



STAT II

$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z)$

Vorwort

Dieser Foliensatz wird durch das Zentrum für Angewandte Ökonomik (kurz ZAÖ) der DHBW Ravensburg bereitgestellt.

Autoren: Prof. Dr. Daniel Blochinger
Illustration: Prof. Dr. Daniel Blochinger
Stand: 10. Juni 2026
Lizenz: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Weitere Lehr- und Lernmaterialien finden Sie auf unserer [Webseite](#).

Fehler gefunden? E-Mail an blochinger@dhbw-ravensburg.de!



Inhaltsverzeichnis

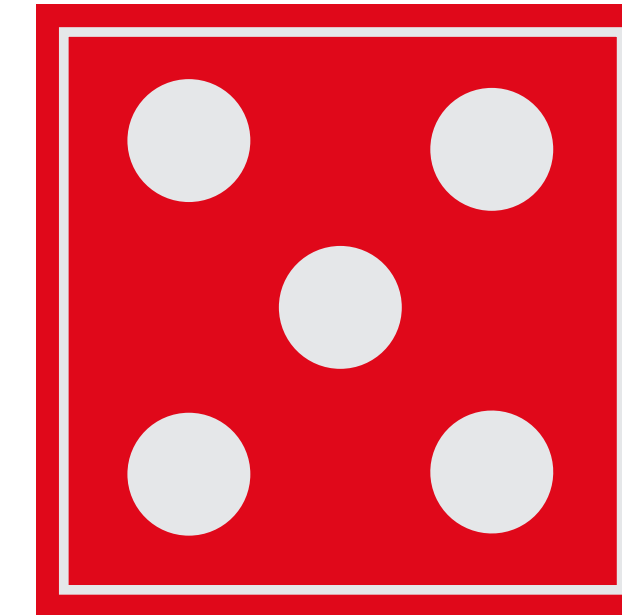
<u>Zufallsvariablen & Verteilungen</u>	4 - 21	<u>Test auf Normalität</u>	130 - 142
<u>Momente von Zufallsvariablen</u>	22 - 44	<u>Die Chi-Quadrat-Testfamilie</u>	143 - 154
<u>Ausgewählte Verteilungen</u>	45 - 69	<u>Lineare Regression & OLS-Schätzung</u>	155 - 182
<u>Zentraler Grenzwertsatz</u>	70 - 76	<u>Herausforderungen bei Regression</u>	183 - 215
<u>Gauss Test (Z-Test)</u>	77 - 92	<u>Maximum Likelihood Schätzung</u>	216 - 226
<u>Induktive Statistik am Beispiel t-Test</u>	93 - 125		
<u>ANOVA als erweiterter t-Test</u>	126 - 129		

Zufallsvariablen

Zufallsvariablen sind Abbildungsvorschriften, die jedem Ereignis der Ereignismenge Ω eine reelle Zahl zuweisen.

$$X: \Omega \rightarrow E \text{ mit } E \subseteq \mathbb{R}$$

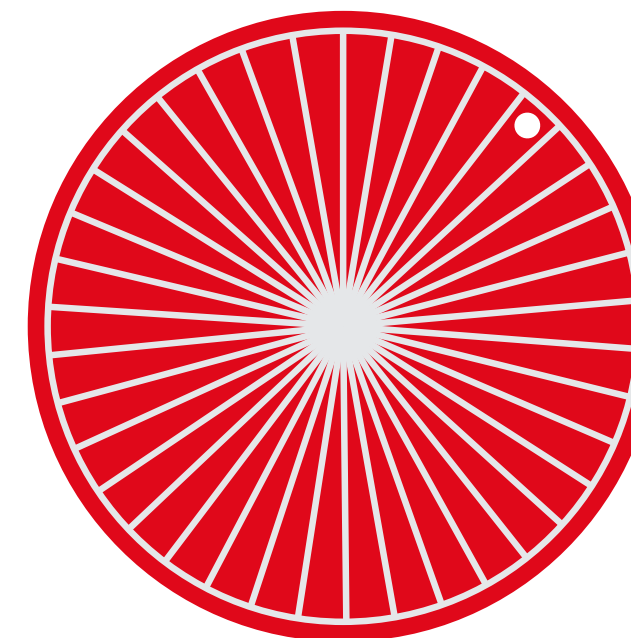
Zufallsvariablen sind mathematische Funktionen, deren Definitionsbereich eine Ereignismenge ist.



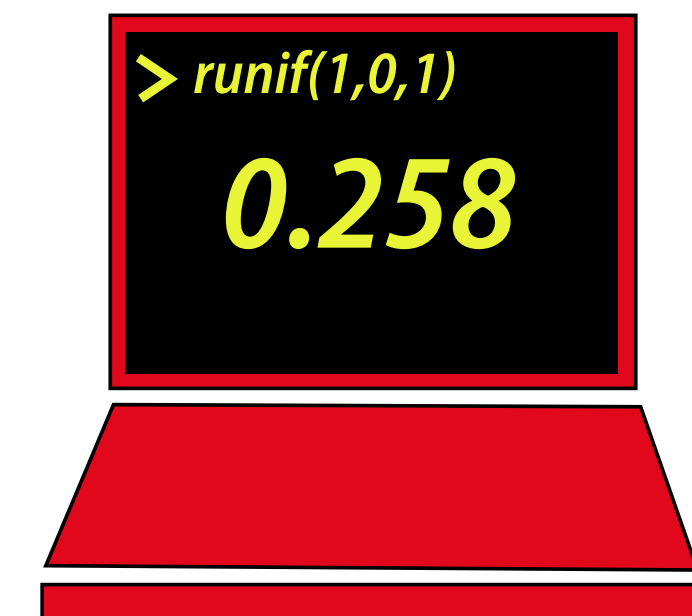
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



$\Omega = \{\text{Kopf, Zahl}\}$
DISKRET



$\Omega = \{00,0,1,2,\dots,36\}$
DISKRET



$\Omega = (0,1)$
KONTINUIERLICH

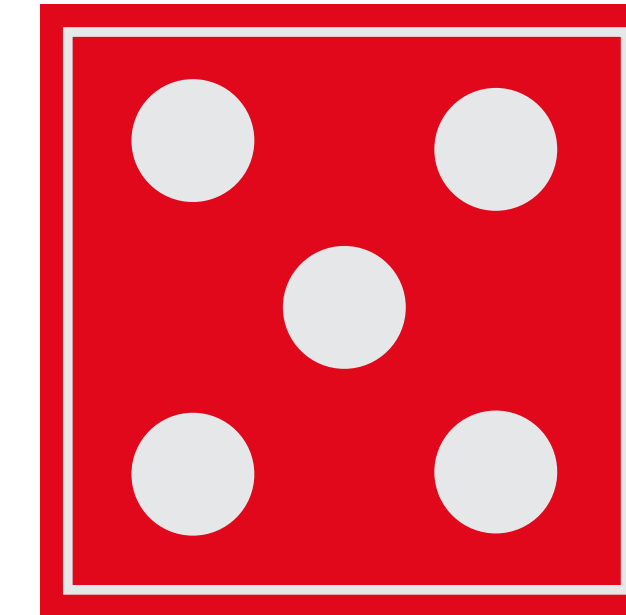
Zufallsvariablen

Sind die Ereignisse bereits reelle Zahlen wie z. B. beim Würfel, kann der zugewiesene Wert einfach das Ereignis sein:

$$X: \{1,2,3,4,5,6\} \rightarrow \{1,2,3,4,5,6\}$$

Ist dies wie im Beispiel der Münze nicht der Fall, müssen wir uns eine Vorschrift überlegen, mit der wir die Ereignisse in reelle Zahlenwerte umwandeln:

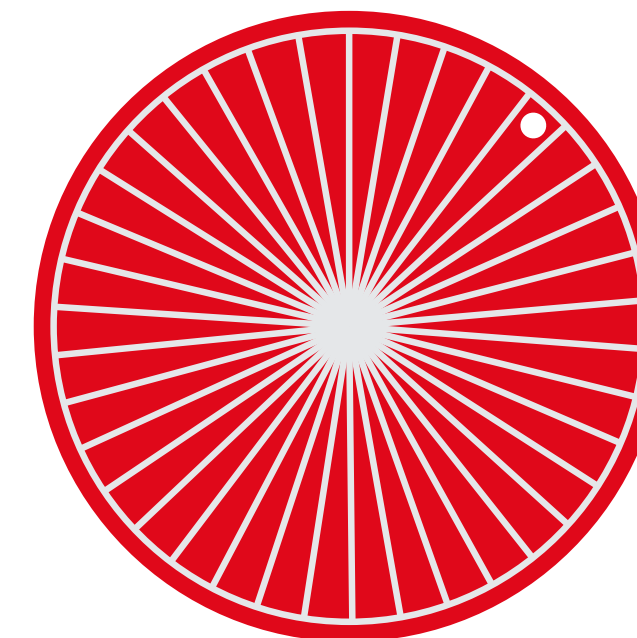
$$X: \{\text{Kopf, Zahl}\} \rightarrow \{0, 1\}$$



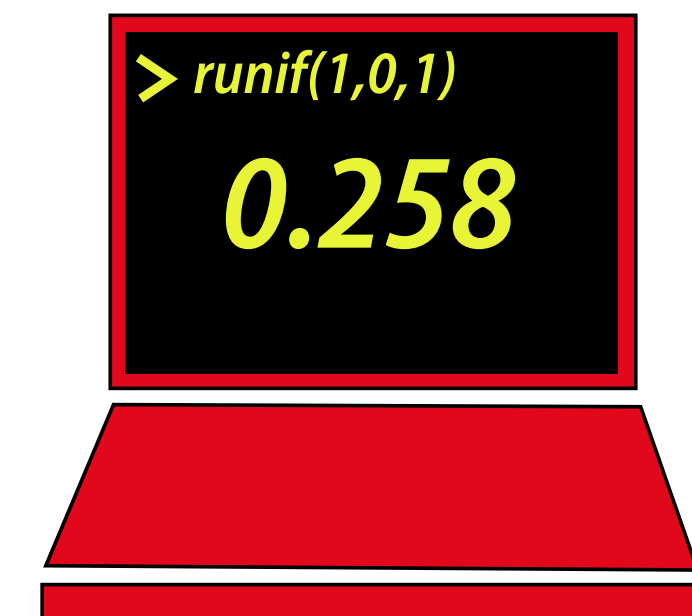
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



$\Omega = \{\text{Kopf, Zahl}\}$
DISKRET



$\Omega = \{00,0,1,2,\dots,36\}$
DISKRET



$\Omega = (0,1)$
KONTINUIERLICH

Zufallsvariablen

Ein Wahrscheinlichkeitsmaß ist eine Abbildung, die jedem Element aus E eine Zahl von 0 bis 1 zuordnet ...

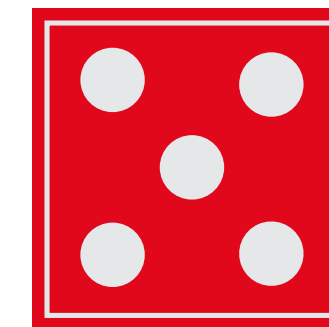
$$p(X): E \rightarrow [0,1]$$

... und die drei Kolmogorov-Axiome erfüllt:

$$P(X=x) \in [0,1] \quad \forall x \in E$$

$$P(X \in E) = 1$$

$$P(X=x_1 \vee X=x_2) = P(X=x_1) + P(X=x_2)$$



$$\Omega = \{1,2,3,4,5,6\}$$

X = Augenzahl eines Würfels

$$E = \{1,2,3,4,5,6\}$$

$$P(X=x) = \frac{1}{6} \quad \forall x \in E$$

$$P(X \in E) = 1$$

$$P(X=x_1 \vee X=x_2) = \frac{2}{6} \quad x_1, x_2 \in E \text{ mit } x_1 \neq x_2$$

Zufallsvariablen

Über die Ereignismenge und die Wahrscheinlichkeiten der enthaltenen Ereignisse entscheidet die **Verteilung**.

Verteilungen werden durch die rechts gezeigten Funktionen definiert.

	Wkt. das $ZV \leq x$	Wkt. das $ZV = x$
diskret	Verteilungsfunktion	Wahrscheinlichkeitsfunktion
kontinuierlich	Verteilungsfunktion	Dichtefunktion

Zufallsvariablen

Ähnlich wie bei Merkmalen in der deskriptiven Statistik unterscheiden wir in diskrete und kontinuierliche Verteilungen!

Diskrete Verteilungen besitzen endlich oder abzählbar unendlich viele Ereignisse.

Kontinuierliche Verteilungen besitzen nicht abzählbar unendlich viele Ereignisse.

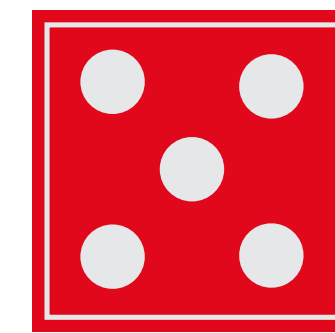
	Wkt. das $ZV \leq x$	Wkt. das $ZV = x$
diskret	Verteilungsfunktion	Wahrscheinlichkeitsfunktion
kontinuierlich	Verteilungsfunktion	Dichtefunktion

Diskrete Verteilungen

Die einfachste Verteilung ist die diskrete Gleichverteilung.

Wie alle diskreten Verteilungen wird sie durch eine Verteilungs- und eine Wahrscheinlichkeitsfunktion definiert.

Schauen wir uns dazu das Beispiel des Würfels an und definieren X , als die Augenzahl die der Würfel anzeigt ...



$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

X = Augenzahl eines Würfels

$$E = \{1, 2, 3, 4, 5, 6\}$$

$$P(X=x) = \frac{1}{6} \quad \forall x \in E$$

$$P(X \in E) = 1$$

$$P(X=x_1 \vee X=x_2) = \frac{2}{6} \quad x_1, x_2 \in E \text{ mit } x_1 \neq x_2$$

Diskrete Verteilungen

Die **Verteilungsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable gleich oder kleiner als x ist.

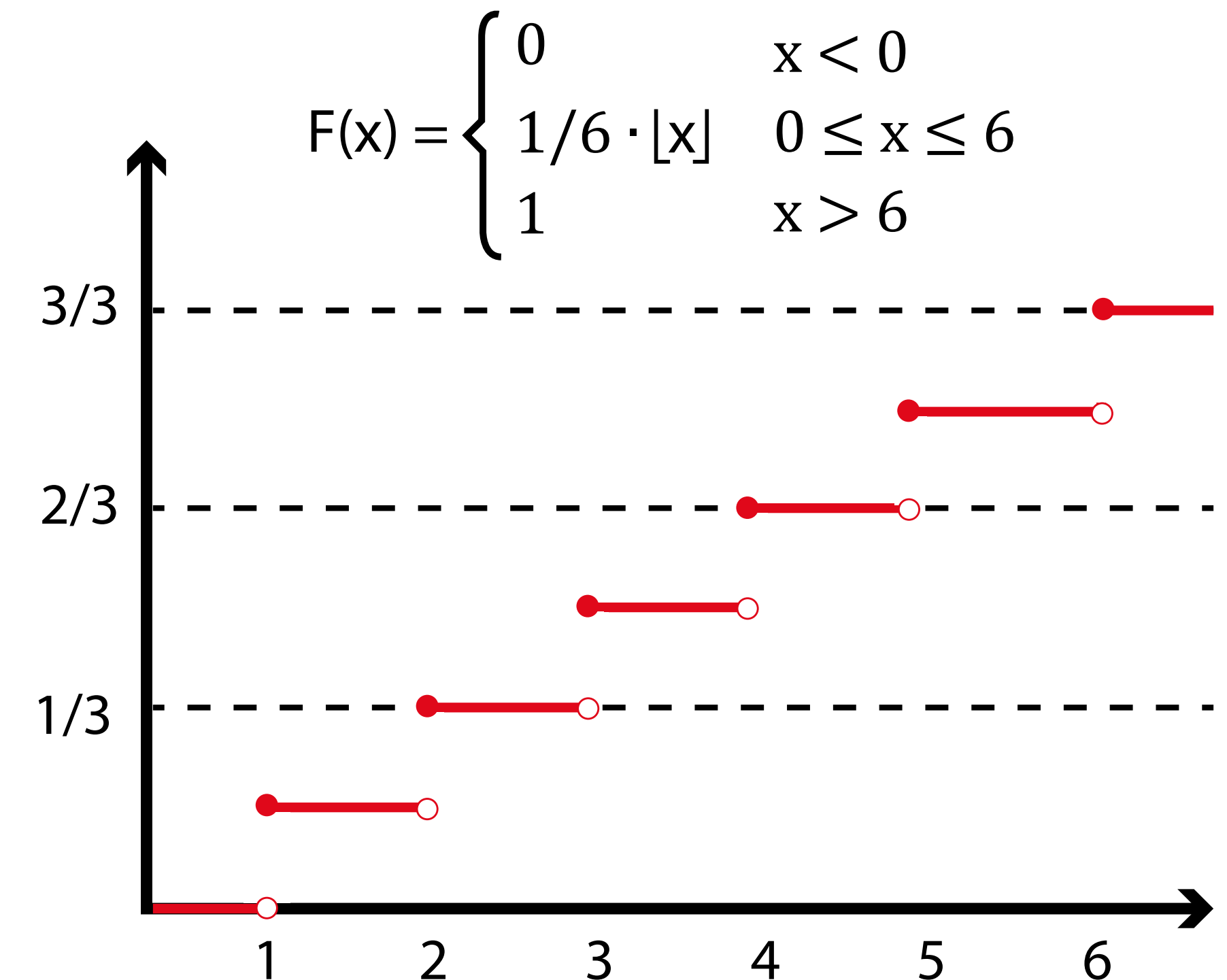
$$F(x) = P(X \leq x)$$

Sie besitzt daher folgende Eigenschaften:

$$F(x) \in [0,1] \quad \forall x \in E$$

$F(x)$ monoton steigend in E

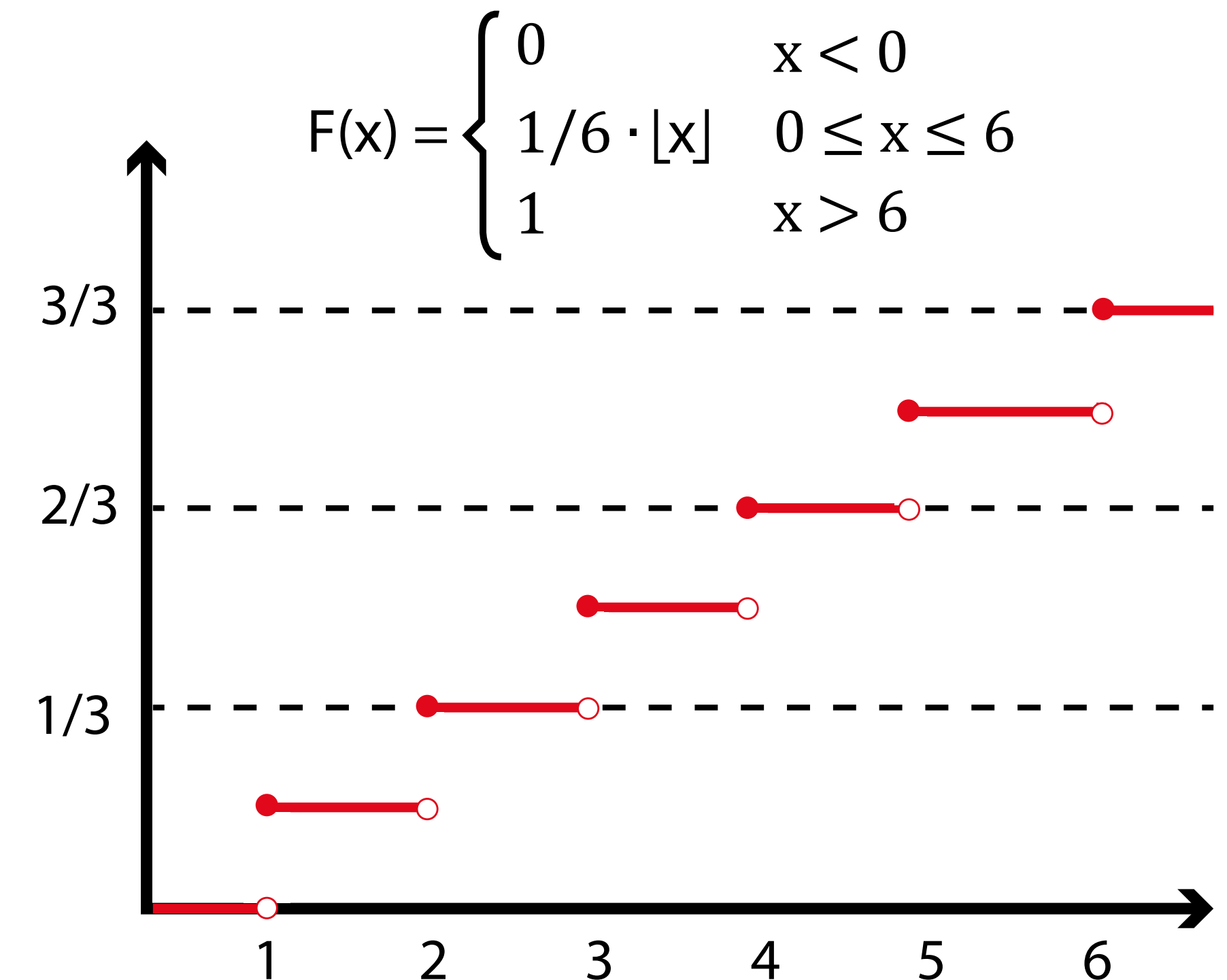
$$\lim_{x \rightarrow -\infty} F(x) = 0 \quad \lim_{x \rightarrow \infty} F(x) = 1$$



Diskrete Verteilungen

Die **Verteilungsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable gleich oder kleiner als x ist.

Beim Würfel ist dies eine Treppe, die bei jeder ganzen Zahl um $1/6$ nach oben geht, bis sie die 1 erreicht.



Der Abrundungsfunktion $[x]$ bzw. auch $\text{floor}(x)$ gibt die nächstkleinere ganze Zahl an. Pendant dazu wäre die Aufrundungsfunktion $\lceil x \rceil$ bzw. $\text{ceil}(x)$.

Diskrete Verteilungen

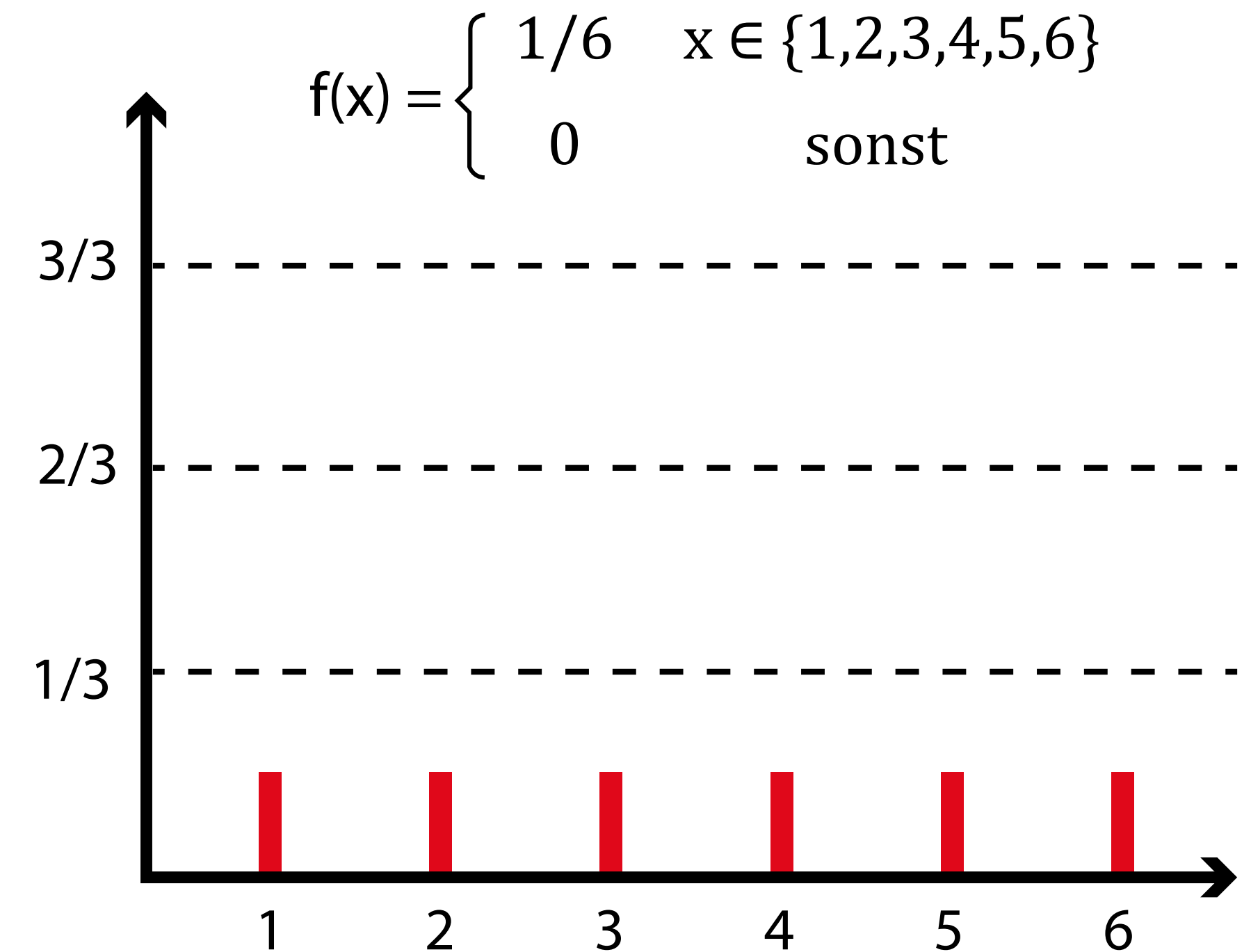
Die **Wahrscheinlichkeitsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable genau x ist.

$$f(x) = P(X = x)$$

Sie besitzt daher folgende Eigenschaften:

$$f(x) \in [0,1] \quad \forall x \in E$$

$$\sum_{x \in E} f(x) = 1$$

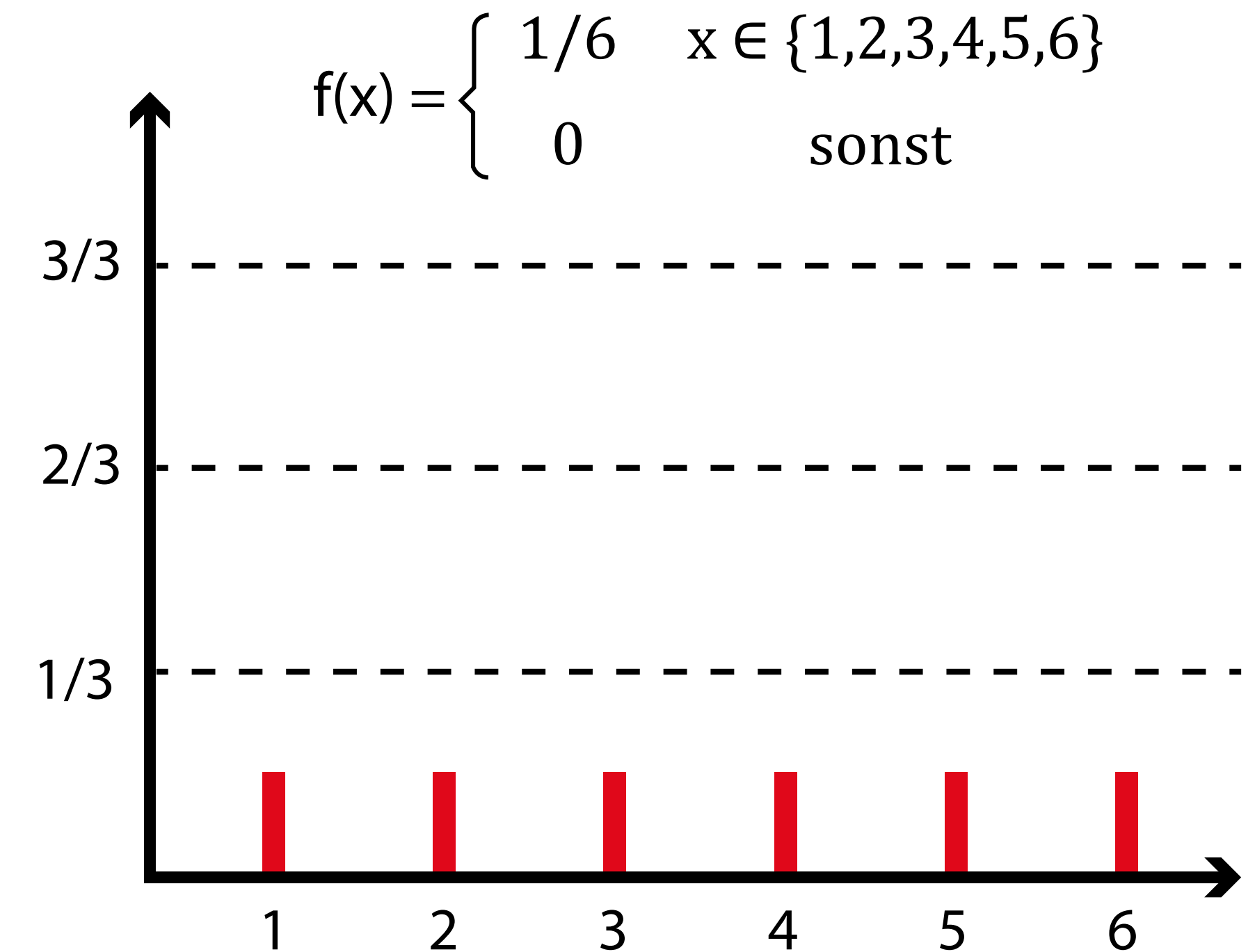


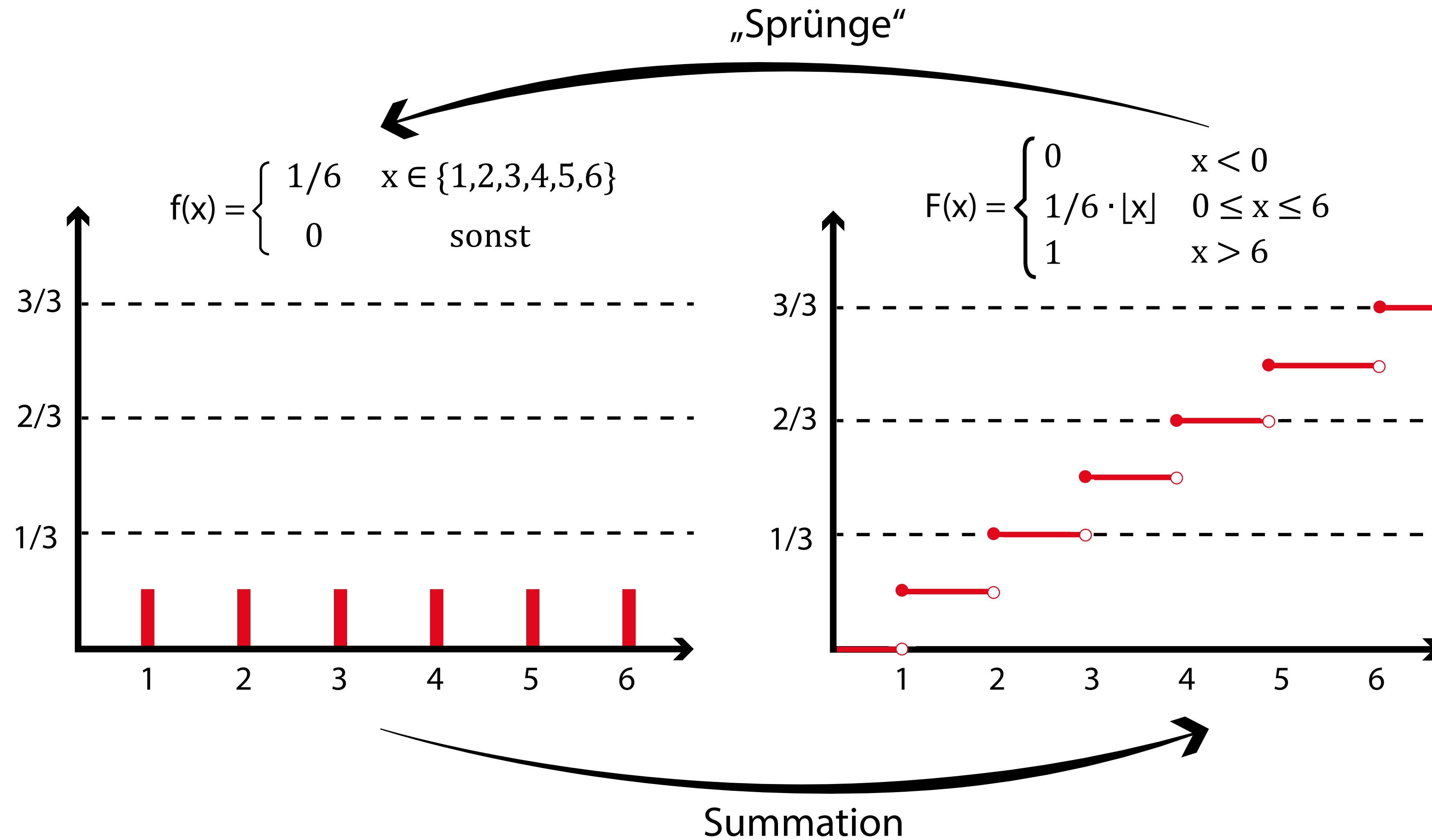
Diskrete Verteilungen

Die **Wahrscheinlichkeitsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable genau x ist.

Beim Würfel kann die Zufallsvariable nur die ganzen Zahlen von 1 bis 6 annehmen. Diese haben die Wahrscheinlichkeit von $1/6$.

Bei allen anderen Werten ist die Wahrscheinlichkeit 0.



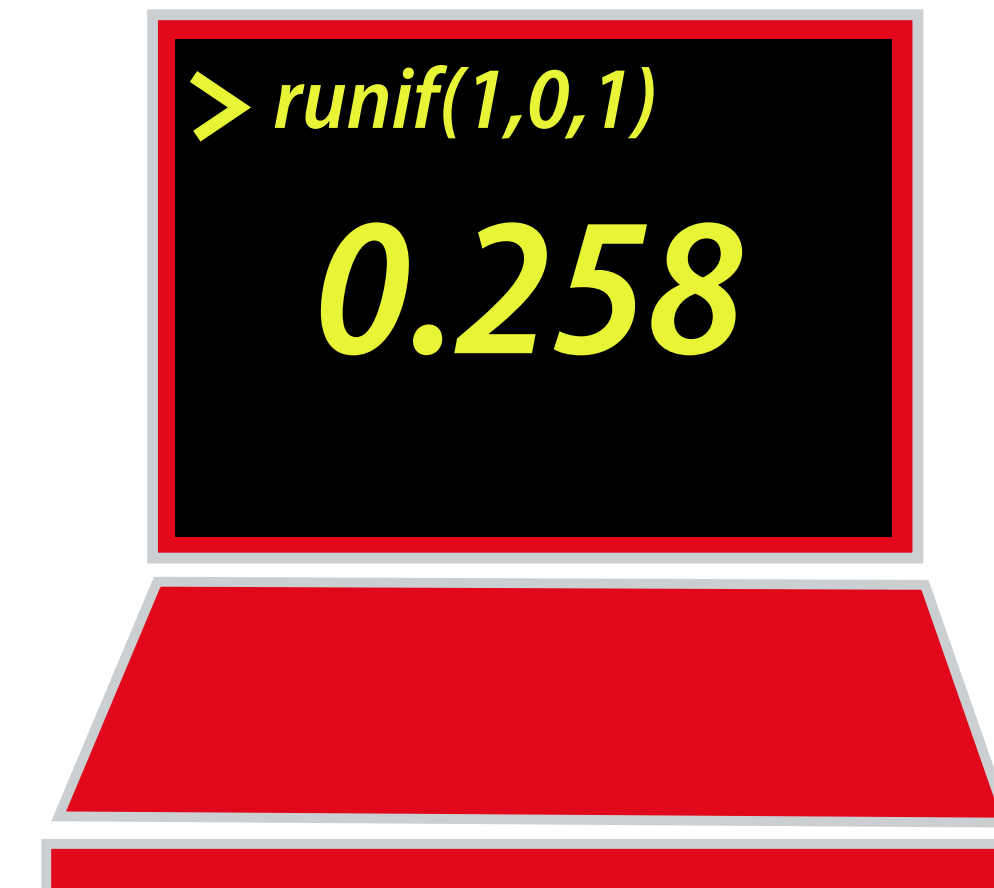


Kontinuierliche Verteilungen

Die Gleichverteilung gibt es auch in kontinuierlich.

Kontinuierliche Verteilungen lassen sich auch über eine Verteilungsfunktion definieren. Statt einer Wahrscheinlichkeitsfunktion haben sie jedoch eine **Dichtefunktion**.

Als Beispiel dazu verwenden wir einen Zufallsgenerator, der eine Zufallszahl zwischen 0 und 1 ausspuckt.



$$\Omega = (0,1)$$

X = Angezeigter Wert

$$E = (0,1)$$

Kontinuierliche Verteilungen

Die **Verteilungsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable gleich oder kleiner als x ist.

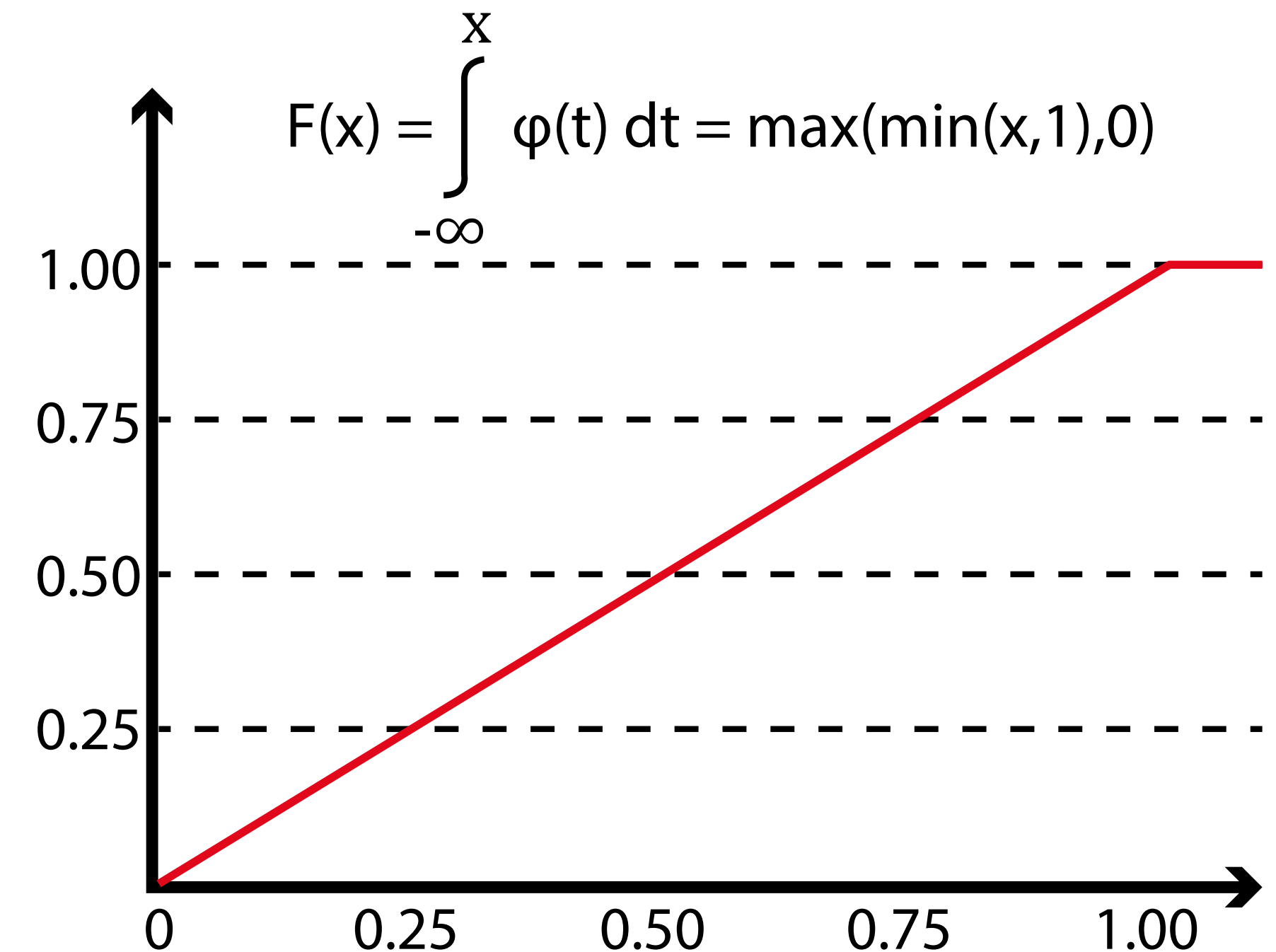
$$F(x) = P(X \leq x)$$

Bei unserem Zufallsgenerator ist diese ...

...0 für alle Werte von x bis zur 0

... x für alle Werte von x zwischen 0 und 1

...1 für alle Werte von x größer als 1



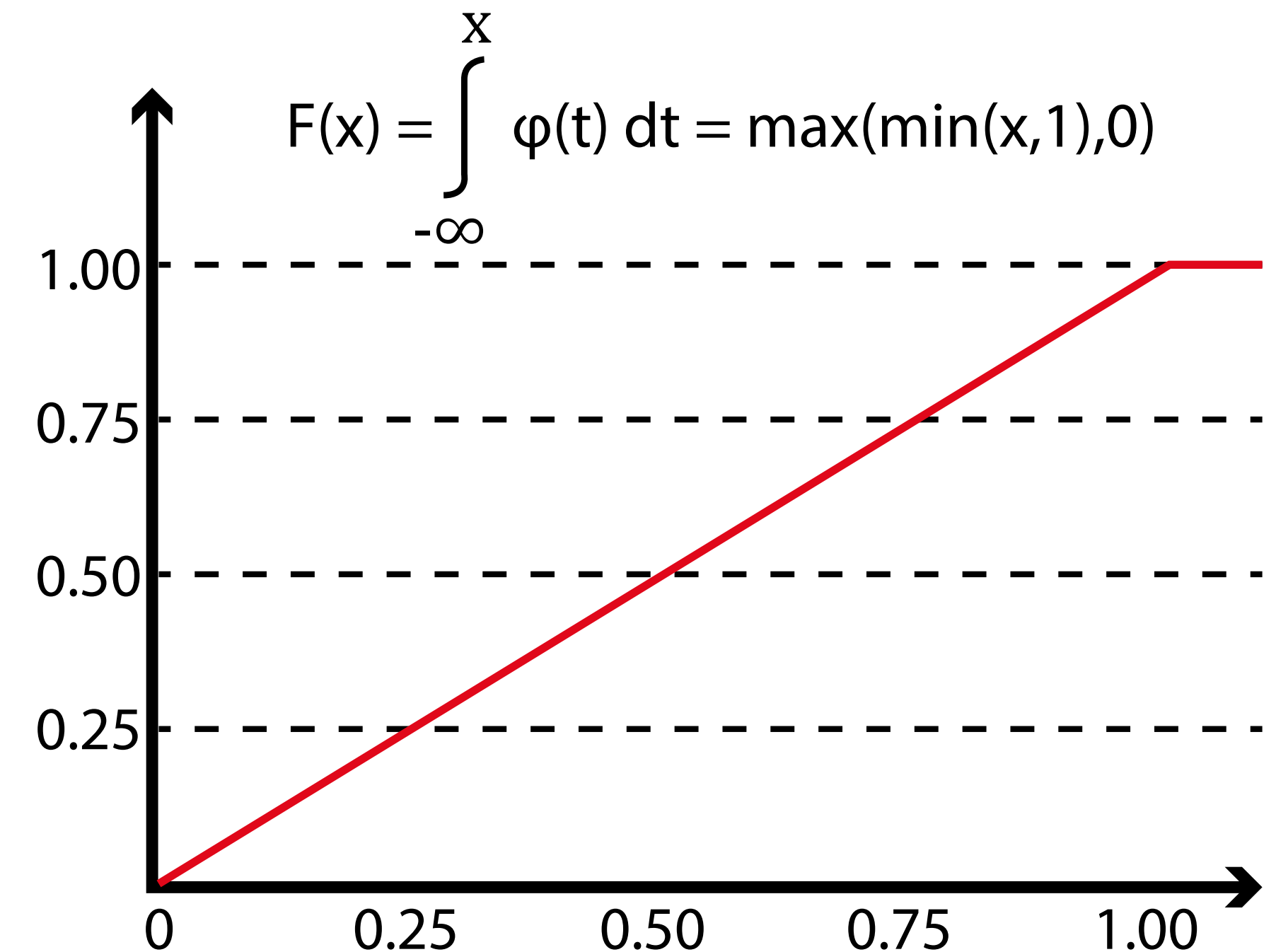
Kontinuierliche Verteilungen

Die **Verteilungsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable gleich oder kleiner als x ist.

$$F(x) = P(X \leq x)$$

Sie besitzt folgenden Zusammenhang zur Dichtefunktion:

$$F(x) = \int_{-\infty}^x \varphi(t) dt$$

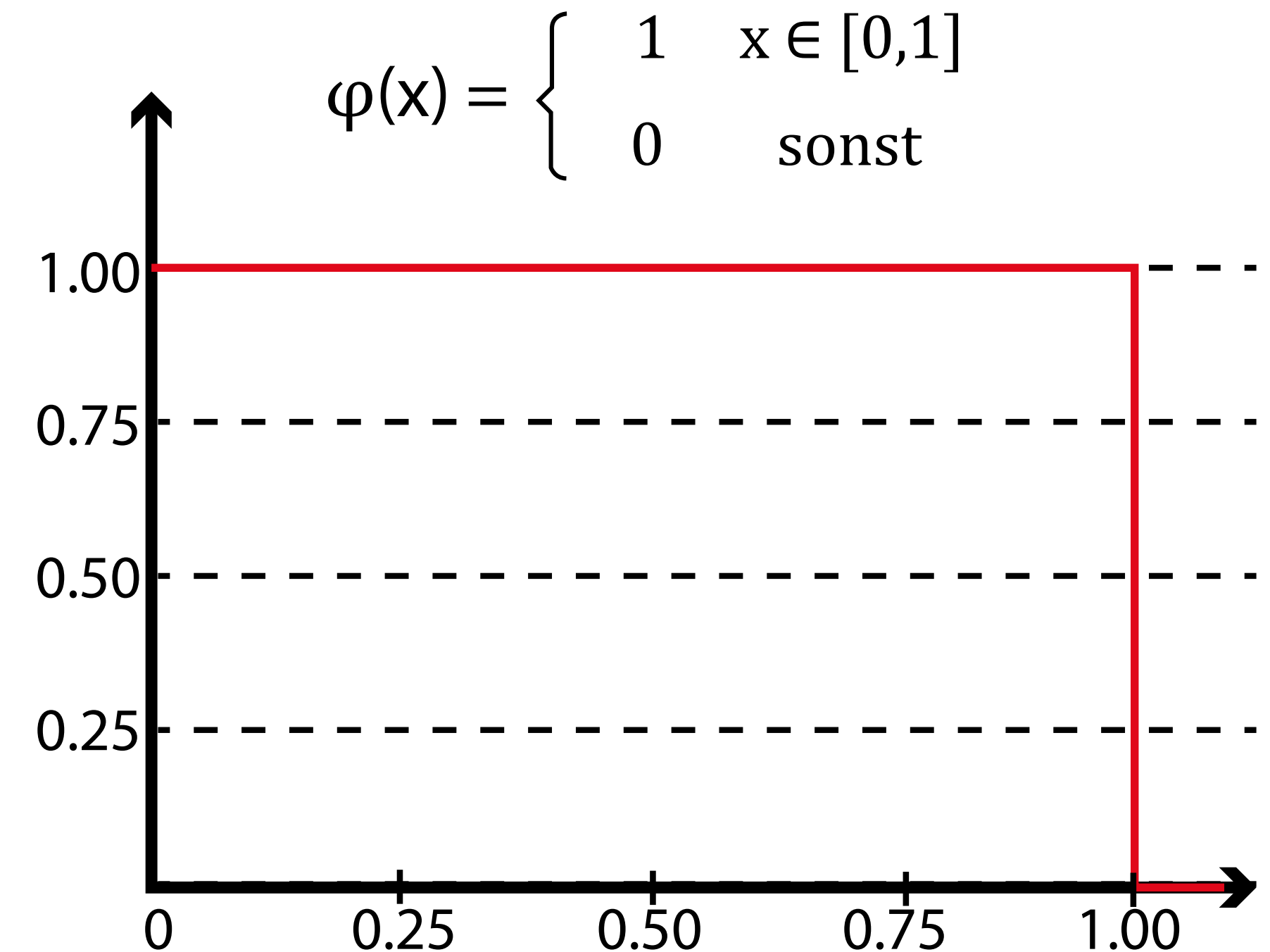


Kontinuierliche Verteilungen

Die **Dichtefunktion** gibt anders als die Wahrscheinlichkeitsfunktion nicht die Wahrscheinlichkeit an, dass x genau einen bestimmten Wert hat.

Stattdessen gilt: Die Wahrscheinlichkeit, dass x einen Wert zwischen a und b annimmt, ist gegeben durch ...

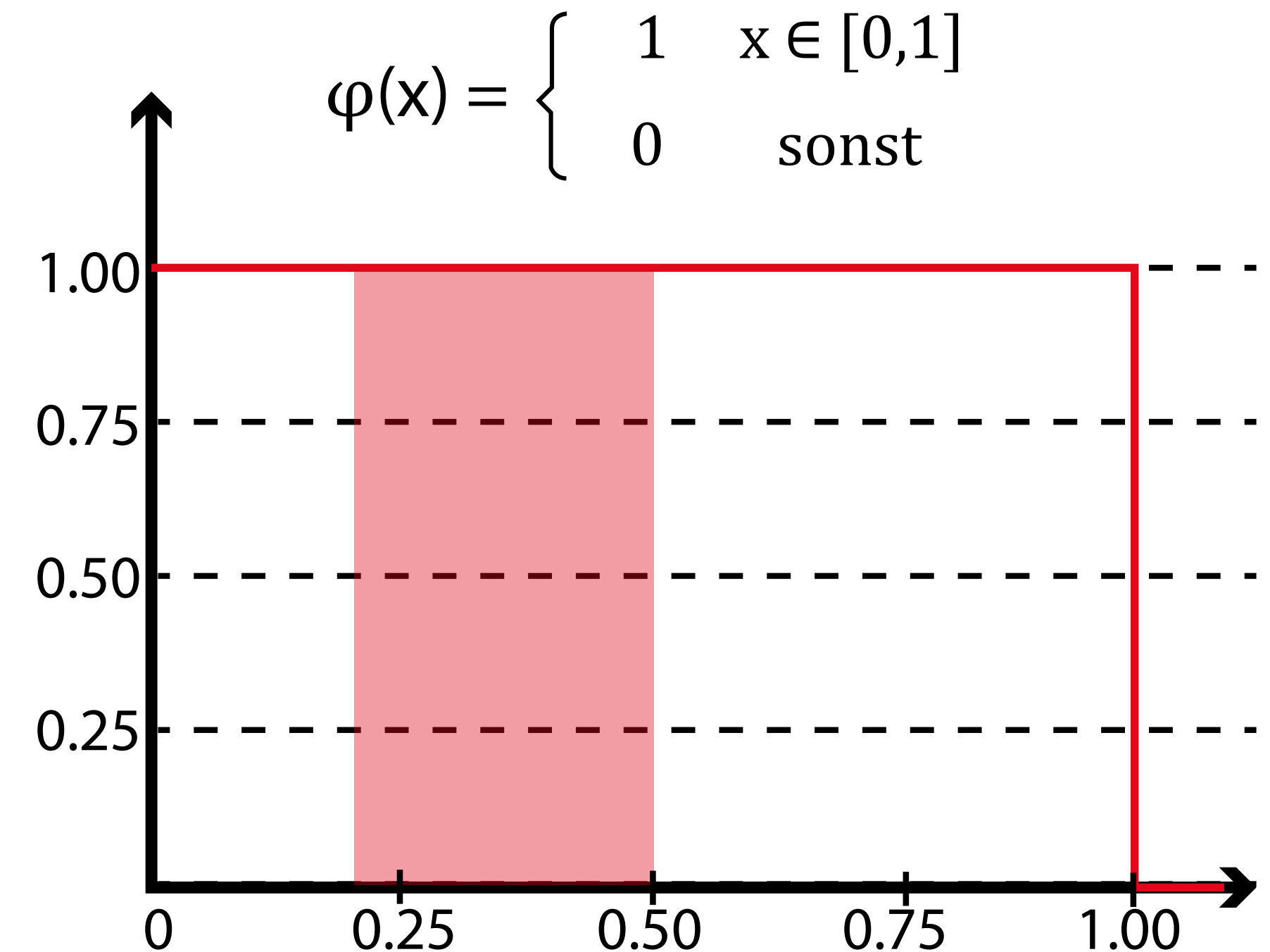
$$P(a \leq X \leq b) = \int_a^b \varphi(x) \, dx$$



Kontinuierliche Verteilungen

Beispiel Die Wahrscheinlichkeit, dass x einen Wert zwischen 0.2 und 0.5 annimmt, ist:

$$P(0.2 \leq X \leq 0.5) = \int_{0.2}^{0.5} \varphi(x) \, dx = \left[x \right]_{0.2}^{0.5} = 0.3$$

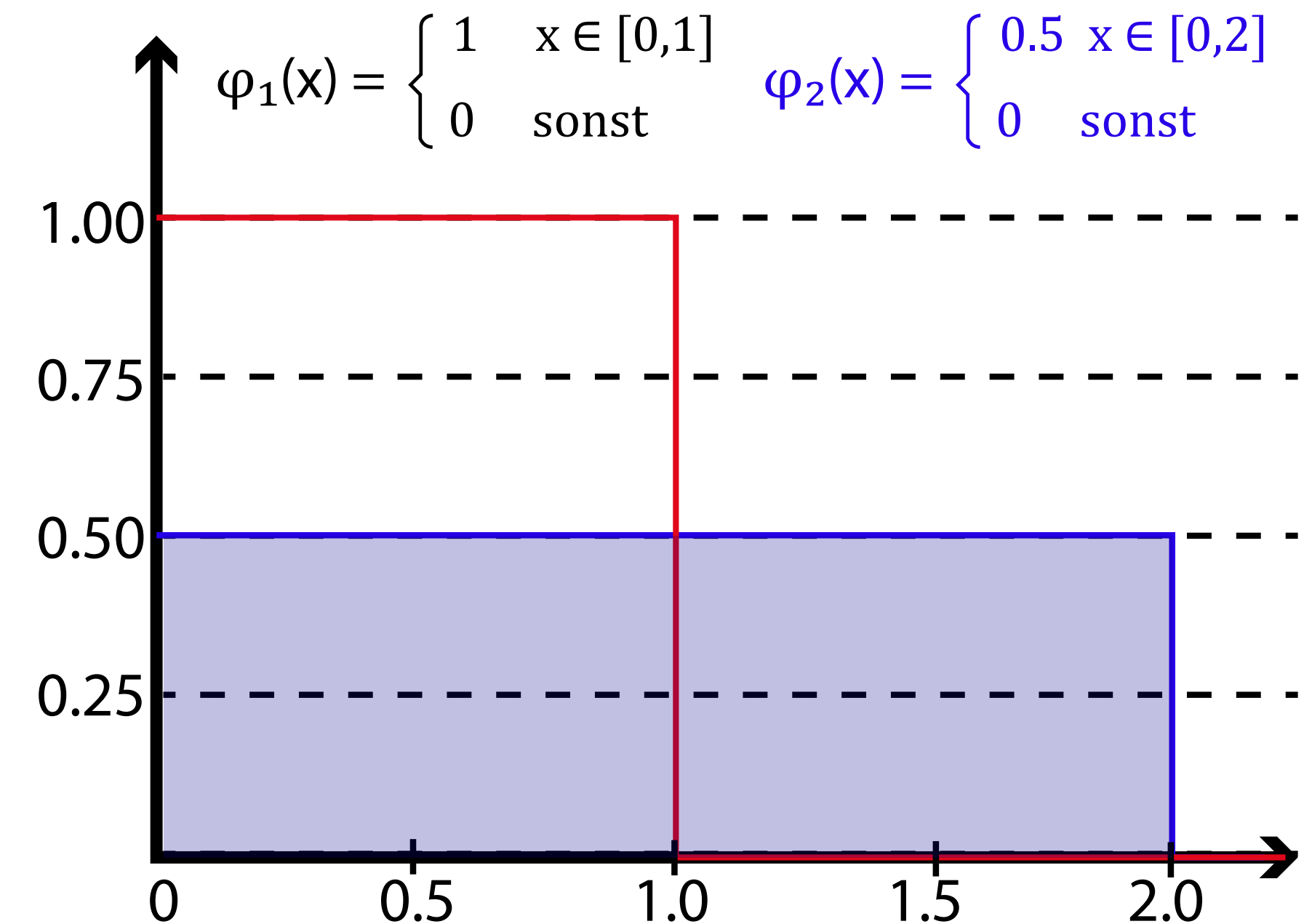


Kontinuierliche Verteilungen

Aus dem zweiten Kolmogorov-Axiom folgt, dass die Gesamtfläche unter der Dichtefunktion genau 1 sein muss.

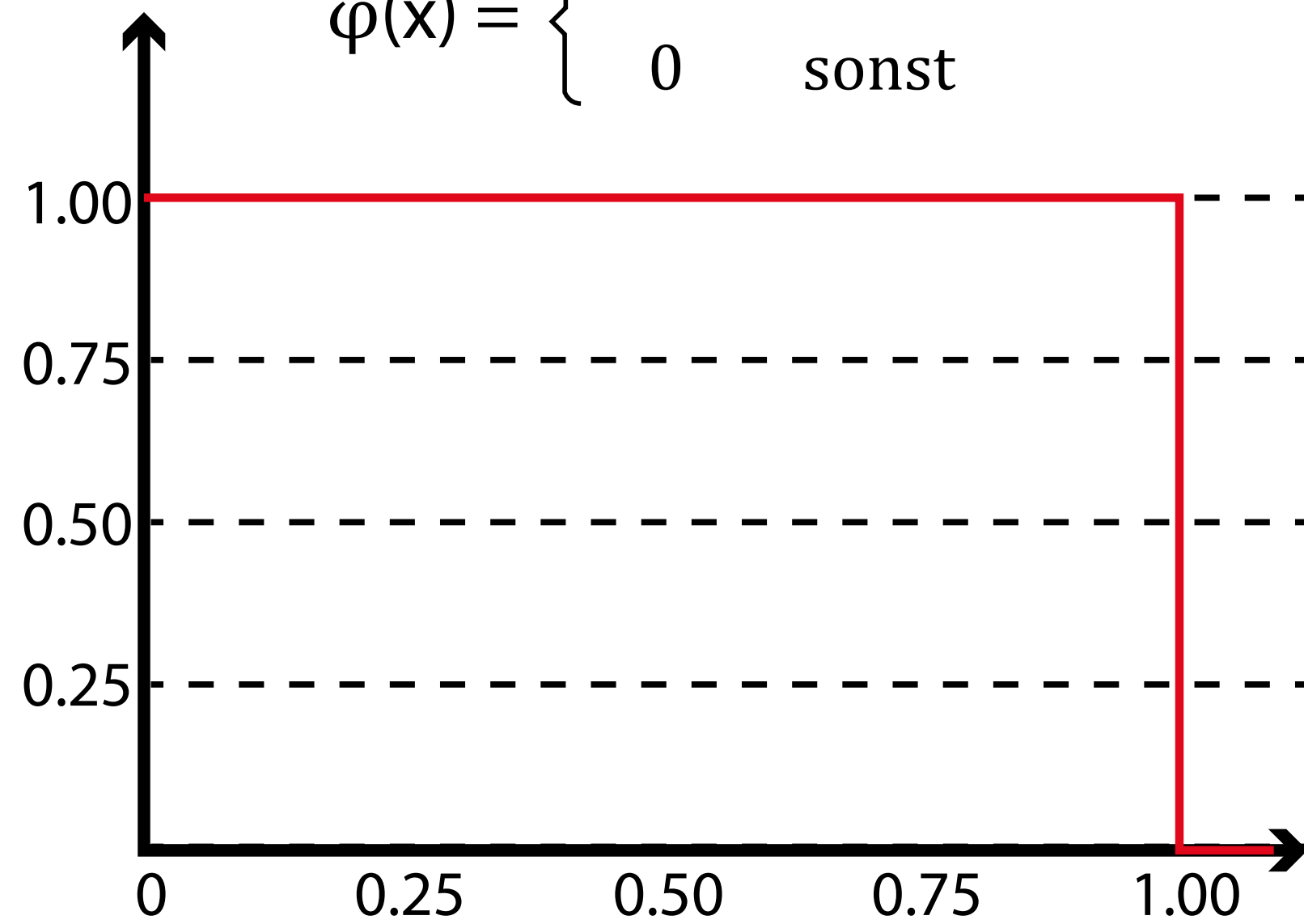
$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} \varphi(x) dx = 1$$

Bei einer Gleichverteilung über das Intervall $[0,2]$ wäre die Höhe der Dichtefunktion z. B. bei konstant 0.5.

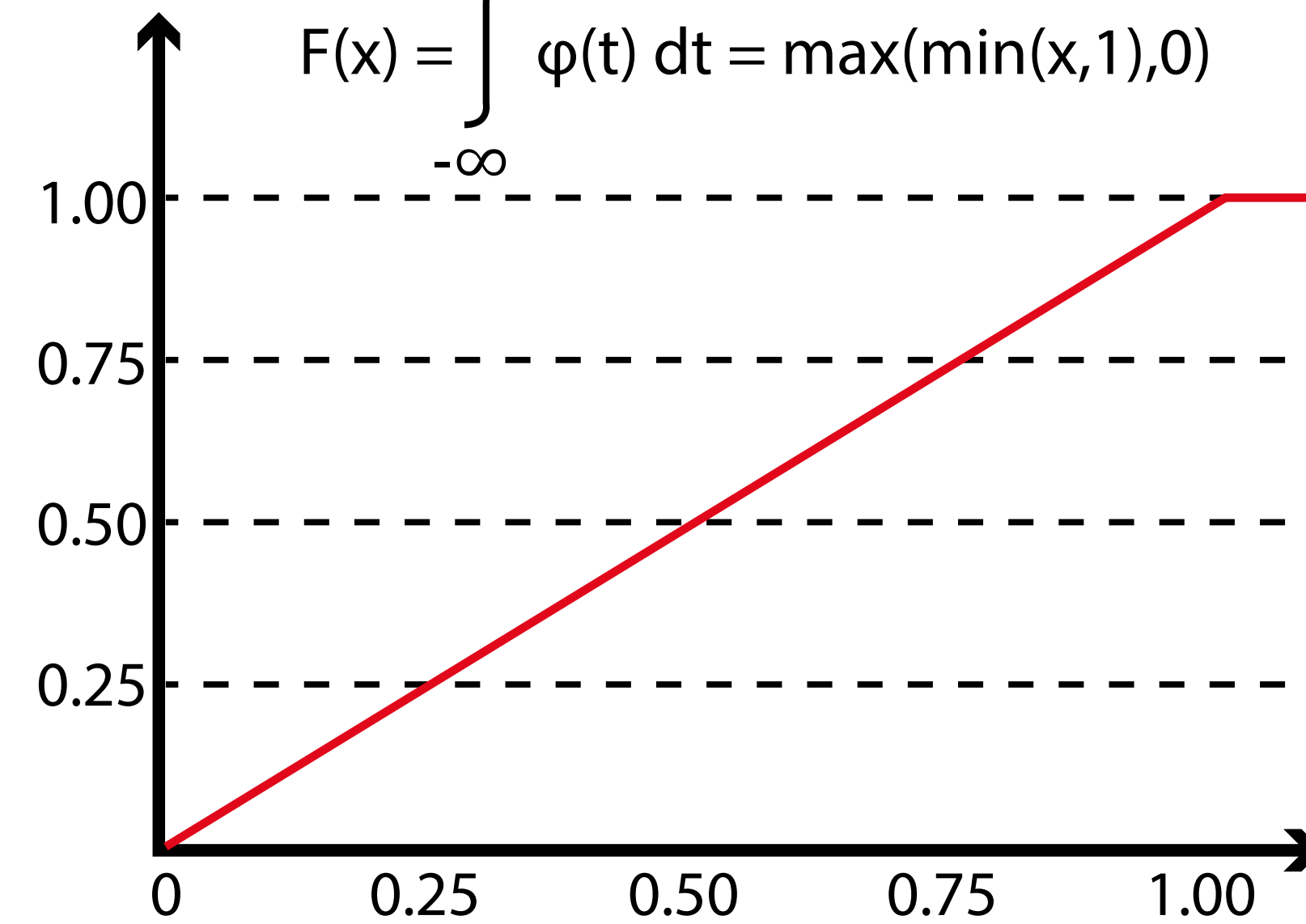


Ableitung

$$\varphi(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{sonst} \end{cases}$$



$$F(x) = \int_{-\infty}^x \varphi(t) dt = \max(\min(x,1),0)$$



Integration

Erwartungswert

Eine wichtige Eigenschaft von Verteilungen ist ihr Erwartungswert μ bzw. $E(X)$. Dieser lässt sich ...

...für diskrete Verteilungen aus der Wahrscheinlichkeitsfunktion berechnen.

... für kontinuierliche Verteilungen aus der Dichtefunktion berechnen.



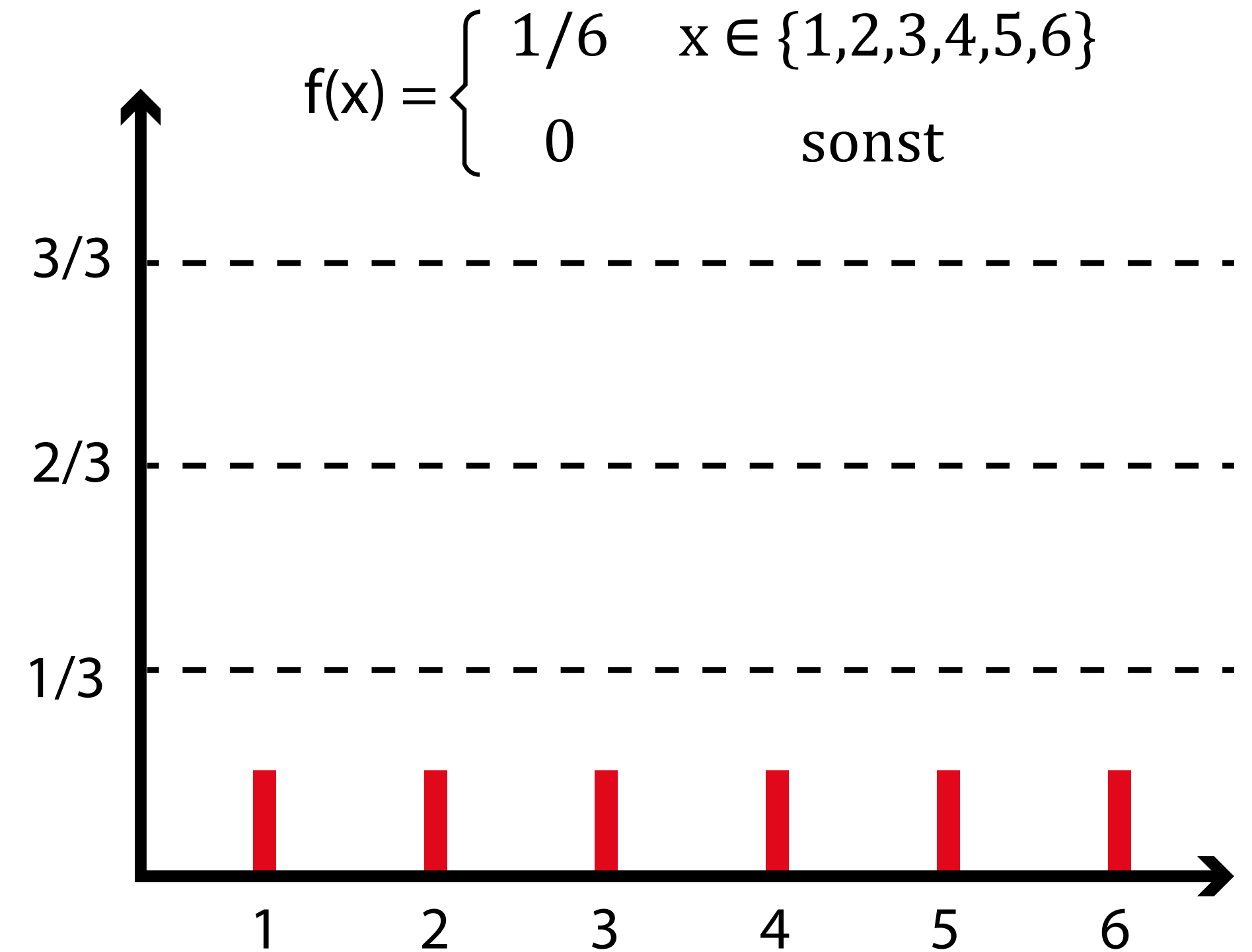
Erwartungswert

Beispiel Würfel Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} 1/6 & x \in \{1,2,3,4,5,6\} \\ 0 & \text{sonst} \end{cases}$$

Wir berechnen den Erwartungswert über die Summe:

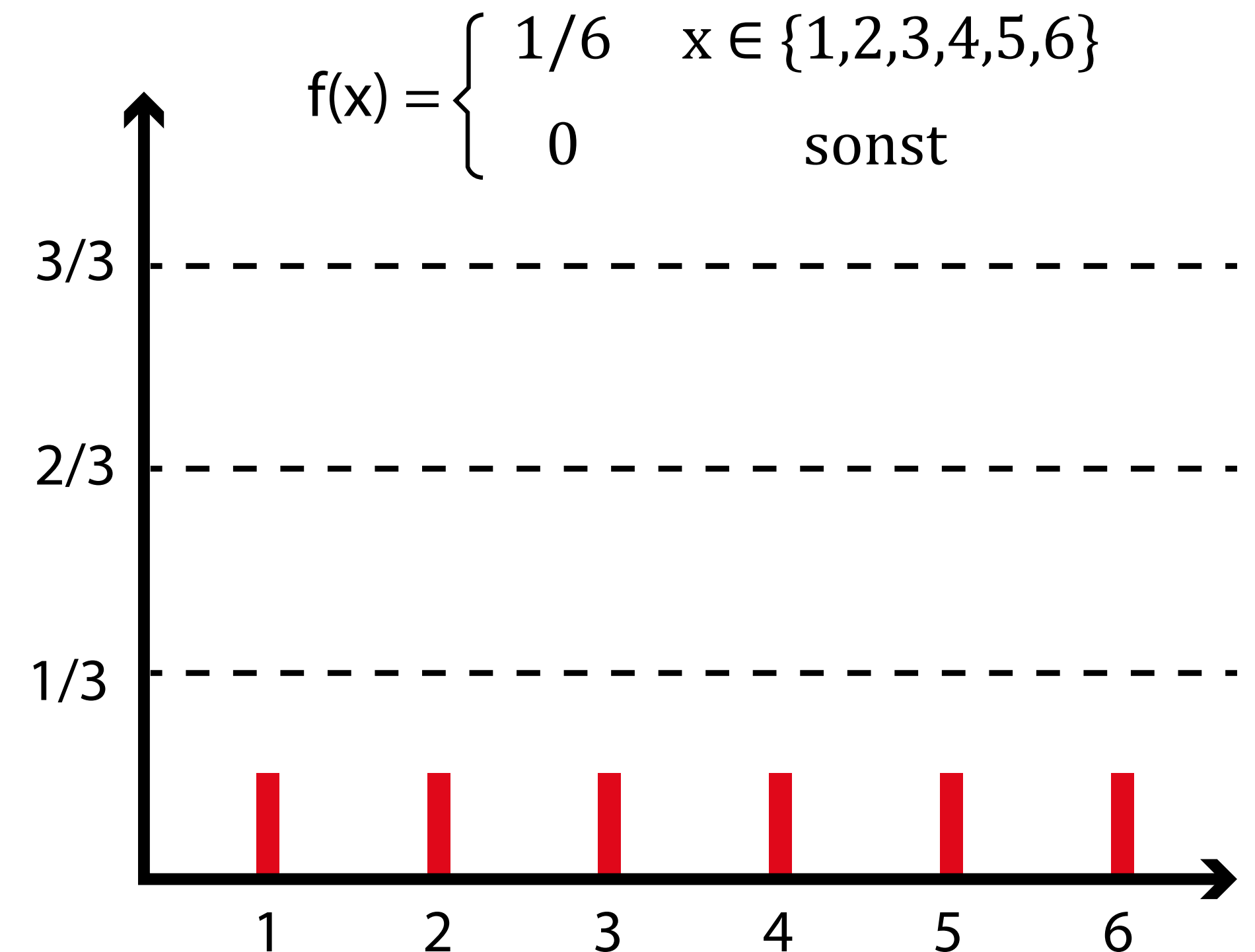
$$\mu = E(X) = \sum_{x \in E} x \cdot f(x)$$



Erwartungswert

Wir berechnen den Erwartungswert über die Summe:

$$\begin{aligned}\mu &= E(X) = \sum_{x \in E} x \cdot f(x) \\ &= 1 \frac{1}{6} + 2 \frac{1}{6} + 3 \frac{1}{6} + 4 \frac{1}{6} + 5 \frac{1}{6} + 6 \frac{1}{6} \\ &= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} \\ &= \frac{21}{6} = 3.5\end{aligned}$$



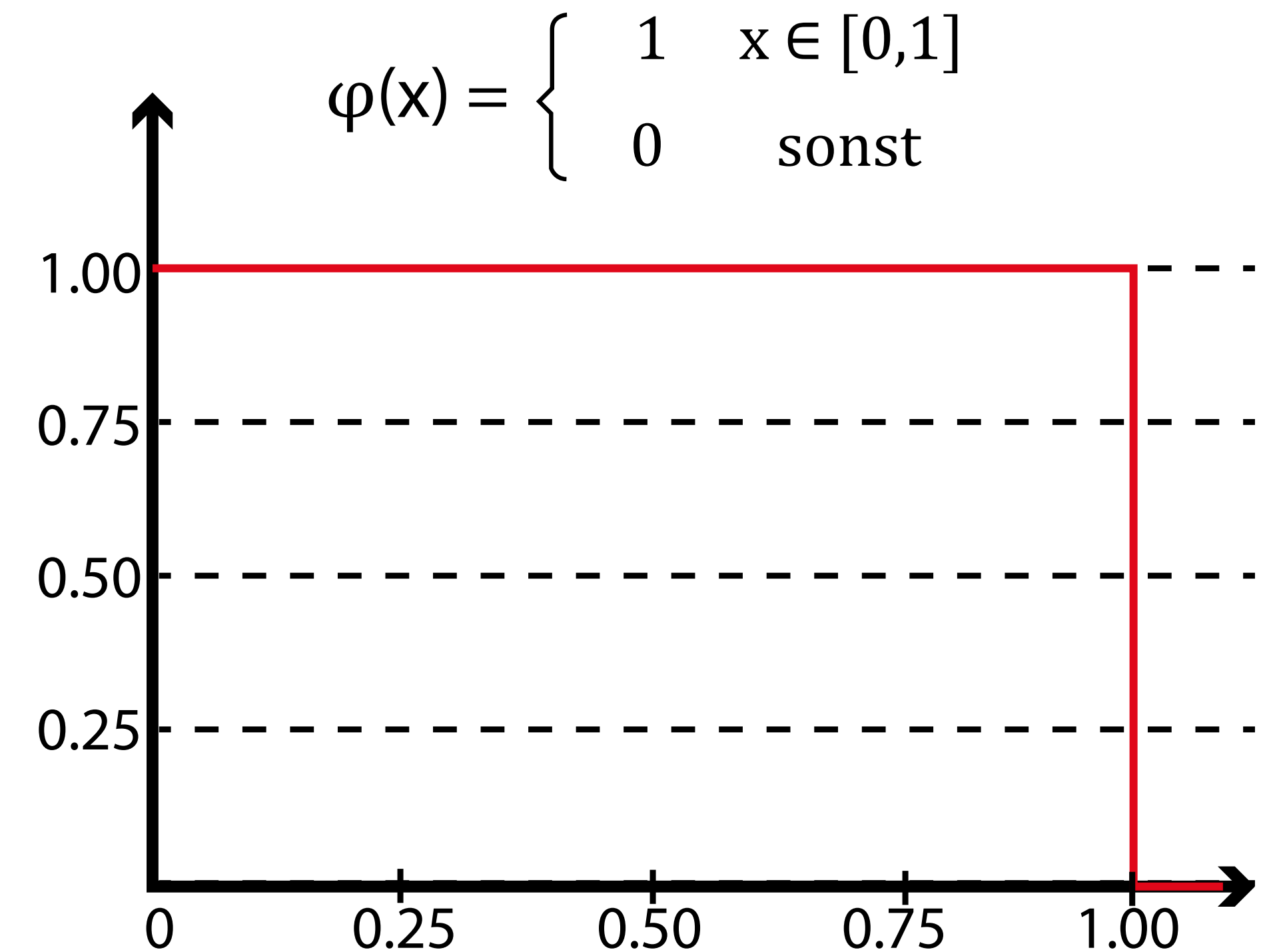
Erwartungswert

Beispiel Zufallsgenerator „0 bis 1“

$$\varphi(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{sonst} \end{cases}$$

Wir berechnen den Erwartungswert über das Integral:

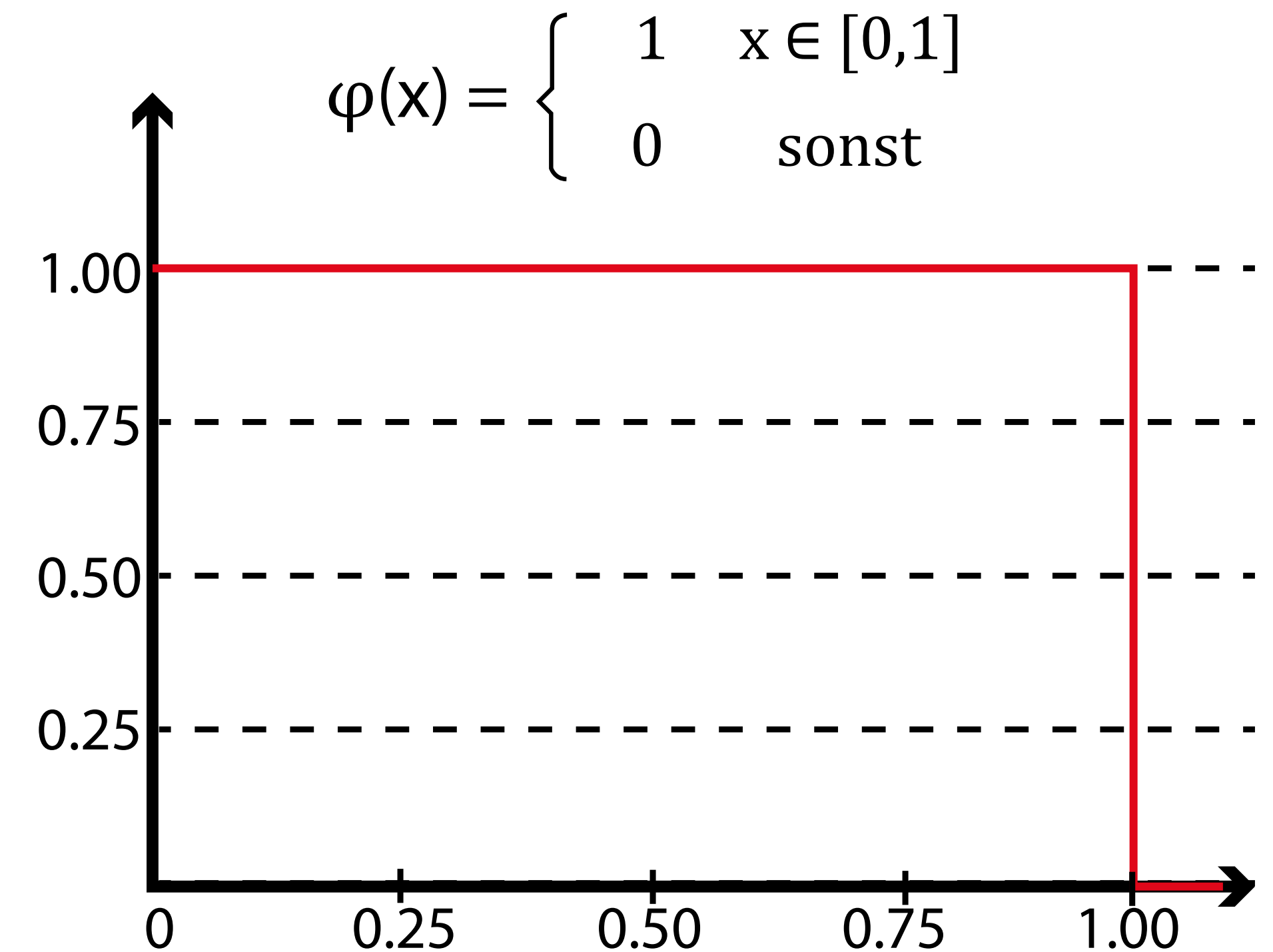
$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot \varphi(x) dx = \int_0^1 x dx$$



Erwartungswert

Wir berechnen den Erwartungswert über das Integral:

$$\begin{aligned}\mu = E(X) &= \int_{-\infty}^{\infty} x \cdot \varphi(x) \, dx = \int_0^1 x \, dx \\ &= \left[0.5x^2 \right]_0^1 \\ &= 0.5 \cdot 1^2 - 0.5 \cdot 0^2 \\ &= 0.5\end{aligned}$$



Erwartungswert

Der Erwartungswert ist linear additiv:

$$E(aX+bY) = a \cdot E(X) + b \cdot E(Y)$$

Beispiel: Wir würfeln mit einem 6-seitigen und einem 20-seitigen Würfel und berechnen die Augensumme. Der D6 zählt dabei dreifach. Der Erwartungswert wäre:

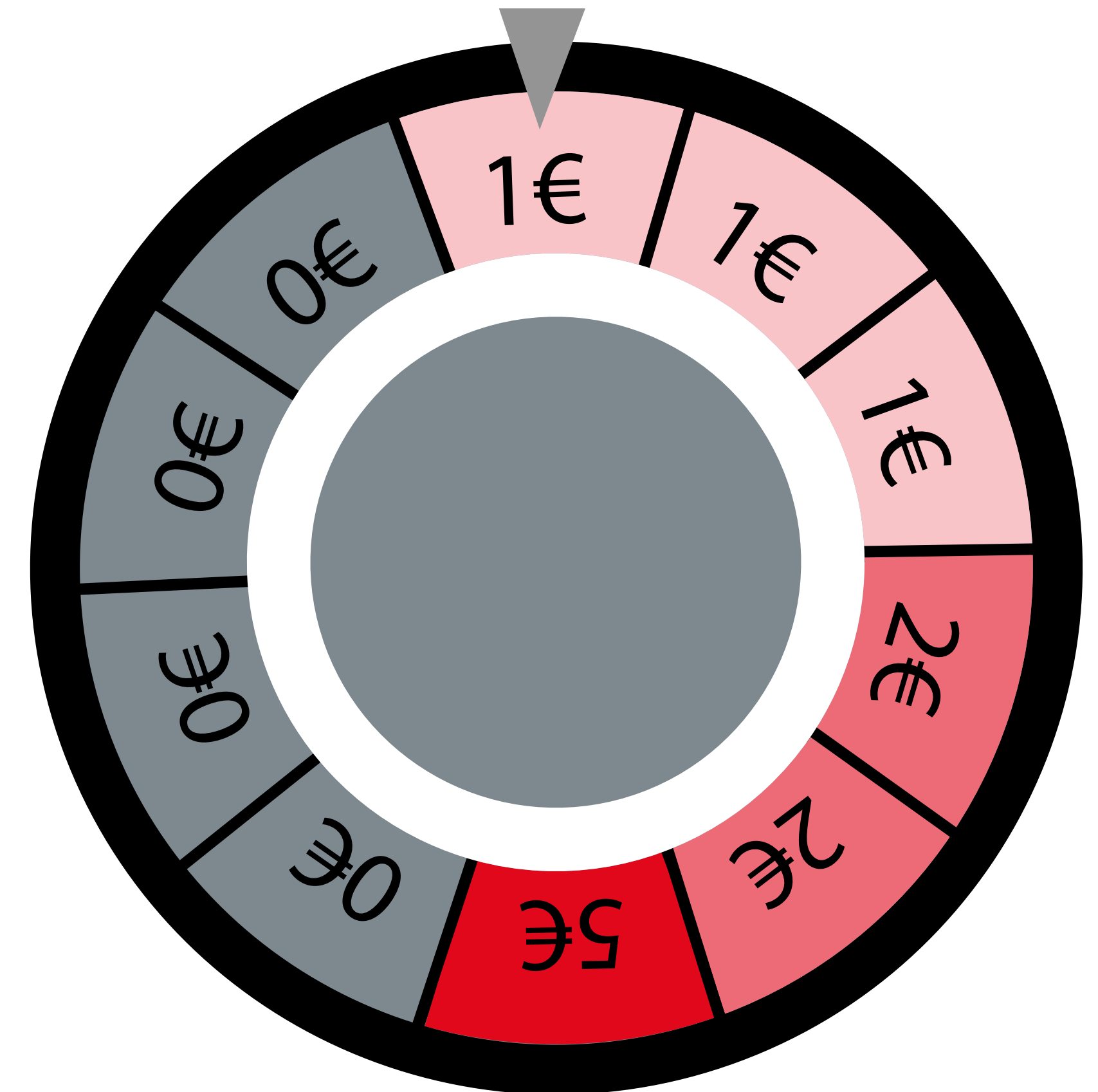
$$E(3X+Y) = 3 \cdot 3.5 + 10.5 = 21$$



Erwartungswert

Wir betrachten das rechts gezeigte Glücksrad und die Zufallsvariable X , welche die erspielte Auszahlung angibt.

- Stelle die Verteilung als Verteilungsfunktion und als Wahrscheinlichkeitsfunktion dar.
- Berechne den Erwartungswert der Auszahlung.



Erwartungswert

Zwei Zufallsgeneratoren erzeugen gleichverteilte Zahlen auf den Intervallen 2 bis 3 und 2 bis 4.

Die Zufallsvariable X (2 bis 3) und Y (2 bis 4) sind die jeweils generierten Zahlen.

- a) Stelle die Verteilung von Y als Verteilungsfunktion und als Dichtefunktion dar.
- b) Berechne den Erwartungswert von $X+Y$

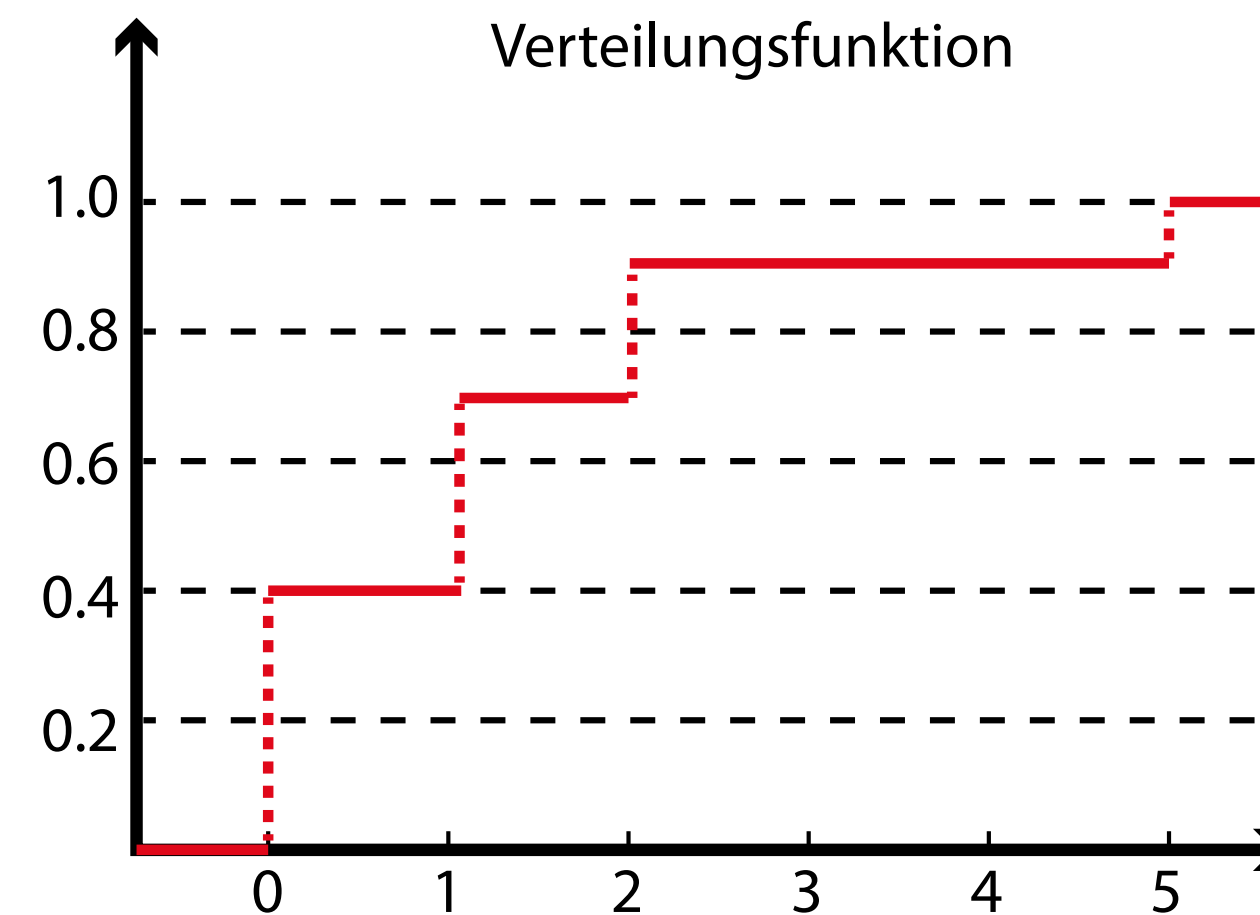
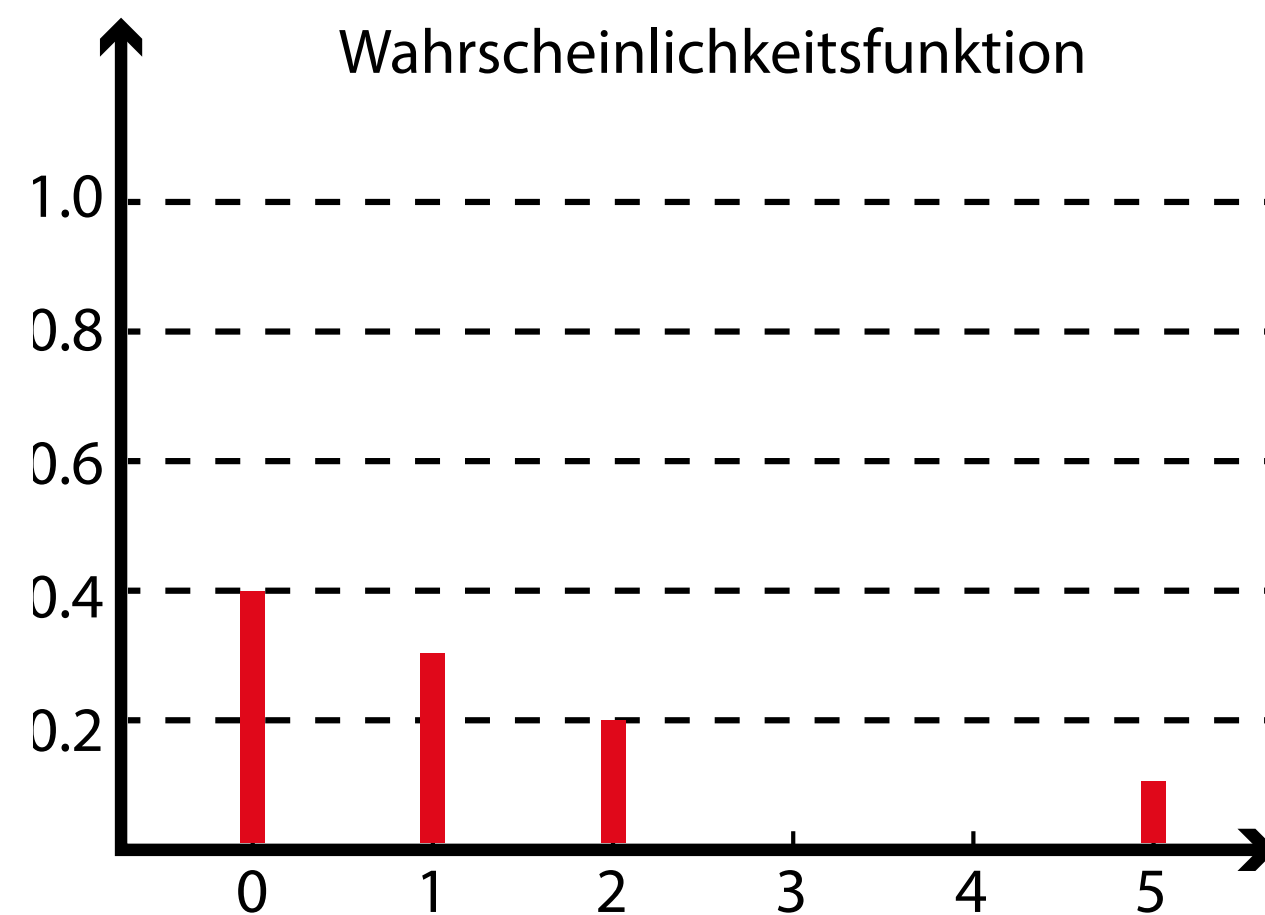
```
> runif(1,2,3)
```

$X = \dots$

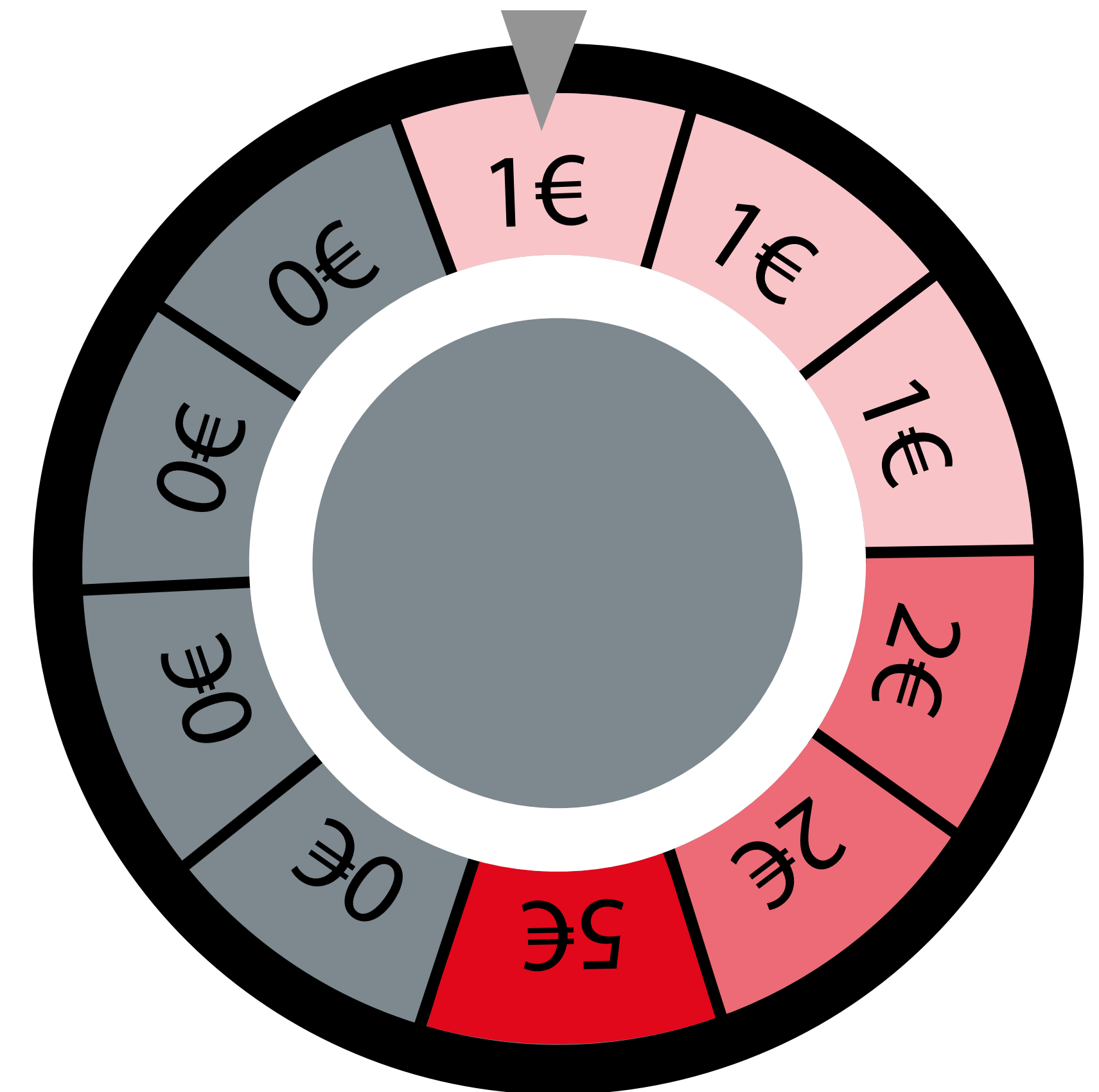
```
> runif(1,2,4)
```

$Y = \dots$

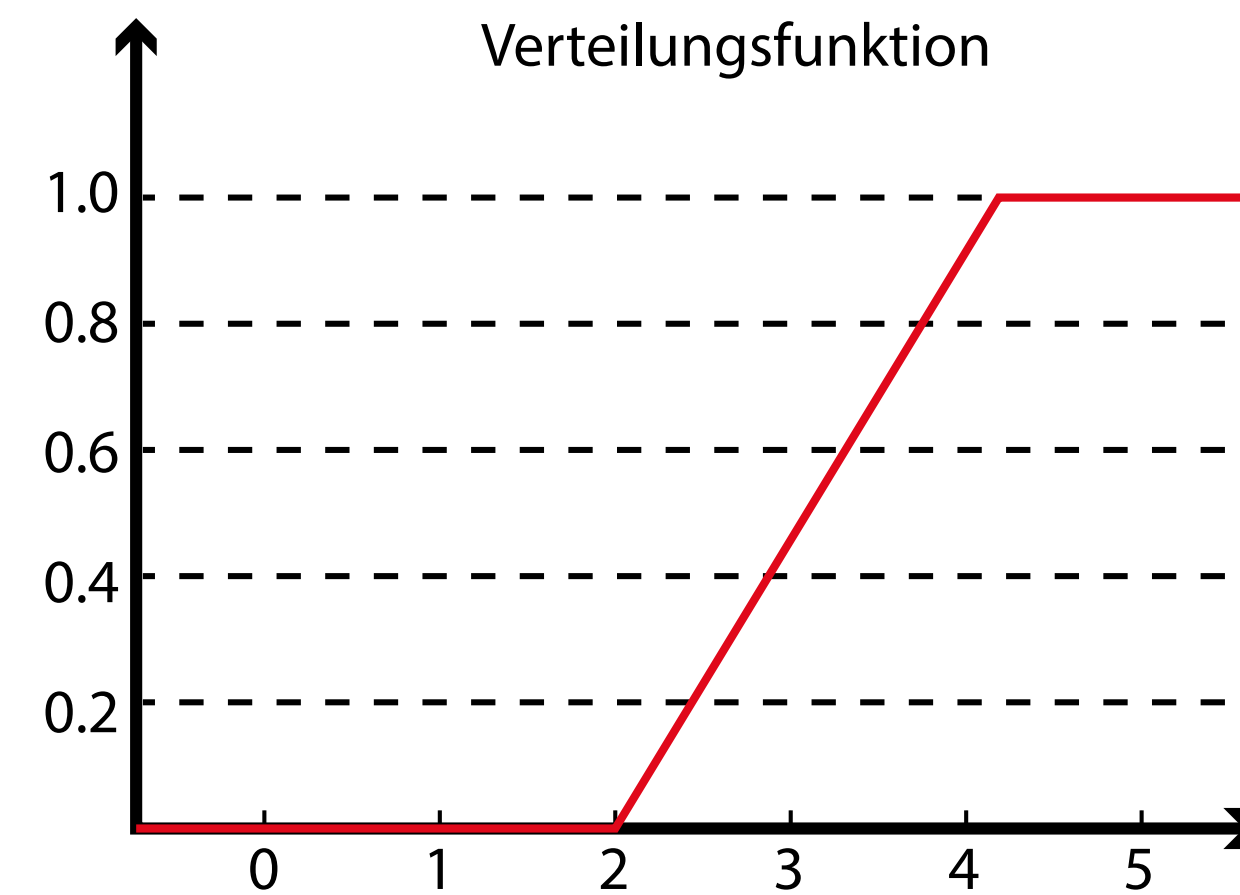
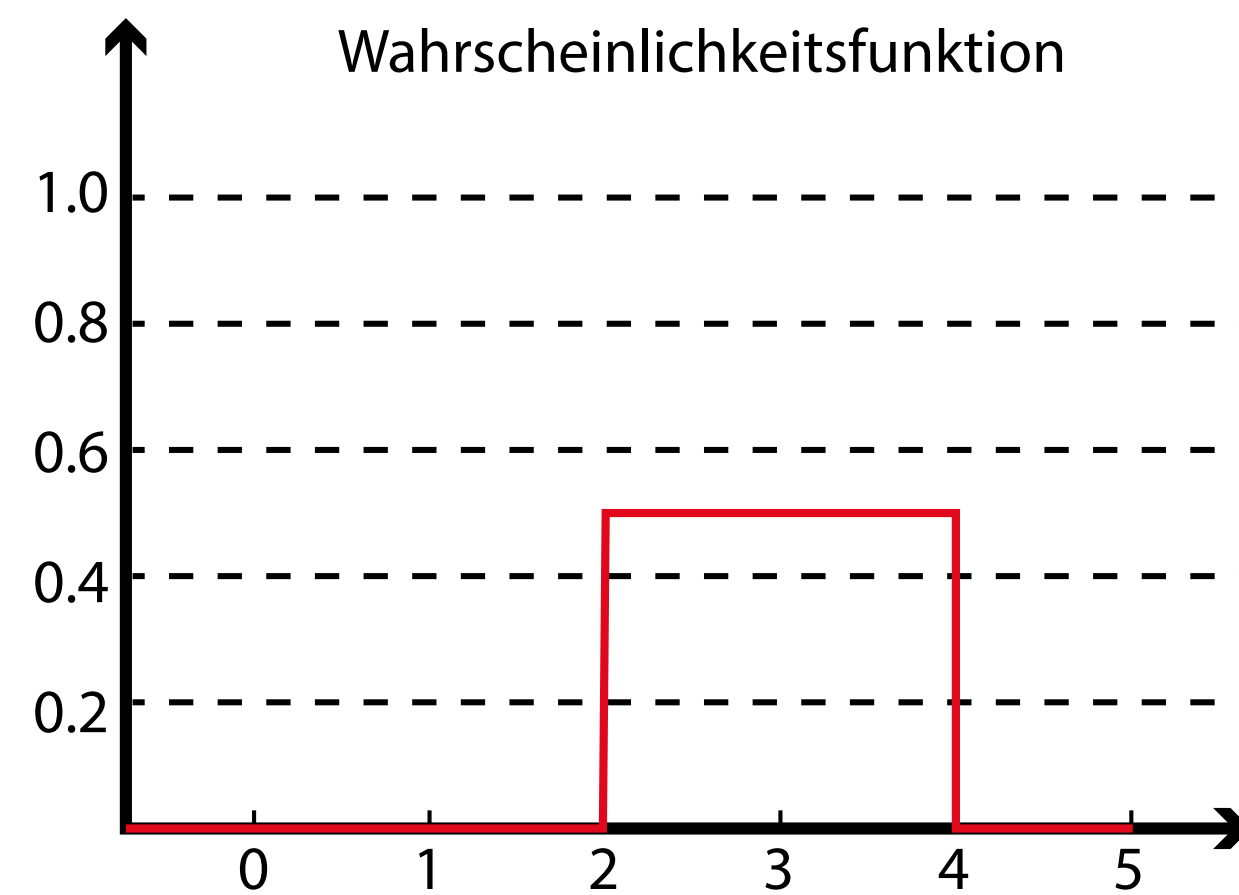
Erwartungswert



$$E(X) = \sum_{x \in E} x \cdot f(x) = 0 \cdot \frac{4}{10} + 1 \cdot \frac{3}{10} + 2 \cdot \frac{2}{10} + 5 \cdot \frac{1}{10} = 1.20\text{€}$$



Erwartungswert



$$E(X) = \int_{-\infty}^{\infty} x \cdot \varphi(x) dx = \int_2^3 x dx = \left[0.5x^2 \right]_2^3 = 0.5 \cdot 3^2 - 0.5 \cdot 2^2 = 2.5$$

$$E(Y) = \int_{-\infty}^{\infty} y \cdot \varphi(y) dy = \int_2^4 0.5y dy = \left[0.25y^2 \right]_2^4 = 3 \quad E(X+Y) = 5.5$$

```
> runif(1,2,3)
```

$X = \dots$

```
> runif(1,2,4)
```

$Y = \dots$

Varianz

Eine weitere wichtige Eigenschaft von Verteilungen ist ihre Varianz σ_x^2 bzw. $\text{Var}(x)$. Diese lässt sich ...

...für diskrete Verteilungen aus der Wahrscheinlichkeitsfunktion berechnen.

... für kontinuierliche Verteilungen aus der Dichtefunktion berechnen.

σ_x^2

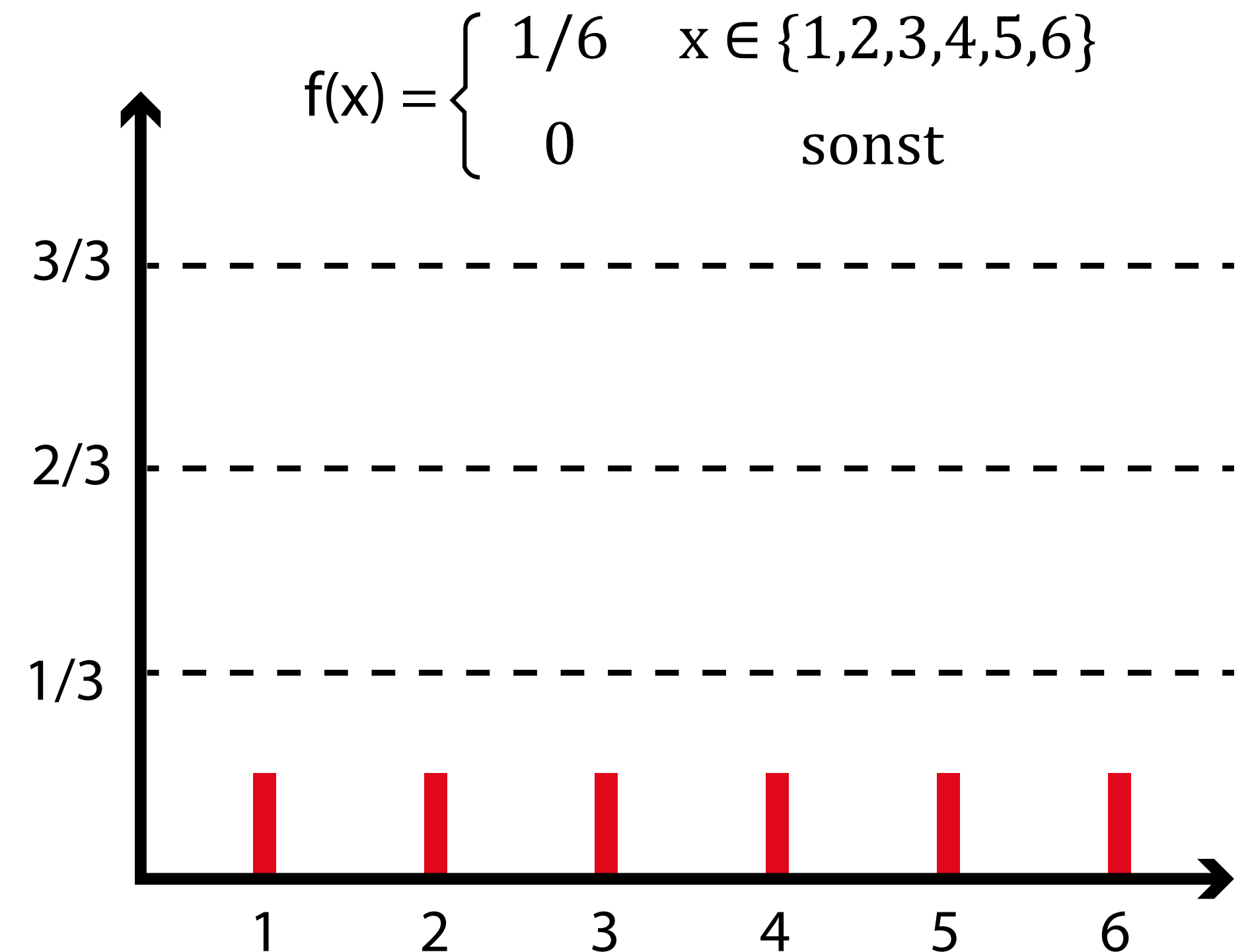
Varianz

Beispiel Würfel: Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} 1/6 & x \in \{1,2,3,4,5,6\} \\ 0 & \text{sonst} \end{cases}$$

Wir berechnen die Varianz über die Summe:

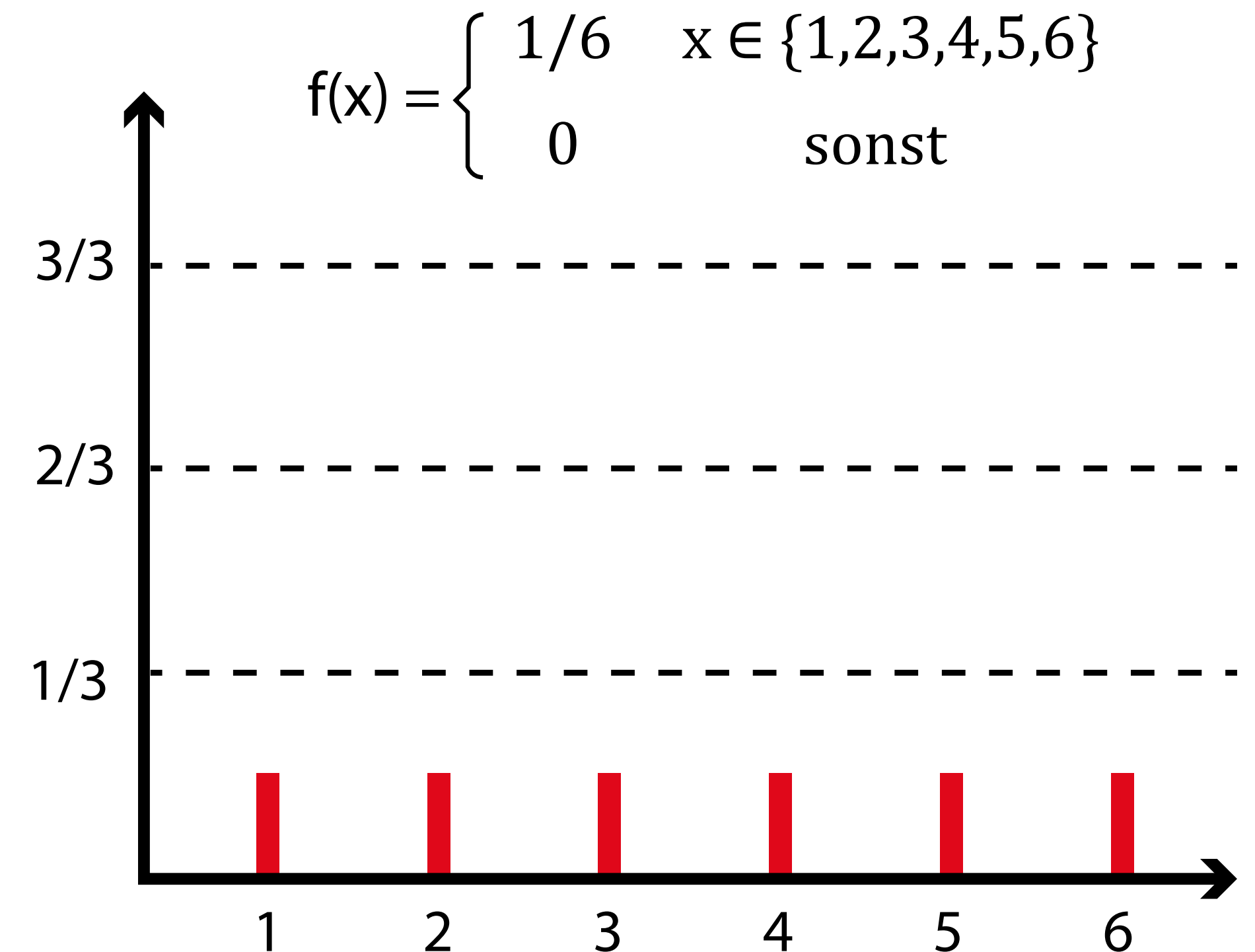
$$\text{Var}(X) = \sum_{x \in E} [x - E(X)]^2 \cdot f(x)$$



Varianz

Wir berechnen die Varianz über die Summe:

$$\begin{aligned}\text{Var}(X) &= \sum_{x \in E} [x - E(X)]^2 \cdot f(x) \\ &= (-2.5)^2 \frac{1}{6} + (-1.5)^2 \frac{1}{6} + (-0.5)^2 \frac{1}{6} \\ &\quad + 0.5^2 \frac{1}{6} + 1.5^2 \frac{1}{6} + 2.5^2 \frac{1}{6} = 2.91\end{aligned}$$



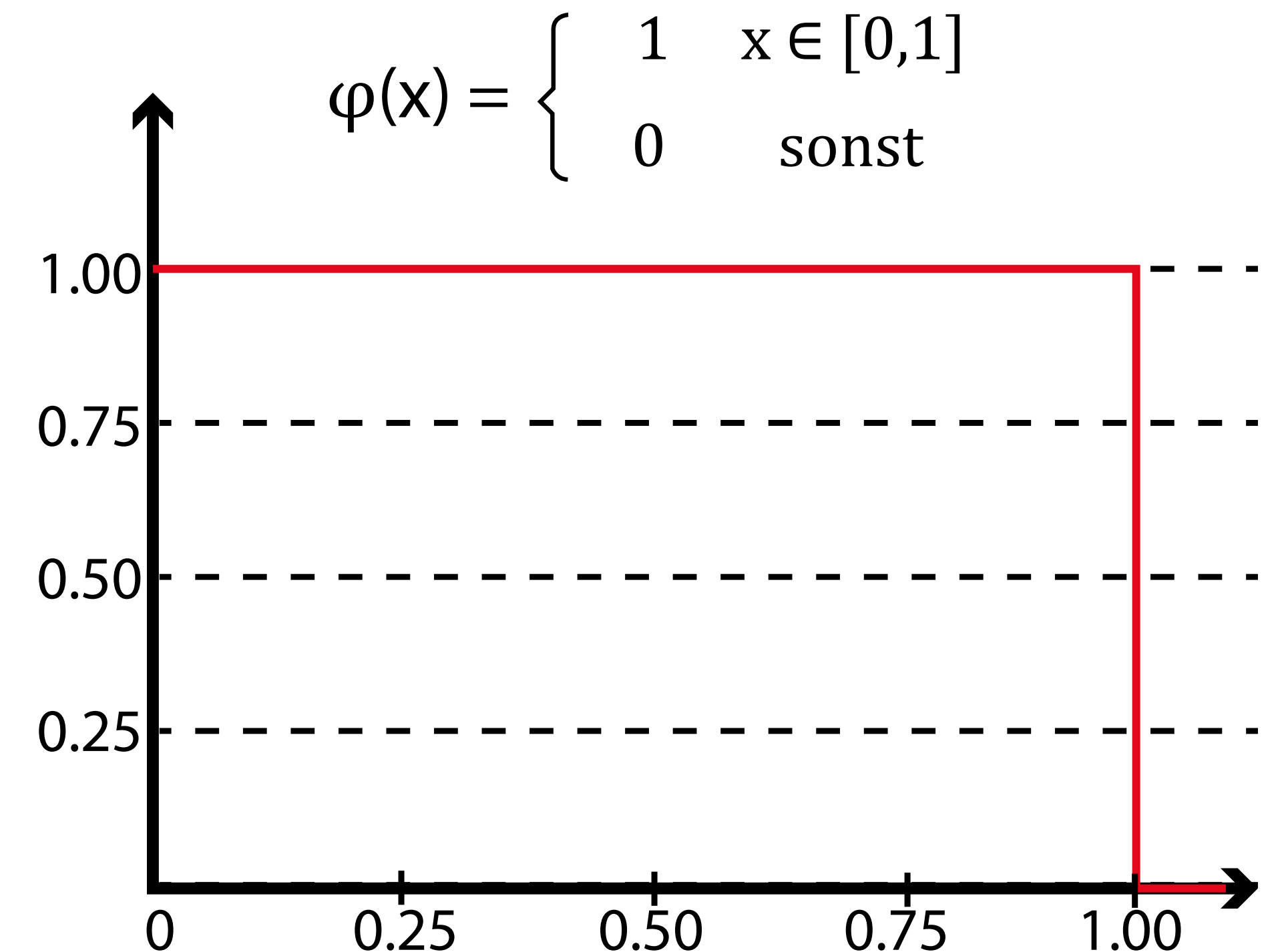
Varianz

Beispiel Zufallsgenerator „0 bis 1“

$$\varphi(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{sonst} \end{cases}$$

Wir berechnen die Varianz über das Integral:

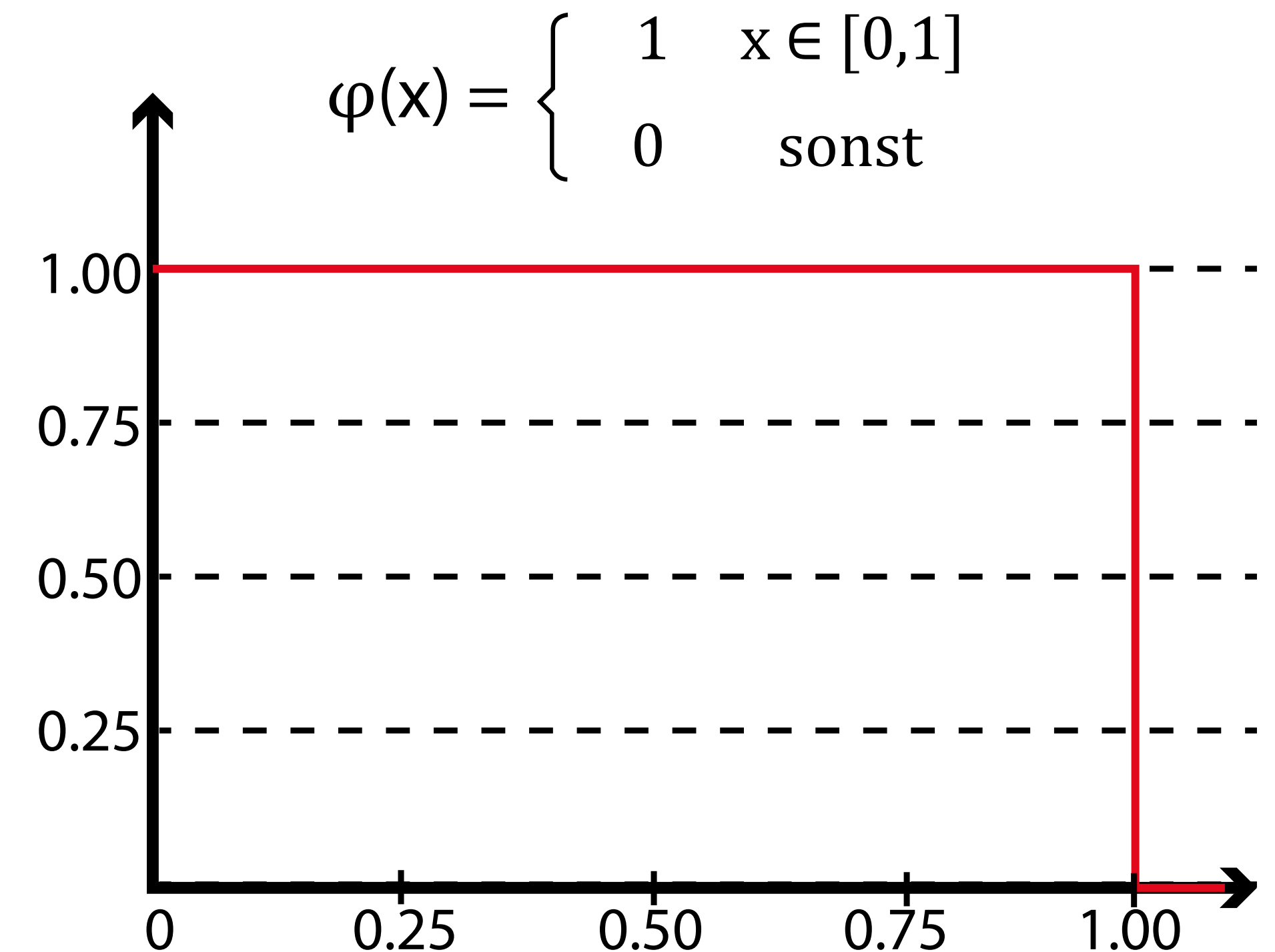
$$\text{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot \varphi(x) \, dx$$



Varianz

Wir berechnen die Varianz über das Integral:

$$\begin{aligned}
 \text{Var}(X) &= \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot \varphi(x) \, dx = \int_0^1 (x-0.5)^2 \, dx \\
 &= \int_0^1 x^2 - x + 0.25 \, dx = \left[\frac{1}{3} x^3 - 0.5x^2 + 0.25x \right]_0^1 \\
 &= 0.08333
 \end{aligned}$$



Varianz

Die Varianz ist nur dann linear additiv, wenn die Zufallsvariablen unabhängig sind. Allgemein gilt:

$$\text{Var}(aX+bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X,Y)$$

Wie in der deskriptiven Statistik können wir auch bei Verteilungen eine Standardabweichung berechnen:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

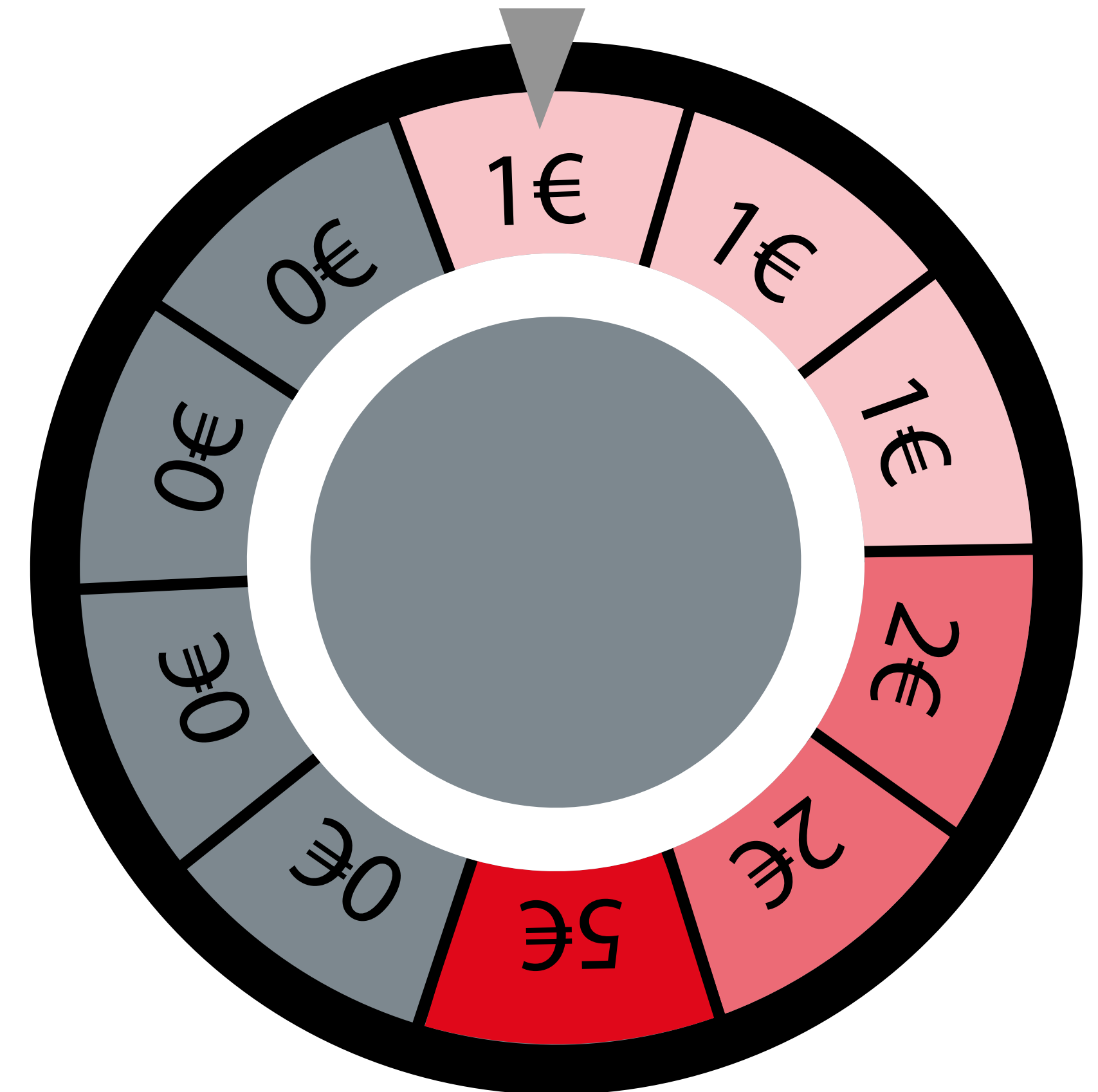
σ_X^2

Varianz

Wir betrachten das rechts gezeigte Glücksrad und die Zufallsvariable X , die die erspielte Auszahlung angibt.

Den Erwartungswert kennen wir bereits: 1.20€.

Berechne nun auch die Varianz und die Standardabweichung der Auszahlung!

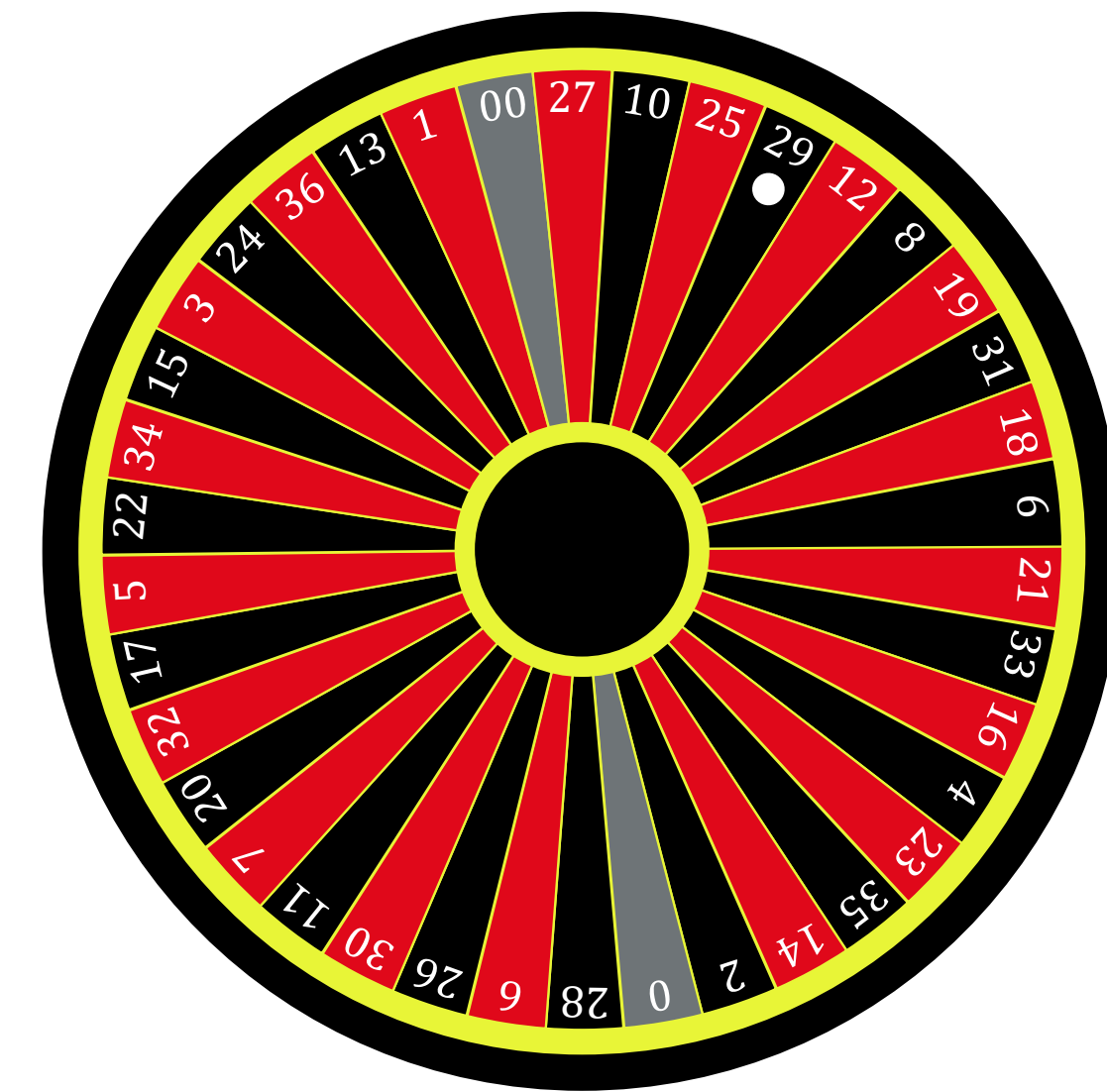


Varianz

Wir betrachten ein amerikanisches Roulettespiel bei dem 10€ auf die Farbe schwarz setzen.

Die Zufallsvariable X beschreibt unsere GuV bei diesem Spiel: Zu 47.36% gewinnen wir 10€ dazu, zu 52.64% verlieren wir 10€.

Berechne Erwartungswert, Varianz und Standardabweichung von X .

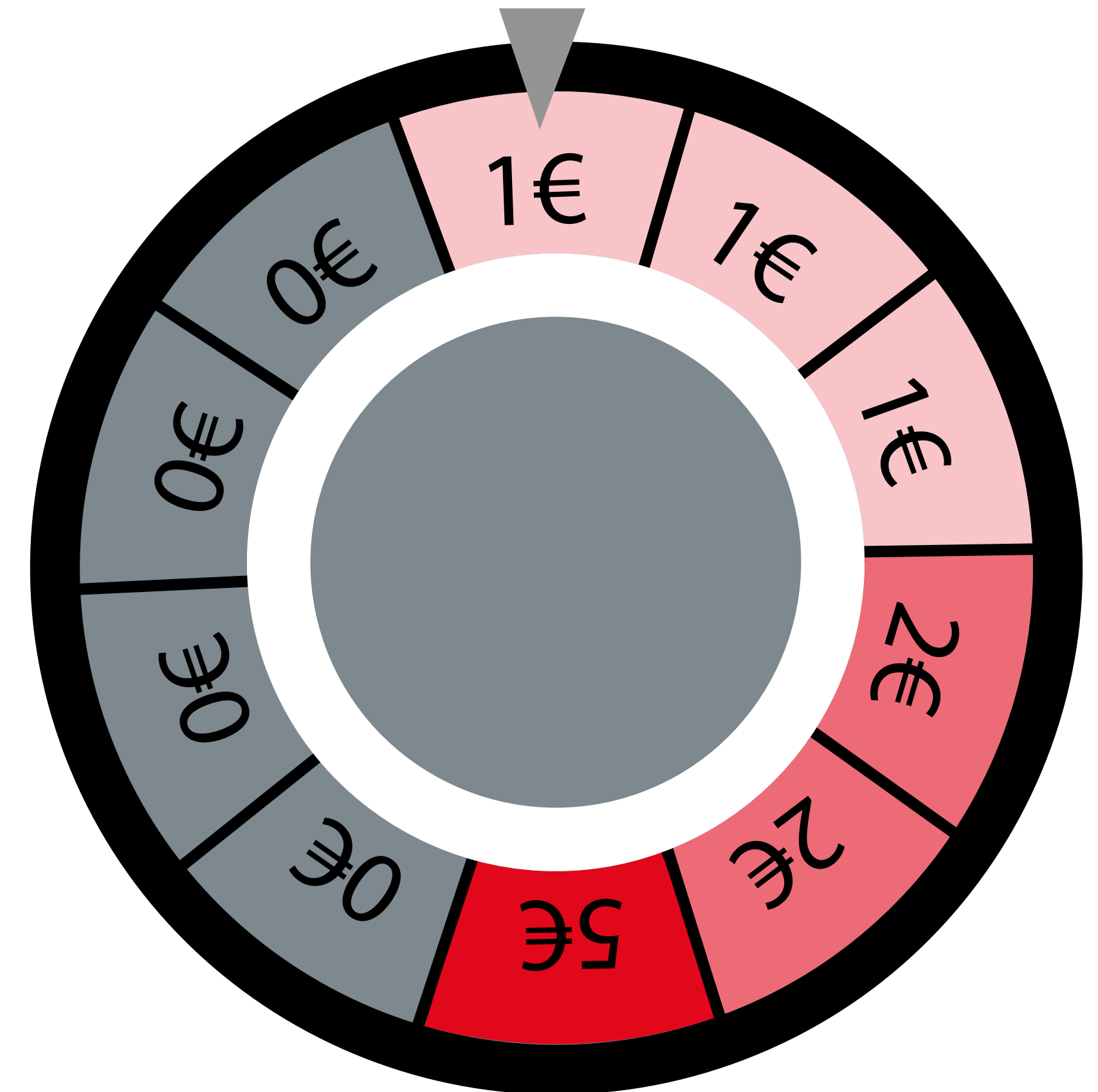


00	0	$\Omega = \{00, 0, 1, 2, \dots, 36\}$									
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

$$E(X) = \sum_{x \in E} x \cdot f(x) = 0 \frac{4}{10} + 1 \frac{3}{10} + 2 \frac{2}{10} + 5 \frac{1}{10} = 1.20$$

$$\begin{aligned} \text{Var}(X) &= \sum_{x \in E} [x - E(X)]^2 \cdot f(x) \\ &= (-1.2)^2 \frac{4}{10} + (-0.2)^2 \frac{3}{10} + 0.8^2 \frac{2}{10} + 3.8^2 \frac{1}{10} \\ &= 2.16 \end{aligned}$$

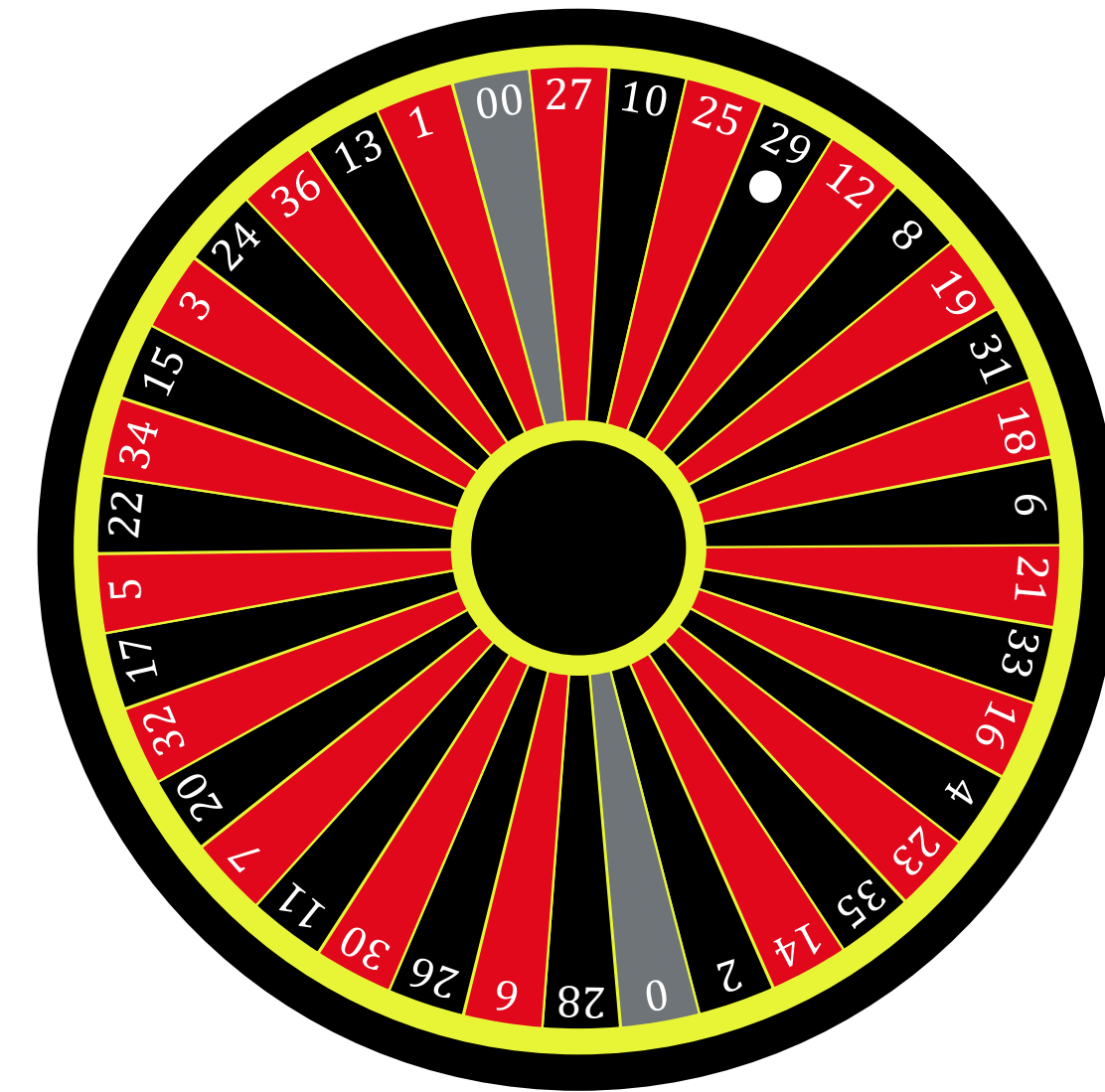
$$\sigma_X = \sqrt{\text{Var}(X)} = 1.47$$



$$E(X) = \sum_{x \in E} x \cdot f(x) = 0.4736 \cdot 10 - 0.5264 \cdot 10 = -0.528$$

$$\begin{aligned} \text{Var}(X) &= \sum_{x \in E} [x - E(X)]^2 \cdot f(x) \\ &= (10.528)^2 \cdot 0.4736 + (-9.472)^2 \cdot 0.5264 \\ &= 99.72 \end{aligned}$$

$$\sigma_x = \sqrt{\text{Var}(X)} = 9.985$$



00	0	$\Omega = \{00, 0, 1, 2, \dots, 36\}$									
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

Momente

Der Erwartungswert ist das wichtigste Moment einer Zufallsvariable. Allgemein sind die Momente definiert als:

$$m_k = E(x^k) = \sum_{x \in E} x^k \cdot f(x) \quad m_k = \int_{-\infty}^{\infty} x^k \cdot \varphi(x) dx$$

Die Varianz ist das wichtigste zentrale Moment einer Zufallsvariable. Allgemein sind diese definiert als:

$$\mu_k = \sum_{x \in E} (x-\mu)^k \cdot f(x) \quad \mu_k = \int_{-\infty}^{\infty} (x-\mu)^k \cdot \varphi(x) dx$$

	Moment	Zentrales Moment
1.	Erwartungswert	0
2.	m_2	Varianz
3.	m_3	(Schiefe)
4.	m_4	(Wölbung)

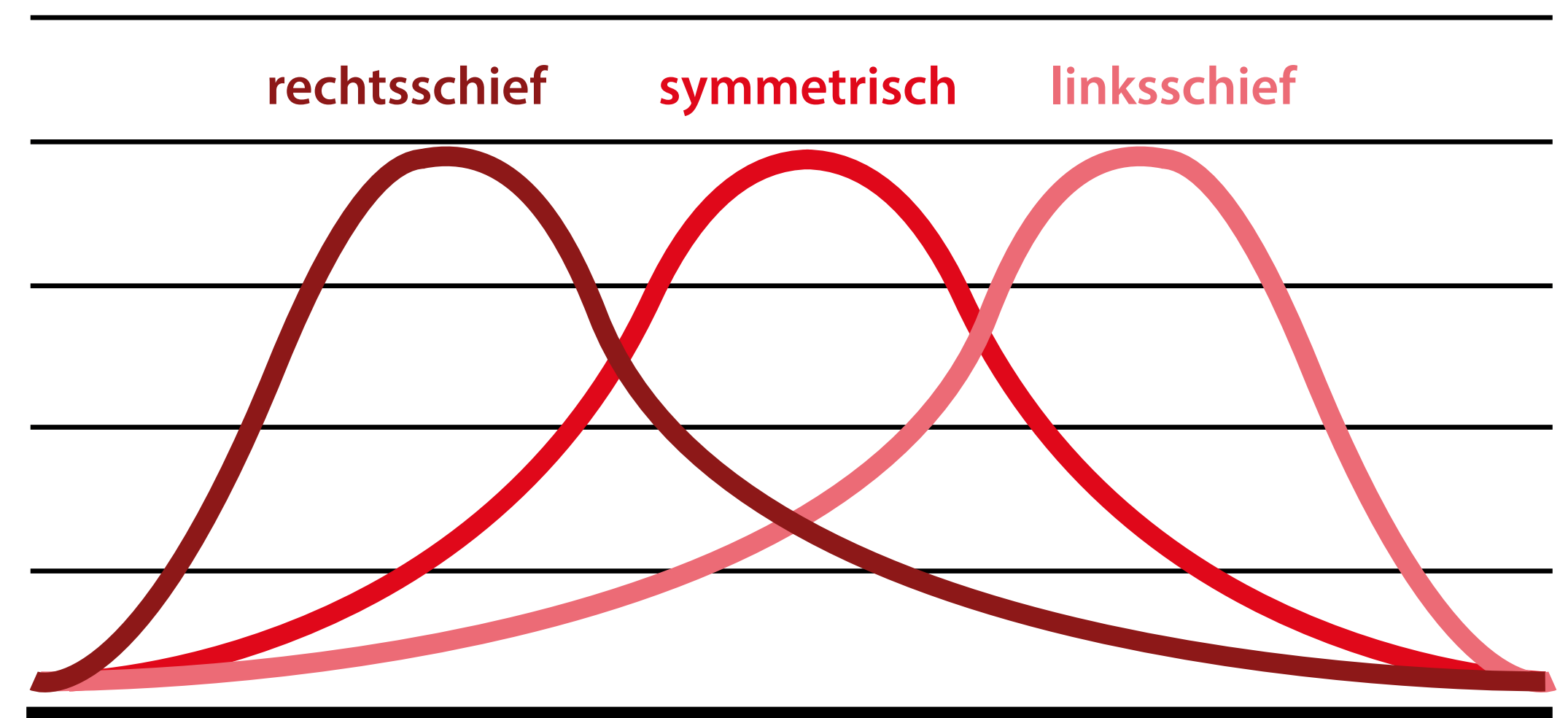
Momente

Die **Schiefe** wird aus dem dritten zentralen Moment berechnet. Die gängigste Variante ist:

$$\gamma_m = \frac{\mu_3}{\sigma^3}$$

Unabhängig von der Normierung gilt:

- Verteilungen mit $\gamma = 0$ sind symmetrisch.
- Verteilungen mit $\gamma > 0$ sind rechtsschief.
- Verteilungen mit $\gamma < 0$ sind linksschief.



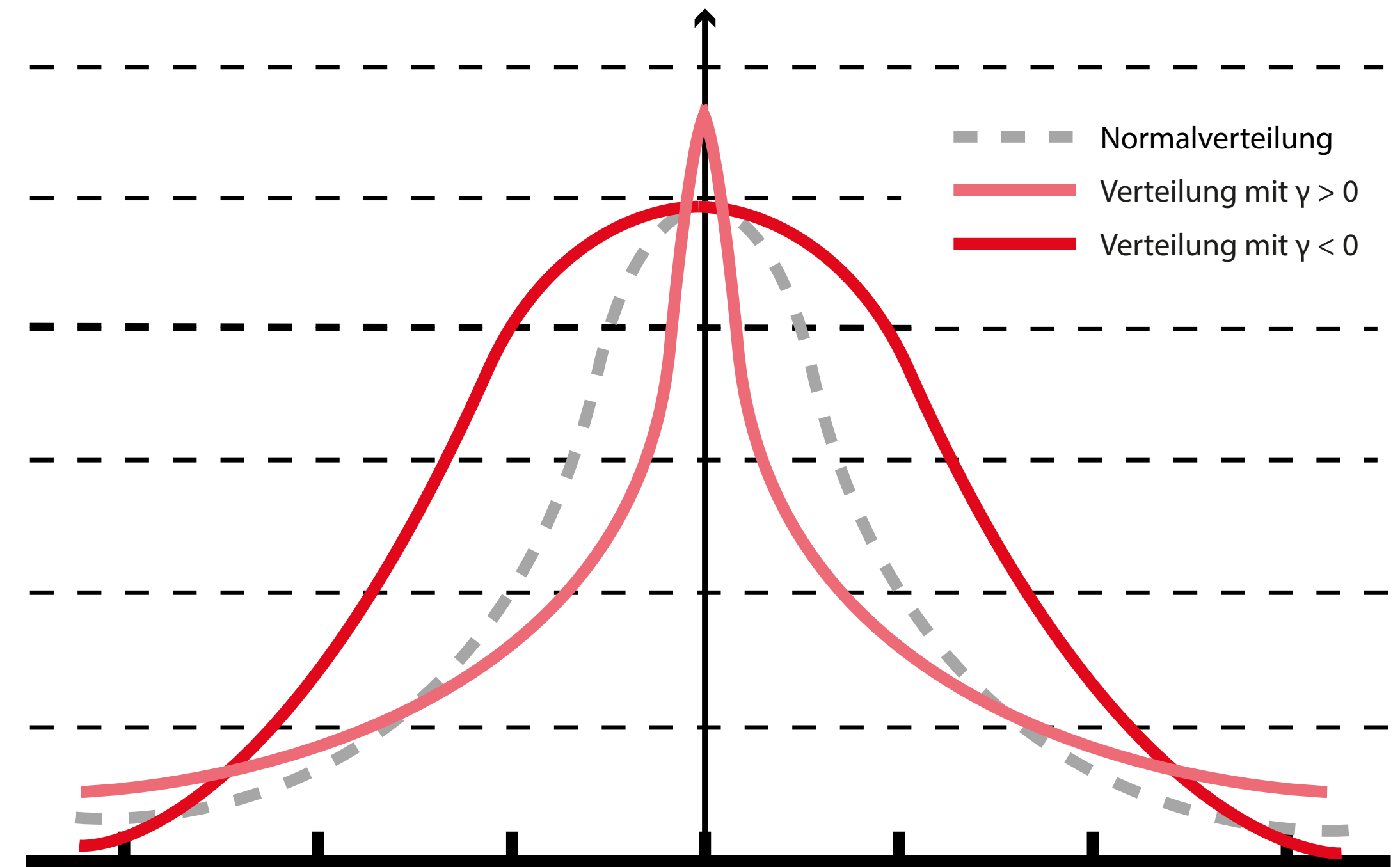
Momente

Die **Wölbung** wird aus dem vierten zentralen Moment berechnet. Die gängigste Variante ist der „Exzess“:

$$\gamma = \frac{\mu_4}{\sigma^4} - 3$$

Durch die -3 wird die „Spitzigkeit“ der Verteilung mit der der Normalverteilung verglichen.

- Verteilungen mit $\gamma > 0$ sind steiler.
- Verteilungen mit $\gamma < 0$ sind breiter.



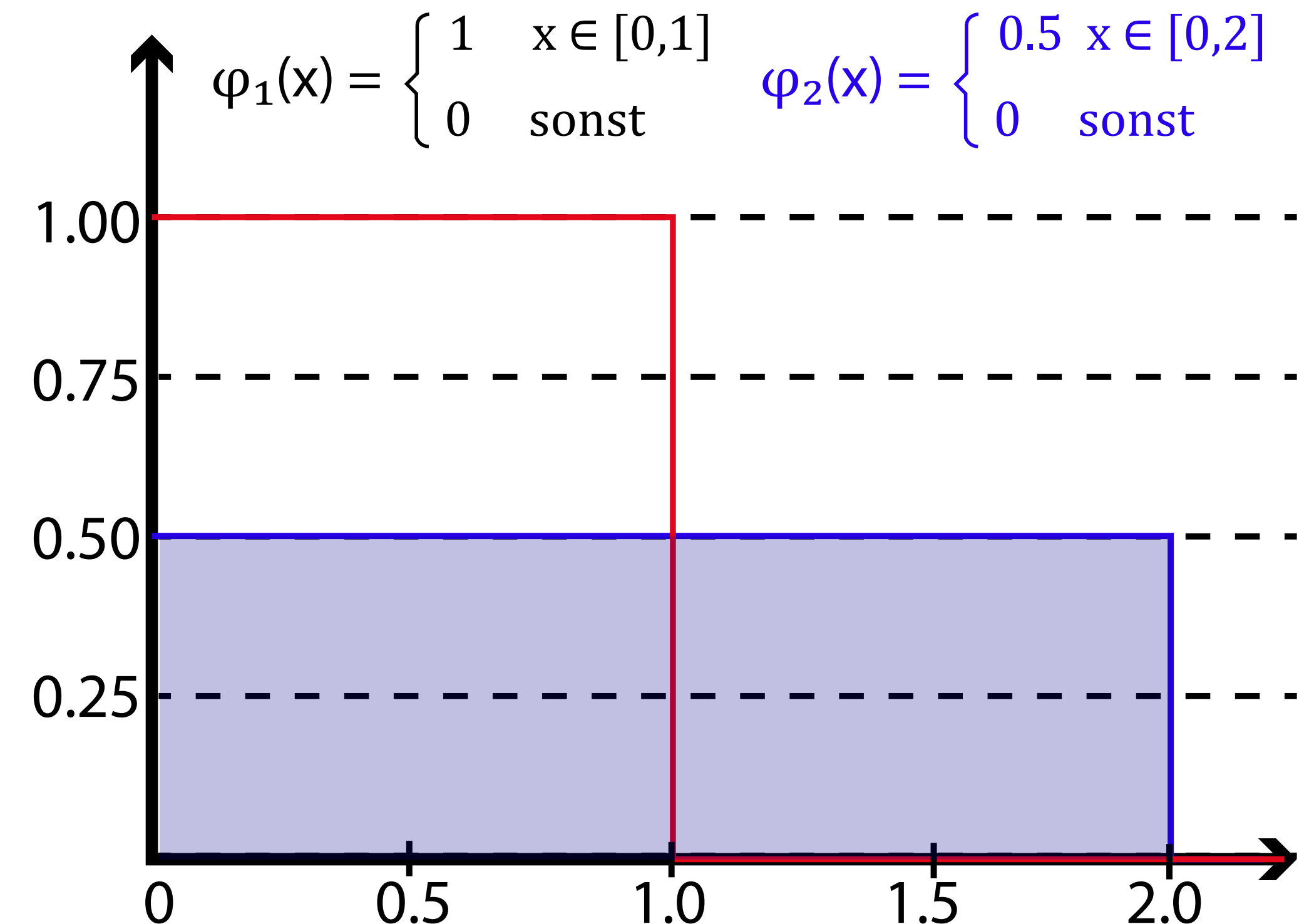
Gleichverteilung

Die Gleichverteilung kennen wir bereits von den Überlegungen zu Erwartungswert und Varianz.

In der kontinuierlichen Variante besitzt sie zwei Parameter: die Untergrenze a und die Obergrenze b .

Erwartungswert und Varianz sind:

$$\mu = \frac{a+b}{2} \quad \sigma^2 = \frac{1}{12} (b-a)^2$$



Gleichverteilung

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot \varphi(x) \, dx = \int_a^b \frac{x}{b-a} \, dx \\ &= \left[\frac{0.5x^2}{b-a} \right]_a^b = \frac{0.5(b^2 - a^2)}{b-a} = 0.5(a + b) \end{aligned}$$

$$\begin{aligned}
 \text{Var}(X) &= \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot \varphi(x) \, dx = \int_a^b (x - 0.5(a+b))^2 \frac{1}{b-a} \, dx \\
 &= \frac{1}{b-a} \int_a^b x^2 - x(a+b) + 0.25(a+b)^2 \, dx \\
 &= \frac{1}{b-a} \left[\frac{x^3}{3} - \frac{x^2(a+b)}{2} + \frac{x(a+b)^2}{4} \right]_a^b \\
 &= \frac{1}{b-a} \left(\frac{a^3}{12} + \frac{a^2b - b^2a}{4} + \frac{b^3}{12} \right) = \frac{1}{b-a} \left(\frac{(b-a)^3}{12} \right) = \frac{(b-a)^2}{12}
 \end{aligned}$$

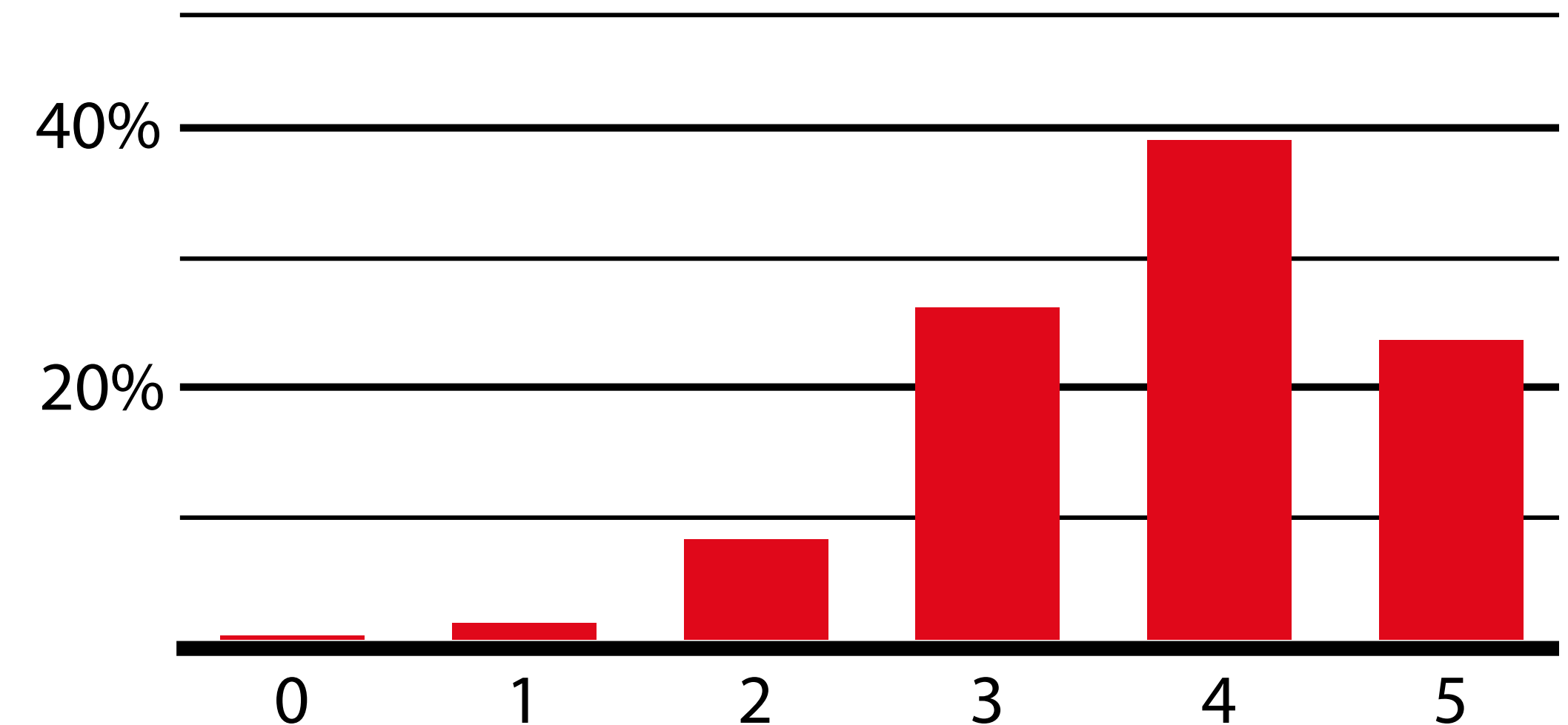
Binomialverteilung

Die Binomialverteilung beschreibt die Wahrscheinlichkeit für k aus n erfolgreichen Versuchen, wenn die Erfolgswahrscheinlichkeit eines einzelnen Versuches p ist.

Wir definieren Sie über ihre Wahrscheinlichkeitsfunktion:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

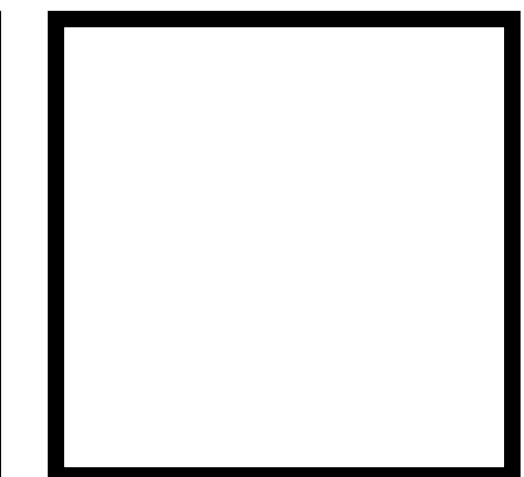
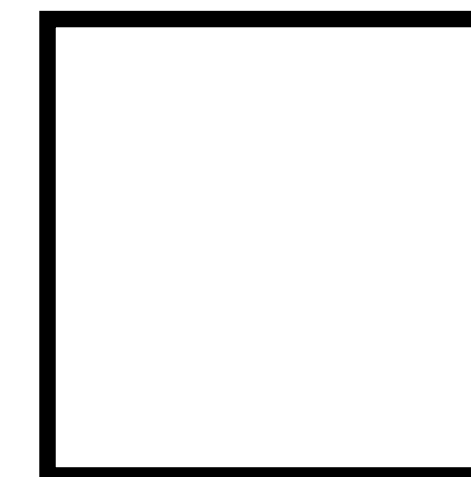
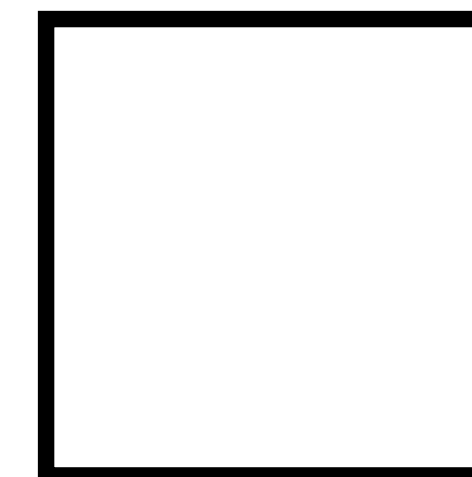
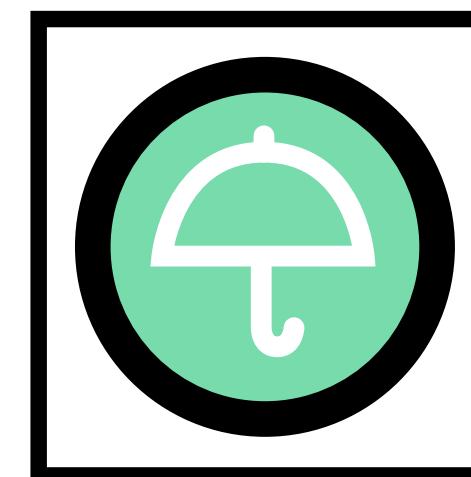
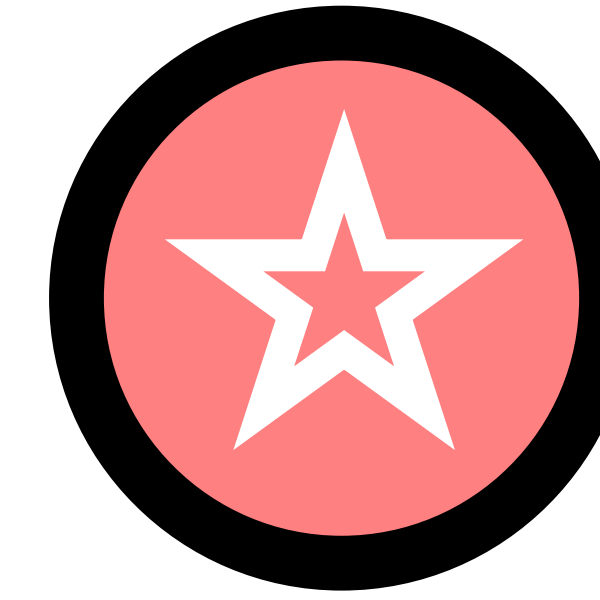
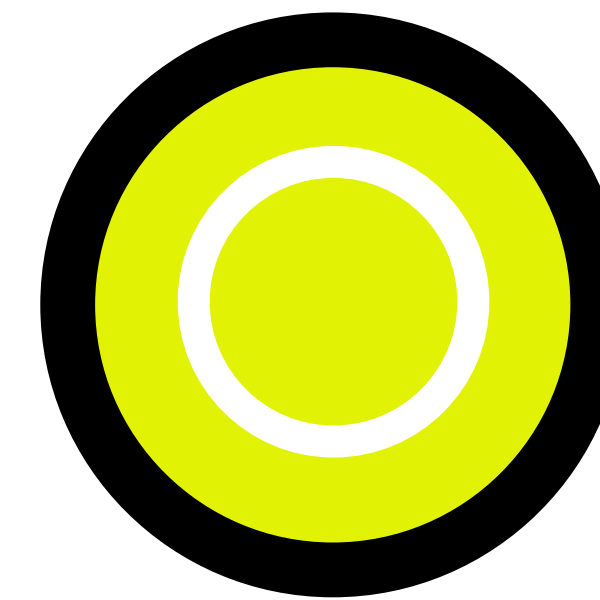
Wahrscheinlichkeit für getroffene Elfmeter aus 5 Versuchen bei einer Trefferwahrscheinlichkeit von 75%



Binomialverteilung

In dieser Wahrscheinlichkeitsfunktion wird die **Fakultät** verwendet. Diese berechnet die Anzahl Möglichkeiten, n verschiedene Dinge anzuordnen:

$$n! = \prod_{i=1}^n i = 1 \cdot 2 \cdot \dots \cdot n$$



4 Möglichkeiten

3 Möglichkeiten

2 Möglichkeiten

1 Möglichkeit

Binomialverteilung

Der **Binomialkoeffizient** „n über k“ verwendet die Fakultät, um die Anzahl Möglichkeiten zu berechnen, mit der wir k aus n Elemente auszuwählen können.

$$\binom{n}{k} = \frac{n!}{k! (n-k)!}$$

Beispiel: 3 aus 5 Elfmetern, die getroffen werden:

$$\binom{5}{3} = \frac{5!}{3! (5-3)!} = \frac{120}{6 \cdot 2} = 10$$



Binomialverteilung

Die Wahrscheinlichkeit von 4 Treffern aus 5 Versuchen ist dann zum Beispiel:

$$\begin{aligned}
 P &= \binom{5}{4} P_{\text{Tor}}^4 \cdot (1 - P_{\text{Tor}}) \\
 &= \frac{5!}{4! (5-4)!} 0.75^4 \cdot 0.25 \\
 &= 5 \cdot 0.75^4 \cdot 0.25 \\
 &= 39.6\%
 \end{aligned}$$

Torchance eines Elfmeters: 75%

Gesucht: Chance auf 4 Treffer aus 5 Versuchen

✓ ✓ ✓ ✓ ✗	$P = 0.75^4 \cdot 0.25$	$= 7.91\%$
✓ ✓ ✓ ✗ ✓	$P = 0.75^3 \cdot 0.25 \cdot 0.75$	$= 7.91\%$
✓ ✓ ✗ ✓ ✓	$P = 0.75^2 \cdot 0.25 \cdot 0.75^2$	$= 7.91\%$
✓ ✗ ✓ ✓ ✓	$P = 0.75^1 \cdot 0.25 \cdot 0.75^3$	$= 7.91\%$
✗ ✓ ✓ ✓ ✓	$P = 0.25 \cdot 0.75^4$	$= 7.91\%$
		$P = 39.6\%$

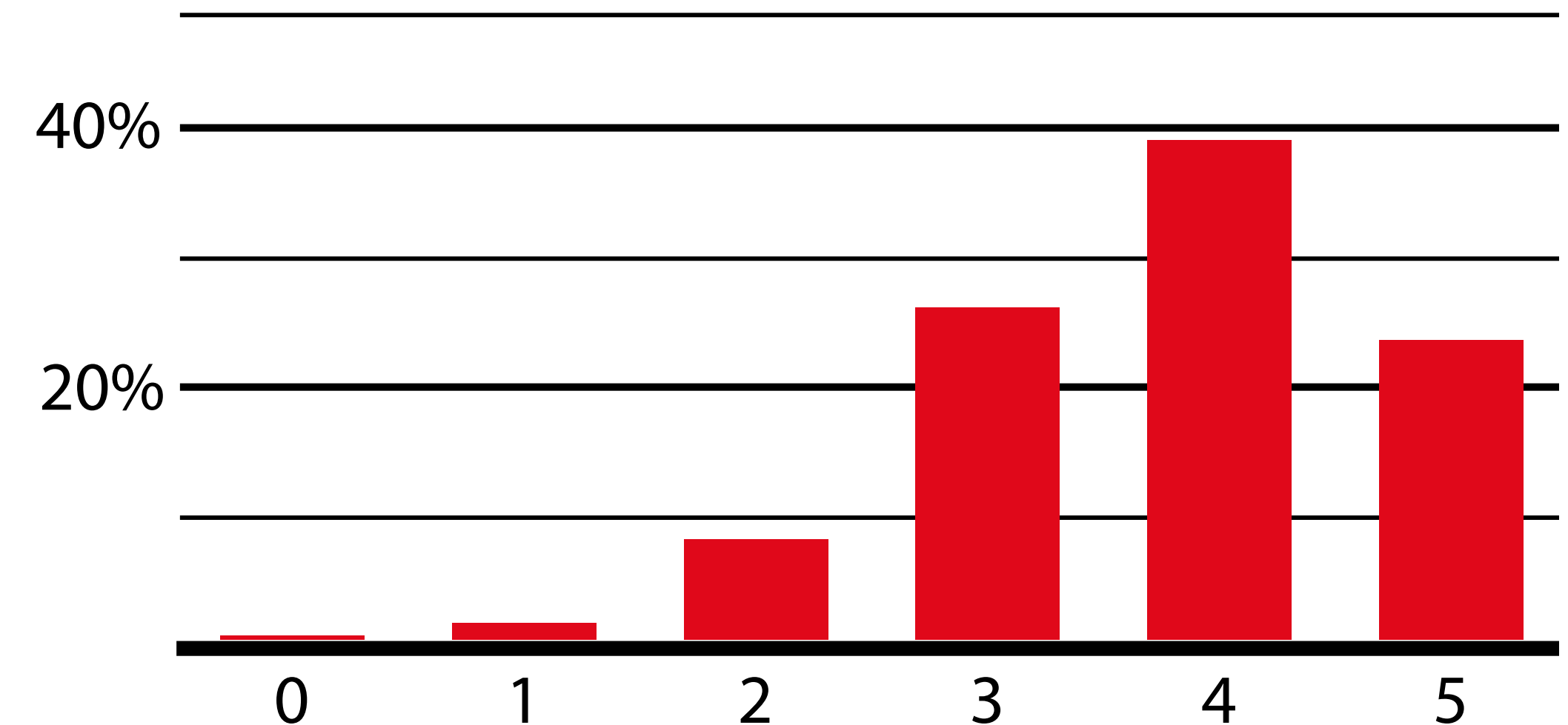
Binomialverteilung

Erwartungswert und Varianz der Binomialverteilung sind einfache und naheliegende Werte, die jedoch schwer herzuleiten sind.

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

Wahrscheinlichkeit für getroffene Elfmeter aus 5 Versuchen bei einer Trefferwahrscheinlichkeit von 75%



Binomialverteilung

$$\begin{aligned}
 \mu &= \sum_{k=0}^n k \cdot \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Summand für } k=0 \text{ ist } 0 \\
 &= \sum_{k=1}^n k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n n \cdot \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
 &= n \sum_{k=1}^n \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad \begin{array}{l} \text{Substitution } j = k-1 \\ \text{Ersetze } k \text{ mit } j+1 \end{array} \\
 &= n \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j+1} (1-p)^{n-j-1} = np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}}_{\text{Ungewichtete Summation über alle Wahrscheinlichkeiten gibt 1}} = np
 \end{aligned}$$

Nebenrechnung

$$\begin{aligned}
 k \binom{n}{k} &= k \frac{n!}{k! (n-k)!} \\
 &= \cancel{k} \frac{n!}{\cancel{k}(k-1)! (n-k)!} \\
 &= n \frac{(n-1)!}{(k-1)! (n-k)!} \\
 &= n \frac{(n-1)!}{(k-1)! ([n-1]-[k-1])!} = n \binom{n-1}{k-1}
 \end{aligned}$$

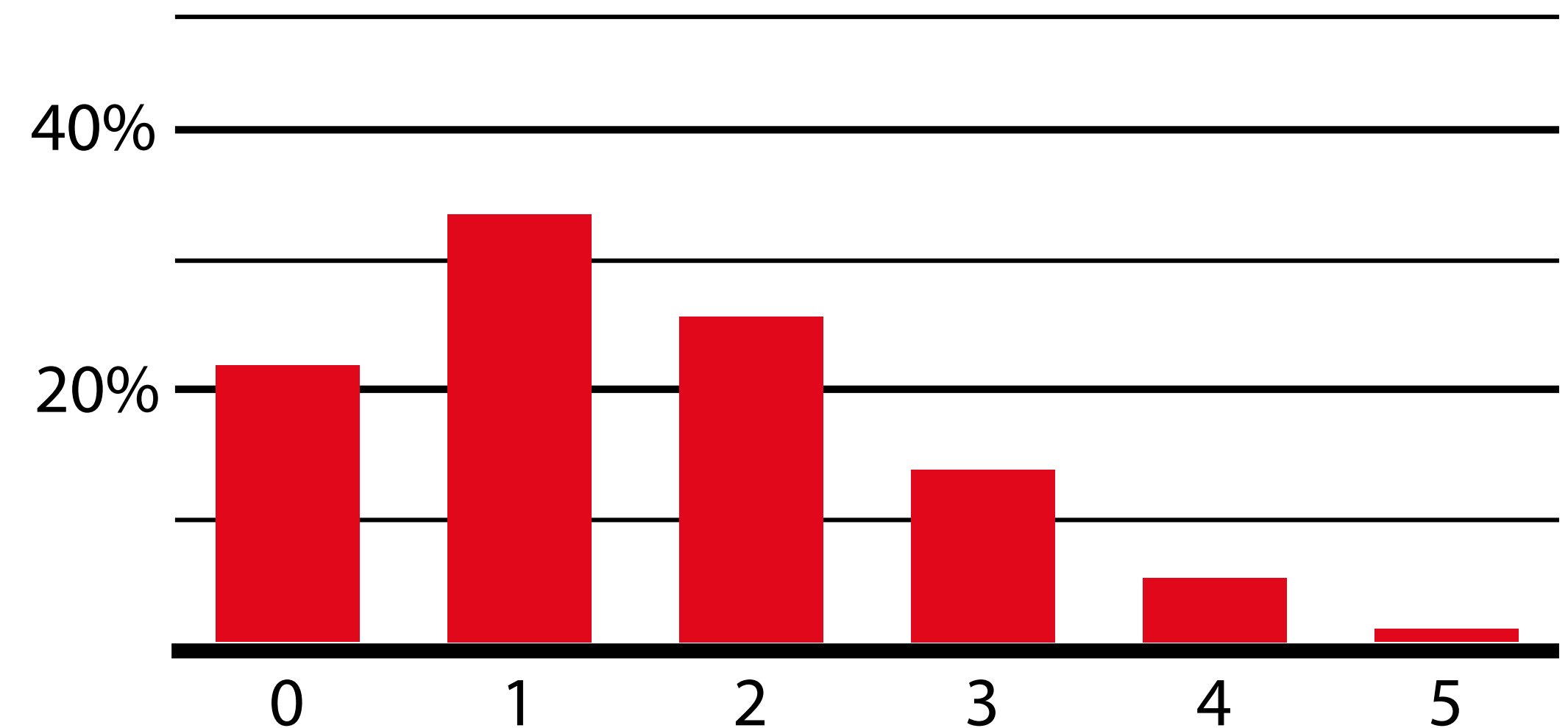
Poisson-Verteilung

Die Poisson-Verteilung beschreibt die Wahrscheinlichkeit von k Ereignissen pro Zeiteinheit, wenn die durchschnittliche Auftrittshäufigkeit λ pro Zeiteinheit beträgt.

Wir definieren Sie über ihre Wahrscheinlichkeitsfunktion:

$$f(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Wahrscheinlichkeit für n IT-Probleme in einer Stunde wenn es an der DHBW RV im Schnitt zu 1.5 IT-Problemen pro Stunde kommt

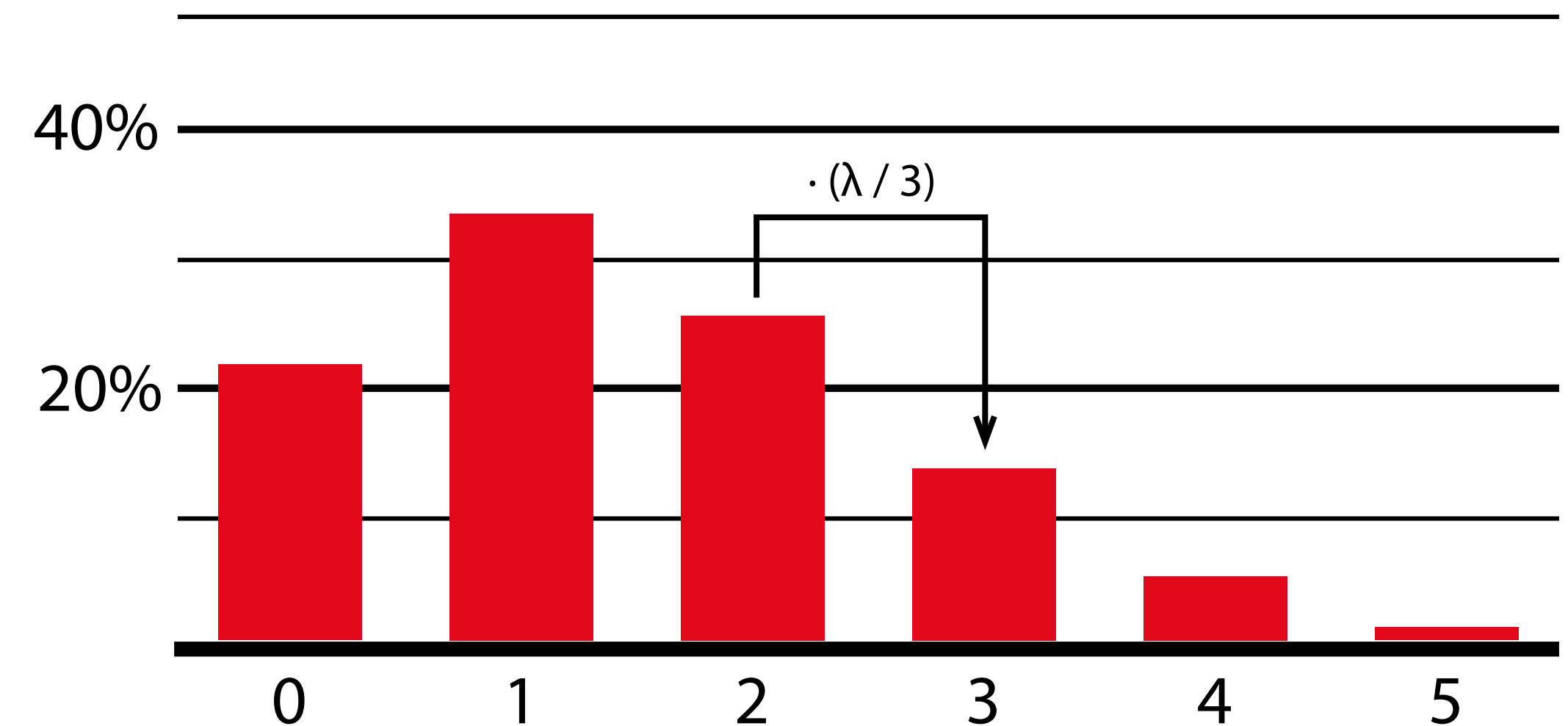


Poisson-Verteilung

Durch die Wahrscheinlichkeitsfunktion hat sie die folgende kuriose Eigenschaft, die bei numerischen Berechnungen Rechenzeit sparen kann.

$$f(k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad f(k+1) = f(k) \frac{\lambda}{k+1}$$

Wahrscheinlichkeit für n IT-Probleme in einer Stunde wenn es an der DHBW RV im Schnitt zu 1.5 IT-Problemen pro Stunde kommt



Poisson-Verteilung

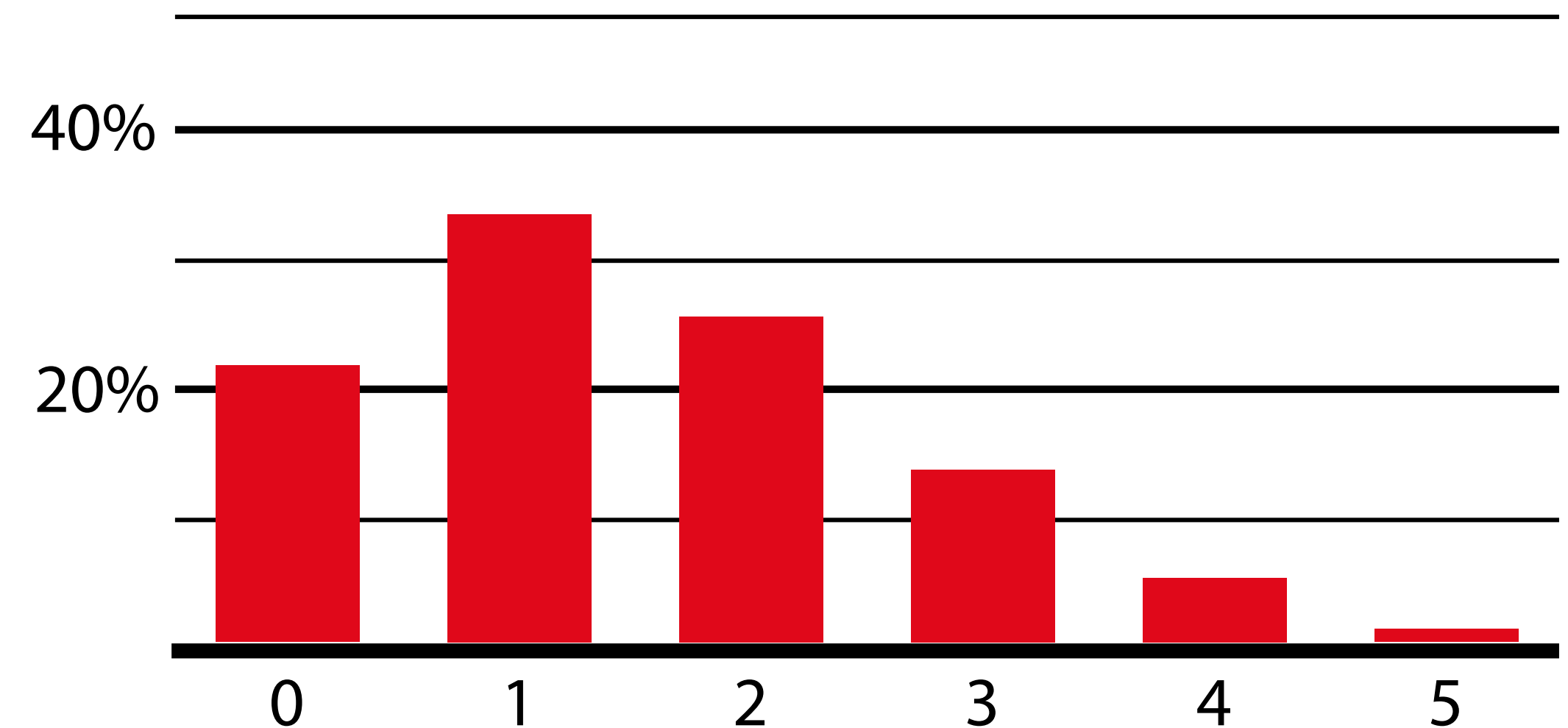
Die Poisson-Verteilung ist mit der Binomialverteilung verwandt. Beide Verteilungen beschreiben die Wahrscheinlichkeiten einer bestimmten Anzahl an "Erfolgen".

Anstelle eines Limits an Versuchen gibt es einen begrenzten Zeitraum, in der die "Erfolge" eintreten sollen.

Mathematisch entspricht das der Grenzwertbetrachtung:

$$\lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{mit } p = \frac{\lambda}{n}$$

Wahrscheinlichkeit für n IT-Probleme in einer Stunde wenn es an der DHBW RV im Schnitt zu 1.5 IT-Problemen pro Stunde kommt



$$\begin{aligned}
 \lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} &= \lim_{n \rightarrow \infty} \frac{n!}{k! (n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &= \left(\frac{\lambda^k}{k!}\right) \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \left(\frac{1}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &\quad \text{KÜRZEN!} \\
 &= \left(\frac{\lambda^k}{k!}\right) \lim_{n \rightarrow \infty} n(n-1)(n-2) \cdots (n-k+1) \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &\quad \text{k Faktoren im Zähler \& Nenner} \\
 &= \left(\frac{\lambda^k}{k!}\right) \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2) \cdots (n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k}
 \end{aligned}$$

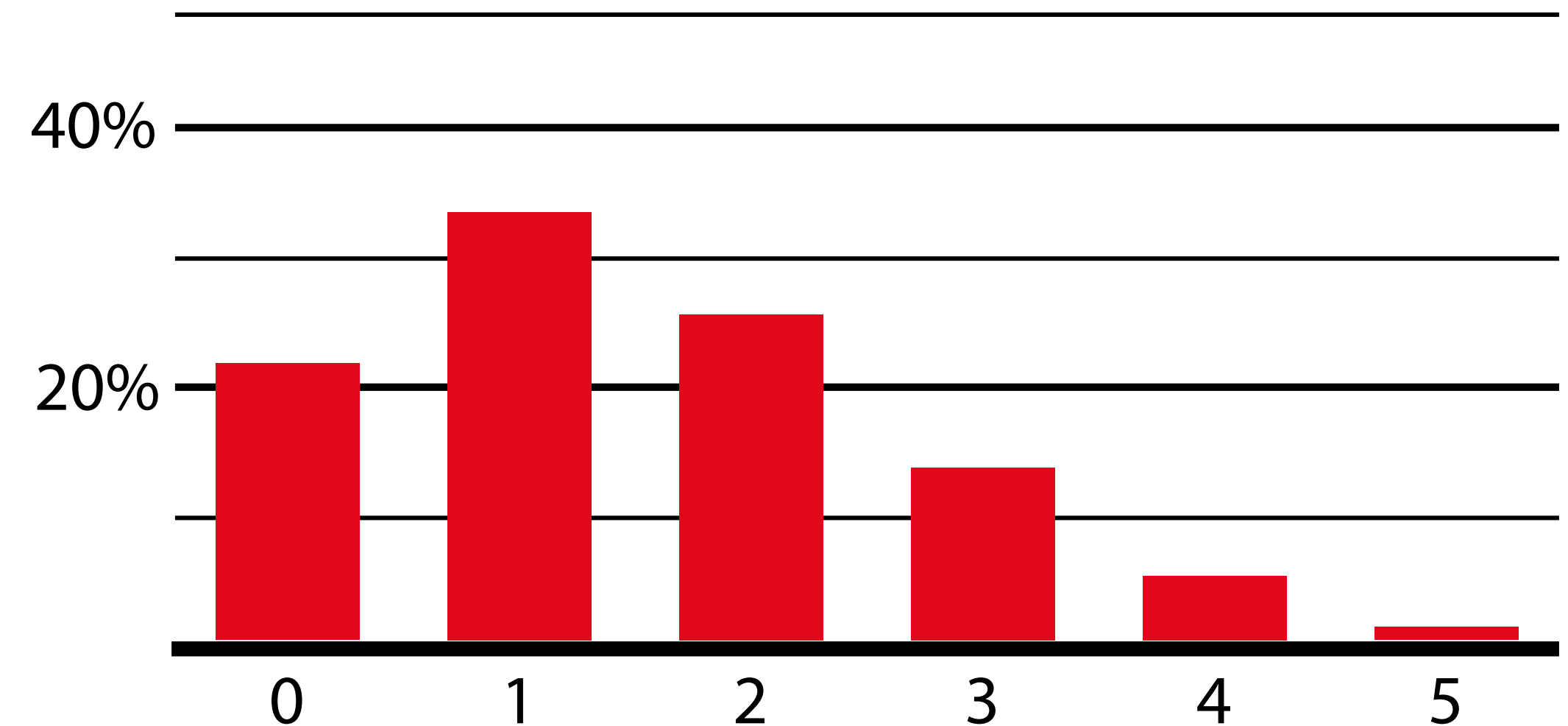
$$\begin{aligned}
 &= \left(\frac{\lambda^k}{k!} \right) \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2) \cdots (n-k+1)}{n^k} \left(1 - \frac{\lambda}{n} \right)^{n-k} \\
 &\quad \text{k Faktoren im Zähler \& Nenner} \\
 &= \left(\frac{\lambda^k}{k!} \right) \lim_{n \rightarrow \infty} \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n} \right)^{n-k} \quad \text{Splitten!} \\
 &\quad \quad \quad a^{b+c} = a^b \cdot a^c \\
 &= \left(\frac{\lambda^k}{k!} \right) \lim_{n \rightarrow \infty} \underbrace{\frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \cdots \frac{n-k+1}{n}}_1 \underbrace{\left(1 - \frac{\lambda}{n} \right)^n}_{e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n} \right)^{-k}}_1 \\
 &= \left(\frac{\lambda^k}{k!} \right) e^{-\lambda}
 \end{aligned}$$

Poisson-Verteilung

Sowohl Erwartungswert als auch Varianz der Poisson-Verteilung sind λ .

$$\mu = \sum_{k=0}^n k \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \lambda \quad \sigma^2 = \lambda$$

Wahrscheinlichkeit für n IT-Probleme in einer Stunde wenn es an der DHBW RV im Schnitt zu 1.5 IT-Problemen pro Stunde kommt

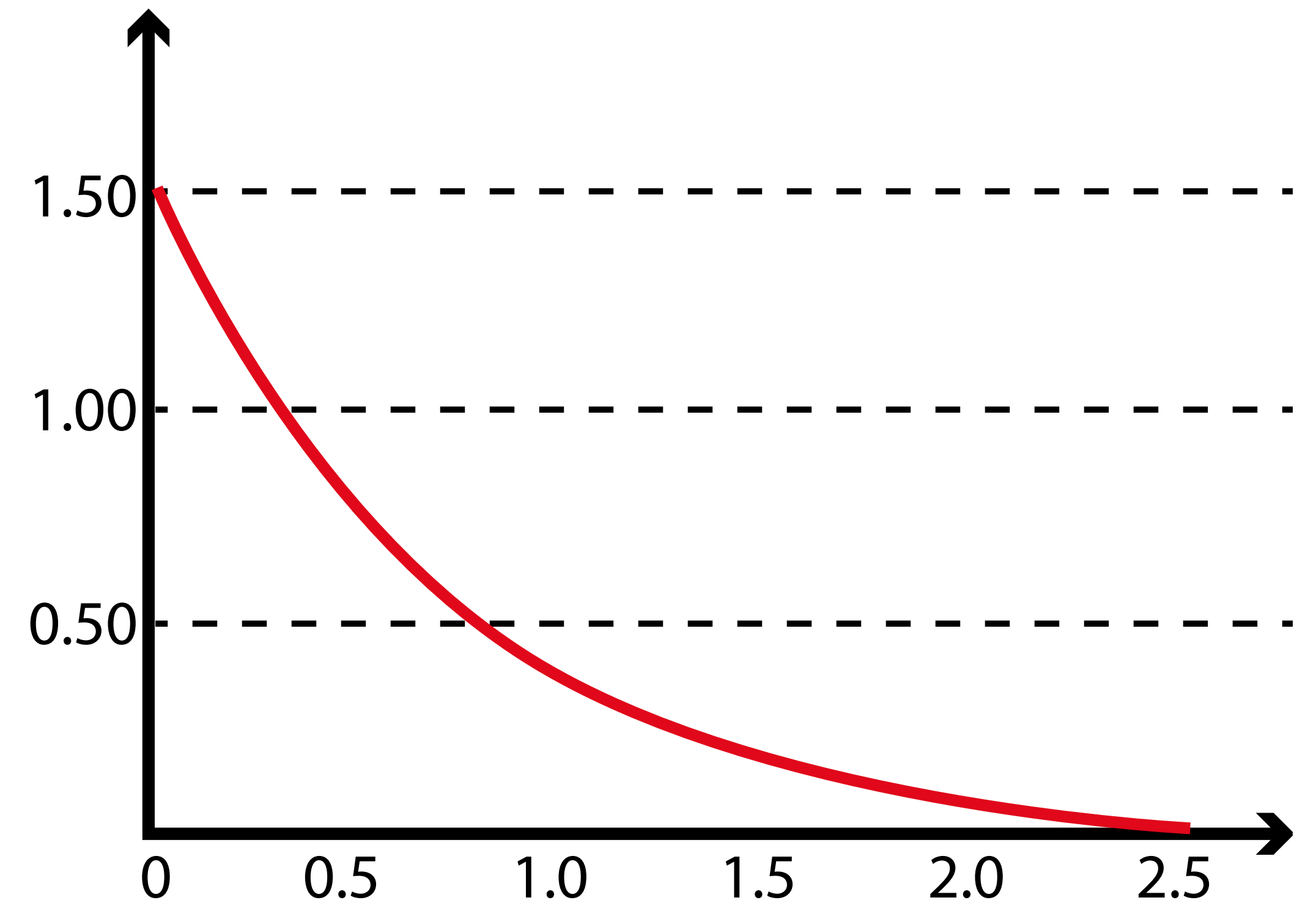


Exponentialverteilung

Die Exponentialverteilung beschreibt die Wahrscheinlichkeit von einer Pause zwischen zwei Ereignissen, wenn die durchschnittliche Auftrittshäufigkeit eines Ereignisses λ pro Zeiteinheit ist.

Sie ist eng verwandt mit der Poisson-Verteilung und wird über ihre Dichtefunktion definiert:

$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



Exponentialverteilung

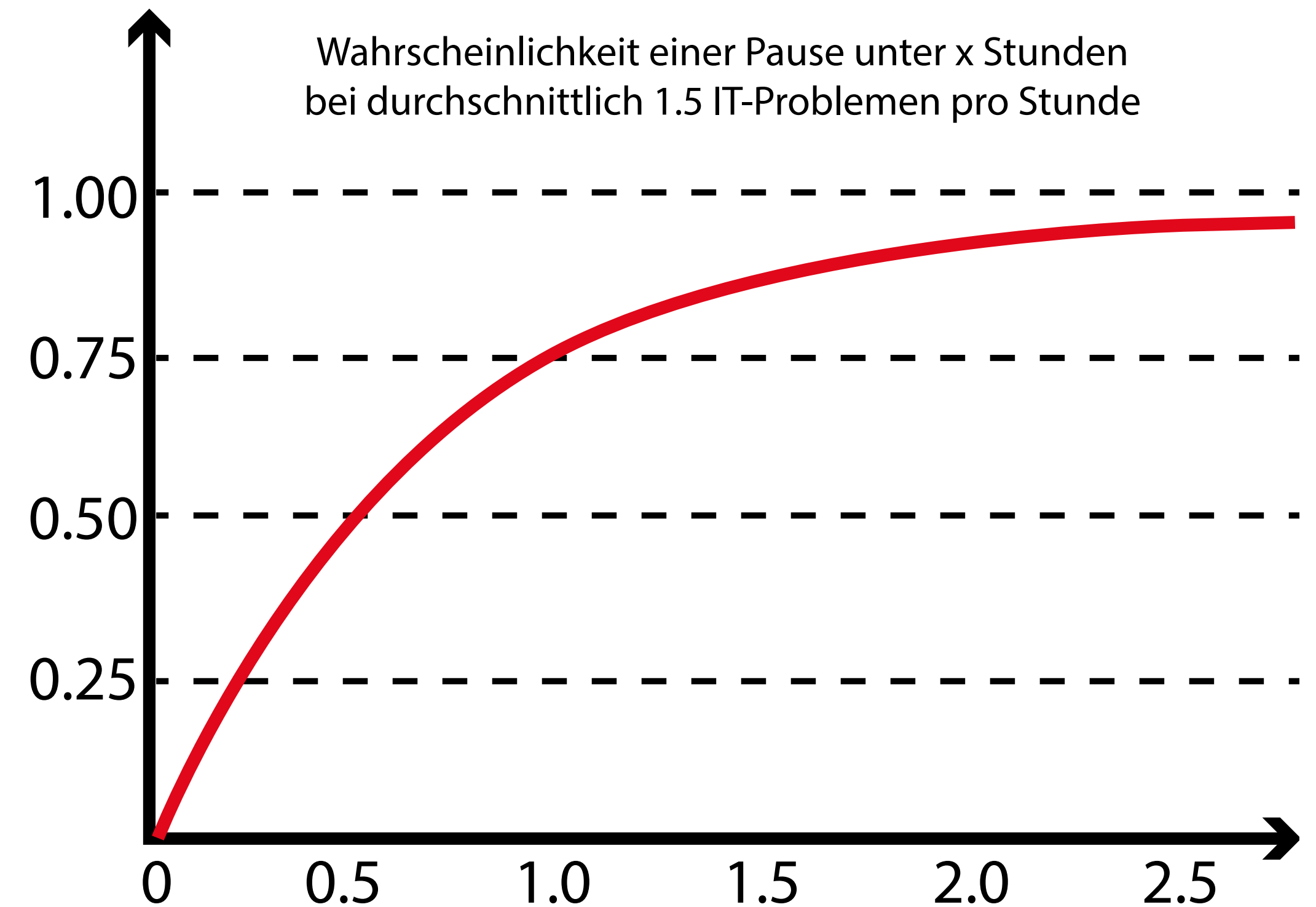
Über Integrale erhalten wir die Verteilungsfunktion, den Erwartungswert und die Varianz.

$$F(x) = \int_0^x \varphi(t) dt = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_0^x$$

Kettenregel!

$$= -e^{-\lambda x} - (-1) = 1 - e^{-\lambda x}$$

$$\mu = \frac{1}{\lambda} \quad \sigma^2 = \frac{1}{\lambda^2}$$



Exponentialverteilung

$$\int_a^b f'(x) \cdot g(x) \, dx = \left[f(x) \cdot g(x) \right] - \int_a^b f(x) \cdot g'(x) \, dx$$

$$\lim_{x \rightarrow \infty} \frac{x}{-e^{\lambda x}} = 0$$

$$\mu = \int_0^{\infty} x \cdot \varphi(x) \, dx = \int_0^{\infty} \underbrace{x}_{g(x)} \cdot \underbrace{\lambda e^{-\lambda x}}_{f'(x)} \, dx = \left[-e^{-\lambda x} \cdot x \right]_0^{\infty} - \int_0^{\infty} -e^{-\lambda x} \, dx$$

f(x) kennen wir bereits!

$$= - \left[\frac{1}{\lambda} e^{-\lambda x} \right]_0^{\infty} = \frac{1}{\lambda}$$

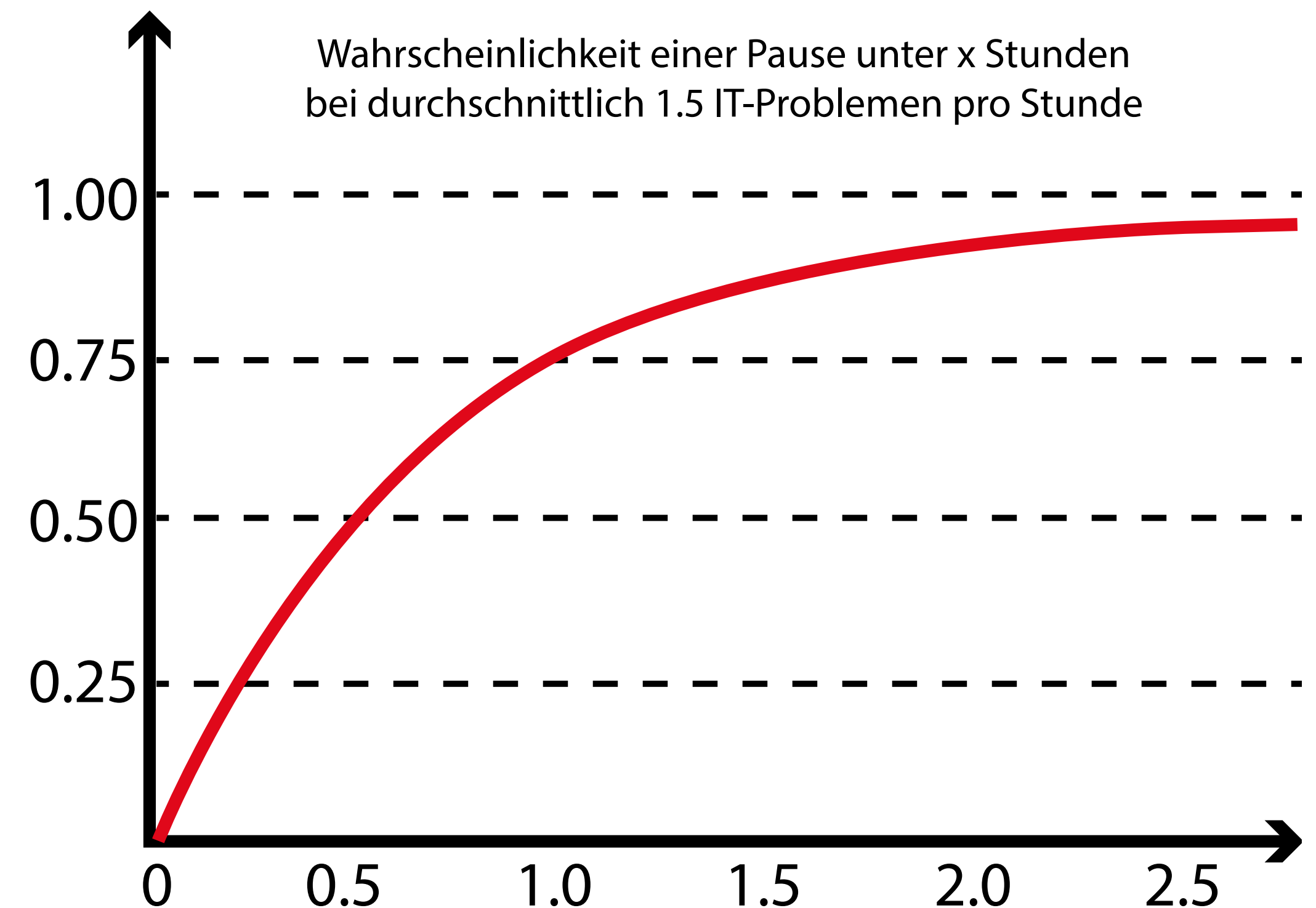
Exponentialverteilung

Für die Interpretation der Exponentialverteilung ist die Gedächtnislosigkeit wichtig.

Eine Roulettekugel landet nicht auf Rot, weil sie in den letzten 3 Spins auf Schwarz gelandet ist.

Ein IT-Problem wird nicht wahrscheinlicher, weil es schon länger keinen Anruf mehr bei der IT.S gab.

$$P(X \geq x+t \mid X \geq x) = P(X \geq t)$$



Exponentialverteilung

$$P(X \geq x) = 1 - P(X \leq x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$$

$$P(X \geq x+t \mid X \geq x) = P(X \geq t) \quad \text{mit} \quad P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$\frac{e^{-\lambda(x+t)}}{e^{-\lambda x}} = e^{-\lambda t}$$

$$\Leftrightarrow \frac{e^{-\lambda x - \lambda t}}{e^{-\lambda x}} = e^{-\lambda t}$$

$$\Leftrightarrow \frac{\cancel{e^{-\lambda x}} e^{-\lambda t}}{\cancel{e^{-\lambda x}}} = e^{-\lambda t}$$

Verteilungen & Modellierung

Ein Dartprofi trainiert die T20. Er wirft nacheinander 3 Pfeile und dokumentiert, wie viele davon die T20 treffen.

Nutzen Sie die Binomialverteilung, um folgende Wahrscheinlichkeiten zu berechnen:

- a) Ein Treffer der T20 bei einem einzelnen Wurf.
- b) In einer Aufnahme mindestens zwei Mal T20 zu treffen.
- c) Ein "180" bei dem alle drei Pfeile T20 treffen.
- d) Eine Serie von 9 perfekten Darts.

Verteilungen & Modellierung

Ein Lehrzentrum möchte eine Hotline für Probleme bei Projektarbeiten einrichten. Der wissenschaftliche Leiter ist naiv und denkt sich, dass ohnehin kaum jemand anrufen wird.

Bei einem Probelauf bekommt er dann aber jede Menge Anrufe. Er protokolliert alle Anrufe und möchte nun Schlüsse daraus ziehen.

Verwenden Sie die Ihnen bekannten Verteilungen, um folgende Fragen zu beantworten! Wie hoch ist die Wahrscheinlichkeit, dass ...

- a) ... in einer Stunde kein einziger Anruf kommt.
- b) ... in einer Stunde mindestens 5 Anrufe kommen.
- c) ... zwischen zwei eingehenden Anrufen max. 5 Minuten liegen.
- d) ... zwischen zwei eingehenden Anrufen mind. 30 Minuten liegen.

Berechnen Sie außerdem den Erwartungswert der Länge der Pause zwischen zwei Gesprächen, sowie die Wahrscheinlichkeit, während eines laufenden Gesprächs das Klopffzeichen zu erhalten. Nehmen Sie dabei an, dass die Länge der Gespräche gleichverteilt ist.

Verteilungen & Modellierung

Die Lösung zu den meisten Aufgabenteilen findet ihr direkt im Code.

Da der Erwartungswert linear ist, gilt für die erwartete Länge der Pause:

$$60/2.5 - 15 = 9 \text{ Minuten}$$

Für die Klopfaufgabe benötigen wir noch folgende Rechnung, die auf dem Gesetz der totalen Wahrscheinlichkeit aufbaut.

$$\begin{aligned}
 p(X < Y) &= \int_{-\infty}^{\infty} p(X < y) \varphi(y) dy \\
 &= \int_{-\infty}^{\infty} F(y) \varphi(y) dy \quad \varphi(y) = \begin{cases} 3 & \text{für } y \in \left[\frac{1}{12}, \frac{5}{12}\right] \\ 0 & \text{für } y \notin \left[\frac{1}{12}, \frac{5}{12}\right] \end{cases} \\
 &= \int_{1/12}^{5/12} (1 - e^{-2.5y}) \cdot 3 dy \\
 &= 3 \cdot \left[\int_{1/12}^{5/12} dy - \int_{1/12}^{5/12} e^{-2.5y} dy \right] = 3 \cdot \left[y + 0.4e^{-2.5y} \right]_{1/12}^{5/12} = 44.9\%
 \end{aligned}$$

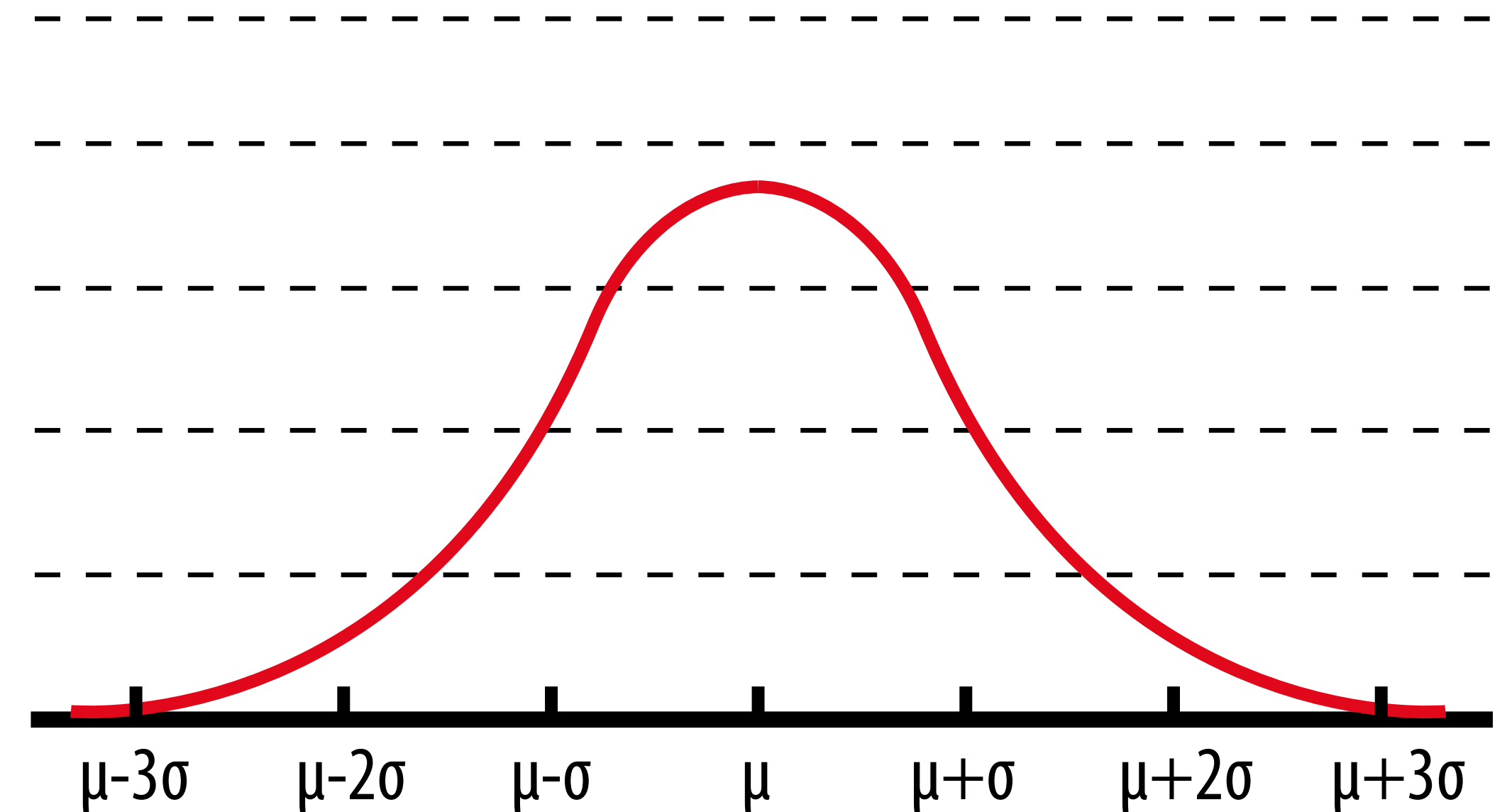
Kehrwert Intervallbreite

Normalverteilung

Die Normalverteilung ist die wichtigste Verteilung der Statistik. Mit dem Erwartungswert μ und der Varianz σ^2 schreiben wir:

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

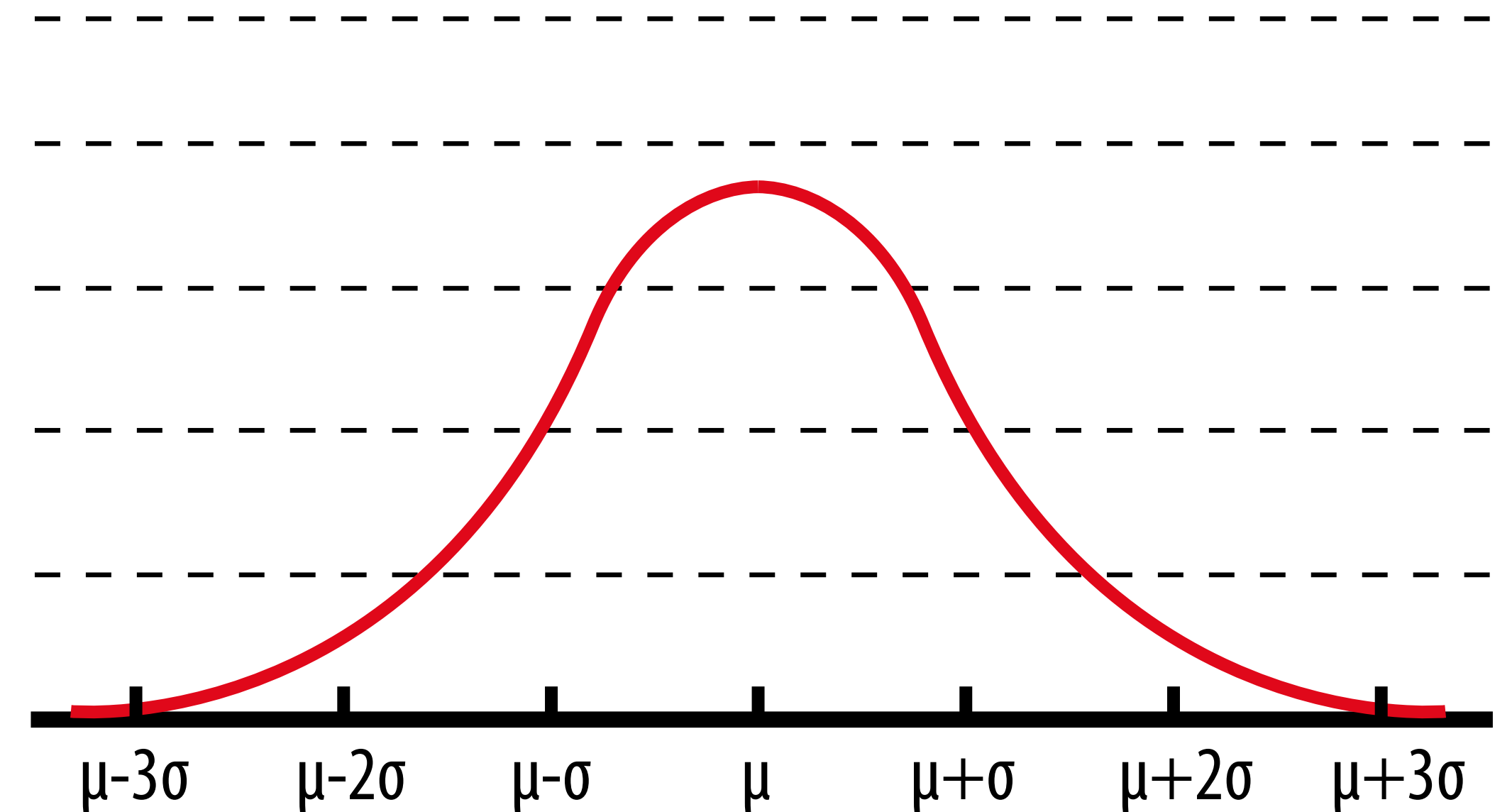


Normalverteilung

Wieso bezeichnen wir diese durchaus komplizierte Verteilung als „normal“?

Zum einen gibt es viele Merkmale in der Natur, deren Verteilung gut zu dieser Normalverteilung passt.

Zum anderen werden wir gleich einen interessanten Zusammenhang zwischen der Normalverteilung und vielen anderen Verteilungen kennenlernen!

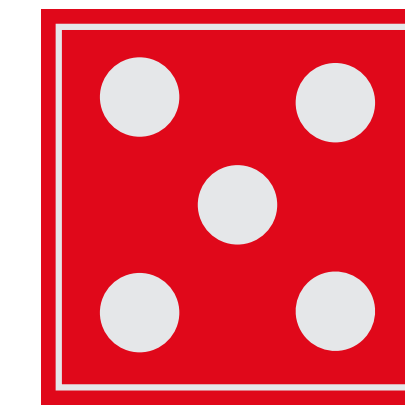


Zentraler Grenzwertsatz

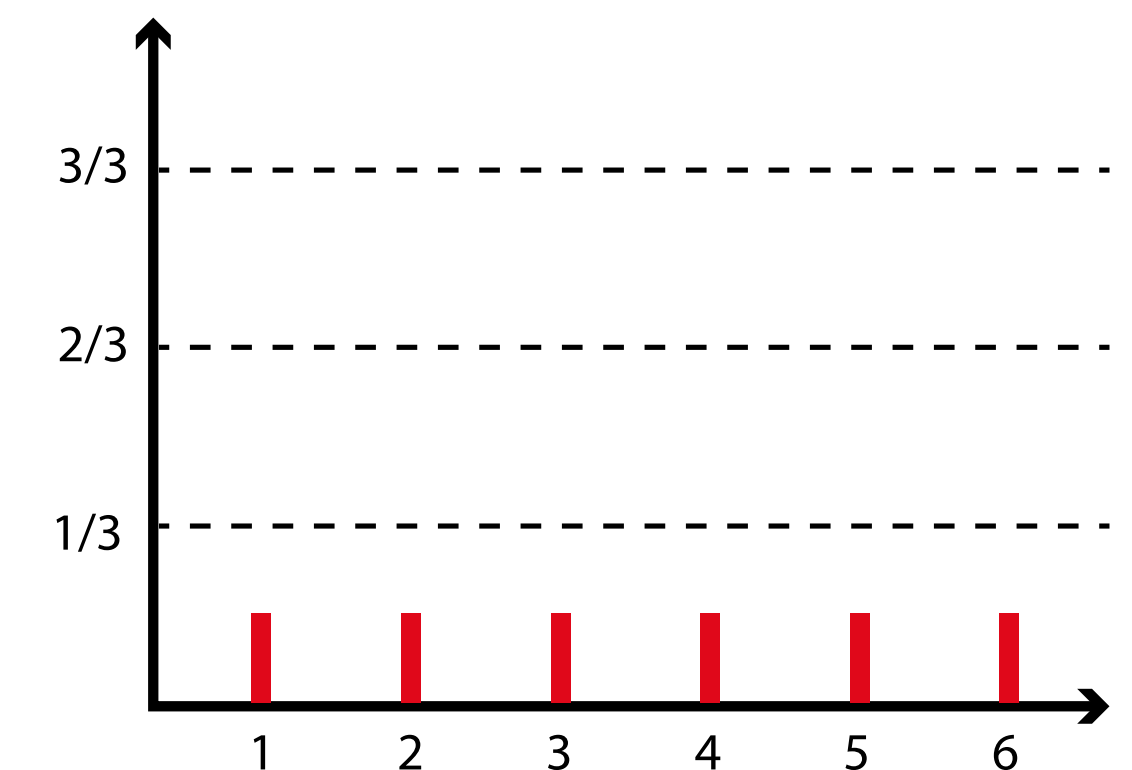
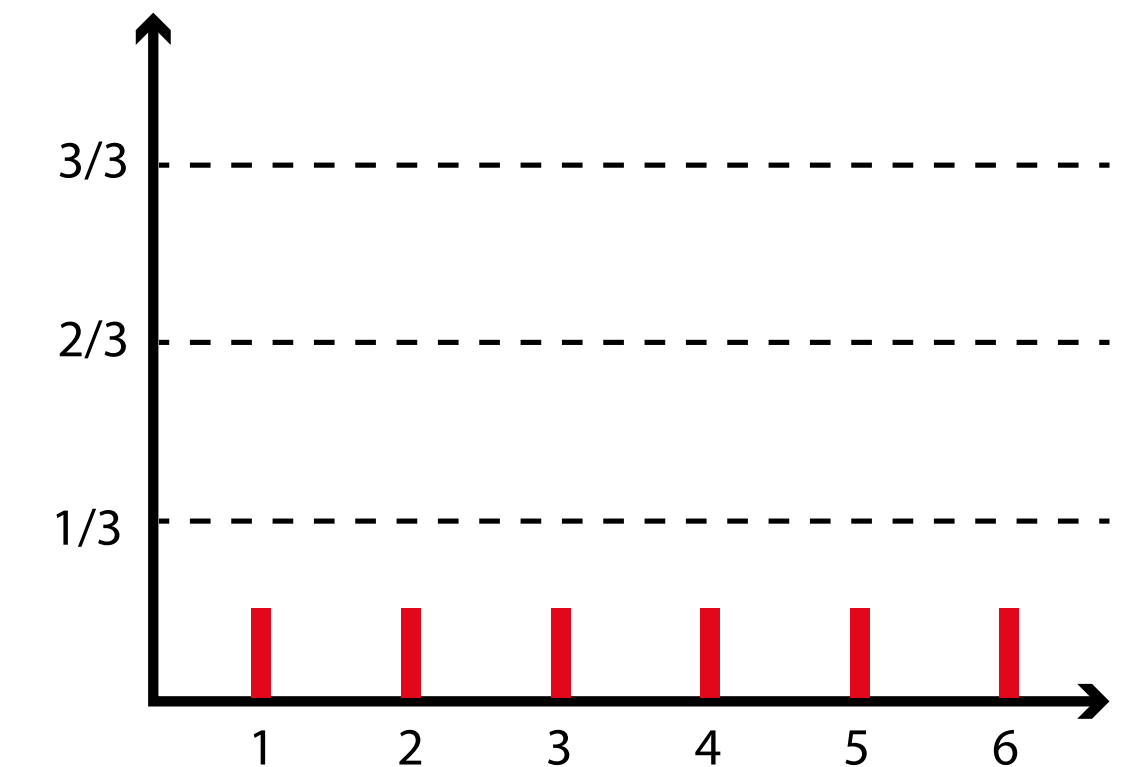
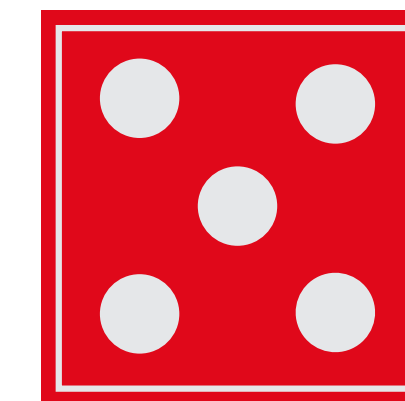
Wir werfen zwei 6-seitige Würfel und definieren X als die Summe der beiden Augenzahlen $X_1 + X_2$

Die einzelnen Augenzahlen X_1 und X_2 sind gleichverteilt mit Erwartungswert 3.5 und Varianz 2.91.

Ist X dann gleichverteilt mit einem Erwartungswert von 7 und einer Varianz von 5.82?



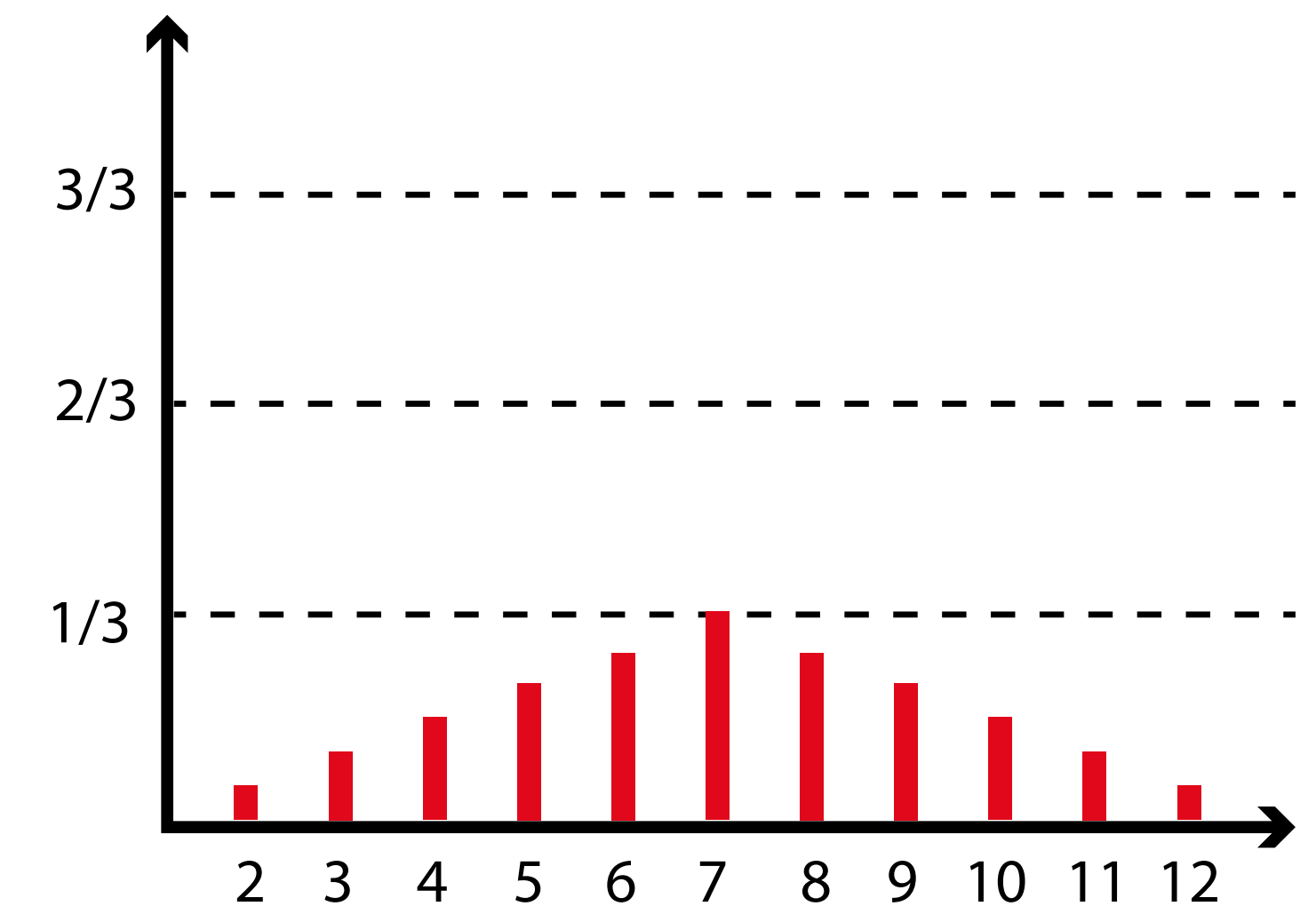
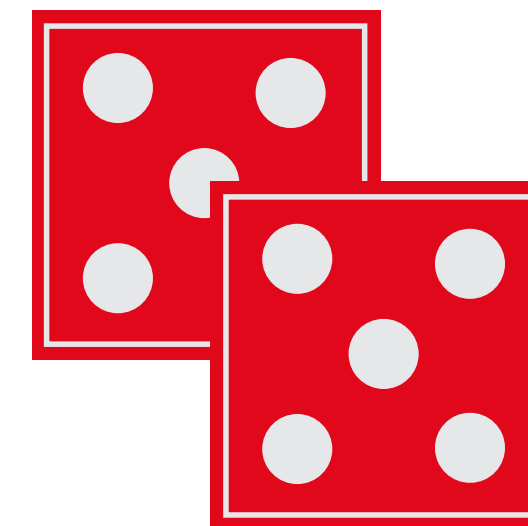
+



Zentraler Grenzwertsatz

Nein, die Werte für den Erwartungswert und die Varianz sind zwar richtig, aber X ist nicht mehr gleichverteilt!

Die mittlere 7 wird zum wahrscheinlichsten Ergebnis, während besonders kleine/große Augenzahlen sehr selten sind!



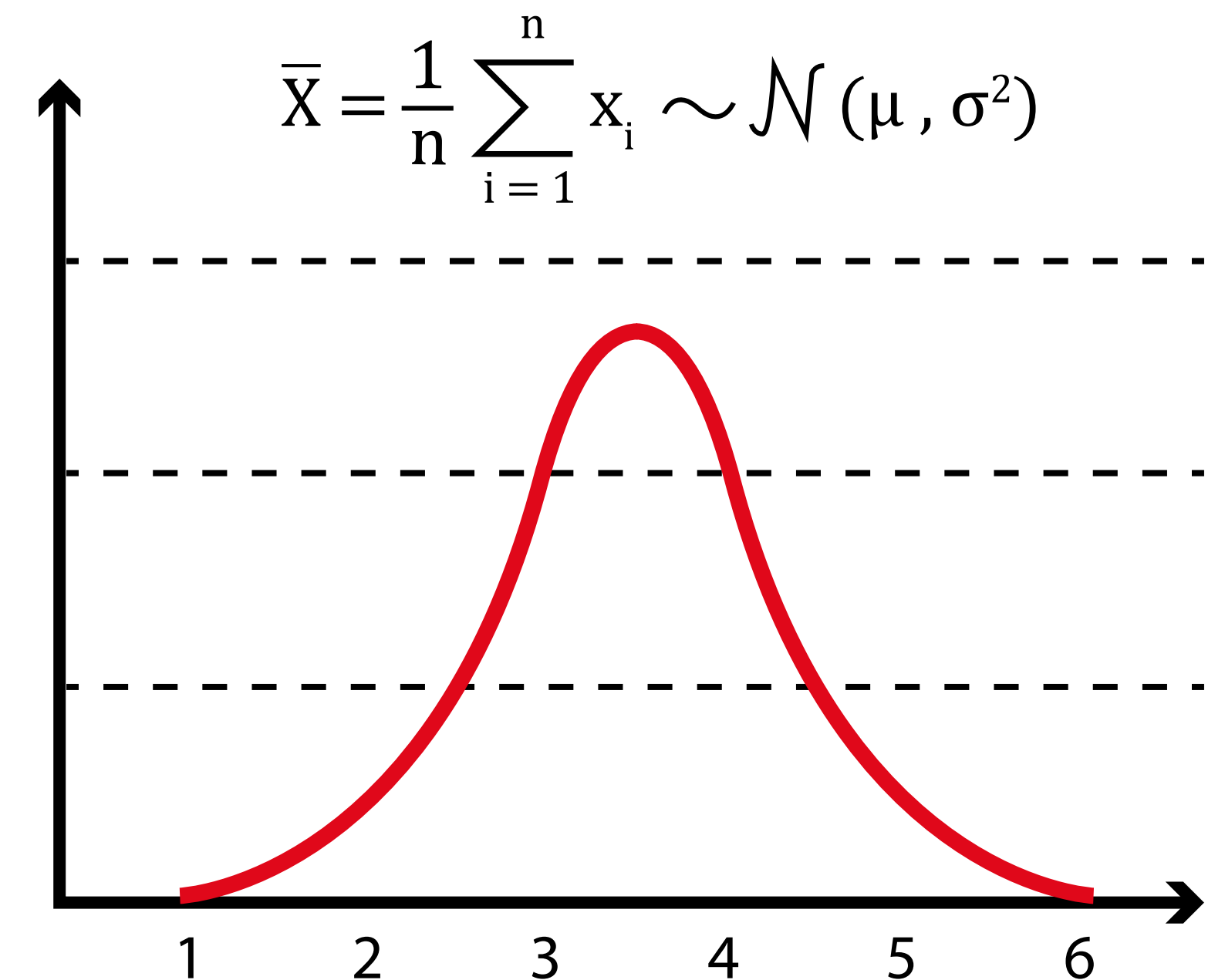
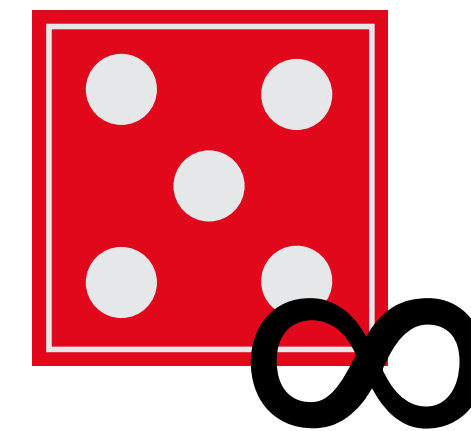
Zentraler Grenzwertsatz

Seien $X_1, X_2, \dots, X_n$ identisch verteilte und unabhängige Zufallsvariablen mit Erwartungswert μ und Varianz σ^2

Für große n ist die Summe der Zufallsvariablen normalverteilt. Im Grenzwert für n gegen unendlich gilt:

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = \Phi(z)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$



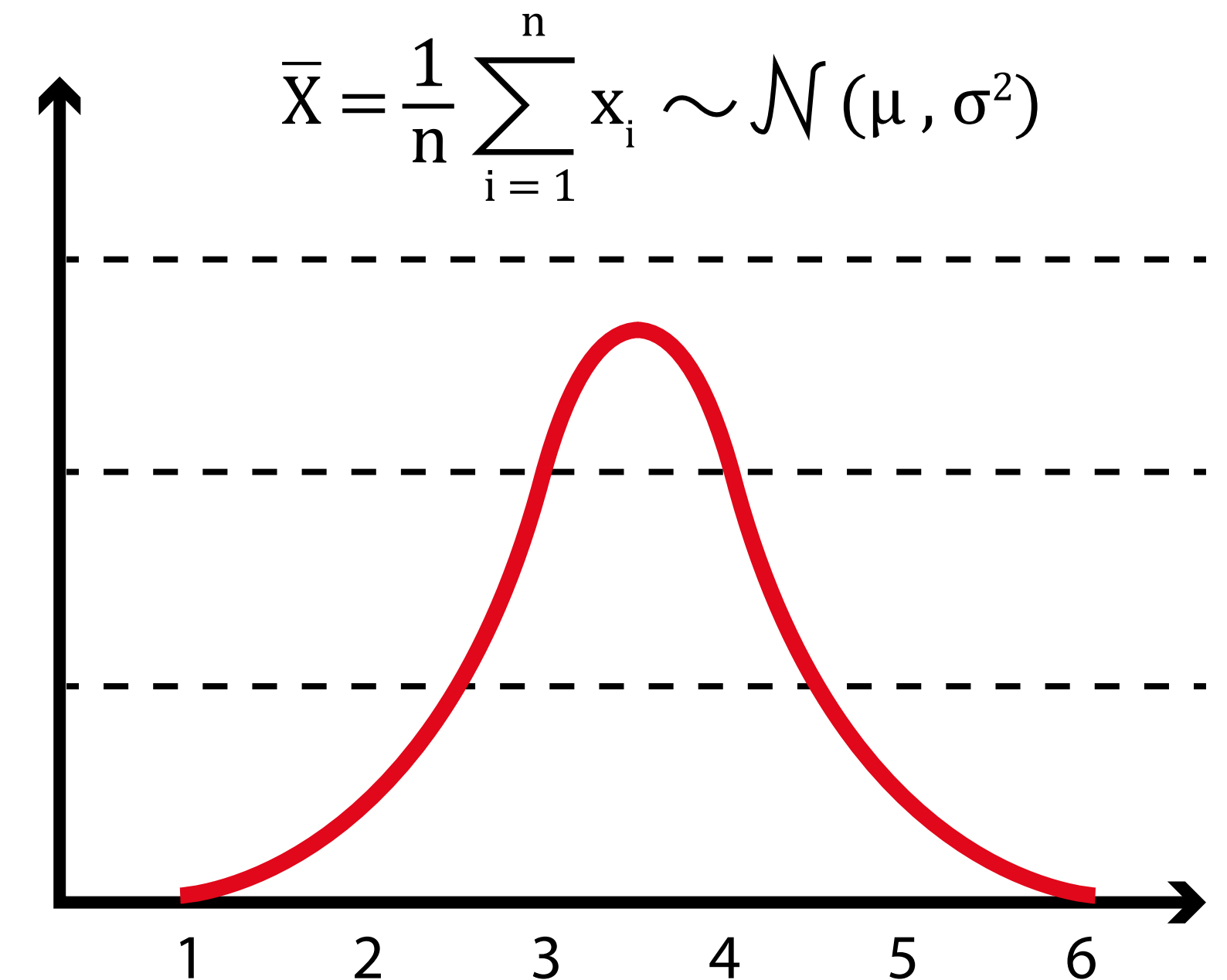
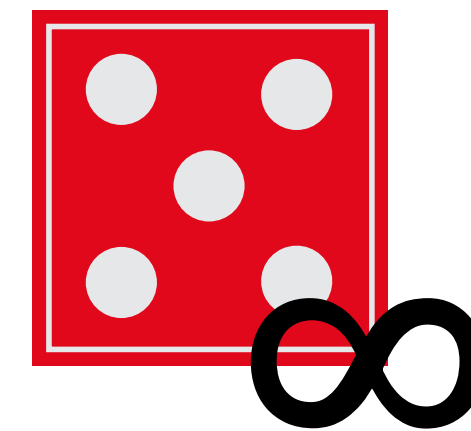
Zentraler Grenzwertsatz

Für die Summe der Zufallsvariablen bedeutet dies Normalverteilung mit Erwartungswert $n\mu$ und Varianz $n\sigma^2$.

$$X = \sum_{i=1}^n x_i \sim N(n\mu, n\sigma^2)$$

Berechnen wir dagegen den Mittelwert X/n erhalten wir:

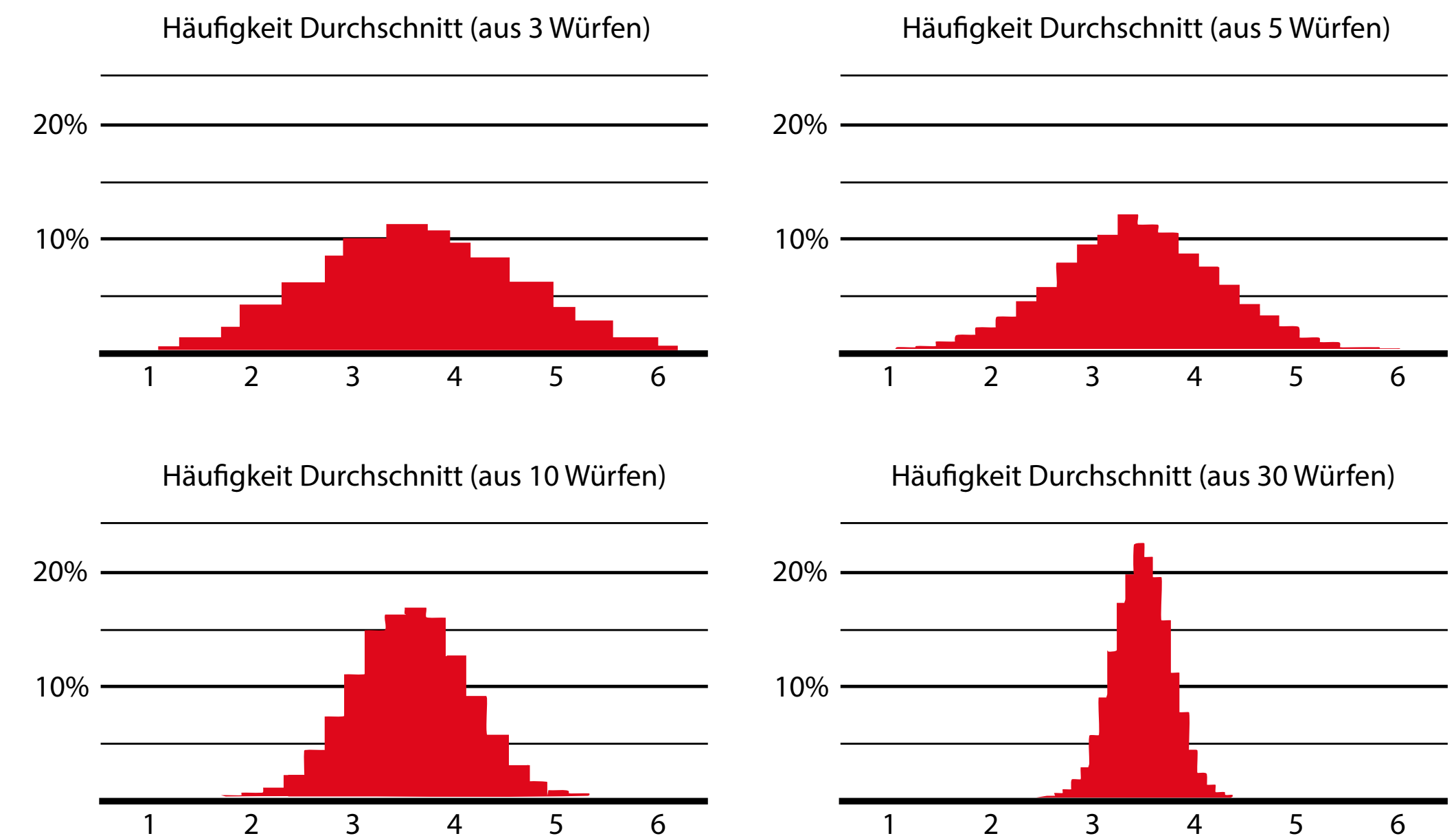
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



Zentraler Grenzwertsatz

Der Grenzwert gegen unendlich vermittelt den Eindruck, dass mit einem großen n eine Zahl in den Tausendern oder Millionen gemeint ist.

Als Faustregel werden aber oft 30 Beobachtungen genannt bzw. 100 falls die Verteilung der einzelnen Zufallsvariablen schwierig ist (z. B. bimodal, multimodal oder sehr schief).



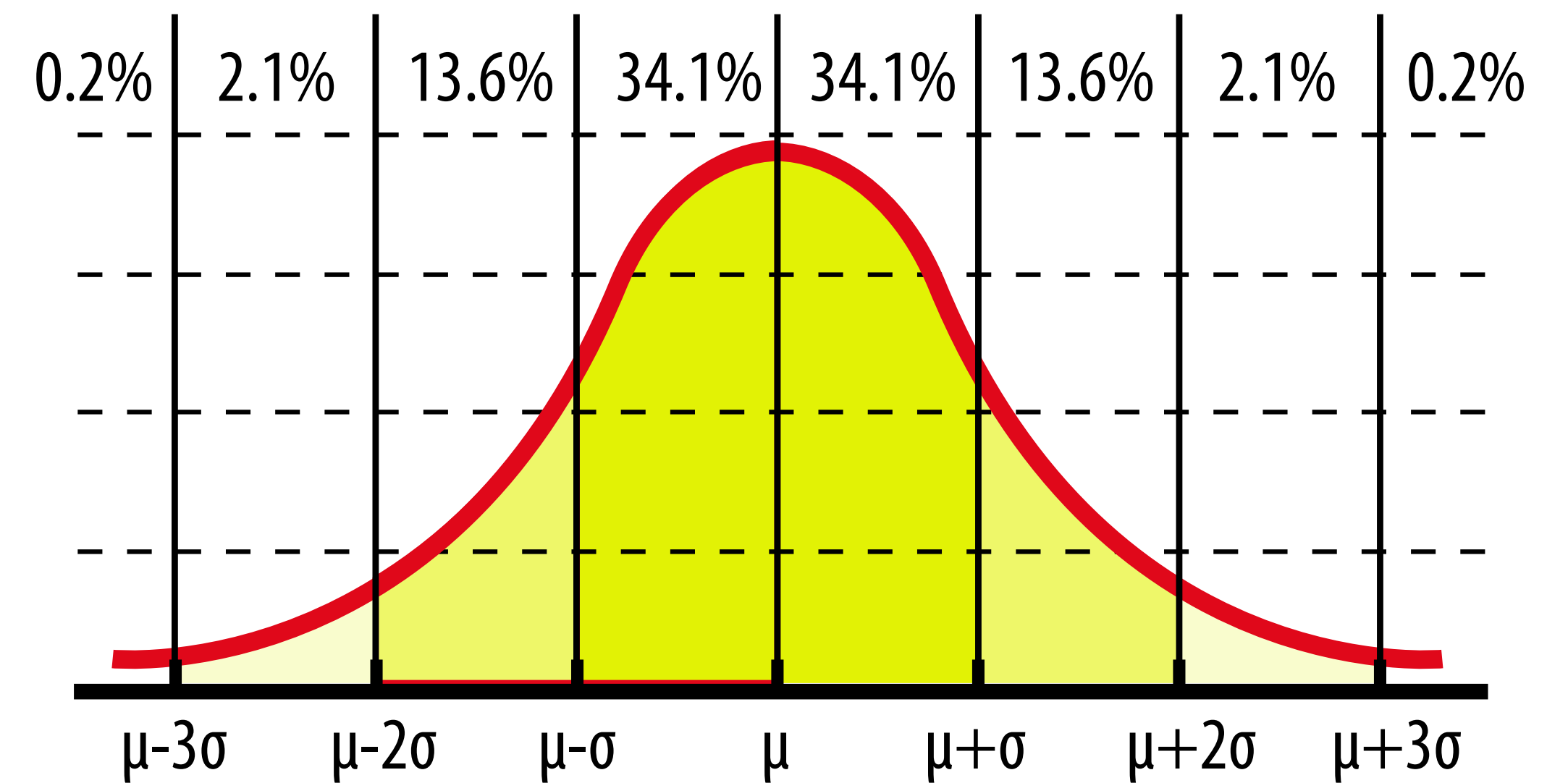
Zentraler Grenzwertsatz

Mithilfe der Dichtefunktion können wir die Wahrscheinlichkeit berechnen, mit dem die Werte in einem bestimmten Bereich liegen.

68.2% liegen innerhalb 1x Standardabweichungen um μ .

95.4% liegen innerhalb 2x Standardabweichungen um μ .

99.6% liegen innerhalb 3x Standardabweichungen um μ .



Zentraler Grenzwertsatz

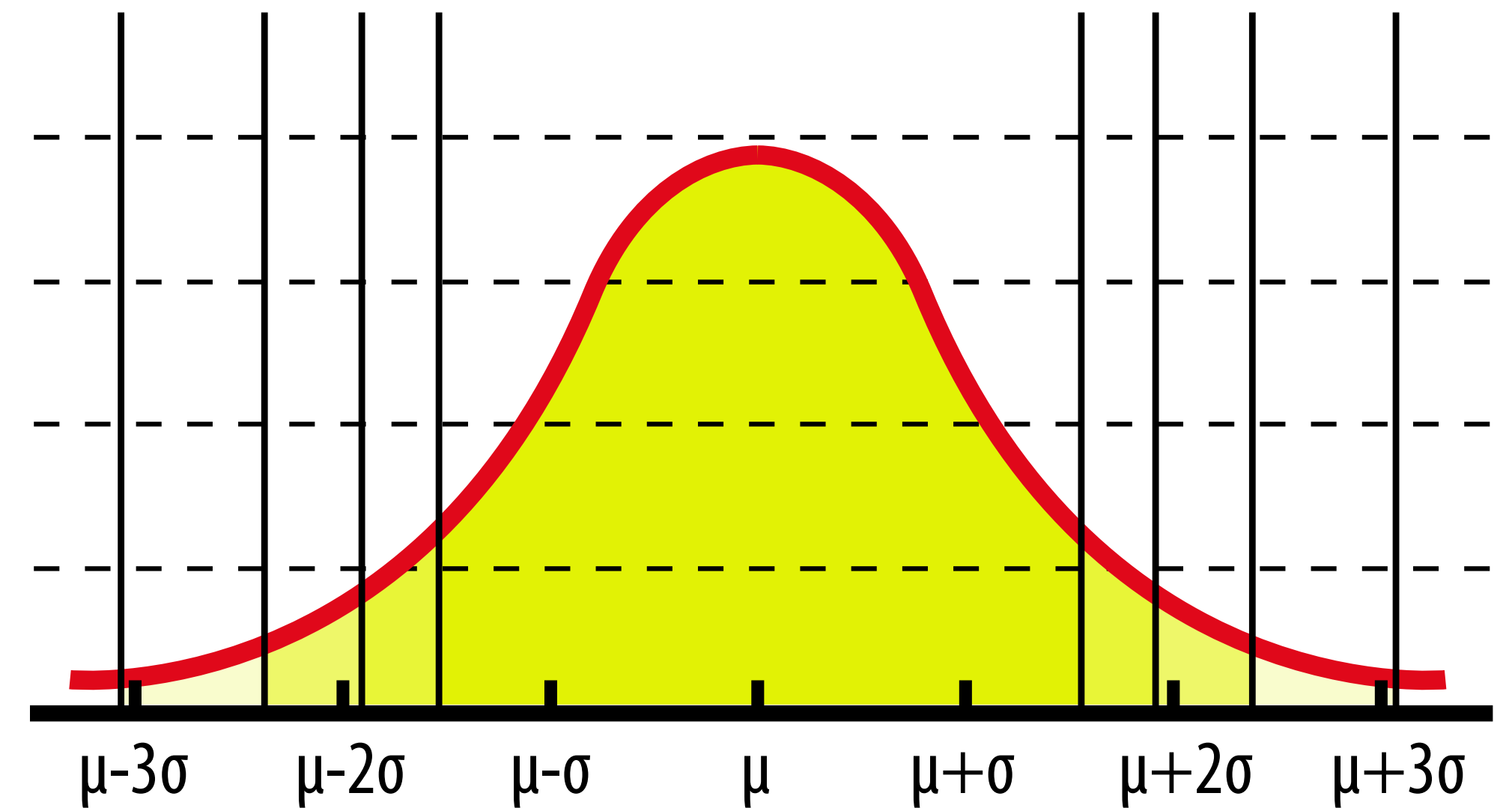
Mithilfe der Dichtefunktion können wir die Wahrscheinlichkeit berechnen, mit dem die Werte in einem bestimmten Bereich liegen.

90.0% liegen innerhalb 1.64x Standardabweichungen um μ .

95.0% liegen innerhalb 1.96x Standardabweichungen um μ .

99.0% liegen innerhalb 2.32x Standardabweichungen um μ .

99.9% liegen innerhalb 3.09x Standardabweichungen um μ .

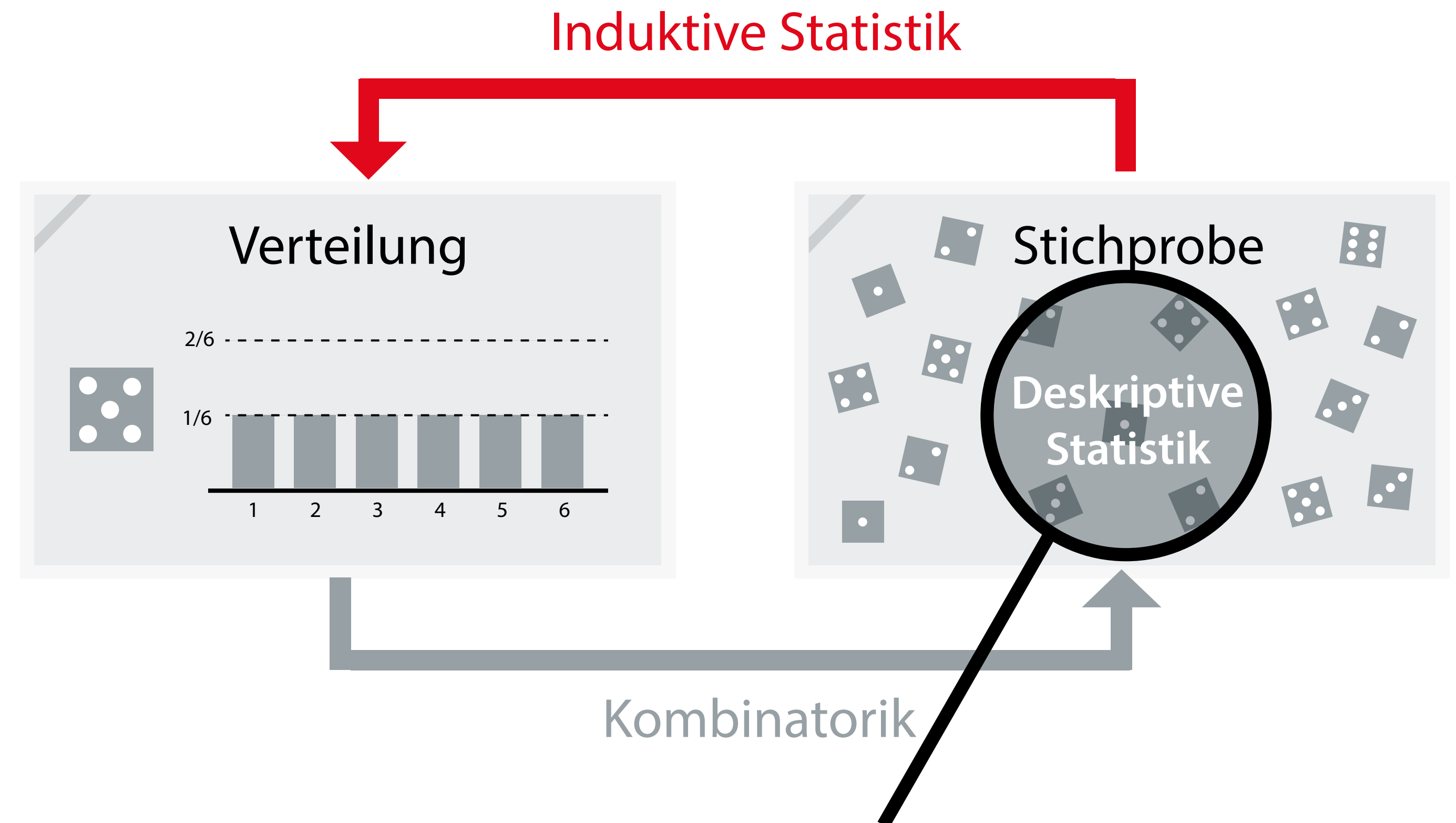


Induktive Statistik

Wir wechseln jetzt in die **induktive Statistik**.

Dort schließen wir von einer Stichprobe auf die zugrunde liegende Verteilung bzw. Grundgesamtheit.

Dabei werden wir statistische Tests anwenden, um Hypothesenpaare zu untersuchen.



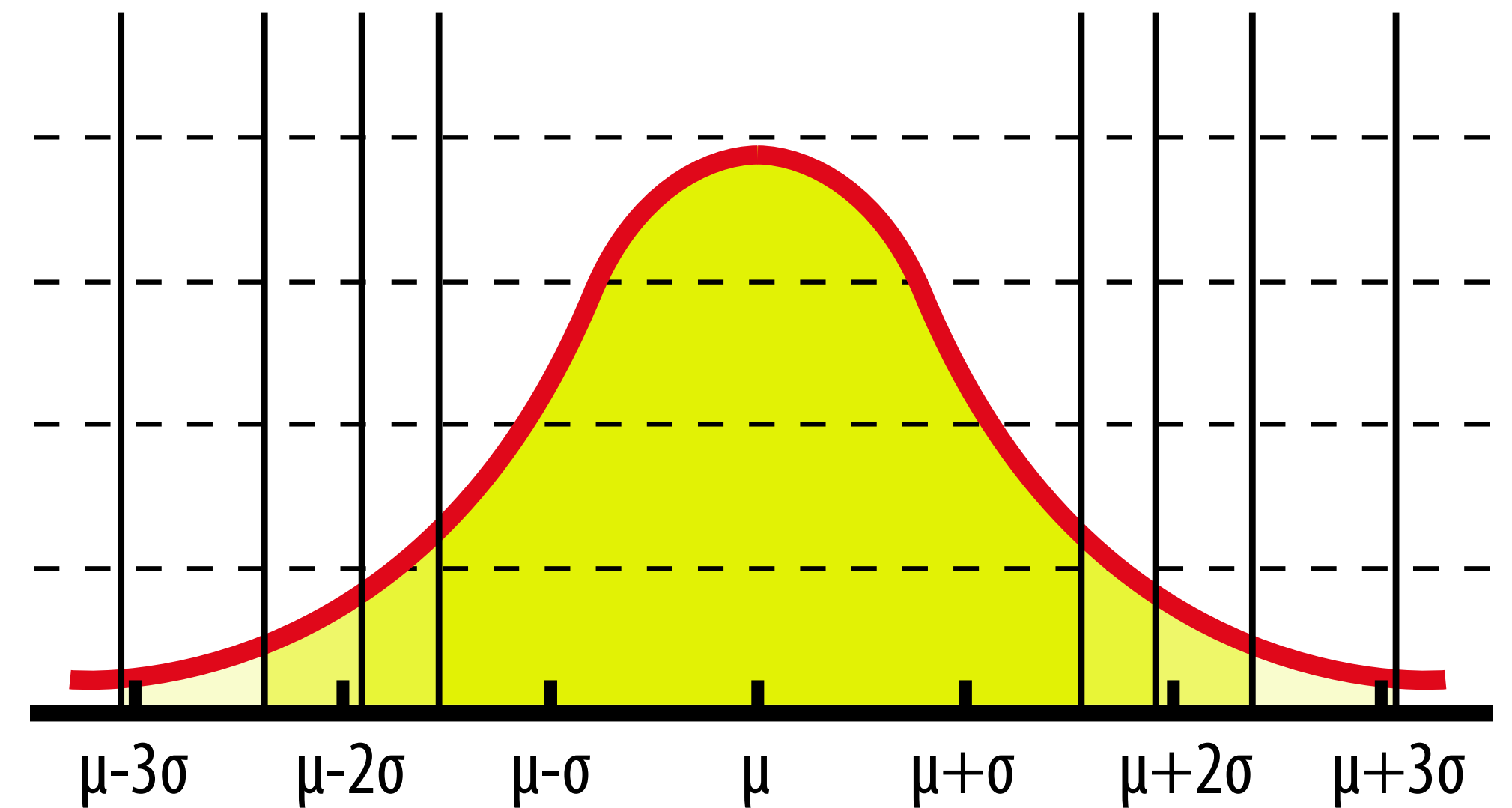
Z-Test

Wir würfeln 40-mal mit einem sechsseitigen Würfel und erhalten die Augensumme 128 bzw. eine durchschnittliche Augenzahl von 3.2.

Wir wissen bereits, dass jede einzelne Augenzahl X_i gleichverteilt ist mit $\mu=3.5$ und $\sigma^2=2.91$

Die durchschnittliche Augenzahl über 40 Würfelversuche ist näherungsweise normalverteilt mit:

$$\mu=3.5 \text{ und } \sigma^2 = \frac{2.91}{40} = 0.07275$$



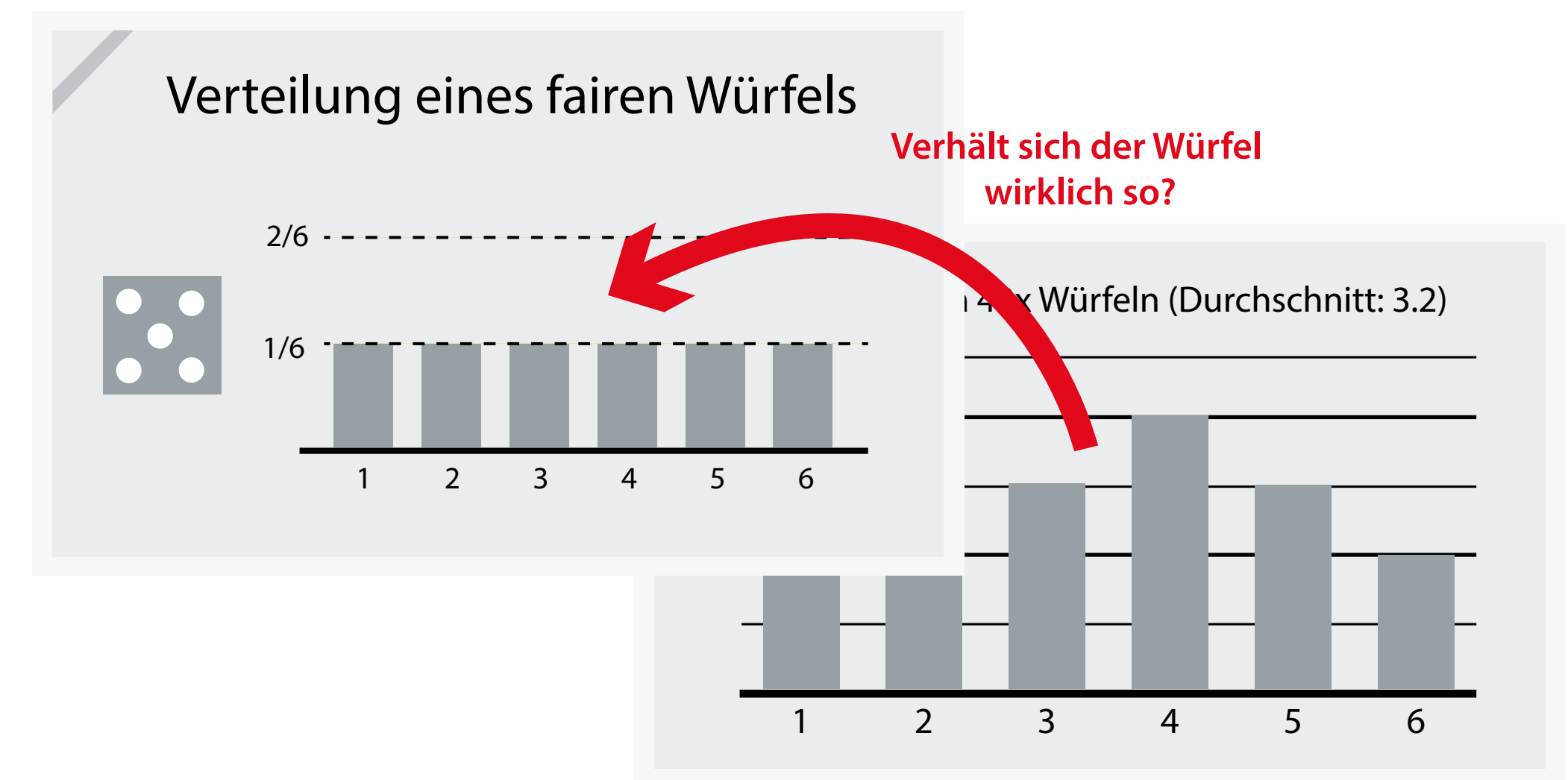
Z-Test

Der Durchschnitt der Stichprobe liegt mit 3.2 unter dem Erwartungswert von 3.5.

Ist die Abweichung von 0.3 Augen zu wenig bei $N=40$ noch plausibel? Wir stellen unser erstes Hypothesenpaar auf!

Nullhypothese Der Würfel zeigt durchschnittlich 3.5 Augen.

Alternativhypothese: Der Würfel zeigt durchschnittlich weniger als 3.5 Augen.



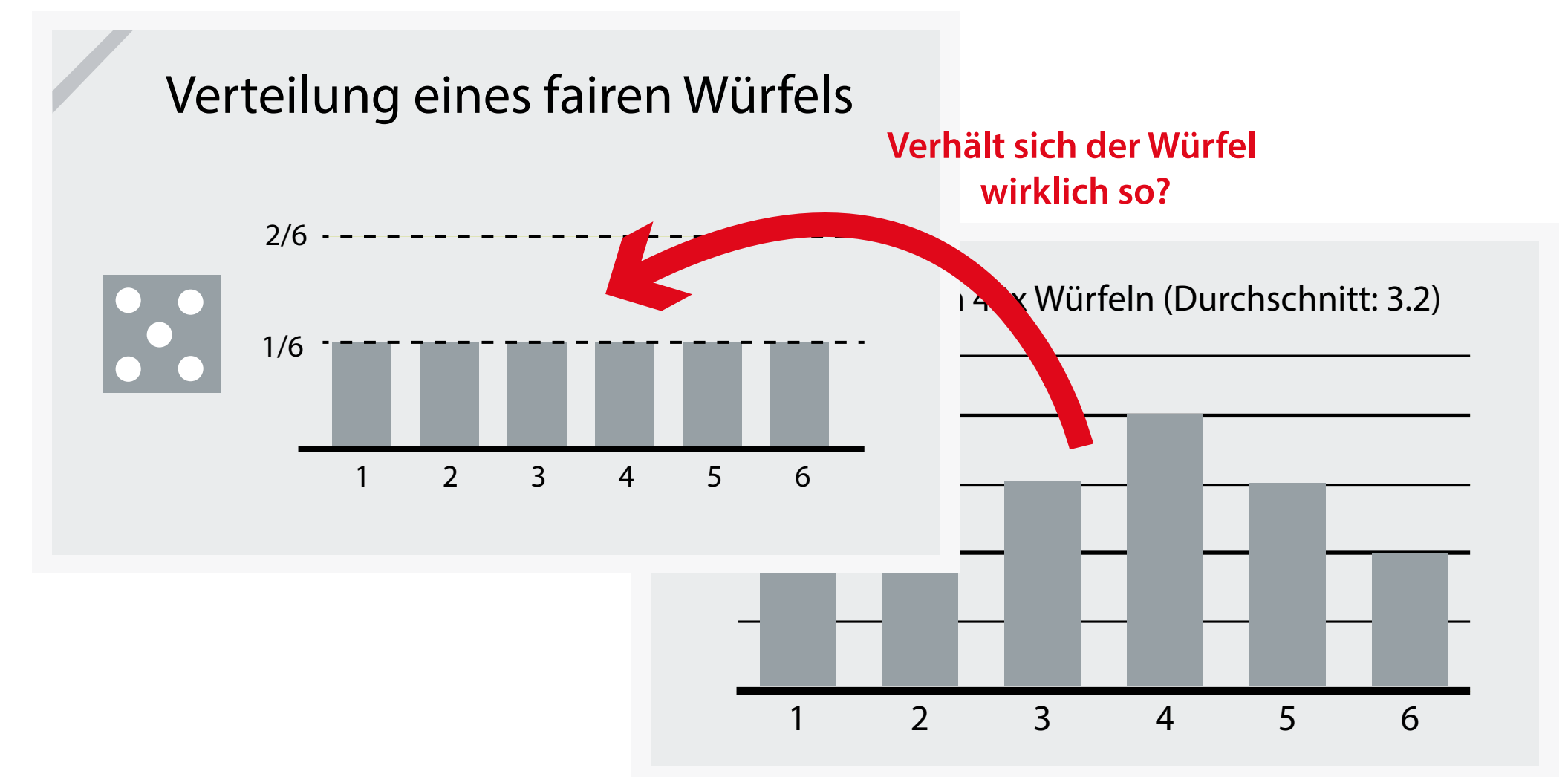
Z-Test

Danach setzen wir die gemessene Abweichung ins Verhältnis zur Standardabweichung. Diese beträgt:

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.07275} = 0.270$$

Unsere Stichprobe liegt 1.11 Standardabweichungen unter dem Erwartungswert.

$$z = \frac{\bar{\mu} - \mu}{\sigma} = -1.112$$

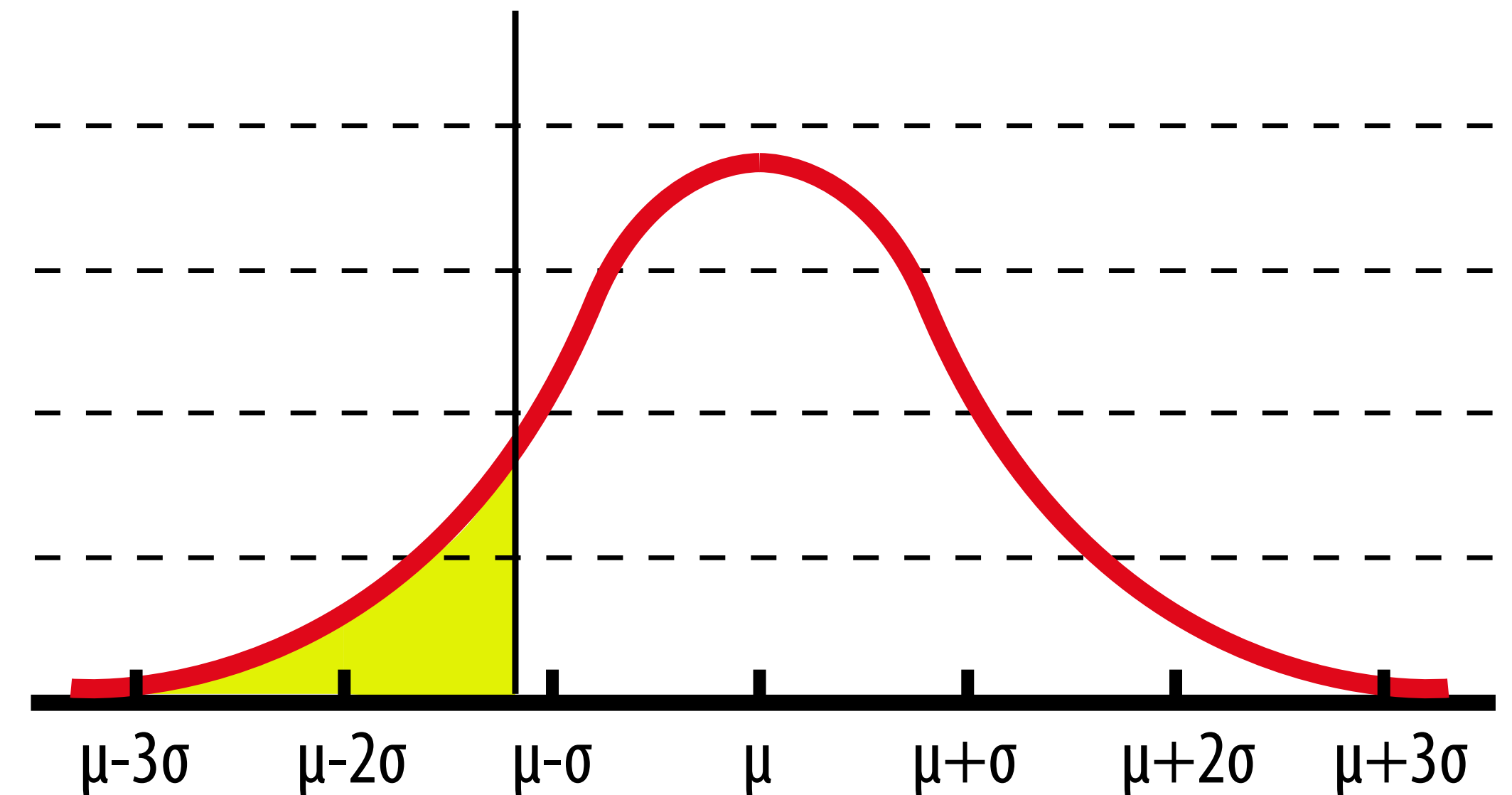


Z-Test

Setzen wir diesen Wert in die Verteilungsfunktion der Standardnormalverteilung ein, erhalten wir den **p-Wert**.

$$p = \Phi(z) = \int_{-\infty}^{-1.112} \varphi(x) dx = 0.133$$

Interpretation: Wäre die Nullhypothese wahr, würden wir bei der gegebenen Stichprobengröße mit Wahrscheinlichkeit von 13.3% eine Abweichung von mindestens 0.3 Augen nach unten erhalten.



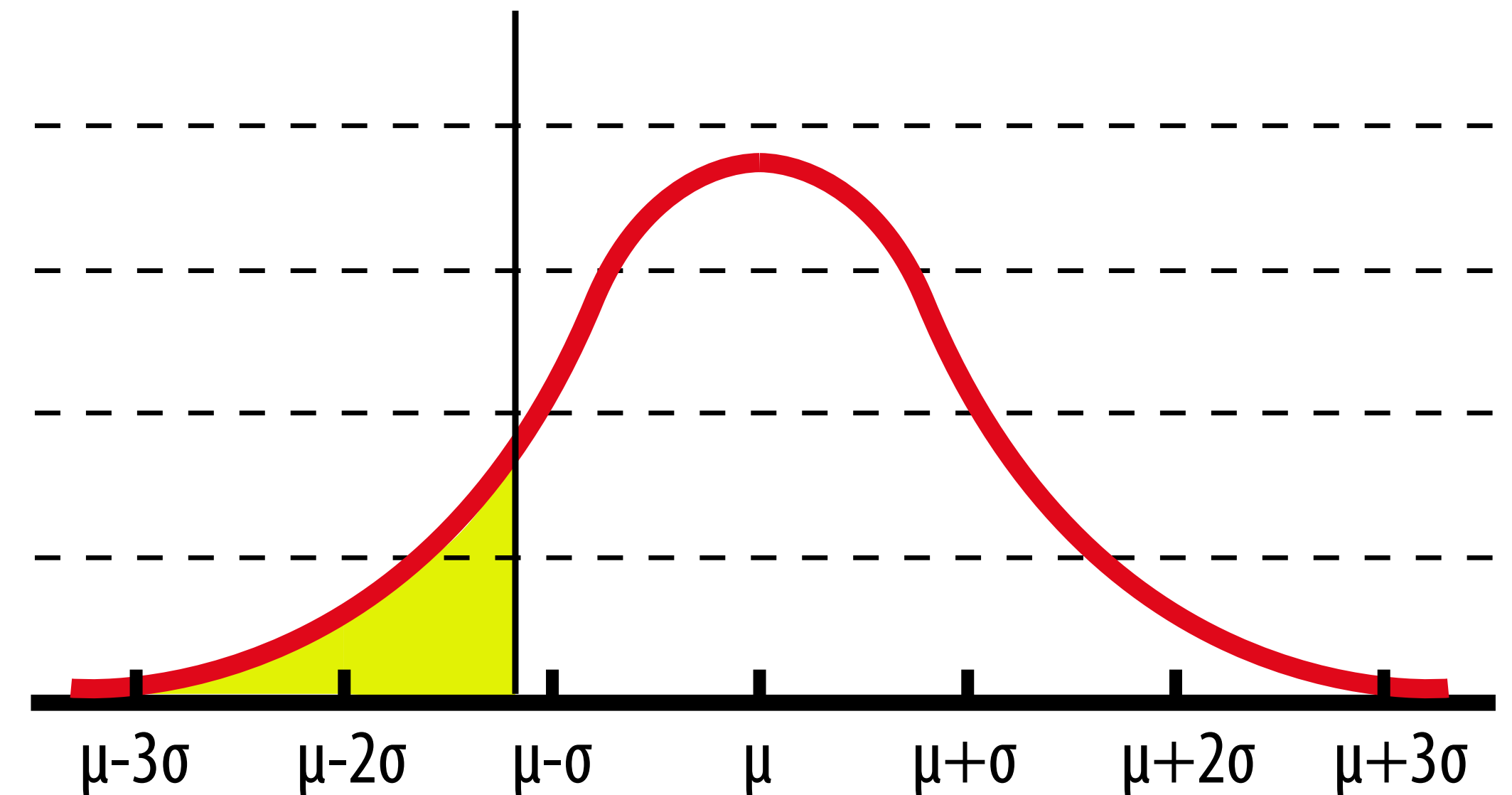
Z-Test

Erst wenn diese Wahrscheinlichkeit kleiner als 5% ist, lehnen wir die Nullhypothese ab.

Hier ist der p-Wert über dieser kritischen Schwelle. Wir behalten die Nullhypothese bei.

Nullhypothese: Der Würfel zeigt im Durchschnitt 3.5 Augen.

Alternativhypothese: Der Würfel zeigt im Durchschnitt weniger als 3.5 Augen.

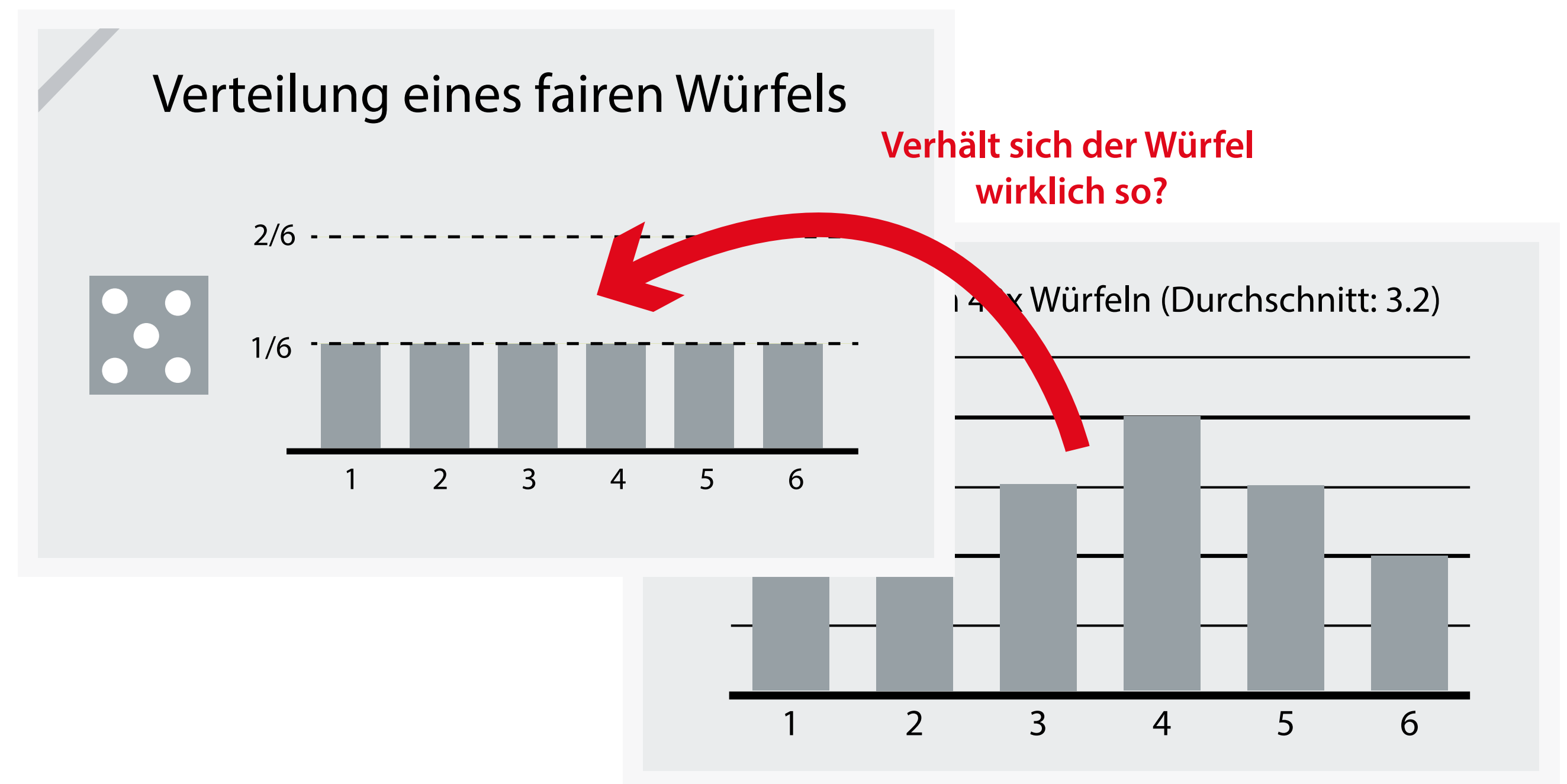


Einseitig vs. Zweiseitig

Gegeben der Datenlage prüfen wir nur, ob der Würfel im Durchschnitt zu wenig Augen zeigt.

Ganz allgemein wäre aber auch ein Würfel, der im Durchschnitt zu viel Augen zeigt, problematisch.

Ein guter Würfel sollte im Durchschnitt im genau 3.5 Augen zeigen, nicht mehr und nicht weniger!

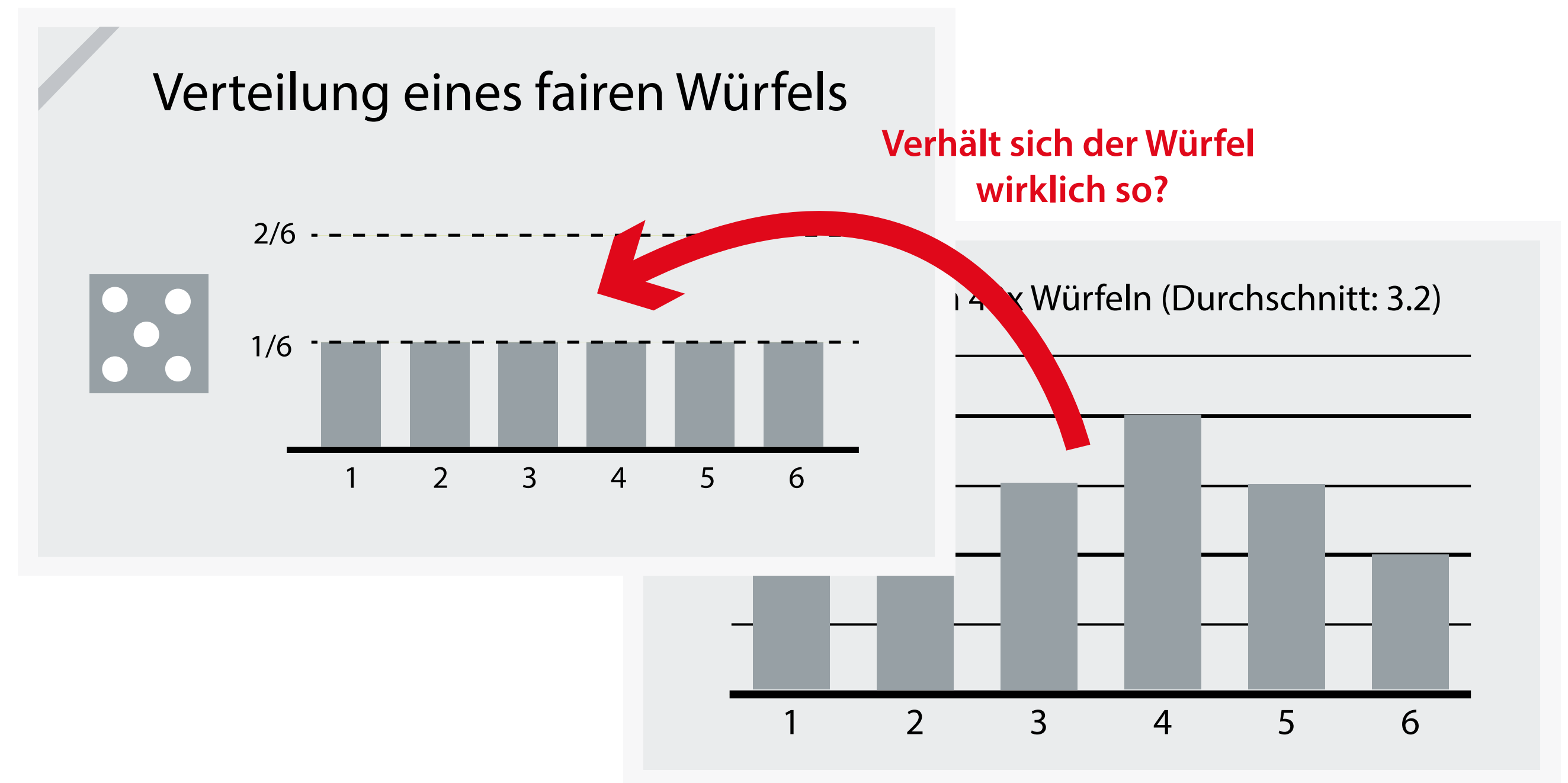


Einseitig vs. Zweiseitig

Unser Test war einseitig, genauer gesagt linksseitig. Wir haben untersucht, ob der Würfel im Schnitt zu wenig Augen zeigt. Ein zweiseitiger Test hätte das Hypothesenpaar:

Nullhypothese: Der Würfel zeigt im Durchschnitt 3.5 Augen.

Alternativhypothese: Der Würfel zeigt im Durchschnitt mehr oder weniger als 3.5 Augen.



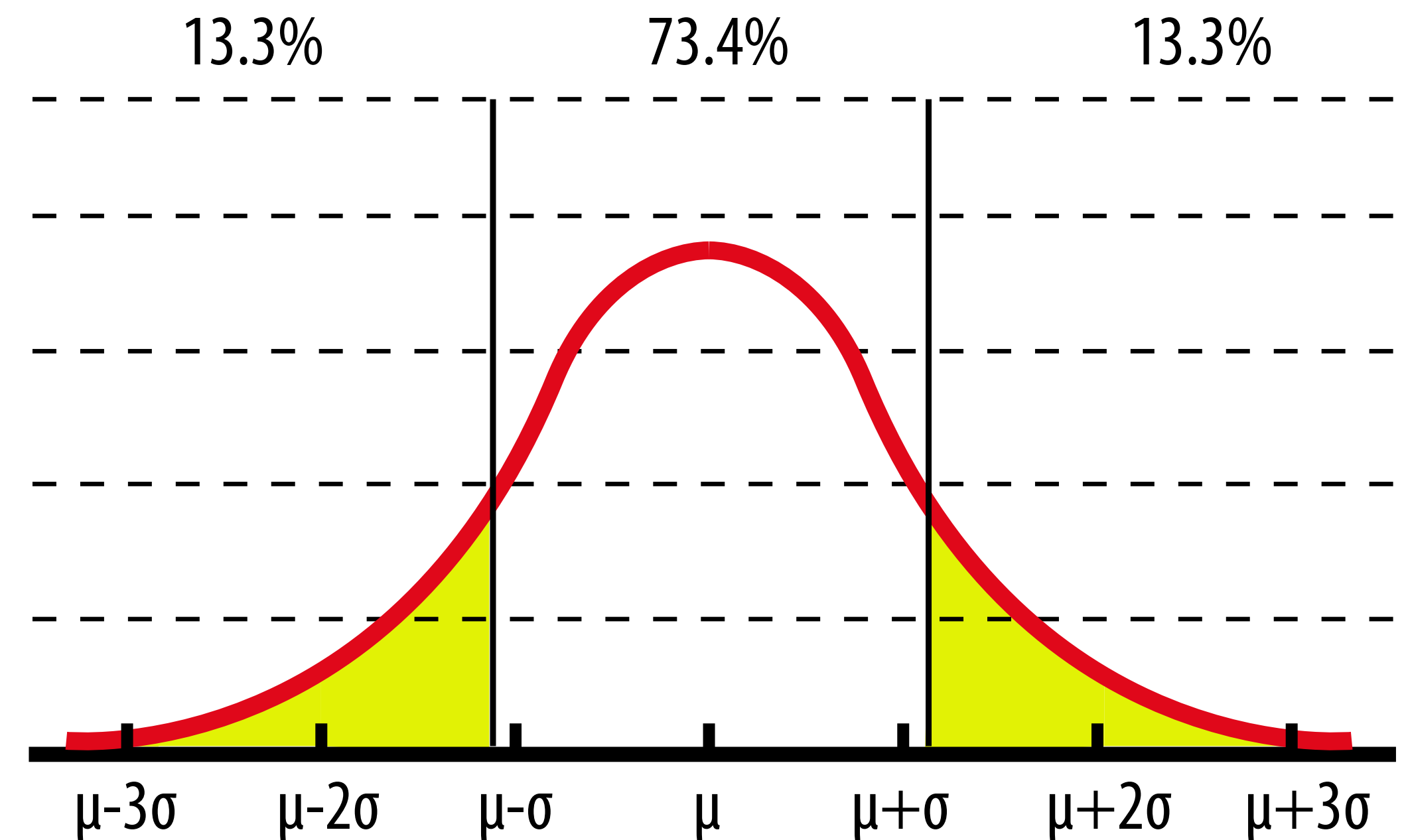
Einseitig vs. Zweiseitig

Die Berechnung der Standardabweichung bleibt dieselbe.
Wir erhalten erneut $\sigma=0.270$ und liegen damit ...

$$z = \frac{\bar{\mu} - \mu}{\sigma} = -1.112$$

... Standardabweichungen unter dem Erwartungswert. Damit berechnen wir:

$$p = 1 - \int_{-1.112}^{1.112} \varphi(x) dx = 0.266$$



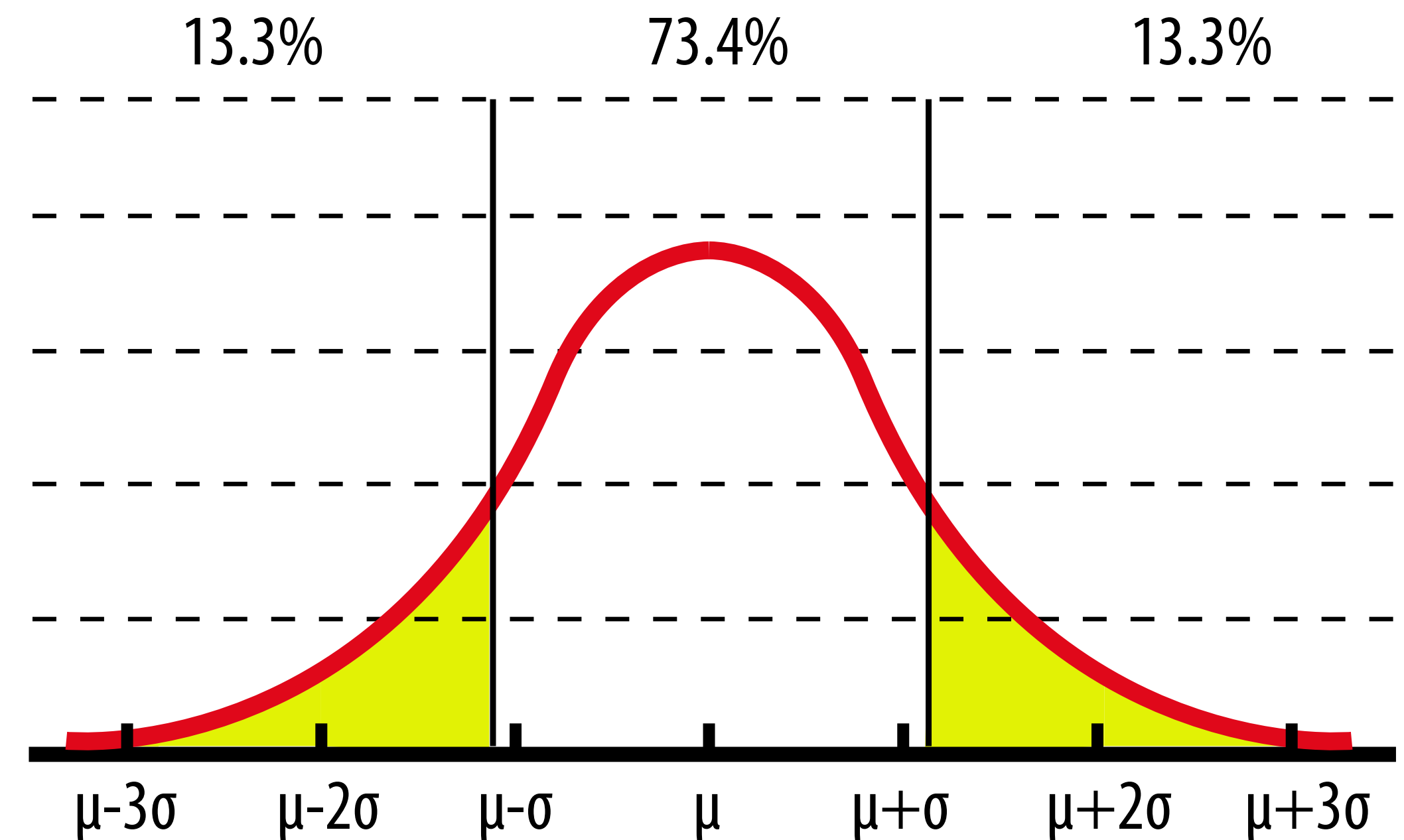
Einseitig vs. Zweiseitig

Bei einem korrekten Würfel wäre die Wahrscheinlichkeit 26.6%, dass die durchschnittliche Augenzahl bei 40 mal Würfeln mehr als 1.11 Standardabweichungen vom Erwartungswert abweicht.

Wir behalten die Nullhypothese erneut bei.

Nullhypothese: Der Würfel zeigt im Schnitt 3.5 Augen.

Alternativhypothese: Der Würfel zeigt im Durchschnitt mehr oder weniger als 3.5 Augen.



Z-Test

Um den Würfel genauer zu untersuchen, würfelt eine Prüfmaschine diesen 10000-mal und erhält die Augensumme 33290.

Wie wahrscheinlich ist ein solches oder noch niedrigeres Ergebnis, gegeben der Würfel ist in Ordnung?

Ein Casino zählt die Häufigkeiten von Rot und Schwarz an einem amerikanischen Roulettetisch mit 0 und 00.

Von 500 Spins sind nur 218 rot.

Wie wahrscheinlich ist ein solcher oder noch niedrigerer Anteil von Rot, gegeben der Tisch ist in Ordnung?

Z-Test

Die durchschnittliche Augenzahl über 10000 Würfelversuche ist ungefähr normalverteilt mit:

$$\mu=3.5 \text{ und } \sigma^2/10000=0.000291$$

Die Standardabweichung beträgt damit:

$$\sigma_X = 0.0171$$

Mit $X = 3.329$ sind wir also 10 Standardabweichungen unter dem Erwartungswert!

Die Wahrscheinlichkeit ist extrem niedrig und weit unterhalb der 5% Schwelle. Wir verwerfen die Nullhypothese!

~~**Nullhypothese:** Der Würfel zeigt im Schnitt 3.5 Augen.~~

Alternativhypothese: Der Würfel zeigt im Durchschnitt weniger als 3.5 Augen.

Z-Test

Wir finden eine Häufigkeit von Rot von:

$$218/500 = 0.436$$

Die theoretische Wahrscheinlichkeit von Rot ist höher:

$$0.473$$

Um die Normalverteilung einsetzen zu können, benötigen wir noch die Varianz der Zufallsvariablen X_i .

Diese ist gegeben als:

$$X_i = \begin{cases} 1 & \text{wenn Spin } i \text{ rot} \\ 0 & \text{wenn Spin } i \text{ nicht rot} \end{cases}$$

Der Erwartungswert ist 0.473 und die Varianz ist:

$$\begin{aligned} \text{Var}(X) &= 0.473 \cdot (0.527)^2 + 0.473 \cdot (-0.473)^2 \\ &+ 0.054 \cdot (-0.473)^2 = 0.249 \end{aligned}$$

Z-Test

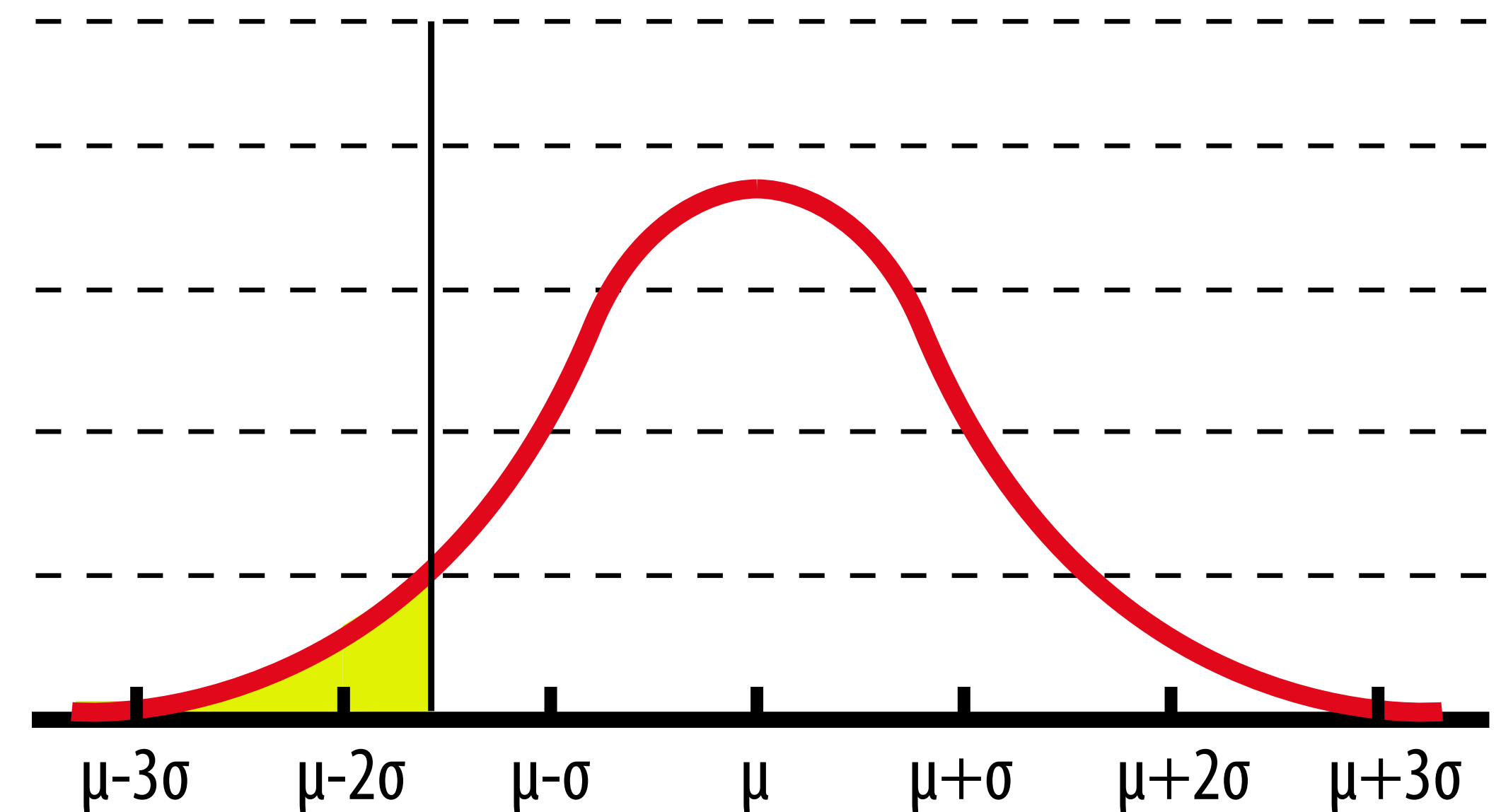
Bei einer Varianz von 0.249 in einem Spin hat der Durchschnitt über 500 Spins die Varianz:

$$0.249/500 = 0.000498$$

Die Standardabweichung ist also:

$$\sigma_x = 0.022315$$

Mit 0.436 sind wir 1.65 Standardabweichungen unter dem Erwartungswert von 0.473

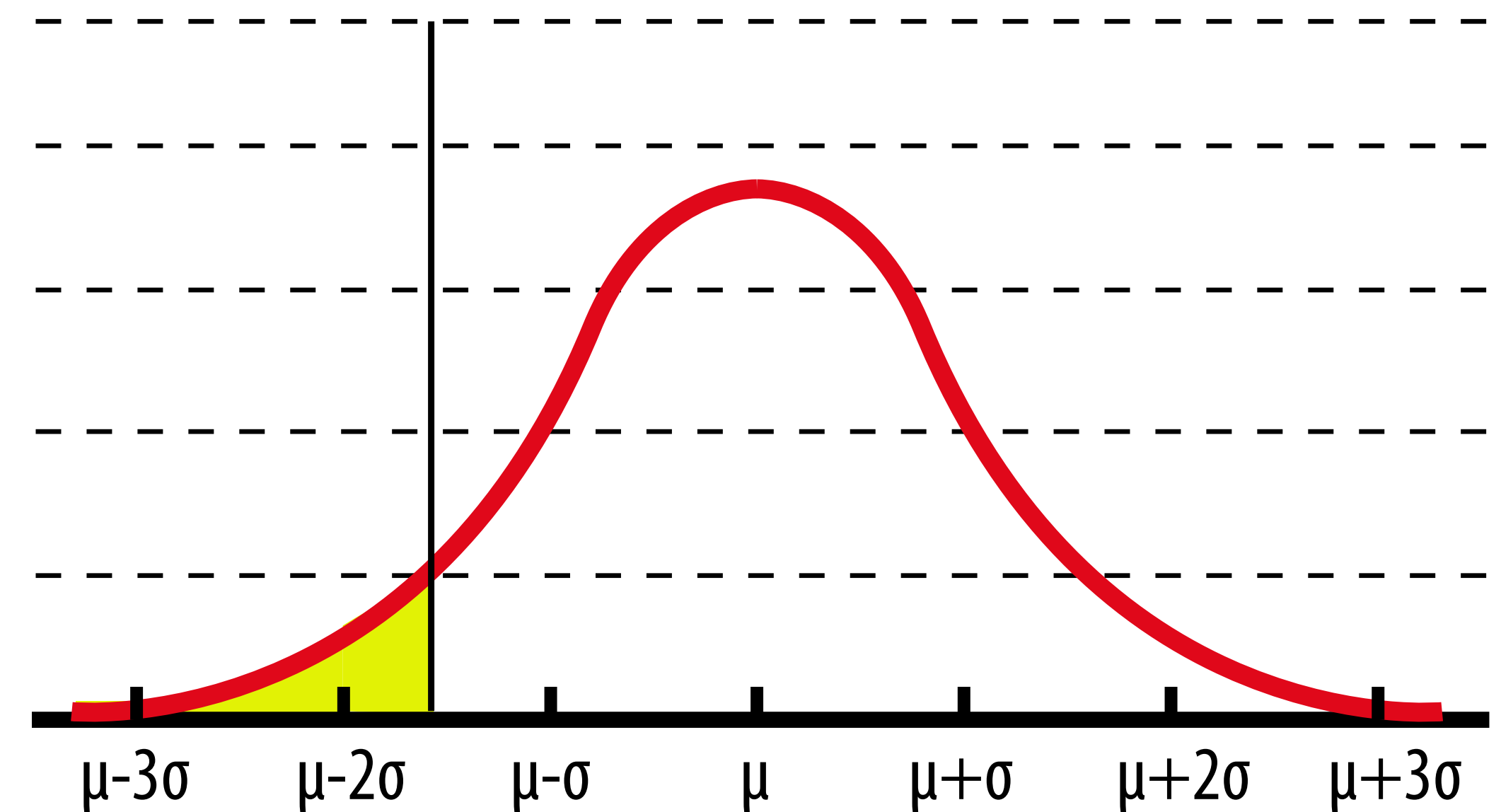


Z-Test

Die Wahrscheinlichkeit ist beim einseitigen Z-Test knapp unterhalb der 5% Schwelle. Wir verwerfen die Nullhypothese!

Nullhypothese: Der Roulettetisch zeigt mit Wahrscheinlichkeit $18/38$ rot.

Alternativhypothese: Der Roulettetisch zeigt mit Wahrscheinlichkeit unter $18/38$ rot.

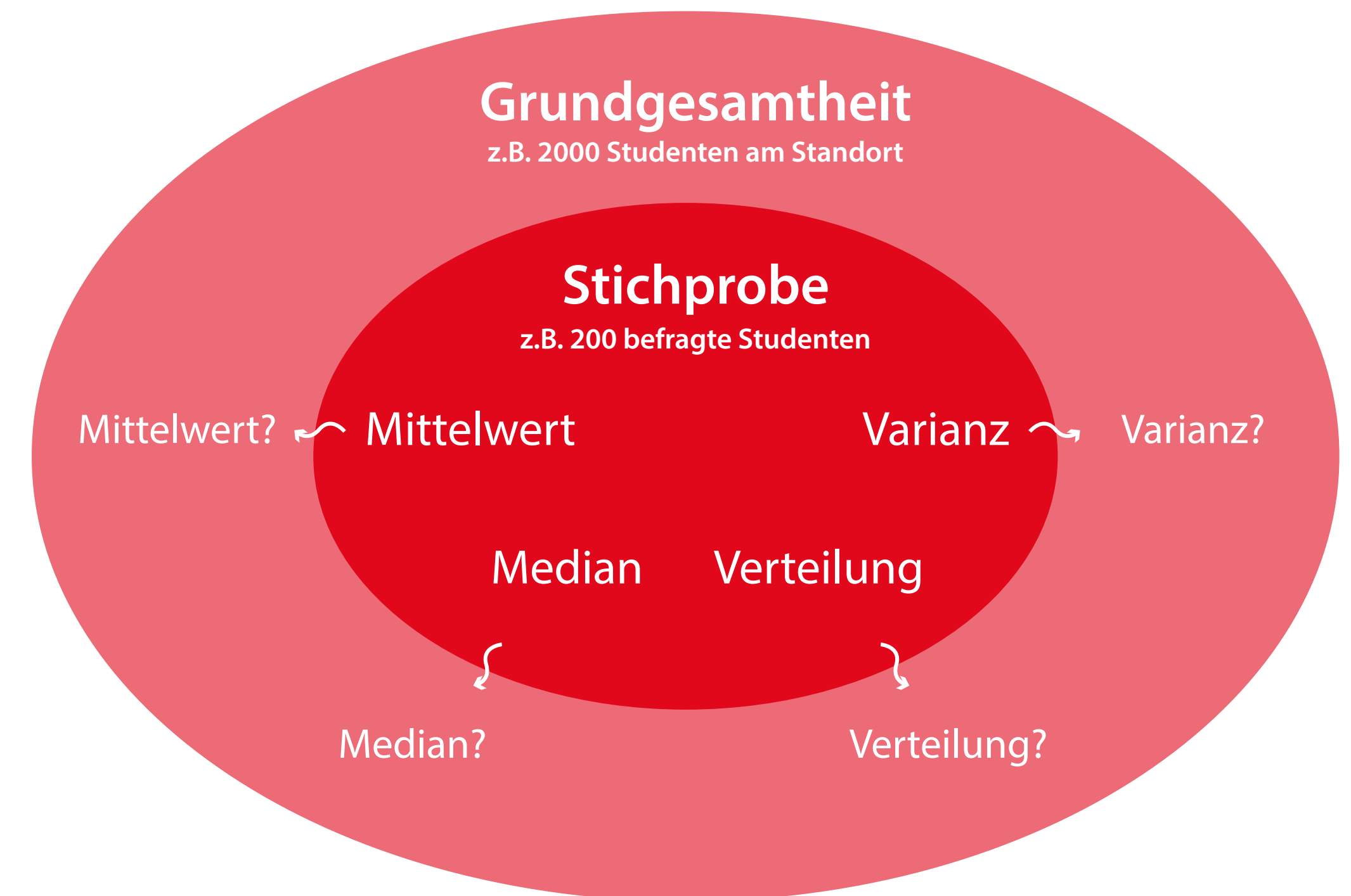


Statistische Tests

Beim Z-Test haben wir die Stichprobe gegen eine Gleichverteilung mit $\mu=3.5$ und $\sigma^2=2.91$ getestet.

Wir können ihn nur anwenden, weil wir wissen, wie sich der Würfel in der Theorie verhalten sollte.

In der empirischen Forschung und auch in Data Science haben wir dieses Wissen oft nicht. Uns stehen nur Informationen über unsere Stichprobe zur Verfügung!



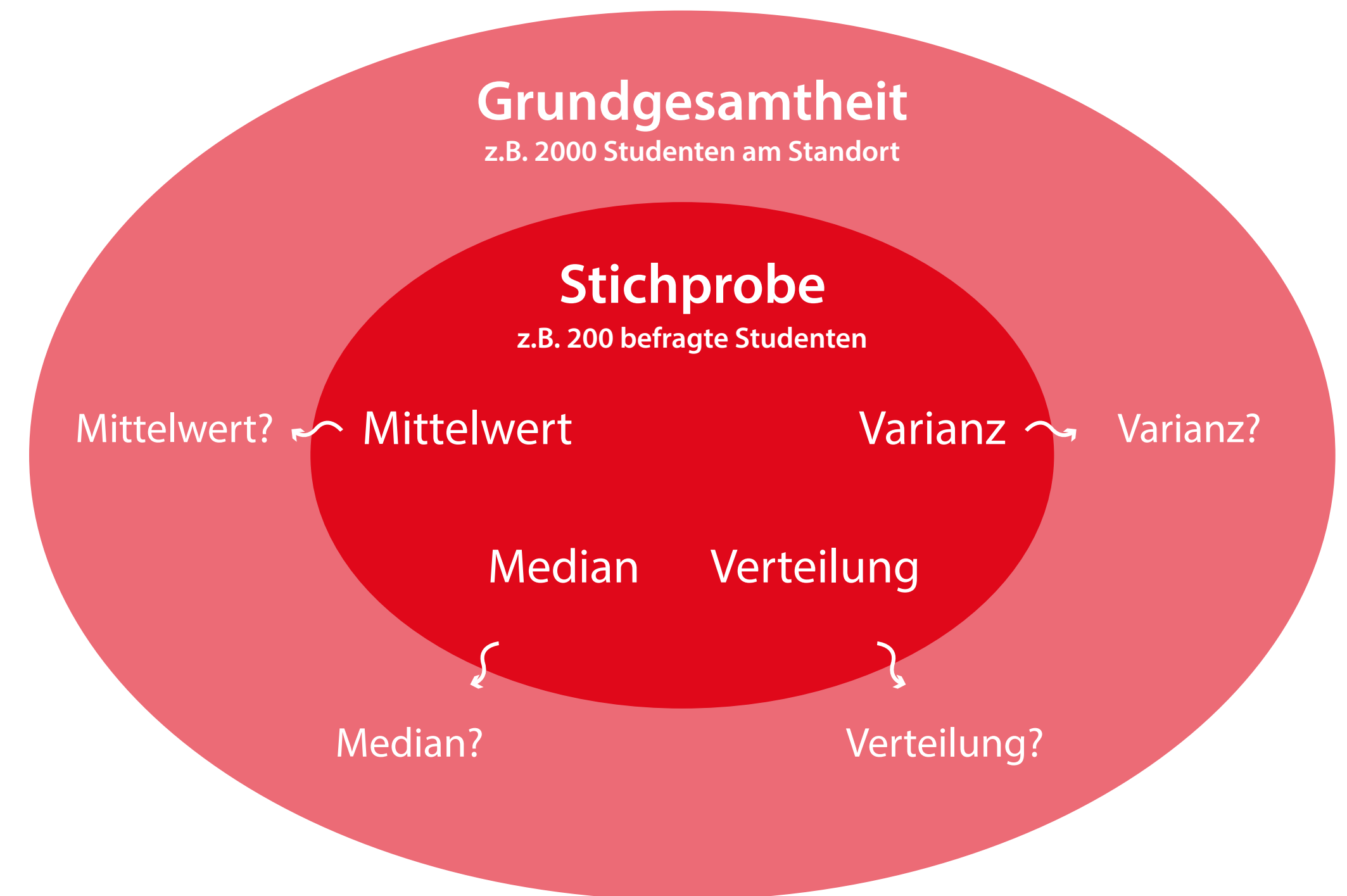
Statistische Tests

Statt die Stichprobe gegen eine aus der Theorie abgeleitete Verteilung zu überprüfen, stellen wir Hypothesen über die Grundgesamtheit auf:

Verteilungshypothesen

Parameterhypothesen

Abhängigkeitshypothesen

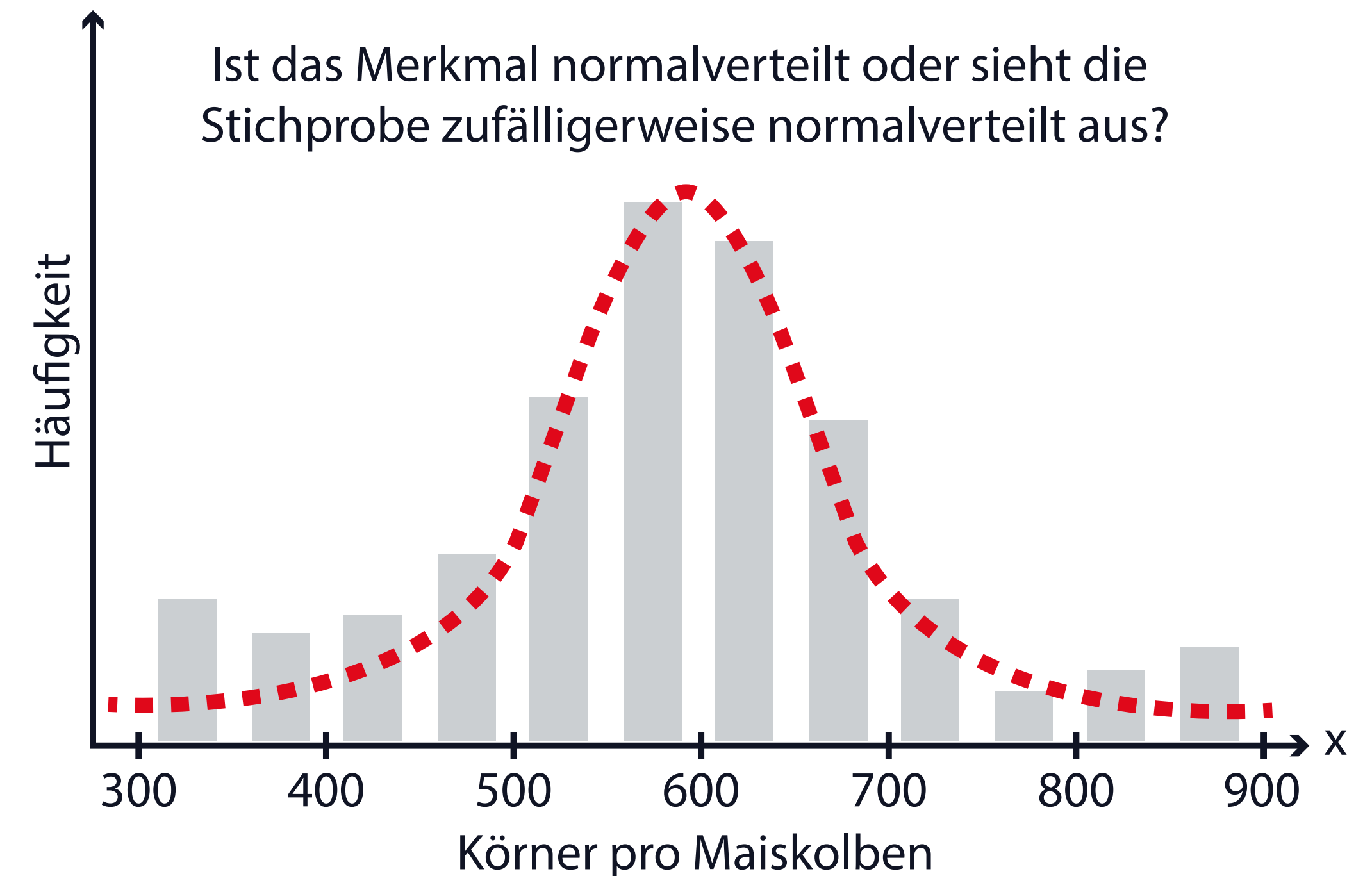


Statistische Tests

Verteilungshypothesen nehmen an, dass ein Merkmal in der Grundgesamtheit einer bestimmten Verteilung folgt.

Da wir nur eine Stichprobe haben, können wir nicht sicher sein, ob deren empirische Verteilung der echten Verteilung in der Grundgesamtheit entspricht.

Mit einem **Verteilungstest** können wir diese Hypothesen überprüfen.

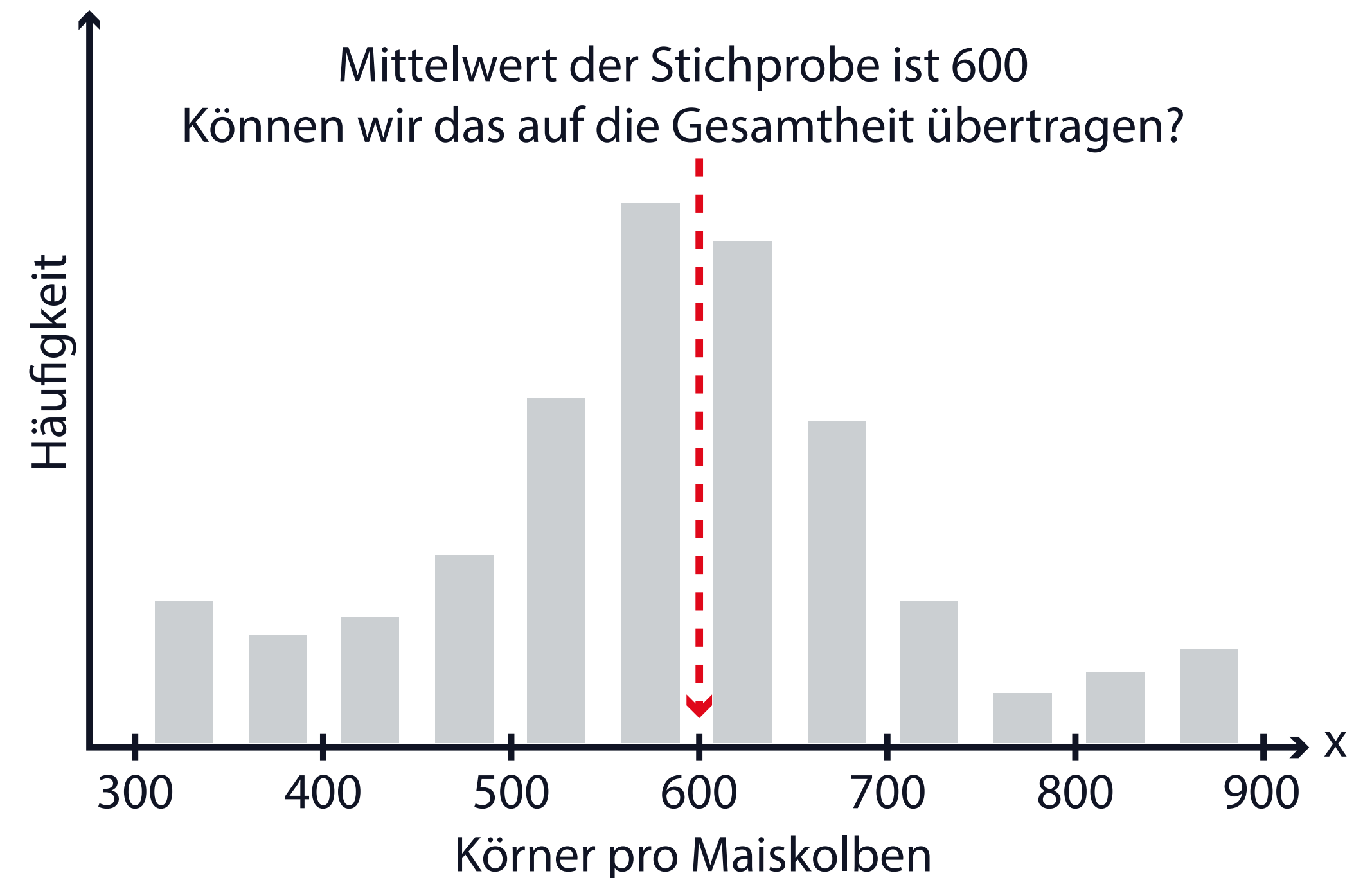


Statistische Tests

Parameterhypothesen nehmen an, dass die Verteilung eines Merkmales in der Grundgesamtheit bestimmte Parameter (Erwartungswert, Varianz, ...) hat.

Da wir nur eine Stichprobe haben, können wir nicht sicher sein, ob deren empirisch gemessene Mittelwerte und Standardabweichungen denen der echten Verteilung entsprechen.

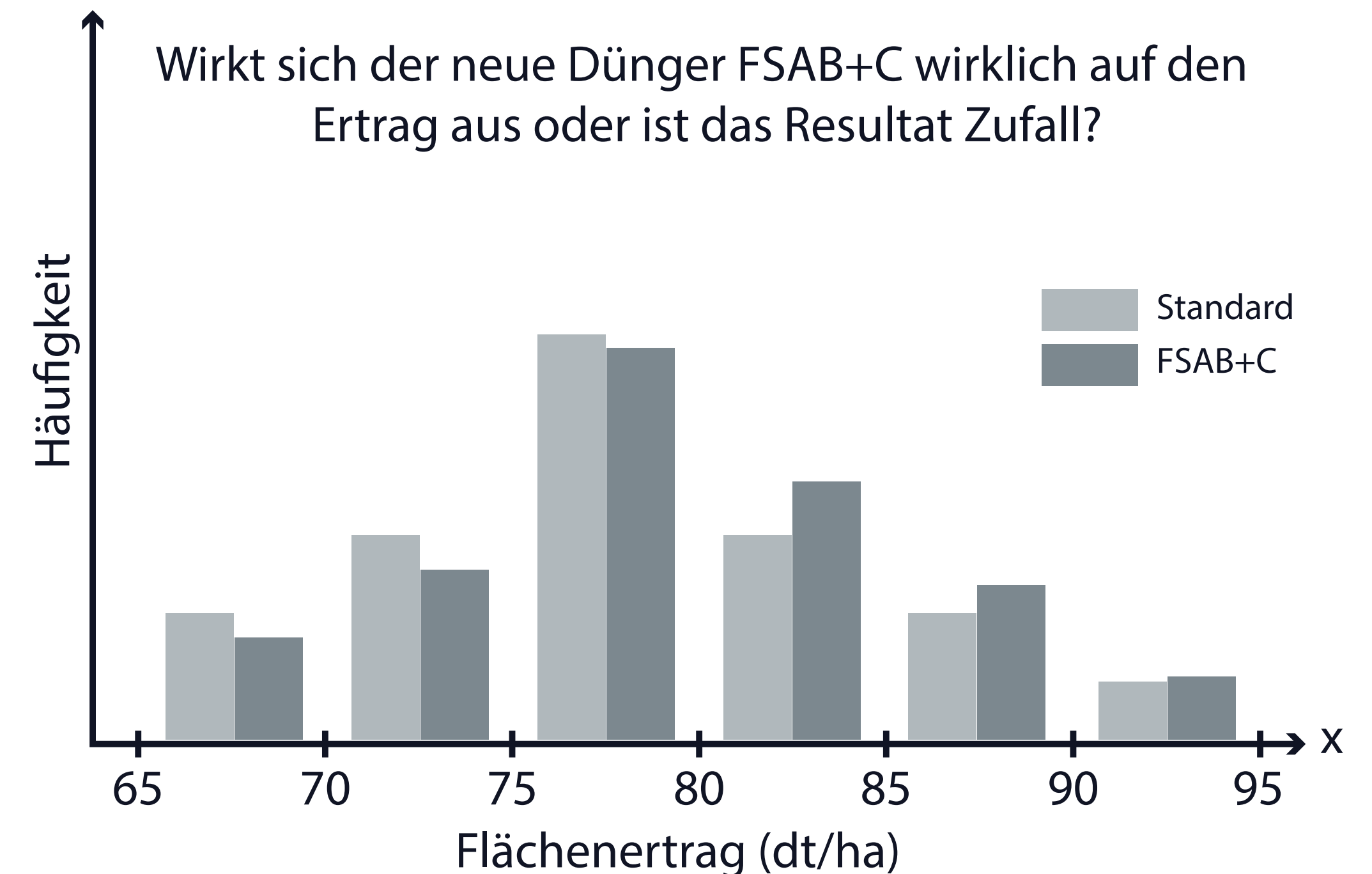
Mit einem **Parametertest** können wir diese Hypothesen überprüfen.



Statistische Tests

Parametertests können auch die Hypothese überprüfen, ob zwei Stichproben aus einer bezüglich eines Merkmals gleichen Grundgesamtheit gezogen wurden.

Fragestellung: Gibt es Unterschiede, z. B. bezüglich des Mittelwerts des Merkmals oder nicht?

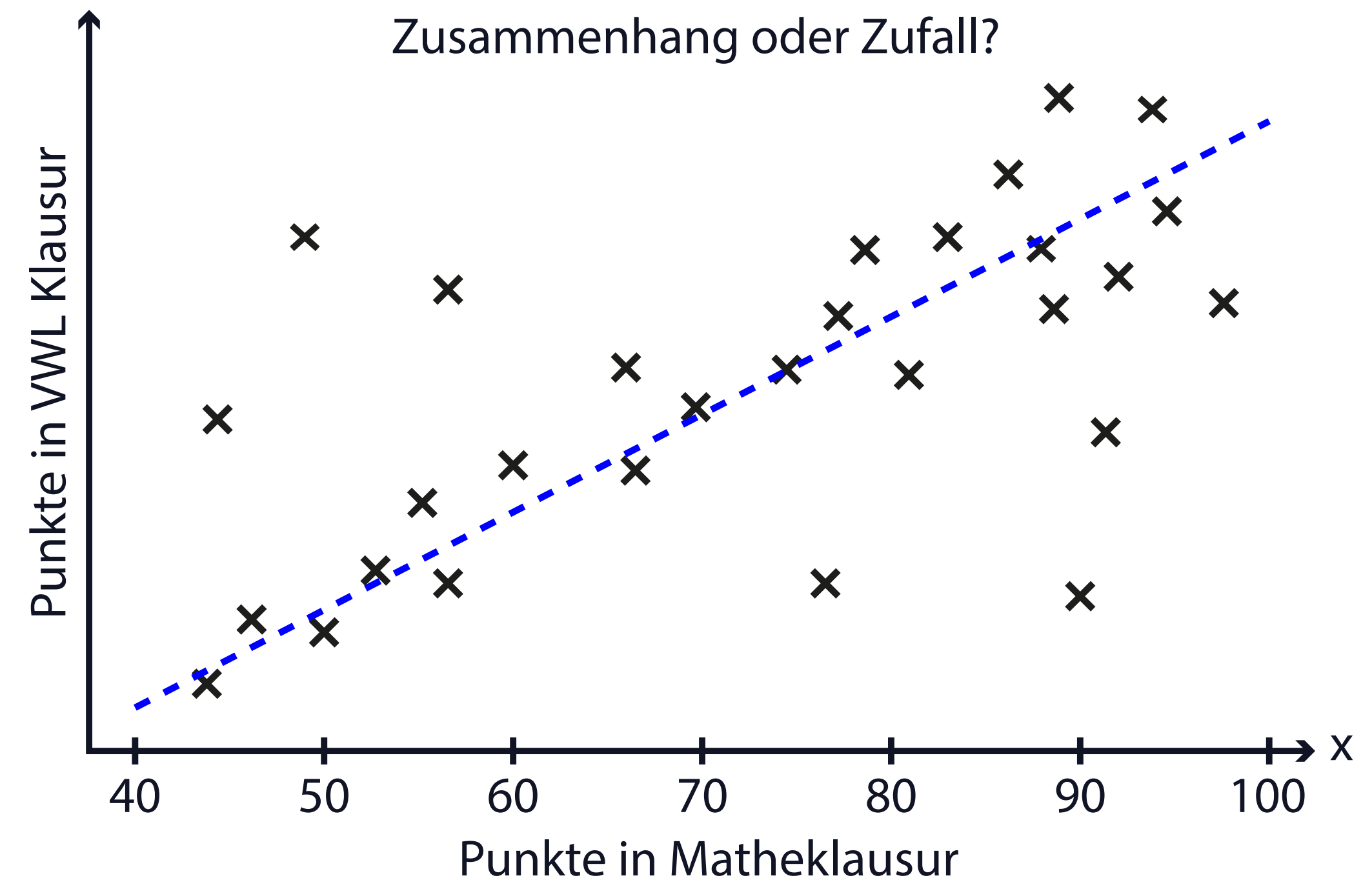


Statistische Tests

Abhängigkeitshypothesen nehmen an, dass zwei Merkmale in der Grundgesamtheit voneinander abhängig sind.

Da wir nur eine Stichprobe haben, können wir nicht sicher sein, ob deren empirisch gemessene Korrelation jener in der Grundgesamtheit entspricht.

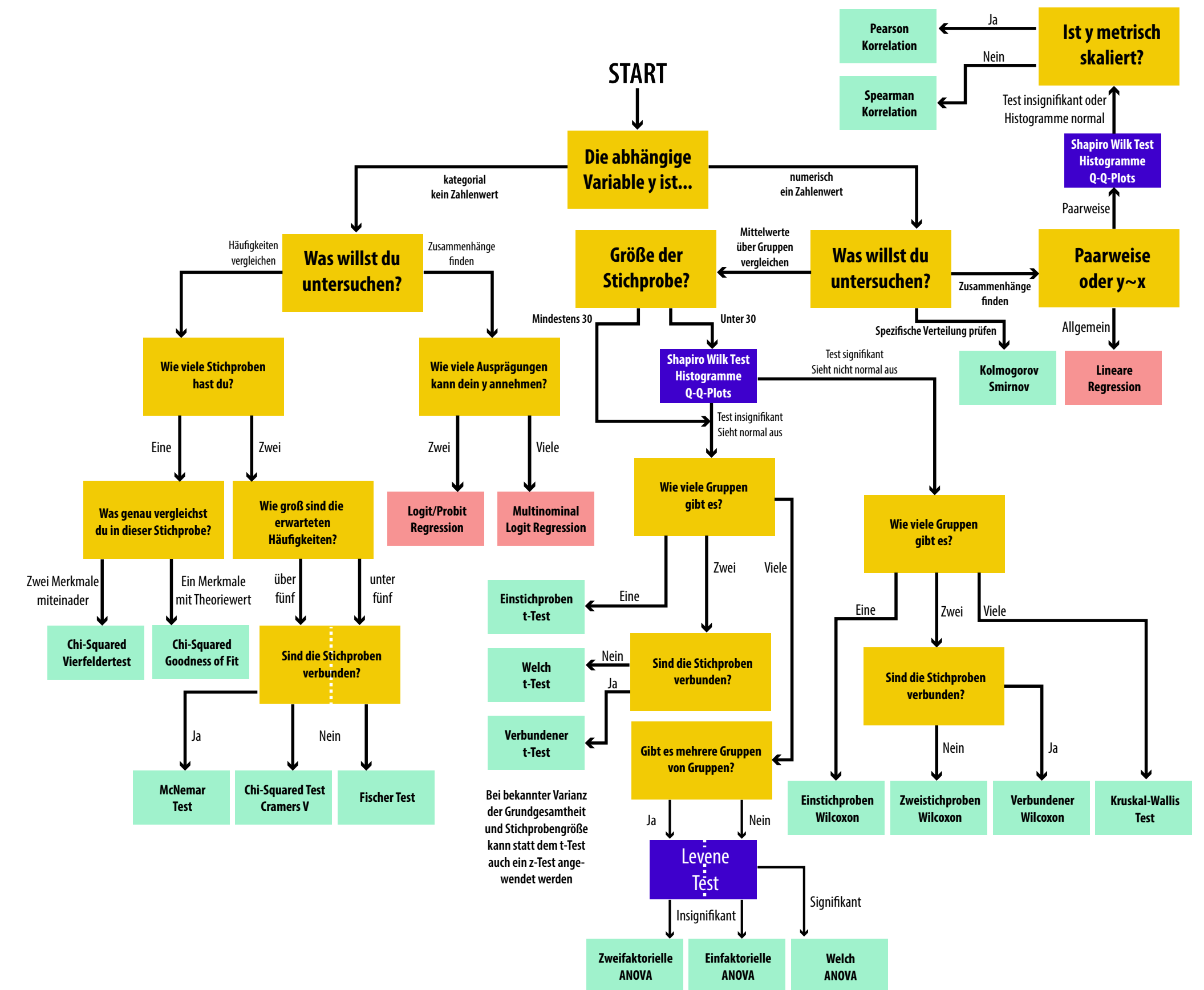
Können wir aus unserer Stichprobe auf eine Unabhängigkeit oder einen Zusammenhang schließen?



Statistische Tests

Es gibt duzende verschiedene statistische Tests! Um die richtige Auswahl zu treffen, müssen wir eine ganze Reihe von Kriterien beachten:

- Wie viele Stichproben haben wir? Eine, zwei mehrere?
- Wie groß ist unsere Stichprobe?
- Auf was wollen wir diese untersuchen?
- Ist unsere abhängige Variable numerisch oder kategorial?
- Ist unsere abhängige Variable normalverteilt?
- Ist unsere unabhängige Variable numerisch oder kategorial?
- Haben wir mehrere unabhängige Variablen?



t-Test

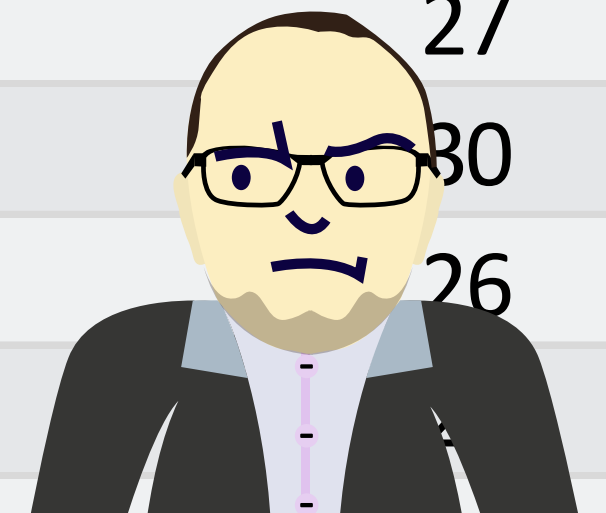
Die Rechenarbeit hinter statistischen Tests ist aufwendig, wird uns jedoch von Statistiksoftware abgenommen. Unsere Aufgabe ist es, den richtigen Test zu finden und die Ergebnisse richtig zu interpretieren!

Beispiel: Ein Studiengangsleiter schaut sich die Leistung von WI- und Data Science Studierenden in verschiedenen Vorlesungen an und möchte ...

... WI- und Data Science Studierenden vergleichen.

... die Schwierigkeit von Marketing und Analysis vergleichen.

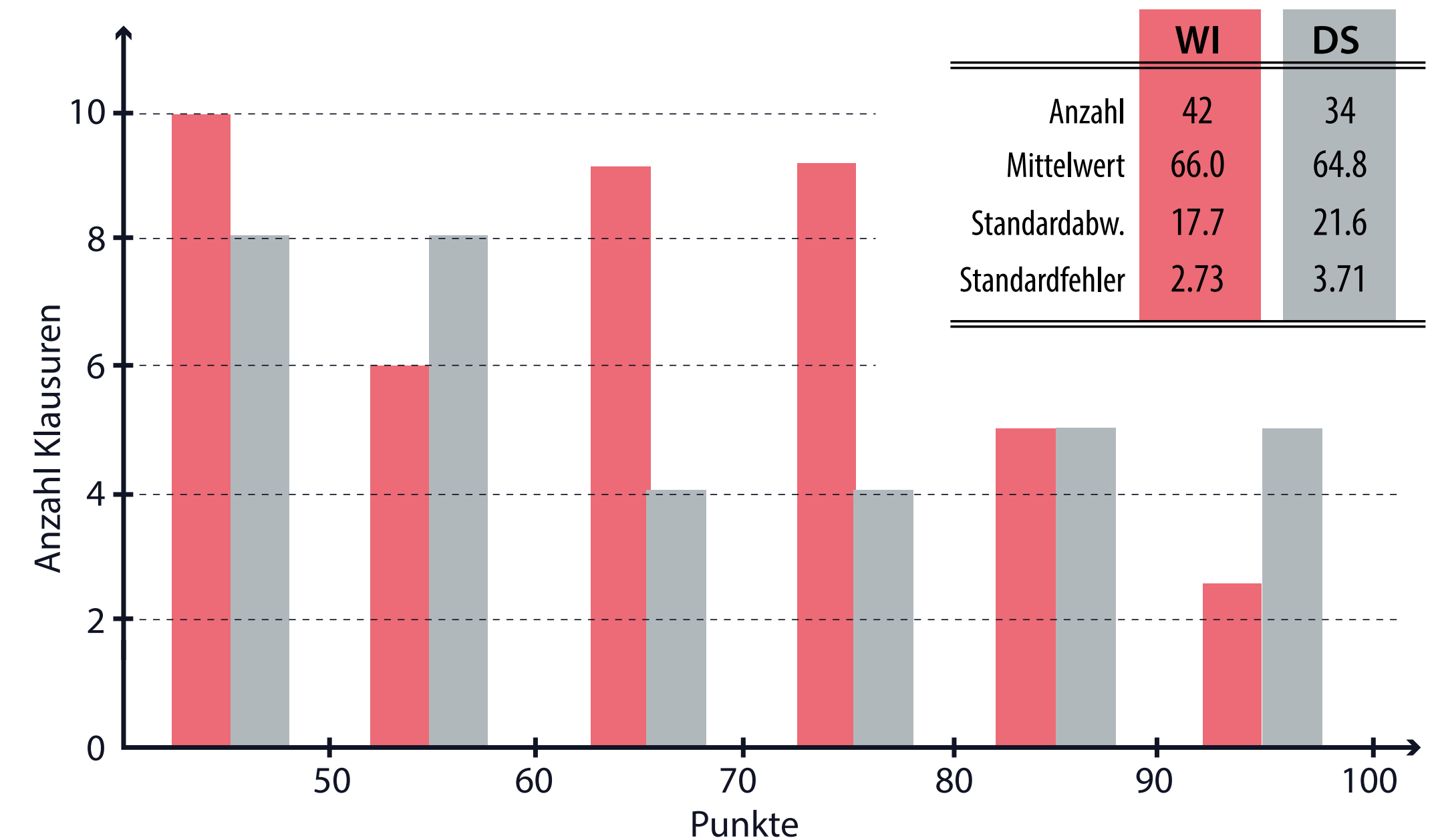
Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



t-Test

Die Wirtschaftsinformatiker zeigen einen Punktedurchschnitt von 66.0 Punkten. Data Science kommt dagegen nur auf 64.8 Punkte.

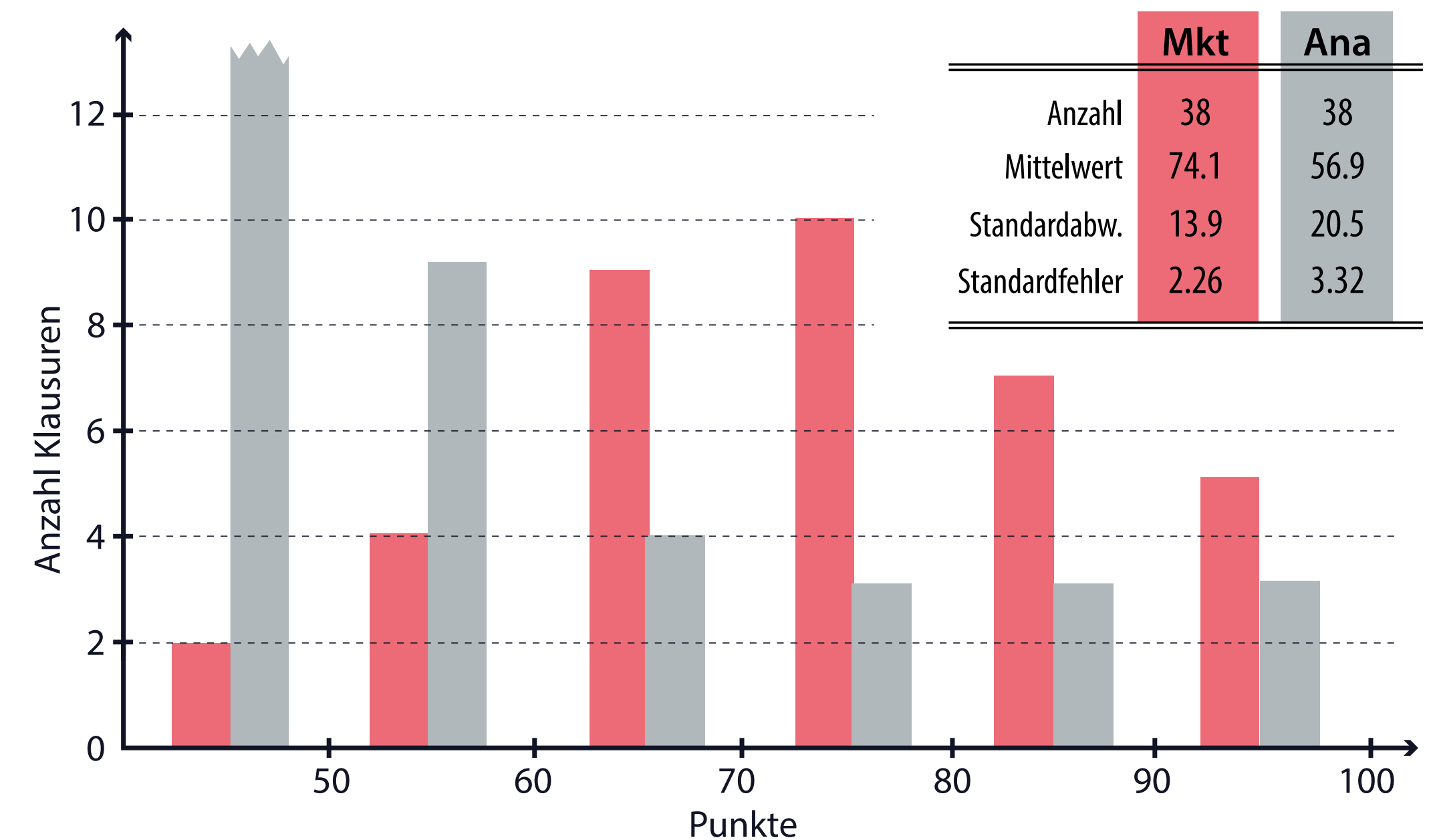
In der Stichprobe sind die Wirtschaftsinformatiker um 1.2 Punkte besser, aber lässt sich das verallgemeinern?



t-Test

In Marketing liegt der Durchschnitt bei 74.1 Punkten. In der Analysis sind es nur 56.9 Punkte.

In der Stichprobe ist der Schnitt bei Marketing um 17.2 Punkte besser, aber lässt sich das verallgemeinern?



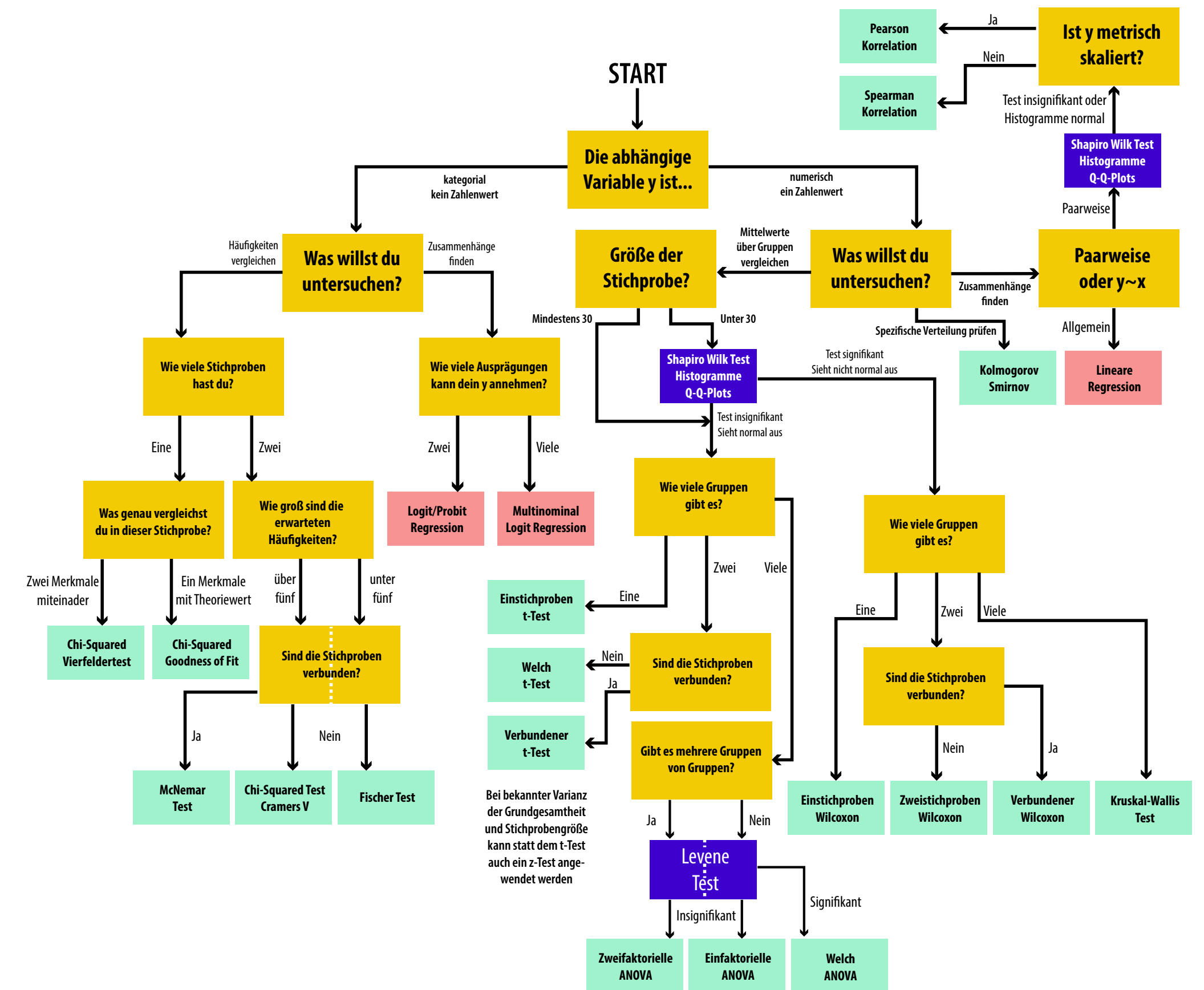
t-Test

Um die Frage der möglichen Verallgemeinerung zu untersuchen, verwenden wir hier den zweistichproben t-Test.

Der zweistichproben t-Test untersucht, ob wir aus den Mittelwerten zweier Stichprobe auf unterschiedliche Erwartungswerte der wahren Verteilungen schließen können.

H0 - Es gibt keinen Unterschied im Erwartungswert der beiden Gruppen.

H1 - Es gibt einen Unterschied im Erwartungswert der beiden Gruppen.



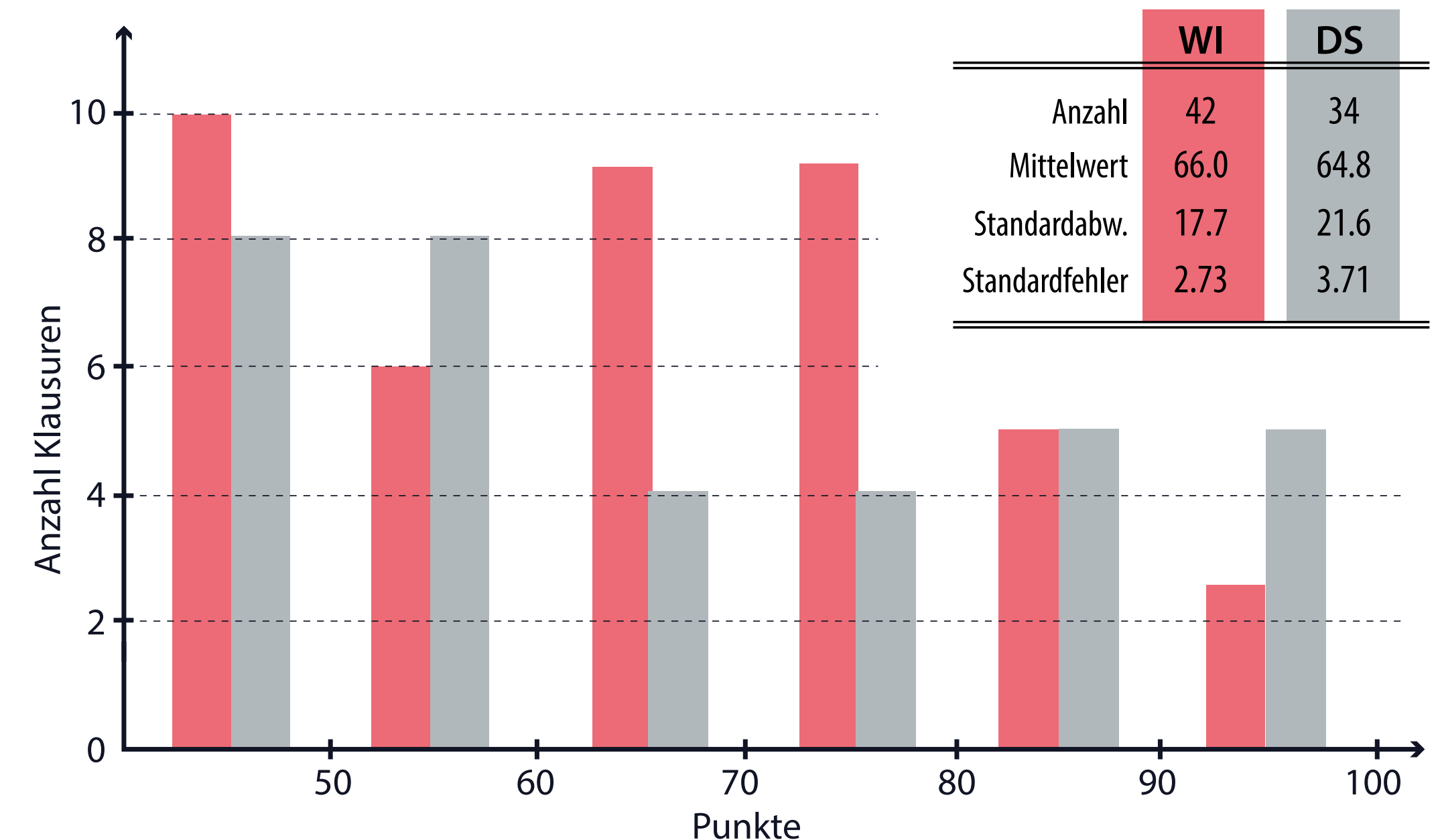
t-Test

Beim Vergleich der Studiengänge haben wir in der Stichprobe ein um 1.2 Punkte besseres Leistungsniveau bei WI festgestellt.

Mit dem zweistichproben t-Test untersuchen wir folgendes Hypothesenpaar:

H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Studiengänge.

H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Studiengänge.



Gepaarte Stichproben

Wir verwenden die Variante für nicht gepaarte Stichproben, weil zwischen den Stichproben kein Zusammenhang besteht.

Würden wir die beiden Stichproben nebeneinanderlegen, gäbe es keinen Zusammenhang zwischen den jeweils ersten, zweiten, dritten usw. Messungen.

Der erste WI-Student und der erste DSKI-Student sind unterschiedliche Personen!

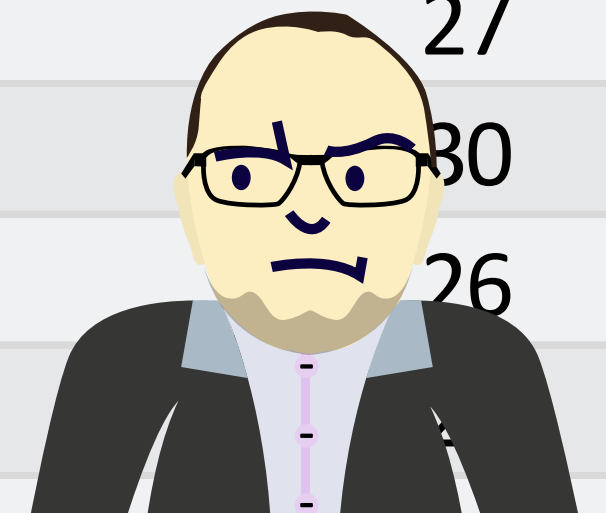
WI	Zusammenhang?	Data Science
50	 X 	58
42	 X 	58
66	 X 	54
62	 X 	60
60	 X 	66
70	 X 	62
66	 X 	74
72	 X 	68

Interpretation p-Wert

Zentrales Ergebnis des t-Tests ist der p-Wert. Für den Vergleich der Studiengänge ist er 0.8091.

Interpretation: Wäre die Nullhypothese wahr, dann erhalten wir bei der gegebenen Stichprobengröße und der gegebenen Varianz mit Wahrscheinlichkeit 80.91% einen Leistungsunterschied von mindestens 1.2 Punkten.

Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



Interpretation p-Wert

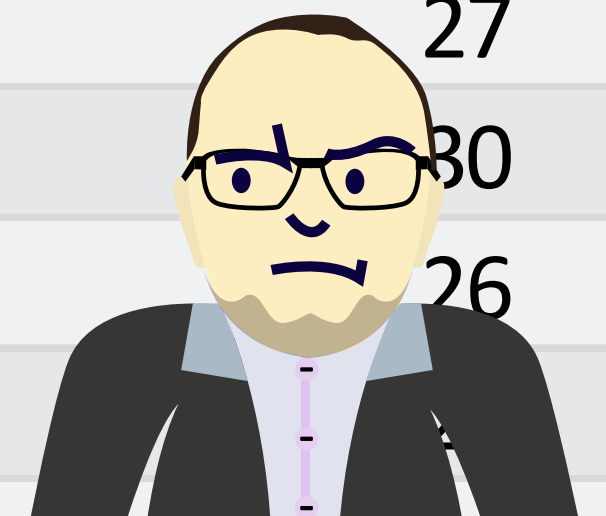
Selbst ohne Unterschiede im Leistungsniveau ist es sehr wahrscheinlich, einen kleinen Unterschied zu messen!

Wir verwerfen die Nullhypothese nur dann, wenn der p-Wert unter 5% liegt. Hier ist er deutlich darüber und deshalb behalten wir die Nullhypothese bei.

H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Studiengänge.

H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Studiengänge.

Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



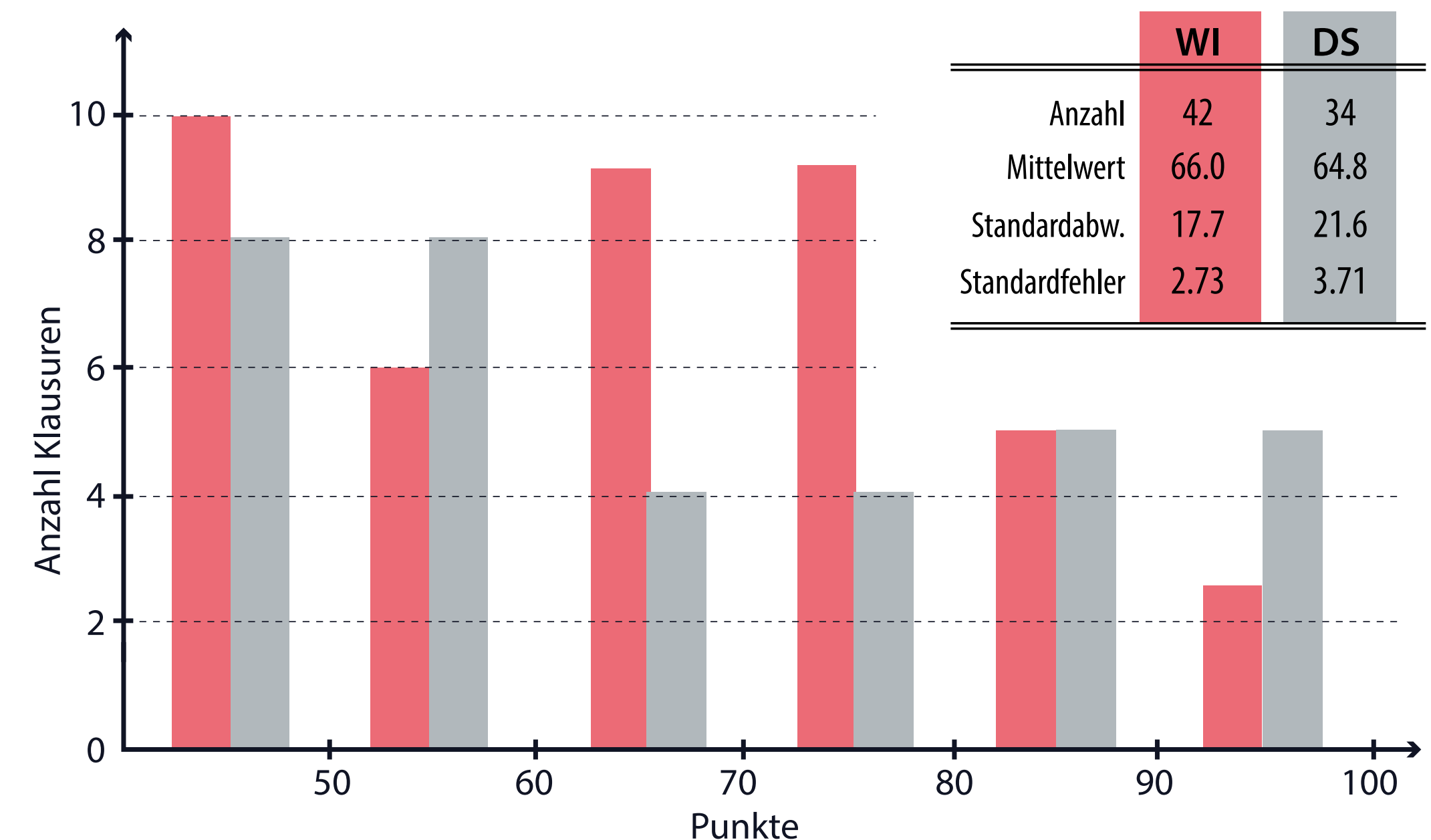
Interpretation p-Wert

Beim Vergleich der Vorlesungen haben wir in der Stichprobe ein um 17.2 Punkte besseres Leistungsniveau in Marketing festgestellt.

Mit dem zweistichproben t-Test untersuchen wir folgendes Hypothesenpaar:

H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Vorlesungen.

H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Vorlesungen.



Gepaarte Stichproben

Wir verwenden die Variante für gepaarte Stichproben, weil zwischen den Stichproben ein Zusammenhang besteht.

Würden wir die beiden Stichproben nebeneinanderlegen, gäbe es einen Zusammenhang zwischen den jeweils ersten, zweiten, dritten usw. Messungen.

Es handelt sich um Prüfungsleistung derselben Person!

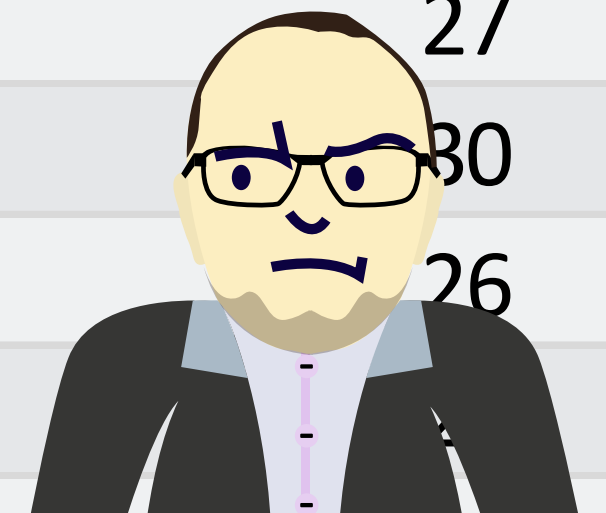
Marketing	Zusammenhang?	Analysis
50		38
42		48
66		30
62		44
60		46
70		40
66		44
72		42

Interpretation p-Wert

Dieses Mal beträgt der p-Wert 0.000000002335.

Wäre die Nullhypothese wahr, dann erhalten wir bei der gegebenen Stichprobengröße und der gegebenen Varianz mit Wahrscheinlichkeit 0.0000002335% einen Leistungsunterschied von mindestens 17.2 Punkten.

Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



Interpretation p-Wert

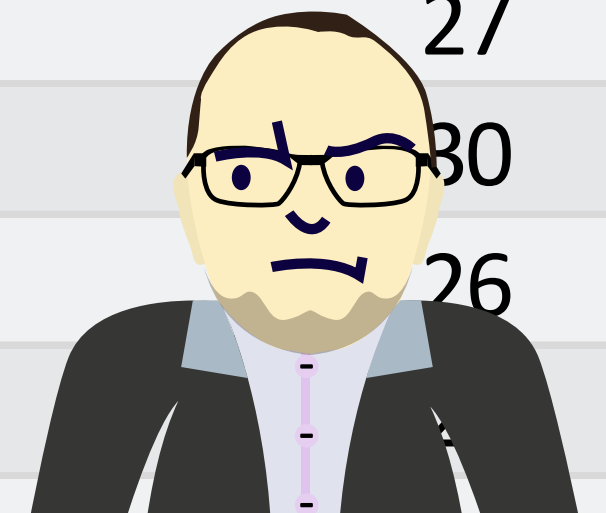
Wir können es nicht ausschließen, aber dass dieser Unterschied durch zufällige Schwankungen zustande kommt, ist extrem unwahrscheinlich.

Wir verwerfen die Nullhypothese, denn, der p-Wert liegt unter 5%.

~~H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Vorlesungen.~~

H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Vorlesungen.

Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



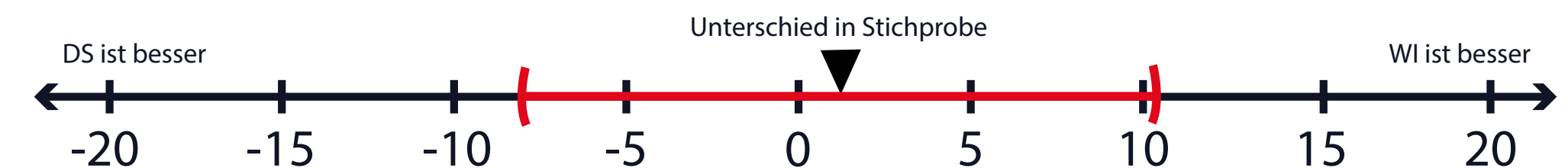
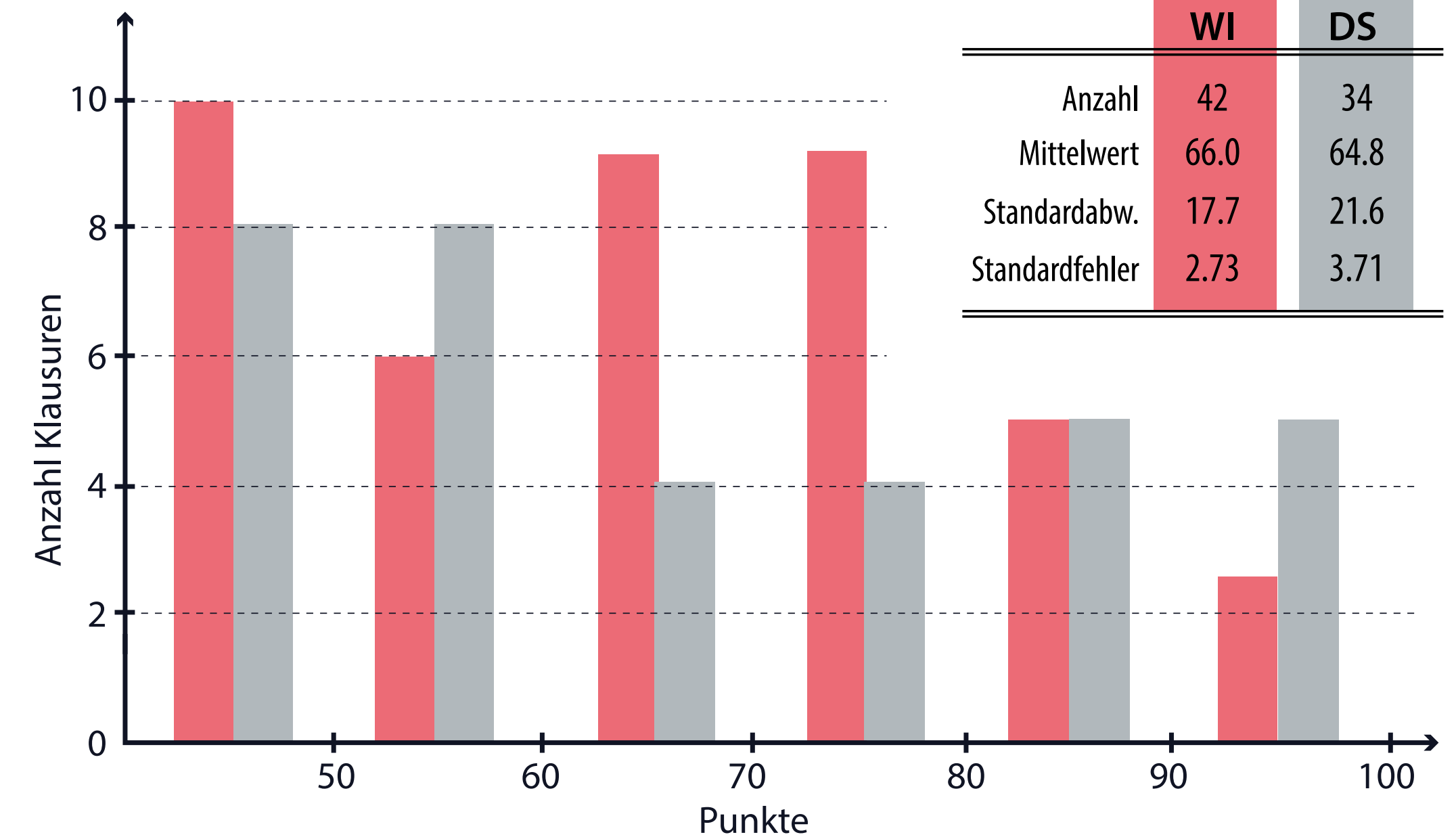
Konfidenzintervalle

Zusätzlich zum p-Wert erhalten wir ein Konfidenzintervall auf einem bestimmten Konfidenzniveau (Standard: 95%).

Beim Vergleich der Studiengänge erhalten wir:

[-8.08 , 10.32]

Der Test ist sich zu 95% sicher, dass der wahre Unterschied in diesem Bereich liegen muss.



Zu 95.0% liegt der wahre Unterschied zwischen -8.08 und 10.32

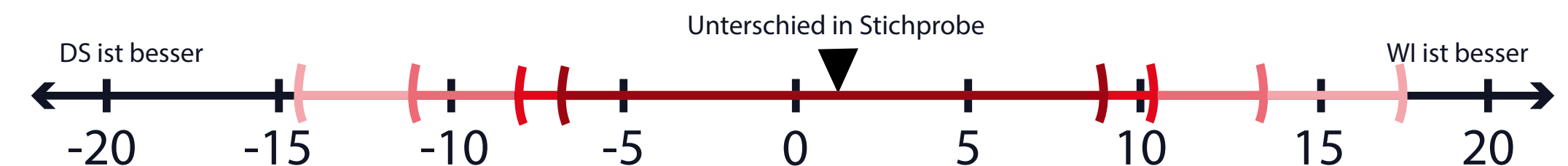
Konfidenzintervalle

Je größer das geforderte Konfidenzniveau, umso breiter wird das Konfidenzintervall.

Der p-Wert ist eins minus das höchste Konfidenzniveau in dem die 0 gerade noch nicht berührt wird.

Wenn das 95% Konfidenzintervall die 0 nicht enthält, dann ist der p-Wert auf jeden Fall kleiner als 5%.

Hier ist dies nicht der Fall, also behalten wir die Nullhypothese bei.



Zu 90.0% liegt der wahre Unterschied zwischen -6.57 und 8.81

Zu 95.0% liegt der wahre Unterschied zwischen -8.08 und 10.32

Zu 99.0% liegt der wahre Unterschied zwischen -11.1 und 13.35

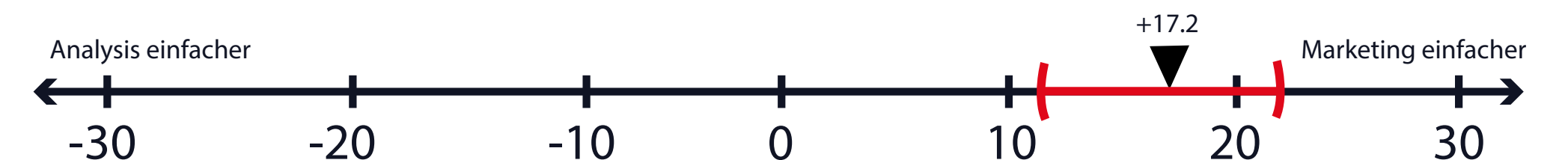
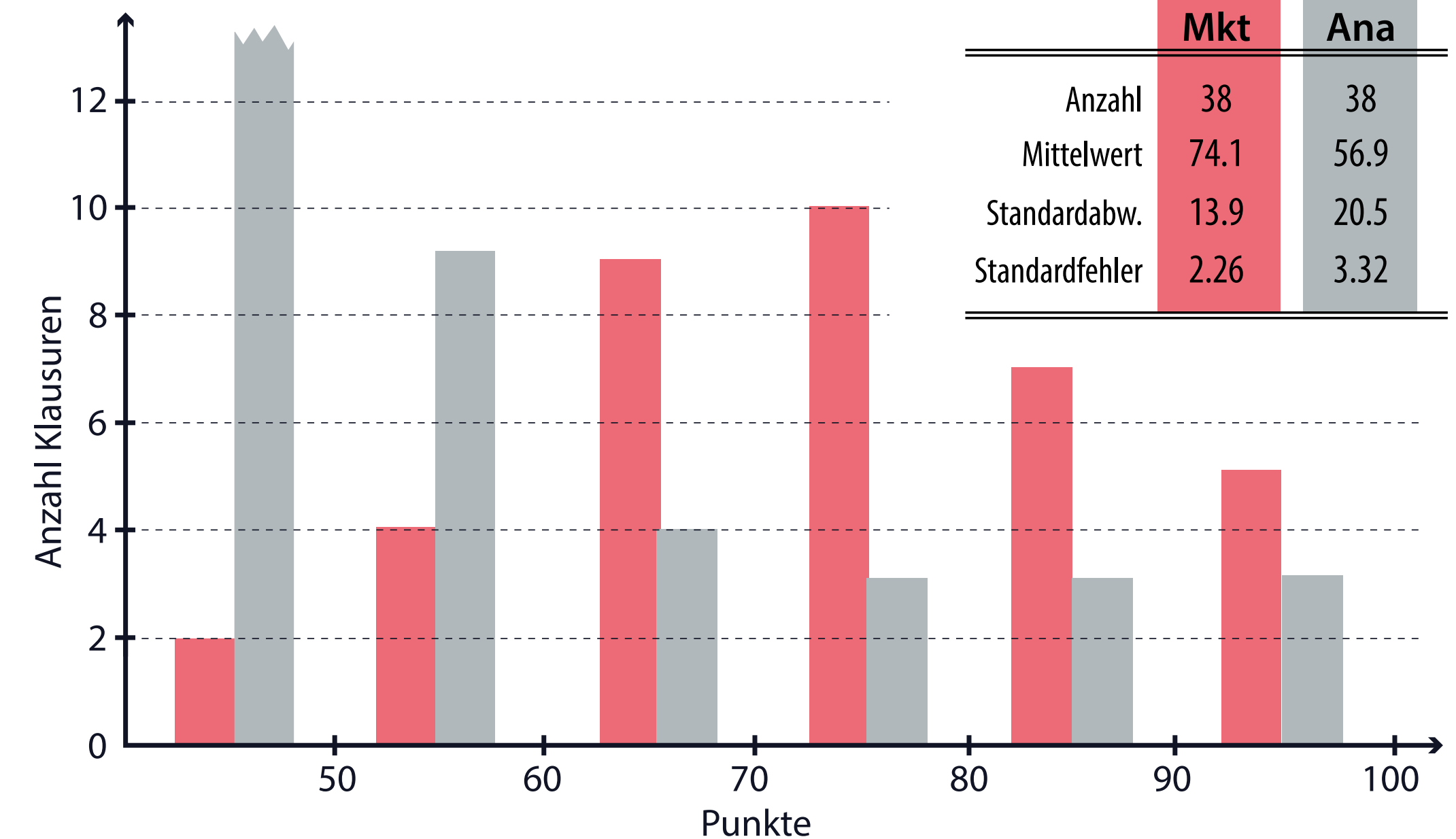
Zu 99.9% liegt der wahre Unterschied zwischen -14.7 und 17.01

Konfidenzintervalle

Beim Vergleich der Vorlesungen ist die 0 nicht im 95% Konfidenzintervall enthalten.

[12.75 , 21.67]

Der Test ist sich zu 95% sicher, dass der wahre Unterschied nicht 0 ist und damit können wir die Nullhypothese verwerfen!



Zu 95.0% liegt der wahre Unterschied zwischen 12.75 und 21.67

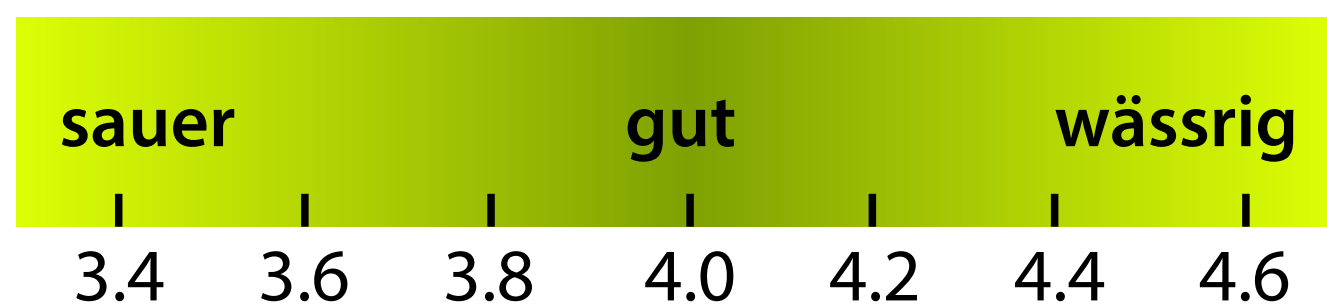
Einstichproben t-Test

Der Einstichproben t-Test überprüft, ob der Mittelwert eines Merkmals einem Zielwert entspricht.

Dazu vergleicht er den Mittelwert einer Stichprobe mit einem exogen vorgegebenen Zielwert. Das generische Hypothesenpaar ist:

H0 - Der wahre Mittelwert entspricht dem Zielwert.

H1 - Der wahre Mittelwert weicht vom Zielwert ab.

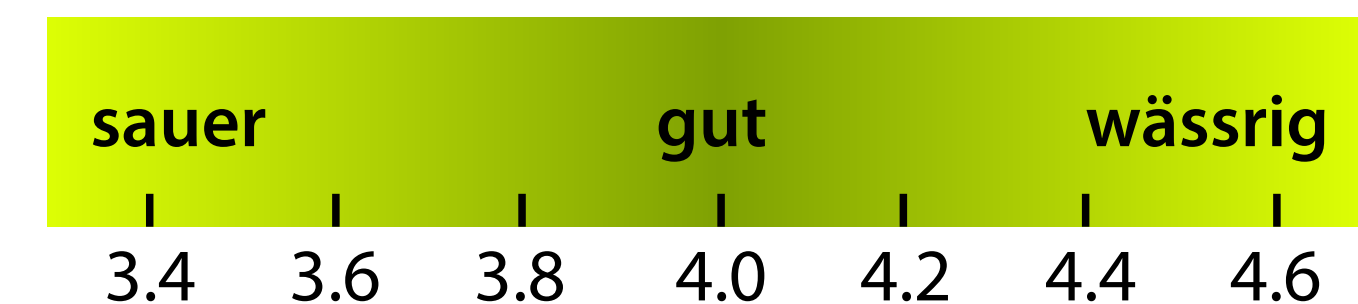


Einstichproben t-Test

Wir betrachten eine Brauerei, welche die Qualität ihres Biers stichprobenartig in einem Labor kontrolliert.

Einer der dort gemessenen Parameter ist der pH-Wert. Der Zielwert ist ein pH-Wert von 4.0.

Bei einem Naturprodukt wie Bier ist eine gewisse Schwankung, um diesen Wert nicht zu verhindern, aber wir wollen wissen, ob wir im Durchschnitt ein Bier mit pH-Wert von 4.0 produzieren!

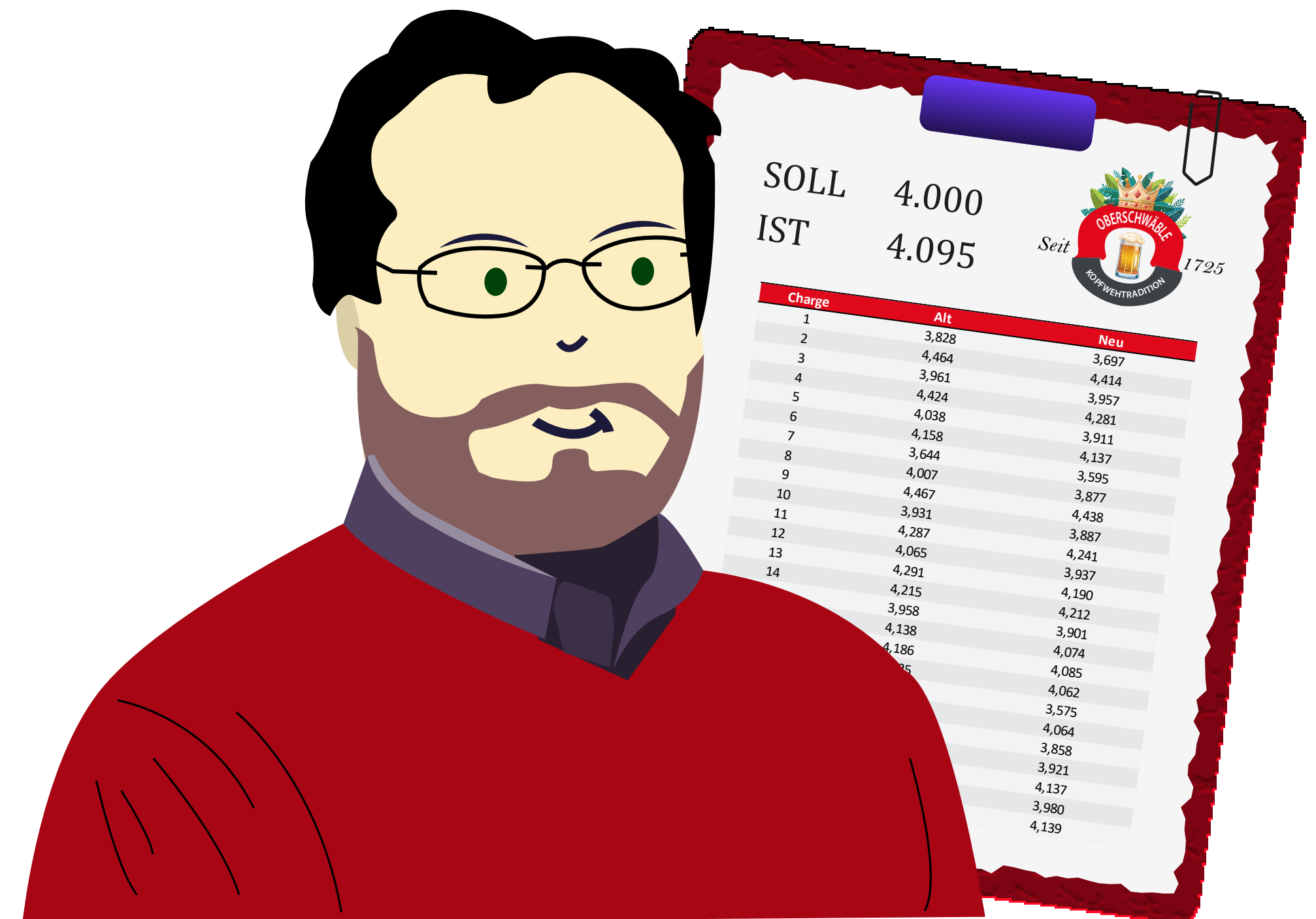


Einstichproben t-Test

Wir nehmen Stichproben aus 25 Chargen unseres Biers und erhalten die rechts gezeigten Werte.

Über alle Chargen gemittelt beträgt der pH-Wert 4.095 und damit etwas mehr als der Sollwert 4.0.

Die Frage ist, ob diese Abweichung signifikant ist. Falls ja, könnten wir unseren Brauprozess in die Richtung mehr Säure anpassen!

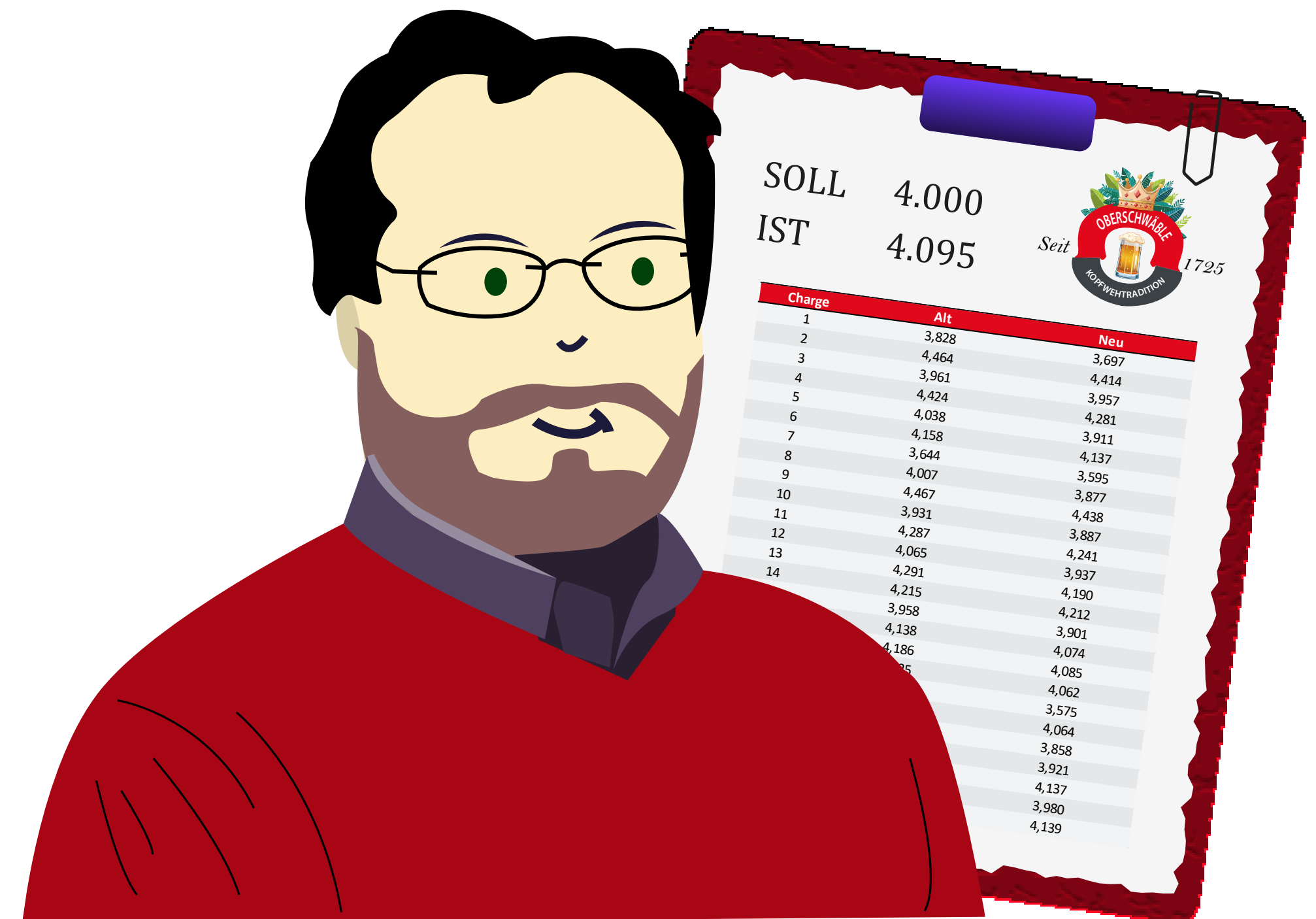


Einstichproben t-Test

Wir wenden einen zweiseitigen einstichproben t-Test an.
 Das Hypothesenpaar dazu wäre:

Nullhypothese H0 Der pH-Wert des gebrauten Bieres entspricht dem Sollwert von 4.0.

Alternativhypothese H1 Der pH-Wert des gebrauten Bieres liegt über oder unter dem Sollwert von 4.0.



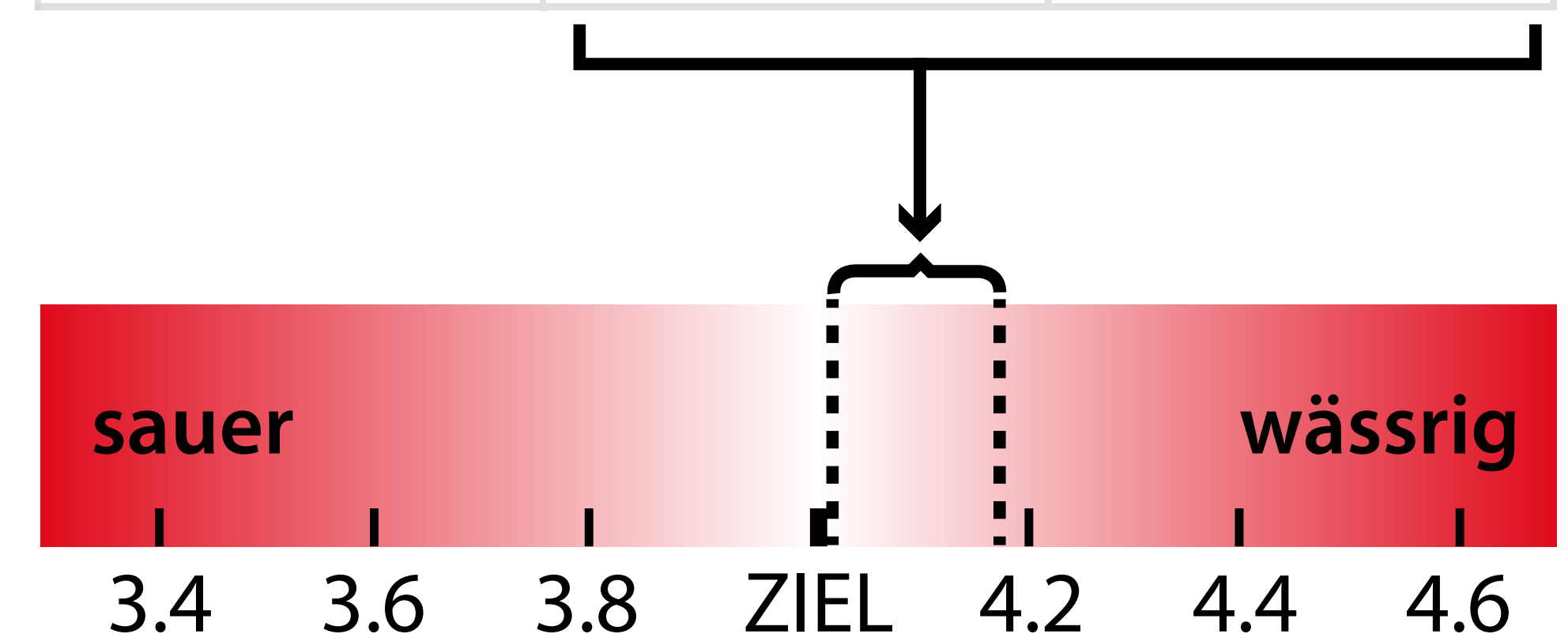
Einstichproben t-Test

Der p-Wert des einstichproben t-Test beträgt 0.0406 und zeigt, dass die gemessene Abweichung von 0.095 signifikant ist.

Wäre die Nullhypothese wahr, erhalten wir gegeben der Stichprobengröße und der Varianz mit Wahrscheinlichkeit von 4.06% eine Abweichung von mindestens 0.095 Skalenpunkten.

Wir können die Nullhypothese verwerfen!

Einstichproben T-Test		
T-Wert	2,16	
P-Wert	0,040618	
Konfid. (95%)	4,00440018	4,18543659



Einstichproben t-Test

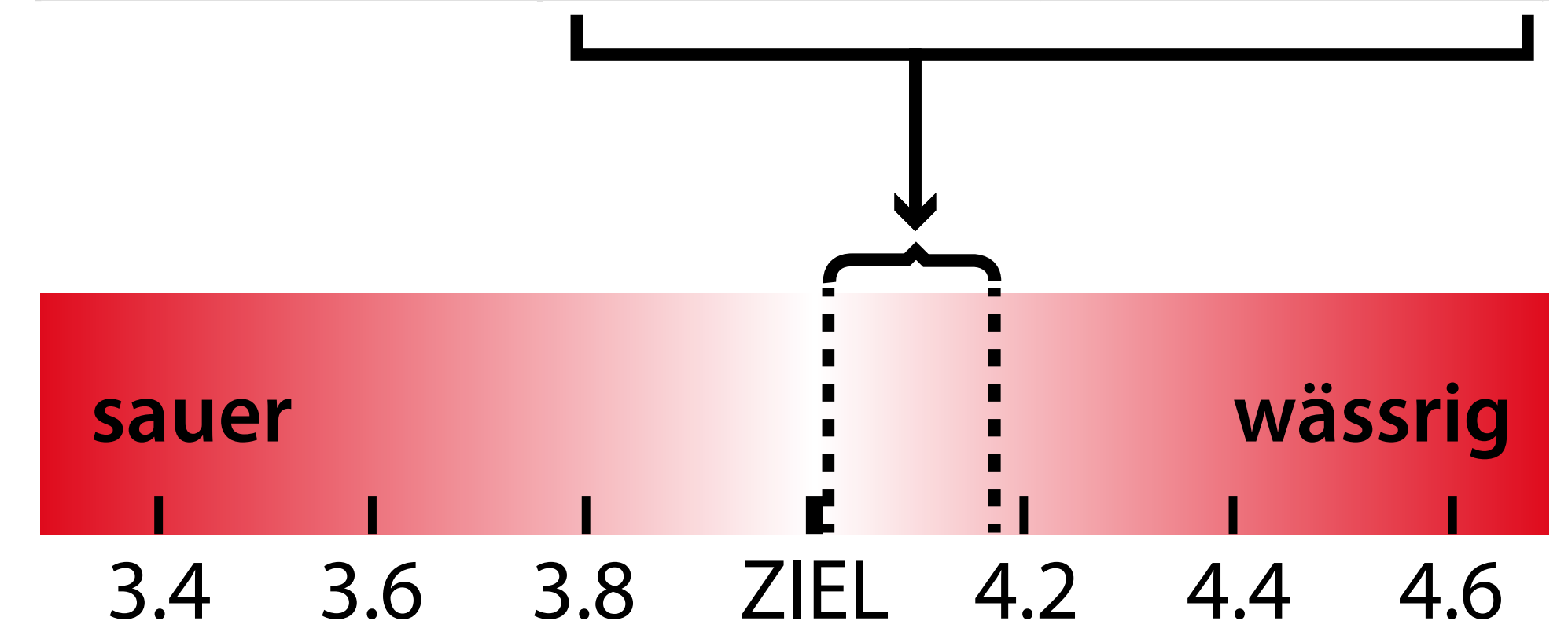
Das 95% Konfidenzintervall beträgt:

[4.004 , 4.185]

Der Test ist sich zu 95% sicher, dass der wahre pH-Wert in diesem Bereich ist und damit nicht dem Zielwert 4.000 entspricht.

Wir können die Nullhypothese verwerfen!

Einstichproben T-Test		
T-Wert	2,16	
P-Wert	0,040618	
Konfid. (95%)	4,00440018	4,18543659



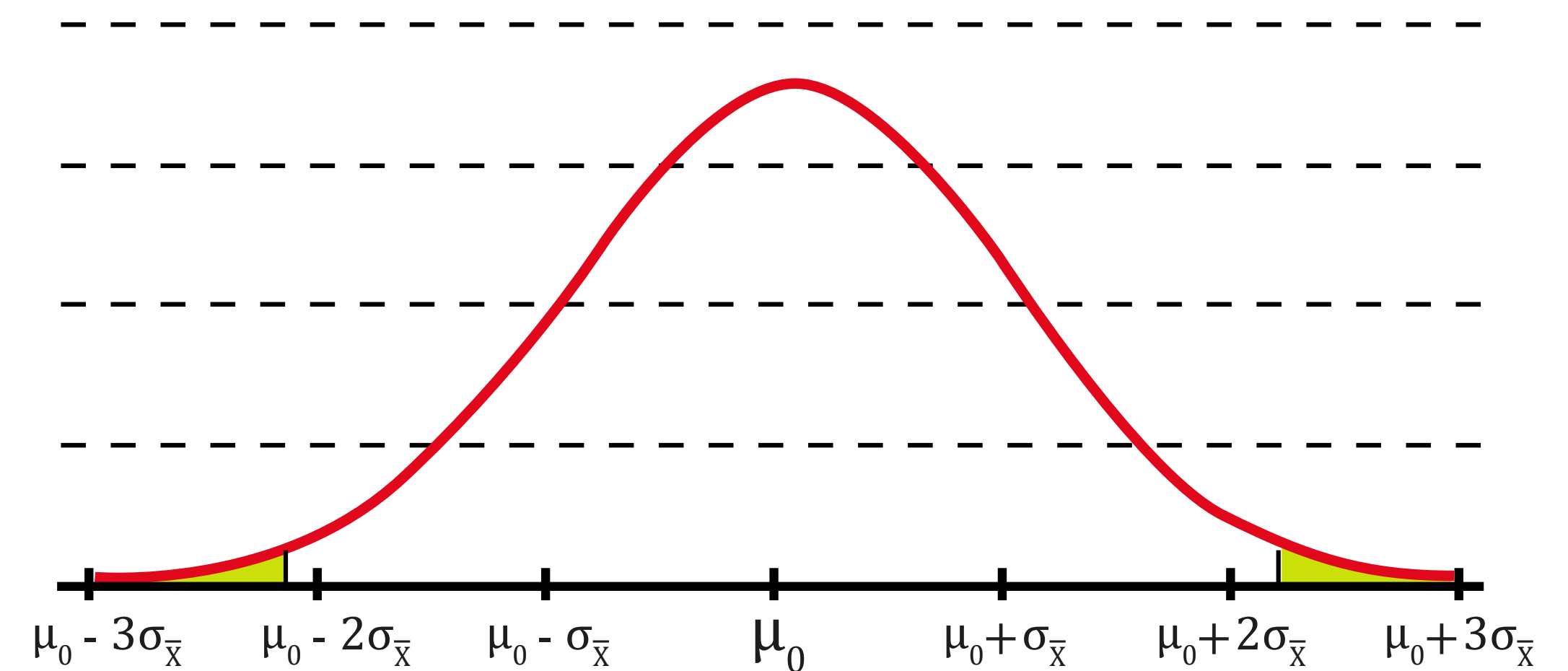
Blick hinter die Kulissen

Der Einstichproben t-Test berechnet zunächst einen t-Wert. Dieser drückt die Abweichung zum Zielwert in Standardfehlern aus:

$$t = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \quad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Um daraus den p-Wert zu berechnen, wird der t-Wert in die Verteilungsfunktion der Studentsche t-Verteilung mit n-1 Freiheitsgraden eingesetzt!

$$p = 2 \cdot (1 - F(t))$$



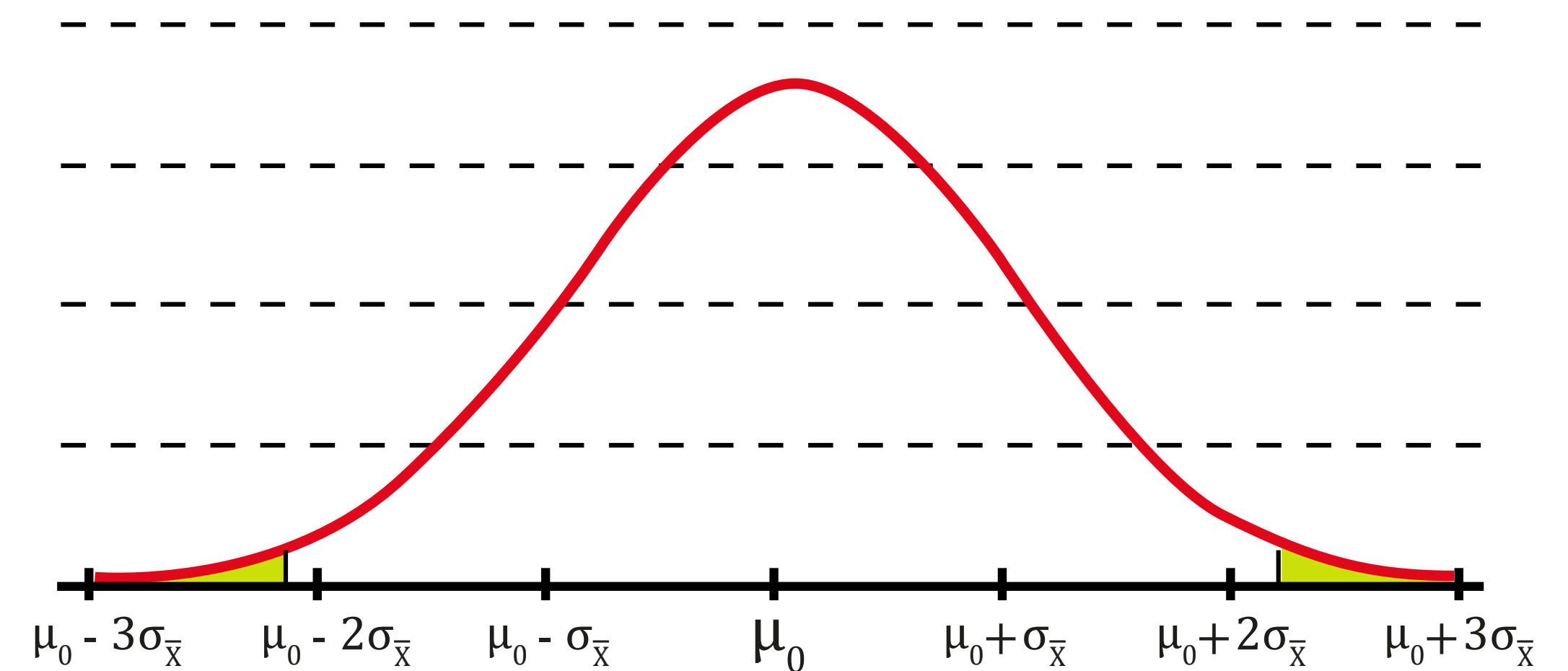
Blick hinter die Kulissen

Die Studentsche t-Verteilung sieht in Schaubildern täuschend ähnlich wie die Normalverteilung aus, ist aber nicht identisch zu ihr.

Der Unterschied zeigt sich auch in der Parametrierung. Statt wie bei der Normalverteilung μ und σ vorzugeben, geben wir eine Anzahl von Freiheitsgraden an.

Einstichproben t-Test Stichprobengröße minus 1

Zweistichproben t-Test Welch Approximation

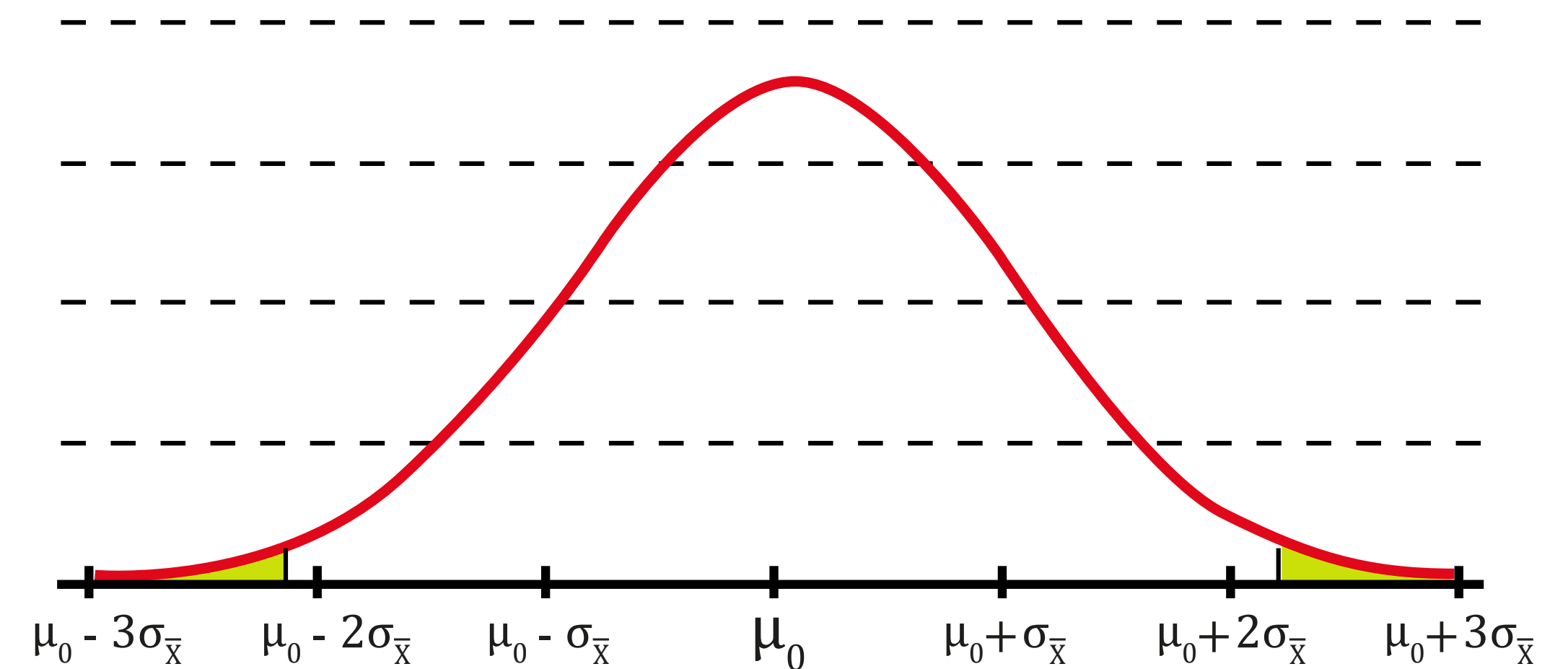


Blick hinter die Kulissen

Warum verwenden wir nicht die Normalverteilung?

Anders als beim z-Test kennen wir die Varianz der wahren Verteilung nicht. Als Schätzwert nehmen wir die Varianz der Stichprobe.

Dieses Vorgehen ist mit Unsicherheiten behaftet und die Studentische t-Verteilung berücksichtigt das.

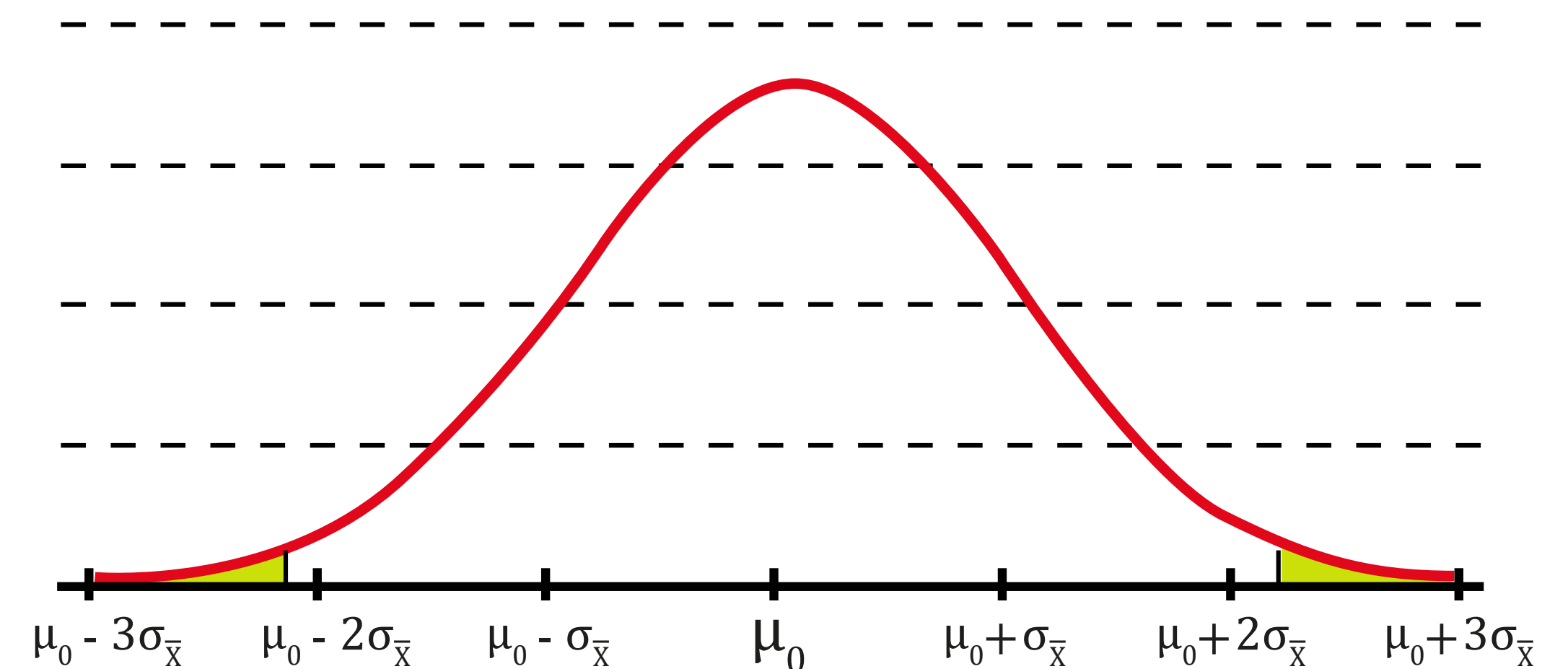


Blick hinter die Kulissen

Was hat es mit den Freiheitsgraden auf sich?

Je größer die Stichprobe, umso kleiner die Unsicherheit. Daher müssen wir die Studentsche t-Verteilung mit der Anzahl an Freiheitsgraden parametrieren.

Für eine hohe Zahl an Freiheitsgraden nähert sich die Studentsche t-Verteilung der Normalverteilung an!



t-Test

Verwende den Mensadatensatz, um folgende Fragen zu beantworten:

- 1. Gibt es Geschlechterunterschiede bei der Bewertung des Ambiente?
- 2. Bewerten Allergiker das Essen im Schnitt gleich gut wie Nicht-Allergiker?
- 3. Werden Ambiente und Essen im Schnitt unterschiedlich gut bewertet?

gender	status	preference	allergy	food_tas	food_hear	food_che	food_spi	food_sat	food_loo	food_var	food_sco
female	student	none	yes	3	4	6	2	5	4	2	3,71
female	student	vegetarian	yes	4	4	6	2	5	3	2	3,71
female	student	vegetarian	no	5	4	6	5	6	5	5	5,14
female	student	none	no	4	5	5	4	5	5	5	4,71
male	student	none	no	4	4	5	4	4	4	4	4,43
male	student	none	no	3	4	4	4	4	3	3	3,86
female	student	none	no	5	4	6	5	6	4	6	5,14
male	student	none	no	4	2	3	1	2	3	2	2,43
male	student	vegetarian	no	4	4	6	5	6	4	4	4,71
male	student	vegetarian	no	2	2	1	2	3	2	4	2,29
male	student	vegetarian	no	4	4	6	3	5	1	3	3,71
female	student	none	no	4	3	5	2	5	3	3	3,57
male	student	vegetarian	no	3	4	2	2	2	3	5	3,00
female	student	none	no	4	4	5	3	5	2	4	3,86
female	student	vegetarian	no	5	3	6	5	6	4	6	5,00
female	student	vegetarian	no	5	4	6	5	6	5	6	5,29
female	student	vegetarian	no	4	4	6	2	5	4	6	4,43
female	student	vegetarian	yes	4	3	6	3	5	5	2	4,00
female	student	none	no	5	4	6	5	5	4	4	4,71
male	student	none	no	4	3	5	4	5	2	3	3,71
female	student	vegetarian	no	3	3	5	4	4	3	4	3,71
female	student	none	no	5	4	6	6	6	5	6	5,43
female	student	vegetarian	yes	2	3	5	3	6	2	3	3,43

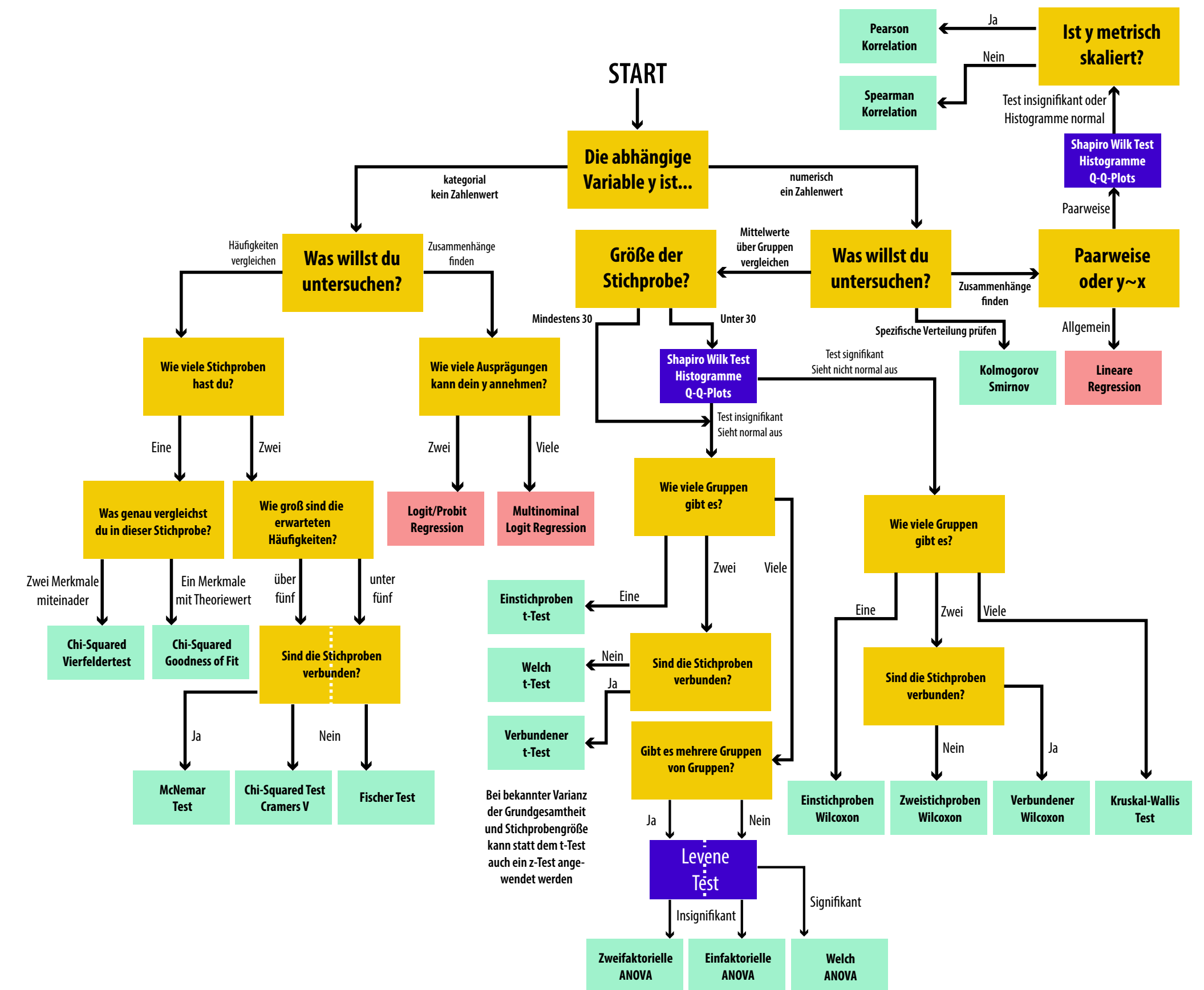
ANOVA

Eine Sache, die der t-Test nicht kann, ist mehr als zwei Stichproben miteinander zu vergleichen!

Mit einer ANOVA ("Analysis of Variance") können wir genau das machen. Das generische Hypothesenpaar lautet:

H0 - Es gibt keinen Unterschied im Erwartungswert der Stichproben.

H1 - Es existiert mindestens eine Paarung von Stichproben mit unterschiedlichen Erwartungswerten.



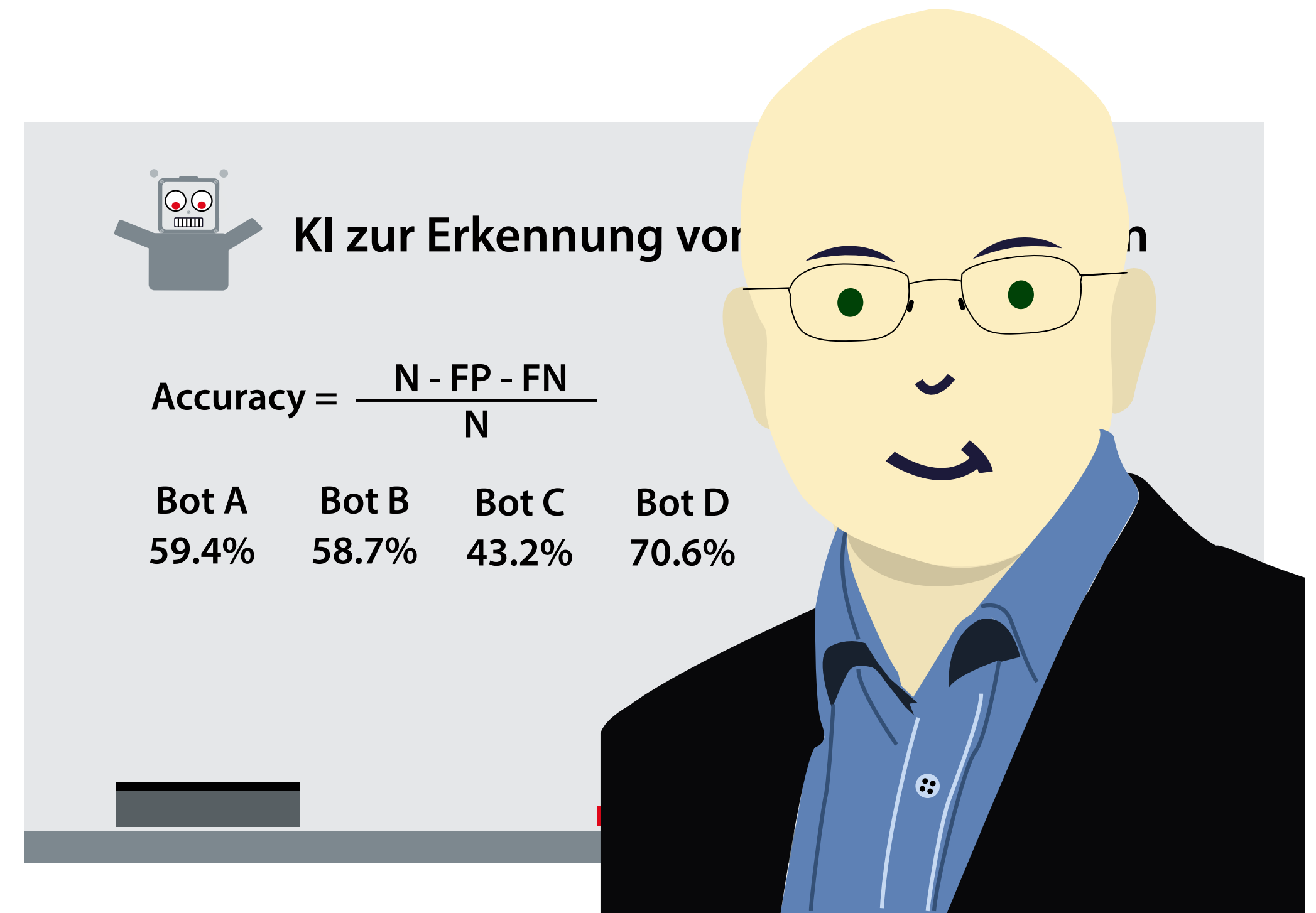
ANOVA

Ein Informatiker testet vier KIs, welche die Literaturangaben von Abschlussarbeiten auf Korrektheit untersuchen sollen.

Jede der KIs wird dazu mit 100 Bachelorarbeiten gefüttert. Gemessen wird die Accuracy, d. h. der Anteil der richtig eingestuften Angaben.

H0 - Es gibt keinen Unterschied in der Accuracy der Bots.

H1 - Es existiert mindestens eine Paarung von Bots unterschiedlicher Accuracy.



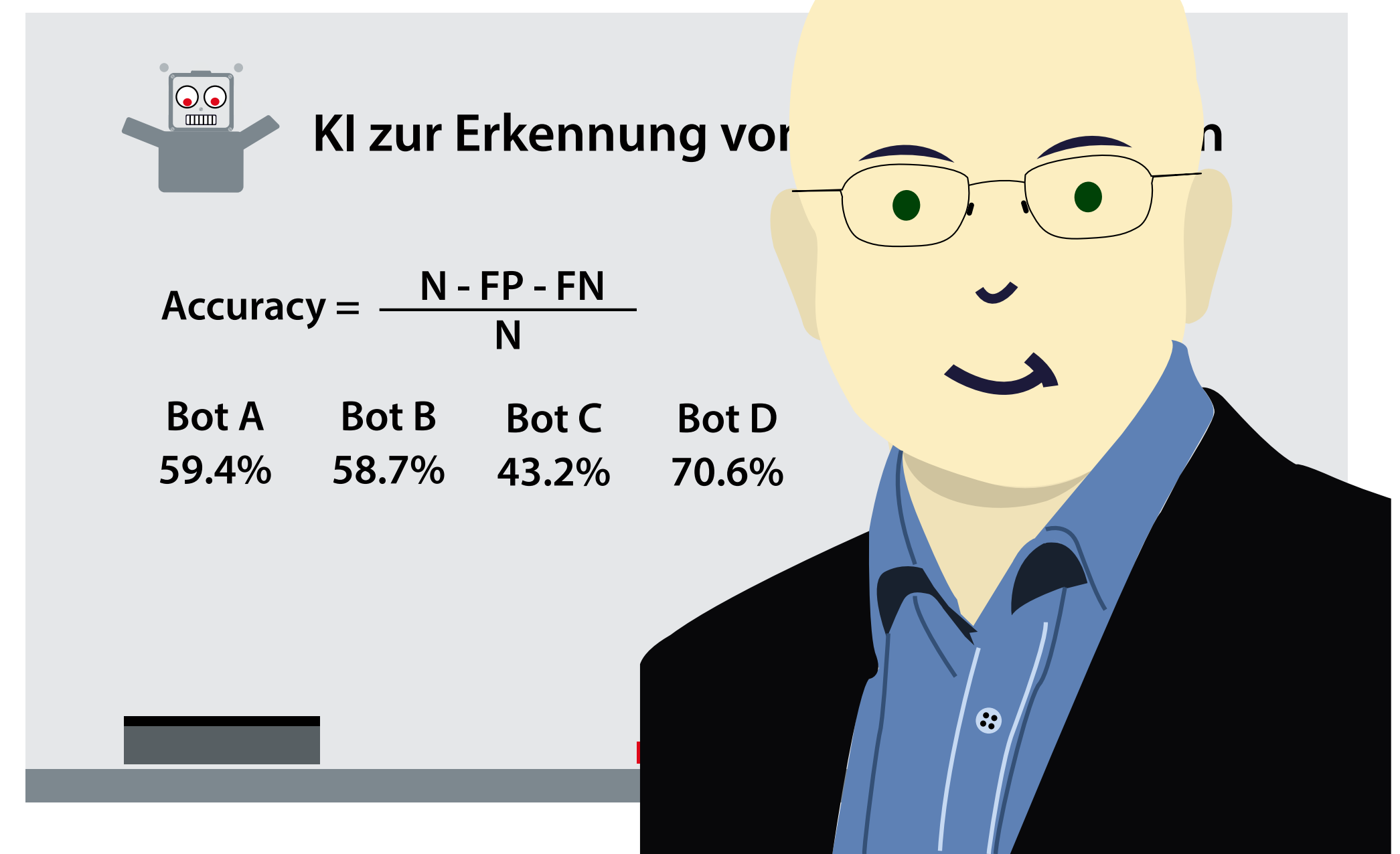
ANOVA

Mit einem p-Wert in der Größenordnung von 10^{-14} können wir die Nullhypothese verwerfen!

~~H0 - Es gibt keinen Unterschied in der Accuracy der Bots.~~

H1 - Es existiert mindestens eine Paarung von Bots unterschiedlicher Accuracy.

Die Alternativhypothese ist aber nicht gerade präzise. Um die Daten genauer zu untersuchen, ergänzen wir die ANOVA um sogenannte Post-Hoc-Tests.



ANOVA

Für den Post-Hoc-Test kommen mehrere Tests infrage:

Tukey HSD

Games-Howell

Paarweise t-Tests mit Bonferroni- / Holm-Korrektur

Am häufigsten wird der Tukey HSD eingesetzt, allerdings setzt dieser homoskedastische Stichproben voraus.

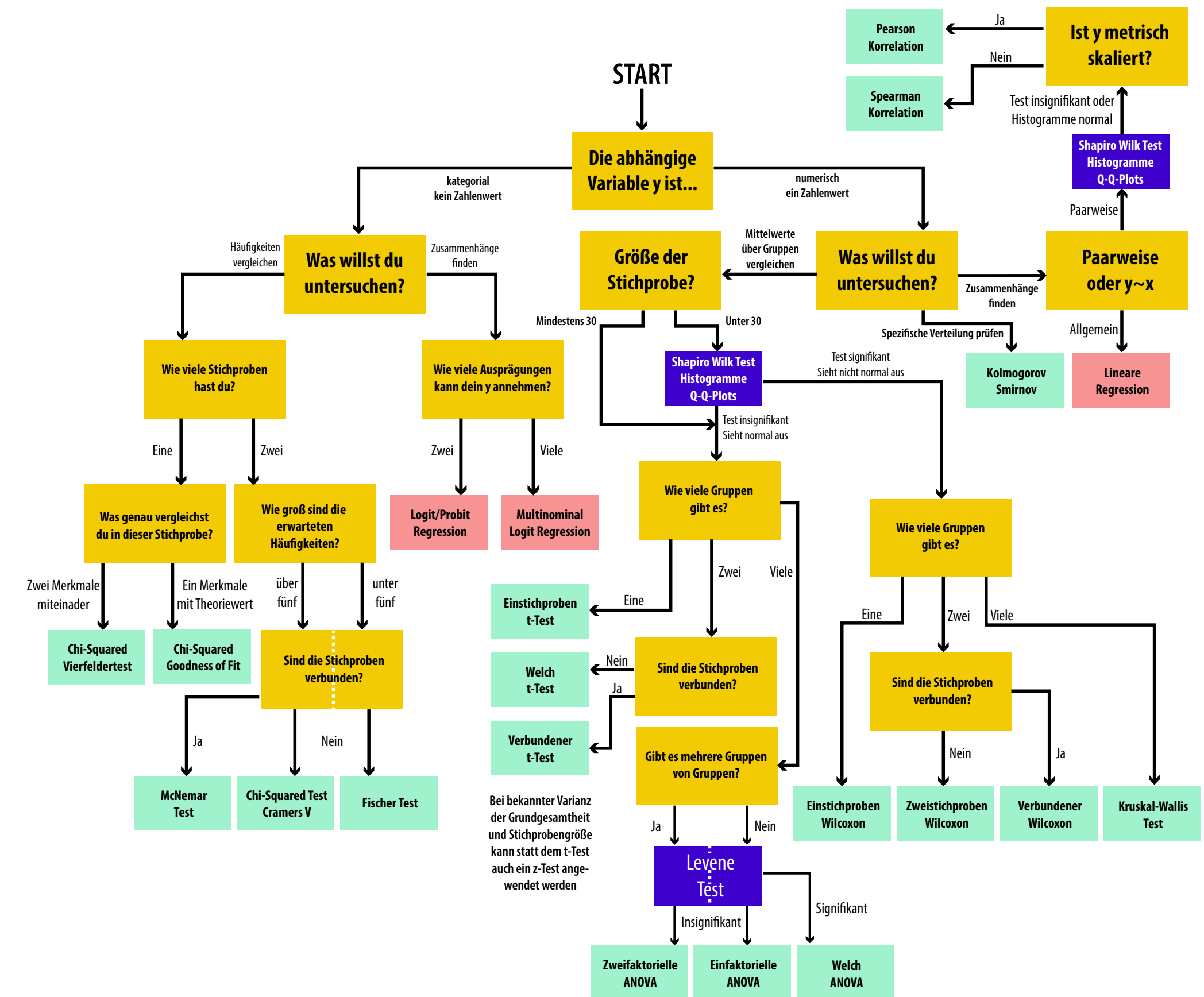
PWT(HOLM)	A	B	C	D
A		0.828	< 0.001	0.002
B	0.828		< 0.001	0.001
C	< 0.001	< 0.001		< 0.001
D	0.002	0.001	< 0.001	

Parametrische Tests

Die t-Tests und die ANOVA sind parametrische Tests.

Parametrische Tests nehmen an, dass ein Merkmal in der Grundgesamtheit einer bestimmten Verteilung folgt.

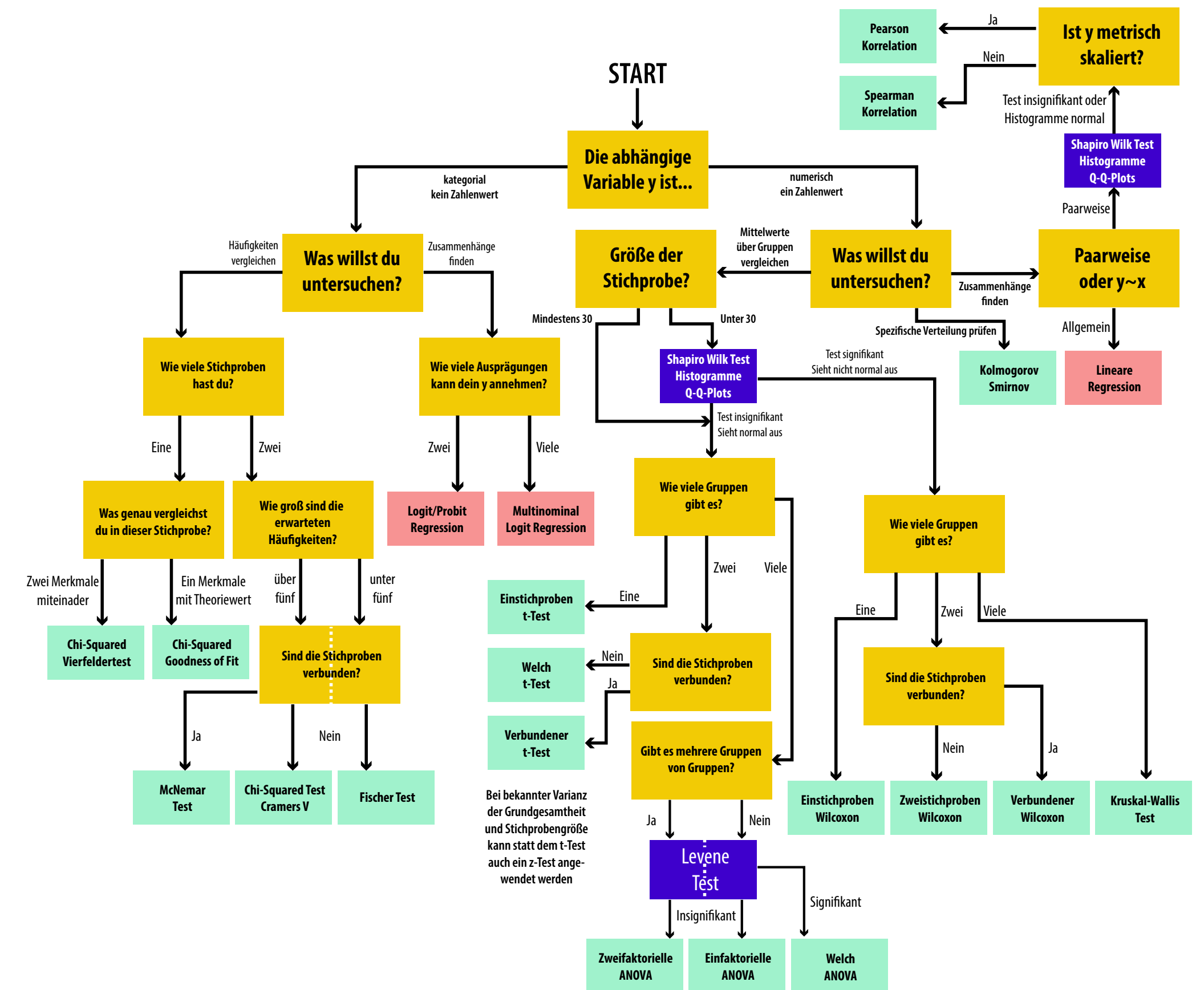
Der parametrische Test sucht dann die zu den Daten am besten passenden Parameter.



Parametrische Tests

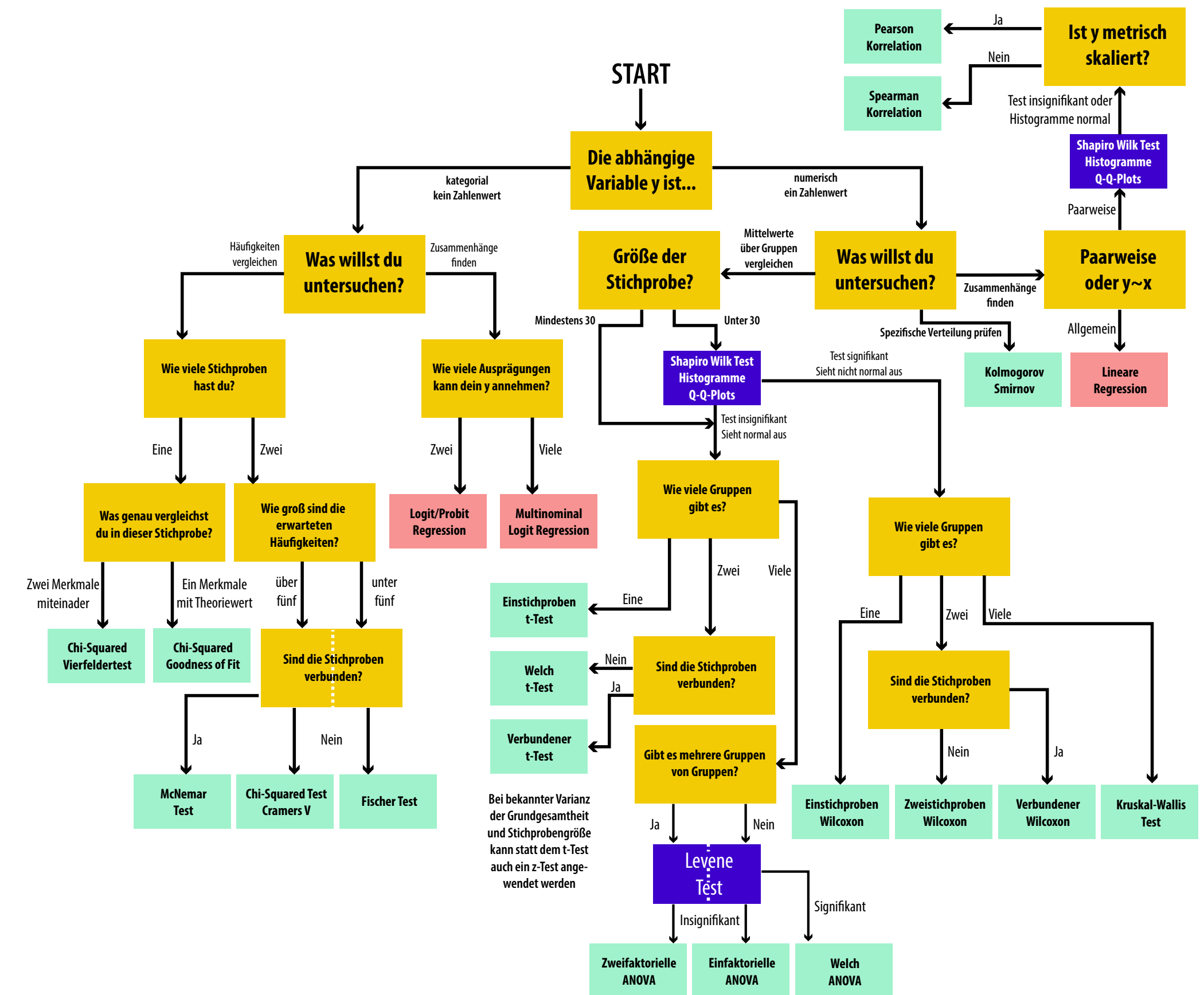
Damit wir den t-Test anwenden dürfen, müssen folgende Voraussetzungen erfüllt sein:

Die Daten der Stichprobe sind metrisch skaliert.
Der Mittelwert der Stichprobe ist normalverteilt.



Sind die Voraussetzungen nicht erfüllt, gibt es die folgenden nicht parametrischen Alternativen zum t-Test:

- Wilcoxon-Rangsummen-Test für 2 ungepaarte Stichproben
- Wilcoxon-Vorzeichen-Rang-Test für 2 gepaarte Stichproben
- Mann-Whitney-U-Test
- Sign-Test für Vergleich Median gegen Zielwert

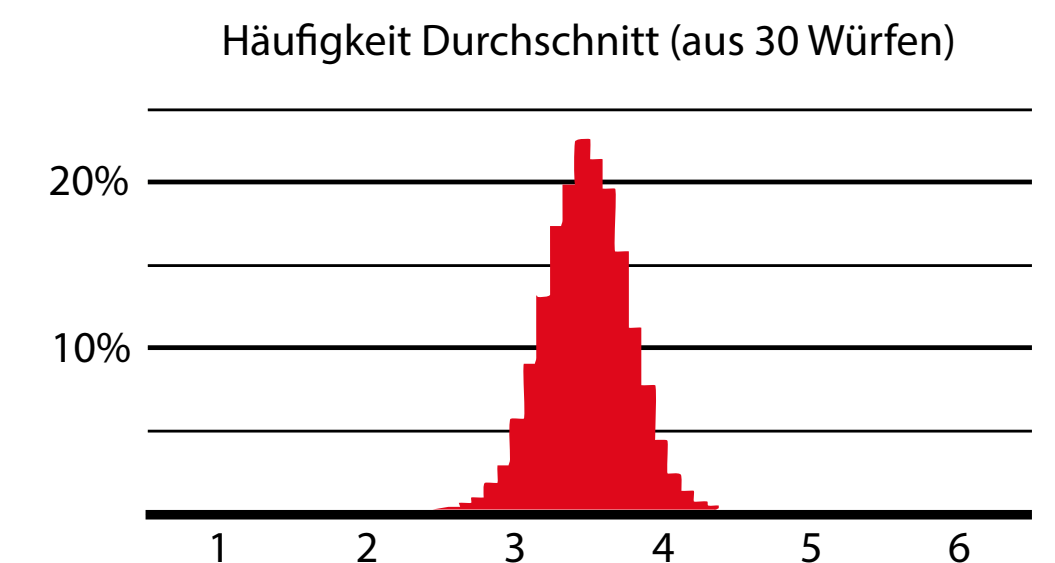
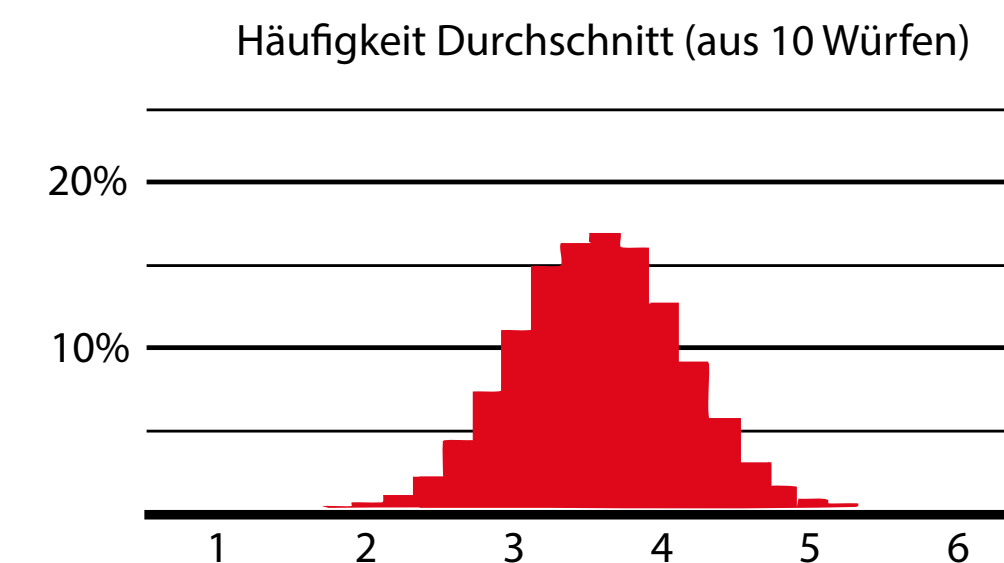
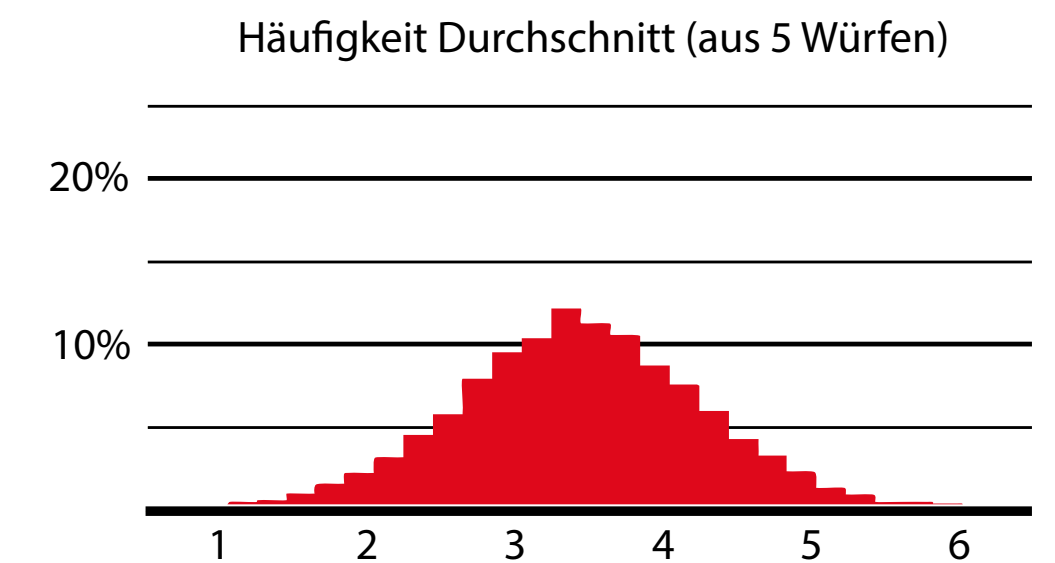
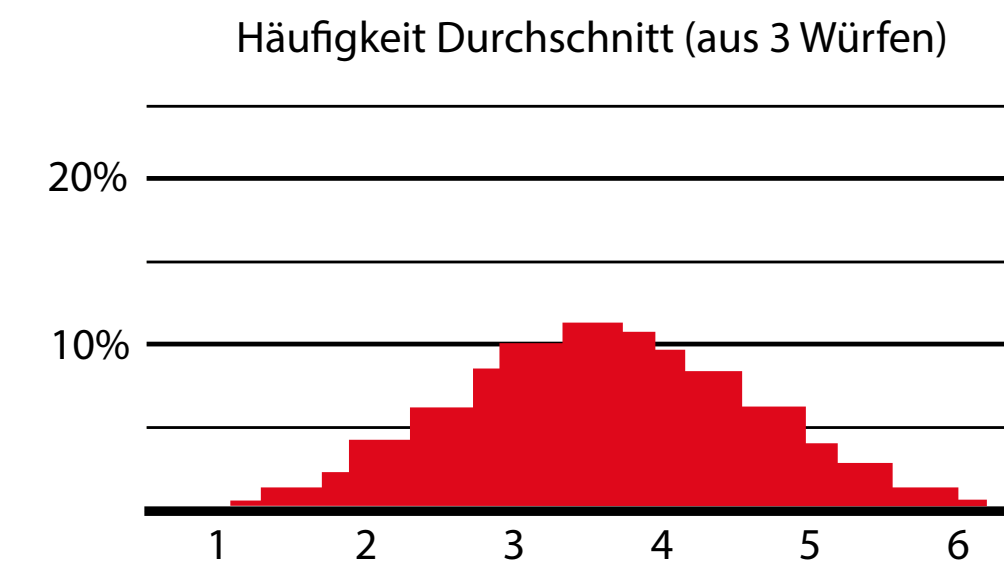


Test auf Normalität

Bei größeren Stichproben (ab $n=30$) können wir die Voraussetzung des normalverteilten Mittelwertes dank des ZGS als automatisch erfüllt annehmen.

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = \Phi(z)$$

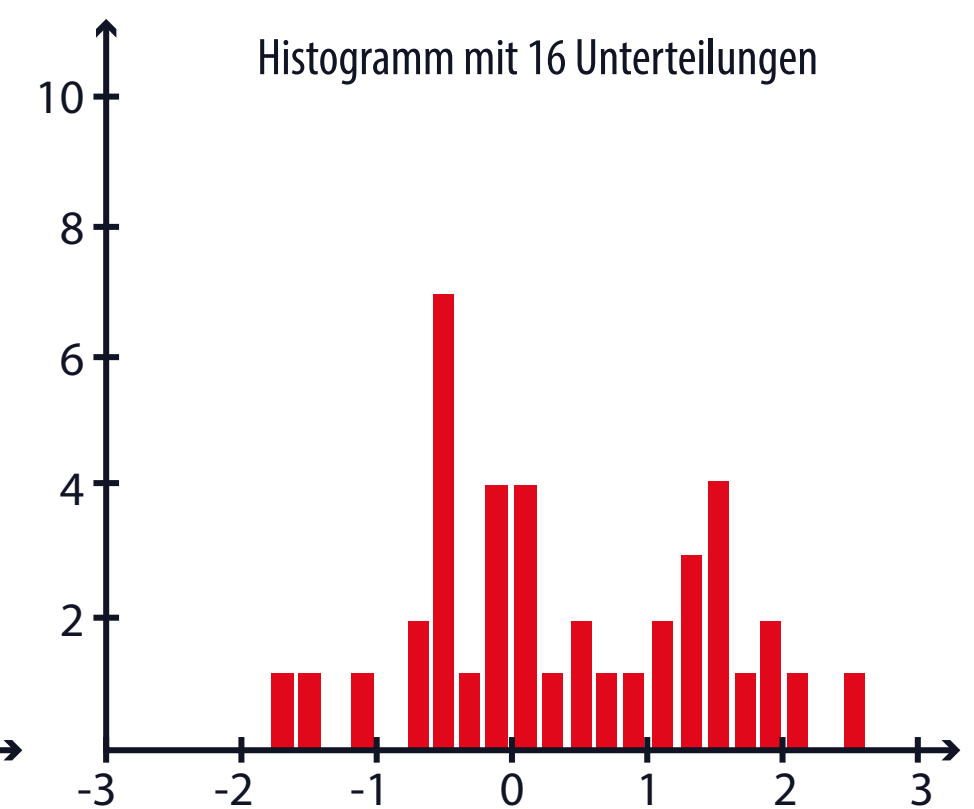
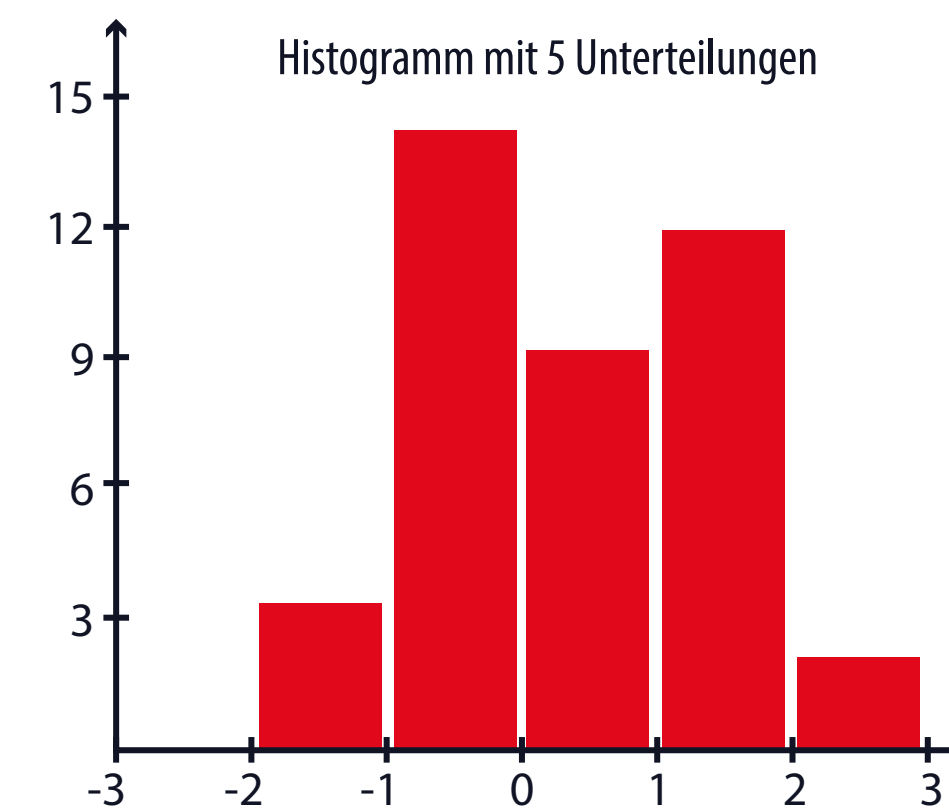
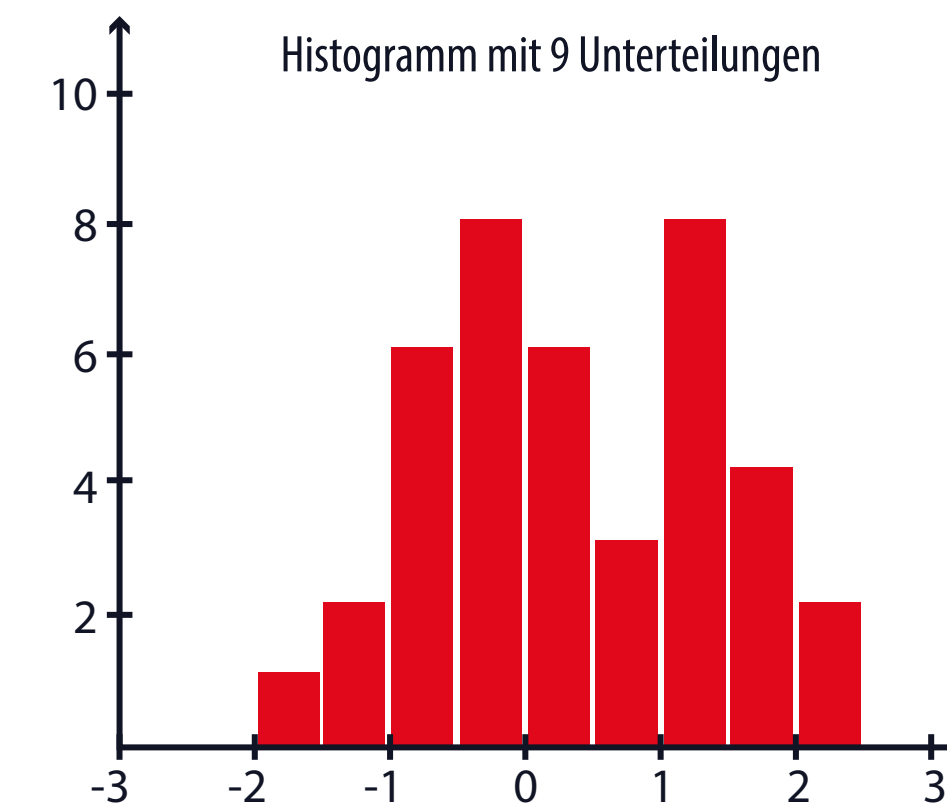
Bei kleineren Stichproben müssen wir genauer hinschauen. Wie erkennen wir, ob eine kleine Stichprobe näherungsweise normalverteilt ist?



Test auf Normalität

Naheliegende Idee: Wir betrachten ein Histogramm der Stichprobe und prüfen, ob es normalverteilt aussieht.

Tatsächlich gibt es mit Q-Q-Plots eine deutlich bessere grafische Methode, um Normalverteilung zu prüfen!



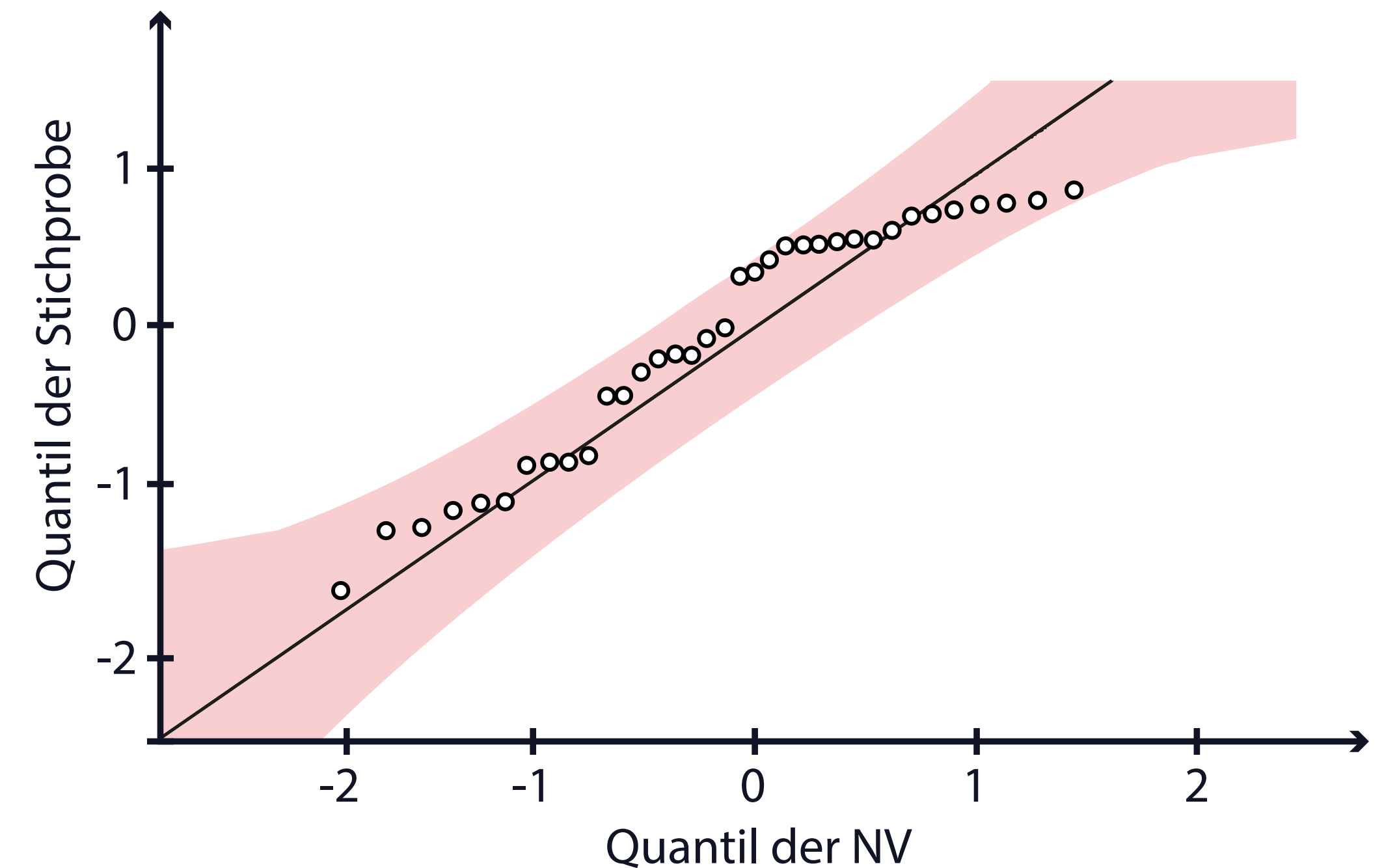
Test auf Normalität

Beim Q-Q-Plot tragen wir die Quantile der Stichprobe gegen die einer Standardnormalverteilung auf.

Die Quantile der Stichprobe entsprechen den aufsteigend sortierten Werten.

Die Quantile der Normalverteilung werden mit folgender Formel berechnet.

$$Q_i = \phi^{-1}(p_i) \quad \text{mit} \quad p_i = \frac{i - 0.5}{n}$$



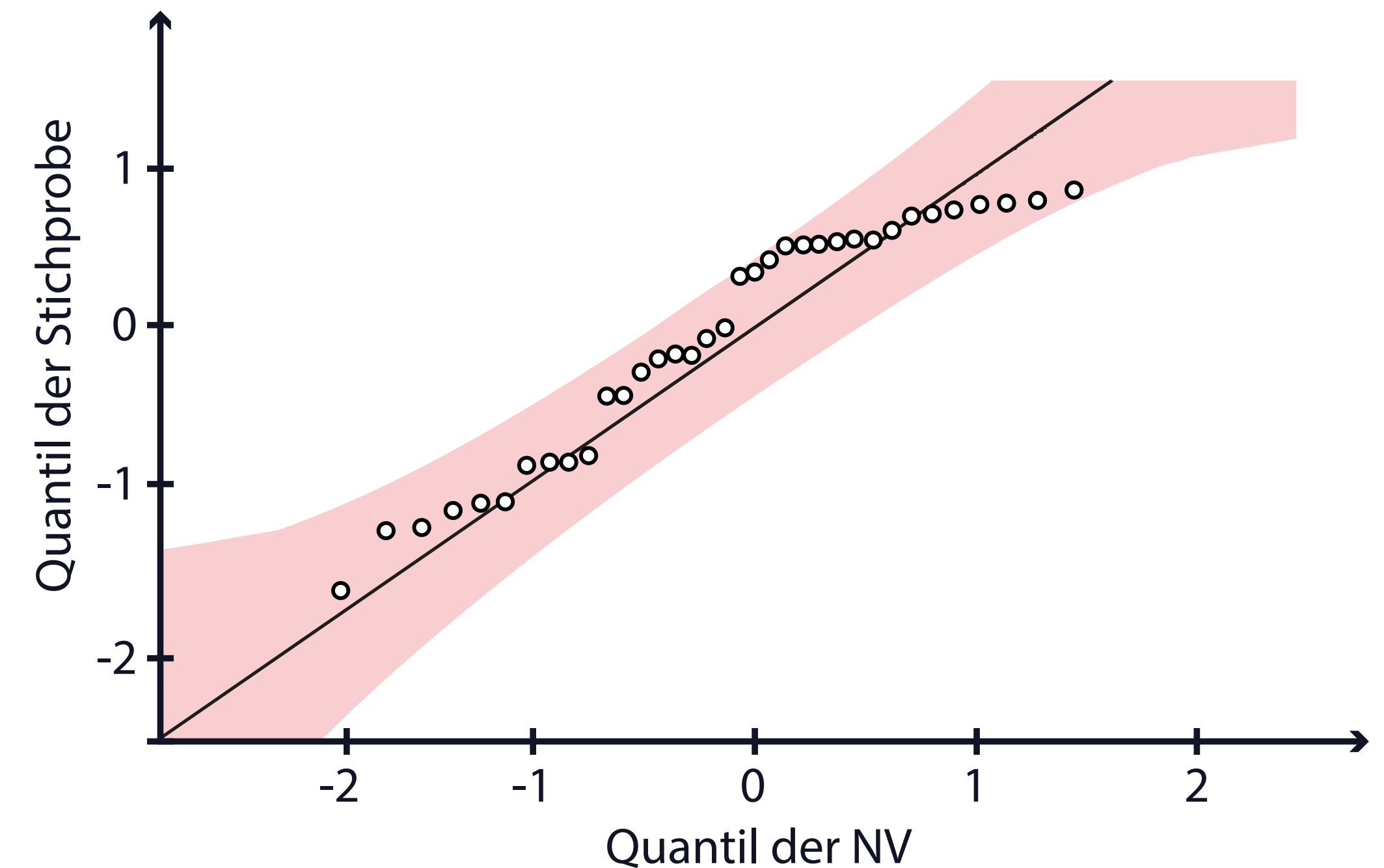
Test auf Normalität

Es gibt natürlich auch statistische Tests die eine Stichprobe, auf eine bestimmte Verteilung überprüfen!

Shapiro-Wilk-Test (Nur Normalverteilung)

Kolmogorov-Smirnov-Test (Mehrere Verteilungen)

Anderson-Darling-Test (Mehrere Verteilungen)



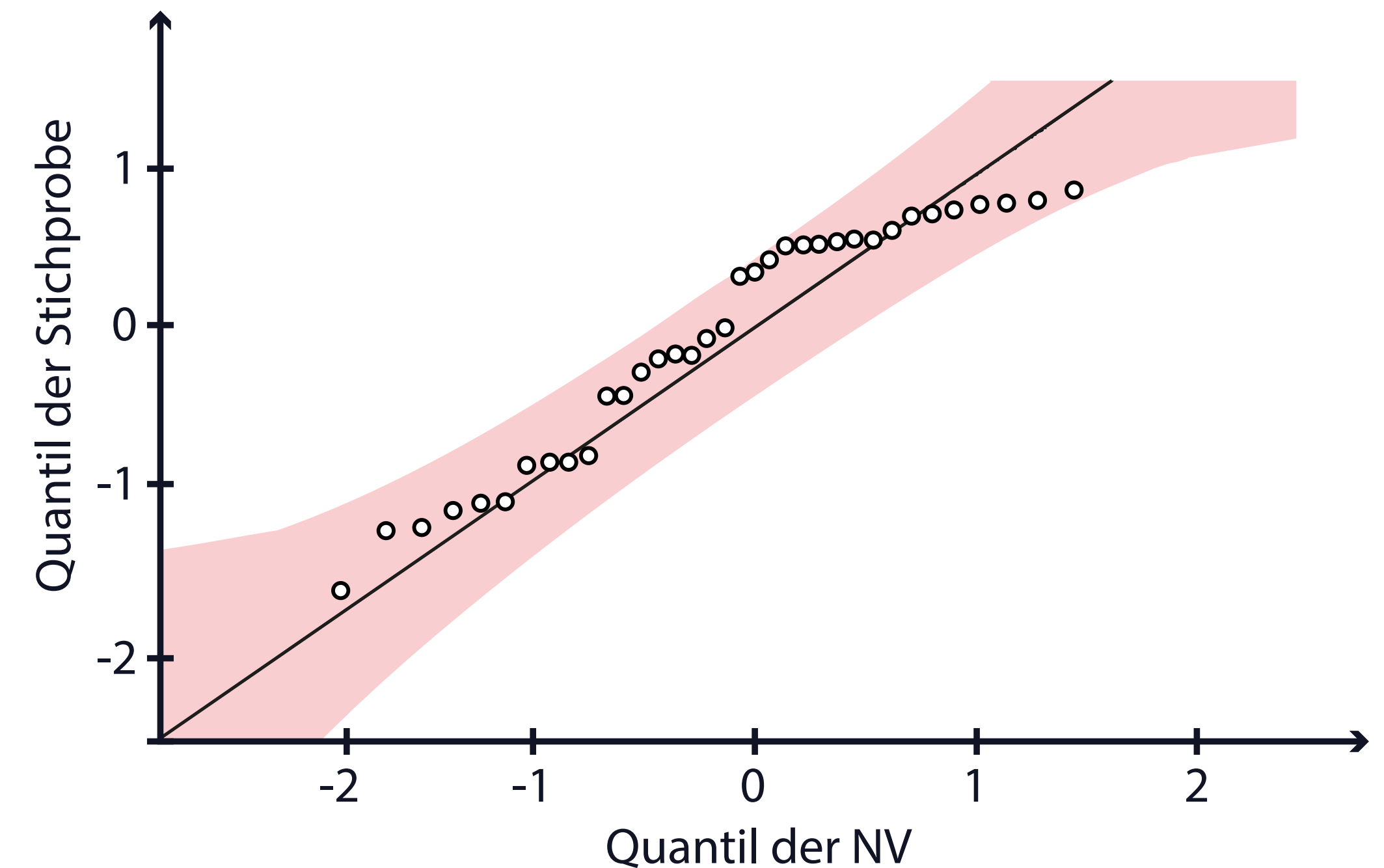
Test auf Normalität

Die Nullhypothese ist, dass das Merkmal in der Grundgesamtheit einer bestimmten Verteilung folgt.

Die Alternativhypothese ist eine signifikante Abweichung von der Verteilung.

Die Tests versuchen, eine Abweichung zu erkennen!

Folglich haben wir unsere Normalverteilung, wenn der p-Wert über 5% liegt und wir die Nullhypothese halten können!



Test auf Normalität

Wir testen die Tests mit **simulierten Daten**. Ihre Aufgabe ist es die Standardnormalverteilung von einer Gleichverteilung über $[-1, 1]$ zu unterscheiden.

Sensitivität Die Wahrscheinlichkeit, dass der Test eine abweichend verteilte Stichprobe auch als solche erkennt und keine false negatives hervorbringt.

Spezifität Die Wahrscheinlichkeit, dass der Test eine nicht abweichend verteilte Stichprobe auch als solche erkennt und keine false positives hervorbringt.

r_{norm} vs r_{unif}	SW-Test	KS-Test	AD-Test
n=15	Sensitivität 12.7% Spezifität 95.3%	Sensitivität 4.3% Spezifität 94.9%	Sensitivität 2.8% Spezifität 95.0%
n=30	Sensitivität 37.7% Spezifität 95.0%	Sensitivität 7.2% Spezifität 95.0%	Sensitivität 17.4% Spezifität 94.8%
n=50	Sensitivität 74.6% Spezifität 95.1%	Sensitivität 23.8% Spezifität 95.2%	Sensitivität 81.6% Spezifität 95.2%
n=100	Sensitivität 99.6% Spezifität 94.7%	Sensitivität 100.0% Spezifität 95.2%	Sensitivität 100.0% Spezifität 94.6%

Test auf Normalität

Gerade bei den kleinen Stichproben ist die Performance nicht gerade überragend. Sehr viele Abweichungen werden nicht erkannt.

Eines ist aber auffällig: Warum ist die Spezifität eigentlich immer bei 95%?

r_{norm} vs r_{unif}	SW-Test	KS-Test	AD-Test
n=15	Sensitivität 12.7% Spezifität 95.3%	Sensitivität 4.3% Spezifität 94.9%	Sensitivität 2.8% Spezifität 95.0%
n=30	Sensitivität 37.7% Spezifität 95.0%	Sensitivität 7.2% Spezifität 95.0%	Sensitivität 17.4% Spezifität 94.8%
n=50	Sensitivität 74.6% Spezifität 95.1%	Sensitivität 23.8% Spezifität 95.2%	Sensitivität 81.6% Spezifität 95.2%
n=100	Sensitivität 99.6% Spezifität 94.7%	Sensitivität 100.0% Spezifität 95.2%	Sensitivität 100.0% Spezifität 94.6%

Test auf Normalität

Die 95% entsprechen der kritischen Schwelle für den p-Wert und damit die Wahrscheinlichkeit eines Fehlers erster Art.

Erhöhen wir die kritische Schwelle, opfern wir Spezifität für Sensitivität.

r_{norm} vs r_{unif}	SW-Test	KS-Test	AD-Test
n=15	Sensitivität 12.7% Spezifität 95.3%	Sensitivität 4.3% Spezifität 94.9%	Sensitivität 2.8% Spezifität 95.0%
n=30	Sensitivität 37.7% Spezifität 95.0%	Sensitivität 7.2% Spezifität 95.0%	Sensitivität 17.4% Spezifität 94.8%
n=50	Sensitivität 74.6% Spezifität 95.1%	Sensitivität 23.8% Spezifität 95.2%	Sensitivität 81.6% Spezifität 95.2%
n=100	Sensitivität 99.6% Spezifität 94.7%	Sensitivität 100.0% Spezifität 95.2%	Sensitivität 100.0% Spezifität 94.6%

Test auf Normalität

Es gibt Varianten der rechts gezeigten Tests, die eine bessere Sensitivität auch bei kleinen Stichproben versprechen, wie z. B. der KS-Test mit Lilliefors Korrektur.

Es gibt auch den Filliben's Test, der das Auswerten der Q-Q-Plots automatisiert.

Am Ende stellt sich aber noch eine andere Frage ...

r_{norm} vs r_{unif}	SW-Test	KS-Test	AD-Test
n=15	Sensitivität 12.7% Spezifität 95.3%	Sensitivität 4.3% Spezifität 94.9%	Sensitivität 2.8% Spezifität 95.0%
n=30	Sensitivität 37.7% Spezifität 95.0%	Sensitivität 7.2% Spezifität 95.0%	Sensitivität 17.4% Spezifität 94.8%
n=50	Sensitivität 74.6% Spezifität 95.1%	Sensitivität 23.8% Spezifität 95.2%	Sensitivität 81.6% Spezifität 95.2%
n=100	Sensitivität 99.6% Spezifität 94.7%	Sensitivität 100.0% Spezifität 95.2%	Sensitivität 100.0% Spezifität 94.6%

Test auf Normalität

Wie schlimm ist es, wenn z. B. beim t-Test die Annahme des normalverteilten Mittelwertes der Stichprobe verletzt ist?

Tatsächlich bedeuten kleinere Abweichungen nicht, dass der Test sofort nutzlos wird. Problematisch sind vor allem folgende Abweichungen:

- Hohe Wahrscheinlichkeit für Ausreißer (Fat Tails)
- Starke Asymmetrie, d. h. eine ausgeprägte Schiefe
- Bimodale Verteilungen

r_{norm} vs r_{unif}	SW-Test	KS-Test	AD-Test
n=15	Sensitivität 12.7% Spezifität 95.3%	Sensitivität 4.3% Spezifität 94.9%	Sensitivität 2.8% Spezifität 95.0%
n=30	Sensitivität 37.7% Spezifität 95.0%	Sensitivität 7.2% Spezifität 95.0%	Sensitivität 17.4% Spezifität 94.8%
n=50	Sensitivität 74.6% Spezifität 95.1%	Sensitivität 23.8% Spezifität 95.2%	Sensitivität 81.6% Spezifität 95.2%
n=100	Sensitivität 99.6% Spezifität 94.7%	Sensitivität 100.0% Spezifität 95.2%	Sensitivität 100.0% Spezifität 94.6%

χ^2 -Unabhängigkeitstest

Mit der Chi-Quadrat-Testfamilie lernen wir Tests für kategoriale Daten bzw. Häufigkeitsdaten kennen.

Am häufigsten wird uns der Unabhängigkeitstests begegnen. Das generische Hypothesenpaar ist:

Nullhypothese H_0 Zwei kategoriale Merkmale weisen keinen Zusammenhang auf.

Alternativhypothese H_1 Zwei kategoriale Merkmale weisen einen Zusammenhang auf.

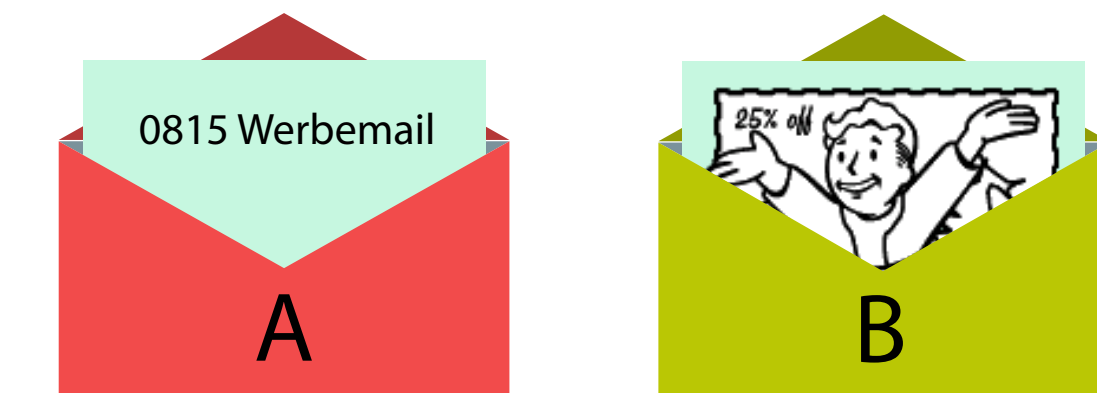
	Männer	Frauen	
Nichtraucher	106 (53%)	122 (61%)	228
Aufgehört	40 (20%)	36 (18%)	76
Raucher	54 (27%)	42 (21%)	96
	200	200	

χ^2 -Unabhängigkeitstest

Wir messen die Performance von zwei verschiedenen Varianten eines Werbemailings: die bisherige Variante A und unsere neu entwickelte Variante B.

Deskriptiv sieht es gut aus: Die neu entwickelte Variante hat eine höhere Öffnungsrate.

Der Chef und der PA-Betreuer sind aber skeptisch: Du hast doch nur einen Glückstreffer gelandet! Mit welchem Test können wir die Sache näher untersuchen?



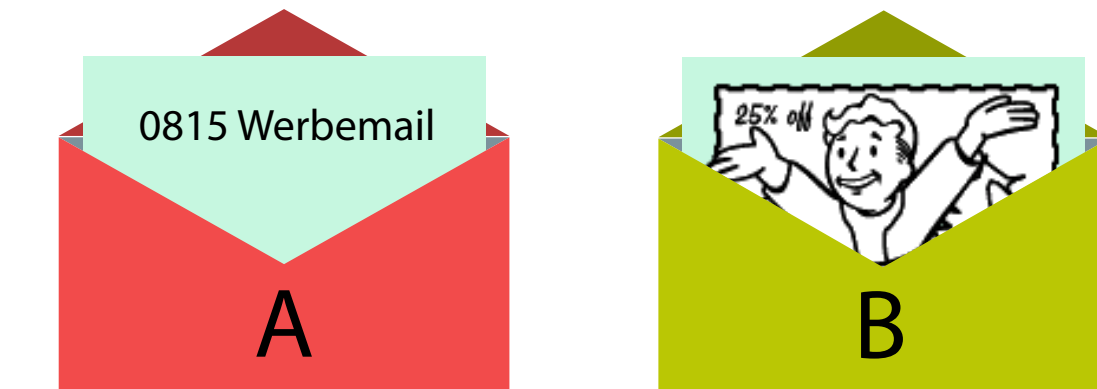
Stichprobe	615	613
Geöffnet	57	71
Ignoriert	558	542
Quote	9.27%	11.58%
		
	Zufall oder signifikant?	

χ^2 -Unabhängigkeitstest

Unser Chi-Quadrat-Unabhängigkeitstest überprüft im Beispiel folgendes Hypothesenpaar:

Nullhypothese H0 Die Häufigkeit geöffneter Werbemails ist bei beiden Varianten identisch.

Alternativhypothese H1 Die Häufigkeit geöffneter Werbemails unterscheidet sich zwischen den beiden Varianten.



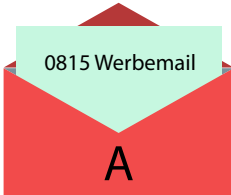
Stichprobe	615	613
Geöffnet	57	71
Ignoriert	558	542
Quote	9.27%	11.58%
		
	Zufall oder signifikant?	

χ^2 -Unabhängigkeitstest

Der Test vergleicht dabei die beobachtete Kontingenztabelle mit einer Kontingenztabelle, die wir erwarten würden, wenn die Nullhypothese ...

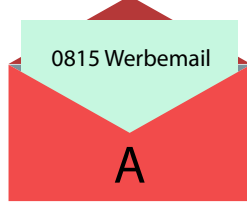
Nullhypothese H0 Die Häufigkeit geöffneter und ignoriert-ter Werbemailings ist bei beiden Varianten identisch.

...zutreffen würde. In unserem Beispiel würden wir dann bei beiden Werbemailings eine Öffnungsrate von 10.4% erwarten!

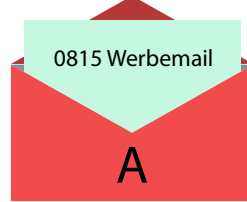



	Variante A	Variante B	
Geöffnet	57 (9.3%)	71 (11.6%)	128 (10.4%)
Ignoriert	558	542	1100
	615	613	1228

Beobachtung

	 A	 B	
	Variante A	Variante B	
Geöffnet	57 (9.3%)	71 (11.6%)	128 (10.4%)
Ignoriert	558	542	1100
	615	613	1228

Erwartung unter H0

	 A	 B	
	Variante A	Variante B	
Geöffnet	$615 \cdot 0.104 = 64$	$613 \cdot 0.104 = 64$	128 (10.4%)
Ignoriert	$615 - 64 = 551$	$613 - 64 = 549$	1100
	615	613	1228

χ^2 -Unabhängigkeitstest

Wir verwenden die Variante für nicht gepaarte Stichproben und erhalten einen p-Wert von 0.2174.

Die Wahrscheinlichkeit eines Häufigkeitsunterschiedes von mindestens 2.3 Prozentpunkten ist selbst unter Gültigkeit der Nullhypothese ...

Nullhypothese H0 Die Häufigkeit geöffneter und ignorerter Werbemailings ist bei beiden Varianten identisch.

...mit 21.74% plausibel. Wir behalten H0 also bei.

```
Console Terminal Background Jobs
R 4.2.2
> m <- matrix(c(57,558,71,542),nrow=2,ncol=2,
+             dimnames=list(c("opened","ignored"),c("A","B")))
>
> rateA <- m[1,1]/(m[1,1]+m[2,1])
> rateB <- m[1,2]/(m[1,2]+m[2,2])
>
> result <- chisq.test(m)

Pearson's Chi-squared test with Yates' continuity
correction

data:  m
X-squared = 1.5216, df = 1, p-value = 0.2174
```

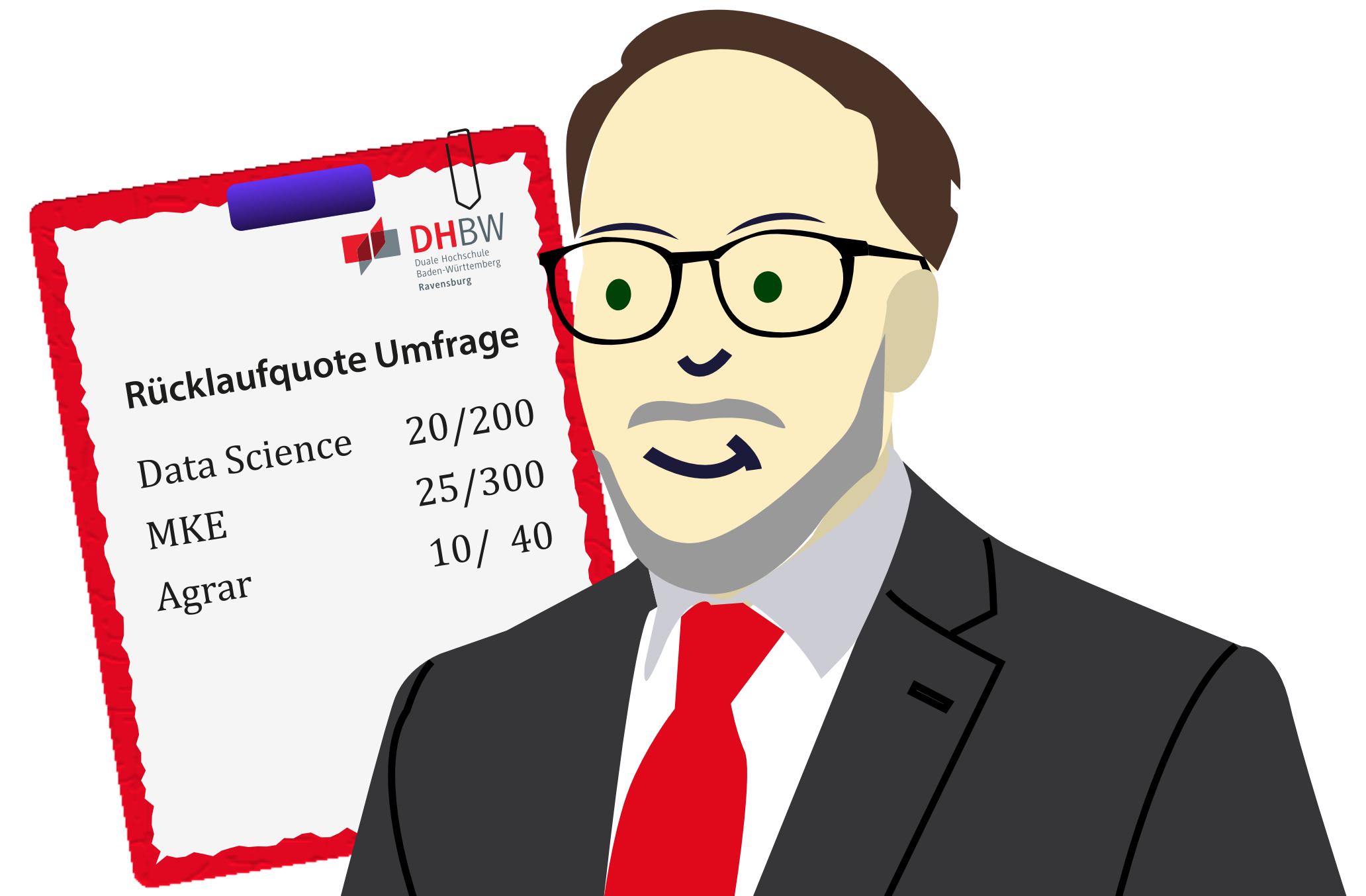
χ^2 -Anpassungstest

Der Chi-Quadrat-Anpassungstest hat eine andere Fragestellung als der Unabhängigkeitstest.

Wir wollen wissen, ob die Verteilung eines kategorialen Merkmals einer erwarteten Verteilung entspricht.

Nullhypothese H_0 Die beobachtete Verteilung passt zu der erwarteten Verteilung.

Alternativhypothese H_1 Die beobachtete Verteilung weicht von der erwarteten Verteilung ab.



χ^2 -Anpassungstest

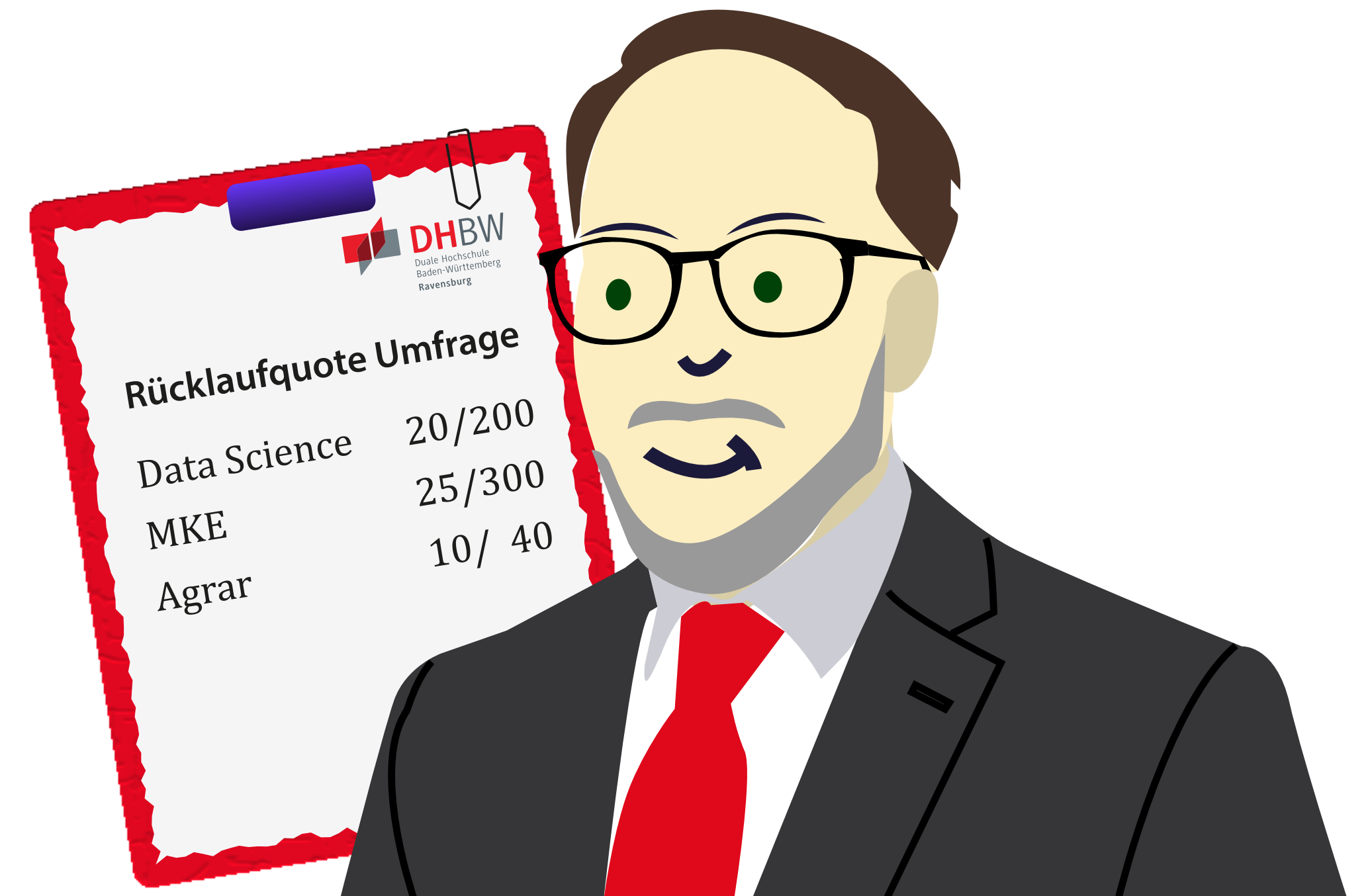
Wir haben eine Umfrage unter Studierenden durchgeführt, bei der auch der Studiengang abgefragt wurde. Wir haben Antworten von:

20 Studierenden aus Data Science

25 Studierenden aus MKE

10 Studierenden aus Agrarwirtschaft

Ist dieses Sample repräsentativ oder sind Studiengänge über- bzw. unterrepräsentiert?



χ^2 -Anpassungstest

Wir testen die beobachtete Verteilung gegen die erwartete Verteilung und erhalten einen p-Wert von 0.00808.

Wäre die Nullhypothese gültig, würden wir mit Wahrscheinlichkeit 0.808% eine mindestens so große Abweichung feststellen.

Das ist unwahrscheinlich und deutlich unter der 5% Hürde für die Signifikanz. Wir verwerfen die Nullhypothese.

```
Console Terminal Background Jobs
R 4.2.2
> sample <- c(20,25,10)
> base <- c(200,300,40)

> chisq.test(sample,p=base,rescale.p = TRUE)

      Chi-squared test for given probabilities

data:  sample
X-squared = 9.6364, df = 2, p-value = 0.008081
```

χ^2 -Testfamilie

Verwende den Mensadatensatz, um Unterschiede in den Essenspräferenzen nach Geschlecht zu untersuchen. Erstelle eine Kontingenztafel, führe einen Chi-Quadrat-Unabhängigkeitstest durch und interpretiere das Resultat.

Prüfe außerdem mit einem Chi-Quadrat-Anpassungstest, ob die Verteilung der Essenspräferenzen dem Bundesdurchschnitt von 10% vegetarisch und 2.5% vegan entspricht. Interpretiere auch hier das Resultat.

preference	allergy	food_tas	food_hear	food_che	food_spi	food_sat	food_loo	food_var	food_sco
none	yes	3	4	6	2	5	4	2	3,71
vegetarian	yes	4	4	6	2	5	3	2	3,71
vegetarian	no	5	4	6	5	6	5	5	5,14
none	no	4	5	5	4	5	5	5	4,11
none	no	4	4	5	4	4	3	3	4,44
none	no	3	4	4	4	6	3	3	3,86
none	no	5	4	6	5	6	4	6	5,14
none	no	4	2	3	1	2	3	2	2,43
vegetarian	no	4	4	6	5	6	4	4	4,71
vegetarian	no	2	2	1	2	3	2	4	2,29
vegetarian	no	4	4	6	3	5	1	3	3,71
none	no	4	3	5	2	5	3	3	3,57
vegetarian	no	3	4	2	2	2	3	5	3,00
none	no	4	4	5	3	5	2	4	3,86
vegetarian	no	5	3	6	5	6	4	6	5,00
vegetarian	no	5	4	6	5	6	5	6	5,29
vegetarian	no	4	4	6	2	5	4	6	4,43
vegetarian	yes	4	3	6	3	5	5	2	4,00
none	no	5	4	6	5	5	4	4	4,71
none	no	4	3	5	4	5	2	3	3,71
vegetarian	no	3	3	5	4	4	3	4	3,71
none	no	5	4	6	6	6	5	6	5,43
vegetarian	yes	2	3	5	3	6	2	3	3,43

Fehler von statistischen Tests

Fehler erster Art Wir lehnen die Nullhypothese ab, obwohl sie eigentlich richtig wäre. Die Wahrscheinlichkeit für diesen Fehler können wir an dem Signifikanzniveau ablesen!

Auch bekannt als: **Falsch Positiv**

Fehler zweiter Art Wir lehnen die Nullhypothese nicht ab, obwohl sie eigentlich falsch wäre.

Auch bekannt als: **Falsch Negativ**



Fehler von statistischen Tests

Datenfehler Wir verwenden Messdaten, die selbst fehlerbehaftet sind.

Modellfehler Wir verwenden einen ungeeigneten Test bzw. treffen Annahmen über Verteilungen und Merkmale, die nicht zutreffen.

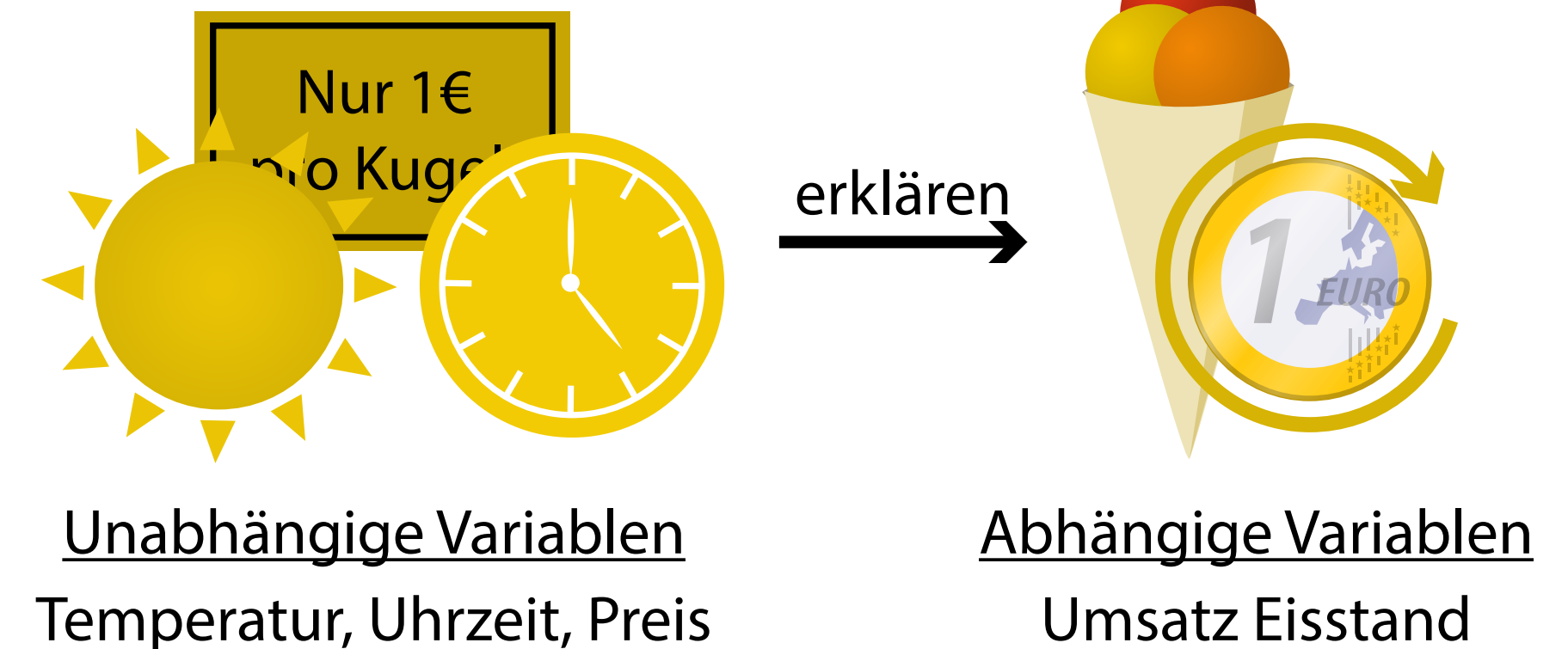


Regressionsanalyse

Bei der Regressionsanalyse suchen wir nach Zusammenhängen zwischen abhängigen und unabhängigen Variablen.

Je nachdem ob wir eine oder mehrere (un-)abhängige Variablen haben, unterscheiden wir zwischen:

Einfacher und multipler Regression
Univariater und multivariater Regression



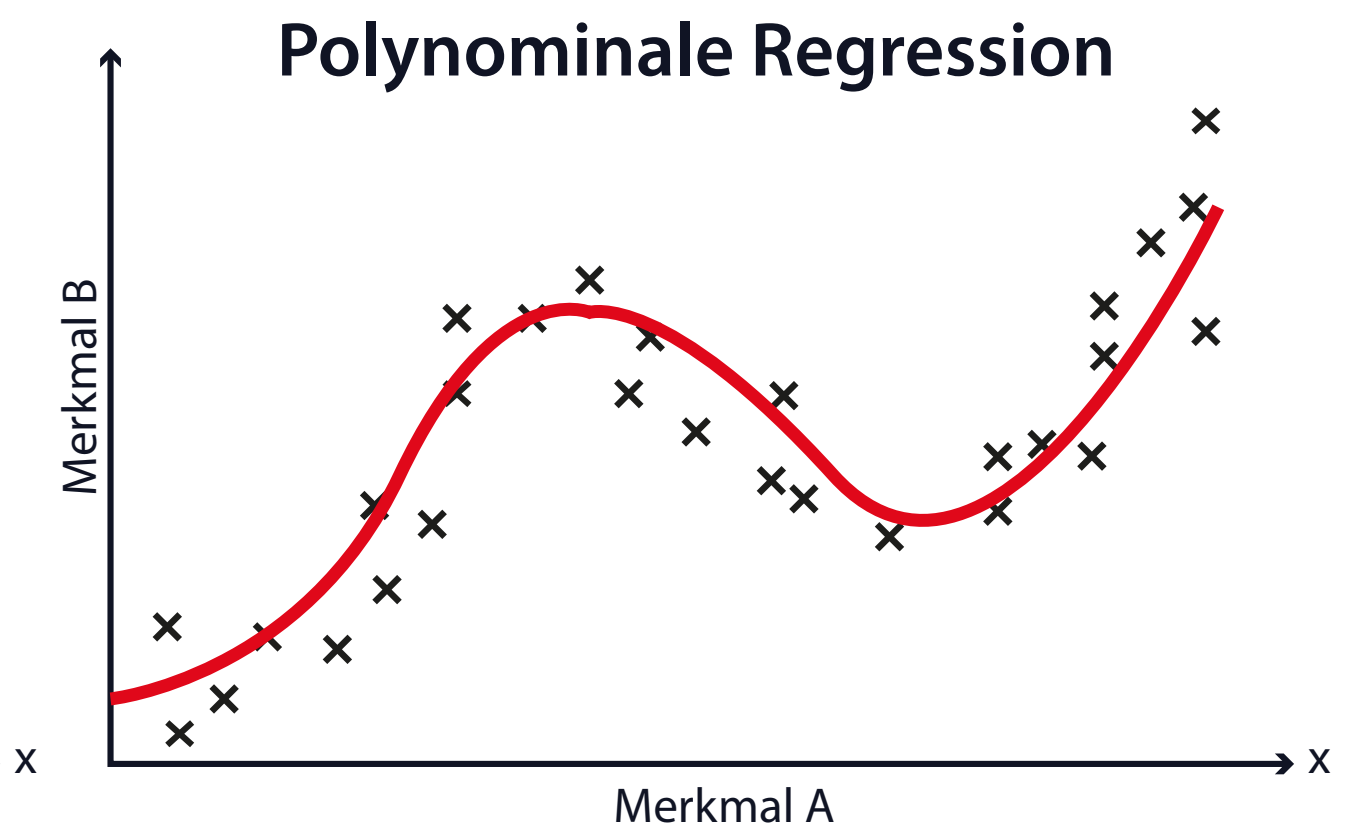
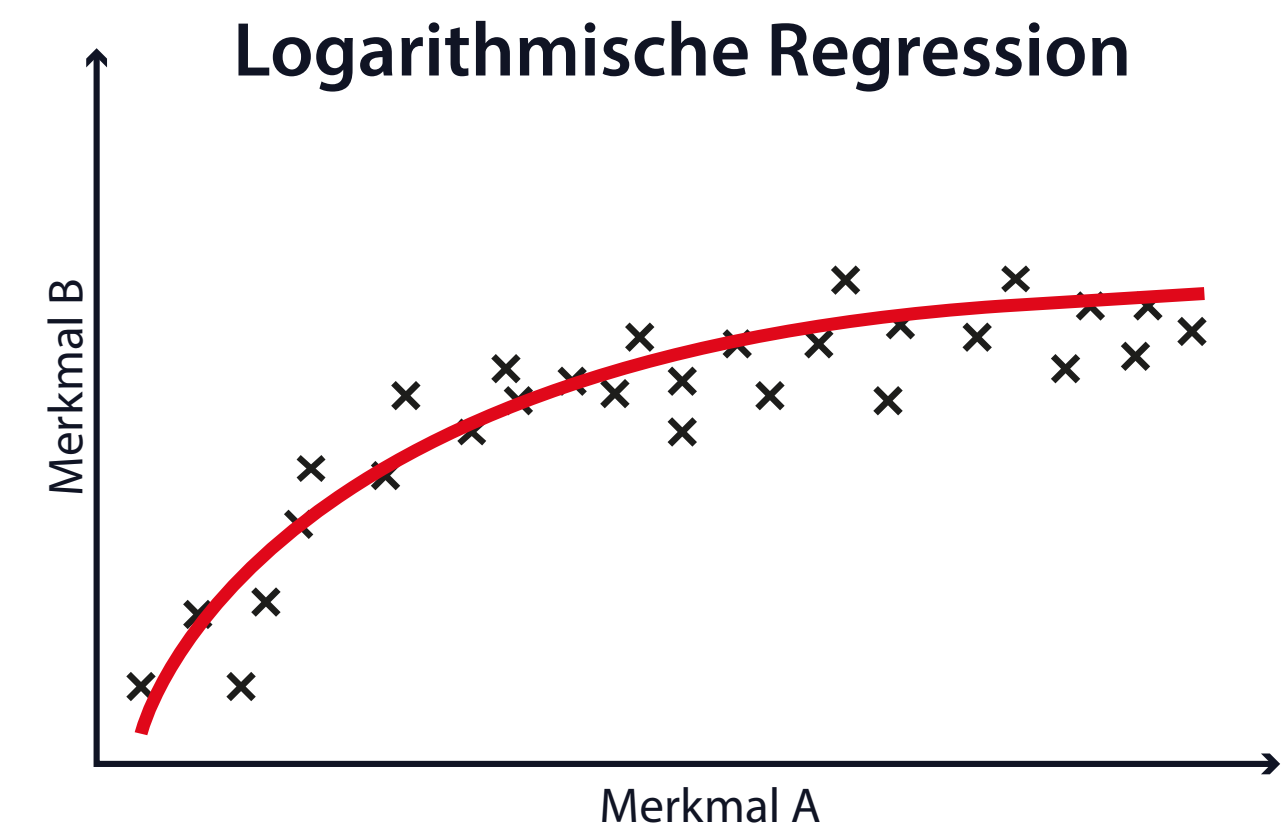
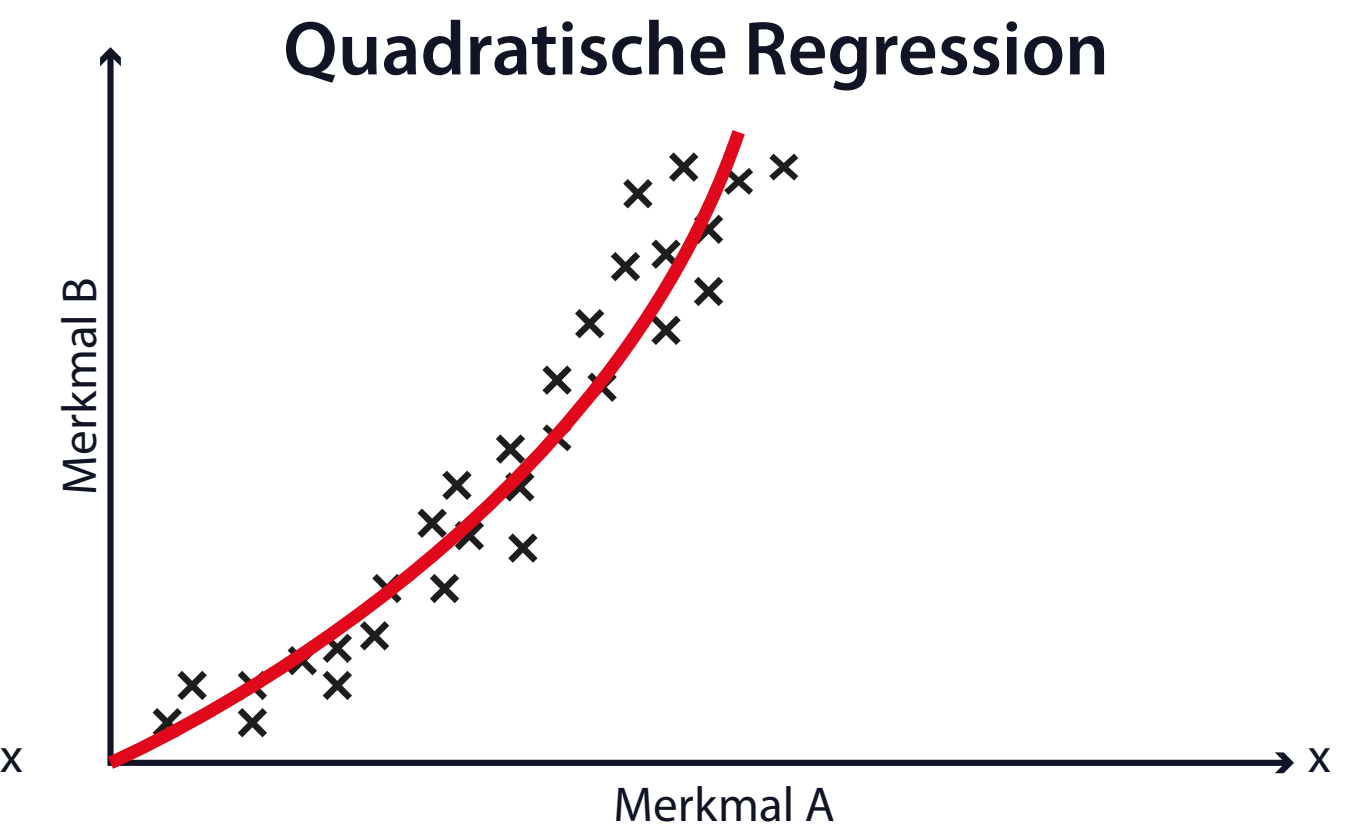
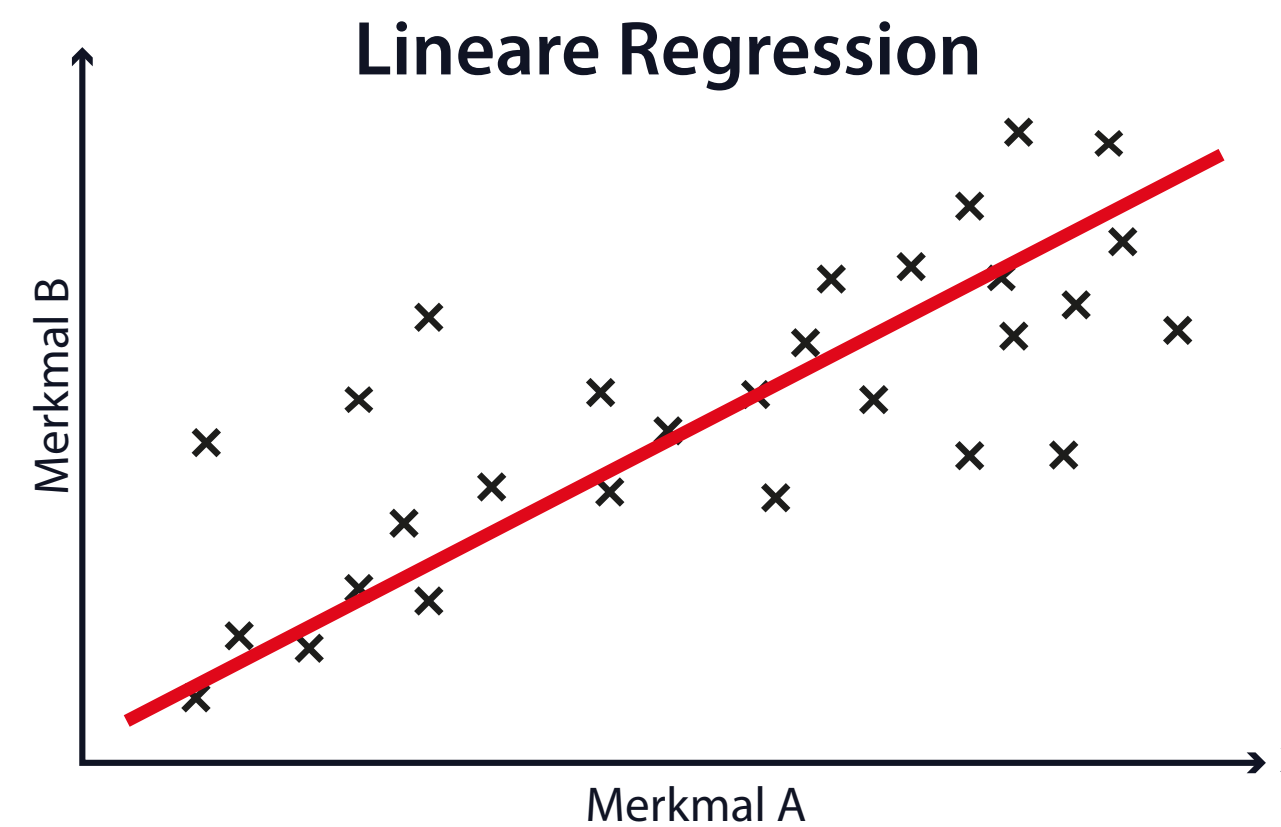
	Eine unabhängige Variable	Mehrere unabhängige Variablen
Eine abhängige Variable	Univariate einfache Regression	Univariate multiple Regression
Mehrere abhängige Variablen	Multivariate einfache Regression	Multivariate multiple Regression

Regressionsanalyse

Bei der linearen Regression suchen wir nach linearen Zusammenhängen.

Es gibt auch nicht lineare Regressionsmodelle: quadratisch, logarithmisch, polynomial usw.

Welches Modell wir wählen, hängt von den Daten und der Theorie hinter den Daten ab.



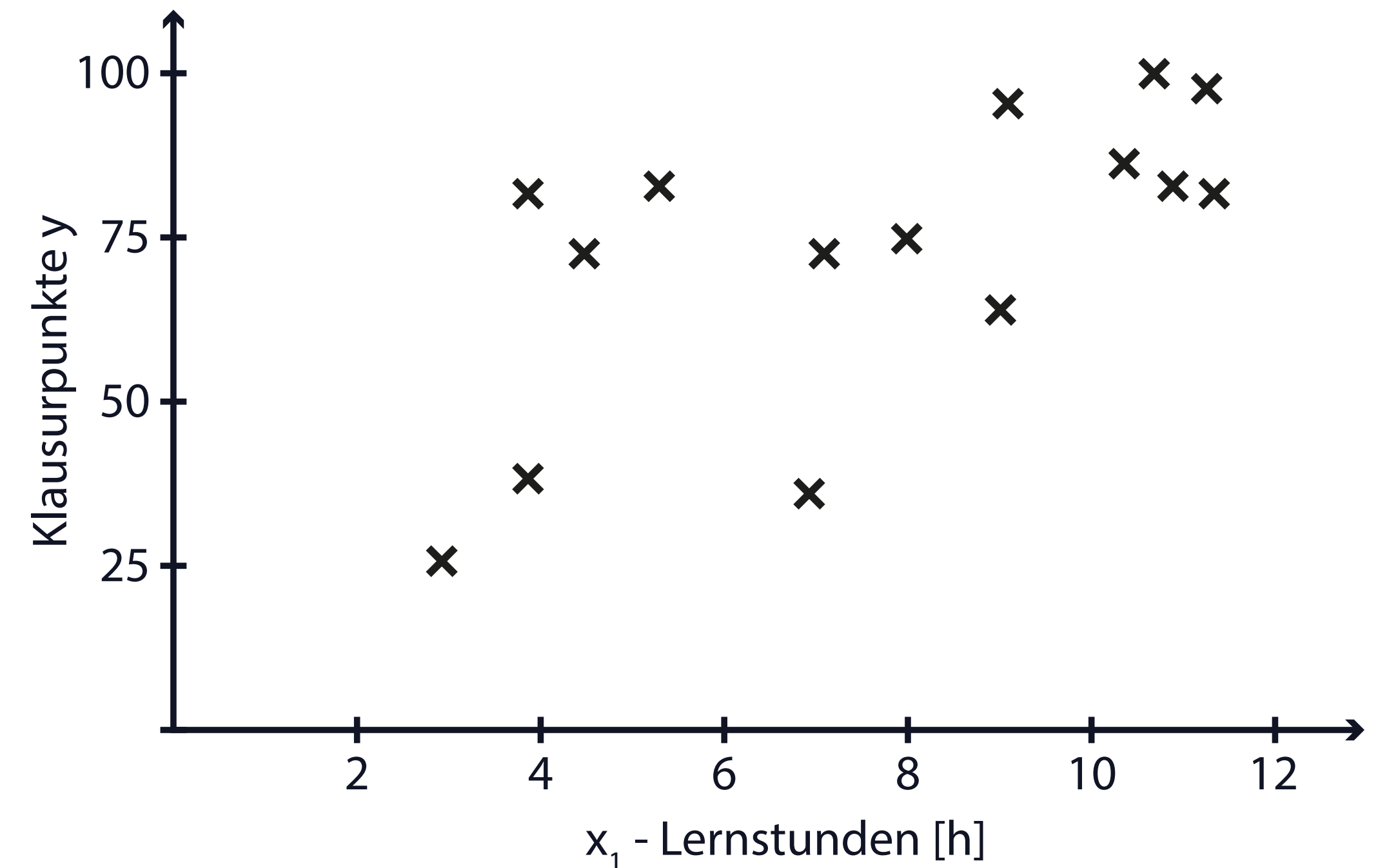
OLS-Regression

Beginnen wir mit einer einfachen, univariaten, linearen Regression: Wir vermuten, dass die in einer Klausur erreichten Punkte durch folgendes Modell erklärt werden können:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Punkte ohne Lernen $\rightarrow \beta_0$
 Punkte pro Lernstunde $\rightarrow \beta_1$
 Lernstunden $\rightarrow x_1$
 Punkte in Klausur $\rightarrow y$
 Alle anderen Einflüsse $\rightarrow \varepsilon$

„Ohne Lernen werden im Schnitt β_0 Punkte erreicht. Eine Stunde mehr Lernen ist im Durchschnitt mit β_1 mehr Punkten assoziiert.“

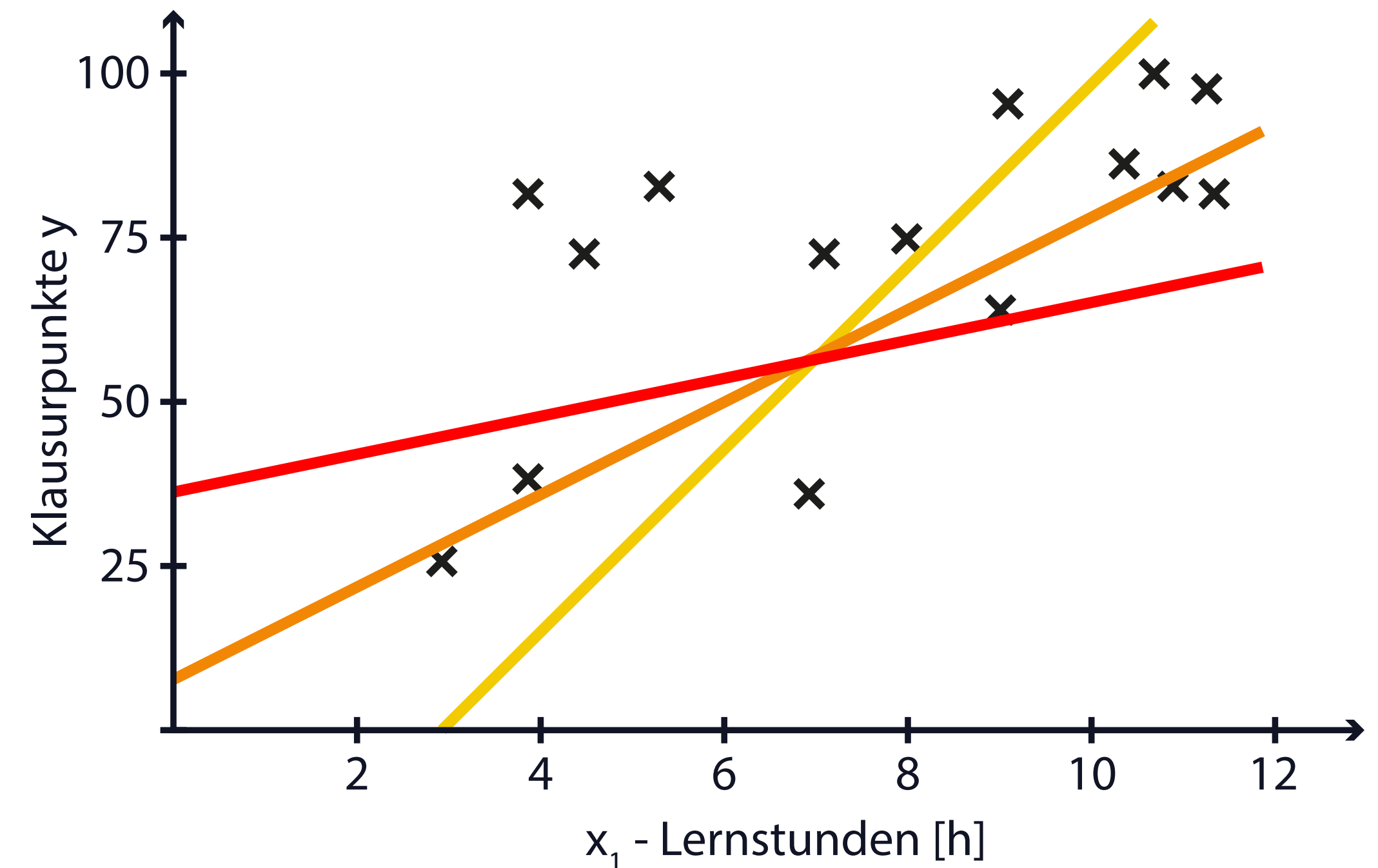


OLS-Regression

Die Aufgabe der Regression ist es Schätzwerte für die Koeffizienten β_1 und β_0 zu finden.

Bei der einfachen Regression können wir diese Schätzer grafisch als Achsenabschnitt β_0 und Steigung β_1 interpretieren!

Aber was entscheidet darüber, ob die Schätzwerte gut zu den Daten passen?



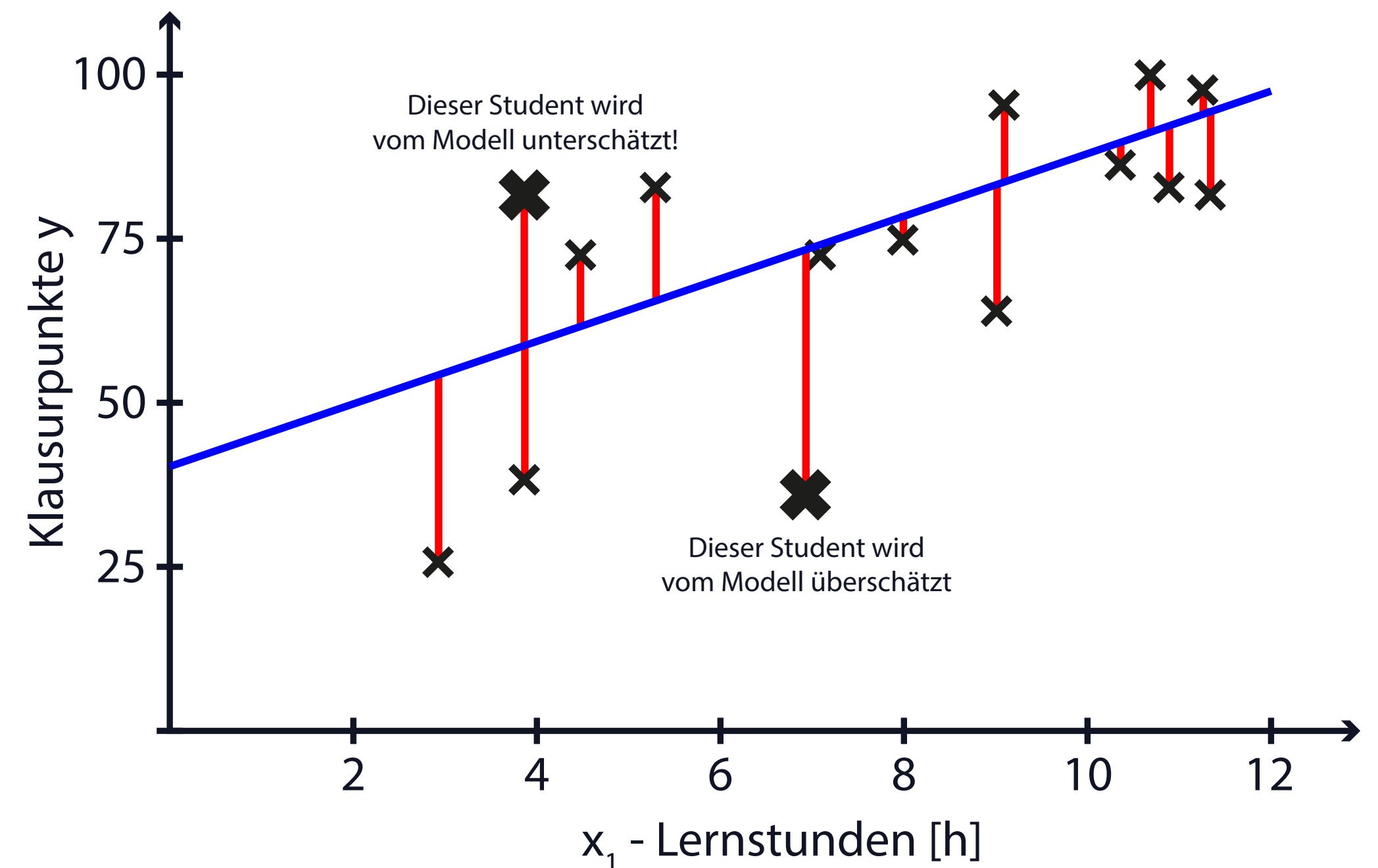
OLS-Regression

Der Algorithmus bewertet seine Modelle durch ein Maß, dass die Abweichung des Modells von der Realität misst.

Diese Abweichungen werden als **Residuen** bezeichnet.

Die OLS-Regression bewertet Modelle über die Summe der quadrierten Residuen. Sie sucht Schätzer, bei denen die Summe der quadrierten Residuen minimal wird.

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



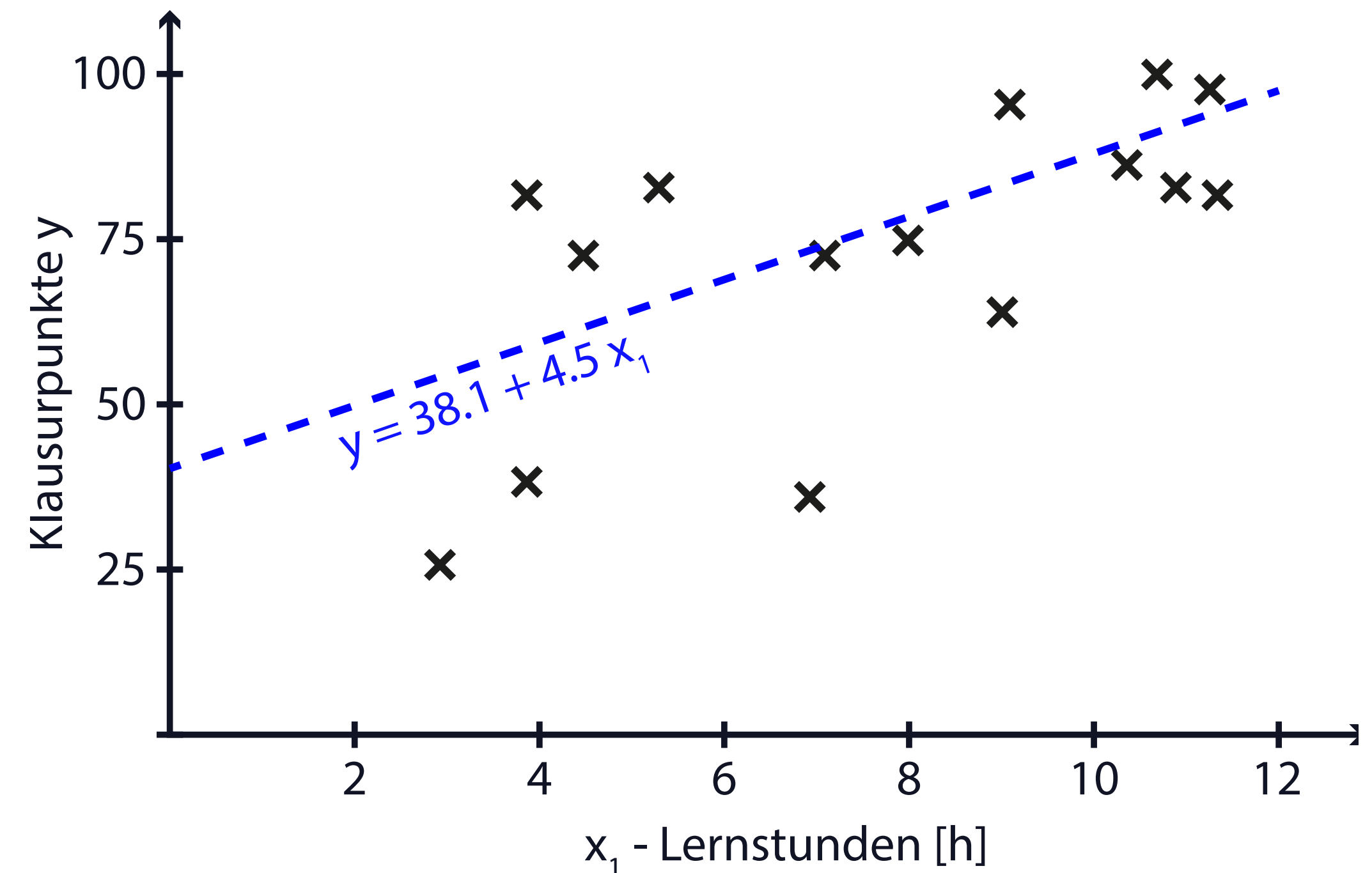
OLS-Regression

Der Hut unterscheidet die Koeffizienten von den dafür gefundenen Schätzern.

$$\hat{\beta}_0 = 38.1, \quad \hat{\beta}_1 = 4.5$$

„Ohne Lernen werden im Schnitt 38.1 Punkte erreicht. Eine Stunde mehr Lernen ist im Durchschnitt mit 4.534 mehr Punkten assoziiert.“

Warum ist der Satz so unnötig kompliziert formuliert?

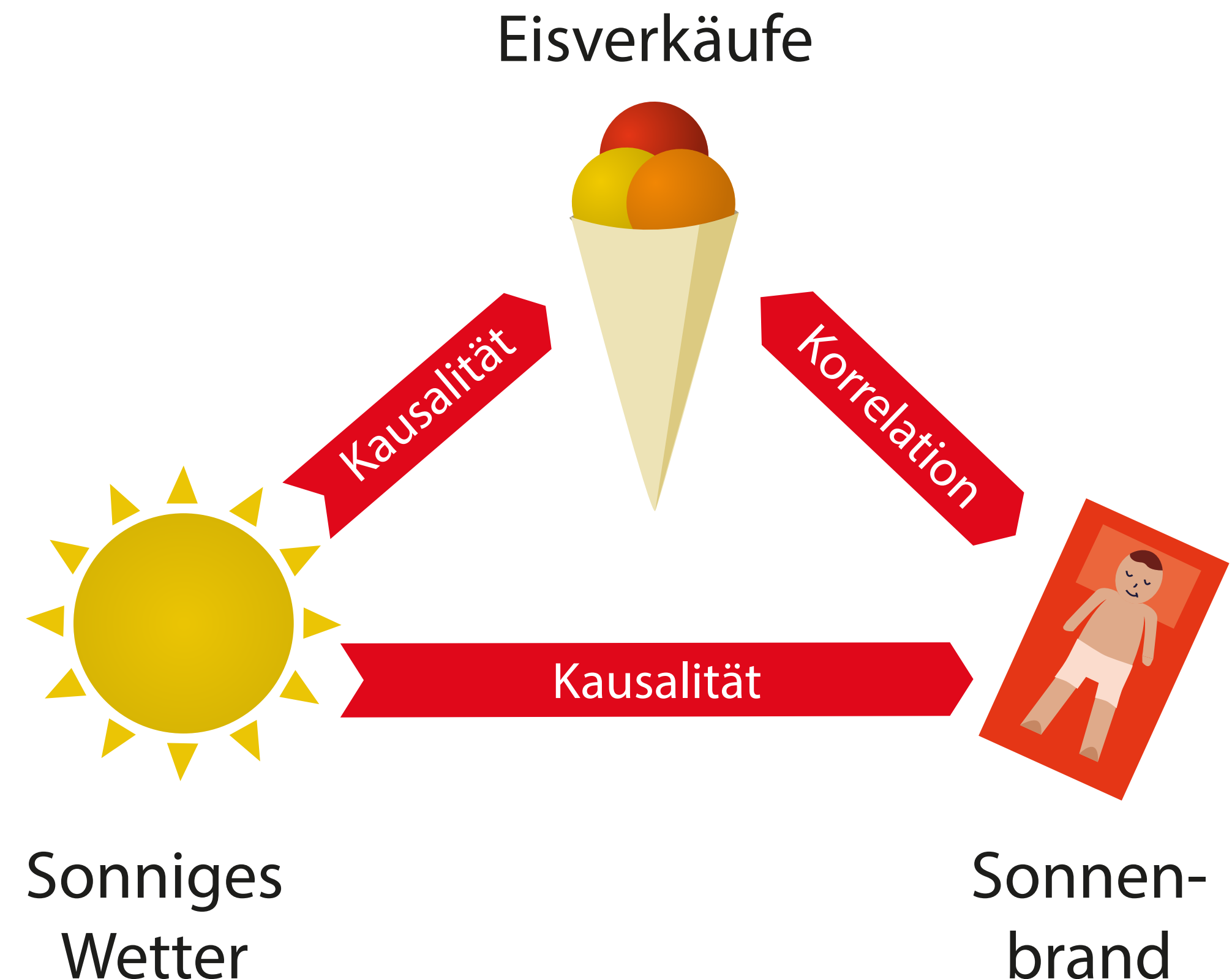


OLS-Regression

Das Ergebnis ist nicht zwingend kausal! Hinter dem gefundenen Zusammenhang kann eine dritte Variable stecken, die wir nicht berücksichtigen und daher zum Störterm ε gehört.

Im Beispiel könnte das z. B. Geschlecht, Gewissenhaftigkeit, Willenskraft oder Intelligenz sein.

Haben wir diese Werte, können wir sie als **Kontrollvariablen** in die Regression mitaufnehmen.



OLS-Regression

Als Ergebnis erhalten wir nicht nur die Schätzer für die beiden Koeffizienten, sondern auch einen Standardfehler für jeden Koeffizienten.

Wie sicher ist sich die Regression mit dem angezeigten Schätzwert?

Je größer der Standardfehler, umso weniger sicher ist sich die Regression!

```

Console | Terminal | Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.100     12.861    2.962   0.0110 *
Lernstunden    4.534      1.583    2.864   0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3398
F-statistic: 8.204 on 1 and 13 DF,  p-value: 0.01329

> confint(result, level=0.95)
              2.5 %    97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363
  
```

OLS-Regression

Hinter jedem Schätzer β_x steht ein Hypothesenpaar:

Nullhypothese Der wahre Wert des Koeffizienten ist 0. Es gibt daher keinen Zusammenhang zwischen der entsprechenden unabhängigen und der abhängigen Variable.

Alternativhypothese Der wahre Wert des Koeffizienten ist nicht 0. Es gibt einen Zusammenhang zwischen der entsprechenden unabhängigen und der abhängigen Variable.

Student	Lernstunden	Punkte
1	11,5	81
2	3,9	37
3	10,7	82
4	3,8	79
5	10,2	100
6	6,6	44
7	4,8	79
8	3,1	24
9	7,8	74
10	6,9	72

OLS-Regression

In unserem Beispiel haben wir folgendes Hypothesenpaar:

Nullhypothese Es gibt keinen Zusammenhang zwischen der Zeit die wir auf die Klausur gelernt haben und den erzielten Punkten ($\beta_1 = 0$).

Alternativhypothese Es gibt einen Zusammenhang zwischen der Zeit, die wir auf die Klausur gelernt haben, und den erzielten Punkten ($\beta_1 \neq 0$).

Student	Lernstunden	Punkte
1	11,5	81
2	3,9	37
3	10,7	82
4	3,8	79
5	10,2	100
6	6,6	44
7	4,8	79
8	3,1	24
9	7,8	74
10	6,9	72

OLS-Regression

Aus Schätzer und Standardfehler berechnet R automatisch den t-Wert und den p-Wert.

Letzterer beantwortet uns folgende Frage:

Gegeben die Nullhypothese ist wahr und das β_1 ist eigentlich 0 - wie hoch ist dann die Wahrscheinlichkeit, dass wir trotzdem einen mindestens so starken Zusammenhang in den Daten sehen?

```

Console | Terminal | Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.100     12.861    2.962   0.0110 *
Lernstunden    4.534      1.583    2.864   0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3398
F-statistic: 8.204 on 1 and 13 DF,  p-value: 0.01329

> confint(result,level=0.95)
              2.5 %      97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363

```

OLS-Regression

Der p-Wert für β_1 beträgt 1.3%.

Wenn die Nullhypothese wahr wäre, dann würden wir mit einer Wahrscheinlichkeit von 1.3% eine Stichprobe ziehen, bei der ein Zusammenhang von mindestens 4.534 Pkt./h besteht.

Die Wahrscheinlichkeit, dass der gefundene Zusammenhang nur Zufall ist, ist also verschwindend gering!

```

Console | Terminal | Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.100     12.861    2.962   0.0110 *
Lernstunden    4.534      1.583    2.864   0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3398
F-statistic: 8.204 on 1 and 13 DF,  p-value: 0.01329

> confint(result,level=0.95)
              2.5 %    97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363

```

OLS-Regression

Üblicherweise verwerfen wir die Nullhypothese, wenn der p-Wert unter 5% liegt. Das ist hier gegeben!

~~H0 für β_1 - Der wahre Wert von β_1 ist null und Lernen hat keinen Effekt auf die erreichte Klausurpunktzahl.~~

H1 für β_1 - Der wahre Wert ist nicht null und Lernen hat einen Effekt auf die erreichte Klausurpunktzahl.

```

Console Terminal Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.100      12.861   2.962   0.0110 *
Lernstunden    4.534       1.583   2.864   0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3398
F-statistic: 8.204 on 1 and 13 DF,  p-value: 0.01329

> confint(result,level=0.95)
              2.5 %      97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363
  
```

OLS-Regression

Wir wissen jetzt, dass es einen positiven Zusammenhang gibt, aber wie stark ist dieser?

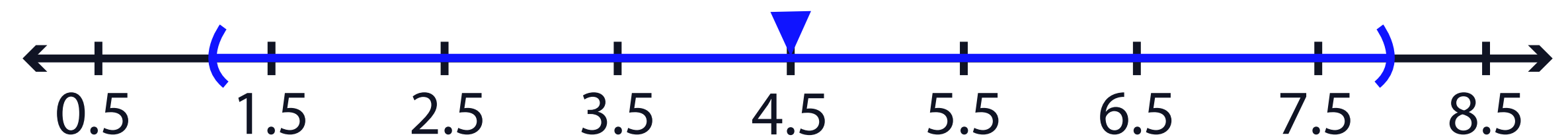
An dieser Stelle kommt das angegebene Konfidenzintervall ins Spiel.

Die Regression ist sich zu 95% sicher, dass der Effekt einer Lernstunde zwischen 1.1 und 7.9 Punkten liegt.

```
(Intercept) 38.100 12.861 2.962 0.0110 *
Lernstunden 4.534 1.583 2.864 0.0133 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared: 0.3869, Adjusted R-squared: 0.3398
F-statistic: 8.204 on 1 and 13 DF, p-value: 0.01329

> confint(result, level=0.95)
                2.5 %    97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363
```



Zu 95% liegt der wahre Wert von β_1 zwischen 1.1 und 7.9 $\pm 1.96 \sigma$

OLS-Regression

Welcher Anteil der Varianz in der abhängigen Variable wird durch unsere unabhängigen Variablen erklärt? Die Antwort gibt der R^2 -Wert!

33.98% der Varianz in den Leistungen werden durch unterschiedliche Länge des Lernens erklärt.

66.02% hängen von anderen Faktoren ab (Tagesform, Begabung, Lerntechnik usw.)

```

Console | Terminal | Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.100      12.861   2.962   0.0110 *
Lernstunden    4.534       1.583   2.864   0.0133 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.94 on 13 degrees of freedom
Multiple R-squared:  0.3869, Adjusted R-squared:  0.3398
F-statistic: 8.204 on 1 and 13 DF, p-value: 0.01329

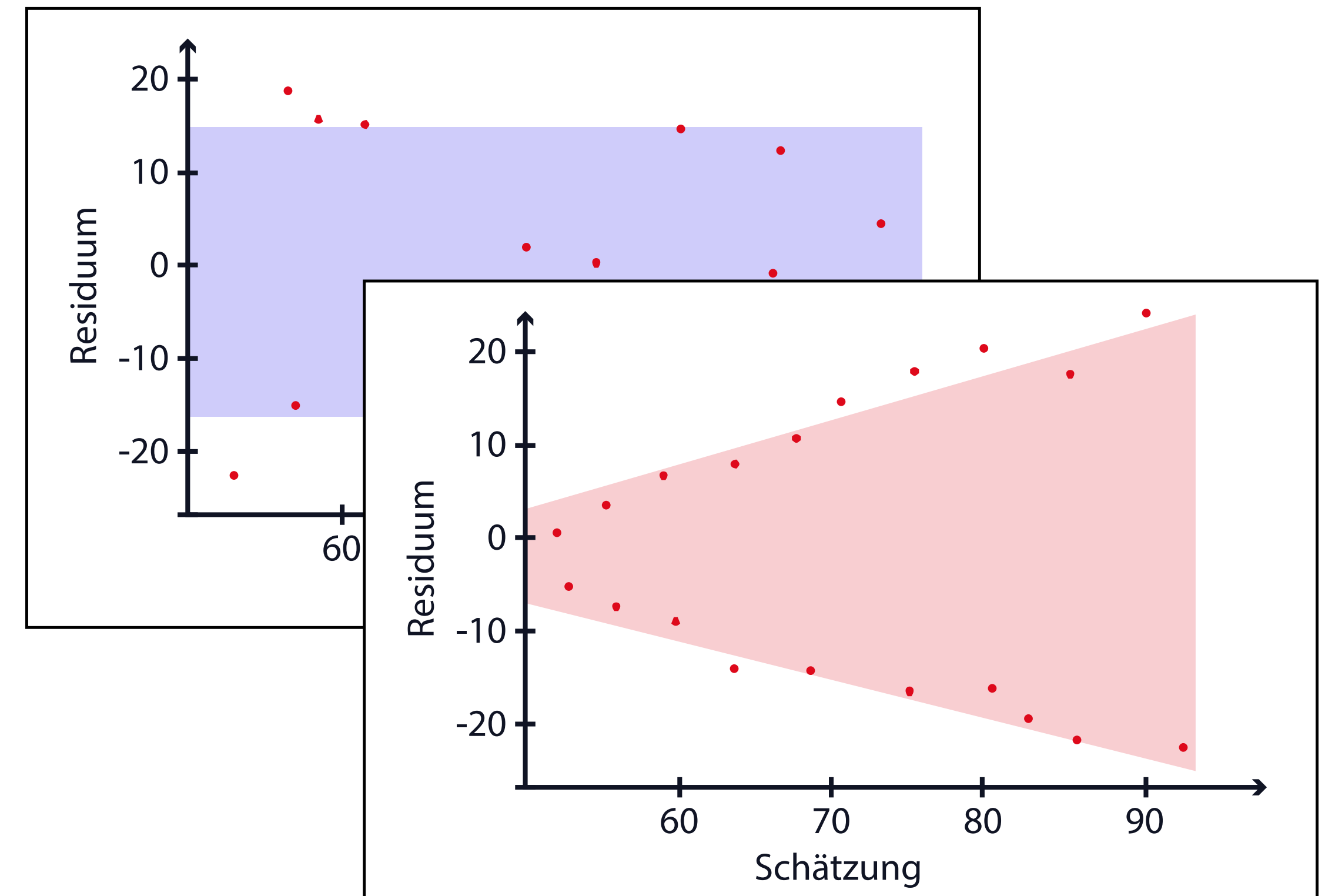
> confint(result, level=0.95)
              2.5 %      97.5 %
(Intercept) 10.315732 65.885123
Lernstunden  1.114246  7.953363
  
```

OLS-Regression

Neben den Schätzern sind auch die Residuen relevant, d. h. die Abweichungen zwischen der Modellschätzung und den wahren Werten der abhängigen Variable.

Wir untersuchen die Residuen üblicherweise auf Normalverteilung und auf Homoskedastizität.

Homoskedastizität liegt vor, wenn die Residuen von der Höhe des geschätzten Merkmals unabhängig sind.

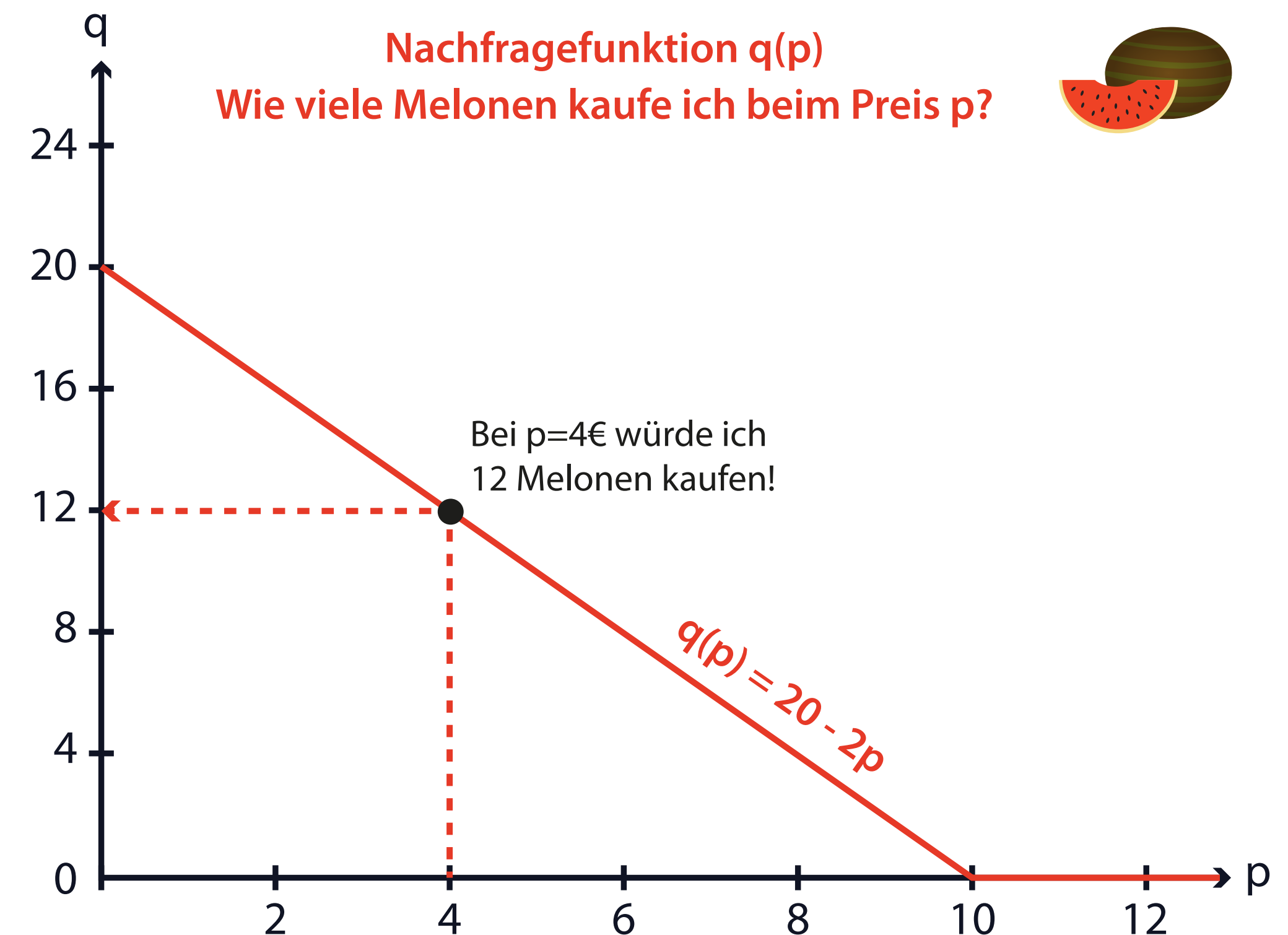


Beispiel Strommarkt

Aus der Mikroökonomik kennen wir Nachfragefunktionen.

Nachfragefunktionen ordnen jedem Preis eine nachgefragte Menge zu, wobei höhere Preise typischerweise zu weniger Nachfrage führen.

Die Parameter dieser Nachfragefunktionen und die Ergebnisse, die wir mit ihnen berechnet haben, waren nicht selten alles andere als realistisch!



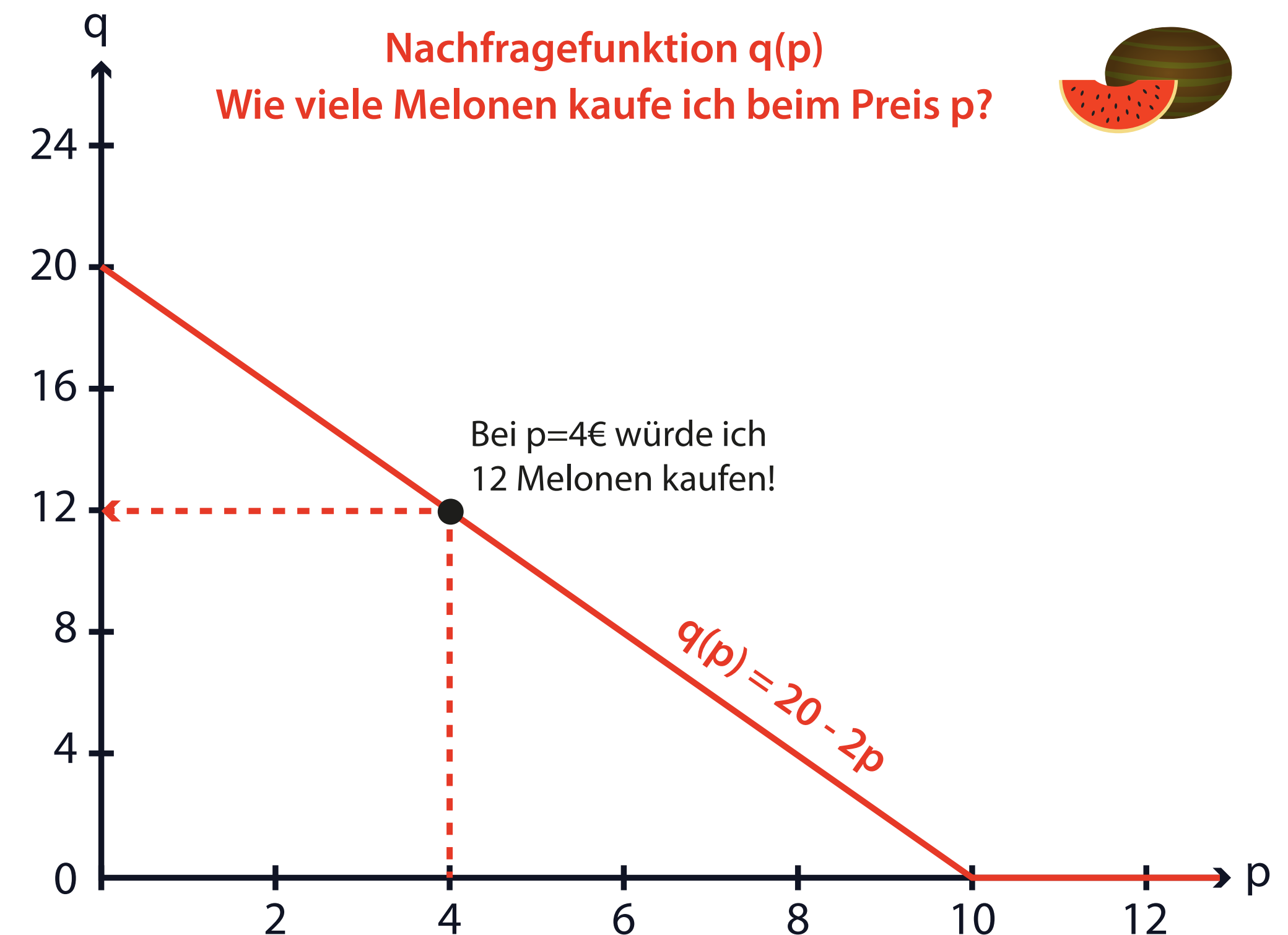
Beispiel Strommarkt

Wir wollen nun eine reale Nachfragefunktion schätzen. Wir gehen dabei von einer linearen Nachfrage aus ...

$$q = \beta_0 + \beta_1 \cdot p + \varepsilon$$

... und wollen Werte für β_0 und β_1 finden, mit denen wir das Nachfrageverhalten aus der Realität so gut wie möglich beschreiben.

Wir benötigen einen Datensatz mit Preisen und Mengen!



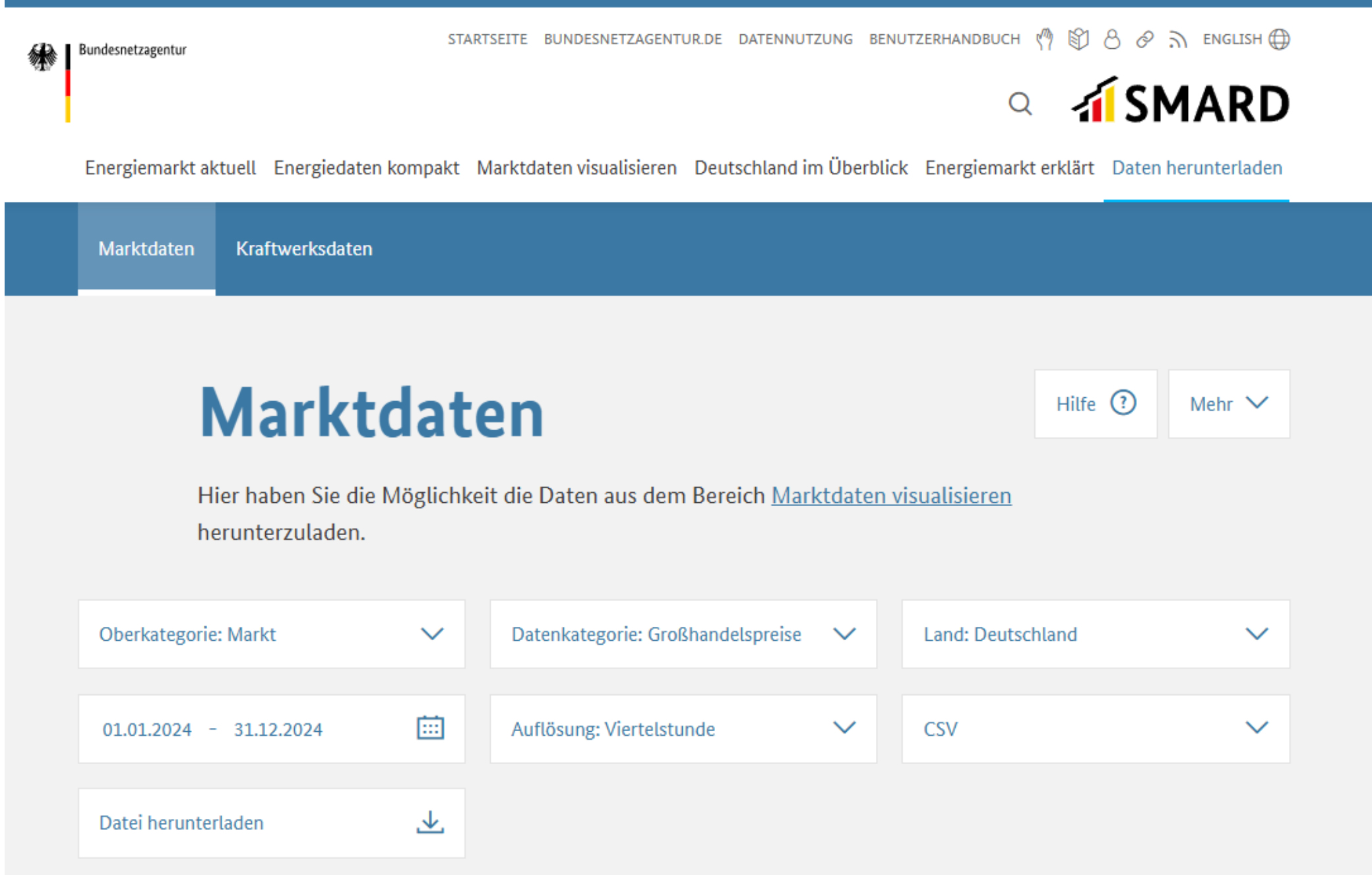
Beispiel Strommarkt

SMARD-Datenbank Für den Strommarkt gibt es eine öffentlich zugängliche Datenbank mit Daten zu Stromerzeugung, Stromverbrauch und Strompreisen.

Wir verwenden viertelstündliche Daten aus 2024 und führen eine Regressionsanalyse durch.

Abhängige Variable: Nachfrage

Unabhängige Variable: Strompreis



Bundesnetzagentur

STARTSEITE BUNDESNETZAGENTUR.DE DATENNUTZUNG BENUTZERHANDBUCH ENGLISH

SMARD

Energiemarkt aktuell Energiedaten kompakt Marktdaten visualisieren Deutschland im Überblick Energiemarkt erklärt Daten herunterladen

Marktdaten Kraftwerksdaten

Marktdaten Hilfe ? Mehr ▾

Hier haben Sie die Möglichkeit die Daten aus dem Bereich [Marktdaten visualisieren](#) herunterzuladen.

Oberkategorie: Markt ▾

Datenkategorie: Großhandelspreise ▾

Land: Deutschland ▾

01.01.2024 - 31.12.2024 📅

Auflösung: Viertelstunde ▾

CSV ▾

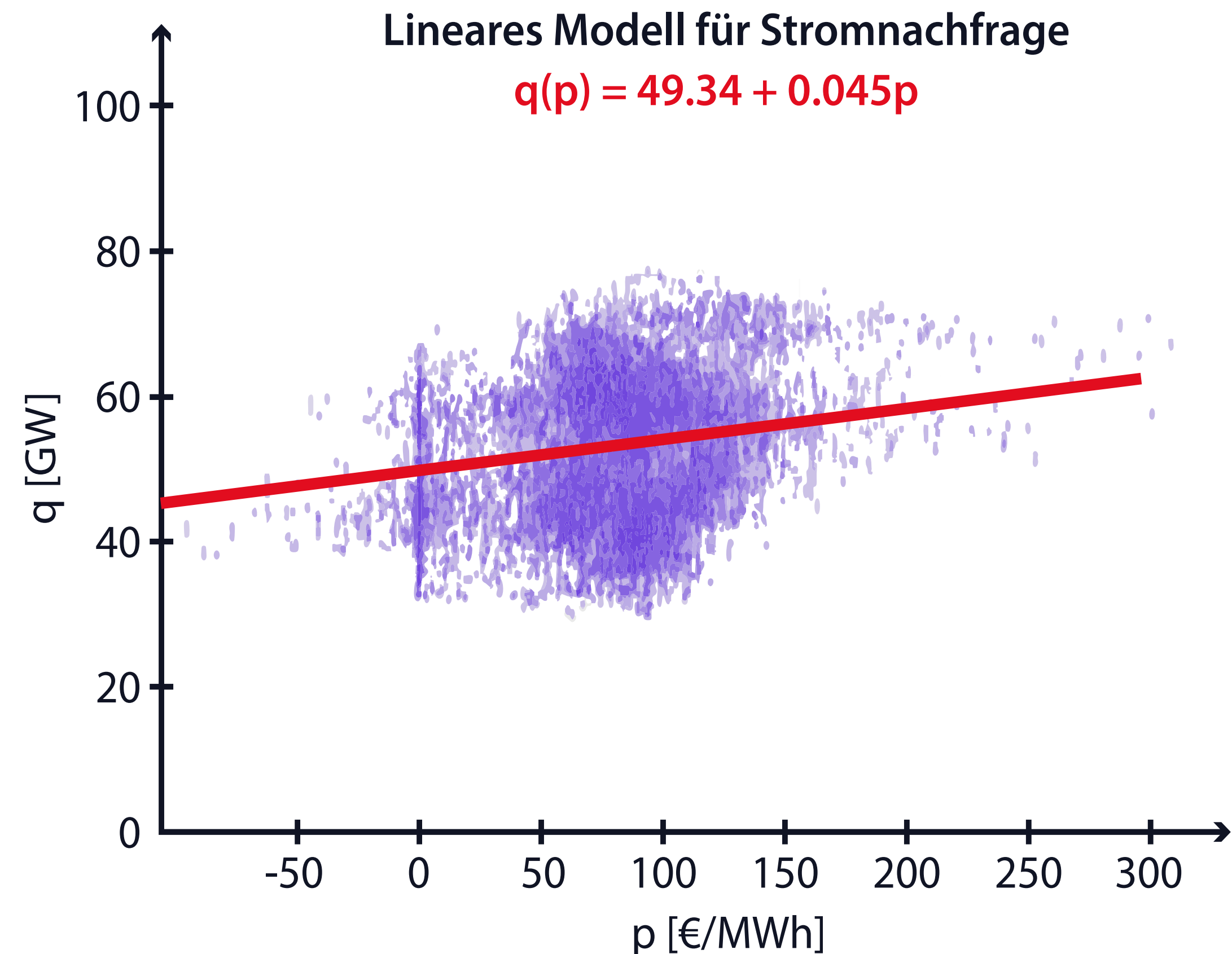
Datei herunterladen ⬇

Beispiel Strommarkt

Die OLS-Regression liefert folgende Schätzer:

$$\hat{\beta}_0 = 49.34, \quad \hat{\beta}_1 = 0.045$$

Da wir eine einfache univariate Regression haben, können wir dieses Ergebnis in einem Liniendiagramm visualisieren!



Beispiel Strommarkt

Darüber hinaus untersuchen wir auch das folgende Hypothesenpaar:

H0 - Der wahre Wert von β_1 ist null. Die Stromnachfrage hängt nicht mit dem Strompreis zusammen.

H1 - Der wahre Wert von β_1 ist nicht null. Die Stromnachfrage hängt mit dem Strompreis zusammen.

```

Console Terminal Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.934e+01  8.420e-02  586.02  <2e-16 ***
Preis       4.545e-02  8.903e-04   51.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.799 on 35134 degrees of freedom
Multiple R-squared:  0.06905,    Adjusted R-squared:  0.06903
F-statistic: 2606 on 1 and 35134 DF,  p-value: < 2.2e-16

> confint(result, level=0.95)
              2.5 %      97.5 %
(Intercept) 49.17592502 49.5059822
Preis       0.04370479  0.0471948
  
```

Beispiel Strommarkt

Der p-Wert ist fast 0 und damit unter 0.05.

Außerdem ist sich die Regression zu 95% sicher, dass der wahre Wert zwischen 0.0437 und 0.0472 liegen muss.

Der gefundene Schätzer ist signifikant und wir können die Nullhypothese verwerfen!

```

Console | Terminal | Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.934e+01  8.420e-02  586.02  <2e-16 ***
Preis       4.545e-02  8.903e-04   51.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.799 on 35134 degrees of freedom
Multiple R-squared:  0.06905,    Adjusted R-squared:  0.06903
F-statistic: 2606 on 1 and 35134 DF,  p-value: < 2.2e-16

> confint(result, level=0.95)
              2.5 %      97.5 %
(Intercept) 49.17592502 49.5059822
Preis       0.04370479  0.0471948
  
```

Beispiel Strommarkt

Irgendwas an unseren Ergebnissen ist seltsam:

Eine um 1€ höherer Großhandelspreis pro MWh geht im Durchschnitt mit einer um 45 MW höheren Nachfrage einher.

Die Stromkunden wollen 45 MW mehr Leistung, wenn der Strom an der Börse 1€ teurer wird?



Doubt

```

Console Terminal Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.934e+01  8.420e-02  586.02  <2e-16 ***
Preis       4.545e-02  8.903e-04   51.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.799 on 35134 degrees of freedom
Multiple R-squared:  0.06905,    Adjusted R-squared:  0.06903
F-statistic: 2606 on 1 and 35134 DF,  p-value: < 2.2e-16

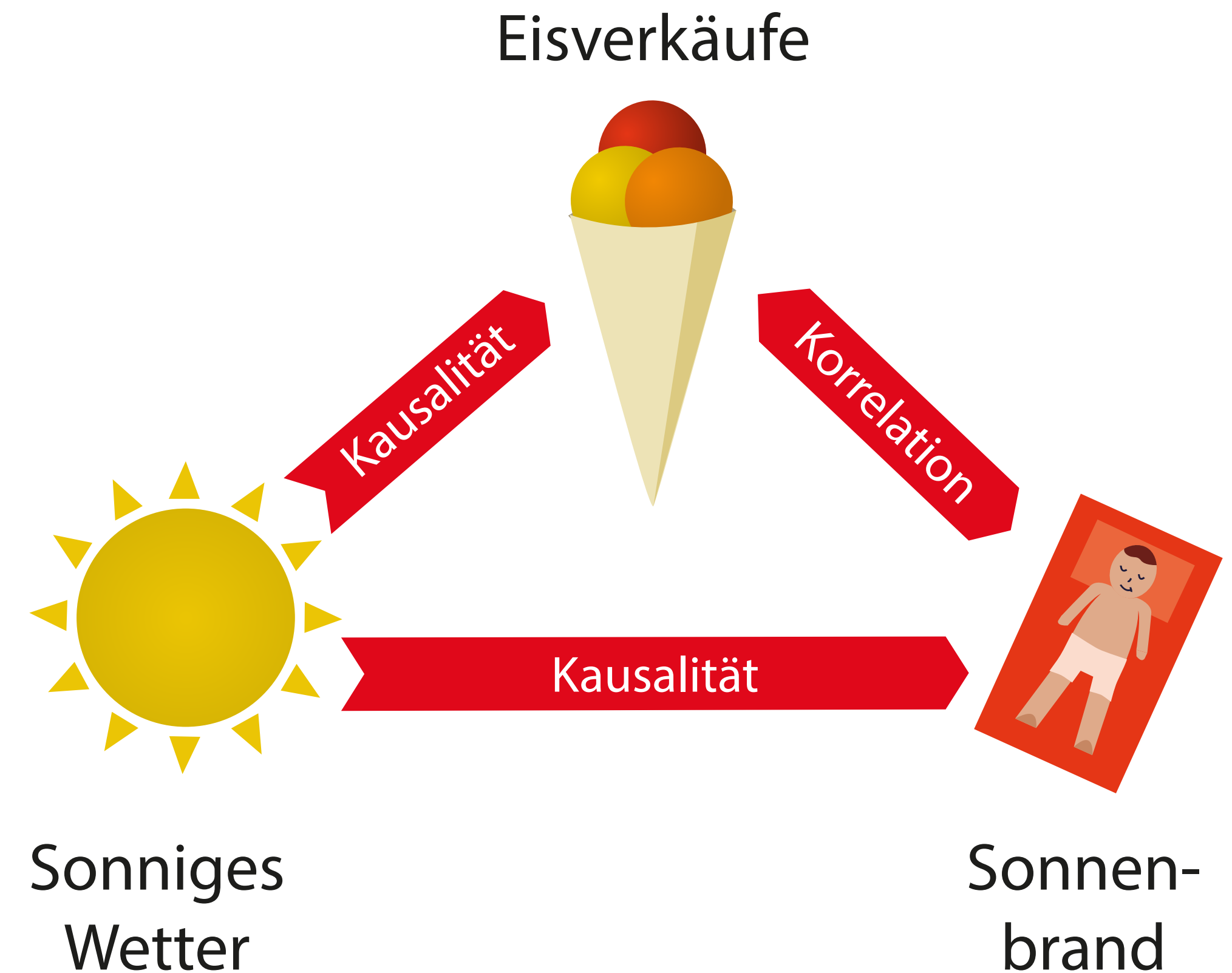
> confint(result, level=0.95)
              2.5 %      97.5 %
(Intercept) 49.17592502 49.5059822
Preis       0.04370479  0.0471948
  
```

Beispiel Strommarkt

Wir erinnern uns: Die OLS-Regression macht ausschließlich Aussagen über Korrelationen!

Trotz minimalem p-Wert beweisen wir weder, ob es eine Kausalität gibt, noch in welche Richtung diese geht.

Der gesunde Menschenverstand legt nahe: Der Börsenpreis wird 1€ teurer, wenn die Konsumenten 45 MW mehr Leistung nachfragen.



Beispiel Strommarkt

Statt der Nachfragefunktion schätzen wir nun die Preis-Absatz-Funktion mit dem Modell:

$$p = \beta_0 + \beta_1 \cdot q + \varepsilon$$

Die Interpretation ist nun einfacher, aber wir haben ein weiteres Problem: Das Modell erklärt nur einen geringen Teil der Preisschwankungen!

6.90% der Preisschwankungen werden durch Nachfrageschwankungen erklärt.

```

Console Terminal Background Jobs
R 4.2.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.87667    1.59791  -1.174    0.24
Netzlaster    1.51937    0.02976  51.050 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.87 on 35134 degrees of freedom
Multiple R-squared:  0.06905, Adjusted R-squared:  0.06903
F-statistic: 2606 on 1 and 35134 DF, p-value: < 2.2e-16

> confint(result, level=0.95)
              2.5 %    97.5 %
(Intercept) -5.008626 1.255278
Netzlaster    1.461033 1.577703
  
```

Beispiel Strommarkt

Wir erweitern das Modell um die EE-Einspeisung:

$$p = \beta_0 + \beta_1 \cdot q + \beta_2 \cdot EE + \varepsilon$$

Logik: wenn günstiger erneuerbarer Strom verfügbar ist, müssen weniger teure Kraftwerke laufen und der Strom wird nach dem Merit-Order-Prinzip günstiger.

Das Modell wird deutlich besser!

Console	Terminal	Background Jobs
R 4.2.2		
<pre> Estimate Std. Error t value Pr(> t) (Intercept) 3.76256 1.03394 3.639 0.000274 *** Netzlast 3.11734 0.02057 151.583 < 2e-16 *** GesamtEE -3.10775 0.01406 -220.982 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 32.91 on 35133 degrees of freedom Multiple R-squared: 0.6105, Adjusted R-squared: 0.6105 F-statistic: 2.753e+04 on 2 and 35133 DF, p-value: < 2.2e-16 </pre>		
<pre> 2.5 % 97.5 % (Intercept) 1.736001 5.789118 Netzlast 3.077026 3.157644 GesamtEE -3.135317 -3.080187 </pre>		

Beispiel Strommarkt

Eine Nachfrageerhöhung um 1 GW geht im Durchschnitt mit einer Erhöhung des Börsenpreises um 3.12€ einher.

Ein Plus von 1 GW EE-Leistung geht im Durchschnitt mit einer Senkung des Börsenpreises um 3.11€ einher.

Console	Terminal	Background Jobs
R 4.2.2		
<pre> Estimate Std. Error t value Pr(> t) (Intercept) 3.76256 1.03394 3.639 0.000274 *** Netzlast 3.11734 0.02057 151.583 < 2e-16 *** GesamtEE -3.10775 0.01406 -220.982 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 32.91 on 35133 degrees of freedom Multiple R-squared: 0.6105, Adjusted R-squared: 0.6105 F-statistic: 2.753e+04 on 2 and 35133 DF, p-value: < 2.2e-16 </pre>		
<pre> 2.5 % 97.5 % (Intercept) 1.736001 5.789118 Netzlast 3.077026 3.157644 GesamtEE -3.135317 -3.080187 </pre>		

OLS-Regression

Verwende, den Mensadatensatz um eine OLS-Regression durchzuführen.

Abhängige Variable ist die Bewertung des Essens in der Spalte "food_score".

Unabhängige Variablen sind Geschlecht, Status, Essenspräferenz, Allergikerstatus sowie die Bewertung von Ambiente und Service.

Interpretiere das Ergebnis.

preference	allergy	food_tas	food_hear	food_che	food_spir	food_sat	food_loo	food_var	food_sco
none	yes	3	4	6	2	5	4	2	3,71
vegetarian	yes	4	4	6	2	5	3	2	3,71
vegetarian	no	5	4	6	5	6	5	5	5,14
none	no	4	5	5	4	5	5	5	4,71
none	no	4	4	5	4	4	3	3	4,43
none	no	3	4	4	4	4	3	3	3,86
none	no	5	4	6	5	6	4	6	5,14
none	no	4	2	3	1	2	3	2	2,43
vegetarian	no	4	4	6	5	6	4	4	4,71
vegetarian	no	2	2	1	2	3	2	4	2,29
vegetarian	no	4	4	6	3	5	1	3	3,71
none	no	4	3	5	2	5	3	3	3,57
vegetarian	no	3	4	2	2	2	3	5	3,00
none	no	4	4	5	3	5	2	4	3,86
vegetarian	no	5	3	6	5	6	4	6	5,00
vegetarian	no	5	4	6	5	6	5	6	5,29
vegetarian	no	4	4	6	2	5	4	6	4,43
vegetarian	yes	4	3	6	3	5	5	2	4,00
none	no	5	4	6	5	5	4	4	4,71
none	no	4	3	5	4	5	2	3	3,71
vegetarian	no	3	3	5	4	4	3	4	3,71
none	no	5	4	6	6	6	5	6	5,43
vegetarian	yes	2	3	5	3	6	2	3	3,43

OLS-Regression

Regressionsmodelle sind sehr mächtig ...

Transformationen und Interaktionen von unabhängigen Variablen sind möglich.

Kategoriale unabhängige Variablen können durch Dummy-Variablen eingebracht werden.

Kategoriale abhängige Variablen können durch Logit- oder Probit-Regression untersucht werden.



OLS-Regression

... aber es gibt eine ganze Reihe an Baustellen!

Durch **Multikollinearität** kann der Einfluss hoch korrelierter unabhängiger Variablen nur schwer untersucht werden.

Durch **Endogenitätsprobleme** können Schätzer systematisch verzerrt werden.

Die manuelle Wahl von unabhängigen Variablen kann zu **Overfitting** führen.



Multikollinearität

Ein Studiengangsleiter möchte ein Vorhersagemodell für die ersten Klausuren.

Er nimmt sich einen Anfängerkurs und lässt sie jeweils einen Kurztest zu Mathe und Coding schreiben.

Die Kurztests geben keine Note, sondern dienen als unabhängige Variablen für ein Regressionsmodell.

Mit den Noten aus dem ersten Studienjahr werden die Modellparameter geschätzt.

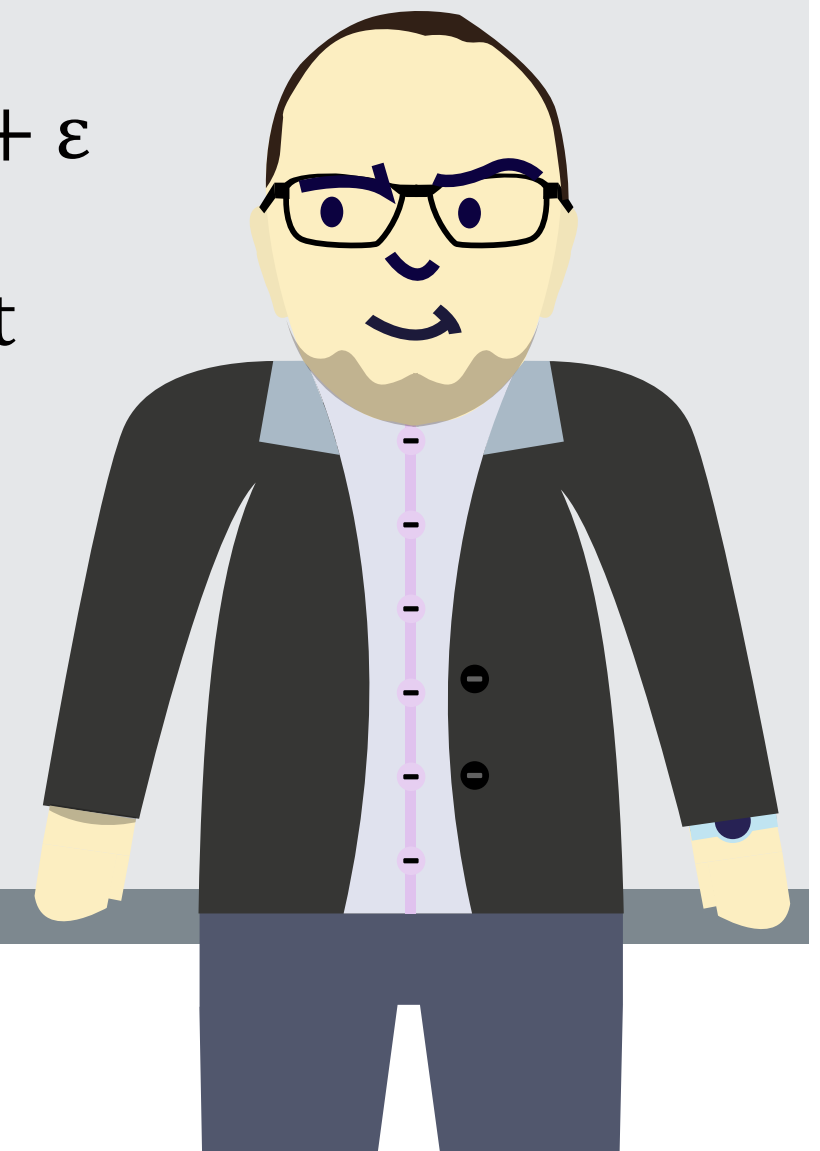
Vorhersage von Klausurergebnissen mit Kurztests zum Studienstart

$$\text{Leistung} = \beta_0 + \beta_1 \cdot \text{MS} + \beta_2 \cdot \text{CS} + \varepsilon$$

MS Ergebnis Mathekurztest

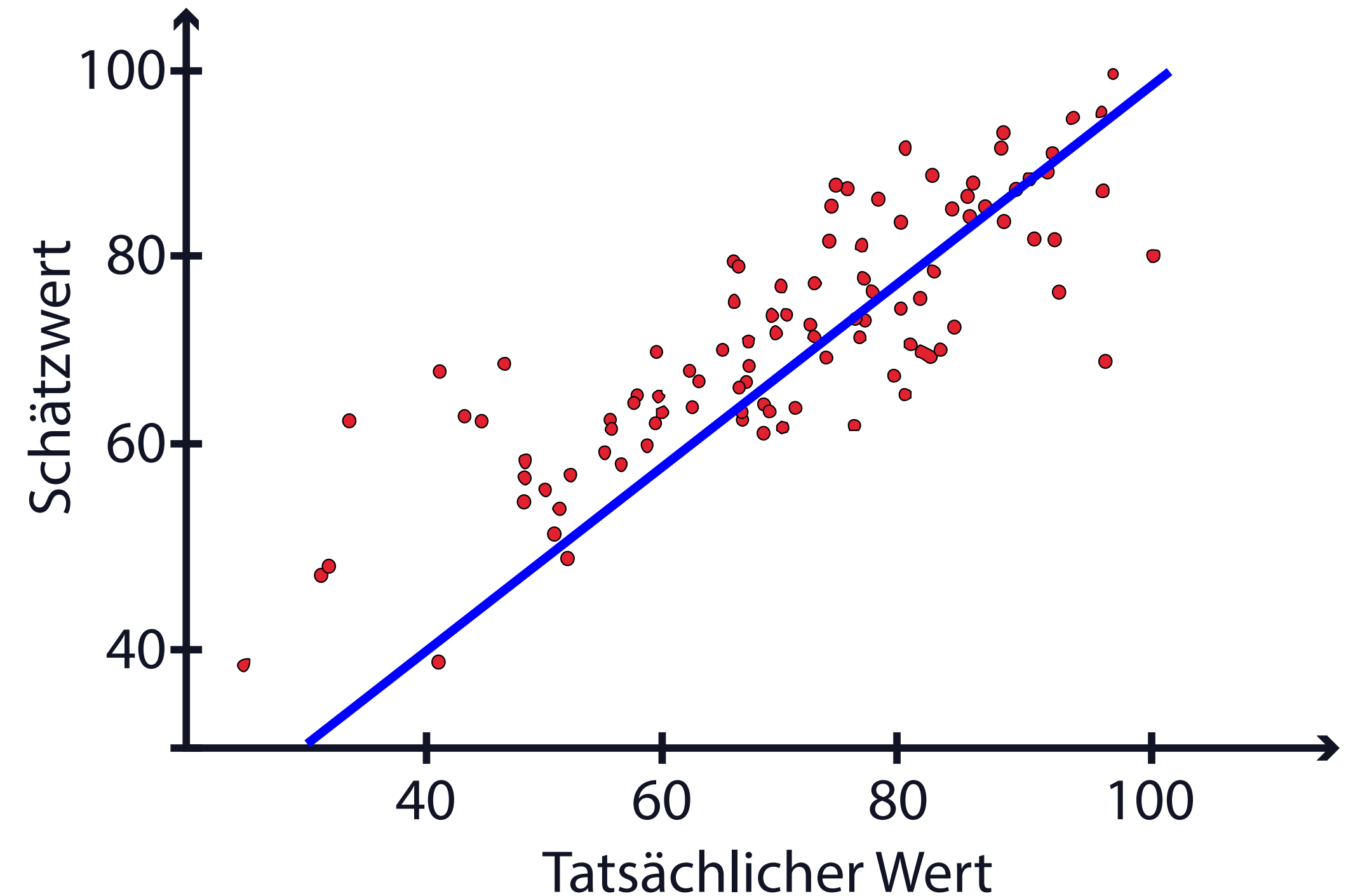
CS Ergebnis Codingtest

ε Sonstige Faktoren



Multikollinearität

Voller Stolz stellt er das geschätzte Modell seinen Kollegen vor. Es ist ziemlich gut darin, die Daten zu erklären!



Multikollinearität

Voller Stolz stellt er das geschätzte Modell seinen Kollegen vor. Es ist ziemlich gut darin, die Daten zu erklären!

Leider ergeben die Koeffizienten überhaupt keinen Sinn.

Ein um 1 Punkt besseres Ergebnis im Codingtest ist mit 0.12 weniger Klausurpunkten assoziiert.

```

Console Terminal Background Jobs
R 4.2.2
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1140     5.9043   2.052   0.0429 *
math_skill    0.9522     0.3668   2.596   0.0109 *
code_skill   -0.1222     0.4226  -0.289   0.7730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.19 on 97 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.6735
F-statistic: 103.1 on 2 and 97 DF,  p-value: < 2.2e-16

```

Multikollinearität

Da wir in diesem Beispiel mit **simulierten Daten** arbeiten, kennen wir die wahren Koeffizienten!

Wir können nicht nur die Residuen $\hat{y} - y$ und davon abgeleitete Fehlermaße, sondern auch die Abweichungen der einzelnen Schätzer $\hat{\beta} - \beta$ betrachten!

Dieser **Schätzfehler** $\hat{\beta} - \beta$ kann als Zufallsvariable mit einer bestimmten Verteilung betrachtet werden.

```

Console Terminal Background Jobs
R 4.2.2
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1140     5.9043   2.052   0.0429 *
math_skill    0.9522     0.3668   2.596   0.0109 *
code_skill   -0.1222     0.4226  -0.289   0.7730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.19 on 97 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.6735
F-statistic: 103.1 on 2 and 97 DF,  p-value: < 2.2e-16

```

Multikollinearität

Der Erwartungswert von $\hat{\beta} - \beta$ ist der **Bias**.

Wenn der Bias = 0 ist, bezeichnen wir den Schätzer als **erwartungstreu**. Ansonsten bezeichnen wir ihn als verzerrt.

Die Standardabweichung von $\hat{\beta} - \beta$ ist der zufällige Schätzfehler. Dieser wird mit größerer Stichprobengröße kleiner.

```

Console Terminal Background Jobs
R 4.2.2
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1140     5.9043   2.052  0.0429 *
math_skill    0.9522     0.3668   2.596  0.0109 *
code_skill   -0.1222     0.4226  -0.289  0.7730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.19 on 97 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.6735
F-statistic: 103.1 on 2 and 97 DF,  p-value: < 2.2e-16

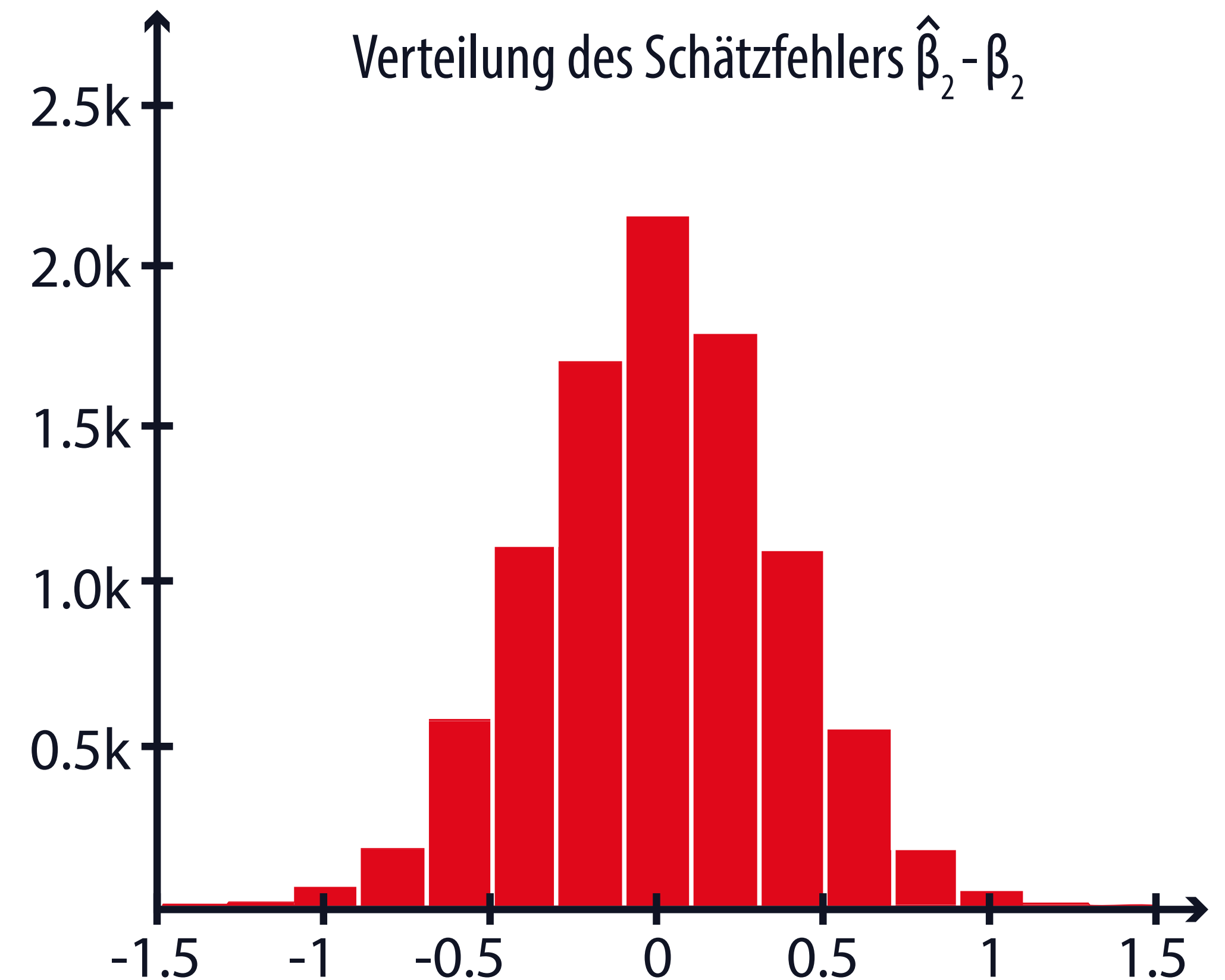
```

Multikollinearität

Jedes Mal wenn wir die Simulation wiederholen, erhalten wir komplett andere Koeffizienten. Wir haben einen sehr hohen zufälligen Schätzfehler.

Der Schätzer ist allerdings trotzdem erwartungstreu. Im Mittel finden wir die richtigen Koeffizienten.

Ursache ist Multikollinearität: Wir haben eine starke Korrelation zwischen mindestens einer Paarung von unabhängigen Variablen.

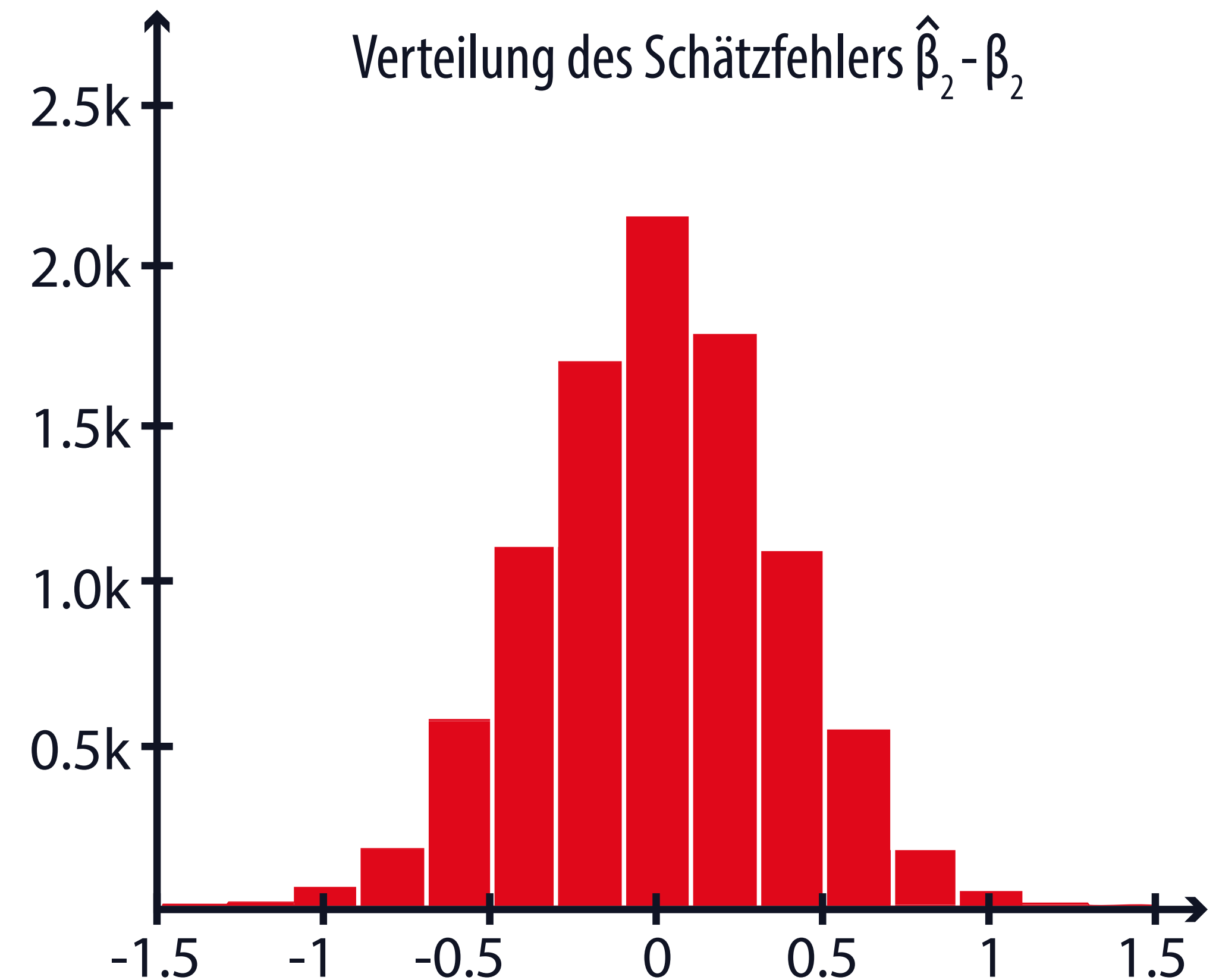


Multikollinearität feststellen

Wir können Multikollinearität auf drei Arten feststellen.

Am naheliegendsten ist der Korrelationskoeffizient zwischen den beiden unabhängigen Variablen MS und CS.

Bei einer größeren Zahl an unabhängigen Variablen können wir eine Korrelationsmatrix ausgeben lassen.



Multikollinearität feststellen

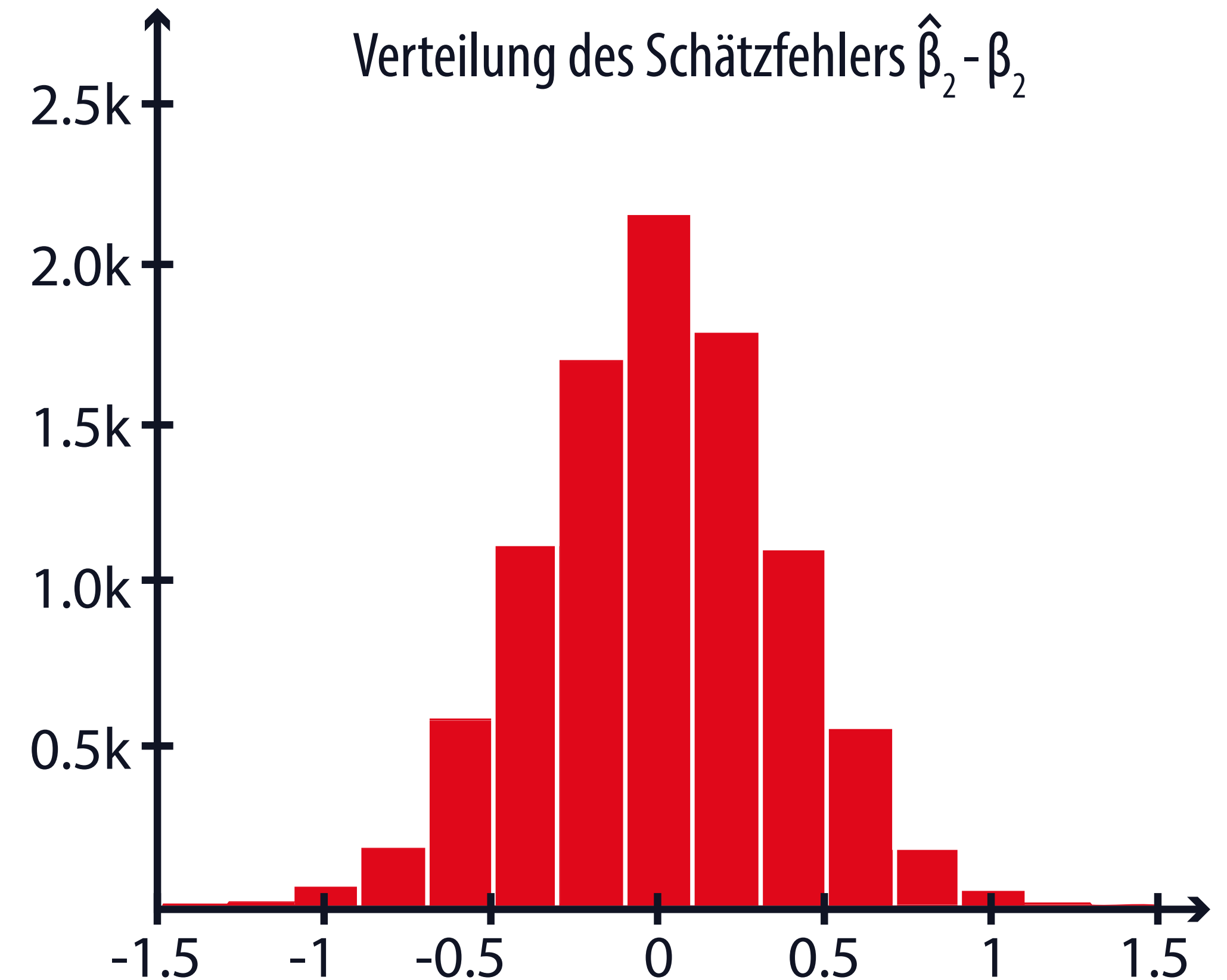
Möglichkeit zwei ist der Variance Inflation Factor VIF.

Wir gehen unsere unabhängigen Variablen durch und regressieren sie gegen alle anderen unabhängigen Variablen.

Dann setzen wir den korrigierten R²-Wert in folgende Formel ein. Ab einem Wert von 10 gehen wir von einer problematischen Multikollinearität aus.

$$VIF = \frac{1}{1 - R^2}$$

Kritischer Wert von: Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill Irwin.



Multikollinearität feststellen

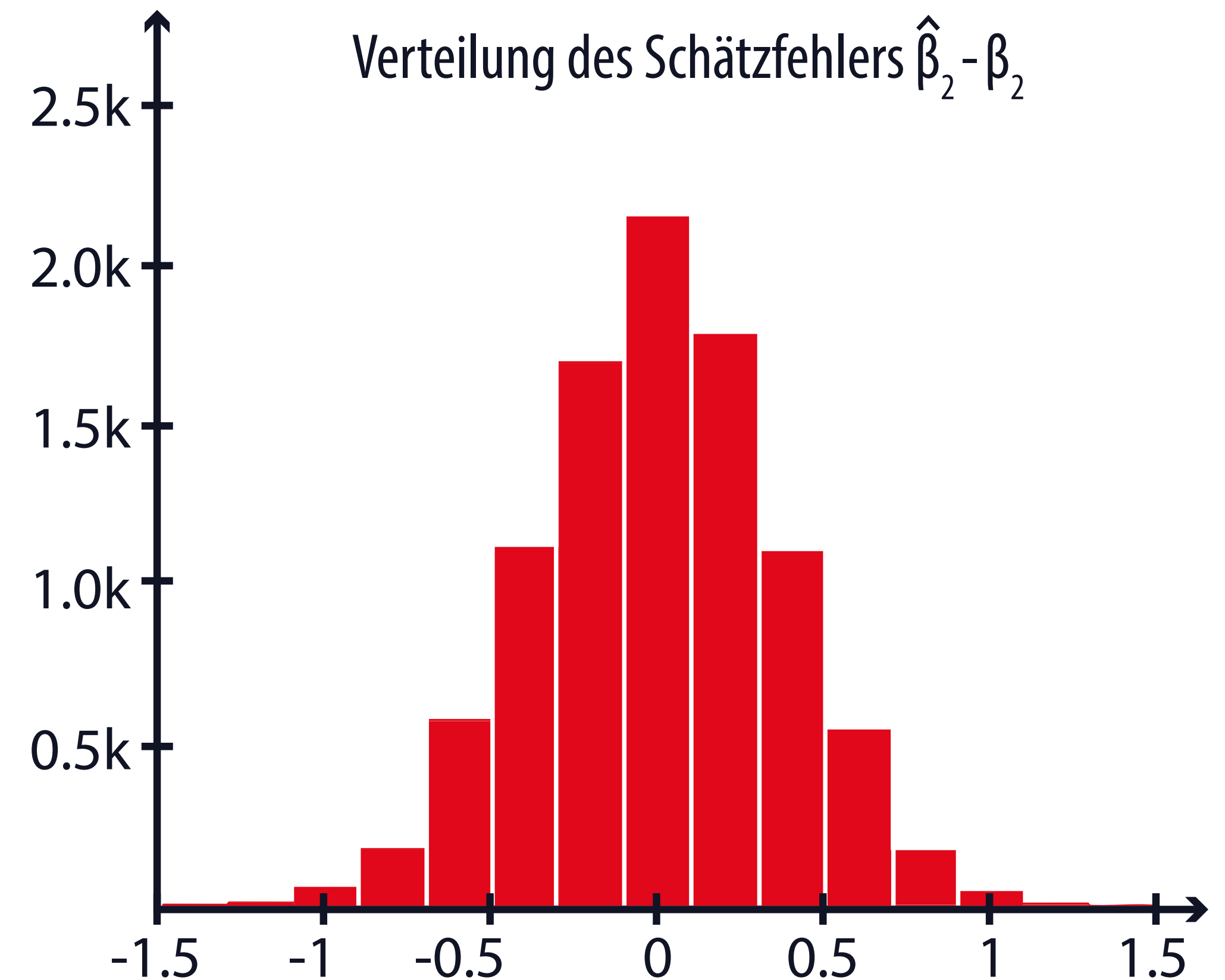
Bei Modellen mit mindestens drei unabhängigen Variablen können wir die Konditionszahl der Korrelationsmatrix berechnen.

$$\kappa = \max(\lambda) - \min(\lambda)$$

$$\det(M - \lambda I) = 0 \quad \text{für } M = \begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \dots \\ \rho_{2,1} & \rho_{2,2} & \\ \vdots & & \ddots \end{pmatrix}$$

Ab $\kappa = 30$ gehen wir von einer problematischen Multikollinearität aus.

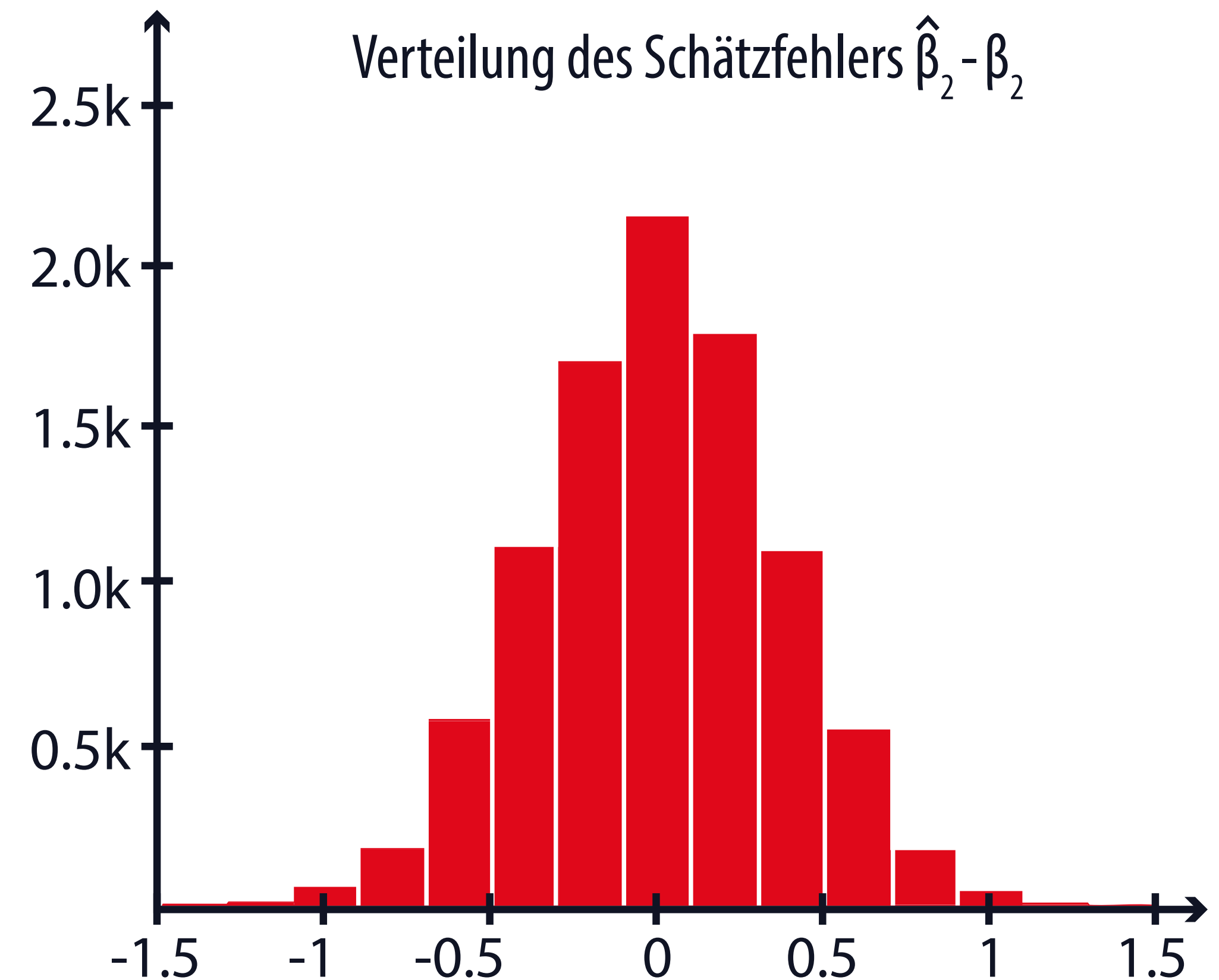
Kritischer Wert von: Belsley, Kuh, & Welsch (1980) – Regression Diagnostics: Identifying Influential Data and Sources of Collinearity



Multikollinearität behandeln

Nach der Diagnose kommt die Behandlung:

- Problematische Variablen entfernen.
- Problematische Variablen zusammenfassen.
- Problematische Variablen weiter zerlegen.
- Alternatives Schätzverfahren (z. B. Ridge-Regression)

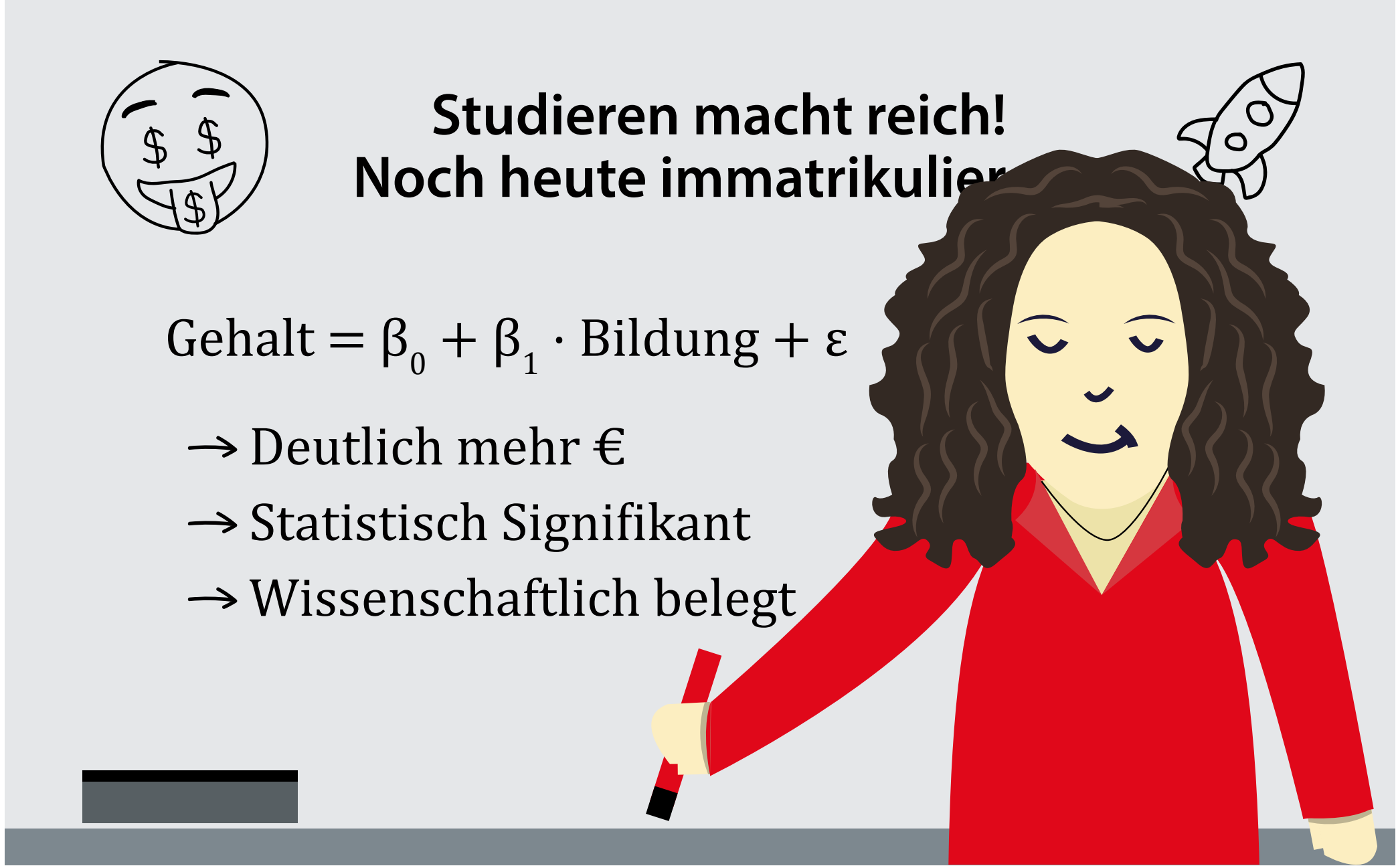


Endogenitätsprobleme

Die DHBW plant eine groß angelegte Studie zum Zusammenhang von Bildungsniveau und Einkommen.

Nach einer umfangreichen Recherche steht ein Fragebogen parat, mit dem das Bildungsniveau sehr detailliert gemessen werden kann.

Am Ende soll eine OLS-Regression zeigen, wie viel €/Jahr mehr ein Punkt auf dieser Bildungsskala bringt.



**Studieren macht reich!
Noch heute immatrikulieren**

$$\text{Gehalt} = \beta_0 + \beta_1 \cdot \text{Bildung} + \varepsilon$$

- Deutlich mehr €
- Statistisch Signifikant
- Wissenschaftlich belegt

Endogenitätsprobleme

Die Ergebnisse sind traumhaft. Der Koeffizient für die Bildung ist so signifikant, dass der p-Wert nicht mehr als double darstellbar ist und es geht um 575€ pro Skalenpunkt.

Rektorat und HoKo wollen das Ergebnis veröffentlichen, doch die Data Science ist skeptisch. Das sieht zu gut aus ...



```

Console Terminal Background Jobs
R 4.2.2
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17392.26    2089.50   8.324 8.27e-16 ***
education    575.57      20.54  28.015 < 2e-16 *****
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

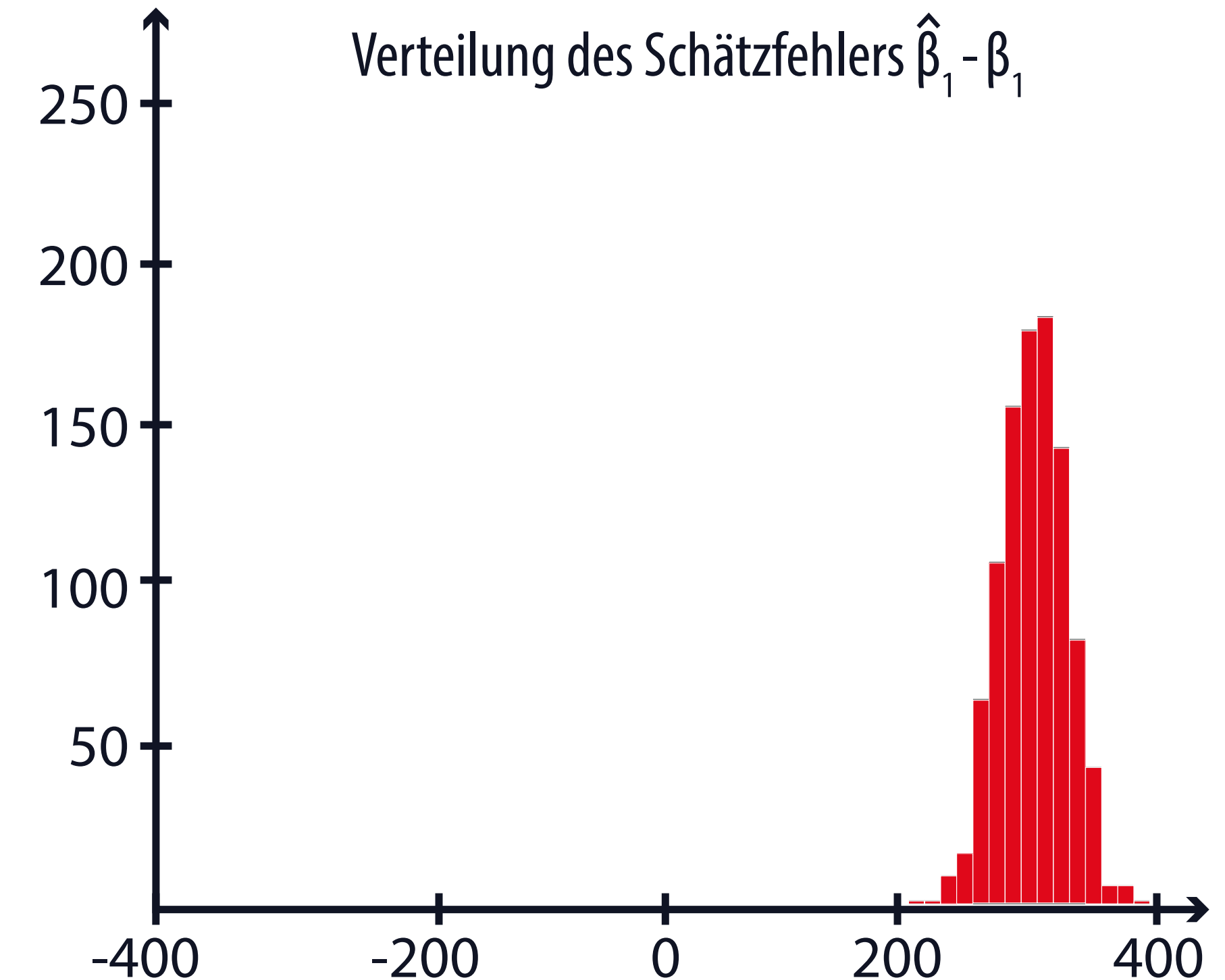
Residual standard error: 6863 on 498 degrees of freedom
Multiple R-squared:  0.6118,    Adjusted R-squared:  0.611
F-statistic: 784.8 on 1 and 498 DF,  p-value: < 2.2e-16
  
```

Endogenitätsprobleme

Wieder arbeiten wir mit simulierten Daten und kennen die wahren Koeffizienten.

Wir erkennen einen starken Bias. Der Schätzer ist nicht erwartungstreu, sondern zeigt uns viel zu hohe Werte an.

Im Gegensatz zum zufälligen Schätzfehler wird der Bias auch nicht mit größerer Stichprobe automatisch kleiner. Wir haben ein massives Problem - aber warum?



Endogenitätsprobleme

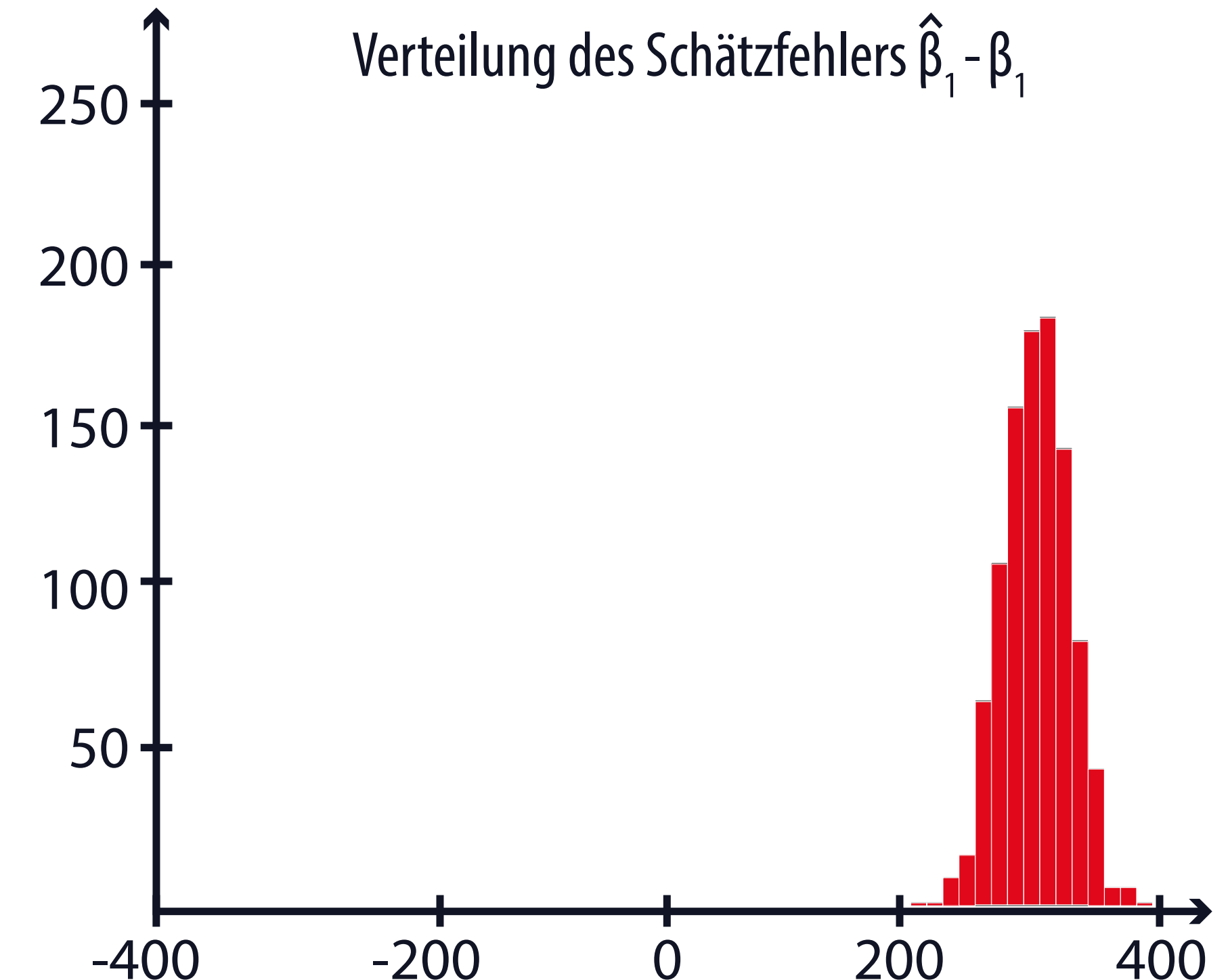
Endogenitätsprobleme entstehen, wenn eine unabhängige Variable vom Störterm statistisch abhängig ist.

$$p(x_{i,j} = x \mid \varepsilon_j) \neq p(x_{i,j} = x)$$

Meistens schreiben wir dieses Kriterium als:

$$\text{cor}(x_i, \varepsilon) \neq 0$$

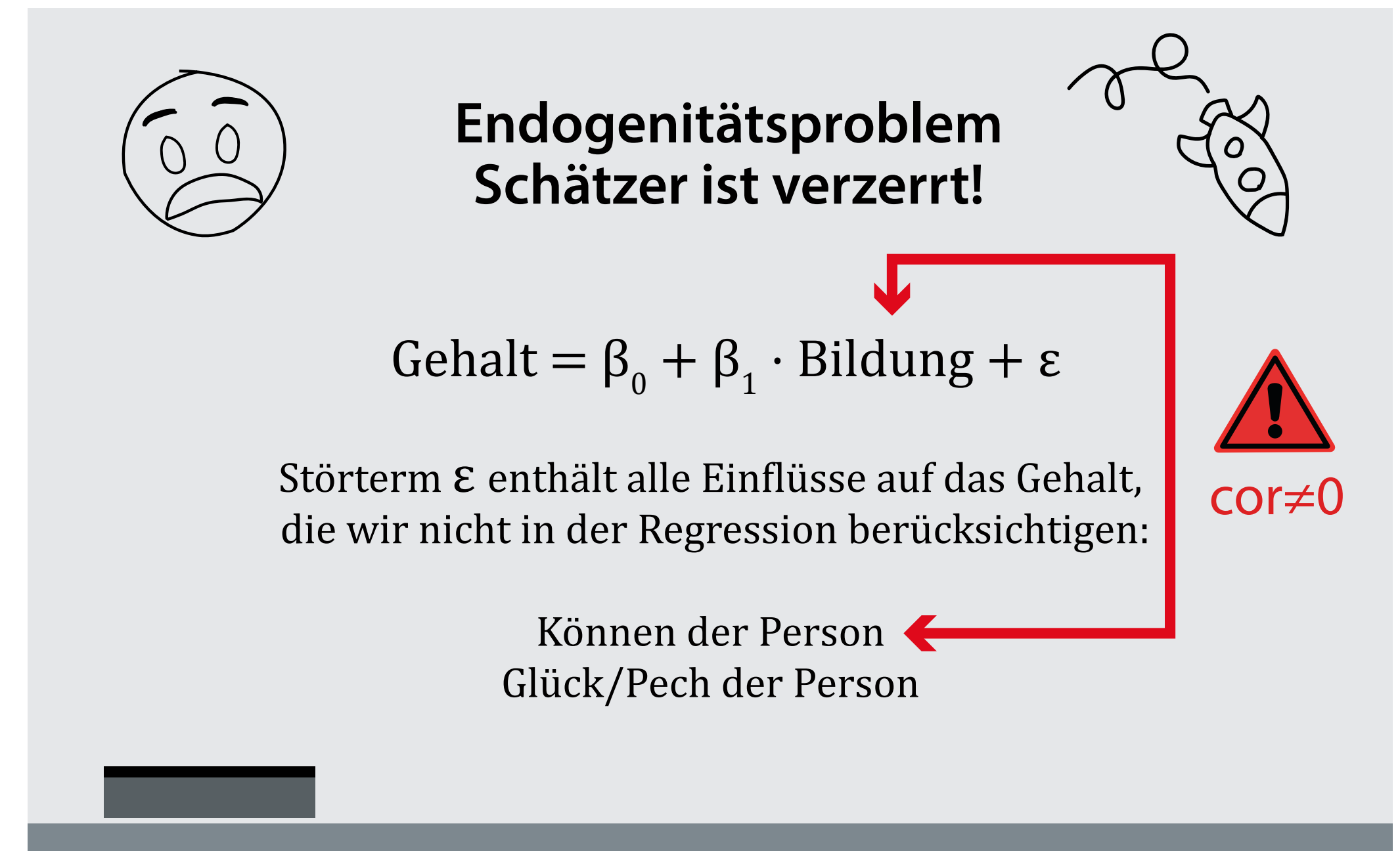
Ein Endogenitätsproblem verzerrt den Schätzer und führt dazu, dass er nicht mehr erwartungstreu ist.



Endogenitätsprobleme

In unserem Beispiel kennen wir den Störterm und können die Korrelation einfach berechnen.

Wir sehen außerdem, dass das Endogenitätsproblem durch den Einfluss der Variable "Ability" auf die Bildung entsteht.



Endogenitätsprobleme

In der Praxis kennen wir den Störterm nicht. In ihm ist alles enthalten, was sich auf die abhängige Variable auswirkt und nicht in die Regression aufgenommen wird.

Wir müssen überlegen, welche Variablen gleichzeitig mit der abhängigen Variable und mit unseren unabhängigen Variablen korrelieren könnten.

Was machen wir, wenn wir problematische Variablen gefunden haben?

WIR SUCHEN DRINGEND! PROBLEMATISCHE VARIABLEN

KORRELIERT MIT ABHÄNGIGER VARIABLE - Die Problemvariable muss mit dem Gehalt einer Person korreliert sein

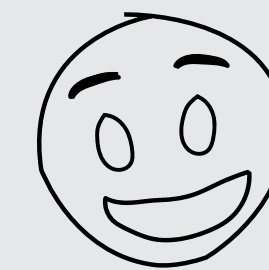
KORRELIERT MIT UNABHÄNGIGER VARIABLE - Die Problemvariable muss mit der Bildung einer Person korreliert sein



Endogenitätsprobleme

Im einfachsten Fall stehen uns Daten zu dieser Variable zur Verfügung. Dann nehmen wir sie in die Regression mit auf und entfernen sie dadurch aus dem Störterm.

Wenn wir die Daten nicht haben, hilft uns ggf. eine besondere Schätzmethode: die Instrumentalvariablenschätzung.



Die einfache Lösung Nur ein Omitted-Variable Problem



$$\text{Gehalt} = \beta_0 + \beta_1 \cdot \text{Bildung} + \beta_2 \cdot \text{Können} + \varepsilon$$

Störterm ε enthält alle Einflüsse auf das Gehalt,
die wir nicht in der Regression berücksichtigen:

~~Können der Person~~
Glück/Pech der Person



Endogenitätsprobleme

Voraussetzung dafür ist, dass wir ein geeignetes Instrument finden, von dem wir Daten haben. Ein Instrument muss ...

... **relevant** sein, d. h., es muss mit der unabhängigen Variable korreliert sein, die das Endogenitätsproblem verursacht.

... **exogen** sein, d. h., es darf nicht mit dem Störterm korreliert sein.

WIR SUCHEN DRINGEND! INSTRUMENT

RELEVANT - Das Instrument muss mit der Bildung unserer Probanden korrelieren.

EXOGEN - Das Instrument darf nicht mit dem Störterm korreliert sein. Insbesondere nicht mit der Fähigkeit einer Person!

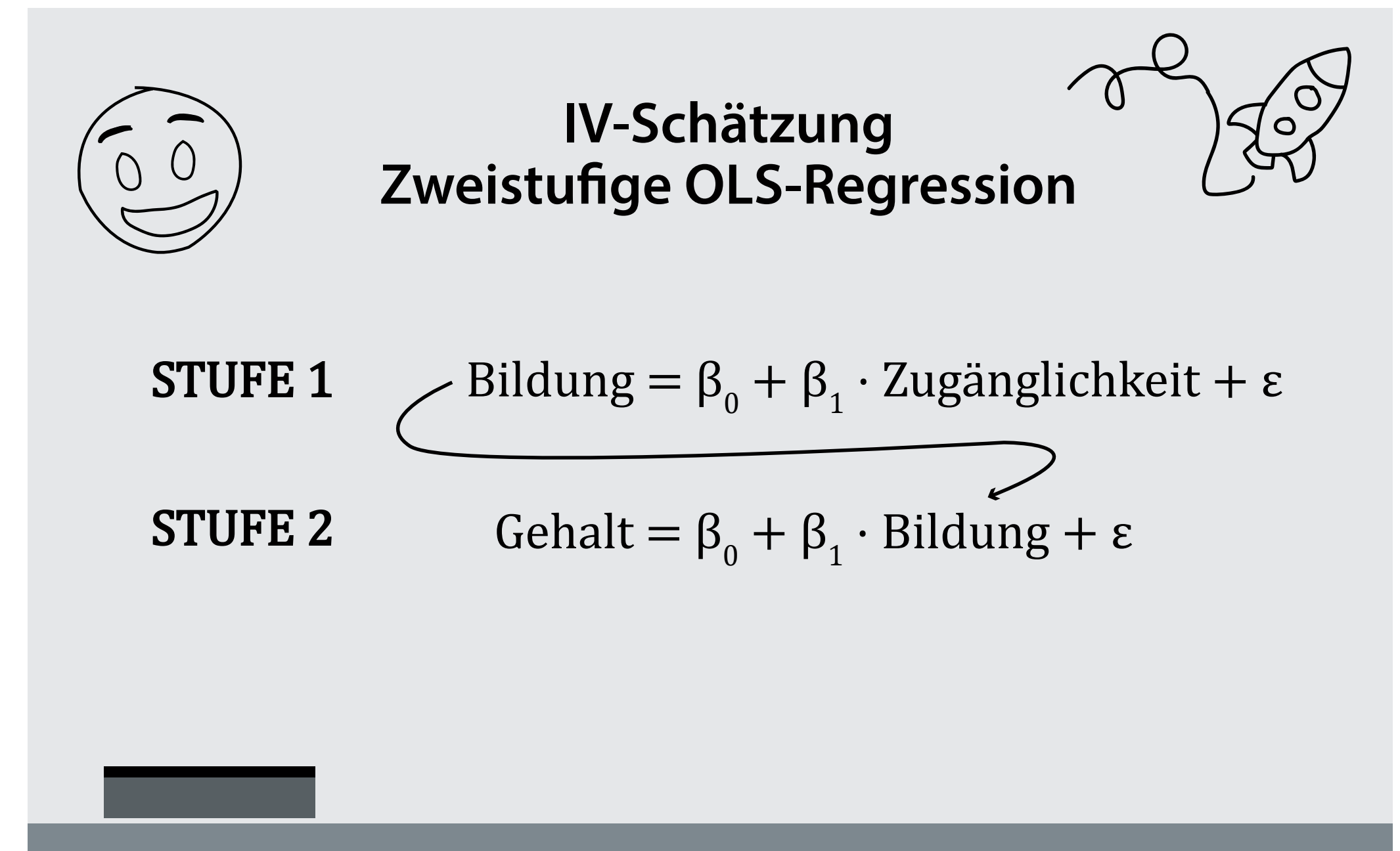


Endogenitätsprobleme

Haben wir ein Instrument gefunden, führen wir eine zweistufige Regression durch.

Stufe 1 Wir machen die unabhängige Variable, die das Endogenitätsproblem verursacht zur abhängigen Variable einer separaten Regression.

Als unabhängige Variable wählen wir unser Instrument.

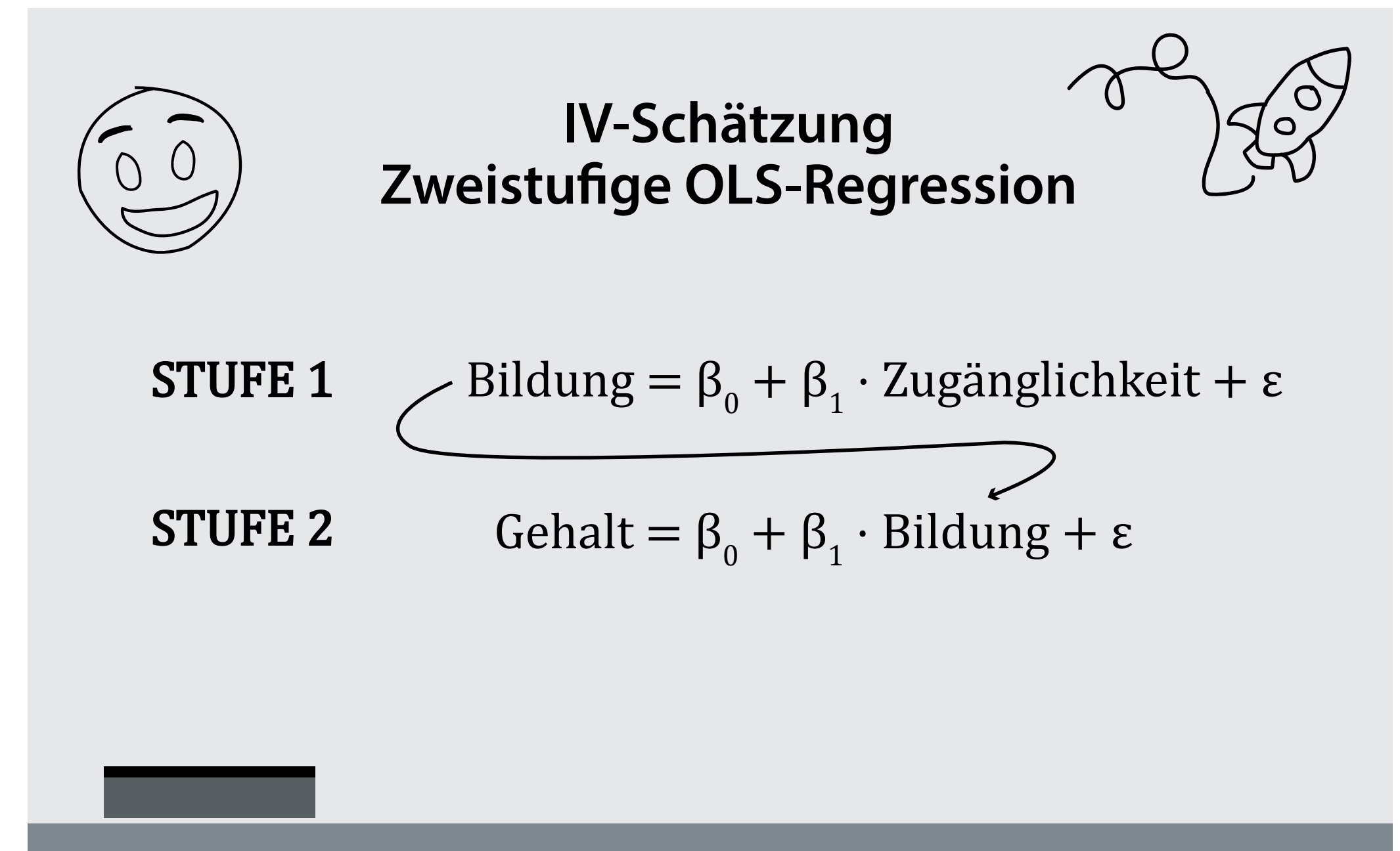


Endogenitätsprobleme

Haben wir ein Instrument gefunden, führen wir eine zweistufige Regression durch.

Stufe 2 Wir führen die ursprüngliche geplante OLS-Regression durch und ersetzen die Werte der problematischen unabhängigen Variable ...

... durch die gefitteten Werte aus der ersten Stufe.

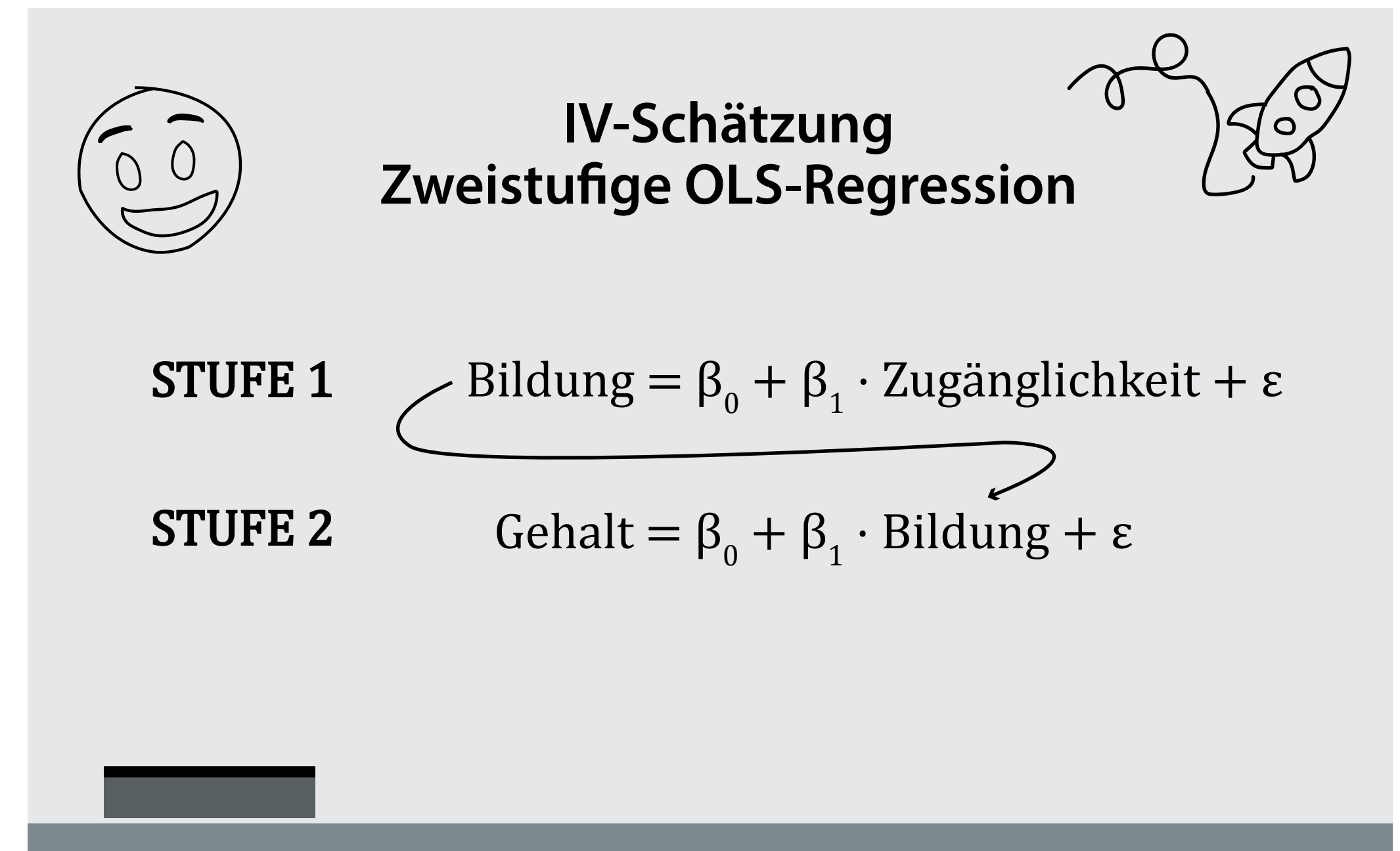


Endogenitätsprobleme

Ein weiterer Vorteil der IV-Schätzung ist, dass wir die Ergebnisse nun auch kausal bewerten dürfen.

Mehr Bildung führt zu einem höheren Gehalt!

Die kausale Interpretation gilt allerdings nur für die instrumentierte Variable und auch nur dann, wenn die Instrumente wirklich relevant und exogen sind.



Overfitting

Die vielen Möglichkeiten der Regressionsanalyse bergen die Gefahr des Overfittings und das sowohl durch den Algorithmus, als auch durch uns als Data Scientists!

Bei ML-Algorithmen tunen wir Parameter, bis wir gute OOS-Werte für Accuracy, Recall, Precision und Co. haben.

Bei der Regression ändern wir Wahl und Transformation von unabhängigen Variablen, bis die "richtigen" Variablen signifikant werden.



Overfitting

Lösung: Wir überlassen die Wahl der unabhängigen Variablen dem Algorithmus oder einem Regelwerk auf die wir uns vor der Datenanalyse festlegen.

Wir probieren es mit der algorithmischen Lösung! Wie bekommen wir den Algorithmus dazu eine "ausgewogene" Wahl der unabhängigen Variablen zu treffen?

Wie verhindern wir, dass der Algorithmus noch mehr overfitted als wir es von Hand tun würden?

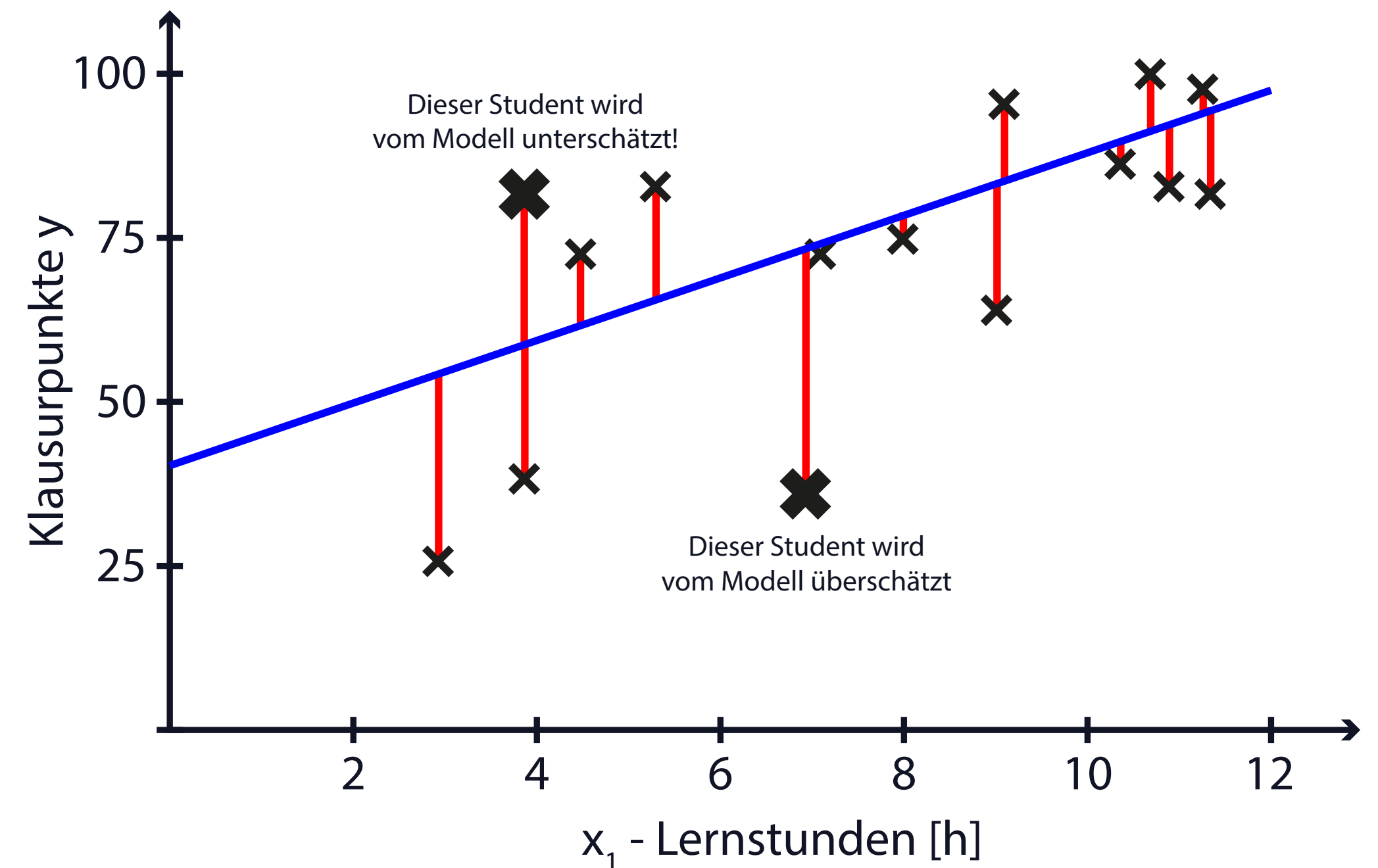


Overfitting

Bisher sucht der Algorithmus nach Schätzern die den MSE, d. h. die Summe der quadrierten Residuen, minimiert.

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Diese Zielfunktion ergänzen wir nun um eine Strafe, die Overfitting verhindern soll.



Overfitting

Grundidee: Die Strafe soll von der Größe der Schätzer und einem Parameter λ abhängen.

Mit λ regeln wir die Strenge der Strafe. Je größer λ , umso stärker wird der Algorithmus für Overfitting bestraft.

Warum wird folgende naheliegende Straffunktion in der Praxis nicht funktionieren?

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^m \hat{\beta}_k$$



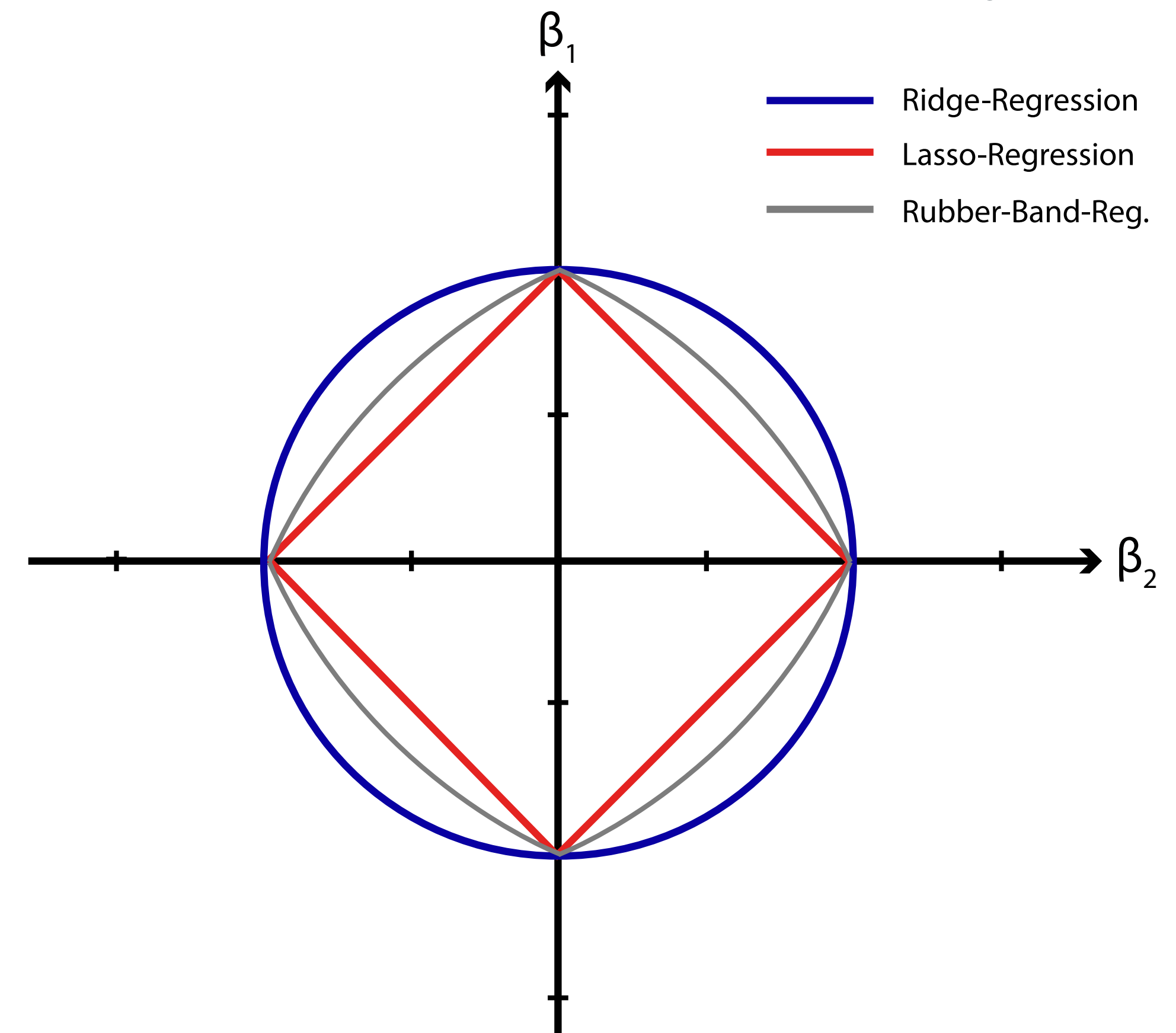
Overfitting

Ansatz der **Ridge-Regression** mit quadrierten Schätzern entspricht der euklidischen Norm (L2).

$$\min \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^m \hat{\beta}_k^2$$

Ansatz der **Lasso-Regression** mit Betrag der Schätzer entspricht der "Taxifahrernorm" (L1).

$$\min \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^m |\beta_k|$$

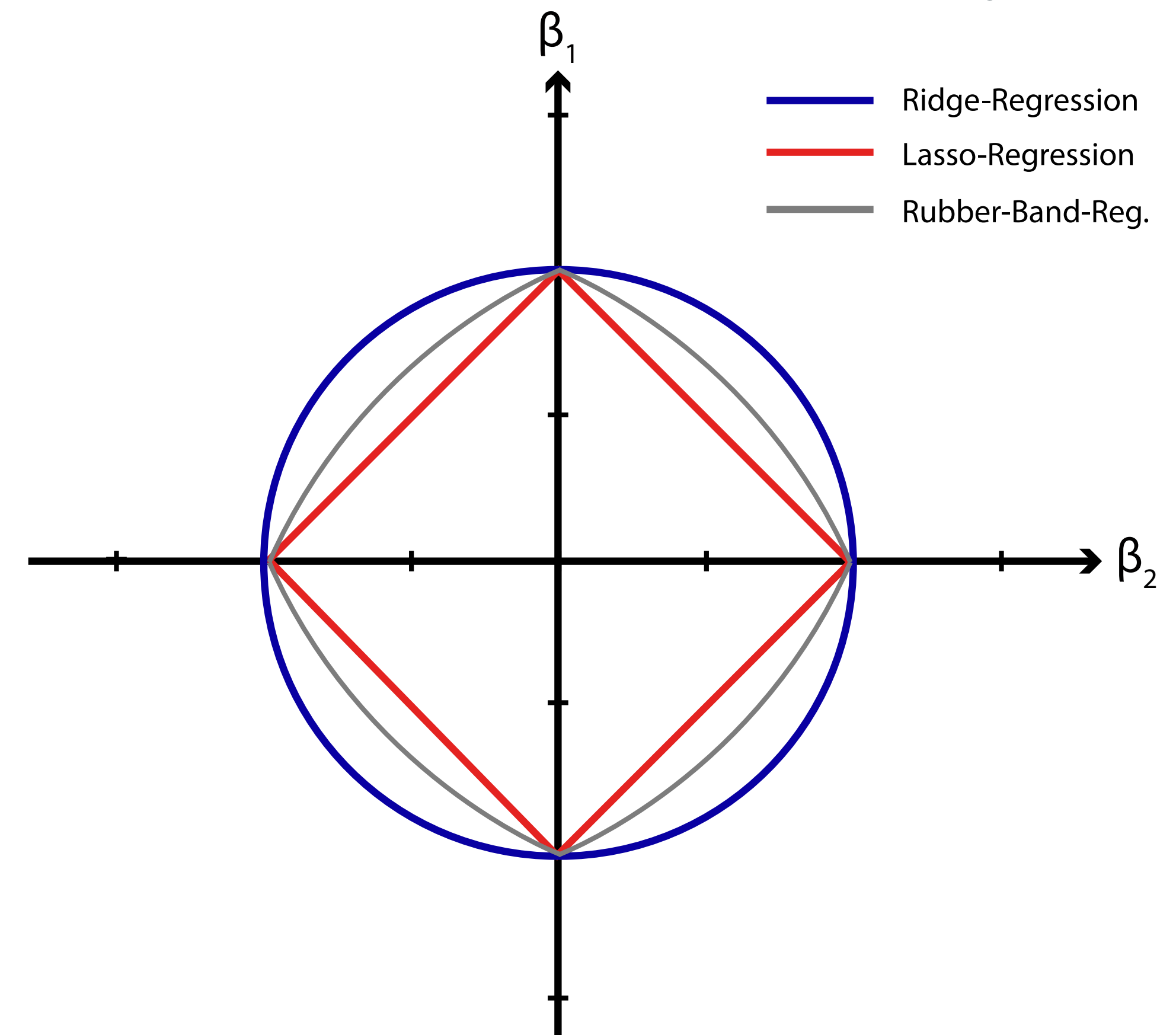


Overfitting

Der vermeintlich kleine mathematische Unterschied hat eine wesentliche Konsequenz:

Bei der **Ridge-Regression** werden unabhängige Variablen selten komplett aus dem Modell entfernt. Bei Schätzern kleiner 1 verkleinert die Quadrierung die Strafe.

Bei der **Lasso-Regression** fehlt dieser Effekt und außerdem ist die Betragsfunktion bei 0 nicht differenzierbar. Hier werden unabhängige Variablen ggf. auch aussortiert.

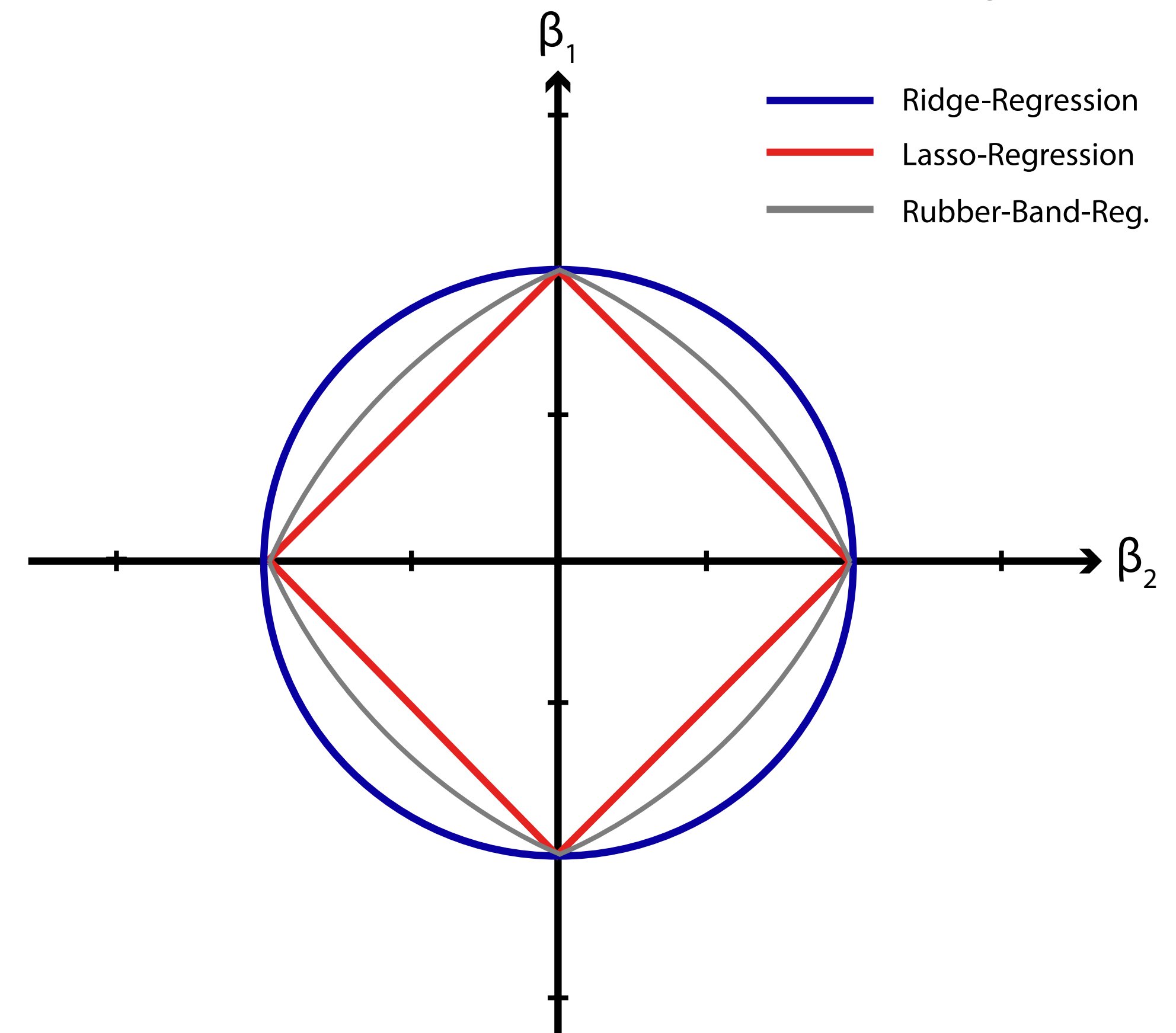


Overfitting

Der vermeintlich kleine mathematische Unterschied hat eine wesentliche Konsequenz:

Die **Ridge-Regression** eignet sich, wenn wir alle unabhängigen Variablen berücksichtigt haben möchten, aber Overfitting verhindern oder Multikollinearität behandeln wollen.

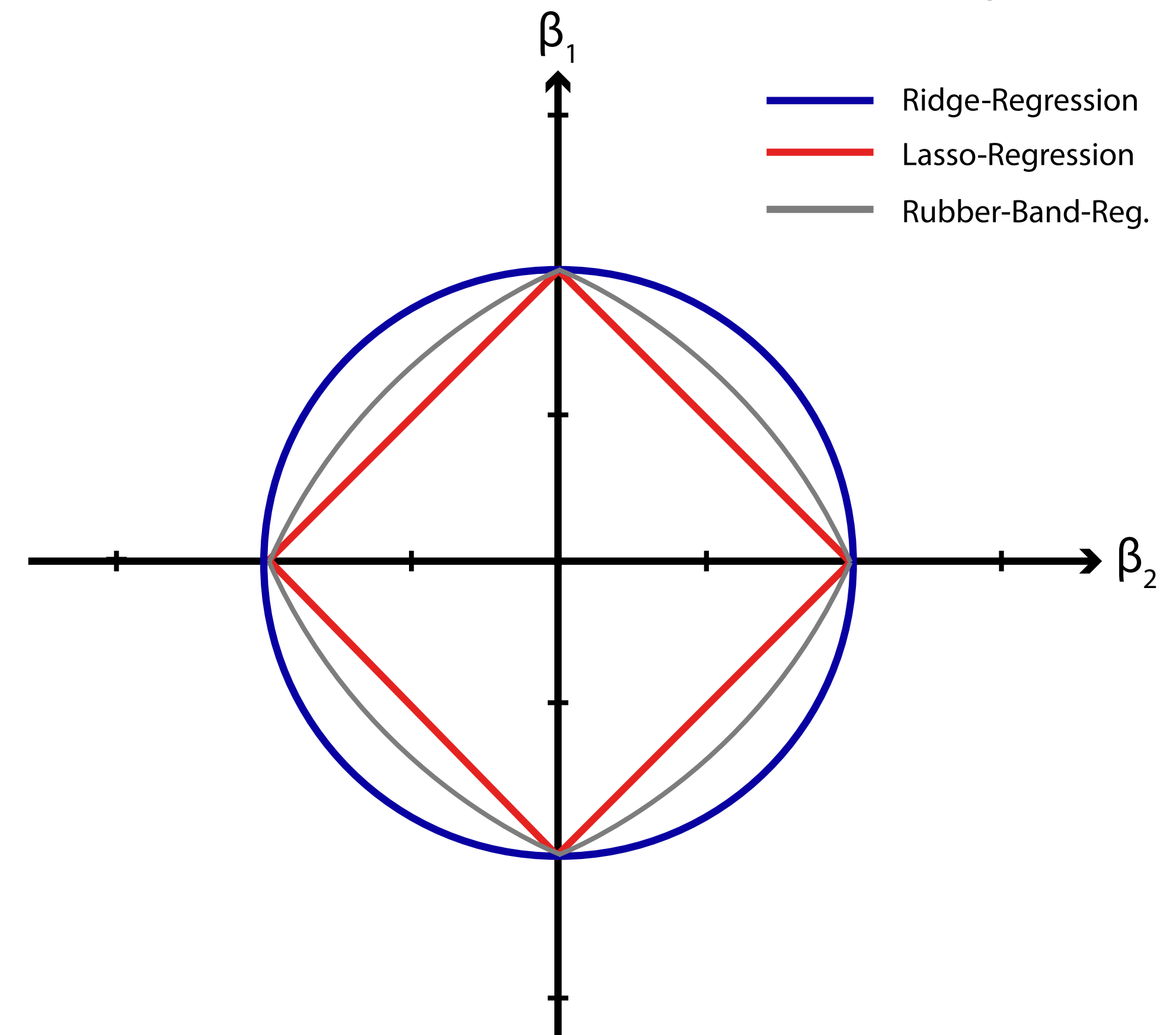
Die **Lasso-Regression** eignet sich, wenn wir herausfinden wollen, welche unabhängigen Variablen wirklich relevant sind.



Overfitting

Es ist möglich, beide Ansätze zu kombinieren. Die **Elastic-Net-Regression** verwendet folgende Straffunktion:

$$\min \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{k=1}^m \hat{\beta}_k^2 + \lambda_2 \sum_{k=1}^m |\beta_k|$$



Overfitting

Woher kommt das Lambda? Wir können es fest vorgeben oder wir überlassen die Wahl dem Algorithmus.

Damit dies nicht zu weiterem Overfitting führt, wird das optimale Lambda mit k-fold-Corssvalidation bestimmt.



Overfitting

Alle diese drei Varianten haben einen Nachteil. Die Schätzer sind nicht mehr erwartungstreu!

Wir nehmen einen höheren Bias in Kauf um Overfitting und damit hohe Varianz und Instabilität zu senken.

Dieser Bias-Variance-Tradeoff gibt es nicht nur in der Regressionsanalyse. Ihr kennt ihn aus dem Machine Learning!

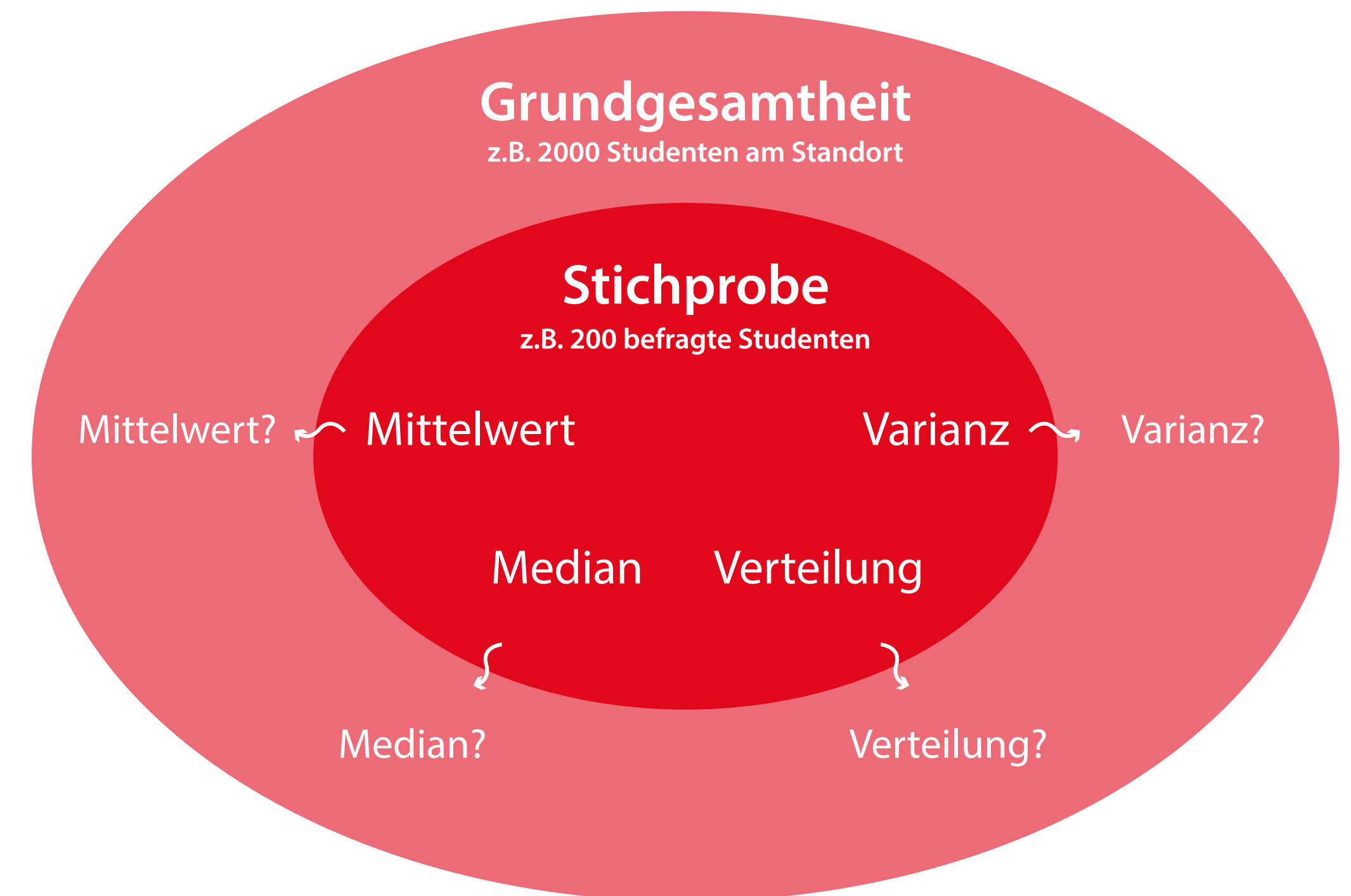


ML-Schätzung

Bei der ML-Schätzung kommen wir zurück zum Grundproblem der statistischen Tests. Wir kennen Eigenschaften unserer Stichprobe, aber nicht die der Grundgesamtheit.

Bei der ML-Schätzung wollen wir von der Stichprobe auf die Verteilung der Grundgesamtheit schließen.

Wir nehmen dabei an, dass die Stichprobe $X_1, \dots, X_n$ aus der gleichen Verteilung gezogen wurde und keine Autokorrelation aufweist, d. h. x_i und x_j sind stochastisch unabhängig für alle $i \neq j$.

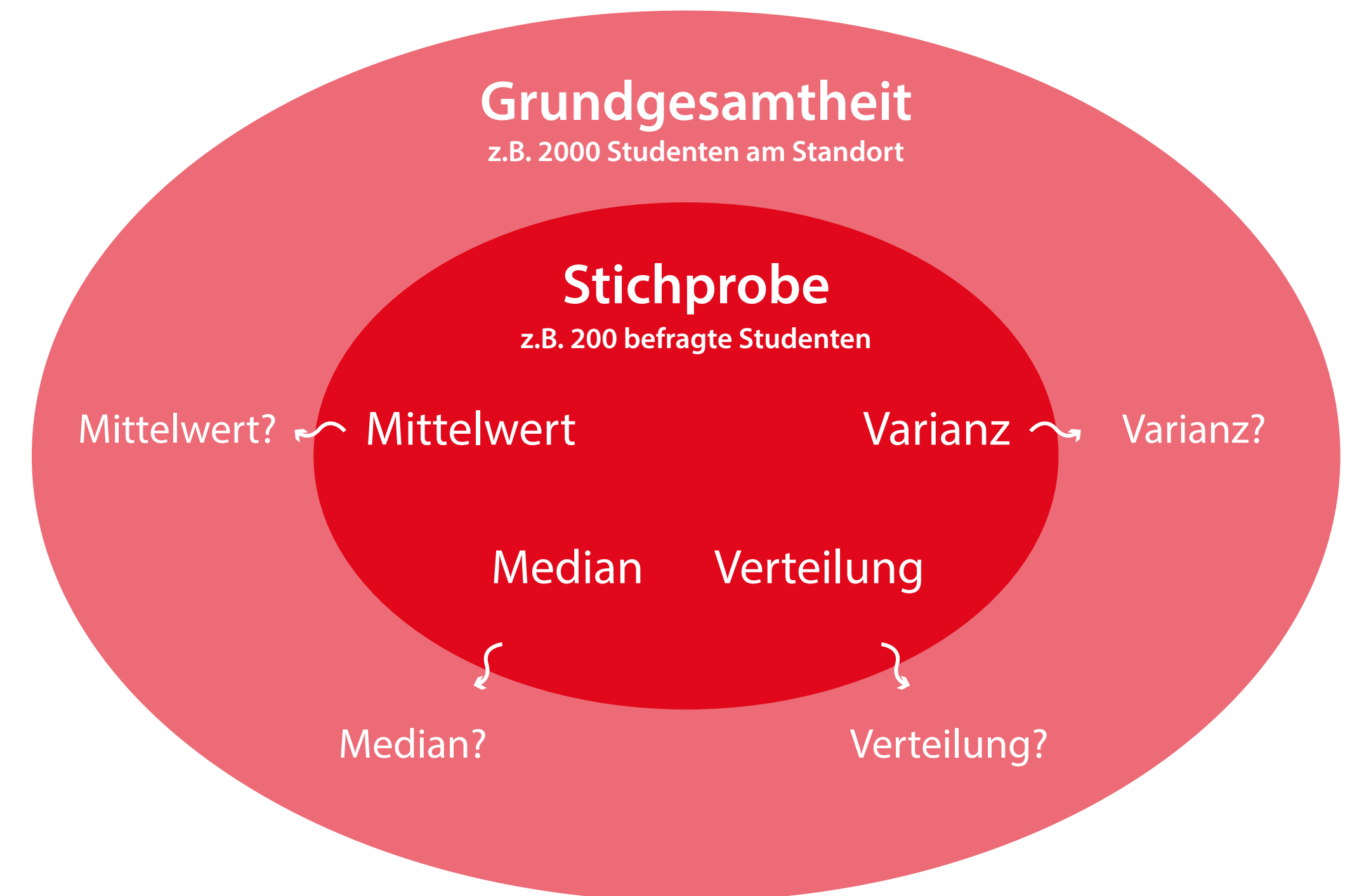


ML-Schätzung

Die Verteilung der Grundgesamtheit geben wir vor. Wir nehmen an, dass die Grundgesamtheit einer bestimmten Verteilung mit den Parametern $\theta_1, \dots, \theta_n$ folgt.

Jetzt suchen wir nach Werten $\varphi_1, \dots, \varphi_n$ für diese Parameter, für die die Wahrscheinlichkeit unsere Stichprobe zu erhalten maximal wird.

$$\max p(x_1, \dots, x_n | \theta_1, \dots, \theta_n)$$



ML-Schätzung

Das die Beobachtungen der Stichprobe als i.i.d. angenommen werden ist:

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)$$

Die Wahrscheinlichkeit, einen bestimmten Wert exakt zu treffen, ist bei kontinuierlichen Verteilungen allerdings unendlich klein. Insbesondere gilt:

$$P(a \leq X \leq b) = \int_a^b \varphi(x) dx$$



ML-Schätzung

Für die Bewertung der Wahrscheinlichkeit, eine bestimmte Stichprobe zu beobachten, wird daher die Likelihood Funktion L eingeführt:

$$L(\theta) = \prod_{i=1}^n \varphi_{\theta}(x_i)$$

Die Log-Likelihood Funktion ist der Logarithmus dieser Funktion. Aus dem Produkt wird eine Summe:

$$\ell(\theta) = \sum_{i=1}^n \ln \varphi_{\theta}(x_i)$$

Erinnerung

$$\ln(a \cdot b) = \ln(a) + \ln(b)$$

$$\begin{aligned} \ln\left(\prod x_i\right) &= \ln(x_1) + \dots \ln(x_n) \\ &= \sum \ln(x_i) \end{aligned}$$

ML-Schätzung

Die ML-Schätzer für die Parameter θ erhalten wir durch Minimieren der negativen Log-Likelihoodfunktion:

$$\min -\ell(\theta)$$

Warum führen wir das Minuszeichen ein? Durch das Minus wird die Log-Likelihoodfunktion zu einer Loss-Funktion, ähnlich wie der (M)SE bei der OLS Regression.

Beispiel: NV

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

ML-Schätzung

Die ML-Schätzer für die Parameter θ erhalten wir durch Minimieren der negativen Log-Likelihoodfunktion:

$$\min -\ell(\theta)$$

Vorgehen aus der mehrdimensionalen Analysis:

Notwendig Alle partiellen Ableitungen 0 setzen.

Hinreichend Hessematrix auf Definitheit untersuchen.

Beispiel: NV

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

$$\begin{aligned}
 L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \\
 \Rightarrow \ell(\mu, \sigma^2) &= \sum_{i=1}^n \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \right] = \sum_{i=1}^n \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \ln \left[e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \right] \\
 &= n \cdot \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] - \sum_{i=1}^n \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \\
 &= n \cdot \ln \left[1 \right] - n \cdot \ln \left[\sqrt{2\pi\sigma^2} \right] - \sum_{i=1}^n \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2
 \end{aligned}$$

Von x_i unabhängig
Konstante mal n

$\ln(ab) = \ln(a) + \ln(b)$

$\ln(a/b) = \ln(a) - \ln(b)$

$\ln(1) = 0$

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^n \frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} \stackrel{!}{=} 0 \Leftrightarrow \sum_{i=1}^n x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^4} \stackrel{!}{=} 0 \Leftrightarrow \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^4} = \frac{n}{2\sigma^2}$$

$$\Leftrightarrow \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = n \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

ML-Schätzung interpretieren

In R oder Python erhalten wir beim Fitten mit Maximum Likelihood Schätzung neben dem Schätzer noch zusätzliche Informationen.

Der Standardfehler gibt an, wie sicher wir uns gegeben der Daten mit unserem Schätzer sein können.

Je größer der Standardfehler, umso mehr ist der Schätzer mit Unsicherheit behaftet.

```
Console Terminal Background Jobs
R 4.2.2
> library(fitdistrplus)
> k <- c(1,2,1,0,4,1,3,2,2,5,2,1,2,3,2,2,1,2,4,4,2)
> summary(fitdist(k,"binom",fix.arg=list(size=5)))

Fitting of the distribution ' binom ' by maximum likelihood
Parameters : prob = 0.4381 (std. err. = 0.0484)
Fixed parameters: size = 5
Loglikelihood: -33.76968 AIC: 69.53937 BIC: 70.58389
```

ML-Schätzung interpretieren

Die Loglikelihood ist der Wert, den $\ell(\theta)$ am gefundenen Maximum annimmt.

Je näher dieser Wert an der 0 ist, umso besser ist das Modell.

Es gibt allerdings keinen Maßstab i.s.V. alles über -100 ist gut. Der Wert eignet sich also eher zum Modellvergleich, als zur direkten Begutachtung eines einzelnen Modells.

```
Console Terminal Background Jobs
R 4.2.2
> library(fitdistrplus)
> k <- c(1,2,1,0,4,1,3,2,2,5,2,1,2,3,2,2,1,2,4,4,2)
> summary(fitdist(k,"binom",fix.arg=list(size=5)))

Fitting of the distribution ' binom ' by maximum likelihood
Parameters : prob = 0.4381 (std. err. = 0.0484)
Fixed parameters: size = 5
Loglikelihood: -33.76968 AIC: 69.53937 BIC: 70.58389
```

ML-Schätzung interpretieren

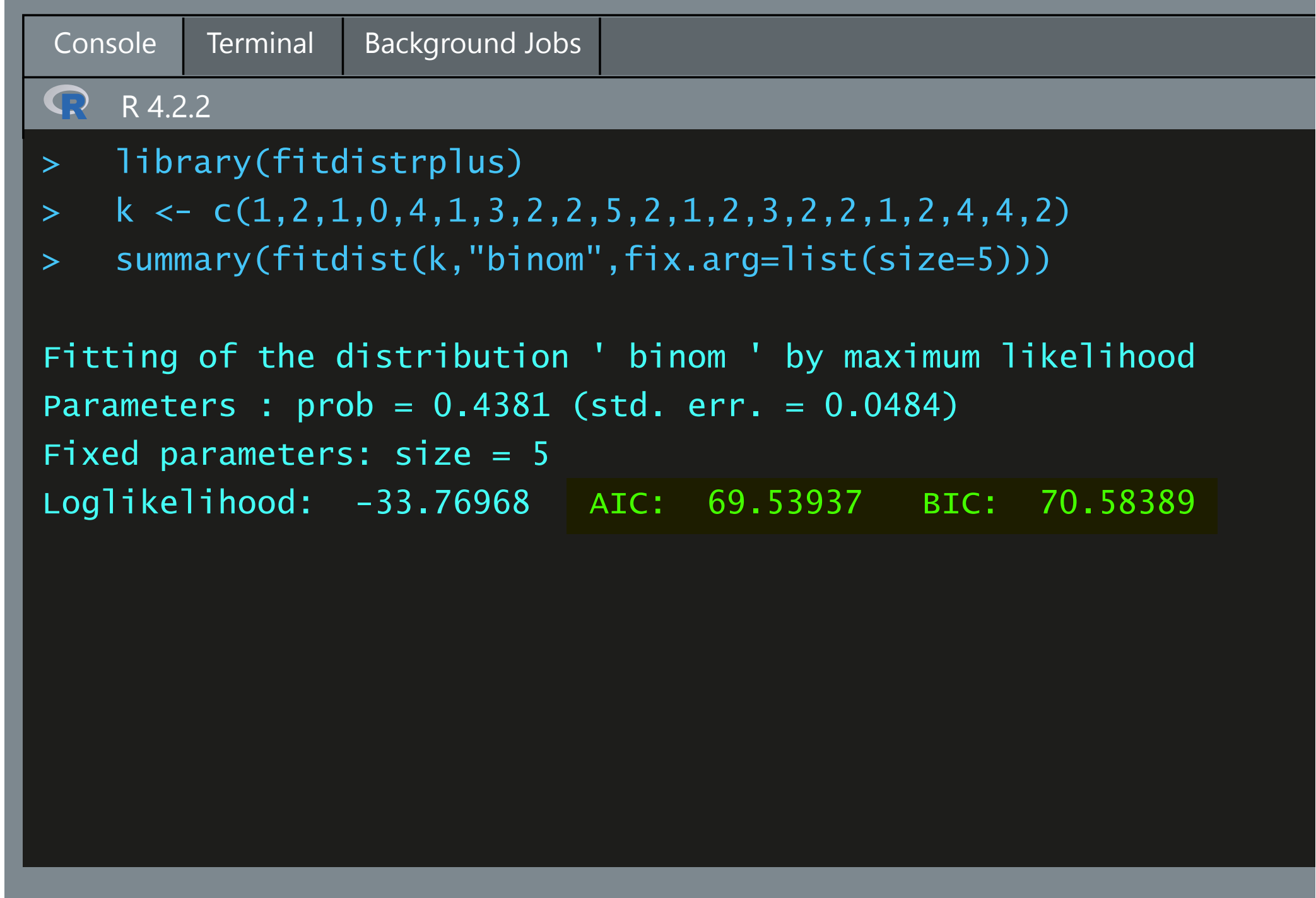
Das Akaike Information Criterion (AIC) und das Bayesian Information Criterion (BIC) ergänzen die Bewertung um Informationen zum Modell.

AIC berücksichtigt die Parameterzahl k :

$$AIC = -2\ell + 2k$$

BIC berücksichtigt auch die Stichprobengröße:

$$BIC = -2\ell + k \cdot \ln(n)$$



```
Console Terminal Background Jobs
R 4.2.2
> library(fitdistrplus)
> k <- c(1,2,1,0,4,1,3,2,2,5,2,1,2,3,2,2,1,2,4,4,2)
> summary(fitdist(k,"binom",fix.arg=list(size=5)))

Fitting of the distribution ' binom ' by maximum likelihood
Parameters : prob = 0.4381 (std. err. = 0.0484)
Fixed parameters: size = 5
Loglikelihood: -33.76968 AIC: 69.53937 BIC: 70.58389
```