Teil 3: Verteilungen (Distributions)

Verteilungen

- Eine Verteilung beschreibt alle wahrscheinlichen Ergebnisse einer **Variablen**.
- In einer **diskreten** Verteilung muss die Summe aller einzelnen Wahrscheinlichkeiten gleich 1 sein
- In einer **kontinuierlichen** Verteilung ist die Fläche unter der Wahrscheinlichkeitskurve gleich 1

Diskrete Wahrscheinlichkeits-verteilung

Diskrete Wahrscheinlichkeitsverteilung

- **Diskrete Wahrscheinlichkeitsverteilungen** werden auch als *Wahrscheinlichkeitsfunktionen* bezeichnet:

Gleichverteilung

Binominalverteilung

Poisson-Verteilung

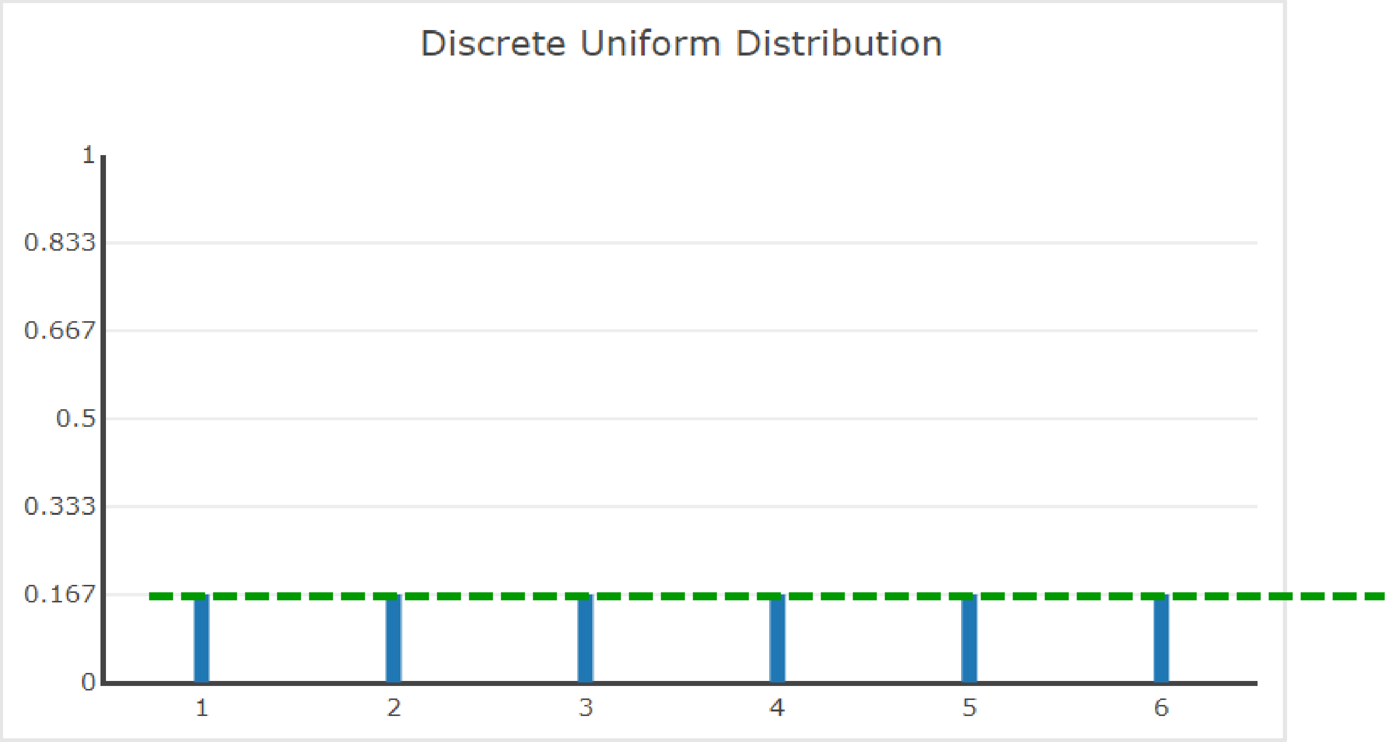
Gleichverteilung (Uniform Distribution)

Gleichverteilung

- Das Werfen eines Würfels hat 6 diskrete, gleichverteilte mögliche Ergebnisse
- Du kannst eine 1 oder 2, aber nicht eine 1,5 werfen
- Die Wahrscheinlichkeiten, dass eine bestimmte Augenzahl gewürfelt wird, ist für alle Zahlen gleich, die Zufallsvariable ist diskret gleichverteilt.

Gleichverteilung

- Wahrscheinlichkeitsverteilung der diskreten Zufallsvariablen “Würfelwurf“.



Alle Balken sind gleich hoch und ergeben zusammen 1

Binominalverteilung

Binominalverteilung

- "Binominal" bedeutet, dass es zwei diskrete, sich gegenseitig ausschließende Ergebnisse eines Zufallsexperiments gibt.

Kopf oder **Zahl**

an oder **aus**

krank oder **gesund**

Erfolg (success) oder **Misserfolg (failure)**

Bernoulli-Experiment

- Ein **Bernoulli-Experiment** ist ein Zufallsexperiment mit genau **zwei** möglichen Versuchsausgängen: *Erfolg* oder *Misserfolg*
- Eine Reihe von Versuchen n wird so lange einer binären Verteilung folgen wie

    a) Die Erfolgswahrscheinlichkeit p konstant ist

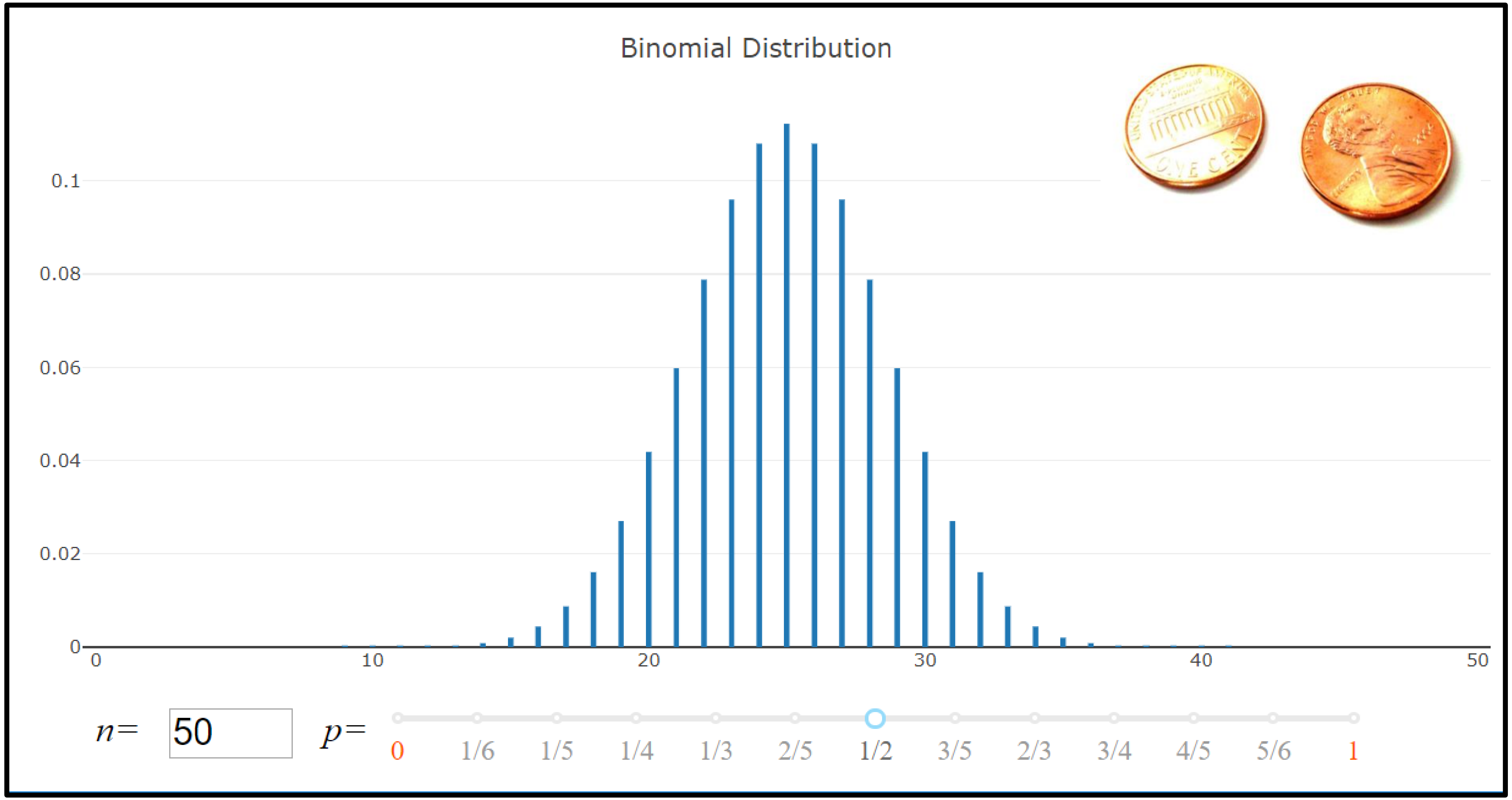
    b) Die Versuche unabhängig voneinander sind

Wahrscheinlichkeitsfunktion einer binominalen Zufallsvariablen

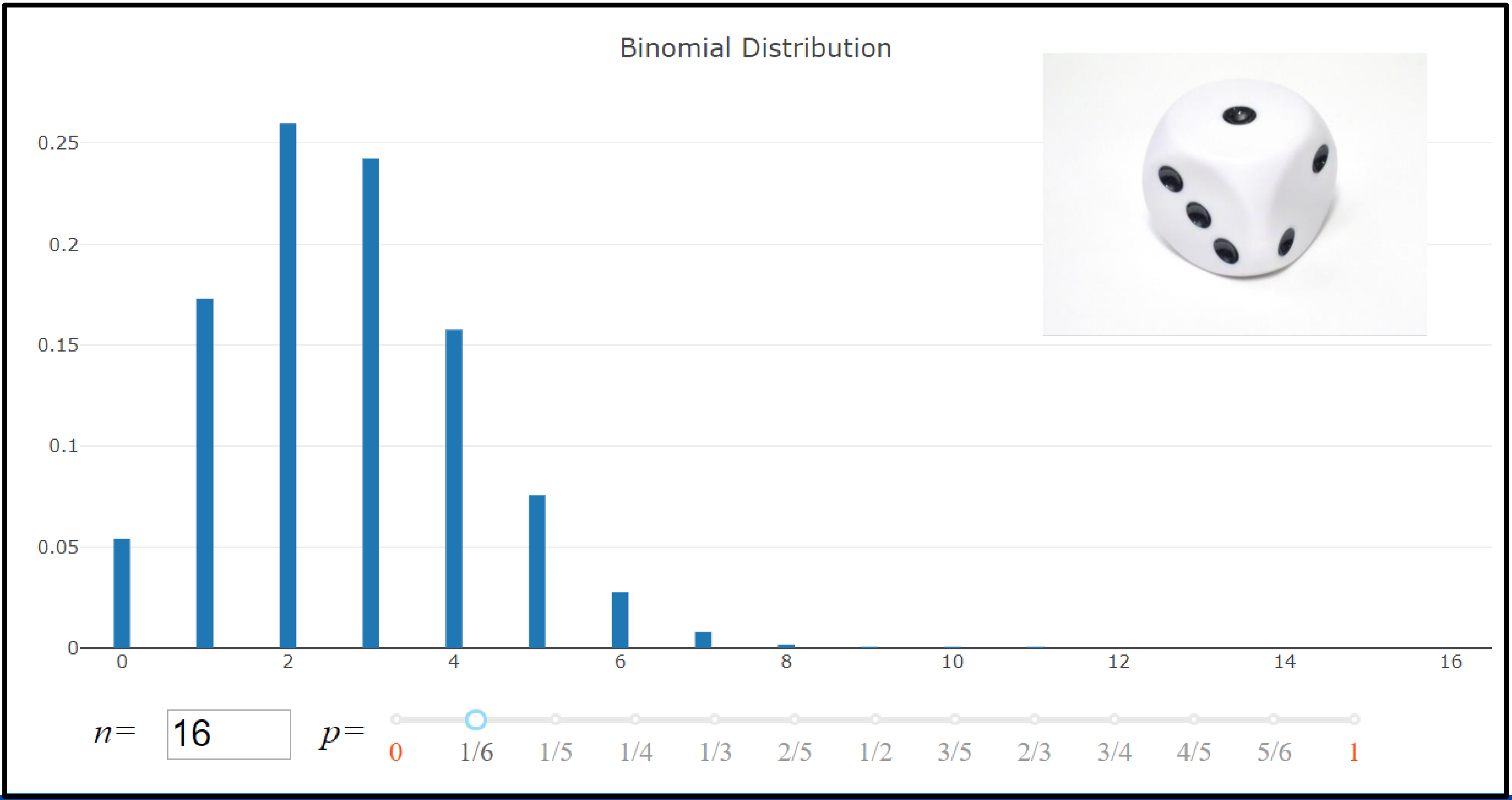
- Gibt die Wahrscheinlichkeit an, x Erfolge in n Versuchen zu beobachten
- Die Erfolgswahrscheinlichkeit für einen einzigen Versuch ist mit p bezeichnet
- Die Wahrscheinlichkeit eines Erfolgs p verändert sich nicht von Versuch zu Versuch

$$P(x\colon n, p) = \binom{n}{x}(p)^{x}(1-p)^{(n-x)}$$

Binominalverteilung

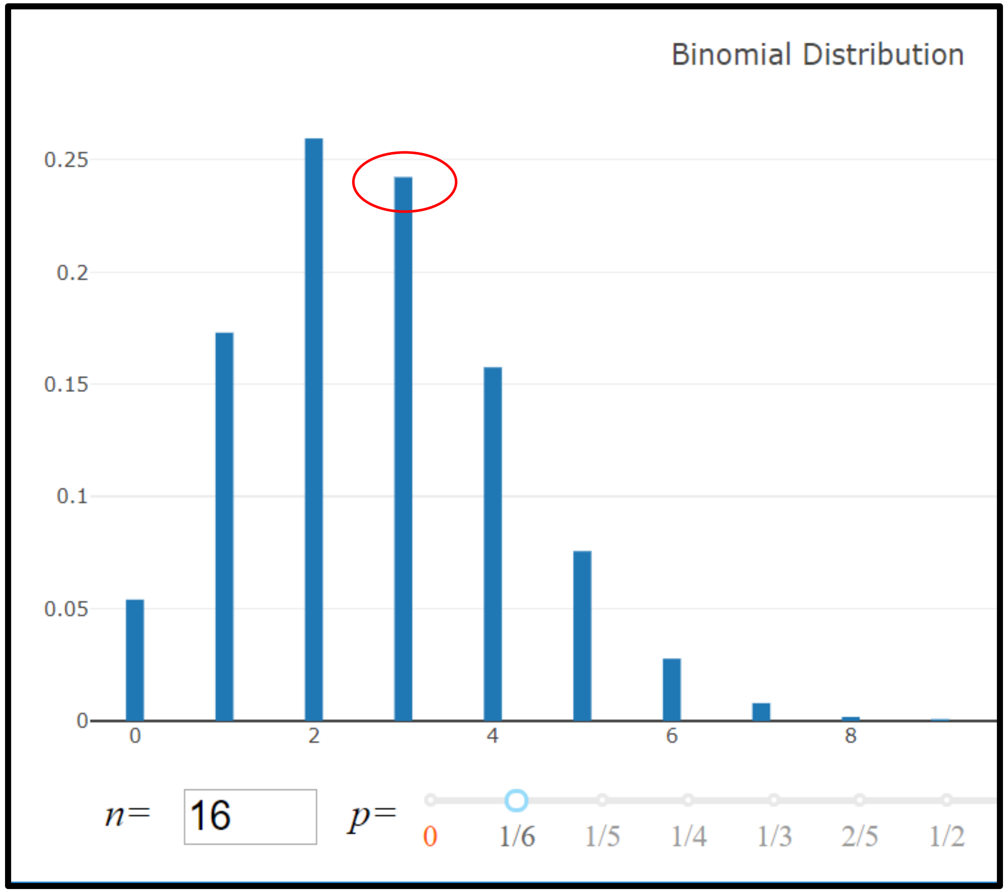


Binominalverteilung



Binominalverteilung Übung

- Wenn du einen Würfel 16 Mal würfelst, wie groß ist dann die Wahrscheinlichkeit, dass eine Fünf 3 Mal auftaucht?
- Basierend auf dem Diagramm, sollte diese bei knapp 0,25 liegen
- $x = 3, n = 16, p = 1/6$



Binominalverteilung Übung

$$P(x\colon n,p) = \binom{n}{x}(p)^x(1-p)^{(n-x)}$$

$$= \left(\frac{n!}{x!\,(n-x)!}\right)(p)^x(1-p)^{(n-x)}$$

$$= \left(\frac{16!}{3!\,(13)!}\right)(1/6)^3(5/6)^{(13)}$$

$$= \left(\frac{16\cdot 15\cdot 14}{3\cdot 2}\right)\left(\frac{1^3}{6^3}\right)\left(\frac{5^{13}}{6^{13}}\right) = 0.242$$

In Excel...

- Wenn du einen Würfel 16 Mal würfelst, wie groß ist dann die Wahrscheinlichkeit, dass eine Fünf 3 Mal auftaucht?

=BINOMDIST(3,16,1/6,FALSE)

***gibt* 0.242313760337131 zurück**

(Deutsch: **=BINOMVERT(3,16,1/6,FALSE)**)

In Python…

- Wenn du einen Würfel 16 Mal würfelst, wie groß ist dann die Wahrscheinlichkeit, dass eine Fünf 3 Mal auftaucht?

```
>>> from scipy.stats import binom
>>> binom.pmf(3,16,1/6)
0.24231376033713251
```

Poisson-Verteilung

Poisson-Verteilung

- Eine Binomialverteilung berücksichtigt die Anzahl der Erfolge aus *n* Versuchen
- Eine **Poisson-Verteilung** berücksichtigt die Anzahl der Erfolge ***pro Zeiteinheit**** im Verlauf mehrerer Ereignisse.

* oder irgendeines anderen Intervalls, z.B. **Entfernung**

Poisson-Verteilung

- Die Berechnung der Poisson-Wahrscheinlichkeitsfunktion beginnt mit einem Erwartungswert:

$$E\,(X) = \mu$$

- Dieser wird dann "Lambda" zugewiesen

$$\lambda = \frac{\#\,occurrences}{interval} = \mu$$

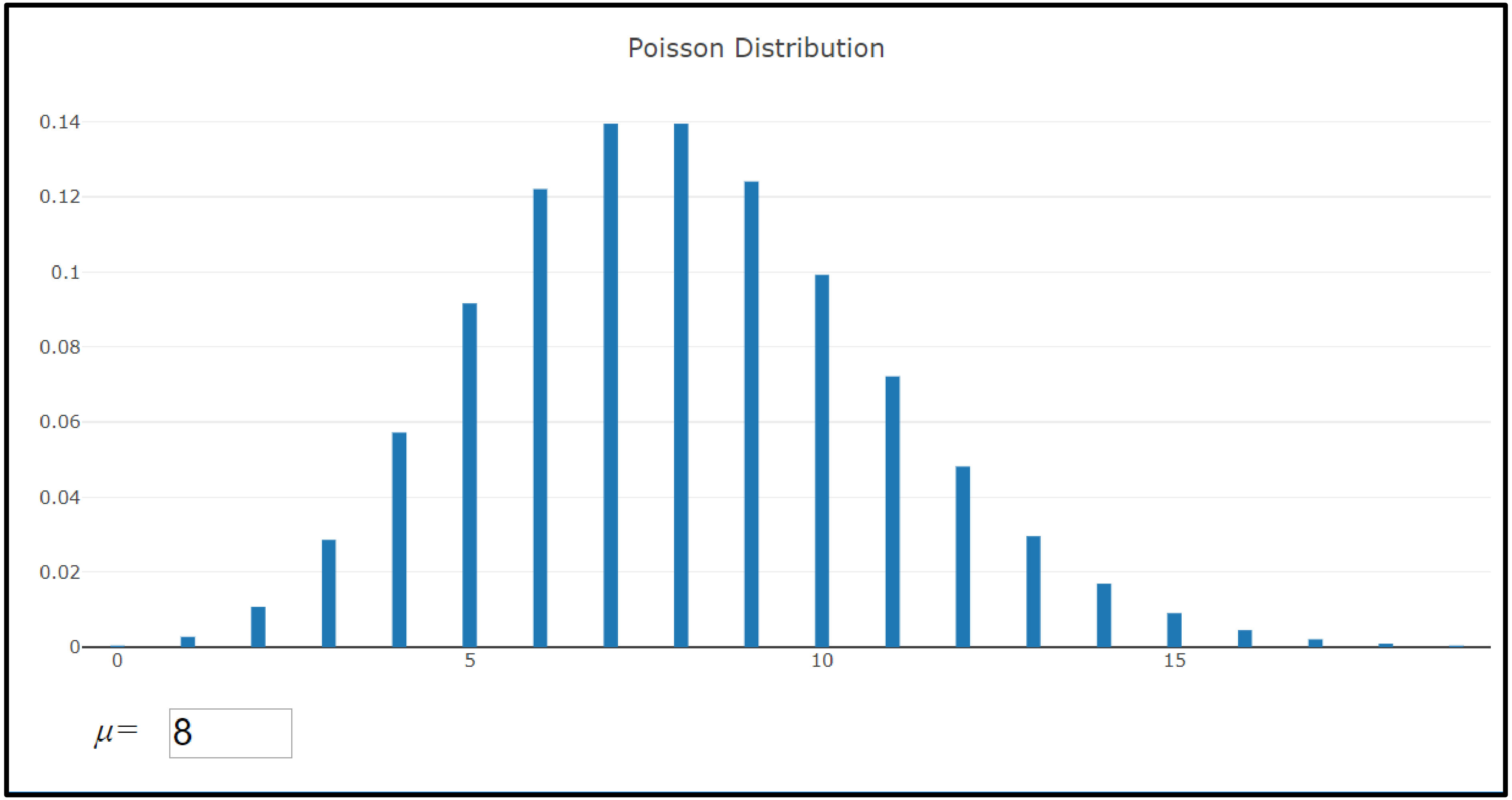
Poisson-Verteilung

- Die Gleichung der Poisson-Verteilung lautet:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

bei e = Eulersche Zahl = 2,71828…

Poisson-Verteilung



Poisson-Verteilung Übung #1

- Ein Lager erhält normalerweise am Freitag zwischen 16 Uhr und 17 Uhr 8 Lieferungen.
- Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **nur 4 Lieferungen** zwischen 16 und 17 Uhr ankommen?

Poisson-Verteilung Übung #1

$$x = 4 \quad \lambda = 8$$

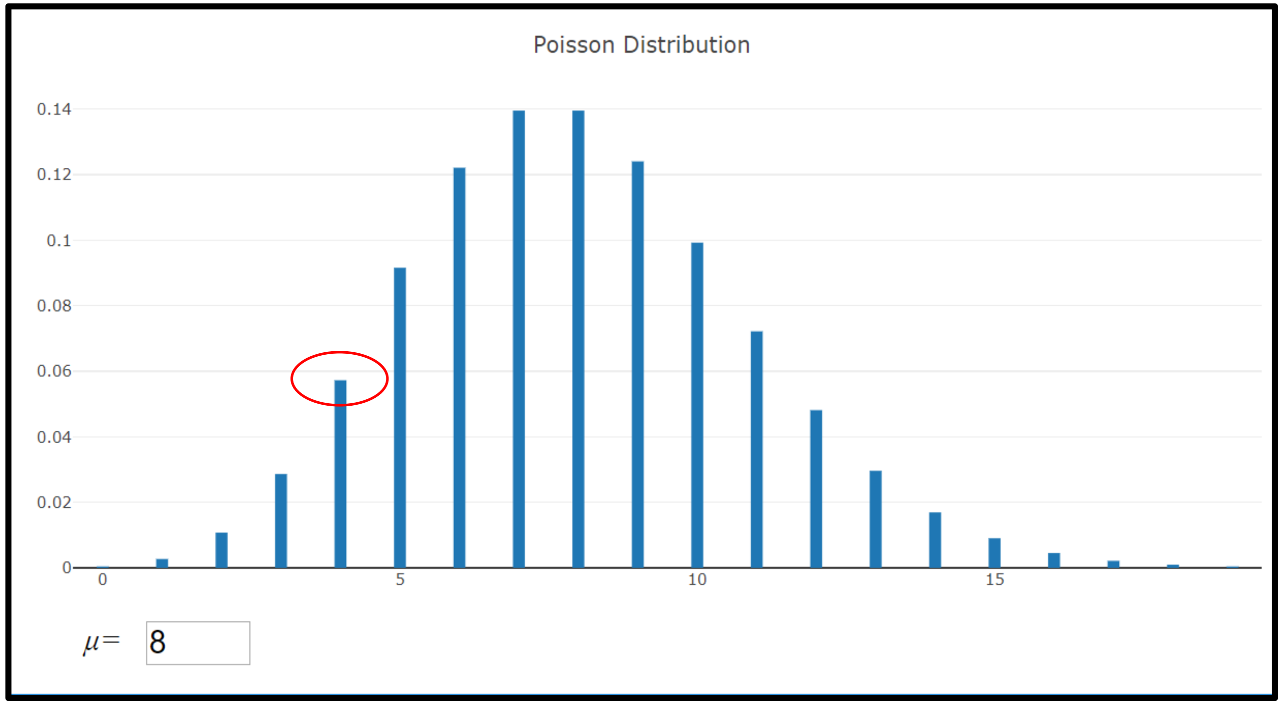
$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{8^4 \cdot 2.71828^{-8}}{4!}$$

$$= \frac{4096 \cdot \left(\frac{1}{2980.96}\right)}{24} = \mathbf{0.0572}$$

Poisson-Verteilung Übung #1

$$= \frac{4096 \cdot \left(\frac{1}{2980.96}\right)}{24} = \mathbf{0.0572}$$

Das passt zu unserem Diagramm!



Poisson-Verteilung

- Die **kumulative Verteilungsfunktion** ist einfach die Summe aller diskreten Wahrscheinlichkeiten
- Die Wahrscheinlichkeit, ***weniger als 4*** Ereignisse in einer Poisson-Verteilung zu sehen, ist:

$$P(X: x < 4) = \sum_{i=0}^{3} \frac{\lambda^i e^{-\lambda}}{i!}$$

$$= \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!}$$

Poisson-Verteilung

- Denke daran, dass die Summe aller Möglichkeiten gleich 1 ist
- Die Wahrscheinlichkeit, **mindestens** 1 Ereignis zu erhalten, ist 1 minus die Wahrscheinlichkeit, keine zu erhalten:

$$P(X: x \geq 1) = 1 - P(X: x = 0)$$

$$= 1 - \frac{\lambda^0 e^{-\lambda}}{0!} = 1 - e^{-\lambda}$$

Poisson-Verteilung Übung #2

- Ein Lager erhält normalerweise am Freitag zwischen 16 und 17 Uhr 8 Lieferungen.
- Wie groß ist die Wahrscheinlichkeit, dass **weniger als 3** an diesem Freitag zwischen 16 und 17 Uhr ankommen?

Poisson-Verteilung Übung #2

$$P(X: x < 3) = \sum_{i=0}^{2} \frac{\lambda^i e^{-\lambda}}{i!} = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!}$$

$$= \frac{8^0 \cdot 2.71828^{-8}}{0!} + \frac{8^1 \cdot 2.71828^{-8}}{1!} + \frac{8^2 \cdot 2.71828^{-8}}{2!}$$

$$= \frac{1 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{8 \cdot \left(\frac{1}{2980.96}\right)}{1} + \frac{64 \cdot \left(\frac{1}{2980.96}\right)}{2}$$

$$= \mathbf{0.0137}$$

Poisson-Verteilung - Teilintervalle

- Die Poisson-Verteilung geht davon aus, dass die Erfolgswahrscheinlichkeit während eines kleinen Zeitintervalls **proportional** zur gesamten Länge des Intervalls ist.
- Wenn du denn Erwartungswert λ über eine Stunde kennst, ist der Erwartungswert für eine Minute dieser Stunde:

$$\lambda_{minute} = \frac{\lambda_{hour}}{60}$$

Poisson-Verteilung Übung #3

- Ein Lager erhält normalerweise am Freitag zwischen 16 und 17 Uhr 8 Lieferungen.
- Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **zwischen 16:00 und 16:05** keine Lieferungen ankommen?

Poisson-Verteilung Übung #3

$$x = 0 \quad \lambda_{1\,hour} = 8$$

$$\lambda_{5\,minutes} = \frac{\lambda_{1\,hour}}{60/5} = \frac{8}{12} = 0.6667$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{0.67^0 \cdot 2.71828^{-0.6667}}{0!}$$

$$= \mathbf{0.5134}$$

In Excel…

1. Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **nur 4 Lieferungen** zwischen 16 und 17 Uhr ankommen?

   =POISSON.DIST(4,8,FALSE) *gibt* **0.057252** *zurück*

   (Deutsch: **=POISSON.VERT(4,8,FALSE)**)

2. Wie groß ist die Wahrscheinlichkeit, dass **weniger als 3** an diesem Freitag zwischen 16 und 17 Uhr ankommen?

   =POISSON.DIST(2,8,TRUE) *gibt* **0.013754** *zurück*

   (Deutsch: **=POISSON.VERT(2,8,TRUE)**)

In Excel…

3. Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **zwischen 16:00 und 16:05** keine Lieferungen ankommen?

   =POISSON.DIST(0,8/12,FALSE) *gibt* **0.513417** *zurück*

   (Deutsch: **=POISSON.VERT(0,8/12,FALSE)**)

In Python...

1. Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **nur 4 Lieferungen** zwischen 16 und 17 Uhr ankommen?

```
>>> from scipy.stats import poisson
>>> poisson.pmf(4,8)
0.057252288495362
```

In Python…

2. Wie groß ist die Wahrscheinlichkeit, dass **weniger als 3** an diesem Freitag zwischen 16 und 17 Uhr ankommen?

```
>>> from scipy.stats import poisson
>>> poisson.cdf(2,8)
0.013753967744002971
```

In Python…

3. Wie hoch ist die Wahrscheinlichkeit, dass an diesem Freitag **zwischen 16:00 und 16:05** keine Lieferungen ankommen?

```
>>> from scipy.stats import poisson
>>> poisson.pmf(0,8/12)
0.513417119032592
```

Kontinuierliche Wahrscheinlichkeits-verteilung

Kontinuierliche Wahrscheinlichkeitsverteilung

- Kontinuierliche Wahrscheinlichkeitsverteilungen werden auch **Wahrscheinlichkeitsdichtefunktionen** (en.:*probability density function PDF*) genannt:

Normalverteilung

Exponentialverteilung

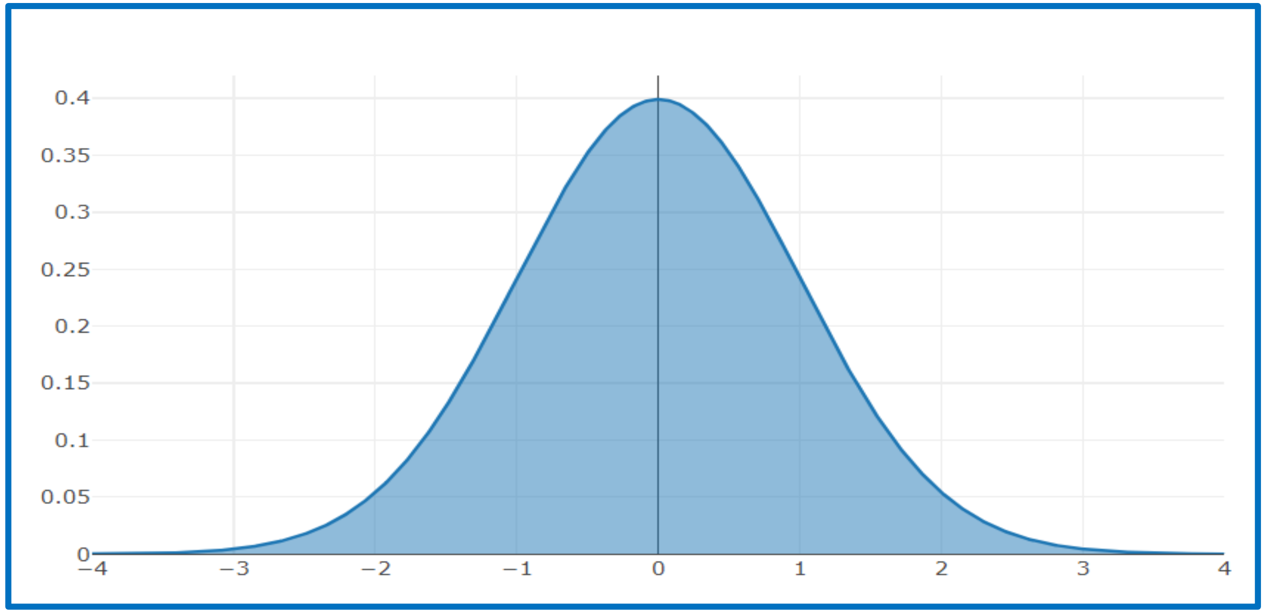
Betaverteilung

Normalverteilung

- Viele reale Datenpunkte folgen einer normalen Verteilung:
  - Größe und Gewicht von Menschen
  - Blutdruck der Bevölkerung
  - Prüfungsergebnisse
  - Messfehler

Normalverteilung

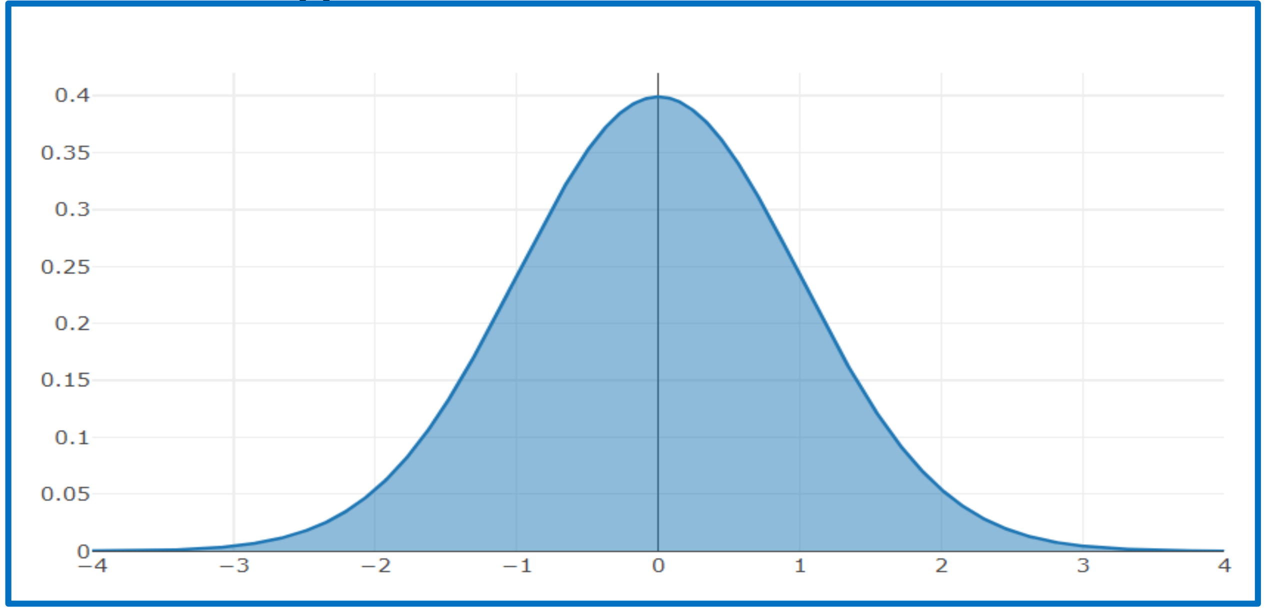
- Diese Datenquellen bewegen sich in der Regel in der Nähe eines zentralen Werts ohne Schiefe nach links oder rechts und nähern sich einer "Normalverteilung" an:



Normalverteilung

Normalverteilung

- Wir verwenden eine **stetige Verteilung**, um das Verhalten dieser Datenquellen zu modellieren.
- Beachten Sie die durchgehende Linie und den Bereich in dieser WDF.

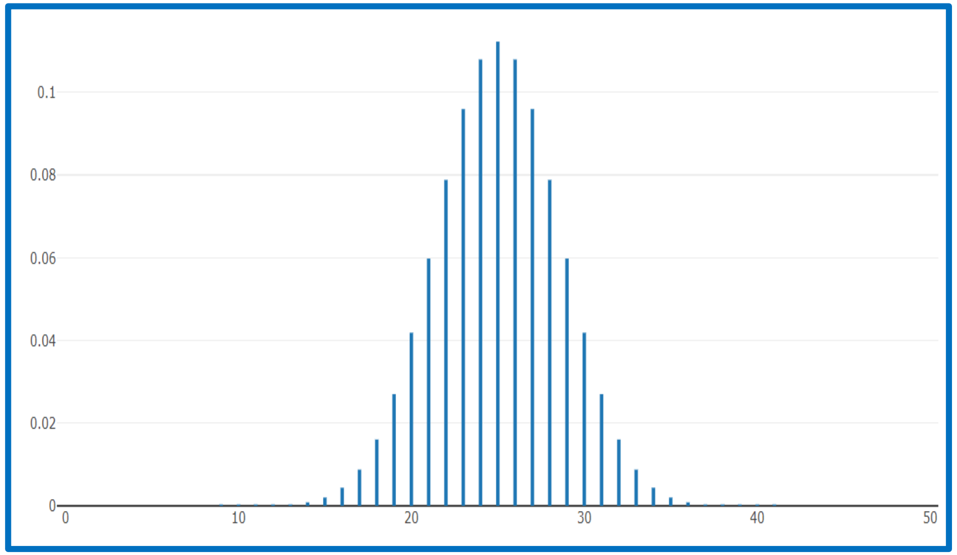


Normalverteilung

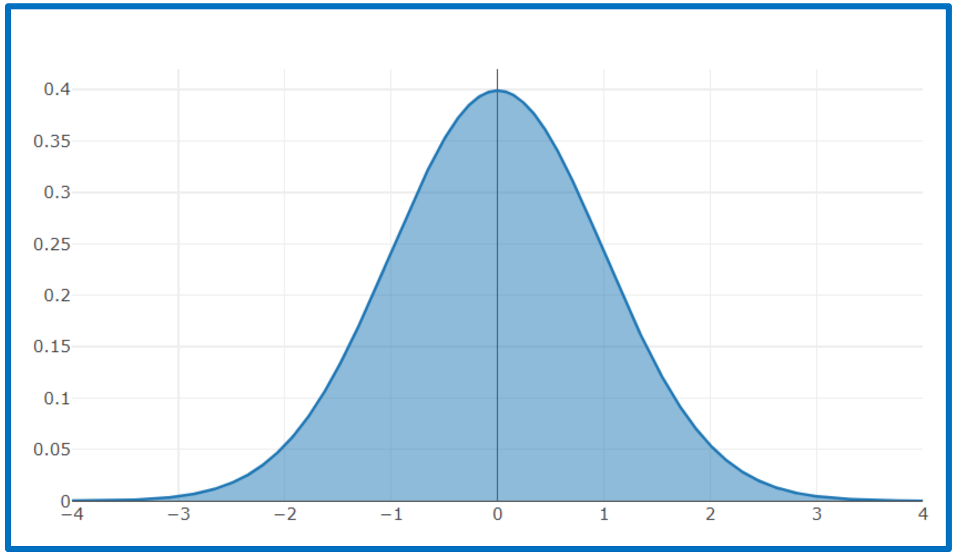
Normalverteilung

- Im Gegensatz zu **diskreten Verteilungen**, bei denen die Summe aller Striche gleich eins ist, ist bei einer **Normalverteilung** die *Fläche unter der Kurve* gleich eins:

Binominalverteilung



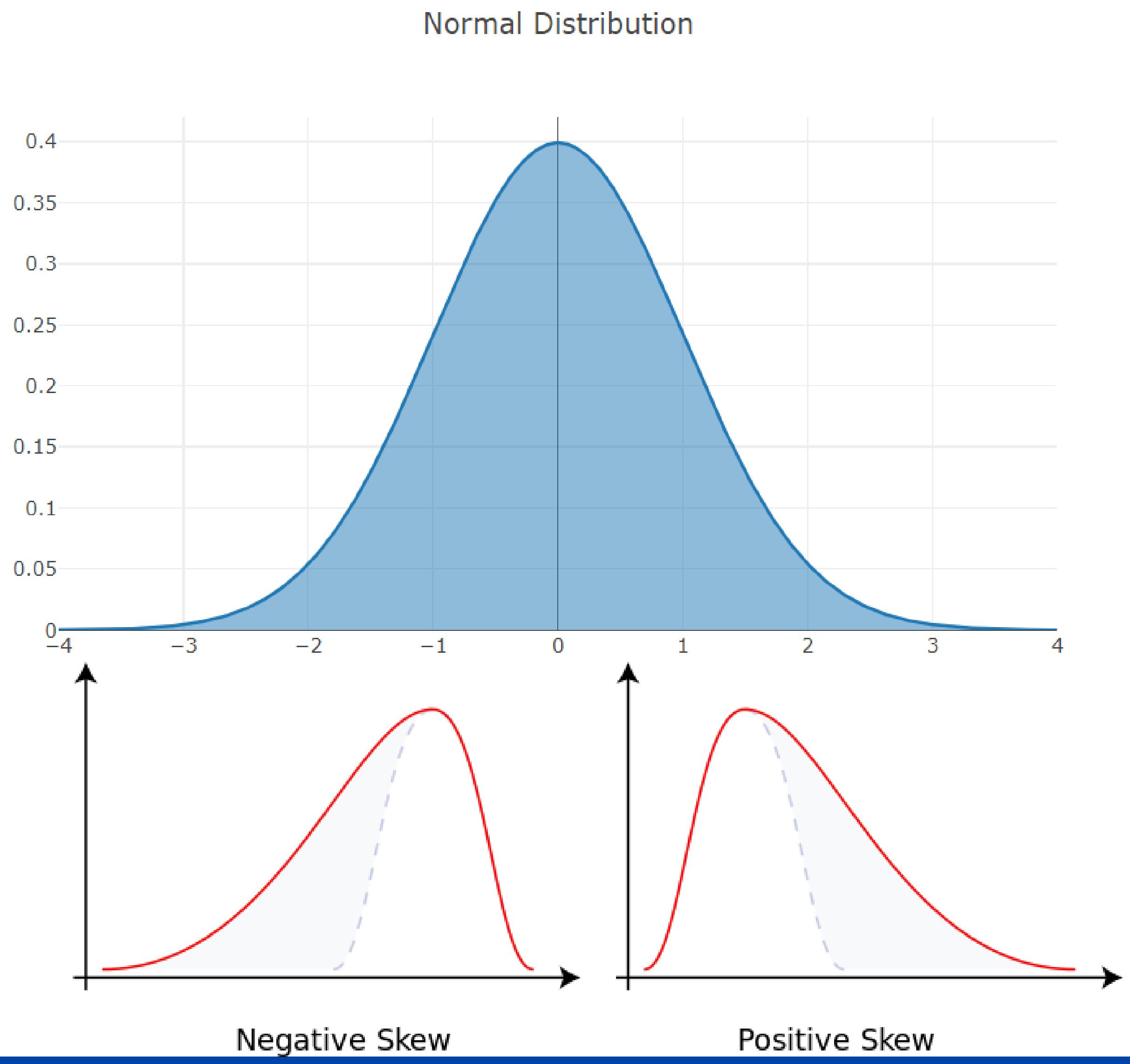
Normalverteilung



Normalverteilung

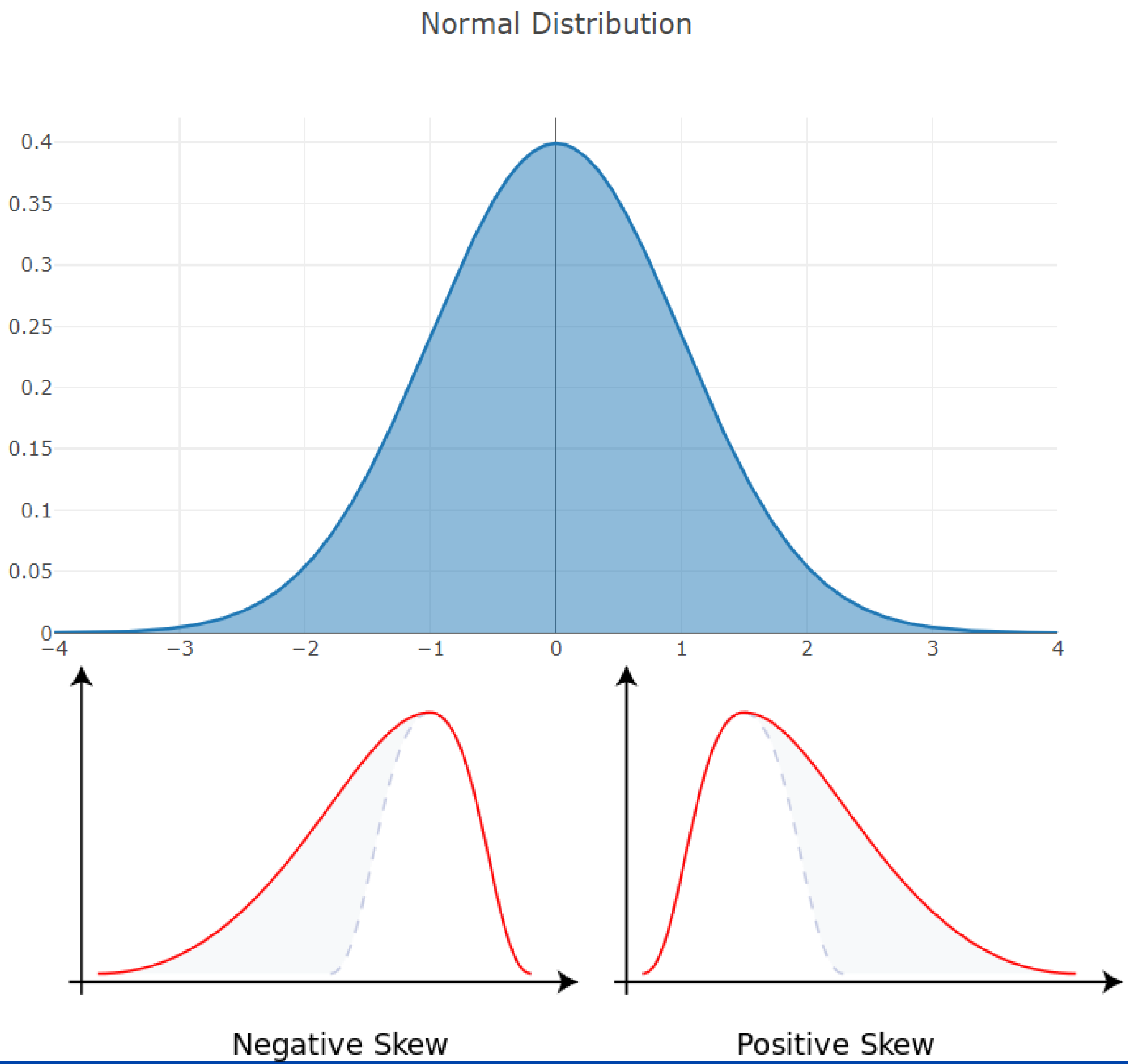
- auch als **Glockenkurve** oder **Gaußsche Normalverteilung** bezeichnet
- immer symmetrisch

Asymmetrische Kurven sind **schief** und **nicht** **normalverteilt**

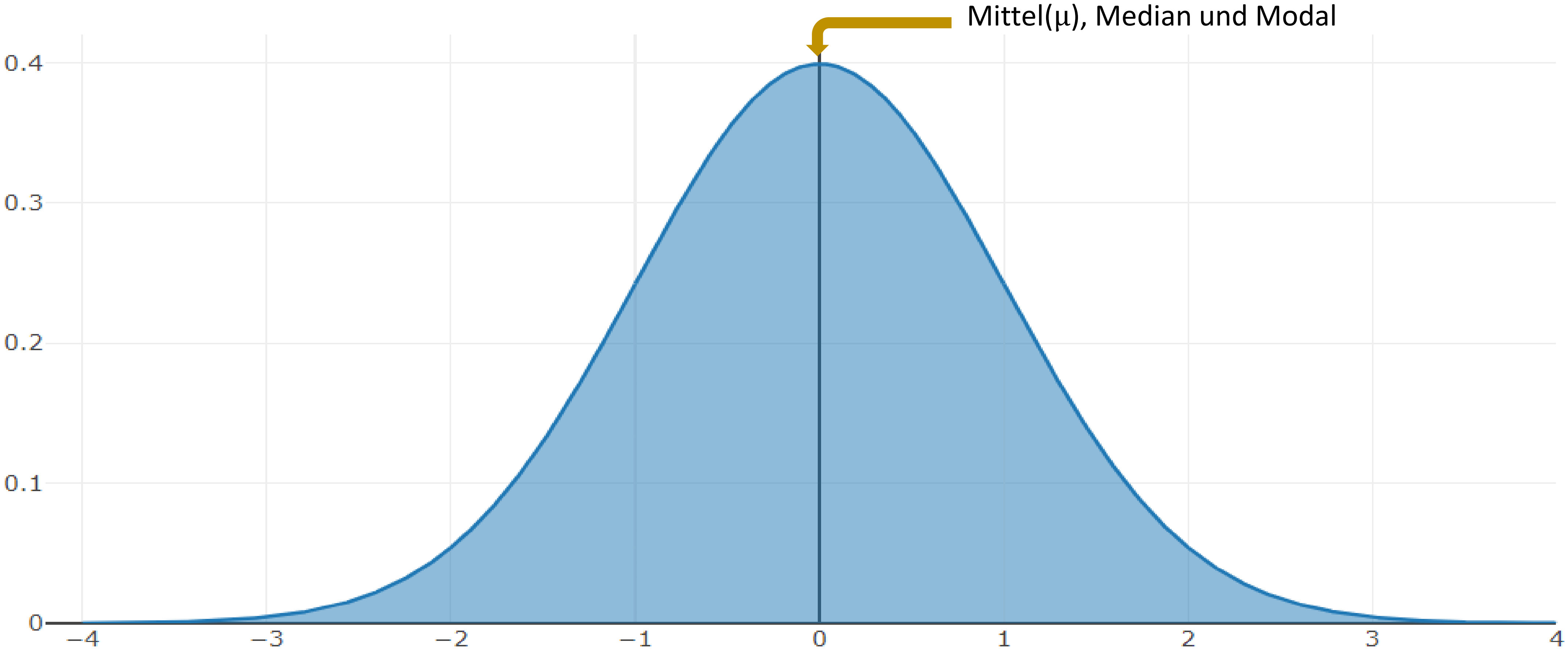


Normalverteilung

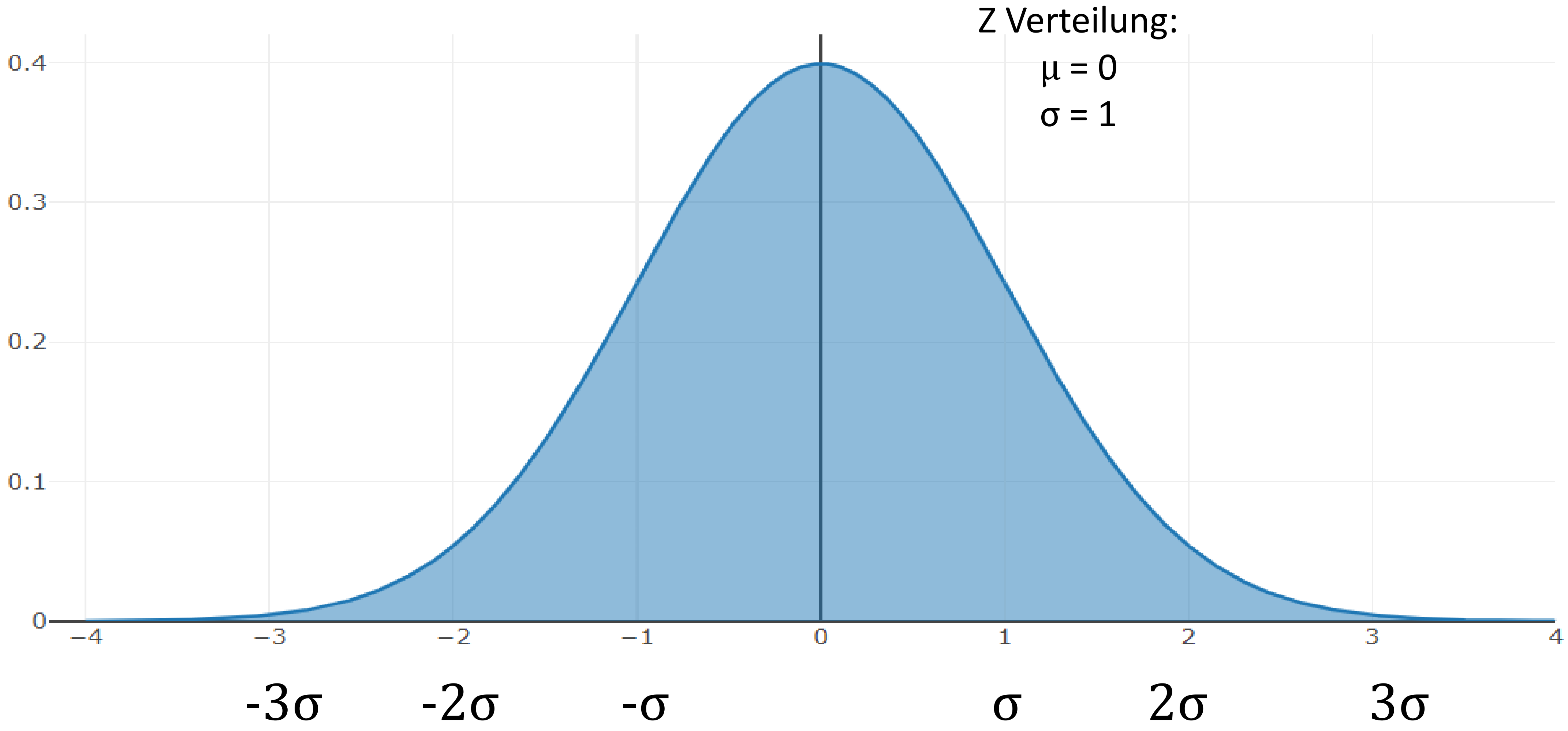
- die **Punktwahrscheinlichkeit** ist null
- Wir können Wahrscheinlichkeiten nur über einen **Wertintervall** oder eine Reihe von Ergebnissen finden



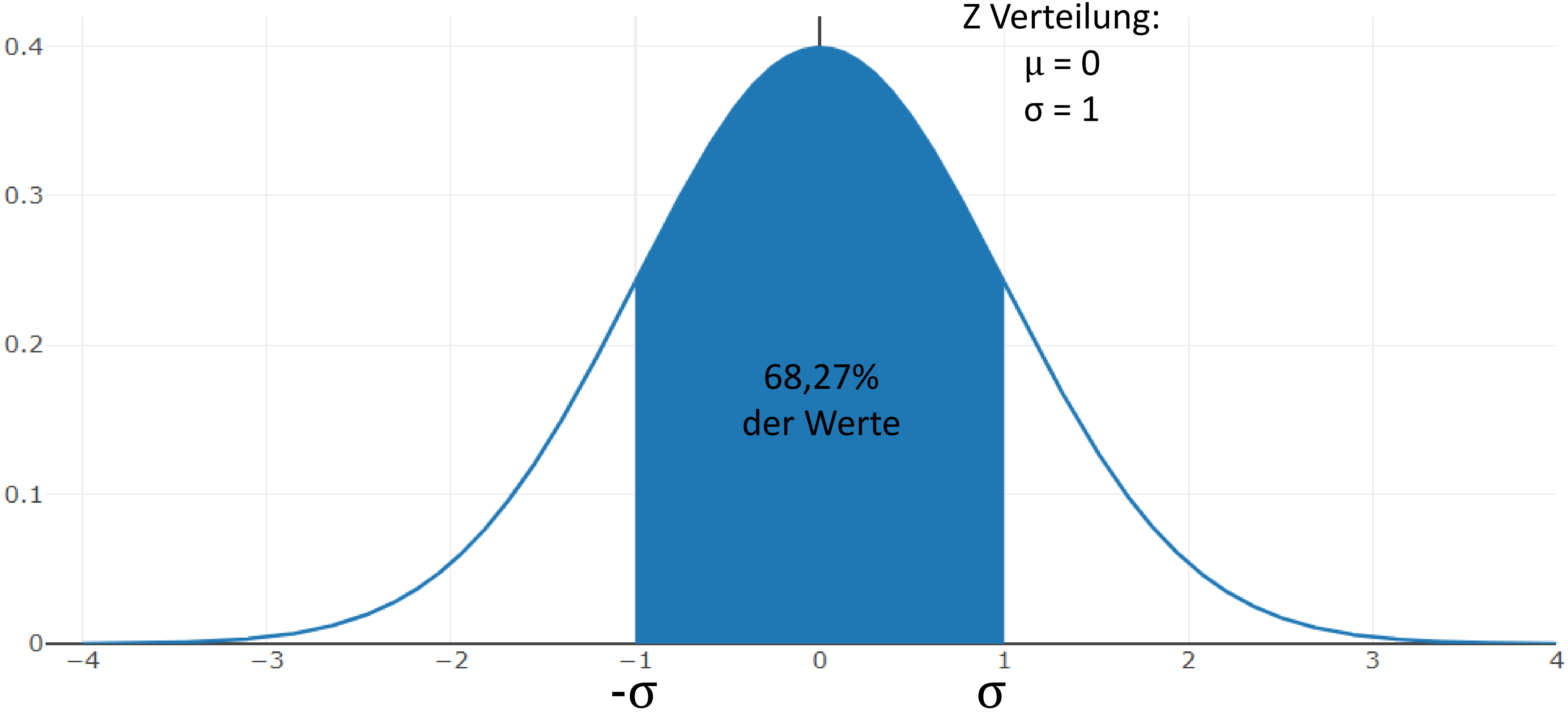
Normalverteilung



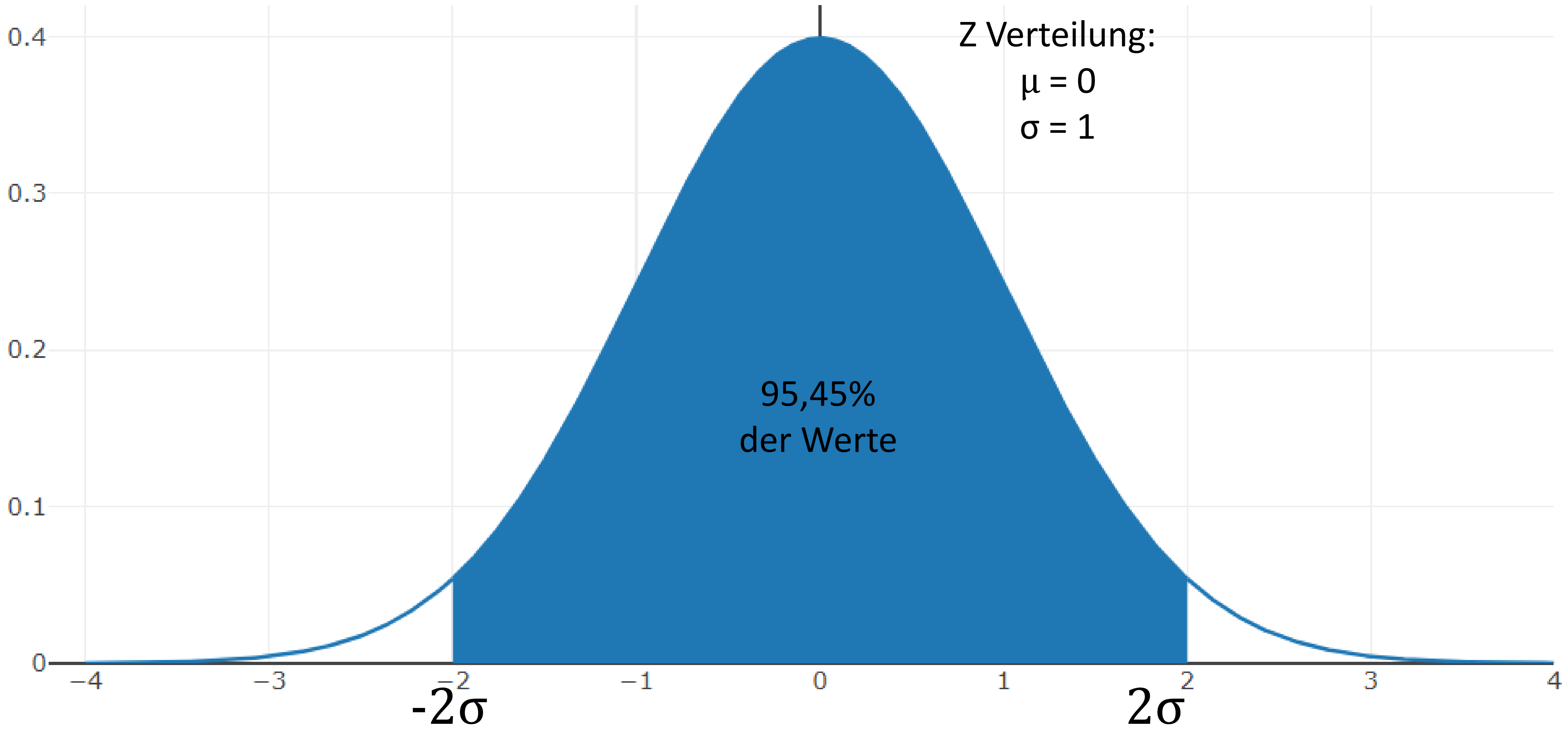
Standardnormalverteilung



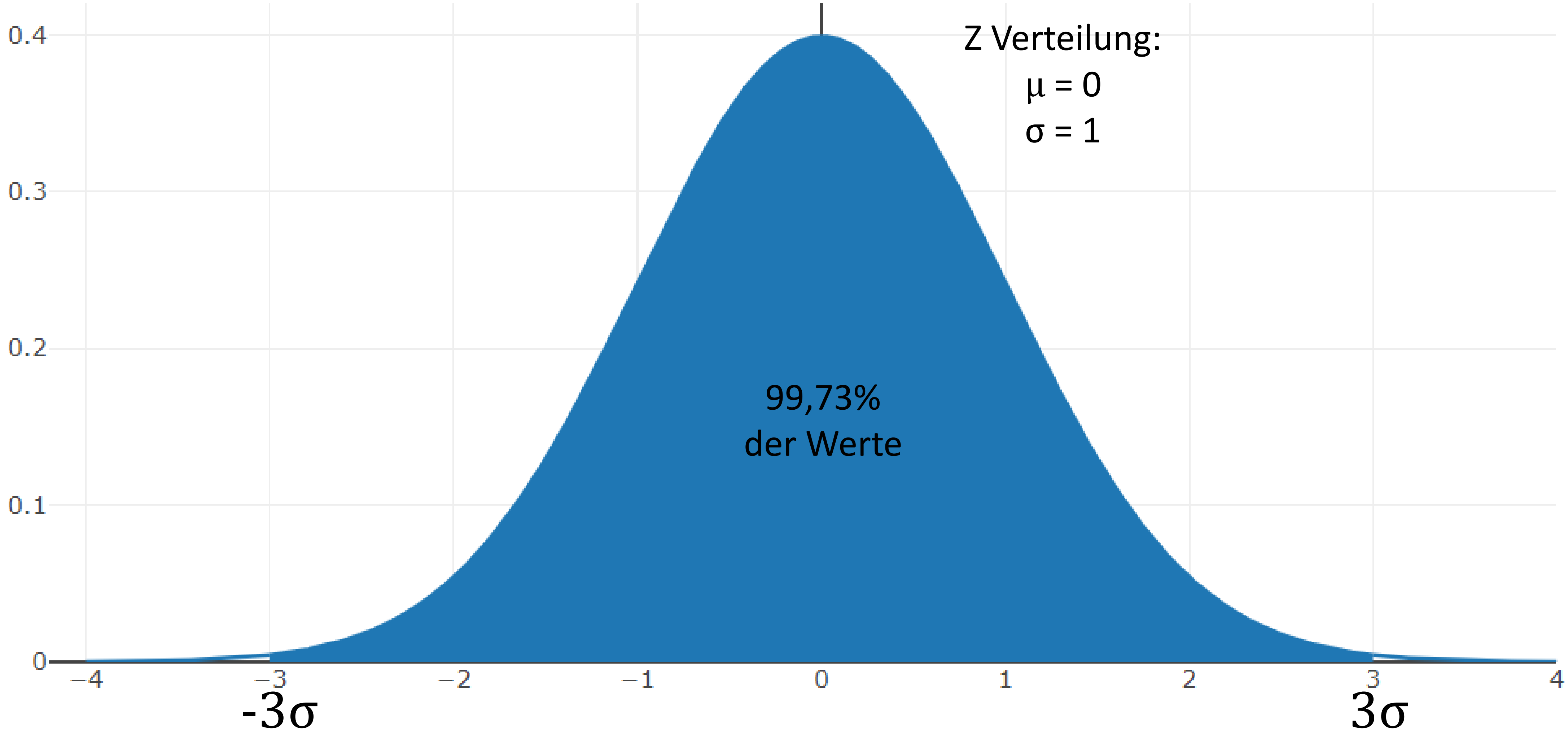
Standardnormalverteilung



Standardnormalverteilung



Standardnormalverteilung

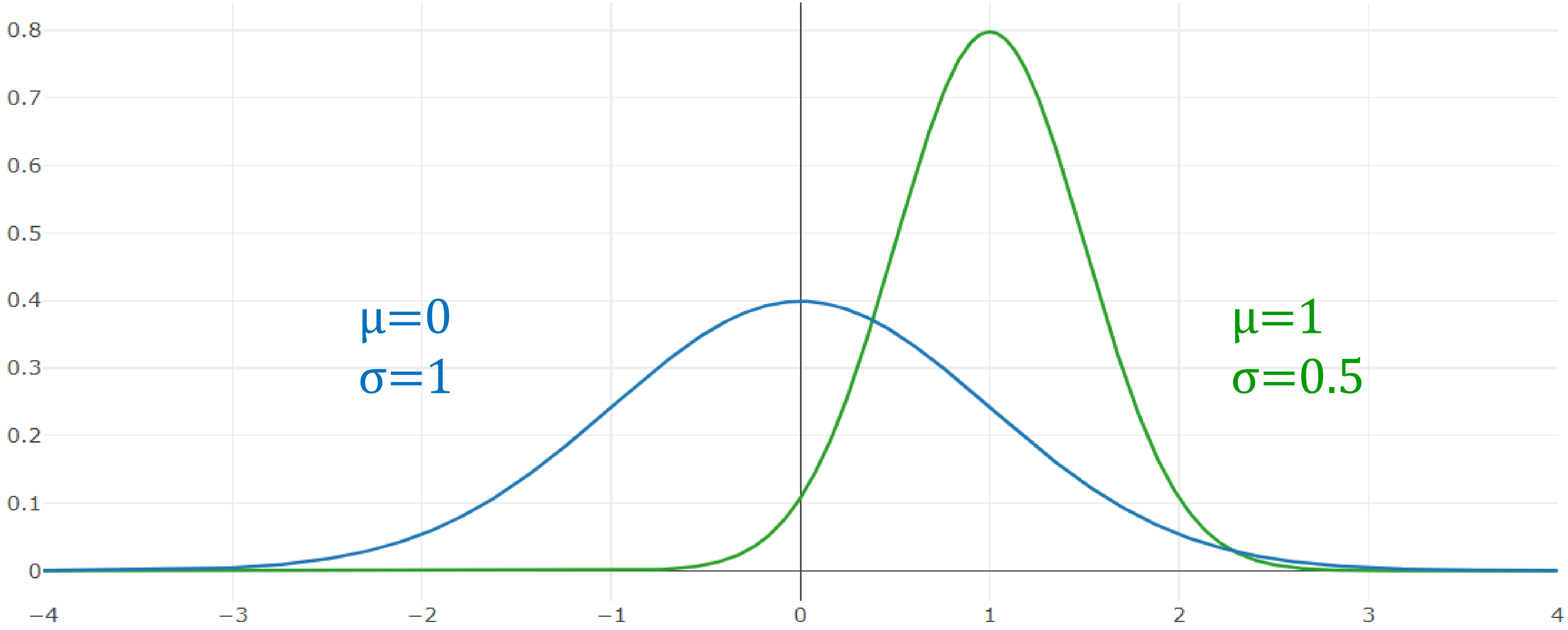


Normalverteilung

- Alle normalverteilten Kurven weisen das gleiche Verhalten auf:
  - Symmetrisch zum Mittelwert (Erwartungswert)
  - 99,73% der Werte liegen innerhalb von drei Standardabweichungen
- Der Mittelwert muss jedoch nicht null sein, und σ muss nicht eins sein.

Formel Normalverteilung

- Andere Populationen können auch normalverteilt sein:



Normalverteilung

- Wenn wir feststellen, dass sich eine Population einer Normalverteilung annähert,
- dann können wir einige wichtige Schlüsse darüber ziehen, wenn wir wissen, dass es sich um den Mittelwert (Erwartungswert) und eine Standardabweichung handelt

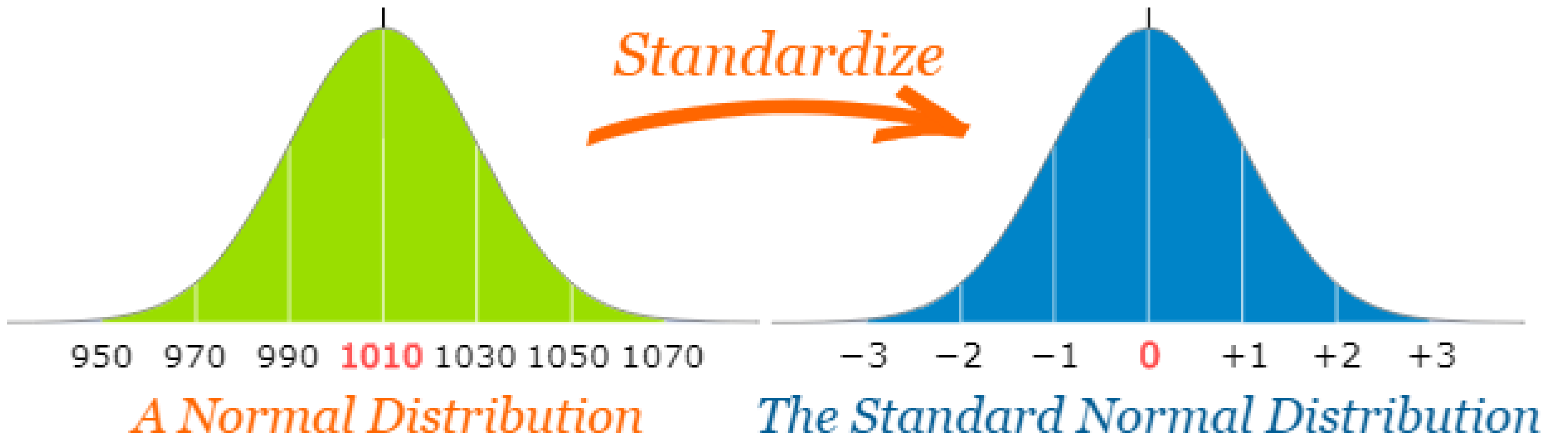
Formel der Normalverteilung und Z-Wert

Normalverteilung

- Im Kapitel über die Statistik dieses Kurses werden Stichproben-, Standardfehler- und Hypothesentests zur Auswertung von Experimente verwendet.
- Ein großer Teil dieser Prozedur besteht darin, zu verstehen, wie man eine Normalverteilung "standardisiert".

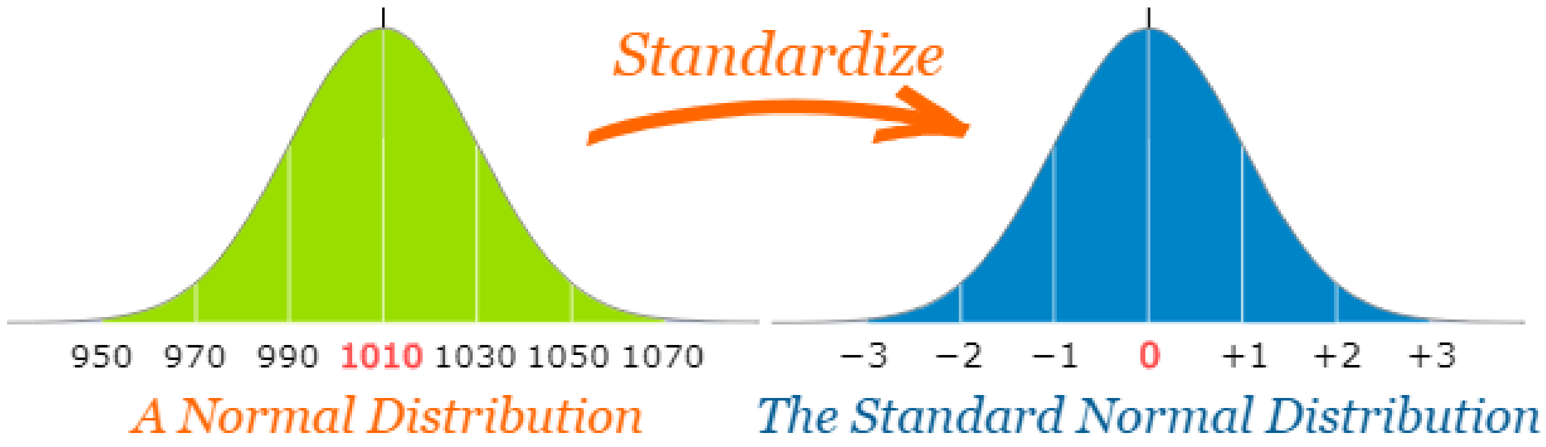
Normalverteilung

- Wir können jede Normalverteilung nehmen und sie auf eine normale Standardnormalverteilung standardisieren.



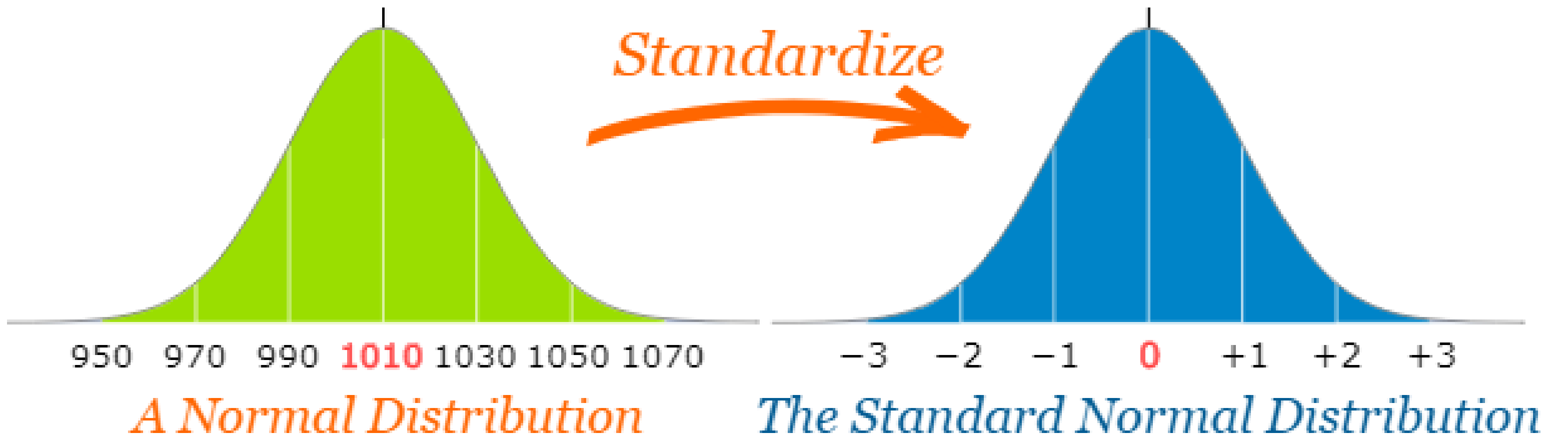
Normalverteilung

- Wir werden in der Lage sein, jeden Wert aus einer Normalverteilung zu nehmen und ihn durch einen Z-Wert zu standardisieren.



Normalverteilung

- Mit diesem Z-Wert können wir dann das Perzentil eines bestimmten x-Wertes berechnen.



Normalverteilung

- Zur kurzen Erinnerung: ein Perzentil sagt aus, "welcher Prozentsatz unter diesen Wert fällt".
- Das bedeutet, dass ein Perzentil-Wert von 95 Prozent angibt, dass 95 Prozent aller anderen Datenpunkte unter diesen Wert fallen.

Normalverteilung

- Zum Beispiel:

  wenn ein Schüler die Note 1,5 erhält und diese Note im Perzentil 90 liegt, dann wissen wir, dass 90% aller anderen Schüler schlechter als eine 1,5 erreicht haben.

Normalverteilung

- Wenn wir unsere Daten als Normalverteilung modellieren können, können wir die Werte in der Normalverteilung in eine Standardnormalverteilung konvertieren, um ein Perzentil zu berechnen.

Normalverteilung

- Zum Beispiel können wir eine Normalverteilung von Testwerten mit einer Mittelwert- und Standardabweichung haben.
- Wir können dann einen Z-Wert verwenden, um das Perzentil eines bestimmten Testergebnisses herauszufinden.

Normalverteilungsformel

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

μ = Erwartungswert

σ = Standardabweichung

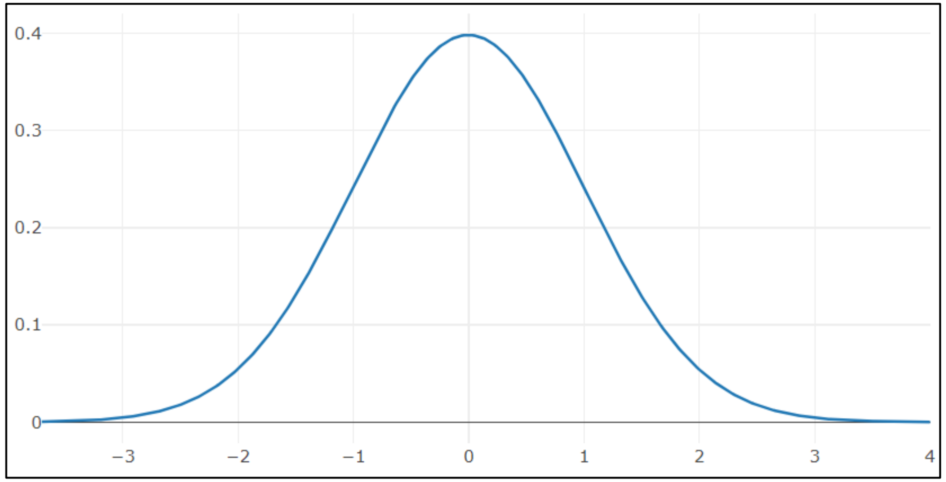
$e = 2.71828$

$\pi = 3.14159$

Normalverteilungsformel

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

So entstand unsere grafische Darstellung
mit einem Erwartungswert von 0
und einer Standardabweichung von 1:



Z-Wert und Z-Tabelle

- Um einen Einblick in einen bestimmten Wert x in einer normalverteilten Populationen zu erhalten, *standardisieren* wir x indem wir einen Z-Wert berechnen:

$$z = \frac{x-\mu}{\sigma}$$

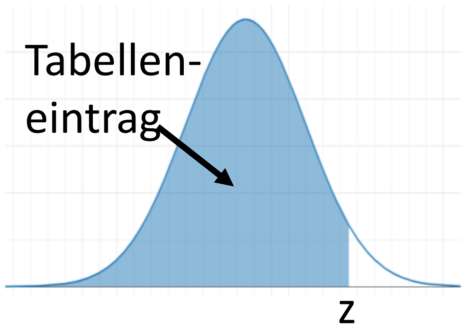
- Wir können dann das Perzentil von x bestimmen, indem wir auf eine Z-Tabelle (Standardnormalverteilungstabelle) zurückgreifen

Wie du eine Z-Tabelle ließt

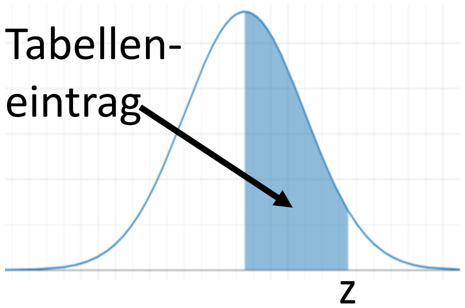
- Eine Z-Tabelle von einer **standardnormalverteilten Wahrscheinlichkeit** bildet einen bestimmten Z-Wert im Bereich links vom Wert unter einer Normalverteilungskurve ab.
- Da die Gesamtfläche unter der Kurve 1 ist, sind die Wahrscheinlichkeiten durch 0 und 1 begrenzt

Wie du eine Z-Tabelle ließt

- Verschiedene Tabellen dienen verschiedenen Zwecken:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141

Z-Werte in Excel

- In Microsoft Excel geben die folgenden Funktionen Z-Scores und Wahrscheinlichkeiten aus:

Input	Input Wert	Formel	Output	Output Wert
z	0.70	=NORMSDIST(B2) =NORMVERT(B2)	p	0.758036
p	0.95	=NORMSINV(B3) =STANDNORMINV(B3)	z	1.644854

Z-Werte in Python

```
>>> from scipy import stats
>>> z = .70
>>> stats.norm.cdf(z)
0.758036347776927
>>> p = .95
>>> stats.norm.ppf(p)
1.6448536269514722
```

Z-Wert Übung

- Ein Unternehmen sucht einen neuen Datenbankadministrator.
- Du gibst den Bewerbern einen standardisierten Test, um ihr technisches Wissen zu testen.
- Deine erste Bewerberin, Maria, bekommt 87 Punkte
- Ist Maria aufgrund ihrer Punktzahl außergewöhnlich qualifiziert?

Z-Wert Übung

- Um zu entscheiden, wie gut ein Bewerber abgeschnitten hat, müssen wir alle Bewerber betrachten.
- Basierend auf Tausenden von früheren Tests, wissen wir, dass der Mittelwert 75 von 100 ist, mit einer Standardabweichung von 7 Punkten.

Z-Wert Übung - Lösung

- Konvertiere zuerst Maria‘s Wert in einen standardisierten Z-Wert mithilfe der Formel:

$$z = \frac{x-\mu}{\sigma}$$

$$= \frac{87-75}{7} = \mathbf{1.7143}$$

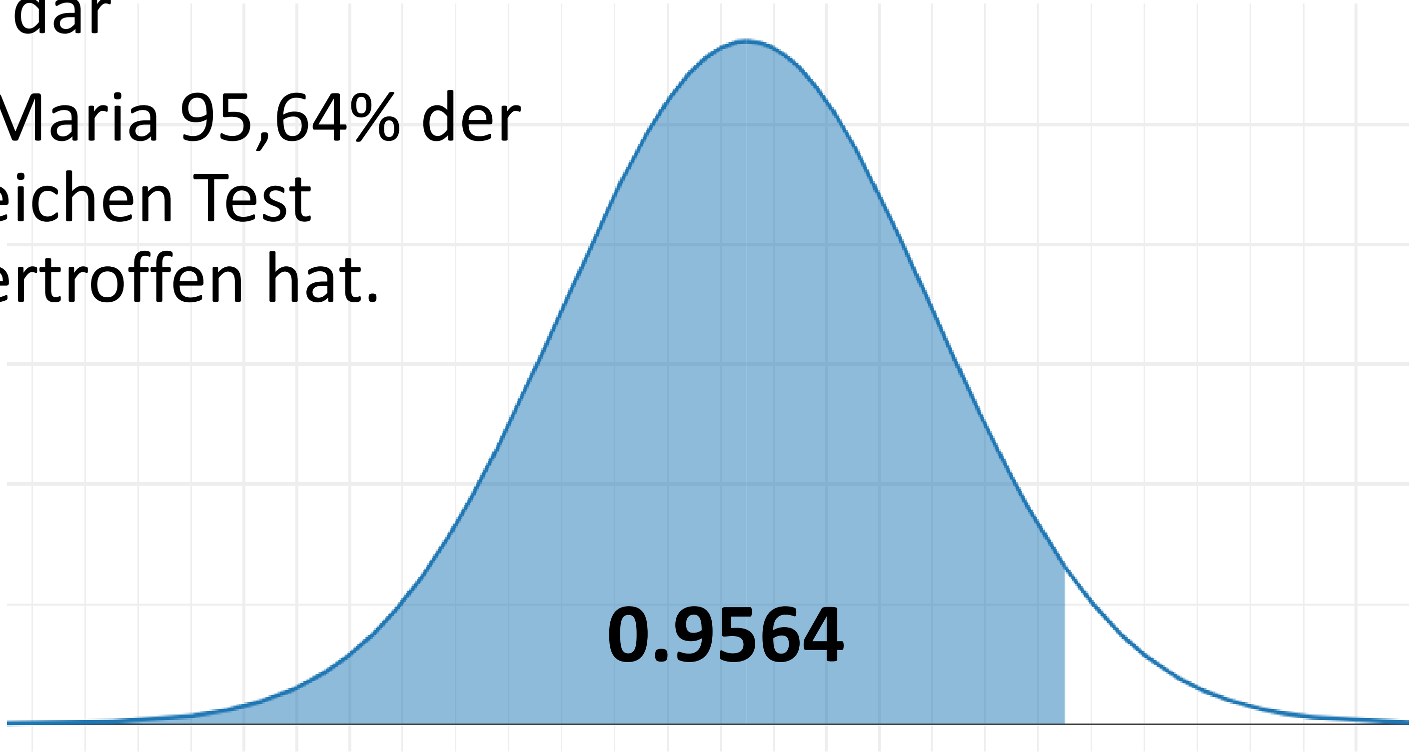
Z-Wert Übung - Lösung

- Suche als nächstes 1,7143 in einer Z-Tabelle:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Z-Wert Übung - Lösung

- 0.9564 stellt den Bereich links von Marias Testergebnis dar
- Das bedeutet, dass Maria 95,64% der anderen, die den gleichen Test gemacht haben, übertroffen hat.



Als nächstes: Statsistik