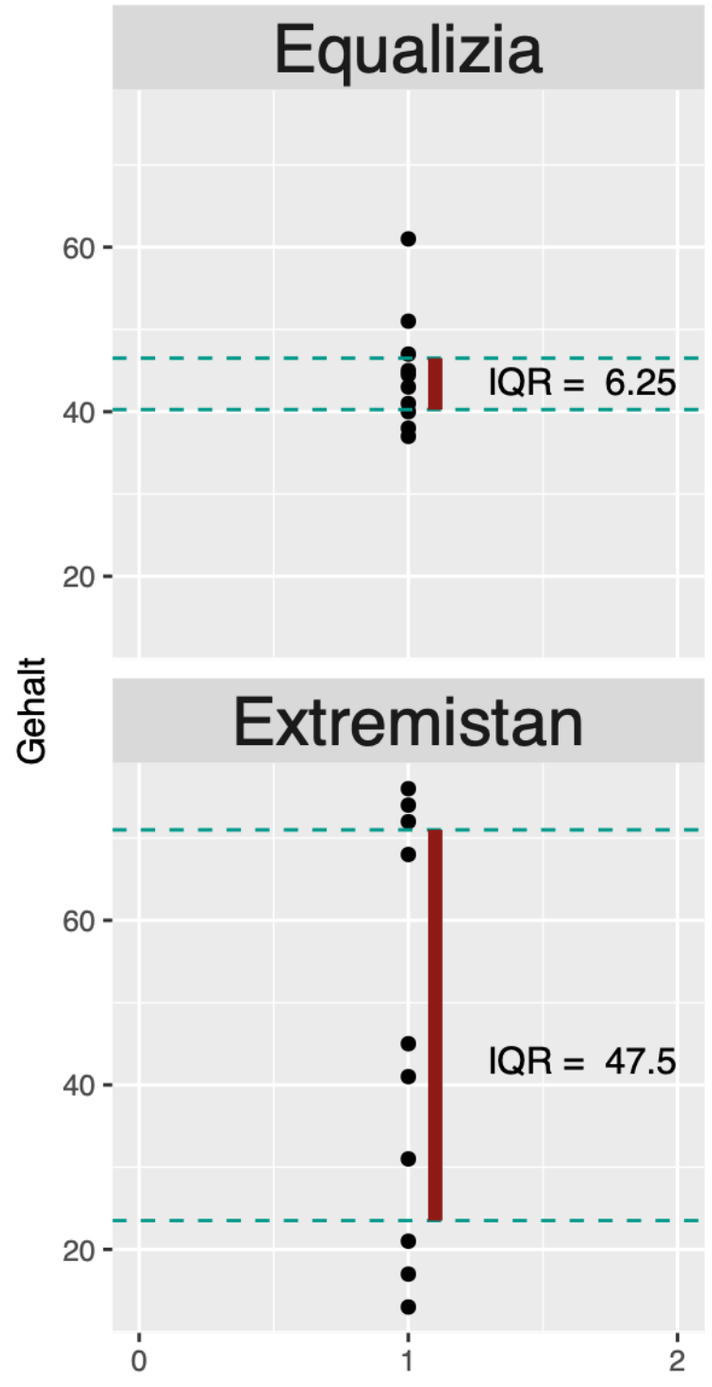
- Ein Quartil …
  - ist eine *bestimmte Art von Quantil*
  - *verallgemeinert den Median*, da der Median dem 2. Quartil entspricht
  - entspricht dem 25. Perzentil

Quantile

- Quantile sind Grenzwerte, die eine Verteilung in Bereiche gleich großer Anteile (oder Wahrscheinlichkeit) schneiden.
- Gängige *Quantile* sind
  - *Quartile* (Viertel)
  - *Quintile* (Fünftel)
  - *Dezile* (Zehntel)
  - *Perzentile* (Hundertstel)
- Ein Quantil ist also ein *Oberbegriff* für die Aufteilung einer Verteilung in eine bestimmte Anzahl an Bereichen gleicher Größe.
- Allgemein ist das p-Quantil definiert, als der Wert, für den gilt, dass er von p Prozent der Beobachtungen (oder, synonym, mit p Prozent Wahrscheinlichkeit) nicht überschritten wird.
- Die Quantilsfunktion q(p) gibt für eine gegebene Wahrscheinlichkeit p aus, welcher Wert q mit dieser Wahrscheinlichkeit nicht überschritten wird.

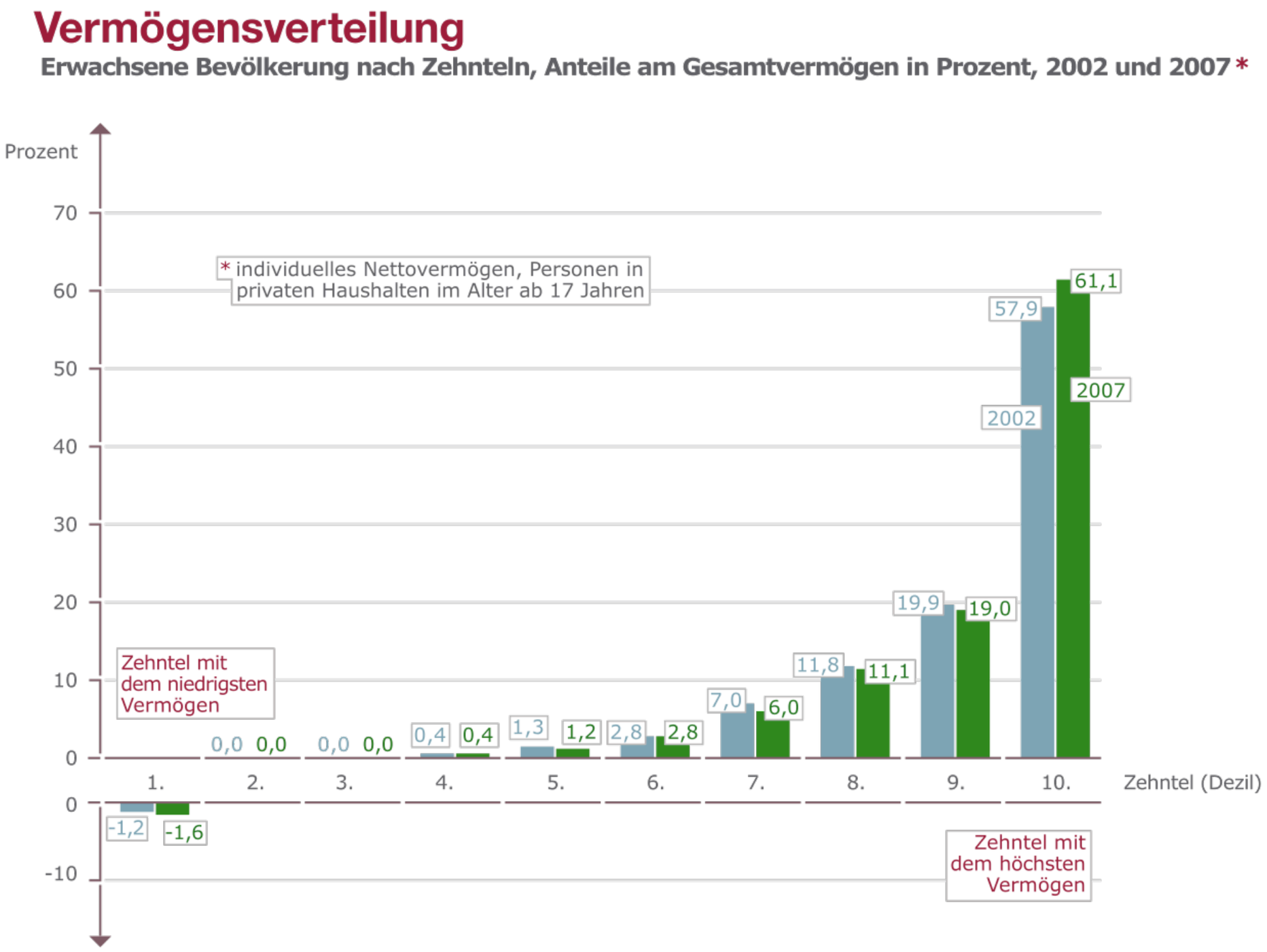
Der Interquartilsabstand als Maß für die Streuung

- Betrachen wir zwei Länder, Equalizia und Extremistan. Im Bild sehen wir je 10 Menschen für jedes der zwei Länder.
- Das mittlere Einkommen scheint ähnlich zu sein.
- Die Streuung ist aber sehr unterschiedlich: In Equalizia verdienen die Menschen alle etwas gleich viel (kleine Streuung); in Extremistan geht die Schere zwischen arm und reich stark auf (große Streuung).
- Die grün gestrichelten Linien im Bild zeigen jeweils das untere und das obere Viertel (1. bzw. 3. Quartil).
- In Equalizia verdient das untere Viertel also höchstens ca. 40 Geldeinheiten; in Extremistan nur ca. 23 GE. Dafür ist das obere Viertel in Extremistan sehr reich; in Equalizia ist das obere Viertel hingen vergleichsweise nah am unteren Viertel.
- Diese Differenz Q3-Q1 bezeichnet man als Interquartilsabstand (engl. inter quartile range; IQR); der IQR ist ein Maß für die Streuung.
- Beachten Sie, dass die Extremwerte (die reichsten und ärmsten Menschen) keinen Einfluss auf die Berechnung des IRQ haben! Daher bezeichnet man den IQR als "robust".



Wo möchten Sie lieber leben?

Dezile der deutschen Vermögensverteilung



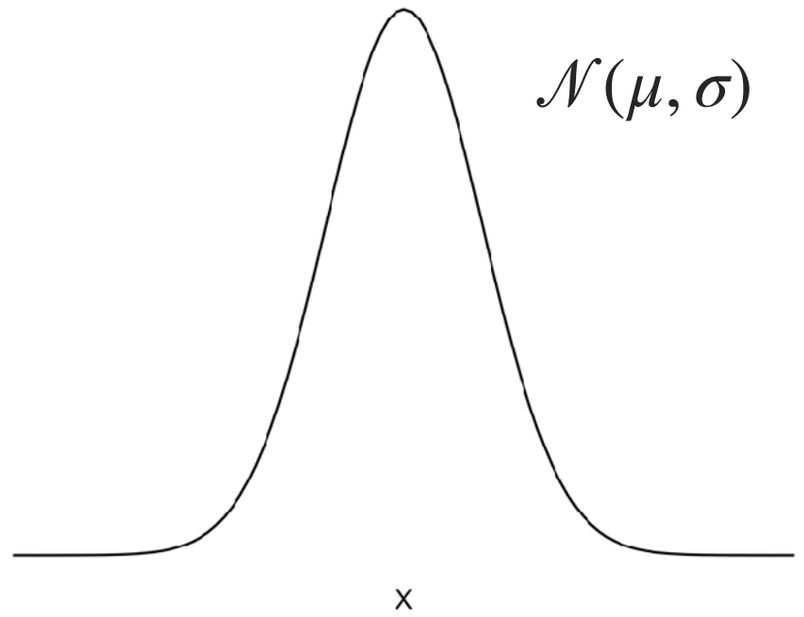
Quelle: SOEP zitiert nach Wikipedia

Quantilfunktion vs. Verteilungsfunktion

Normalverteilung

- Die Normalverteilung ist eine Verteilung mit folgenden Eigenschaften:
  - Die Daten verteilen sich *symmetrisch* um das Zentrum aller Werte.
  - Die Form erinnert an eine *Glocke*.
  - Mittelwert = Median = Modus
  - Normalverteilungen sind durch *zwei Größen* (Parameter) komplett determiniert: Mittelwert (μ) und Standardabweichung (σ, sd).
  - Es gibt unendlich viele verschiedene Normalverteilungen, die sich (nur) im Mittelwert und/oder Standardabweichung unterscheiden.
  - Alle Normalverteilungen sind *sich ähnlich* in dem Sinne, dass ihre Form ähnlich ist: Das Verhältnis der Breite von „Mittelbereich“ zu „Randbereichen“ ist immer gleich.
- Viele Größen sind normalverteilt: z.B. IQ, Körpergröße und -gewicht von Erwachsenen, Messfehler, Gewichts eines maschinenproduzierten Gegenstands, …
- Andere Größen sind nicht normalverteilt: Einkommen, Vermögen, Erfolg, Zitationen, Bekanntheit, …
- Ob eine Größe normalverteilt ist, kann (und muss) empirisch übergeprüft werden.
- Eine Normalverteilung mit μ = 1 und σ = 1, nennt man *Standardnormalverteilung*.

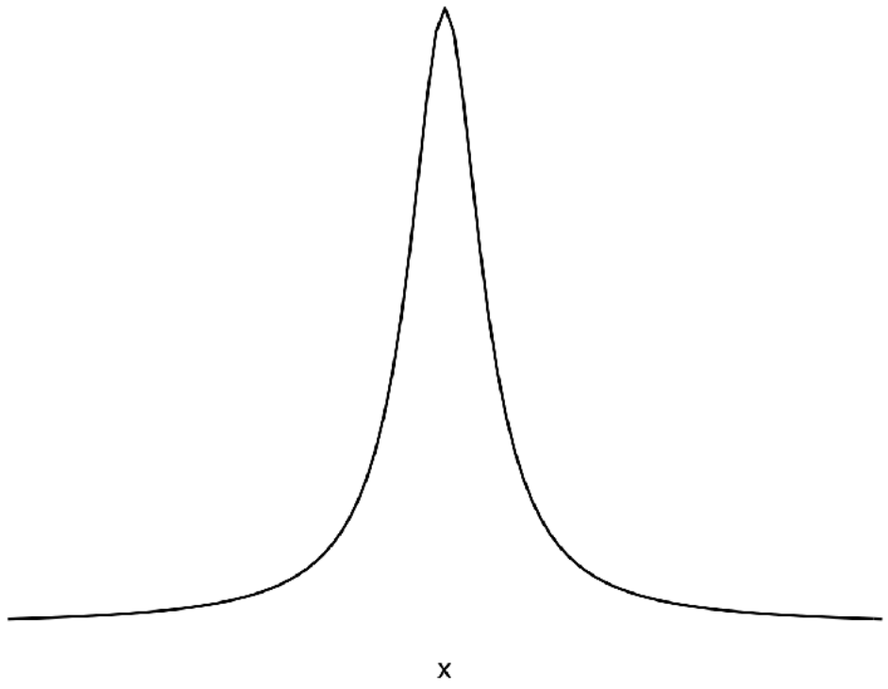
Normalverteilungen, Beispiele



Wölbung (Kurtosis) im Vergleich zur Normalverteilung

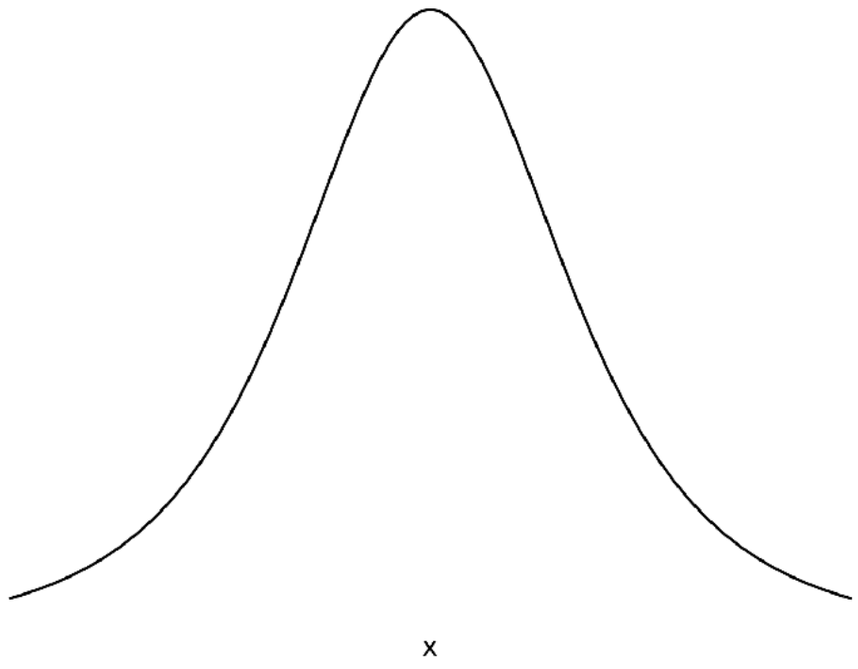
Steilgipflige Verteilungen

- leptokurtisch
- Die Werte verteilen sich eng um den Mittelwert
- Die Verteilung erscheint „spitz“ und „schmalschulterig“



Flachgipflige Verteilungen

- platykurtisch
- Die Werte verteilen sich weit um den Mittelwert in die „Ränder“ hinaus
- Die Verteilung erscheint „platt“ und „breitschulterig“

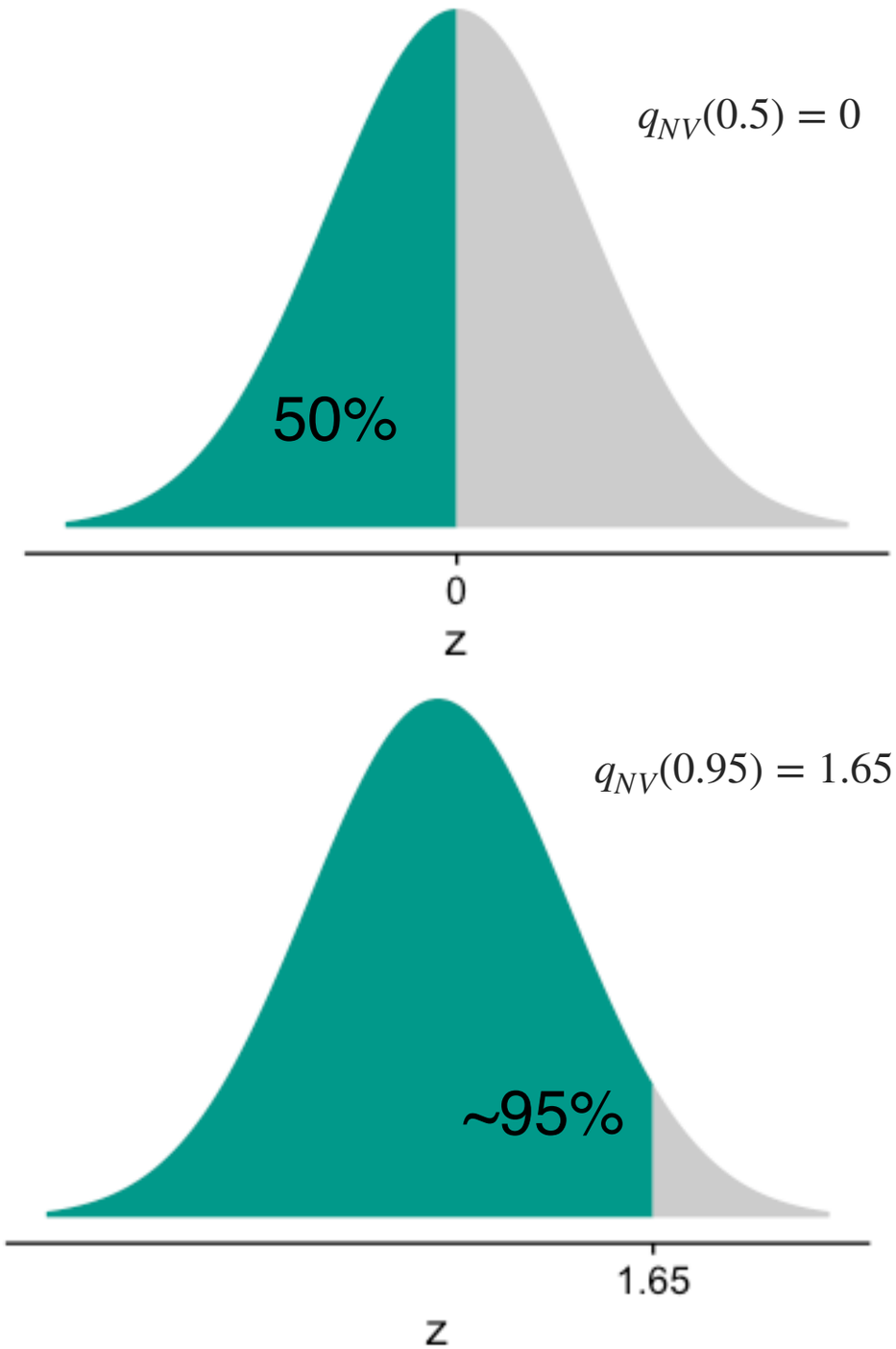


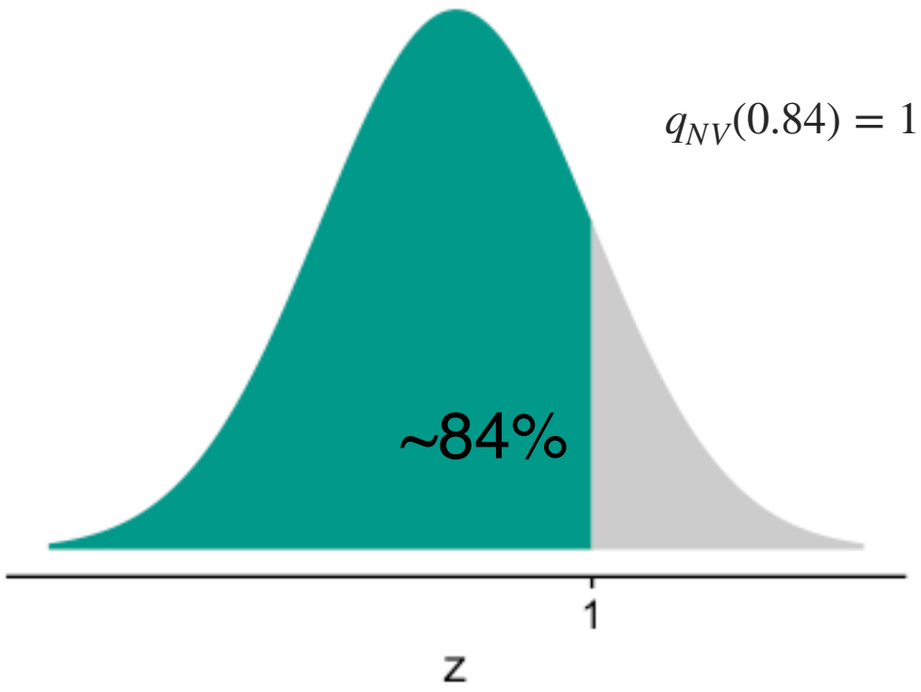
Standardisierung mit der z-Transformation

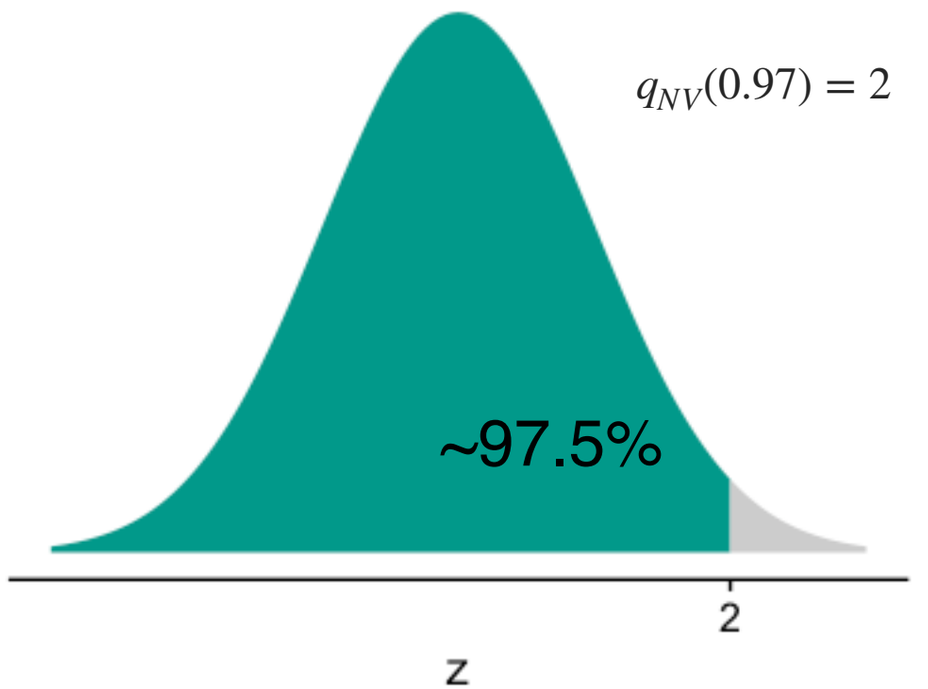
- Kennt man Mittelwert μ und Standardabweichung σ einer normalverteilten Variablen X, so kann man jeden Punkt auf dieser Verteilung (Kurve) bestimmen; damit kann man dann auch die Flächenanteile bestimmen.
- Alle Normalverteilungen sind verwandt in dem Sinne, dass die Flächenanteile unter der Kurve immer dem gleichen Abstand zum Mittelwert (in SD-Einheiten) entsprechen.
- Daher reicht es, wenn jemand einmal für eine einzige Normalverteilung alle Flächenabschnitte bestimmt (wem's Spaß macht). Wir schauen dann diese Werte nach.
- Die einfachste Normalverteilung ist die mit μ = 0 und σ = 1; man nennt sie daher Standardnormalverteilung.
- Anhand der Standardnormalverteilung können Sie die Wahrscheinlichkeiten jedes Werts jeder Normalverteilung einfach bestimmen; man z-transformiert dazu einen Wert xi der Person i aus einer beliebigen Normalverteilung in einen Wert z_{xi} aus der Standardnormalverteilung.
- Z-Transformiert man eine Verteilung, so resultiert μ = 0 und σ = 1.
  - Zieht man von jedem Wert den MW ab, so ist der MW um MW kleiner und damit 0.
  - Teilt man jeden Wert durch SD, so ist die SD um den Faktor SD geringer und damit 1.

$$z_i = \frac{x_i - \bar{x}}{sd_x}$$

Einige Quantile (q) der Normalverteilung (NV)







Eigenschaften von z-verteilten Werte

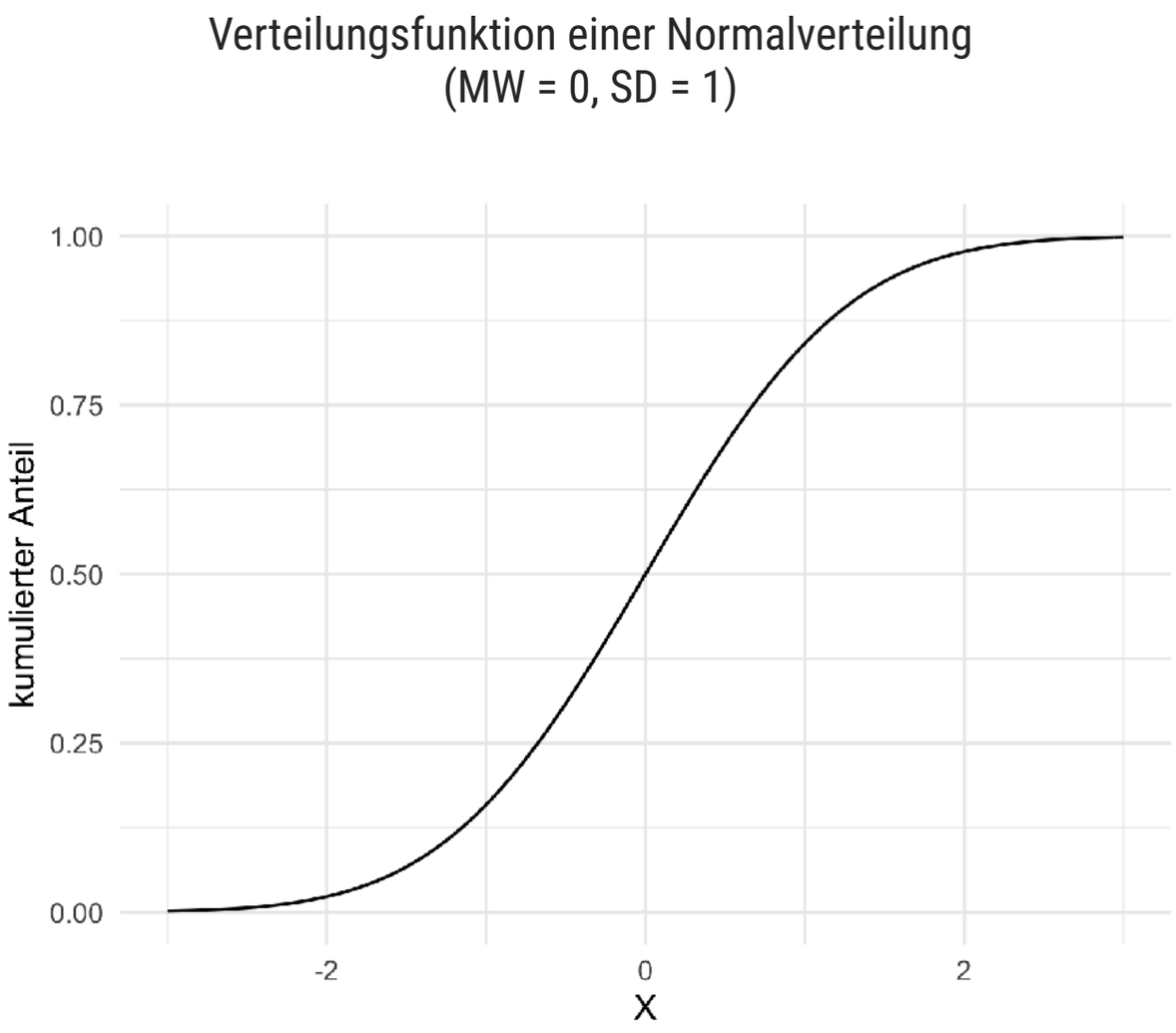
- z-Transformiert man jedes normalverteilte Xi, so ist die resultierende Verteilung z-verteilt.
- Die Verteilung hat den Mittelwert **$\mu_z = \bar{z} = 0$** und die Streuung **$sd_z = Var_z = 1$.**
- Wenn die Varianz 1 ist, ist auch die Standardabweichung notwendig 1.

$$Var_z = \frac{1}{n}\sum (z_i - \bar{z})^2 \qquad \text{Formel der Varianz}$$

da aber $\bar{z} = 0,$ ergibt sich:

$$\begin{aligned} Var_z &= \frac{1}{n}\sum {z_i}^2 \\ &= \frac{1}{n}\sum \frac{(x_i - \bar{x})^2}{sd_X^2} && \text{Formel für z-Werte einsetzen und quadrieren} \\ &= \frac{\frac{1}{n}\sum (x_i - \bar{x})^2}{Var_X} && \text{Im Zähler steht die Formel für die Varianz} \\ &= \frac{Var_X}{Var_X} = 1 \end{aligned}$$

Die Verteilungsfunktion



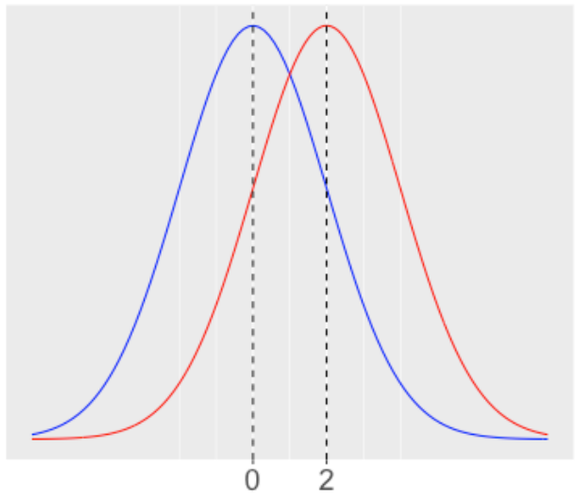
- Die empirische Verteilungsfunktion $F_e(x)$ gibt an, welcher Anteil der Beobachtungen kleiner oder gleich x sind.
- Sie sagt aus, wie wahrscheinlich es ist, einen Wert kleiner oder gleich x zu beobachten, liefert also eine Wahrscheinlichkeit als Funktionswert.
- Die theoretische Verteilungsfunktion F(x) gibt für eine bestimmte Verteilung (wie eine bestimmte Normalverteilung NV) an, wie wahrscheinlich es ist, einen Wert kleiner oder gleich x zu beobachten gegeben der Modellparameter.
- Für die Verteilungfunktion der Normalverteilung wird auch der Buchstabe Φ (großes Phi) verwendet.

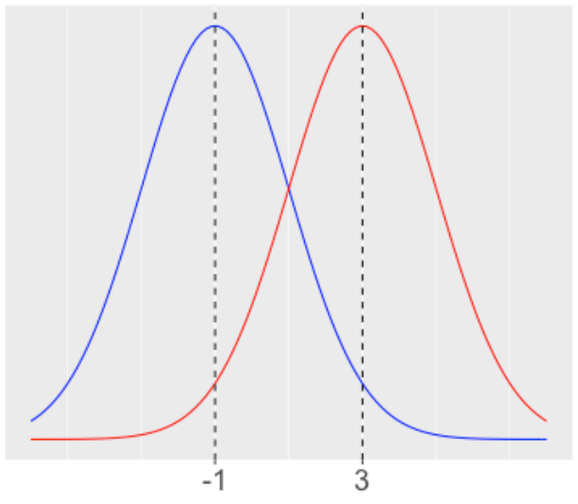
$$F_e(x) = \frac{\text{Anzahl Beobachtungen} \leq x}{n}$$

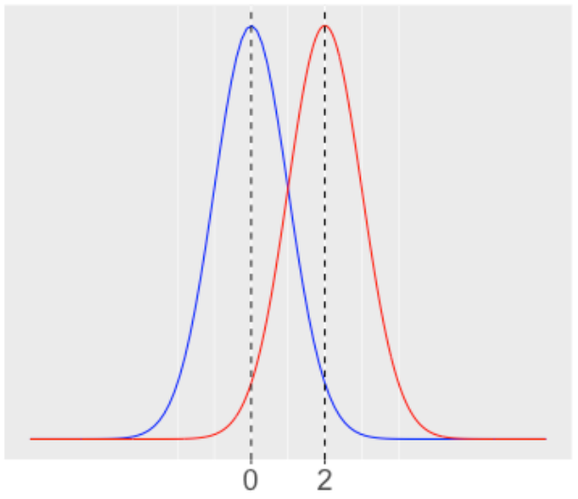
$$F_{NV}(x) = \Phi(\mu, \sigma, x)$$

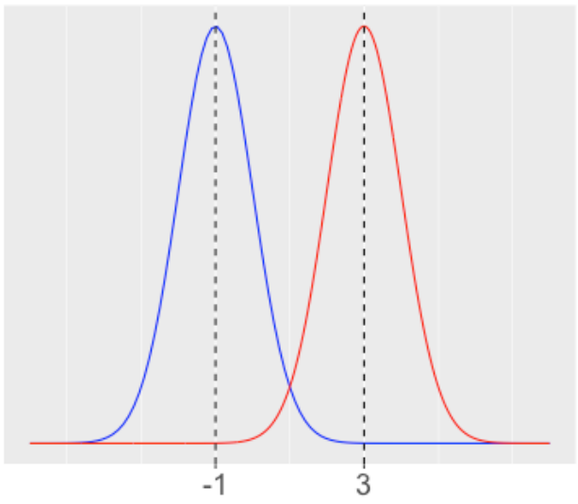
Cohens d

Verteilung der Einpark-Zeiten: Männer (blau) vs. Frauen (rot)









sehr viel „Überlappung“: *schwacher* Effekt

sehr wenig „Überlappung“: *starker* Effekt

- Anhand der „Überlappung“ der Kurven lässt sich die Stärke des „Einpark-Effekts“ (Unterschied zwischen den Gruppen) veranschaulichen.
- Die Größe des Unterschieds (der Überlappung) hängt nicht nur von der Differenz der Mittelwerte ($\bar{X}1$-$\bar{X}2$) ab, sondern auch von der Streuung der Verteilungen.
- Das Verhältnis der Mittelwertsunterschiede zu mittlerer Streuung nennt man Cohens d.
- Cohens d ist ein Maß der Effektstärke, das die Differenz der Mittelwerte zweier Gruppen in Bezug zur Streuung setzt.
- Grobe Faustregel: kleiner / mittlerer / großer Unterschied: d = 0.2 / 0.5 / 0.8.

$$d = \frac{\overline{X_1} - \overline{X_2}}{sd}$$

Abschluss

Hinweise

- Dieses Dokument steht unter der Lizenz CC-BY 3.0.
- Autor: Sebastian Sauer
- Für externe Links kann keine Haftung übernommen werden.
- Dieses Dokument entstand mit reichlicher Unterstützung vieler Kolleginnen und Kollegen aus der FOM. Vielen Dank!
- Dieses Dokument baut in Teilen auf auf dem Skript zu quantitative Methoden des ifes-Instituts der FOM-Hochschule.