Statistik

Corona - Notversion

M. Oettinger

03.06.2020

Lineare Regression

Bei vielen Beobachtungswerten stetiger Merkmale ist die grafische Darstellung als Streudiagramm praktisch, um einen Hinweis auf einen möglichen Zusammenhang zwischen den Daten zu erkennen.

In einer solchen Darstellung zeigt sich eine Abhängigkeit der Merkmale voneinander in einer Häufung von Datenpunkten entlang von Geraden (bei linearer Abhängigkeit) oder allgemeinerer Funktionen. Der Zusammenhang kann durch Anpassung einer Modellfunktion (meist linear) über die lineare Regression geschätzt werden.

Bestimmtheitsmaß

Die lineare Regression beschreibt den Zusammenhang zwischen einer oder mehreren unabhängigen (oder erklärenden) Variablen und einer abhängigen Variablen. Handelt es sich um eine Regression mit einer unabhängigen Variablen, so spricht man von einer einfachen Regression, bei mehreren unabhängigen Variablen von einer multiplen Regression.

Das (empirische) Bestimmtheitsmaß (auch Determinationskoeffizient) ist eine Kennzahl zur Beurteilung der Anpassungsgüte einer Regression. Sie bewegt sich zwischen Null (bei einem vollkommen ungeeigneten Modell) und 1 (bei perfekter Anpassung des Modells an eine Stichprobe), beispielsweise um zu bewerten, wie gut gemessene Werte zu einem Modell passen.

Bestimmtheitsmaß

In bivariater Verteilungen gibt es meist eine Größe (die abhängige Variable y), deren Schwankung (Varianz) über eine zweite Größe (die die unabhängige Variable x) erklärt werden soll. Das ist beispielweise sinnvoll, um über den geschätzten Zusammenhang eine Vorhersage (oder Prognose) für die abhängige Variable zu erstellen. Sind n Wertepaare (x_i, y_i) gegeben, ist jedem Wert x_i ein Merkmals- oder Messwert y_i zugeordnet. Die Zuordnung wird über eine Modellfunktion, in der Statistik auch als Schätzer bezeichnet, beschrieben, die für jedes x_i einen Schätzwert $\hat{y}_i$ bestimmt. Für den linearen Zusammenhang einer linearen Regression gilt also

$$\hat{y}_i = b \cdot x_i + a.$$

mit dem Schätzer $\hat{y}(x) = a \cdot x + b$.

Bestimmtheitsmaß

Jede einzelne Beobachtung y_i lässt sich dann berechnen als die Summe aus dem vorhergesagten Wert (der Schätzung) $\hat{y}_i$ und der Abweichung zwischen beobachtetem und vom Modell vorhergesagten Wert e_i (Fehler bzw. Residuum):

$$y_i = \hat{y}_i + e_i$$

Jetzt lässt sich aus den geschätzten Werten analog zur empirischen Varianz bzw. Variation (der Summe der quadrierten Abweichungen der Merkmalswerte y_i vom arithmetischen Mittel)

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

natürlich auch eine Varianz oder Variation der Schätzwerte bestimmen, indem die Abweichungen *der Schätzwerte* $\hat{y}_i$ vom arithmetischen Mittel quadriert und summiert werden.

Bestimmtheitsmaß

Definition: erklärte Variation

Die Variation der Schätzwerte $\hat{y}_i$ bezeichnet man als erklärte Variation. Sie ist typischerweise geringer als die Gesamtvariation, weil diese zusätzlich die Beiträge der Residuen enthält (es ist $y_i = \hat{y}_i + e_i$):

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

ist die erklärte Variation der Stichprobe $\{x_i; y_i\}$ mit dem Schätzer $\hat{y}_i = f(x_i)$ (meist $\hat{y}_i = a \cdot x_i + b$).

Die Varianz enthält zusätzlich zur erklärten Variation Beiträge der Residuen. Den Quotient aus erklärter Variation und Gesamtvariation bezeichnet man alsBestimmtheitsmaß

Bestimmtheitsmaß

Definition: das Bestimmtheitsmaß R^2

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

wird als das Bestimmtheitsmaß der Regression $\hat{y}_i(x_i)$ bezeichnet.

Zusätzlich gilt mit den Residuen e_i (die Gleichheit ist nicht trivial zu zeigen, deshalb wird hier darauf verzichtet)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

mit der unerklärten Variation

$$\sum_{i=1}^{n}(e_i)^2$$

.

Bestimmtheitsmaß

Was bedeutet das Bestimmtheitsmaß anschaulich? Die naive Näherung für die Werte y in der Stichprobe ist das arithmetische Mittel

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Durch Einbeziehen der unabhängigen Größe x über die lineare Regression kann diese sehr simple Näherung deutlich verbessert werden - um den Beitrag der erklärten Variation (die Abweichungen der Schätzwerte $\hat{y}_i$ vom arithmetischen Mittel - kennt man den Schätzwert in Abhängigkeit von x, hat man genau diese Abweichungen eliminiert). Übrig bleiben aber noch immer die Beiträge der Residuen, der Abweichungen zwischen den Schätzwerten und den gemessenen Werten $e_i = y_i - \hat{y}_i$. Sie werden durch die Regression nicht erklärt, man bezeichnet sie deshalb als unerklärte Variation.

Bestimmtheitsmaß

Das Bestimmtheitsmaß lässt sich über eine Betrachtung zweier Grenzfälle interpretieren:

- für eine vollkommen ungeeignete Modellfunktion erwarten wir, dass sich die Abweichungen nicht über das Modell erklären lassen. In diesem Fall wird die erklärte Variation verschwinden (die Residuen werden dabei maximal). Damit muss sich für R^2 als Quotient aus erklärter Variation und Gesamtvariation der Wert Null ergeben.
- für eine perfekte Modellfunktion verschwinden die Residuen (weil jeder Wert y_i genau durch den Schätzwert $\hat{y}_i$ modelliert wird. In diesem Fall verschwindet die unerklärte Variation und es ist $R^2 = 1$.

Bestimmtheitsmaß

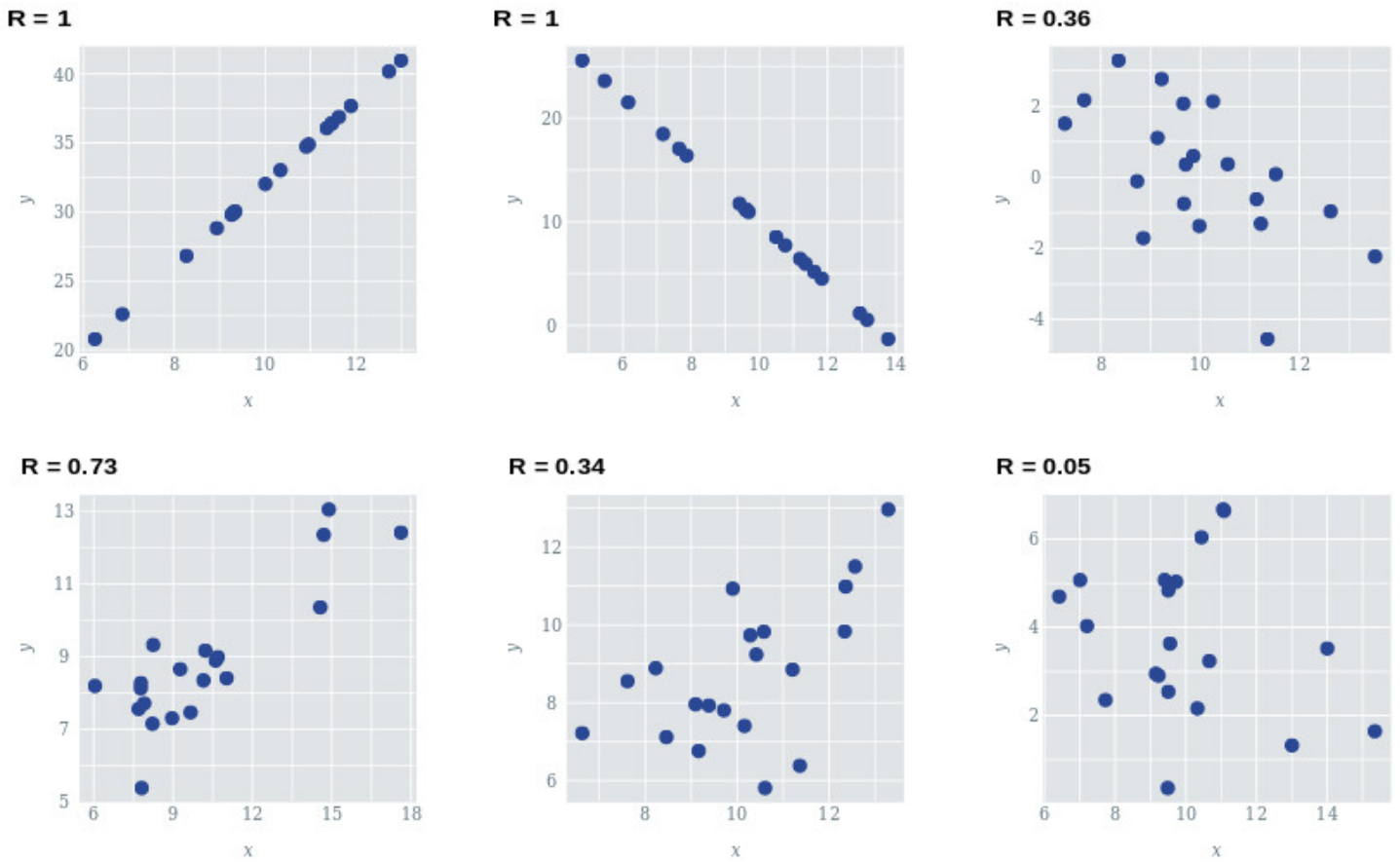


Abbildung: Streudiagramme für unterschiedliche Werte von R^2. Je besser die Datenpunkte auf einer Linie liegen, desto höher ist das Bestimmtheitsmaß. Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt R^2 nahe Null.

Beispiel: Hersteller von Dingen

Für das Beispiel des Herstellers war die gefundene Steigung der angepassten Geraden $a = -0,98$, der Achsenabschnitt $b = 19,73$, die Modellfunktion also

$$y = -0,98 \cdot x + 19,73.$$

Die Schätzwerte lassen sich jetzt direkt berechnen:

$$\hat{y}_i = -0,98 \cdot x_i + 19,73$$

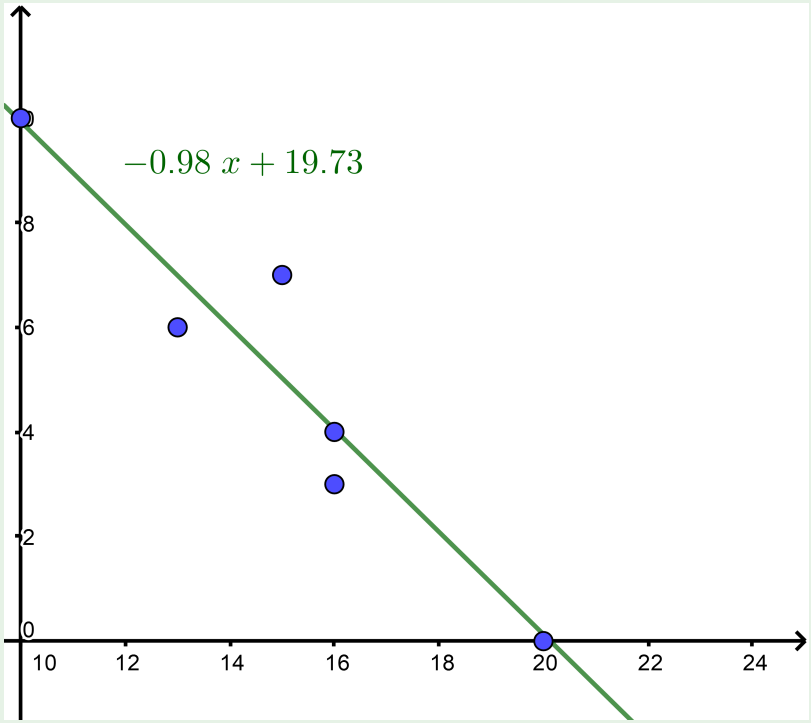


Abbildung: Lineare Regression am Beispiel.

Beispiel: Hersteller von Dingen

Die benötigten Daten zur Berechnung sind die arithmetischen Mittel $\bar{x} = 15$ und $\bar{y} = 5$ und die Werte (die letzte Zeile enthält die Summen der Spalten)

x_i	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
20	0	-5	25	-4,911	24,115
16	3	-2	4	-0,982	0,965
15	7	2	4	0,000	0,000
16	4	-1	1	-0,982	0,965
13	6	1	1	1,964	3,858
10	10	5	25	4,911	24,115
			60		54,018

Tabelle: Daten zur Berechnung von R^2

Bestimmtheitsmaß

Beispiel: Hersteller von Dingen

Die Variation der verkauften Einheiten enthält zusätzlich zur erklärten Variation die Beiträge der Abweichungen vom Schätzer

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 60,$$

die erklärte Variation ist nur etwas geringer

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 54,02.$$

Der Quotient der beiden ist das Bestimmtheitsmaß $R^2 = \frac{54,02}{60} = 0,90$. Der Wert von 90% bedeutet, dass die Daten sehr gut durch das berechnete Modell beschrieben werden.

Bestimmtheitsmaß

Anmerkungen

- Der bereits bekannte Korrelationskoeffizient misst die Qualität eines linearen Zusammenhangs zwischen x_i und y_i, das Bestimmtheitsmaß aber die Qualität der Modellfunktion (des Schätzers).
- Bei linearen Zusammenhängen $\hat{y}_i = ax_i + b$ entspricht das Bestimmtheitsmaß dem Quadrat des Pearson-Korrelationskoeffizienten: $R^2 = r_{xy}^2$.

Beispiel: Hersteller

Der Korrelationskoeffizient im Beispiel war $r_{xy} = -0,95$ (gerundet). Das Bestimmtheitsmaß ist also $R^2 = r_{xy}^2 = (-0,95)^2 = 0,90$.

Ausflug in die Stochastik

Wirft man einen idealen Würfel n-mal, kann man bei den einzelnen Würfen kein Ergebnis vorhersagen (der Würfel fällt aus geometrischen Gründen vollkommen zufällig). Wird das Experiment aber wiederholt, erwartet man eine (etwa) gleichmäßige Verteilung der Augenzahlen auf die Ergebnisse (etwa in einem Sechstel der Würfe jede Augenzahl).

Bei unendlich vielen Würfen gleichen sich die Unterschiede zwischen den verschiedenen Augenzahlen aus, man erwartet jedes Ergebnis mit derselben Häufigkeit.

Definition: Elementarereignis

Wenn sich die Mathematik mit dem Zufall beschäftigt, benötigt sie Modelle von Situationen mit unsicherem Ausgang (die sich mathematisch beschreiben lassen). Diese Modelle nennen wir (ideale) Zufallsexperimente (oder Zufallsversuche) Jedes (ideale) Zufallsexperiment besitzt eine Menge möglicher Versuchsausgänge. Jeder Versuchsausgang wird als Elementarereignis genannt. Die Menge aller Elementarereignisse nennt man den Ereignisraum (oft als Ω abgekürzt).

Beispiel: Würfel

Für einen idealen Würfel ist jede Augenzahl $(1 \ldots 6)$ ein Elementarereignis, die Menge aller Elementarereignisse ist der Ereignisraum

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

Definition: Ereignis

Ein Ereignis ist eine Zusammenfassung von Versuchsausgängen (Elementarereignissen). Mathematisch ausgedrückt ist ein Ereignis eine Teilmenge des Ereignisraums. Jedes Elementarereignis ist selbst ein Ereignis, aber auch Kombinationen von Elementarereignissen.

Beispiel:

Für einen idealen Würfel ist jede Augenzahl ein Ereignis (Teilmenge des Ereignisraums). Das Ereignis 'die Augenzahl ist gerade', mathematisch zu beschreiben über die Teilmenge $B = \{2, 4, 6\}$ des Ereignisraums Ω, ist aber kein Elementarereignis.

Definition: Wahrscheinlichkeit

Die Wahrscheinlichkeit ist ein Maß für die Sicherheit oder Unsicherheit eines Ereignisses in einem Zufallsexperiment. Jedem möglichen Ausgang eines Zufallsexperimentes wird eine reelle Zahl zwischen 0 und 1 zugeordnet, die Wahrscheinlichkeit für das jeweilige Ereignis.

Für die Wahrscheinlichkeit, dass das Ereigniss A eintritt schreibt man $P(A)$ (von probability). Je höher $P(A)$ ist, desto wahrscheinlicher ist das Ereignis A bei einem Zufallsexperiment.

- Tritt A mit Sicherheit ein, so gilt $P(A) = 1$
- Tritt A mit Sicherheit nicht ein, so gilt $P(A) = 0$

Satz: Gesetz der großen Zahlen

Die relative Häufigkeit eines Zufallsergebnisses nähert sich beliebig nahe an die (theoretische) Wahrscheinlichkeit des Ergebnisses an, wenn man das Zufallsexperiment nur oft genug wiederholt.

Das Gesetz der großen Zahlen trifft keine Aussagen über absolute Häufigkeiten! Der Ausgang eines Zufallsexperiments ist bei jeder neuen Wiederholung unbestimmt.

Beispiel: Würfelexperimente

Beim Werfen eines idealen Würfels erwartet man, dass jede mögliche Augenzahl *etwa* gleich oft auftritt - bei mehrfacher Wiederholung werden Abweichungen zwischen den relativen Häufigkeiten geringer. Bei sehr vielen Wiederholungen wird der Einfluss des Zufalls geringer, weil sich zufällige Abweichungen in den absoluten Häufigkeiten niederschlagen. Durch die große Zahl von Wiederholungen wird der Beitrag zu den relativen Häufigkeiten klein - sie liegen nahe bei 1/6.

Bei vielen Zufallsexperimenten ist es schwierig, die Wahrscheinlichkeit von Ereignissen direkt zu bestimmen. In solchen Fällen wird für P(A) das Experiment sehr oft wiederholt und die Wahrscheinlichkeit des Ereignisses A als Grenzwert der relativen Häufigkeiten $h(A)$ des Ereignisses A für $n \to \infty$ geschätzt.

Für den idealen Würfel (gleichverteilt mit 6 Seiten) ist die erwartete Wahrscheinlichkeit für jede der Augenzahlen A

$$P(A) = \lim_{n\to\infty} h(A) = \lim_{n\to\infty} \frac{\frac{n}{6} \pm \Delta}{n} = \frac{1}{6}$$

Definition: Laplace-Experiment

Ein Laplace-Experiment ist ein Zufallsversuch, bei dem die Wahrscheinlichkeiten aller möglichen Ergebnisse gleich sind. Ein typisches Beispiel für ein Laplace-Experiment ist das Werfen einer Münze. Für Laplace-Experimente gilt für die Wahrscheinlichkeit $P(E)$, dass das Ergebnis E eintritt

$$P(E) = \frac{\text{Zahl der Fälle, die das Ergebnis } E \text{ liefern}}{\text{Zahl aller möglichen Fälle}}$$

Ist die Wahrscheinlichkeit gesucht, mit einem idealen Würfel mit 6 Seiten eine 3 zu werfen, ist die Zahl der möglichen Ergebnisse 6, in einem Fall ist das Ergebnis 3, also gilt

$$P(E) = \frac{1}{6}$$

Bei zwei idealen (und unterscheidbaren) Würfeln ist die Augensumme eine aus zusammengesetzte Größe, die in einzelne Laplace-Experimente zerlegt werden kann: es gibt für den ersten Würfel 6 mögliche Augenzahlen, der zweite Würfel wird davon aber nicht beeinflusst, er hat in jedem Fall ebenfalls 6 Möglichkeiten. Insgesamt gibt es also $n = 6 \cdot 6$ mögliche Kombinationen. Die Zahl der möglichen Kombinationen n_i für eine bestimmte Augensumme i unterscheidet sich je nach Summe, die Wahrscheinlichkeit ist

$$P(A_i) = \frac{\text{Zahl der günstigen Fälle}}{\text{Zahl der möglichen Fälle}} = \frac{n_i}{n}$$

Beispiel: Augensumme zweier Würfel

Die Wahrscheinlichkeit, mit zwei Würfeln eine 7 oder eine 4 zu erwürfeln ist

$$P(A_7) = \frac{n_7}{n} = \frac{6}{36} = \frac{1}{6}.$$

$$P(A_4) = \frac{n_4}{n} = \frac{3}{36} = \frac{1}{12}.$$

Die Wahrscheinlichkeit dafür, dass beide Würfel dieselbe Zahl zeigen (Pasch) ist

$$P = \frac{6}{36} = \frac{1}{6}$$

Beispiel: Augensumme zweier Würfel

Die beiden Grenzfälle:
die Wahrscheinlichkeit, mit zwei Würfeln eine Augensumme von 1 zu erwürfeln ist

$$P(A_1) = \frac{n_1}{n} = \frac{0}{36} = 0$$

- das wird mit Sicherheit nicht eintreten.
Die Wahrscheinlichkeit, mit zwei Würfeln eine Augensumme zwischen 2 und 12 zu erwürfeln, ist

$$P(E) = \frac{36}{36} = 1$$

- irgendeine Augensumme müssen die Würfel ja zeigen.

Beispiel:

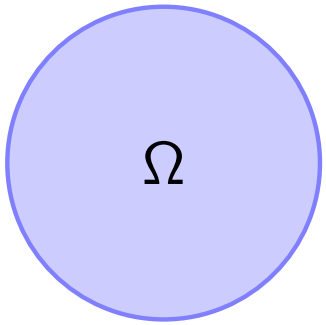
zieht man aus einer Urne mit 10 roten, 15 blauen und 5 grünen Kugeln eine Kugel zufällig, ist das kein Laplace-Experiment. Die Versuchsausgänge rot, blau und grün (für die herausgegriffene Kugel) haben nicht die gleiche Chance, einzutreten.

Es lässt sich aber leicht auf ein Laplace-Experiment zurückführen: nummeriert man die Kugeln ('heimlich') durch, besitzt jede der 30 Kugeln ihre eigene Identität. Nun wird jede Nummer mit der gleichen Wahrscheinlichkeit gezogen - wir haben aus dem Urnenbeispiel vorübergehend ein Laplace-Experiment gemacht. Die Wahrscheinlichkeit dafür, eine rote Kugel zu ziehen ist dann

$$P(\text{rot}) = \frac{10 \text{ rote Kugeln}}{30 \text{ mögliche Kugeln}} = \frac{1}{3}$$

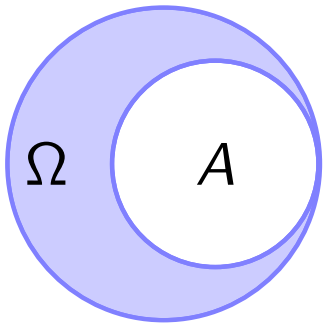
Rechnen mit Wahrscheinlichkeiten

Sind A und B zwei Ergebnismengen eines Zufallsexperiments und Ω die Menge aller möglichen Ergebnisse, so können die Wahrscheinlichkeiten für die Ergebnisse miteinander verknüpft werden:



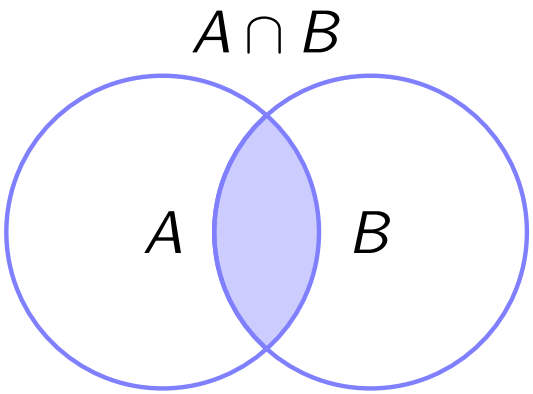
$P(\Omega) = 1$, wenn Ω alle möglichen Versuchsausgänge umfasst.

$\Omega - A$

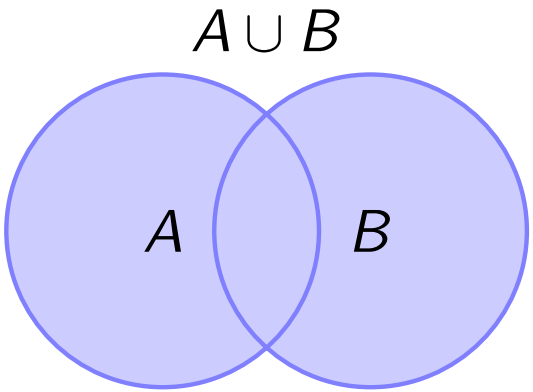


Die Gegenwahrscheinlichkeit zu A ist die Wahrscheinlichkeit dafür, dass nicht Ereignis A eintritt: $P(\bar{A}) = 1 - P(A)$

Rechnen mit Wahrscheinlichkeiten



Die Wahrscheinlichkeit für das Ereignis 'A und B' ist
$P(A \cap B) = P(A) \cdot P(B)$,
wenn die Ereignisse A und B voneinander unabhängig sind



Die Wahrscheinlichkeit für 'A oder B' ist
$P(A \cup B) = P(A) + P(B) - P(A) \cap P(B)$

Rechnen mit Wahrscheinlichkeiten

Beispiel:

Eine Stichprobe liefert (geschätzte) Wahrscheinlichkeiten dafür, welche Speisen und Getränke beim Italiener bestellt werden: 60% der befragten Kunden bestellten Pizza (Ausgang A), 30% Spaghetti (B) und 10% andere Gerichte (C).

Wenn keiner der Kunden mehrere Gerichte bestellt, ist die Wahrscheinlichkeit dafür, dass irgendein Gericht bestellt wird natürlich $P(\Omega) = P(A) + P(B) + P(C) = 1$.

Die Wahrscheinlichkeit dafür, dass nichts gegessen wurde, kann über die Gegenwahrscheinlichkeit berechnet werden:
$1 - (P(A) + P(B) + P(C)) = 0$.

Rechnen mit Wahrscheinlichkeiten

Beispiel:

Wurde von den Gästen gleichzeitig bei 60% der Besuche Rotwein (D) und bei 40% (kaltes!) Pils getrunken (E), ist die Wahrscheinlichkeit dafür, dass Spaghetti mit Pils bestellt werden

$$P(A) \cap P(E) = P(A) \cdot P(E) = \frac{6}{10} \cdot \frac{3}{1}0 = \frac{18}{100} = 18\%$$

(unter der Annahme, dass die Wahl des Getränks von der gewählten Speise unabhängig ist!), die Wahrscheinlichkeit dafür, dass ein Kunde Spaghetti oder Pils bestellt ist

$$P(A) \cup P(E) - P(A) \cap P(E) = \frac{6}{10} + \frac{3}{10} - \frac{18}{100} = 72\%$$

Binomialkoeffizient

Binomialkoeffizienten geben an, auf wie viele verschiedene Arten man k Objekte aus einer Menge von n verschiedenen Objekten auswählen kann. Das Experiment wird dabei ohne Zurücklegen und ohne Beachtung der Reihenfolge durchgeführt. Es entspricht dem Ziehen von Kugeln aus einer Urne

Definition: Binomialkoeffizient

Für den Binomialkoeffizienten 'n über k' gilt

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

Binomialkoeffizient

Beispiel: Lotto

Beim vereinfachten Lotto '6 aus 49' werden $k = 6$ Zahlen aus einer Menge von $n = 49$ Zahlen gezogen. Die Zahl der möglichen Kombinationen ist

$$\binom{49}{6} = \frac{n!}{k!(n-k)!} = \frac{49!\cdot}{6! \cdot 43!}$$
$$= \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} = 13983816$$

Nur in einem dieser Fälle liegen sechs Richtige vor (Laplace-Experiment), also ist die Wahrscheinlichkeit dafür, die richtige Kombination zu erraten $P = \frac{1}{13983816}$.