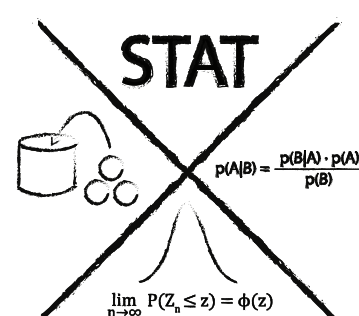


Grundlagen Statistik



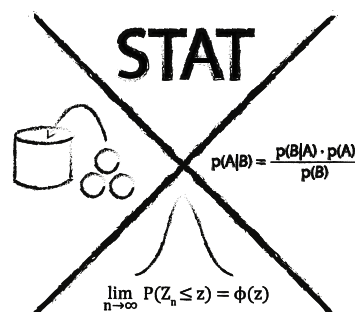
Vorwort

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 2025
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](#).

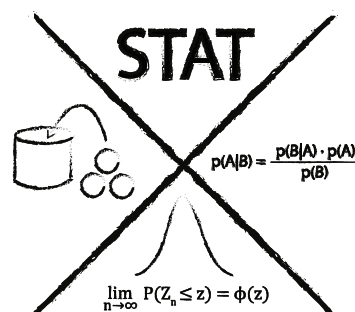
Fehler gefunden? E-Mail an blochinger@dhbw-ravensburg.de!



Inhaltsverzeichnis

<u>Einführung & Begrifflichkeiten</u>	<u>5 - 31</u>
<u>Lageparameter</u>	<u>32 - 54</u>
<u>Streuungsparameter</u>	<u>55 - 68</u>
<u>Kovarianz & Korrelation</u>	<u>69 - 88</u>
<u>Visualisierung</u>	<u>89 - 129</u>
<u>Wahrscheinlichkeiten</u>	<u>130 - 151</u>
<u>Permutationen</u>	<u>152 - 173</u>

<u>Urnenmodelle</u>	<u>174 - 206</u>
<u>Bedingte Wahrscheinlichkeiten</u>	<u>207 - 222</u>
<u>Zufallsvariablen & Verteilungen</u>	<u>223 - 268</u>
<u>Z-Test</u>	<u>269 - 282</u>
<u>T-Test</u>	<u>283 - 312</u>
<u>Chi-Quadrat Test</u>	<u>313 - 324</u>
<u>Lineare Regression</u>	<u>325 - 355</u>



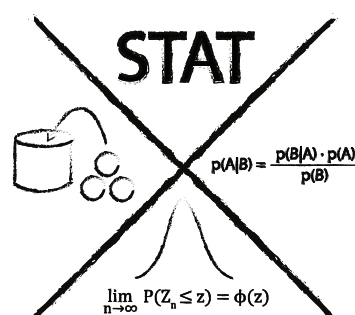
Software

In der Praxis führen wir statistische Berechnungen und Tests nicht von Hand, sondern mit Statistiksoftware (R, SPSS, Stata) oder Tabellenkalkulationssoftware (Excel) durch.

In dieser Veranstaltung schauen wir zunächst hinter die Kulissen, um die Funktionen und Ausgaben von Statistiksoftware später besser verstehen zu können.

Danach werden wir Excel für die Berechnungen verwenden!

Zum Erlernen von R, Python & SPSS gibt es Kurse von ZDI/ZEK.



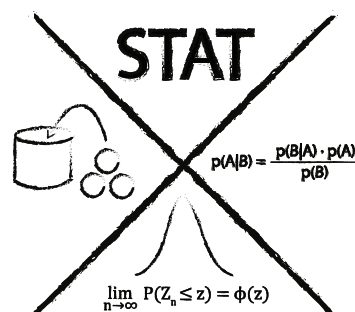
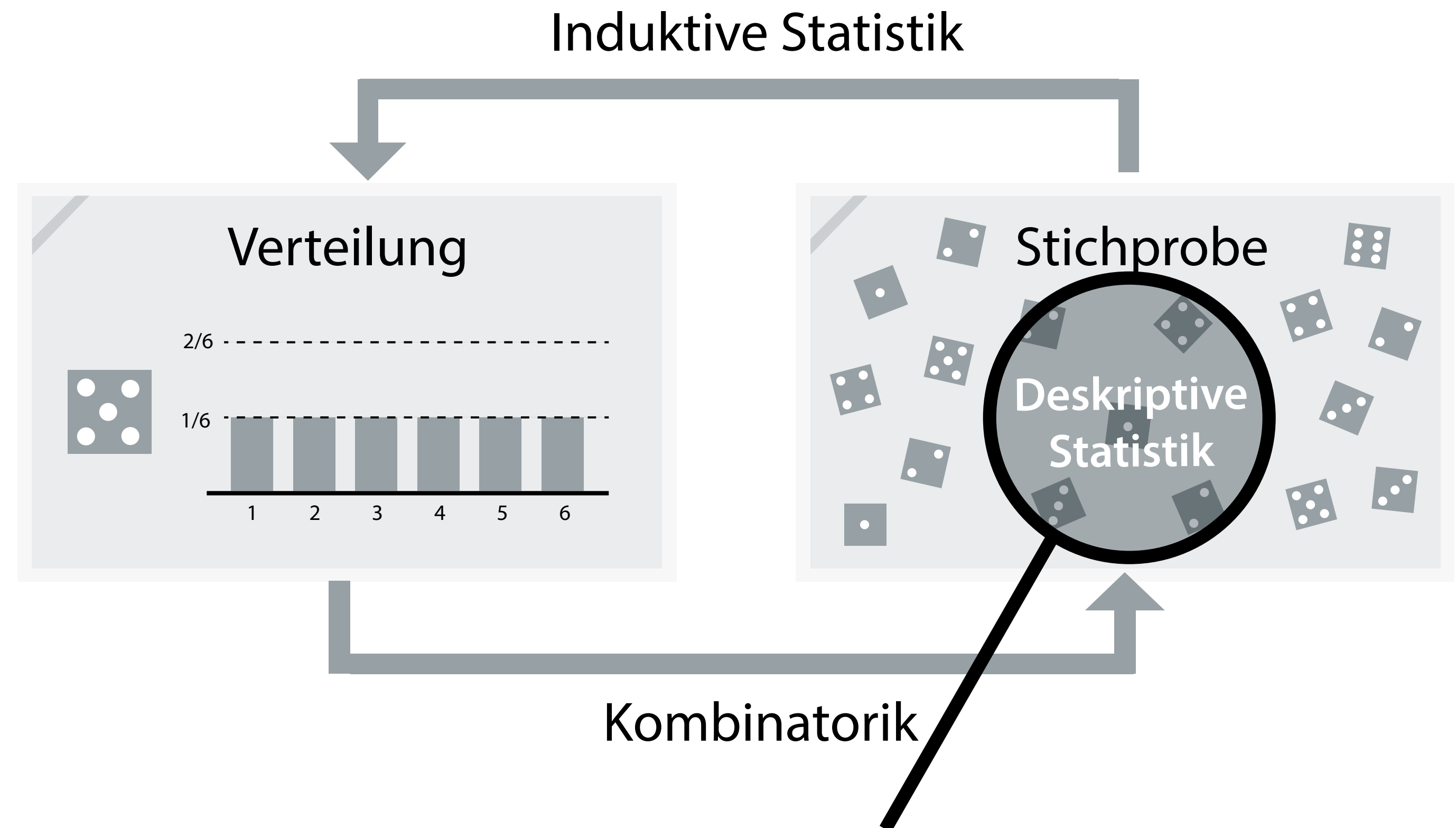
Teilgebiete der Statistik

Was ist Statistik? Wir weichen der Frage ein Stück weit aus, denn ...

...Statistik besteht aus mehreren Teilgebieten, die sich stark unterscheiden.

...allgemeine Definitionen von Statistik sind daher oft nicht sonderlich hilfreich.

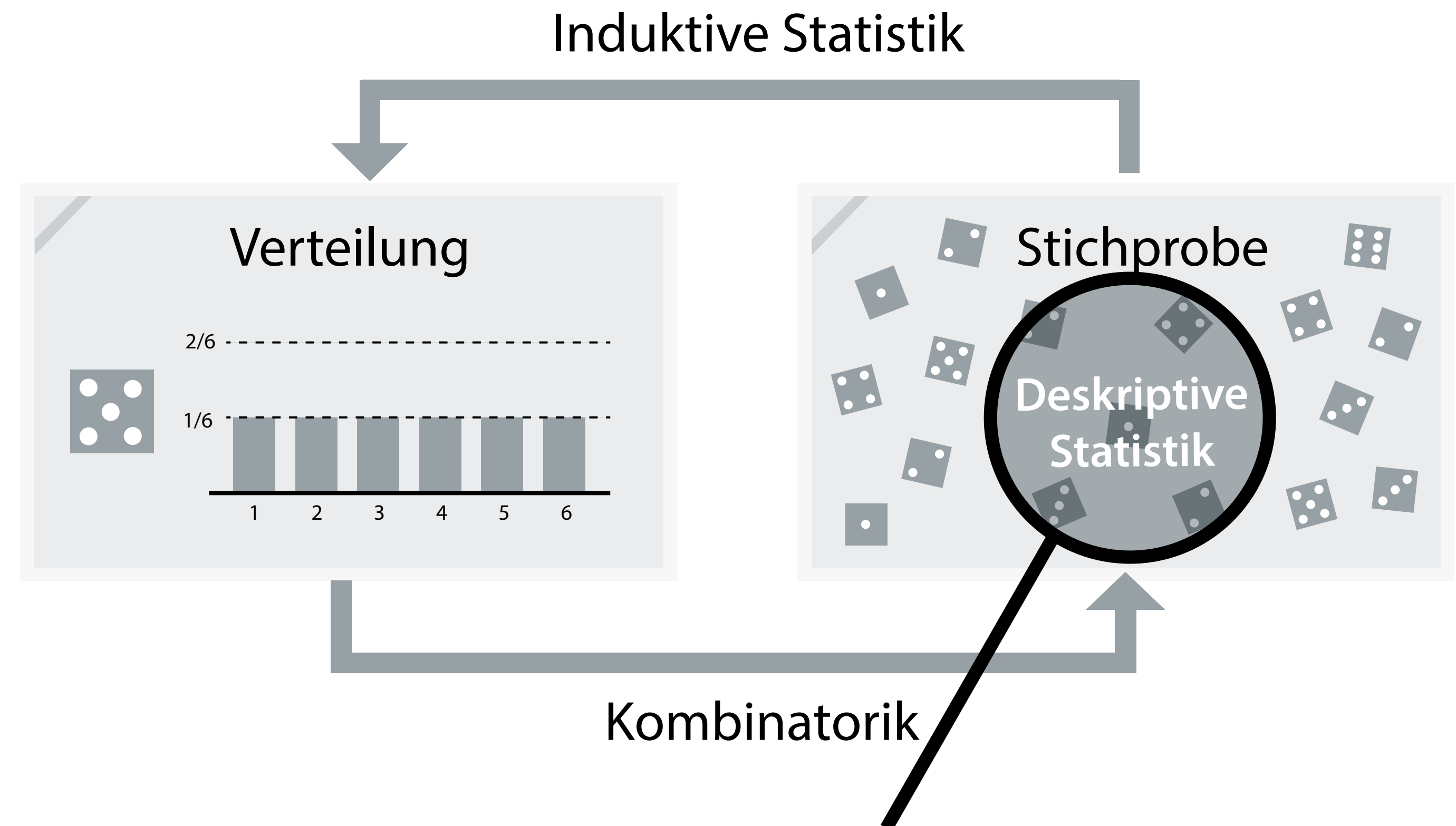
...wir verstehen mehr, wenn wir die einzelnen Teilgebiete und ihren Zusammenhang verstehen.



Teilgebiete der Statistik

Statistik besteht aus den drei Teilgebieten:

- Deskriptive Statistik
- Kombinatorik
- Induktive Statistik



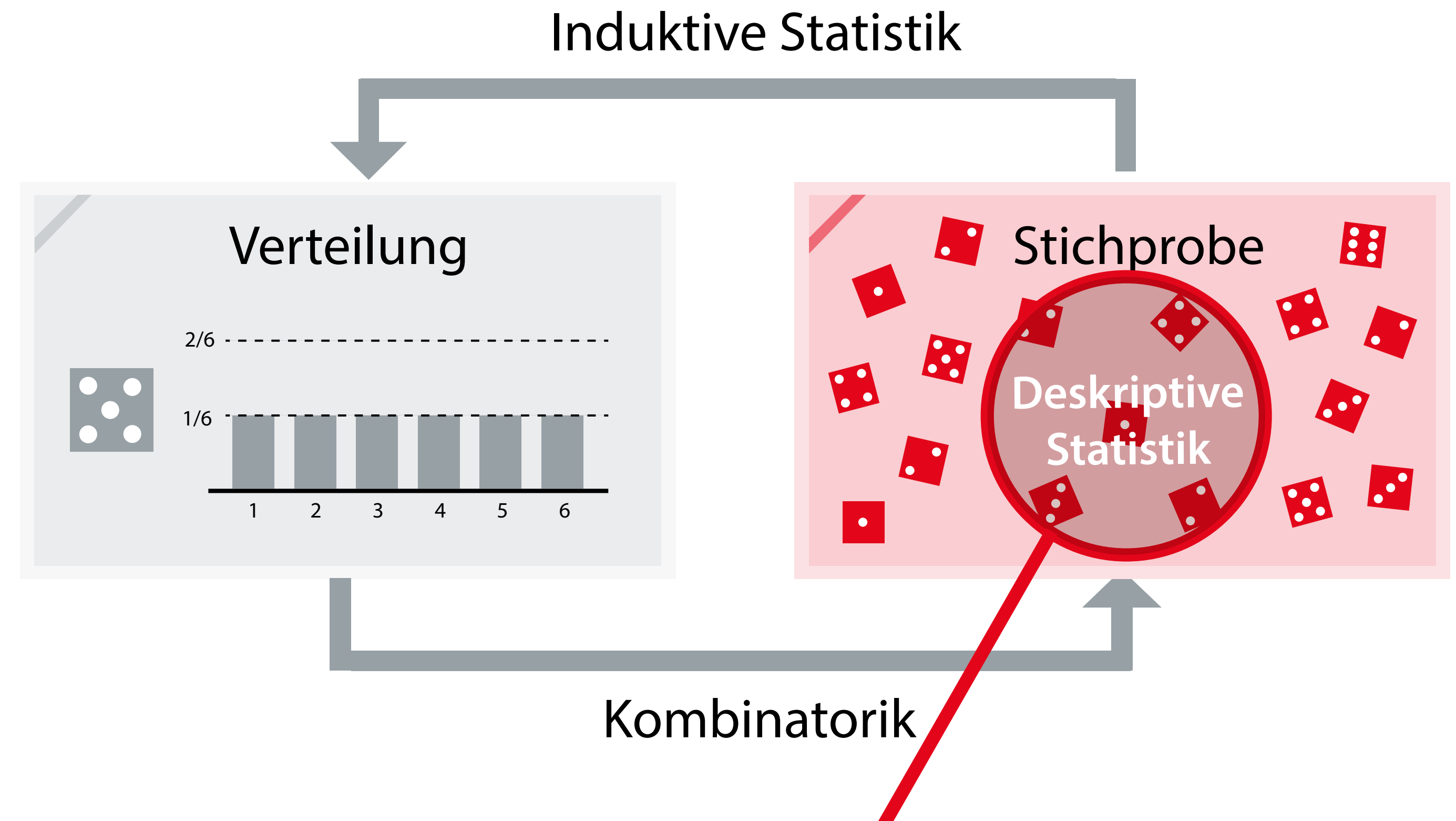
Teilgebiete der Statistik

Deskriptive Statistik beschreibt eine Stichprobe mit diversen Kennzahlen wie z. B.:

Lagemaße (Mittelwert, Median, Modalwert)

Streuemaße (Varianz, Standardabweichung)

Neben reinen Wertangaben kommen hier auch Tabellen und Schaubilder zum Einsatz!



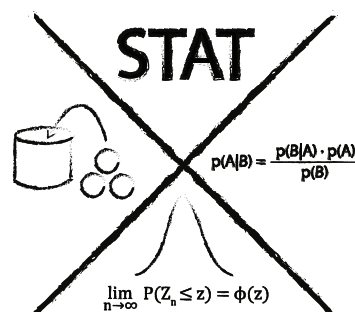
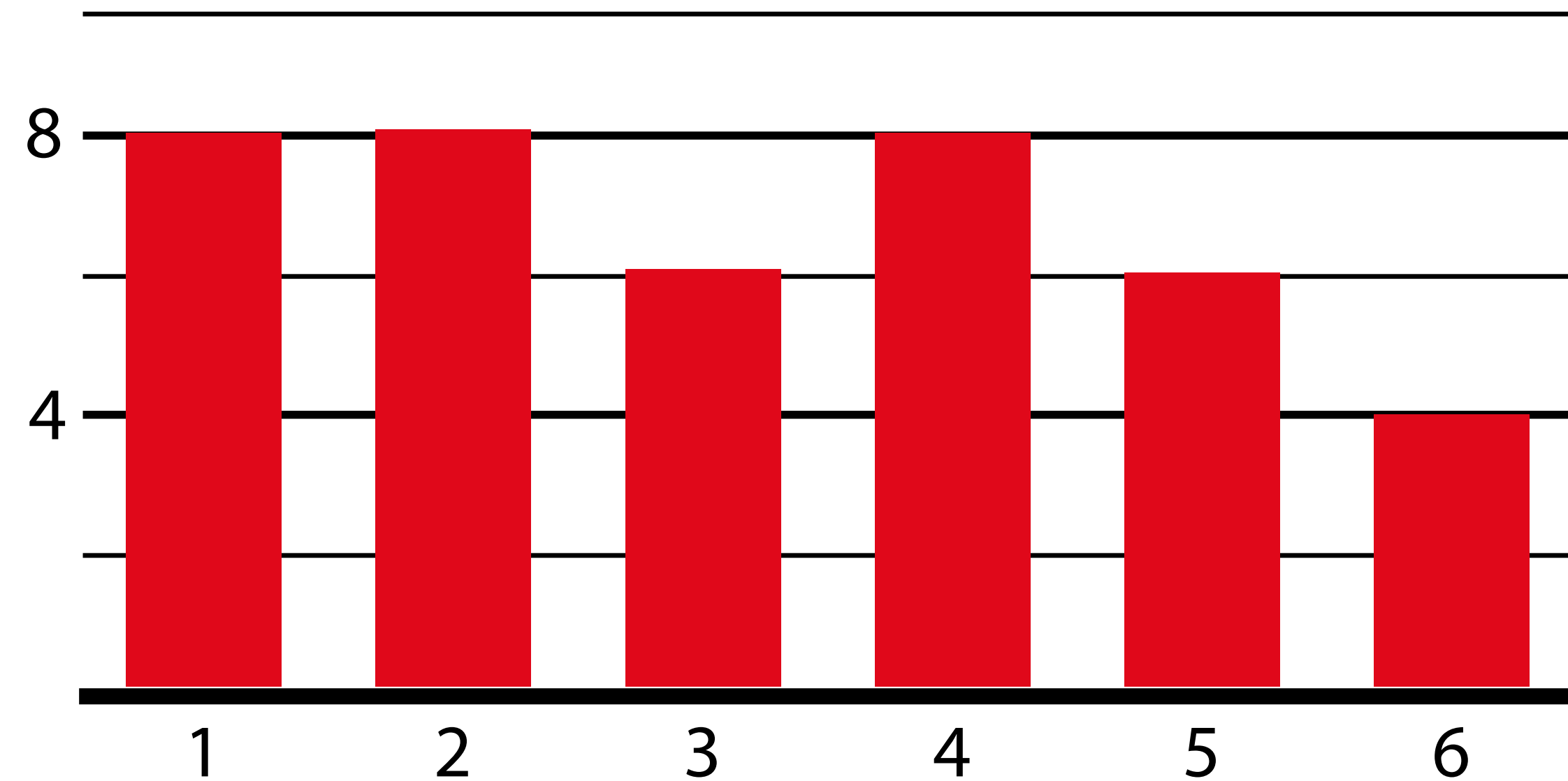
Teilgebiete der Statistik

Deskriptive Statistik - Beispiel

Wir erheben eine Stichprobe, indem wir 40 mal mit einem Würfel würfeln.

Wir beschreiben diese Stichprobe durch ein Histogramm und geben die durchschnittlich gewürfelte Augenzahl an.

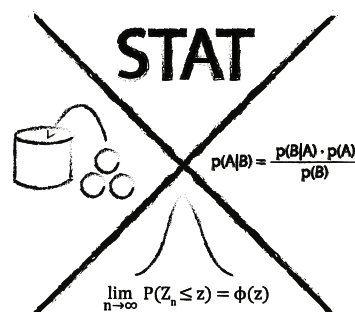
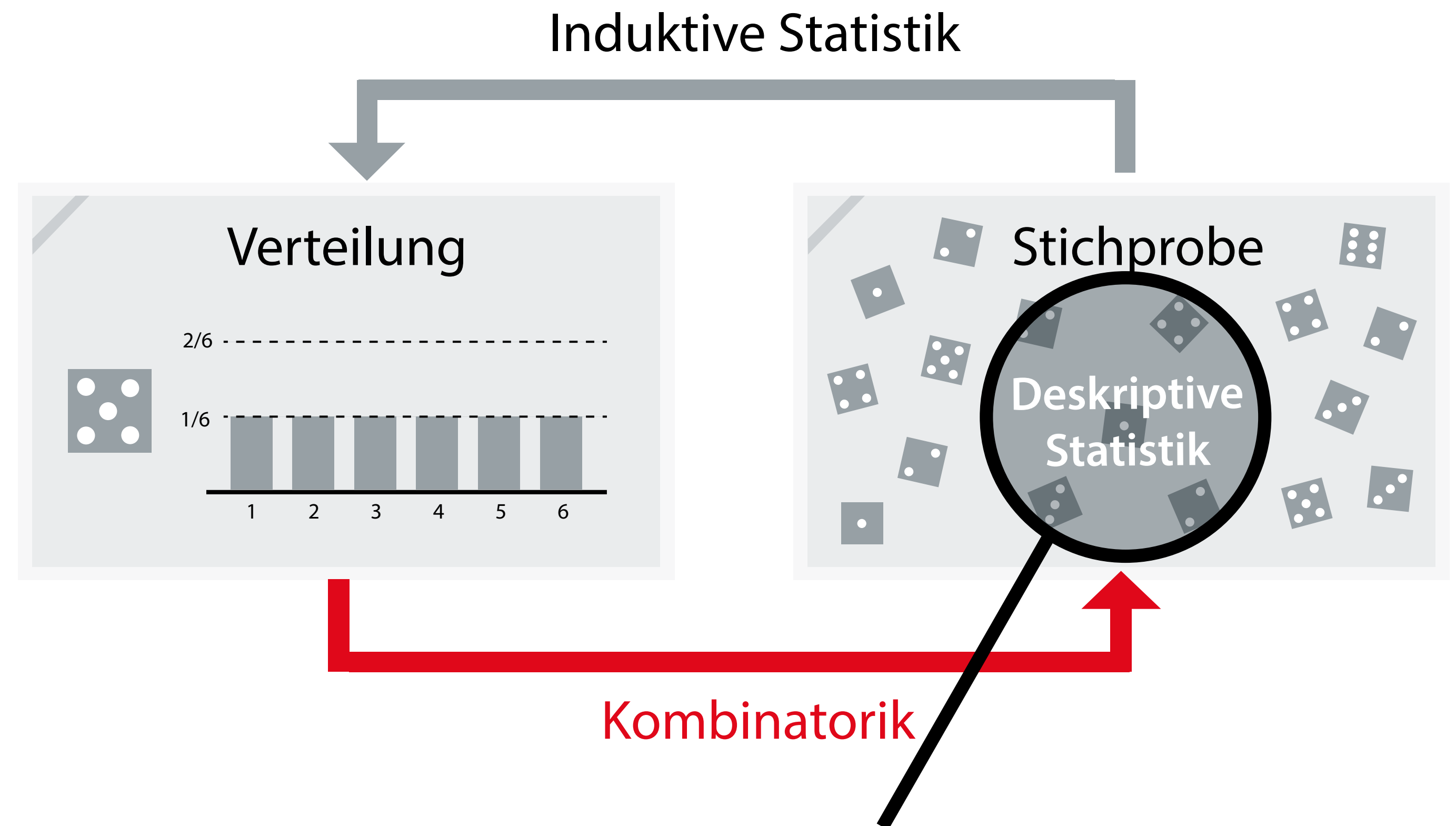
Histogramm 40x Würfeln (Durchschnitt: 3.2)



Teilgebiete der Statistik

Kombinatorik berechnet aus einer gegebenen Verteilung bzw. Grundgesamtheit Wahrscheinlichkeiten.

Einige der grundlegenden Wahrscheinlichkeitsmodelle (z. B. Urnenmodelle) sind oft schon aus der Schulmathe bekannt.



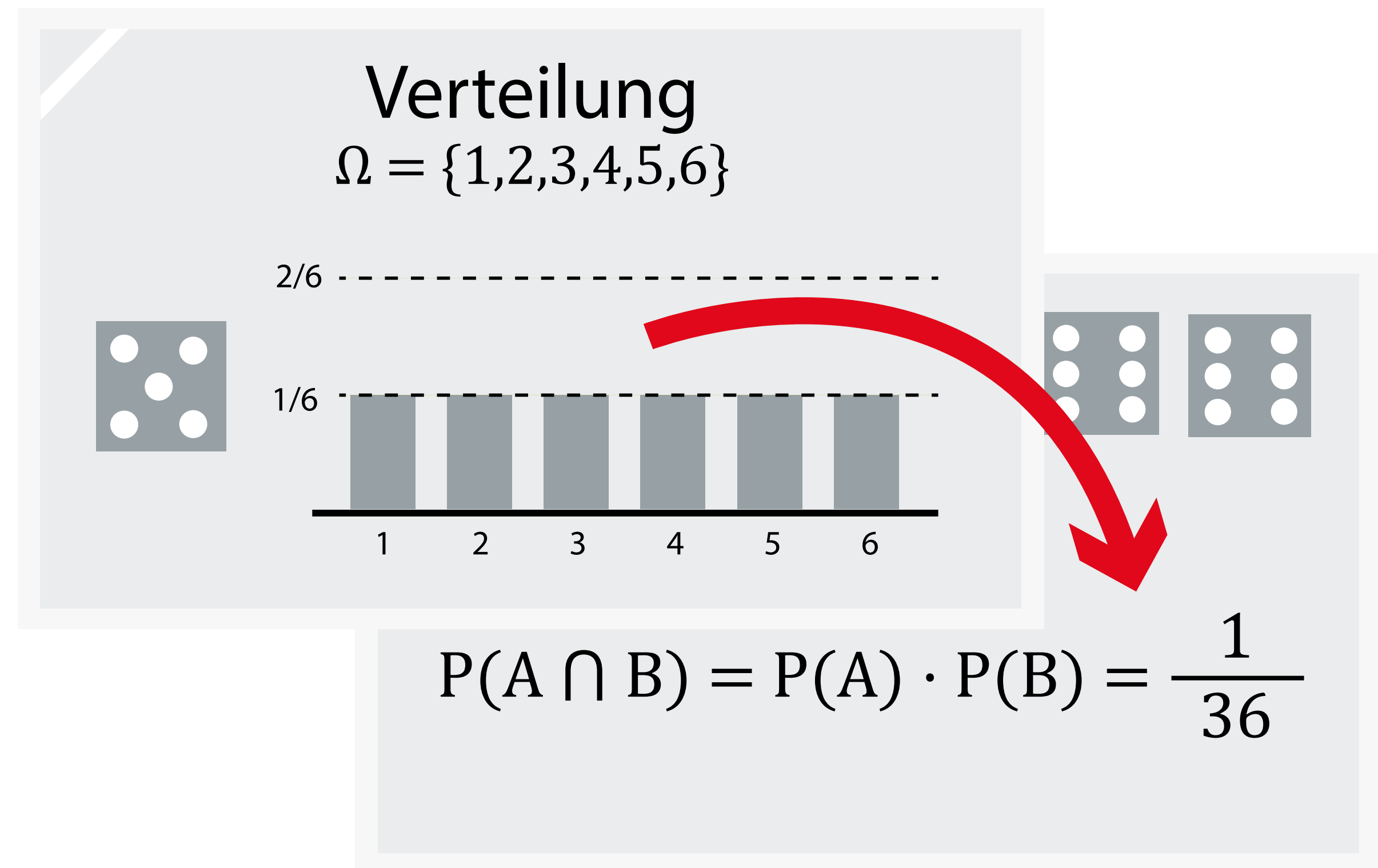
Teilgebiete der Statistik

Kombinatorik - Beispiel

Wie hoch ist die Wahrscheinlichkeit weniger als 4 Augen zu würfeln?

Wie hoch ist die Wahrscheinlichkeit zweimal hintereinander eine 6 zu würfeln?

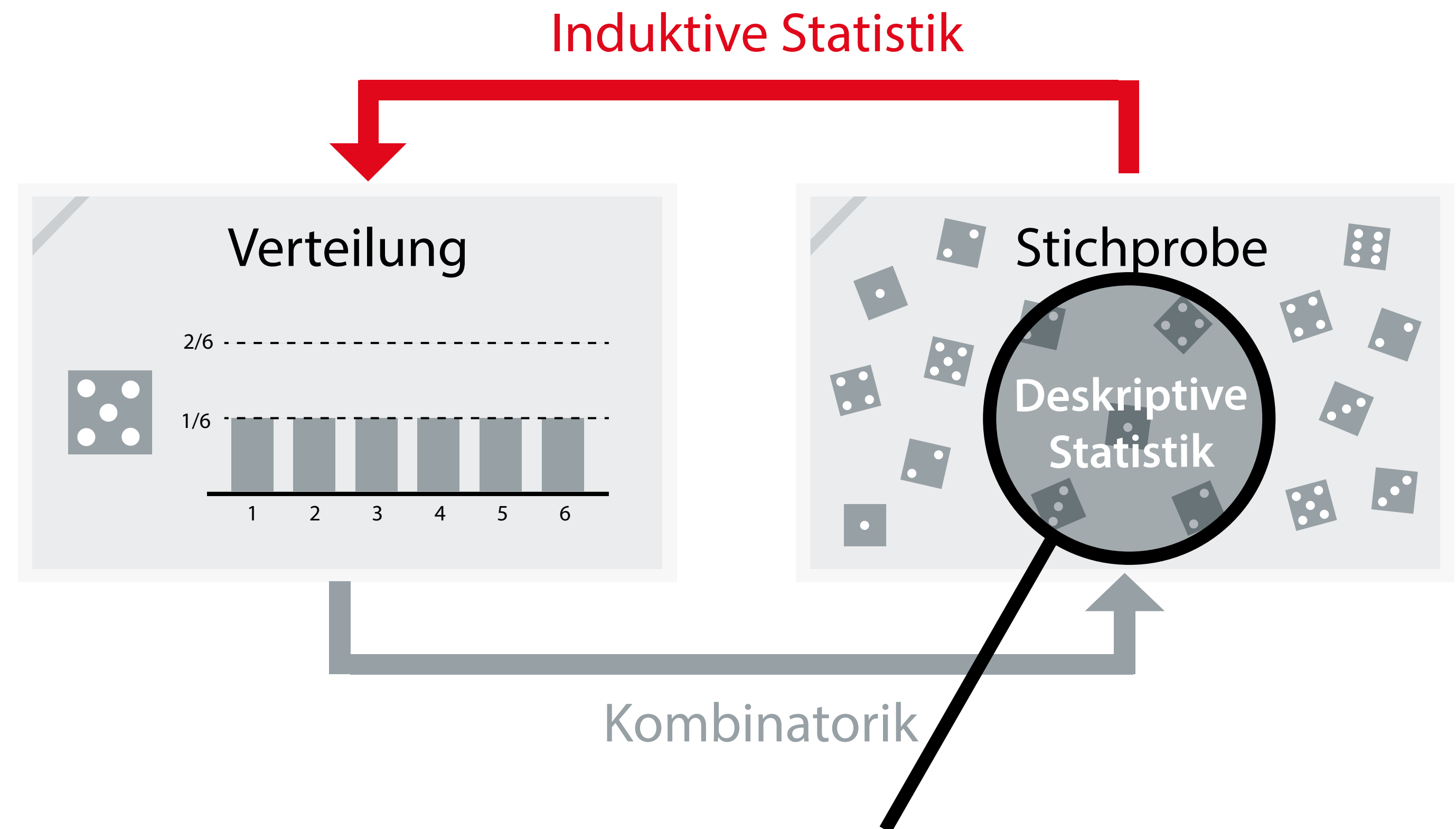
Wie hoch ist die Wahrscheinlichkeit bei 3 Würfeln keine einzige 1 zu würfeln?



Teilgebiete der Statistik

Induktive Statistik schließt von einer Stichprobe auf die zugrunde liegende Verteilung in der Grundgesamtheit.

Bei diesem anspruchsvollen Teilgebiet werden wir mit Hypothesen und statistischen Tests arbeiten!



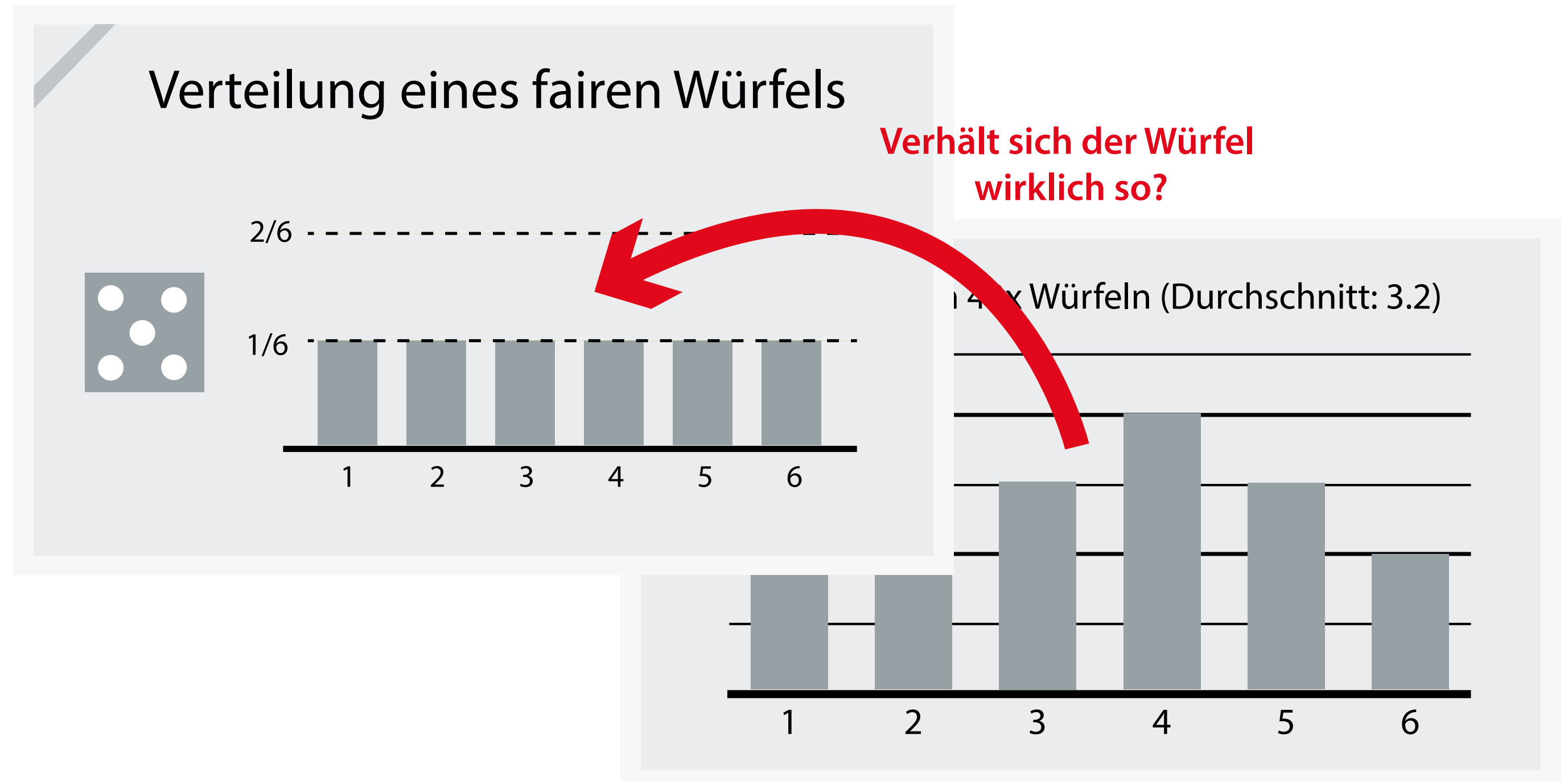
Teilgebiete der Statistik

Beispiel: Induktive Statistik mit Würfel

Wir vergleichen unsere Stichprobe mit der theoretisch erwarteten Verteilung.

Statistischer Test auf Gleichverteilung mit Werten von 1 bis 6

Statistischer Test auf durchschnittliche Augenzahl von 3.5

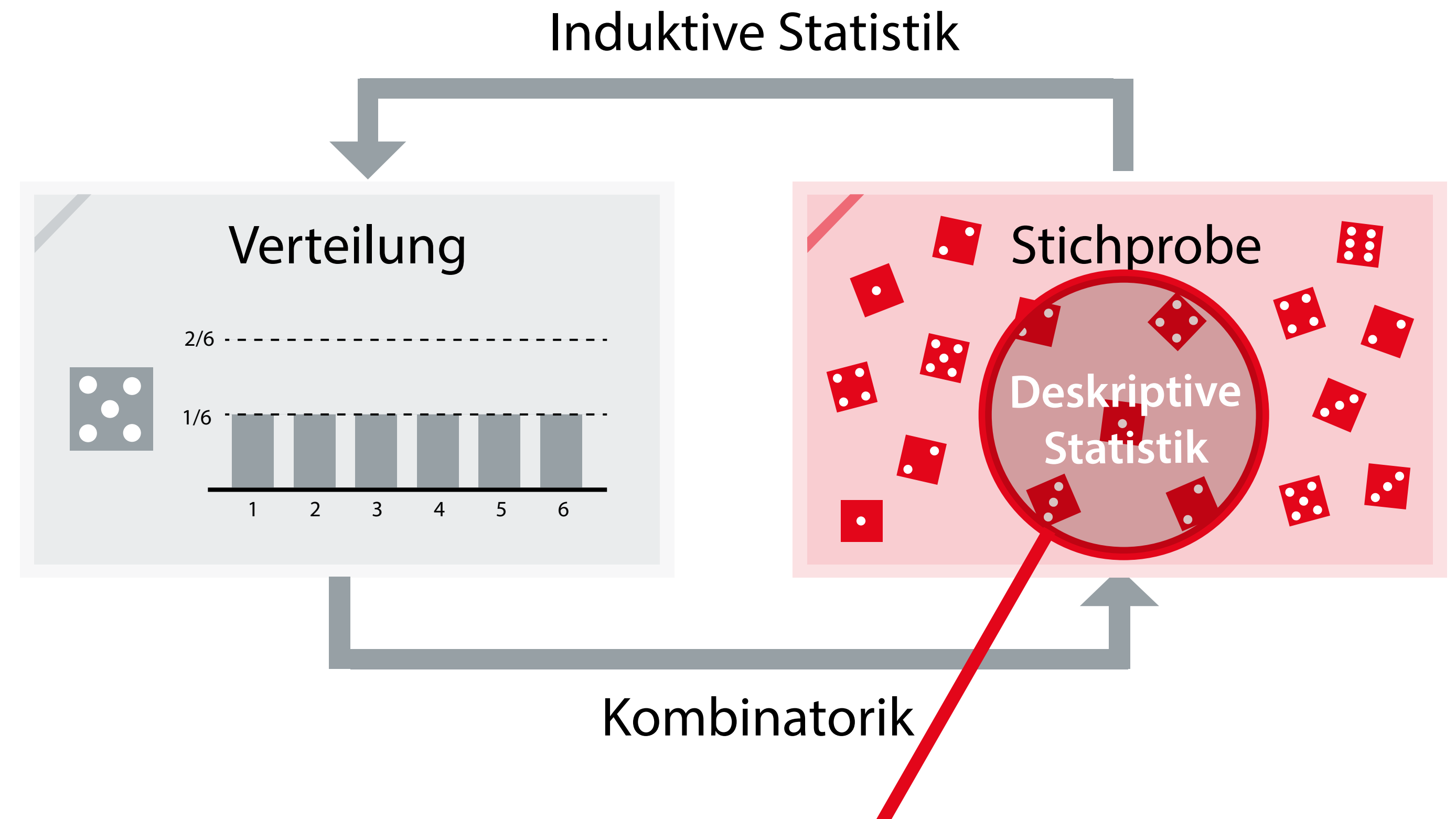


Grundlegende Begriffe

In der **deskriptiven Statistik** möchten wir stichprobenartig erhobene Merkmale beschreiben und visualisieren.

Was ist eine Stichprobe?

Was sind Merkmale?



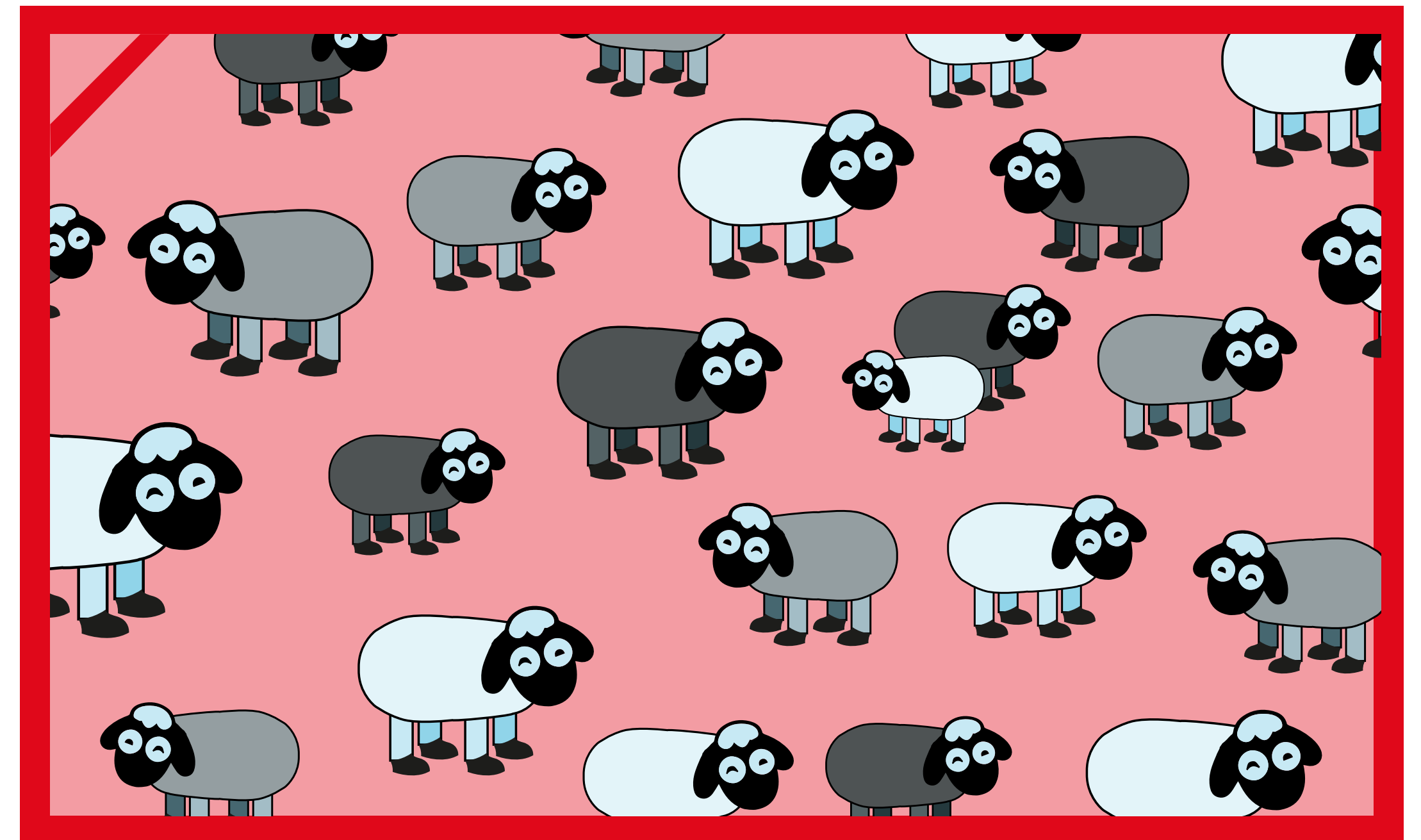
Grundlegende Begriffe

Wir betrachten eine große Herde von 10.000 Schafen und wollen diese statistisch bzgl. ihren Fellfarben und ihrem Gewicht untersuchen.

Die Herde ist unsere **Grundgesamtheit**.

Jedes einzelne Schaf ist ein **Merkmalsträger**.

Die **Merkmale** sind Fellfarbe und Gewicht.

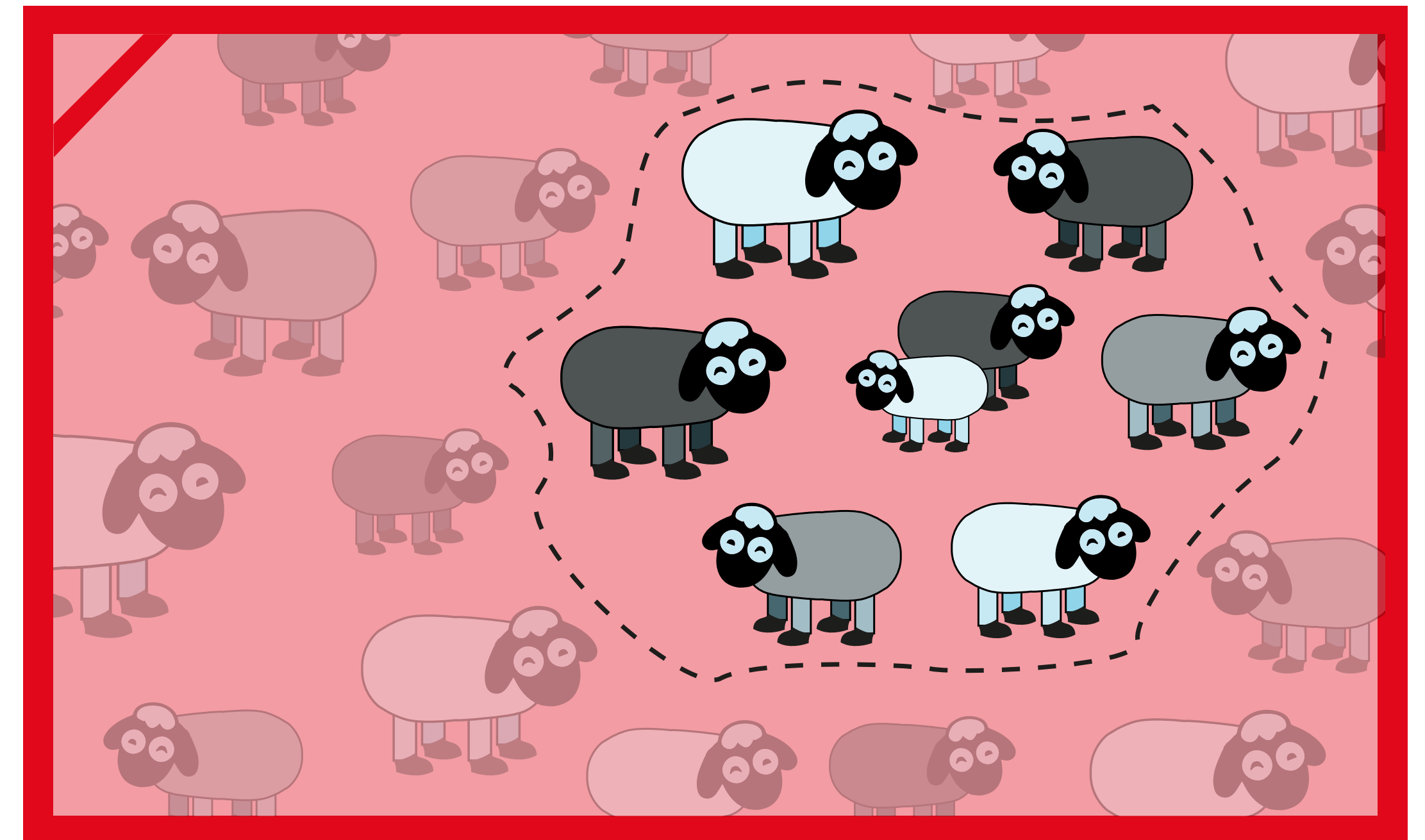


Grundlegende Begriffe

Würden wir von allen Schafen Fellfarbe und Gewicht erfassen, würden wir eine **Vollerhebung** durchführen.

Hier wäre das möglich; in anderen Beispielen sprechen wirtschaftliche und technische Gründe dagegen.

Wir ziehen eine zufällige **Stichprobe** an Merkmalsträgern aus der Grundgesamtheit und erfassen die Merkmale nur von diesen!

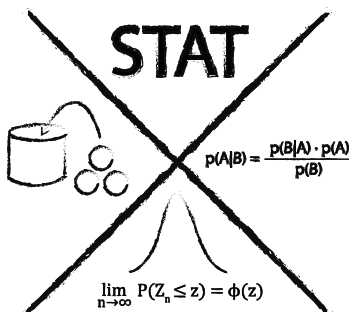


Grundlegende Begriffe

Jedes Merkmal nimmt bei jedem Merkmalsträger eine bestimmte **Ausprägung** an.

Die oft tabellarisch vorgenommene Aufzählung von Merkmalen und den zu diesen Merkmalen erhobenen Ausprägungen nennt man **Urliste**.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0



Grundlegende Begriffe

Bei den Merkmalen können wir unterscheiden ...

Die Ausprägungen von **quantitativen** Merkmalen sind Zahlenwerte.

Die Ausprägungen von **qualitativen** Merkmalen sind keine Zahlenwerte, sondern Kategorien.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0

Diagramm zur Klassifizierung der Merkmale:

- Das Wort **quantitativ** hat zwei rote Pfeile, die auf die Spalten **Art** und **Gewicht (kg)** zeigen.
- Das Wort **qualitativ** hat zwei rote Pfeile, die auf die Spalten **Fellfarbe** und **Gewicht (kg)** zeigen.

Grundlegende Begriffe

Viele Merkmale lassen sich grundsätzlich sowohl quantitativ als auch qualitativ angeben.

Wenn qualitative Merkmale nur bestimmte Ausprägungen annehmen können, werden diese manchmal als Zahlenwert **codiert** angegeben.

Um den Datensatz zu verstehen, benötigt man dann eine **Legende** bzw. ein **Codebuch**.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	1	1	87.5
04257	1	0	92.4
05100	0	0	
00047	1		
09432	0		
08254	0		

Codebuch	
Art	Fellfarbe
0 - ♂ Bock	0 ○ weiß
1 - ♀ Schaf	1 ● grau
	2 ● schwarz

Grundlegende Begriffe

Quantitative Merkmale können dagegen **klassiert**, d. h. in Klassen eingeteilt, werden. Diese Klassen werden entweder ...

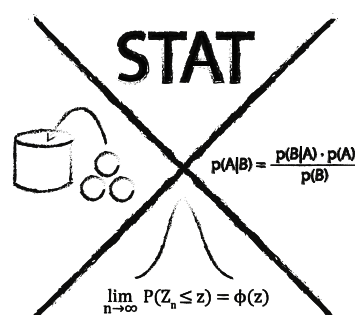
...durch eine Ober- und eine Untergrenze definiert.

...durch eine Mitte und eine Breite definiert.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	S
04257	♀	○ weiß	S
05100	♂	○ weiß	
00047	♀	●	
09432	♂	●	
08254	♂	●	

Gewichtsklassen

XS	bis zu 79.9kg
S	80.0kg bis 99.9kg
M	100.0kg bis 119.9kg
L	120.0kg bis 139.9kg
XL	ab 140.0kg



Grundlegende Begriffe

Bei den quantitativen Merkmalen können wir zwischen drei Skalenniveaus unterscheiden:

- Ordinalskala
- Intervallskala
- Proportionalskala

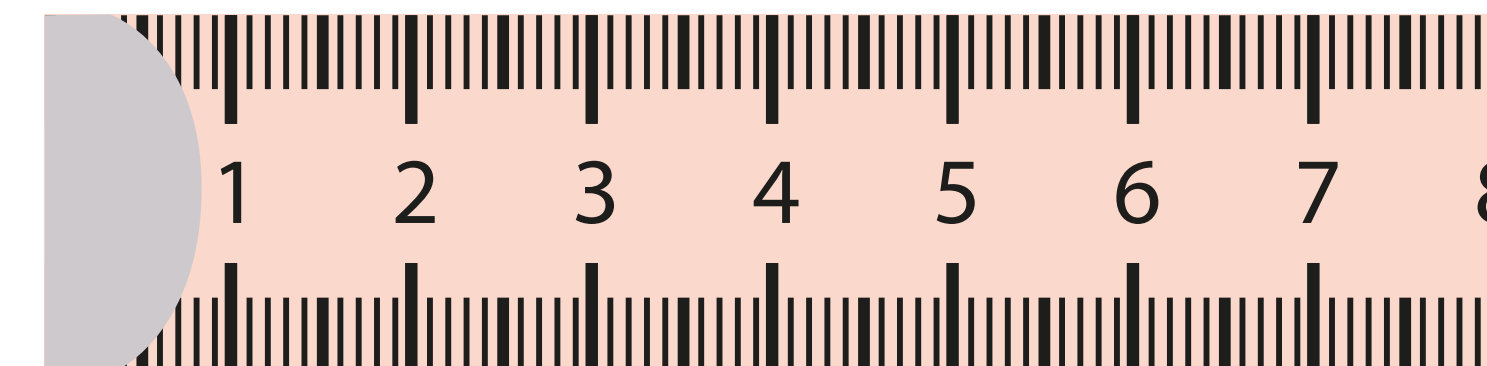
Intervall- und Proportionalskala werden oft auch als Kardinalskala oder als metrische Skala bezeichnet.



Ordinalskala



Kardinalskala
(Intervall)



Kardinalskala
(Proportional)

Grundlegende Begriffe

Ordinalskalen geben nur eine Rangfolge vor und erlauben keine Aussagen über Abstände und Proportionen. Klassisches Beispiel sind Schulnoten:

Eine 2.3 ist besser als eine 3.1



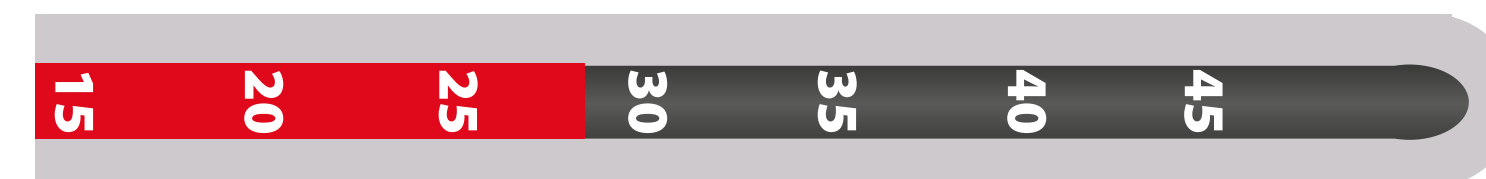
Zwischen 1.0 und 2.0 ist der gleiche Leistungsabstand wie zwischen 2.0 und 3.0



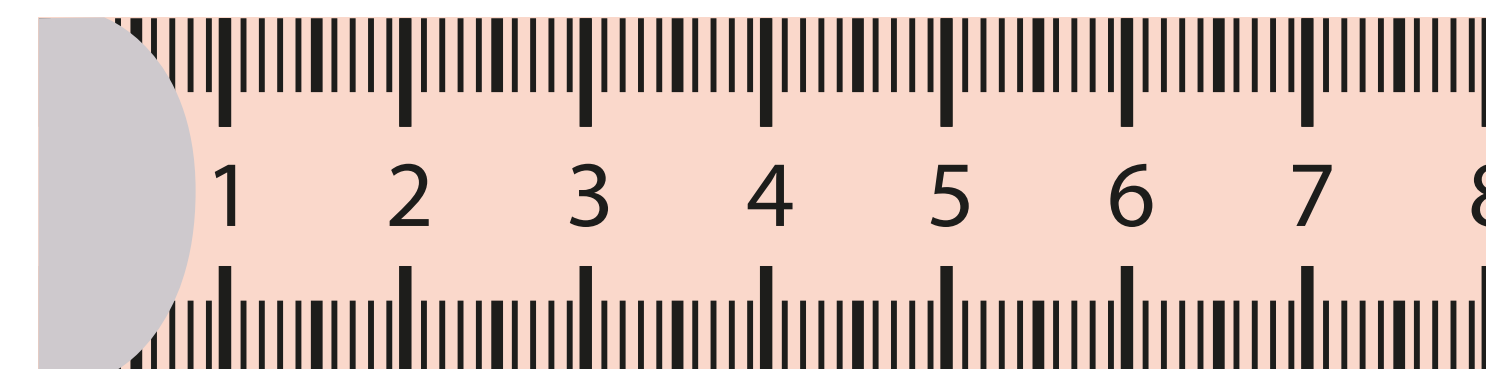
Eine 2.0 ist halb so schlecht wie eine 4.0



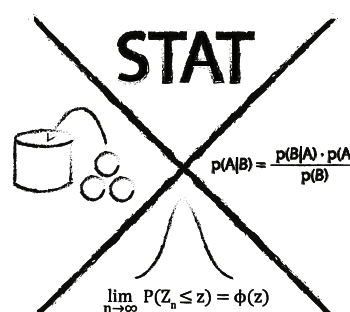
Ordinalskala



Kardinalskala
(Intervall)



Kardinalskala
(Proportional)



Grundlegende Begriffe

Intervallskalen geben eine Rangfolge vor und erlauben zusätzlich Aussagen über Abstände. Aussagen über Proportionen sind weiterhin nicht möglich. Beispiel Temperatur in Celsius:

15°C ist wärmer als -5°C



Zwischen 15°C und -5°C ist der gleiche Abstand wie zwischen 30°C und 10°C



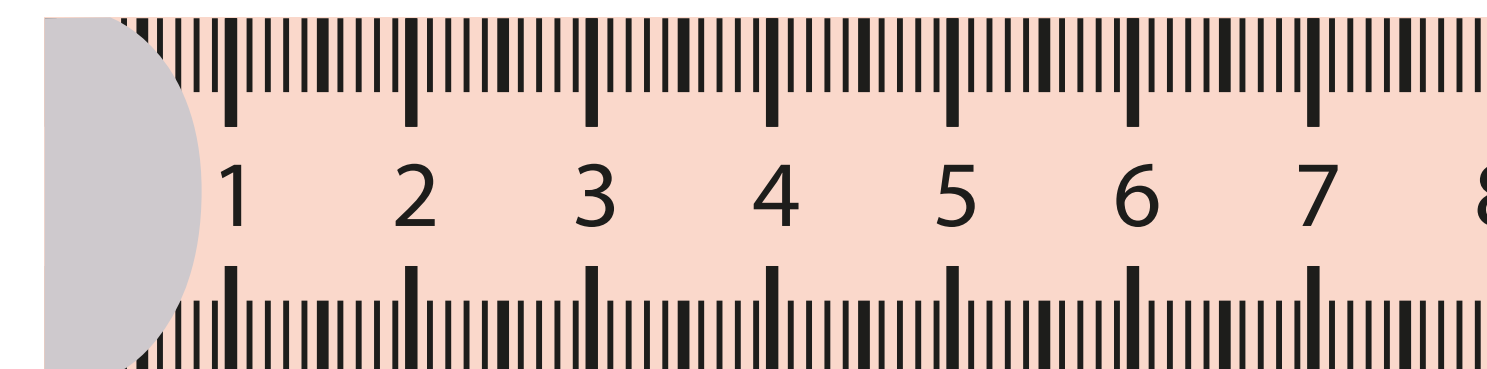
15°C ist -3 mal so warm wie -5°C



Ordinalskala



Kardinalskala
(Intervall)



Kardinalskala
(Proportional)

Grundlegende Begriffe

Proportionalskalen geben eine Rangfolge vor und erlauben Aussagen über Abstände und Proportionen. Beispiel Längenmaß:

800m sind länger als 200m



Zwischen 800m und 200m ist der gleiche Unterschied wie zwischen 4000m und 3400m



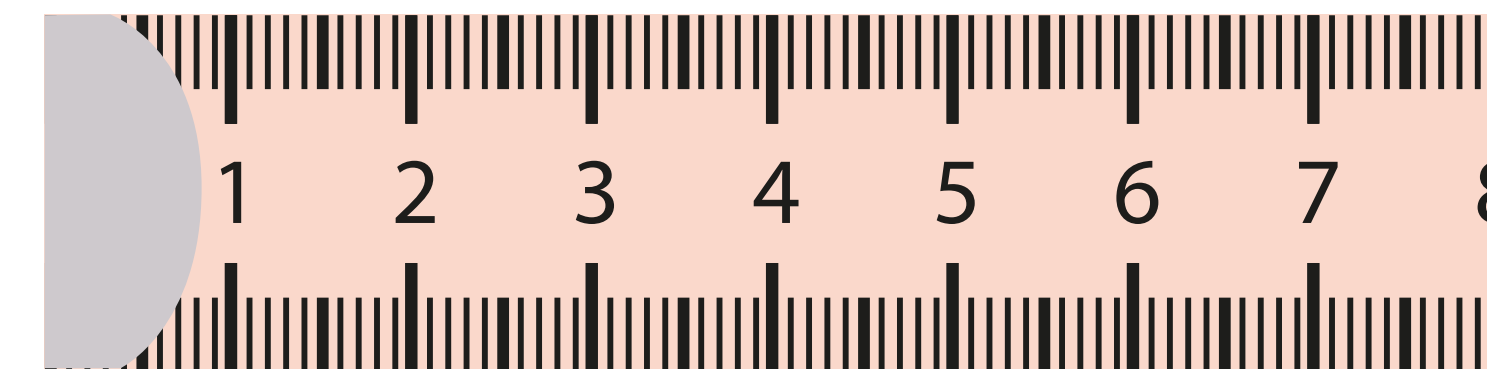
800m sind vier mal so lang wie 200m



Ordinalskala



Kardinalskala
(Intervall)



Kardinalskala
(Proportional)

Grundlegende Begriffe

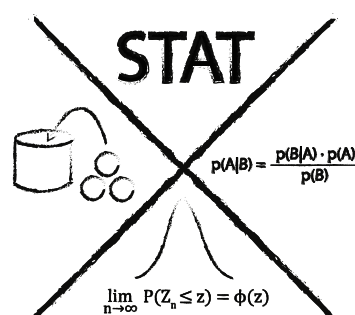
Eine weitere Unterscheidung für quantitative Merkmale:

Wenn nur bestimmte Ausprägungen möglich sind, bezeichnet man das Merkmal als **diskret**.

Die Kennnummer, die wir den Schafen geben, ist z. B. ein diskretes Merkmal. Es gibt Schaf Nr. 1 und Schaf Nr. 2 aber nichts dazwischen.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0

diskret (pointing to Schaf) kontinuierlich (pointing to Gewicht (kg))
 binär (pointing to Art) kategorial (pointing to Fellfarbe)



Grundlegende Begriffe

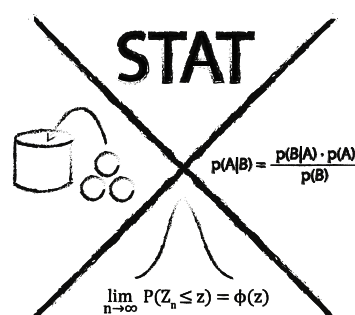
Das würde auch gelten, wenn die Herde nicht auf 10000 Schafe begrenzt ist.

In diesem Fall gibt es zwar unendlich viele Werte, die jedoch an ein festes Raster gebunden sind.

Mathematisch bezeichnet man das als Merkmal mit abzählbar unendlich vielen möglichen Ausprägungen.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0

diskret (pointing to Schaf)
 kontinuierlich (pointing to Gewicht (kg))
 binär (pointing to Art)
 kategorial (pointing to Fellfarbe)



Grundlegende Begriffe

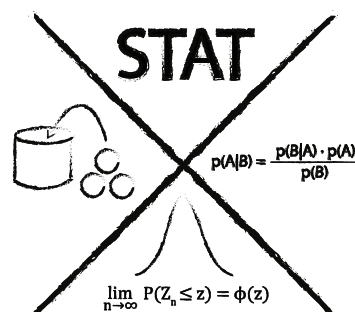
Wenn nicht-abzählbar unendlich viele Ausprägungen möglich sind, bezeichnet man das Merkmal als **kontinuierlich**.

Das Gewicht der Schafe ist ein Beispiel! Zwischen den Ausprägungen 87.0 kg und 88.0 kg ist z. B. auch 87.5 kg als Ausprägung möglich.

Hinweis: Auch wenn wir hier immer nur eine Nachkommastelle angeben, könnten wir (eine genaue Waage vorausgesetzt) immer feinere Zwischenwerte finden.

Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0

diskret (pointing to Schaf)
 kontinuierlich (pointing to Gewicht (kg))
 binär (pointing to Art)
 kategorial (pointing to Fellfarbe)



Grundlegende Begriffe

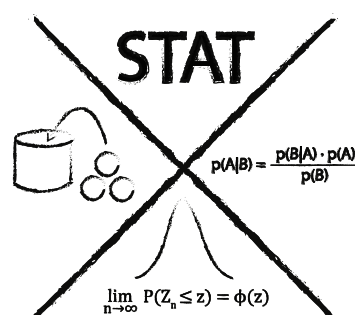
Bei qualitativen Merkmalen gibt es eine ähnliche Unterscheidung:

Wenn nur zwei Ausprägungen möglich sind, nennt man das Merkmal **binär** bzw. **dichotomisch**.

Wenn mehrere Ausprägungen möglich sind, die keine Ordnung vorgeben, nennt man das Merkmal **kategorial**.

Wenn mehrere Ausprägungen möglich sind und diese eine Ordnung vorgeben, nennt man das Merkmal **ordinal**.

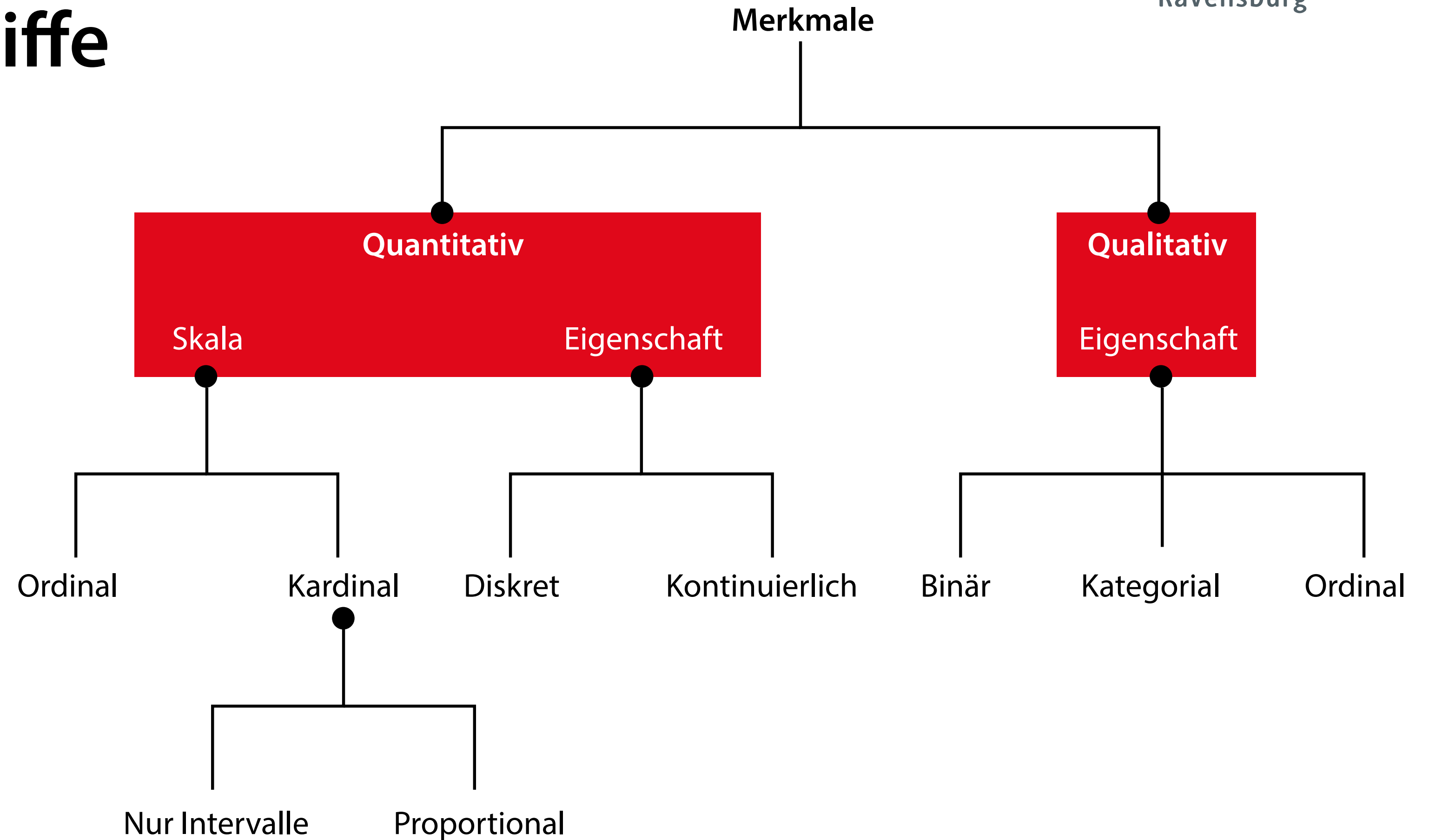
diskret		kontinuierlich	
Schaf	Art	Fellfarbe	Gewicht (kg)
00137	♀	● grau	87.5
04257	♀	○ weiß	92.4
05100	♂	○ weiß	77.3
00047	♀	● grau	61.5
09432	♂	● schwarz	117.4
08254	♂	● grau	84.0
binär		kategorial	



Grundlegende Begriffe

Hier eine Übersicht über die vielen Adjektive, die wir auf den letzten Folien kennengelernt haben.

Sie werden uns im Folgenden immer wieder begegnen!



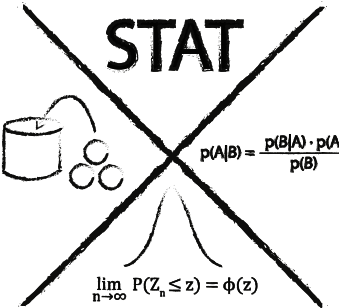
Grundlegende Begriffe

Überlege bei folgenden Datensätzen, ob die Merkmale quantitativ oder qualitativ sind und welche Eigenschaften sie haben (binär, diskret usw.)

Gebe bei quantitativen Merkmalen zusätzlich an, ob sie auf einer Ordinal- oder Kardinalskala gemessen werden.

Käsesorte	Würze	Im Angebot	Fett(%)
Butterkäse	Mild	Ja	60
Gouda	Normal	Nein	48
Edamer	Normal	Nein	45
Emmentaler	Normal	Nein	45
Tilsiter	Würzig	Ja	45
Bergkäse	Würzig	Nein	45

Gericht	Veggi	Preis	Energie
Linsen	Ja	1.90€	174 kcal
Schnitzel	Nein	2.60€	192 kcal
Pommes	Ja	1.40€	321 kcal
Wurstsalat	Nein	2.00€	251 kcal



Grundlegende Begriffe

Die Käsesorte ist ein qualitativ kategoriales Merkmal. Es sind mehr als zwei mögliche Ausprägungen möglich und es wird keine Ordnung vorgegeben.

Die Würze ist dagegen ein qualitativ ordinales Merkmal, da hier eine Ordnung vorgegeben wird.

Die dritte Spalte zeigt ein qualitativ binäres Merkmal, da nur die Ausprägungen „Ja“ und „Nein“ möglich sind.

Der Fettgehalt ist ein quantitatives Merkmal, das auf einer Kardinalskala gemessen wird.

Käsesorte	Würze	Im Angebot	Fett(%)
Butterkäse	Mild	Ja	60
Gouda	Normal	Nein	48
Edamer	Normal	Nein	45
Emmentaler	Normal	Nein	45
Tilsiter	Würzig	Ja	45
Bergkäse	Würzig	Nein	45

Grundlegende Begriffe

Der Name des Gerichts ist ein qualitatives Merkmal. Es ist kategorial, da es mehr als 2 mögliche Ausprägungen gibt und keine Ordnung vorgegeben wird.

Die zweite Spalte ist ein qualitatives und binäres Merkmal, da es nur zwei mögliche Merkmalsausprägungen gibt.

Sowohl Preis als auch Energie sind quantitative Merkmale, die auf einer Kardinalskala gemessen werden.

Gericht	Veggi	Preis	Energie
Linsen	Ja	1.90€	174 kcal
Schnitzel	Nein	2.60€	192 kcal
Pommes	Ja	1.40€	321 kcal
Wurstsalat	Nein	2.00€	251 kcal

Lageparameter

In Formeln wird die **Stichprobengröße** mit n und die für ein Merkmal gemessenen Werte mit x_i bezeichnet.

















Das **Subskript** i kann Werte von 1 bis n annehmen:

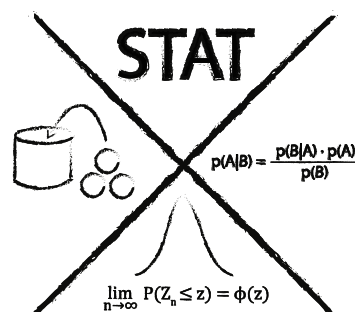
$$x_1 = 123$$

$$x_4 = 128$$

$$x_5 = 112$$

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h



Lageparameter

Die Lageparameter eines Merkmals geben eine Ausprägung an, zu dem dieses Merkmal tendiert.

















Die drei wichtigsten Lageparameter sind:

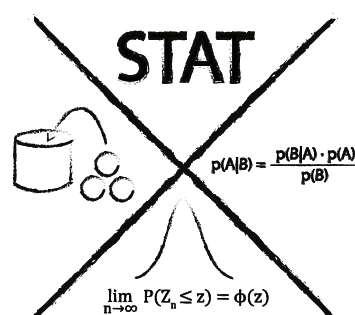
Mittelwert

Median

Modalwert

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h










Lageparameter

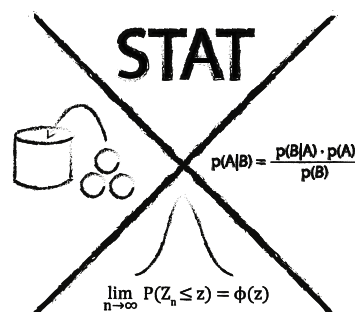
Der Mittelwert eines Merkmals ist die Summe über die gemessenen Werte geteilt durch die Stichprobengröße:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} [x_1 + x_1 + \dots + x_n]$$

In dem Beispiel rechts wäre der Mittelwert 140.5 km/h.

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h



















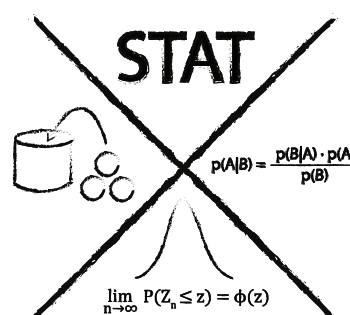
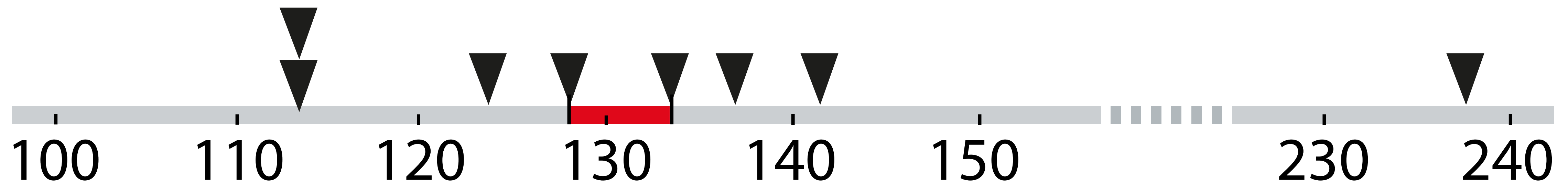
Lageparameter

Der Median ist ein Wert, von dem aus gesehen es gleich viele größere und kleinere Werte gibt.

Im Beispiel rechts suchen wir einen Wert, von dem aus gesehen es jeweils 4 langsamere und schnellere Fahrer gibt.

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h



















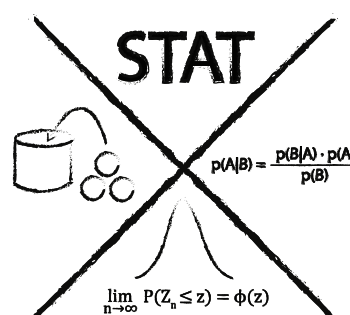
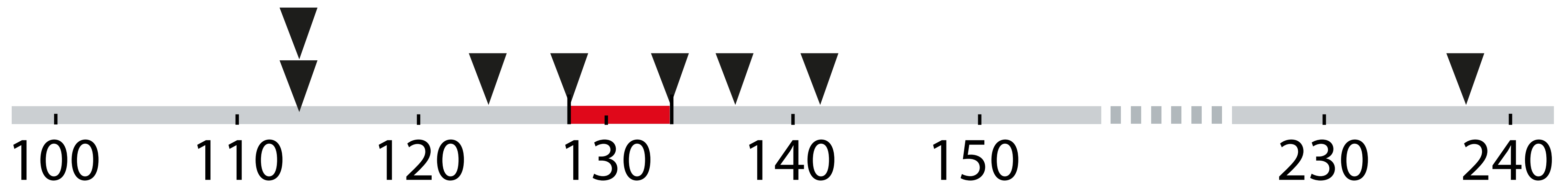
Lageparameter

Der Median muss zwischen 128 und 133 km/h liegen.

Alle Werte dazwischen erfüllen das geforderte Kriterium. Meistens verwenden wir in diesem Fall den Mittelwert der Ober- und Untergrenze: $\tilde{x} = 130.5$ km/h.

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h



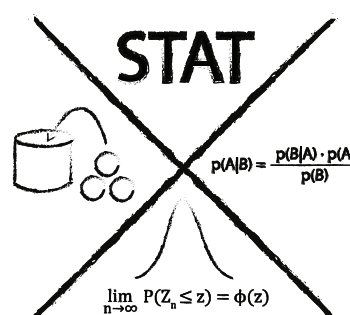
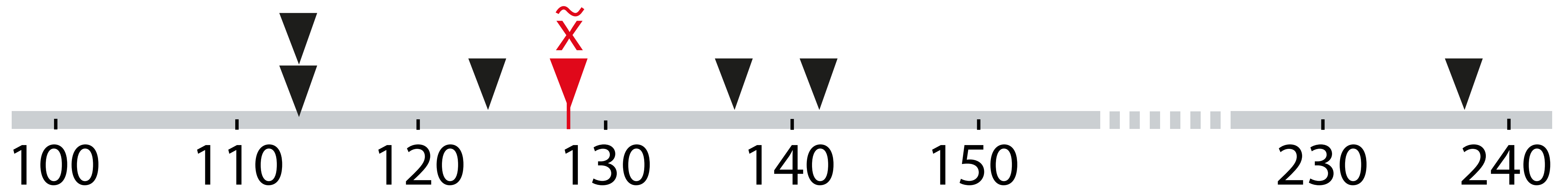
Lageparameter

Bei einer ungeraden Stichprobengröße ist der Wert des Medians eindeutig.

Reiht man die Werte von klein nach groß auf, ist der Wert in der Mitte gleichzeitig der Median: $\tilde{x} = 128 \text{ km/h}$.

Geschwindigkeitsmessung

Messung 1	ES EL 0815	123 km/h	Messung 5	RV XY 1234	112 km/h
Messung 2	RV AU 4321	112 km/h	Messung 6	RT BD1990	141 km/h
Messung 3	RV SO 2134	137 km/h	Messung 7	BC XX 666	238 km/h
Messung 4	RV SD 0741	128 km/h	Messung 8	GR ED 888	138 km/h











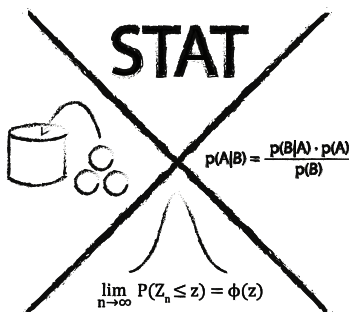
Lageparameter

Der **Modalwert** ist die Merkmalsausprägung mit der größten Häufigkeit. Hier wäre dies 112 km/h - der einzige Wert, der zweimal gemessen wurde.

Ausprägung	112	123	128	133	137	141	238
Häufigkeit (abs.)	2	1	1	1	1	1	1
Häufigkeit (rel.)	25%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%

Geschwindigkeitsmessung

Messung 1	 123 km/h	Messung 5	 112 km/h
Messung 2	 112 km/h	Messung 6	 141 km/h
Messung 3	 137 km/h	Messung 7	 238 km/h
Messung 4	 128 km/h	Messung 8	 133 km/h



Lageparameter

















Die drei Lageparameter für unser Beispiel sind sehr unterschiedlich. Wie ist das Ergebnis zu interpretieren?

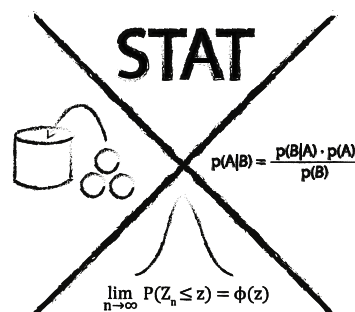
Mittelwert $\bar{x} = 140.5 \text{ km/h}$

Median $\tilde{x} = 130.5 \text{ km/h}$

Modalwert $x_m = 112.0 \text{ km/h}$

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h











Lageparameter

Die meisten Autofahrer in der Stichprobe fahren 112 km/h, allerdings fährt eine große Mehrheit deutlich schneller!

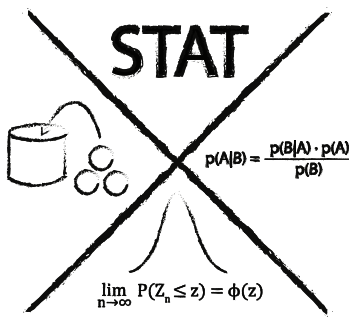
Der Modalwert ist zwar sehr einfach zu interpretieren, aber bei kleinen Stichproben auf kontinuierlichen Skalen stark vom Zufall abhängig.

Kommt jede Ausprägung nur einmal vor oder gibt es einen Gleichstand zwischen zwei Ausprägungen, ist der Modalwert sogar nicht definiert.

Geschwindigkeitsmessung

Messung 1	 123 km/h	Messung 5	 112 km/h
Messung 2	 112 km/h	Messung 6	 141 km/h
Messung 3	 137 km/h	Messung 7	 238 km/h
Messung 4	 128 km/h	Messung 8	 133 km/h

Ausprägung	112	123	128	133	137	141	238
Häufigkeit (abs.)	2	1	1	1	1	1	1
Häufigkeit (rel.)	25%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%



















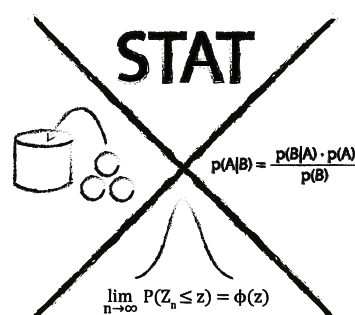
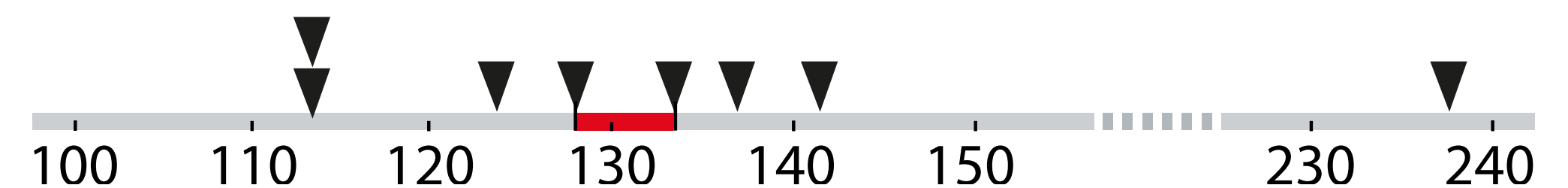
Lageparameter

Jeweils die Hälfte der Autofahrer fahren langsamer und schneller als 130.5 km/h.

Der Medianwert unterschlägt aber, dass es einen Autofahrer gibt, der fast doppelt so schnell fährt, aber genau das kann auch ein Vorteil sein, wenn man einen „typischen Autofahrer“ beschreiben möchte.

Geschwindigkeitsmessung

Messung 1	Messung 5
 ES  EL 0815 123 km/h	 RV  XY 1234 112 km/h
Messung 2	Messung 6
 RV  AU 4321 112 km/h	 RT  BD1990 141 km/h
Messung 3	Messung 7
 RV  SO 2134 137 km/h	 BC  XX 666 238 km/h
Messung 4	Messung 8
 RV  SD 0741 128 km/h	 GP  RD 888 133 km/h



















Lageparameter

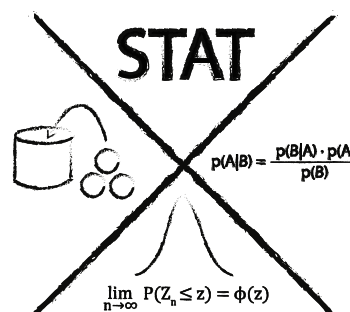
Im arithmetischen Mittel fahren die Autofahrer 140.5 km/h, wobei die meisten Autofahrer deutlich langsamer fahren!

Beim arithmetischen Mittel werden die meisten Informationen der einzelnen Werte transportiert, aber dadurch ist es anfällig für Ausreißer!

Ausreißer sind ungewöhnlich hohe oder niedrige Werte. In diesem Beispiel ist der Biberacher mit 238 km/h ein solcher Ausreißer.

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h

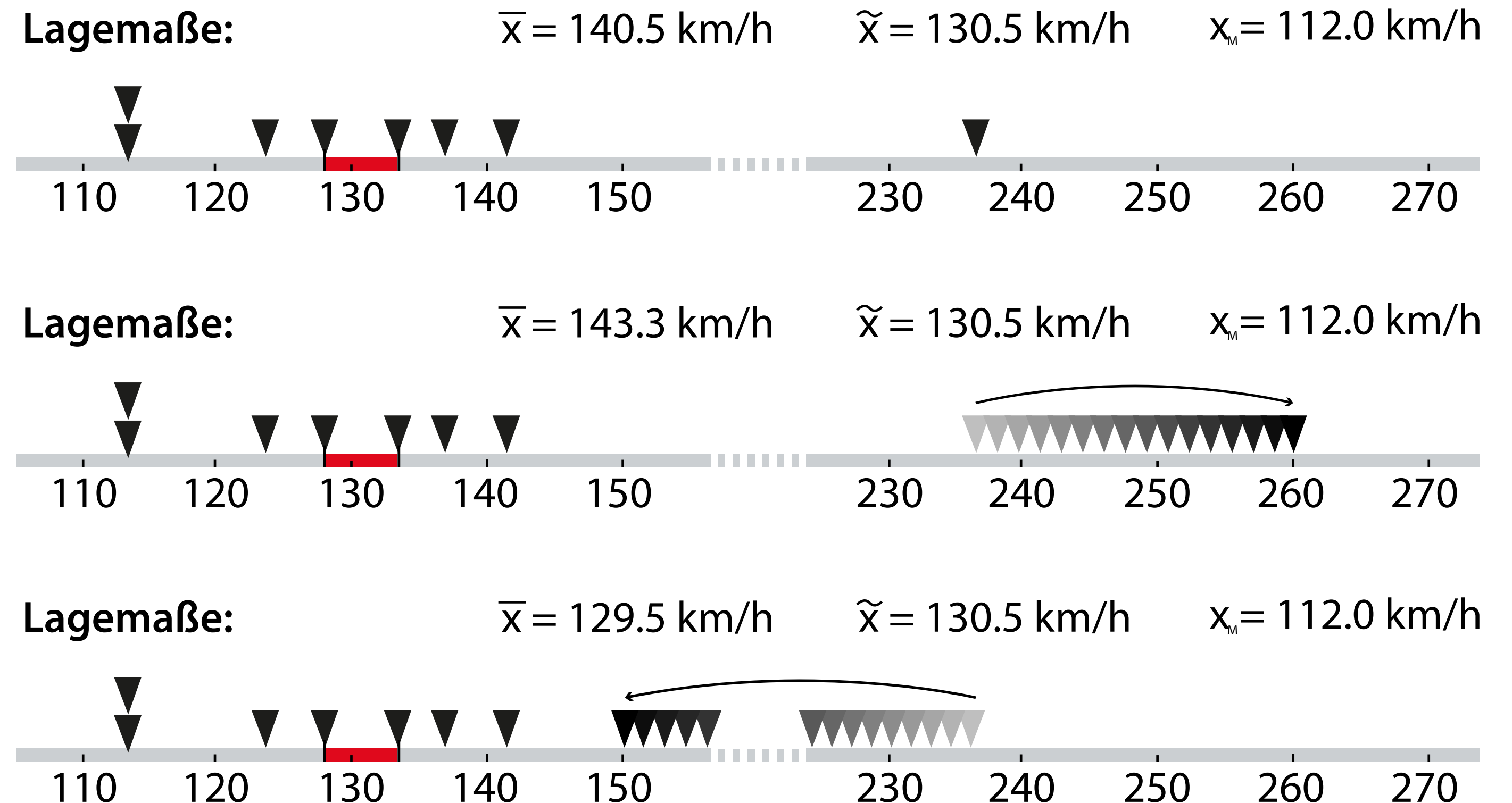


Lageparameter

Würde der Biberacher noch schneller fahren, hätte dies keine Auswirkung auf den Median.

Würde er auf 150 km/h abbremsen, hätte dies ebenfalls keine Auswirkung auf den Median.

Der Median ist robust gegenüber Ausreißern, aber unterschlägt auch deren Einfluss.



















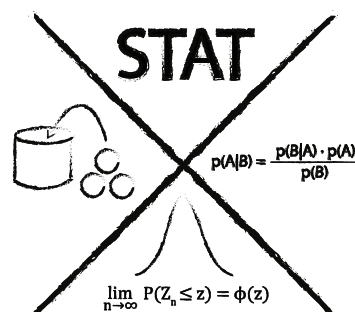
Lageparameter

Die reine Betrachtung des Mittelwertes von 140.5 km/h würde eine hohe durchschnittliche Geschwindigkeit suggerieren, obwohl nur eine Minderheit so schnell fährt.

Die reine Betrachtung des Medians von 130.5 km/h würde dagegen den Raser unter den Tisch fallen lassen. Aus Sicht des Medians ist es nur ein Fahrer der schneller als 130.5 km/h fährt.

Geschwindigkeitsmessung

Messung 1	 ES  EL 0815	123 km/h	Messung 5	 RV  XY 1234	112 km/h
Messung 2	 RV  AU 4321	112 km/h	Messung 6	 RT  BD1990	141 km/h
Messung 3	 RV  SO 2134	137 km/h	Messung 7	 BC  XX 666	238 km/h
Messung 4	 RV  SD 0741	128 km/h	Messung 8	 GP  RD 888	133 km/h

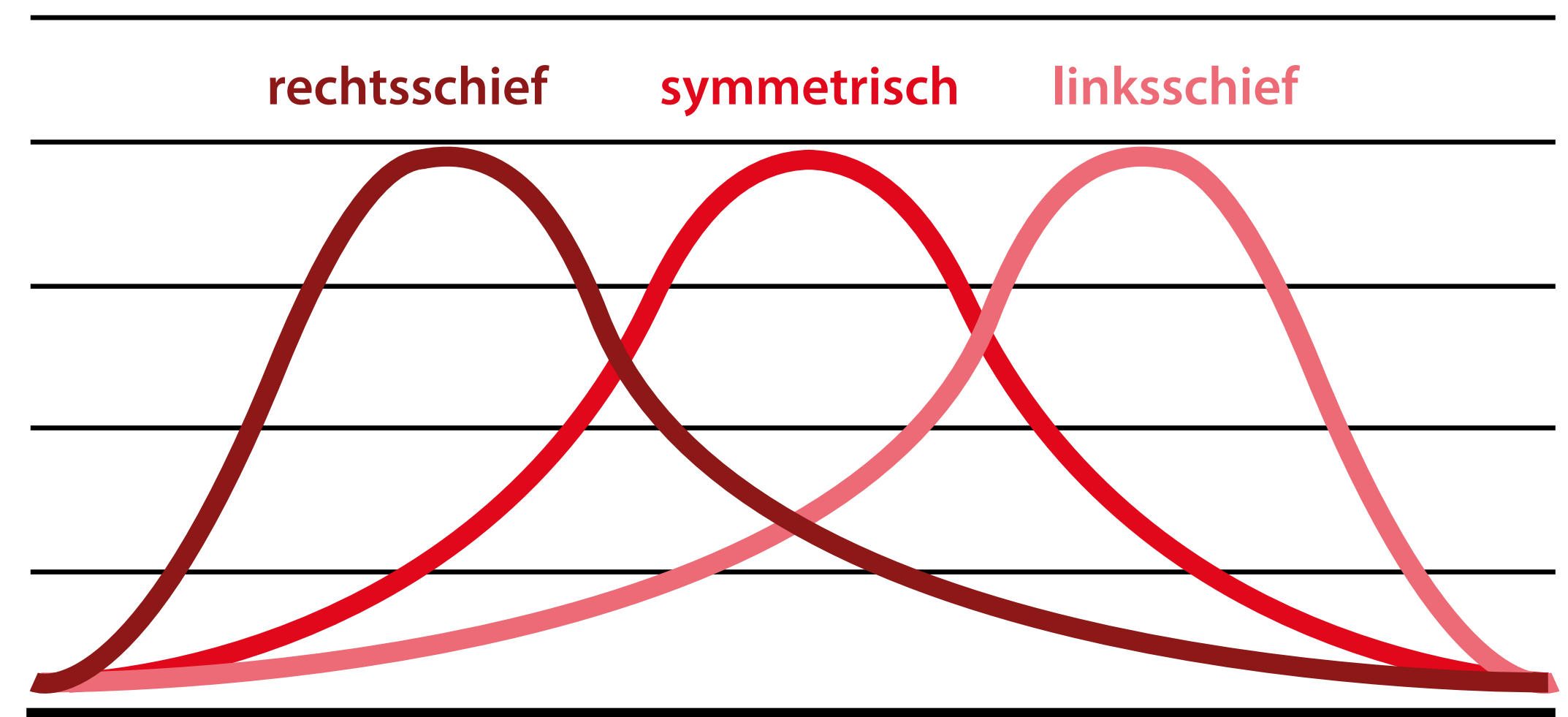


Lageparameter

Median > Mittelwert - Daten sind linksschief oder enthalten Ausreißer nach unten.

Median < Mittelwert - Daten sind rechtsschief oder enthalten Ausreißer nach oben.

Median = Mittelwert - Daten sind symmetrisch und enthalten keine Ausreißer in eine spezifische Richtung.



Lageparameter

Abhängig vom Skalenniveau kommen bestimmte Lageparameter überhaupt nicht in Frage.

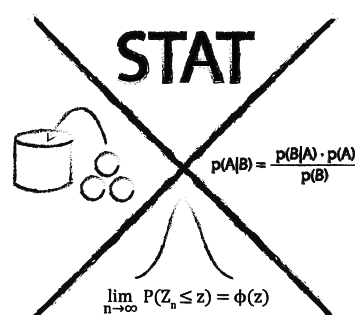
Im Beispiel haben wir Daten auf einer metrischen Skala und dürfen alle drei Lagemaße verwenden.

Bei kategorialen Daten machen dagegen weder ein Median noch ein Mittelwert Sinn.

	Modalwert	Median	Mittelwert
kategorial	Ja	-	-
ordinal	Ja	Ja	**
metrisch*	Ja	Ja	Ja

* Egal ob nur intervall- oder auch proportional skaliert

** Wird häufig trotzdem berechnet - Vorsicht bei der Interpretation!



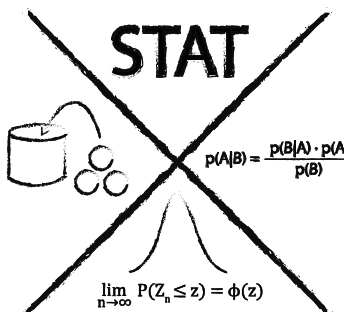
Lageparameter

Grenzfall sind ordinale Daten. Insbesondere im Zusammenhang mit Umfragen und Likert-Skalen werden diese gerne als metrisch behandelt.

Statistisch ist dies nicht 100% korrekt, aber gängig.

	Modalwert	Median	Mittelwert
kategorial	Ja	-	-
ordinal	Ja	Ja	**
metrisch*	Ja	Ja	Ja

* Egal ob nur intervall- oder auch proportional skaliert
** Wird häufig trotzdem berechnet - Vorsicht bei der Interpretation!



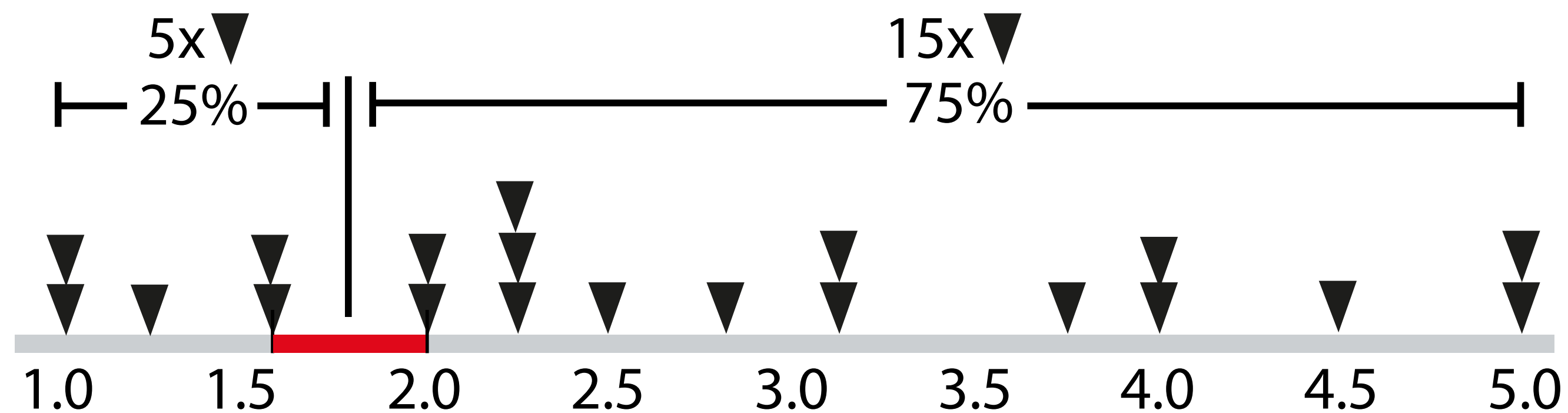
Lageparameter

Neben Mittelwert, Median und Modalwert gibt es noch die **Quantile**: eine ganze Familie von Lageparametern!

Das xx% - Quantil ist ein Wert, für den gilt: xx% der Werte sind gleich oder kleiner als dieser Wert.



Beispiel: 20 Klausurnoten
25% Quantil

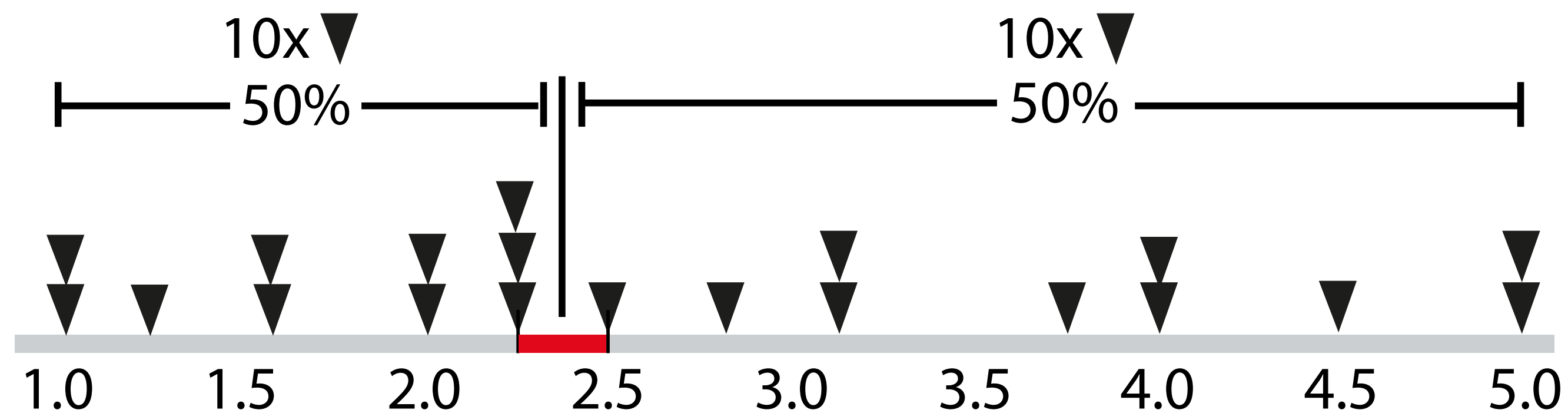


Lageparameter

Der Median ist auch ein **Quantil** - das 50% Quantil! Ein Wert für den gilt: 50% der Werte sind gleich oder kleiner als dieser Wert.



Beispiel: 20 Klausurnoten
 50% Quantil = Median



Lageparameter

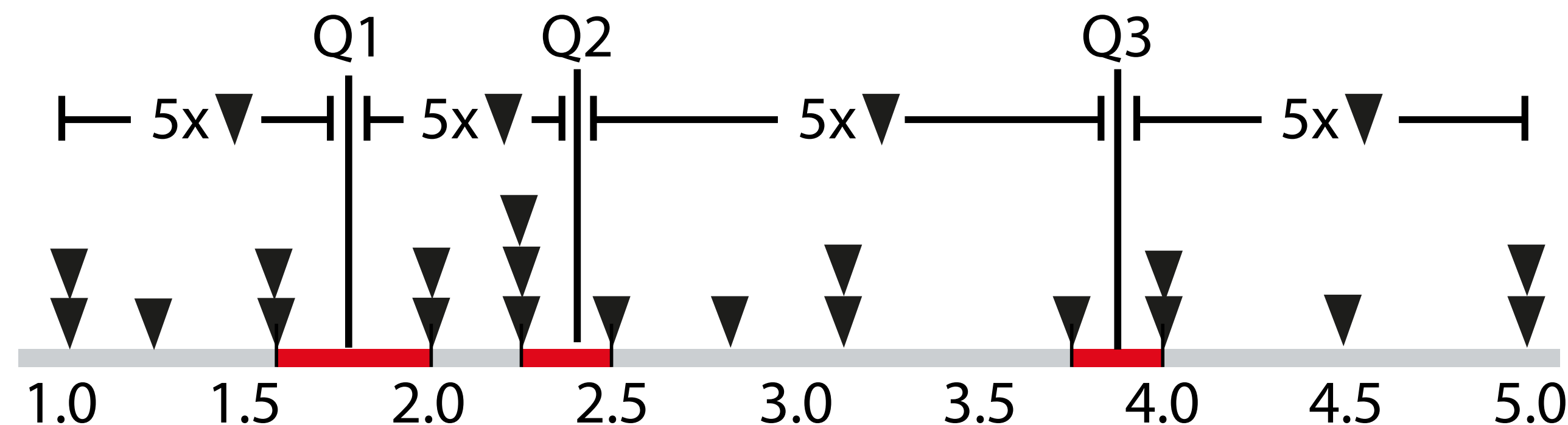
Es gibt auch **Terzile, Quartile, Dezile ...**

Der lateinische Name gibt die Anzahl Unterteilungen an. Quartile teilen die Daten in 4 Abschnitte mit einer gleich großen Anzahl von Werten (hier: $20/4=5$).



Beispiel: 20 Klausurnoten

1. Quartil = 25% besser als...
2. Quartil = 50% besser als...
3. Quartil = 75% besser als...



Lageparameter

Gib für die drei Merkmale Feinstaub, Temperatur und Luftfeuchtigkeit wenn möglich folgende Lageparameter an:

Mittelwert

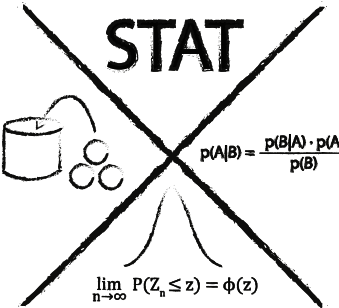
Median

Modalwert

Erstes und drittes Quartil

Tag	PM10	°C	RFM
01.02.23	13	5	78
02.02.23	9	6	82
03.02.23	7	7	85
04.02.23	14	7	82
05.02.23	16	4	94
06.02.23	25	1	85
07.02.23	35	3	69
08.02.23	37	3	65

Feinstaubdaten Stuttgart/Echterdingen. Datenquelle: Deutscher Wetterdienst (<https://cdc.dwd.de/portal/202209231028/searchview>) und Stadtklima Stuttgart (https://www.stadtklima-stuttgart.de/index.php?luft_messdaten_feinstaubwerte)



Lageparameter

Gib für die drei Merkmale Feinstaub, Temperatur und Luftfeuchtigkeit wenn möglich folgende Lageparameter an:

$$\overline{PM} = 19.5$$

$$\widetilde{PM} = 15$$

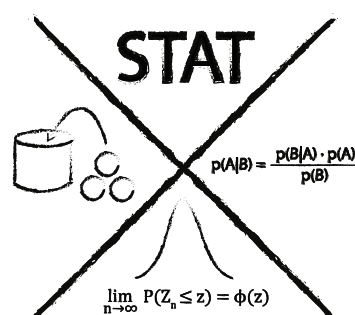
$$PM_m = -$$

1. Quartil zwischen 9 und 13

3. Quartil zwischen 25 und 35

Tag	PM10	°C	RFM
01.02.23	13	5	78
02.02.23	9	6	82
03.02.23	7	7	85
04.02.23	14	7	82
05.02.23	16	4	94
06.02.23	25	1	85
07.02.23	35	3	69
08.02.23	37	3	65

Feinstaubdaten Stuttgart/Echterdingen. Datenquelle: Deutscher Wetterdienst (<https://cdc.dwd.de/portal/202209231028/searchview>) und Stadtklima Stuttgart (https://www.stadtklima-stuttgart.de/index.php?luft_messdaten_feinstaubwerte)



Lageparameter

Gib für die drei Merkmale Feinstaub, Temperatur und Luftfeuchtigkeit wenn möglich folgende Lageparameter an:

$$\bar{T} = 4.50$$

$$\tilde{T} = 4.50$$

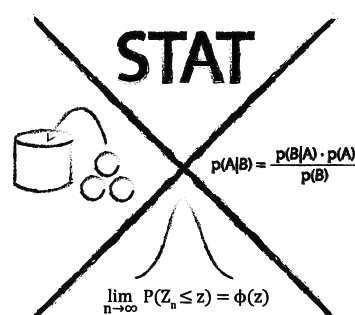
$$T_m = -$$

1. Quartil bei 3

3. Quartil zwischen 6 und 7

Tag	PM10	°C	RFM
01.02.23	13	5	78
02.02.23	9	6	82
03.02.23	7	7	85
04.02.23	14	7	82
05.02.23	16	4	94
06.02.23	25	1	85
07.02.23	35	3	69
08.02.23	37	3	65

Feinstaubdaten Stuttgart/Echterdingen. Datenquelle: Deutscher Wetterdienst (<https://cdc.dwd.de/portal/202209231028/searchview>) und Stadtklima Stuttgart (https://www.stadtklima-stuttgart.de/index.php?luft_messdaten_feinstaubwerte)



Lageparameter

Gib für die drei Merkmale Feinstaub, Temperatur und Luftfeuchtigkeit wenn möglich folgende Lageparameter an:

$$\overline{\text{RFM}} = 80.0$$

$$\widetilde{\text{RFM}} = 82.0$$

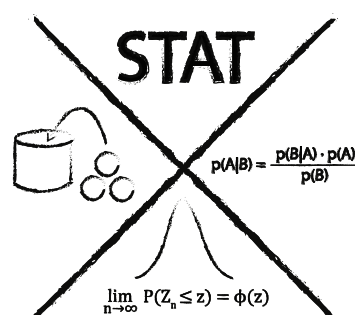
$$\text{RFM}_m = -$$

1. Quartil zwischen 69 und 78

3. Quartil bei 85

Tag	PM10	°C	RFM
01.02.23	13	5	78
02.02.23	9	6	82
03.02.23	7	7	85
04.02.23	14	7	82
05.02.23	16	4	94
06.02.23	25	1	85
07.02.23	35	3	69
08.02.23	37	3	65

Feinstaubdaten Stuttgart/Echterdingen. Datenquelle: Deutscher Wetterdienst (<https://cdc.dwd.de/portal/202209231028/searchview>) und Stadtklima Stuttgart (https://www.stadtklima-stuttgart.de/index.php?luft_messdaten_feinstaubwerte)



Streuungsparameter

Streuungsparameter geben an, wie stark die Werte um den Mittelwert streuen. Auch hier gibt es mehrere wichtige Maße:

Varianz

Standardabweichung

Lineare Abweichung

Spannweite

Preisspiegel Diesel in Ravensburg



JET Gartenstr.
170ct. / Liter



ARAL Friedr.-Str.
172ct. / Liter



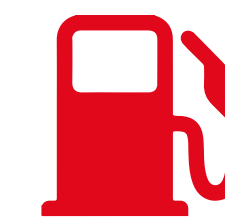
RAN Abt-Hyller
170ct. / Liter



ESSO Wangenerstr.
176ct. / Liter



Schindele
169ct. / Liter



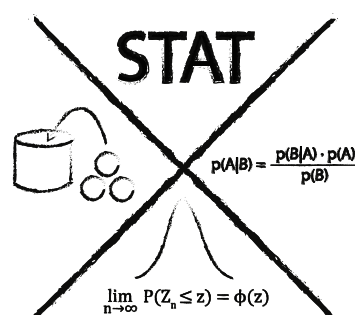
Roth Wangenerstr.
175ct. / Liter



Schell MBS
172ct. / Liter



ARAL Jahnstraße
176ct. / Liter



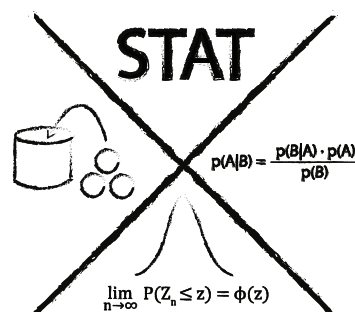
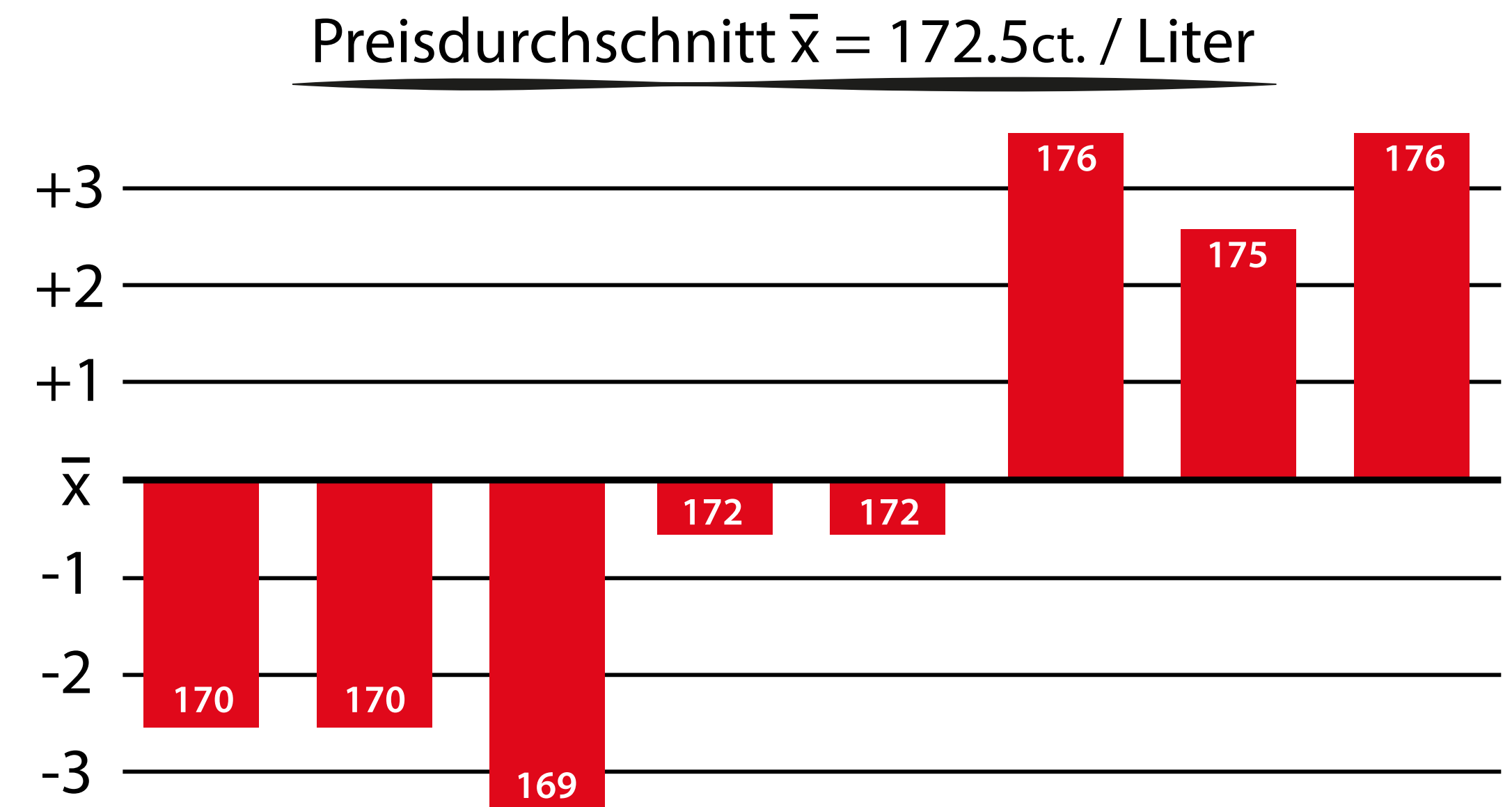
Streuungsparameter

Ansatz: Wir bestimmen die Streuung aus der Abweichung zum Mittelwert.

$$x_i - \bar{x}$$

Problem: Wenn wir einfach den Mittelwert über diese Abweichungen berechnen, mitteln sich Abweichungen nach oben und unten genau aus.

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











Streuungsparameter

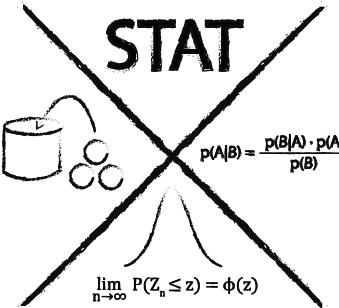
Die **Varianz** löst dieses Problem durch Quadrierung. Jede reelle Zahl wird durch Quadrierung positiv:

$$(x_i - \bar{x})^2 \geq 0$$

Die **Varianz** σ^2 ist definiert als:

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

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0











Streuungsparameter

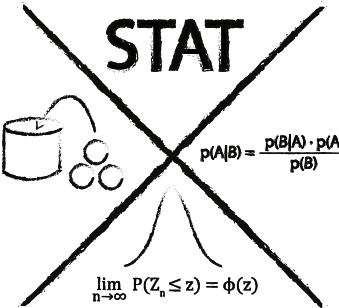
Im Beispiel erhalten wir eine Varianz von genau 8ct.

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{56}{7} = 8.00$$

Warum schreiben wir Sigma zum Quadrat?

Warum teilen wir die Summe durch n-1?

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0



Streuungsparameter









Im Beispiel erhalten wir eine Varianz von genau 8ct.

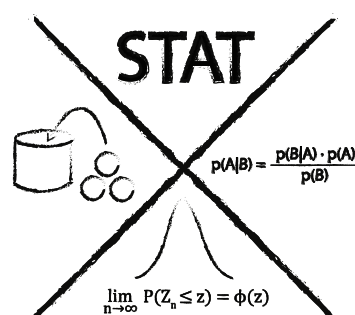
$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{56}{7} = 8.00$$

Sigma steht für die **Standardabweichung**:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 2.83$$

Die „minus eins“ im Nenner ist die s. g. Bessel-Korrektur.
Um diese zu verstehen, fehlt uns einiges an Vorwissen!

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0











Streuungsparameter

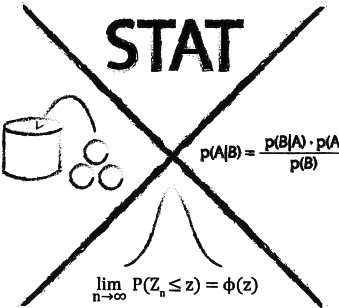
Wie interpretieren wir $\sigma^2 = 8.00$ und $\sigma=2.83$?

Je größer die Varianz/Standardabweichung umso stärker streuen die Daten um den Mittelwert.

Die Standardabweichung ist darüber hinaus **ein Maß für** den durchschnittlichen Abstand der Werte zum Mittelwert.

Bessere Interpretationsmöglichkeiten haben wir, wenn das Merkmal einer bestimmten Verteilung folgt ...

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0











Streuungsparameter

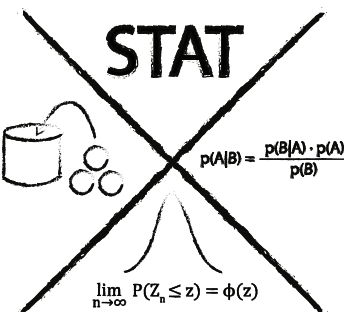
Die Standardabweichung ist nicht genau der durchschnittliche Abstand der Werte zum Mittelwert.

Durch die Summenbildung über die Quadrate werden starke Abweichungen stärker gewichtet.

Im Allgemeinen gilt:

$$\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \neq x_1 + x_2 + \dots + x_n$$

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0











Streuungsparameter

Um den durchschnittlichen Abstand der Werte zum Mittelwert zu berechnen, addieren wir die Beträge der Werte:

$$MD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{19}{8} = 2.375$$

Ähnlich wie beim Vergleich Mittelwert vs. Median ist die Standardabweichung anfälliger für Ausreißer als der MD.

	Werte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	JET Gartenstr. 170ct. / Liter	- 2.5	6.25
	RAN Abt-Hyller 170ct. / Liter	- 2.5	6.25
	Schindele 169ct. / Liter	- 3.5	12.25
	Schell MBS 172ct. / Liter	- 0.5	0.25
	ARAL Friedr.-Str. 172ct. / Liter	- 0.5	0.25
	ESSO Wangenerstr. 176ct. / Liter	+3.5	12.25
	Roth Wangenerstr. 175ct. / Liter	+2.5	6.25
	ARAL Jahnstraße 176ct. / Liter	+3.5	12.25
	Summe	0.0	56.0

Streuungsparameter

Die Spannweite R (engl.: range) einer Stichprobe gibt den Abstand zwischen dem größten und kleinsten Wert an:

$$R = \max(x_i) - \min(x_i) = 7\text{ct}$$

Statt der Spannweite wird oft auch der Wertebereich als Intervall angegeben. Hier z. B.:

$$x_i \in [169, 176]$$

Preisspiegel Diesel in Ravensburg



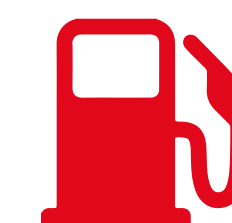
JET Gartenstr.
170ct. / Liter



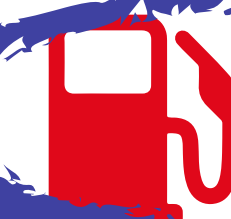
ARAL Friedr.-Str.
172ct. / Liter



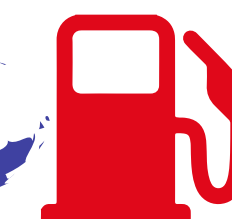
RAN Abt-Hyller
170ct. / Liter



ESSO Wangenerstr.
176ct. / Liter



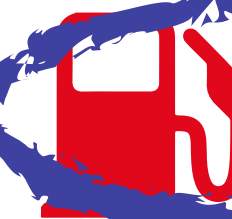
Schindele
169ct. / Liter



Roth Wangenerstr.
175ct. / Liter



Schell MBS
172ct. / Liter



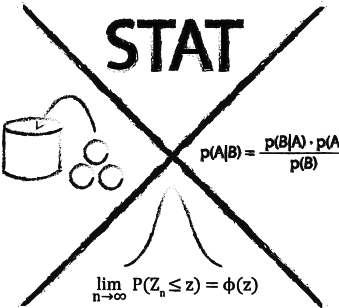
ARAL Jahnstraße
176ct. / Liter

Streuungsparameter

Berechne folgende Kennzahlen für die drei Merkmale Preis, Motorleistung und Leergewicht:

- Varianz
- Standardabweichung
- Range

Auto	Kosten (€)	Leistung (PS)	Gewicht (kg)
Polo	21000	110	1200
Golf 1.4	24000	150	1400
ID.3	37500	150	1800
Golf GTI	40000	290	1600
Up!	15000	100	1000



Streuungsparameter

Grundlage für die Varianz und die Standardabweichung ist der Mittelwert:

$$\begin{aligned}\bar{c} &= \frac{1}{5} \sum_{i=1}^5 f_i \\ &= \frac{1}{5} (21000 + 24000 + 37500 + 40000 + 15000) \\ &= 27500\end{aligned}$$

Auto	Kosten (€)	Leistung (PS)	Gewicht (kg)
Polo	21000	110	1200
Golf 1.4	24000	150	1400
ID.3	37500	150	1800
Golf GTI	40000	290	1600
Up!	15000	100	1000

Streuungsparameter

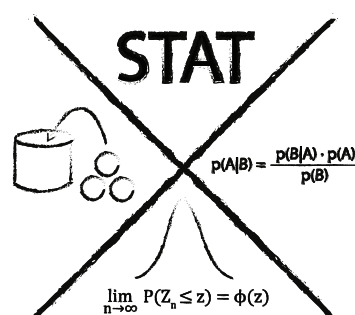
Mittelwerte der drei Merkmale:

$$\bar{c} = 27500 \quad \bar{p} = 160 \quad \bar{m} = 1400$$

Damit können wir die Varianzen σ^2 berechnen:

$$\begin{aligned} \sigma_c^2 &= \frac{1}{4} \sum_{i=1}^5 (c_i - \bar{c})^2 \\ &= \frac{1}{4} [(21000 - 27500)^2 + (24000 - 27500)^2 + \dots] \\ &= 116.750.000 \end{aligned}$$

Auto	Kosten (€)	Leistung (PS)	Gewicht (kg)
Polo	21000	110	1200
Golf 1.4	24000	150	1400
ID.3	37500	150	1800
Golf GTI	40000	290	1600
Up!	15000	100	1000



Streuungsparameter

Mittelwerte der drei Merkmale:

$$\bar{c} = 27500 \quad \bar{p} = 160 \quad \bar{m} = 1400$$

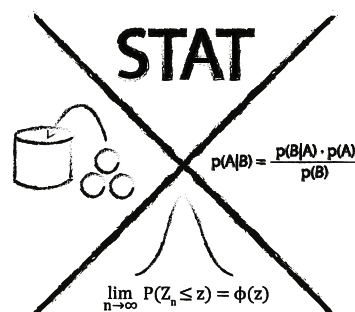
Damit können wir die Varianzen σ^2 berechnen:

$$\sigma_c^2 = 116750000 \iff \sigma_c = 10805$$

$$\sigma_p^2 = 5800 \iff \sigma_p = 76.2$$

$$\sigma_m^2 = 100000 \iff \sigma_m = 316.2$$

Auto	Kosten (€)	Leistung (PS)	Gewicht (kg)
Polo	21000	110	1200
Golf 1.4	24000	150	1400
ID.3	37500	150	1800
Golf GTI	40000	290	1600
Up!	15000	100	1000



Streuungsparameter

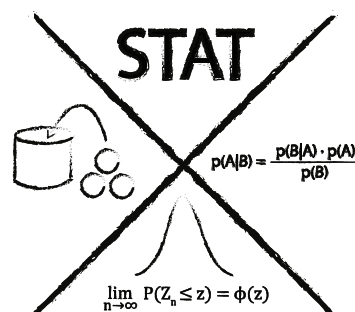
Die Spannweiten berechnen sich deutlich einfacher:

$$\begin{aligned} R_c &= \max(c_i) - \min(c_i) \\ &= 40000 - 15000 \\ &= 25000 \end{aligned}$$

$$R_p = 290 - 100 = 190$$

$$\begin{aligned} R_m &= 1800 - 1000 \\ &= 800 \end{aligned}$$

Auto	Kosten (€)	Leistung (PS)	Gewicht (kg)
Polo	21000	110	1200
Golf 1.4	24000	150	1400
ID.3	37500	150	1800
Golf GTI	40000	290	1600
Up!	15000	100	1000



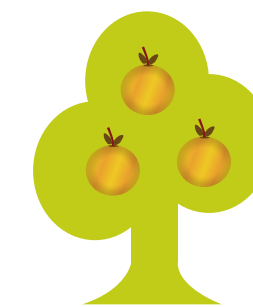
Kovarianz & Korrelation

Die bisherigen Kennzahlen beziehen sich immer auf ein Merkmal. Jetzt lernen wir auch Kennzahlen kennen, die zwei Merkmale verknüpfen.

Zentrale Frage: Hängen zwei Merkmale voneinander ab?

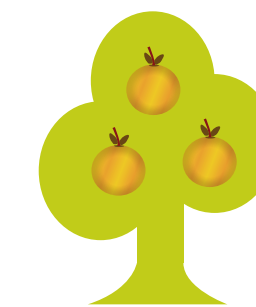
Beispiel: Wir erheben eine Stichprobe von 6 Apfelbäumen mit den Merkmalen Kronhöhe x und Ertrag y .

Erträge von Apfelbäumen



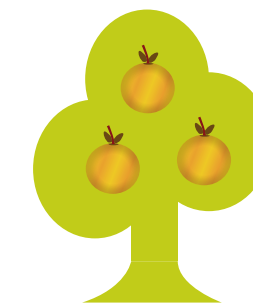
$$x_1 = 5.83 \text{ Meter}$$

$$y_1 = 317 \text{ kg}$$



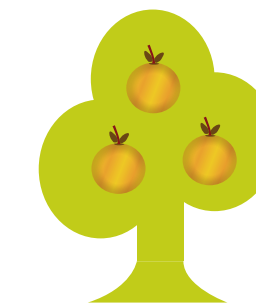
$$x_4 = 6.05 \text{ Meter}$$

$$y_4 = 345 \text{ kg}$$



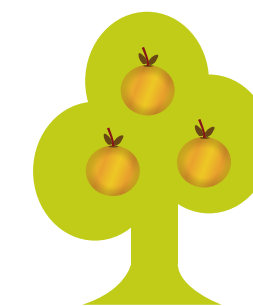
$$x_2 = 4.73 \text{ Meter}$$

$$y_2 = 245 \text{ kg}$$



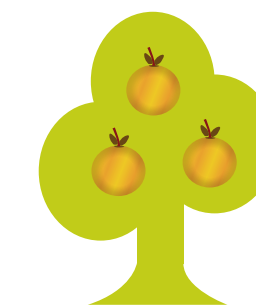
$$x_5 = 4.19 \text{ Meter}$$

$$y_5 = 190 \text{ kg}$$



$$x_3 = 6.10 \text{ Meter}$$

$$y_3 = 298 \text{ kg}$$



$$x_6 = 4.90 \text{ Meter}$$

$$y_6 = 195 \text{ kg}$$

Kovarianz & Korrelation

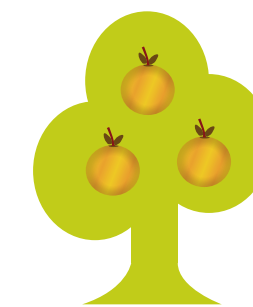
Die Kovarianz zwischen zwei Merkmalen ist definiert als:

$$q_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

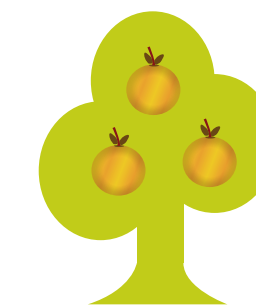
Wir multiplizieren die Abweichung vom Mittelwert bei Merkmal x mit derer bei Merkmal y.

Warum machen wir das? Wie kommt man darauf?

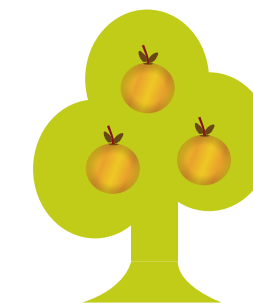
Erträge von Apfelbäumen



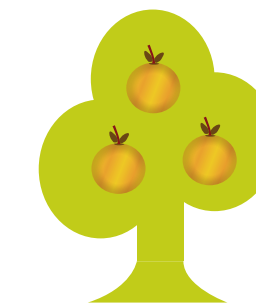
$x_1 = 5.83$ Meter
 $y_1 = 317$ kg



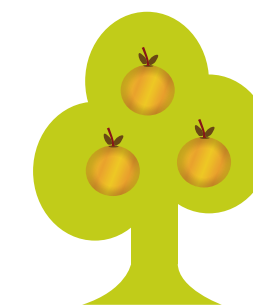
$x_4 = 6.05$ Meter
 $y_4 = 345$ kg



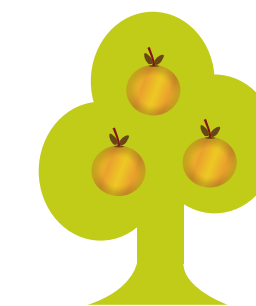
$x_2 = 4.73$ Meter
 $y_2 = 245$ kg



$x_5 = 4.19$ Meter
 $y_5 = 190$ kg



$x_3 = 6.10$ Meter
 $y_3 = 298$ kg



$x_6 = 4.90$ Meter
 $y_6 = 195$ kg

Kovarianz & Korrelation

Betrachten wir dazu zwei stilisierte Merkmale mit Mittelwerten von jeweils 2.5

Fallen überdurchschnittlich hohe Werte des einen Merkmals mit überdurchschnittlich hohen Werten des anderen zusammen ...

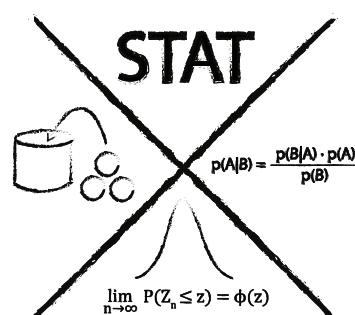
...ist der Wert positiv.

Merkmals x	Merkmals y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	1	-1.5	-1.5	2.25
2	2	-0.5	-0.5	0.25
3	3	+0.5	+0.5	0.25
4	4	+1.5	+1.5	2.25
SUMME				5.00

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i = 2.5$$

$$\bar{y} = \frac{1}{4} \sum_{i=1}^4 y_i = 2.5$$

$$q_{x,y} = \frac{1}{4-1} \sum_{i=1}^4 (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{3} 5 = 1.67$$



Kovarianz & Korrelation

Betrachten wir dazu zwei stilisierte Merkmale mit Mittelwerten von jeweils 2.5

Fallen überdurchschnittlich hohe Werte des einen Merkmals mit überdurchschnittlich niedrigen Werten des anderen zusammen

...

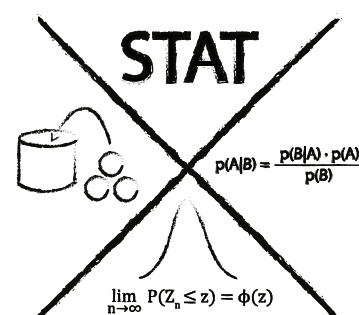
...ist der Wert negativ.

Merkmal x	Merkmal y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	4	-1.5	+1.5	-2.25
2	3	-0.5	+0.5	-0.25
3	2	+0.5	- 0.5	-0.25
4	1	+1.5	- 1.5	-2.25
SUMME				-5.00

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i = 2.5$$

$$\bar{y} = \frac{1}{4} \sum_{i=1}^4 y_i = 2.5$$

$$q_{x,y} = \frac{1}{4-1} \sum_{i=1}^4 (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{3} (-5) = -1.67$$



Kovarianz & Korrelation

Betrachten wir dazu zwei stilisierte Merkmale mit Mittelwerten von jeweils 2.5

Gibt es kein Muster zwischen den Abweichungen vom Mittelwert bei den beiden Merkmalen ...

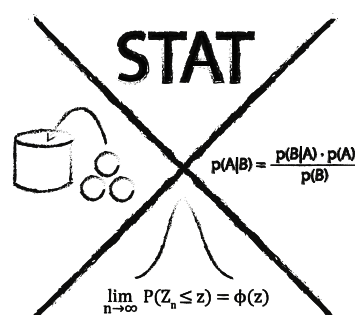
...ist der Wert in der Nähe von 0

Merkmal x	Merkmal y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	3	-1.5	+0.5	-0.75
2	1	-0.5	-1.5	+0.75
3	4	+0.5	+1.5	+0.75
4	2	+1.5	-0.5	-0.75
SUMME				0.00

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i = 2.5$$

$$\bar{y} = \frac{1}{4} \sum_{i=1}^4 y_i = 2.5$$

$$q_{x,y} = \frac{1}{4-1} \sum_{i=1}^4 (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{3} \cdot 0 = 0$$



Kovarianz & Korrelation

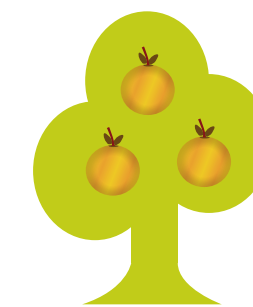
Die Kovarianz zwischen zwei Merkmalen ist definiert als:

$$q_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Oft ist diese Kovarianz nur ein Zwischenergebnis zur Berechnung des Bravais-Pearson-Korrelationskoeffizienten:

$$\rho_{x,y} = \frac{q_{x,y}}{\sigma_x \cdot \sigma_y}$$

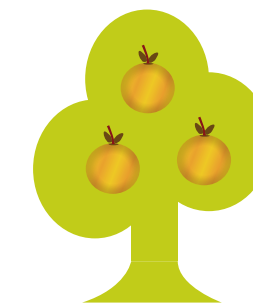
Erträge von Apfelbäumen



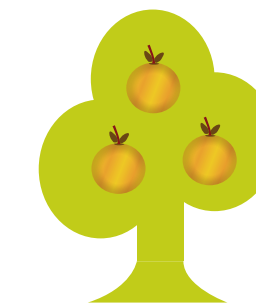
$x_1 = 5.83$ Meter
 $y_1 = 317$ kg



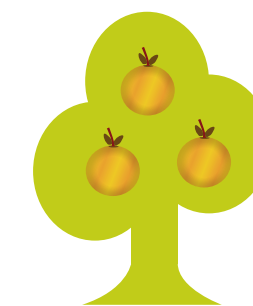
$x_4 = 6.05$ Meter
 $y_4 = 345$ kg



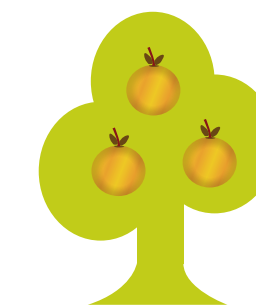
$x_2 = 4.73$ Meter
 $y_2 = 245$ kg



$x_5 = 4.19$ Meter
 $y_5 = 190$ kg



$x_3 = 6.10$ Meter
 $y_3 = 298$ kg



$x_6 = 4.90$ Meter
 $y_6 = 195$ kg

Kovarianz & Korrelation

Um diesen Korrelationskoeffizienten in unserem Baumbeispiel zu berechnen, benötigen wir ...

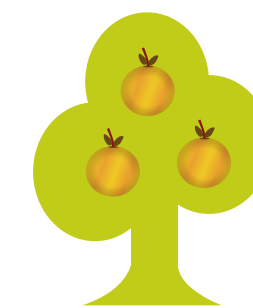
... die Mittelwerte von x und y

... die Standardabweichungen von x und y

... die Kovarianz von x und y

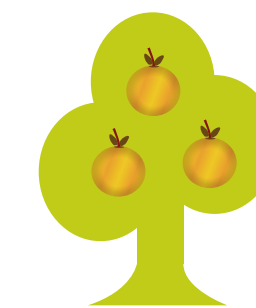
Selbst bei einem kleinen Datensatz ist das eine Menge Schreib- und Tipparbeit!

Erträge von Apfelbäumen



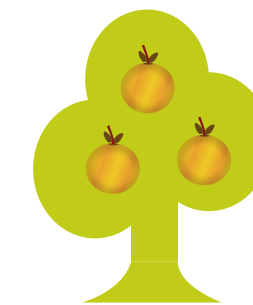
$$x_1 = 5.83 \text{ Meter}$$

$$y_1 = 317 \text{ kg}$$



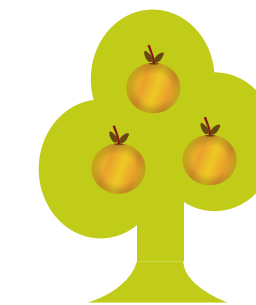
$$x_4 = 6.05 \text{ Meter}$$

$$y_4 = 345 \text{ kg}$$



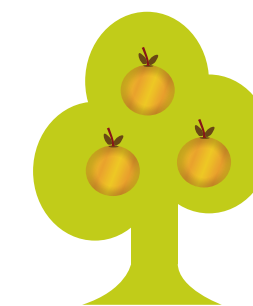
$$x_2 = 4.73 \text{ Meter}$$

$$y_2 = 245 \text{ kg}$$



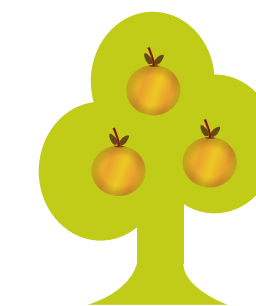
$$x_5 = 4.19 \text{ Meter}$$

$$y_5 = 190 \text{ kg}$$



$$x_3 = 6.10 \text{ Meter}$$

$$y_3 = 298 \text{ kg}$$



$$x_6 = 4.90 \text{ Meter}$$

$$y_6 = 195 \text{ kg}$$

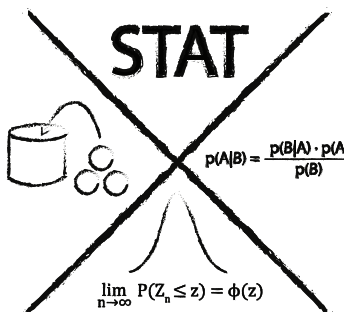
Kovarianz & Korrelation

Wir überschreiten an dieser Stelle eine Grenze. Auch wenn die Rechnung mit Stift, Papier und Taschenrechner möglich wäre, wäre sie vor allem eines: zeitaufwendig!

Wir wollen uns jetzt zunächst allgemein Gedanken über die Interpretation des Korrelationskoeffizienten machen.

Danach wenden wir uns Excel zu ...

Baum	Kronhöhe (m)	Ertrag (kg)
1	5,83	317,0
2	4,73	245,0
3	6,10	298,0
4	6,05	345,0
5	4,19	190,0
6	4,90	195,0
Mittelwert	5,30	265,00
Median	5,37	271,50
Stabw.	0,80	64,99
Kovarianz	47,322	
Korrelation	0,910	



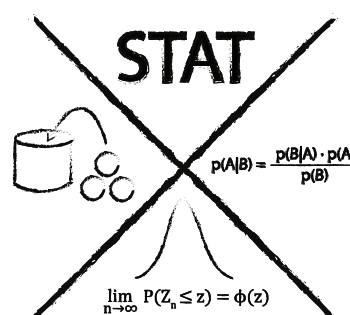
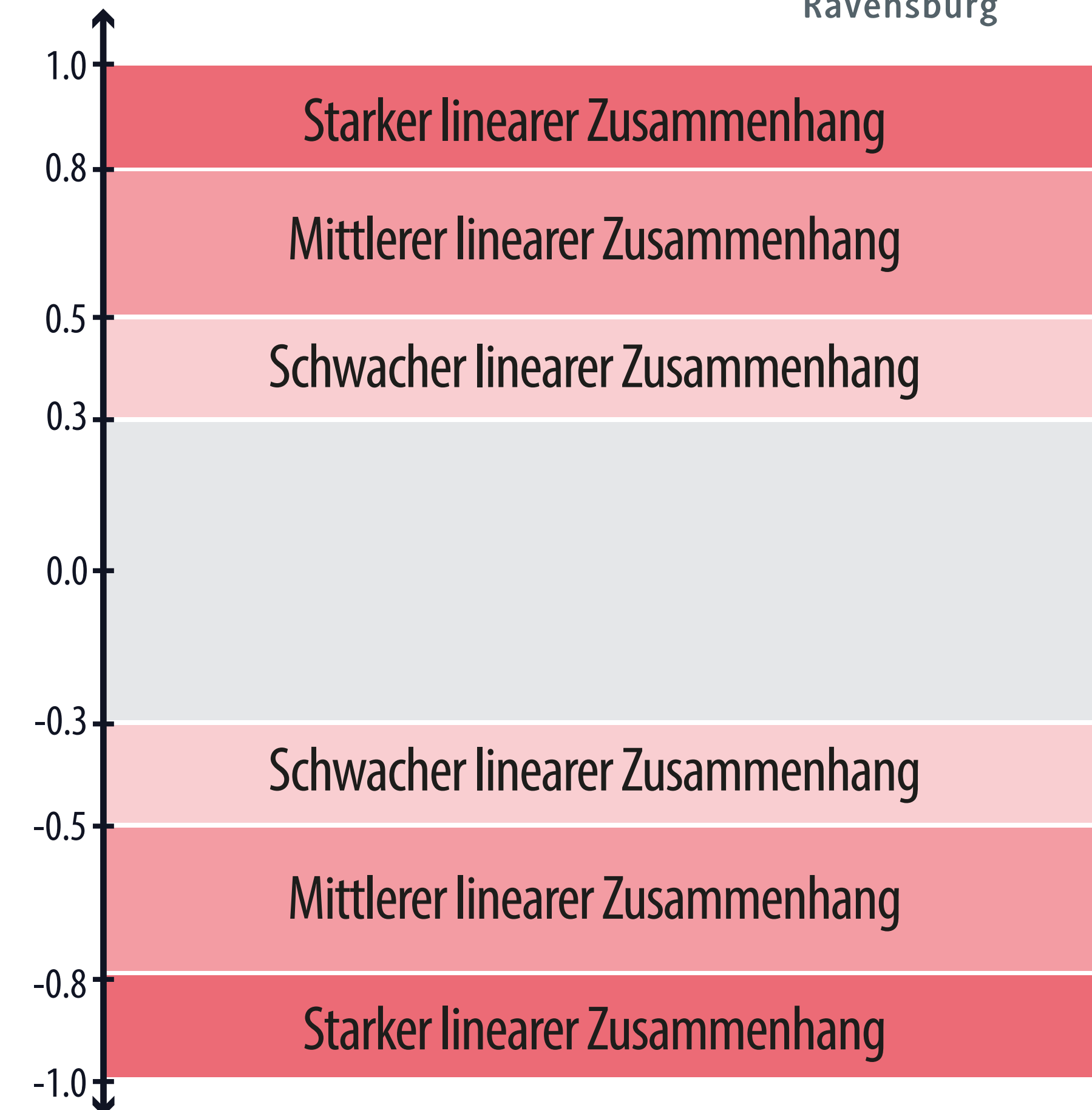
Kovarianz & Korrelation

Der Korrelationskoeffizient ρ kann Werte zwischen minus und plus 1 annehmen:

$$\rho \in [-1,1]$$

In Lehrbüchern finden sich diverse Klassifizierungsschemas und Faustregeln, ab welchen Werten der Zusammenhang als schwach oder stark gilt.

Wir wollen uns hier aber eher auf die generelle Bedeutung der Werte konzentrieren.

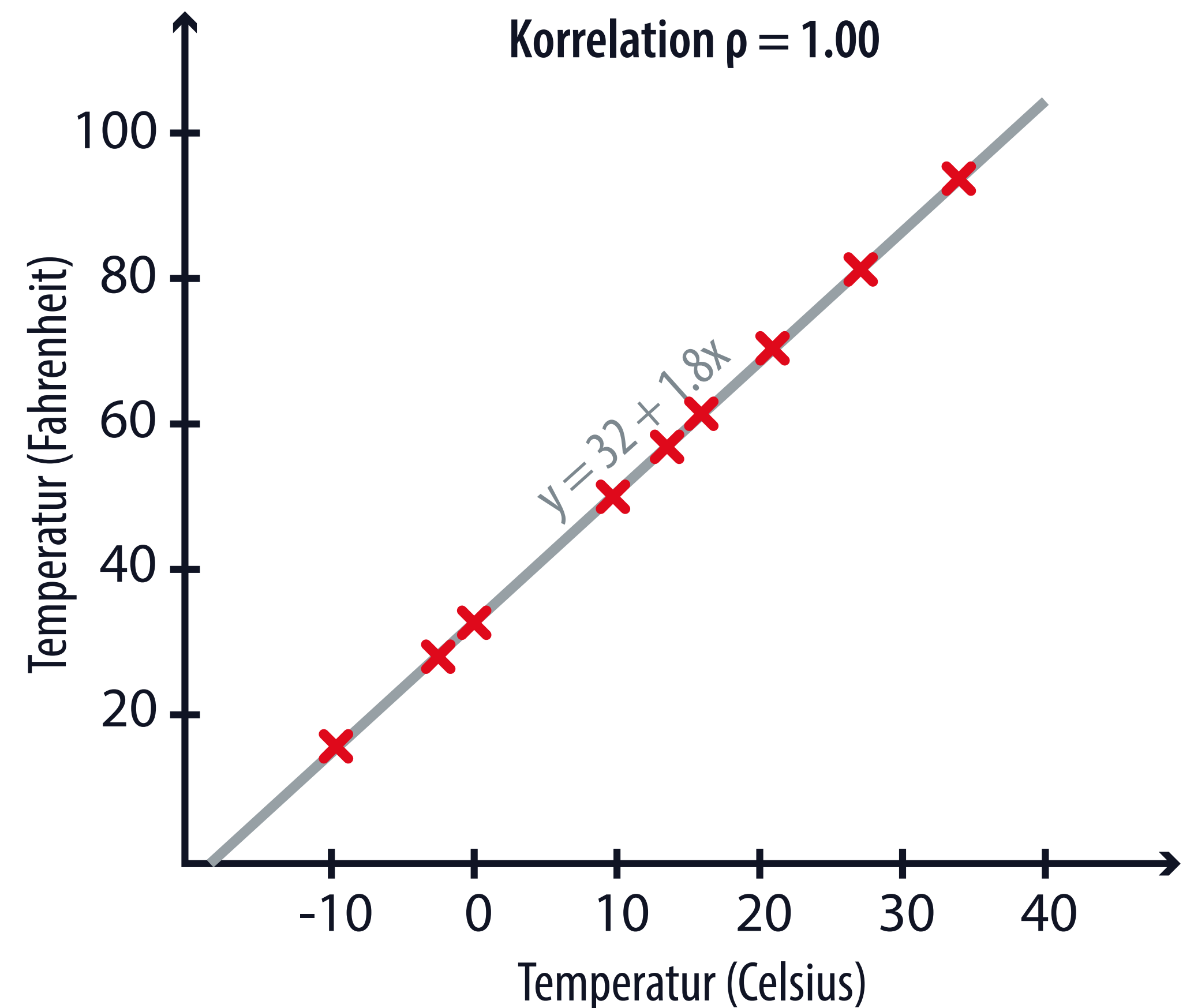


Kovarianz & Korrelation

Ein Wert von $\rho = 1.00$ bedeutet lineare Abhängigkeit bzw. perfekte Korrelation.

Dies bedeutet, dass wir den Zusammenhang der beiden Merkmale durch ein lineares Modell beschreiben können. Der eine Wert ergibt sich zwingend aus dem anderen!

$$y_i = \beta_0 + \beta_1 x_i \text{ mit } \beta_1 > 0$$

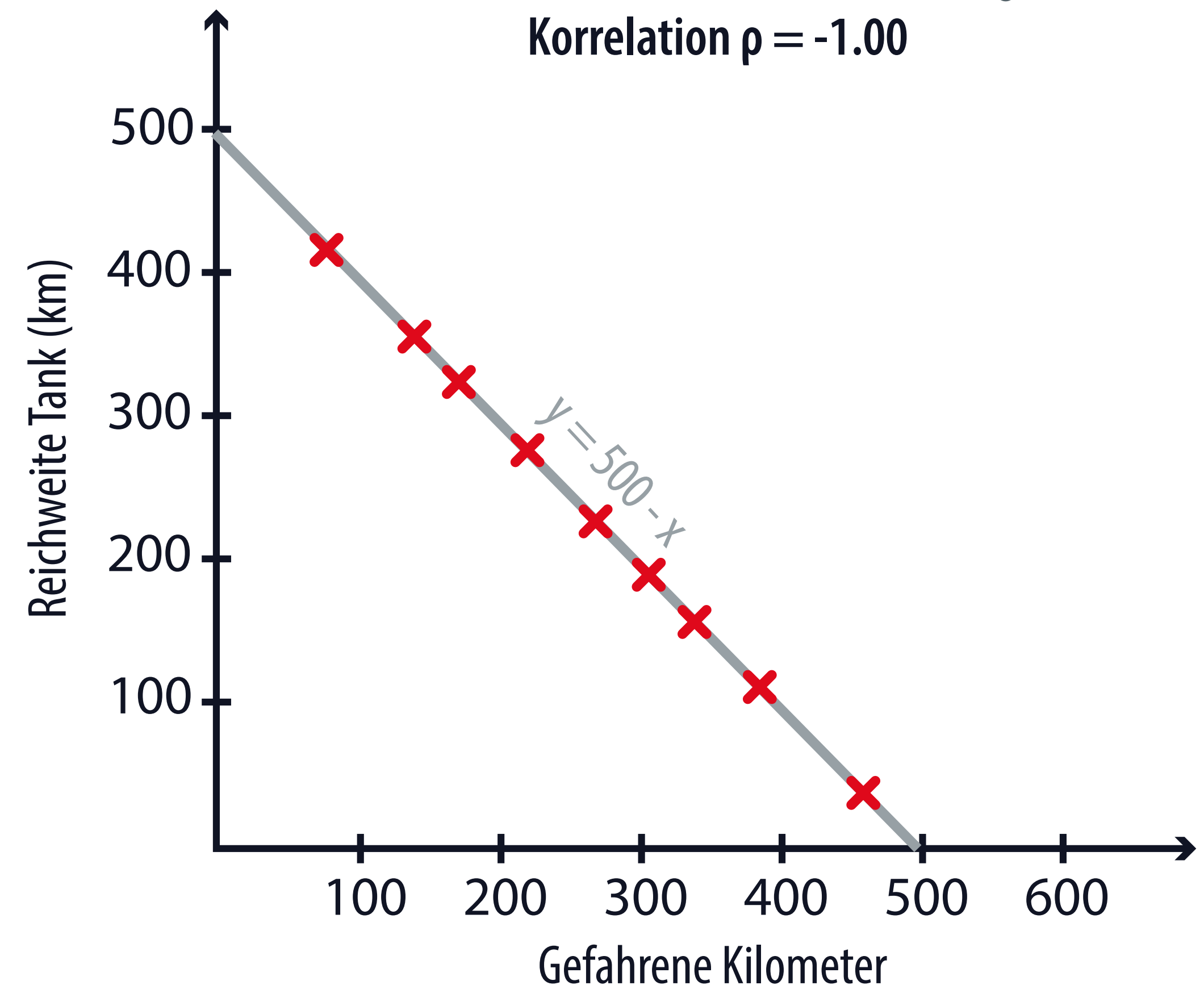


Kovarianz & Korrelation

Ein Wert $\rho = -1.00$ bedeutet ebenfalls lineare Abhängigkeit bzw. perfekte Korrelation.

Wieder können wir den Zusammenhang der Merkmale mit einem linearen Modell beschreiben. Der Faktor ist jedoch negativ!

$$y_i = \beta_0 + \beta_1 x_i \text{ mit } \beta_1 < 0$$

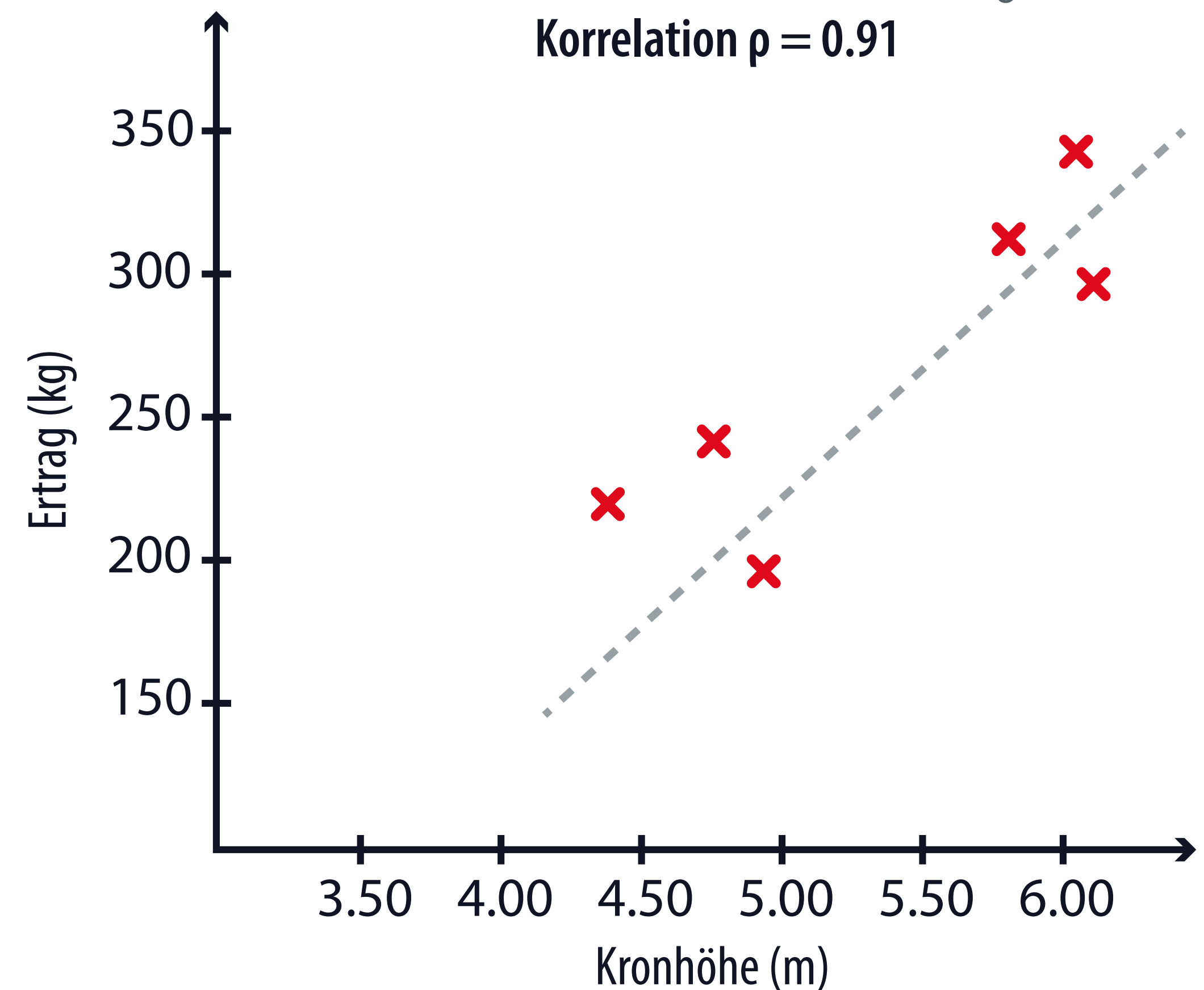


Kovarianz & Korrelation

Ein Wert $\rho \in (0,1)$ bedeutet positive Korrelation.

Ist eines der Merkmale größer als sein Durchschnitt, ist das andere tendenziell auch größer als sein Durchschnitt.

Je näher das ρ an der 1 liegt, umso stärker ist der Zusammenhang. Wir erinnern uns an die Tabelle auf Folie 77.

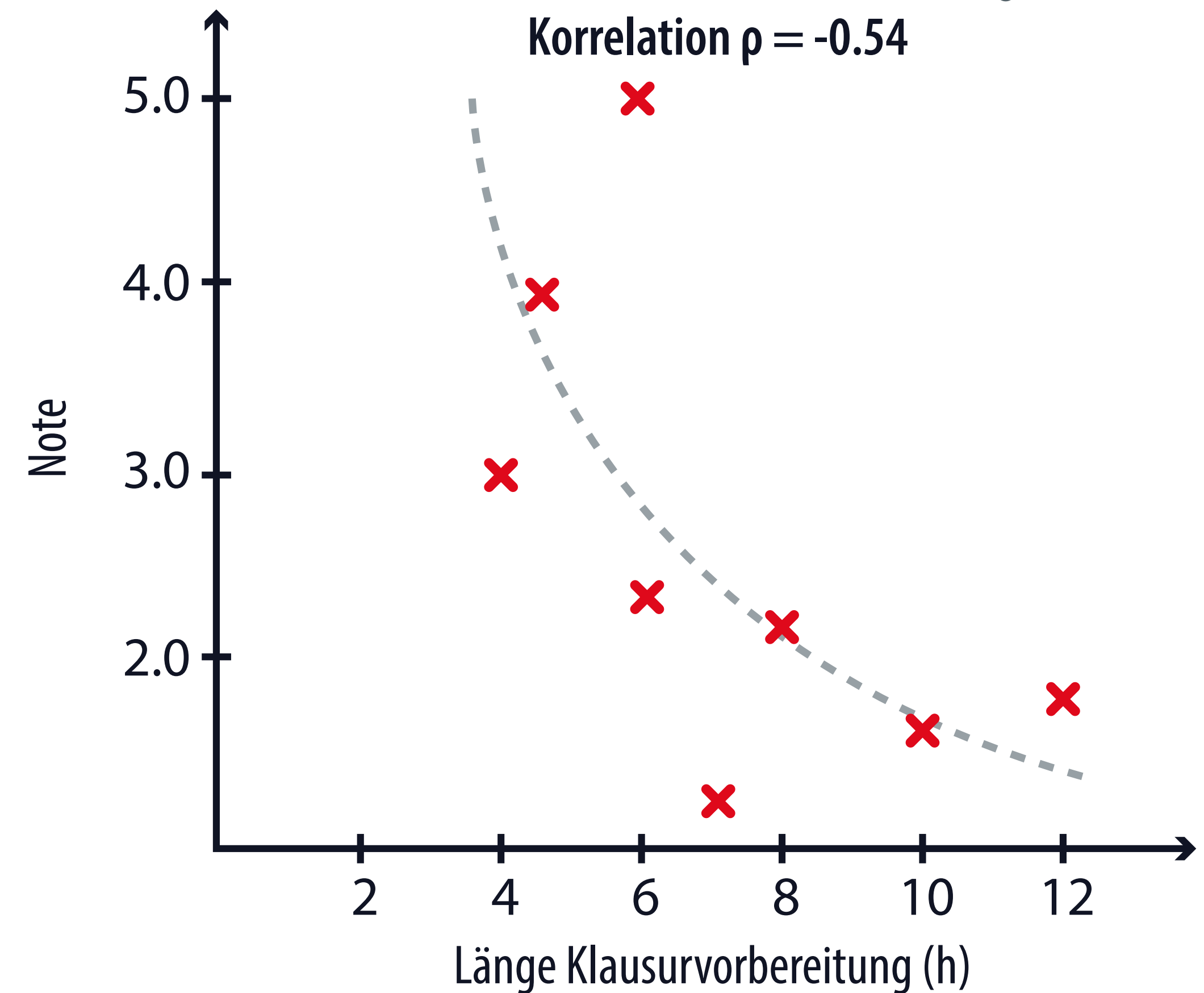


Kovarianz & Korrelation

Ein Wert $\rho \in (-1,0)$ bedeutet negative Korrelation.

Ist eines der Merkmale größer als sein Durchschnitt, ist das andere tendenziell kleiner als sein Durchschnitt.

Je näher das ρ an der -1 liegt, umso stärker ist der Zusammenhang. Wir erinnern uns an die Tabelle auf Folie 77.

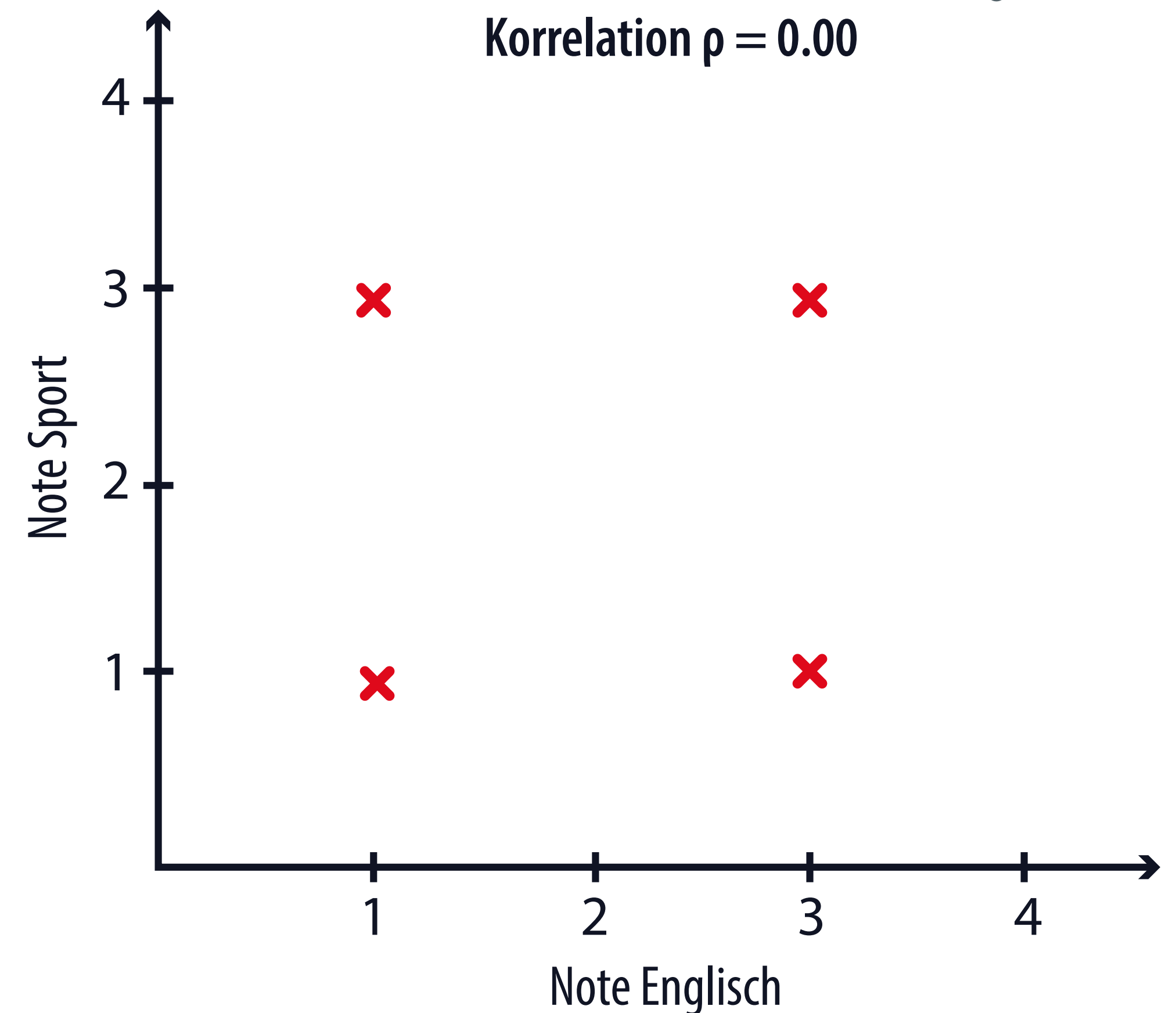


Kovarianz & Korrelation

Ein Wert von $\rho=0.00$ bedeutet keine Korrelation zwischen den beiden Merkmalen.

Es ist allerdings unwahrscheinlich diesen Wert exakt zu treffen. Durch reinen Zufall ergeben sich Werte von ρ knapp über oder unter 0.

Aus diesem Grund würden wir auch Werte wie $\rho=0.0172$ oder $\rho=-0.0218$ als keine oder allenfalls sehr schwache Korrelation bezeichnen.

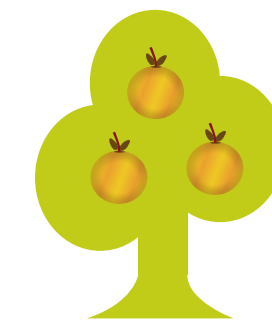


Kovarianz & Korrelation

Die Merkmale Kronhöhe und Ertrag sind mit $\rho = 0.91$ stark positiv korreliert!

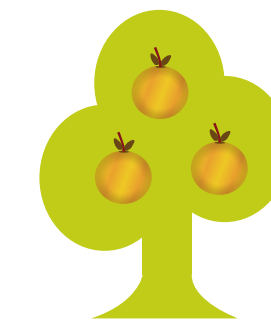
Ein überdurchschnittlich hoher Baum gibt tendenziell eine überdurchschnittlich große Ernte.

Erträge von Apfelbäumen



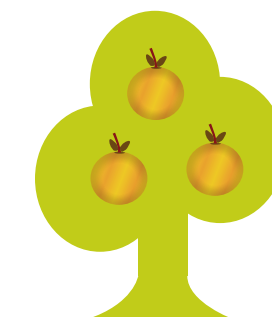
$$x_1 = 5.83 \text{ Meter}$$

$$y_1 = 317 \text{ kg}$$



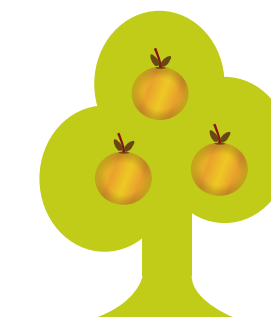
$$x_4 = 6.05 \text{ Meter}$$

$$y_4 = 345 \text{ kg}$$



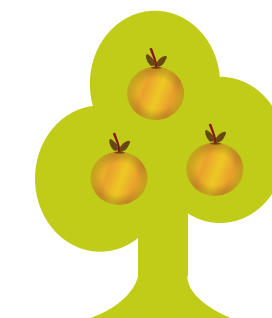
$$x_2 = 4.73 \text{ Meter}$$

$$y_2 = 245 \text{ kg}$$



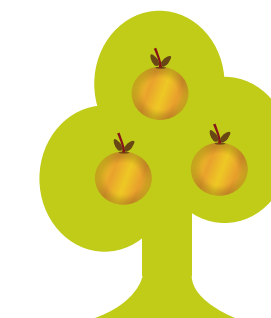
$$x_5 = 4.19 \text{ Meter}$$

$$y_5 = 190 \text{ kg}$$



$$x_3 = 6.10 \text{ Meter}$$

$$y_3 = 298 \text{ kg}$$



$$x_6 = 4.90 \text{ Meter}$$

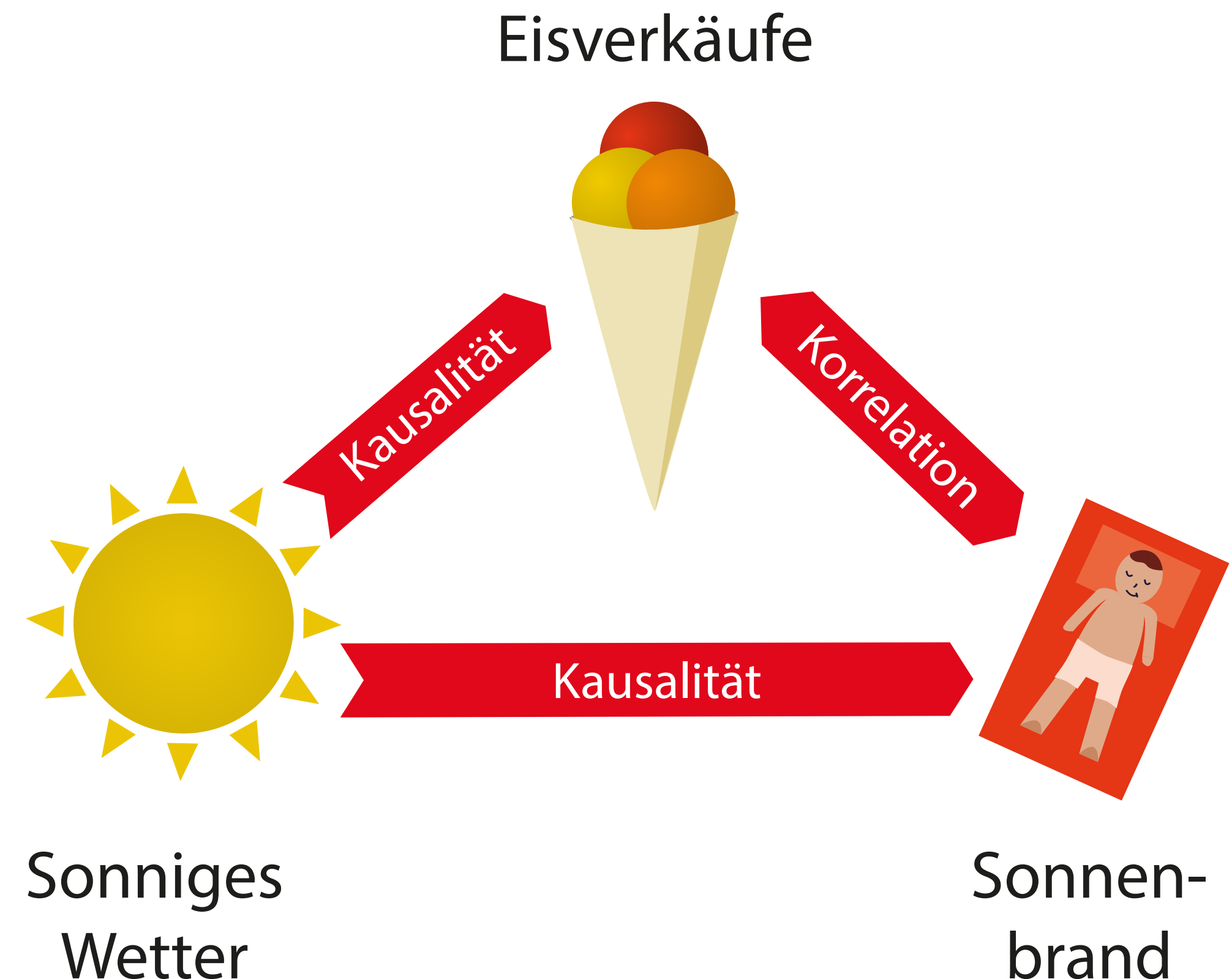
$$y_6 = 195 \text{ kg}$$

Kovarianz & Korrelation

Vorsicht Aus Korrelation folgt nicht zwingend Kausalität!

Zwei Datenreihen können einen Korrelationskoeffizienten von fast 1 haben, obwohl es keine Ursache-Wirkung Beziehung zwischen ihnen gibt.

Umgekehrt kann es trotz eines Korrelationskoeffizienten von nahezu $\rho=0.00$ eine Ursache-Wirkung Beziehungen zwischen zwei Merkmalen geben.



Kovarianz & Korrelation

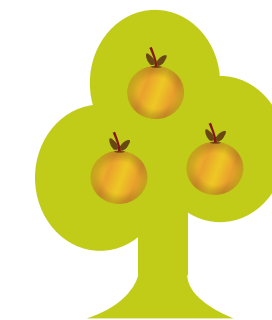
Die Merkmale Kronhöhe und Ertrag sind mit $\rho = 0.91$ stark positiv korreliert!

Möglichkeit 1 - die Bäume tragen viele Äpfel, weil sie groß sind.

Möglichkeit 2 - die Bäume sind groß, weil sie viele Äpfel tragen.

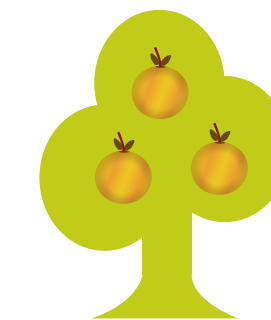
Möglichkeit 3 - es gibt dritte Faktoren, die sowohl die Kronhöhe als auch die Ertragsstärke fördern.

Erträge von Apfelbäumen



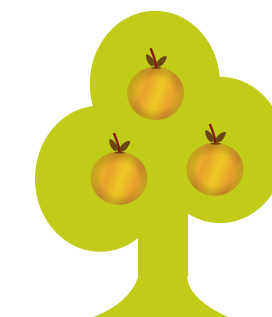
$$x_1 = 5.83 \text{ Meter}$$

$$y_1 = 317 \text{ kg}$$



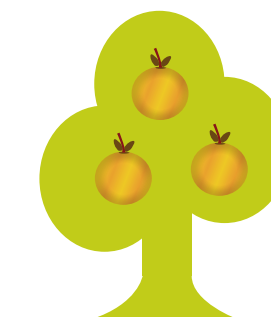
$$x_4 = 6.05 \text{ Meter}$$

$$y_4 = 345 \text{ kg}$$



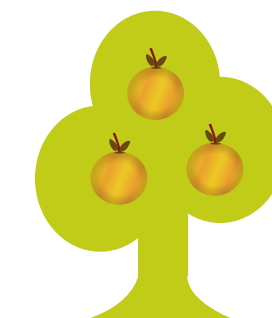
$$x_2 = 4.73 \text{ Meter}$$

$$y_2 = 245 \text{ kg}$$



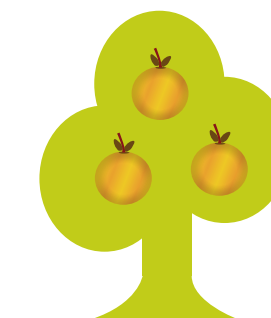
$$x_5 = 4.19 \text{ Meter}$$

$$y_5 = 190 \text{ kg}$$



$$x_3 = 6.10 \text{ Meter}$$

$$y_3 = 298 \text{ kg}$$



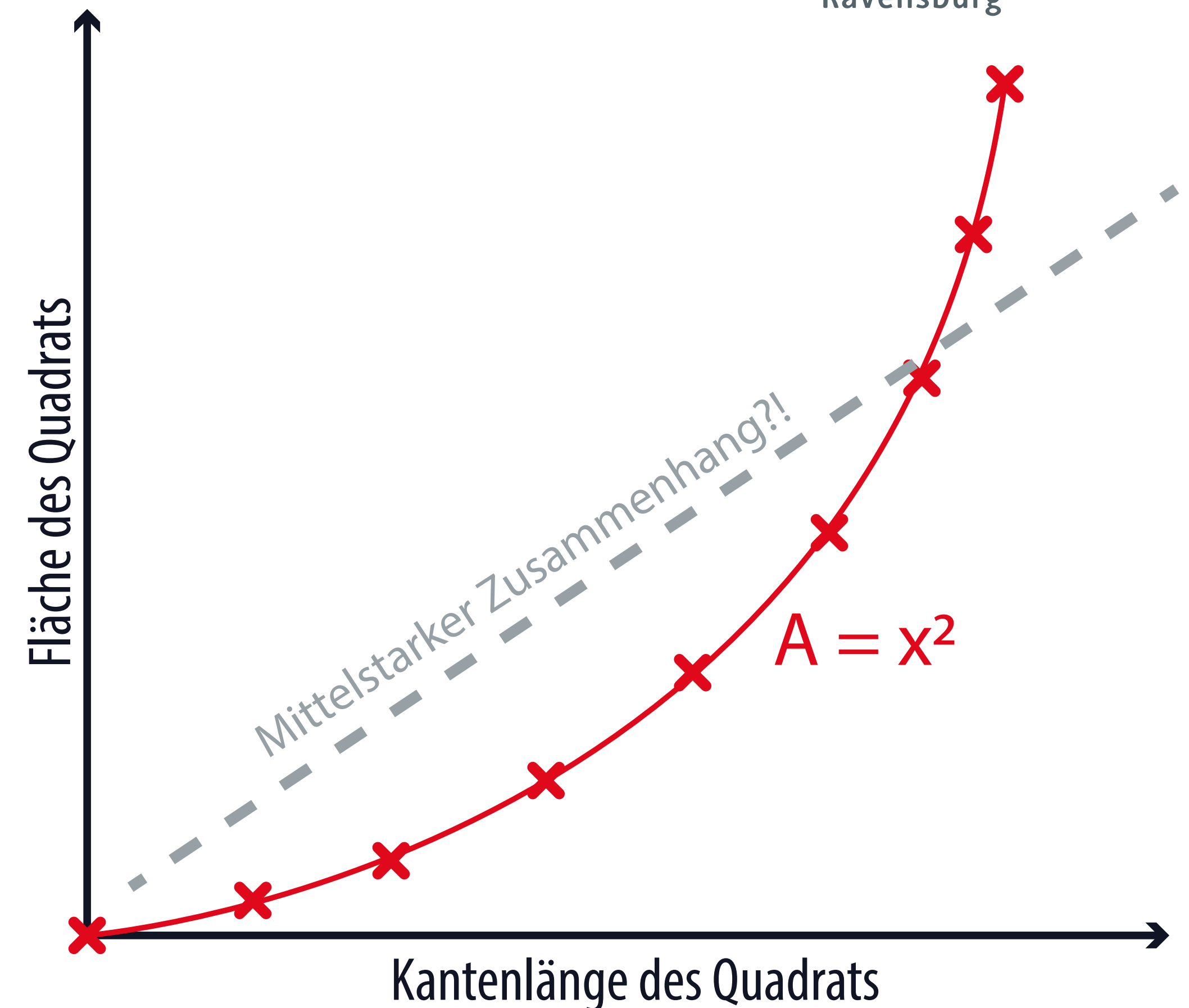
$$x_6 = 4.90 \text{ Meter}$$

$$y_6 = 195 \text{ kg}$$

Kovarianz & Korrelation

Vorsicht Der Bravais-Pearson Korrelationskoeffizient untersucht die Daten nur auf lineare Abhängigkeit.

Bei Daten die eine andere Art von Abhängigkeit aufweisen (z. B. quadratisch, logarithmisch, exponentiell, log-linear) unterschätzt der Korrelationskoeffizient den Zusammenhang.

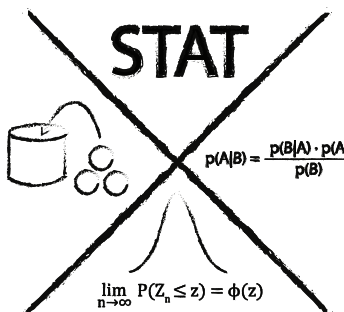


Kovarianz & Korrelation

Vorsicht Genau wie bei den Lagemaßen, benötigen wir für den Bravais-Pearson Korrelationskoeffizienten Daten mit bestimmten Eigenschaften.

Die Daten müssen metrisch sein. Bei ordinalen Daten sollten wir eigentlich einen anderen Koeffizienten (Spearman) verwenden und bei kategorialen Daten müssen wir zwingend auf andere Maße zurückgreifen!

	kategorial	ordinal	metrisch
Modalwert	✓	✓	✓
Median	-	✓	✓
Mittelwert	-	-	✓
Standardabweichung	-	-	✓
Standardfehler	-	-	✓
Range	-	✓	✓
KK (Pearson)	-	-	✓
KK (Spearman)	-	✓	-



Deskriptive Statistik in Excel

Die bisher kennengelernten Datensätze waren sehr klein und künstlich konstruiert, um bestimmte Kennzahlen oder Zusammenhänge zu zeigen.

Wir wollen uns jetzt einen ersten „echten“ Datensatz in Excel anschauen!

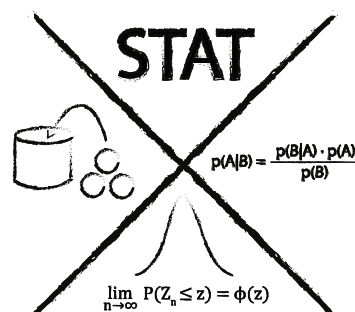
**Datensätze in
Statistik Vorlesungen**



**Datensätze in
der Praxis**



Bildquelle: Pixabay unter Pixabay Content Licence (<https://pixabay.com/photos/things-items-car-old-decrepit-2609870/>) und (<https://pixabay.com/illustrations/ai-generated-cars-showroom-garage-8593983/>)



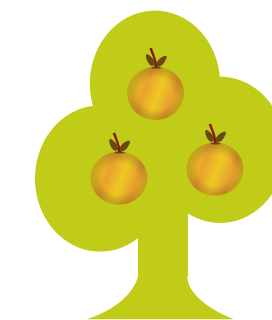
Visualisierung

Wie visualisiere ich eine Stichprobe und die dazu berechneten Kennzahlen?

Es kommt darauf an, welche Informationen wir weitergeben möchten, wer unsere Zielgruppe ist und welches Medium wir verwenden.

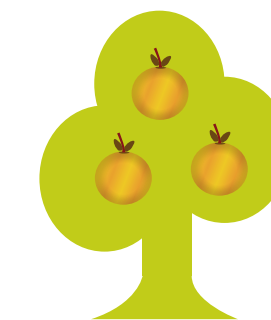
Annahme im Folgenden: Eine Projektarbeit.

Erträge von Apfelbäumen



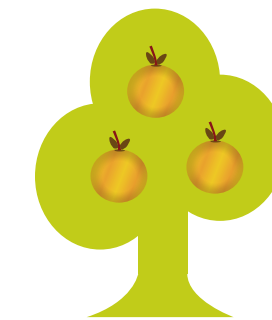
$$x_1 = 5.83 \text{ Meter}$$

$$y_1 = 317 \text{ kg}$$



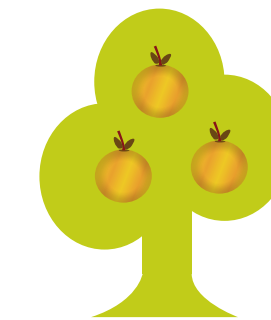
$$x_4 = 6.05 \text{ Meter}$$

$$y_4 = 345 \text{ kg}$$



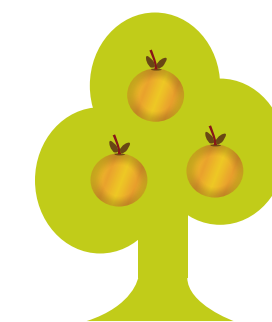
$$x_2 = 4.73 \text{ Meter}$$

$$y_2 = 245 \text{ kg}$$



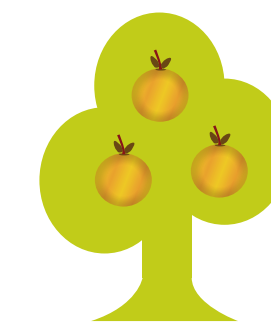
$$x_5 = 4.19 \text{ Meter}$$

$$y_5 = 190 \text{ kg}$$



$$x_3 = 6.10 \text{ Meter}$$

$$y_3 = 298 \text{ kg}$$



$$x_6 = 4.90 \text{ Meter}$$

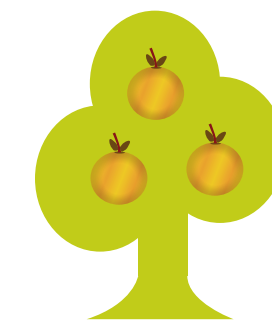
$$y_6 = 195 \text{ kg}$$

Einzelne Werte & Kennzahlen

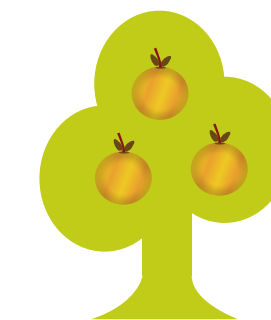
Sind nur bestimmte Kennzahlen einiger Merkmale relevant, ist eine Visualisierung nice to have, aber nicht notwendig. Ein Text ist ausreichend:

Bei einer Untersuchung von 6 Apfelbäumen bezüglich Kronhöhe ($\bar{x} = 5.30$, $\sigma_x = 0.80$) und Ertrag ($\bar{y} = 265$, $\sigma_y = 65.0$) wurde eine hohe Korrelation von $\rho = 0.91$ zwischen diesen Merkmalen festgestellt.

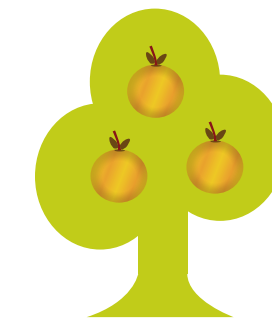
Erträge von Apfelbäumen



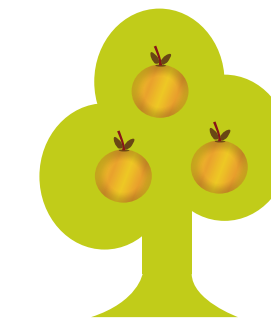
$x_1 = 5.83$ Meter
 $y_1 = 317$ kg



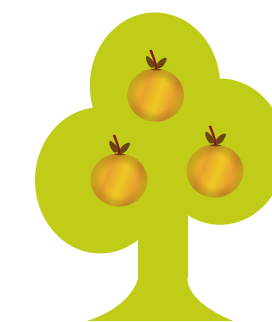
$x_4 = 6.05$ Meter
 $y_4 = 345$ kg



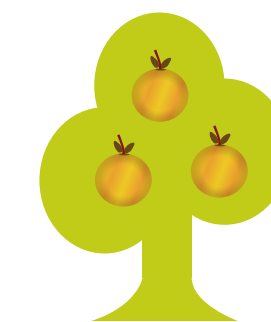
$x_2 = 4.73$ Meter
 $y_2 = 245$ kg



$x_5 = 4.19$ Meter
 $y_5 = 190$ kg



$x_3 = 6.10$ Meter
 $y_3 = 298$ kg



$x_6 = 4.90$ Meter
 $y_6 = 195$ kg

Einzelne Werte & Kennzahlen

Möchte man Kennzahlen mehrerer Merkmale oder mehrerer Gruppen von Merkmalsträgern vergleichen, sollte eine Tabelle oder eine Abbildung verwendet werden.

Auf keinen Fall seitenlang Werte aufzählen!

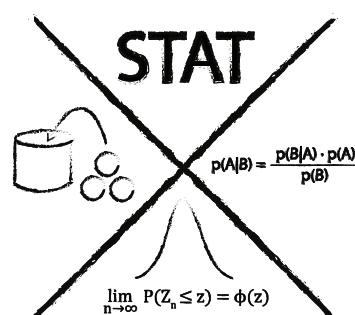
~~Bei der Untersuchung kosteten Äpfel in Deutschland im Durchschnitt 0.64€ pro kg, Birnen in Österreich 0.78€ pro kg, Bananen in der Schweiz 1.37€ pro kg, Orangen in Deutschland dagegen 3.17€ pro kg. Ferner gibt es auch in Österreich Bananen, mit 1.32€ billiger als in den Nachbarländern Österreich (1.32€) und Schweiz (1.42€)...~~

Obst	Äpfel	Birnen	Bananen	Orangen	Pfirsiche	Zitronen
Mittelw.	0.64€	0.79€	1.37€	3.17€	2.27€	1.92€
Stabw.	0.27€	0.31€	0.42€	0.77€	0.40€	0.51€

Tabelle 1 - Mittlere Preise ausgesuchter Obstsorten

Land	Äpfel	Birnen	Bananen	Orangen	Pfirsiche	Zitronen
DE	0.64€ (0.27€)	0.79€ (0.31€)	1.37€ (0.42€)	3.17€ (0.77€)	2.27€ (0.40€)	1.92€ (0.51€)
AU	0.63€ (0.28€)	0.78€ (0.35€)	1.32€ (0.43€)	3.14€ (0.81€)	2.37€ (0.37€)	1.62€ (0.50€)
CH	0.59€ (0.24€)	0.74€ (0.30€)	1.42€ (0.55€)	3.37€ (1.04€)	2.17€ (0.43€)	2.04€ (0.47€)

Tabelle 2 - Preise ausgesuchter Obstsorten in DACH-Region



Diagramme

Diagramme sind Abbildungen, die dem Leser unserer Arbeit eine schnelle Übersicht über die Daten geben. Die wichtigsten Typen:

Balkendiagramm für kategoriale Vergleiche

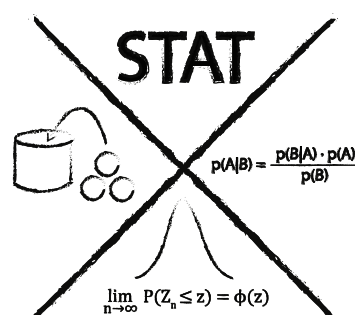
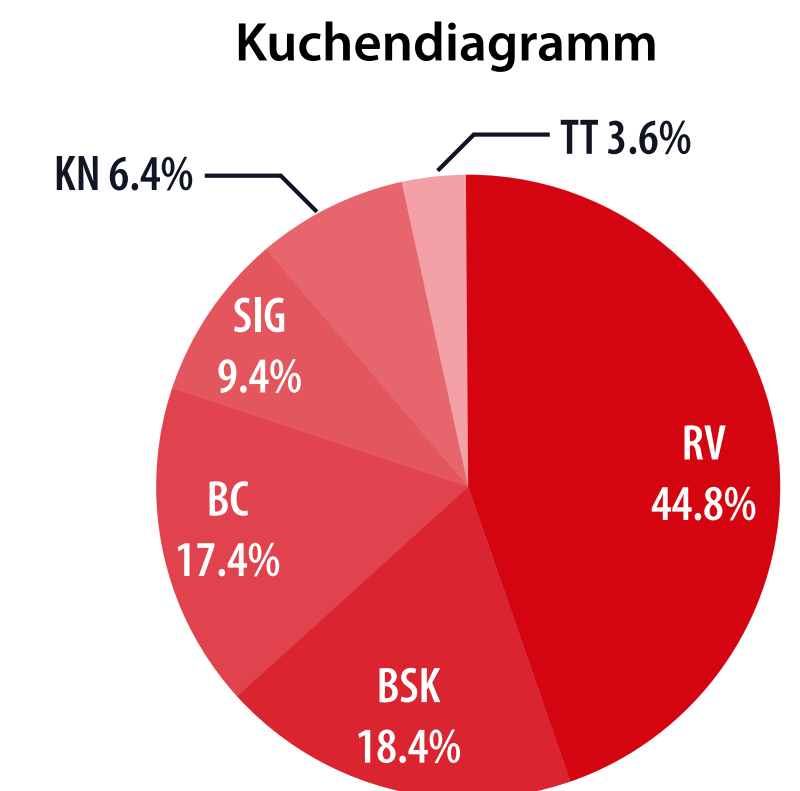
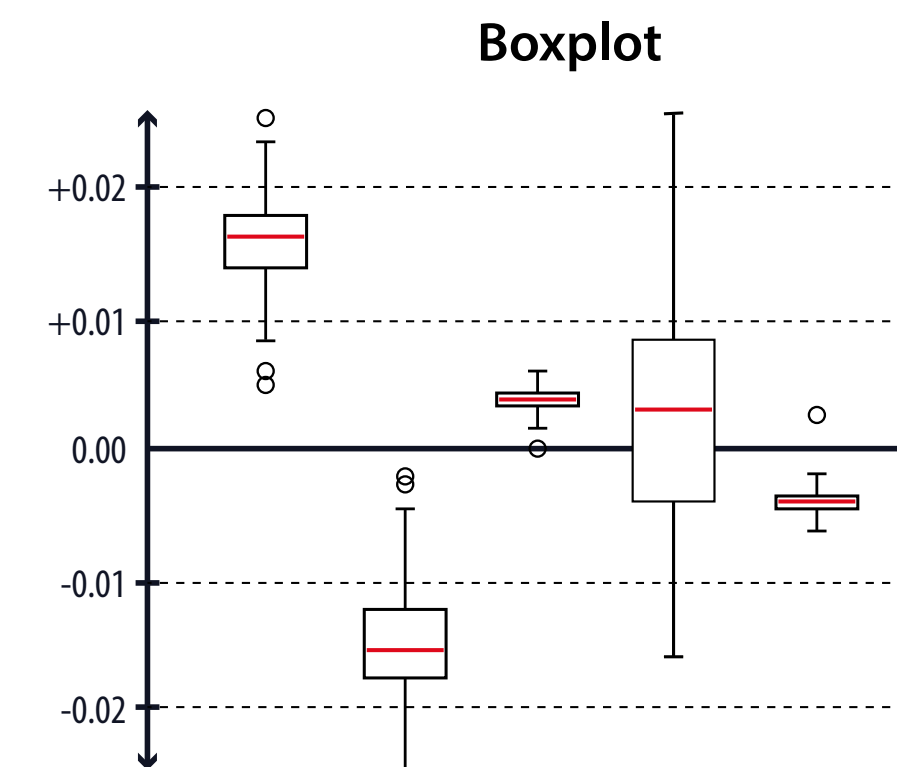
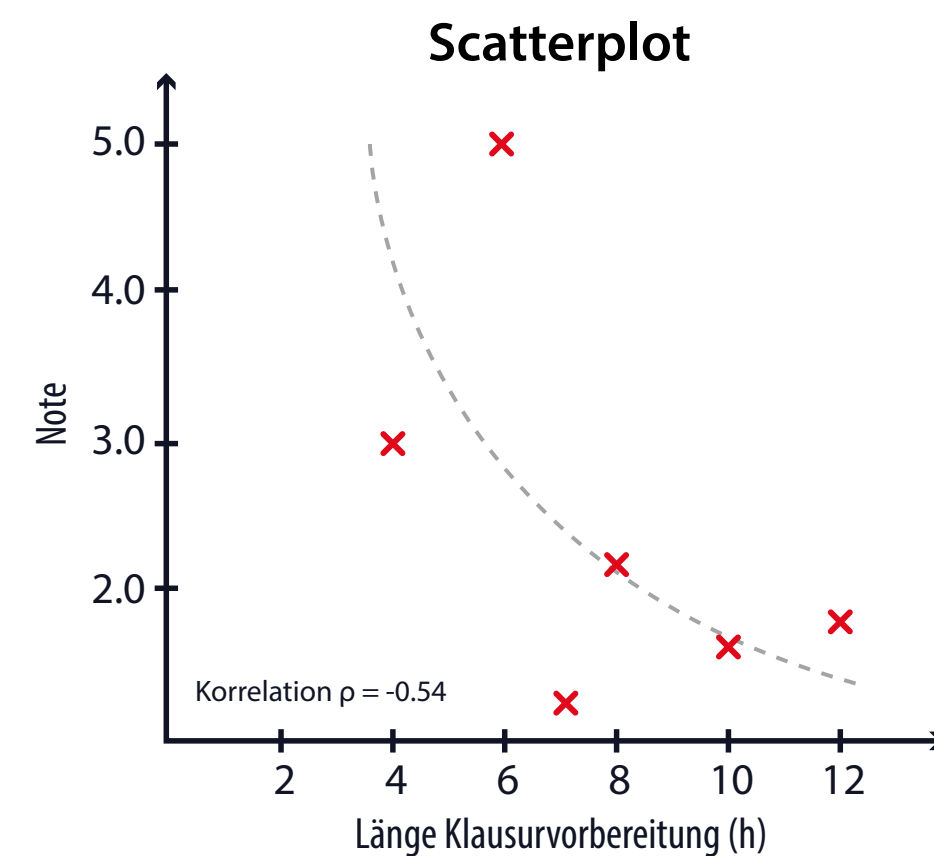
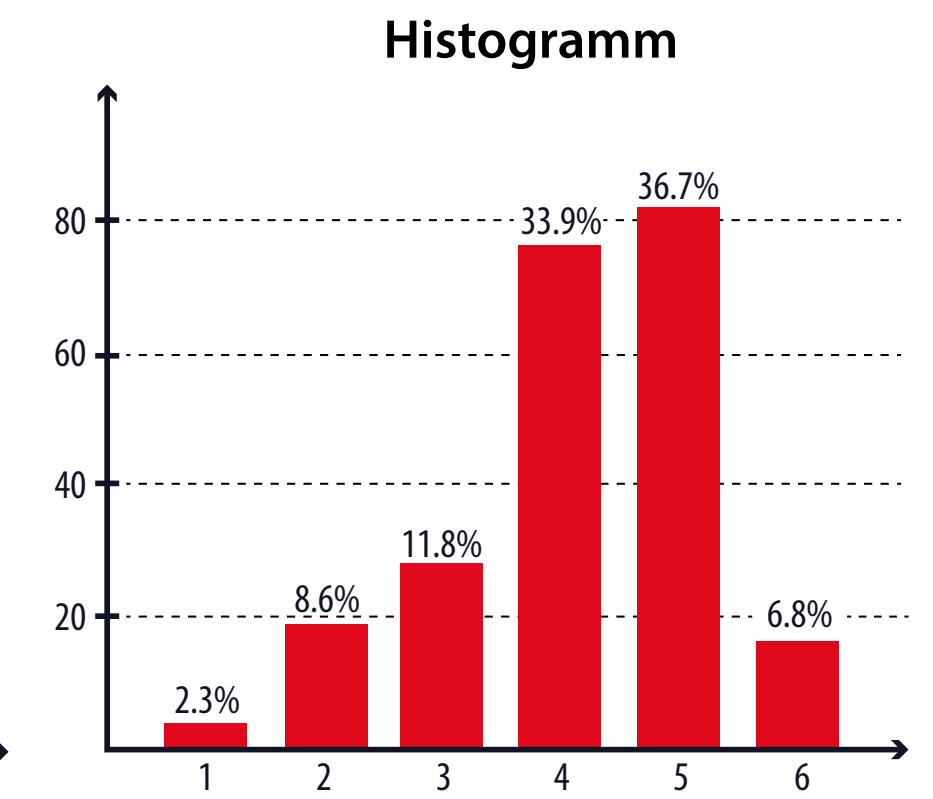
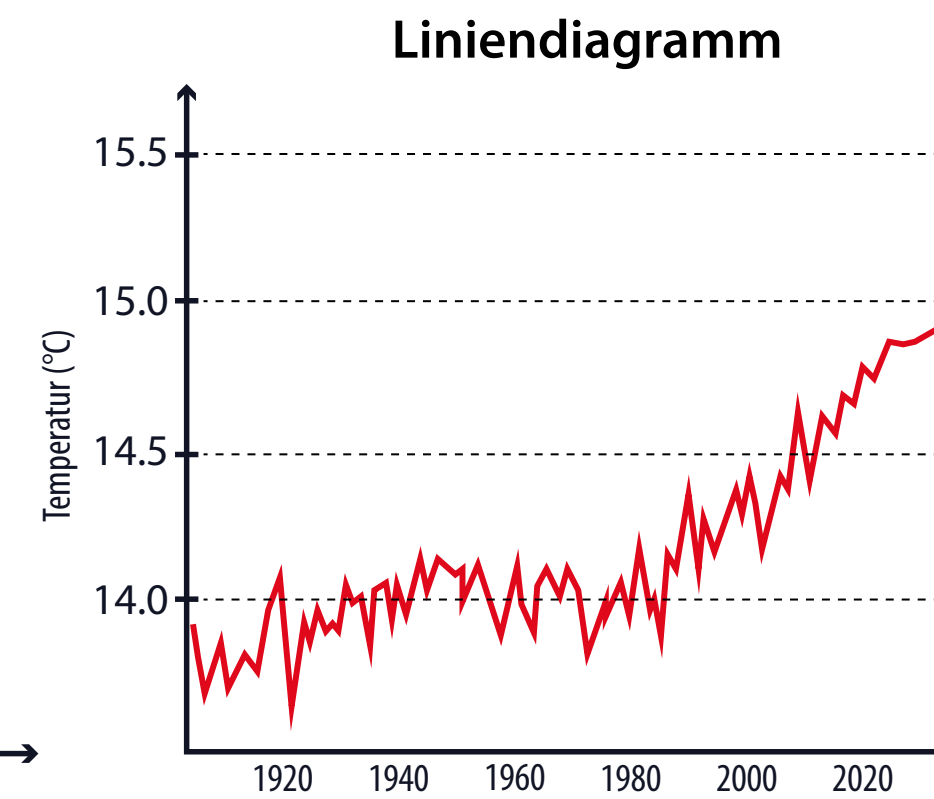
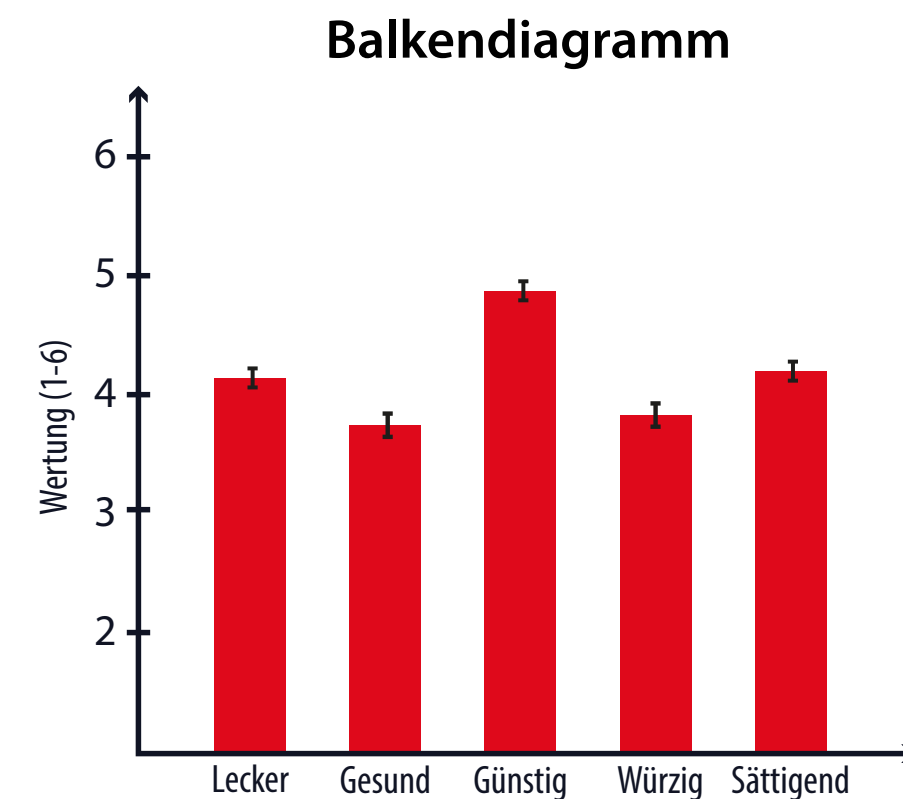
Histogramme für Verteilungen

Boxplots zum Vergleich von Verteilungen

Liniendiagramme für Zeitreihen

Scatterplots für Korrelationen

Kuchendiagramme als Alternative zu Balken



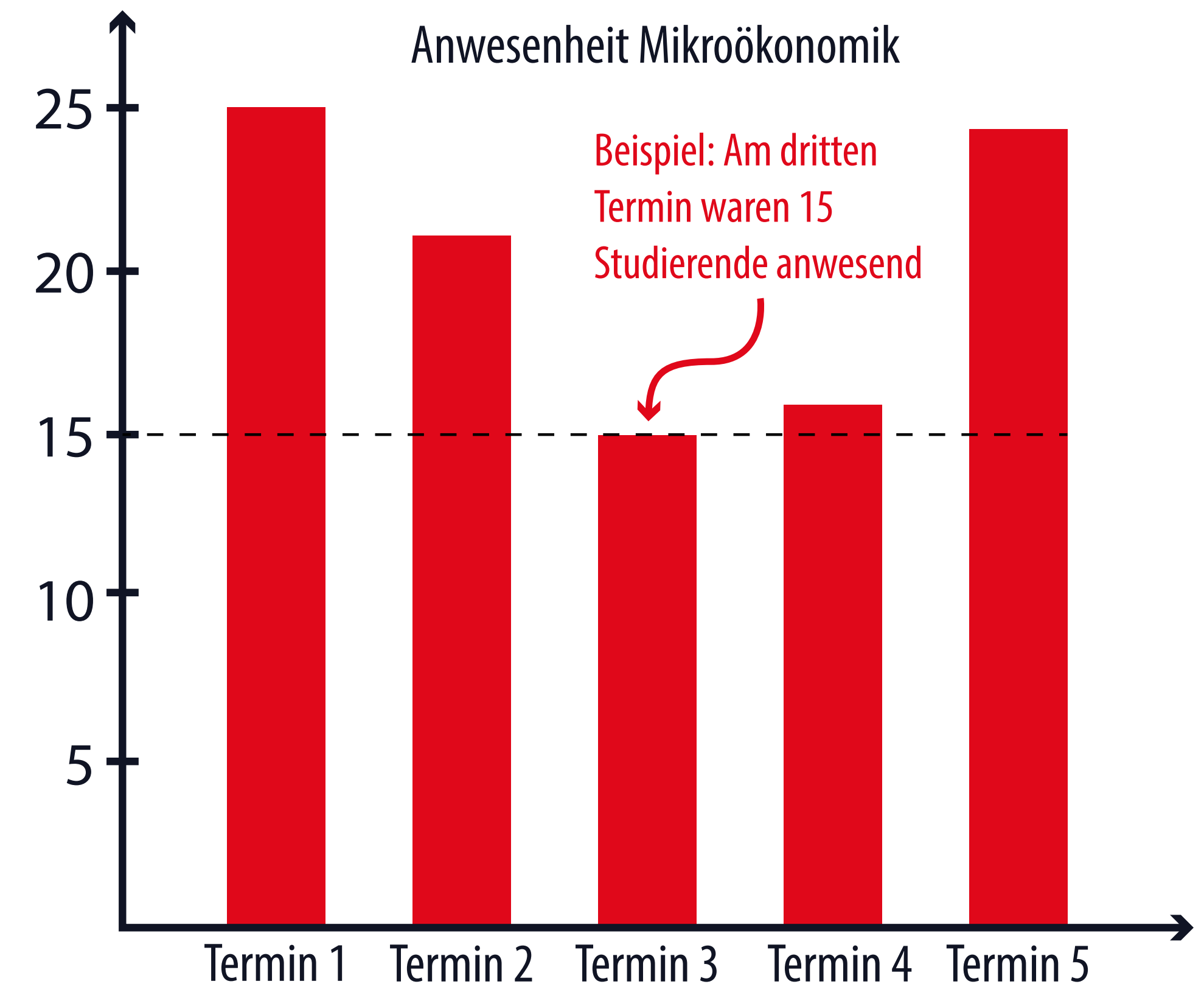
Balkendiagramme

Balkendiagramm eignen sich, um numerische Merkmale verschiedener Merkmalsträger zu vergleichen.

Die x-Achse ist kategorial und listet die Merkmalsträger auf.

Die y-Achse ist numerisch und ist passend zu den Merkmalsausprägungen skaliert.

Wir lesen die Merkmalsausprägungen an der Höhe der Balken ab.



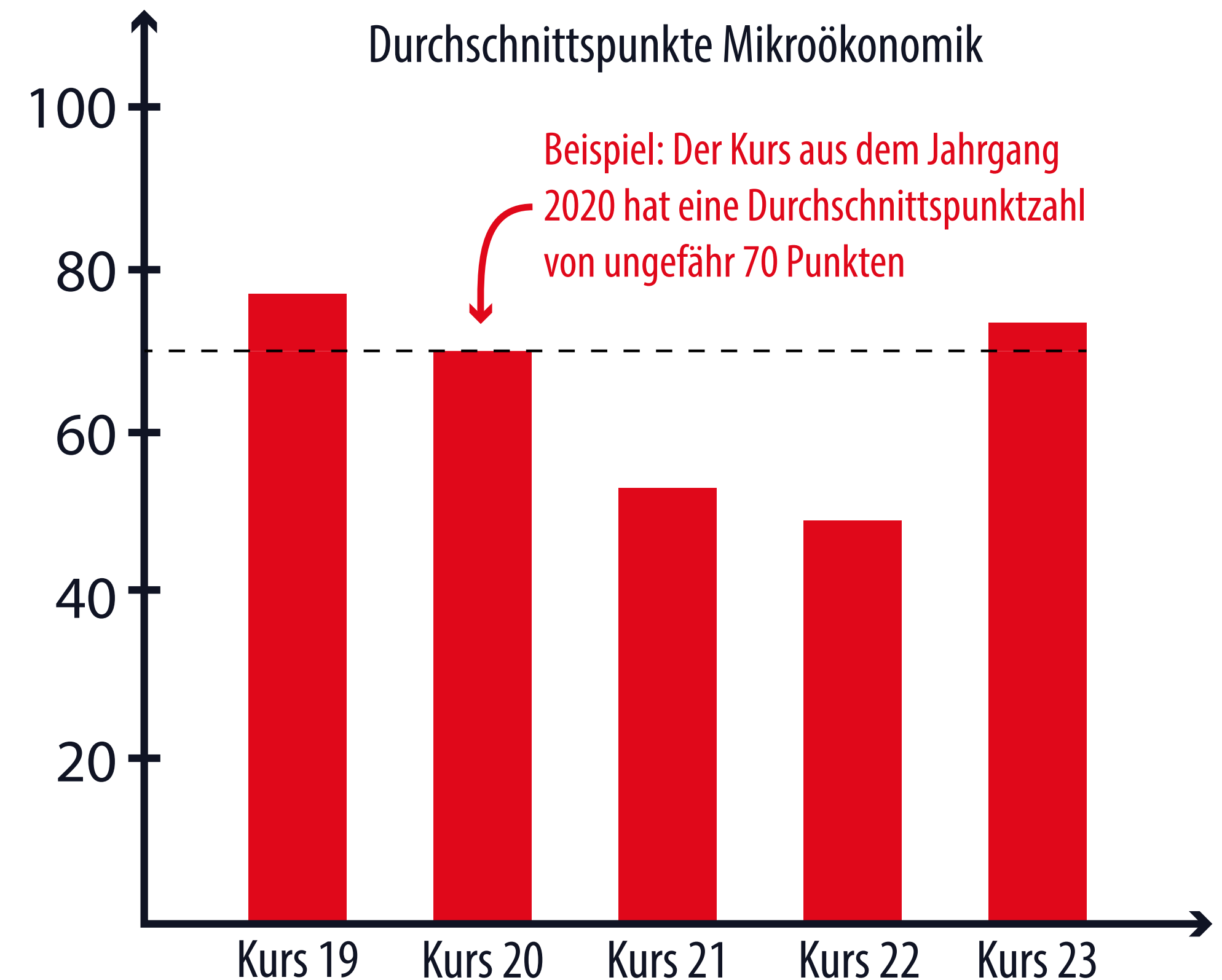
Balkendiagramme

Balkendiagramm eignen sich auch, um Mittelwerte von Gruppen zu vergleichen.

Die x-Achse ist kategorial und listet die Gruppen.

Die y-Achse ist numerisch und ist passend zu den Mittelwerten skaliert.

Wir lesen die Mittelwerte an der Höhe der Balken ab.



Balkendiagramme

Balkendiagramm eignen sich auch, um Mittelwerte von Umfrage-Items zu vergleichen.

Die x-Achse ist kategorial und listet die Items auf.

Die y-Achse ist numerisch und ist entsprechend der in der Umfrage verwendeten Skala skaliert.

Wir lesen die Mittelwerte an der Höhe der Balken ab.



Balkendiagramme

Mit Fehlerbalken können wir zusätzlich zu den Mittelwerten auch die Standardabweichungen oder die Standardfehler angegeben werden.

Da die Bedeutung der Fehlerbalken nicht standardisiert ist, sollte in der Bildunterschrift oder der Legende erklärt werden, was die Fehlerbalken bedeuten.

Variante mit **Standardabweichung**

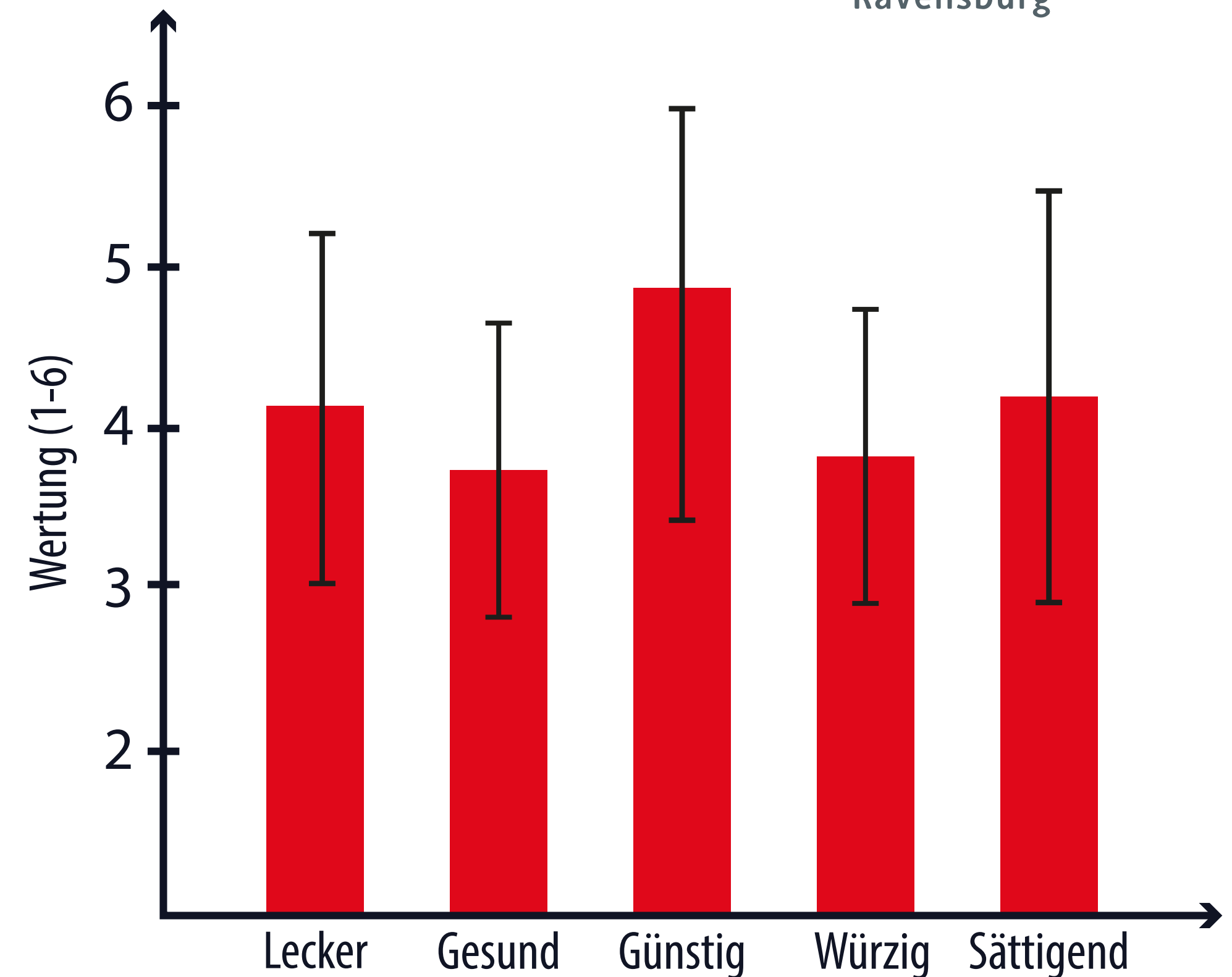


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den **Standardabweichungen**

Balkendiagramme

Mit Fehlerbalken können wir zusätzlich zu den Mittelwerten auch die Standardabweichungen oder die Standardfehler angegeben werden.

Da die Bedeutung der Fehlerbalken nicht standardisiert ist, sollte in der Bildunterschrift oder der Legende erklärt werden, was die Fehlerbalken bedeuten.

Variante mit **Standardfehlern**

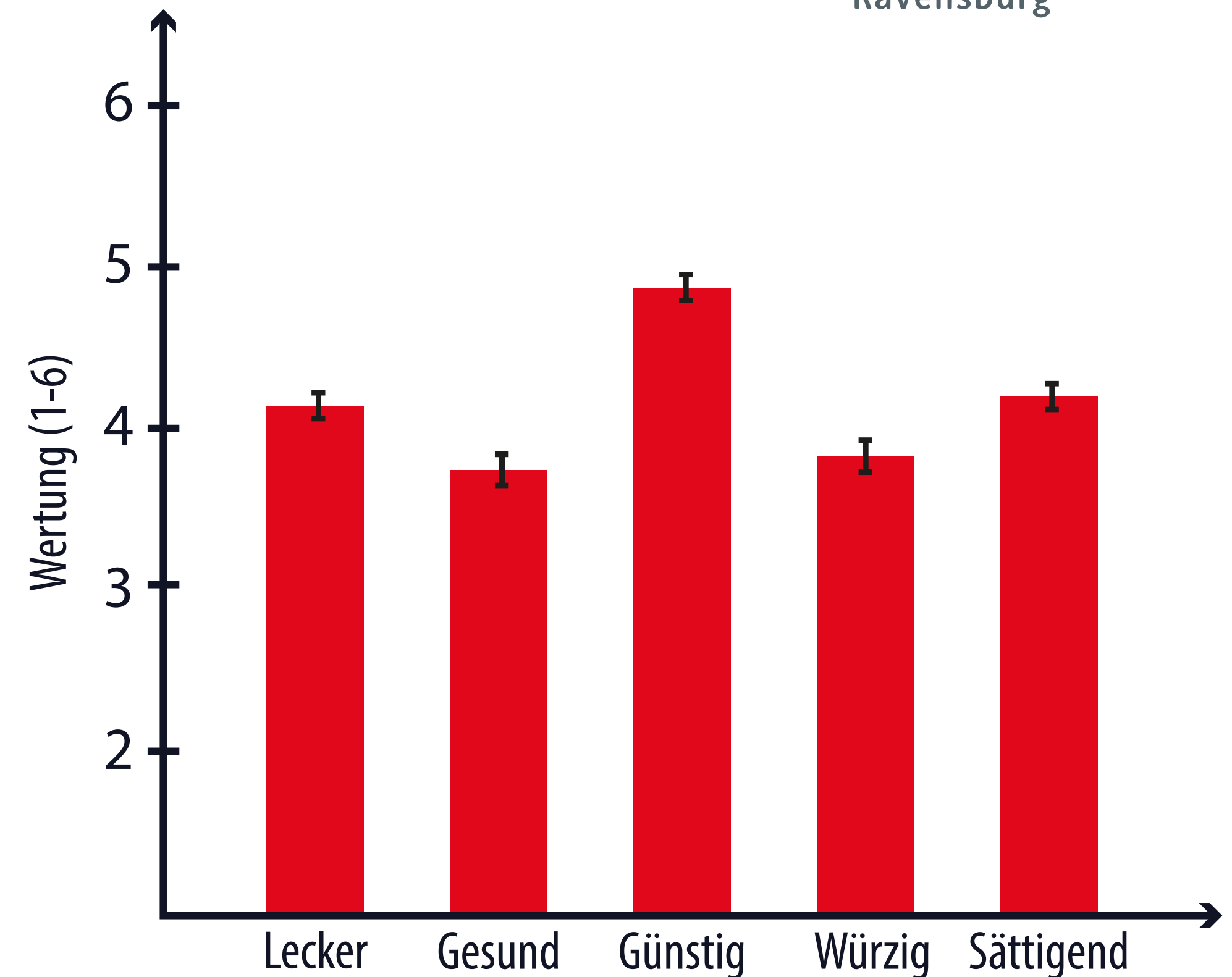


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den **Standardfehlern**

Balkendiagramme

Was bedeutet Standardfehler? Der Standardfehler ist der Quotient aus der Standardabweichung und der Wurzel der Stichprobengröße:

$$\sigma(\bar{X}) = \frac{\sigma_x}{\sqrt{n}}$$

Im Unterschied zur Standardabweichung wird er mit größerer Stichprobe automatisch kleiner!

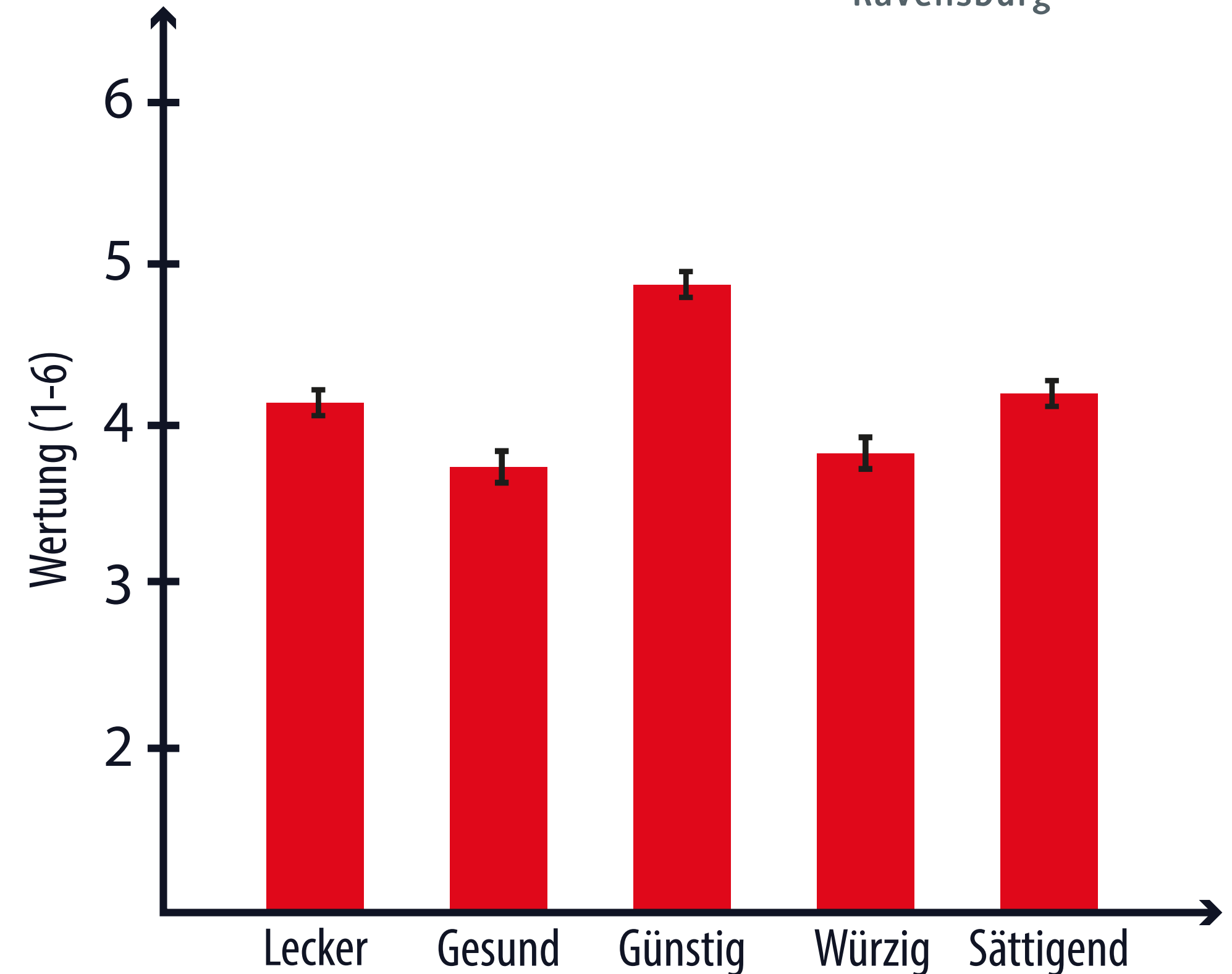


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den **Standardfehlern**

Balkendiagramme

Standardabweichung beschreibt die mittlere Abweichung der Werte zum Mittelwert. Verwende diese, um die Streuung innerhalb der Daten zu zeigen, aus denen der Mittelwert gebildet wurde.

Standardfehler beschreibt die „Ungenauigkeit“ des aus der Stichprobe berechneten Mittelwerts. Verwende diesen, wenn du betonen möchtest:

- Wie zuverlässig der Mittelwert ist.
- Ob sich Mittelwerte der Gruppen/Items unterscheiden.

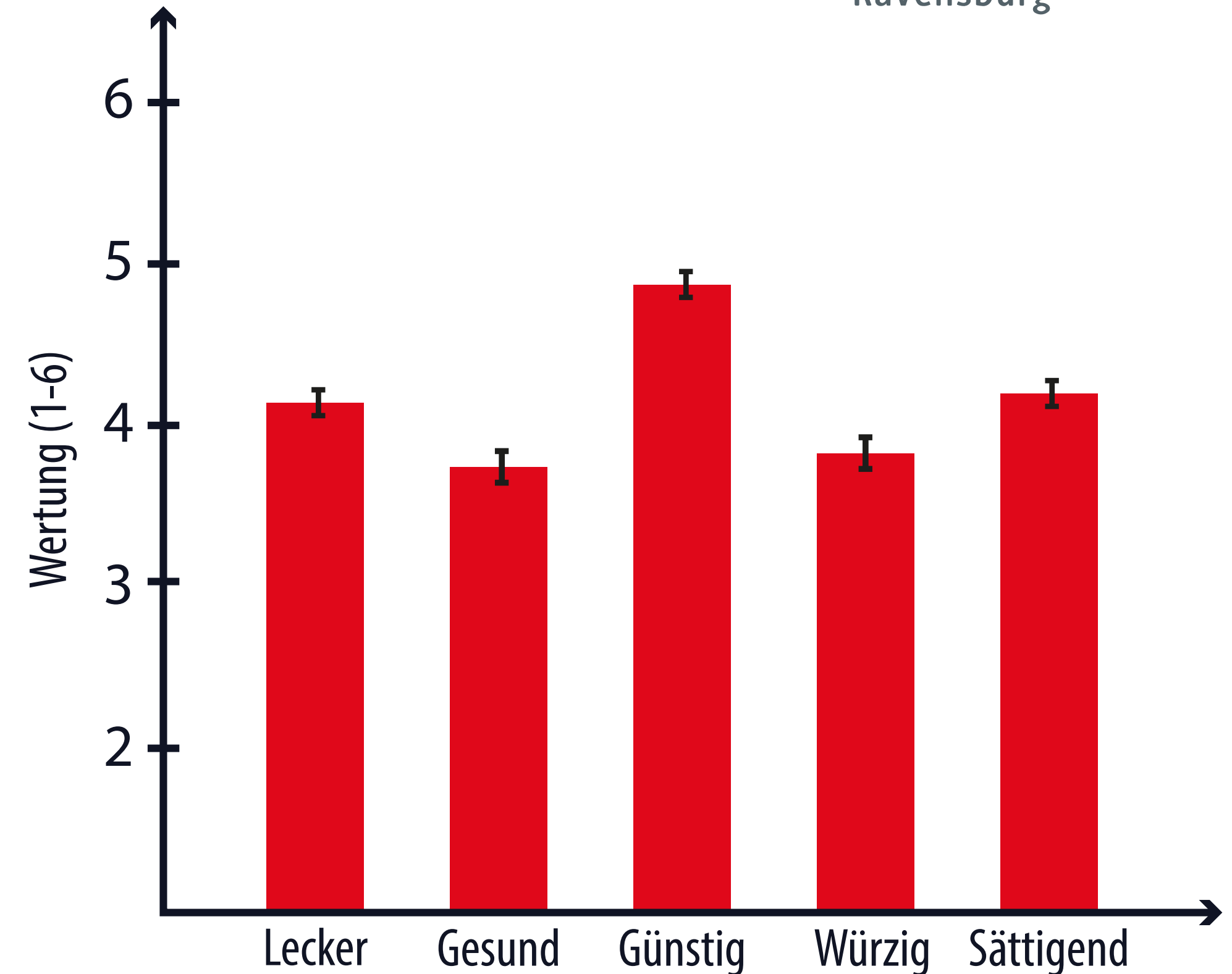


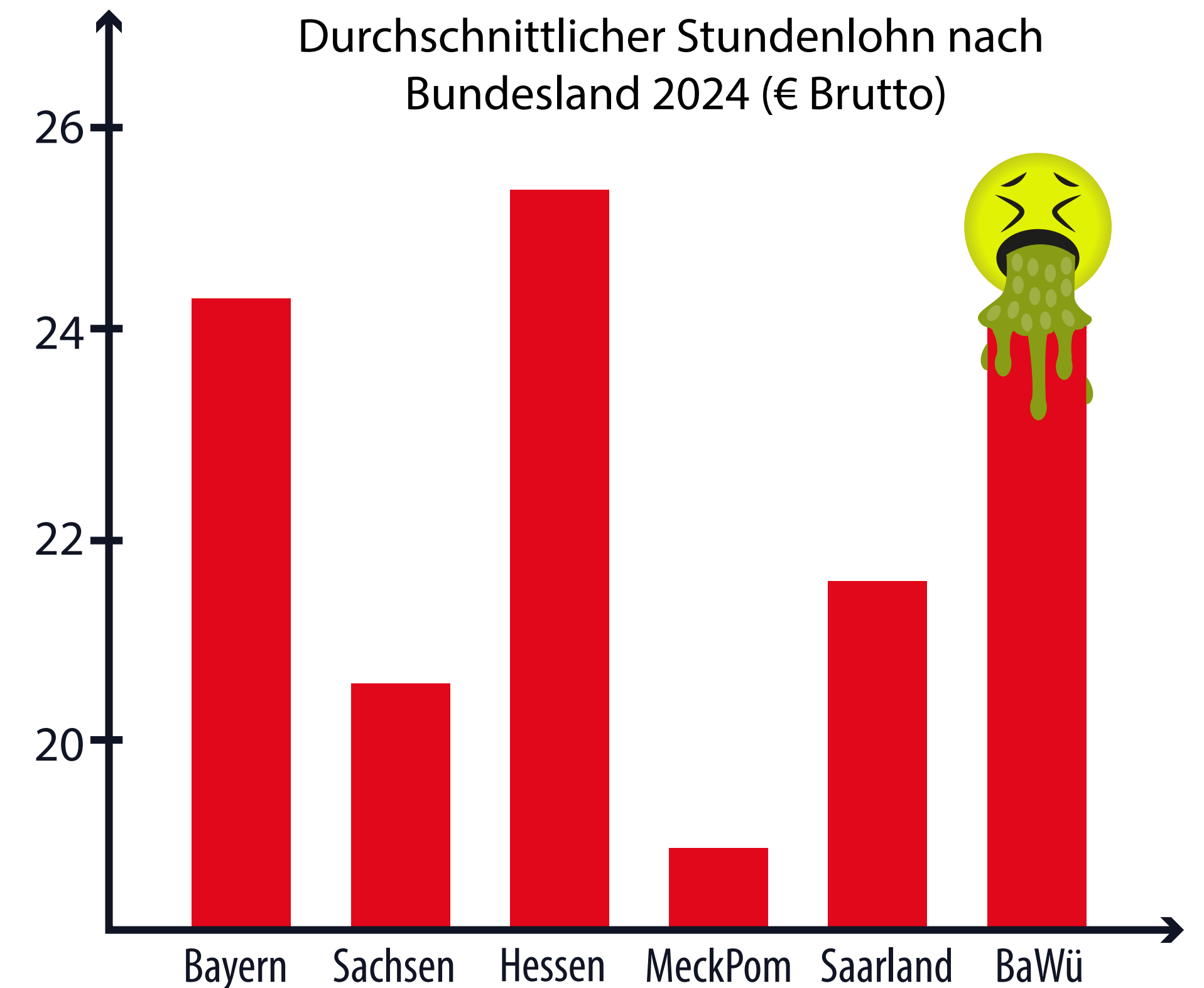
Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den **Standardfehlern**

Balkendiagramme

Auch wenn Balkendiagramme intuitiv und selbsterklärend sind, gibt es trotzdem Regeln, auf die wir achten müssen.

In dem Beispiel rechts hat der Autor zwei Fehler gemacht. Einen davon ggf. sogar absichtlich!

Wie können wir das Balkendiagramm besser machen?

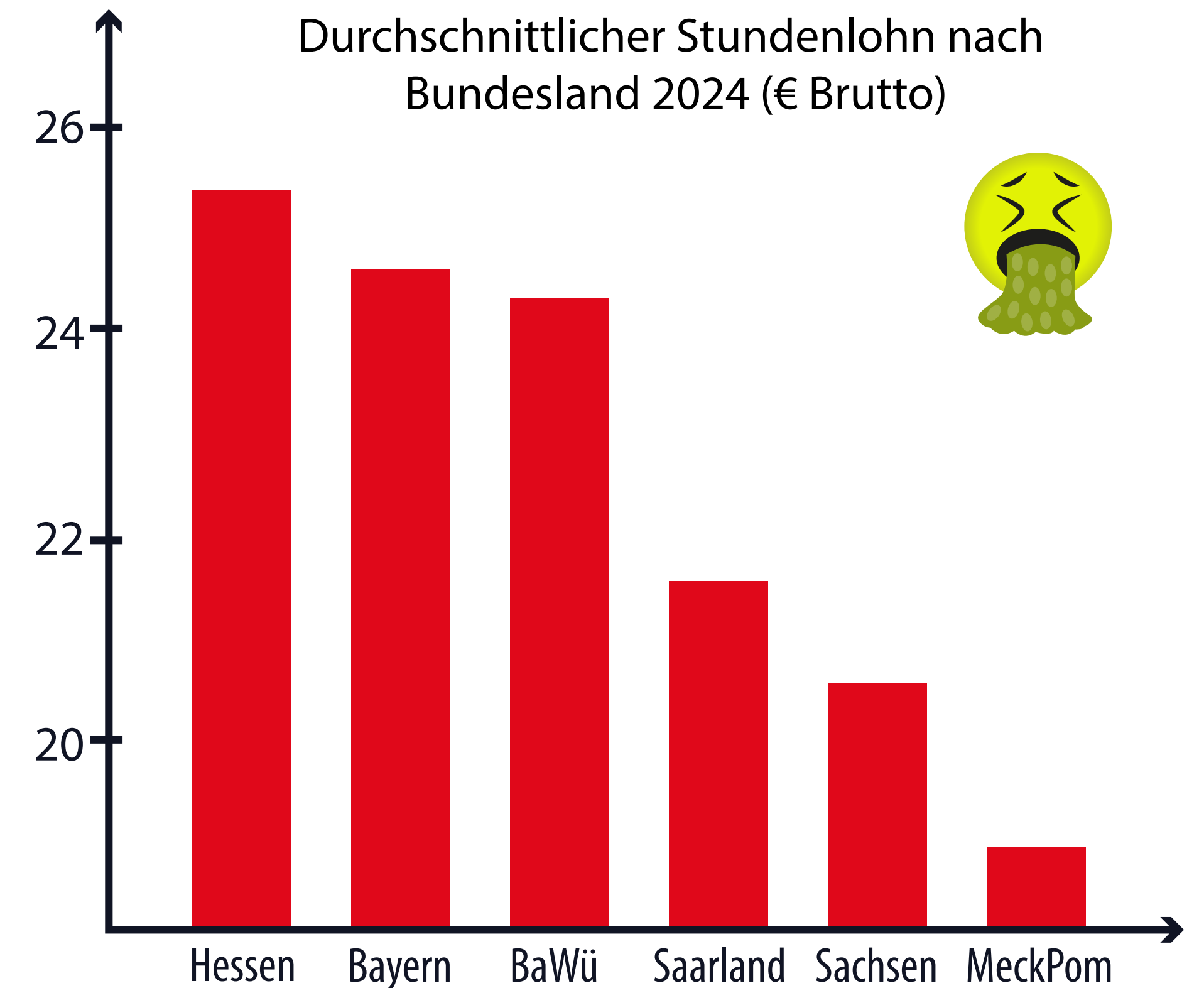


Datenquelle: Statistisches Bundesamt (<https://www-genesis.destatis.de/datenbank/online/statistic/62361/table/62361-0051>)

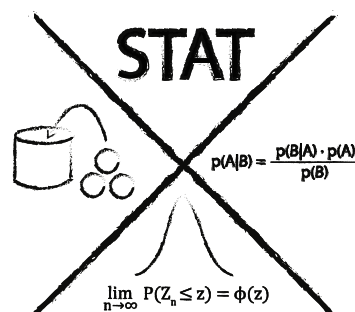
Balkendiagramme

Sortierung Wenn keine zwingende oder naheliegende Sortierung vorliegt, sortieren wir die Balken nach ihren Werten absteigend.

Durch die Sortierung nach Werten können wir die Werte der Bundesländer schneller vergleichen und auch optisch sieht das Diagramm gleich besser aus!



Datenquelle: Statistisches Bundesamt (<https://www-genesis.destatis.de/datenbank/online/statistic/62361/table/62361-0051>)

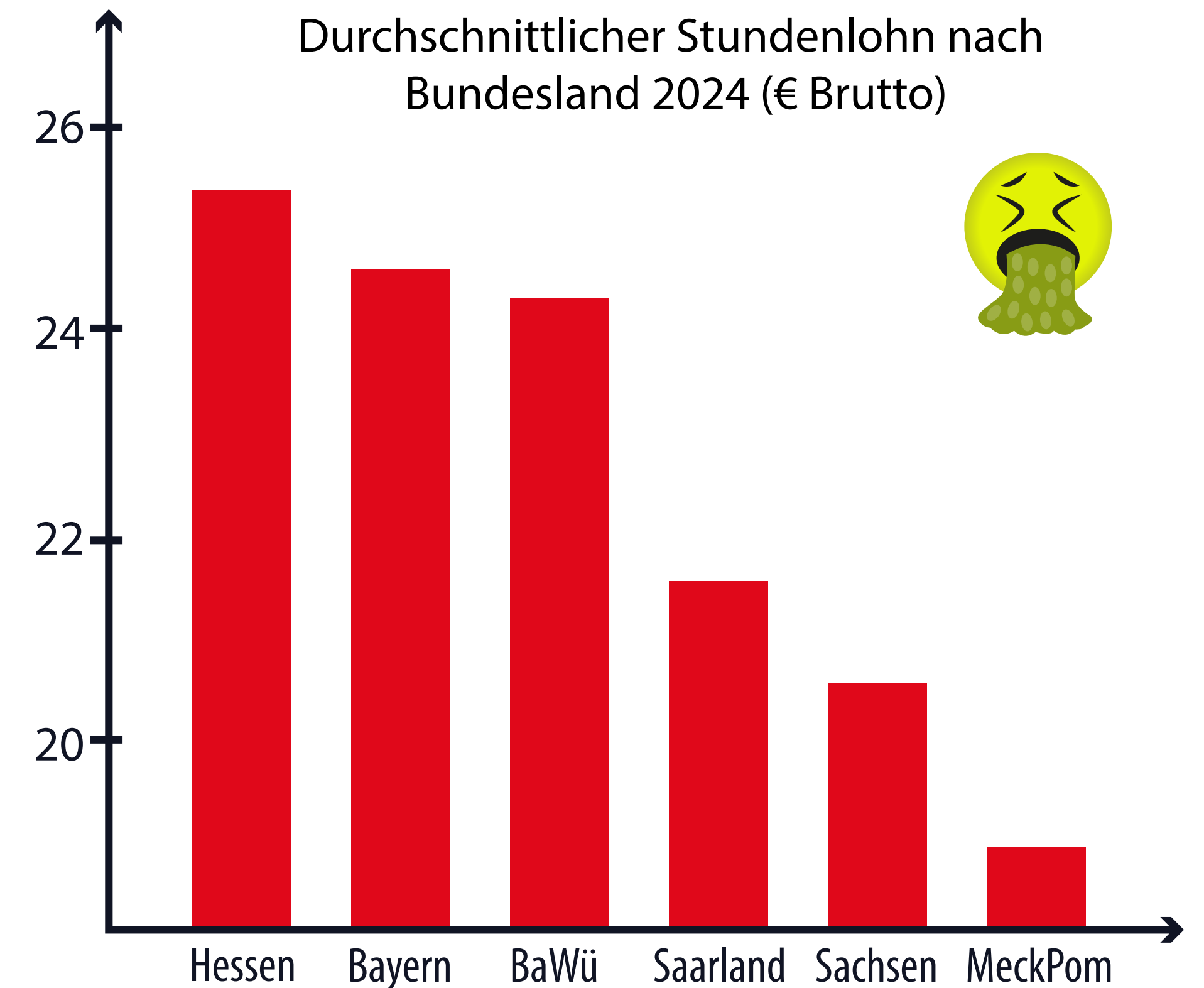


Balkendiagramme

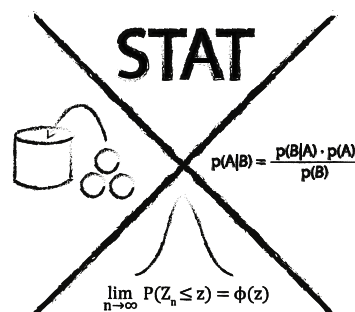
Skalierung Die Skalierung der y-Achse wurde bewusst so gewählt, dass die Unterschiede möglichst groß aussehen.

Das kann gut gemeint sein: Mit dieser Skalierung können die Werte besser abgelesen und verglichen werden.

Das kann aber auch manipulativ sein! Durch die Skalierung bestimmen wir die Botschaft, welche das Diagramm auf den ersten Blick übermittelt.



Datenquelle: Statistisches Bundesamt (<https://www-genesis.destatis.de/datenbank/online/statistic/62361/table/62361-0051>)

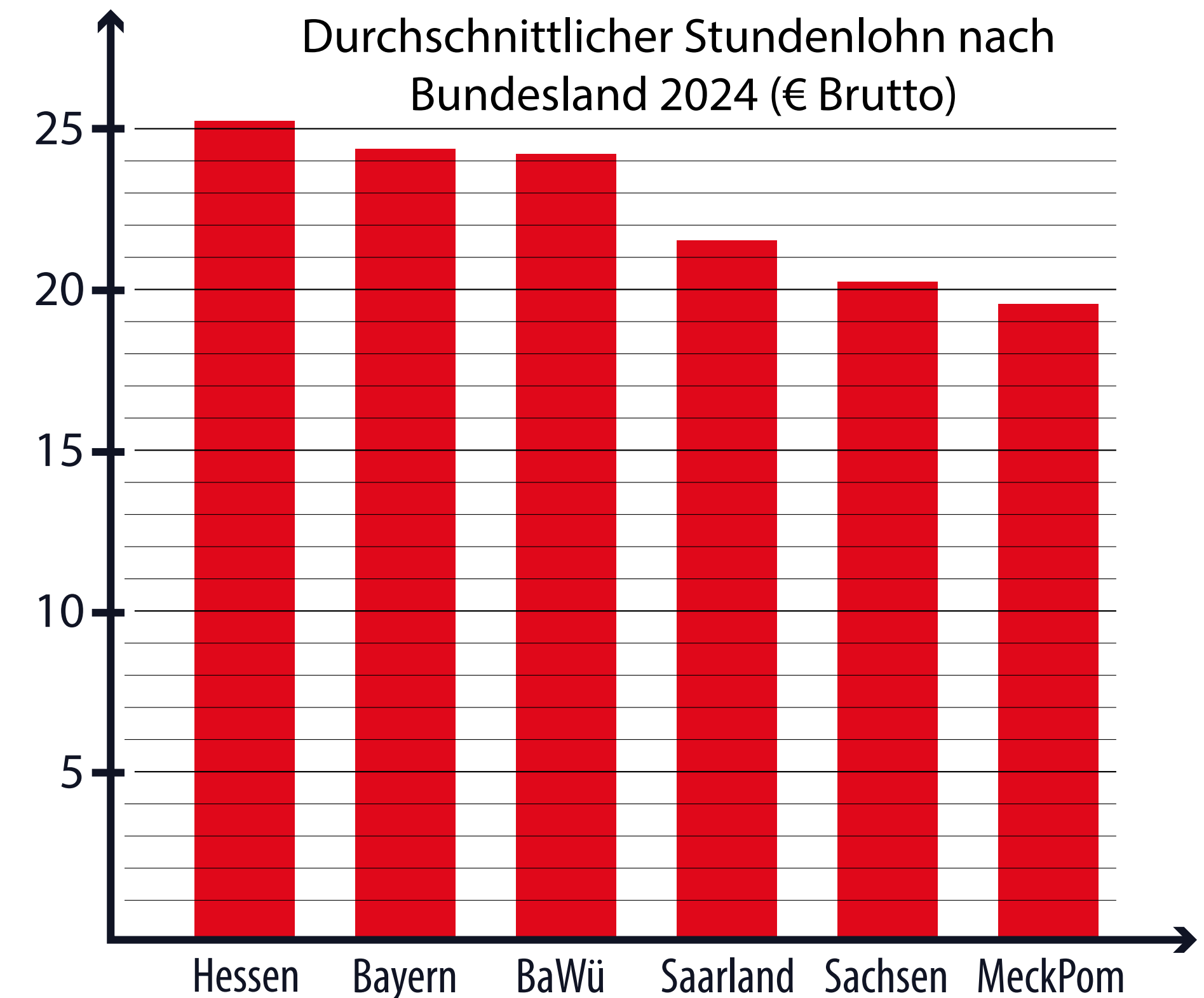


Balkendiagramme

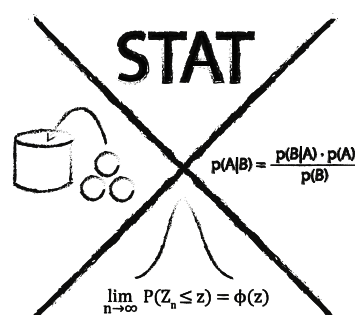
In wissenschaftlichen Arbeiten wollen wir neutral die Datenlage zeigen. Aktives Framing ist nicht erwünscht!

Law of Proportional Ink: Die Größe der Farbfläche soll proportional zum Datenwert sein. Dazu muss die y-Achse bei dem Wert 0 beginnen.

Für das einfache Ablesen können wir Hilfslinien einsetzen.



Datenquelle: Statistisches Bundesamt (<https://www-genesis.destatis.de/datenbank/online/statistic/62361/table/62361-0051>)



Balkendiagramme

Es gibt natürlich Ausnahmen vom **Law of Proportional Ink**

Bei unseren Merkmalen beinhaltet die Skala die Wertungsstufen von 1 bis 6. Würden wir die y-Achse bei 0 beginnen lassen, würden wir die Wertungen besser erscheinen lassen, als sie sind!



Balkendiagramme

Eine weitere Ausnahme liegt vor, wenn die Spannweite der dargestellten Werte sehr gering ist.

Hier kann das **Law of Proportional Ink** dazu führen, dass die Unterschiede in den Balkenhöhen nicht erkennbar sind.

Wir müssen dann entweder die y-Achse abweichend skalieren oder statt dem Merkmal eine Differenz abbilden!

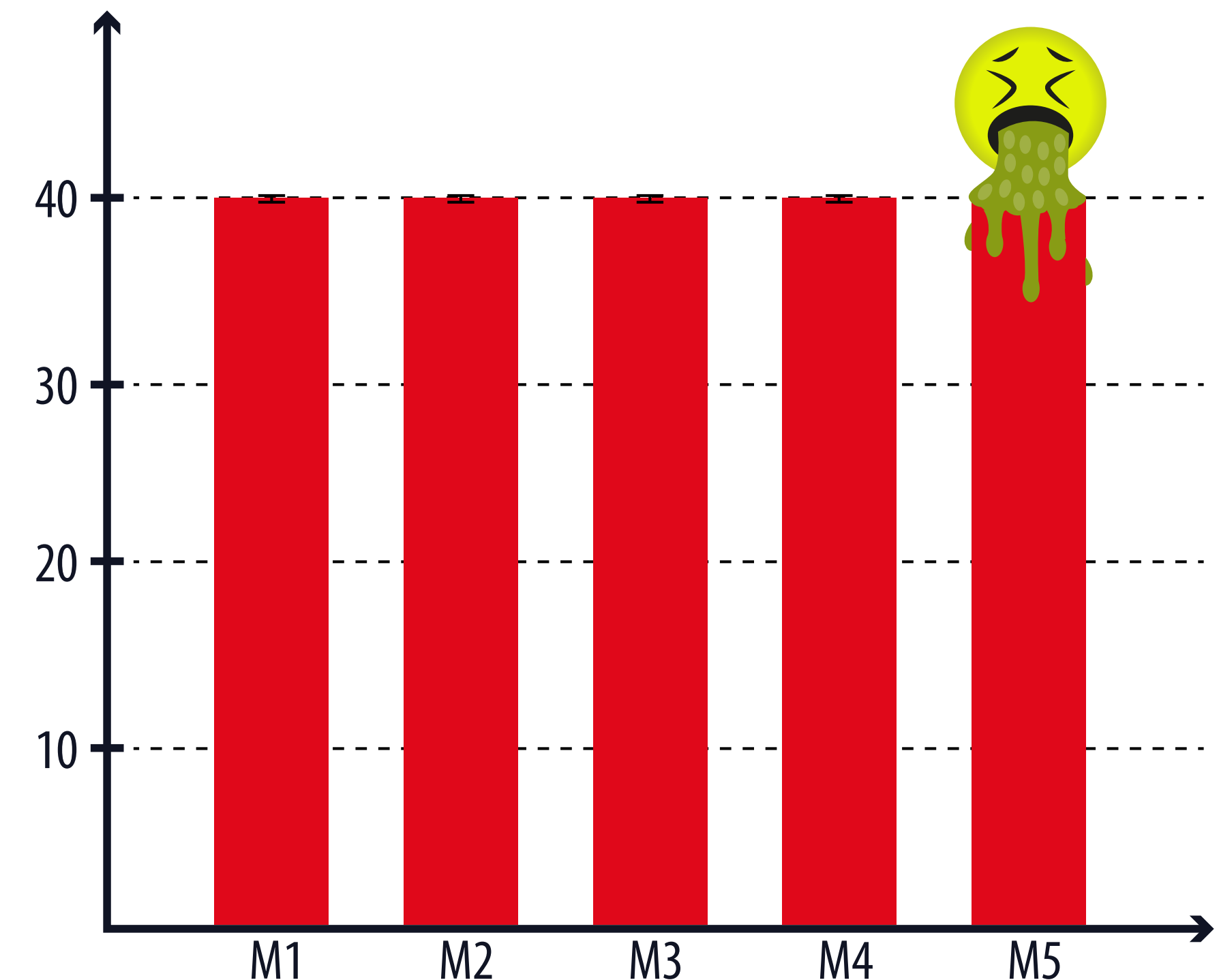


Abbildung 1 - Durchschnittliche Länge von M6x40mm Schrauben aus verschiedenen Drahtschneidemaschinen

Balkendiagramme

Statt die durchschnittliche Länge der Schrauben könnten wir die durchschnittliche Abweichung zum Zielwert 40mm angeben. Wir erkennen jetzt, dass:

- die Maschine 1 grundsätzlich zu lange Stücke schneidet.
- die Maschine 2 grundsätzlich zu kurze Stücke schneidet.
- die Maschinen 3 und 5 am genauesten arbeiten.
- die Länge bei Maschine 4 am stärksten variiert.

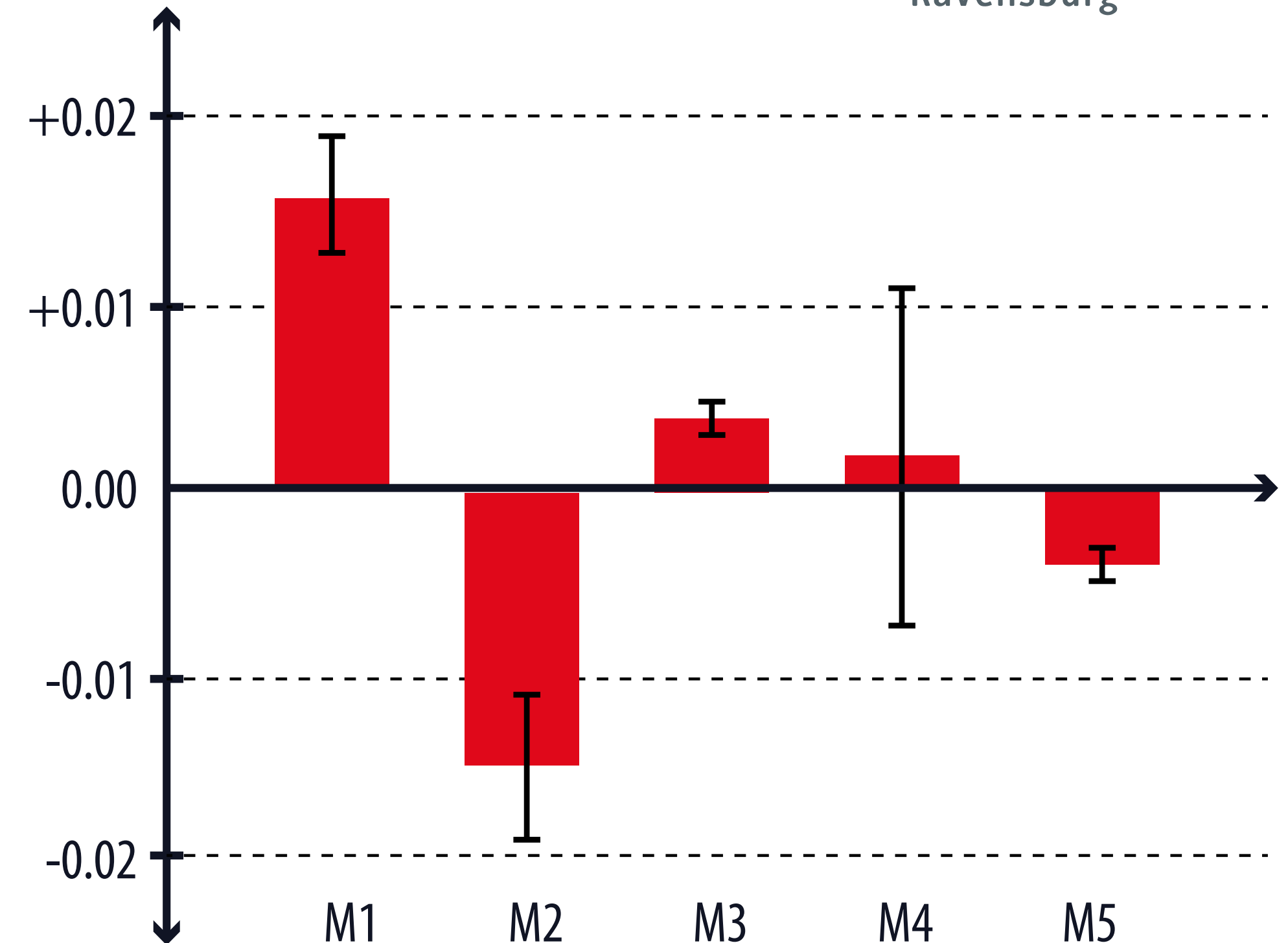


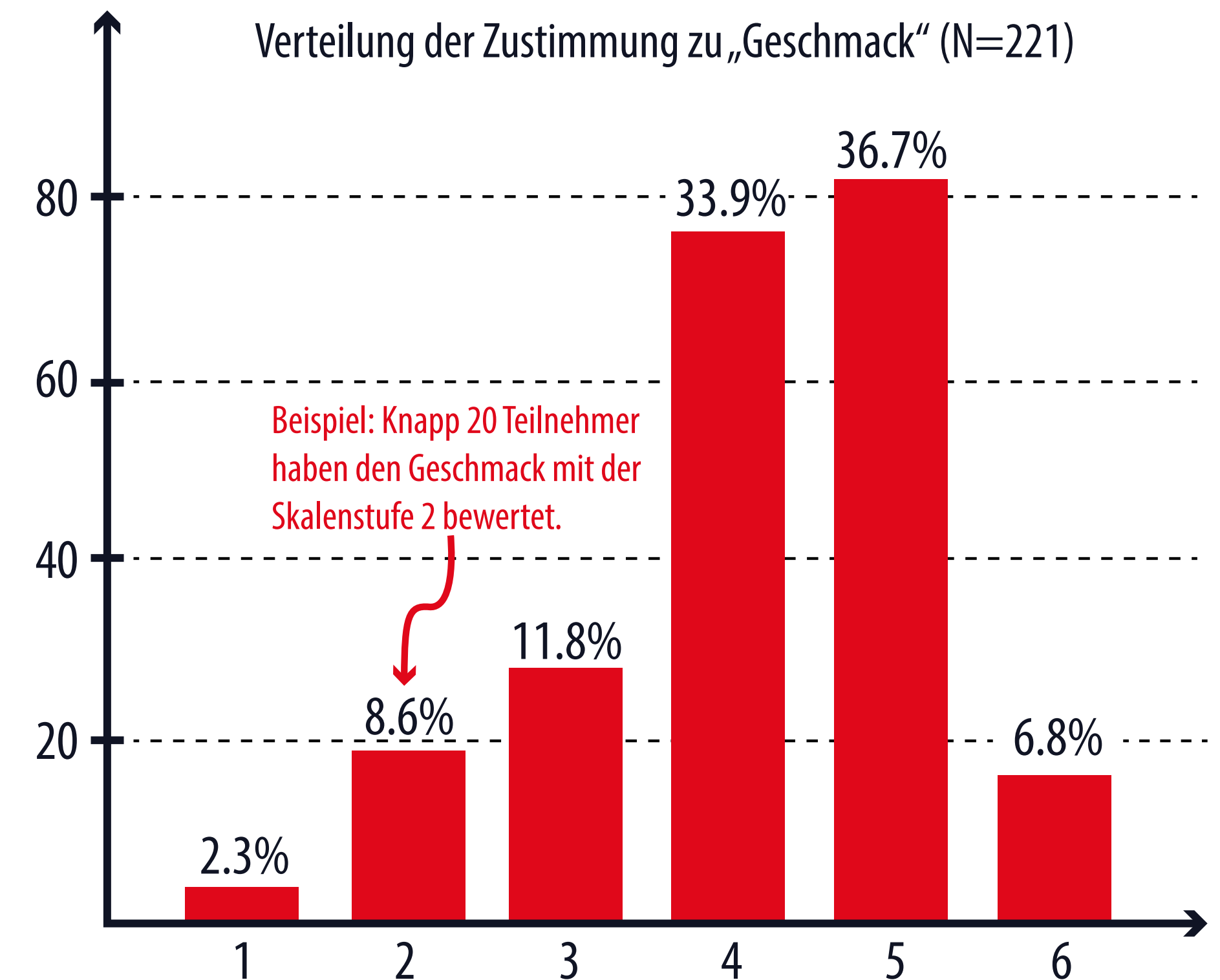
Abbildung 1 - Systematische und zufällige Fehler bei M6x40mm Schrauben aus verschiedenen Drahtschneidemaschinen. Fehlerbalken zeigen Standardabweichungen der Länge

Histogramme

Histogramme sind besondere Balkendiagramme, bei denen die Häufigkeit gegen die Merkmalsausprägungen aufgetragen wird.

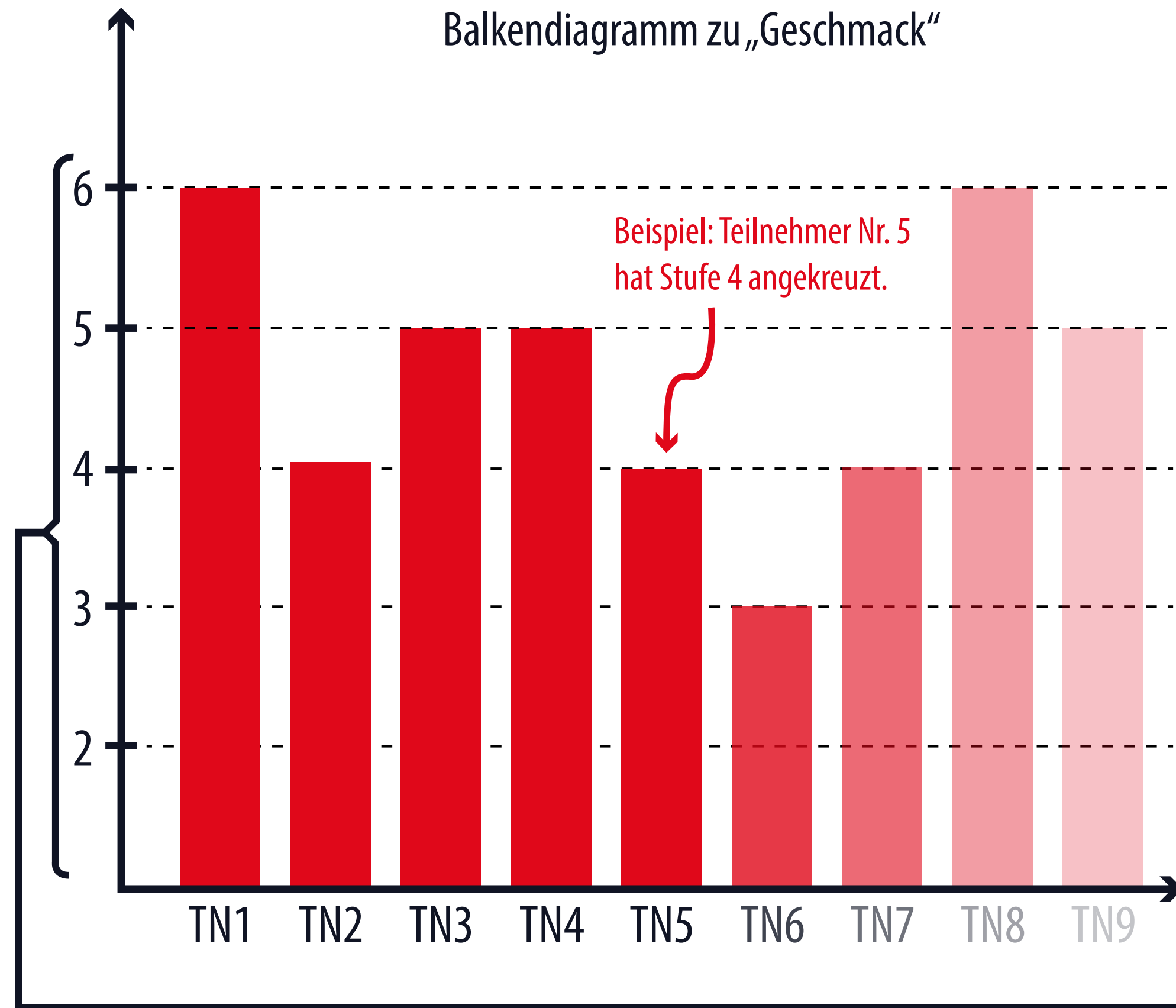
Die x-Achse enthält die Merkmalsausprägungen.

Die y-Achse die Häufigkeit



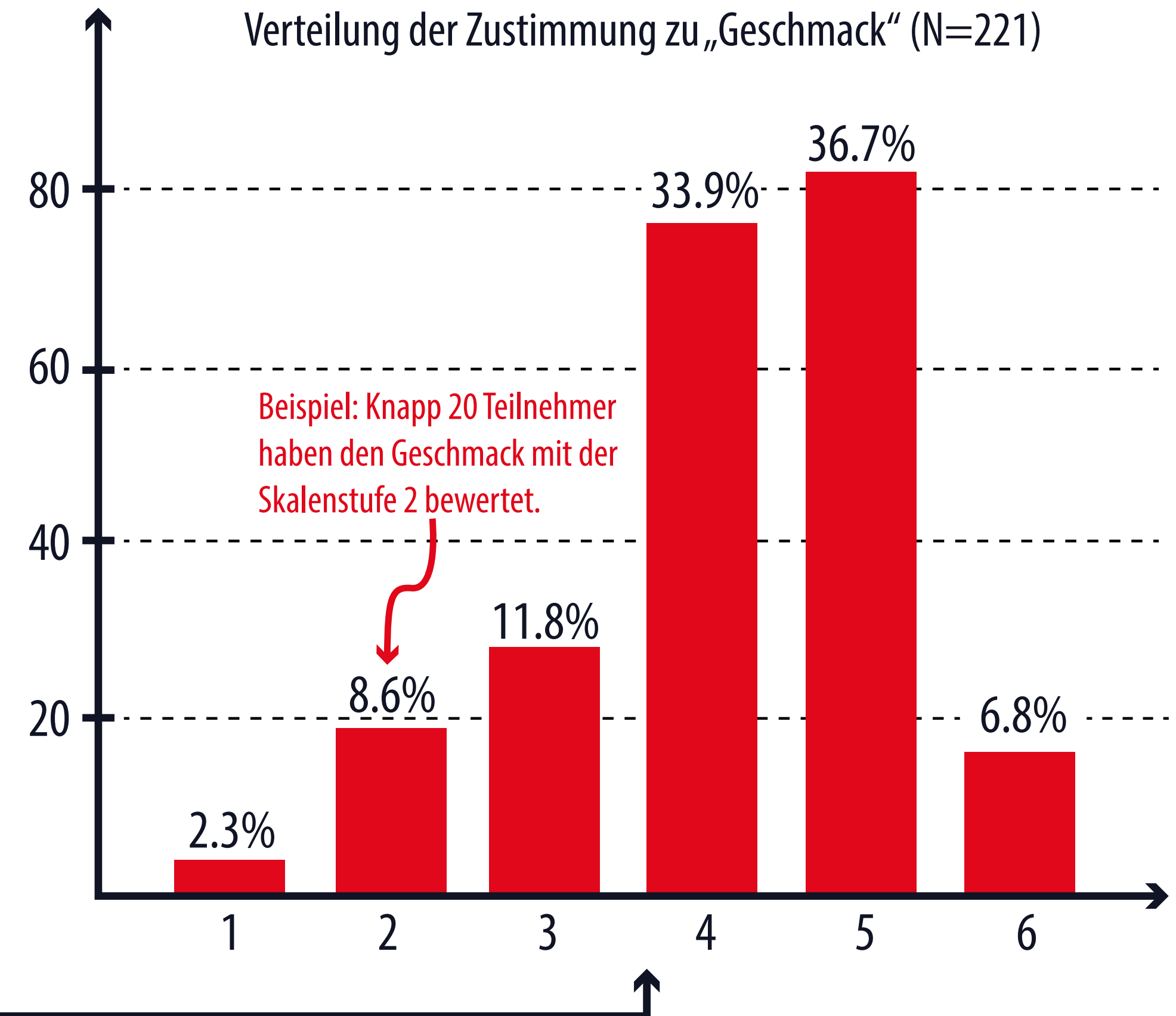
Datenquelle: Erhebung des Kurses WMKMM21B im Rahmen der Vorlesung „Empirische Methoden“

Balkendiagramm zu „Geschmack“



Die Merkmalsausprägungen kommen jetzt auf die x-Achse!

Verteilung der Zustimmung zu „Geschmack“ (N=221)



Histogramme

Wir unterscheiden die absolute und die relative Häufigkeit.

Absolute Häufigkeit - eine Zählung wie oft Ausprägungen eines Merkmals in der Stichprobe vorkommen.

Relative Häufigkeit - relativer Anteil von Ausprägungen in der Stichprobe.

Zustimmung	Teilnehmer	Anteil
1 - gar nicht	5	2.3%
2 -	19	8.6%
3 -	26	11.8%
4 -	75	33.9%
5 -	81	36.7%
6 - voll und ganz	15	6.8%

Histogramme

Im vorherigen Beispiel haben wir ein diskretes Merkmal mit insgesamt 6 verschiedene Merkmalsausprägungen betrachtet.

Dadurch gab es im Histogramm auch genau 6 Balken.

Wie können wir ein Histogramm für kontinuierliche Merkmale zeichnen? Schauen wir uns ein fiktives Beispiel an!

Umfrage Bayrischer Bierkonsum

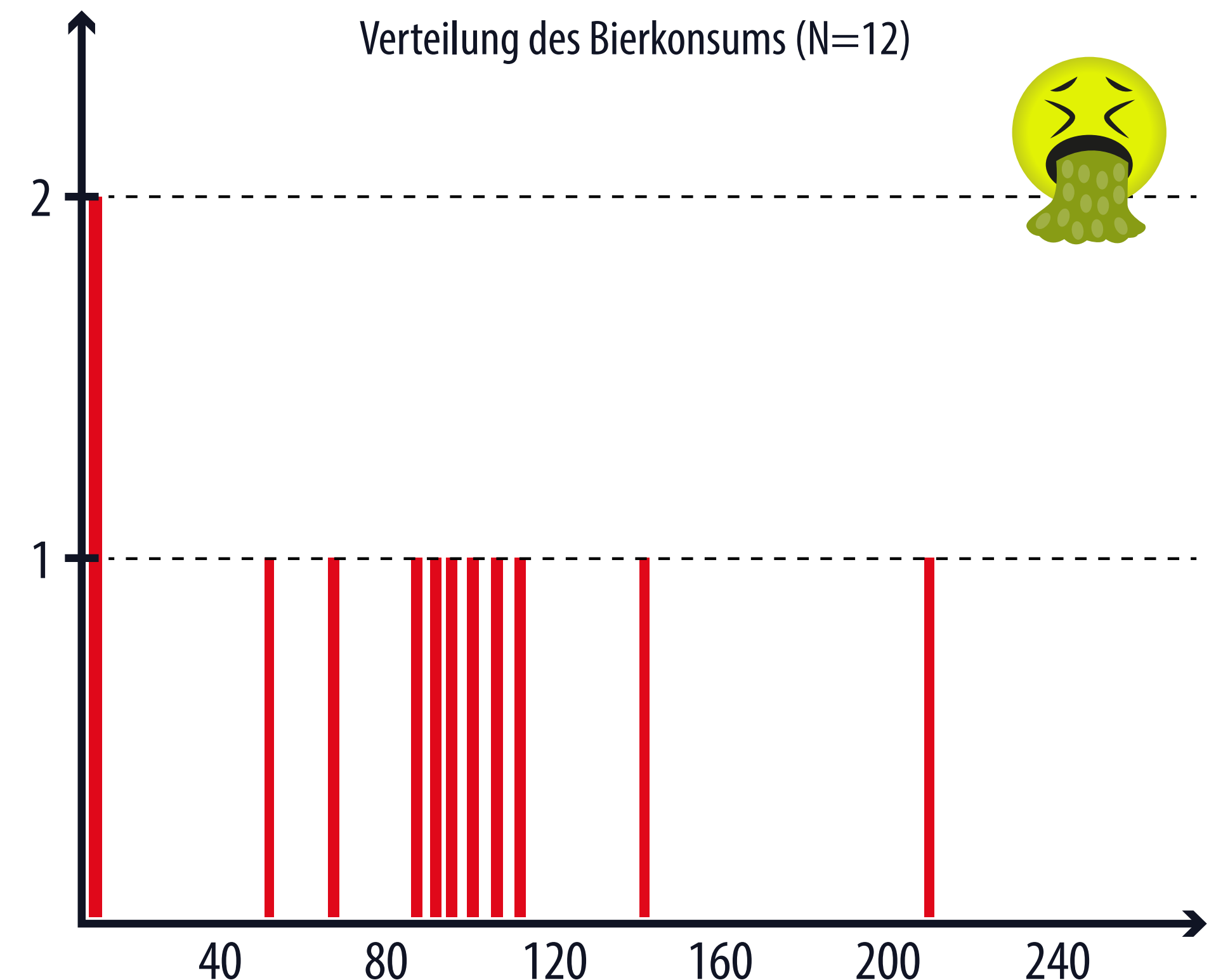
	117l / Jahr		140l / Jahr		205l / Jahr
	104l / Jahr		47l / Jahr		87l / Jahr
	98l / Jahr		0l / Jahr		96l / Jahr
	0l / Jahr		68l / Jahr		109l / Jahr

Histogramme

Würden wir jede Merkmalsausprägung als eigenen Balken zeichnen, erhalten wir das Schaubild rechts.

Es ist mathematisch gesehen nicht falsch, aber schlecht abzulesen ist. Die Balken müssen sehr dünn gewählt werden, da selbst kleinste Wertunterschiede zu einem separaten Balken führen.

Teilen wir dagegen die Ausprägungen in Klassen ein ...



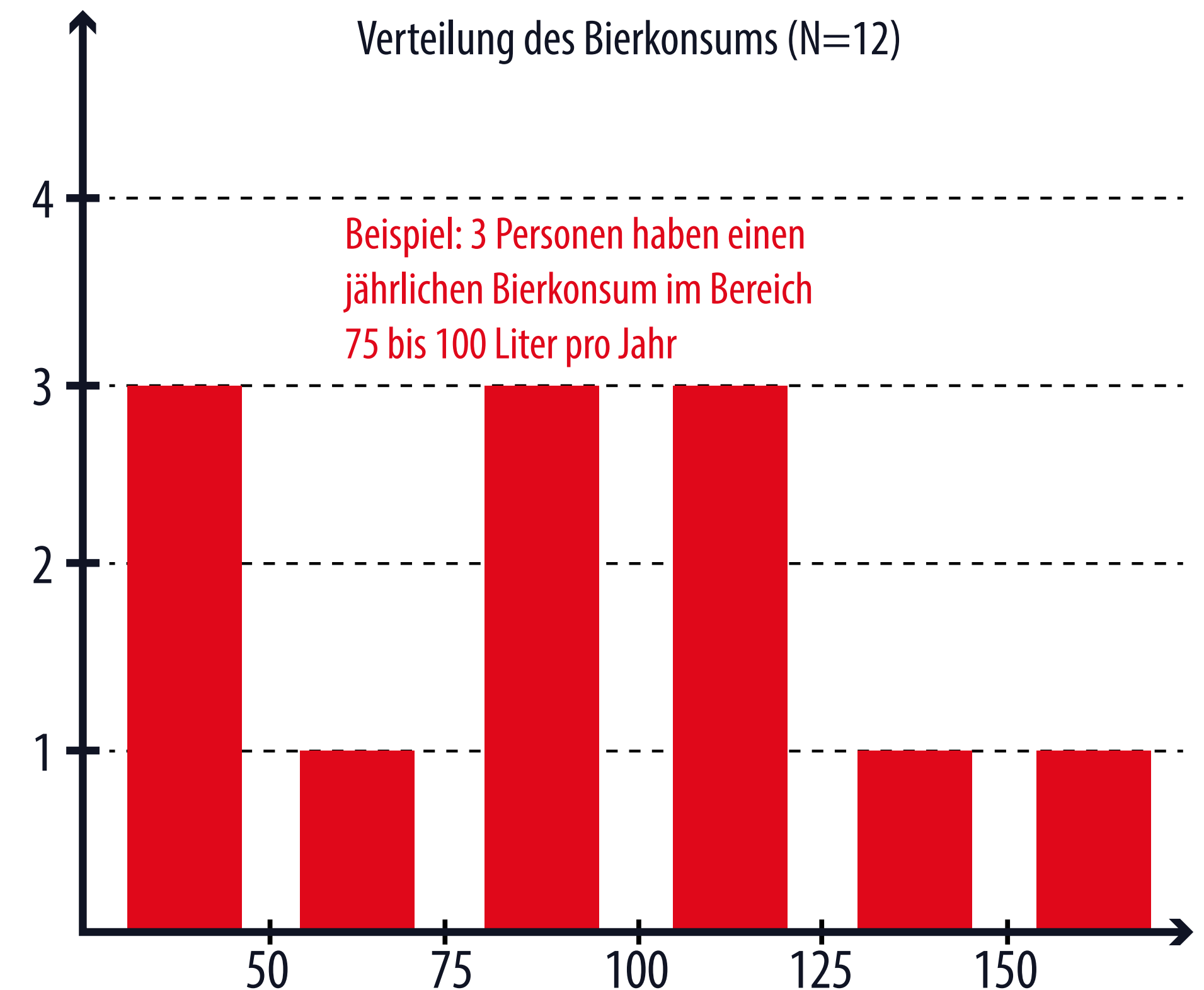
Histogramme

...ist das Schaubild deutlich besser abzulesen! Bei der Wahl der Klassen sind folgende Dinge zu beachten:

Die Intervalle sollten dieselbe Breite haben, hier z. B. 25

Am linken Rand kann eine Unterlauf-Klasse definiert werden. Im Beispiel: „Alle Werte unter 50l pro Jahr“

Am rechten Rand kann eine Überlauf-Klasse definiert werden. Im Beispiel: „Alle Werte über 150l pro Jahr“

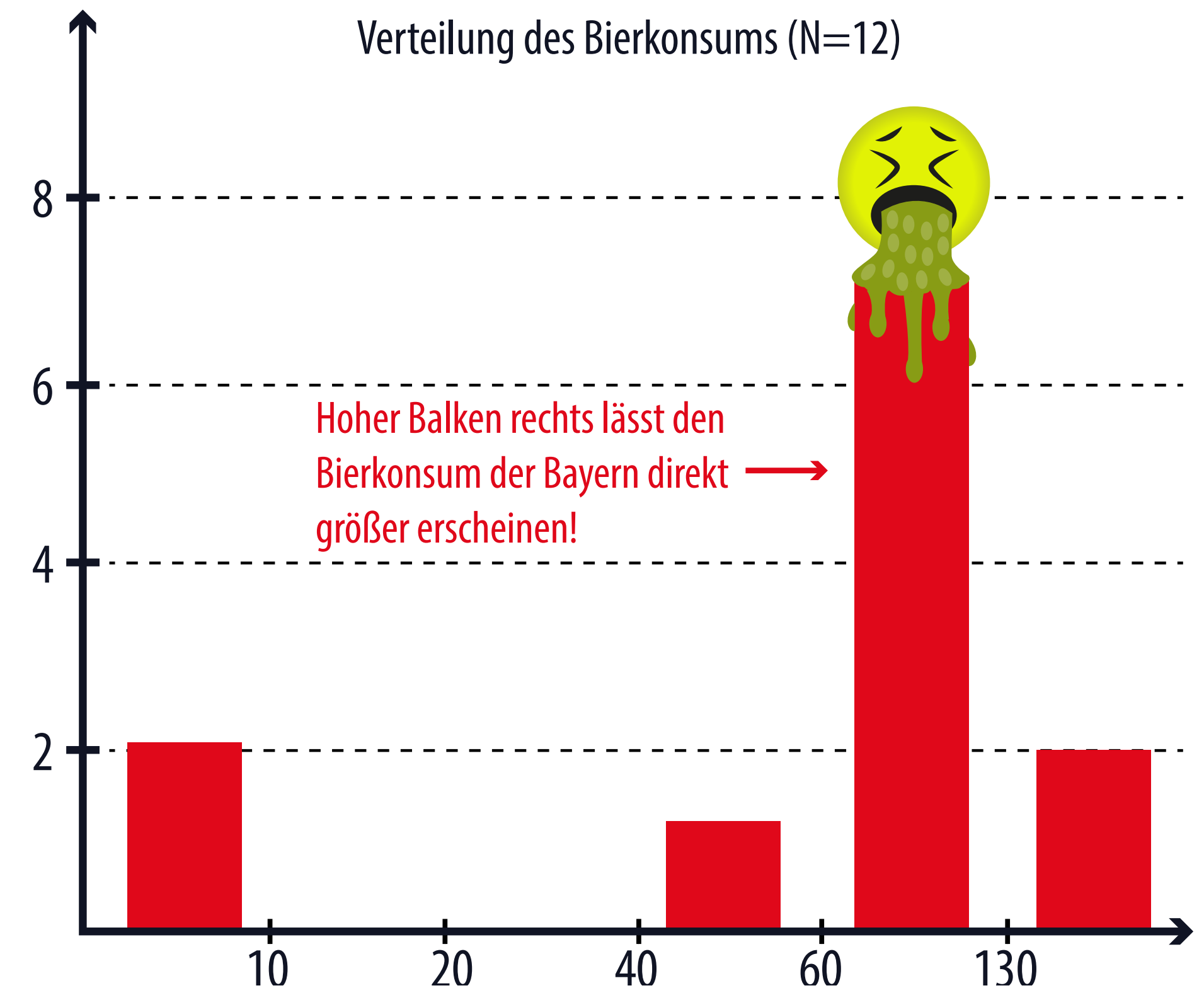


Histogramme

Unterschiedliche Klassenbreiten sollten vermieden werden!

Diese wirken im besten Fall als schlecht lesbar und im schlechtesten Fall als manipulativ.

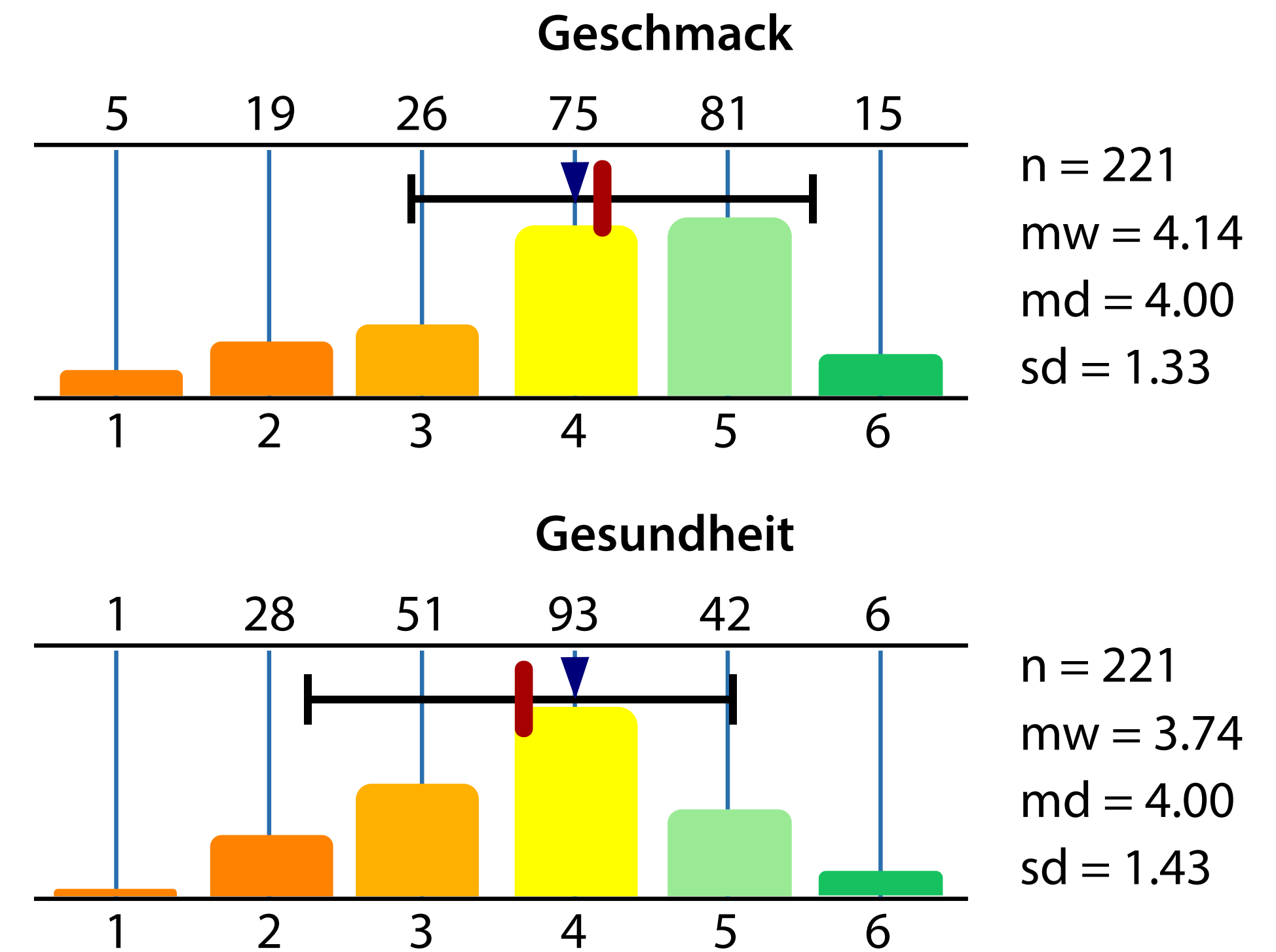
Ausnahmen sind natürlich die Unter- und Überlaufklassen, wobei diese nur Ausreißer einfangen sollten.



Histogramme

Histogramme können durch die Integration von Lage- und Streuungsparameter aufgewertet werden.

Die konkrete Umsetzung in Excel/Word ist mit den Funktionen Beschriftung und Textbox möglich, aber umständlich.



Box-Plot

Das Histogramm zeigt uns die Verteilung eines Merkmals.

Mit Box-Plots bzw. Box-Whisker-Plots können wir die Verteilungen mehrerer Merkmale vergleichen. Der Box-Plot zeigt zwar nicht die ganze Verteilung, aber ...

...das 1. und 3. Quartil als Kasten

...den Median als Linie im Kasten

...Ausreißer als Kreise

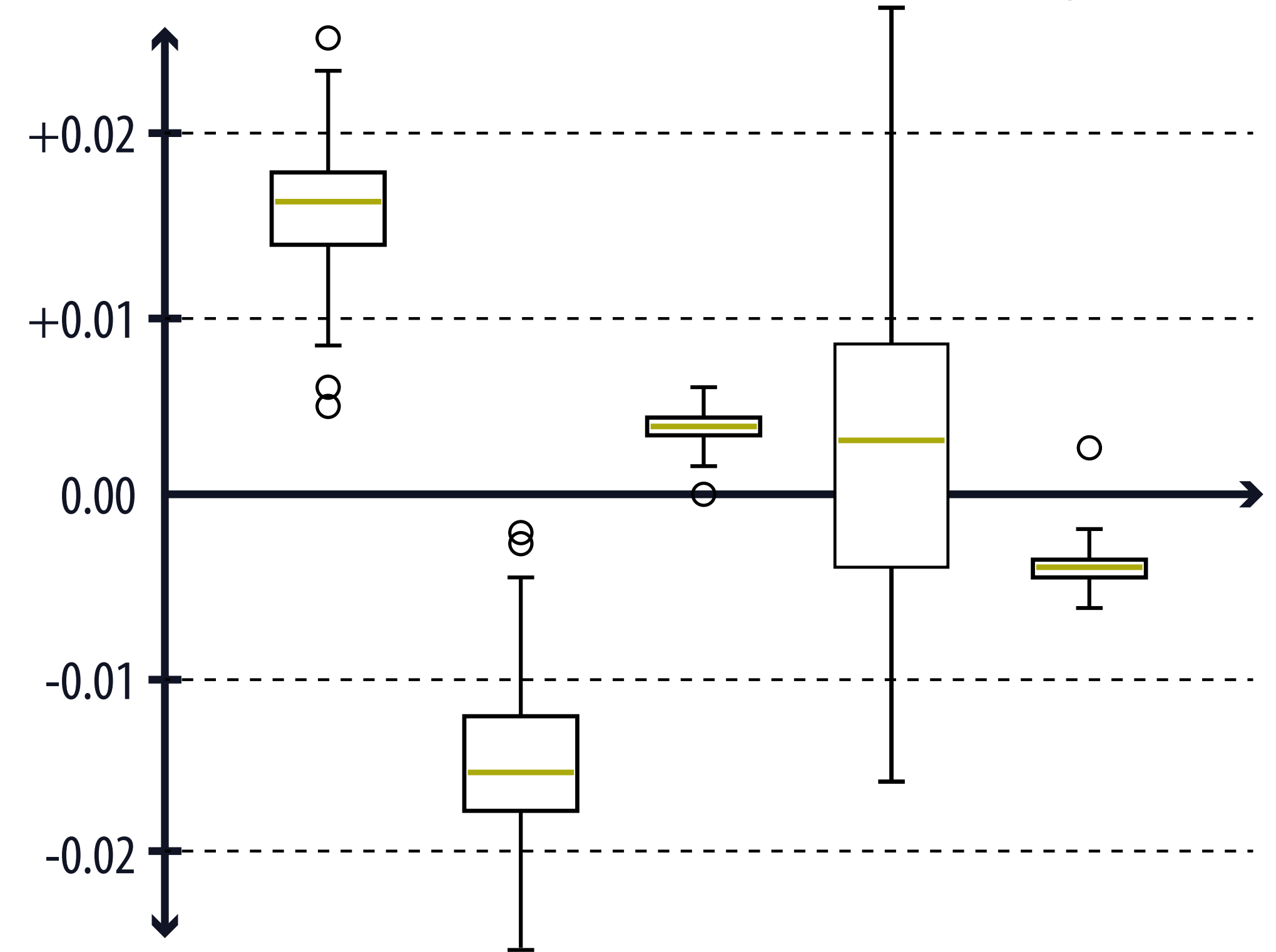


Abbildung 1 - Länge von M6x40mm Schrauben aus verschiedenen Drahtschneidemaschinen

Box-Plot

Für die namensgebenden Whisker (Antennen) gilt:

Sie erstrecken sich vom Ende der Box nach oben und unten.

Ihre Länge kann bis zum 1.5-fachen der Höhe der Box gehen, wird jedoch vom Wertebereich der Stichprobe begrenzt.

Alle Datenpunkte, die außerhalb der Whisker liegen, gelten als Ausreißer und werden mit Kreisen gezeigt.

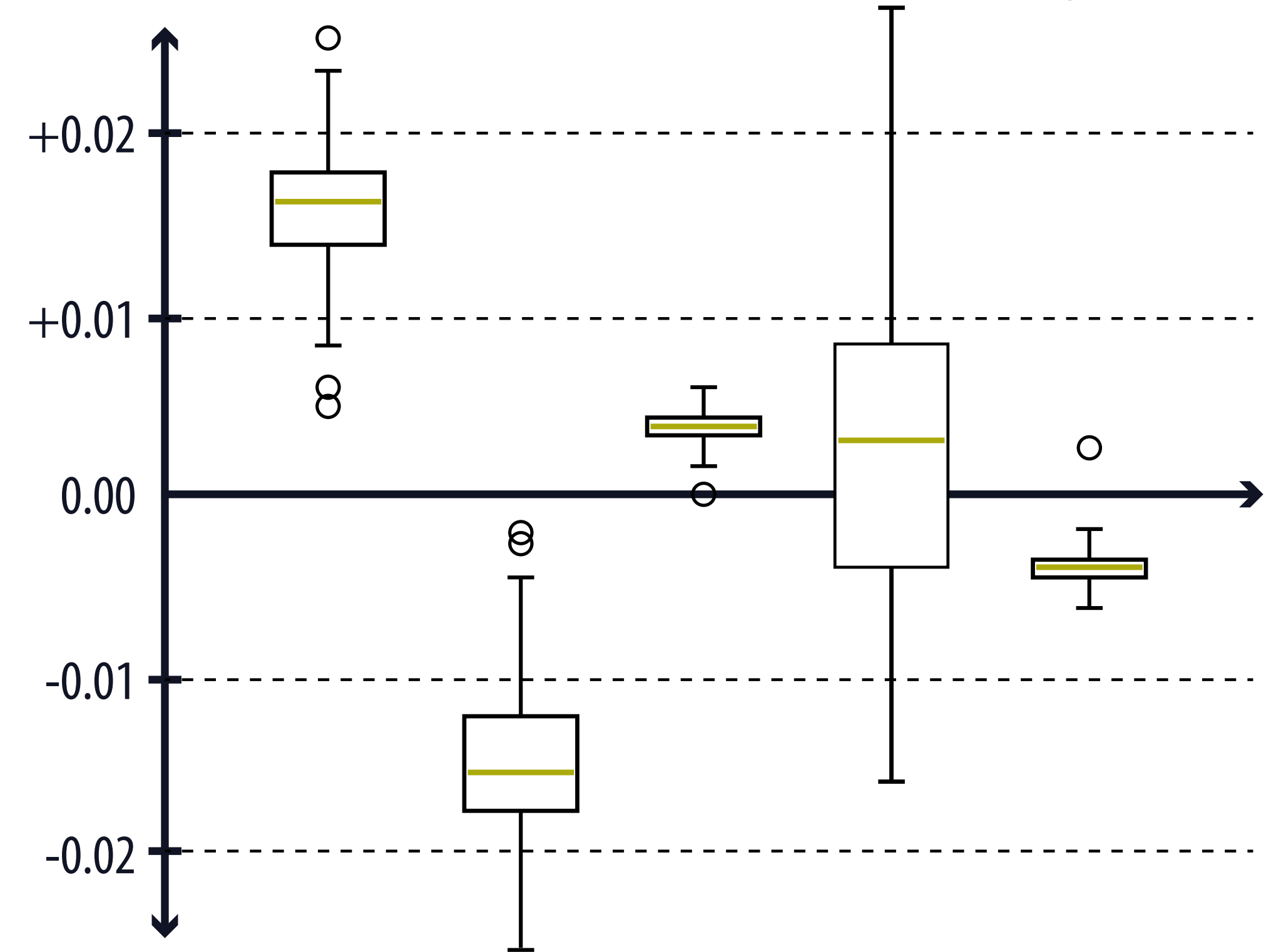


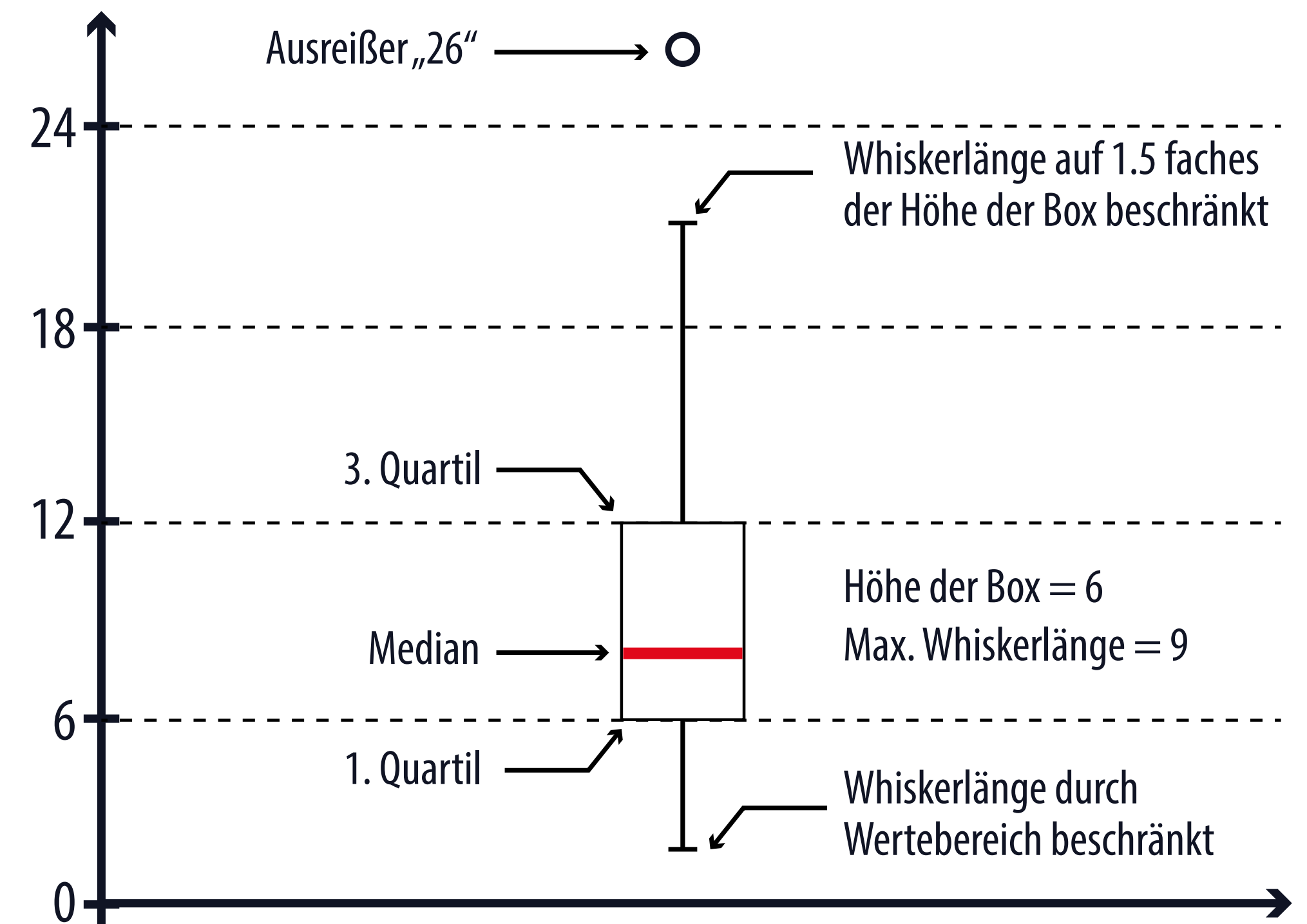
Abbildung 1 - Länge von M6x40mm Schrauben aus verschiedenen Drahtschneidemaschinen

Box-Plot

Kurzbeispiel: wir betrachten einen Datensatz mit einem Merkmal und einer Stichprobengröße von 9

4 8 6 10 2 7 14 26 12

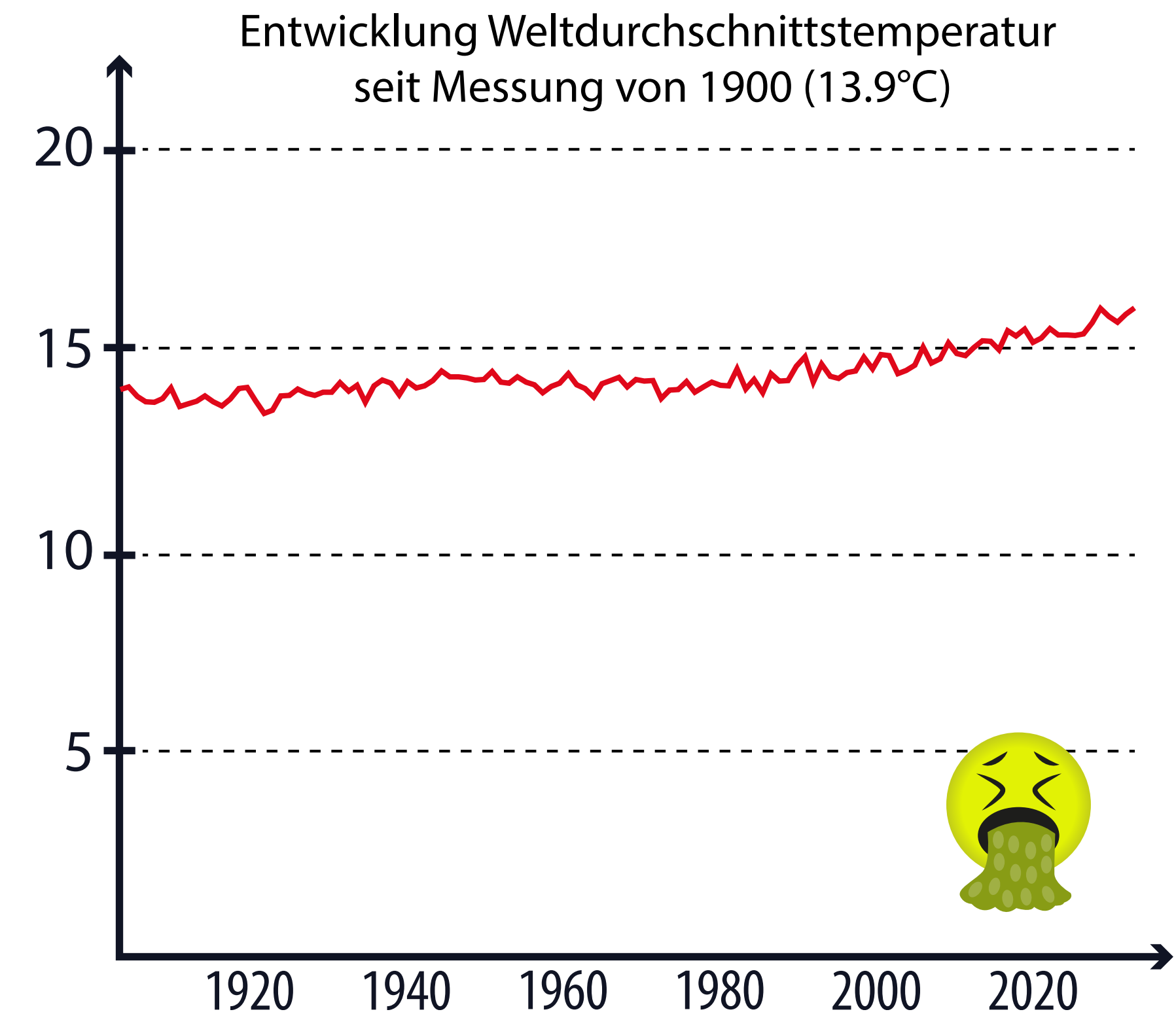
Wir erhalten die Quartile 6,8 und 12.



Liniendiagramme

Bei **Zeitreihen** sollten Liniendiagramme verwendet werden. Die Sortierung und Skalierung der x-Achse dürfte selbsterklärend sein.

Die Skalierung der y-Achse wird passend zu den Daten skaliert und muss nicht zwingend die 0 enthalten!

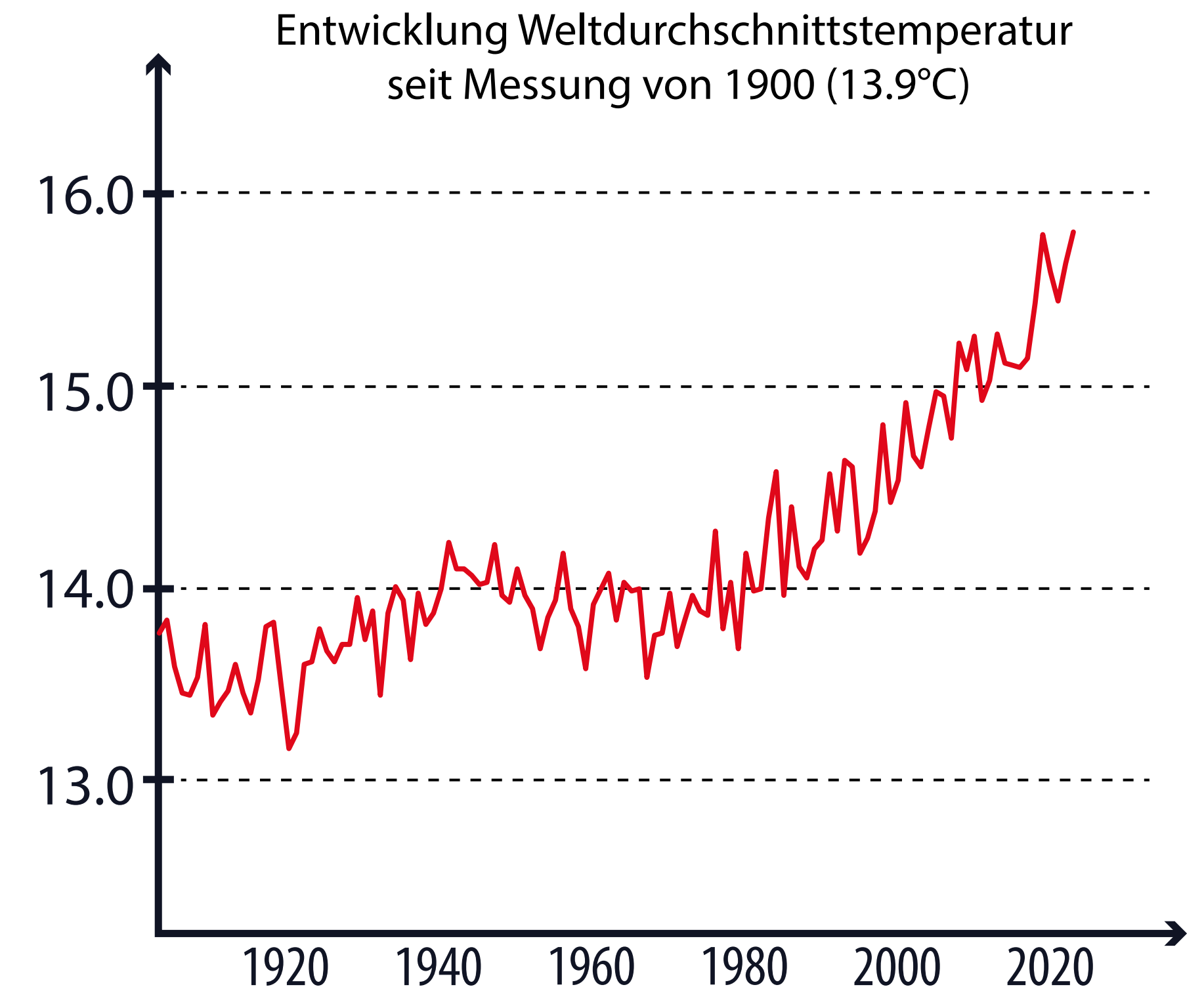


Datenquelle: NCEI (<https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/globe/tavg/land/ytd/12/1900-2024>)

Liniendiagramme

Bei **Zeitreihen** sollten Liniendiagramme verwendet werden. Die Sortierung und Skalierung der x-Achse dürfte selbsterklärend sein.

Die Skalierung der y-Achse wird passend zu den Daten skaliert und muss nicht zwingend die 0 enthalten!



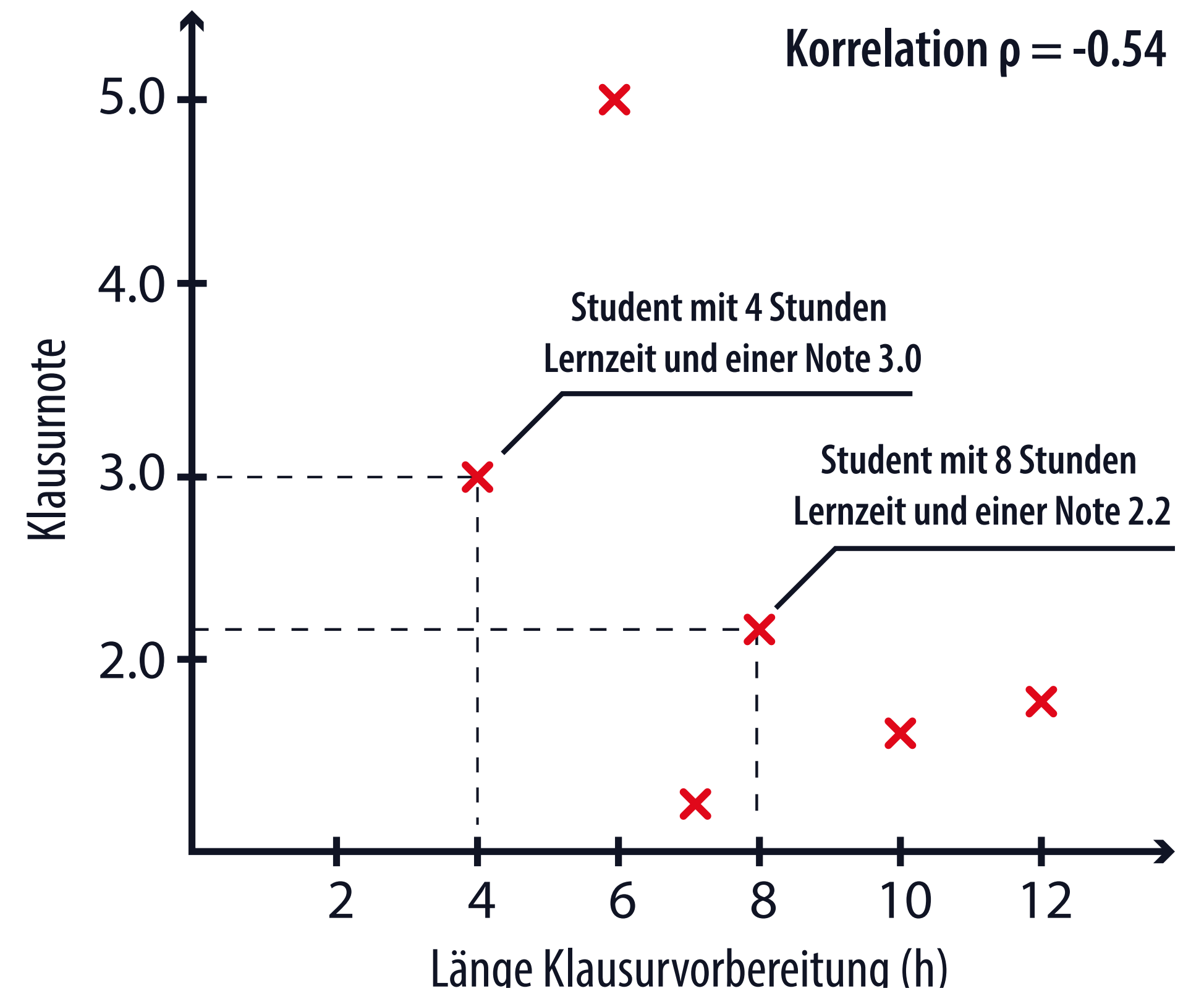
Datenquelle: NCEI (<https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/globe/tavg/land/ytd/12/1900-2024>)

Scatterplots

Mit Scatterplots können wir die Korrelation von zwei numerisch kontinuierlichen Merkmalen aufzeigen.

Der x- und y-Achse des Schaubilds wird je eines der Merkmale zugewiesen.

Jeder Merkmalsträger wird als Punkt eingezeichnet. Seine Merkmalsausprägungen definieren die Koordinaten.



Scatterplots

Bei diskreten Merkmalen sind Scatterplots zwar möglich, aber schwierig zu darstellen. Sehr viele Punkte liegen direkt übereinander.

Die Kontingenztabelle ist dann eine gute Alternative!

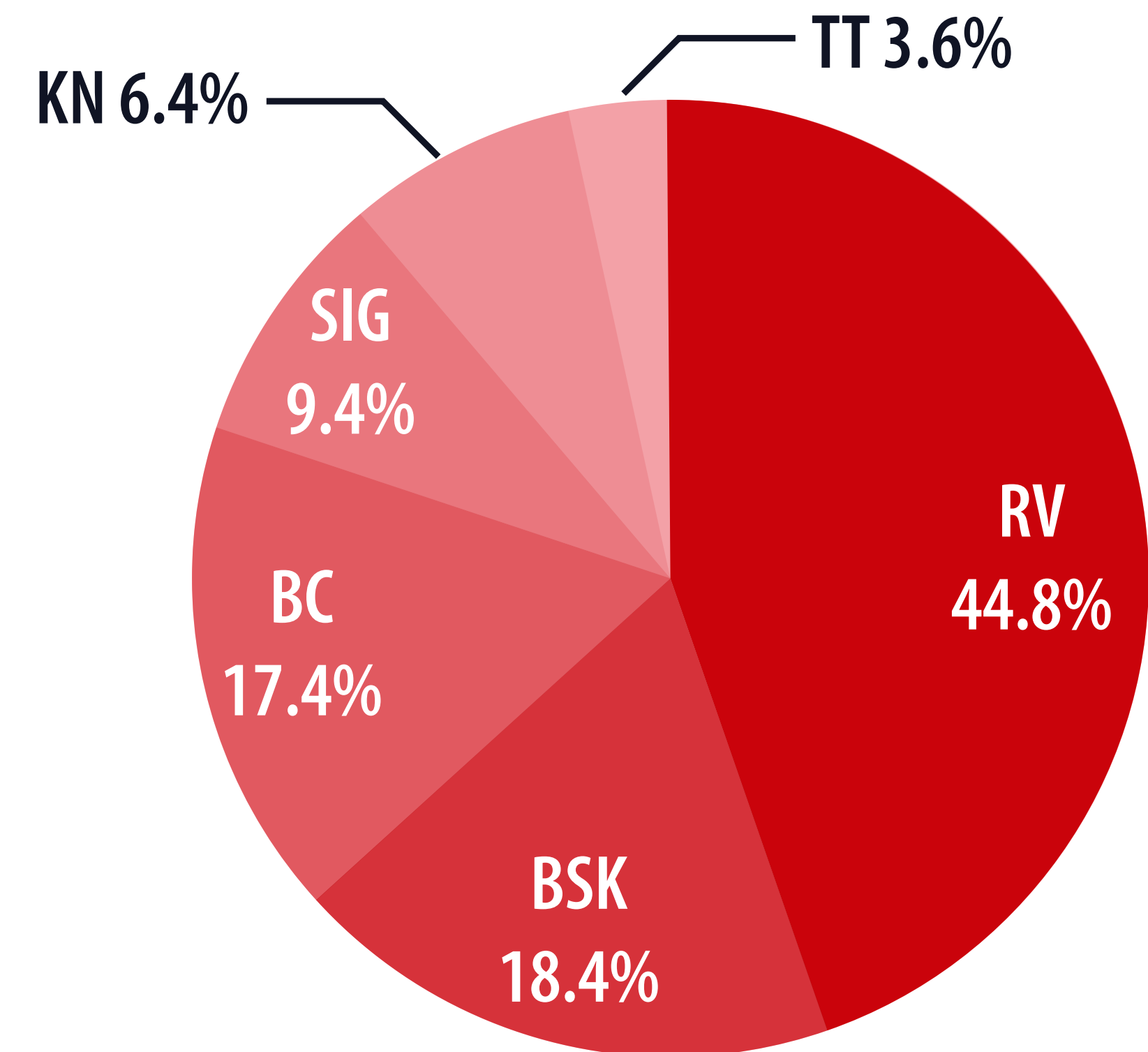
		Geschmack					
		1 Schlecht	2	3	4	5	6 Sehr gut
Gesundheit	1 Schlecht	0	0	0	1	0	0
	2	3	7	7	9	2	0
	3	0	10	6	22	13	0
	4	1	2	12	32	39	6
	5	1	0	1	10	24	7
	6 Sehr gut	0	0	0	1	3	2

Kuchendiagramm

Kuchendiagramme können zur Visualisierung von Häufigkeiten eingesetzt werden, sind aber weniger gut abzulesen wie ein Balkendiagramm.

Auch die Beschriftung ist oft schwieriger als bei einem Balkendiagramm, insbesondere bei seltenen Ausprägungen.

Glaubt mir keiner? Schauen wir uns das Beispiel mit der Stichprobe nach Landkreisen noch mal an ...



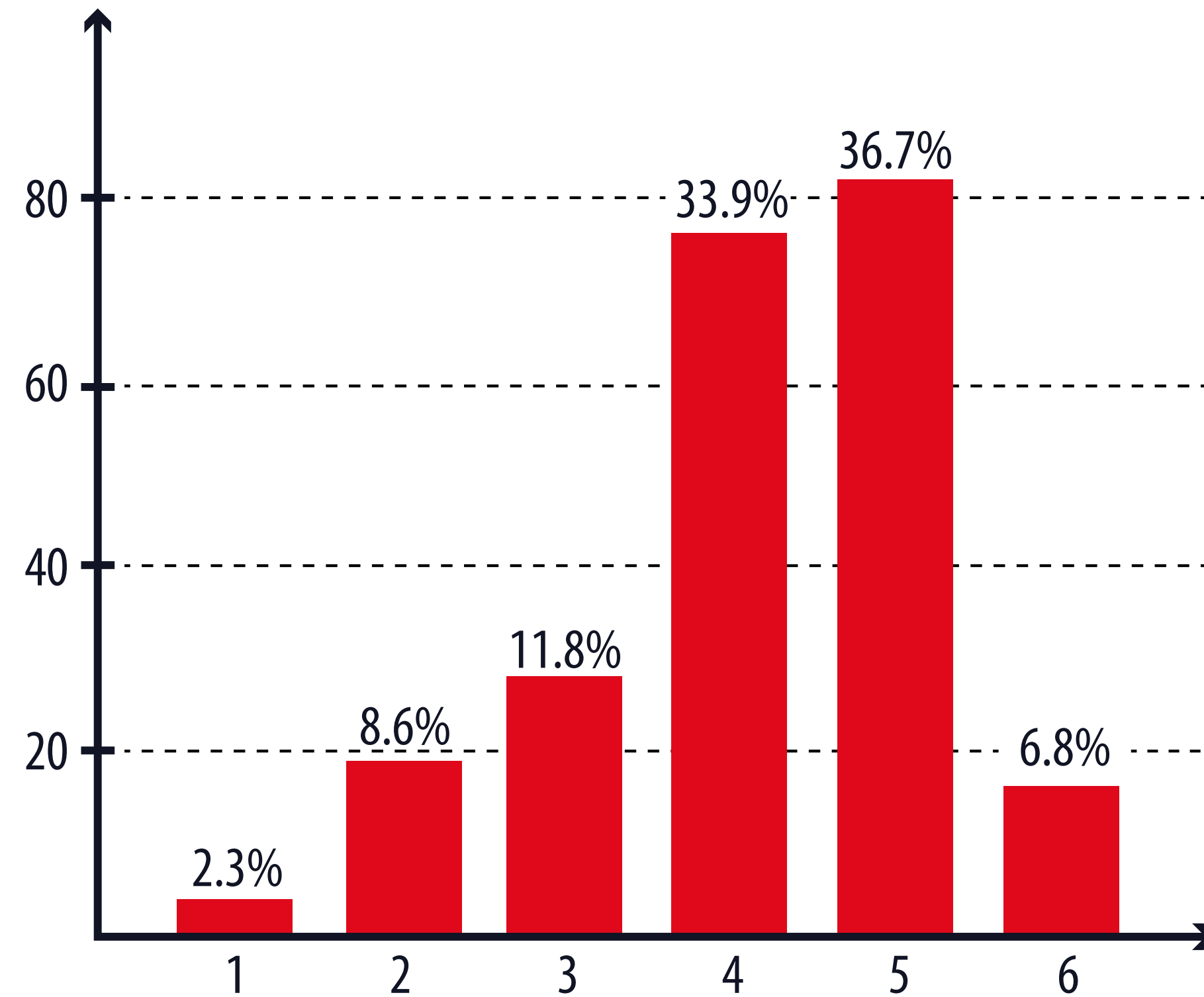


Abbildung 1 - Histogramm der Bewertungen
im Kriterium Geschmack (N=221)

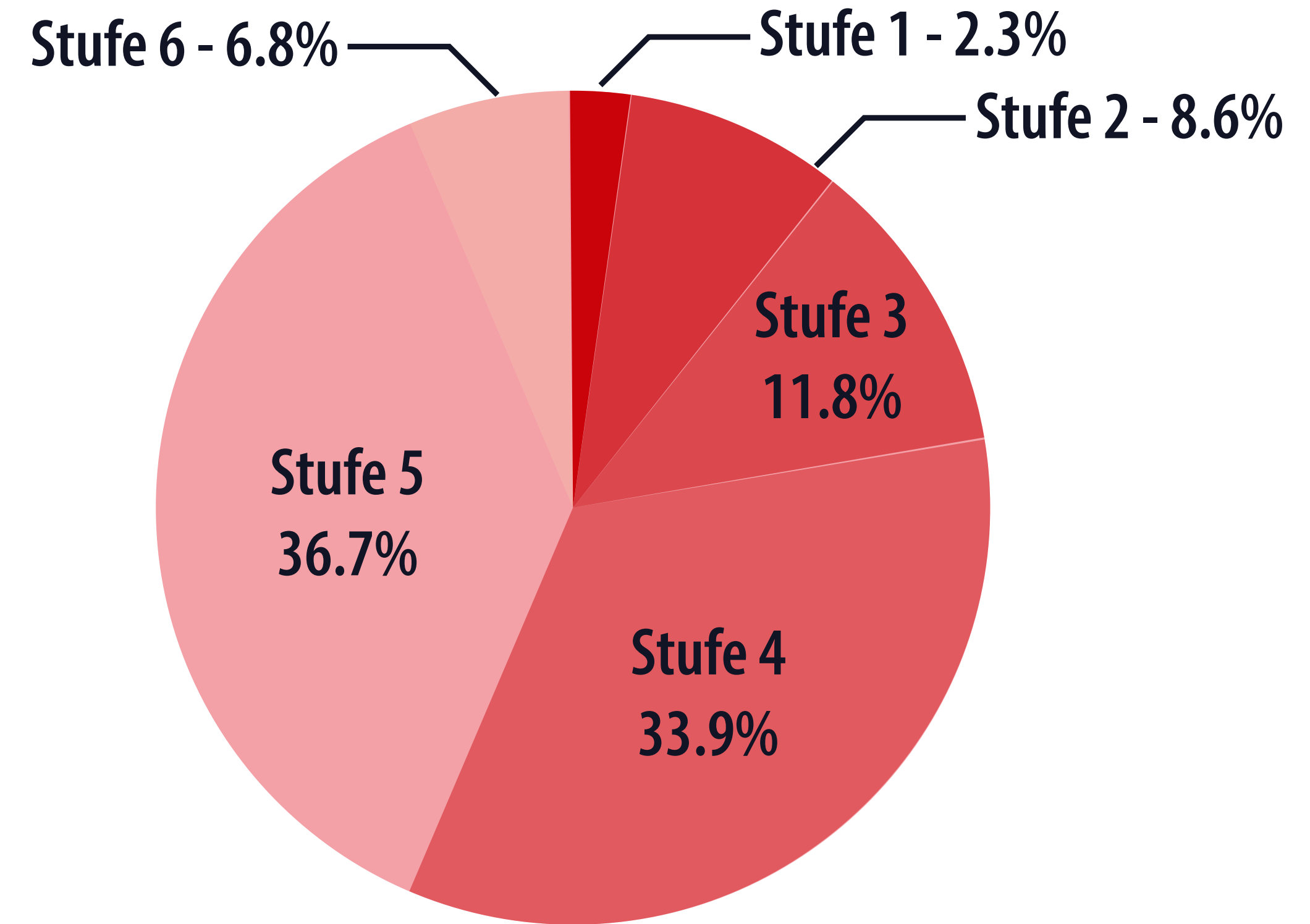


Abbildung 2 - Das gleiche als Kuchendiagramm.
Nicht schöner, aber seltener!

Anwendung in PA/BA

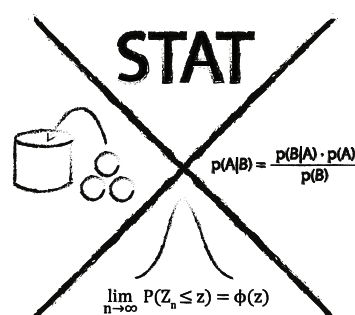
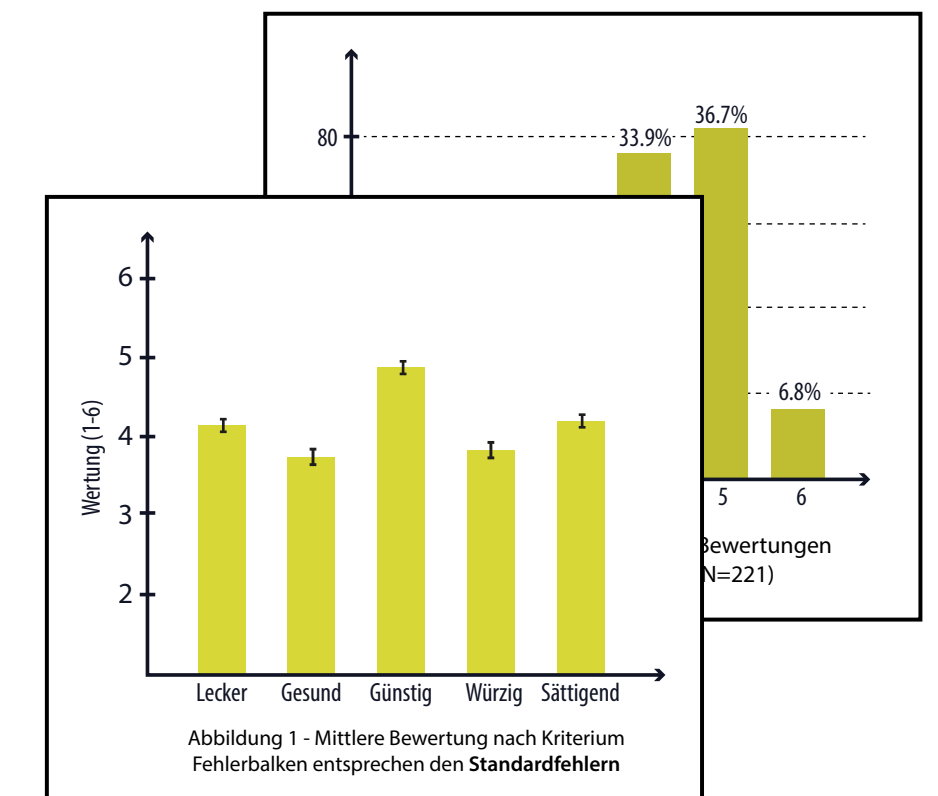
Wie integriere ich meine Exceltabellen und Diagramme in meine Projekt- oder Bachelorarbeit? **Auf keinen Fall über Screenshots oder Snipping Tool!**

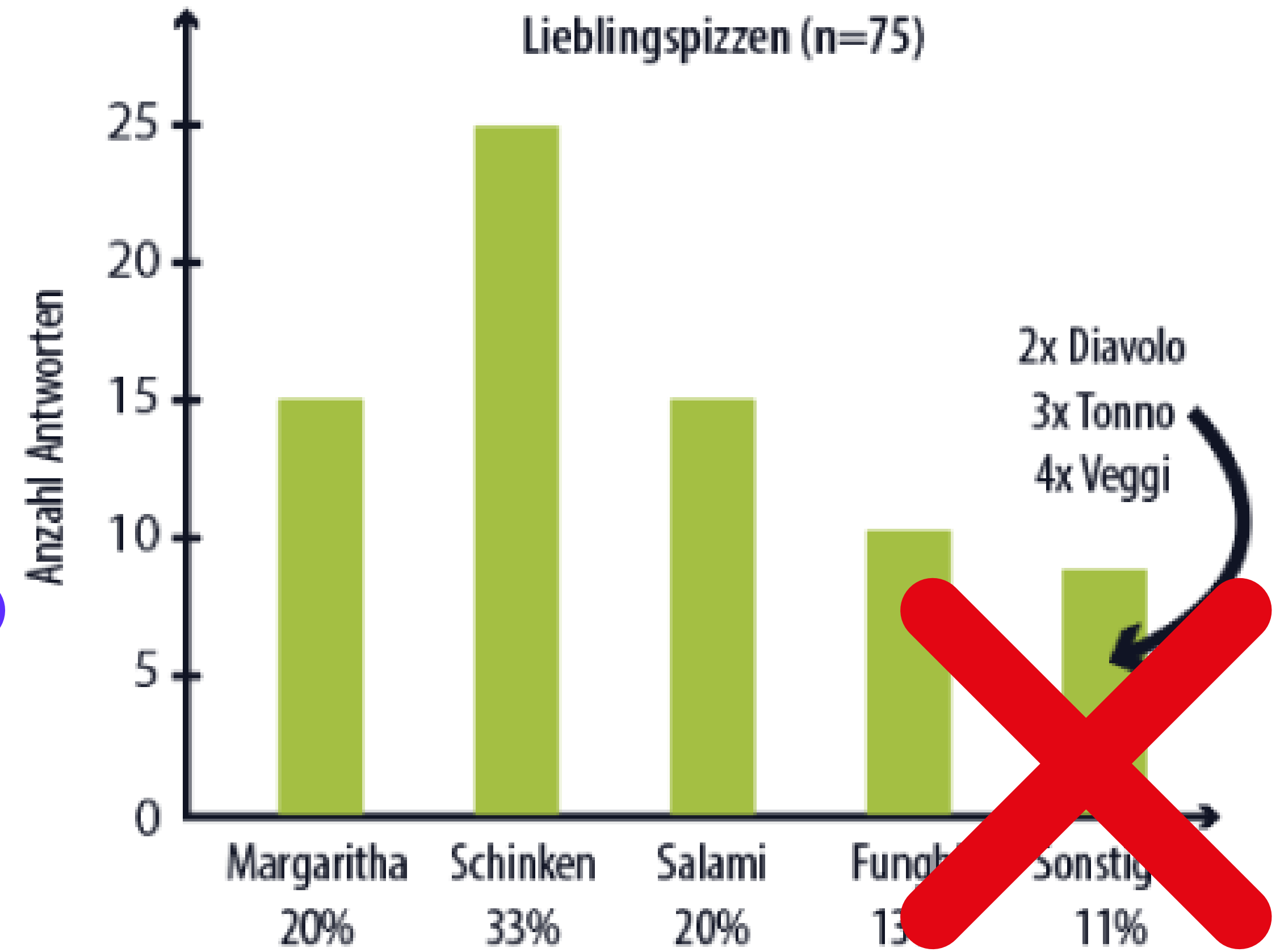
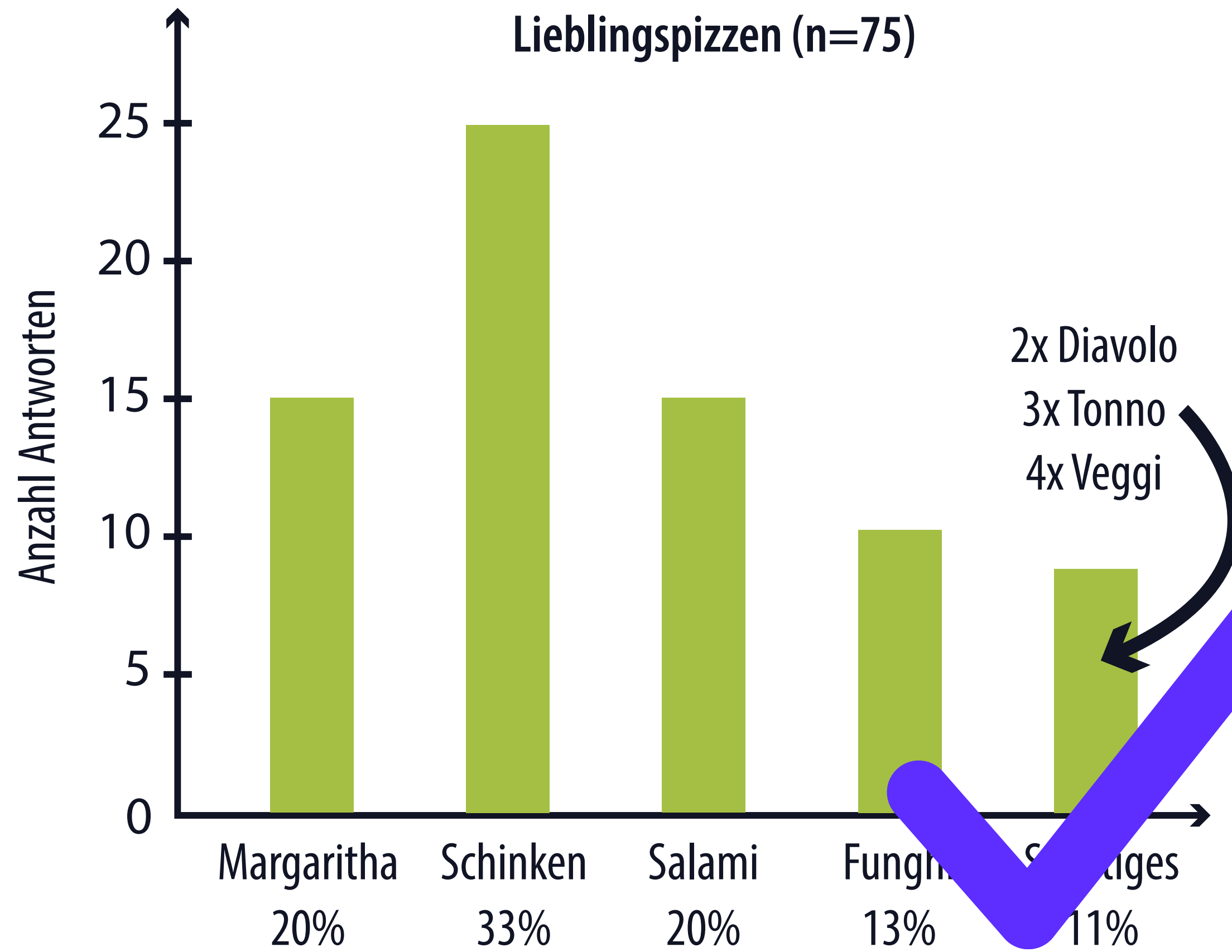
Word Direkt per Copy & Paste aus Excel

Creative Cloud Copy & Paste in Illustrator, Upload in Library, Einbindung in Frame in InDesign

Latex Export als PDF und Einbindung über den LaTeX Befehl `\includegraphics` in einer figure Umgebung

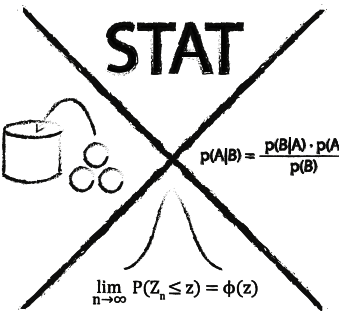
???





Nr.	Fellfarbe	Tage seit Kalbung	Gewicht	Milchleistung
1	Gefleckt	58	867,5	32,7
2	Gefleckt	178	660,6	24,3
3	Schwarz	2	777,6	18,1
4	Gefleckt	14	737,8	21,4
5	Gefleckt	48	781,3	30,3
6	Braun	112	786,6	7,8
7	Braun	222	839,8	2,8
8	Schwarz	153	746,3	14,7
9	Braun	44	766,8	19,2
10	Braun	222	748,8	1,0
11	Gefleckt	93	738,8	30,6
12	Braun	194	762,8	2,0
13	Braun	183	810,3	3,3
14	Braun	236	742,1	0,6
15	Schwarz	178	723,8	11,5
16	Gefleckt	167	711,3	26,1
17	Schwarz	60	732,0	27,7
18	Braun	194	802,7	2,7
19	Schwarz	78	771,6	15,9
20	Braun	120	629,5	1,1
21	Braun	248	706,2	-0,1

Nr.	Fellfarbe	Tage seit Kalbung	Gewicht	Milchleistung
1	Gefleckt	58	867,5	32,7
2	Gefleckt	178	660,6	24,3
3	Schwarz	2	777,6	18,1
4	Gefleckt	14	737,8	21,4
5	Gefleckt	48	781,3	30,3
6	Braun	112	786,6	7,8
7	Braun	222	839,8	2,8
8	Schwarz	153	746,3	14,7
9	Braun	44	766,8	19,2
10	Braun	222	748,8	1,0
11	Gefleckt	93	738,8	30,6
12	Braun	194	762,8	2,0
13	Braun	183	810,3	3,3
14	Braun	236	742,1	0,6
15	Schwarz	178	723,8	11,5
16	Gefleckt	167	711,3	26,1
17	Schwarz	60	732,0	27,7
18	Braun	194	802,7	2,7
19	Schwarz	78	771,6	15,9
20	Braun	120	629,5	1,1
21	Braun	248	706,2	-0,1



Anwendung in PA/BA

Auf was sollte ich sonst noch achten?

Nummerierung Jede Tabelle/Abbildung benötigt eine Bildunterschrift mit dem Schlüsselwort, einer fortlaufenden Nummer und einem kurzen Erklärungstext.

Querverweise Jede Tabelle oder Abbildung sollte mindestens ein Mal vom Text aus referenziert werden.

Quellenangaben Keine Fußnoten in der Bildunterschrift! Die Bildunterschrift muss etwaige Quellenangaben direkt enthalten.

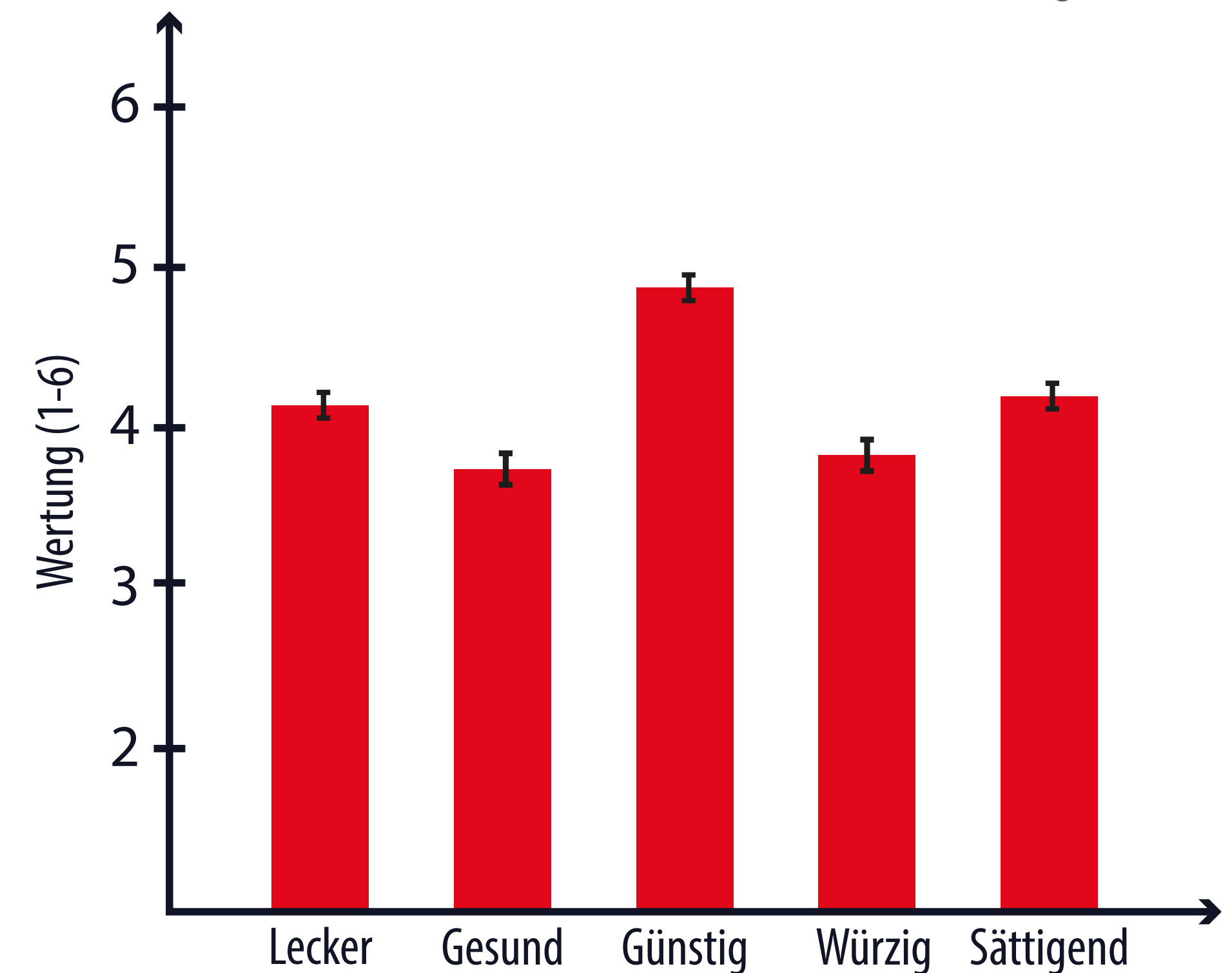


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den **Standardfehlern**

Anwendung in PA/BA

Konsistenz Die Abbildungen sollten einheitliche Gestaltungsmerkmale (Größe, Schriftart, Farben) aufweisen.

Aspect Ratio Beim Vergrößern und Verkleinern muss das Längenverhältnis erhalten bleiben, ansonsten sehen insbesondere Beschriftungen verzerrt aus.

Platzierung Die Abbildungen sollte nicht über den Textblock herausragen und in der Nähe der Textstelle sein, von der aus auf sie verwiesen wird.

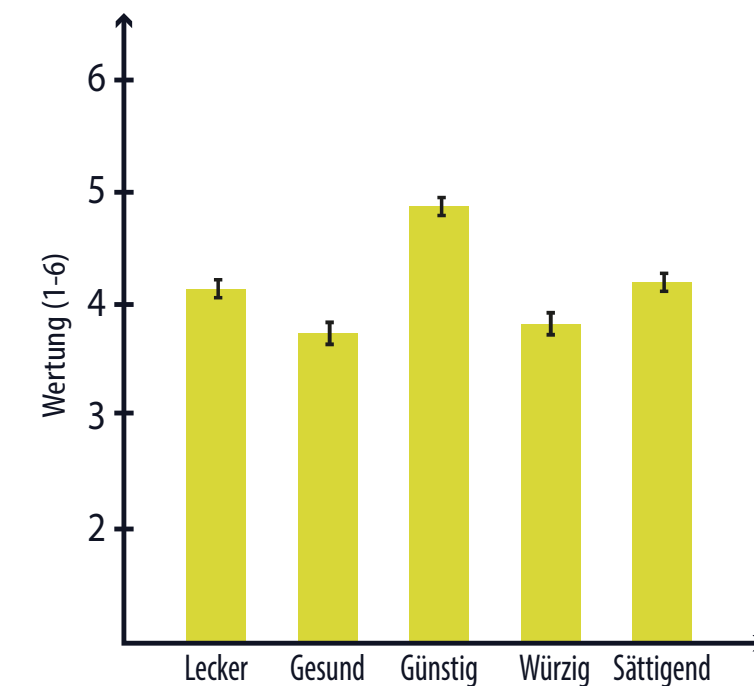
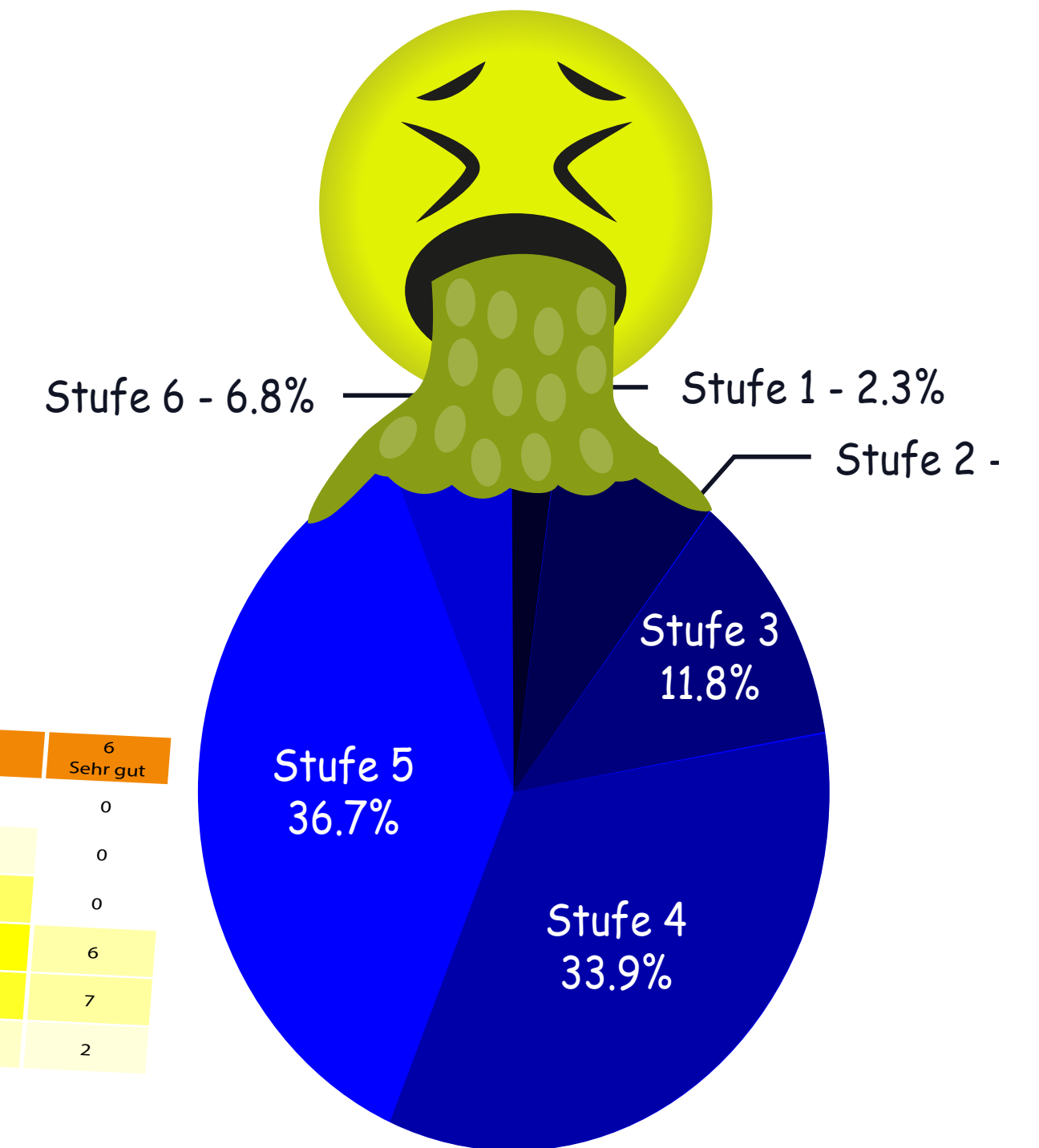


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den Standardfehlern

Geschmack						
	1 Schlecht	2	3	4	5	6 Sehr gut
1 Schlecht	0	0	0	1	0	0
2	3	7	7	9	2	0
3	0	10	6	22	13	0
4	1	2	12	32	39	6
5	1	0	1	10	24	7
6 Sehr gut	0	0	0	1	3	2



Anwendung in PA/BA

Anhang Tabellen und Schaubilder, die für das Verständnis der Arbeit wichtig sind, gehören auf keinen Fall in den Anhang!

Versetzen wir uns dazu in unseren Leser bzw. Gutachter hinein. Möchte er beim Lesen unserer statistischen Auswertung ständig mehrere Seiten nach hinten und wieder zurück blättern?

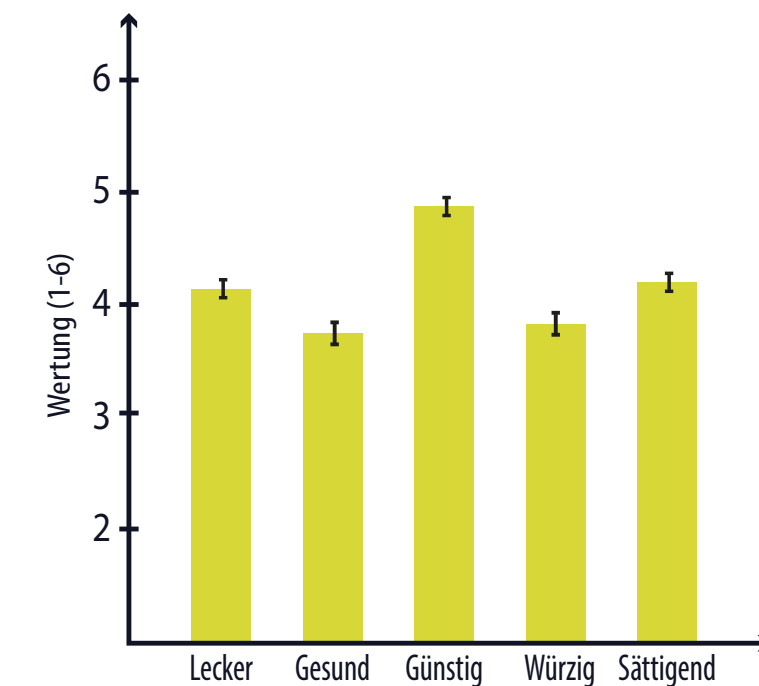
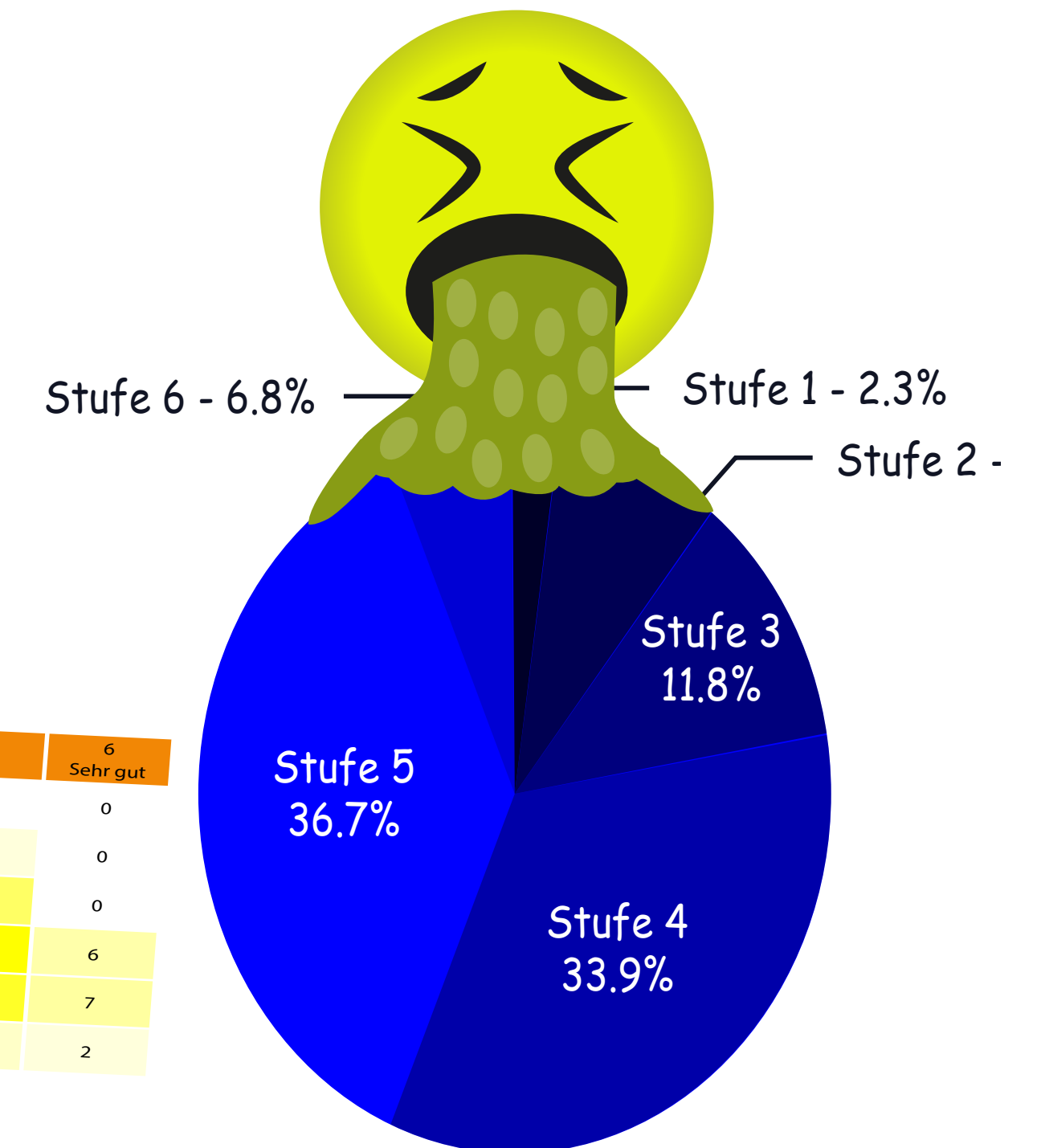


Abbildung 1 - Mittlere Bewertung nach Kriterium
Fehlerbalken entsprechen den Standardfehlern

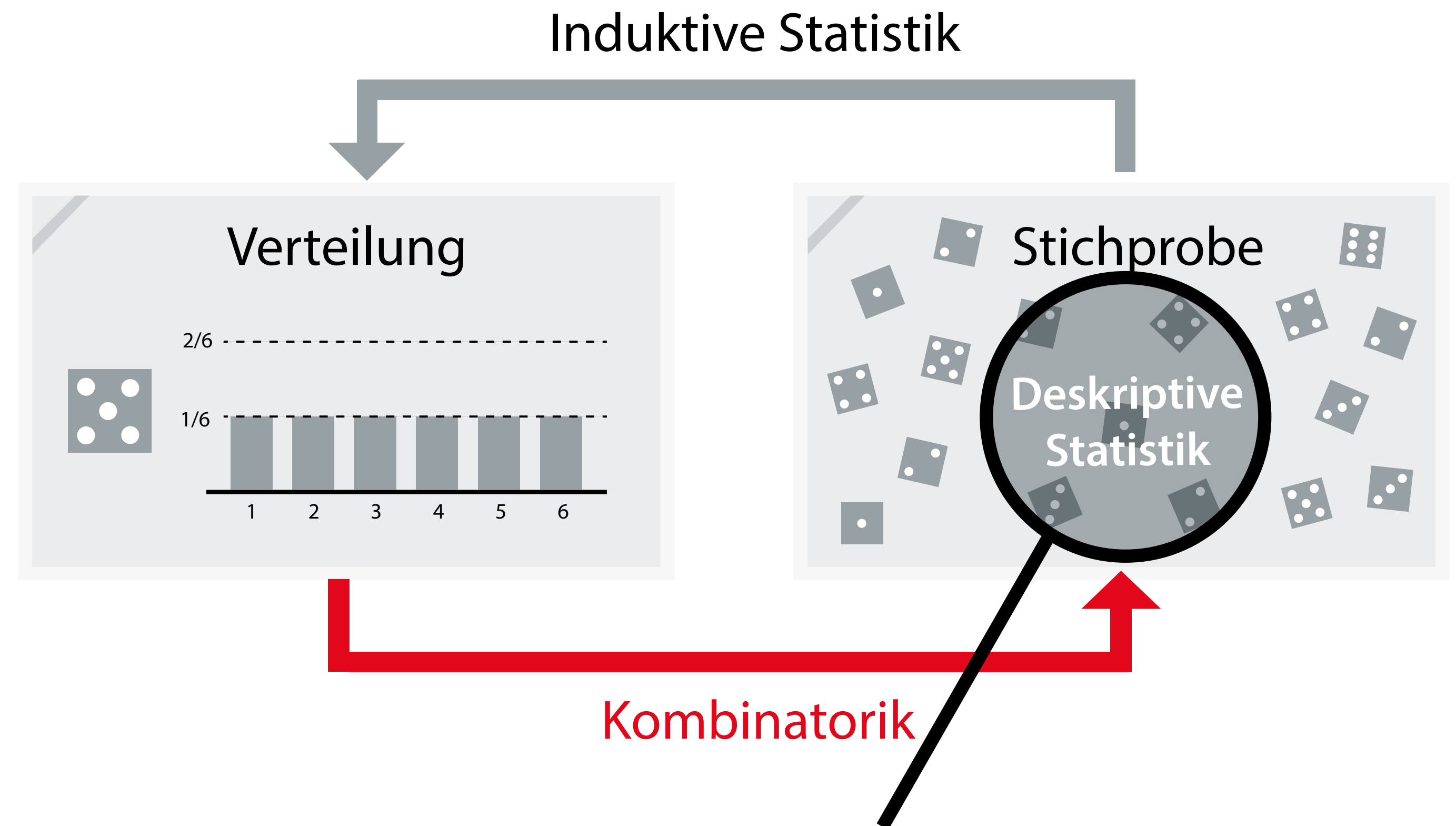
	Geschmack					
	1 Schlecht	2	3	4	5	6 Sehr gut
1 Schlecht	0	0	0	1	0	0
2	3	7	7	9	2	0
3	0	10	6	22	13	0
4	1	2	12	32	39	6
5	1	0	1	10	24	7
6 Sehr gut	0	0	0	1	3	2



Kombinatorik

Kombinatorik berechnet aus einer gegebenen Verteilung Wahrscheinlichkeiten.

Wir beginnen auch hier mit einigen Begrifflichkeiten ...



Kombinatorik

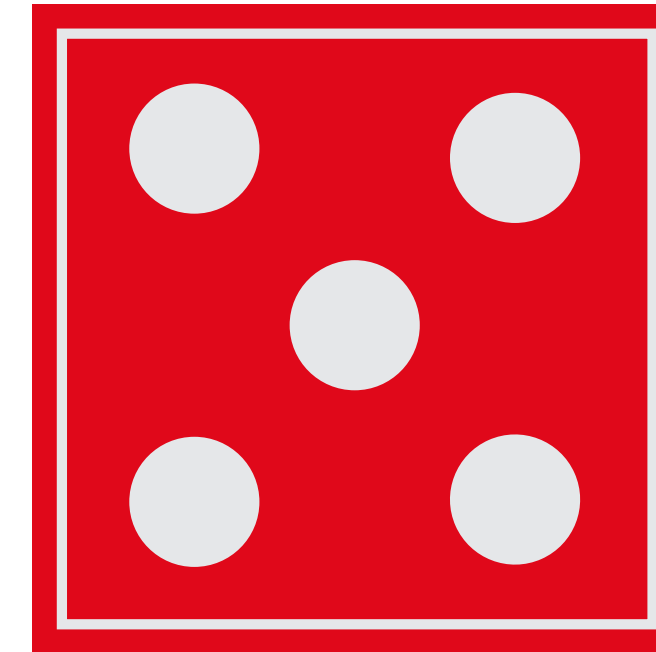
Als Erstes lernen wir die **Ereignismenge** Ω kennen. Diese enthält alle Elementarereignisse, die eintreten können.

Würfel $\Omega = \{1,2,3,4,5,6\}$

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

Roulette $\Omega = \{00,0,1,2,3,\dots,36\}$

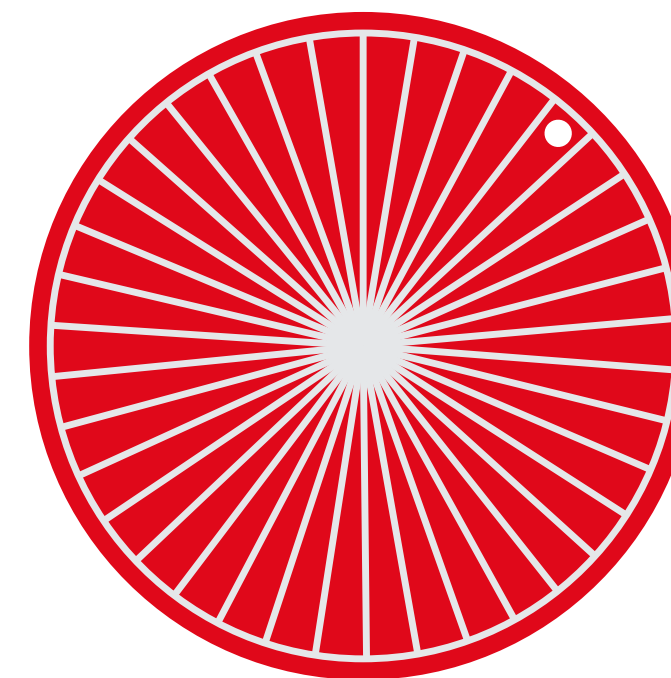
Elementarereignisse sind so gesehen das Pendant zu den Merkmalsausprägungen der deskriptiven Statistik!



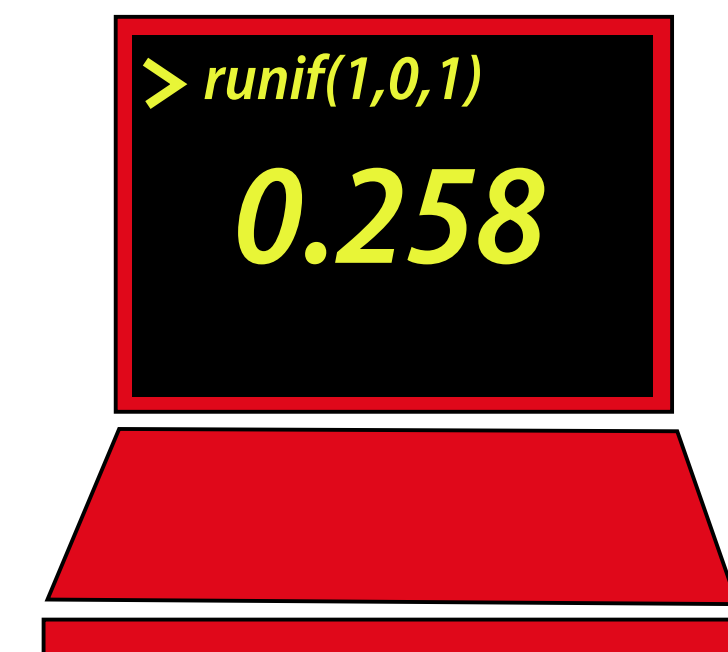
$$\Omega = \{1,2,3,4,5,6\}$$



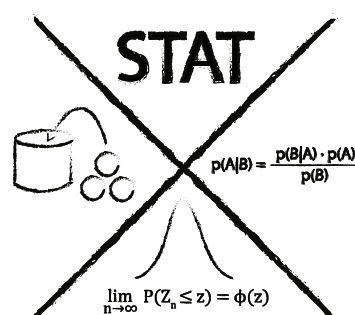
$$\Omega = \{\text{Kopf, Zahl}\}$$



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



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



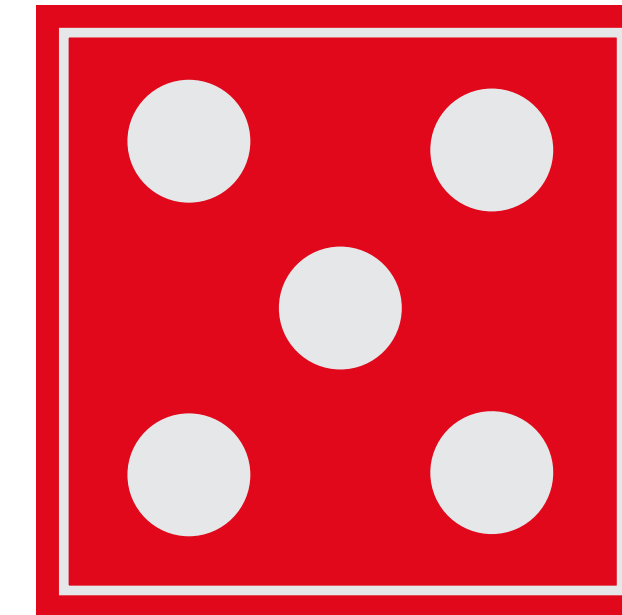
Kombinatorik

Ereignismengen können diskret oder kontinuierlich sein.

Diskrete Ereignismengen enthalten endlich oder abzählbar unendlich viele Elementarereignisse.

Kontinuierliche Ereignismengen enthalten nicht-abzählbar unendlich viele Elementarereignisse.

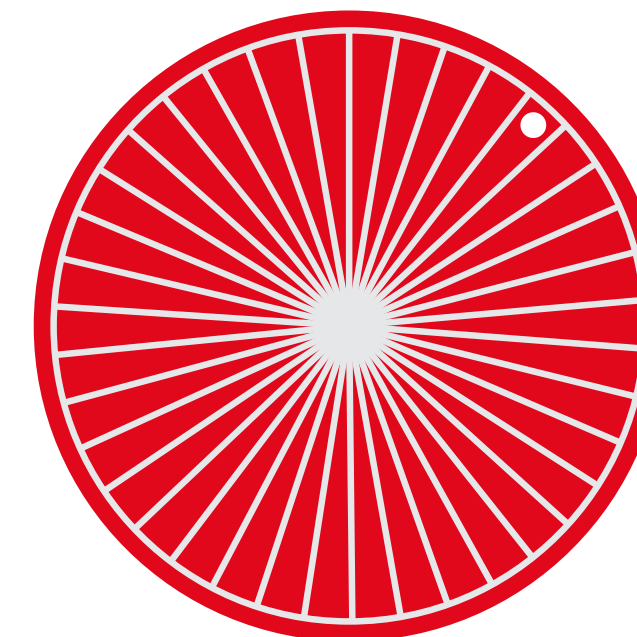
Bei Letzterem kann die Ereignismenge oft durch ein Intervall angegeben werden, z. B. $\Omega = (0,1)$.



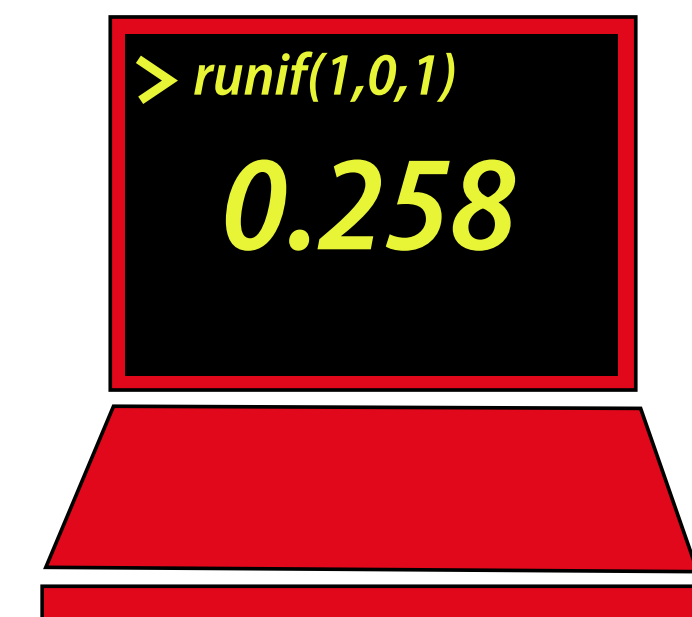
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



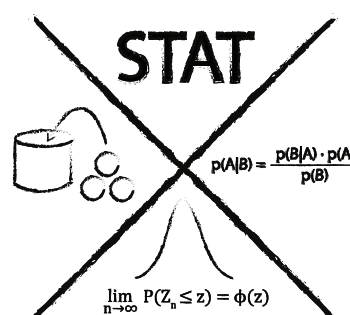
$\Omega = \{\text{Kopf, Zahl}\}$
DISKRET



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



$\Omega = (0,1)$
KONTINUIERLICH



Kombinatorik

Die Wahrscheinlichkeit ordnet jedem Elementarereignis eine Zahl zwischen 0 und 1 zu ...

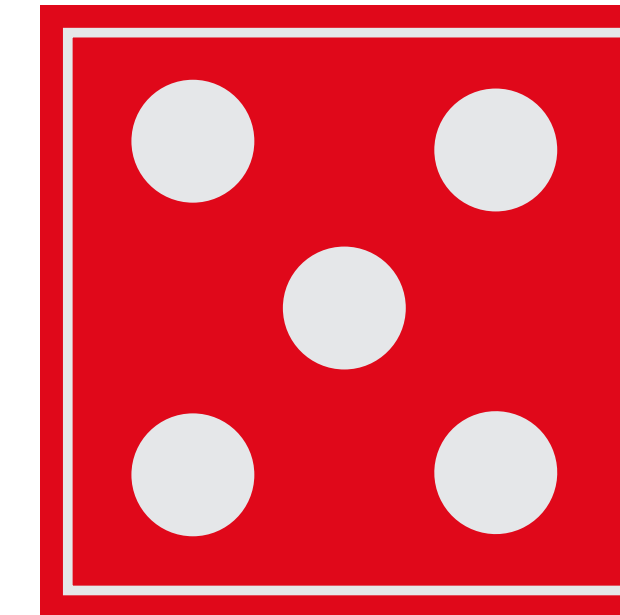
$$P: \Omega \rightarrow [0,1]$$

...und erfüllt die Kolmogorov-Axiome:

$$P(\omega) \in [0,1] \quad \forall \omega \in \Omega$$

$$P(\Omega) = 1$$

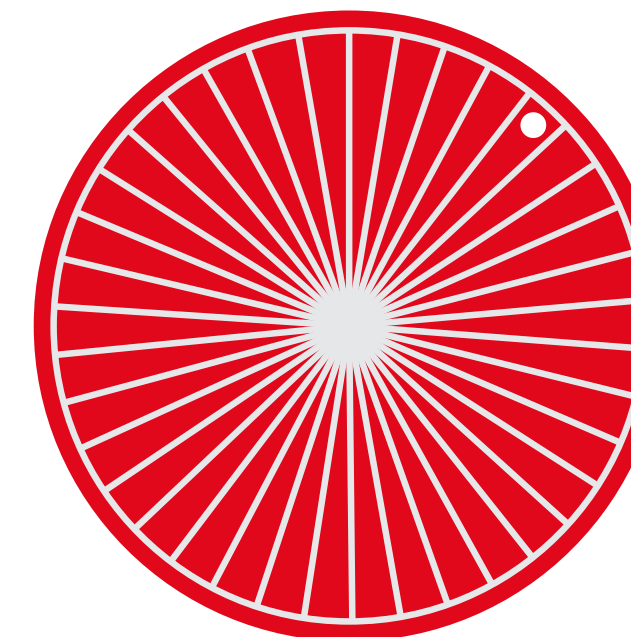
$$P(\omega_1 \cup \omega_2) = P(\omega_1) + P(\omega_2) \quad \omega_1 \neq \omega_2$$



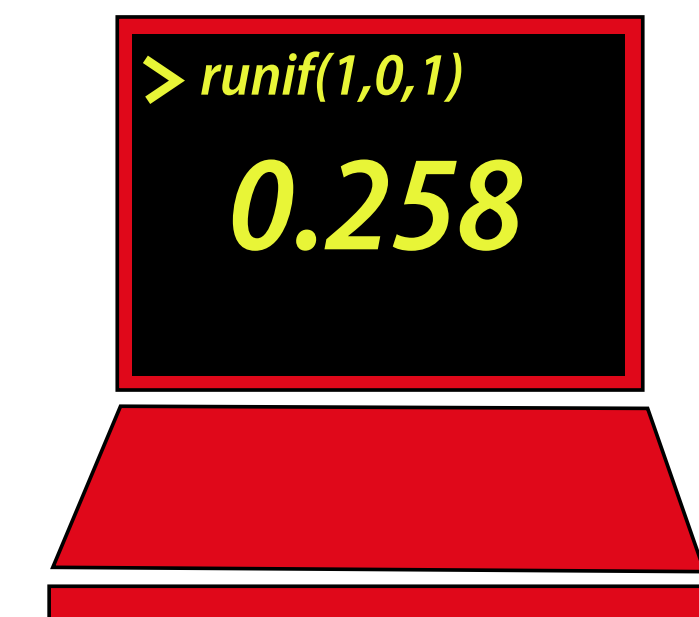
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



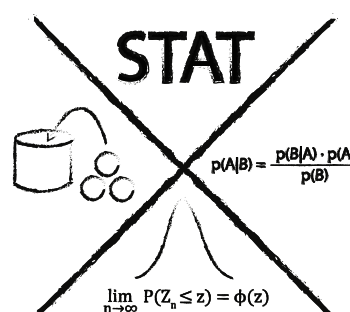
$\Omega = \{\text{Kopf, Zahl}\}$
DISKRET



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



$\Omega = (0,1)$
KONTINUIERLICH

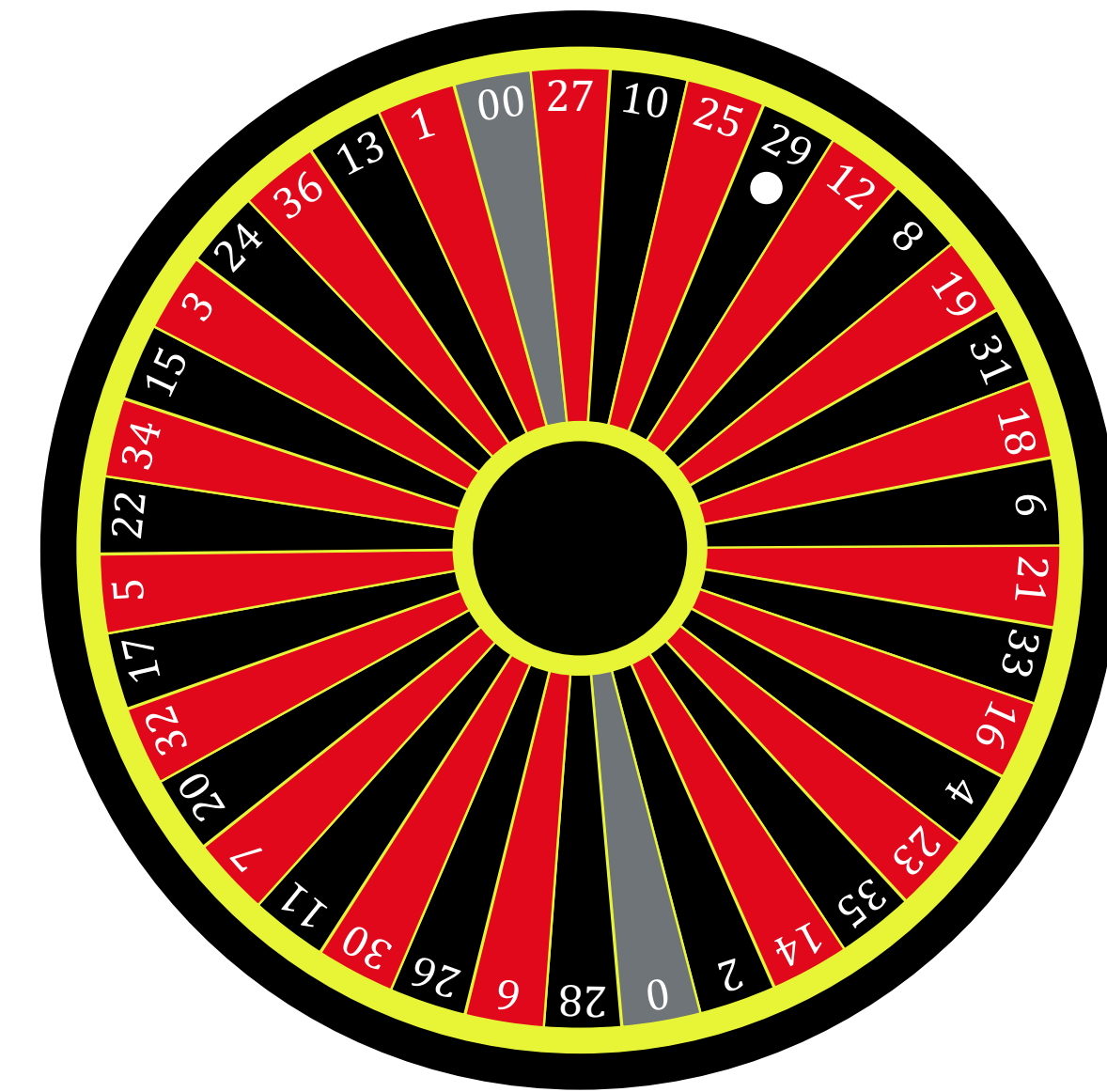


Beispiel Roulette

Roulette besitzt eine diskrete Ereignismenge. Jedes Elementarereignis besitzt dieselbe Wahrscheinlichkeit:

$$P(\omega) = \frac{1}{38} = 2.631\% \quad \forall \omega \in \Omega$$

Mit dieser Verteilung können wir die Wahrscheinlichkeit von **Ereignissen** am Roulettetisch berechnen.



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

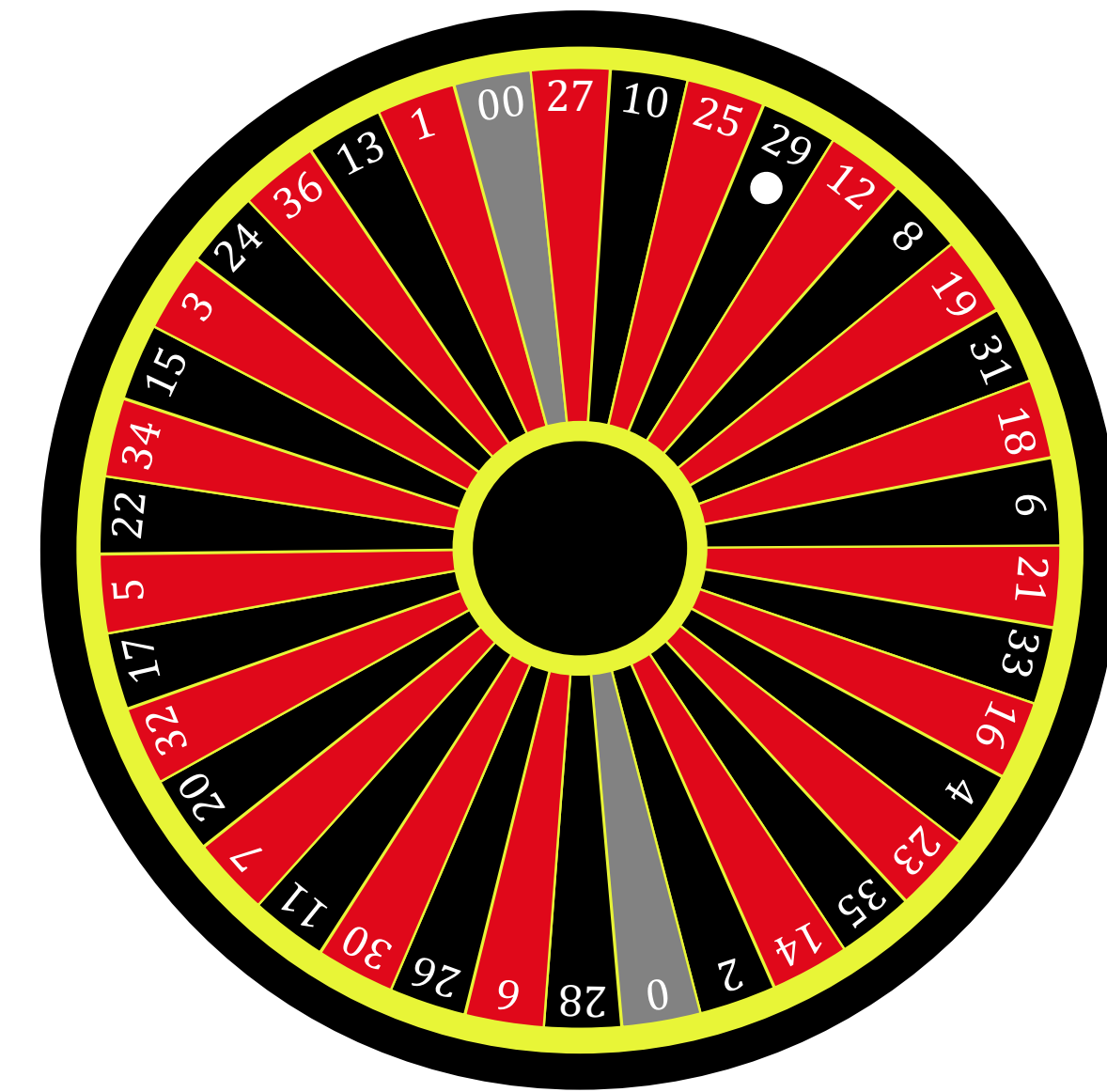
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

Beispiel Roulette

Ereignis A - Die Kugel fällt auf ein rotes Feld.

Wir können dieses Ereignis in Form einer Menge darstellen, die alle „roten Elementarereignisse“ enthält:

$$A = \{1,3,5,7,9,12, \dots, 36\}$$



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

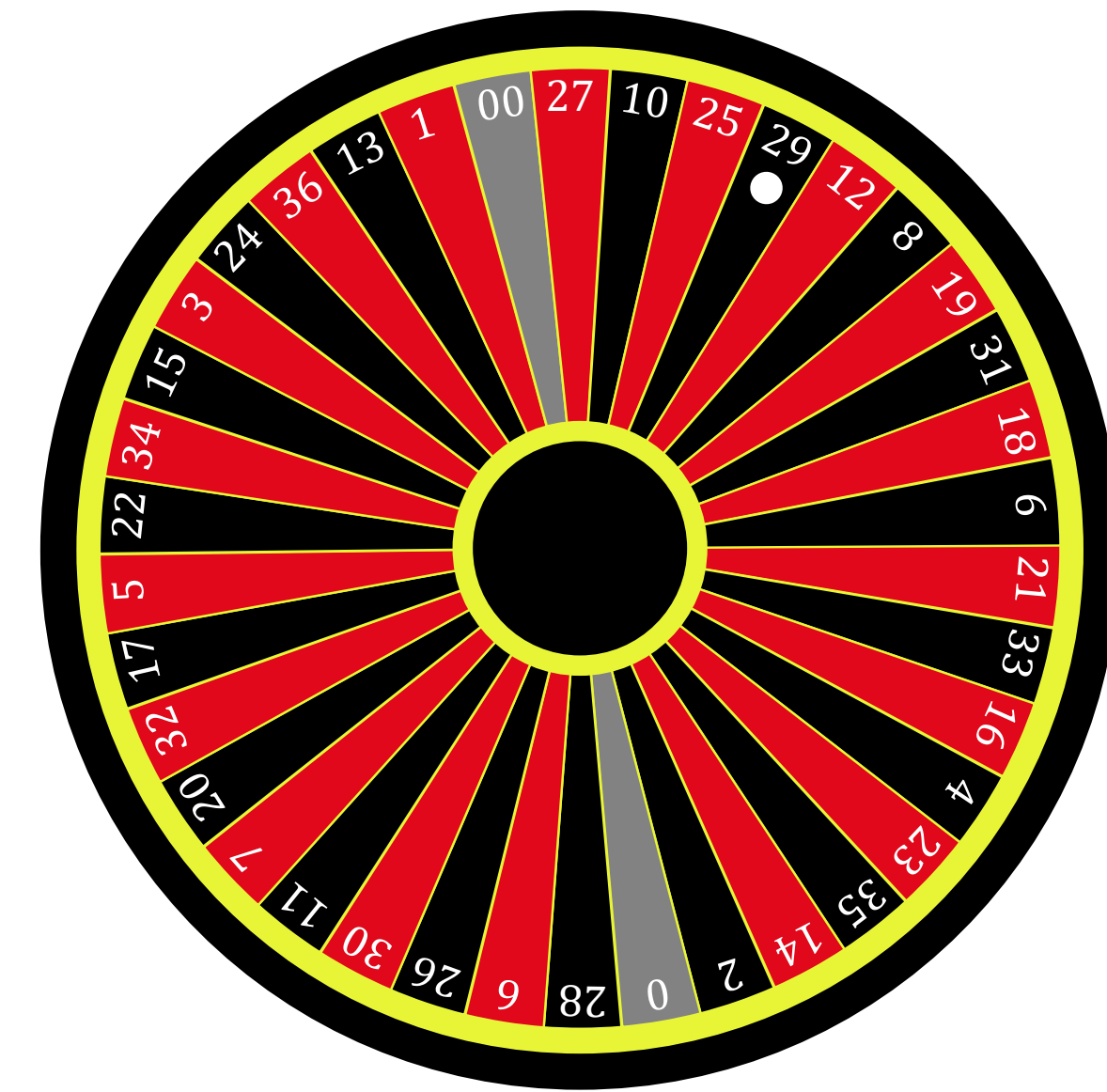
Beispiel Roulette

Ereignis A - Die Kugel fällt auf ein rotes Feld.

$$A = \{1,3,5,7,9,12, \dots, 36\}$$

Wir können das dritte Kolmogorov-Axiom anwenden.

$$\begin{aligned} P(A) &= P(\{1\}) + P(\{3\}) + P(\{5\}) + \dots \\ &= 18 \frac{1}{38} = \frac{18}{38} = 47.368\% \end{aligned}$$



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

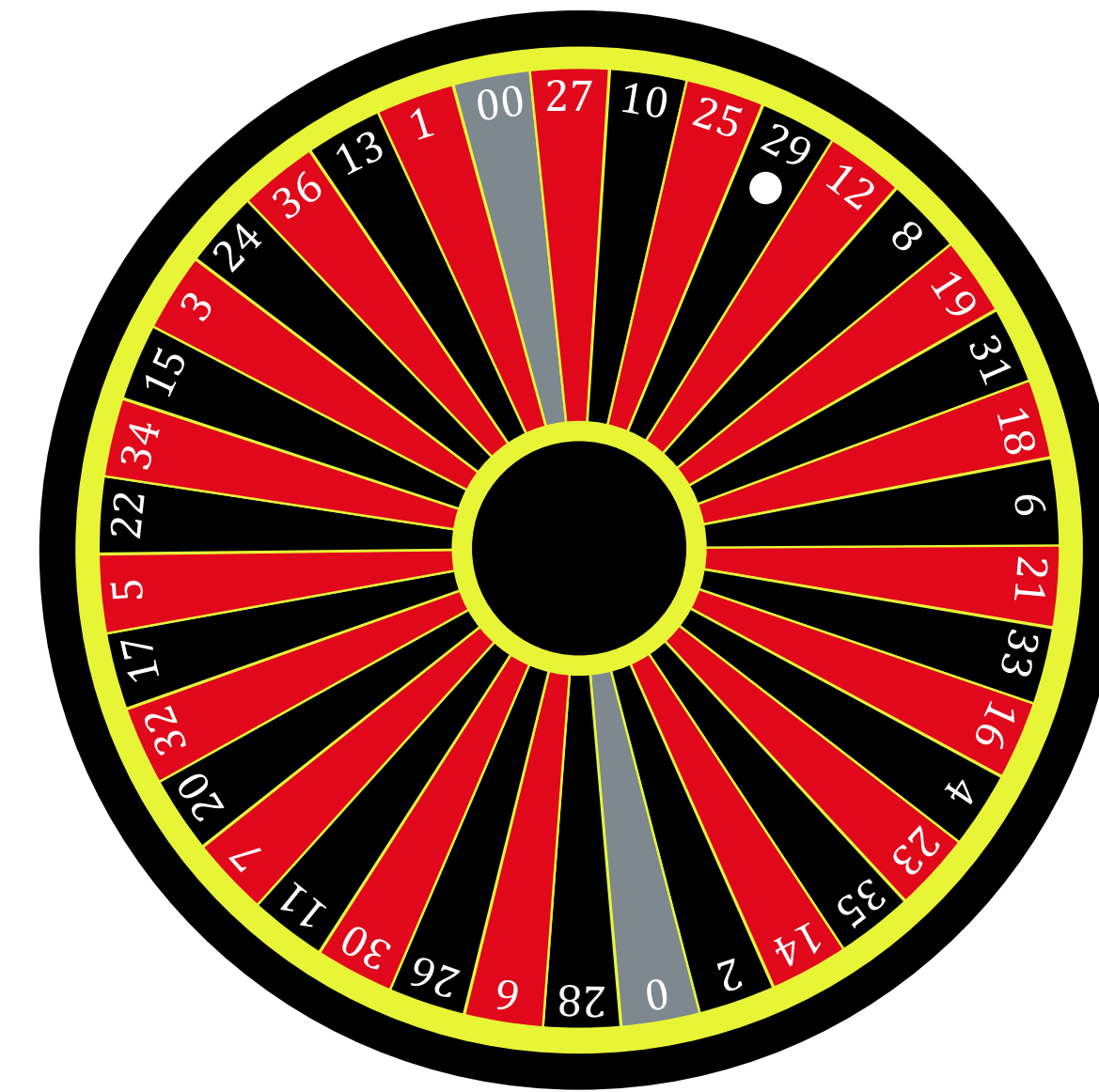
Beispiel Roulette

Ereignis B - Die Kugel fällt auf die „first five“

$$B = \{00, 0, 1, 2, 3\}$$

Wir können das dritte Kolmogorov-Axiom anwenden.

$$\begin{aligned} P(B) &= P(\{00\}) + P(\{0\}) + P(\{1\}) + \dots \\ &= 5 \cdot \frac{1}{38} = \frac{5}{38} = 13.158\% \end{aligned}$$



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

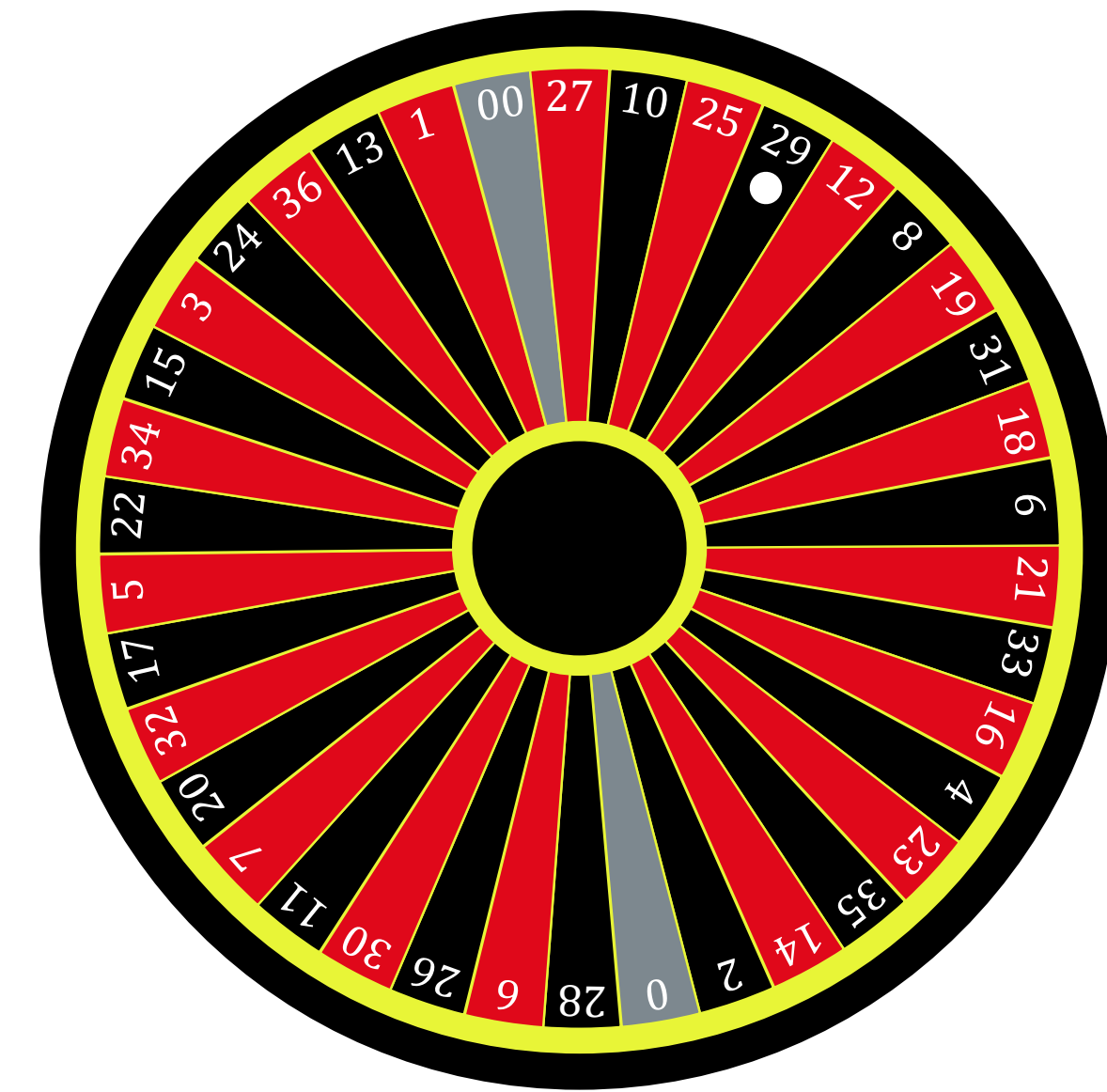
Beispiel Roulette

Ereignis C - Die Kugel fällt auf das erste Duzend.

$$C = \{1,2,3,4,5,6,7,8,9,10,11,12\}$$

Wir können das dritte Kolmogorov-Axiom anwenden.

$$\begin{aligned} P(C) &= P(\{1\}) + P(\{2\}) + P(\{3\}) + \dots \\ &= 12 \frac{1}{38} = \frac{12}{38} = 31.579\% \end{aligned}$$



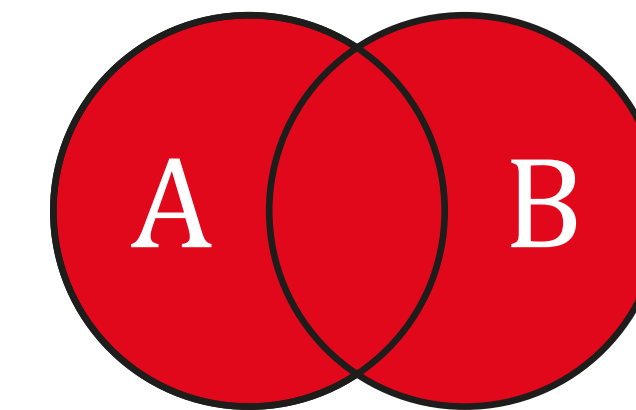
00	0	$\Omega = \{00,0,1,2,\dots,36\}$									
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

Verknüpfung von Ereignissen

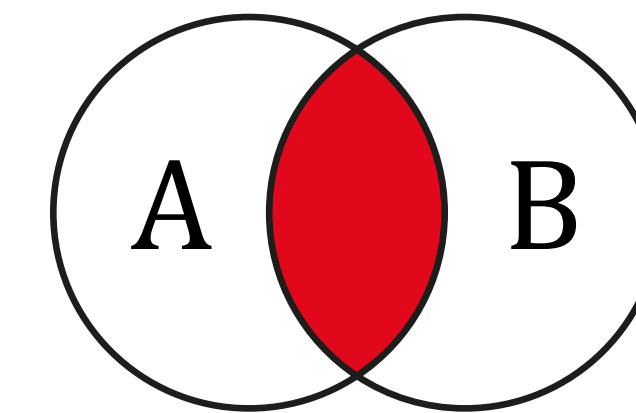
Ereignisse können miteinander kombiniert werden. Die Regeln hängen davon ab, ob Ereignisse stochastisch voneinander abhängig sind oder nicht.

Stochastisch unabhängige Ereignisse beeinflussen sich nicht. Das Eintreten des einen macht das andere nicht mehr oder weniger wahrscheinlich.

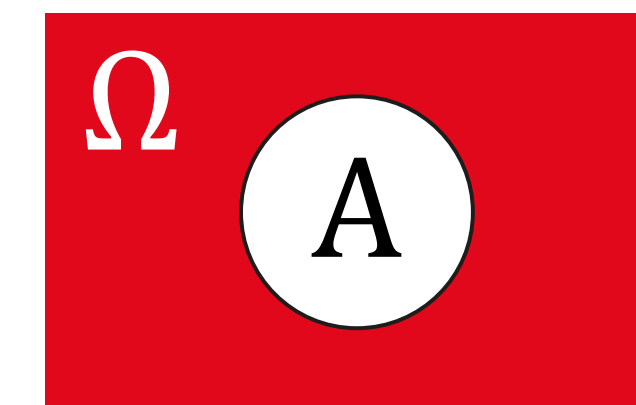
Stochastisch abhängige Ereignisse beeinflussen sich gegenseitig in ihrer Wahrscheinlichkeit.



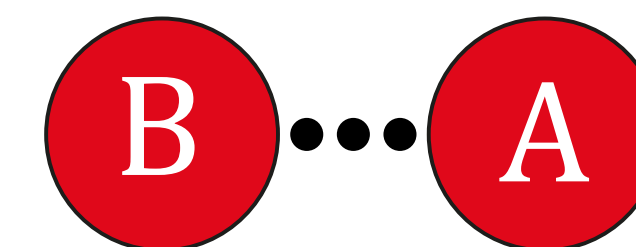
Vereinigung $A \cup B$
A oder B



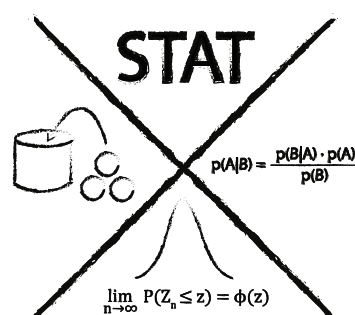
Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



Bedingte Wkt. $A | B$
A nach B



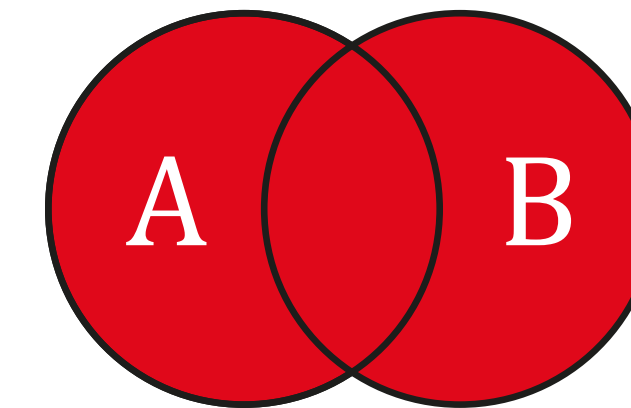
Unabhängige Ereignisse

Für **stochastisch unabhängige** Ereignisse gelten die folgenden Rechenregeln:

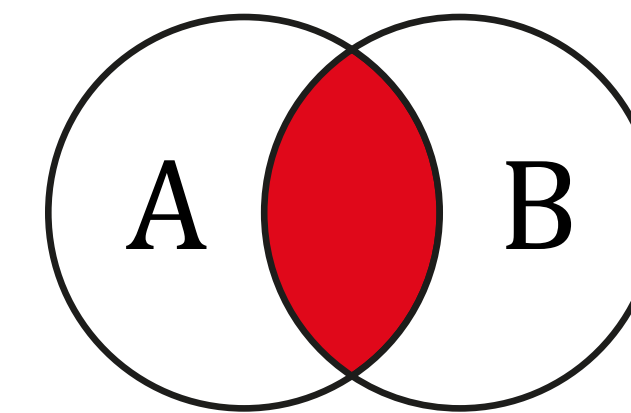
$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A \cup B) = 1 - (P(\bar{A}) \cdot P(\bar{B}))$$

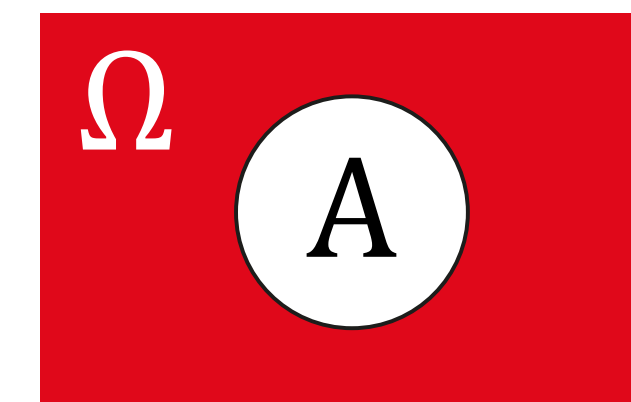
$$P(A | B) = P(A)$$



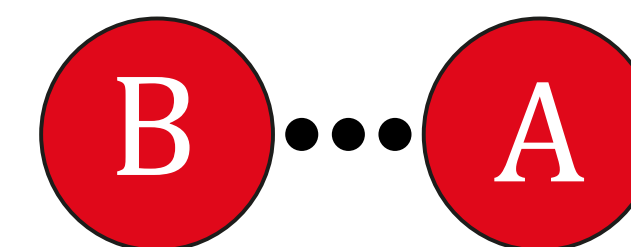
Vereinigung $A \cup B$
A oder B



Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



Bedingte Wkt. $A | B$
A nach B

Unabhängige Ereignisse

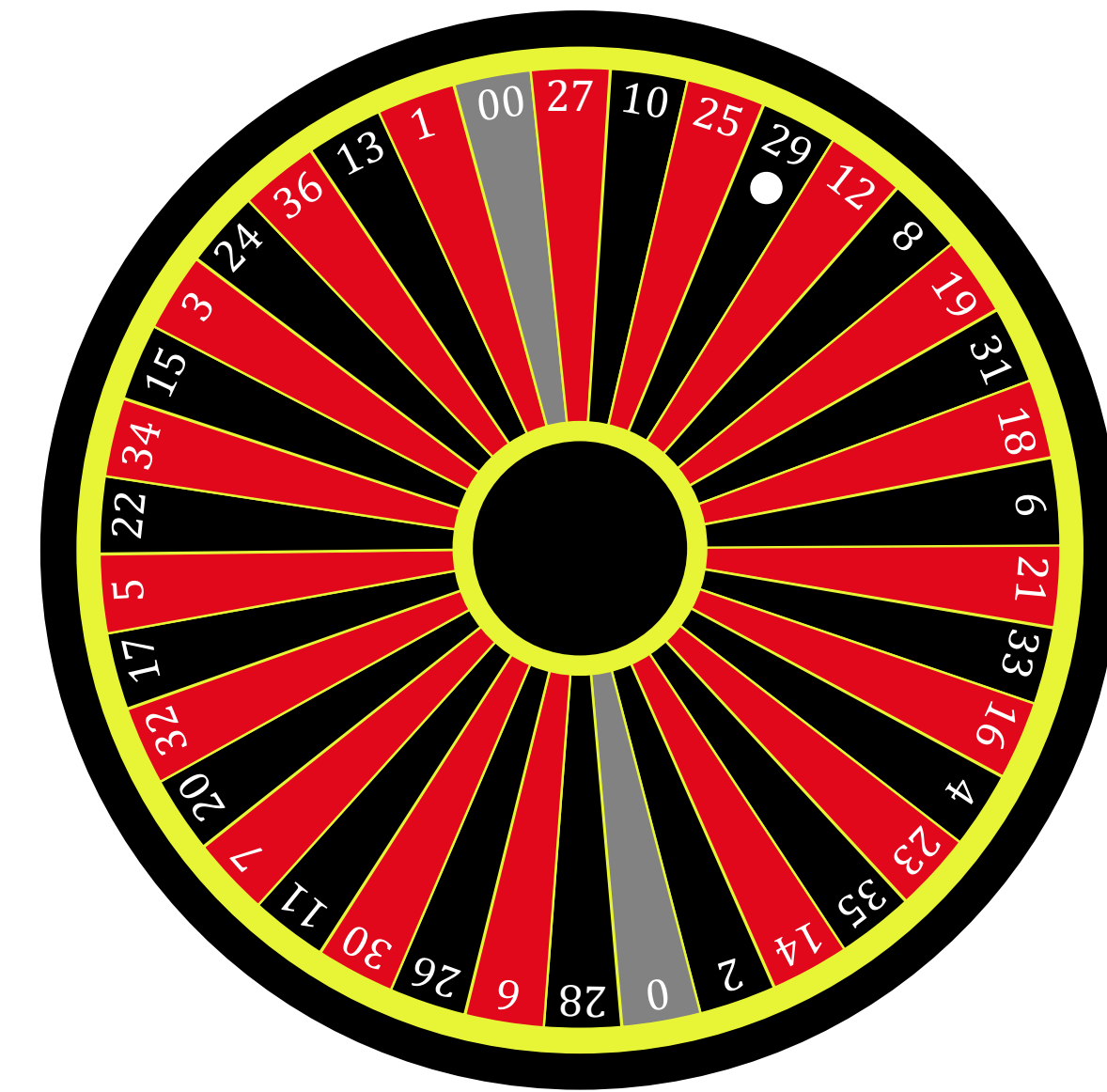
Ereignis A_i - Die Kugel fällt im i-ten Spin auf ein rotes Feld.

$$A_i = \{1, 3, 5, 7, 9, 12, \dots, 36\}$$

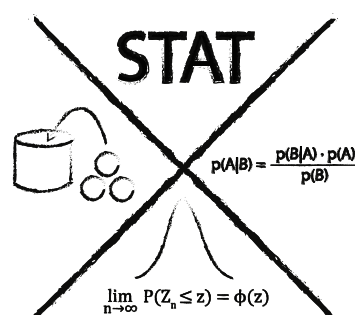
Wie hoch ist die Wahrscheinlichkeit von 3 mal rot in Folge?

$$P(A_1 \cap A_2 \cap A_3) = \frac{18}{38} \cdot \frac{18}{38} \cdot \frac{18}{38} = \frac{5832}{54872}$$

$$= 10.628\%$$



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



Unabhängige Ereignisse

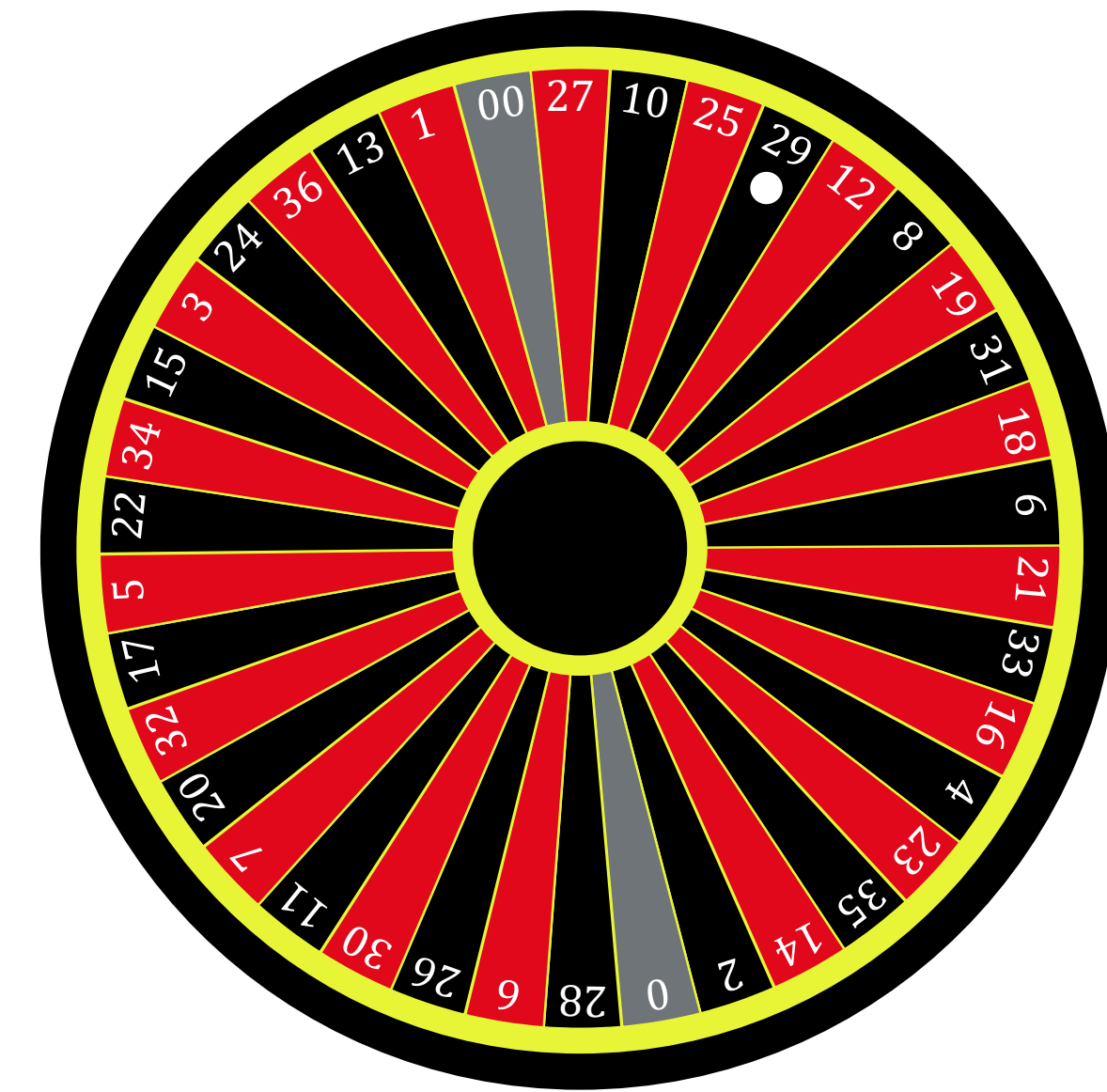
Wie hoch ist die Wahrscheinlichkeit, dass in 100 Spins keine einzige „0“ oder „00“ vorkommt?

Wir definieren das Ereignis $N_i = \{00, 0\}$ und rechnen:

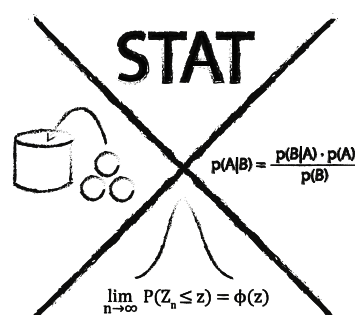
$$P(N_i) = \frac{2}{38} = 5.263\%$$

$$P(\bar{N}_i) = 1 - \frac{2}{38} = 94.737\%$$

$$P(\bar{N}_1 \cap \bar{N}_2 \cap \dots \cap \bar{N}_{100}) = \left(1 - \frac{2}{38}\right)^{100} = 0.449\%$$



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



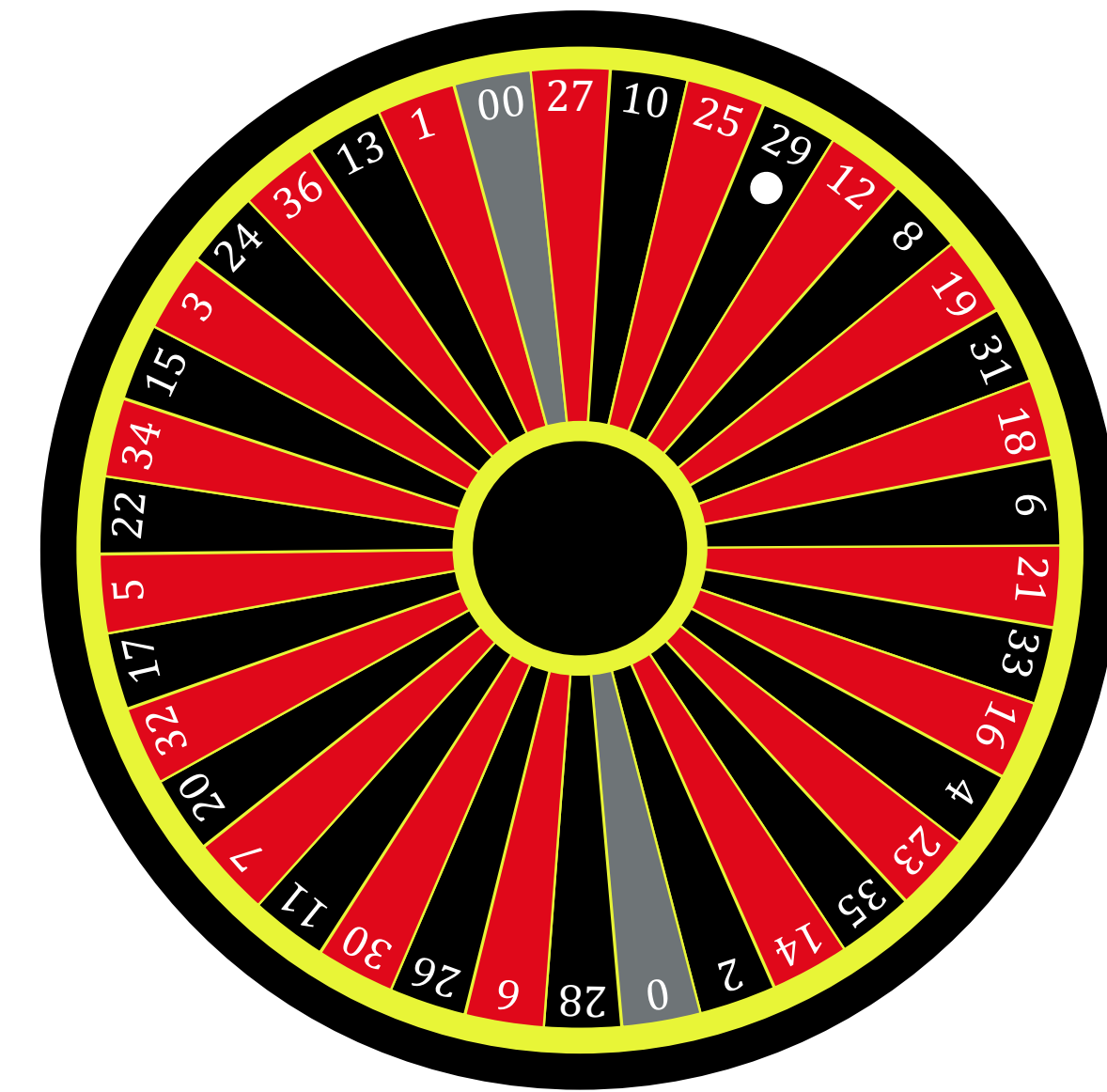
Unabhängige Ereignisse

Wie hoch wäre im Vergleich dazu die Wahrscheinlichkeit, dass in 100 Spins mindestens eine „0“ oder „00“ vorkommt?

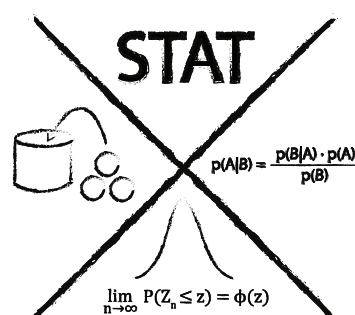
$$P(N_i) = \frac{2}{38} = 5.263\%$$

$$P(\bar{N}_i) = 1 - \frac{2}{38} = 94.737\%$$

$$\begin{aligned} P(N_1 \cup N_2 \cup \dots \cup N_{100}) &= 1 - P(\bar{N}_1) \cdot \dots \cdot P(\bar{N}_{100}) \\ &= 1 - P(\bar{N}_i)^{100} = 99.551\% \end{aligned}$$



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



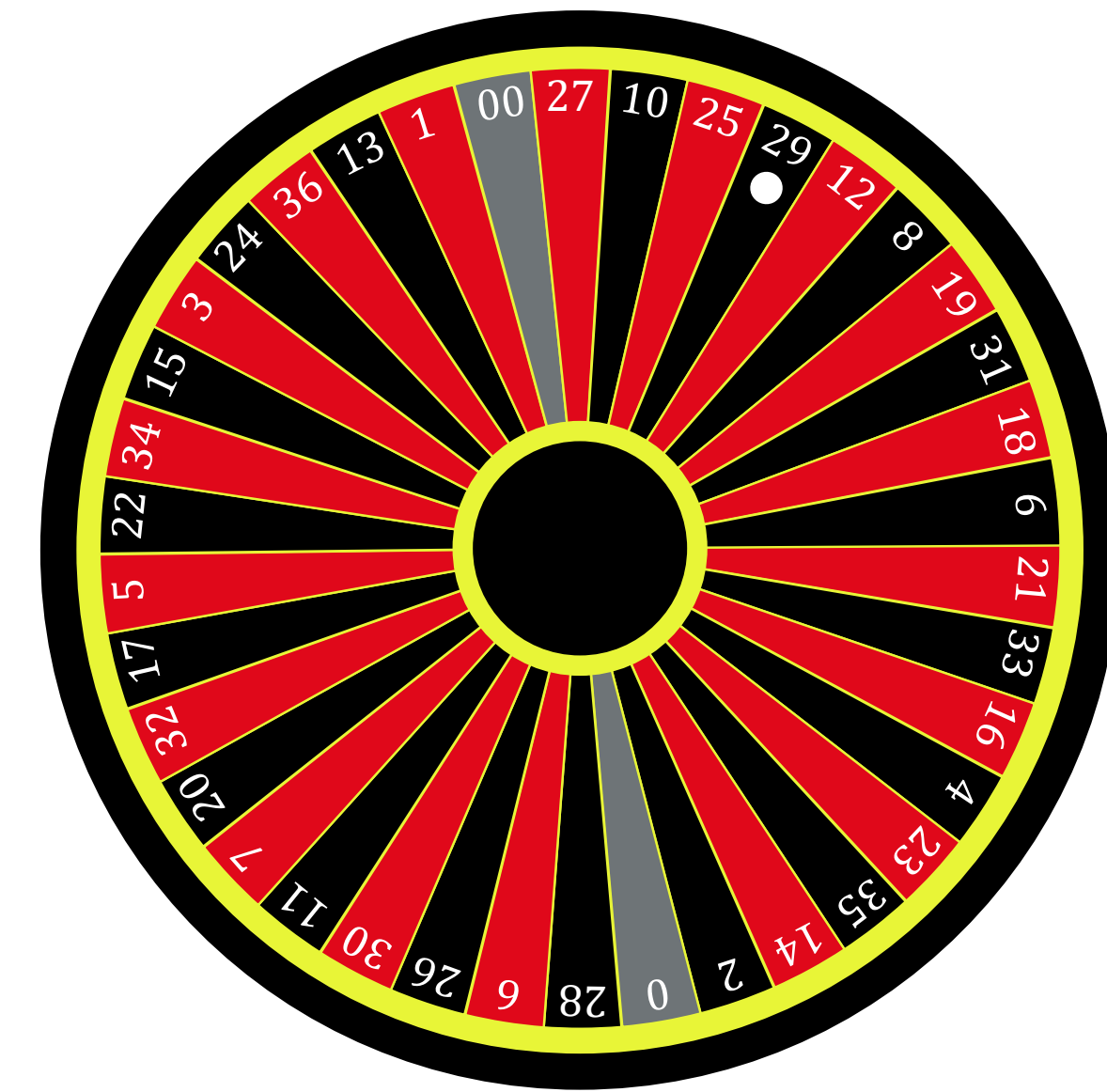
Unabhängige Ereignisse

Ereignis A_i - Die Kugel fällt im i-ten Spin auf ein rotes Feld.

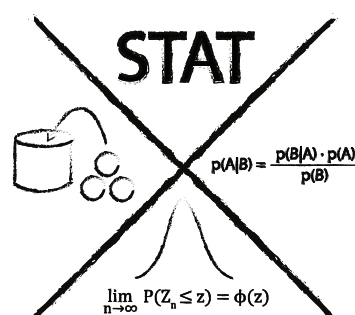
$$A_i = \{1, 3, 5, 7, 9, 12, \dots, 36\}$$

Wie hoch ist die Wahrscheinlichkeit von Rot im zweiten Spin, nachdem im ersten bereits rot kam?

$$P(A_2 | A_1) = P(A_2) = \frac{18}{38} = 47.36\%$$



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



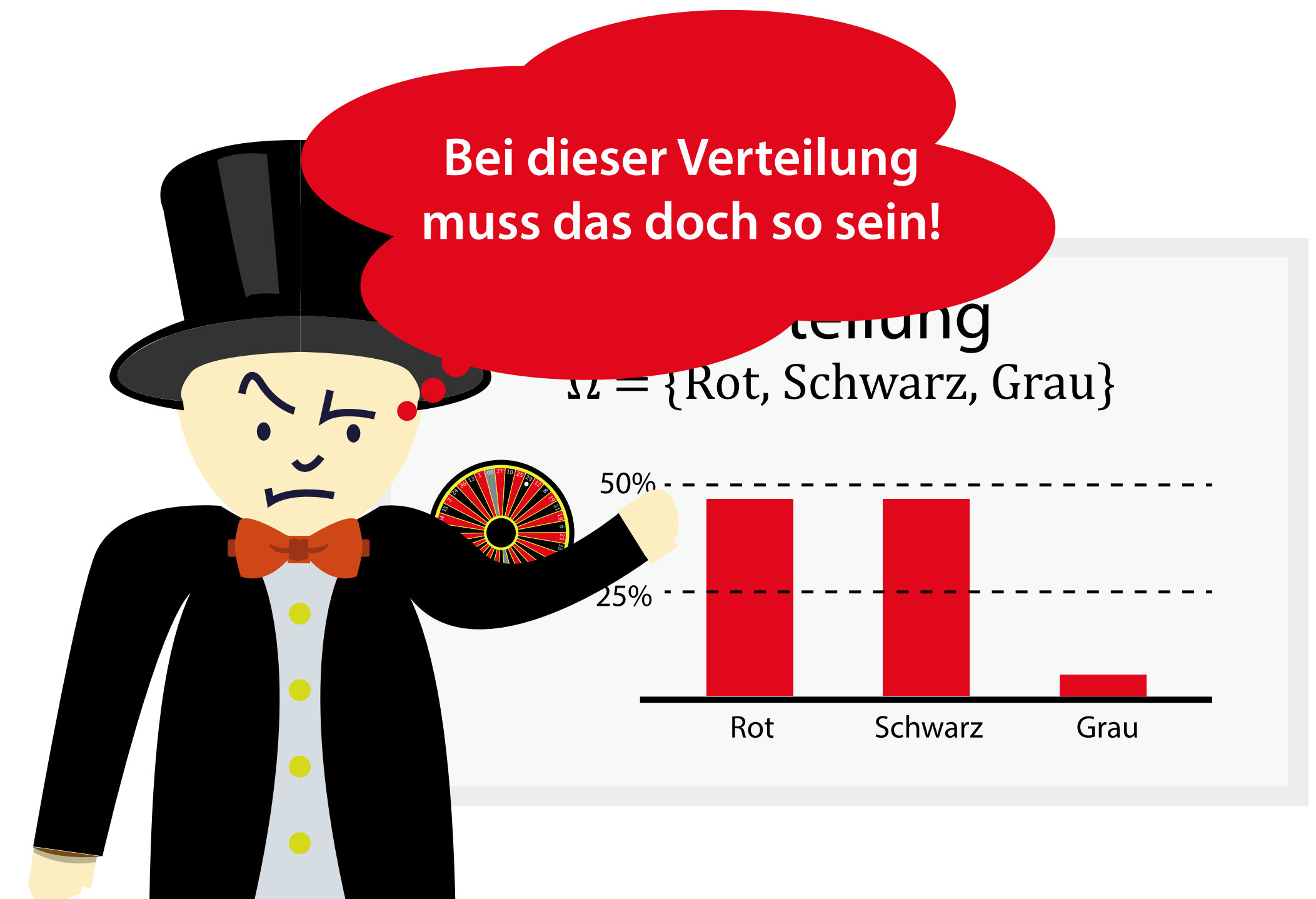
Dieses Ergebnis klingt trivial, aber Menschen neigen tatsächlich dazu, stochastische Abhängigkeiten zu sehen, wo keine sind.

Gamblers Fallacy

Der Grundgedanke hinter dem **Gamblers Fallacy** ist gar nicht so falsch: Laut Verteilung kommt „rot“ in 47.368% aller Fälle.

Überwachen wir den Roulettetisch über eine sehr große Zahl an Spins, sollte also auch ca. 47.368% rot sein.

Wenn die Farbe schwarz eine Zeit lang überproportional häufig kommt, sollte es auch Phasen geben, in denen rot überproportional häufig kommt!

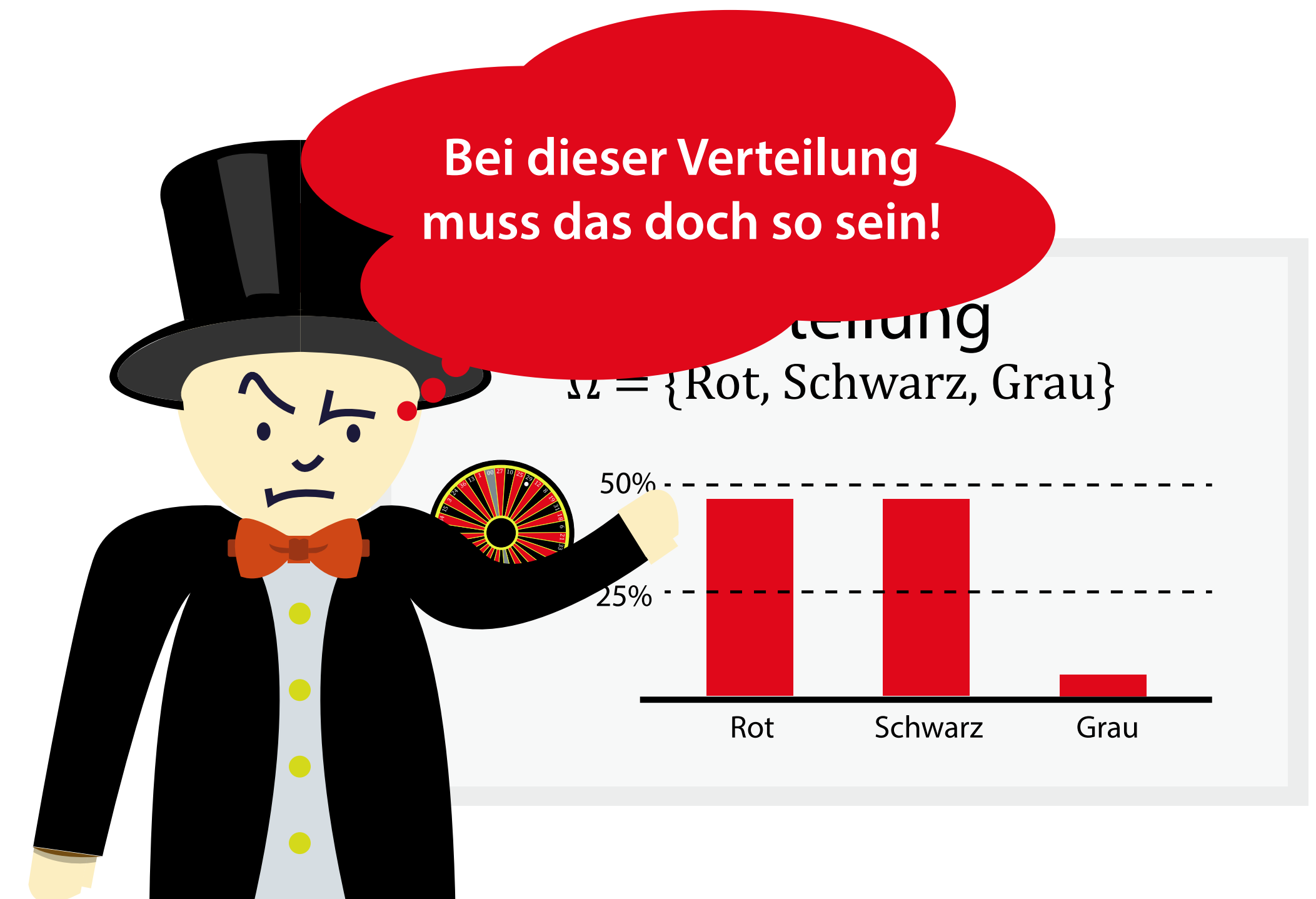


Gamblers Fallacy

Korrekt, aber wir wetten beim Roulette nicht auf die langfristige Verteilung nach unendlich vielen Spins, sondern auf den nächsten Spin.

Dieser Spin ist von den vorherigen Spins unabhängig.

Anders gedacht: Woher soll die Kugel auch wissen, wo sie vorher gelandet ist? Sie hat weder ein Gedächtnis noch einen freien Willen!



Unabhängige Ereignisse

Betrachten Sie einen Studierenden, der ein Kahoot-Quiz mit mehreren Fragen spielt. Zu jeder Frage gibt es 4 Antwortmöglichkeiten, von denen eine richtig ist.

Da der Studierende in der Vorlesung nicht aufgepasst hat, wählt er zufällig eine der Antwortmöglichkeiten aus.

Verwenden Sie die Ergebnismenge $\Omega = \{\text{Falsch}, \text{Richtig}\}$ um folgende Aufgaben zu bearbeiten.

- a) Geben Sie die Wahrscheinlichkeit für die beiden Elementarereignisse an
- b) Wie hoch ist die Wahrscheinlichkeit von 3 richtigen in Folge?
- c) Wie viele Fragen muss das Quiz haben, damit der Studierende zu 99% Wahrscheinlichkeit mindestens eine richtig hat?
- d) Am Ende verliert der Studierende mit 6 von 15 richtig beantworteten Fragen. Hatte er eher Pech oder Glück?

Unabhängige Ereignisse

a) Geben Sie die Wahrscheinlichkeit für die beiden Elementarereignisse an

$$P(\text{Richtig}) = 0.25 \text{ bzw. } 25\%$$

$$P(\text{Falsch}) = 0.75 \text{ bzw. } 75\%$$

b) Wie hoch ist die Wahrscheinlichkeit von 3 richtigen in Folge?

Definiere das Ereignis A_i , dass die i -te Frage richtig beantwortet wird, wobei i die Werte 1, 2 und 3 annimmt.

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= (0.25)^3 \\ &= 0.0156 \\ &= 1.56\% \end{aligned}$$

Unabhängige Ereignisse

c) Wie viele Fragen muss das Quiz haben, damit der Studierende zu 99% Wahrscheinlichkeit mindestens eine richtig hat?

Wahrscheinlichkeit mindestens eines Treffers:

$$P(A_1 \cup \dots \cup A_n) = 1 - P(\bar{A}_1) \cdot \dots \cdot P(\bar{A}_n) \stackrel{!}{=} 0.99$$

Dies ist erfüllt, wenn der blaue Term 0.01 beträgt!

Wahrscheinlichkeit einer Negativserie über n Fragen:

$$P(\bar{A}_1) \cdot \dots \cdot P(\bar{A}_n) = 0.75^n$$

Suche nach dem Exponenten n für den die Wahrscheinlichkeit 1% unterschreitet.

$$0.75^n \stackrel{!}{=} 0.01$$

$$\iff n = \log_{0.75}(0.01)$$

$$\iff n = 16.01$$

Unabhängige Ereignisse

d) Am Ende verliert der Studierende mit 6 von 15 richtig beantworteten Fragen. Hatte er eher Pech oder Glück?

Bei einer Trefferquote von 25% liegt der Erwartungswert bei ...

$$15 \cdot 0.25 = 3.75$$

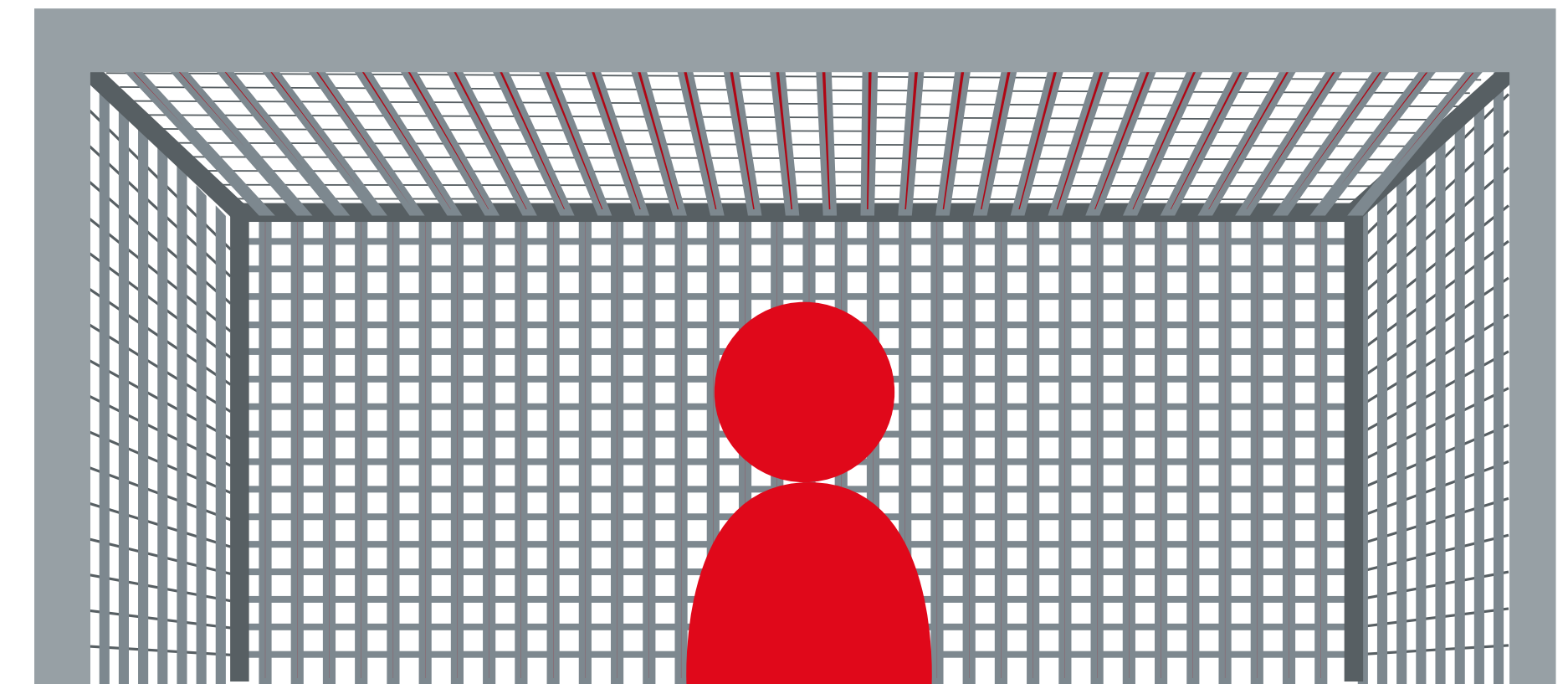
...Treffern. Der Studierende hatte also ziemliches Glück mit seinen 6 Richtigen!

Permutationen

Mit den bisherigen Regeln können wir nur bestimmte Arten von Kombinatorikaufgaben lösen.

Bereits bei folgender Aufgabe kämen wir nicht mehr weiter:

Ein Elfmeterschütze im Profifußball erzielt laut DFB zu 75% ein Tor. Wie hoch ist die Wahrscheinlichkeit, dass er von 5 Elfmetern genau 4 verwandelt?



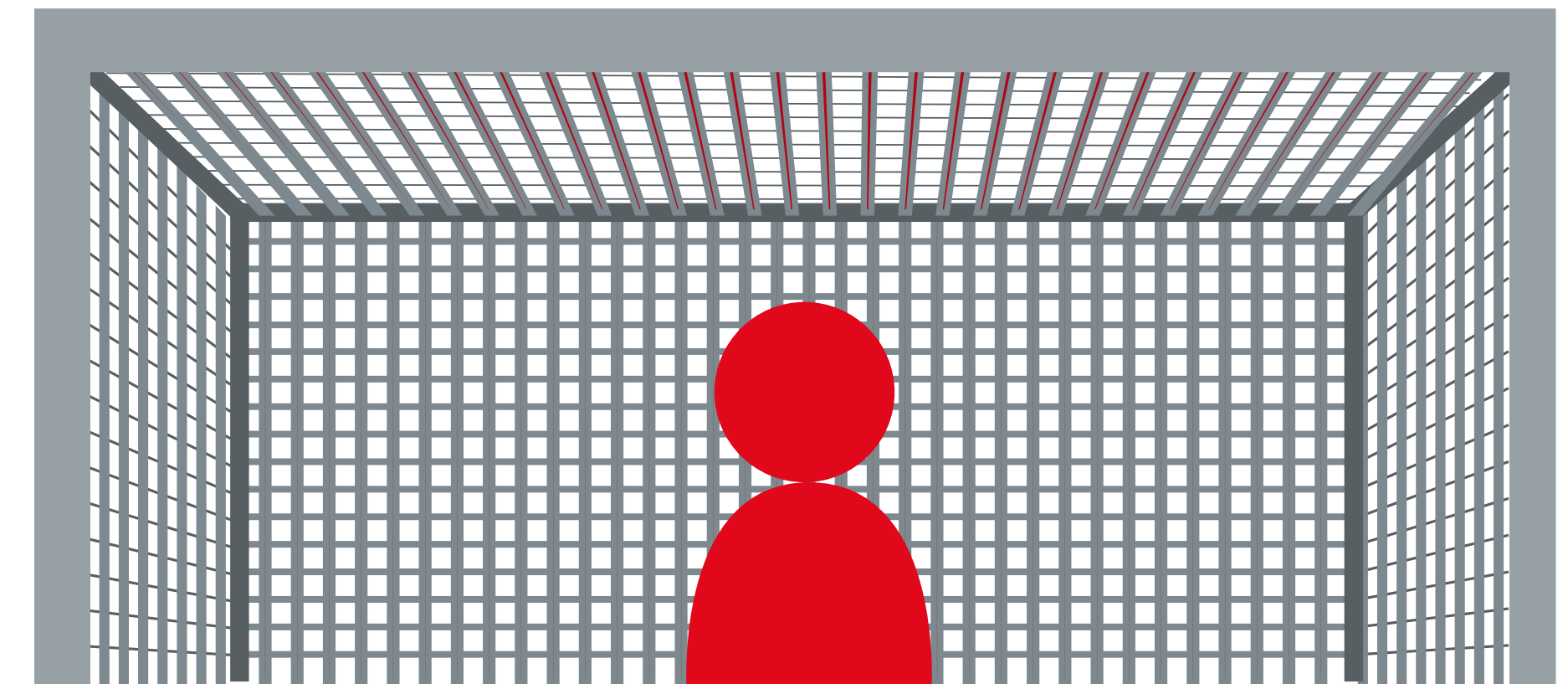
Torchance eines Elfmeters: $P_G = 75\%$
Gesucht: Chance auf 4 Treffer aus 5 Versuchen

Permutationen

Naiver Ansatz: Wir berechnen die Wahrscheinlichkeit mit der folgenden Formel:

$$\begin{aligned}
 P &= P_G^4 \cdot (1-P_G)^1 \\
 &= 0.75^4 \cdot (1-0.75)^1 \\
 &= 7.91\%
 \end{aligned}$$

Der Wert ist viel zu niedrig; irgendwas stimmt da nicht!



Torchance eines Elfmeters: $P_G = 75\%$
 Gesucht: Chance auf 4 Treffer aus 5 Versuchen

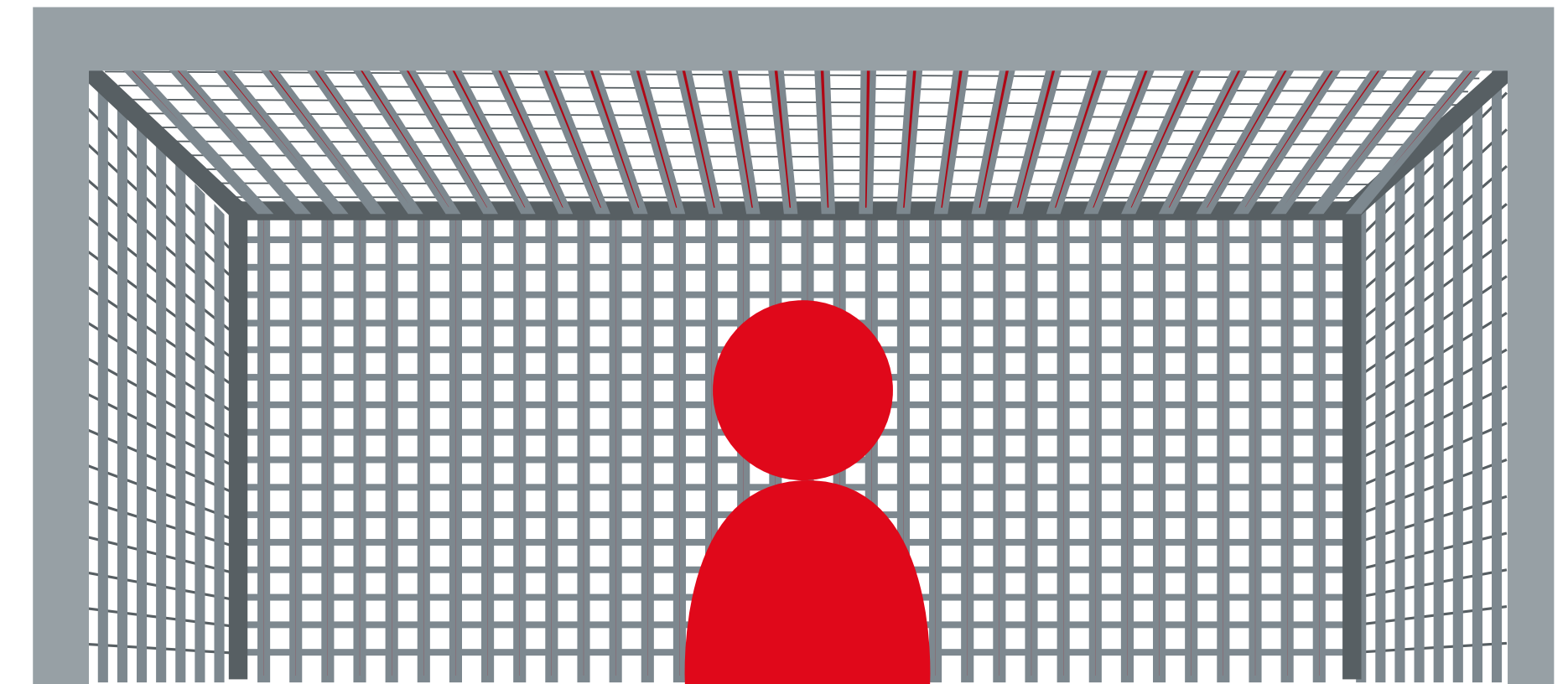
Permutationen

Mit unserer Formel haben wir die Wahrscheinlichkeit berechnet, dass der Schütze die ersten vier trifft und den fünften verschießt.

$$P = P_G^4 \cdot (1 - P_G)^1 = 7.91\%$$

Die Formel bezieht sich auf eine ganz bestimmte Reihenfolge von Treffern und Fehlschüssen.

Wie machen wir sie von der Reihenfolge unabhängig?



Torchance eines Elfmeters: 75%

Gesucht: Chance auf 4 Treffer aus 5 Versuchen



Permutationen

Grundgedanke: Wir gehen alle Möglichkeiten durch, wie der Schütze auf insgesamt 4 Tore kommen kann.

Es gibt 5 Möglichkeiten, um einen von fünf Elfmeter zu verfehlen: den ersten, den zweiten, ...

Es gibt 5 Möglichkeiten - alle sind gleich wahrscheinlich!

Torchance eines Elfmeters: 75%

Gesucht: Chance auf 4 Treffer aus 5 Versuchen

$$\checkmark \checkmark \checkmark \checkmark \times \quad P = 0.75^4 \cdot 0.25 = 7.91\%$$

$$\checkmark \checkmark \checkmark \times \checkmark \quad P = 0.75^3 \cdot 0.25 \cdot 0.75 = 7.91\%$$

$$\checkmark \checkmark \times \checkmark \checkmark \quad P = 0.75^2 \cdot 0.25 \cdot 0.75^2 = 7.91\%$$

$$\checkmark \times \checkmark \checkmark \checkmark \quad P = 0.75^1 \cdot 0.25 \cdot 0.75^3 = 7.91\%$$

$$\times \checkmark \checkmark \checkmark \checkmark \quad P = 0.25 \cdot 0.75^4 = 7.91\%$$

$$P = 39.6\%$$

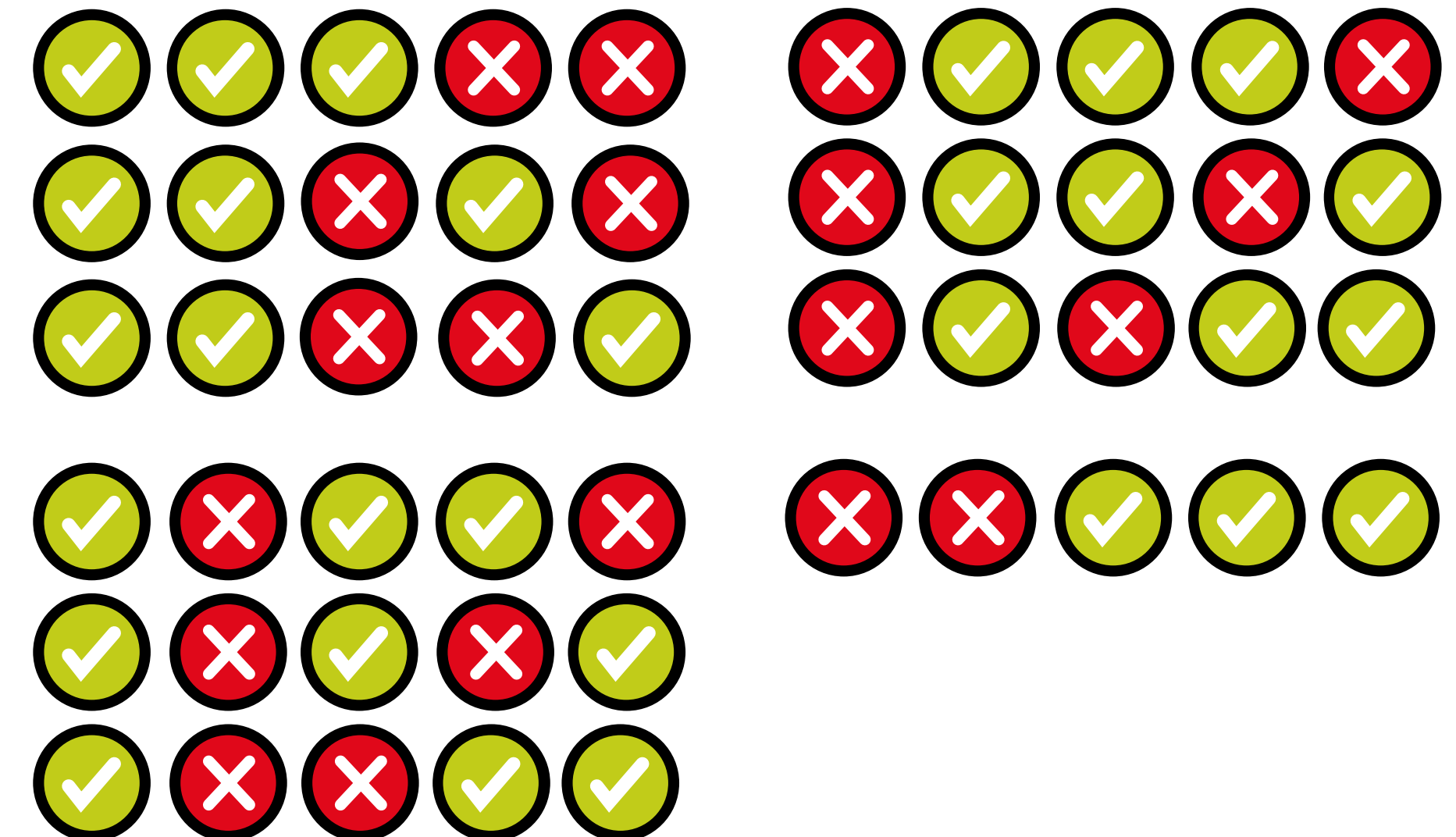
Permutationen

Problem: Nicht immer sind die Möglichkeiten so einfach abzuzählen.

Wenn wir nach Möglichkeiten 2 von 5 Elfmetern zu verfehlen suchen, wird es schon kniffliger ...

Gibt es eine Formel, mit der wir dies ausrechnen können statt es durchdenken zu müssen?

10 Möglichkeiten 2 von 5 Elfmetern zu verschießen

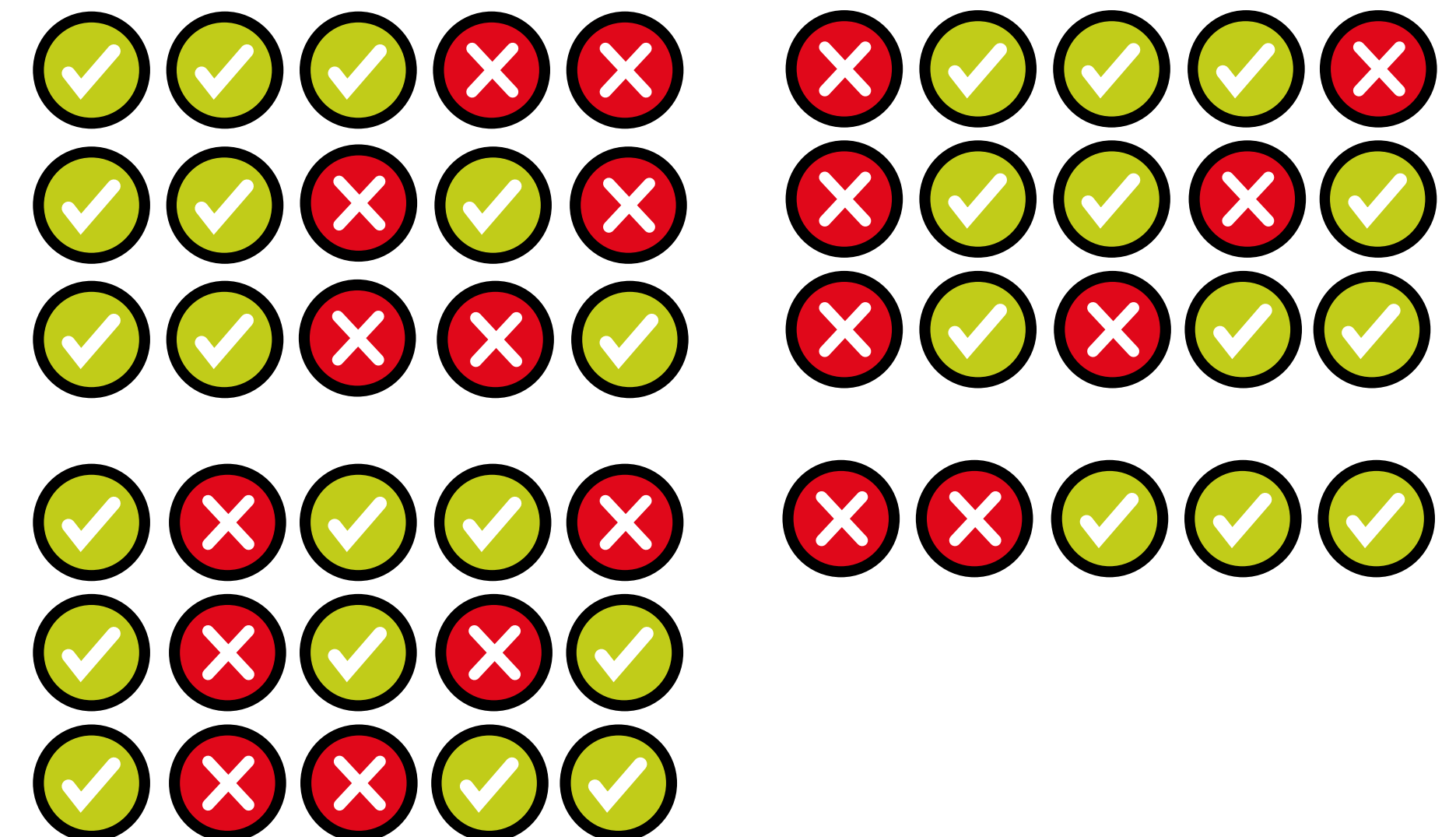


Permutationen

Hier kommt die Überschrift ins Spiel: Die Permutation kommt vom lateinischen „permutare“ vertauschen.

Jede rechts gezeigte Anordnung von Toren/Fehlschüssen ist eine von insgesamt 10 Permutationen, die unsere Vorgabe 3 Treffer aus 5 Versuchen erfüllt.

10 Möglichkeiten 2 von 5
Elfmetern zu verschießen

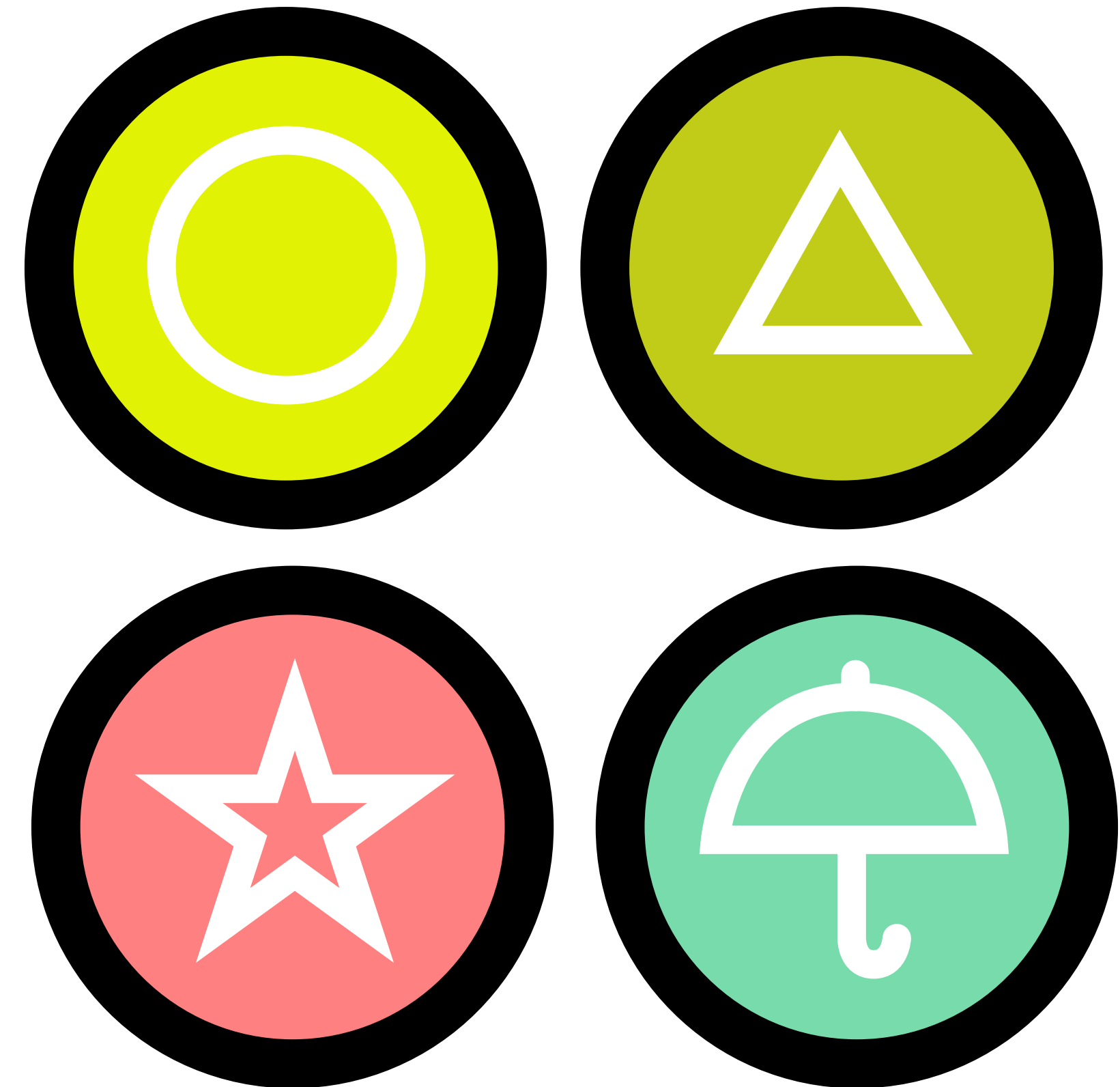


Permutationen

Wie berechnen wir diese Anzahl an Permutationen? Wir benötigen dazu zwei mathematische Funktionen: Die Fakultät und den Binomialkoeffizienten.

Um diese Funktionen besser zu verstehen, stellen wir das Elfmeterbeispiel zurück und betrachten eine einfachere Fragestellung!

Wie viele Möglichkeiten gibt es 4 verschiedene Dinge anzuordnen?



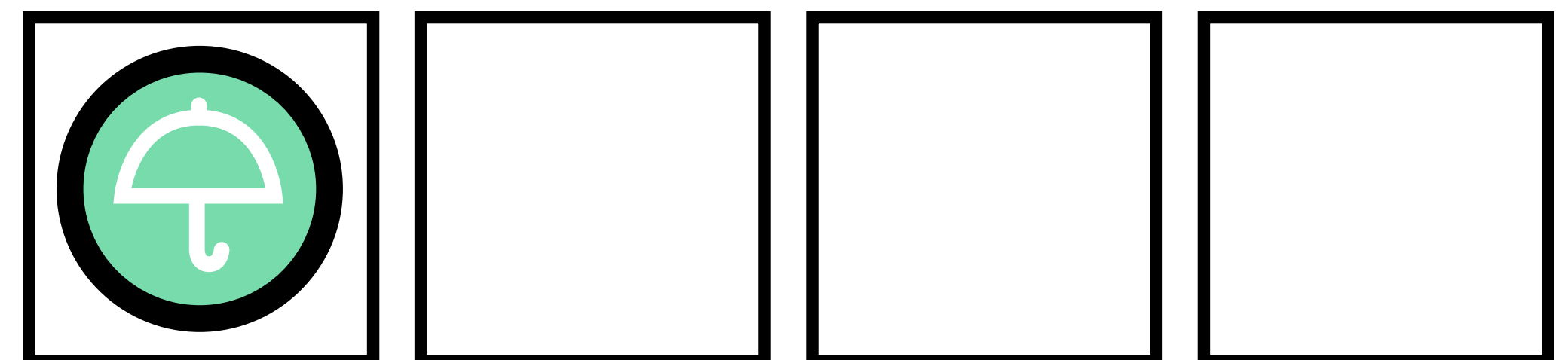
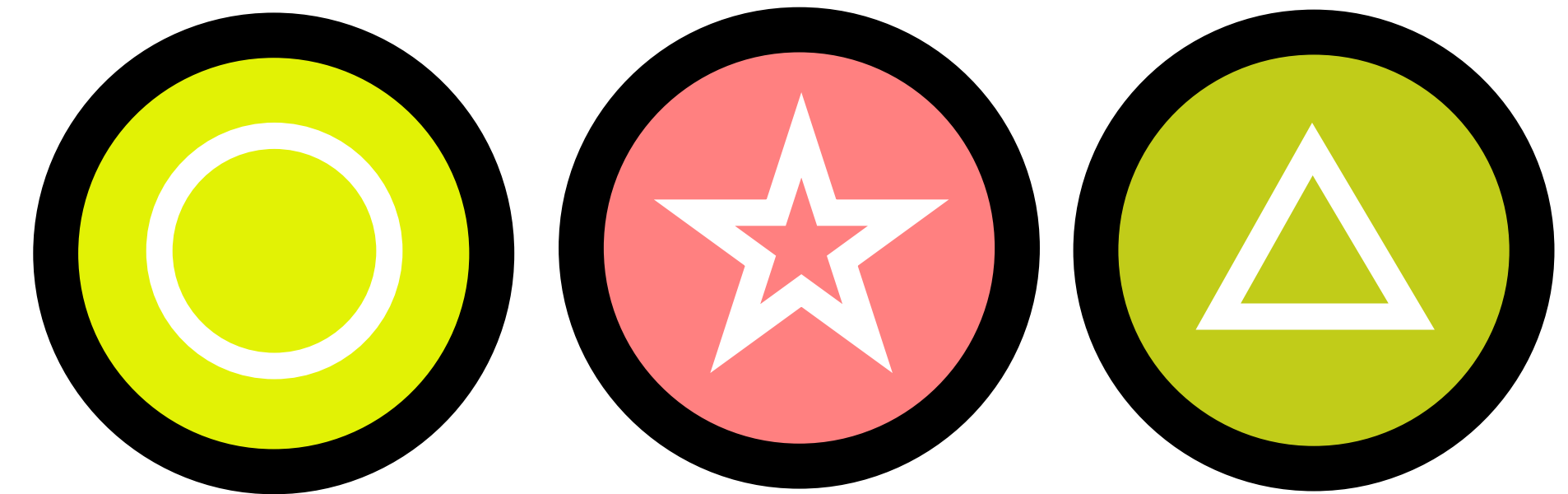
Permutationen

Wir gehen die möglichen Anordnungen im Kopf durch und berechnen dabei folgenden Wert:

$$1 \cdot 2 \cdot 3 \cdot 4 = 24$$

Fakultät die Anzahl Möglichkeiten um n verschiedene Dinge anzuordnen ist:

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



4 Möglichkeiten

3 Möglichkeiten

2 Möglichkeiten

1 Möglichkeit



DHBW-Taschenrechner: Shift + [x⁻¹] Taste links oben
Excel Funktion: =FAKULTÄT(N)
Wolfram/Google: einfach „!“ eingeben

Permutationen

Aber wie sieht es aus, wenn nicht alle Dinge gleich sind?
In unserem Elfmeterbeispiel haben wir nicht 5 verschiedene Dinge sondern:

3x Tor

2x Fehlschuss

Mit der Fakultät $5!$ erhalten wir 120 Anordnungen und damit viel zu viele. Jetzt kommt die zweite Funktion ins Spiel!



Permutationen

Binomialkoeffizient „n über k“ gibt die Anzahl Möglichkeiten an, um k Elemente aus n auszuwählen.

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

Wir wählen 3 aus 5 Elfmetern, die getroffen werden:

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



DHBW-Taschenrechner: $n \dots [nCr] \dots k$
Excel Funktion: $=\text{KOMBINATIONEN}(n;k)$
Wolfram: $\text{BINOMIAL}(n,k)$
Google: $n \text{ choose } k$

Permutationen

Für den Binomialkoeffizienten gilt:

$$\binom{n}{k} = \binom{n}{n-k}$$

Es spielt daher keine Rolle, ob wir 3 Treffer oder 2 Fehlschüsse aus 5 Elfm Metern wählen.

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

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

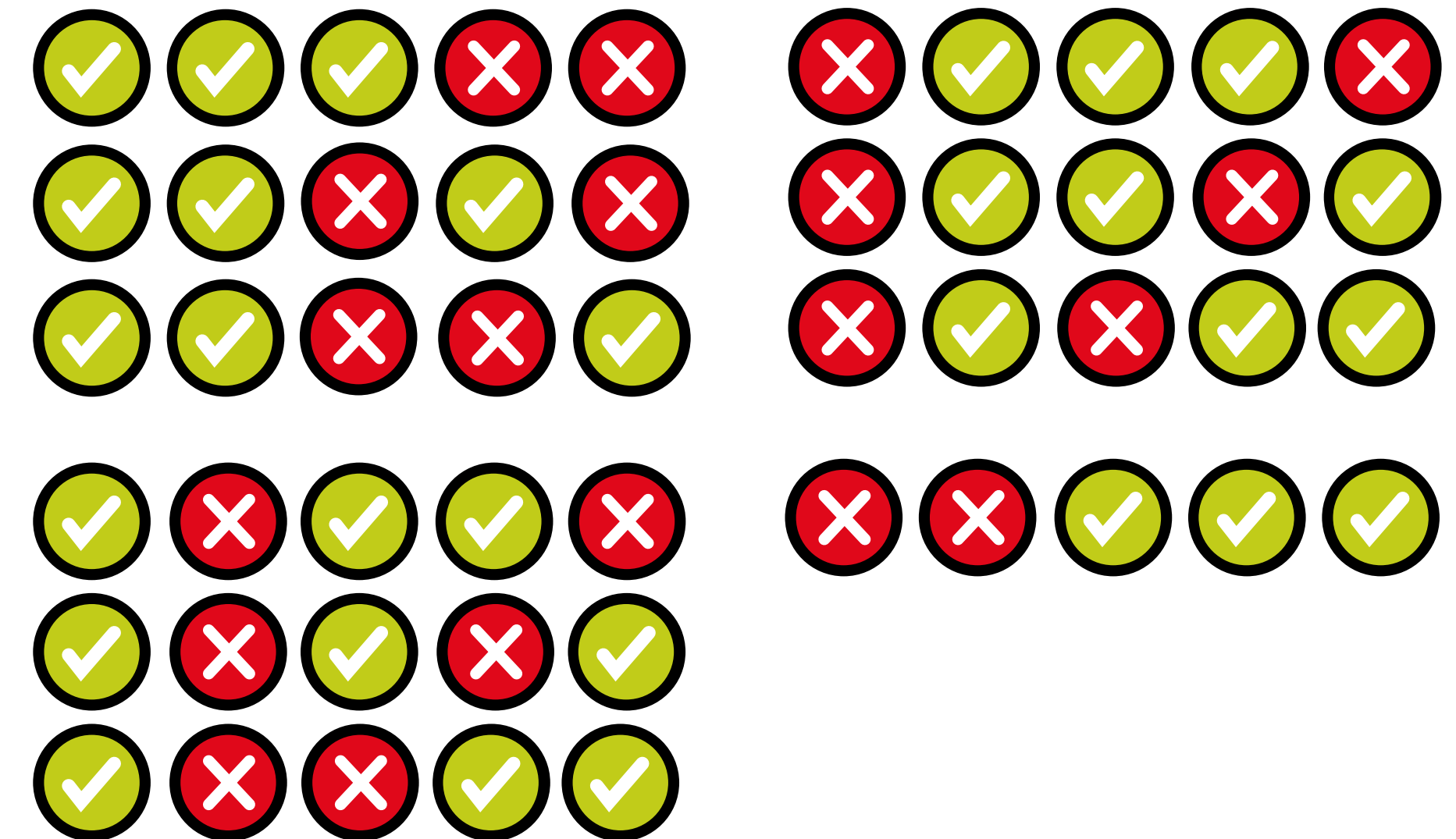


Permutationen

Die Wahrscheinlichkeit von 3 Treffern aus 5 Versuchen ist damit:

$$\begin{aligned}
 P &= \binom{5}{3} P_{\text{Tor}}^3 \cdot (1 - P_{\text{Tor}})^2 \\
 &= \frac{5!}{3! (5-3)!} 0.75^3 \cdot 0.25^2 \\
 &= 10 \cdot 0.75^3 \cdot 0.25^2 \\
 &= 26.4\%
 \end{aligned}$$

10 Möglichkeiten 2 von 5
Elfmetern zu verschießen



Permutationen

Die Wahrscheinlichkeit von 4 Treffern aus 5 Versuchen ist damit:

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

Torchance eines Elfmeters: 75%

Gesucht: Chance auf 4 Treffer aus 5 Versuchen

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

$$P = 39.6\%$$







Permutationen

Mit dem Binomialkoeffizienten können wir nicht nur die Wahrscheinlichkeit von k Erfolgen aus n Versuchen berechnen ...

...sondern auch die Wahrscheinlichkeit von mindestens bzw. maximal k Erfolgen!

Torchance eines Elfmeters: 75%

Gesucht: Chance auf k Erfolge aus 5 Versuchen

	$k=0$	$P = 0.1\%$
	$k=1$	$P = 1.5\%$
	$k=2$	$P = 8.8\%$
	$k=3$	$P = 26.4\%$
	$k=4$	$P = 39.6\%$
	$k=5$	$P = 23.6\%$







Permutationen

Wahrscheinlichkeit von mindestens 3 Treffer? Summiere über die Wahrscheinlichkeiten für $k=3$, $k=4$ und $k=5$ Treffer!

$$P = 0.264 + 0.396 + 0.236 = 89.6\%$$

Torchance eines Elfmeters: 75%

Gesucht: Chance auf k Erfolge aus 5 Versuchen

	$k=0$	$P = 0.1\%$
	$k=1$	$P = 1.5\%$
	$k=2$	$P = 8.8\%$
	$k=3$	$P = 26.4\%$
	$k=4$	$P = 39.6\%$
	$k=5$	$P = 23.6\%$







Permutationen

Wahrscheinlichkeit von maximal 3 Treffer? Summiere über die Wahrscheinlichkeiten für $k=3$, $k=2$, $k=1$ und $k=0$ Treffer!

$$P = 0.001 + 0.015 + 0.088 + 0.264 = 36.8\%$$

Torchance eines Elfmeters: 75%

Gesucht: Chance auf k Erfolge aus 5 Versuchen

	$k=0$	$P = 0.1\%$
	$k=1$	$P = 1.5\%$
	$k=2$	$P = 8.8\%$
	$k=3$	$P = 26.4\%$
	$k=4$	$P = 39.6\%$
	$k=5$	$P = 23.6\%$

Permutationen

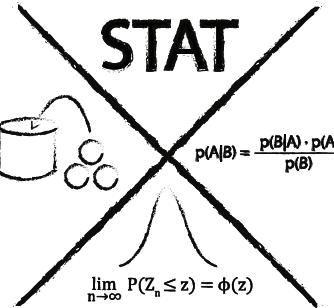
Mit Software können wir uns viel Rechenarbeit ersparen.
In Excel verwenden wir:

=BINOM.VERT(Zahl_Erfolge;Versuche;Wkt;Kummuliert)

Zahl Erfolge: Wie viel Erfolge gewünscht sind.
Versuche: Wie viel Versuche erlaubt sind.
Wkt: Einzelwahrscheinlichkeit Erfolg

Genau n Erfolge Kumuliert auf FALSCH
Mindestens n Erfolge Kumuliert auf WAHR

Erfolge	Genau	Maximal	Minimal
0	0,1%	0,1%	100,0%
1	1,5%	1,6%	99,9%
2	8,8%	10,4%	98,4%
3	26,4%	36,7%	89,6%
4	39,6%	76,3%	63,3%
5	23,7%	100,0%	23,7%



Permutationen

Die Polizei lasert 10 Fahrzeuge auf der B30 zwischen Biberach und Ravensburg.

Die Wahrscheinlichkeit, dass ein Fahrzeug zu schnell ist, beträgt 20%

a) Wie hoch ist die Wahrscheinlichkeit, dass die Polizei genau 4 Bußgelder verhängt?

b) Wie hoch ist die Wahrscheinlichkeit, dass die Polizei nicht mehr als ein Bußgeld verhängt?

Der Zoll kontrolliert bei einer Stichprobe 3 Lkw auf Schmuggelware.

Die Wahrscheinlichkeit, dass ein Lkw Schmuggelware transportiert ist 5%

a) Stelle die Wahrscheinlichkeiten, k Treffer zu landen, in einer Tabelle dar!

b) Wie hoch ist die Wahrscheinlichkeit, dass der Zoll mindestens eine Verhaftung tätigt?

Permutationen

Die Polizei lasert 10 Fahrzeuge auf der B30 zwischen Biberach und Ravensburg.

Die Wahrscheinlichkeit, dass ein Fahrzeug zu schnell ist, beträgt 20%

a) Wie hoch ist die Wahrscheinlichkeit, dass die Polizei genau 4 Bußgelder verhängt?

Verwende $P_B = 0.2$ als Symbol für Wahrscheinlichkeit, dass ein Fahrzeug zu schnell ist.

$$\begin{aligned}
 P_4 &= \binom{10}{4} P_B^4 \cdot (1-P_B)^6 \\
 &= \frac{10!}{4! (10-4)!} 0.20^4 \cdot 0.80^6 \\
 &= \frac{3628800}{24 \cdot 720} 0.000419 \\
 &= 8.8\%
 \end{aligned}$$

Permutationen

Die Polizei lasert 10 Fahrzeuge auf der B30 zwischen Biberach und Ravensburg.

Die Wahrscheinlichkeit, dass ein Fahrzeug zu schnell ist, beträgt 20%

b) Wie hoch ist die Wahrscheinlichkeit, dass die Polizei nicht mehr als ein Bußgeld verhängt?

Wir addieren die Wahrscheinlichkeiten von null oder einem Bußgeld: 37.6%

$$\begin{aligned}
 P_1 &= \binom{10}{9} P_B^1 \cdot (1-P_B)^9 \\
 &= \frac{10!}{9! (10-9)!} 0.20^1 \cdot 0.80^9 = 26.84\%
 \end{aligned}$$

$$\begin{aligned}
 P_0 &= \binom{10}{10} P_B^0 \cdot (1-P_B)^{10} \\
 &= \frac{10!}{10! (10-10)!} 0.80^{10} = 10.74\%
 \end{aligned}$$

Permutationen

Der Zoll kontrolliert bei einer Stichprobe 3 Lkw auf Schmuggelware.

Die Wahrscheinlichkeit, dass ein Lkw Schmuggelware transportiert ist 5%

a) Stelle die Wahrscheinlichkeiten, k Treffer zu landen, in einer Tabelle dar!

$$P_k = \binom{3}{k} 0.05^k \cdot (1-0.05)^{3-k}$$

Erfolge	Wahrscheinlichkeit
0	85.74%
1	13.54%
2	0.71%
3	0.01%

Permutationen

Der Zoll kontrolliert bei einer Stichprobe 3 Lkw auf Schmuggelware.

Die Wahrscheinlichkeit, dass ein Lkw Schmuggelware transportiert ist 5%

b) Wie hoch ist die Wahrscheinlichkeit, dass der Zoll mindestens eine Verhaftung tätigt?

Addiere die unteren 3 Werte zu 14.26%

$$P_k = \binom{3}{k} 0.05^k \cdot (1-0.05)^{3-k}$$

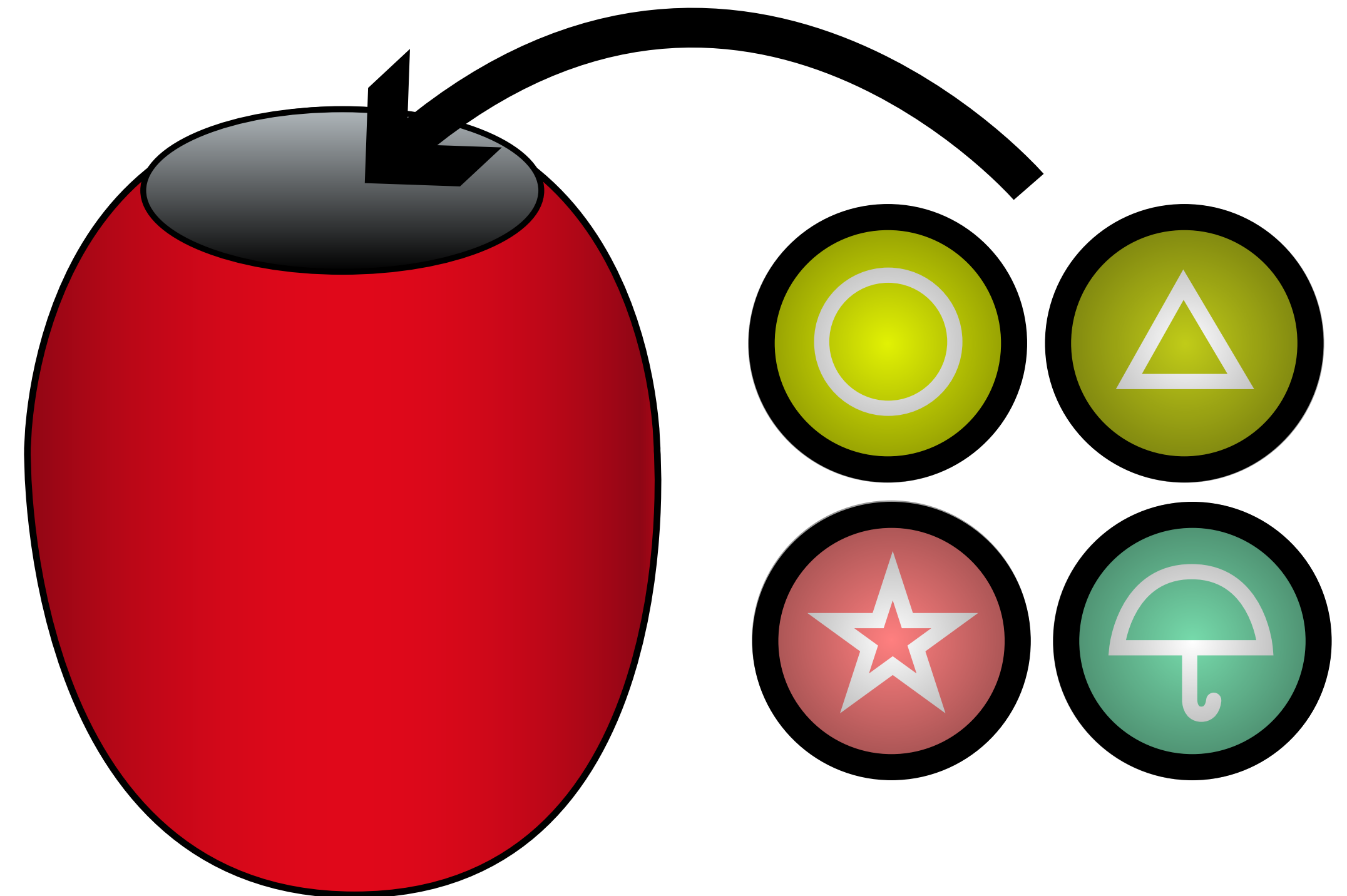
Erfolge	Wahrscheinlichkeit
0	85.74%
1	13.54%
2	0.71%
3	0.01%

Urnenmodelle

Der hier entdeckte **Binomialkoeffizient** „n über k“ erlaubt uns die Berechnung der Wahrscheinlichkeit von k Erfolgen bei n Versuchen.

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

Er ermöglicht, uns aber auch die vier Urnenmodelle anzuwenden und damit noch mehr Kombinatorikaufgaben zu lösen!



Urnenmodelle

Wir haben eine Urne mit n verschiedenen Kugeln.

Wir ziehen nacheinander k dieser n Kugeln und notieren uns das Resultat.

Die vier Varianten ergeben sich aus den beiden Dimensionen **Zurücklegen** und **Reihenfolge**.

Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Die Tabelle rechts zeigt die Anzahl an möglichen Ergebnissen in den einzelnen Varianten.

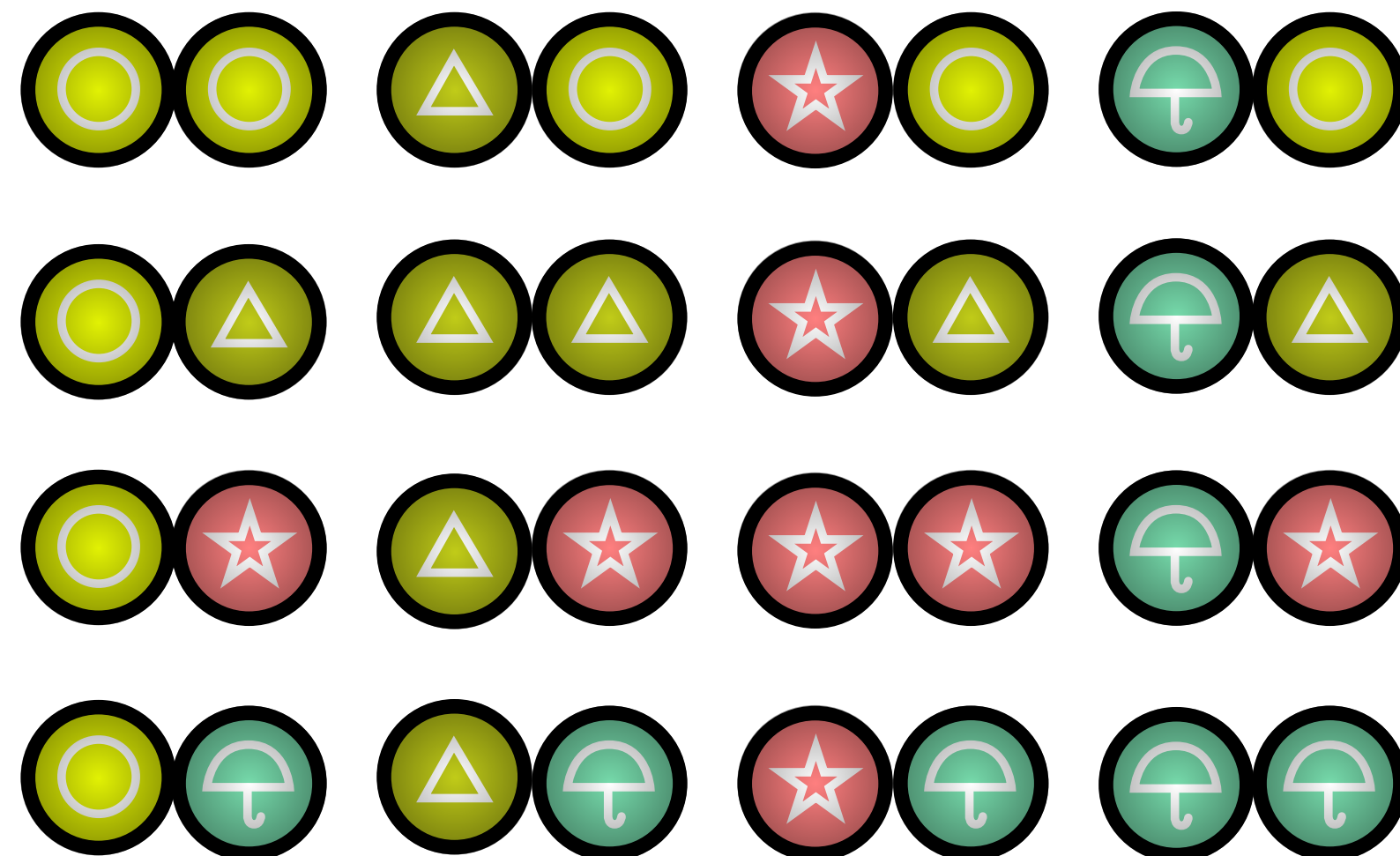
Wie hoch wären diese, bei $n=4$ und $k=2$?

Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Ziehen mit Zurücklegen und Reihenfolge relevant:

$$n^k = 4^2 = 16$$

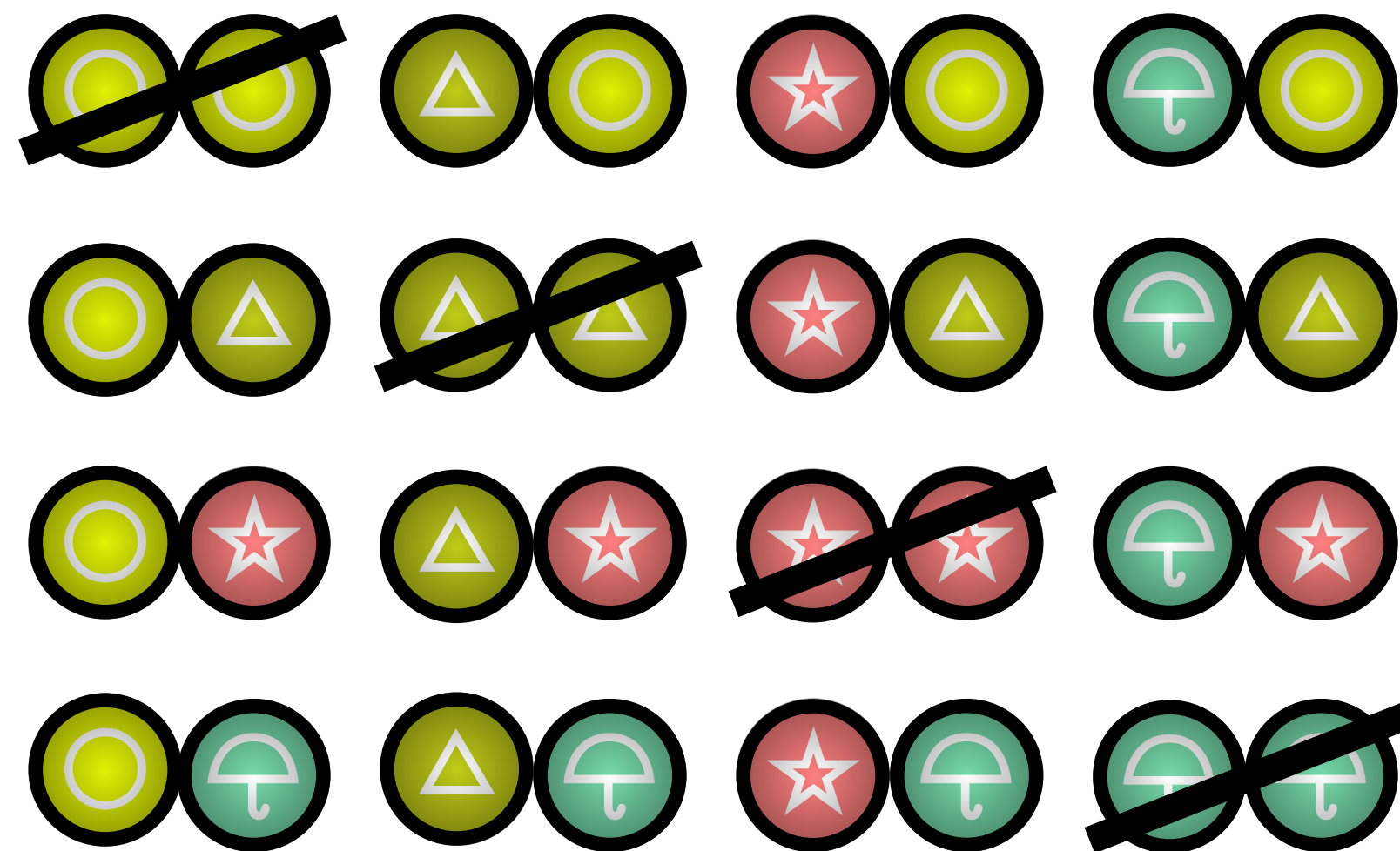


Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Ziehen ohne Zurücklegen und Reihenfolge relevant:

$$\frac{n!}{(n-k)!} = \frac{4!}{(4-2)!} = 12$$

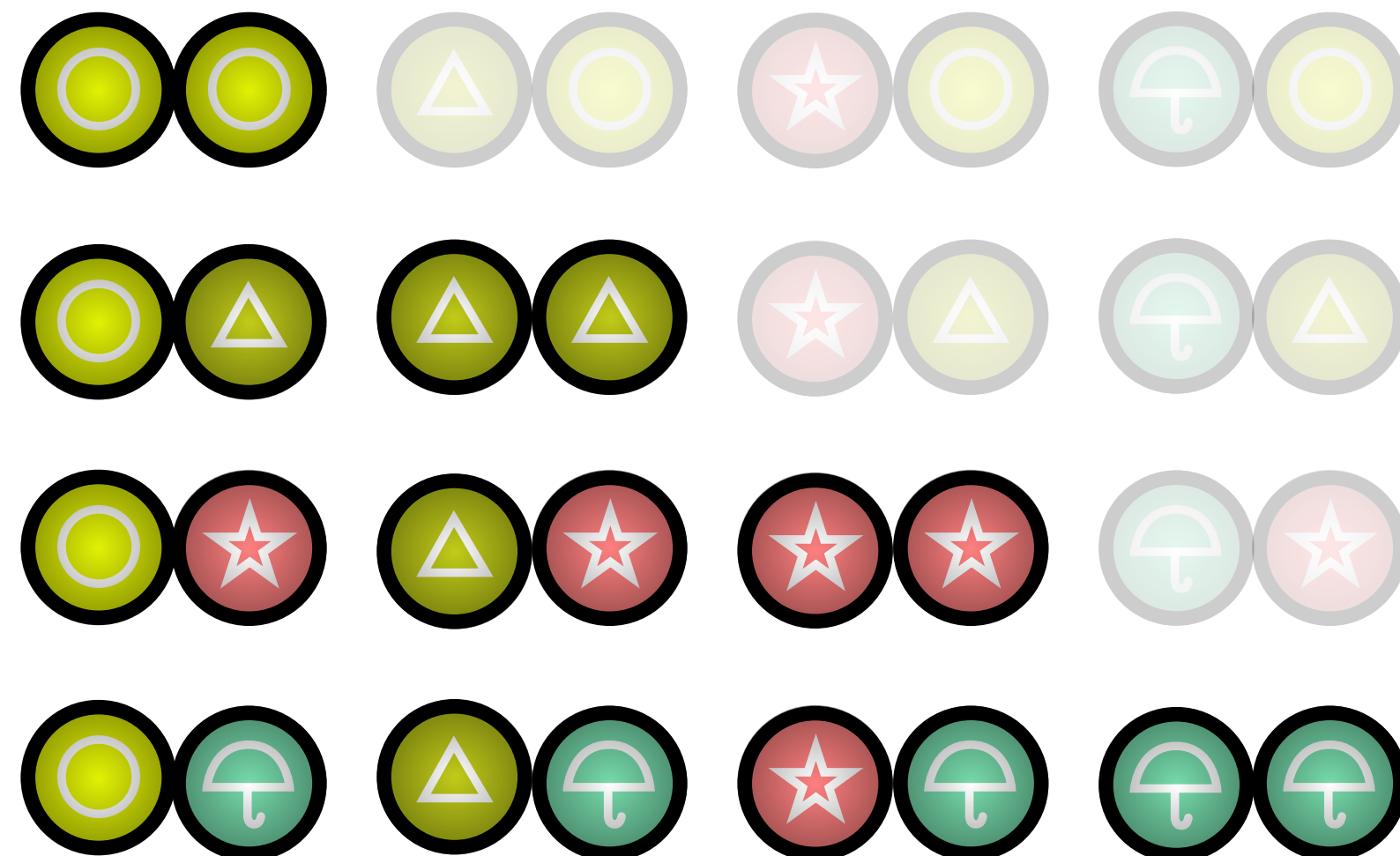


Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Ziehen mit Zurücklegen und Reihenfolge irrelevant:

$$\binom{n+k-1}{k} = \binom{4+2-1}{2} = \frac{5!}{2!(5-2)!} = 10$$

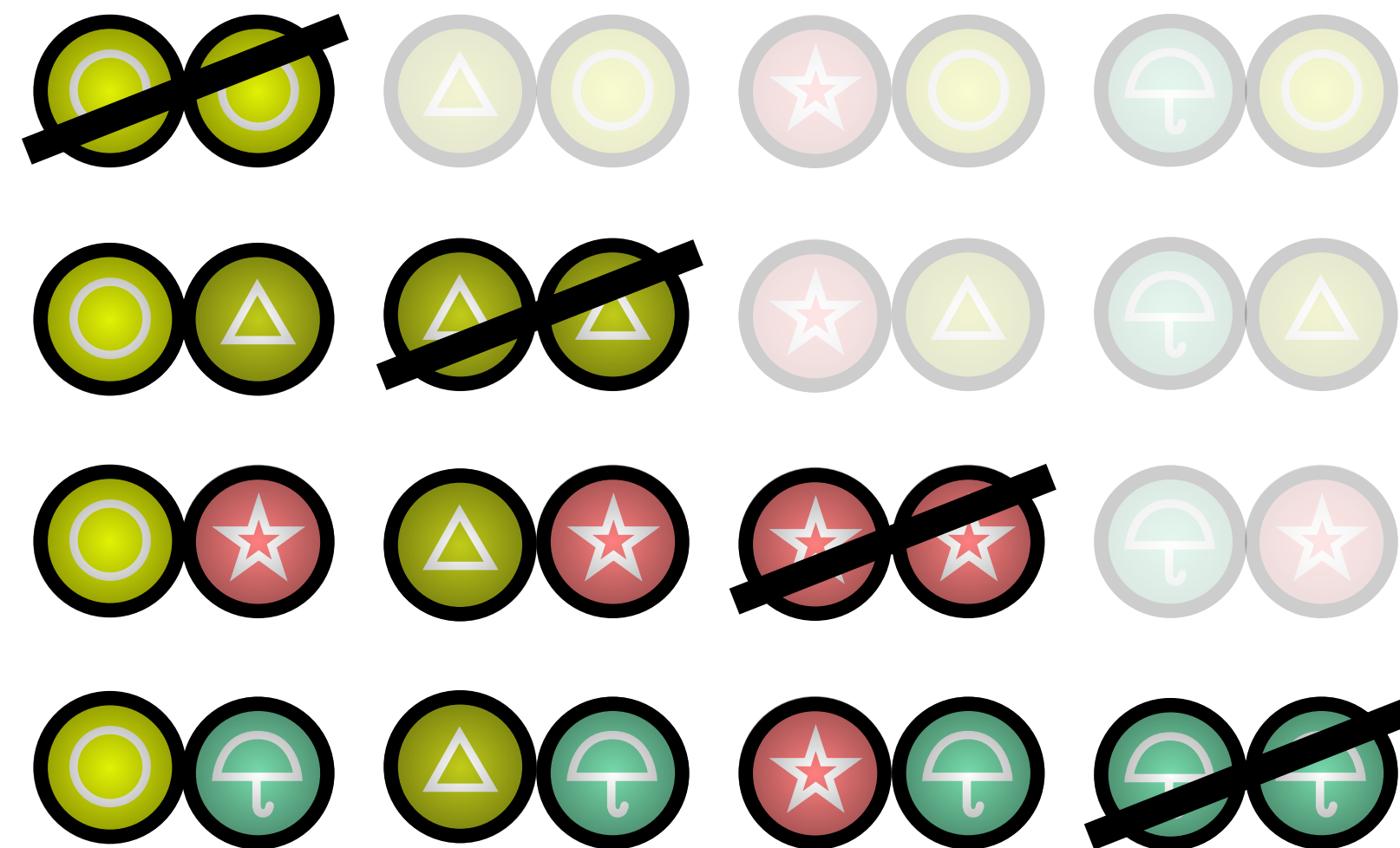


Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Ziehen ohne Zurücklegen und Reihenfolge irrelevant:

$$\binom{n}{k} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$$



Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

Was machen wir mit diesen Werten? Wie berechnen wir damit Wahrscheinlichkeiten?

Grundidee: Wir definieren Ereignisse und berechnen deren Wahrscheinlichkeit als Quotienten:

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$


Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Urnenmodelle

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$

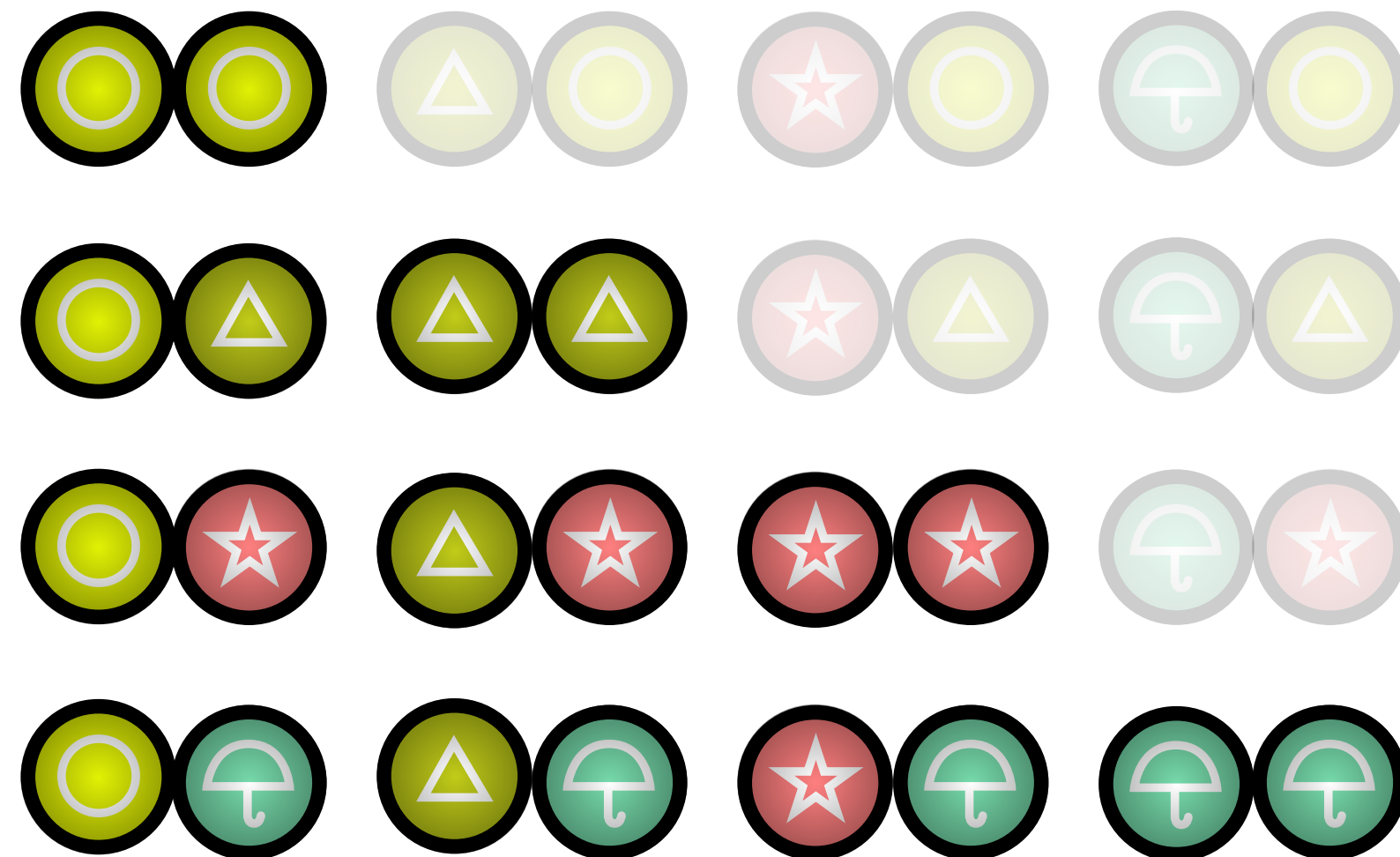
Vorsicht! Diese Vorgehensweise funktioniert nur, wenn jedes mögliche Ergebnis mit derselben Wahrscheinlichkeit eintritt.


Schauen wir uns noch mal die Folien 177 bis 180 an, dann erkennen wir, dass dies bei einem Urnenmodell nicht immer der Fall ist!

Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$ 	$\binom{n}{k}$

Urnenmodelle

Auch wenn wir „Kreis Dreieck“ und „Dreieck Kreis“ als dasselbe Ereignis behandeln ist es doppelt so wahrscheinlich wie die Ereignisse auf der Diagonalen.



Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$ 	$\binom{n}{k}$

Urnenmodelle (6 aus 49)

Ziehen ohne Zurücklegen und Reihenfolge irrelevant.

Beispiel Lotto 6 aus 49

Ereignis 5 Richtige aus 49, Superzahl irrelevant

Wie wahrscheinlich ist das Ereignis? Dazu müssen wir Zähler und Nenner unseres Quotienten berechnen!

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$



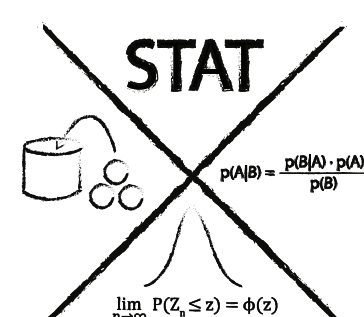
1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X X X X X X X							X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X



LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (6 aus 49)

Ziehen ohne Zurücklegen und Reihenfolge irrelevant.

Beispiel Lotto 6 aus 49

Ereignis 5 Richtige aus 49, Superzahl irrelevant

Im Nenner brauchen wir die Anzahl der Ergebnisse, die beim Ziehen von 6 aus 49 entstehen können:

$$\binom{n}{k} = \binom{49}{6} = \frac{49!}{6! (49-6)!} = 13.983.816$$

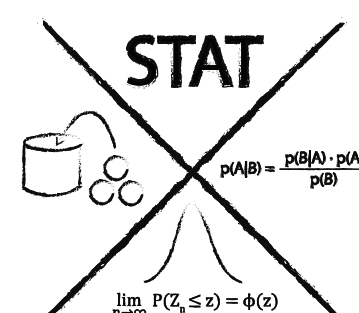


1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X	X	X	X	X	X	7	X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X

LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (6 aus 49)

Ziehen ohne Zurücklegen und Reihenfolge irrelevant.

Beispiel Lotto 6 aus 49

Ereignis 5 Richtige aus 49, Superzahl irrelevant

Im Zähler brauchen wir die Anzahl der Ergebnisse, bei denen man 5 von 6 Richtigen auswählt:

$$\binom{6}{5} \binom{43}{1} = \frac{6!}{5!(6-5)!} \frac{43!}{1!(43-1)!} = 258$$

\swarrow 1 von 43 Falschen
 \swarrow 5 von 6 Richtigen

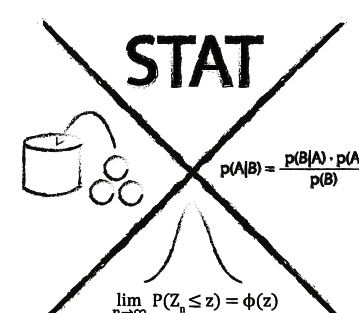


1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
<hr/>													
X	X	X	X	X	X	7	X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X

LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (6 aus 49)

Ziehen ohne Zurücklegen und Reihenfolge irrelevant.

Beispiel Lotto 6 aus 49

Ereignis 5 Richtige aus 49, Superzahl irrelevant

Da von 13.9 Millionen möglichen Ergebnissen 258 zur geforderten Gewinnklasse führen gilt:

$$P(A) = \frac{258}{13.983.816} = 0.001844 \%$$



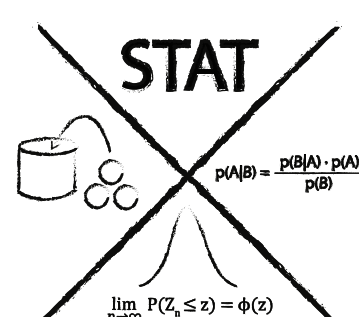
1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X X X X X X X							X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X



LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (Spiel 77)

Ziehen mit Zurücklegen und Reihenfolge relevant.

Beispiel Spiel 77

Ereignis 5 richtige Endziffern

Wie wahrscheinlich ist das Ereignis? Dazu müssen wir Zähler und Nenner unseres Quotienten berechnen!

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$



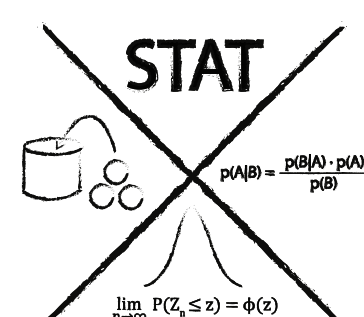
1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X X X X X X X							X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X



LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (Spiel 77)

Ziehen mit Zurücklegen und Reihenfolge relevant.

Beispiel Spiel 77

Ereignis 5 richtige Endziffern

Nenner: Anzahl der Ergebnisse, die beim Ziehen der 7-stelligen Losnummer entstehen können:

$$n^k = 10^7 = 1.000.000$$



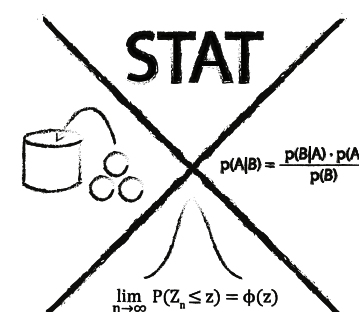
1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X X X X X X X							X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X



LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (Spiel 77)

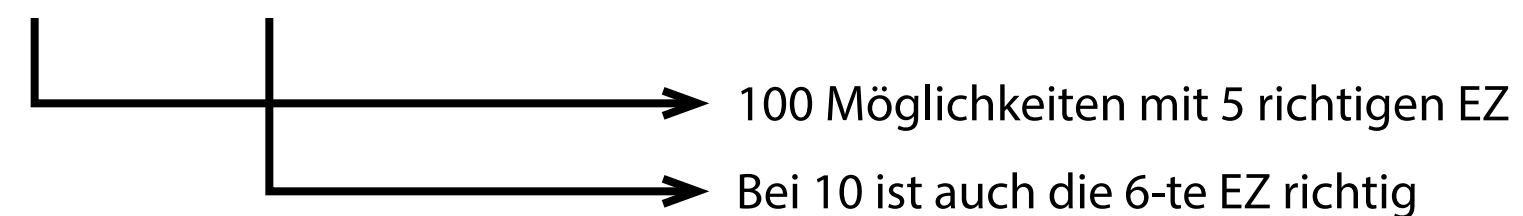
Ziehen mit Zurücklegen und Reihenfolge relevant.

Beispiel Spiel 77

Ereignis 5 richtige Endziffern

Zähler: Anzahl der Ergebnisse, bei denen 5 Endziffern richtig sind und die sechste Endziffer falsch ist.

$$10^2 - 10 = 90$$

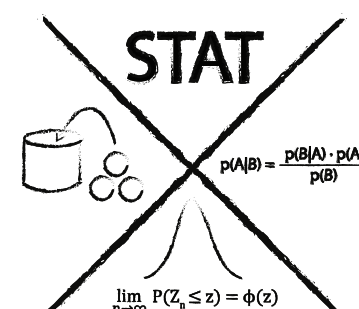


1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
<hr/>													
X	X	X	X	X	X	7	X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X

LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (Spiel 77)

Ziehen ohne Zurücklegen und Reihenfolge irrelevant.

Beispiel Spiel 77

Ereignis 5 richtige Endziffern

Da von 10 Millionen möglichen Ergebnissen 90 zur geforderten Gewinnklasse führen gilt:

$$P(A) = \frac{90}{10000000} = 0.0009 \%$$



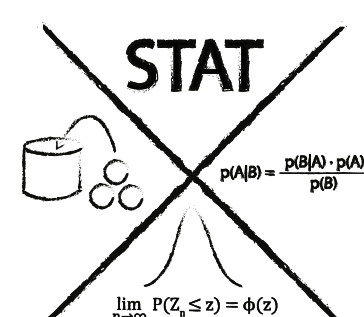
1	2	3	4	5	6	7	1	X	3	4	5	6	7
8	9	X	11	12	13	14	8	9	X	11	X	13	X
15	16	17	18	X	20	21	15	X	17	18	19	20	21
22	X	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	X	33	34	35	29	30	31	32	33	34	35
X	37	38	39	40	X	42	36	X	38	39	40	41	42
43	44	45	46	47	48	49	43	44	45	46	47	48	49
X X X X X X X							X	2	3	4	5	6	X
8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	X	20	21
22	23	23	25	26	27	28	22	23	23	25	26	27	28
29	30	31	32	33	34	35	29	30	X	32	33	34	35
36	37	38	39	40	41	42	36	37	38	39	40	41	42
43	44	45	46	47	48	49	X	44	45	46	47	48	X



LOSNUMMER

↳ **7 5 1 2 9 1 9**

Superzahl



Urnenmodelle (Kniffel)

Ziehen mit Zurücklegen und Reihenfolge irrelevant.

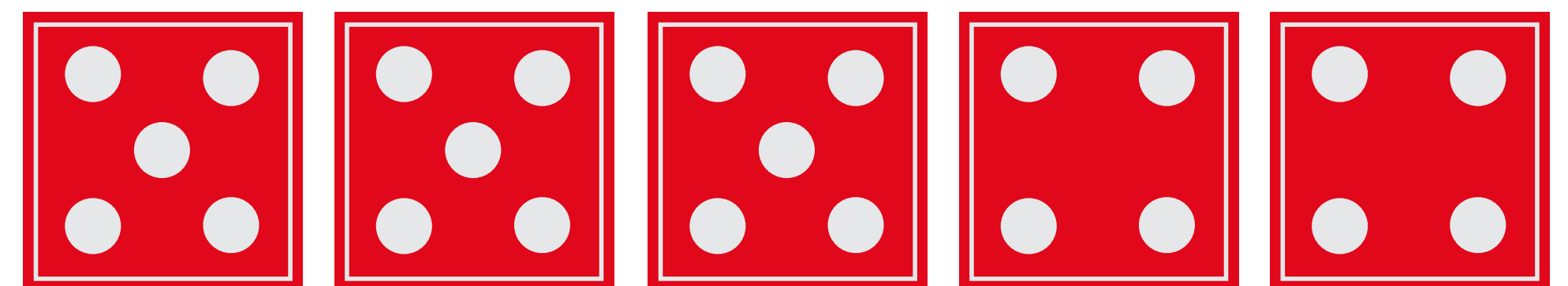
Beispiel Kniffel (5 Würfel)

Ereignis Full House im ersten Versuch

Wie wahrscheinlich ist das Ereignis? Dazu müssen wir Zähler und Nenner unseres Quotienten berechnen!

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$

Kniffel




KNIFFEL-BLOCK	Spieler 1	Spieler 2	Spieler 3
Full-House	✓		
Viererpasch			

Urnenmodelle (Kniffel)

Vorsicht Falle: Die Anzahl der regeltechnischen Ergebnisse, die beim Würfeln von 5 Würfeln entstehen können, ist zwar ...

$$\binom{6+5-1}{5} = \frac{10!}{5!(10-5)!} = 252$$

... aber nicht jedes dieser Ergebnisse ist gleich wahrscheinlich. Für die Berechnung des Nenners verwenden wir deshalb die Formel n^k .

Mögliche Ergebnisse	Mit Zurücklegen	Ohne Zurücklegen
Reihenfolge relevant	n^k	$\frac{n!}{(n-k)!}$
Reihenfolge irrelevant	$\binom{n+k-1}{k}$ 	$\binom{n}{k}$

Urnenmodelle (Kniffel)

Ziehen mit Zurücklegen und Reihenfolge irrelevant.

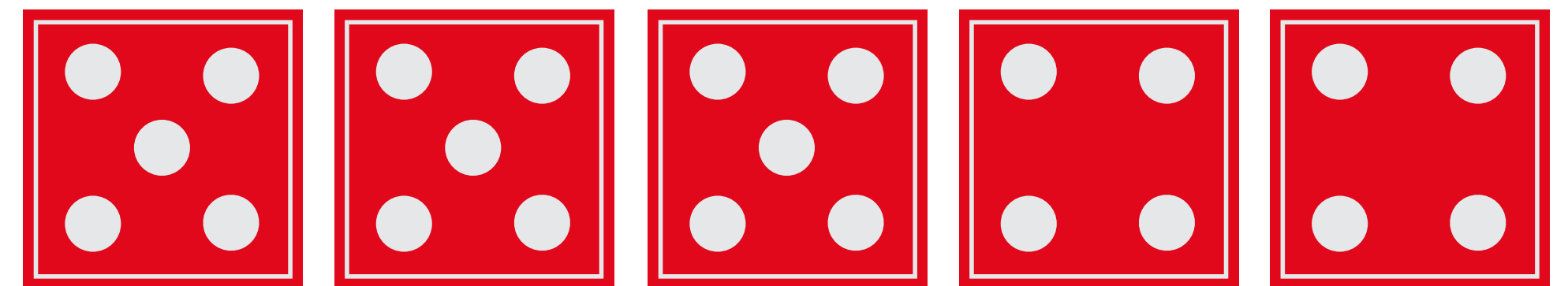
Beispiel Kniffel (5 Würfel)

Ereignis Full House im ersten Versuch

Nenner: Anzahl der möglichen Ergebnisse bei fünf 6-seitigen Würfeln unter Berücksichtigung mehrerer Reihenfolgen:

$$6^5 = 7776$$

Kniffel



KNIFFEL-BLOCK	Spieler 1	Spieler 2	Spieler 3
Full-House	✓		
Viererpasch			

Urnenmodelle (Kniffel)

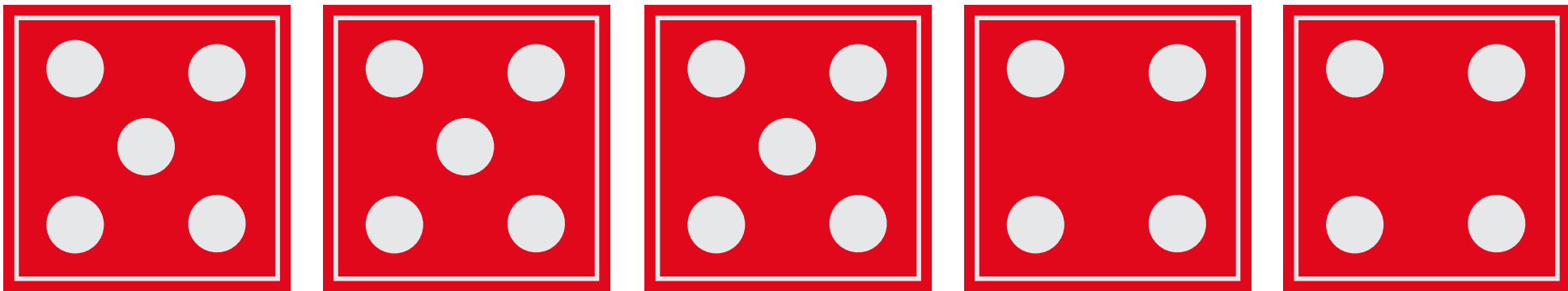
Ziehen mit Zurücklegen und Reihenfolge irrelevant.

Beispiel Kniffel (5 Würfel)
Ereignis Full House im ersten Versuch

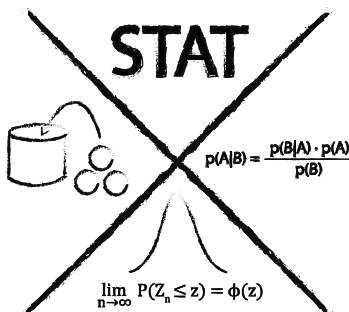
Zähler: Als Erstes berechnen wir die Anzahl an verschiedenen Full Houses. Wir haben 6 Möglichkeiten für den Drilling und 5 Möglichkeiten für das Paar.

$6 \cdot 5 = 30$

Kniffel



KNIFFEL-BLOCK	Spieler 1	Spieler 2	Spieler 3
Full-House	✓		
Viererpasch			



Urnenmodelle (Kniffel)

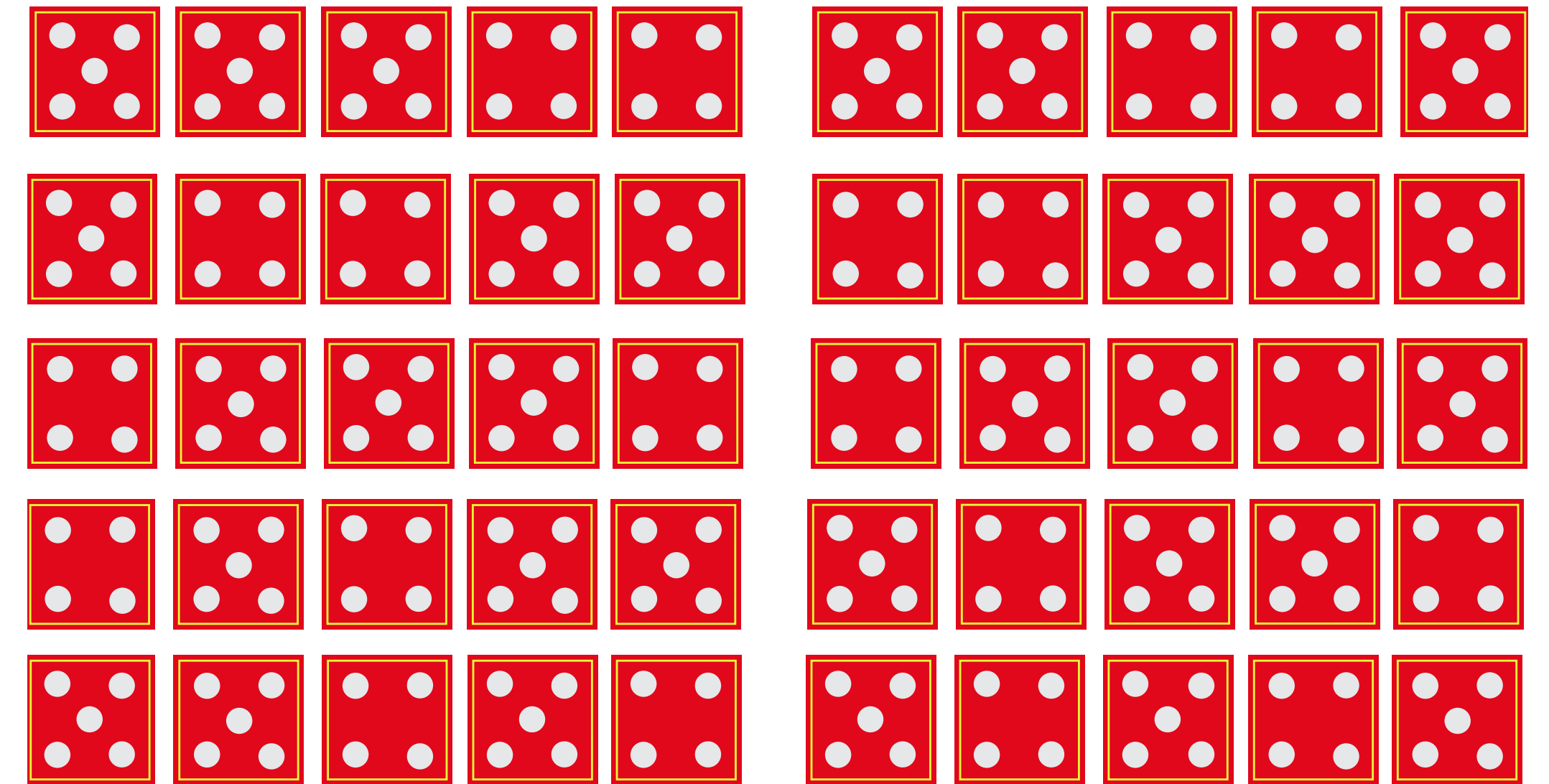
Ziehen mit Zurücklegen und Reihenfolge irrelevant.

Beispiel Kniffel (5 Würfel)

Ereignis Full House im ersten Versuch

Zähler: Diese 30 Varianten multiplizieren wir mit der Anzahl Anordnungen eines Fullhouse auf 5 Würfel. Es gibt ...

$$\binom{5}{3} = \binom{5}{2} = 10$$



10 Kombinationen die spieltechnisch alle als
4er Pash und 5er Drilling zählen

Urnenmodelle (Kniffel)

Ziehen mit Zurücklegen und Reihenfolge irrelevant.

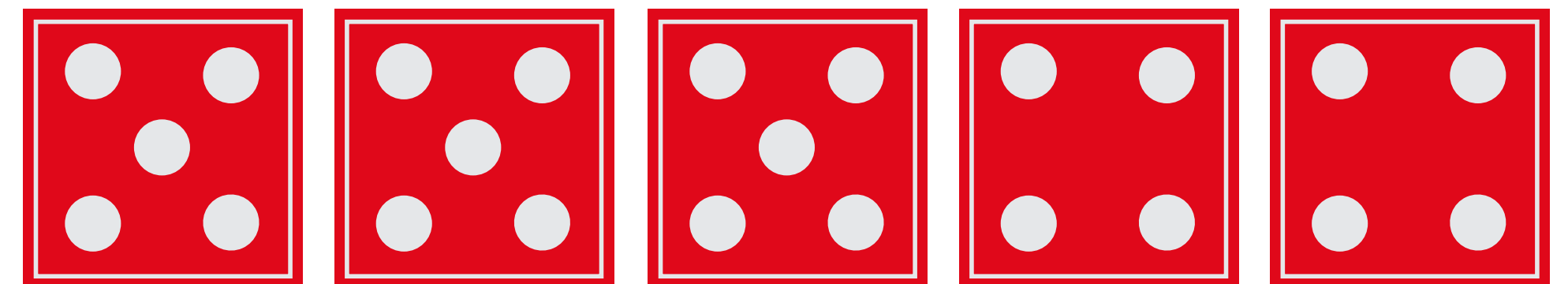
Beispiel Kniffel (5 Würfel)

Ereignis Full House im ersten Versuch

Da von 7776 möglichen Ergebnissen 300 ein Full House sind, gilt:

$$P(A) = \frac{300}{7776} = 3.85 \%$$

Kniffel



KNIFFEL-BLOCK	Spieler 1	Spieler 2	Spieler 3
Full-House	✓		
Viererpasch			

Urnenmodelle (ABC-Würfel)

Ziehen ohne Zurücklegen und Reihenfolge relevant.

Beispiel ABC-Würfel (26 Buchstaben ohne Umlaute)

Ereignis Mit vier zufälligen Würfeln eine Zahl schreiben.

Wie wahrscheinlich ist das Ereignis? Dazu müssen wir Zähler und Nenner unseres Quotienten berechnen!

$$P(A) = \frac{\text{Zum Ereignis A passende Ergebnisse}}{\text{Mögliche Ergebnisse}}$$



Urnenmodelle (ABC-Würfel)

Ziehen ohne Zurücklegen und Reihenfolge relevant.

Beispiel ABC-Würfel (26 Buchstaben ohne Umlaute)

Ereignis Mit vier zufälligen Würfeln eine Zahl schreiben.

Nenner: Anzahl der Ergebnisse, die beim Anordnen von 4 aus 26 Würfeln entstehen können.

$$\frac{n!}{(n-k)!} = \frac{26!}{(26-4)!} = 358800$$



Urnenmodelle (ABC-Würfel)

Ziehen ohne Zurücklegen und Reihenfolge relevant.

Beispiel ABC-Würfel (26 Buchstaben ohne Umlaute)

Ereignis Mit vier zufälligen Würfeln eine Zahl schreiben.

Da von 358800 möglichen Ergebnissen 6 eine korrekt geschriebene Zahl sind, gilt:

$$P(A) = \frac{6}{358.800} = 0.001672 \%$$



Urnenmodelle (Eurojackpot)

Beim Eurojackpot müssen zwei unabhängige Tipps abgegeben werden: 5 aus 50 und 2 aus 12 Eurozahlen.

- Berechne die Wahrscheinlichkeit beim 5 aus 50 genau 4 Richtige zu treffen.
- Berechne die Wahrscheinlichkeit beim 2 aus 12 beide Eurozahlen zu treffen.
- Wie hoch ist die Wahrscheinlichkeit für die Gewinnklasse 4 Richtige plus 2 Eurozahlen (Gewinn ca. 5000€).

Hier **5 aus 50** ankreuzen

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	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Hier **2 aus 12** ankreuzen

①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
⑪	⑫								

Urnenmodelle (Geburtstage)

Wie hoch ist die Wahrscheinlichkeit, dass in einer 6er WG jeder in einem anderen Monat Geburtstag hat?

Wie hoch ist die Wahrscheinlichkeit, dass in einer 6er WG genau 3 Personen im selben Monat Geburtstag haben?

Nehme dabei an, dass jeder Geburtsmonat mit $P=1/12$ gleich wahrscheinlich ist.



Urnenmodelle (Eurojackpot)

$$a) \binom{n}{k} = \binom{50}{5} = \frac{50!}{5!(50-5)!} = 2.118.760$$

$$\binom{5}{4} \binom{45}{1} = \frac{5!}{4!(5-4)!} \frac{45!}{1!(45-1)!} = 225$$

$\binom{45}{1}$ → 1 von 45 Falschen
 $\binom{5}{4}$ → 4 von 5 Richtigen

$$P(A) = \frac{225}{2118760} = 0.0106\%$$

$$b) \binom{n}{k} = \binom{12}{2} = \frac{12!}{2!(12-2)!} = 66$$

$$P(B) = \frac{1}{66} = 1.51\%$$

c) A und B sind unabhängig, also gilt:

$$P(A \cap B) = P(A) \cdot P(B) = 0.000160\%$$

Urnenmodelle (Geburtstage)

Es gibt 12376 verschiedene Häufigkeitsverteilungen...

$$\binom{n+k-1}{k} = \binom{12+6-1}{6} = \frac{17!}{6!(17-6)!} = 12376$$

...aber nicht alle sind gleich wahrscheinlich. Unser Ansatz ist daher die folgende Zahl an Ergebnissen:

$$n^k = 12^6 = 2985984$$



Urnenmodelle (Geburtstage)

Die Anzahl Kombinationen bei denen alle Bewohner in verschiedenen Monaten Geburtstag haben ist:

$$\binom{12}{6} \cdot 6! = 665280$$

Dabei wählen wir zunächst 6 aus 12 Monaten und verteilen sie auf 6 Personen in beliebiger Anordnung

$$p = \frac{665280}{2985984} = 22.28\%$$



Urnenmodelle (Geburtstage)

Die Anzahl Kombinationen bei denen genau drei Bewohner im gleichen Monat Geburtstag haben ist:

$$\binom{6}{3} \cdot 12 \cdot 11^3 = 319440$$

Dabei wählen wir zunächst 3 aus 6 Personen und weisen diesen einen der 12 Monate zu. Die übrigen Personen werden auf die 11 anderen Monate verteilt.

$$p = \frac{319440}{2985984} = 10.7\%$$

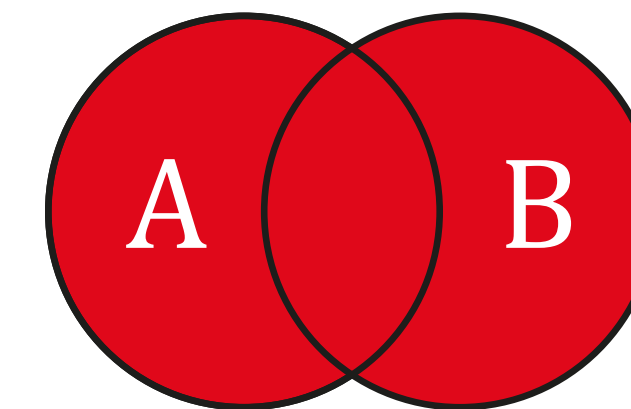


Bedingte Wahrscheinlichkeit

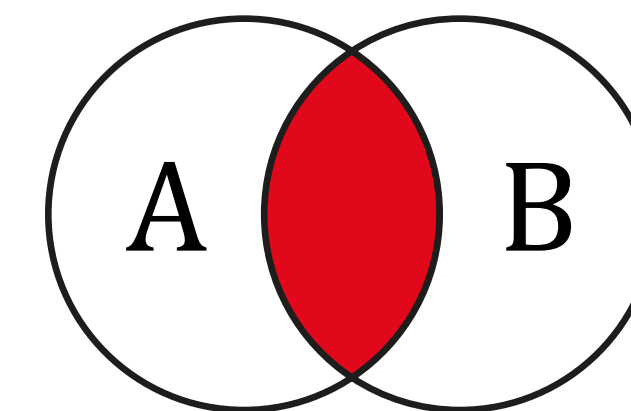
Erinnerung: Die Rechenregeln für kombinierte Ereignisse hängen davon ab, ob sie stochastisch abhängig sind:

Stochastisch unabhängige Ereignisse beeinflussen sich nicht. Das Eintreten des einen macht das andere nicht mehr oder weniger wahrscheinlich.

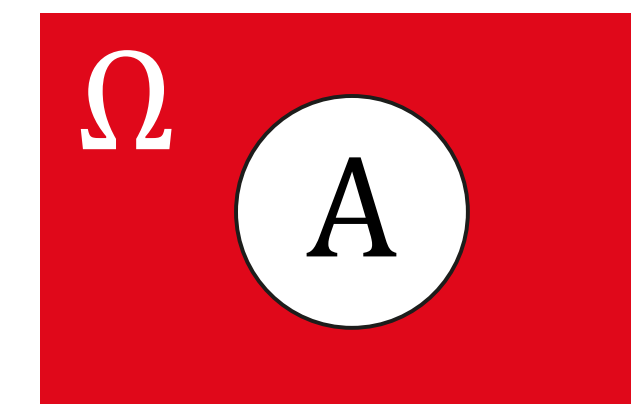
Stochastisch abhängige Ereignisse beeinflussen sich gegenseitig in ihrer Wahrscheinlichkeit.



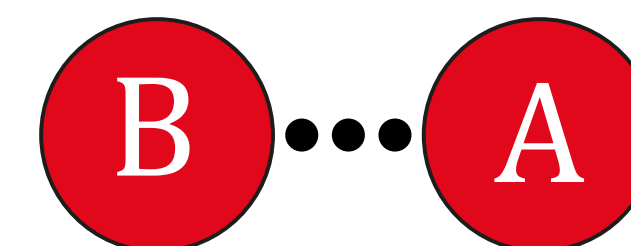
Vereinigung $A \cup B$
A oder B



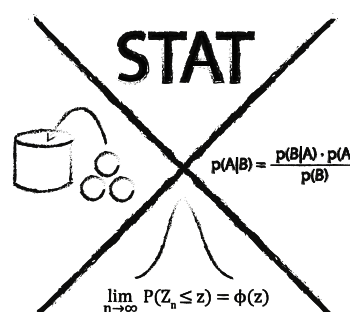
Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



Bedingte Wkt. $A | B$
A nach B



Bedingte Wahrscheinlichkeit

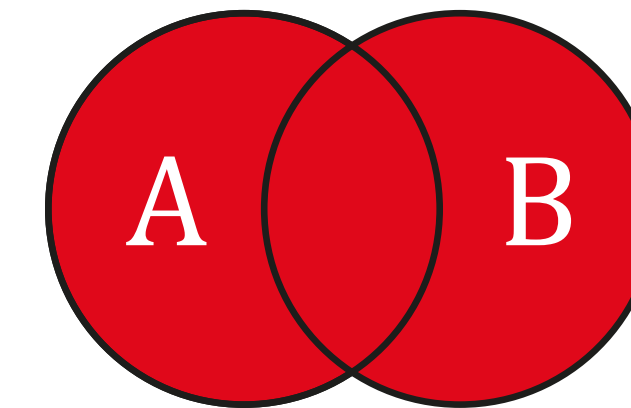
Für **stochastisch abhängige** Ereignisse gelten die folgenden Rechenregeln:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

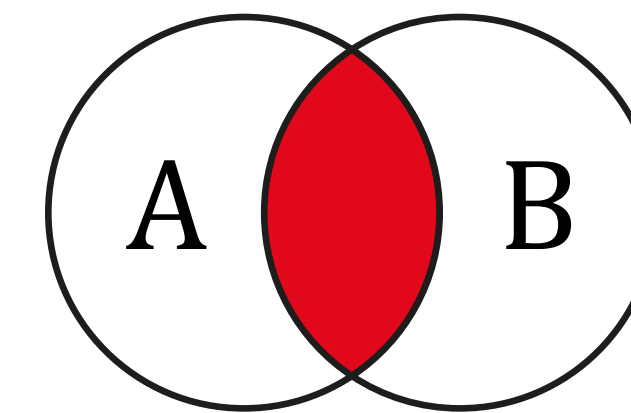
$$P(A \cap B) = P(B) \cdot P(A | B)$$

$$P(A) = P(A|B) P(B) + P(A|\bar{B}) P(\bar{B})$$

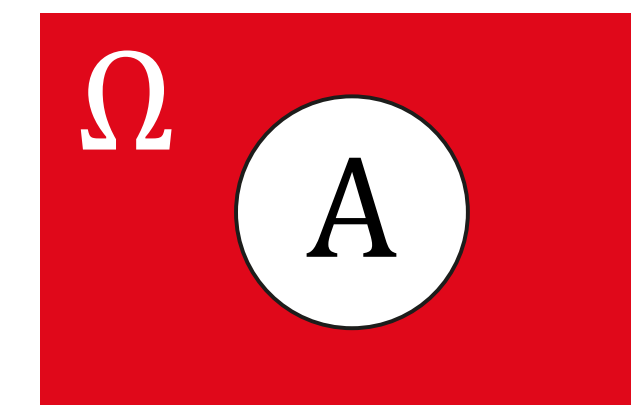
$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$



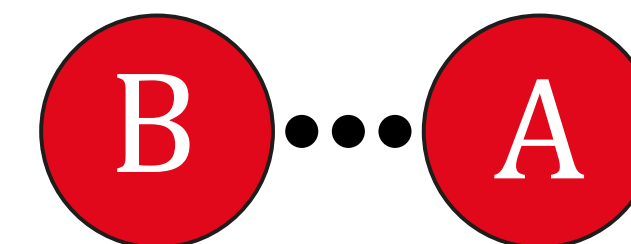
Vereinigung $A \cup B$
A oder B



Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



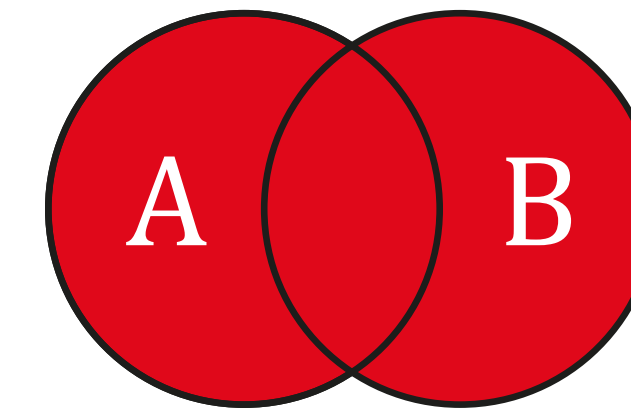
Bedingte Wkt. $A | B$
A nach B

Bedingte Wahrscheinlichkeit

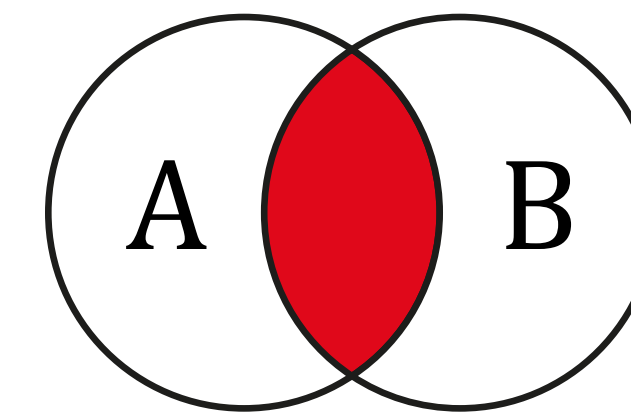
Die ersten beiden Formeln zeigen die **Definition** der bedingten Wahrscheinlichkeit und den direkt daraus hergeleiteten **Multiplikationssatz**.

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

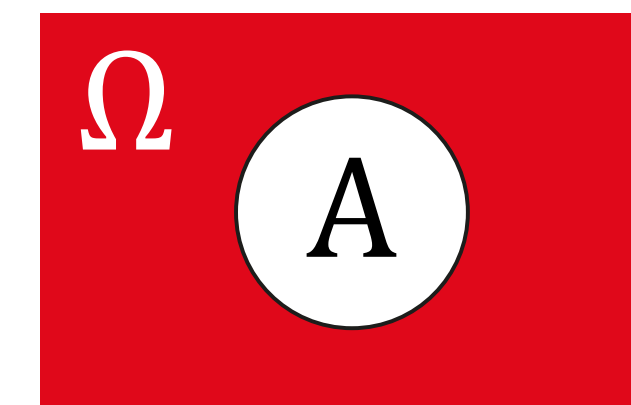
$$P(A \cap B) = P(B) \cdot P(A | B)$$



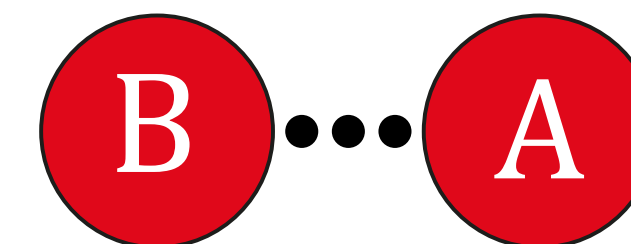
Vereinigung $A \cup B$
A oder B



Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



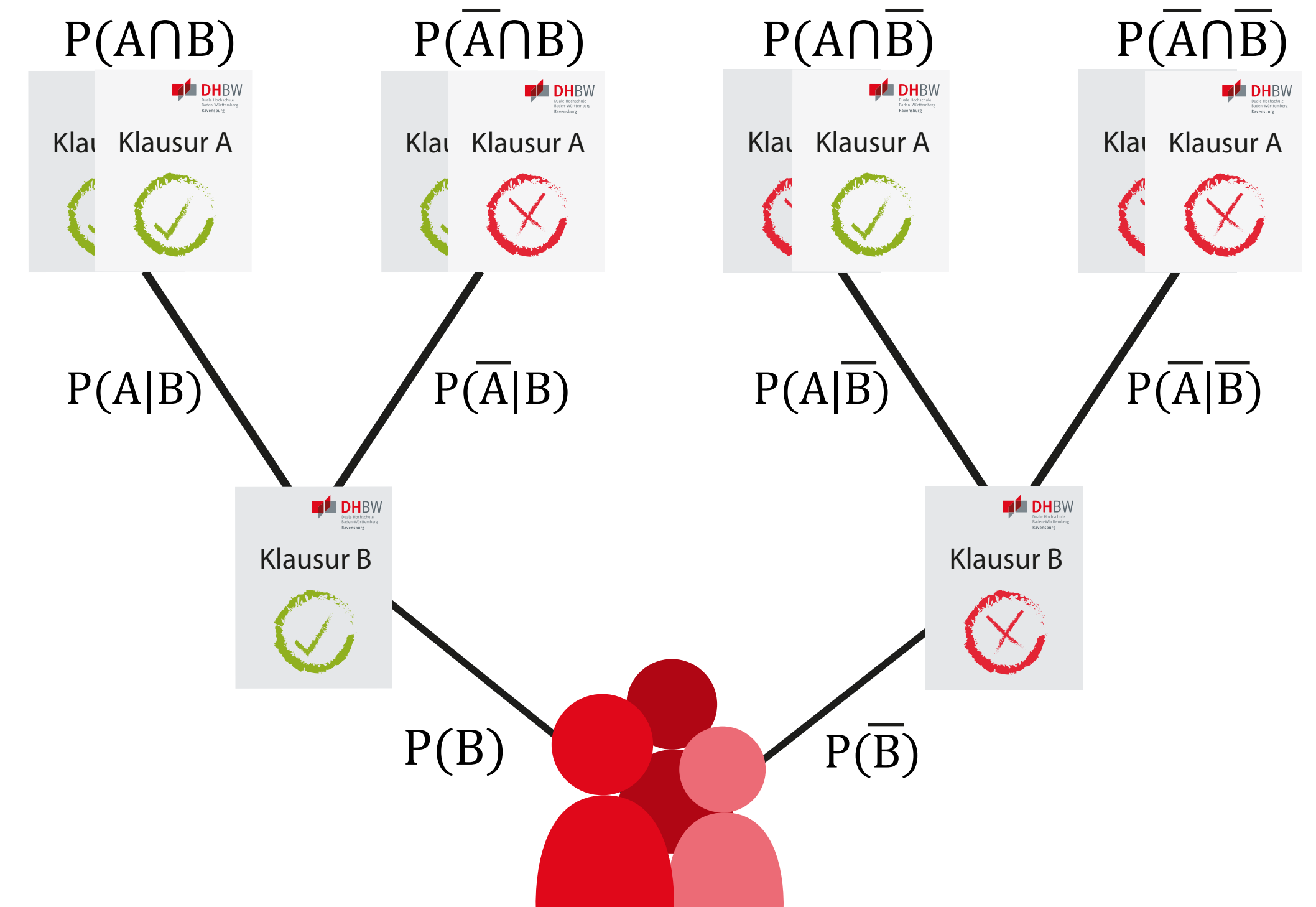
Bedingte Wkt. $A | B$
A nach B

Bedingte Wahrscheinlichkeit

Ein Studierender hat 2 Klausuren in einer Woche. Die erste Klausur besteht er zu 90% Wahrscheinlichkeit.

Besteht er diese, dann besteht er auch die zweite zu 90% Wahrscheinlichkeit. Ansonsten wird er nervös und besteht die zweite nur noch zu 50%.

Wie hoch ist die Wahrscheinlichkeit eine bestimmte Kombination der Klausuren zu bestehen?



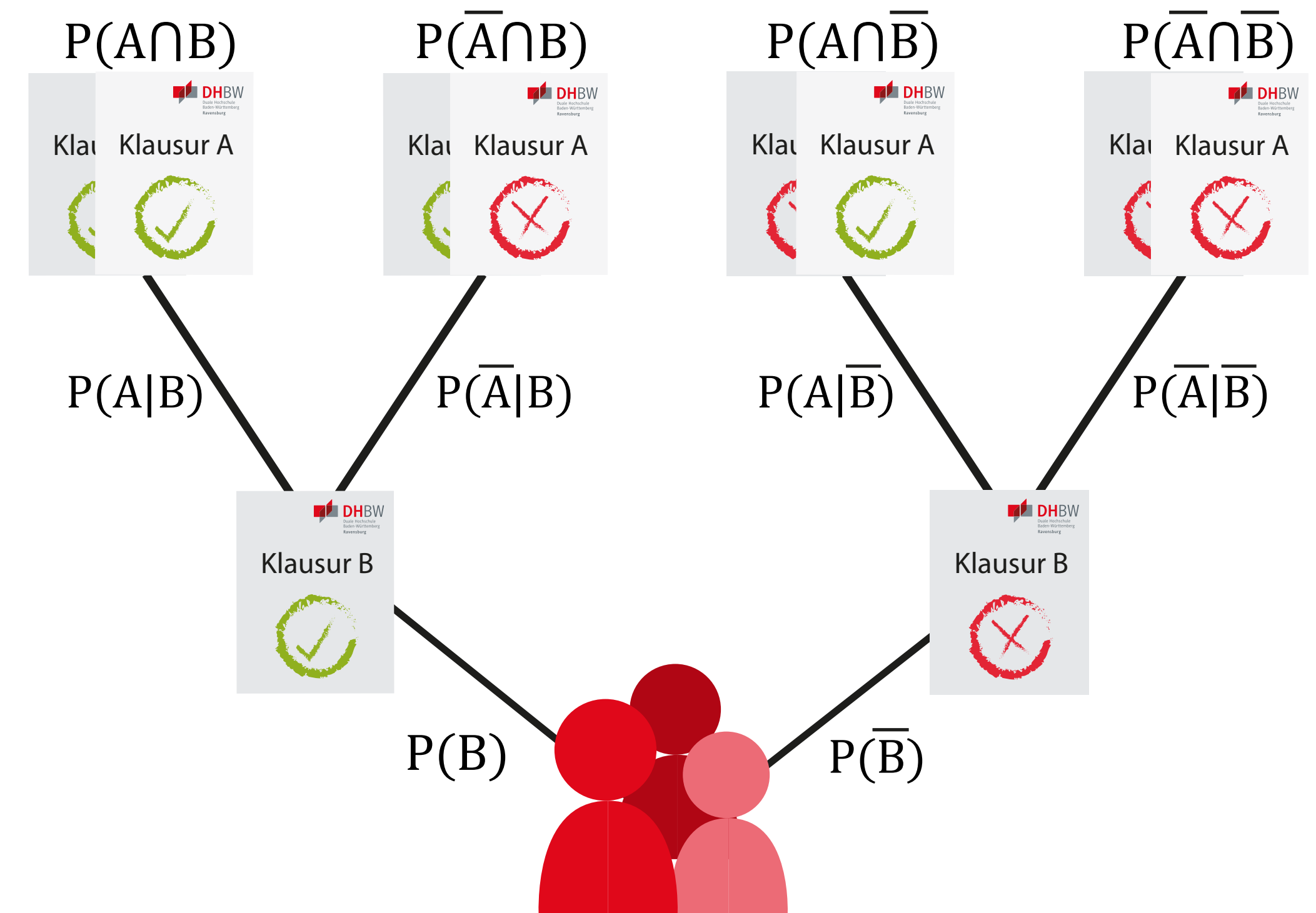
Bedingte Wahrscheinlichkeit

$$\begin{aligned} P(A \cap B) &= P(B) \cdot P(A | B) \\ &= 0.90 \cdot 0.90 = 81\% \end{aligned}$$

$$\begin{aligned} P(\bar{A} \cap B) &= P(B) \cdot P(\bar{A} | B) \\ &= 0.90 \cdot 0.10 = 9\% \end{aligned}$$

$$\begin{aligned} P(A \cap \bar{B}) &= P(\bar{B}) \cdot P(A | \bar{B}) \\ &= 0.10 \cdot 0.50 = 5\% \end{aligned}$$

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\bar{B}) \cdot P(\bar{A} | \bar{B}) \\ &= 0.10 \cdot 0.50 = 5\% \end{aligned}$$

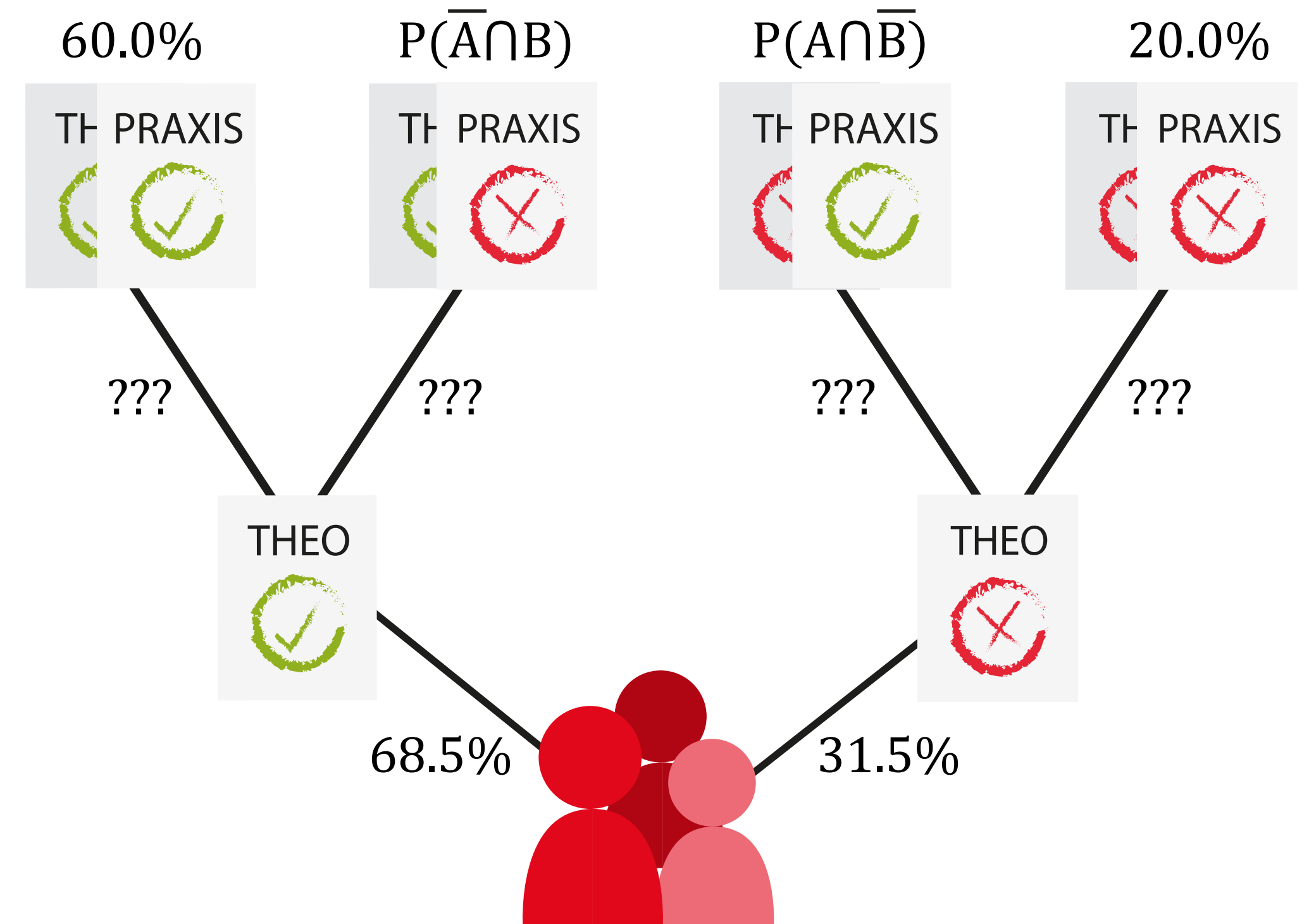


Bedingte Wahrscheinlichkeit

Laut Kraftfahrtbundesamt schaffen 68.5% die theoretische Fahrprüfung in BaWü auf den ersten Versuch.

Angenommen 60% schaffen beide Prüfungsteile (Theorie und Praxis) in einem Versuch und 20% brauchen für beide Prüfungsteile mehrere Versuche ...

...wie hoch sind dann die bedingten Wahrscheinlichkeiten?



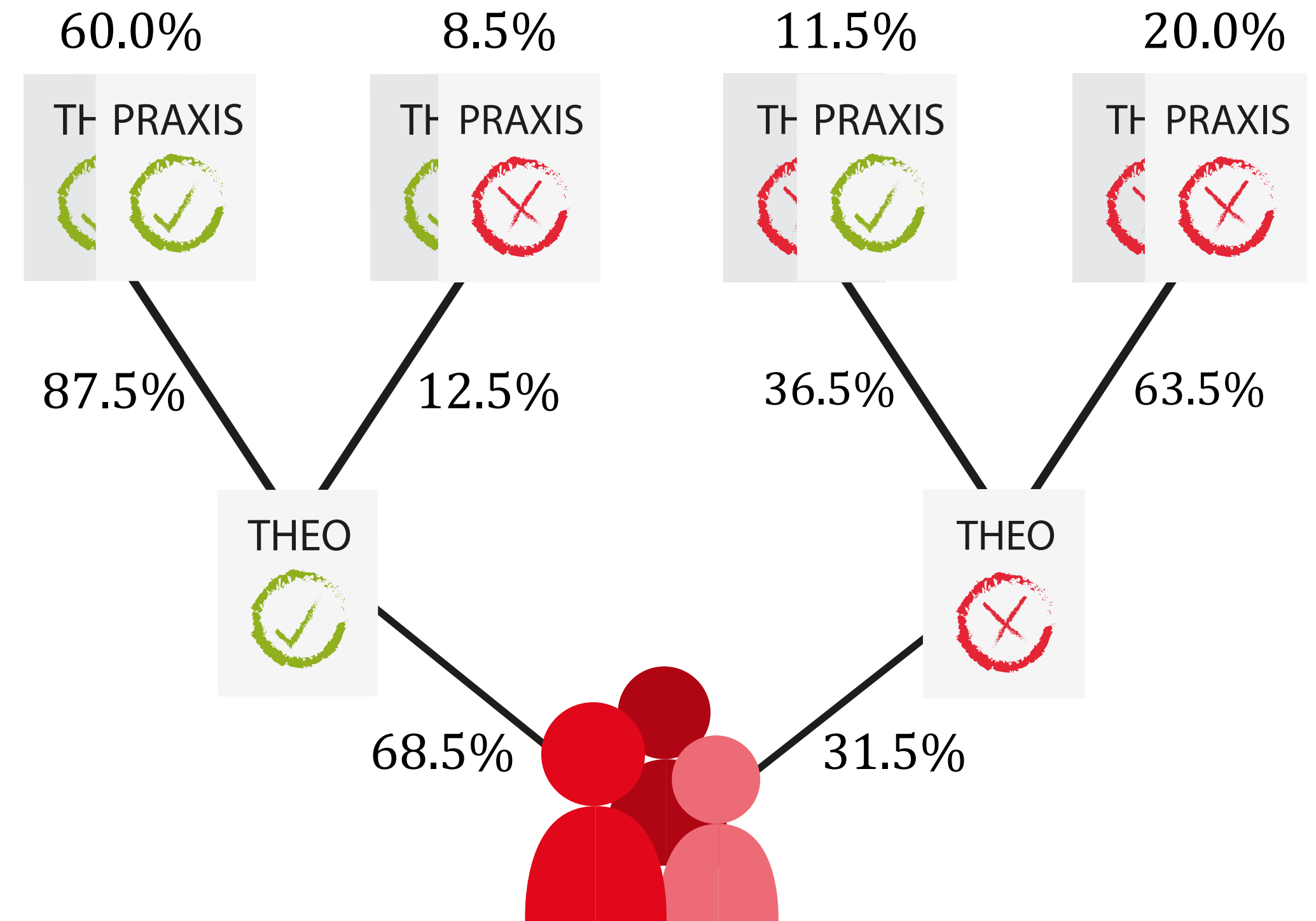
Bedingte Wahrscheinlichkeit

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.600}{0.685} = 87.5\%$$

$$P(\bar{A} | B) = 1 - P(A | B) = 12.5\%$$

$$P(\bar{A} | \bar{B}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{B})} = \frac{0.200}{0.315} = 63.5\%$$

$$P(A | \bar{B}) = 1 - P(\bar{A} | \bar{B}) = 36.5\%$$

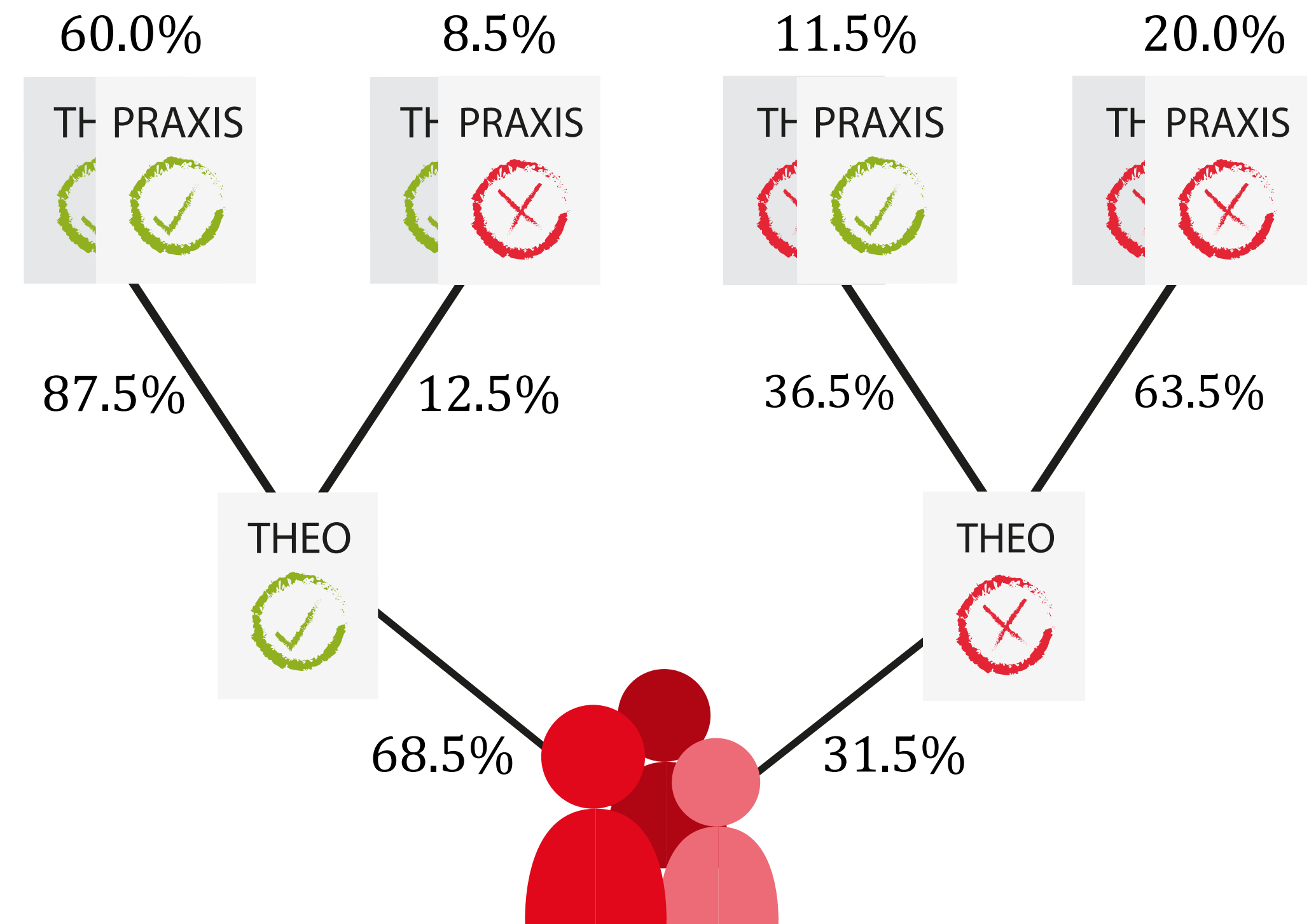


Bedingte Wahrscheinlichkeit

Die dritte Gleichung ist der **Satz der totalen Wahrscheinlichkeit**. Er verknüpft die bedingten Wahrscheinlichkeiten von Ereigniskombinationen.

$$P(A) = P(A|B) P(B) + P(A|\bar{B}) P(\bar{B})$$

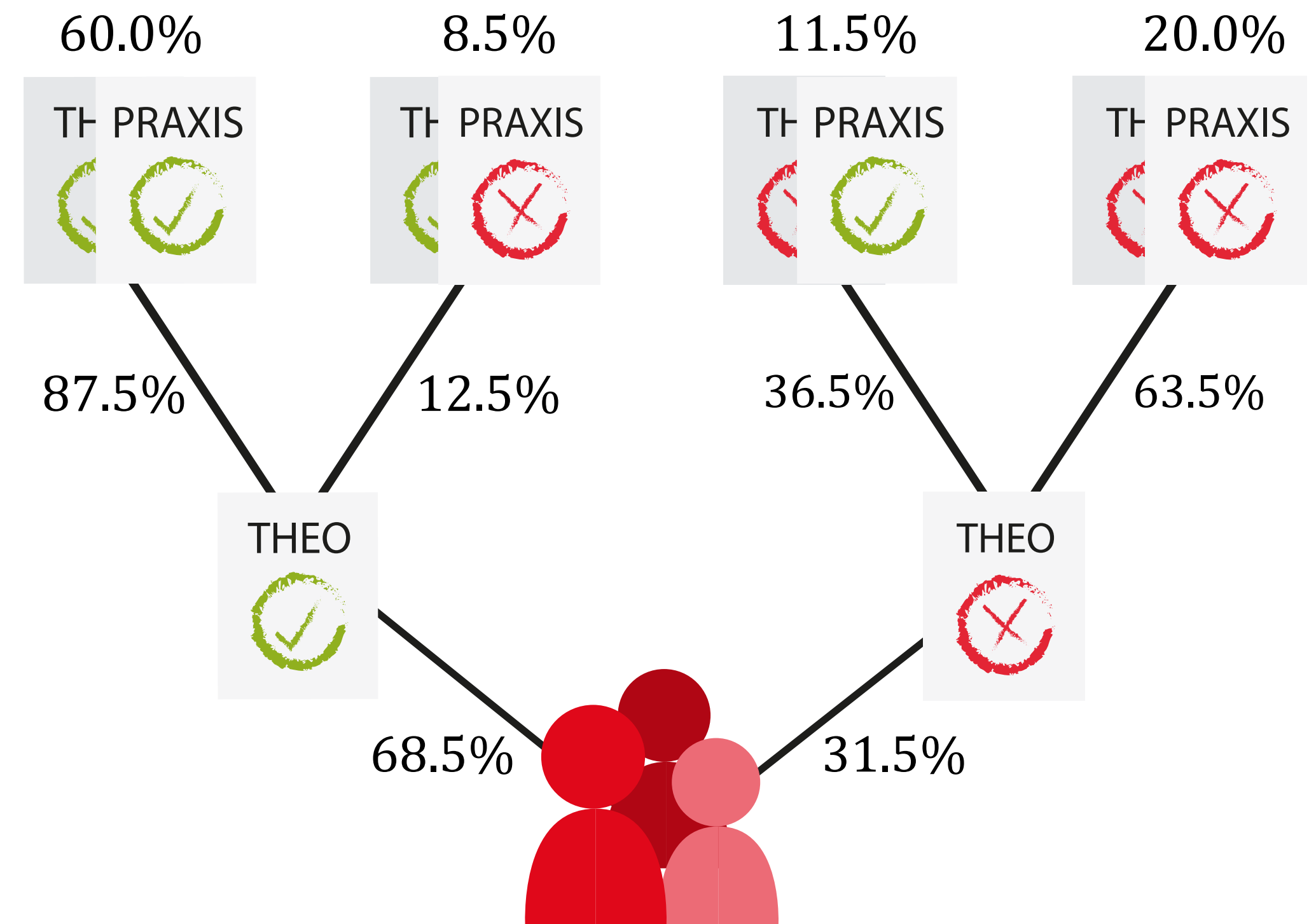
Er erlaubt uns die Berechnung der s. g. a priori Wahrscheinlichkeit des Ereignisses A.



Bedingte Wahrscheinlichkeit

Hier liefert er uns die a priori Wahrscheinlichkeit, die praktische Prüfung direkt zu bestehen, d. h., bevor wir wissen, ob wir die Theorieprüfung aufs erste Mal schaffen.

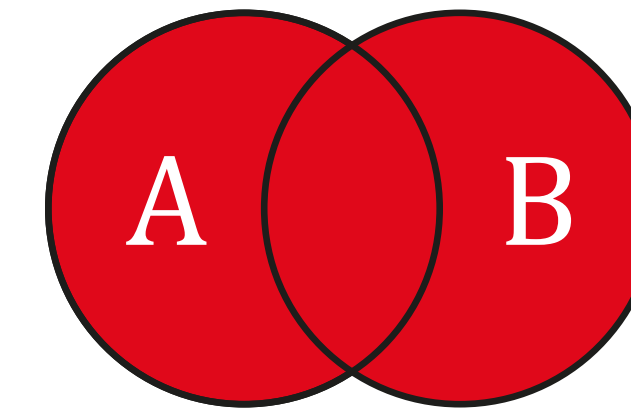
$$P(A) = 0.875 \cdot 0.685 + 0.365 \cdot 0.315 = 71.5\%$$



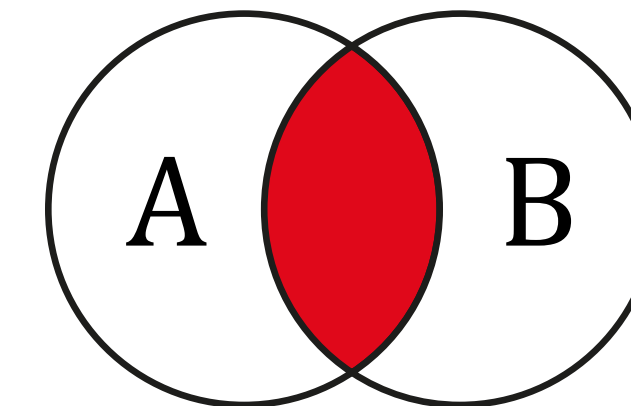
Bedingte Wahrscheinlichkeit

Die letzte Gleichung ist der **Satz von Bayes**. Er erlaubt es uns Bedingung und Folge zu „vertauschen“.

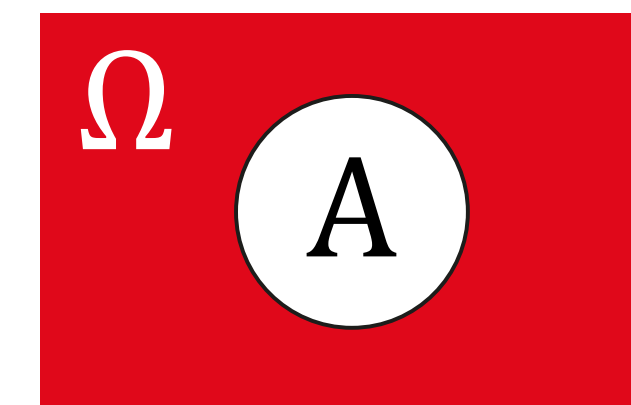
$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$



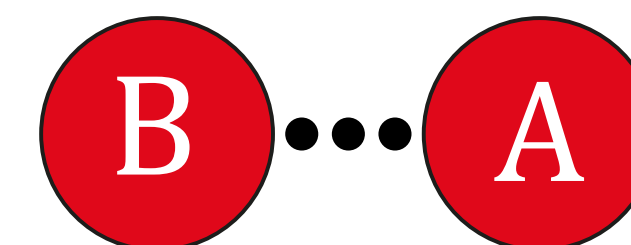
Vereinigung $A \cup B$
A oder B



Schnittmenge $A \cap B$
A und B



Komplement \bar{A}
Alles außer A



Bedingte Wkt. $A | B$
A nach B

Bedingte Wahrscheinlichkeit

Wir definieren die stochastisch abhängigen Ereignisse:

Ereignis A Eine Coronainfektion
 Ereignis B Ein positiver Antigentest

Wir betrachten einen Antigentest mit einer Sensitivität von 92.5% und einer Spezifität von 99.0%

$$P(B | A) = 92.5\%$$

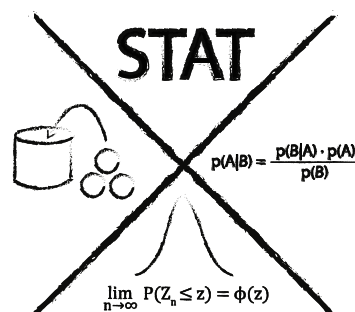
$$P(\bar{B} | \bar{A}) = 99.0\%$$

Wir gehen darüber hinaus von einer Inzidenz von 500 aus, d. h. von 100000 Personen sind 500 positiv:

$$P(A) = 0.005 = 0.5\%$$

Wie hoch ist die Wahrscheinlichkeit, dass ein Test an einer beliebigen Person positiv ist?

Wie hoch ist die Wahrscheinlichkeit, dass ich positiv bin, wenn mein Test anschlägt?



Bedingte Wahrscheinlichkeit

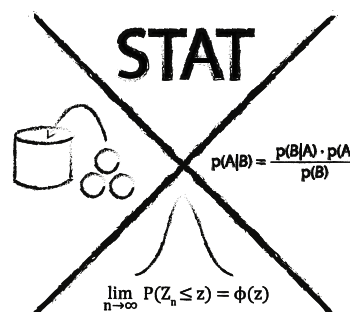
Wie hoch ist die Wahrscheinlichkeit, dass ein Test an einer beliebigen Person positiv ist?

$$P(B|\bar{A}) = 1 - P(\bar{B}|\bar{A}) = 1 - 0.990 = 0.010$$

$$\begin{aligned} P(B) &= P(B|A) P(A) + P(B|\bar{A}) P(\bar{A}) \\ &= 0.925 \cdot 0.005 + 0.010 \cdot 0.995 \\ &= 0.014575 = 1.4575\% \end{aligned}$$

Wie hoch ist die Wahrscheinlichkeit, dass ich positiv bin, wenn mein Test anschlägt?

$$\begin{aligned} P(A | B) &= \frac{P(B | A) \cdot P(A)}{P(B)} \\ &= \frac{0.925 \cdot 0.005}{0.014575} \\ &= 31.73\% \end{aligned}$$



Bedingte Wahrscheinlichkeit

Die Wahrscheinlichkeit, dass es an einem Tag im Blaubeurer Ring zu mindestens einem Unfall kommt, beträgt 15%.

An nebeligen Tagen sind es sogar 20% und das ist in Ulm leider auch an jedem zweiten Tag der Fall.

a) Definiere geeignete Ereignisse und visualisiere die gegebene Information als Wahrscheinlichkeitsbaum.

b) Wie hoch ist die Wahrscheinlichkeit, dass es an einem Tag ohne Nebel einen Unfall im Blaubeurer Ring gibt?

Die Ravensburger Polizei sucht einen gefährlichen Verbrecher. Bekannt ist nur, dass er einer von 50.000 Ravensburgern ist.

Um die Ermittlung zu beschleunigen, wird ein DNA-Test eingesetzt. Dieser hat eine Sensitivität von 100.0% und einer Spezifität von 99.999%

Der Test schlägt bei einer x-beliebigen Person an. Wie hoch ist die Wahrscheinlichkeit, dass diese auch schuldig ist?

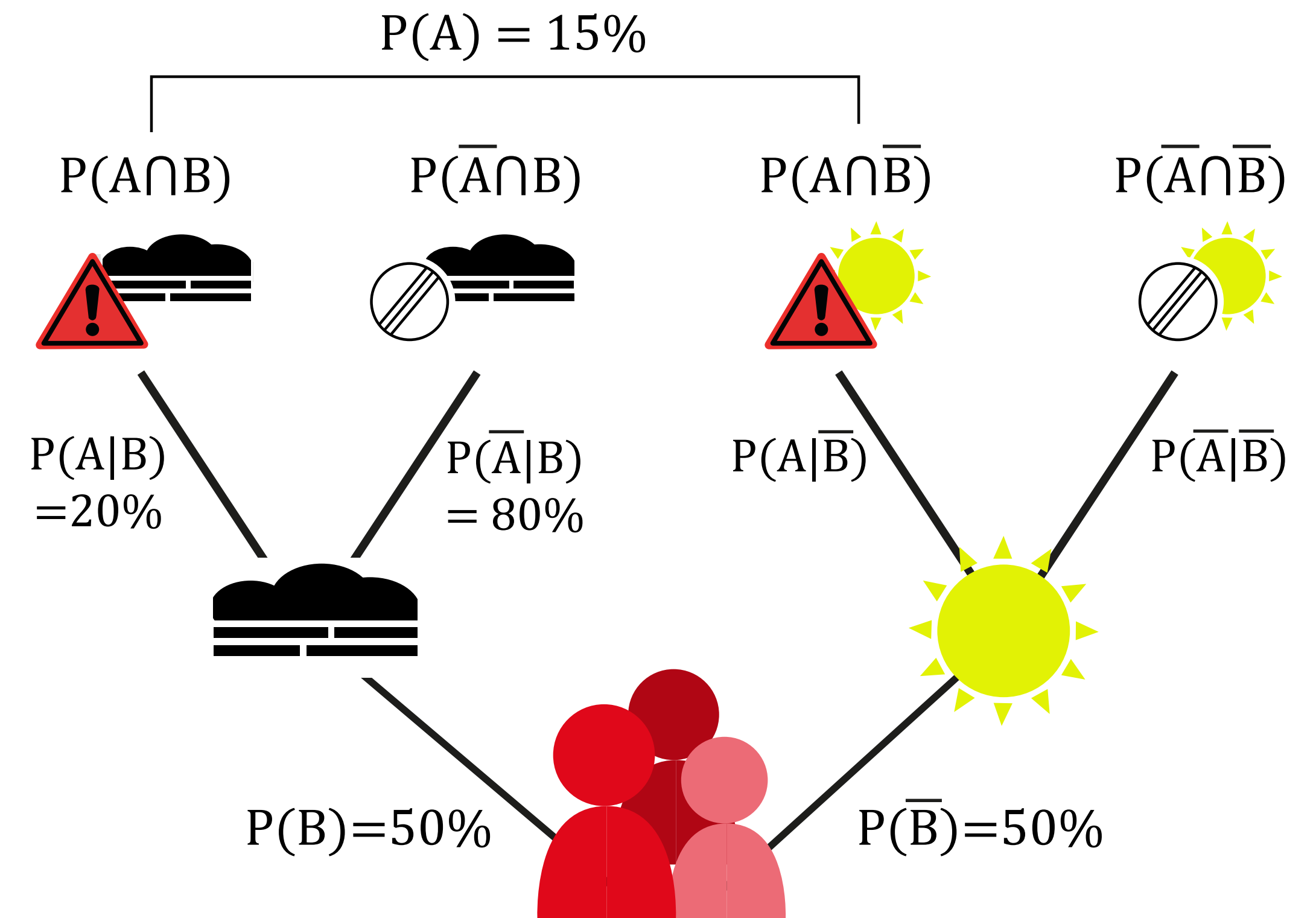
Bedingte Wahrscheinlichkeit

Die Wahrscheinlichkeit, dass es an einem Tag im Blaubeurer Ring zu mindestens einem Unfall kommt beträgt 15%.

An nebeligen Tagen sind es sogar 20% und das ist in Ulm leider auch an jedem zweiten Tag der Fall.

a) Definiere geeignete Ereignisse und visualisiere die gegebene Information als Wahrscheinlichkeitsbaum.

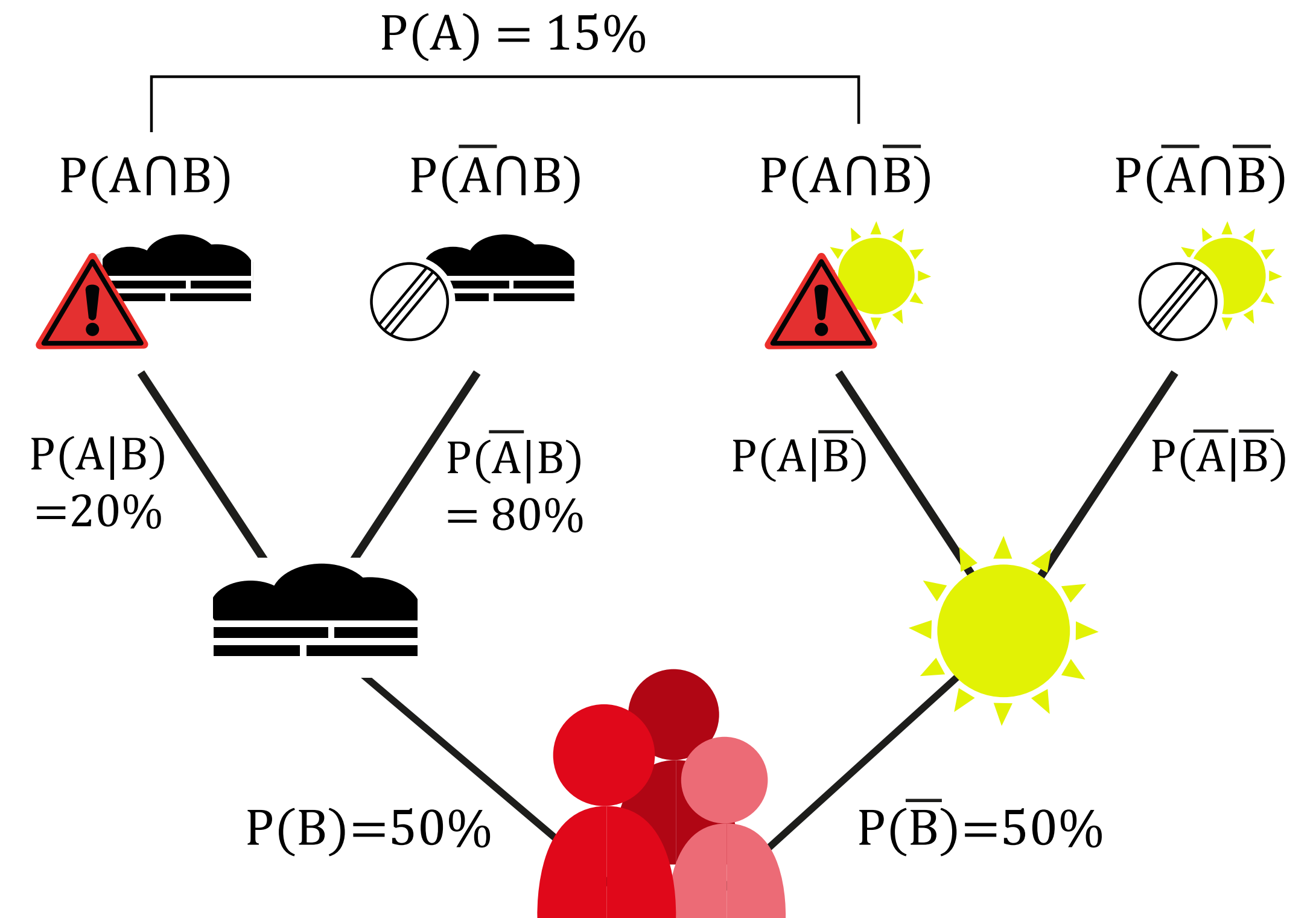
Ereignis A Es gibt einen Unfall im Blaubeurer Ring.
Ereignis B Es ist nebelig in Ulm.



Bedingte Wahrscheinlichkeit

b) Wie hoch ist die Wahrscheinlichkeit, dass es an einem Tag ohne Nebel einen Unfall im Blaubeurer Ring gibt?

$$\begin{aligned}
 P(A) &= P(A|B) P(B) + P(A|\bar{B}) P(\bar{B}) \stackrel{!}{=} 0.150 \\
 \Leftrightarrow 0.200 \cdot 0.500 + x \cdot 0.500 &= 0.150 \\
 \Leftrightarrow 0.100 + 0.500x &= 0.150 \\
 \Leftrightarrow 0.500x &= 0.050 \\
 \Leftrightarrow x &= 0.100 = 10\%
 \end{aligned}$$



Bedingte Wahrscheinlichkeit

Wie hoch ist die Wahrscheinlichkeit, dass ein Test an einer beliebigen Person eine Übereinstimmung feststellt?

$$P(B|\bar{A}) = 1 - P(\bar{B}|\bar{A}) = 1 - 0.99999 = 0.00001$$

$$P(B) = P(B|A) P(A) + P(B|\bar{A}) P(\bar{A})$$

$$= 1 \cdot \frac{1}{50000} + 0.00001 \frac{49999}{50000}$$

$$= 0.00003$$

Wie hoch ist die Wahrscheinlichkeit, dass eine Person, bei welcher der Test anschlägt, schuldig ist?

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

$$= \frac{1 \cdot 0.00002}{0.00003}$$

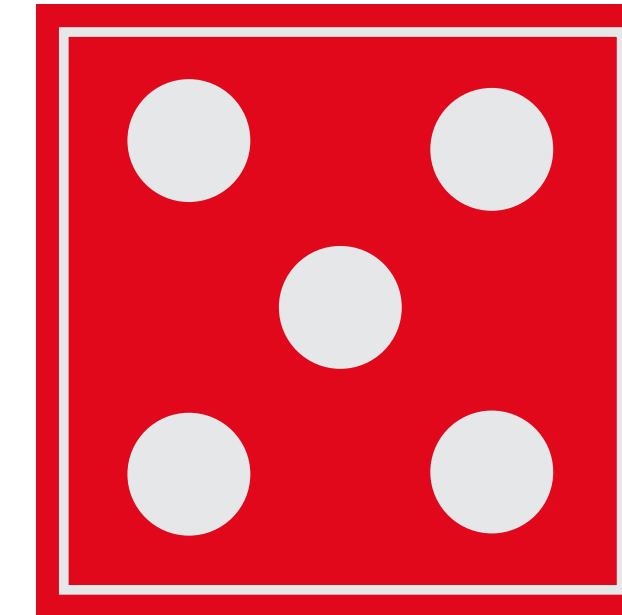
$$= 66.7\%$$

Zufallsvariablen

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

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

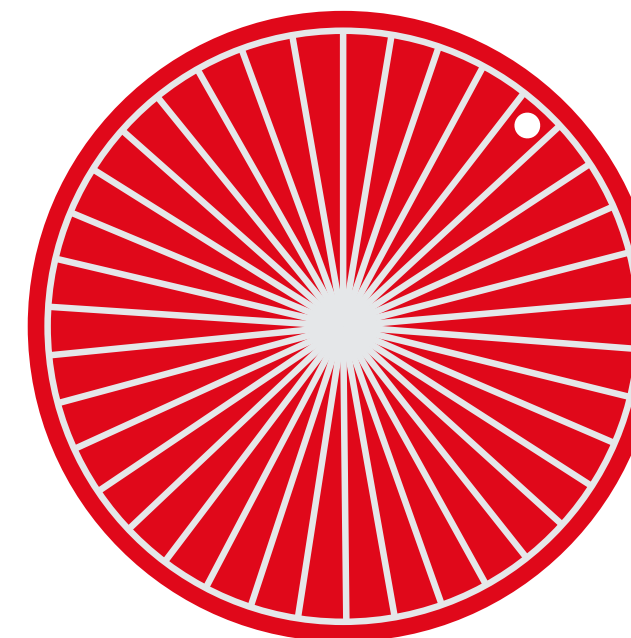
Zufallsvariablen sind mathematische Funktionen, deren Definitionsbereich eine Ereignismenge ist.



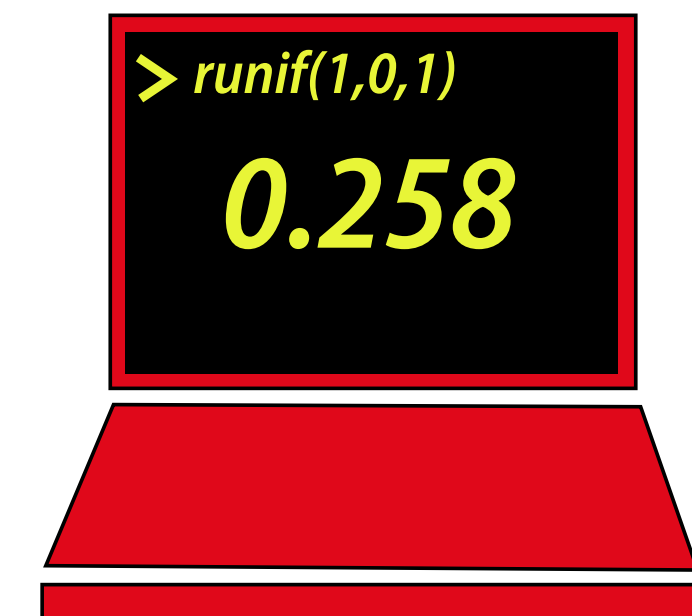
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



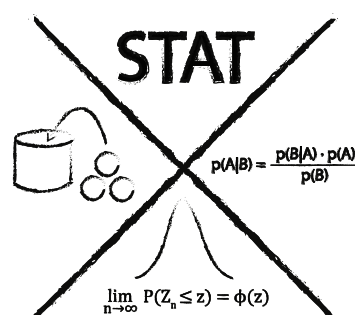
$\Omega = \{\text{Kopf, Zahl}\}$
DISKRET



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



$\Omega = (0,1)$
KONTINUIERLICH



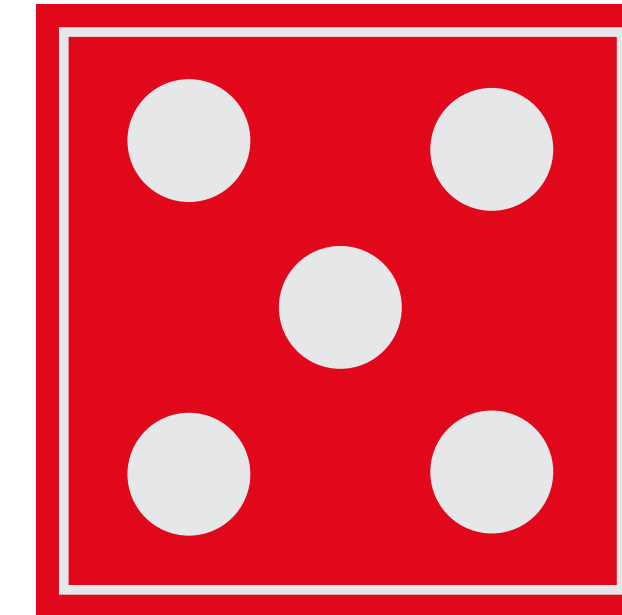
Zufallsvariablen

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 wie wir die Ereignisse in reelle Zahlenwerte umsetzen:

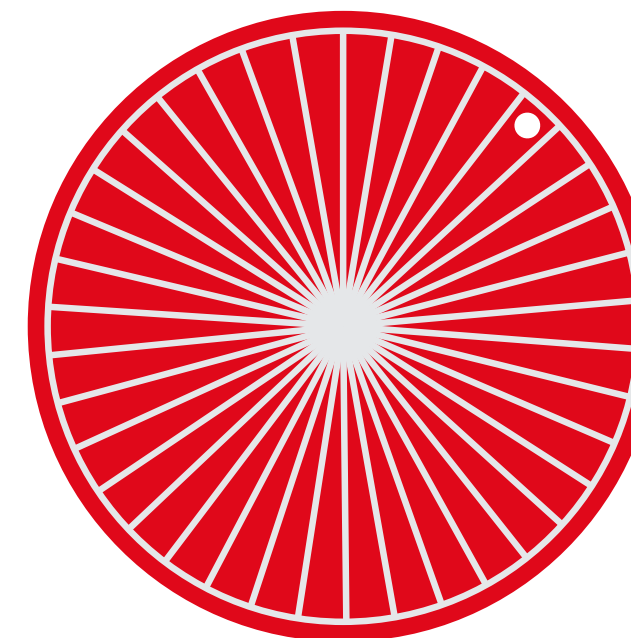
$$X: \{\text{Kopf, Zahl}\} \rightarrow \{0, 1\}$$



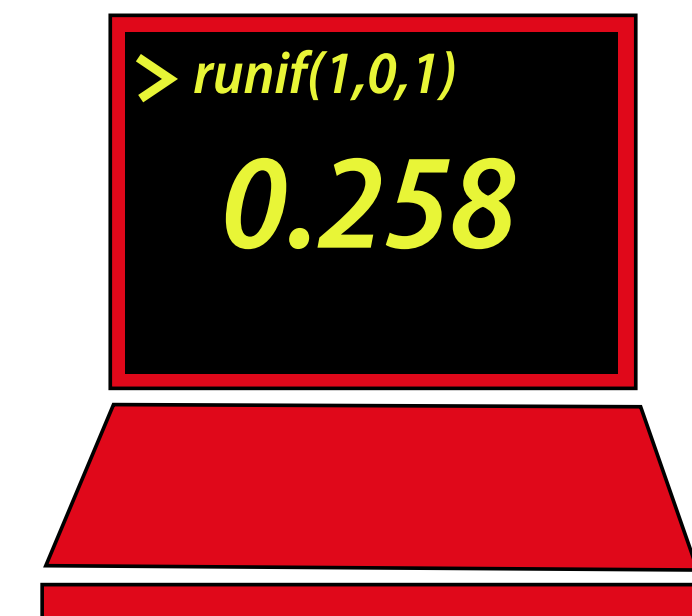
$\Omega = \{1,2,3,4,5,6\}$
DISKRET



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

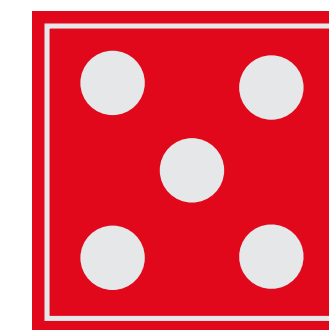
$$p(X): E \rightarrow [0,1]$$

... 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

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

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

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

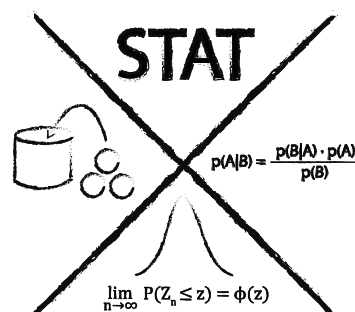
$$P(X=x_1 \vee X=x_2) = \frac{2}{6} \quad x_1, x_2 \in E \text{ mit } x_1 \neq x_2$$

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*



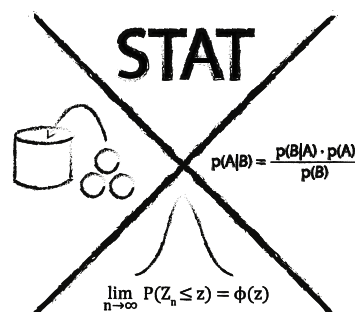
Zufallsvariablen

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

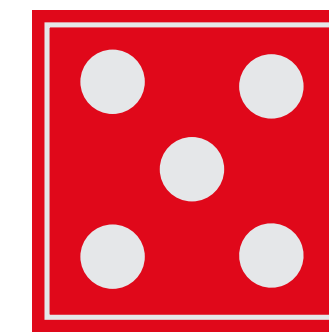


Gleichverteilung

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

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

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

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

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

Gleichverteilung

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

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

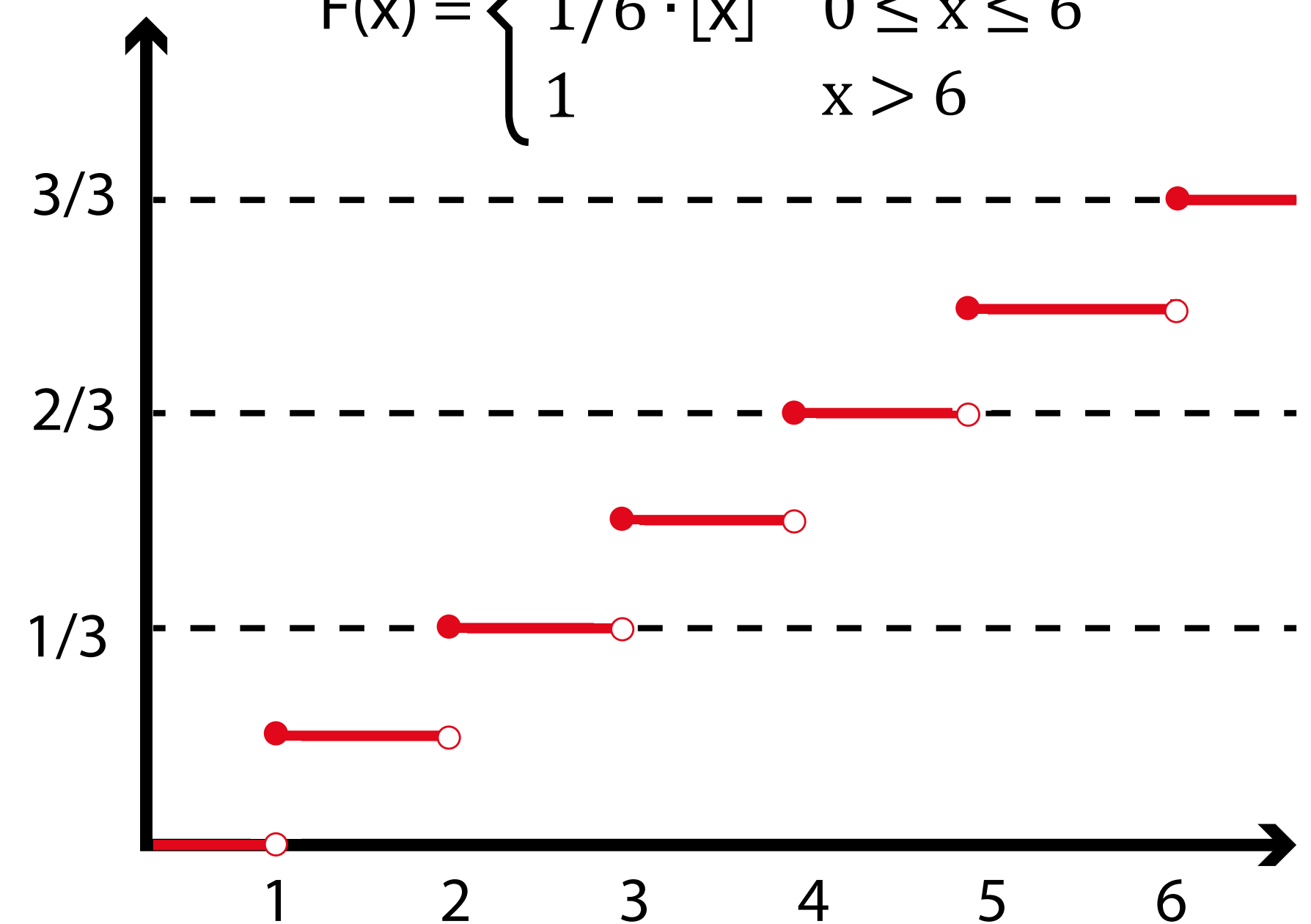
Sie besitzt daher folgende Eigenschaften:

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

$F(x)$ monoton steigend in E

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

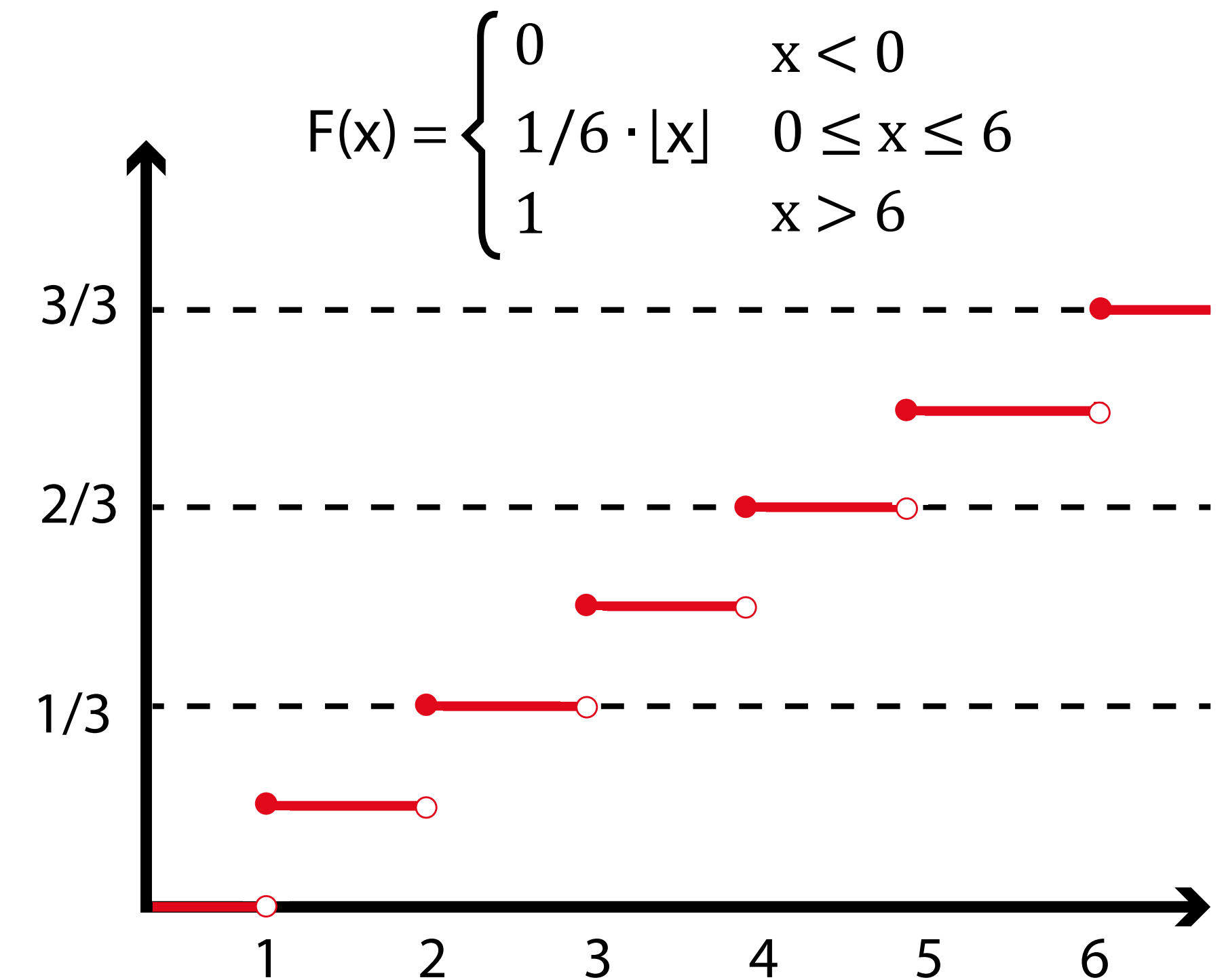
$$F(x) = \begin{cases} 0 & x < 0 \\ 1/6 \cdot [x] & 0 \leq x \leq 6 \\ 1 & x > 6 \end{cases}$$



Gleichverteilung

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 x bzw. $\text{ceil}(x)$.

Gleichverteilung

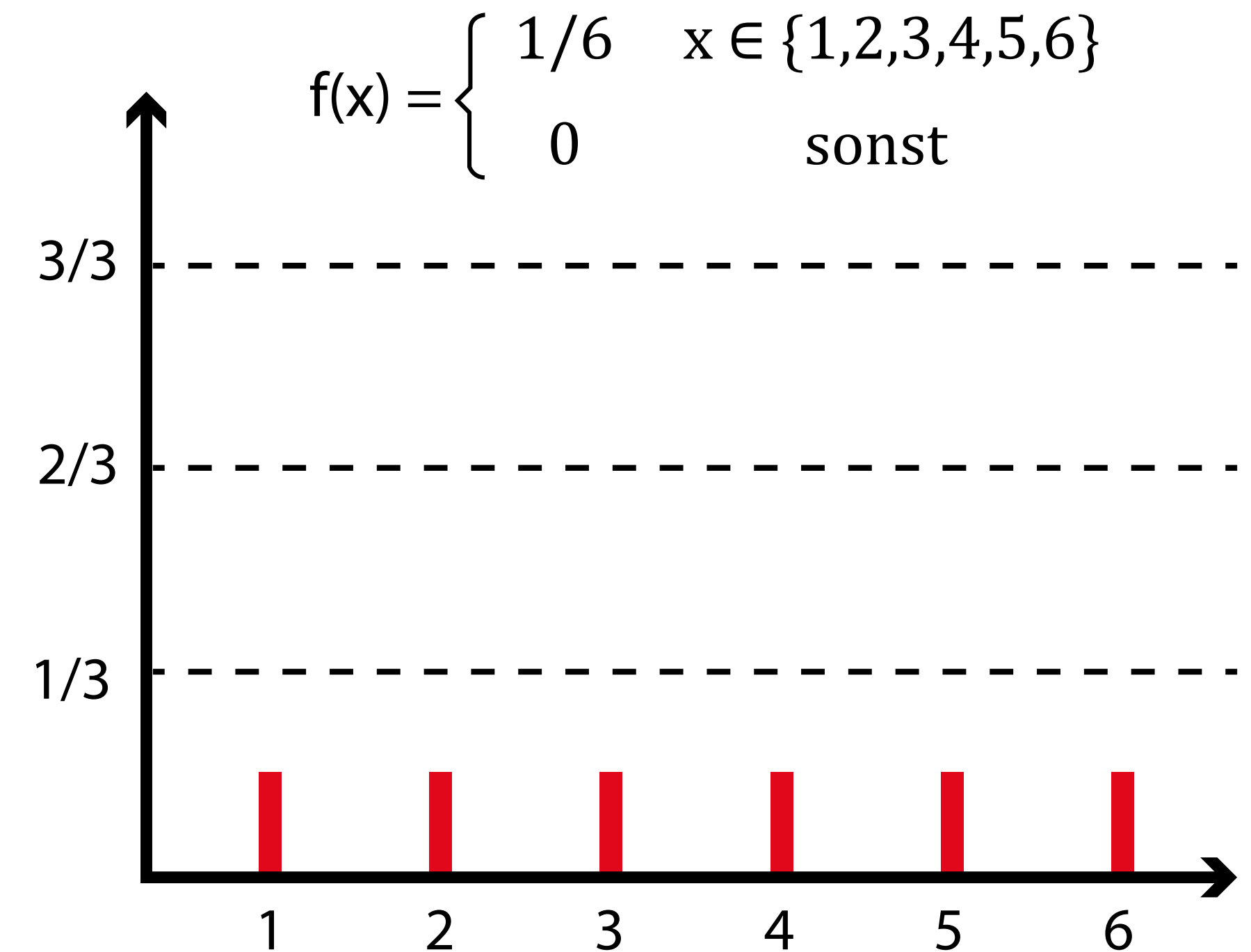
Die **Wahrscheinlichkeitsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable genau x ist.

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

Sie besitzt daher folgende Eigenschaften:

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

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

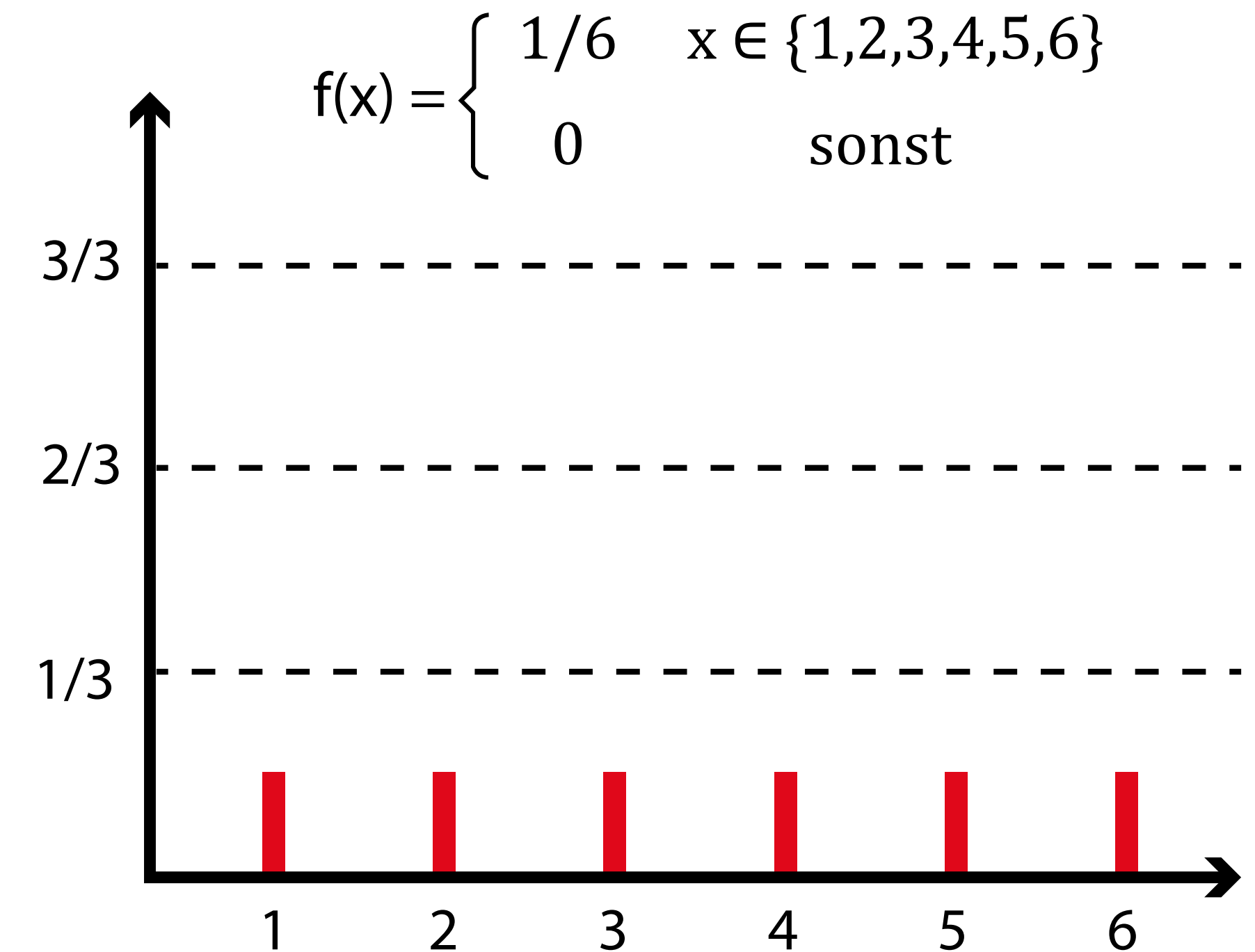


Gleichverteilung

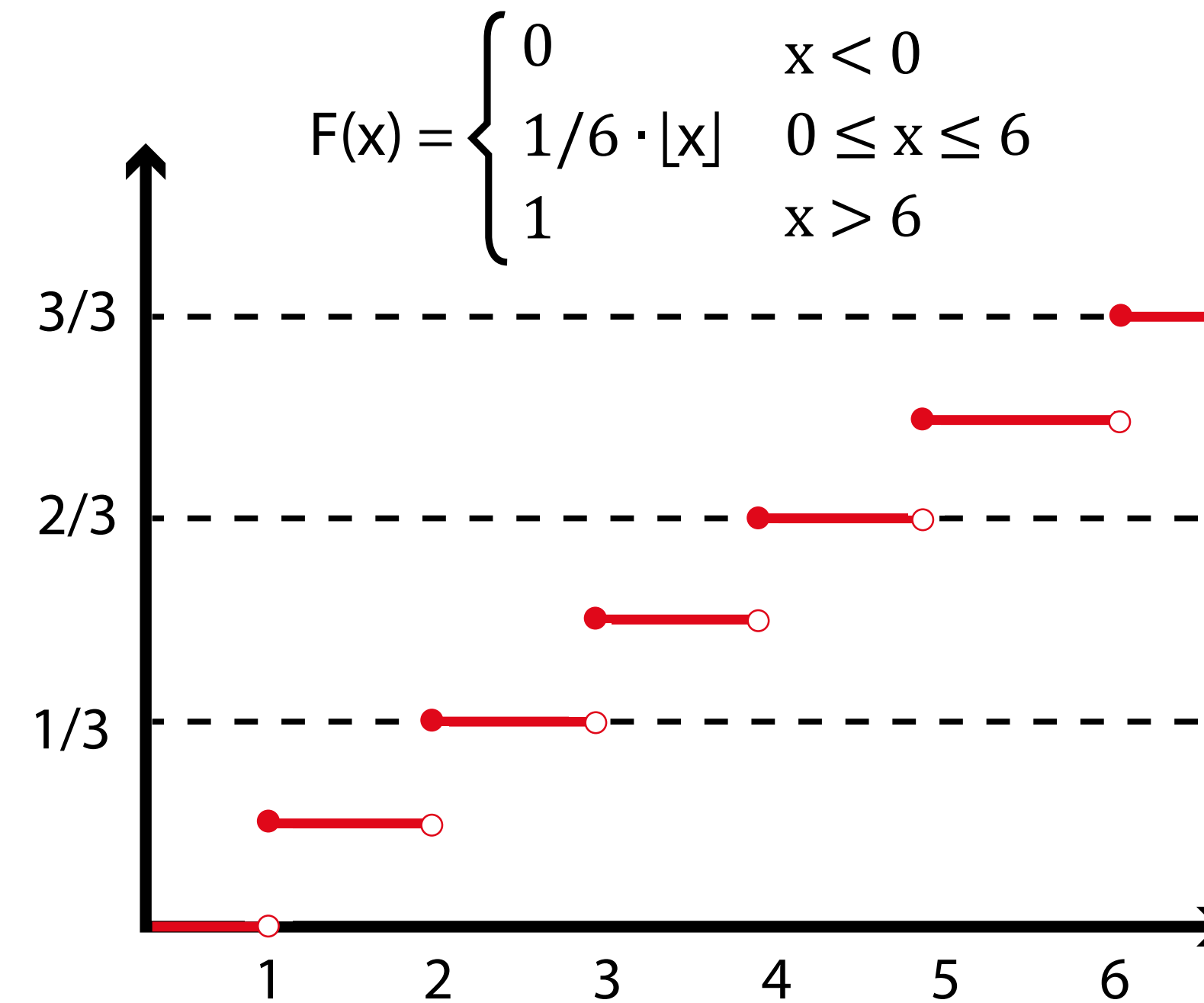
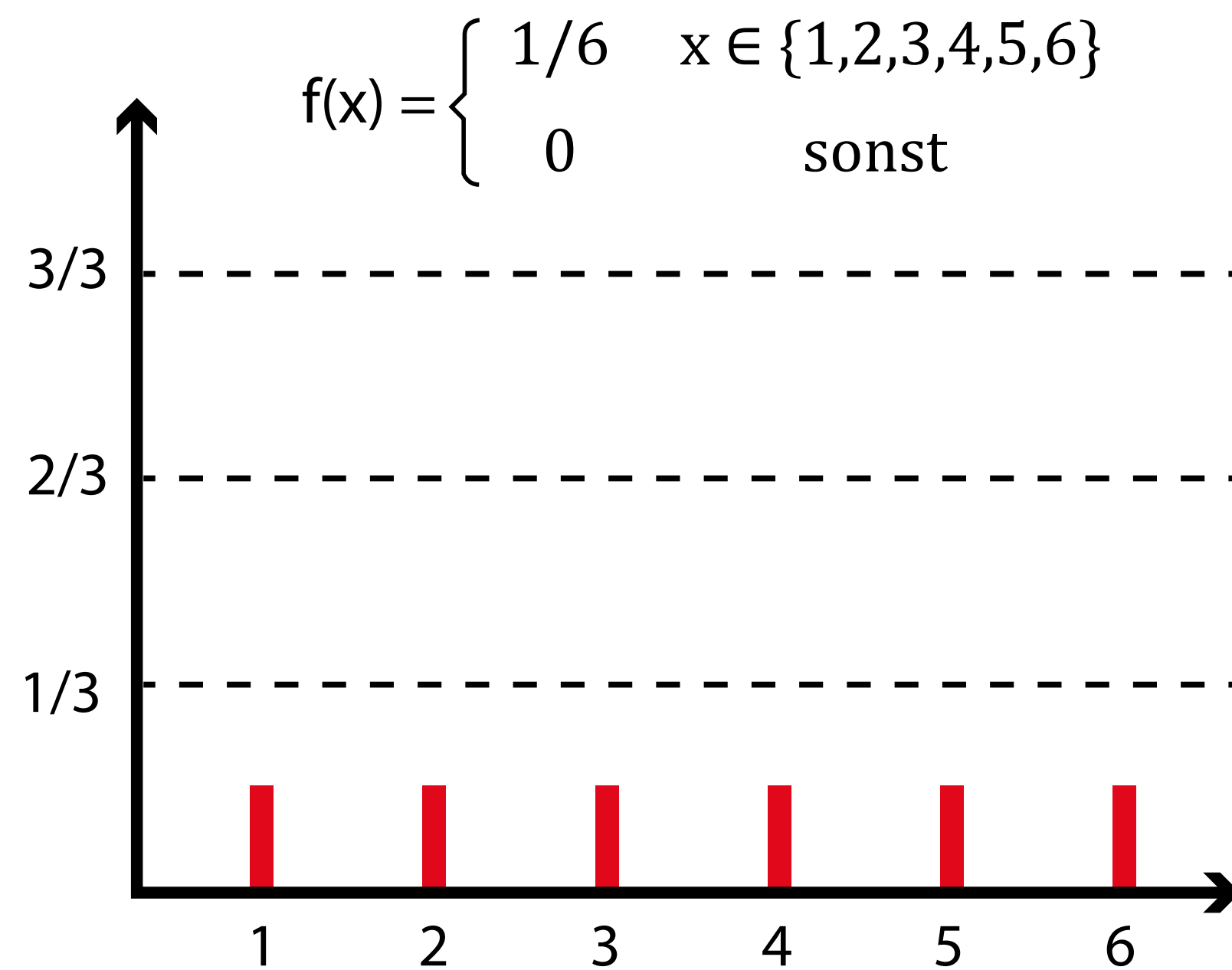
Die **Wahrscheinlichkeitsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable genau x ist.

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.



„Sprünge“



Summation

Gleichverteilung

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.



```
> runif(1,0,1)
```

0.258

$\Omega = (0,1)$

$X = \text{Angezeigter Wert}$

$E = (0,1)$

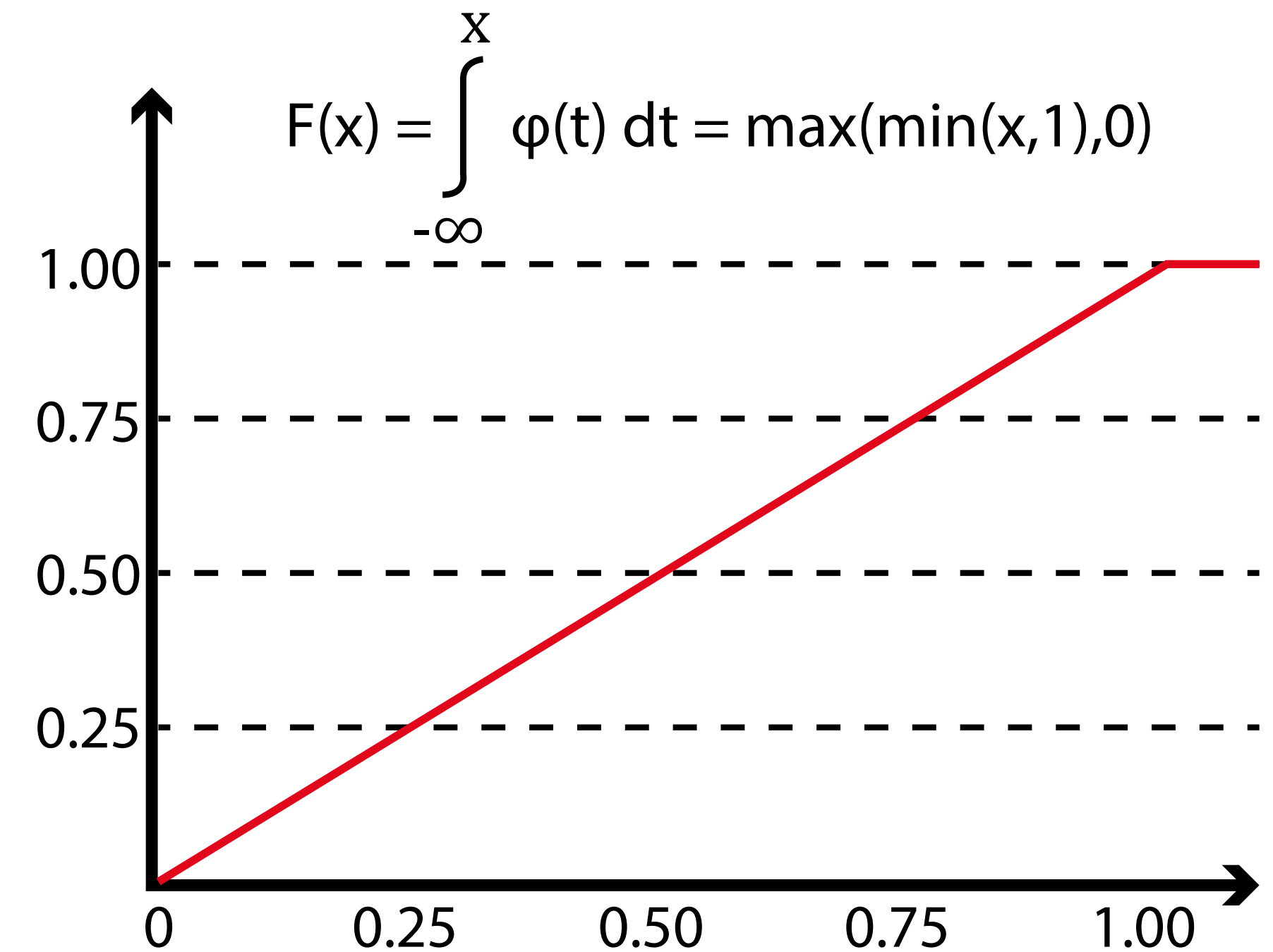
Gleichverteilung

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

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

Sie besitzt daher folgende Eigenschaft:

$$\int_{-\infty}^{\infty} \varphi(t) dt = 1$$



Gleichverteilung

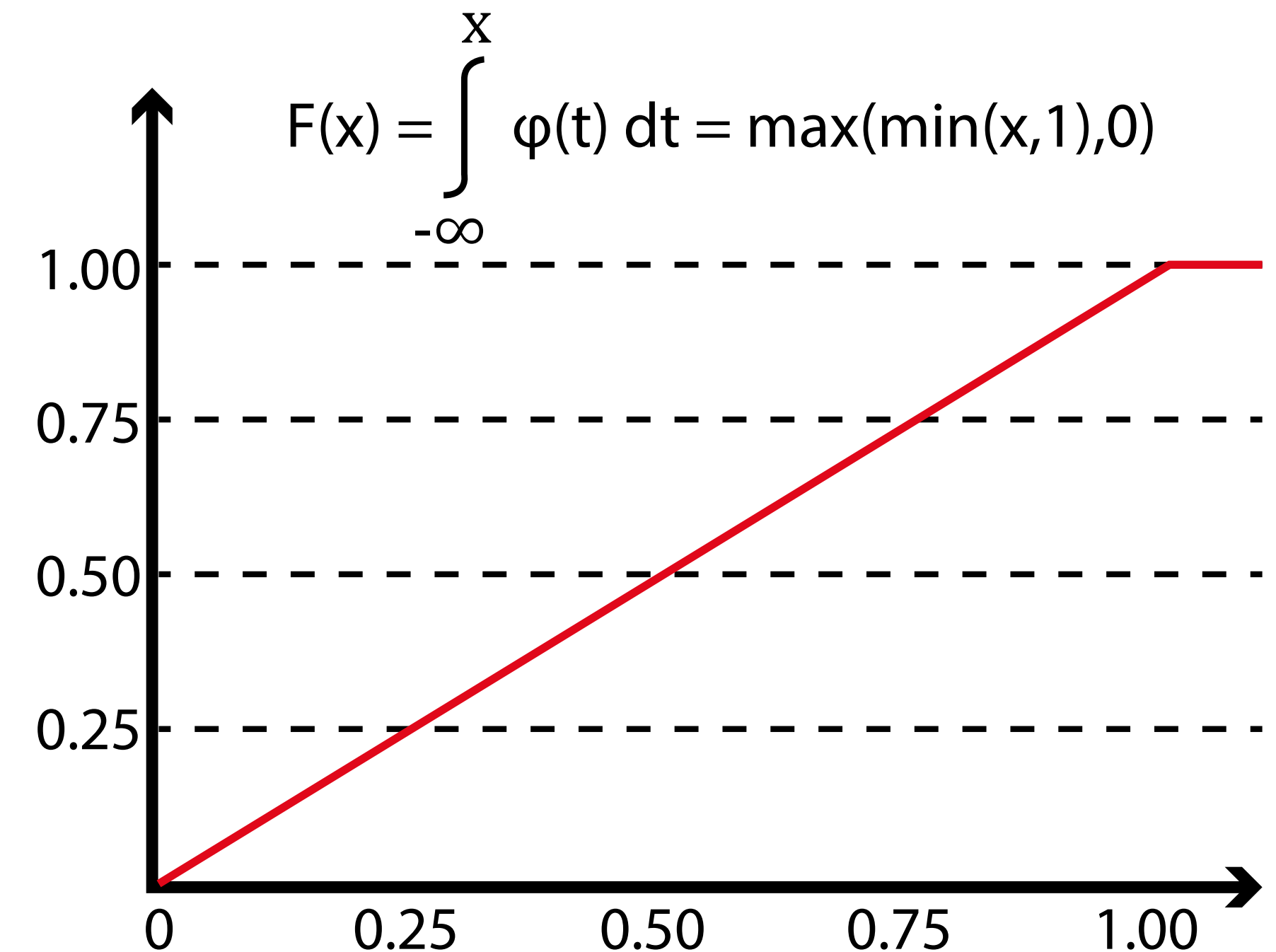
Die **Verteilungsfunktion** gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable gleich oder kleiner als x ist.

Beim 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



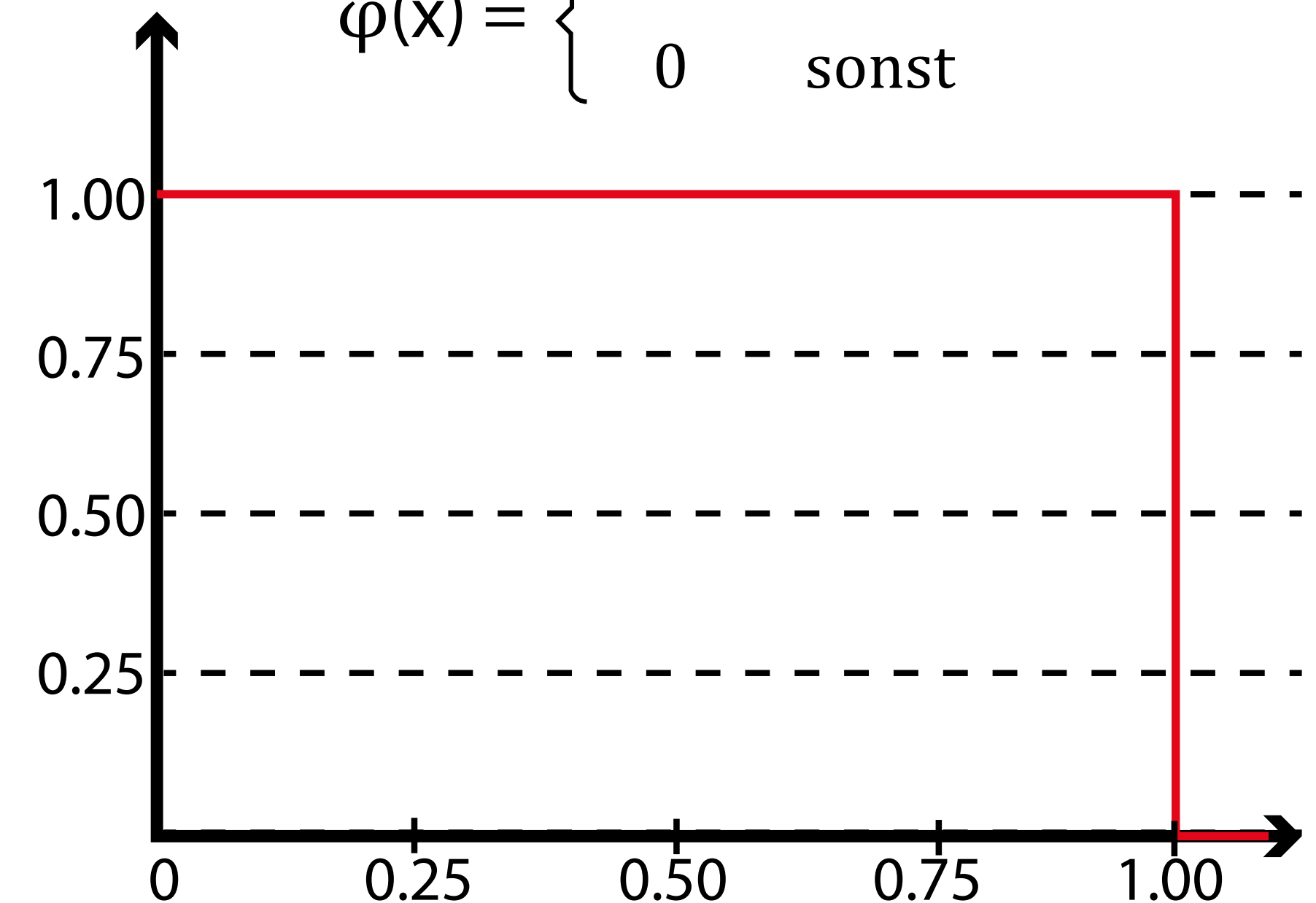
Gleichverteilung

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 ...

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

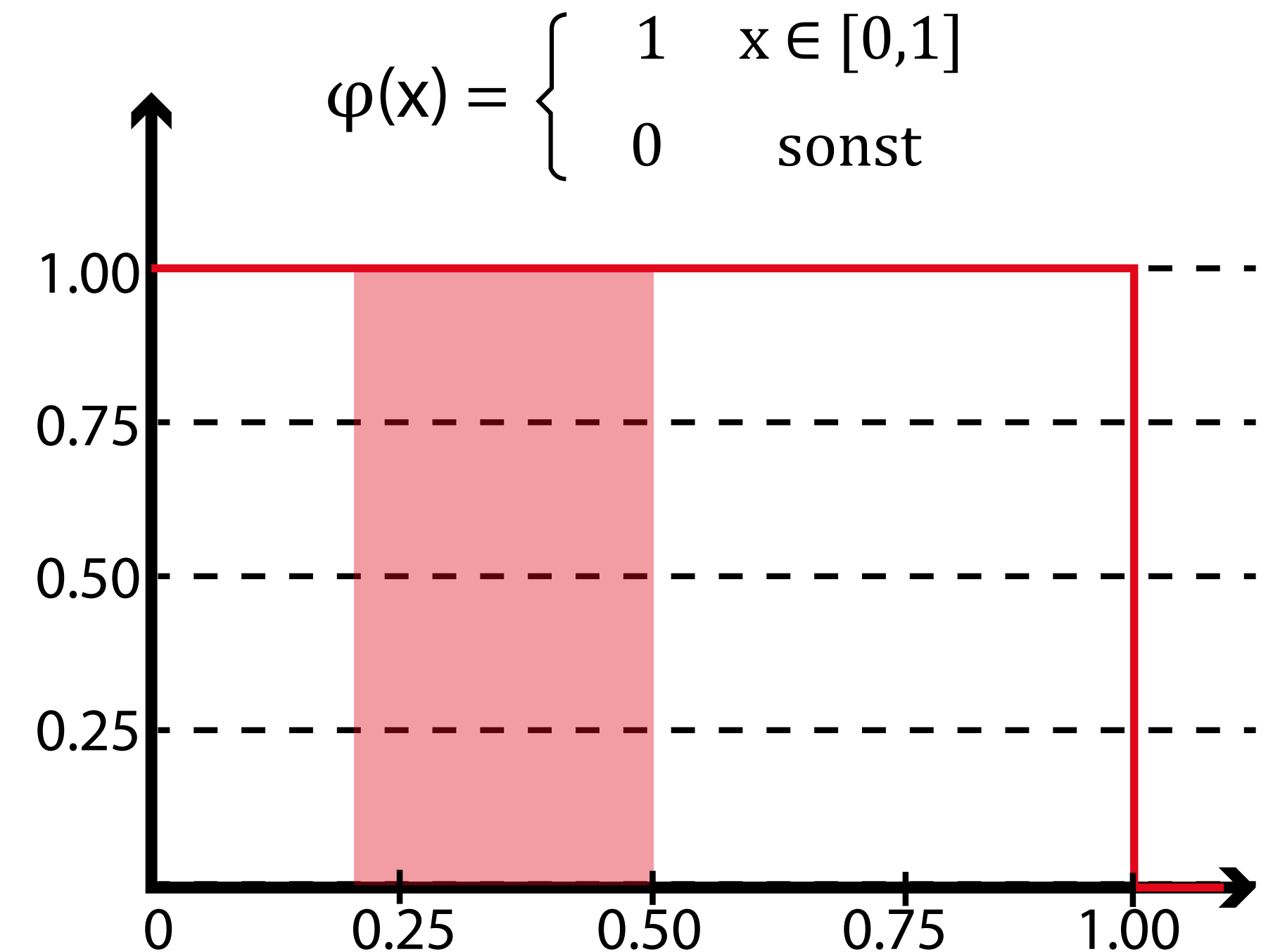
$$\varphi(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{sonst} \end{cases}$$



Gleichverteilung

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

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

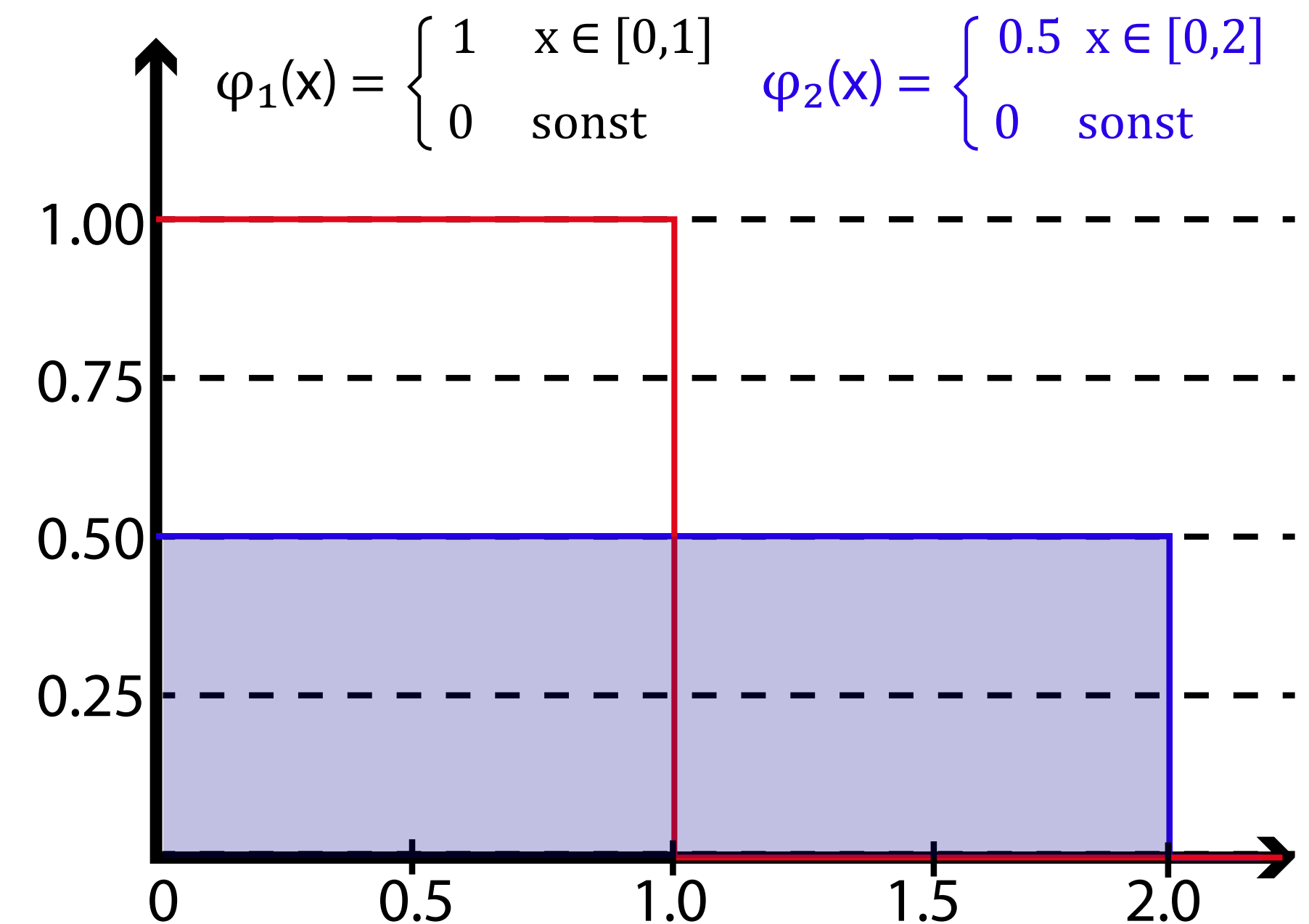


Gleichverteilung

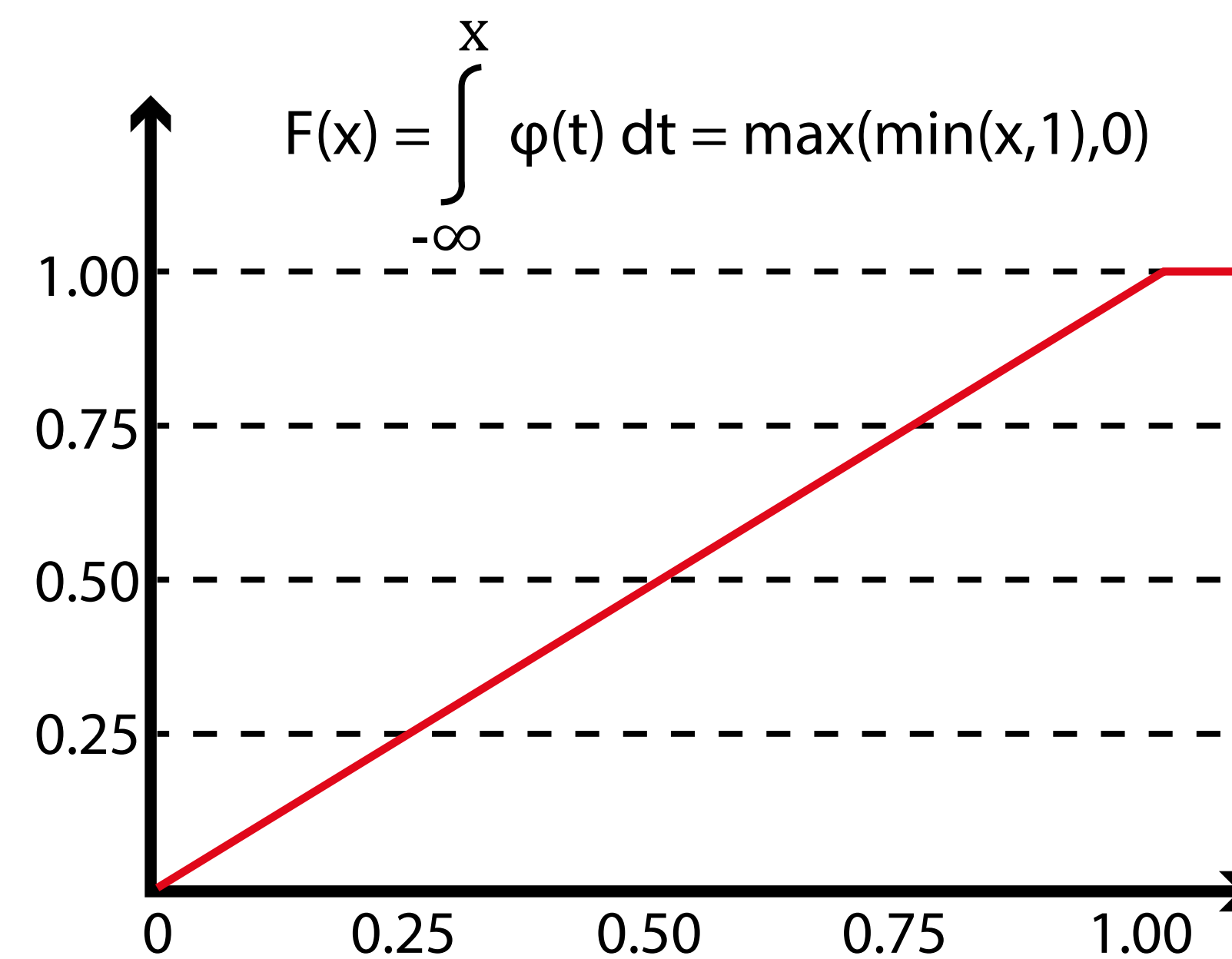
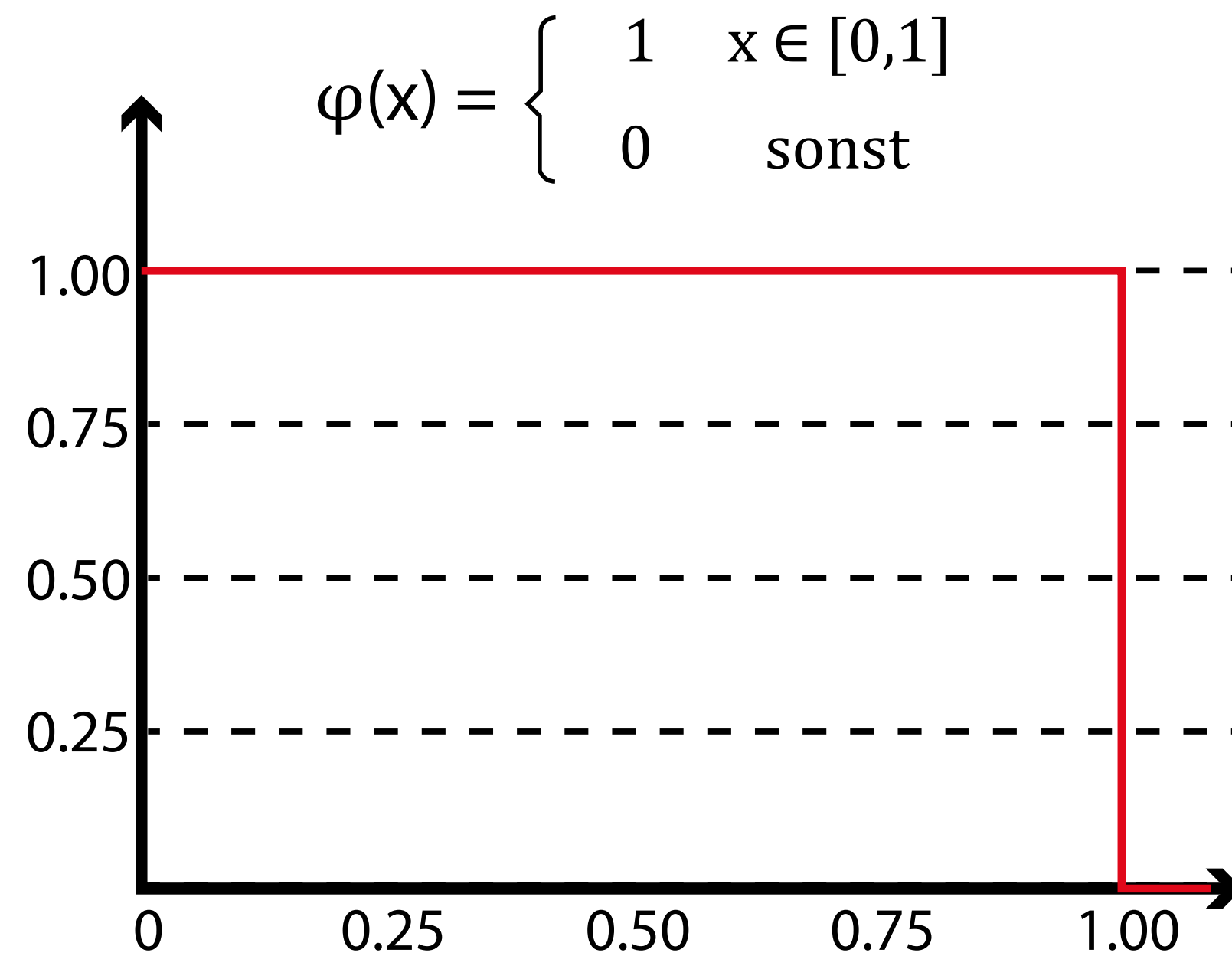
Aus dem zweiten Kolmogorov-Axiom folgt, dass die Gesamtfläche unter der Dichtefunktion genau 1 sein muss.

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

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



Ableitung



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.



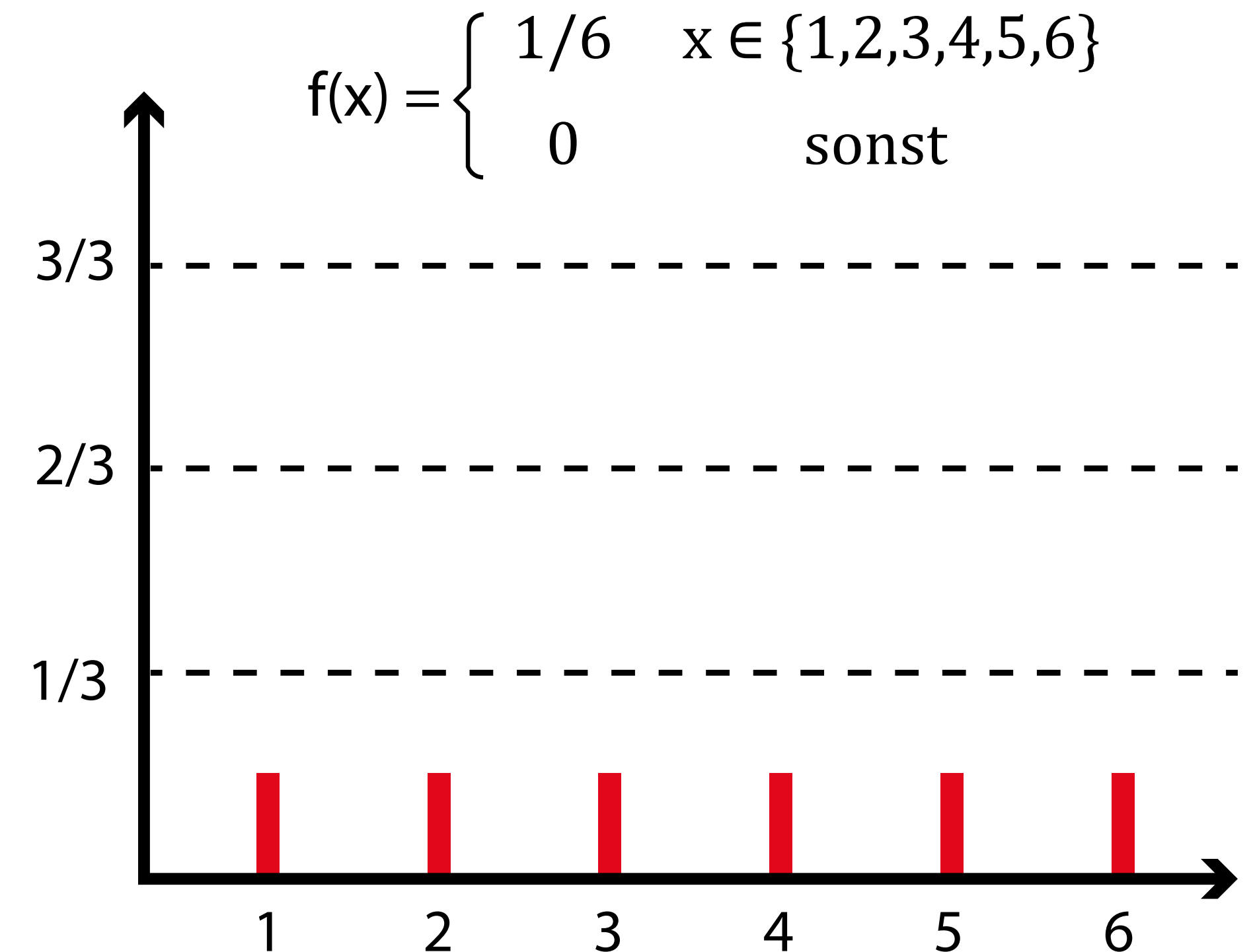
Erwartungswert

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:

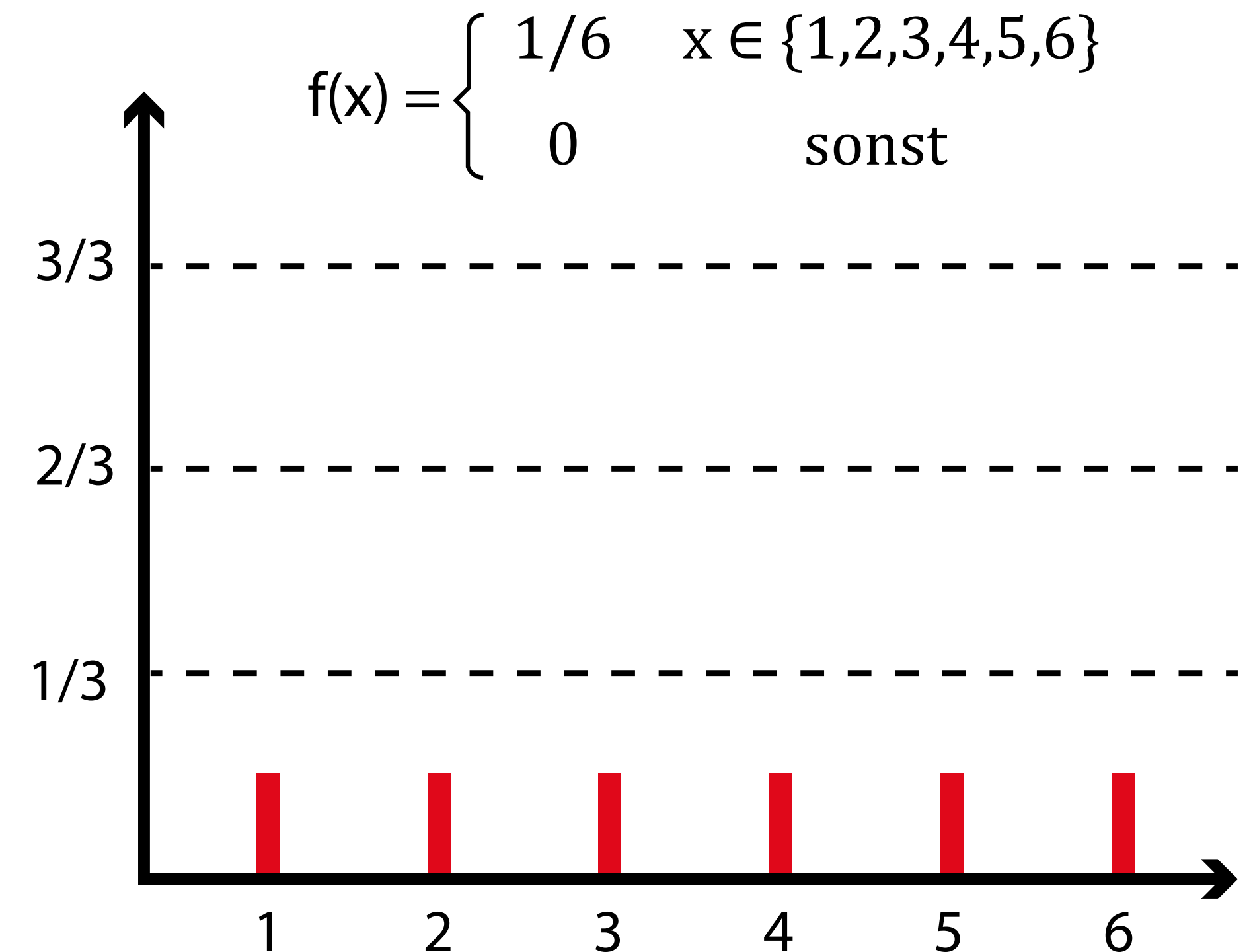
$$\mu = E(X) = \sum_{x \in E} x \cdot f(x)$$



Erwartungswert

Wir berechnen den Erwartungswert über die Summe:

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



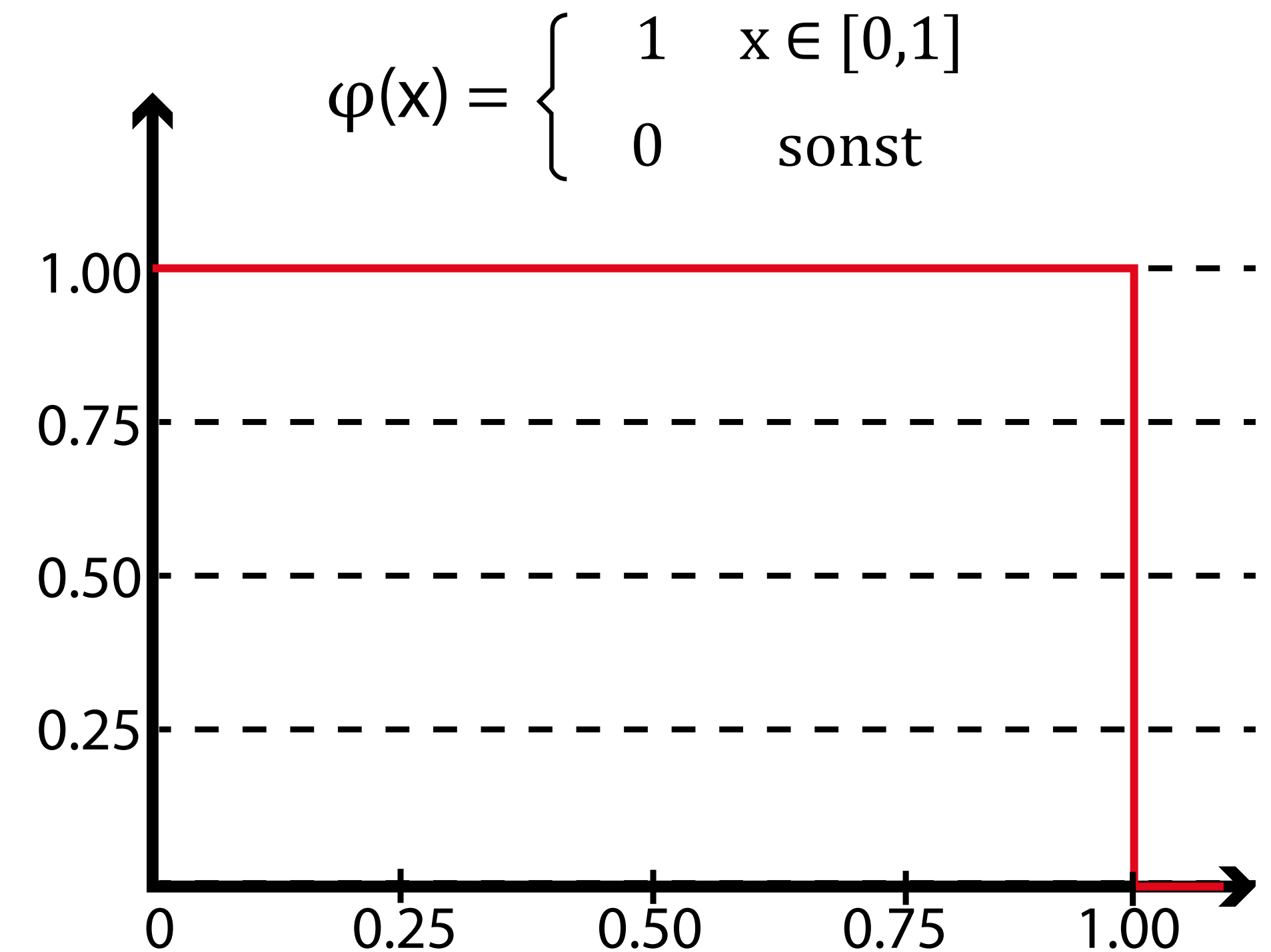
Erwartungswert

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:

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

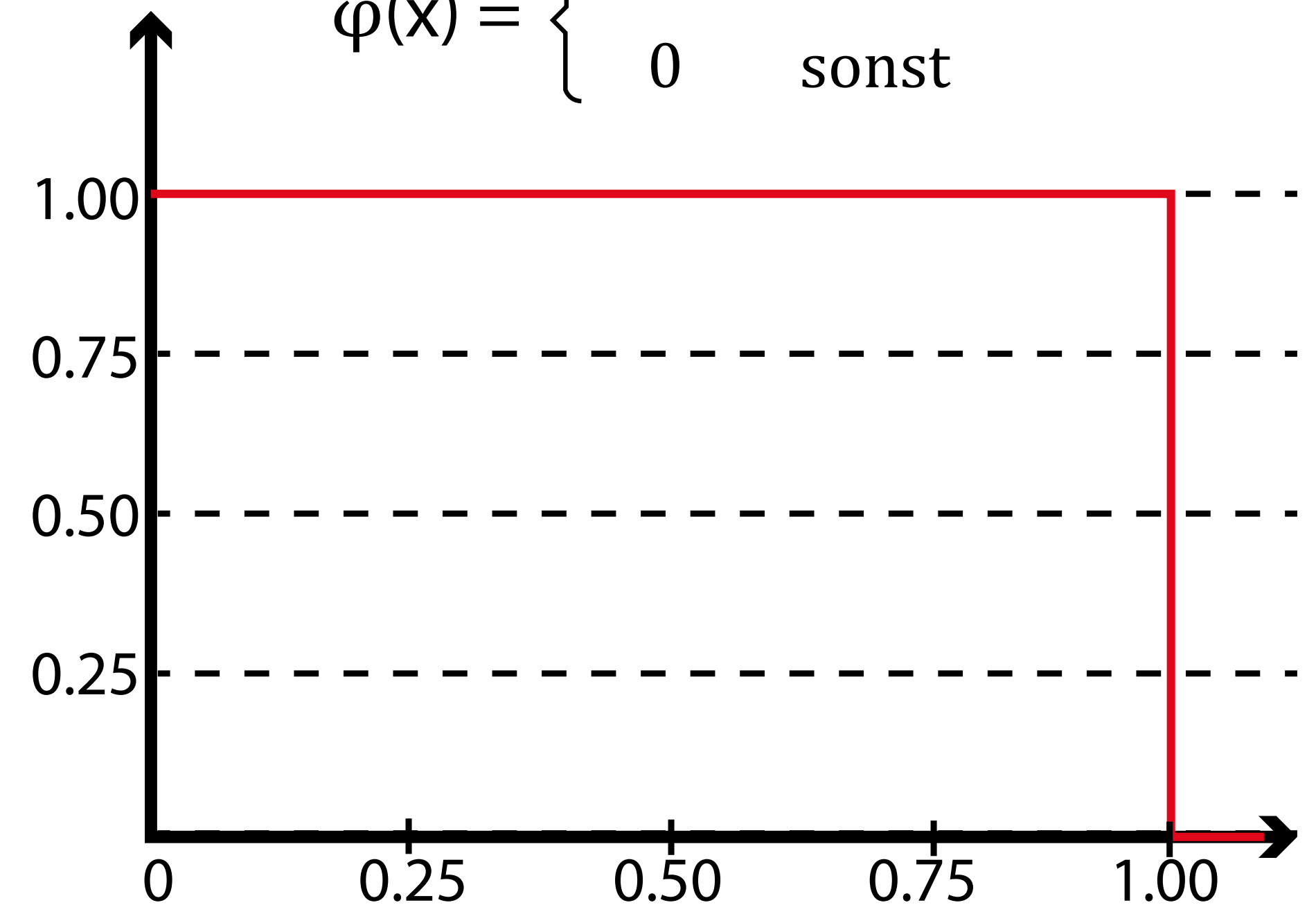


Erwartungswert

Wir berechnen den Erwartungswert über das Integral:

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

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



Erwartungswert

Der Erwartungswert ist linear additiv:

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

Beispiel: Wir würfeln mit drei 6-seitigen und einem 20-seitigen Würfel und berechnen die Augensumme. Der Erwartungswert wäre:

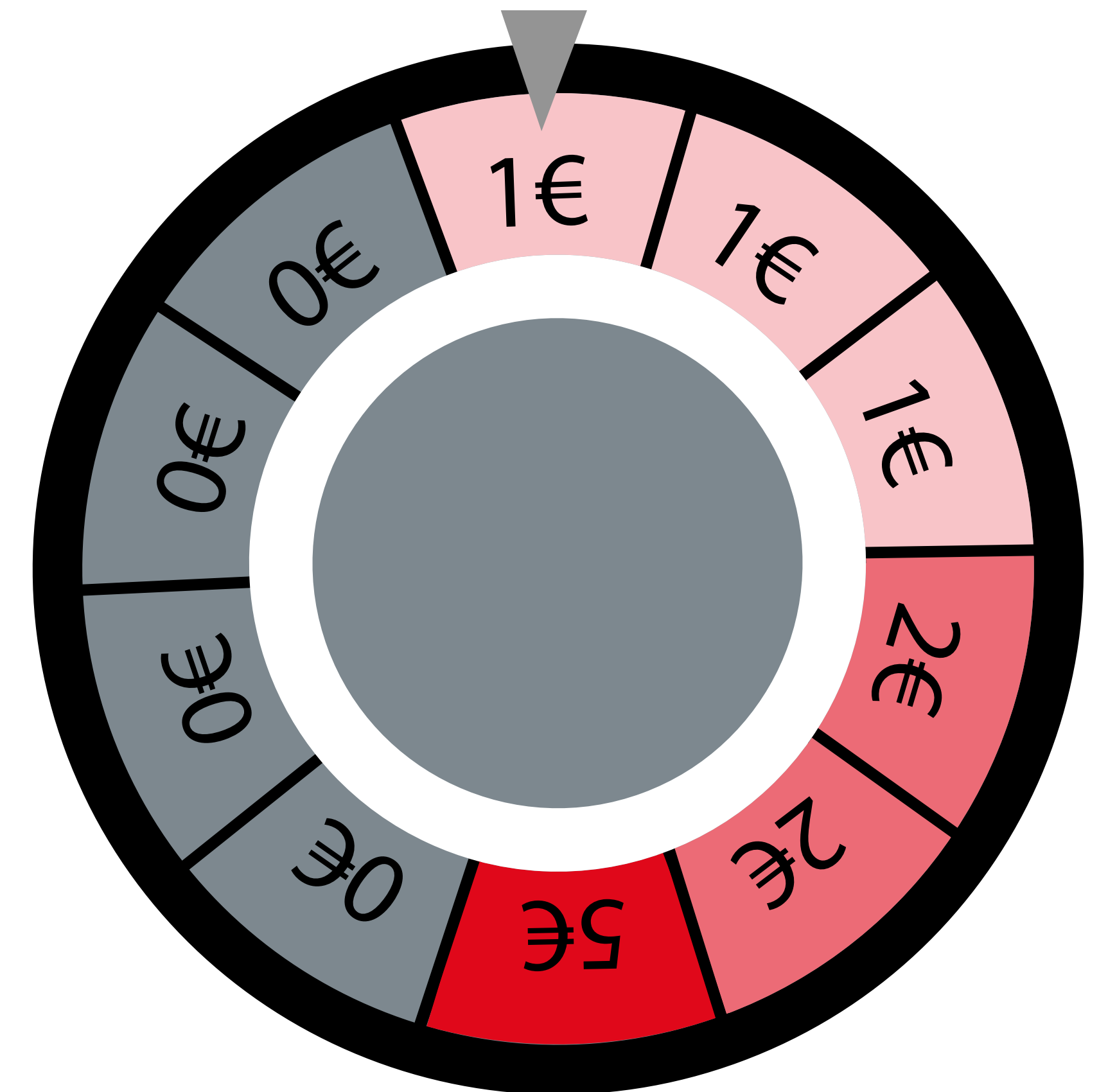
$$E(3X+Y) = 3 \cdot 3.5 + 10.5 = 21$$



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 $X+Y$

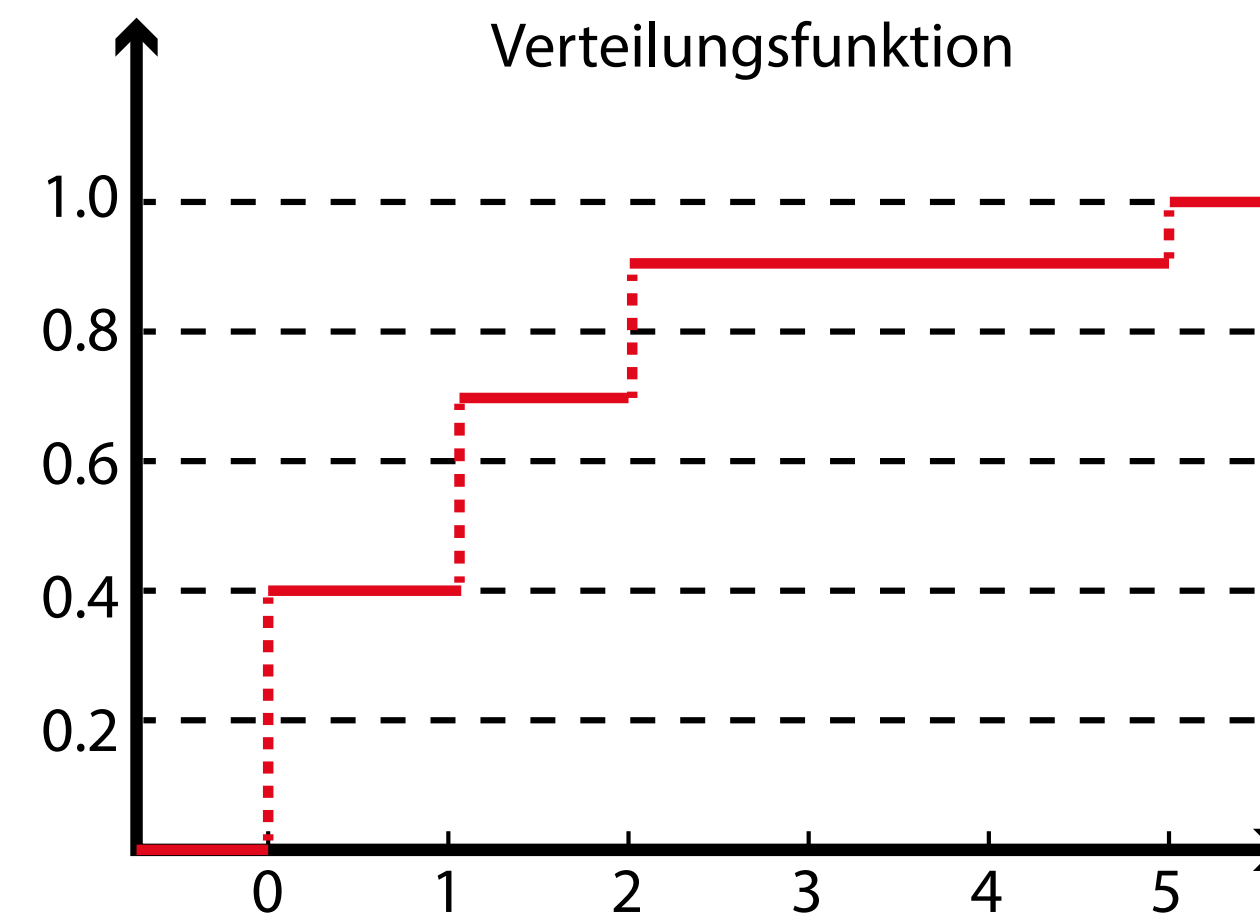
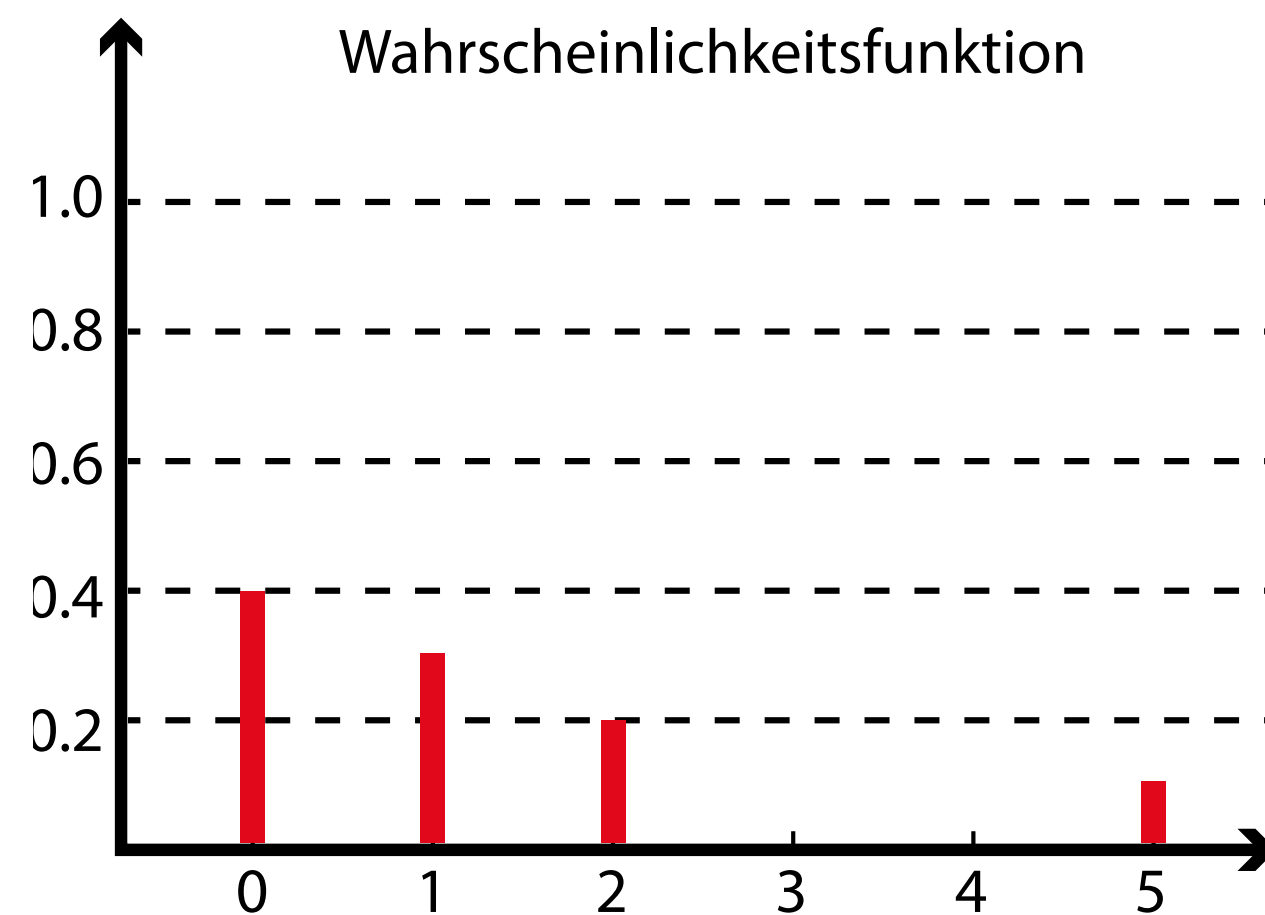
```
> runif(1,2,3)
```

$X = \dots$

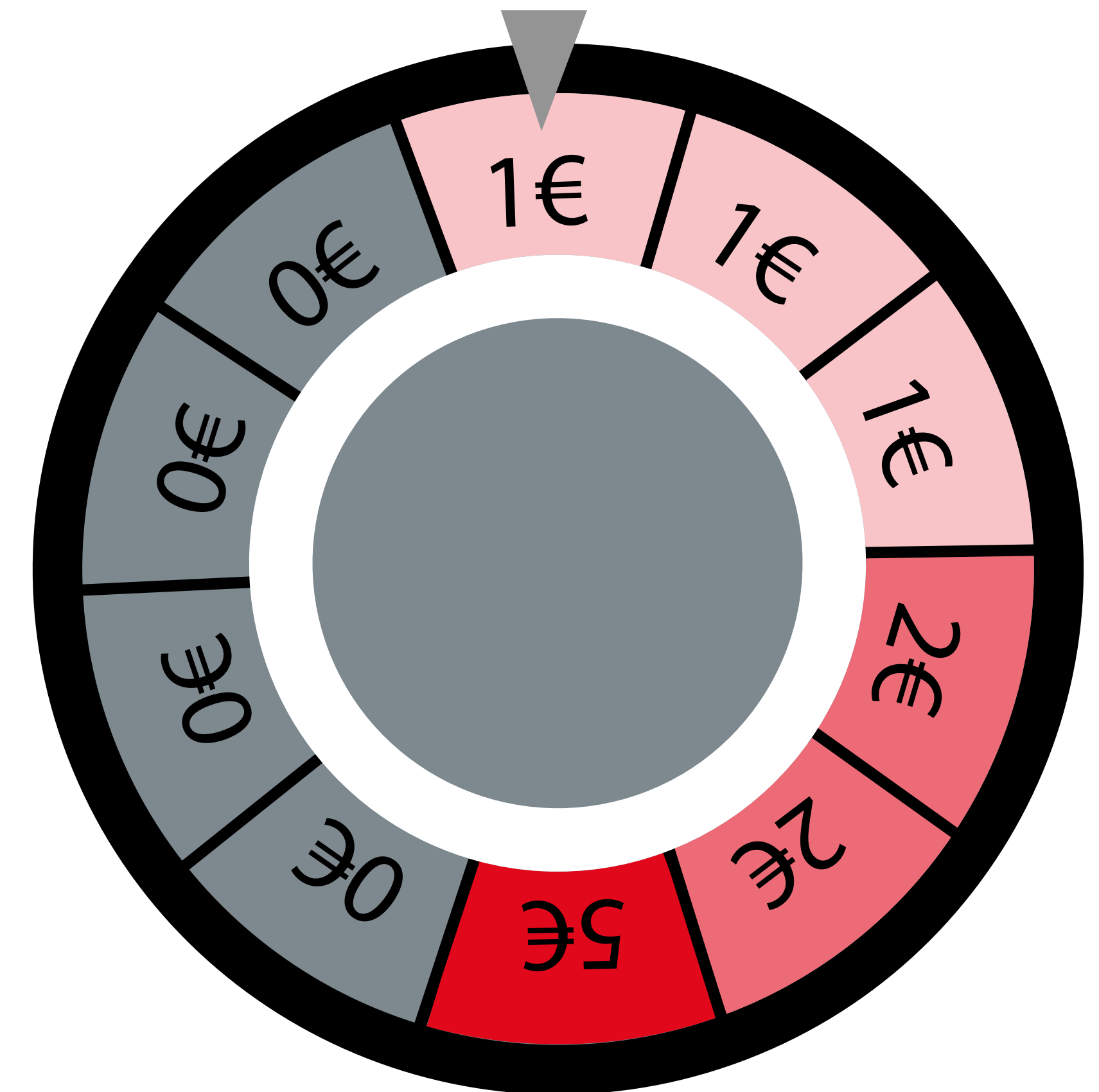
```
> runif(1,2,4)
```

$Y = \dots$

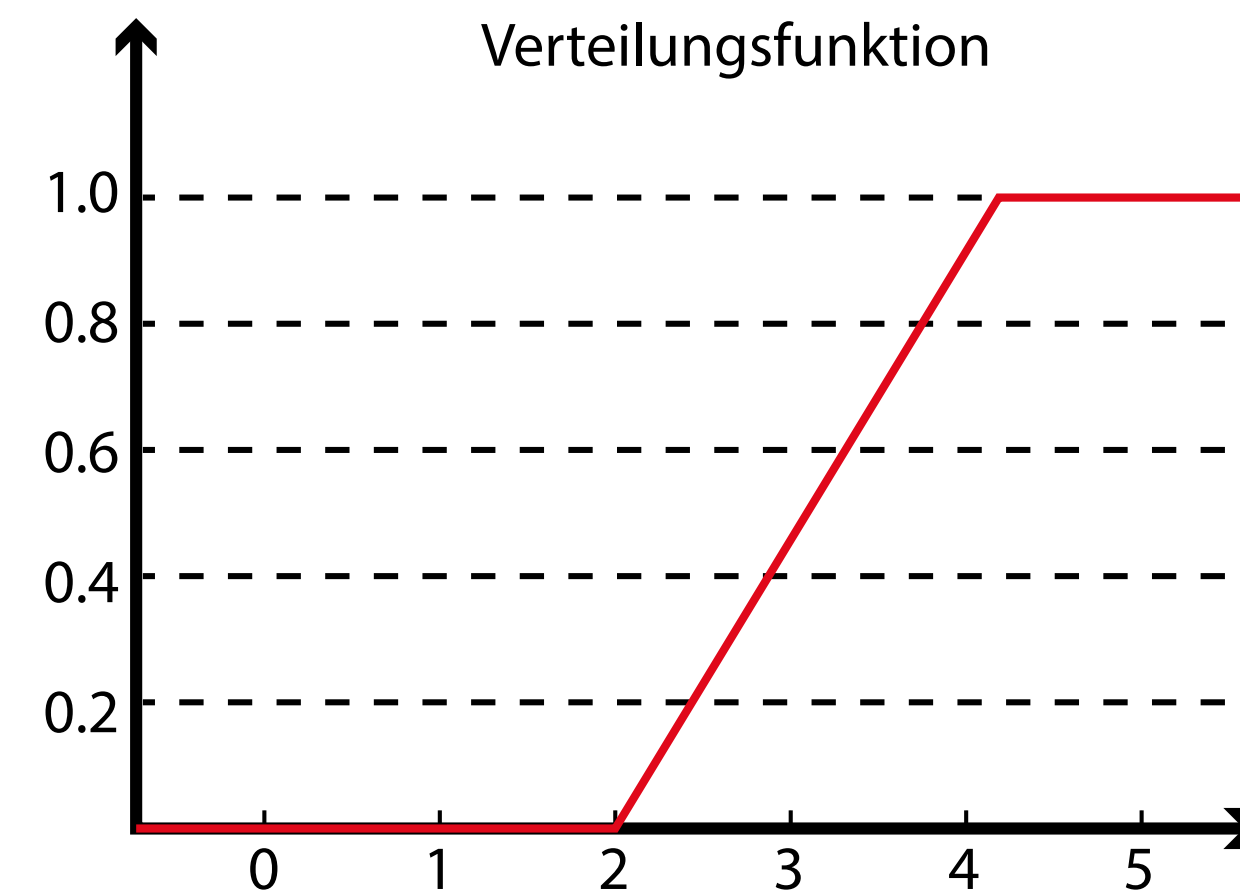
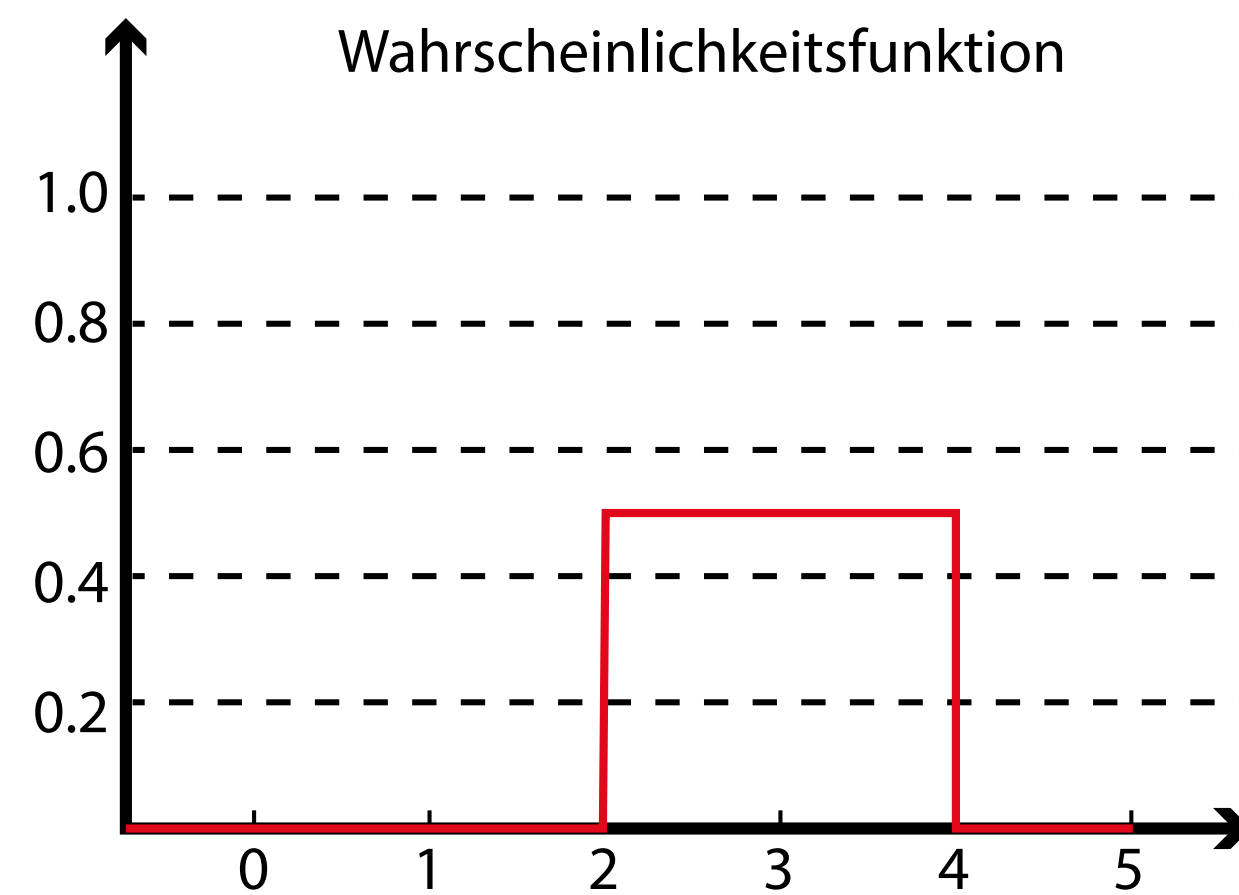
Erwartungswert



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



Erwartungswert



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

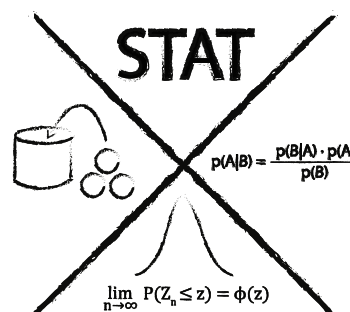
$$E(Y) = \int_{-\infty}^{\infty} y \cdot \varphi(y) dy = \int_2^4 0.5y dy = \left[0.25y^2 \right]_2^4 = 3 \quad E(X+Y) = 5.5$$

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

$X = \dots$

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

$Y = \dots$



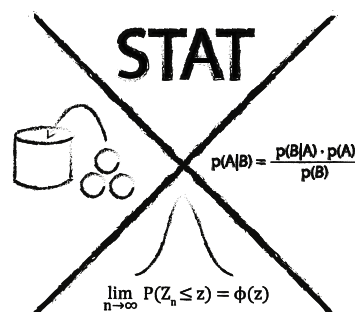
Varianz

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



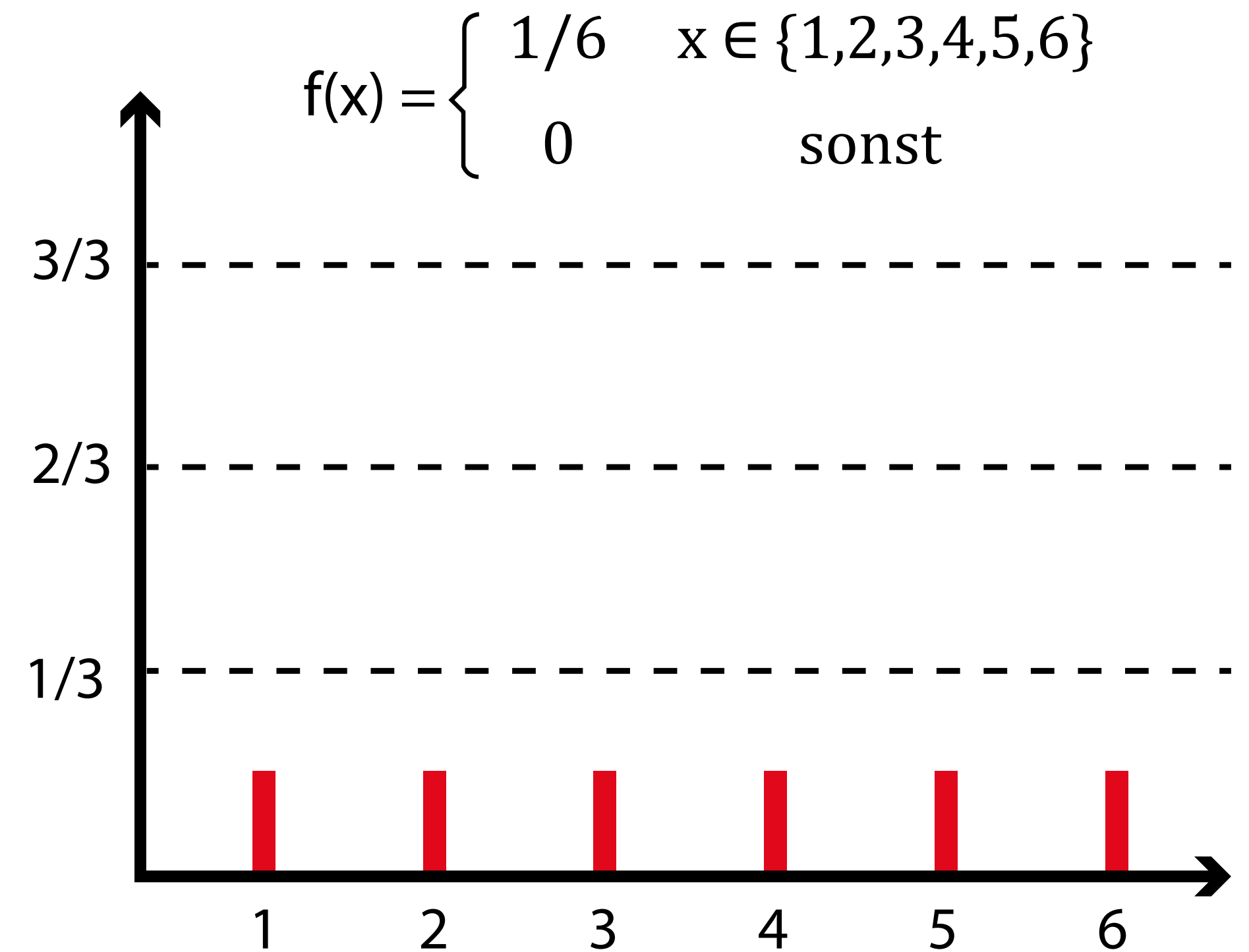
Varianz

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:

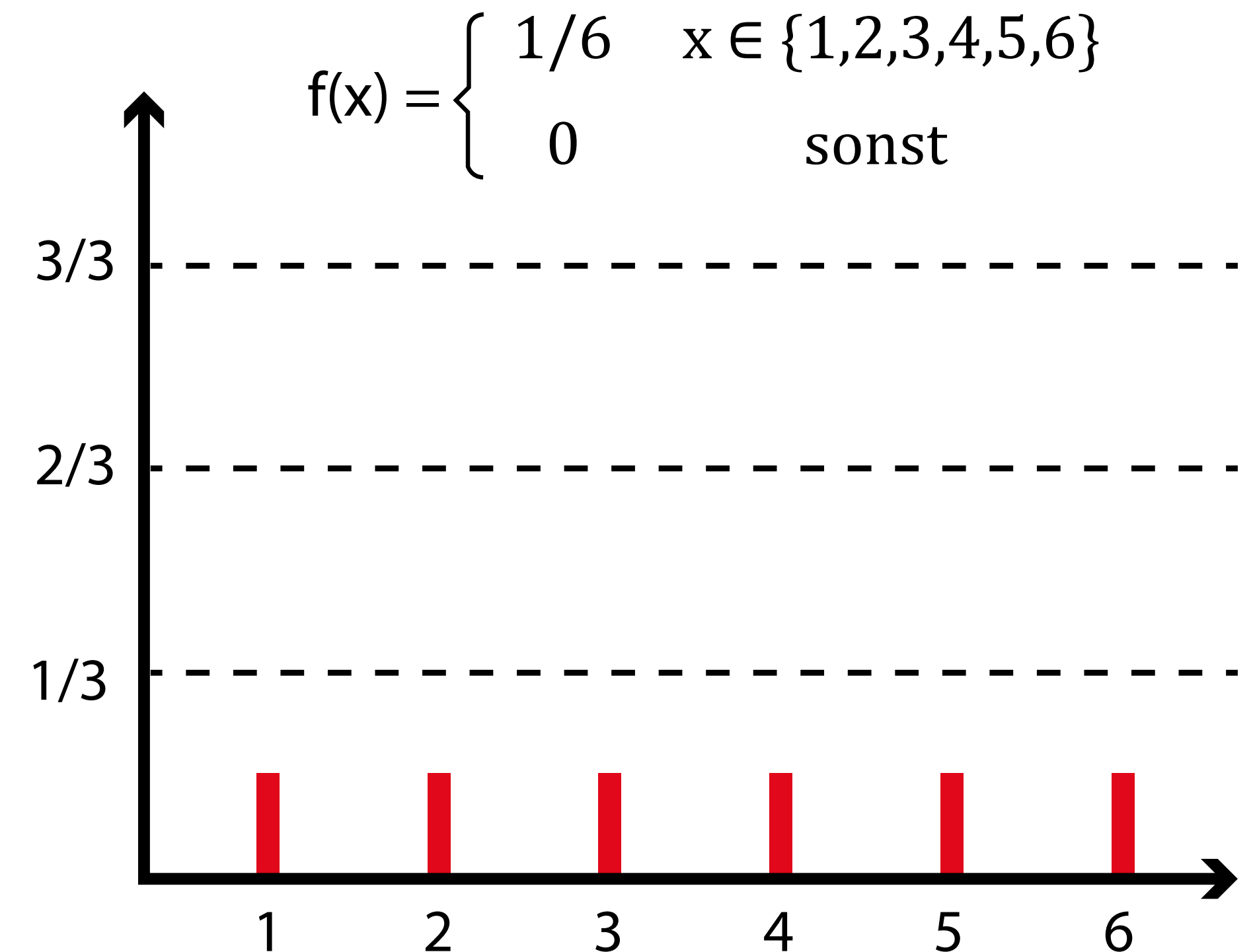
$$\text{Var}(X) = \sum_{x \in E} [x - E(X)]^2 \cdot f(x)$$



Varianz

Wir berechnen die Varianz über die Summe:

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



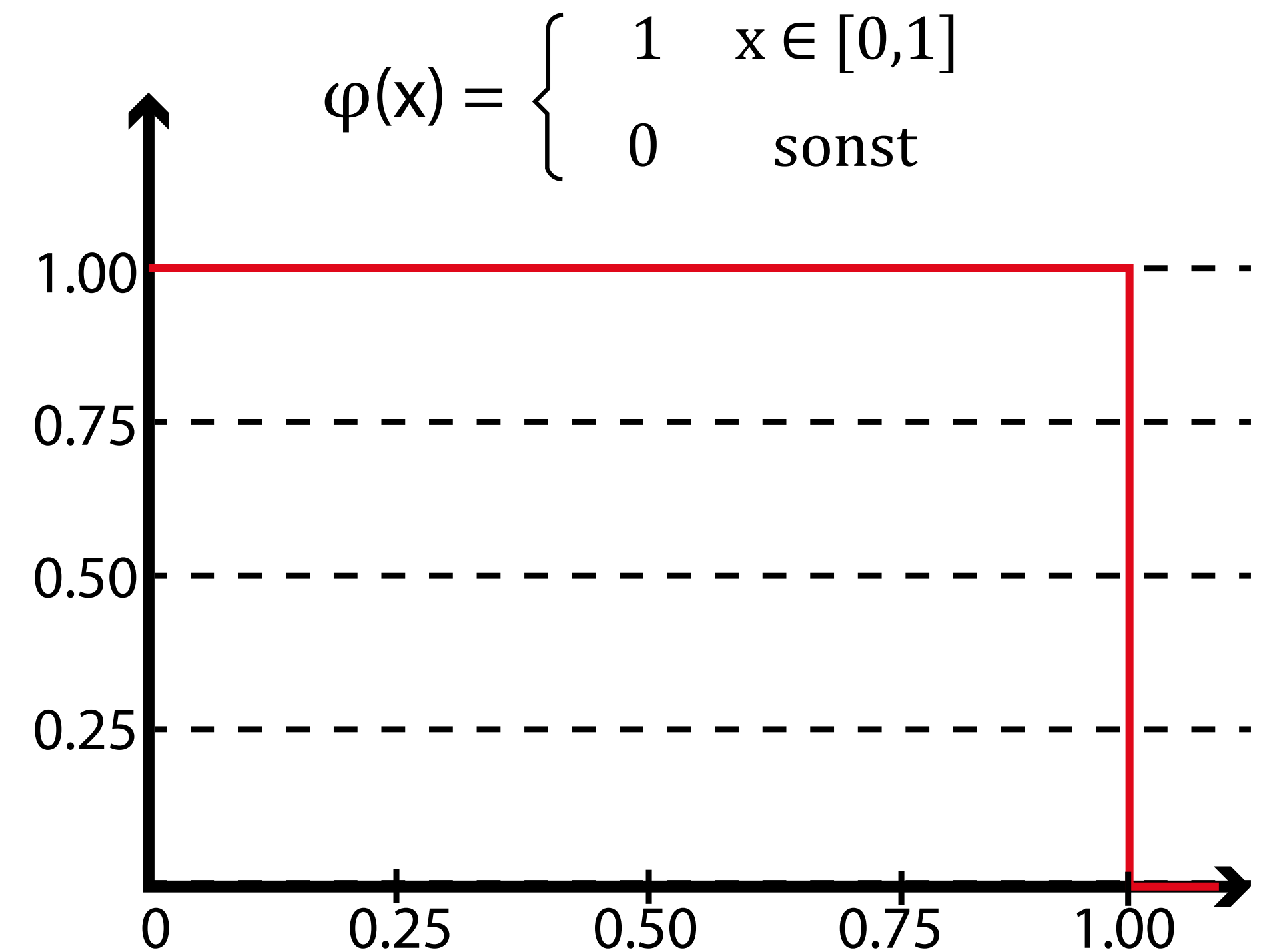
Varianz

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:

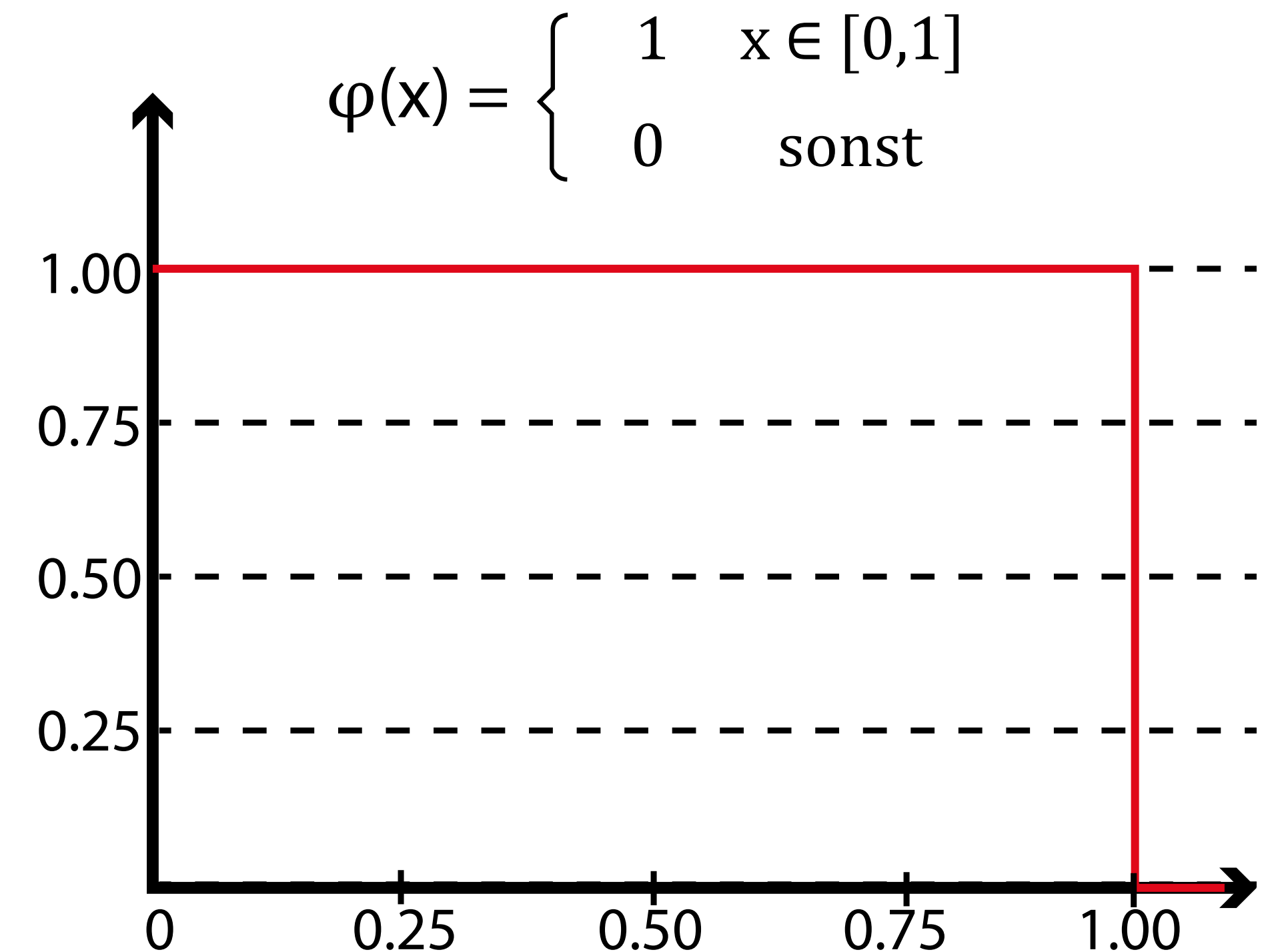
$$\text{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot \varphi(x) \, dx$$



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

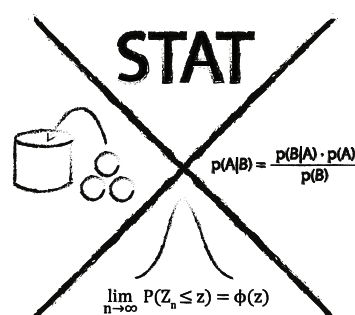
Die Varianz ist nur dann linear additiv, wenn die Zufallsvariablen unabhängig sind. Allgemein gilt:

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

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

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

σ_X^2

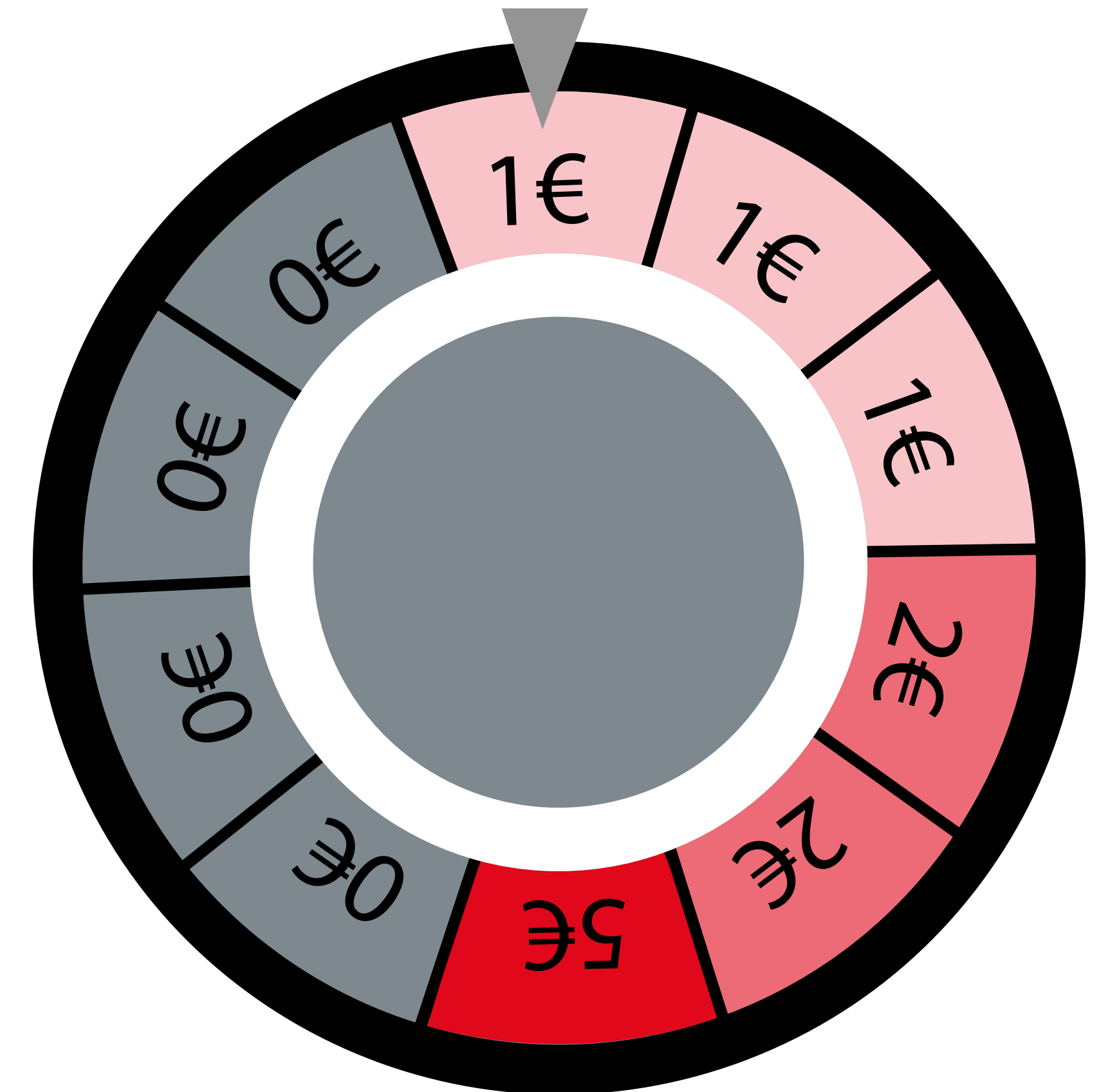


Varianz

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!

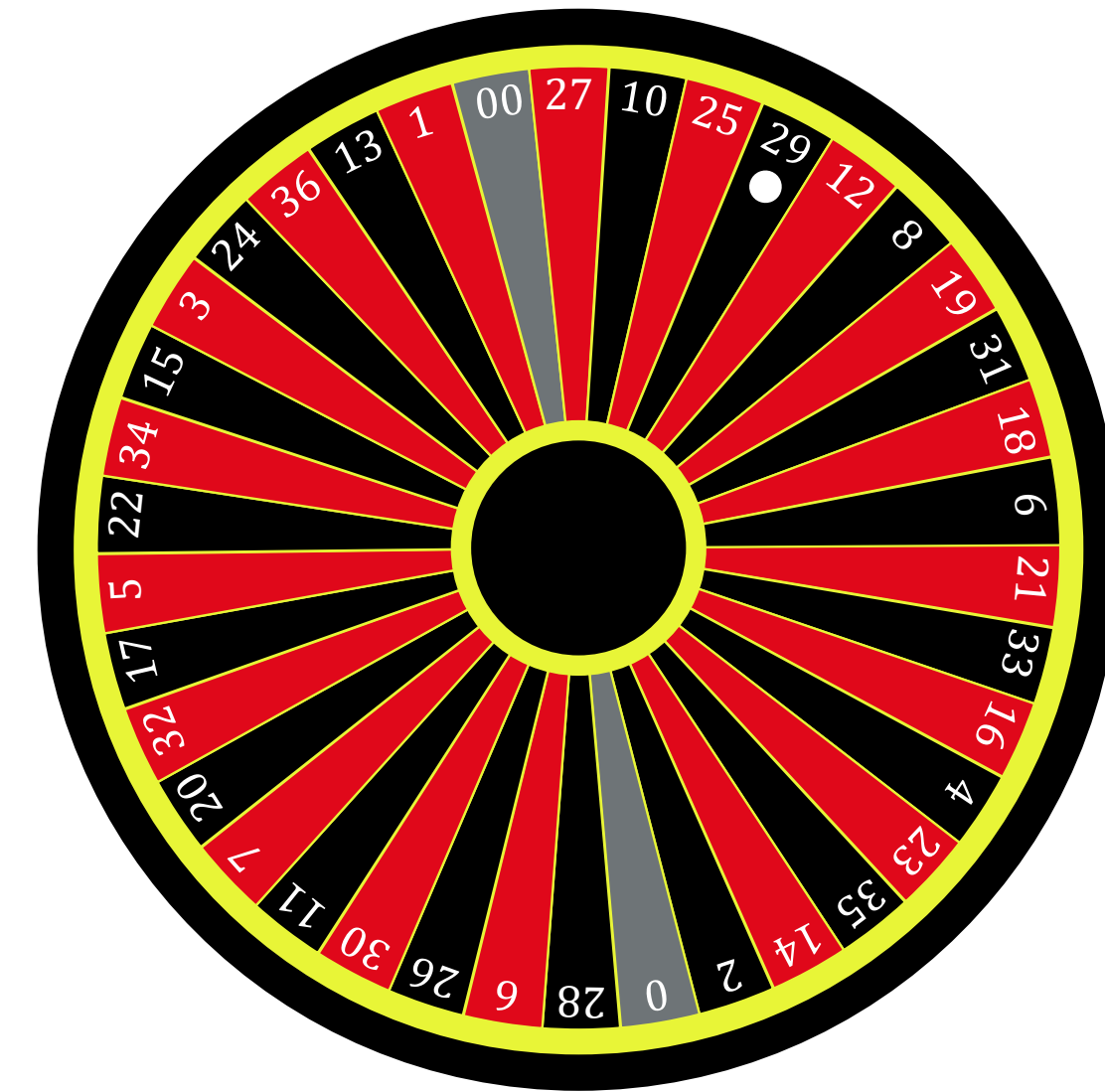


Varianz

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 .

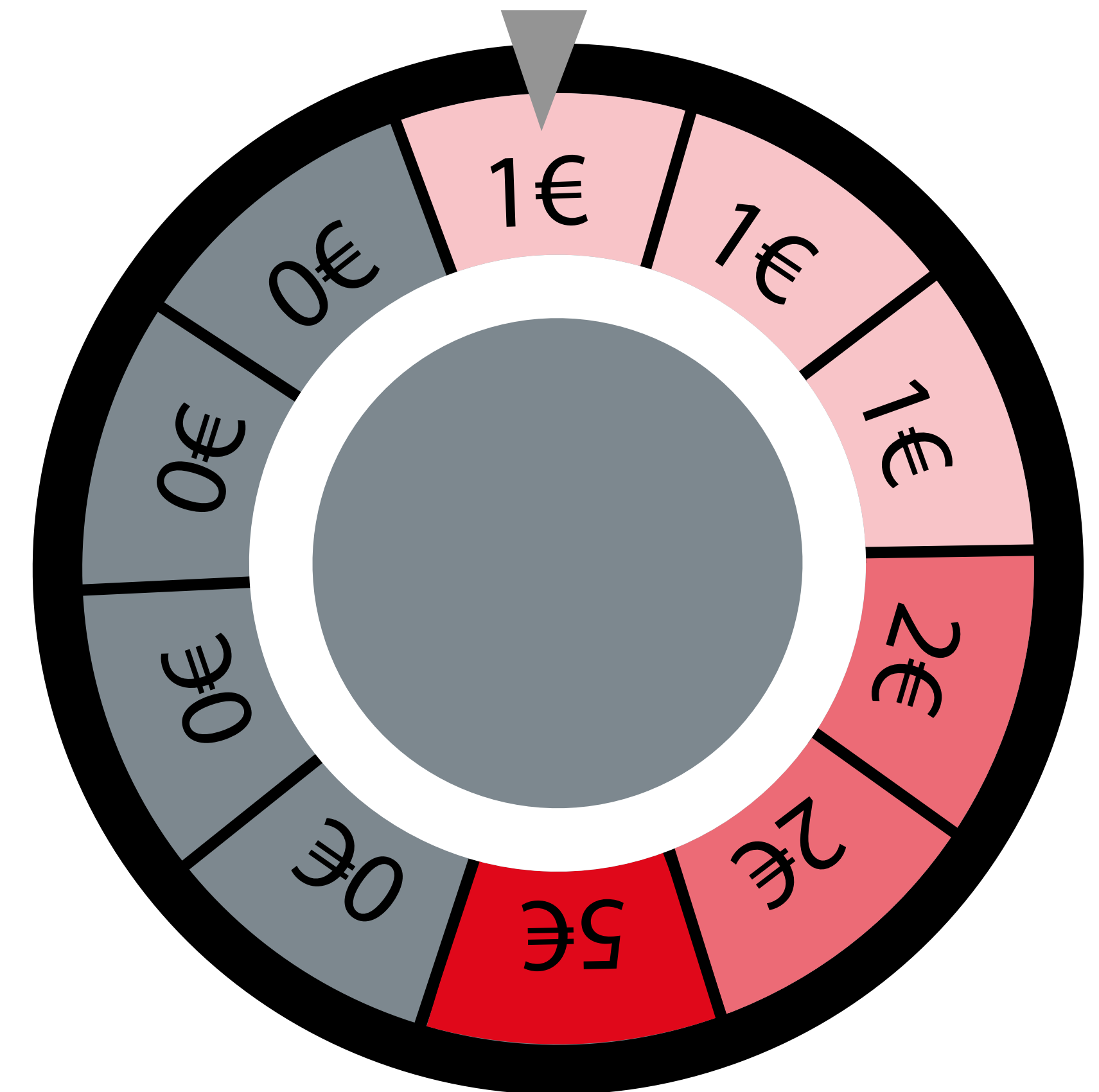


00	0	$\Omega = \{00, 0, 1, 2, \dots, 36\}$									
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

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

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

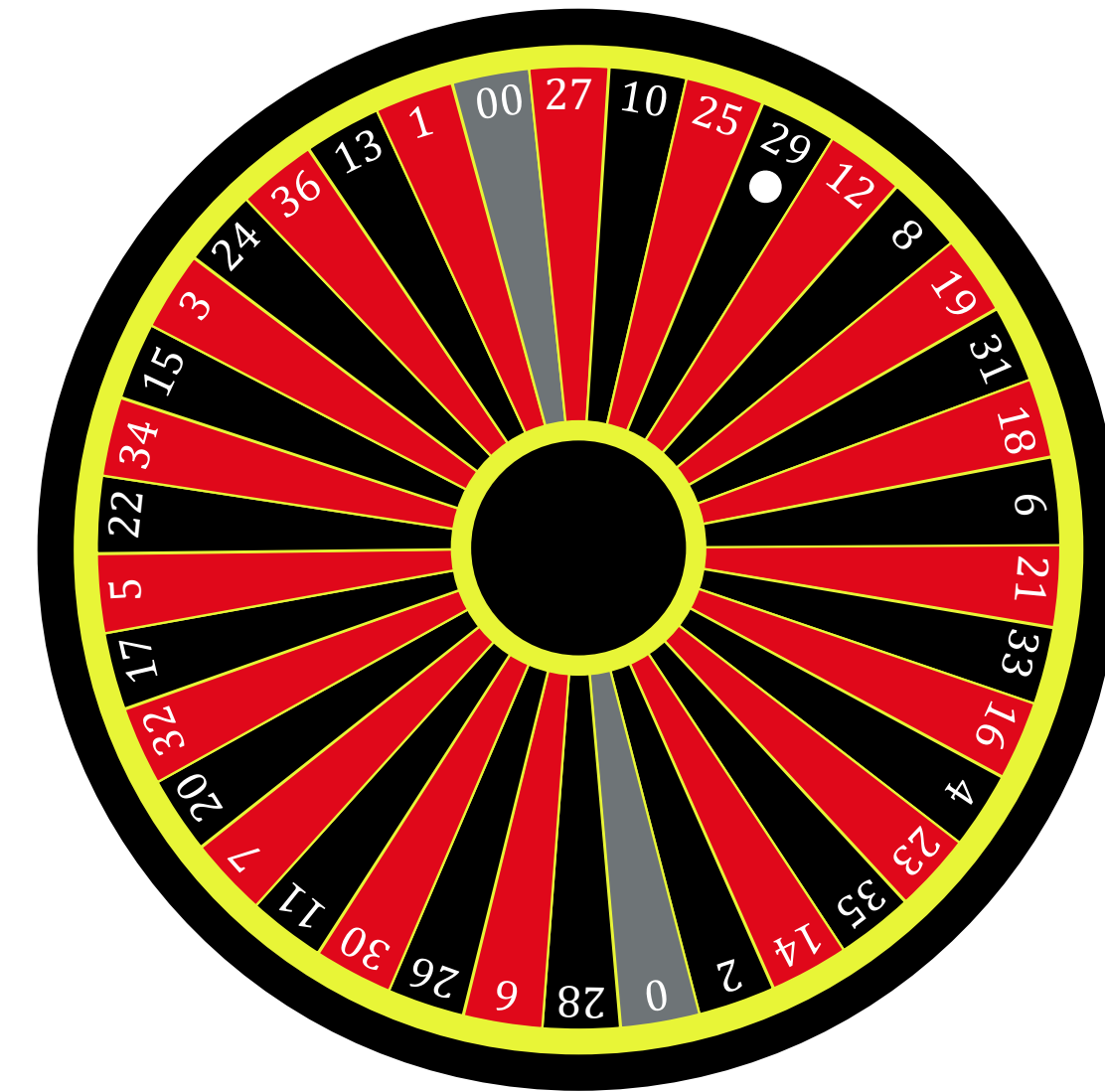
$$\sigma_X = \sqrt{\text{Var}(X)} = 1.47$$



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

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

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



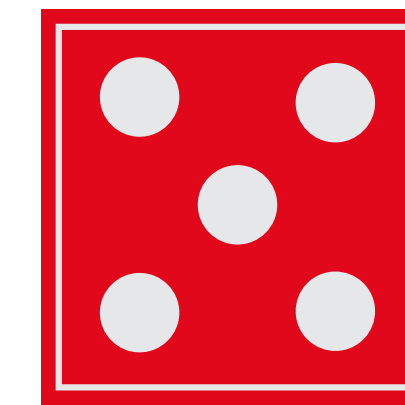
00	0	$\Omega = \{00, 0, 1, 2, \dots, 36\}$									
1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36

Zentraler Grenzwertsatz

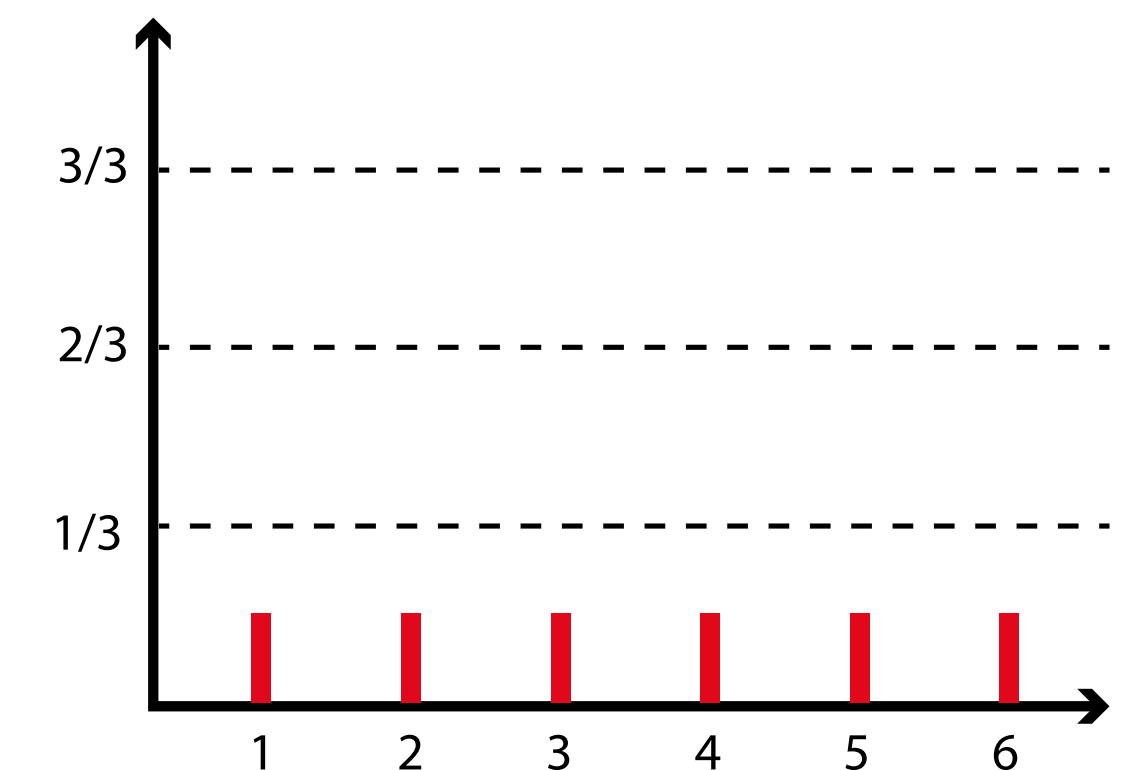
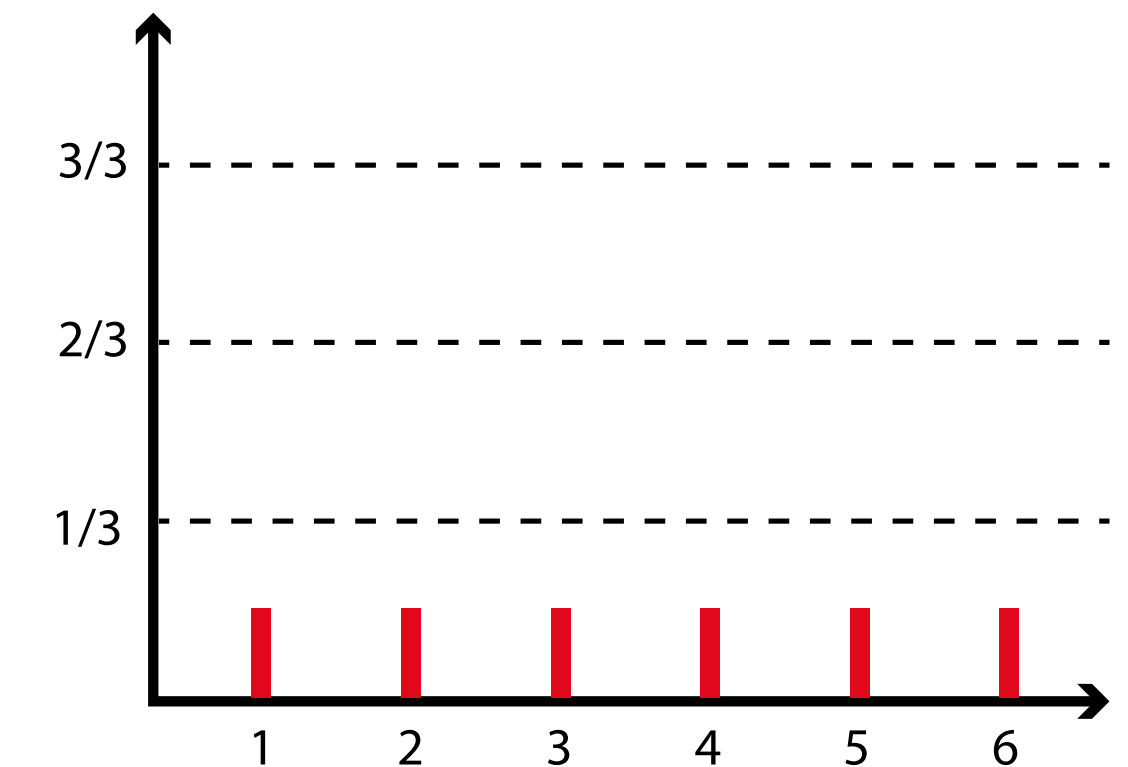
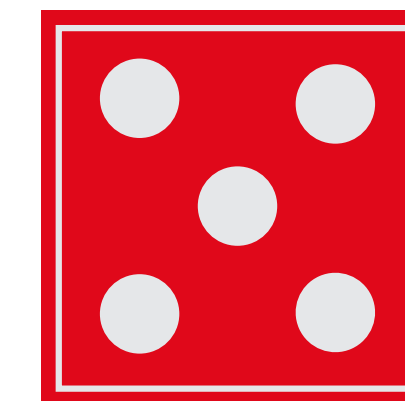
Wir werfen zwei 6-seitige Würfel und definieren X als die Summe der beiden Augenzahlen $X_1 + X_2$

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?



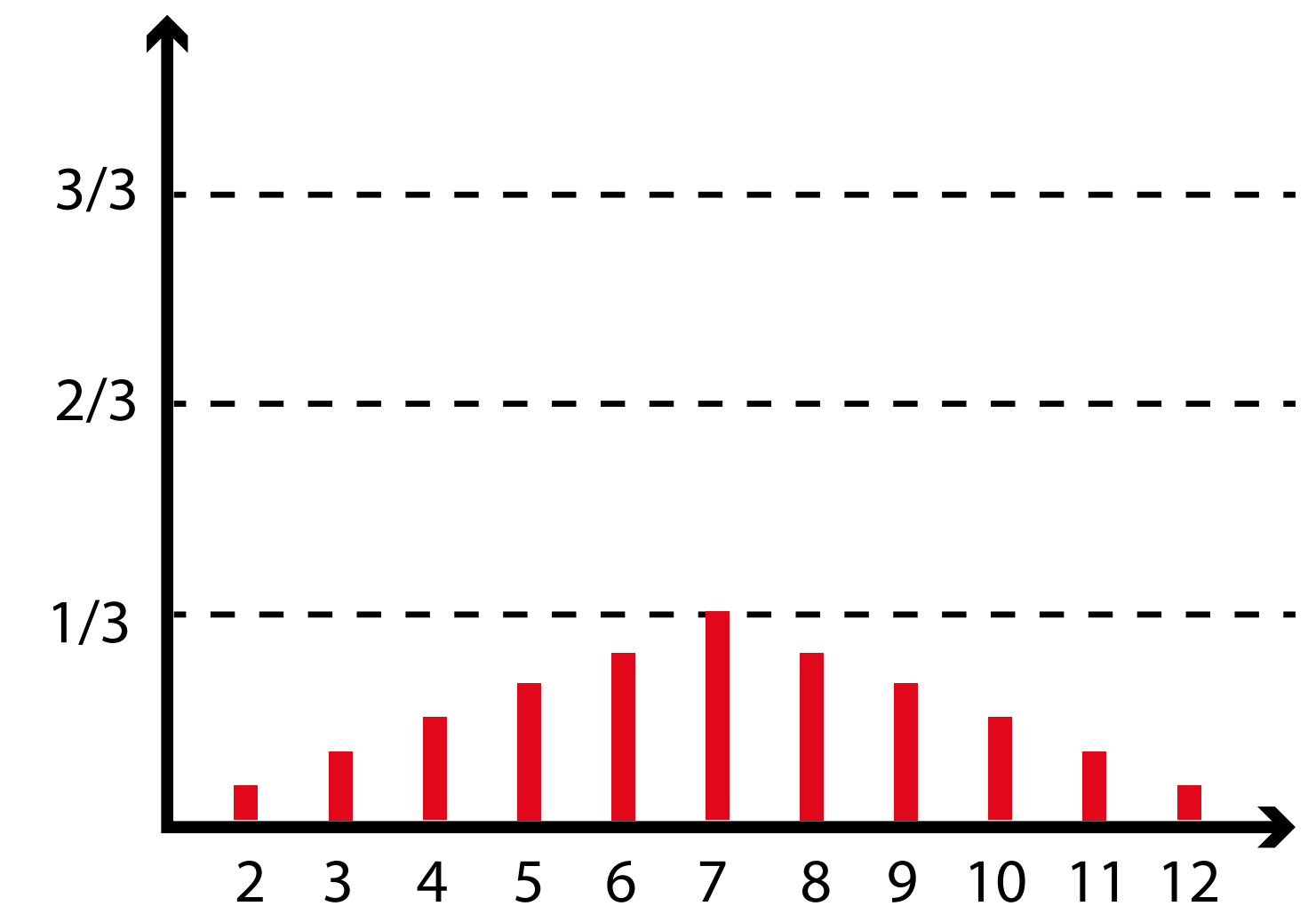
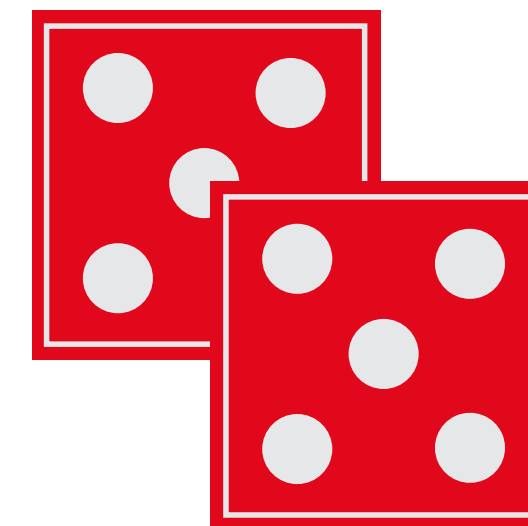
+



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!



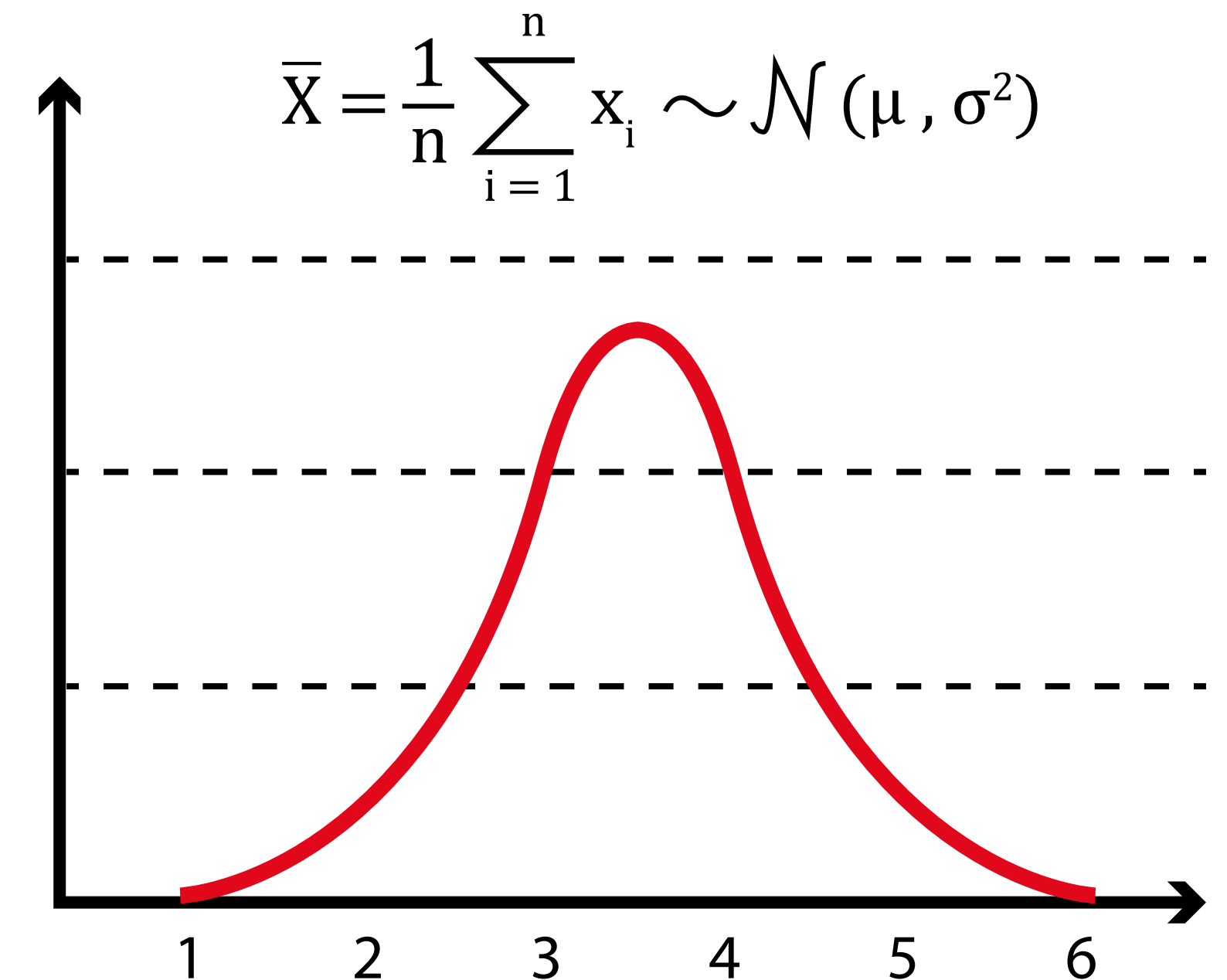
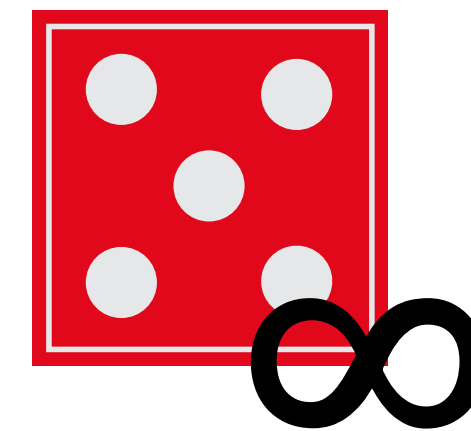
Zentraler Grenzwertsatz

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:

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

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

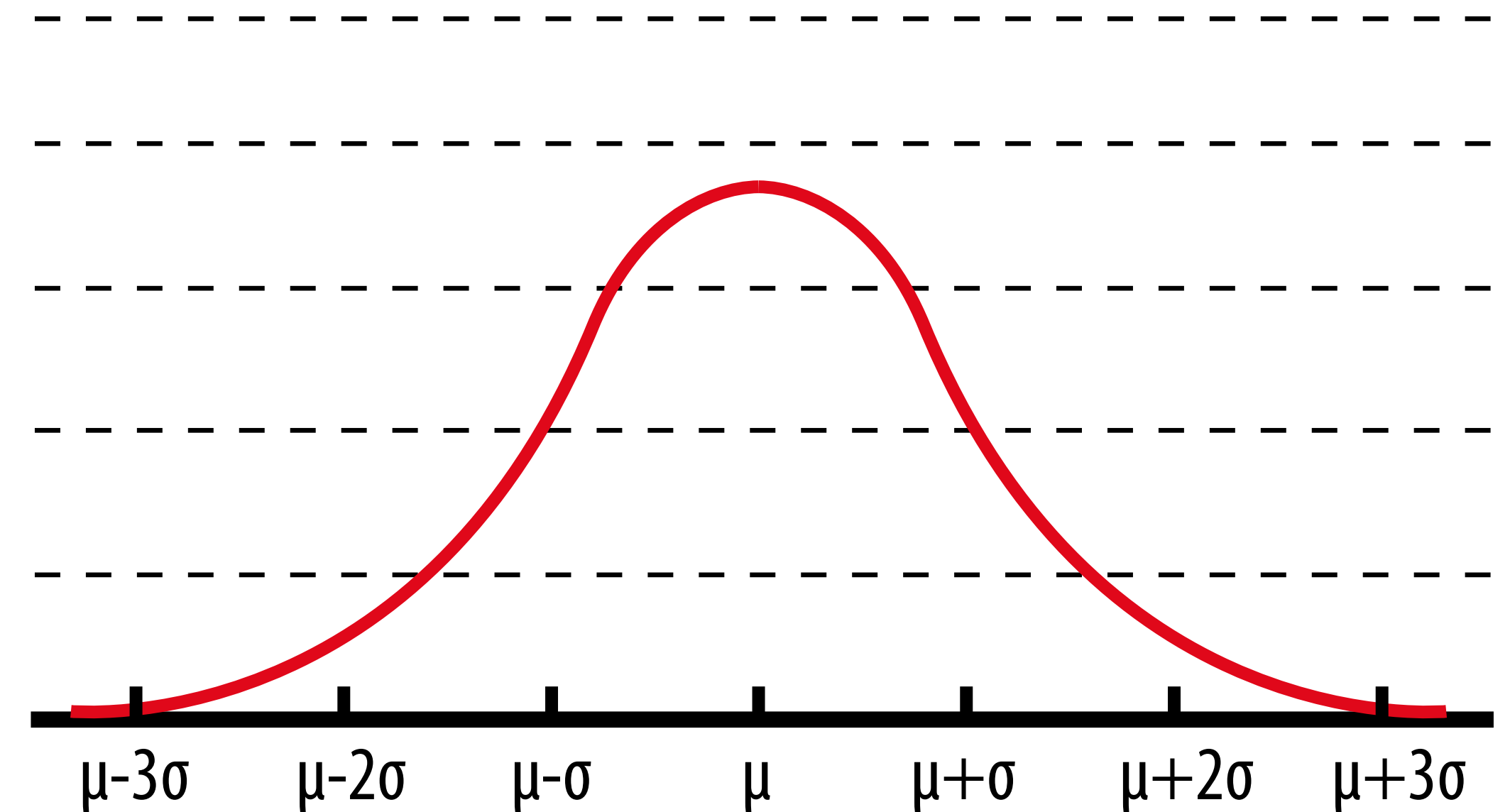


Normalverteilung

Mit der Normalverteilung lernen wir die wichtigste Verteilung kennen. Ihre Verteilungs- und Dichtefunktionen sind leider sehr kompliziert:

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

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



In Excel verfügbar mit der Funktion:

=NORM.VERT(x;μ;σ;WAHR) für Verteilungsfunktion
=NORM.VERT(x;μ;σ;FALSCH) für Dichtefunktion

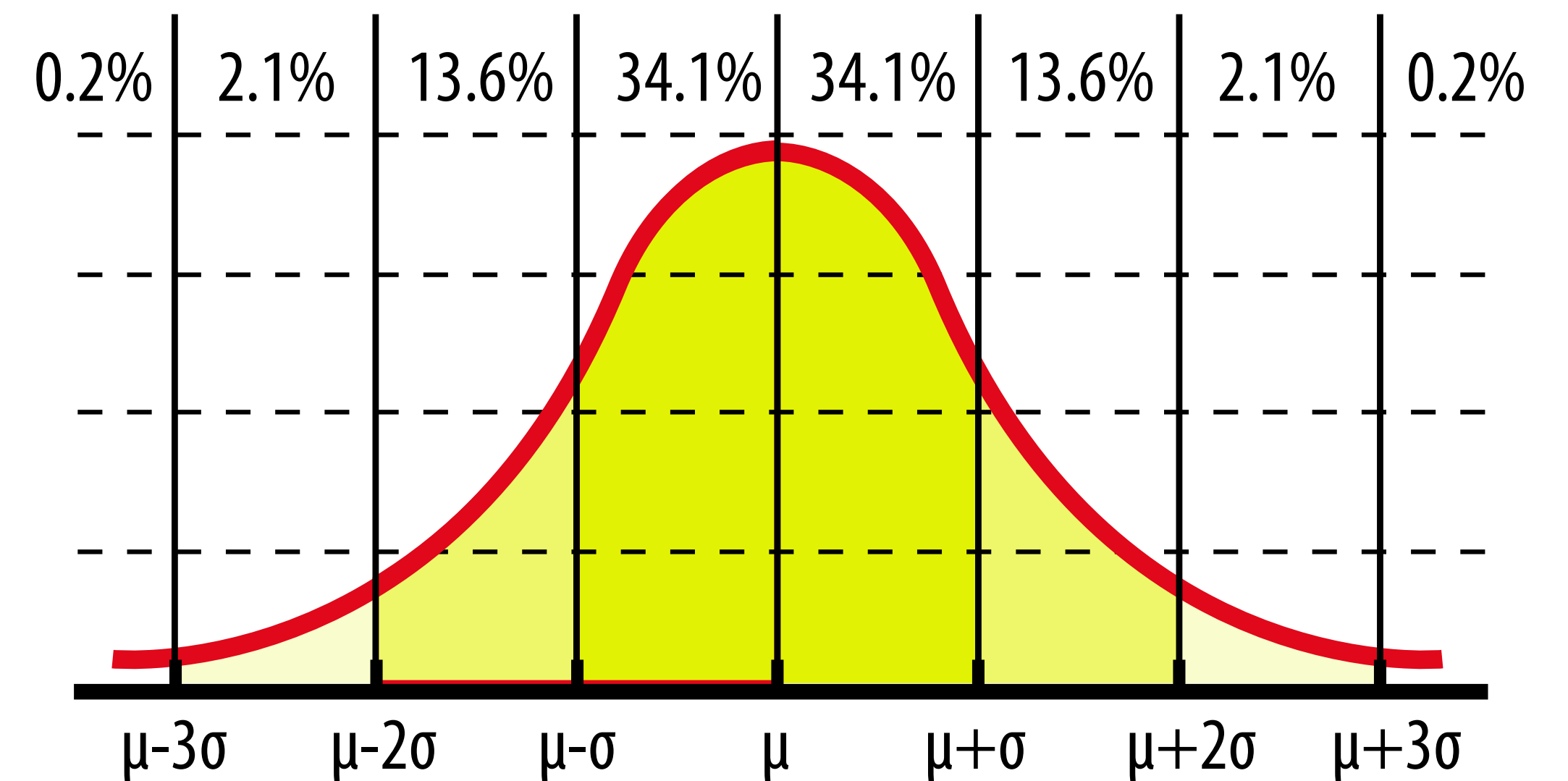
Normalverteilung

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 .



Normalverteilung

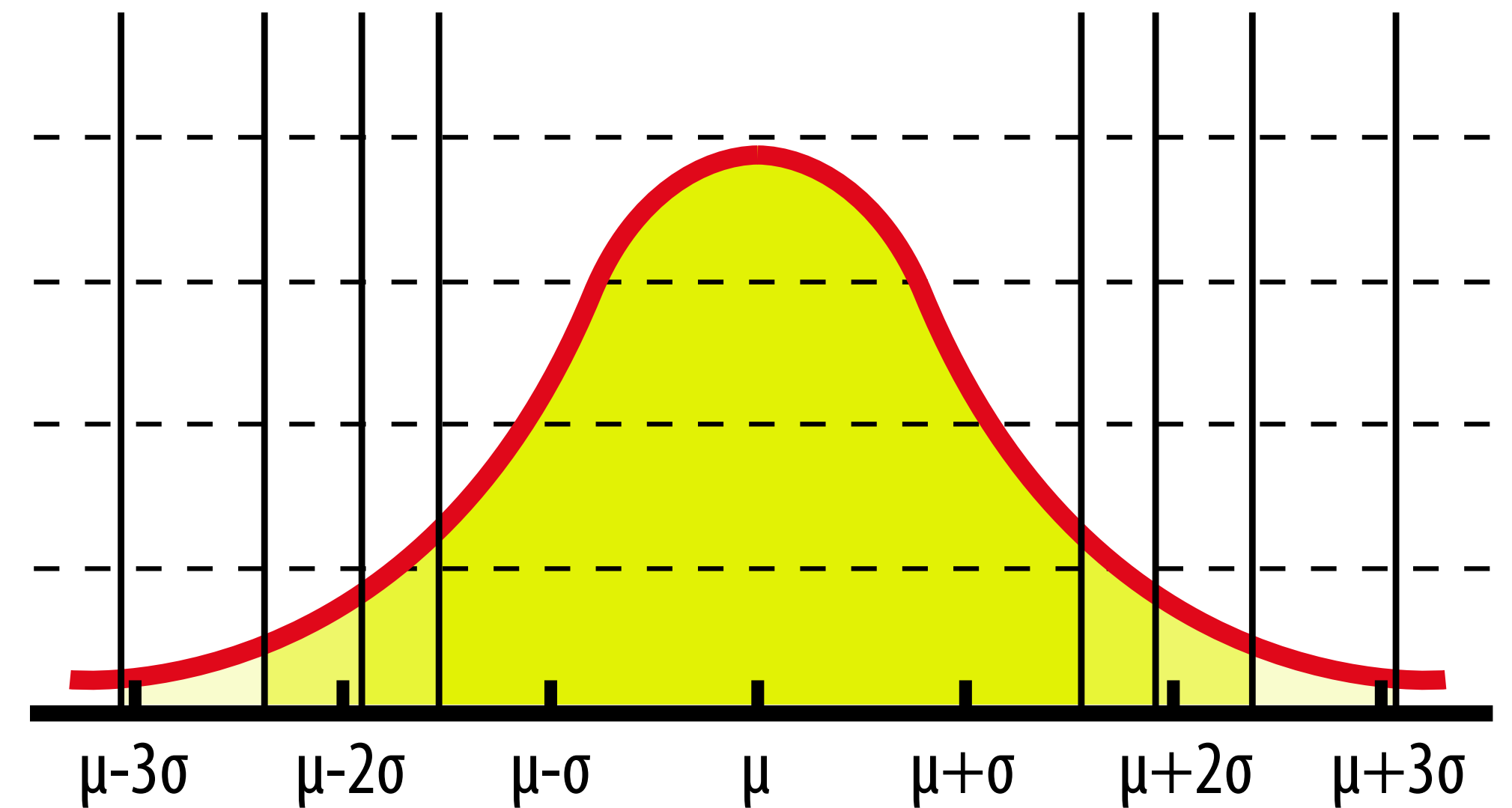
Mithilfe der Dichtefunktion können wir die Wahrscheinlichkeit berechnen, mit dem die Werte in einem bestimmten Bereich liegen.

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 .



In Excel berechenbar mit der Funktion:

`=NORM.INV(Prozentsatz; μ ; σ)`

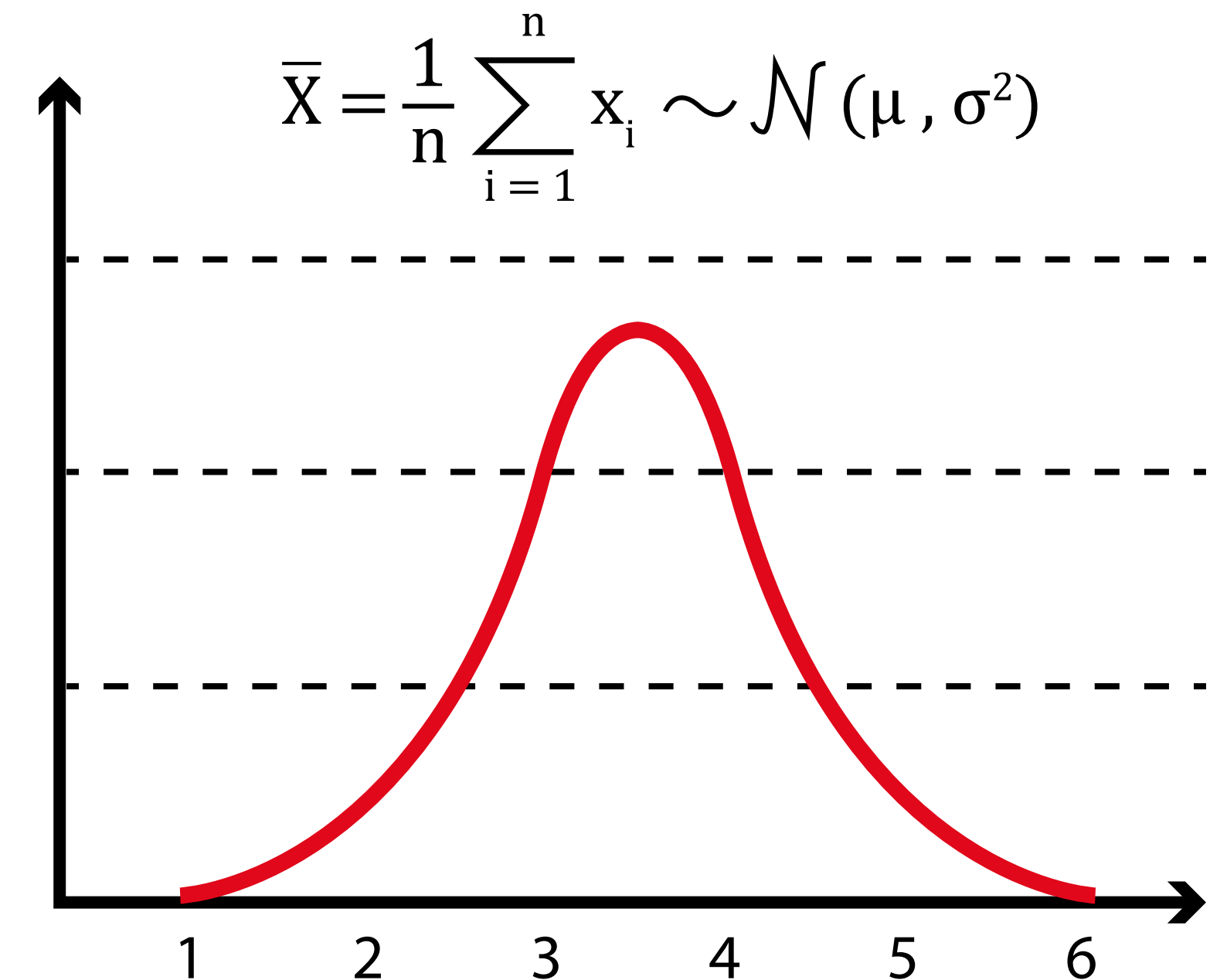
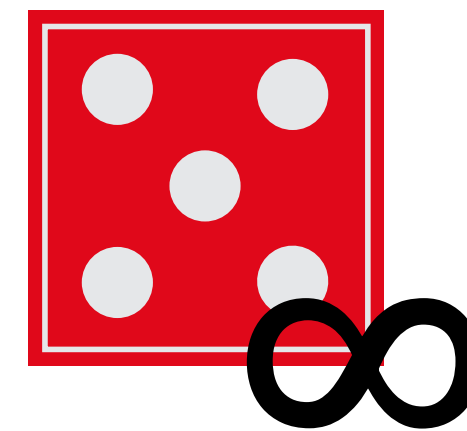
Zentraler Grenzwertsatz

Die Summe der Zufallsvariablen ist, wenn n groß genug ist, normalverteilt mit Erwartungswert $n\mu$ und Varianz $n\sigma^2$.

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

Der „Durchschnittswert“ von X ist somit:

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

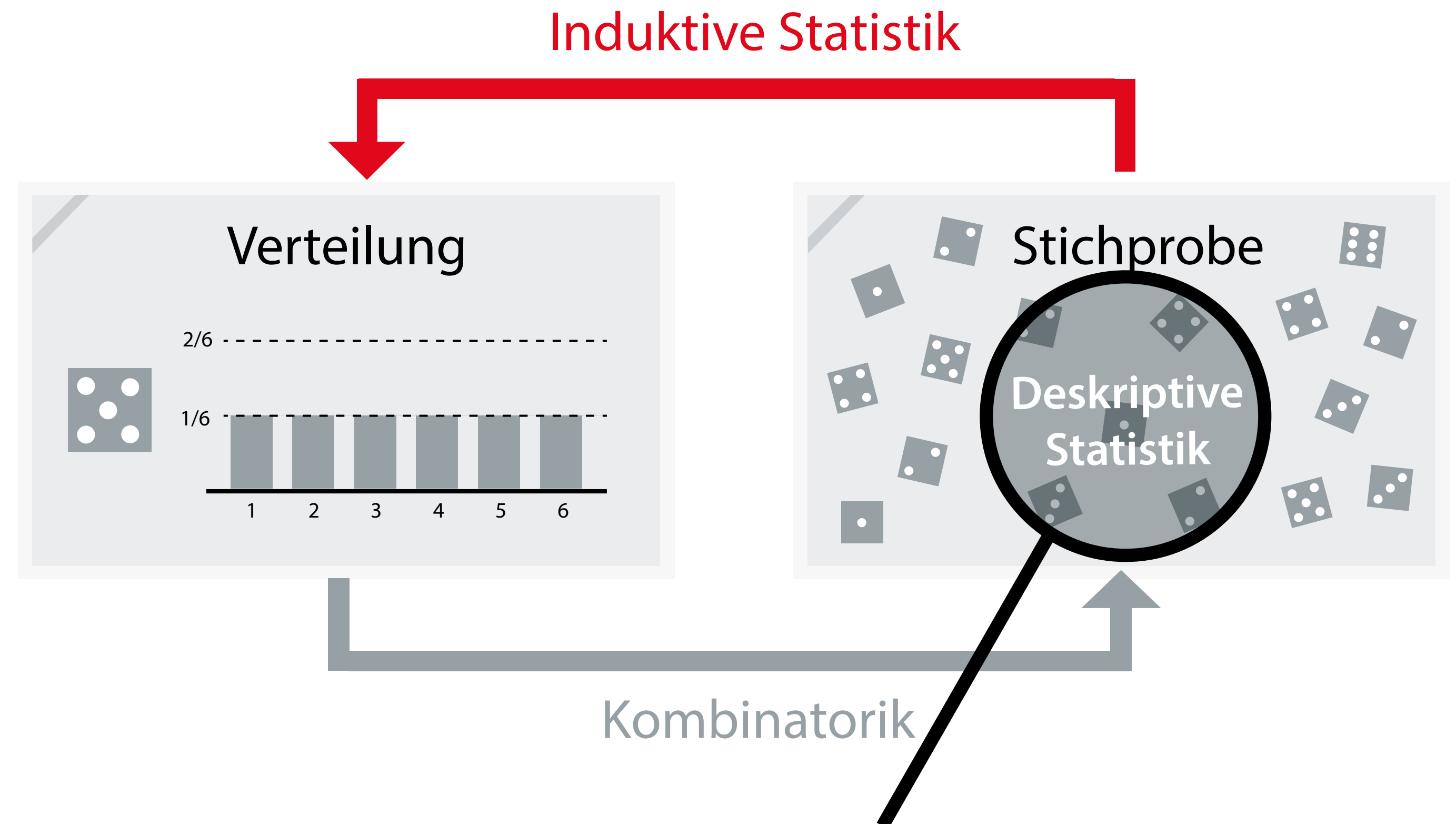


Induktive Statistik

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.



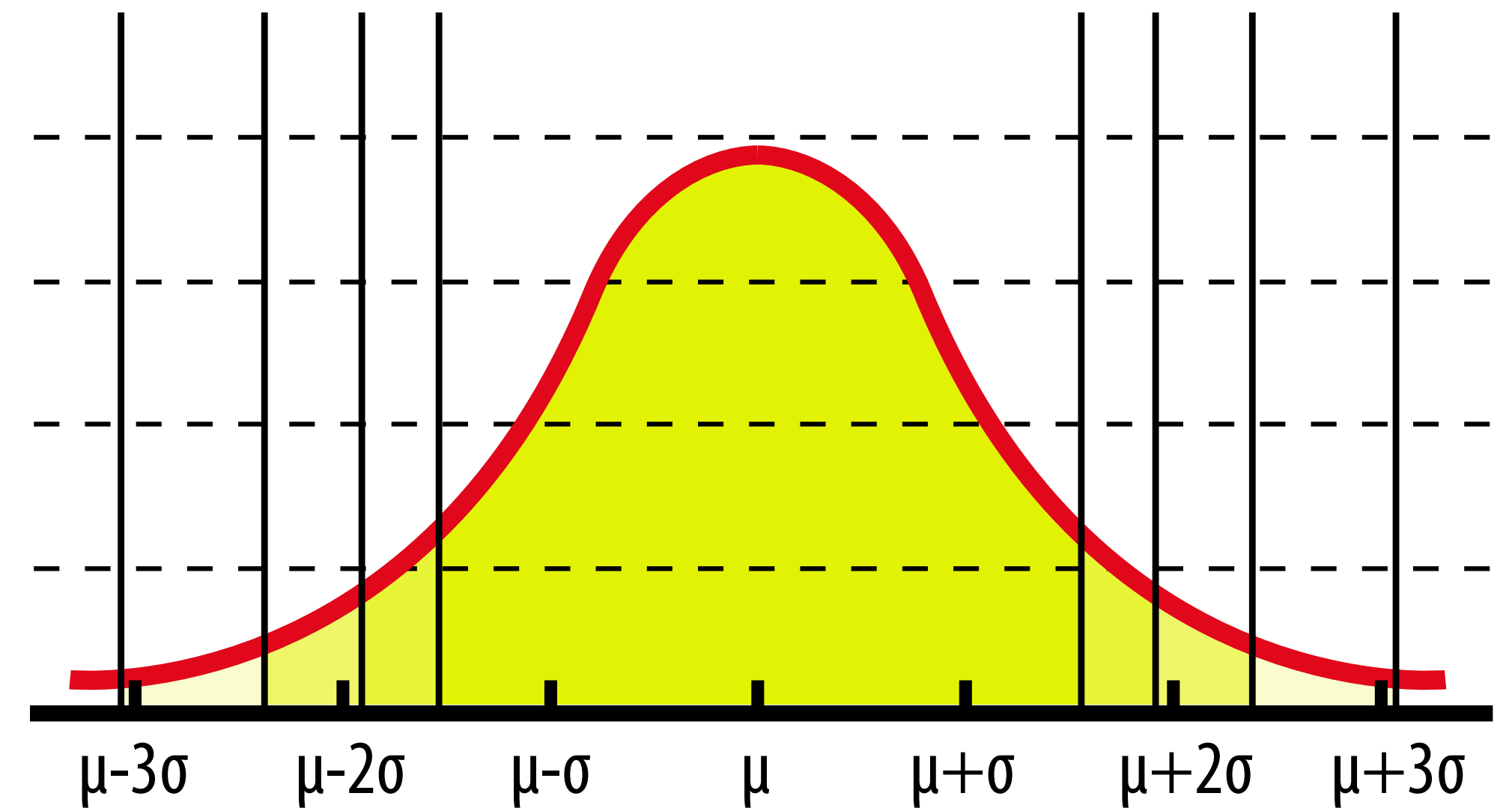
Z-Test

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:

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



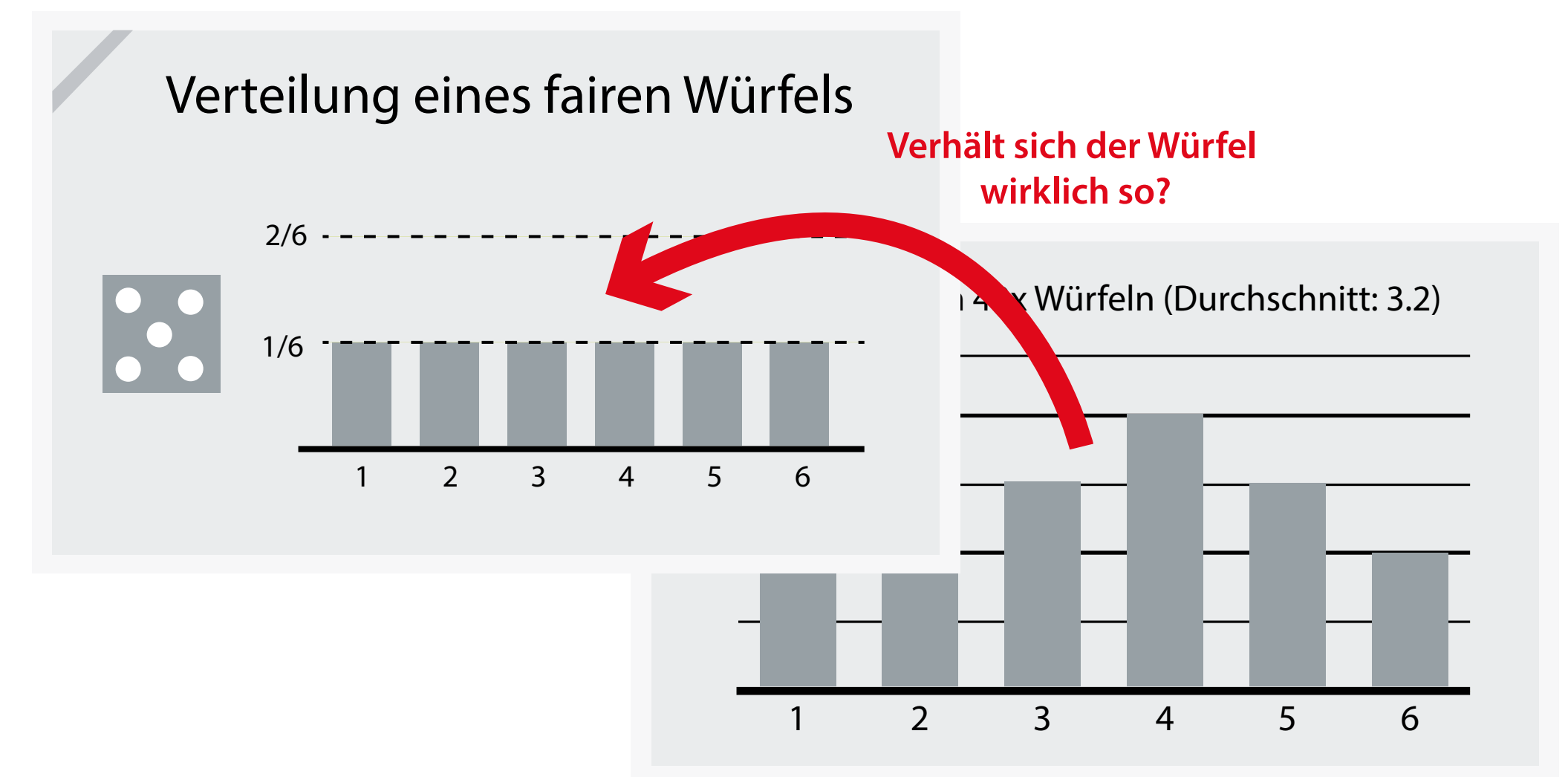
Z-Test

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.



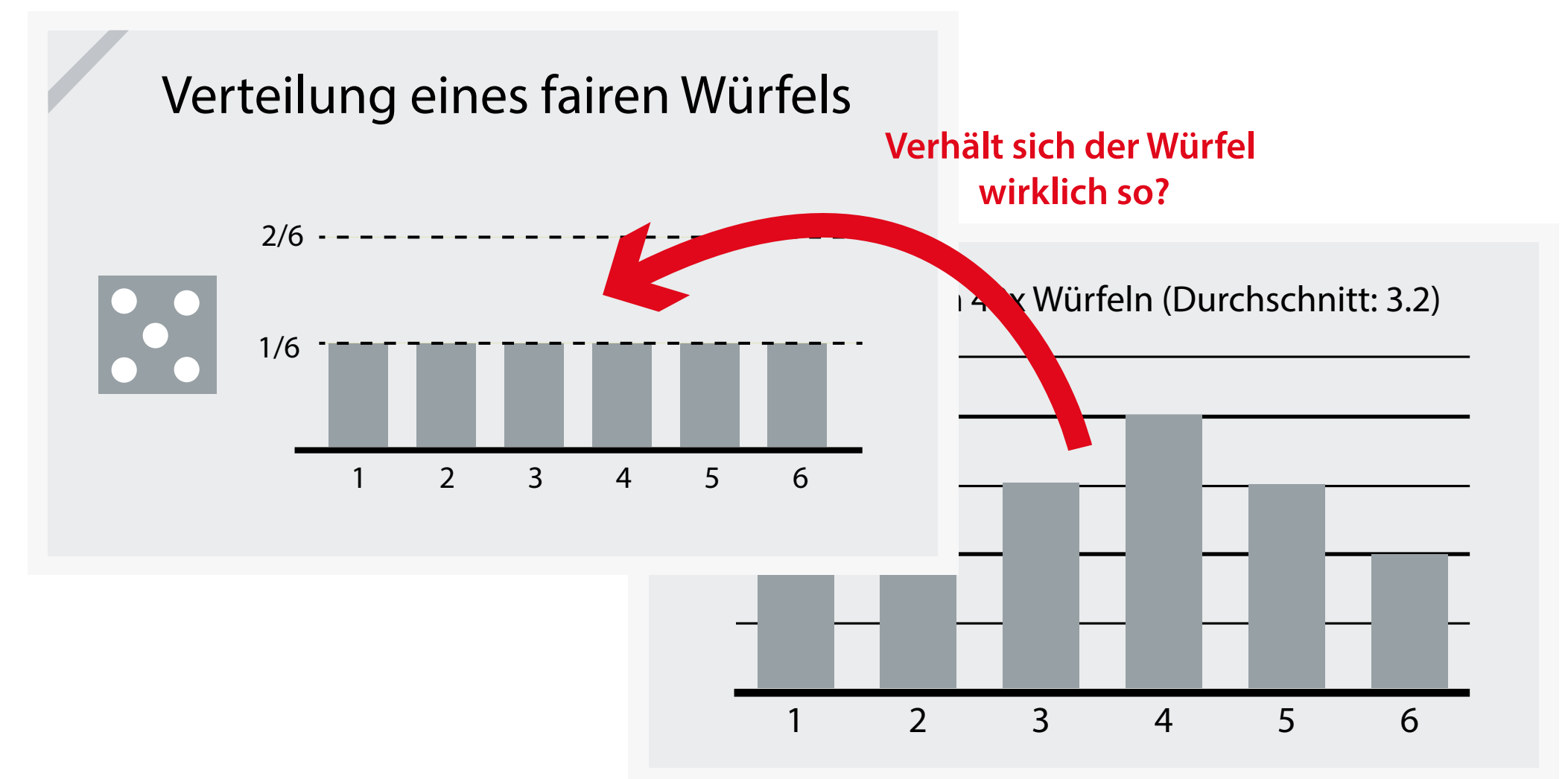
Z-Test

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.

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

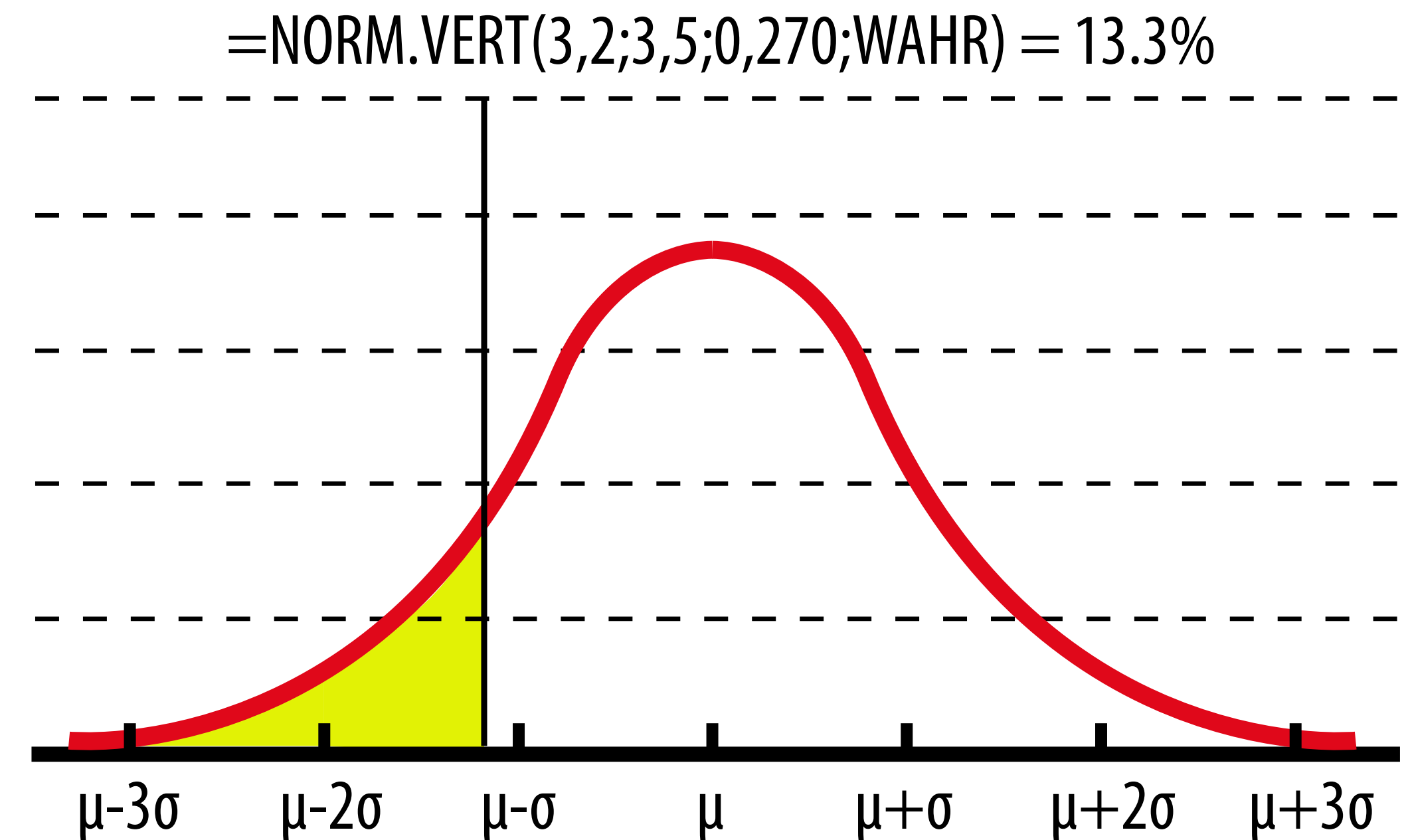


Z-Test

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.



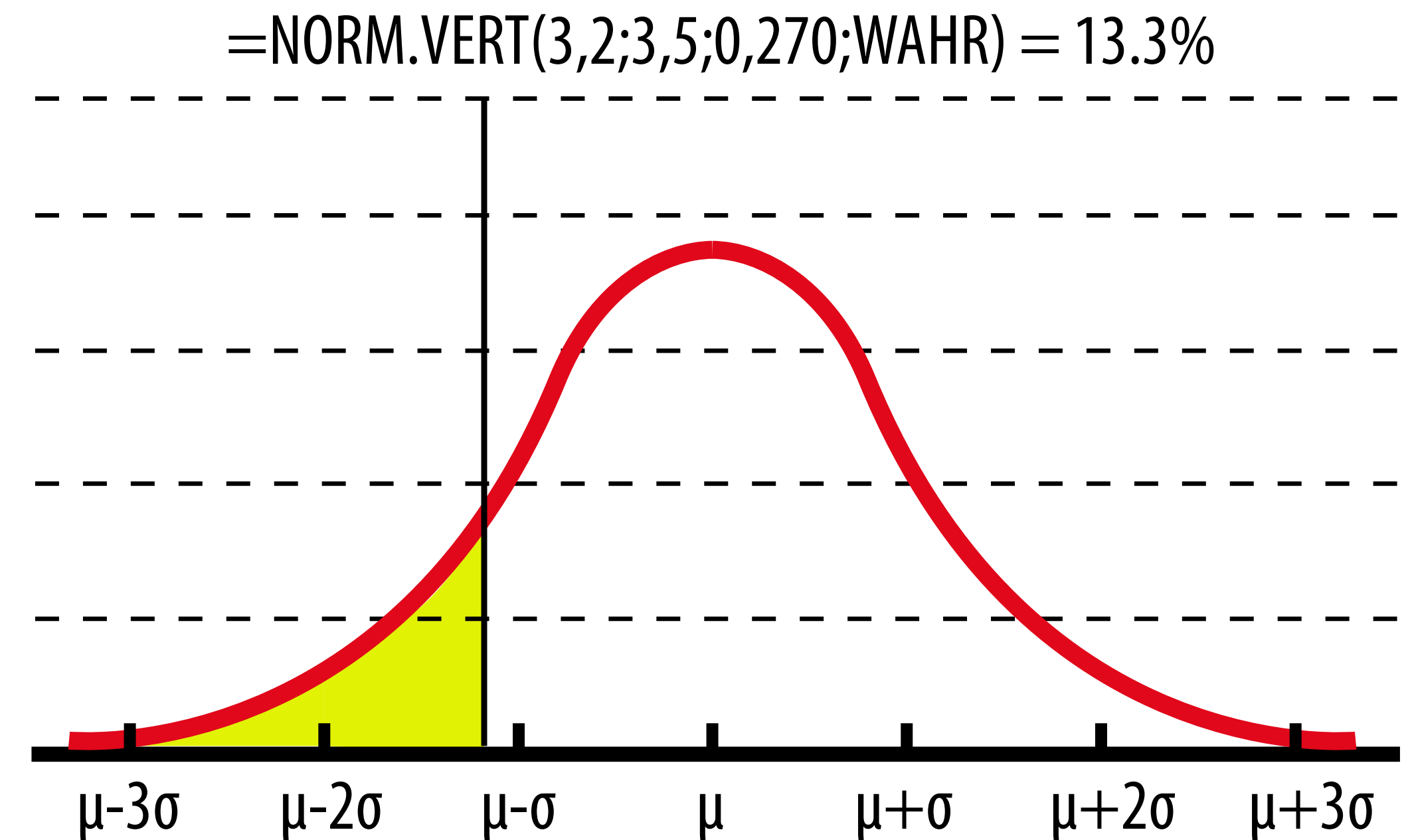
Z-Test

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.

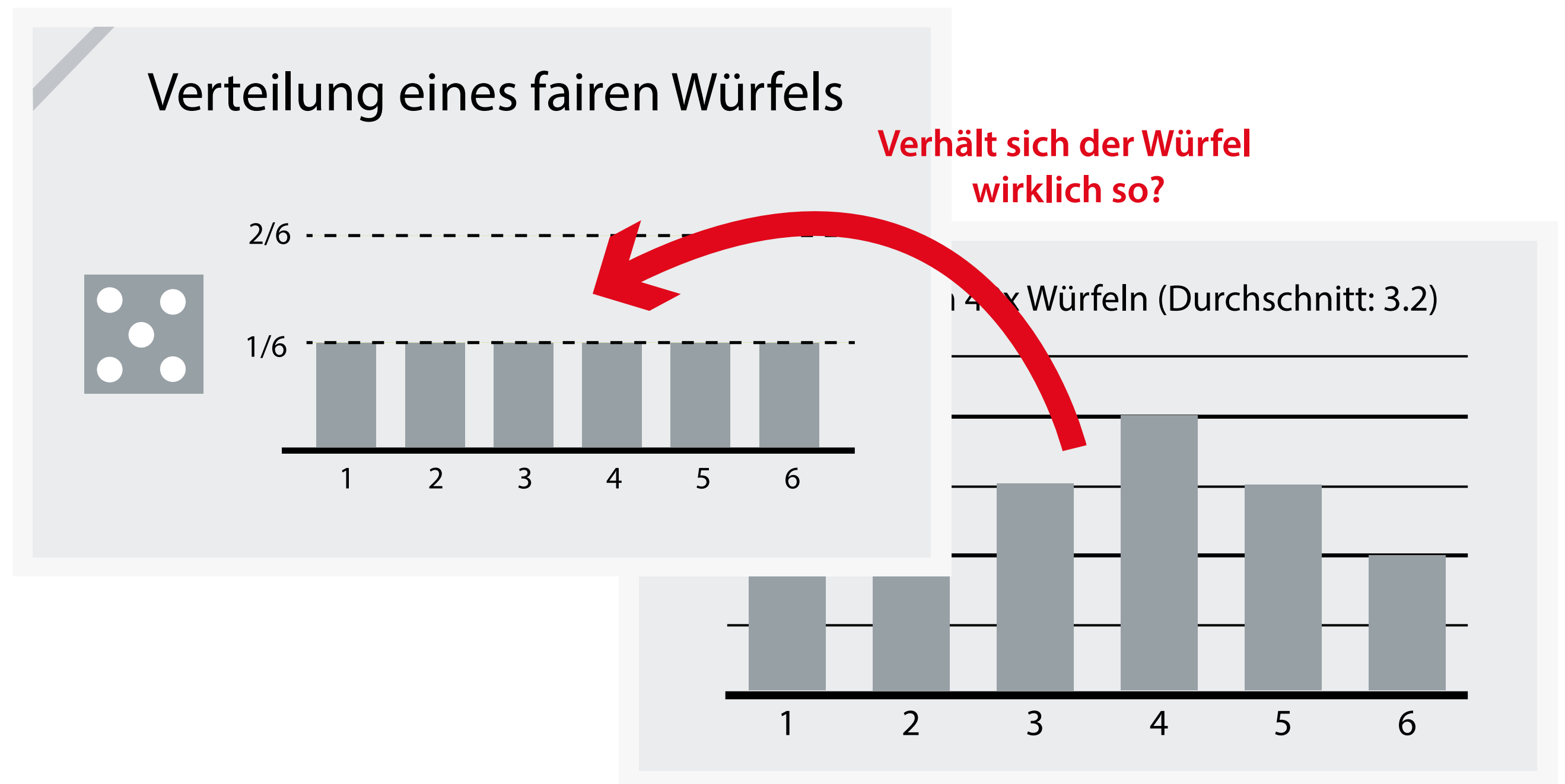


Einseitig vs. Zweiseitig

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!

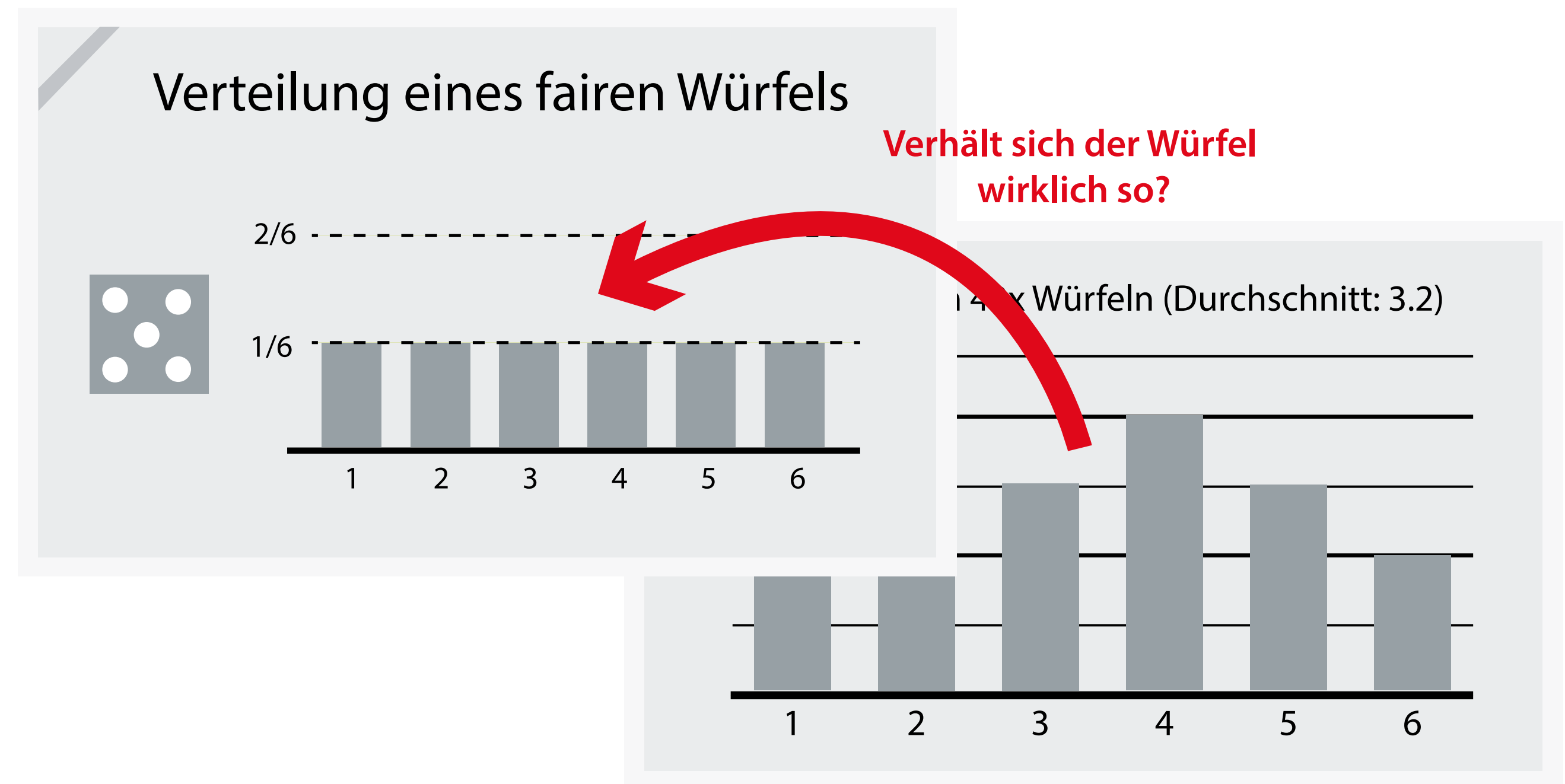


Einseitig vs. Zweiseitig

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.



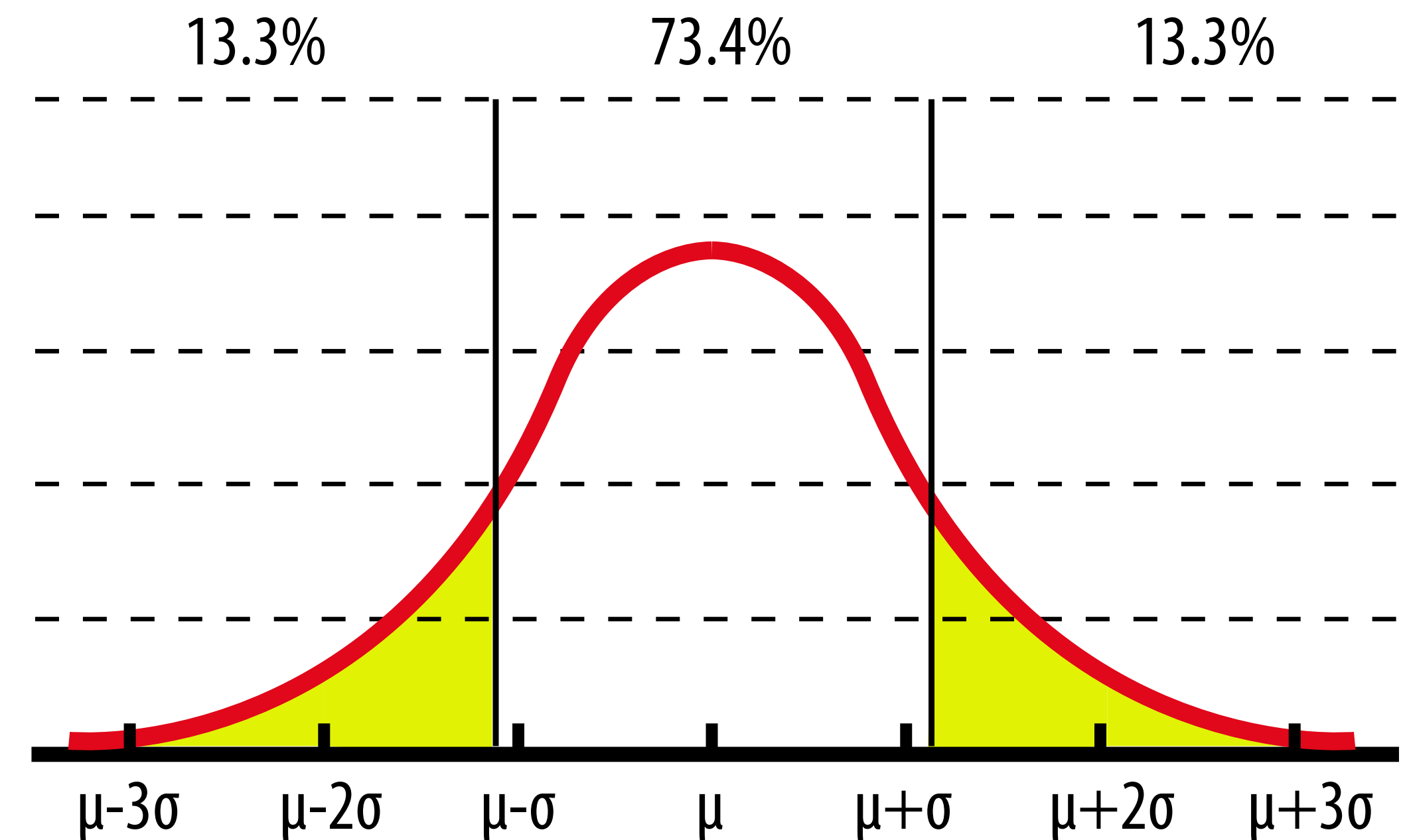
Einseitig vs. Zweiseitig

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.

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.



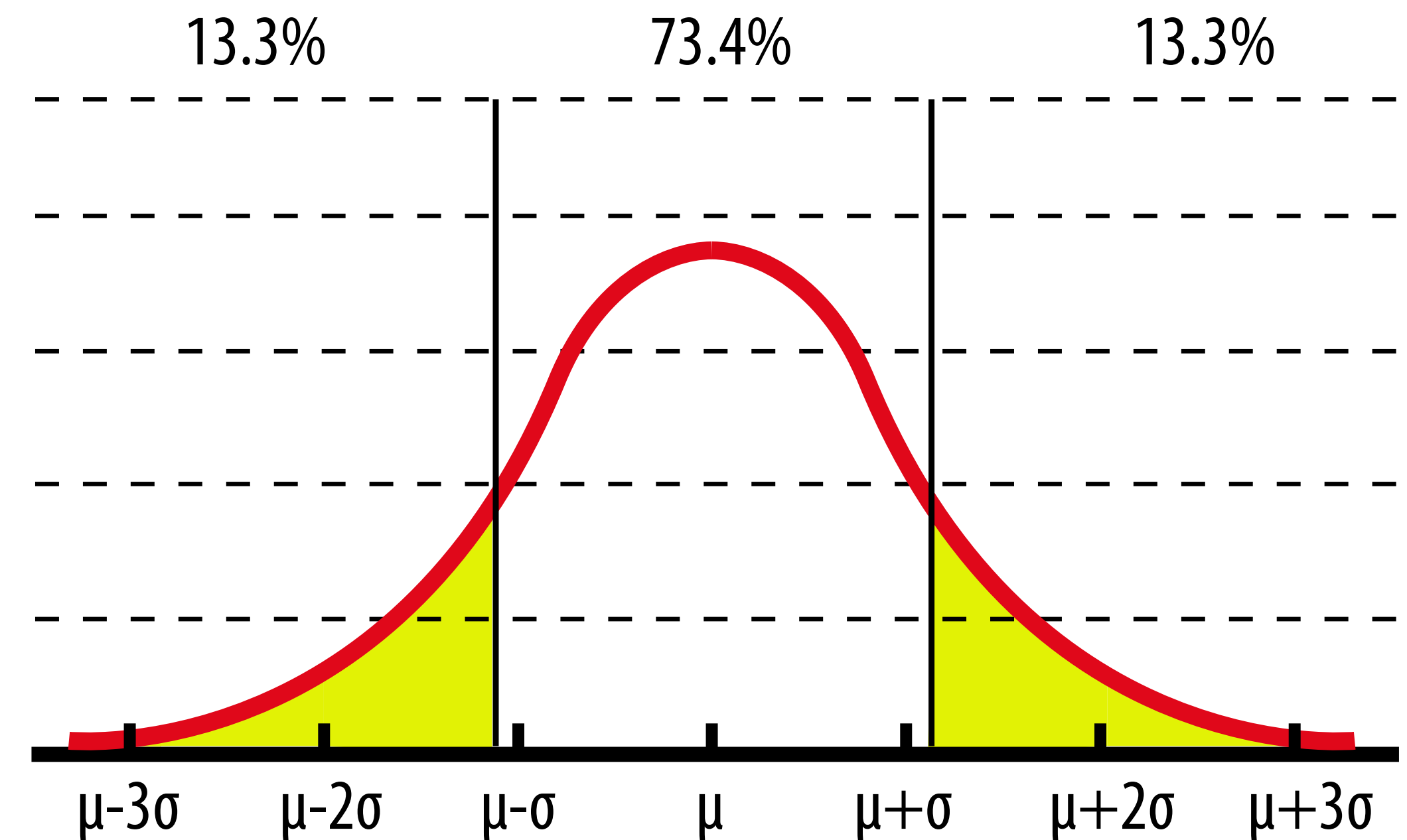
Einseitig vs. Zweiseitig

Die Abweichung vom Erwartungswert 3.5 ist in dieser zweiseitigen Betrachtung nochmals plausibler geworden.

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 Roulettetisch.

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:

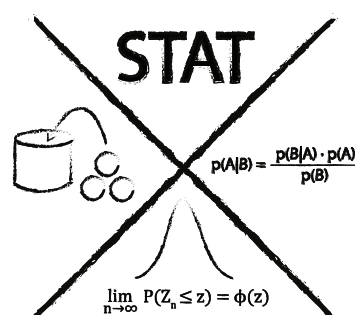
$$\sigma_X = 0.0171$$

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:

$$0.473$$

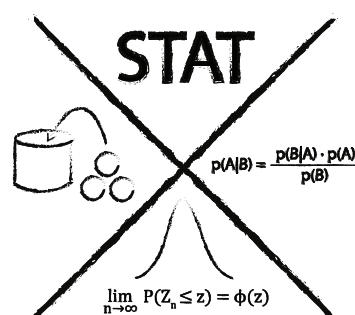
Um die Normalverteilung einsetzen zu können, benötigen wir noch die Varianz der ZVen X_i

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

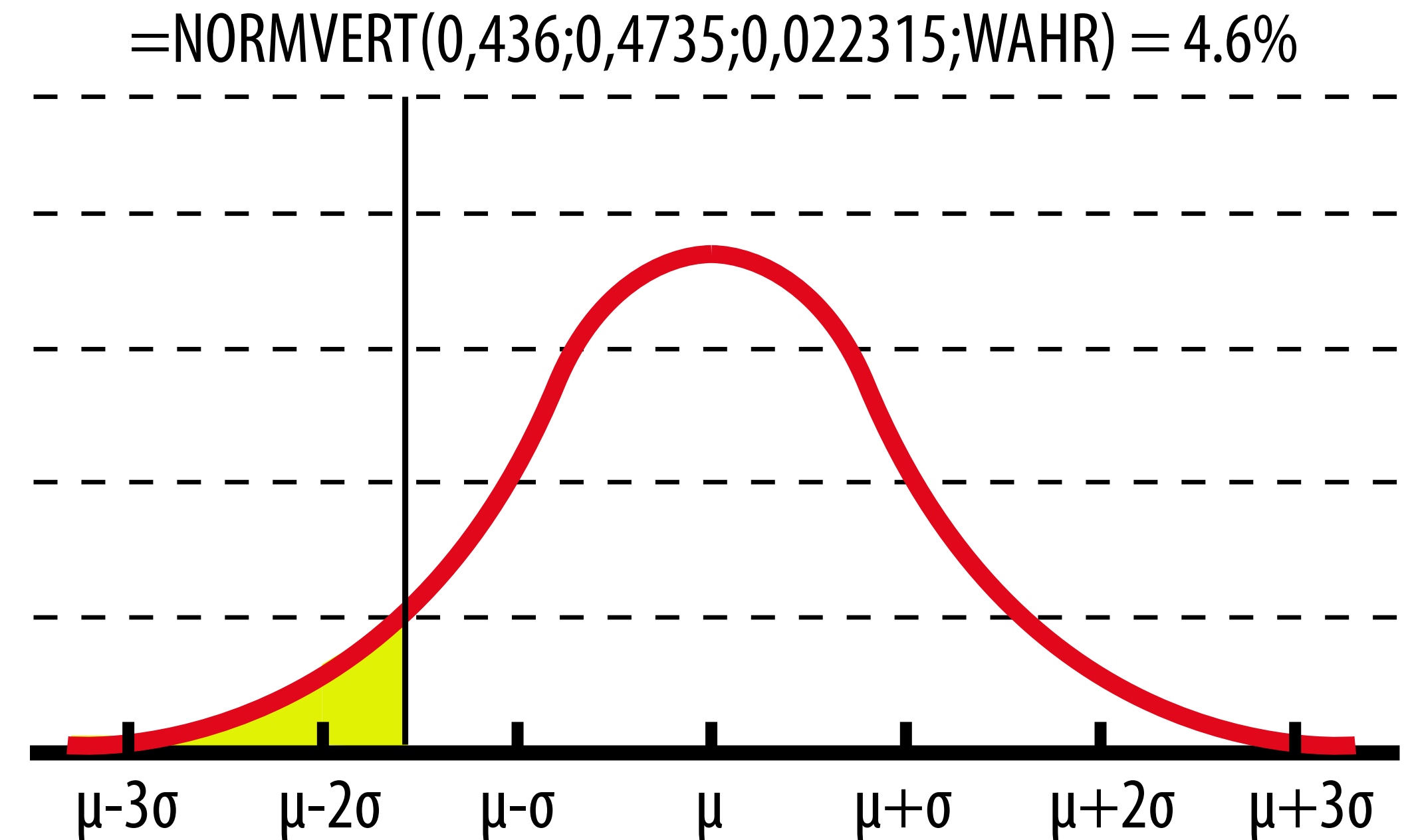
Bei einer Varianz von 0.249 in einem Spin hat der Durchschnitt über 500 Spins die Varianz:

$$0.249/500 = 0.000498$$

Die Standardabweichung ist also:

$$\sigma_x = 0.022315$$

Mit 0.436 sind wir 1.65 Standardabweichungen unter dem Erwartungswert von 0.473

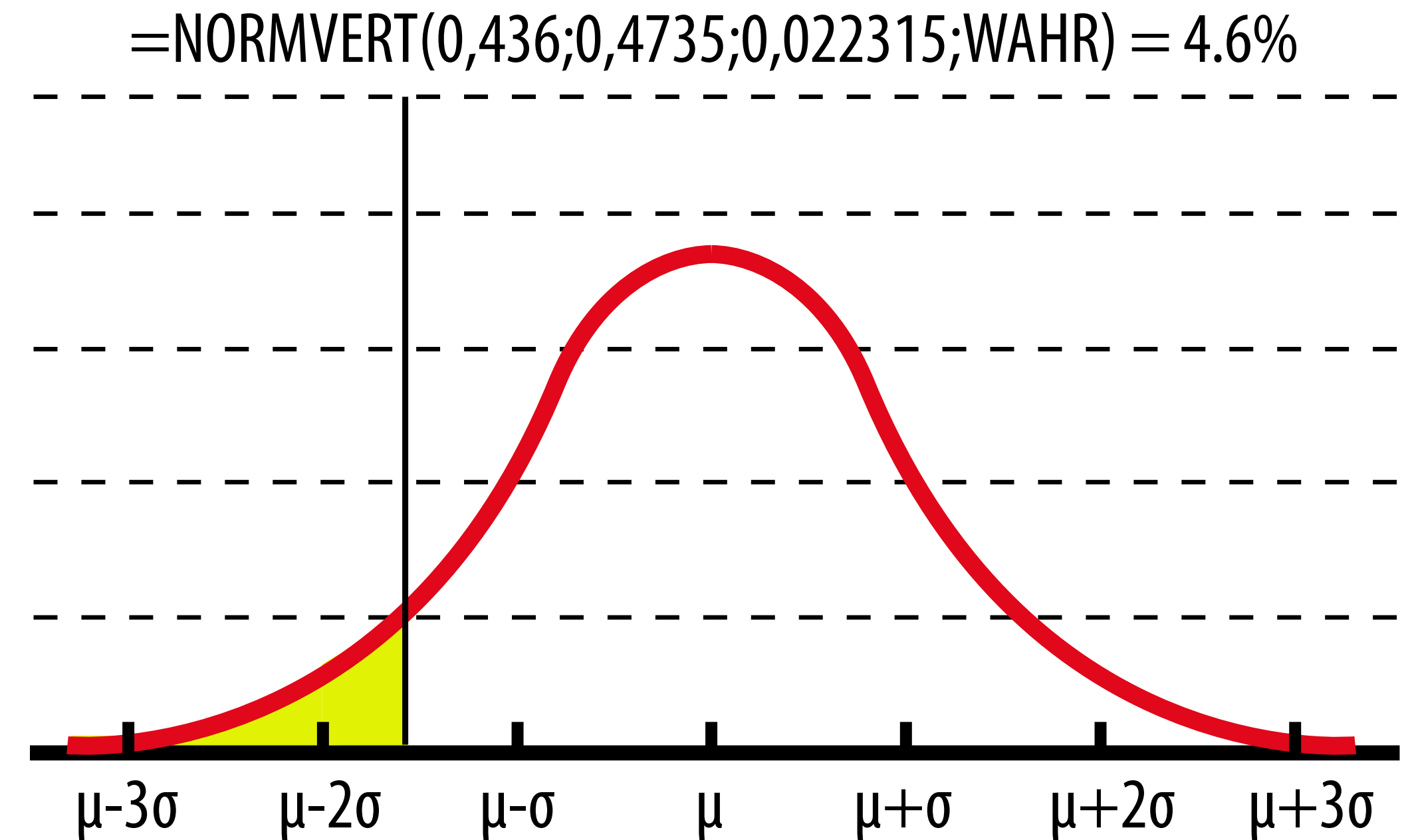


Z-Test

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.

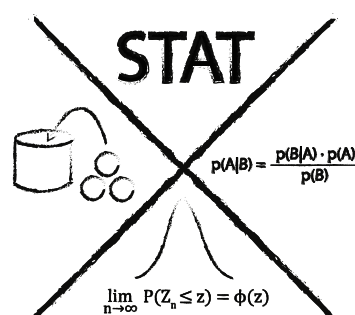
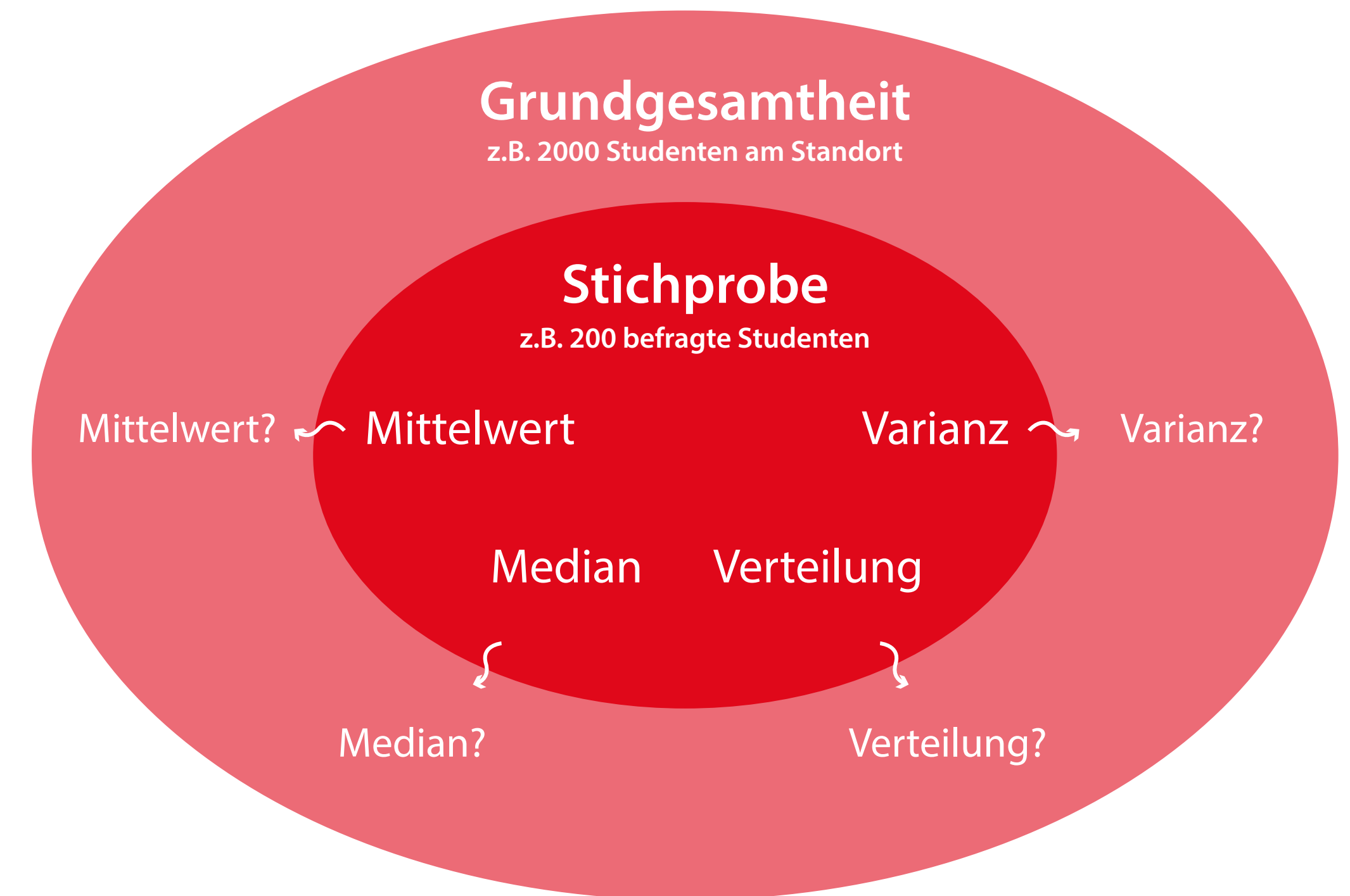


Statistische Tests

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 haben wir dieses Wissen oft nicht. Uns stehen nur Informationen über unsere Stichprobe zur Verfügung!



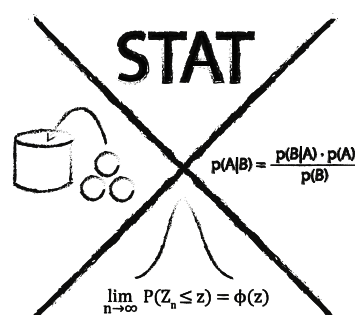
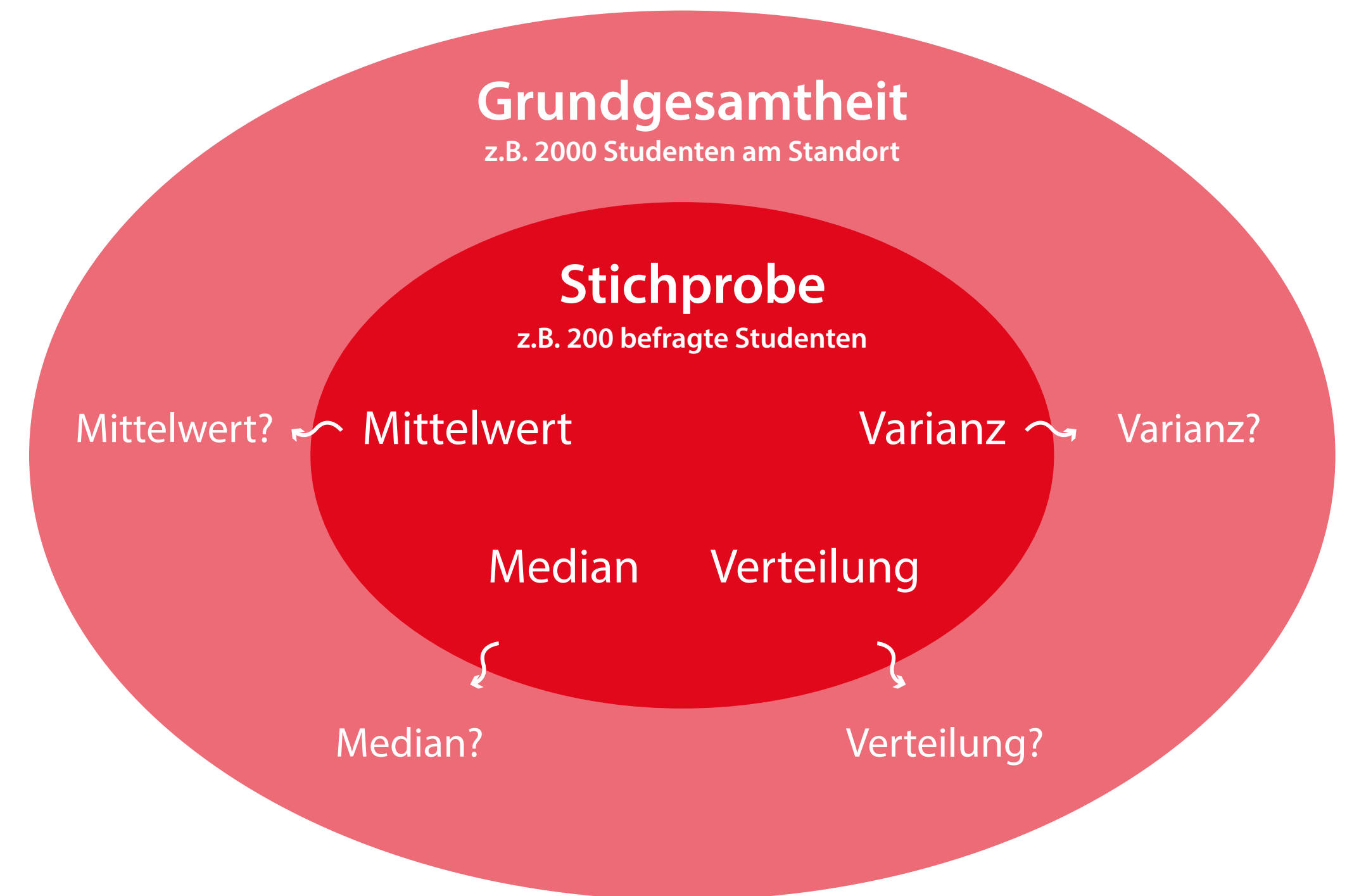
Statistische Tests

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

Verteilungshypothesen

Parameterhypothesen

Abhängigkeitshypothesen

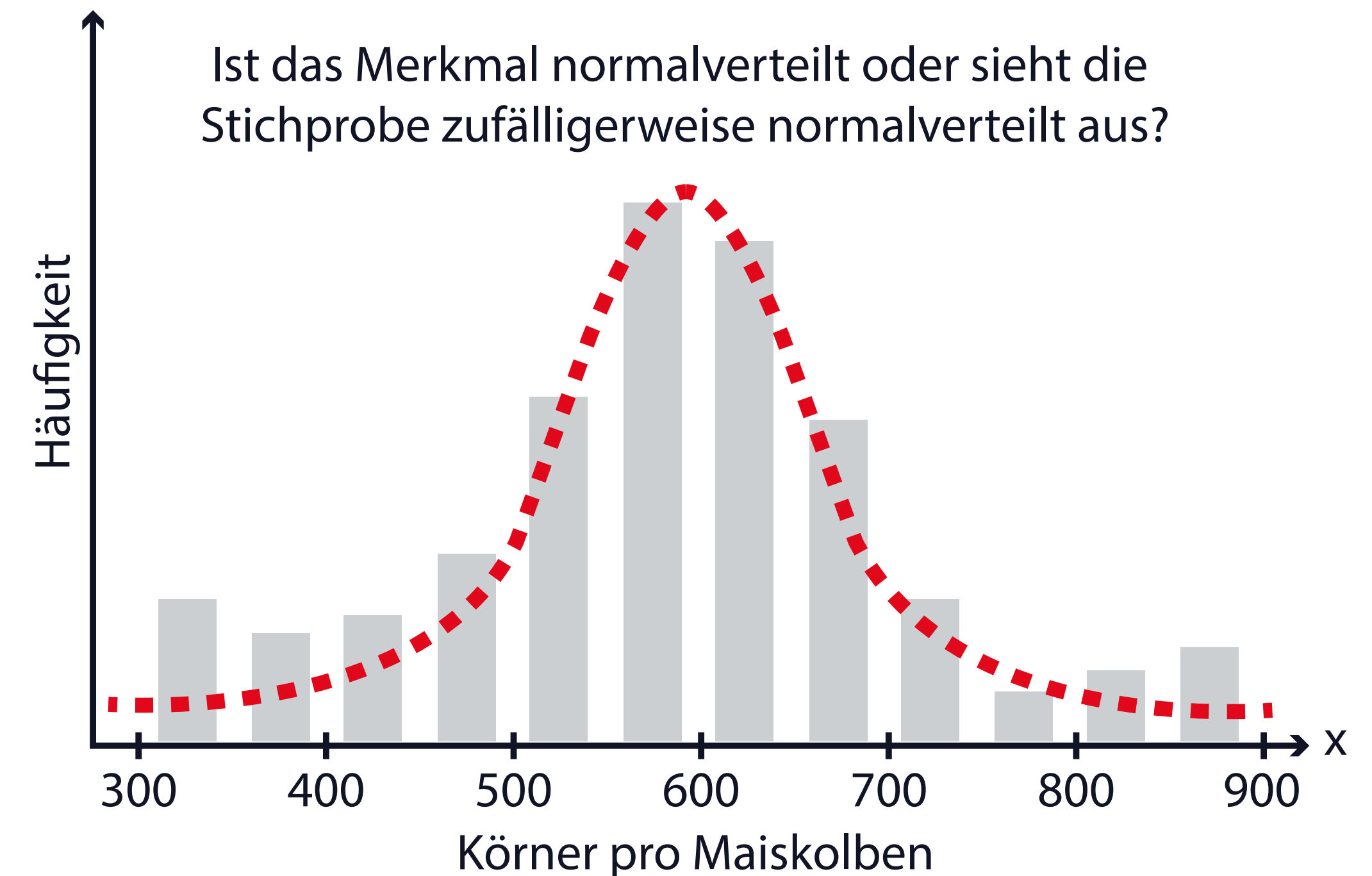


Statistische Tests

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.

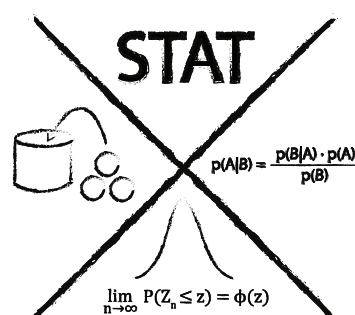
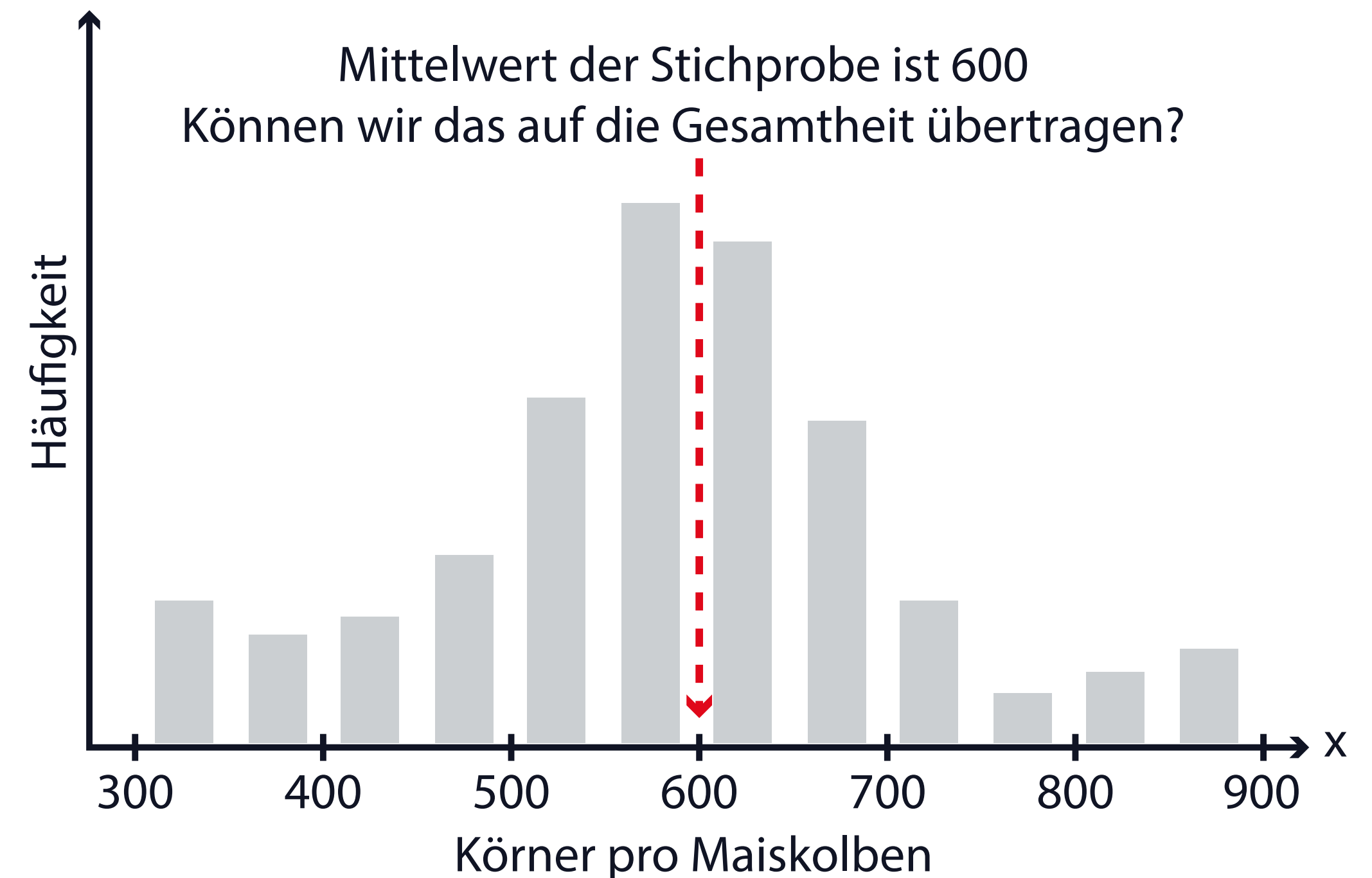


Statistische Tests

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.

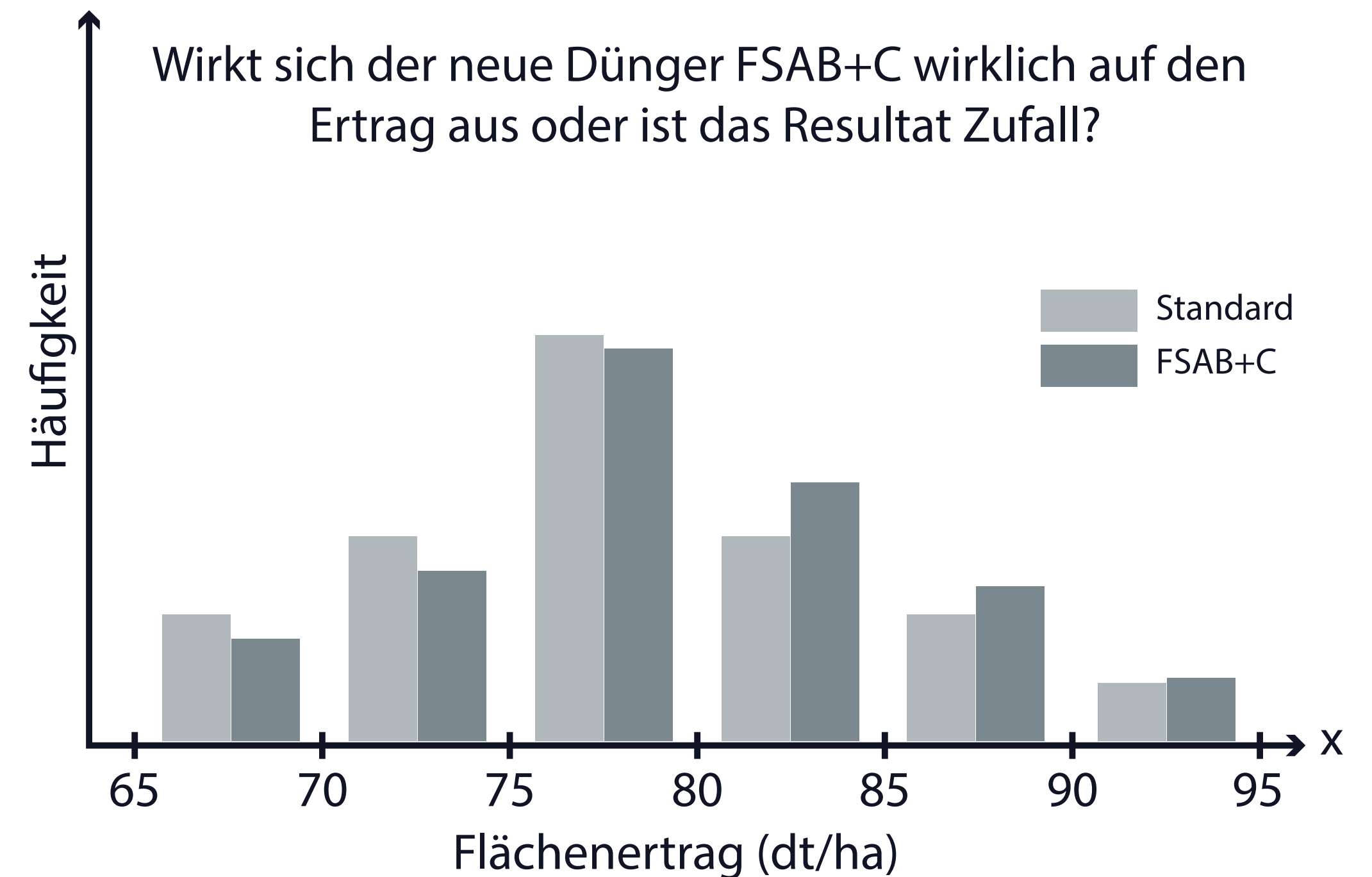
Mit einem **Parametertest** können wir diese Hypothesen überprüfen.



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?

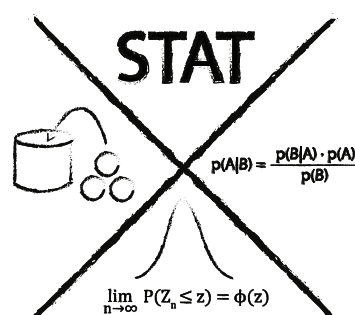
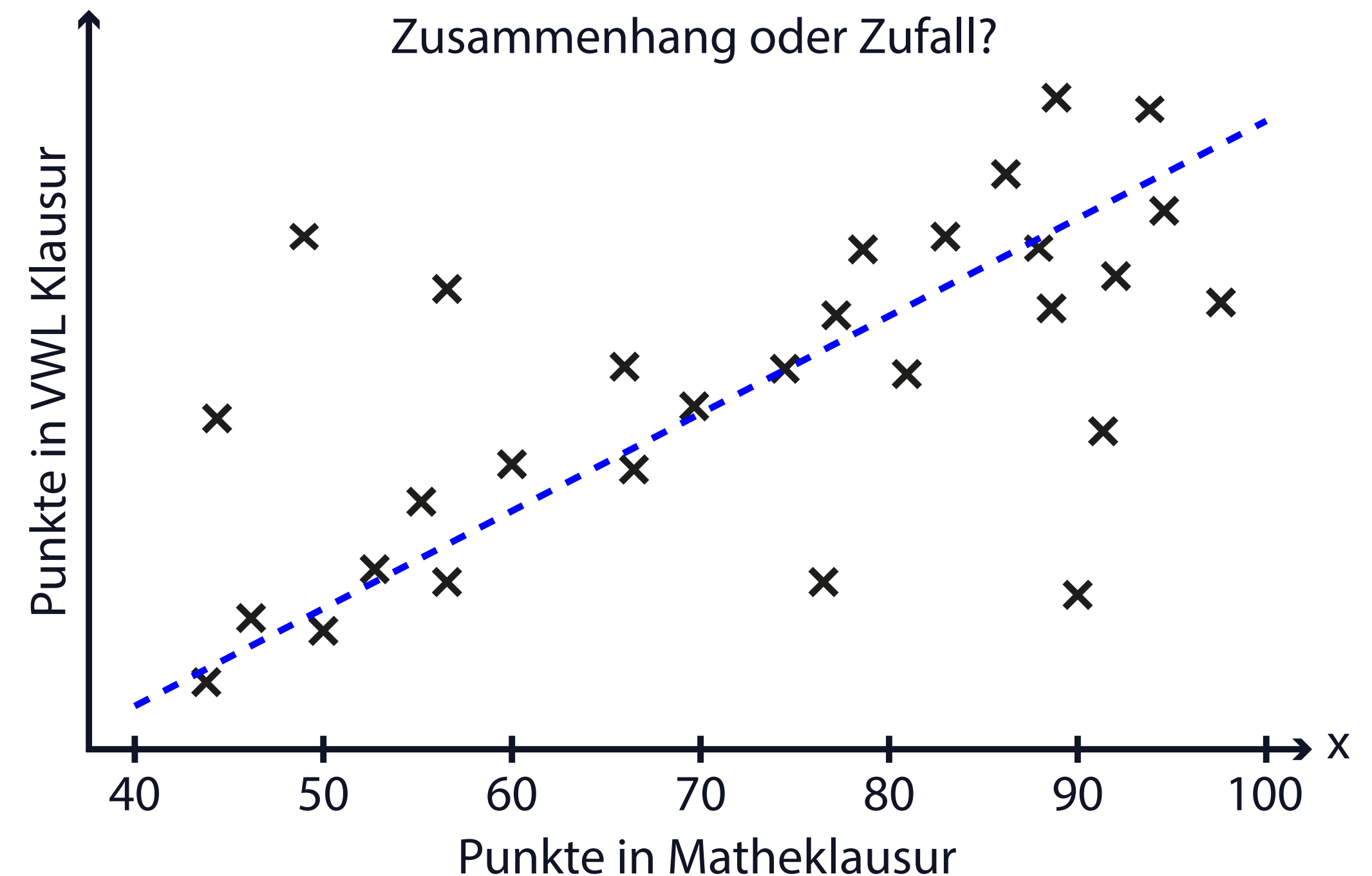


Statistische Tests

Unabhängigkeitshypothesen nehmen an, dass zwei Merkmale in der Grundgesamtheit voneinander unabhängig sind.

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

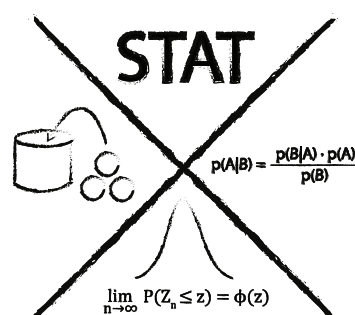
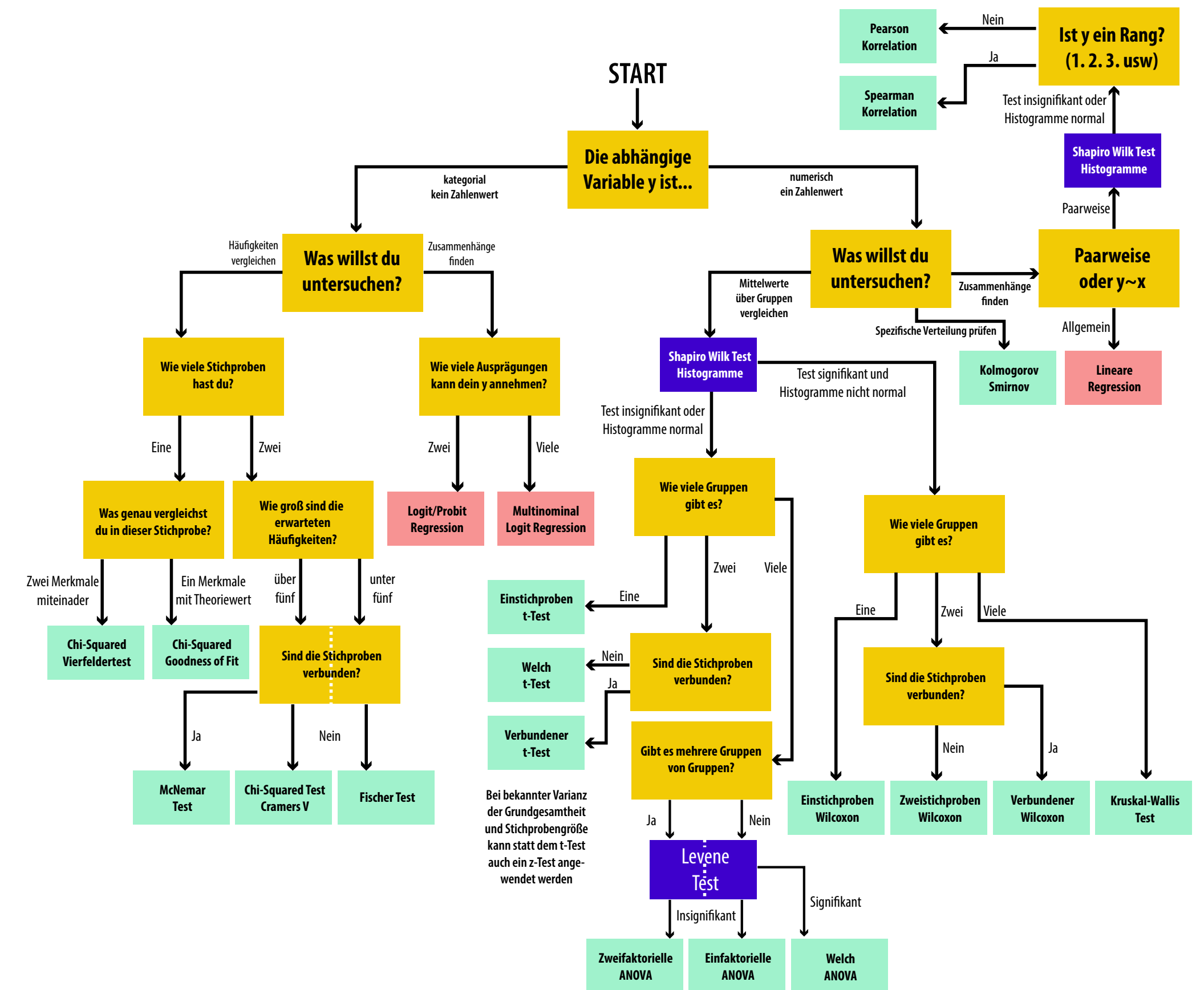
Können wir aus unserer Stichprobe auf eine Unabhängigkeit oder einen Zusammenhang schließen?



Es gibt duzende verschiedene statistische Tests!

Um diese 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?



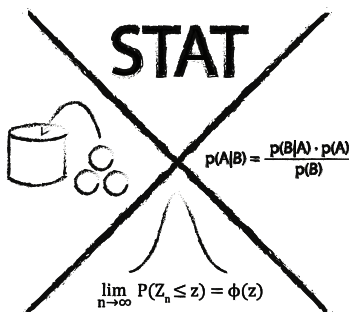
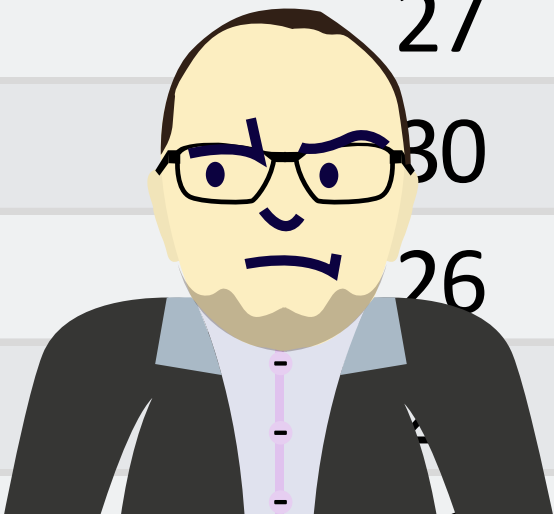
t-Test

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.

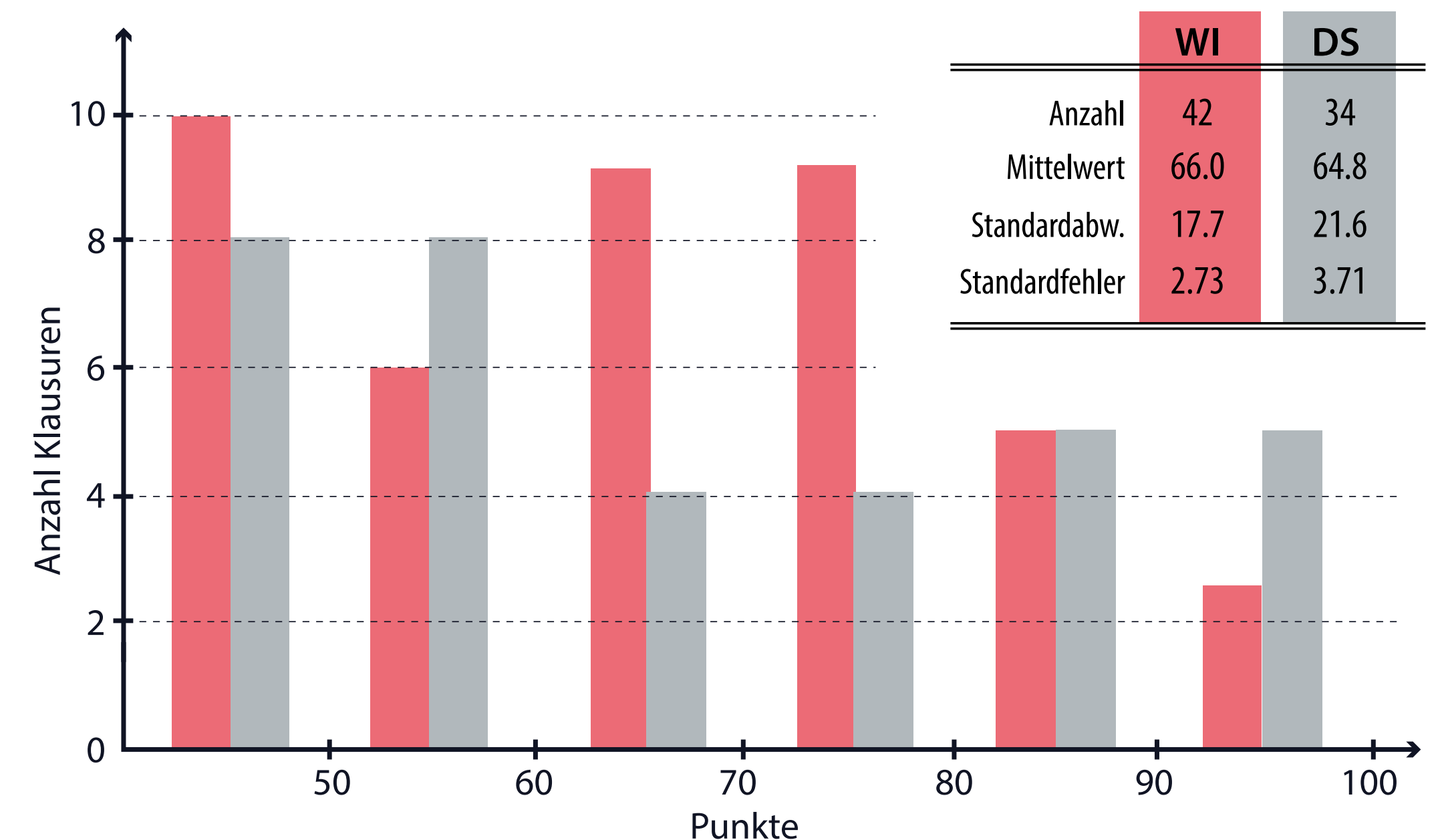
Nummer	Kurs	Marketing	Analysis
1	WI	25	19
2	WI	21	24
3	WI	33	15
4	WI	31	22
5	WI	30	23
6	WI	35	20
7	WI	33	22
8	WI	36	21
9	Data Science	31	27
10	Data Science	32	30
11	Data Science	37	26
12	Data Science	38	



t-Test

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

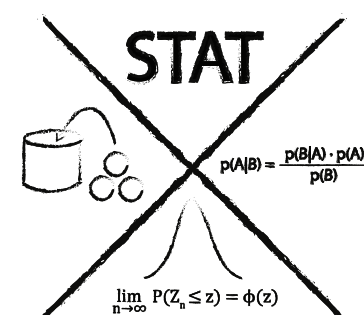
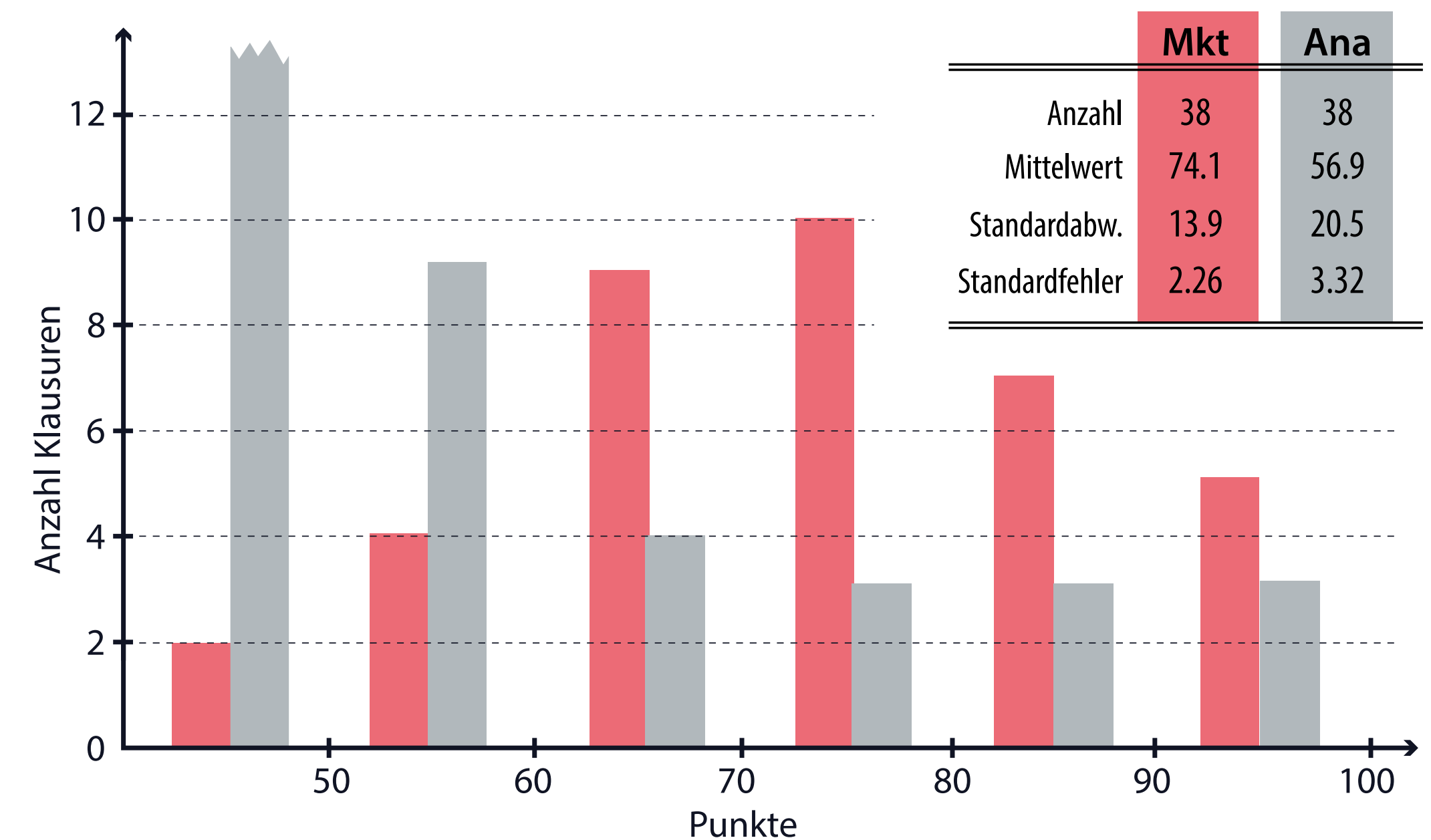
In der Stichprobe sind die Wirtschaftsinformatiker um 1.2 Punkte besser, aber lässt sich das verallgemeinern?



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?



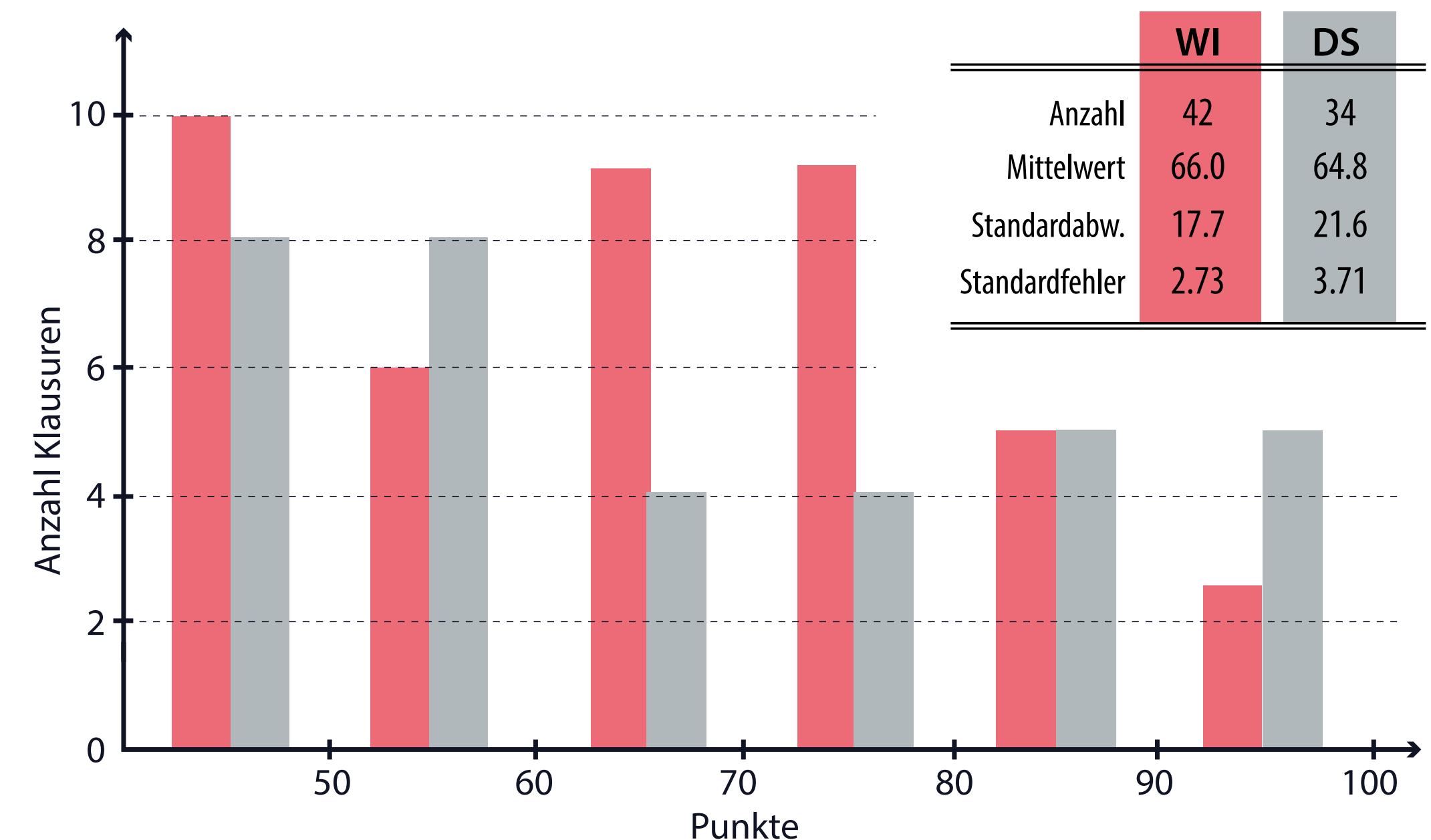
t-Test

Beim Vergleich der Studiengänge haben wir in der Stichprobe ein um 1.12 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 DS-Student sind unterschiedliche Personen!

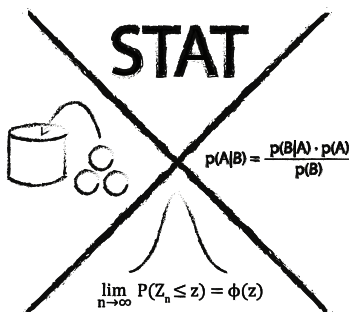
WI	Zusammenhang?	Data Science
50	 X 	58
42	 X 	58
66	 X 	54
62	 X 	60
60	 X 	66
70	 X 	62
66	 X 	74
72	 X 	68

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 1.12 Punkten.

Vergleich der Studiengänge					
Mittelwert WI	66,0				
Mittelwert DS	64,9				
Abweichung	1,12				
H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Studiengänge.					
H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Studiengänge.					
p-Wert	0,809073208				
Nullhypothese wird beibehalten					



Interpretation p-Wert

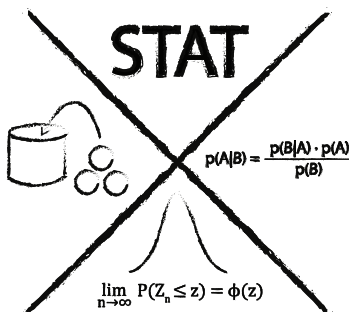
Selbst ohne Unterschiede im Leistungsniveau ist es sehr wahrscheinlich einen kleinen Unterschied zu messen!

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.

Vergleich der Studiengänge					
Mittelwert WI	66,0				
Mittelwert DS	64,9				
Abweichung	1,12				
H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Studiengänge.					
H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Studiengänge.					
p-Wert	0,809073208				
Nullhypothese wird beibehalten					



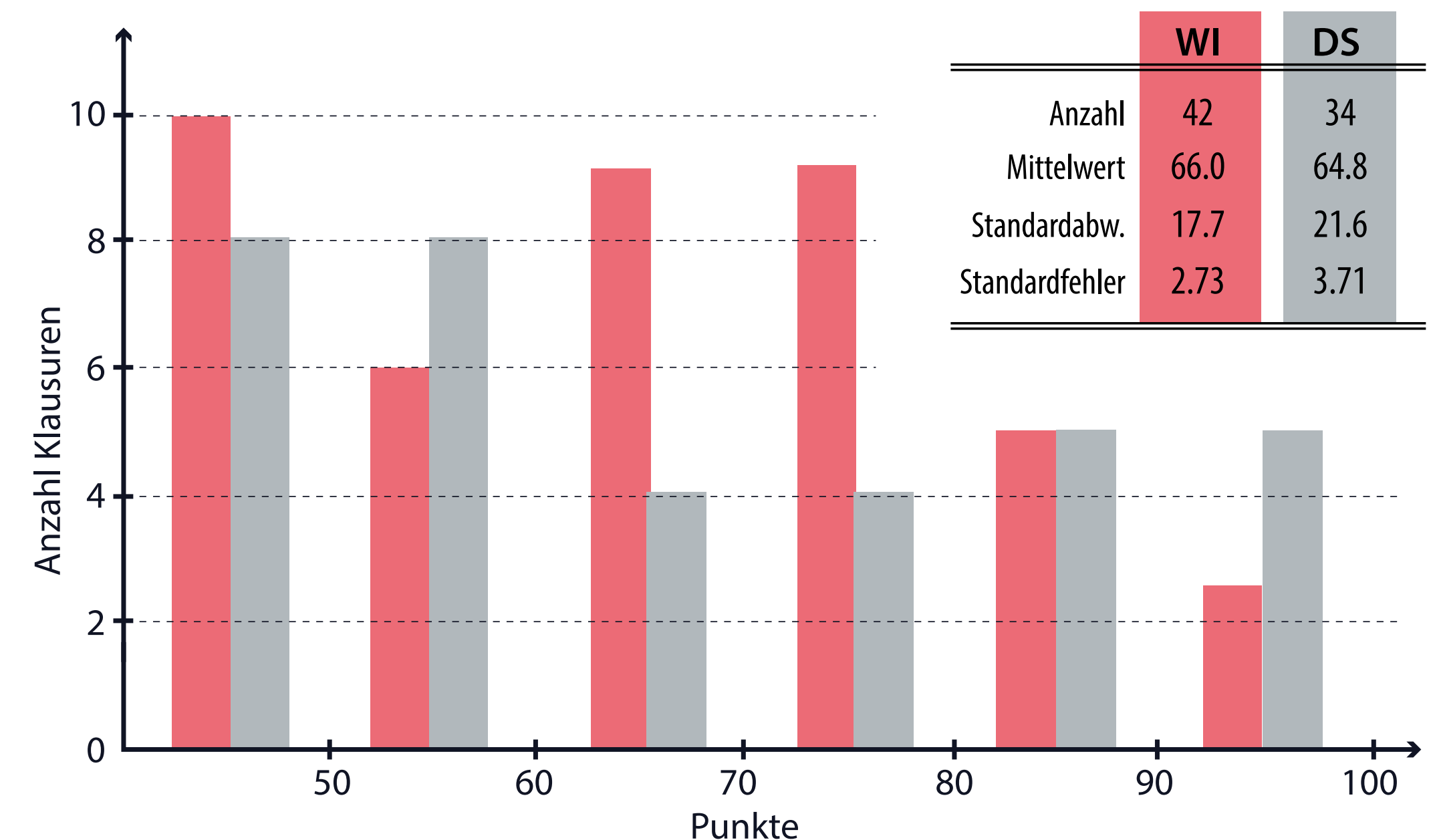
Interpretation p-Wert

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!

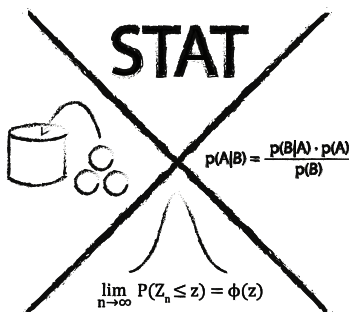
Marketing	Zusammenhang?	Analysis
50		38
42		48
66		30
62		44
60		46
70		40
66		44
72		42

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.

Vergleich der Vorlesungen					
Mittelwert MKT	74,1				
Mittelwert ANA	56,9				
Abweichung	17,21				
H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Vorlesungen.					
H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Vorlesungen.					
p-Wert	2,33513E-09	Gepaarte Stichprobe!			
Nullhypothese wird verworfen					



Interpretation p-Wert

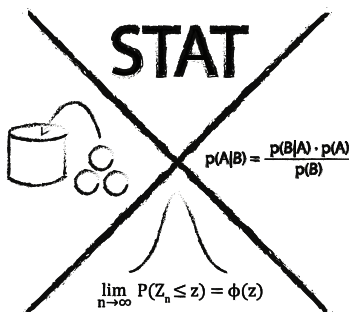
Wir können es nicht ausschließen, aber dass dieser Unterschied durch zufällige Schwankungen zustande kommt, ist extrem unwahrscheinlich.

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.

Vergleich der Vorlesungen					
Mittelwert MKT	74,1				
Mittelwert ANA	56,9				
Abweichung	17,21				
H0 - Es gibt keinen Unterschied im Leistungsniveau der beiden Vorlesungen.					
H1 - Es gibt einen Unterschied im Leistungsniveau der beiden Vorlesungen.					
p-Wert	2,33513E-09		Gepaarte Stichprobe!		
Nullhypothese wird verworfen					



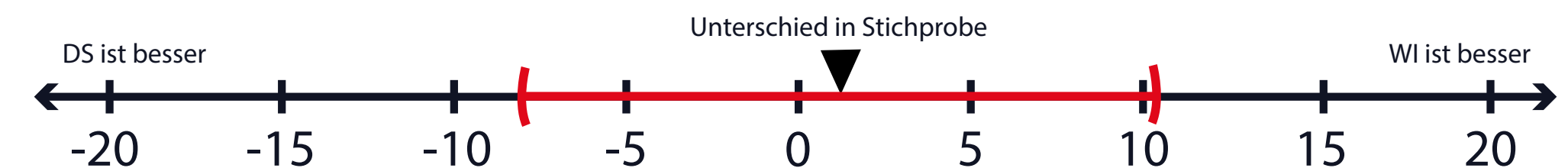
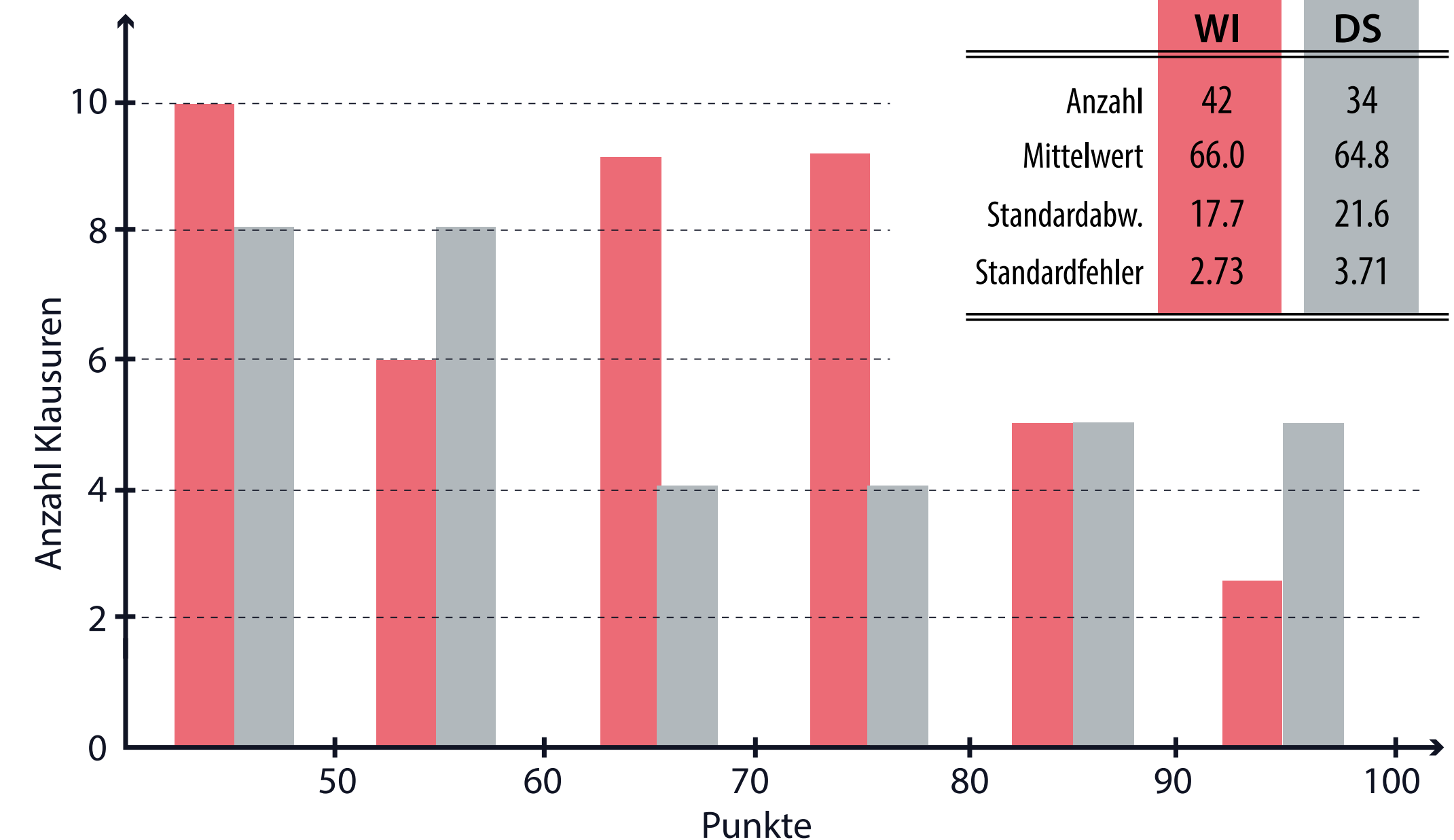
Konfidenzintervalle

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

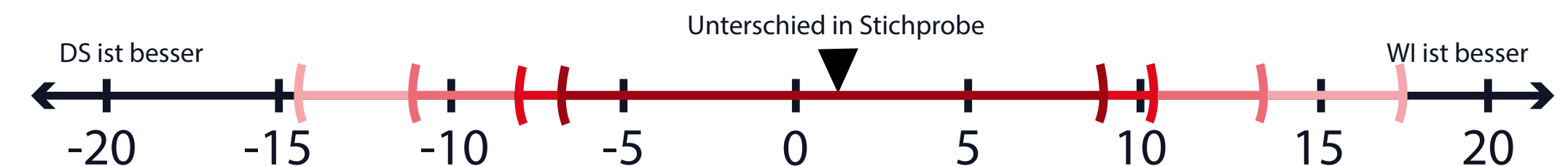
Konfidenzintervalle

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

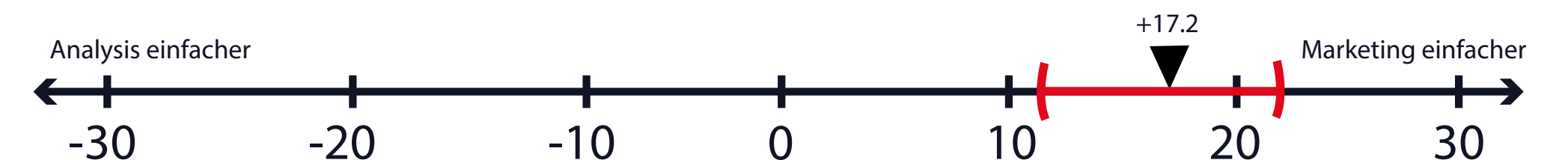
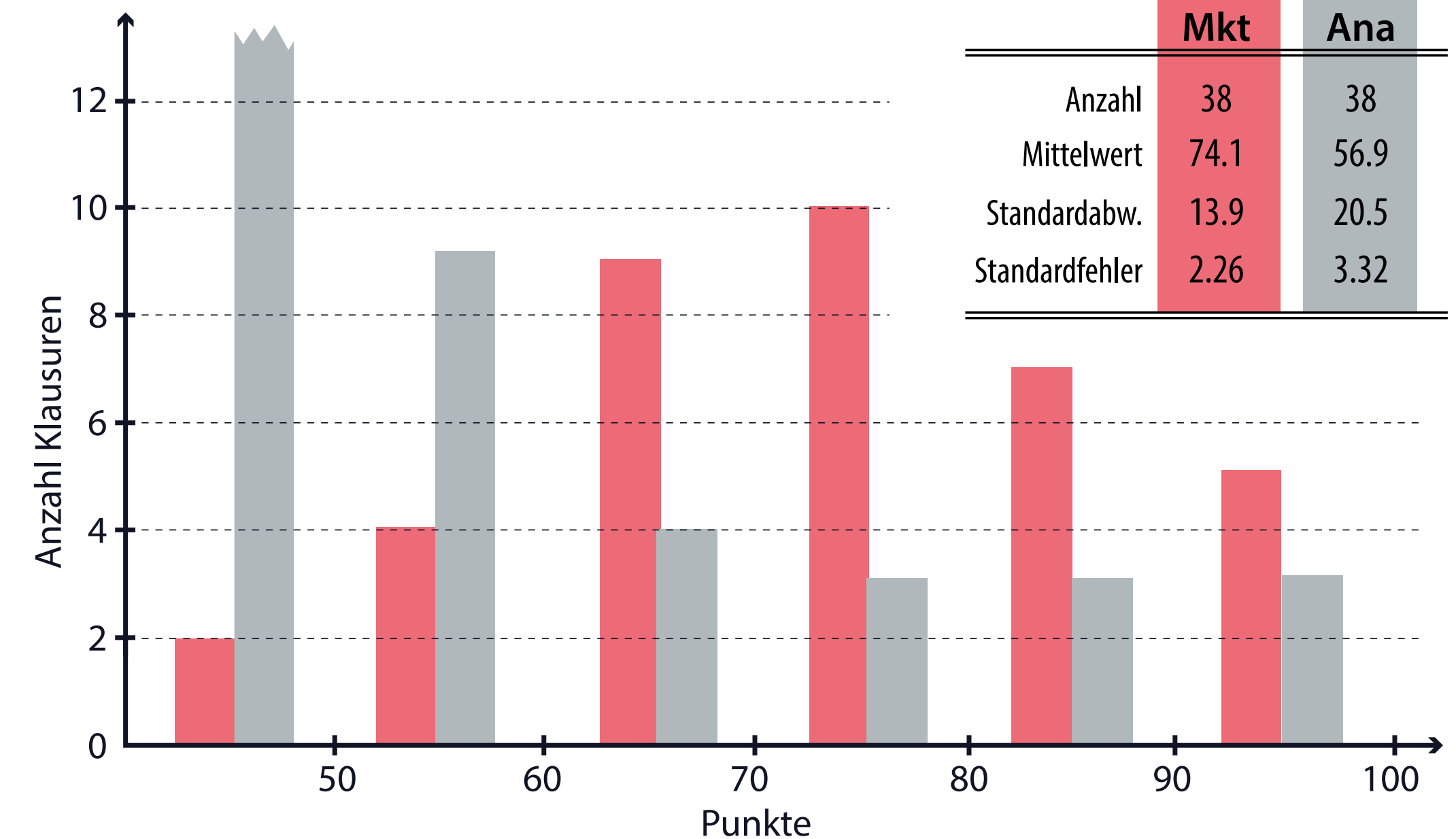
Zu 99.9% liegt der wahre Unterschied zwischen -14.7 und 17.01

Konfidenzintervalle

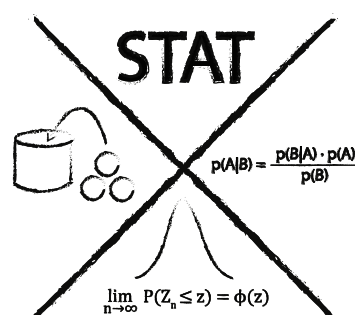
Beim Vergleich der Vorlesungen ist die 0 nicht im 95% Konfidenzintervall enthalten.

[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



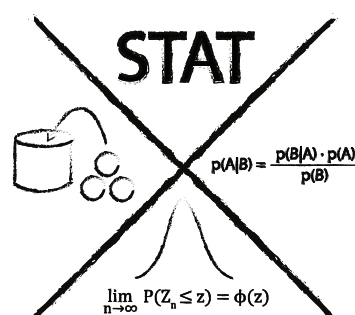
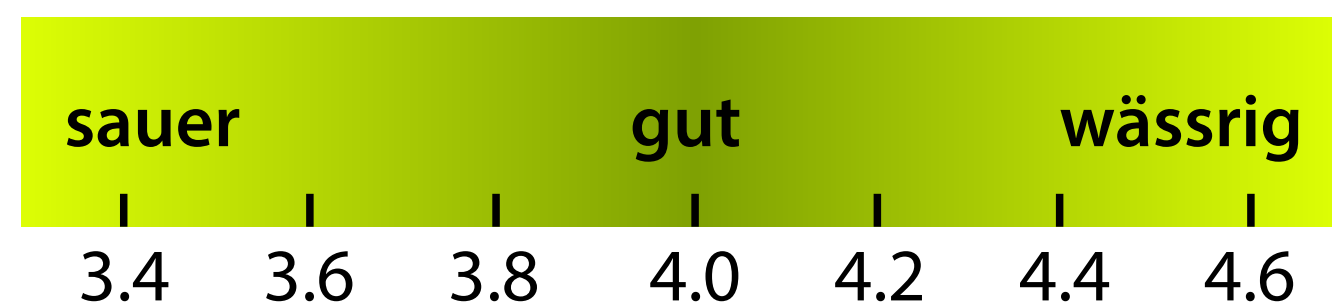
Einstichproben t-Test

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.

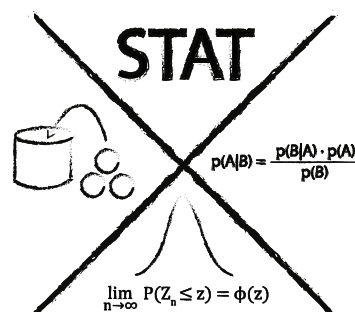
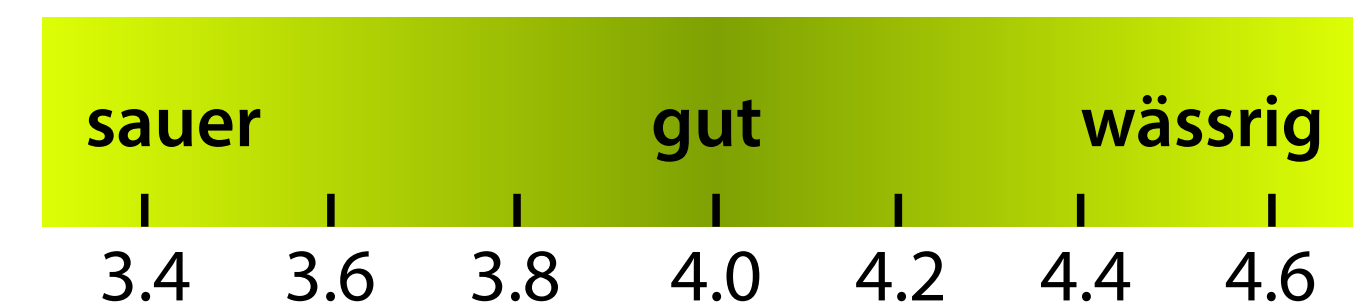


Einstichproben t-Test

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!

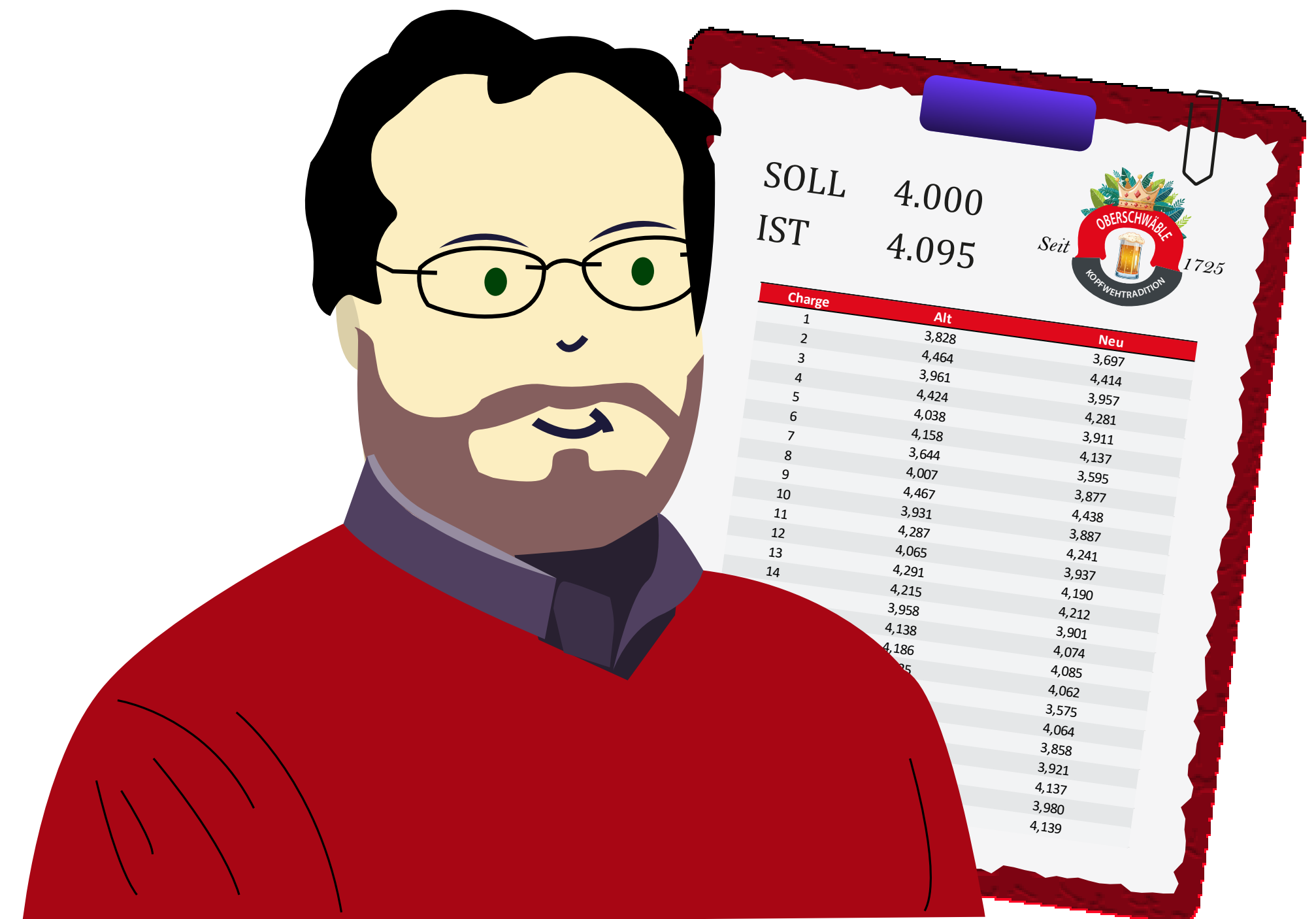


Einstichproben t-Test

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!

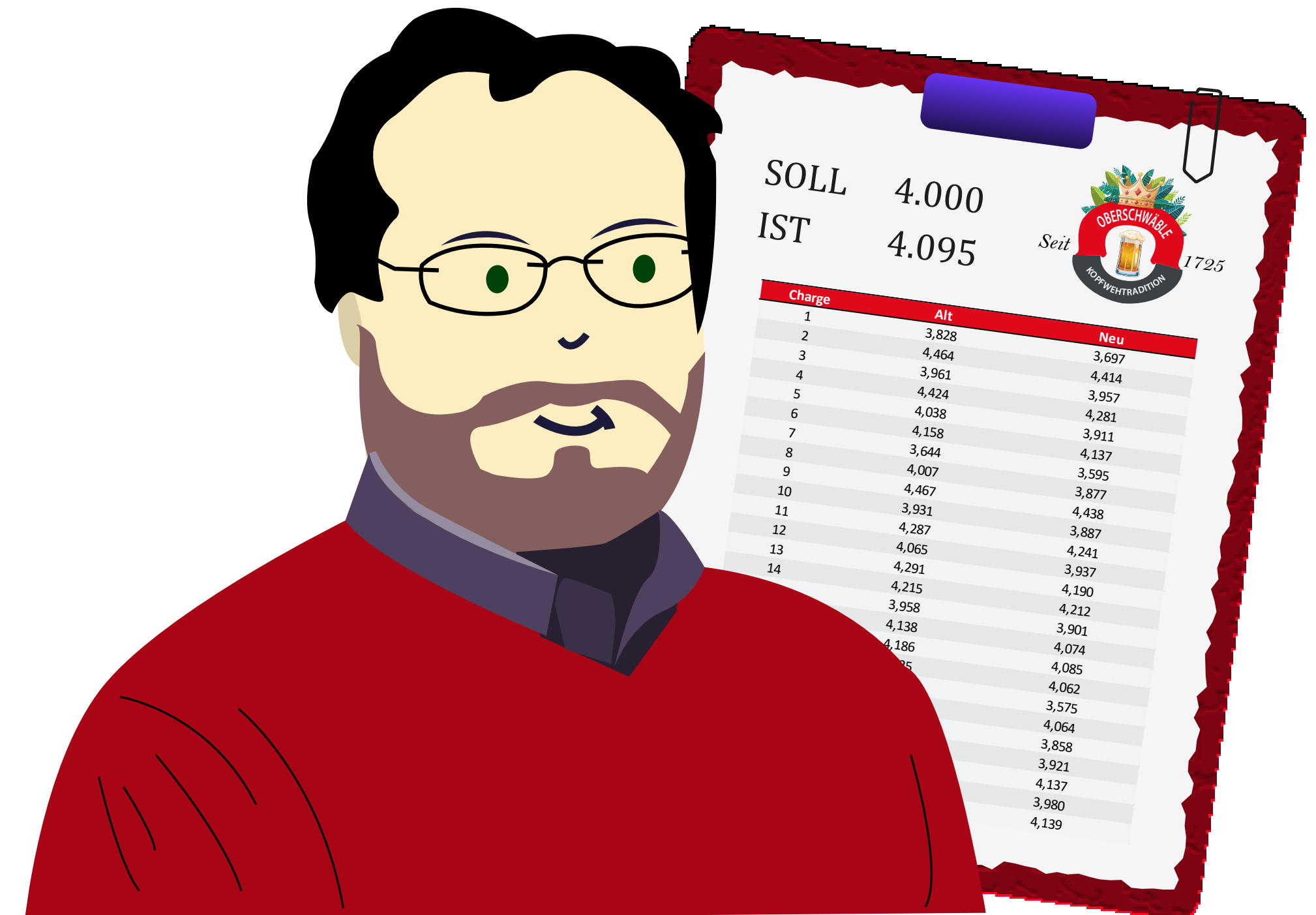


Einstichproben t-Test

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.



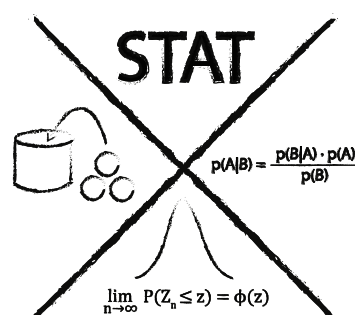
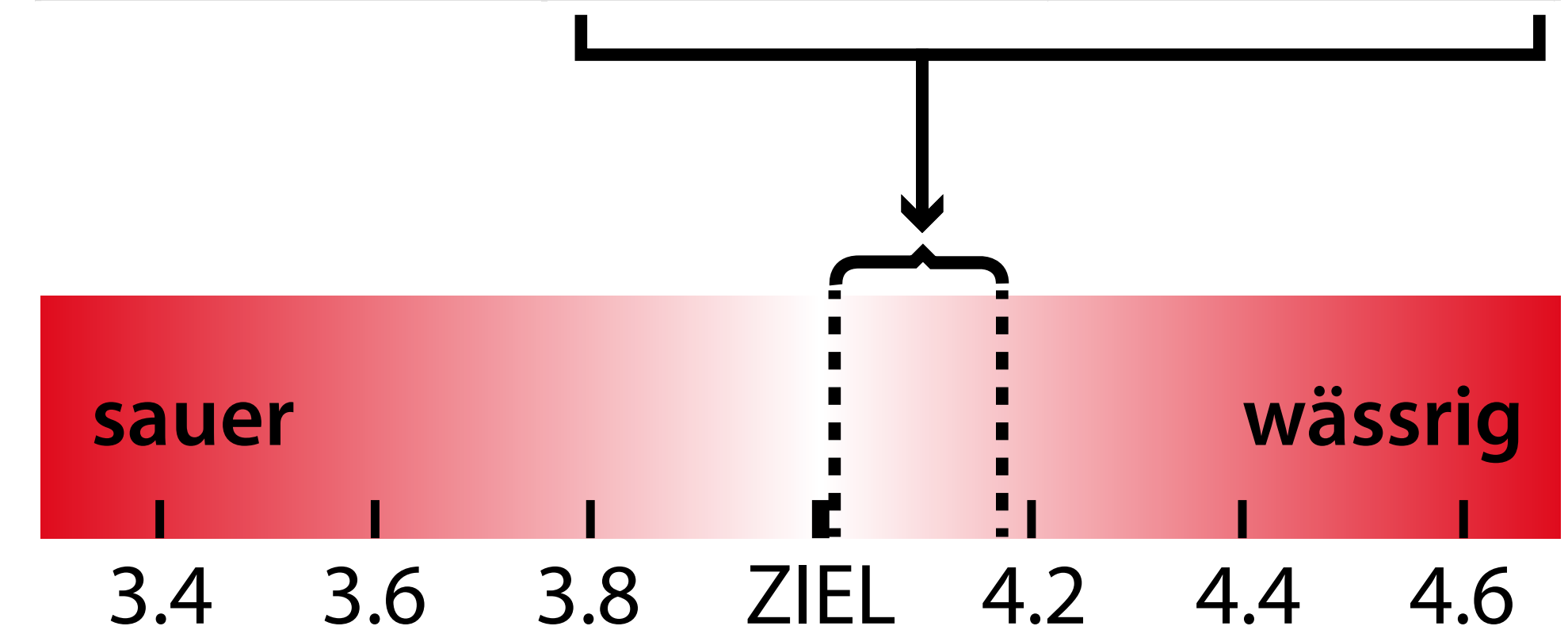
Einstichproben t-Test

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

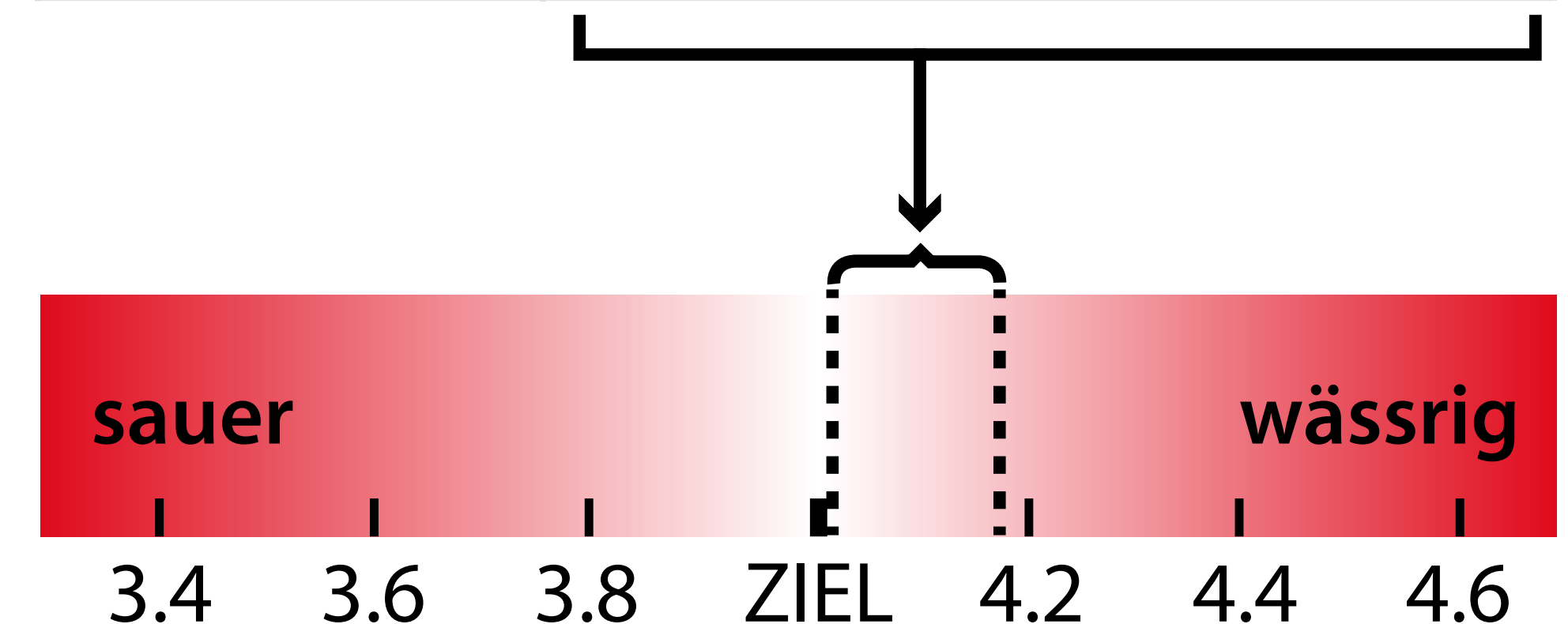
Das 95% Konfidenzintervall beträgt:

[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



t-Test

Die Verbraucherschutzzentrale untersucht den Zucker-
gehalt von 20 Flaschen Kako-Calo. Auf dem Etikett wird
ein Zuckergehalt von 106g pro Liter angegeben.

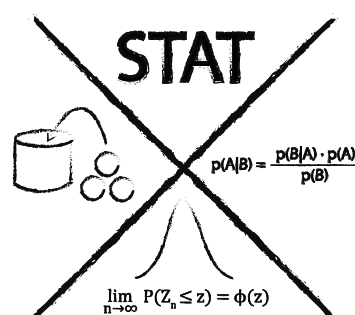
In der Stichprobe werden 111g pro Liter gemessen. Skan-
dal oder Zufall?

a) Prüfe mit einem einstichproben t-Test, ob der Zucker-
gehalt zum Sollwert 106g/l passt. Stelle das Hypothe-
senpaar auf, berechne und interpretiere den P-Wert und
gib das 95% Konfidenzintervall an!

Der Getränkehersteller reagiert auf die Verbraucher-
schutzbeschwerde und passt seinen Produktionspro-
zess an. In einer neuen Studie werden 50 Flaschen des
Getränkeherstellers untersucht.

b) Führe einen zweistichproben t-Test durch, um zu prü-
fen, ob der Prozess tatsächlich geändert wurde. Stelle
das Hypothesenpaar auf, berechne und interpretiere
den P-Wert.

c) Führe einen einstichproben t-Test durch, um zu prü-
fen, ob der Zuckergehalt jetzt zum Sollwert 106g/l passt.

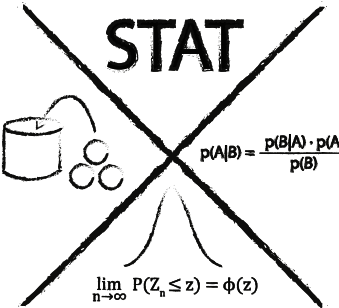


t-Test

Verwende den t-Test, um im Mensadatensatz 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



χ^2 -Unabhängigkeitstest

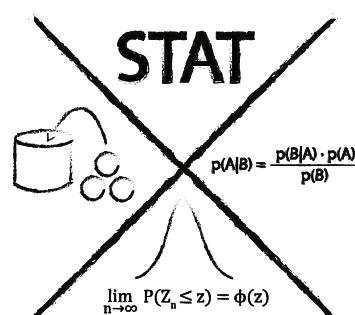
Mit der Chi-Quadrat-Testfamilie lernen wir Tests für kategoriale Daten bzw. Häufigkeitsdaten kennen.

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

Nullhypothese H0 Zwei kategoriale Merkmale weisen keinen Zusammenhang auf.

Alternativhypothese H1 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	

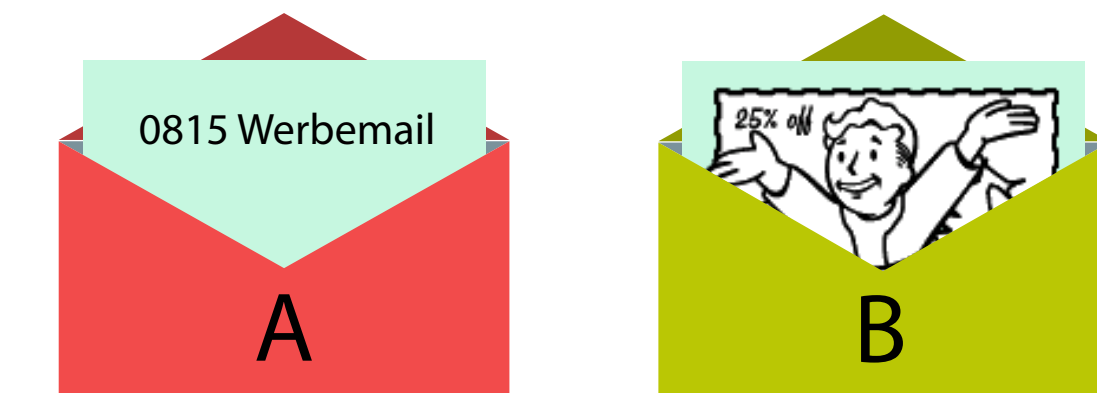



χ^2 -Unabhängigkeitstest

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?



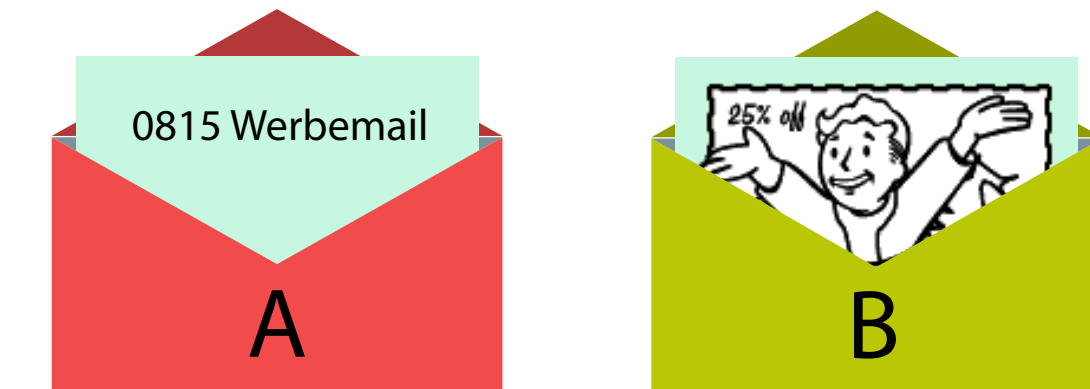
Stichprobe	615	613
Geöffnet	57	71
Ignoriert	558	542
<hr/>		
Quote	9.27%	11.58%
		
	Zufall oder signifikant?	


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.



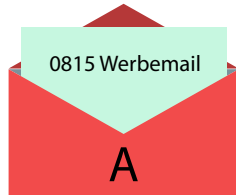

Stichprobe	615	613
Geöffnet	57	71
Ignoriert	558	542
<hr/>		
Quote	9.27%	11.58%
		
	Zufall oder signifikant?	

χ^2 -Unabhängigkeitstest

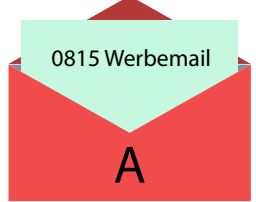
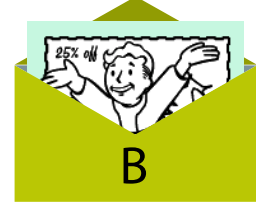
Der Test vergleicht dabei die beobachtete Kontingenztabelle mit einer Kontingenztabelle, die wir erwarten würden, wenn die Nullhypothese ...

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

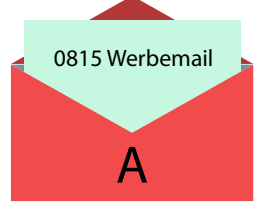
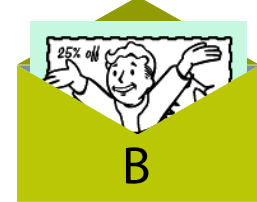
...zutreffen würde. In unserem Beispiel würden wir dann bei beiden Werbemailings eine Öffnungsrate von 10.4% erwarten!

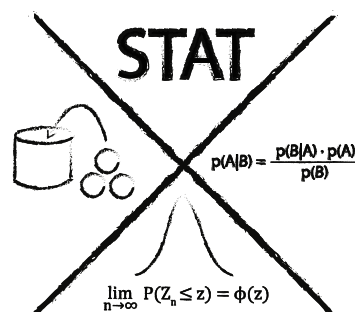
	 A	 B	
	Variante A	Variante B	
Geöffnet	57 (9.3%)	71 (11.6%)	128 (10.4%)
Ignoriert	558	542	1100
	615	613	1228

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

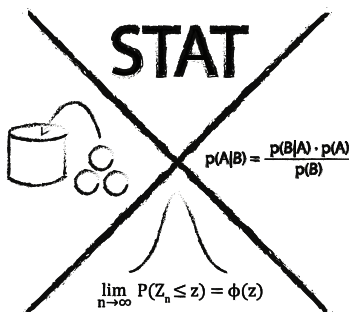
Wir verwenden die Variante für nicht-gepaarte Stichproben und erhalten einen p-Wert von 0.1845.

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

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

...mit 18.45% plausibel. Wir behalten H0 also bei.

Chi-Quadrat Test				
Nullhypothese	Es gibt keinen Unterschied in den Öffnungsraten			
Alternativhypothese	Es gibt einen Unterschied in den Öffnungsraten			
Messung	Offen	Ignoriert	Gesamt	Öffnungsrate
Newsletter A	57	558	615	9,27%
Newsletter B	71	542	613	11,58%
Gesamt	128	1100	1228	10,42%
Erwartung unter H0	Offen	Ignoriert	Gesamt	Öffnungsrate
Newsletter A	64	551	615	10,42%
Newsletter B	64	549	613	10,42%
Abweichung	2,31%			
P-Wert	0,1845	nicht signifikant, könnte Zufall sein		



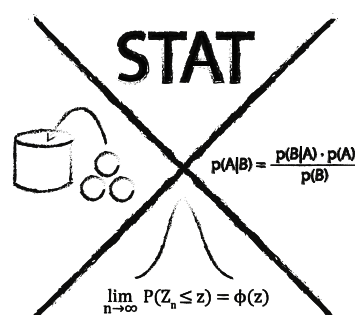
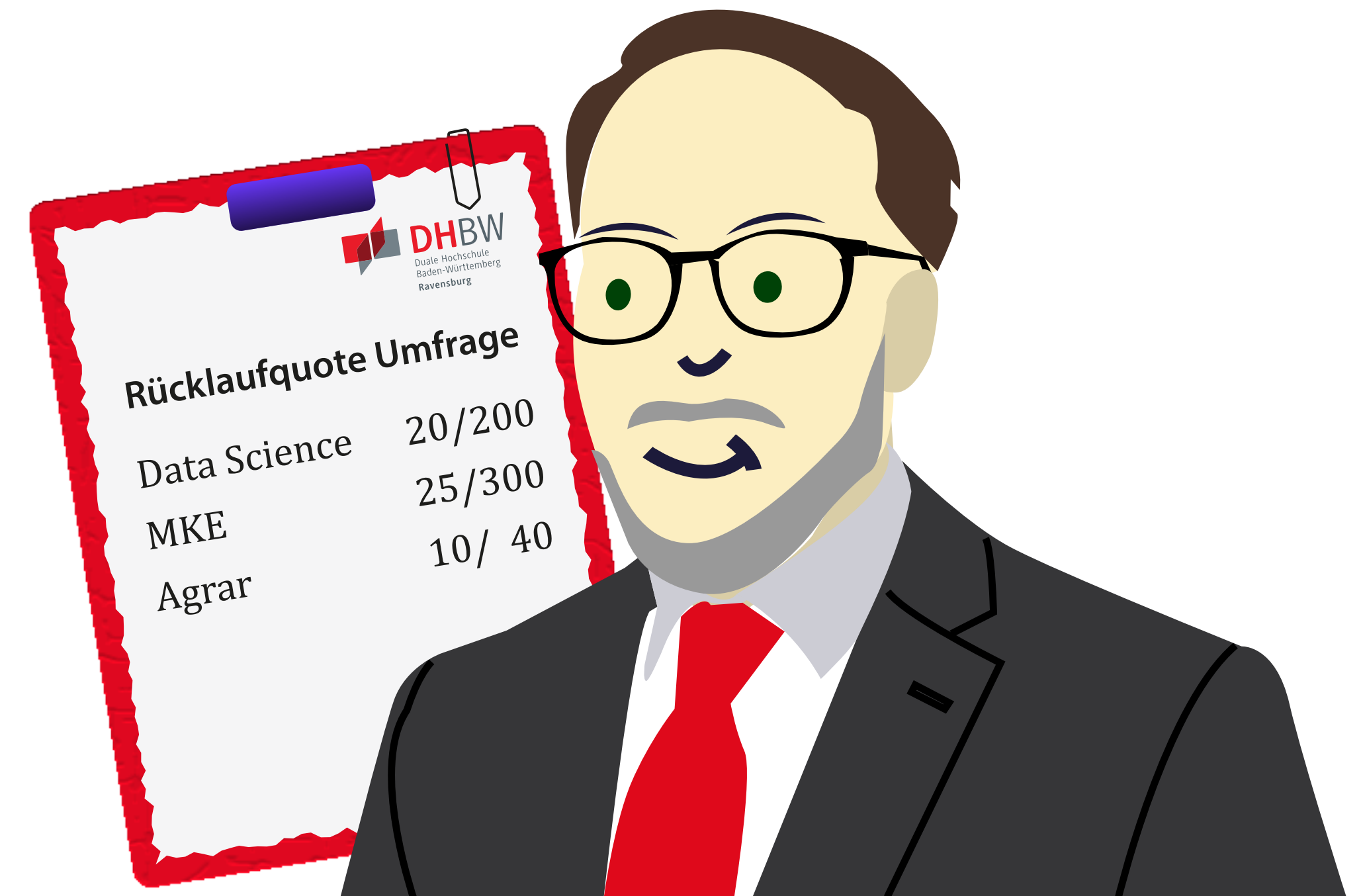
χ^2 -Anpassungstest

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

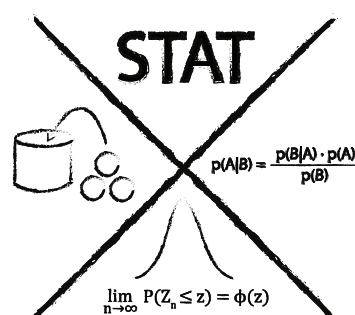
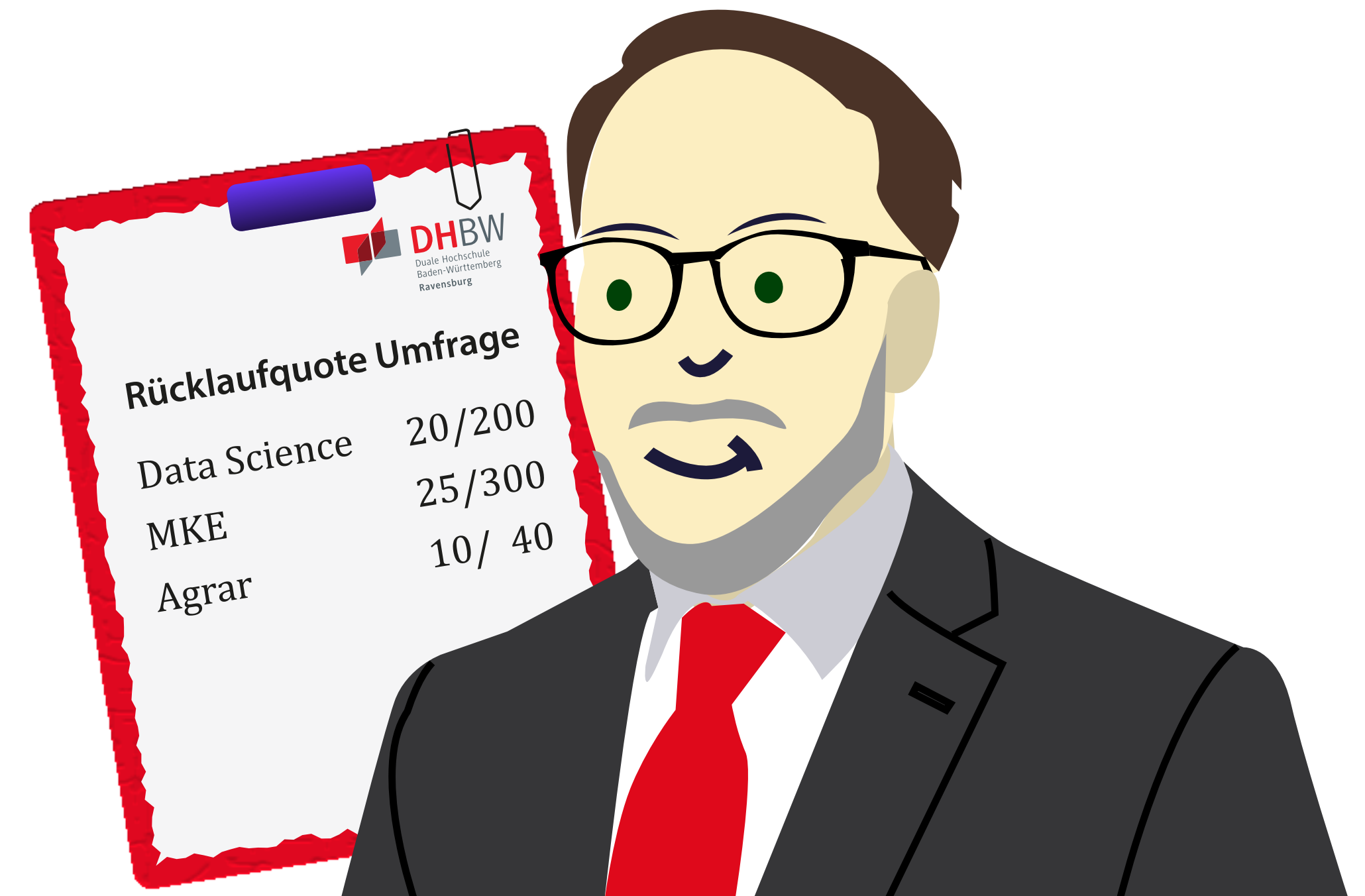
Wir haben eine Umfrage unter Studierenden durchgeführt, bei der auch der Studiengang abgefragt wurde. Wir haben Antworten von:

20 Studierenden aus Data Science

25 Studierenden aus MKE

10 Studierenden aus Agrarwirtschaft

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



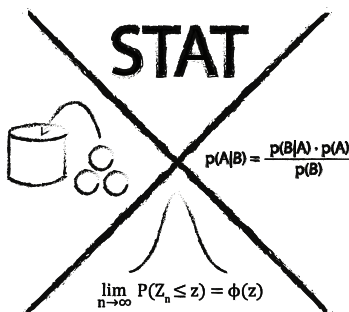
χ^2 -Anpassungstest

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.

Chi-Quadrat-Anpassungs-Test				
Nullhypothese	Die Verteilung in der Stichprobe entspricht der Grundgesamtheit			
Alternativhypothese	Die Verteilung in der Stichprobe weicht von der Grundgesamtheit ab			
Messung	Data Sciece	MKE	Agrar	Gesamt
Umfrage	20	25	10	55
DHBW	200	300	40	540
Rücklaufquote	10,0%	8,3%	25,0%	10,2%
Erwartung unter H0	Data Sciece	MKE	Agrar	
Umfrage	20	31	4	55
P-Wert	0,81%			

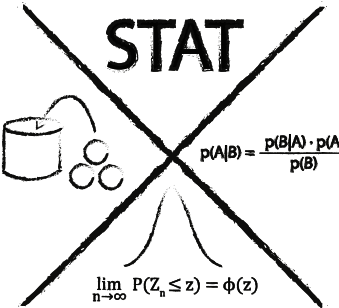


χ²-Testfamilie

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% vegi 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

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

Der t-Test setzt voraus, dass der Mittelwert der Stichprobe normalverteilt ist. Ansonsten muss eine nicht-parametrische Alternative wie der Wilcoxon-Test eingesetzt werden.

Datenfehler Wir verwenden Messdaten, die selbst fehlerbehaftet sind.



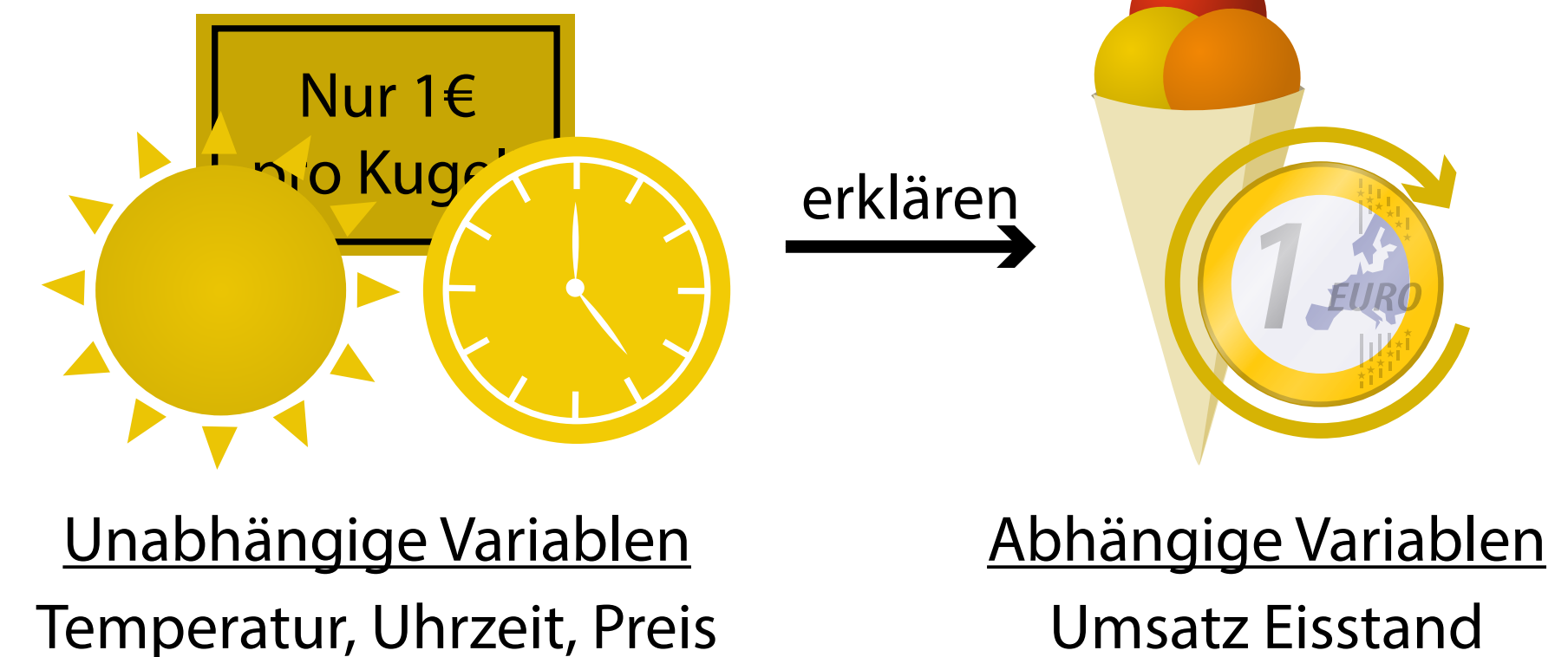
Regressionsanalyse

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



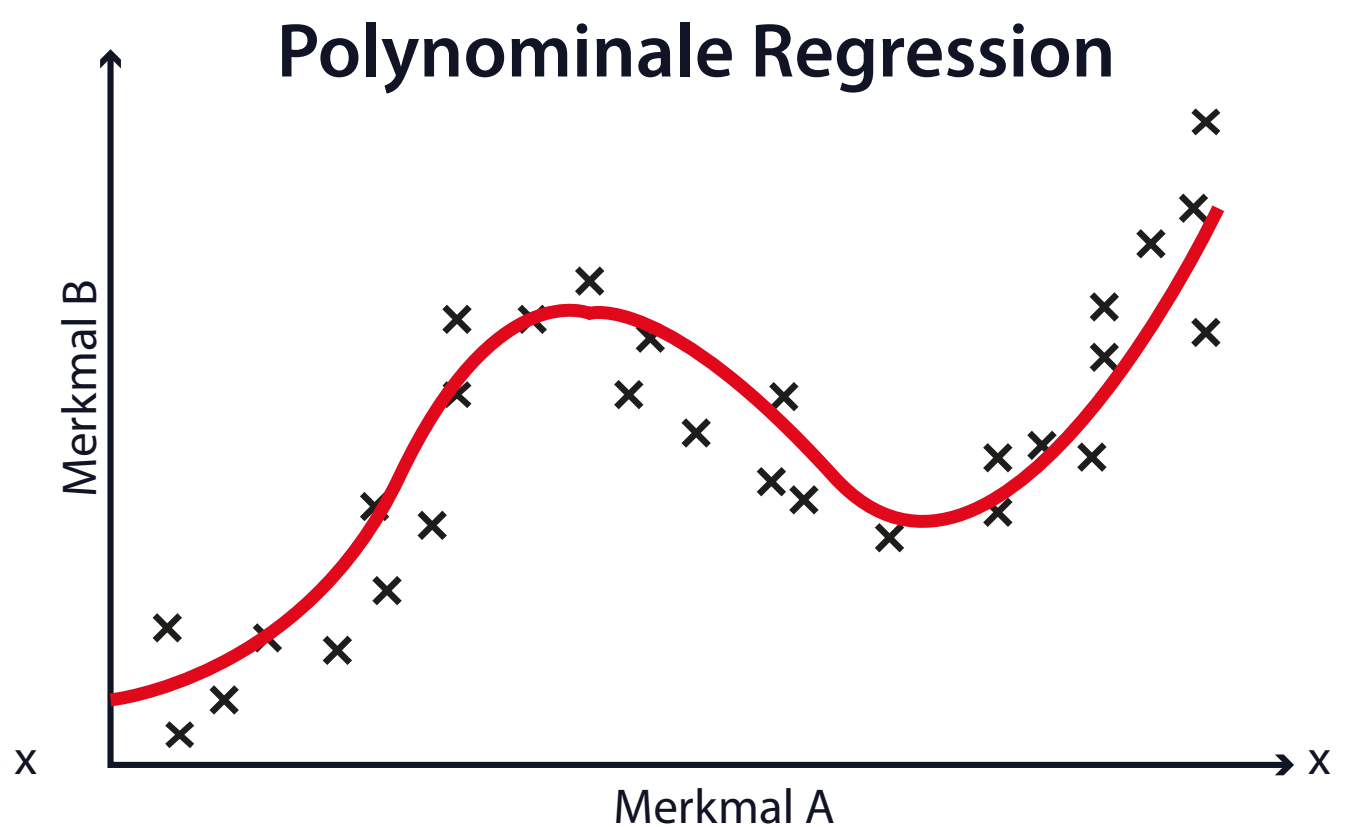
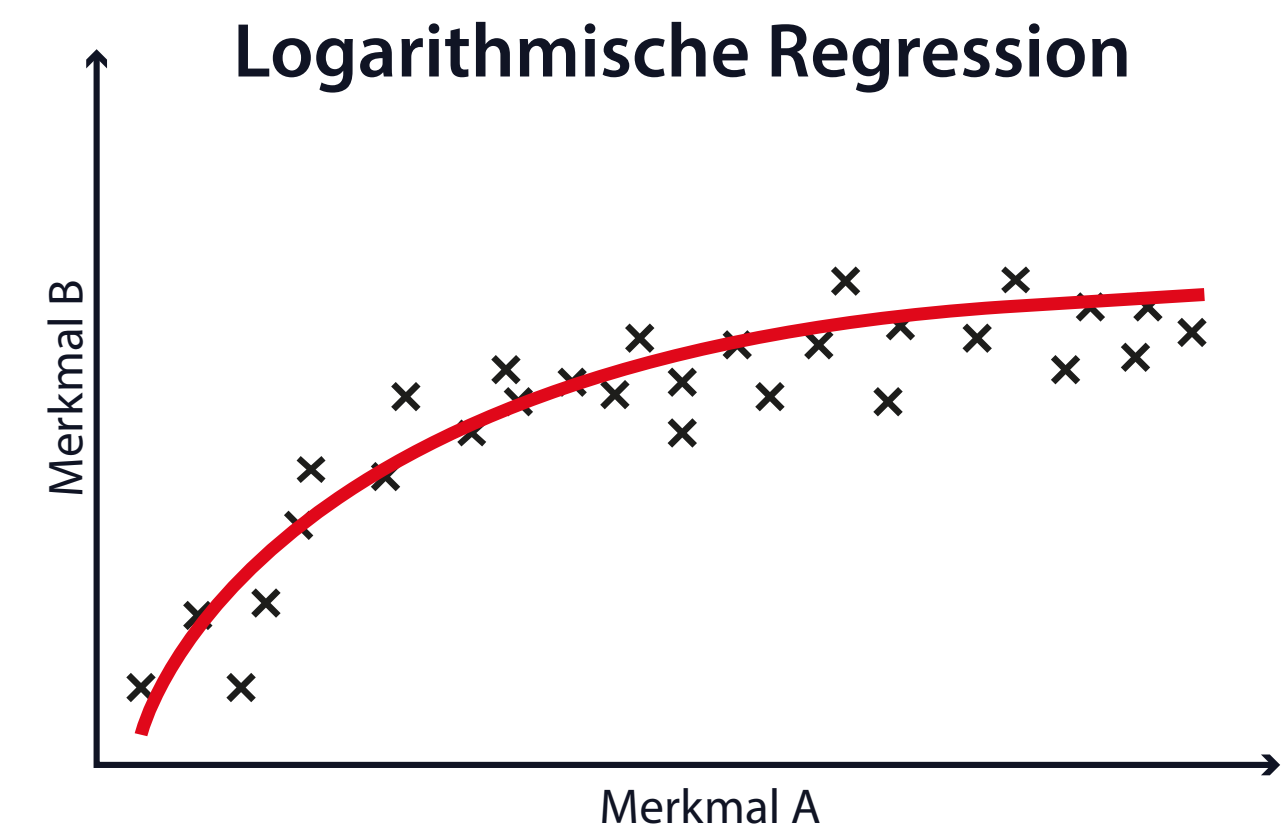
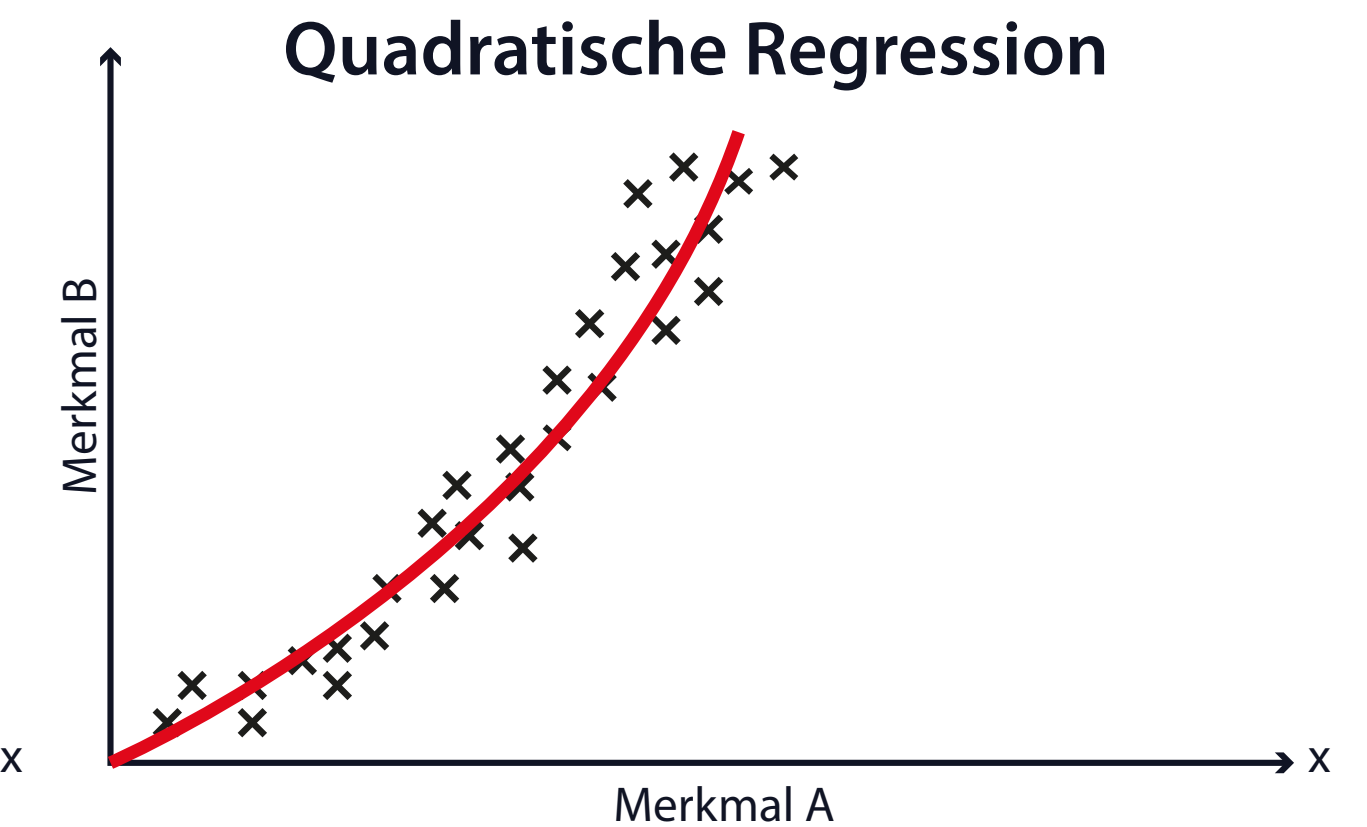
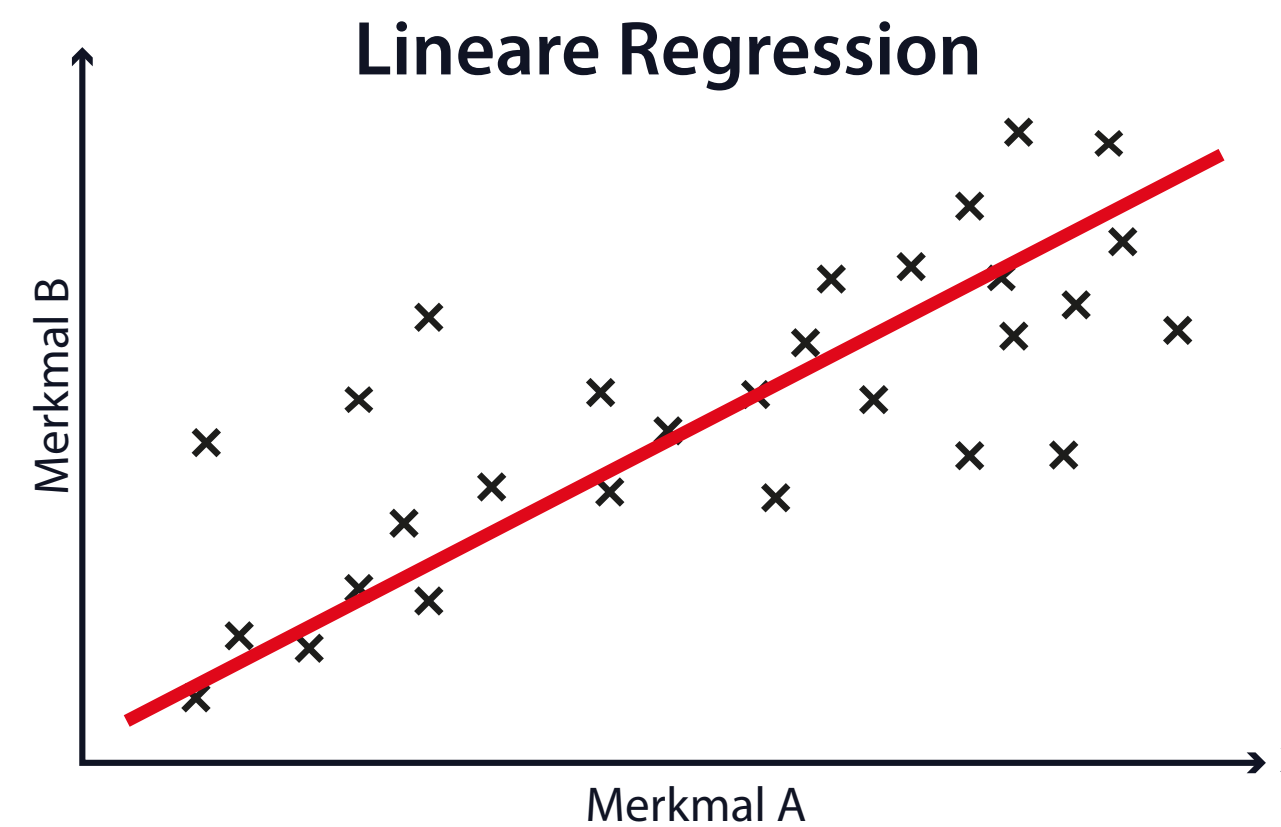
	Eine unabhängige Variable	Mehrere unabhängige Variablen
Eine abhängige Variable	Univariate einfache Regression	Univariate multiple Regression
Mehrere abhängige Variablen	Multivariate einfache Regression	Multivariate multiple Regression

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.



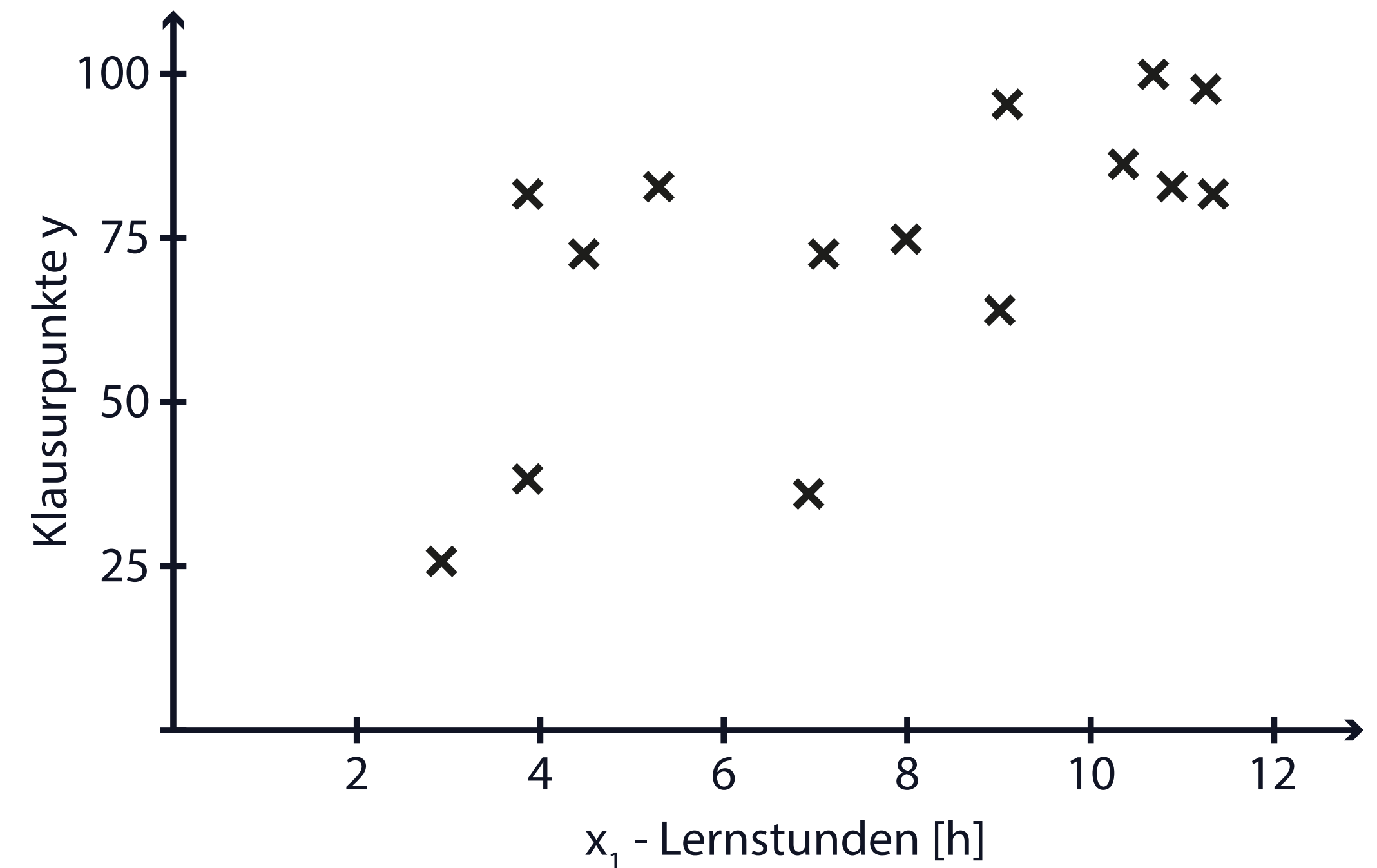
Lineare Regression

Beginnen wir mit einer einfachen, univariaten und 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 führt im Durchschnitt zu β_1 mehr Punkten in der Klausur“.

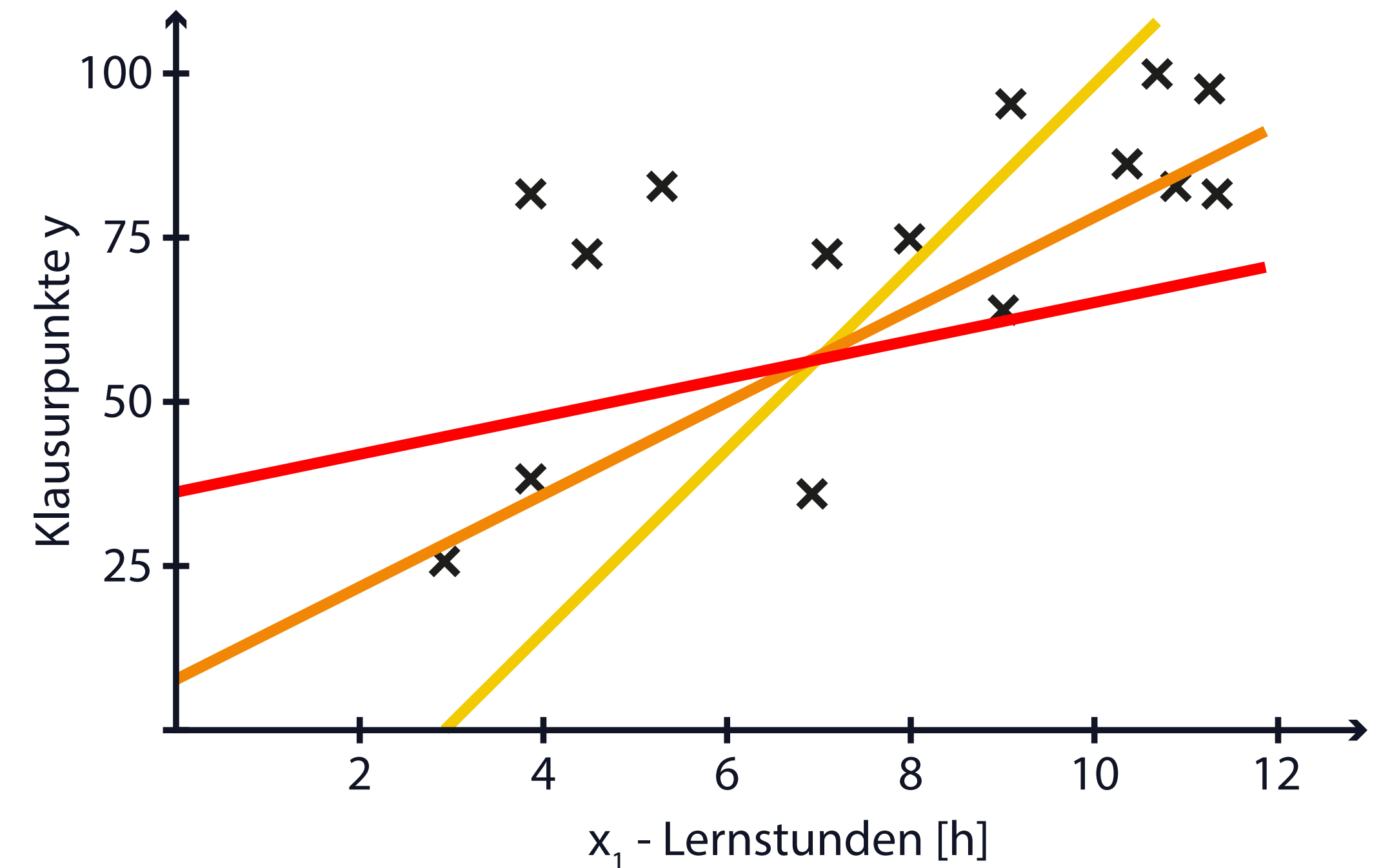


Lineare Regression

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?



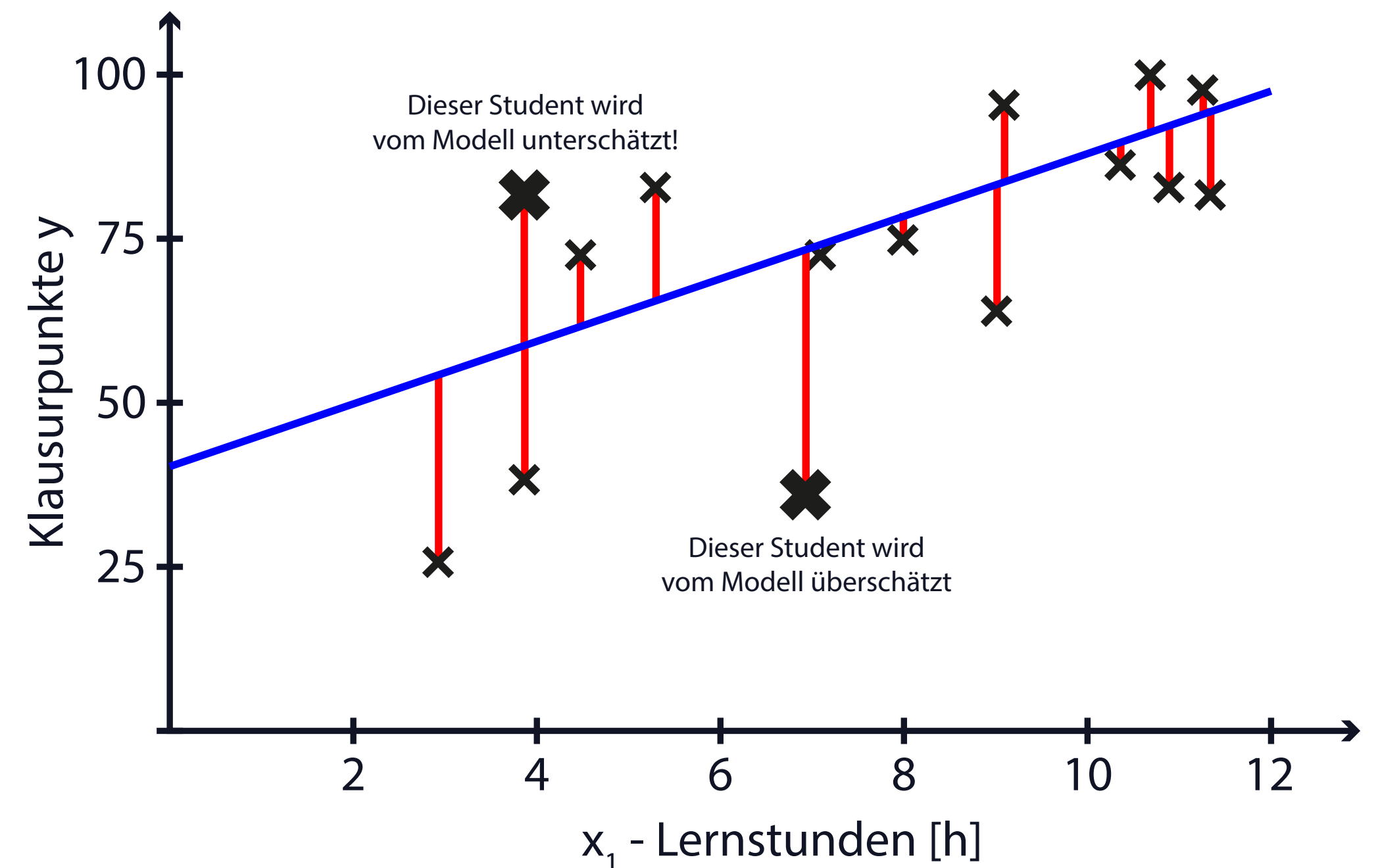
Lineare Regression

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.

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



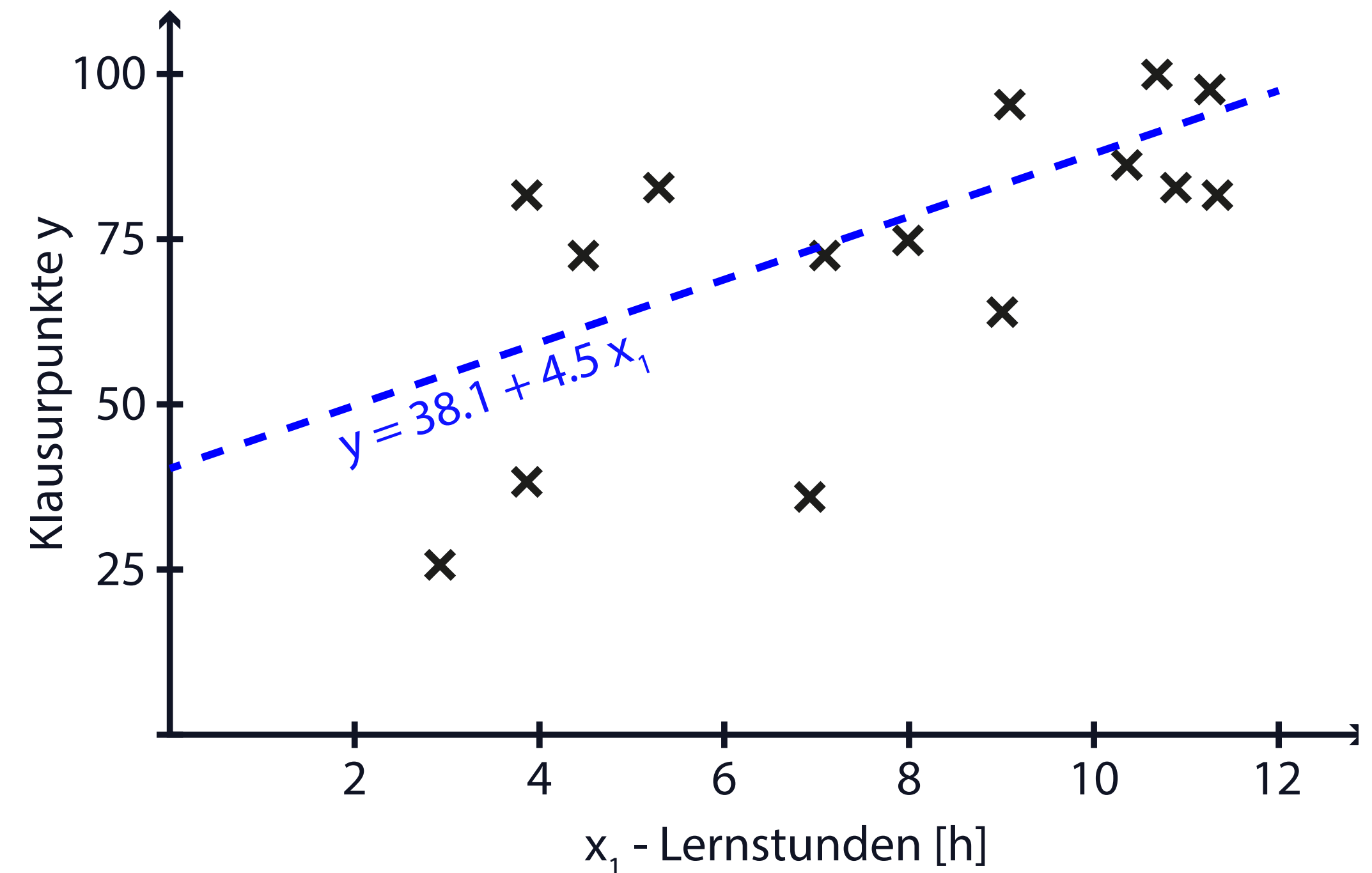
Lineare Regression

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?

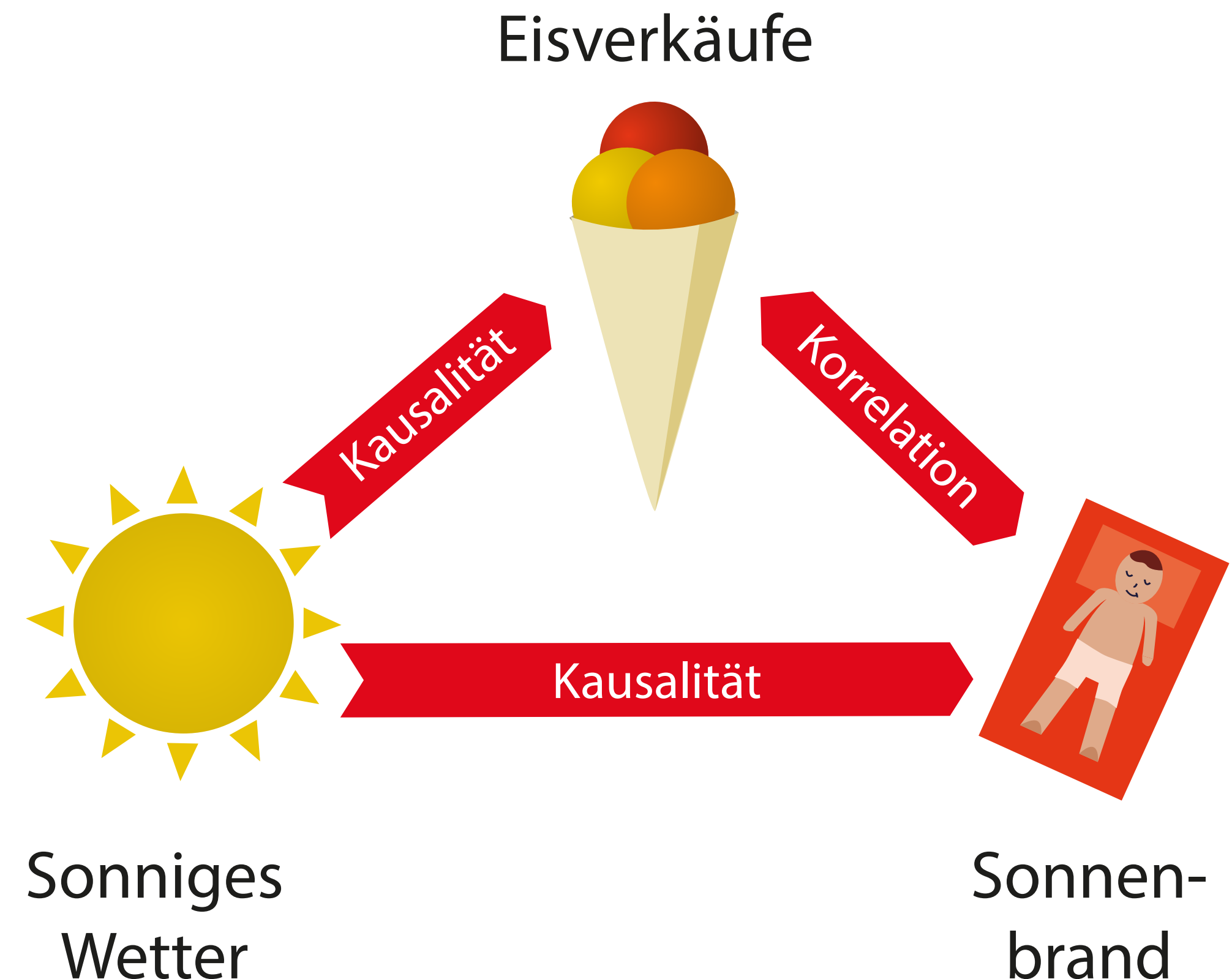


Lineare Regression

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.



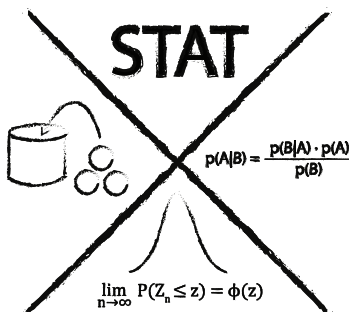
Lineare Regression

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!

AUSGABE: ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
	Koeffizienten	Std.Err	t-Statistik	P-Wert	Untere 95%	Obere 95%
Schnittpunkt	38,100	12,861	2,962	0,011	10,316	65,885
Lernstunden	4,534	1,583	2,864	0,013	1,114	7,953



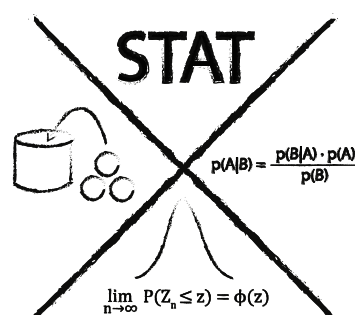
Lineare Regression

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



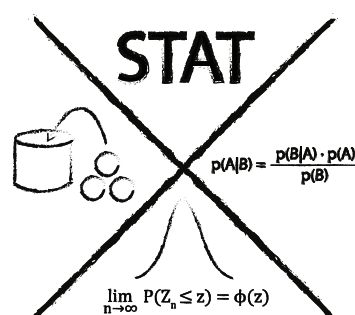
Lineare Regression

In unserem Beispiel haben wir folgendes Hypothesenpaar:

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

Alternativhypothese Es gibt einen Zusammenhang zwischen der Zeit, die wir auf die Klausur gelernt haben, und den dann 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



Lineare Regression

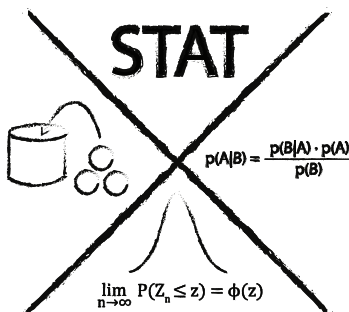
Aus Schätzer und Standardfehler berechnet Excel automatisch den p-Wert.

Dieser 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 der Stichprobe sehen?

* Hier: 4.5 Punkte/Lernstunde oder mehr.

AUSGABE: ZUSAMMENFASSUNG						
<i>Regressions-Statistik</i>						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
Koeffizienten		Std.Err	t-Statistik	P-Wert	Untere 95%	Obere 95%
Schnittpunkt	38,100	12,861	2,962	0,011	10,316	65,885
Lernstunden	4,534	1,583	2,864	0,013	1,114	7,953



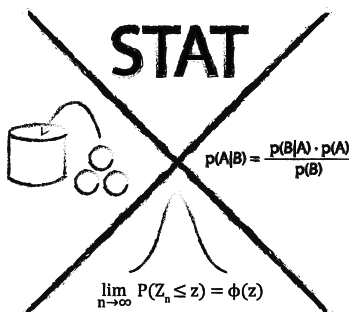
Lineare Regression

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!

AUSGABE: ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
	Koeffizienten	Std.Err	t-Statistik	P-Wert	Untere 95%	Obere 95%
Schnittpunkt	38,100	12,861	2,962	0,011	10,316	65,885
Lernstunden	4,534	1,583	2,864	0,013	1,114	7,953



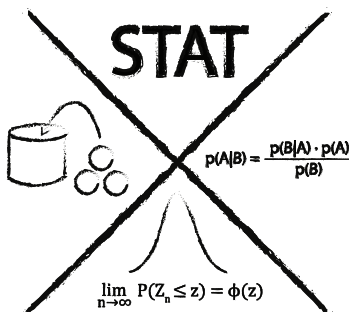
Lineare Regression

Wir verwerfen 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.

AUSGABE: ZUSAMMENFASSUNG						
<i>Regressions-Statistik</i>						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
	Koeffizienten	Std.Err	t-Statistik	P-Wert	Untere 95%	Obere 95%
Schnittpunkt	38,100	12,861	2,962	0,011	10,316	65,885
Lernstunden	4,534	1,583	2,864	0,013	1,114	7,953



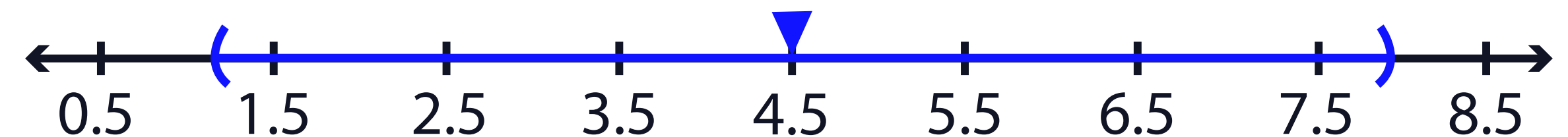
Lineare Regression

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

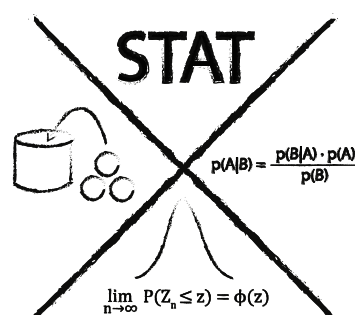
An dieser Stelle kommt das angegebene Konfidenzintervall ins Spiel.

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

AUSGABE: ZUSAMMENFASSUNG						
<i>Regressions-Statistik</i>						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
	<i>Koeffizienten</i>	<i>Std.Err</i>	<i>t-Statistik</i>	<i>P-Wert</i>	<i>Untere 95%</i>	<i>Obere 95%</i>
Schnittpunkt	38,100	12,861	2,962	0,011	10,316	65,885
Lernstunden	4,534	1,583	2,864	0,013	1,114	7,953



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



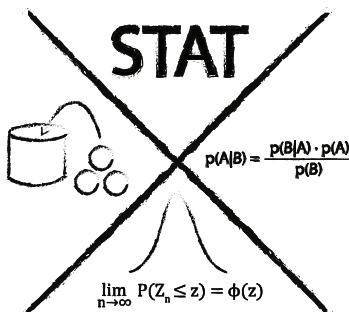
Lineare Regression

Welcher Anteil der Varianz in der abhängigen Variable wird durch unsere unabhängigen Variablen erklärt? Die Antwort gibt der R²-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.)

AUSGABE: ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,6220					
Bestimmtheitsmaß	0,3869					
Adjustiertes Bestimmtheitsmaß	0,3398					
Standardfehler	17,9390					
Beobachtungen	15					
		Koeffizienten	Std.Err	t-Statistik	P-Wert	Untere 95% Obere 95%
Schnittpunkt		38,100	12,861	2,962	0,011	10,316 65,885
Lernstunden		4,534	1,583	2,864	0,013	1,114 7,953

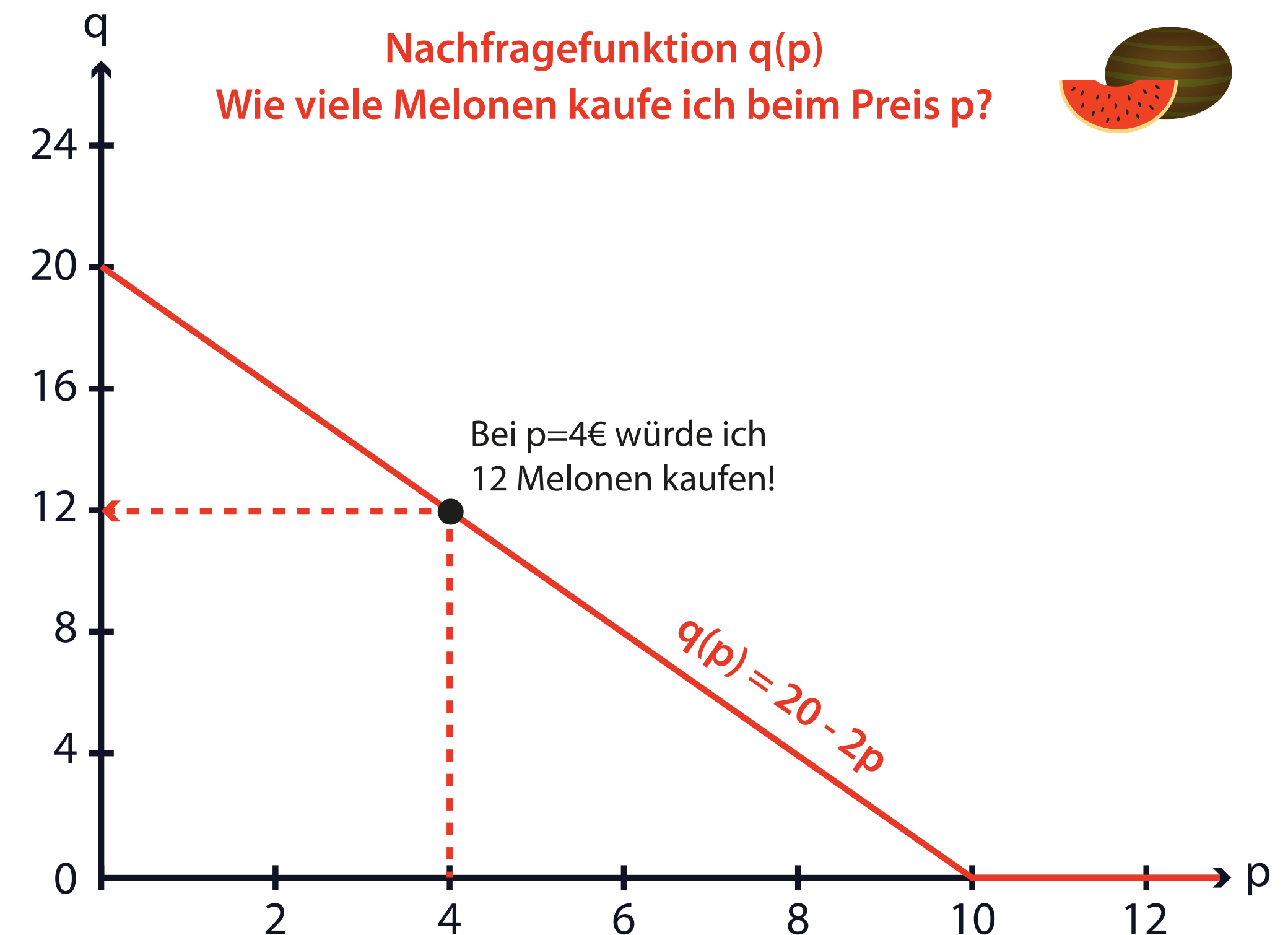


Beispiel Strommarkt

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!



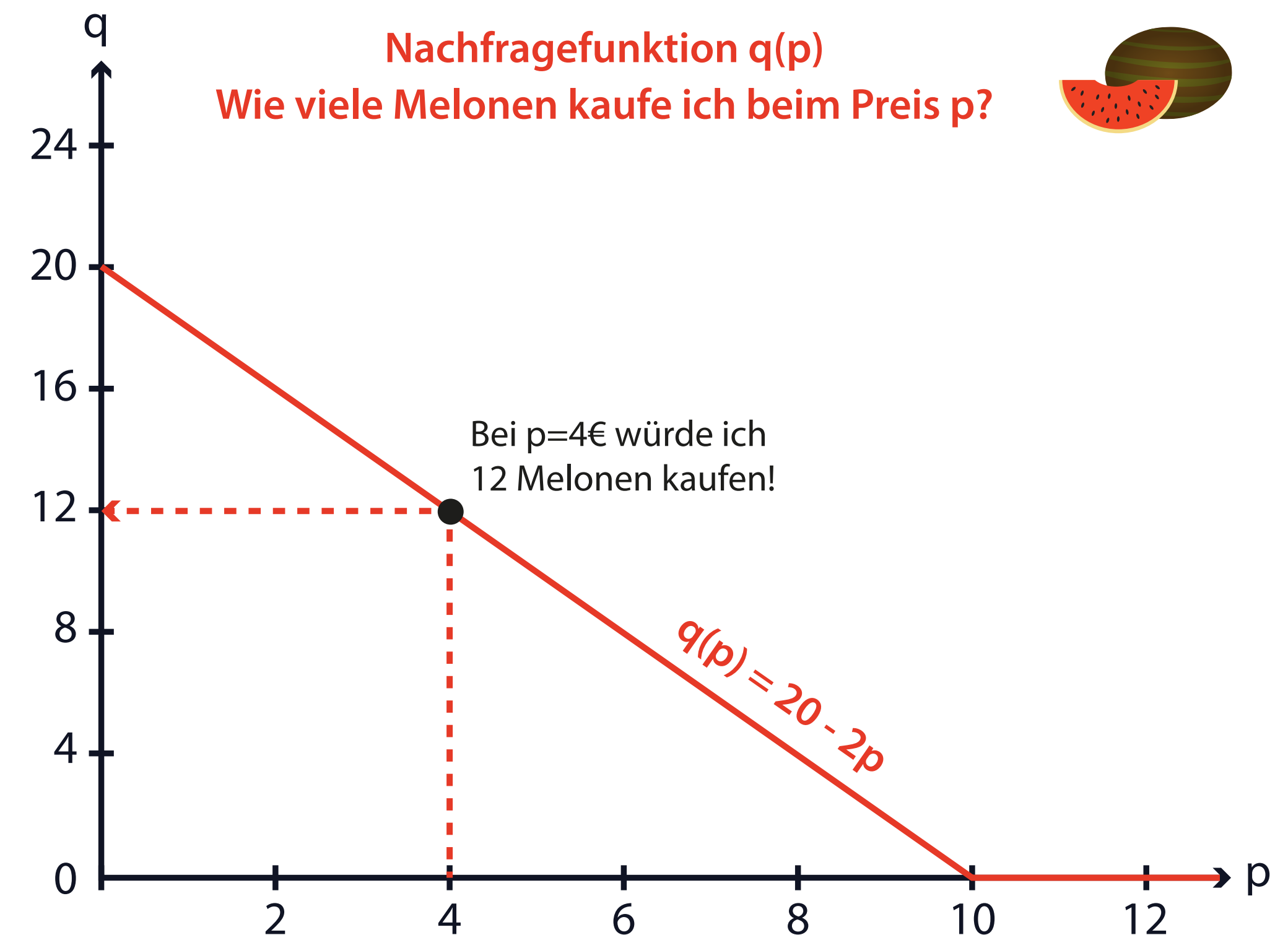
Beispiel Strommarkt

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

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

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

Dazu benötigen wir aber einen Datensatz der Preise und Mengen enthält.



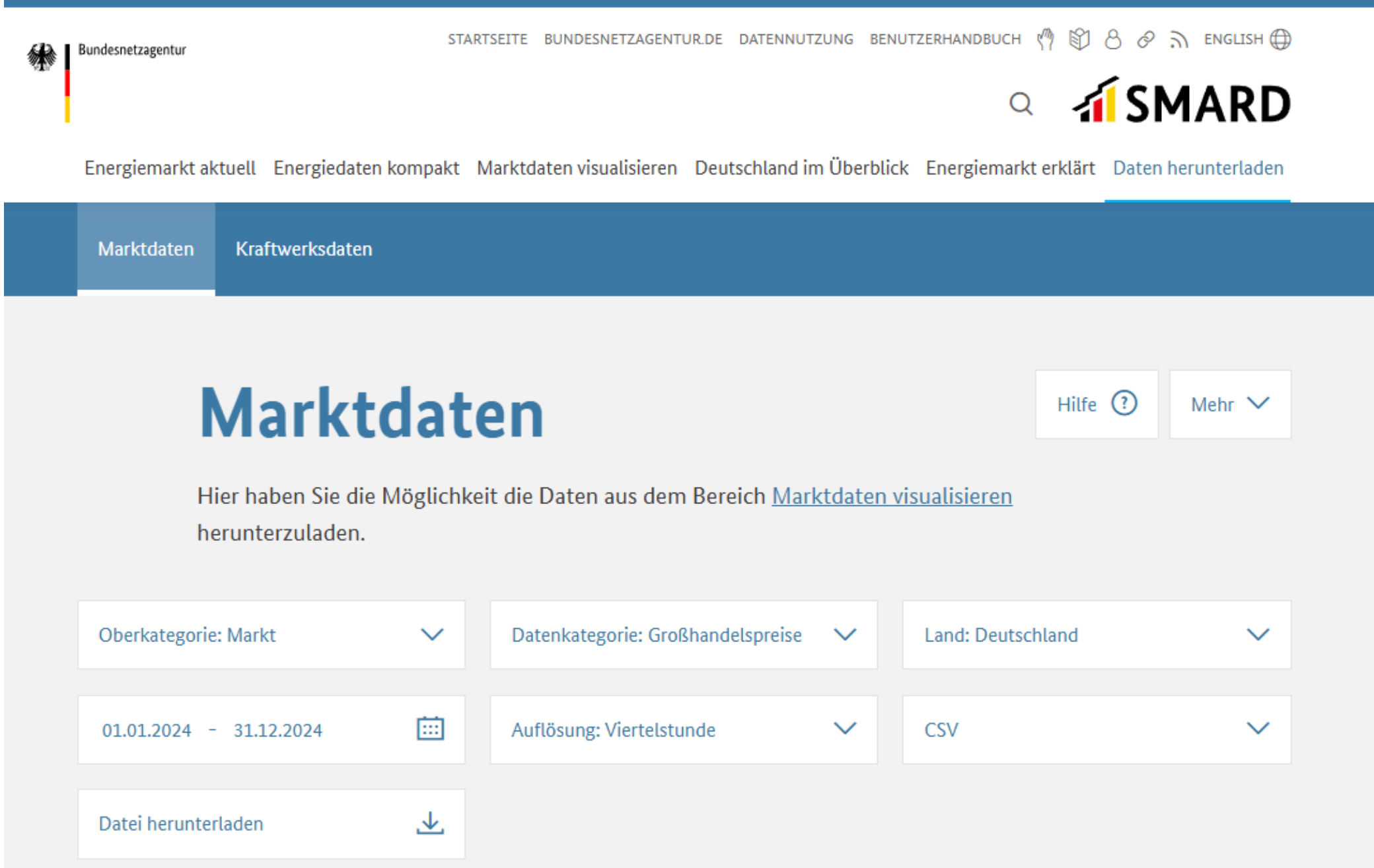
Beispiel Strommarkt

SMARD-Datenbank Für den Strommarkt gibt es eine öffentlich zugängliche Datenbank mit Strompreisen und Nachfragewerten.

Aus dieser Datenbank laden wir den rechts gezeigten Datensatz herunter und führen eine Regressionsanalyse in Excel 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

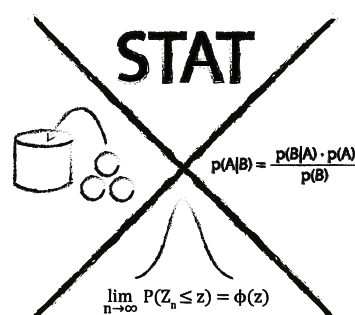
Marktdaten Hilfe ? Mehr ▾

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 ⬇

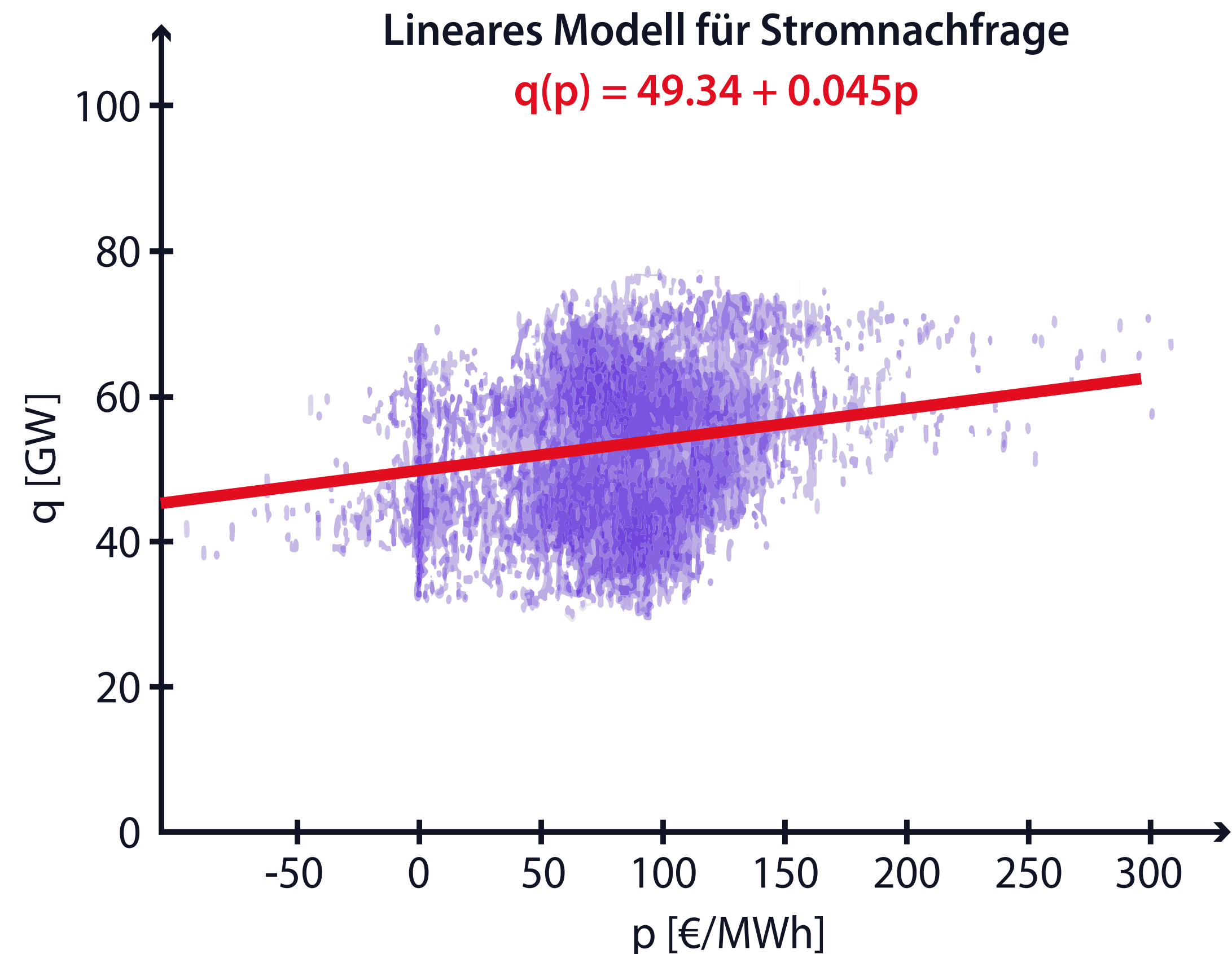


Beispiel Strommarkt

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

Der Schätzer für β_1 ist 0.0454 mit einem Standardfehler von 0.0009.

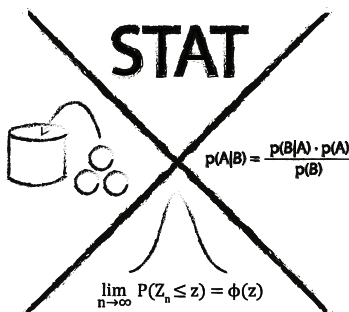
Unter Gültigkeit der Nullhypothese erhalten wir mit Wahrscheinlichkeit $< 0.00001\%$ einen Schätzer von mindestens ± 0.0454 .

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

AUSGABE ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,263					
Bestimmtheitsmaß	0,069					
Adjustiertes Bestimmtheitsmaß	0,069					
Standardfehler	8,799					
Beobachtungen	35136					
	Koeff	STABW	t	p	Untere 95%	Obere 95%
Schnittpunkt	49,3410	0,0842	586,0186	0,0000	49,1759	49,5060
Preis	0,0454	0,0009	51,0504	0,0000	0,0437	0,0472



Zu 95% liegt der wahre Wert von β_1 zwischen 0.0437 und 0.0472



Beispiel Strommarkt

Der gefundene Schätzer gilt damit als signifikant!

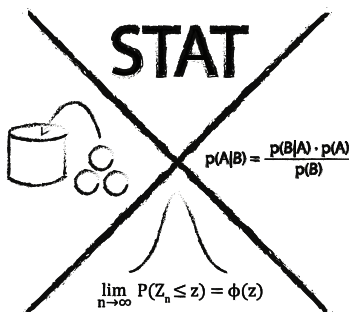
~~H0 für β_+ - Der wahre Wert von β_+ ist null. Die Stromnachfrage hängt nicht mit dem Strompreis zusammen.~~

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

AUSGABE ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,263					
Bestimmtheitsmaß	0,069					
Adjustiertes Bestimmtheitsmaß	0,069					
Standardfehler	8,799					
Beobachtungen	35136					
	Koeff	STABW	t	p	Untere 95%	Obere 95%
Schnittpunkt	49,3410	0,0842	586,0186	0,0000	49,1759	49,5060
Preis	0,0454	0,0009	51,0504	0,0000	0,0437	0,0472



Zu 95% liegt der wahre Wert von β_1 zwischen 0.0437 und 0.0472



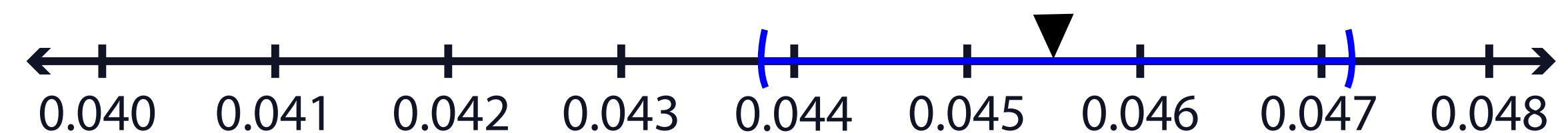
Rückwärtskausalität

Irgendwas an unseren Ergebnissen ist seltsam:

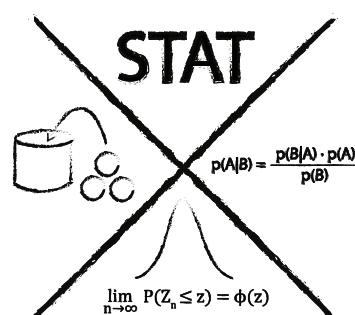
Eine um 1€ höherer Großhandelspreis pro MWh geht im Durchschnitt mit einer um 45.4 MW höheren Nachfrage einher.

Die Stromkunden wollen 45.4 MW mehr Leistung, wenn der Strom an der Börse 1€ teurer wird?

AUSGABE ZUSAMMENFASSUNG						
<i>Regressions-Statistik</i>						
Multipler Korrelationskoeffizient	0,263					
Bestimmtheitsmaß	0,069					
Adjustiertes Bestimmtheitsmaß	0,069					
Standardfehler	8,799					
Beobachtungen	35136					
	<i>Koeff</i>	<i>STABW</i>	<i>t</i>	<i>p</i>	<i>Untere 95%</i>	<i>Obere 95%</i>
Schnittpunkt	49,3410	0,0842	586,0186	0,0000	49,1759	49,5060
Preis	0,0454	0,0009	51,0504	0,0000	0,0437	0,0472



Zu 95% liegt der wahre Wert von β_1 zwischen 0.0437 und 0.0472



Rückwärtskausalität

Eine um 1€ höherer Großhandelspreis pro MWh geht im Durchschnitt mit einer um 45.4 MW höheren Nachfrage einher.

~~Die Stromkunden wollen 45.4 MW mehr Leistung, wenn der Strom an der Börse 1€ teurer wird.~~

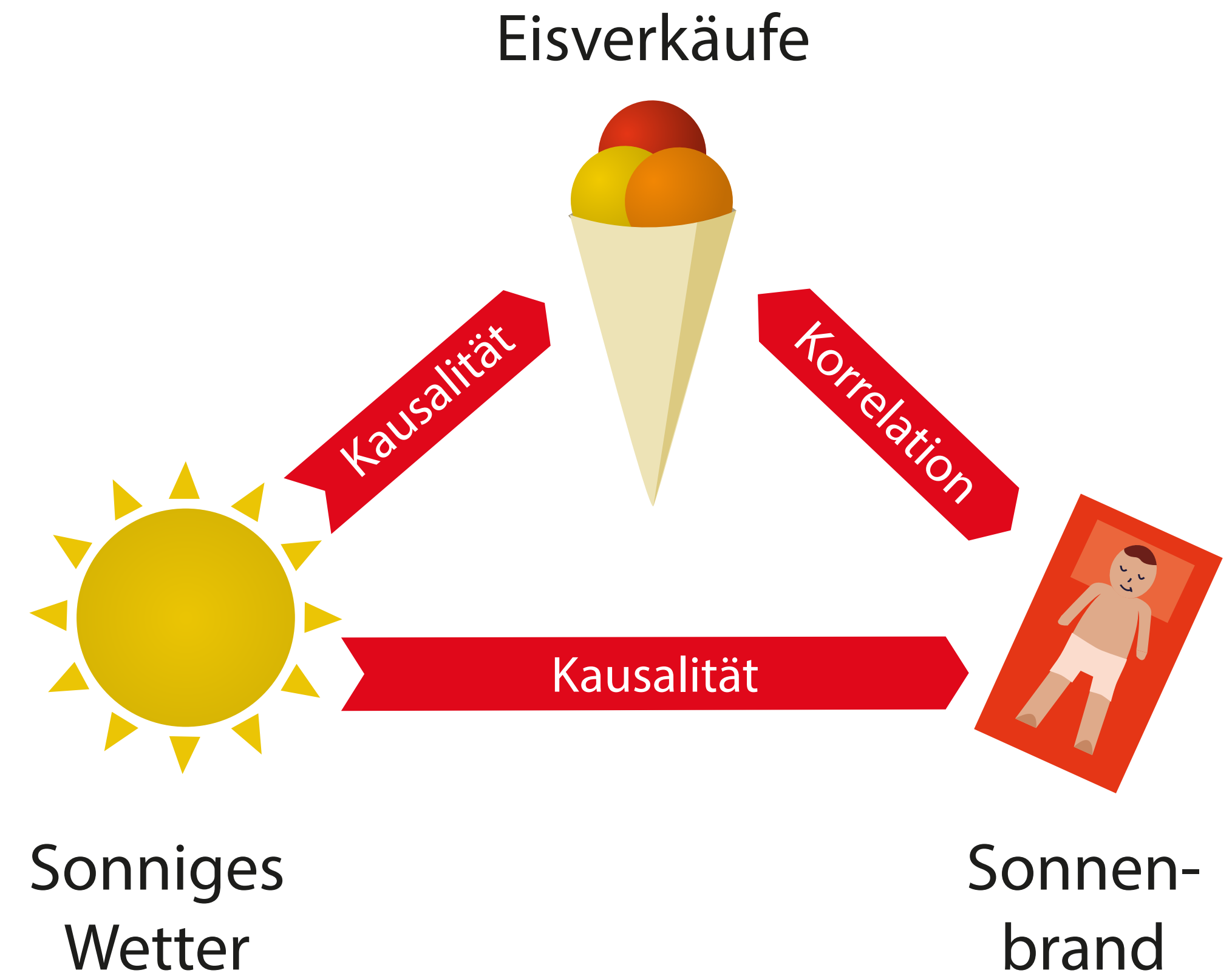
Das macht überhaupt keinen Sinn! Das Gegenteil sollte der Fall sein!



Rückwärtskausalität

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.

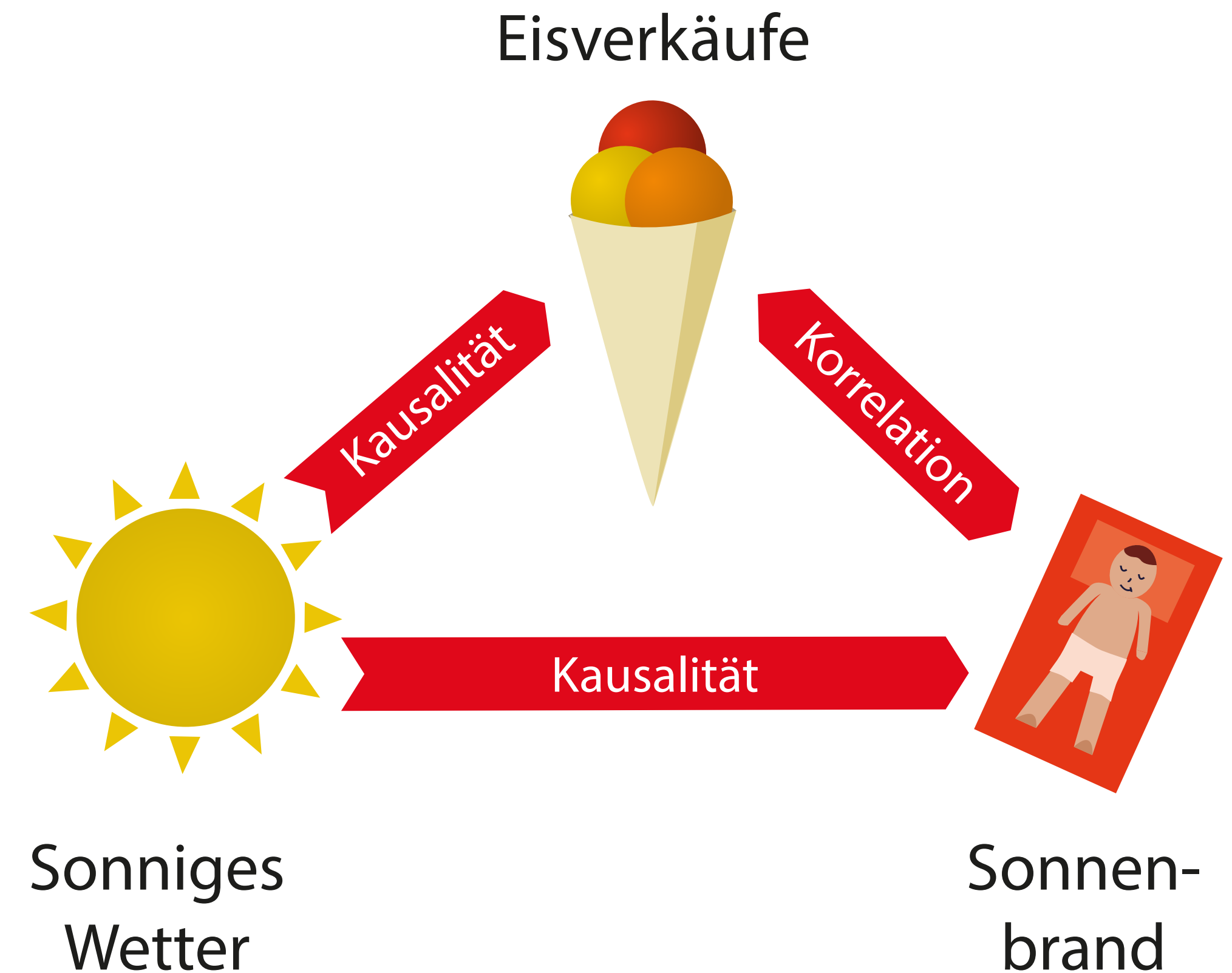


Rückwärtskausalität

Eine um 1€ höherer Großhandelspreis pro MWh geht im Durchschnitt mit einer um 45.4 MW höheren Nachfrage einher.

~~Die Stromkunden wollen 45.4 MW mehr Leistung, wenn der Strom an der Börse 1€ teurer wird.~~

Der Börsenpreis wird 1€ teurer, wenn die Konsumenten 45.4 MW mehr Leistung nachfragen.



Fortsetzung des Beispiels

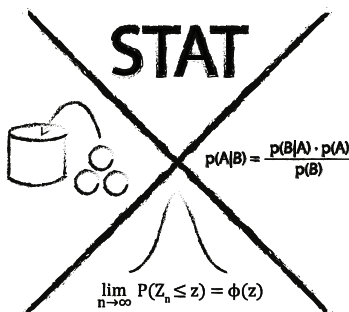
Statt der Nachfragefunktion schätzen wir nun die Preis-Absatz-Funktion mit dem Modell:

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

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

6.9% der Preisschwankungen werden durch Nachfrageschwankungen erklärt.

AUSGABE ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,263					
Bestimmtheitsmaß	0,069					
Adjustiertes Bestimmtheitsmaß	0,069					
Standardfehler	8,799					
Beobachtungen	35136					
	Koeff	STABW	t	p	Untere 95%	Obere 95%
Schnittpunkt	49,3410	0,0842	586,0186	0,0000	49,1759	49,5060
Preis	0,0454	0,0009	51,0504	0,0000	0,0437	0,0472



Fortsetzung des Beispiels

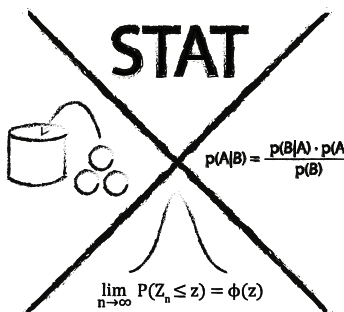
Wir erweitern das Modell um die EE-Einspeisung:

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

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!

AUSGABE ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,781					
Bestimmtheitsmaß	0,610					
Adjustiertes Bestimmtheitsmaß	0,610					
Standardfehler	32,907					
Beobachtungen	35136					
	Koeff	STABW	t	p	Untere 95%	Obere 95%
Schnittpunkt	3,7626	1,0339	3,6390	0,0003	1,7360	5,7891
GesamtEE	-3,1078	0,0141	-220,9821	0,0000	-3,1353	-3,0802
Netzlaster	3,1173	0,0206	151,5827	0,0000	3,0770	3,1576

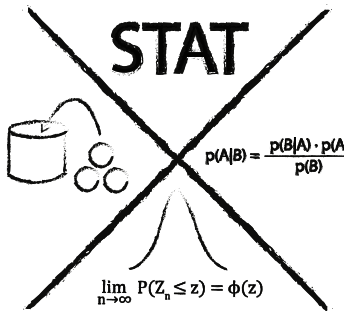


Fortsetzung des Beispiels

Eine Nachfrageerhöhung um 1 GW geht im Durchschnitt mit einer Erhöhung des Börsenpreises um 4.55€ einher.

Ein Plus von 1 GW EE-Leistung geht im Durchschnitt mit einer Senkung des Börsenpreises um 4.78€ einher.

AUSGABE ZUSAMMENFASSUNG						
Regressions-Statistik						
Multipler Korrelationskoeffizient	0,781					
Bestimmtheitsmaß	0,610					
Adjustiertes Bestimmtheitsmaß	0,610					
Standardfehler	32,907					
Beobachtungen	35136					
	Koeff	STABW	t	p	Untere 95%	Obere 95%
Schnittpunkt	3,7626	1,0339	3,6390	0,0003	1,7360	5,7891
GesamtEE	-3,1078	0,0141	-220,9821	0,0000	-3,1353	-3,0802
Netzlaster	3,1173	0,0206	151,5827	0,0000	3,0770	3,1576



Lineare 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.



Lineare 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.

