KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Datenbeschaffung und -aufbereitung:
04_03Aufbereitung_Wahrscheinlichkeitstheorie

Einführung in die Wahrscheinlichkeitstheorie

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele

Nach diesem Video kannst du ...

- die Begriffe Wahrscheinlichkeit, Verteilung, Erwartungswert, Standardabweichung und Quantil an einem Beispiel erklären.
- Eigenschaften der Normalverteilung nennen.
- vorgegebene Daten mit einem Boxplot visualisieren.

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

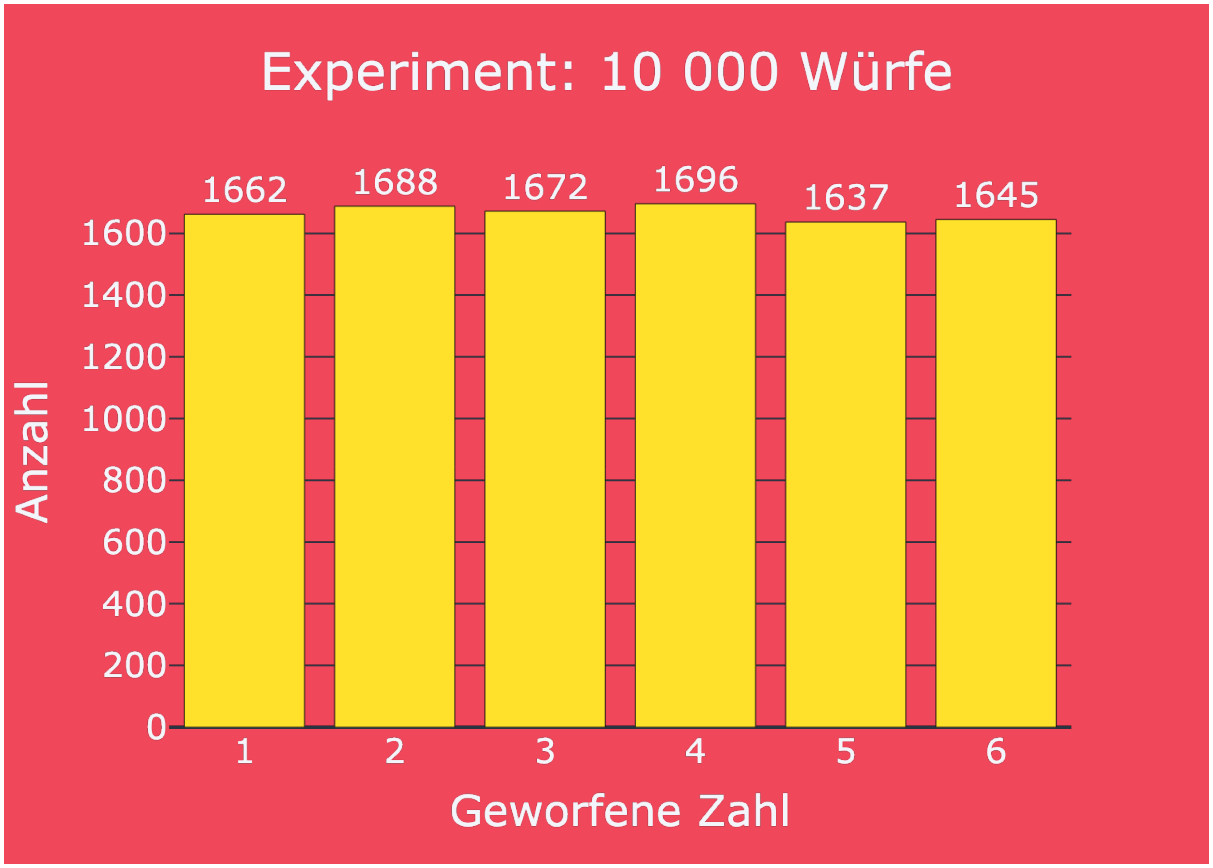
Inhalt

Mit dem Wort „Wahrscheinlichkeit” kannst du sicher aus deinem Alltag schon etwas anfangen. Doch was versteht man in der Mathematik und Statistik eigentlich unter Wahrscheinlichkeit? Und wie hängt das mit unseren Daten zusammen?

Wahrscheinlichkeiten

Eine Wahrscheinlichkeit ist mathematisch gesehen erst einmal nur eine reelle Zahl zwischen 0 und 1. Wenn du diese Zahl mit 100 multiplizierst, kommst du dann zu der dir wohl bekannten Darstellung in Prozent. Dabei bedeutet eine Wahrscheinlichkeit von 1, dass ein Ereignis garantiert eintreten wird, und 0, dass es garantiert nicht eintreten wird.

Wenn wir 10 000-mal mit einem 6-seitigen Würfel würfeln und zählen, wie häufig jede Zahl auftaucht, kann das Ergebnis zum Beispiel so aussehen.

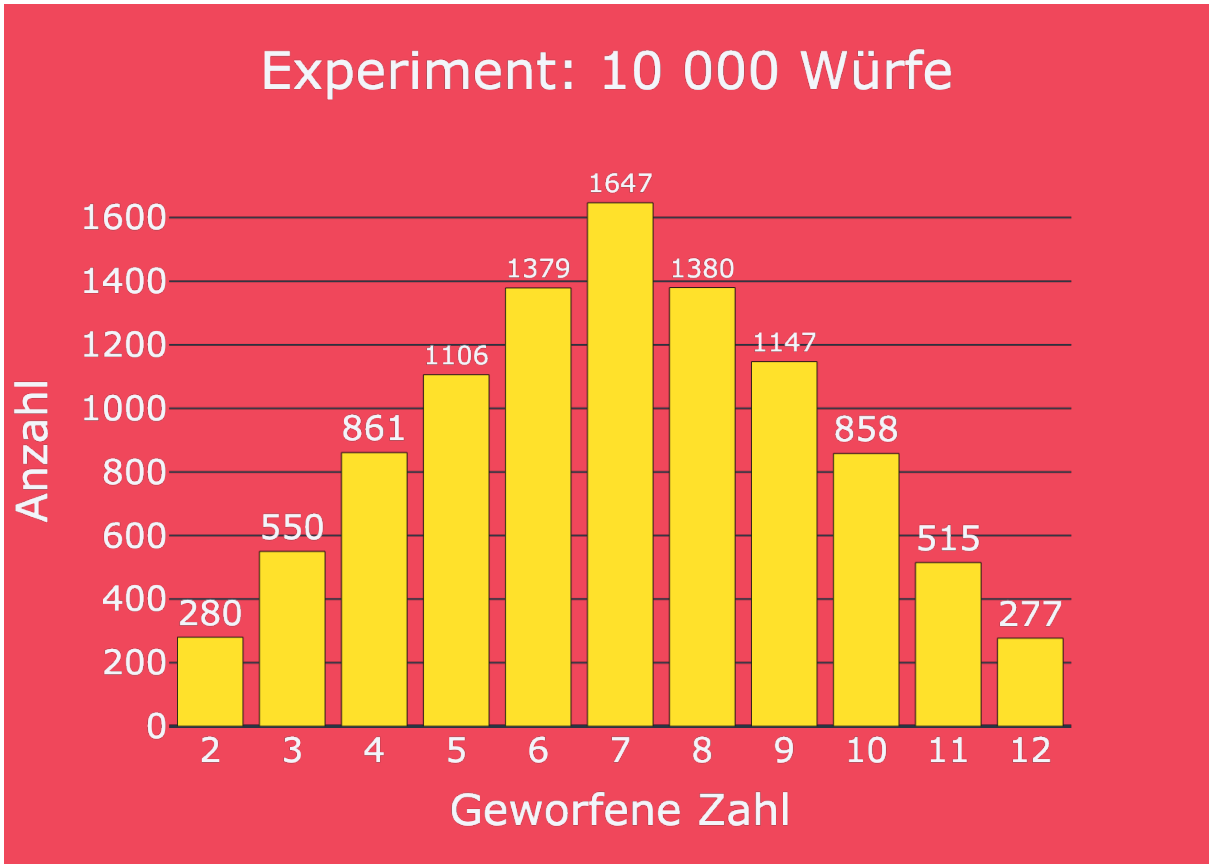


Bei diesem Diagramm handelt es sich um ein Histogramm. Die Werte auf der y-Achse werden mit einer Aggregierungsfunktion gebildet, hier die Anzahl der jeweils geworfenen Zahlen. Zu erwarten ist, dass jede Zahl ungefähr gleich häufig vorkommt, da z. B. die Wahrscheinlichkeit, eine 1 zu würfeln genauso hoch sein sollte, wie eine 6 zu würfeln. Andernfalls wäre der Würfel gezinkt. Die Wahrscheinlichkeit für jede Zahl beträgt also bei einem 6-seitigen Würfel 1/6. Du siehst am Histogramm, dass tatsächlich jeder Wert auf dem Würfel ungefähr gleich häufig gewürfelt wurde. Wenn wir aber vorab die

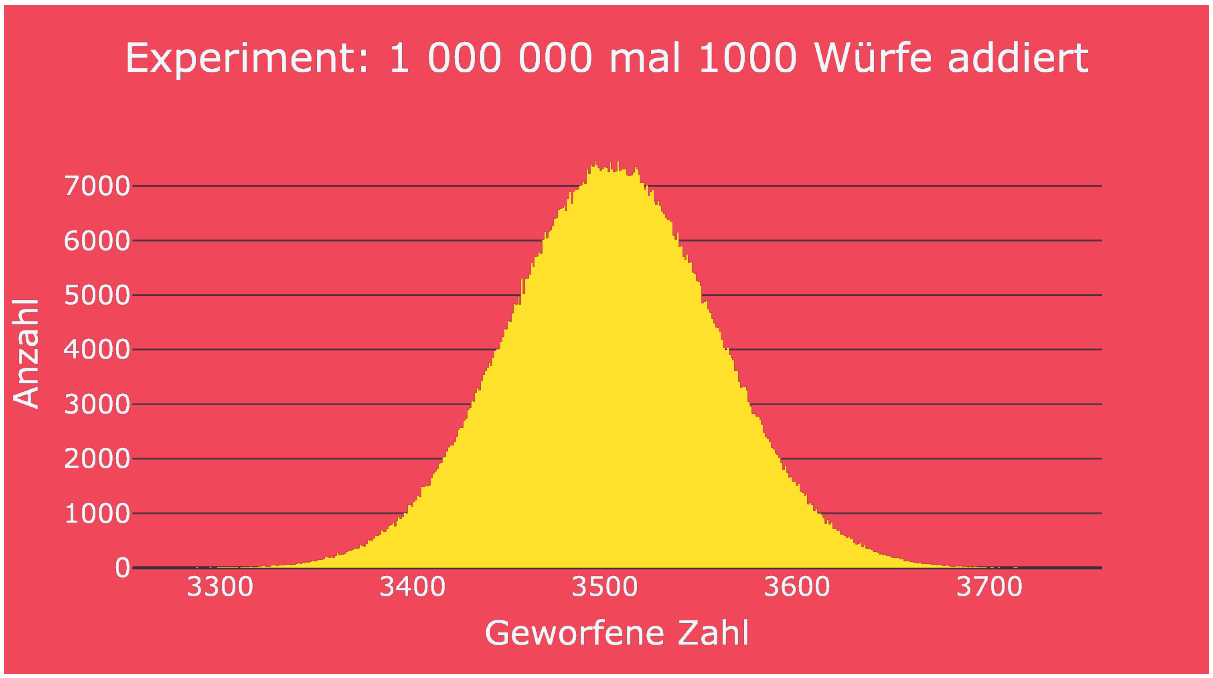
Wahrscheinlichkeiten nicht wissen und die Wahrscheinlichkeit, eine 1 zu würfeln, aus den beobachteten Häufigkeiten im Histogramm berechnen wollen, müssen wir die Anzahl an gewürfelten Einsen (hier 1662) durch die Anzahl an Würfen insgesamt (10 000) teilen, was zu einem Ergebnis von 0,1662 oder 16,62 % führt. Je häufiger wir würfeln, desto mehr nähert sich die empirisch gemessene Wahrscheinlichkeit der „wahren" Wahrscheinlichkeit von 1/6 = 0, 167 an.

Normalverteilung

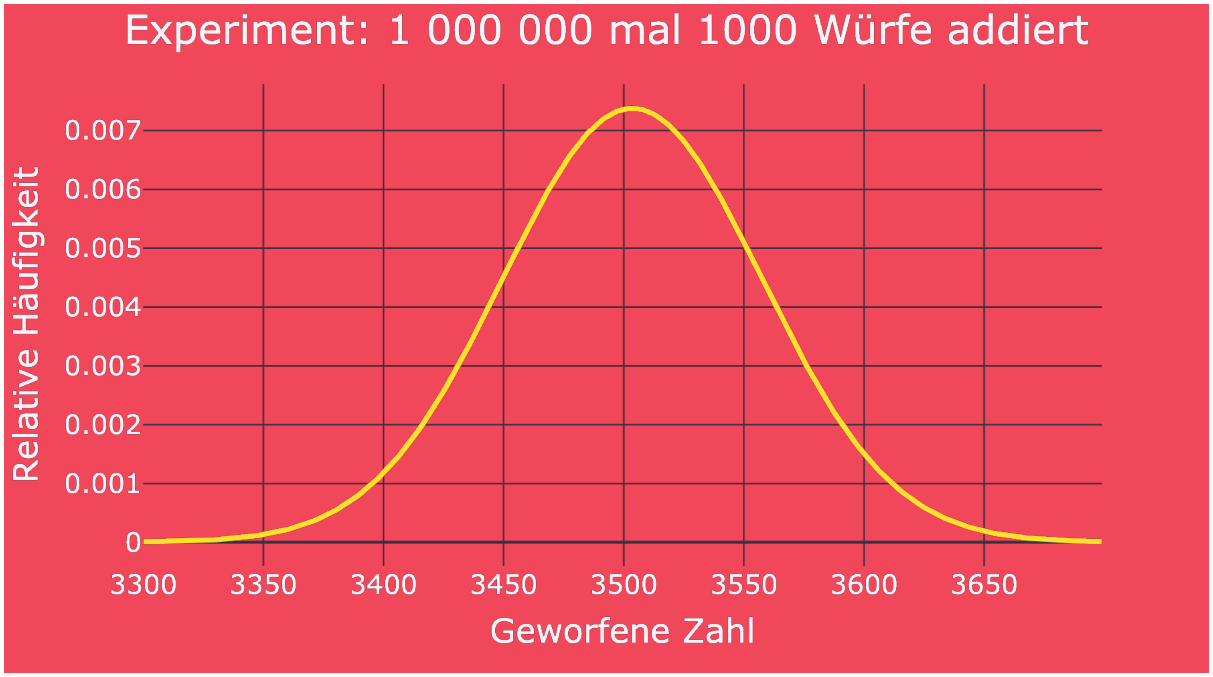
Starten wir doch jetzt mal ein neues Experiment, in dem wir zwei Würfel nehmen und die Summe berechnen. Unsere Ereignisse bzw. Messwerte sind also die Summe zweier geworfener Würfel. Das Diagramm sieht schon anders aus, oder?



Hier ist nicht mehr jedes Ereignis gleich wahrscheinlich. So kann zum Beispiel eine Summe von 7 durch die Würfelpaare (1,6), (2,5), (3,4), (4,3), (5,2) und (6,1) erreicht werden, also durch sechs Kombinationen, wohingegen 12 nur durch (6,6), also eine einzige Kombination erreicht wird. Mit derselben Rechnung wie eben beträgt die Wahrscheinlichkeit für eine 7 0,1647 oder 16,47 %, die Wahrscheinlichkeit für die Summe 12 hingegen nur 0,0277, also 2,77 %.
Je häufiger wir würfeln und desto mehr Würfel wir gleichzeitig verwenden, desto mehr nähert sich unser Diagramm einer Glockenform an.



Tauschen wir jetzt noch die absoluten Häufigkeiten der y-Achse durch relative Häufigkeiten aus d. h., teilen alle Werte der y-Achse durch die Anzahl der Gesamtwürfe, und glätten die entstehende Kurve, dann erhalten wir diese Kurve.



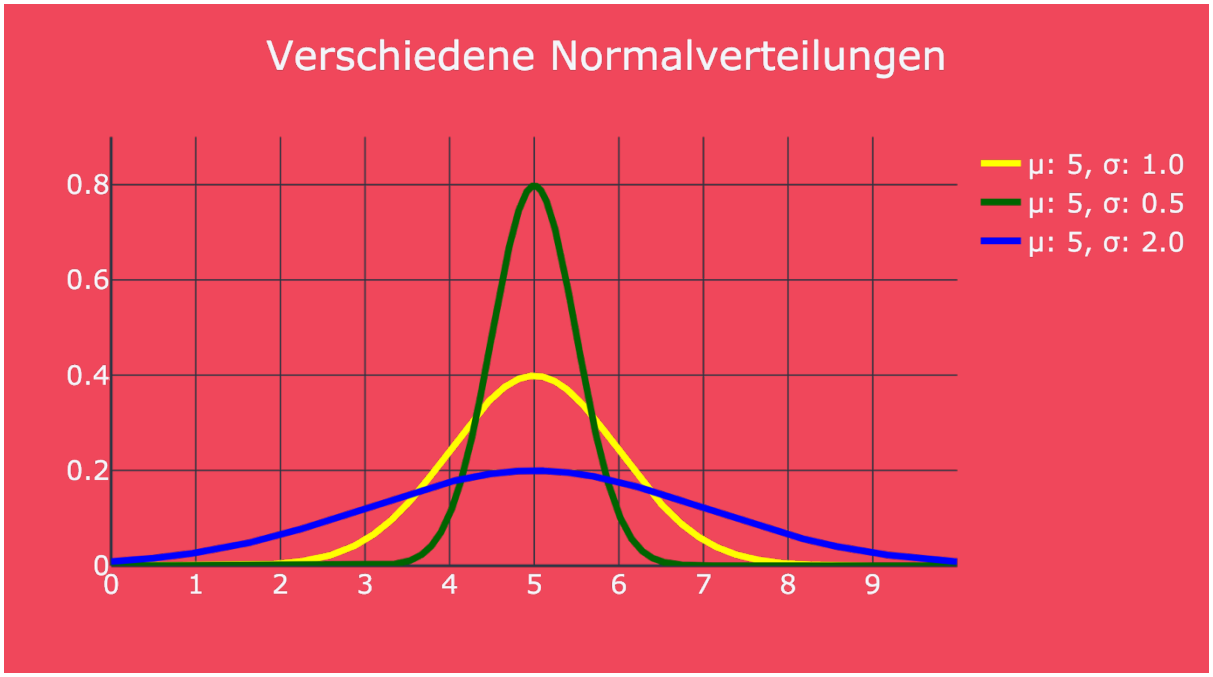
Diese sogenannte „Glockenkurve” beschreibt eine Normalverteilung oder Gaußsche Verteilung, benannt nach dem deutschen Mathematiker Carl Friedrich Gauß. Eine Verteilung

ist eine Funktion, die jedem Ereignis (hier: Summe der Würfe) eine Wahrscheinlichkeit zuordnet. Bei der Normalverteilung sind Ereignisse in der Mitte wahrscheinlicher als an den Rändern.

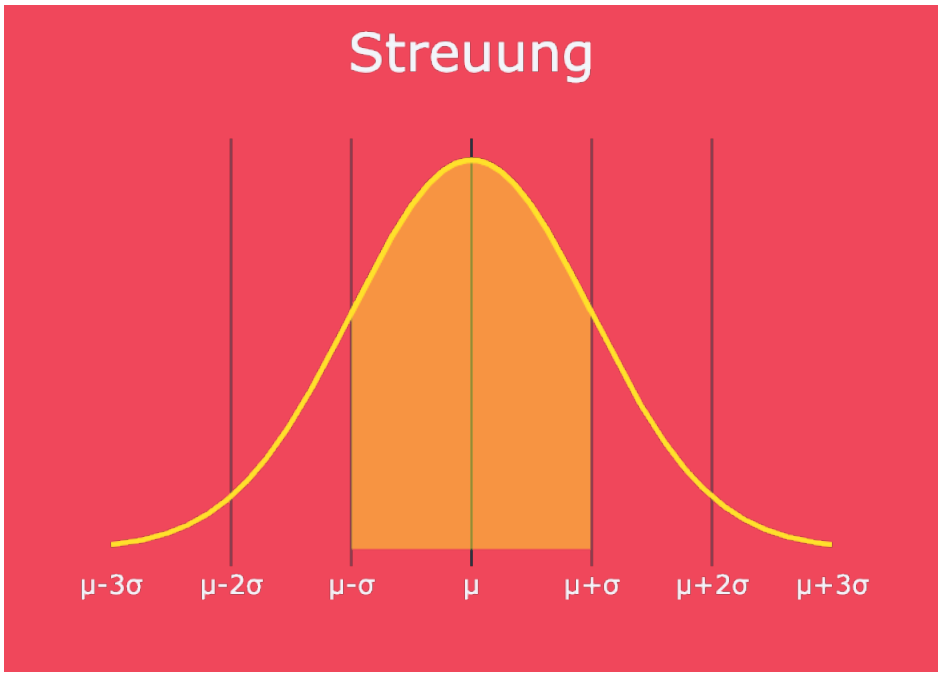
Mittelwert und Standardabweichung

Wenn wir ein Experiment wie die Würfelwürfe häufig wiederholen, immer das Ergebnis notieren und am Ende den Mittelwert bilden, erhalten wir den Wert, den wir im Schnitt beim Experiment erwarten können. Dieser Mittelwert nennt sich passenderweise auch Erwartungswert und wird mit μ („mü") bezeichnet. Bei der Normalverteilung befindet er sich am höchsten Punkt der Glocke.

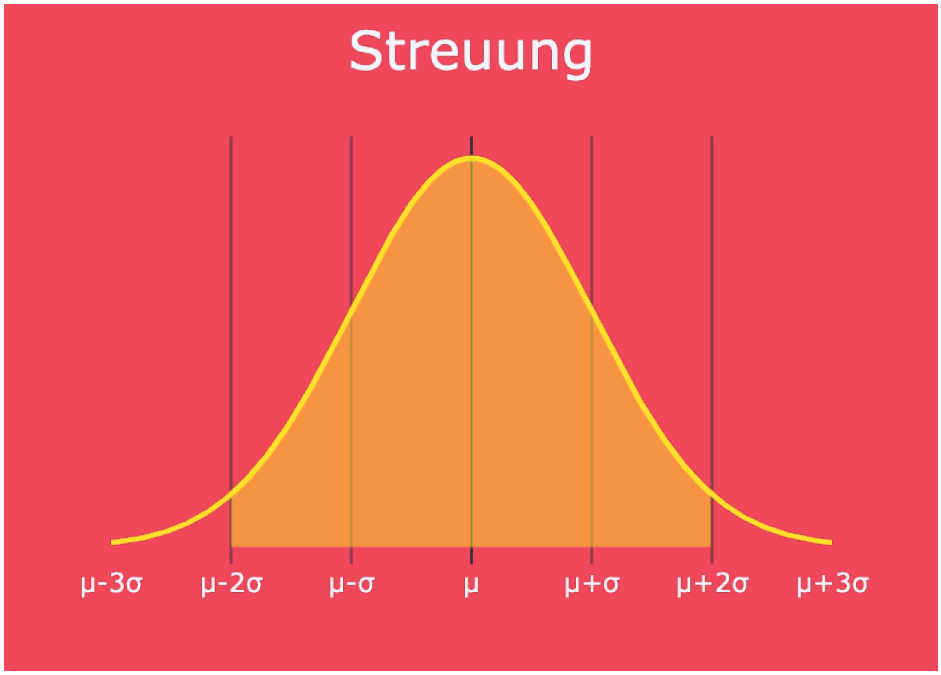
Ein zweiter wichtiger Begriff ist die Standardabweichung σ („sigma"). Sie gibt an, wie weit die Messwerte im Durchschnitt vom Erwartungswert entfernt sind. Hier siehst du eine Normalverteilung mit Erwartungswert 5 und Standardabweichung 1. Bei einer niedrigeren Standardabweichung erhältst du die grüne und bei einer höheren Standardabweichung die blaue Kurve.



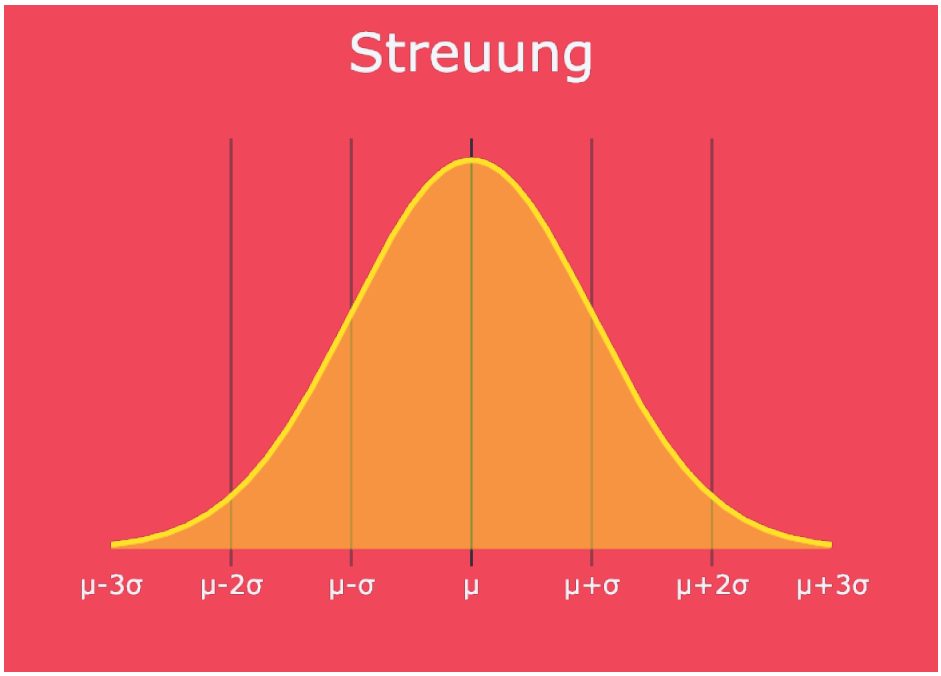
Die Standardabweichung gibt also die Streuung der Messwerte um den Erwartungswert an. Dabei enthält der Bereich (μ–σ) bis (μ+σ), also der Bereich aller Werte, die maximal um die Standardabweichung σ vom Mittelwert μ abweichen, rund 68 % der Messwerte.



Lassen wir es zu, dass die Werte bis um das Zweifache der Standardabweichung vom Mittelwert abweichen, also (μ–2σ) bis (μ+2σ), so erhalten wir einen Bereich mit rund 95 % aller Messwerte.



Ein Bereich mit bis zu dreifacher Abweichung vom Mittelwert (μ–3σ) bis (μ+3σ) enthält rund 99,7 % aller Messwerte.

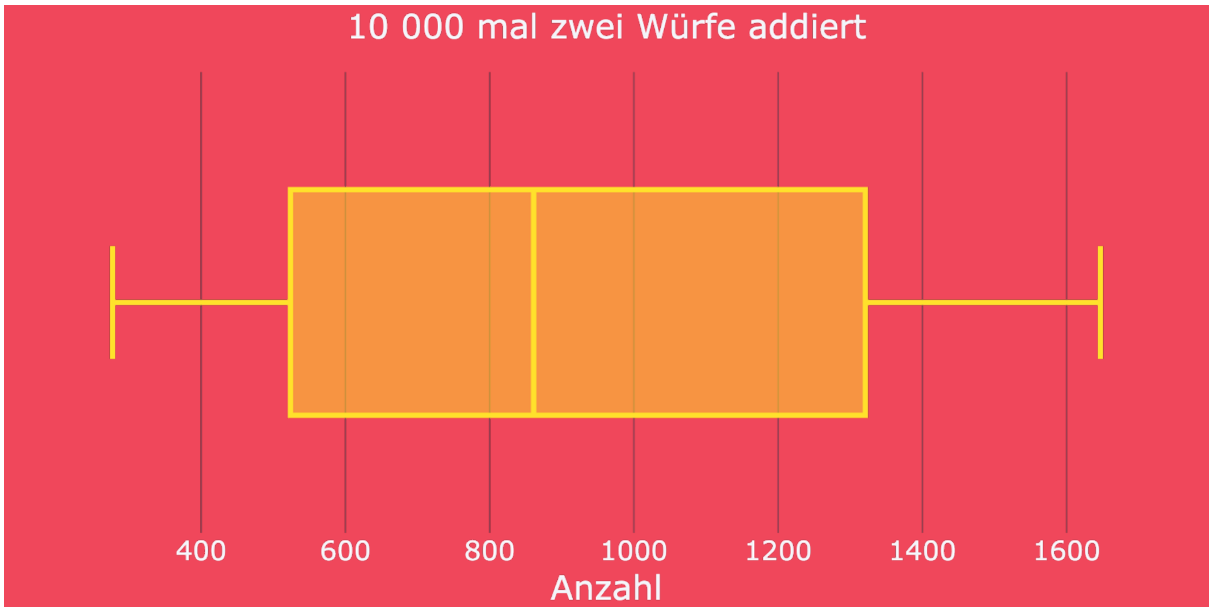


Häufig wird als Streuungsmaß auch das Quadrat der Standardabweichung verwendet, die sogenannte Varianz.

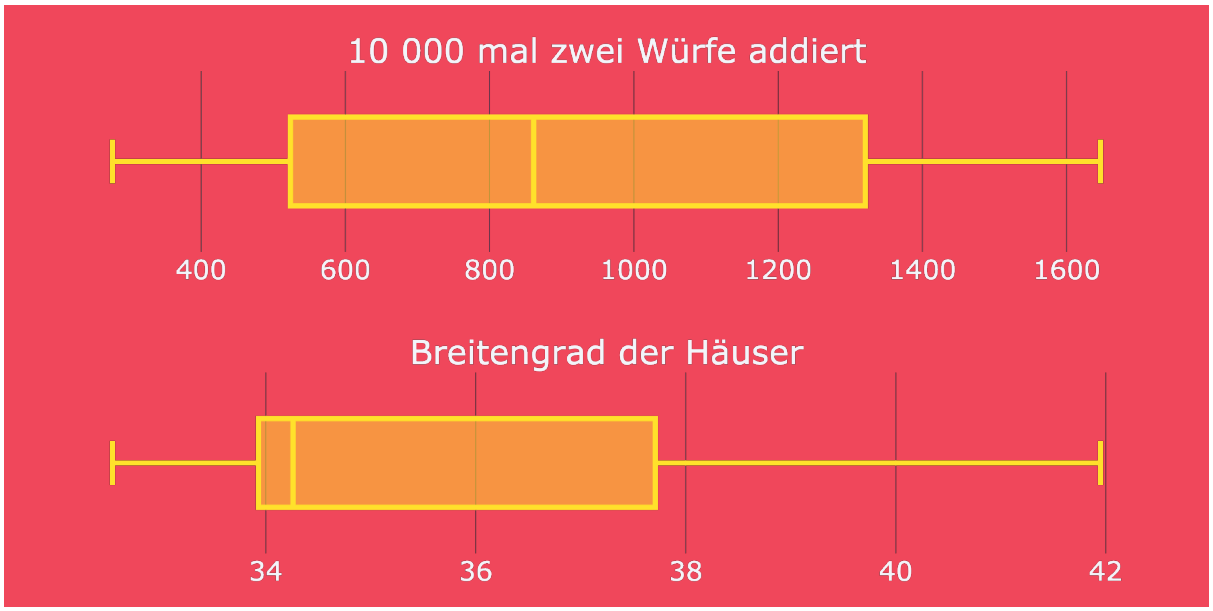
Quantile und Boxplots

Wenn du deine Daten betrachtest, kann es auch wichtig sein zu wissen, wie viel Prozent deiner Daten kleiner als ein bestimmter Wert ist. Hierfür benutzen wir den Begriff des Quantils: Hat das 25 %-Quantil deiner Daten den Wert 13,1, dann sind genau 25 % deiner Werte kleiner oder gleich 13,1. Das 50 %-Quantil ist dabei nichts anderes als der Median deiner Daten, also der Wert, bei dem die Hälfte deiner Daten kleiner oder gleich diesem Wert sind.

Ein Boxplot wie hier im Bild gezeigt erlaubt es dir, einen Überblick über die Verteilung deiner Daten zu erhalten. Er gehört zu unserem Würfelbeispiel mit jeweils zwei geworfenen Würfeln.



Boxplots bilden wichtige Größen deiner Daten ab und erlauben dir so, deine Daten besser zu analysieren. Falls sich in deinen Daten keine Ausreißer befinden, gibt der Strich ganz links den Minimalwert und der ganz rechts den Maximalwert deiner Daten an. Das linke Ende der Box gibt das 25 %-Quantil an, das rechte Ende das 75 %-Quantil. Der Strich in der Mitte zeigt dir den Median. Ausreißer würden als einzelne Punkte angezeigt werden.



Als Kontrast dazu können wir uns auch den Boxplot zum Feature „Breitengrad" des Datensatzes California Housing ansehen, bei dem die Proportionen des Boxplots anders sind. Hier fällt vor allem auf, wie klein der Abstand zwischen dem 25 %-Quantil und dem Median ist. Daraus können wir schließen, dass sich viele der betrachteten Häuser um den 34. Breitengrad herum befinden und sich oberhalb dieses Breitengrades die Häuser stärker streuen.

Die Visualisierung mit Boxplots hilft also dabei, die Verteilung deiner Daten besser einschätzen zu können.

Abschluss

Bei der Datenaufbereitung benutzen wir häufig Wahrscheinlichkeiten und Verteilungen. Wir können zum Beispiel feststellen wollen, ob es sich bei einem Datenpunkt um einen Ausreißer handelt, sprich, ob er zu weit von den erwarteten Werten abweicht. Oder wir wollen berechnen, ob eine unserer Klassen in den Trainingsdaten unterrepräsentiert ist, wir also weniger Datenpunkte dieser Klasse haben, als wir von der Verteilung der Werte eigentlich erwarten würden. In diesem Video hast du die Grundlagen der Wahrscheinlichkeitstheorie kennen gelernt, mit deren Hilfe du deine Daten besser verstehen und analysieren kannst.

Weiterführendes Material

https://studyflix.de/mathematik-schueler/thema/stochastik-153

Disclaimer