

Unmöglich oder Unwahrscheinlich?

Geometrie in vielen Dimensionen hilft statistische Probleme zu lösen.

Seit es Statistik gibt, stellt sich die Frage "Unmöglich oder Unwahrscheinlich?". Statistiker versuchen, Daten mit Hilfe von statistischen Modellen zu fassen. Statt wie in der Physik zu berechnen, was geschehen wird, weist man Wahrscheinlichkeiten zu. Beim Würfeln ist der nächste Wurf nicht vorhersagbar, aber langfristig wird jede Seite gemäß den Wahrscheinlichkeiten fallen. Was aber, wenn man nicht lange Würfeln kann? Wenn in zehn Würfeln keine Sechs gefallen ist, sollte man annehmen der Würfel sei gezinkt, oder hat man einfach nur Pech gehabt? Meine Forschung beschäftigt sich mit Methoden zur Unterscheidung von unmöglichen und unwahrscheinlichen Ereignissen.

In meiner Doktorarbeit habe ich statistische Modelle für so genannte kategorielle Daten untersucht. Kategorielle Daten tauchen immer dann auf, wenn in Kategorien einsortiert werden kann, z.B. in Studien zur Wirksamkeit von Medikamenten. Dort werden die Patienten nach Alter, Geschlecht, Behandlungserfolg und weiteren Kategorien sortiert. Anhand einer großen Tabelle soll nun eine Aussage über die Wirksamkeit eines Medikaments gemacht werden. Ein statistisches Modell besteht aus der Gesamtheit dieser Kategorien zusammen mit zusätzlichen Annahmen über ihre Zusammenhänge und Wahrscheinlichkeiten für das Auftreten bestimmter Merkmalskombinationen. Eine Aufgabe der Statistik ist es, Hypothesen über das Modell zu testen. Unter der Annahme, dass ein Medikament wirksam ist, fragt man nach der Wahrscheinlichkeit, die in den realen Daten beobachteten Werte zu erhalten. Ist diese Wahrscheinlichkeit zu klein, muss die Hypothese der Wirksamkeit verworfen werden. Probleme mit dieser Methode treten immer dann auf, wenn die große Tabelle Nullen enthält, es also z.B. keine Patientin in einem gewissen Alter gab, für die sich ein Behandlungserfolg eingestellt hat. Es muss unterschieden werden, ob Kombinationen prinzipiell unmöglich oder nur unwahrscheinlich sind. Wenn dieser Unterschied auch klein erscheint, ist er für die Funktionsweise der Tests entscheidend.

Als Mathematiker stehen für mich nicht die Daten selbst, sondern grundlegende, theoretische Fragen im Vordergrund. Bei komplizierten Modellen mit vielen Kategorien kann man unter Umständen nicht voraussagen, welche Kombinationen unmöglich sind. Zugehörige Hypothesentests können nur dann eingesetzt werden, wenn genug Daten vorliegen, um sicherzustellen, dass auch unwahrscheinliche Kombinationen beobachtet wurden. Genau das kann in der Medikamentenforschung nicht vorausgesetzt werden. Es

ist daher entscheidend, genau zu wissen, welche Kombinationen möglich sind. Um dieses Problem zu bearbeiten, habe ich es in ein geometrisches Problem umformuliert. Damit stand mir der vielfältige Werkzeugkasten der algebraischen Geometrie zur Verfügung.

Die Stärke der Mathematik besteht in der Fähigkeit zur Abstraktion. Manchmal ist erst, wenn ein Problem völlig von seinem Originalgewand entkleidet wurde, der Blick frei auf das Wesentliche und damit die Lösung. Man sagt gerne, ein Problem sei schon halb gelöst, sowie es verstanden ist. Die Abstraktion bestand für mich darin, das statistische Modell als geometrisches Objekt zu betrachten. Dazu wird jede mögliche Zuweisung von Wahrscheinlichkeiten zu den Merkmalskombination als Punkt in einem Koordinatensystem aufgefasst. Die Gesamtheit aller solcher Zuweisungen ist ein geometrisches Abbild des statistischen Modells. In einigen winzigen Beispielen erhält man sogar bekannte geometrische Objekte. Gibt es nur zwei Merkmale die je zwei mögliche Werte haben, so findet man etwa einen dreidimensionalen Tetraeder. Die Dimension des Raumes in dem Geometrie betrieben wird, entspricht der Anzahl der Tabelleneinträge minus eins und kann somit sehr groß werden.

Wie kann man nun mit dieser Geometrie arbeiten und verstehen welche Merkmalskombinationen unmöglich sind? Hier kommt der Werkzeugkasten der algebraischen Geometrie ins Spiel. Es stellt sich heraus, dass statistische Modelle durch Polynomgleichungen beschrieben werden können. Man findet einen Satz von Gleichungen, so dass das Modell genau die Lösungsmenge dieser Gleichungen ist. Diese Beschreibung begegnet einem schon in der Schule: Der Kreis ist die Lösungsmenge einer einzigen quadratischen Gleichung in zwei Variablen. (Die Summe der Quadrate der Koordinaten ist gleich dem Quadrat des Radius.) Ein statistische Modell hingegen kann die Lösungsmenge von tausenden von Gleichungen in hunderten von Variablen sein. Niemand würde je diese Gleichungen auf ein Blatt Papier schreiben. Hier wird stets mit Computerhilfe gearbeitet. Die Gleichungen zu finden, ist oft so kompliziert, dass es sich sogar lohnt, sie zu tabellieren! Zusammen mit meinem Kollegen Johannes Rauh habe ich eine Internetdatenbank für Gleichungen von statistischen Modellen erstellt.

Ausgestattet mit der richtigen Abstraktion, geometrischen Ideen und einem Computer reduziert sich das Problem der unmöglichen Kombinationen auf das Lösen von Gleichungen, ein grundlegendes Problem der Mathematik. Um zu testen, ob eine Merkmalskombination wirklich unmöglich ist, wird ein Testpunkt konstruiert und geprüft, ob er die Gleichungen des Modells erfüllt. Zunächst habe ich versucht, die allgemeine Lösungstheorie für Polynomgleichungen zu verwenden, aber die Anzahl der Gleichungen und Variablen ist einfach zu groß. Um hier weiter zu kommen, war es unabdingbar, eine

spezielle Struktur auszunutzen. Meine Modelle sind beschrieben durch Binomgleichungen, d.h. Polynome mit genau zwei Termen. Ein Experte für diese Gleichungen ist Prof. Bernd Sturmfels von der University of California in Berkeley. Er gab mir den entscheidenden Tipp: Statt mit den Gleichungen selbst sollte man nur mit den zwei Exponenten jedes Binoms arbeiten. Dann verhalten sich Binomgleichungssysteme fast wie die linearen Gleichungssysteme, die man schon in der Schule löst. Ich habe Spezialalgorithmen, die auf dieser Idee basieren, in einem Computerprogramm implementiert. Damit können nun sehr große Binomgleichungssysteme gelöst werden. Für die statistischen Modelle hatte ich somit sowohl Werkzeuge zum Finden der Gleichungen als auch für ihre Lösung zur Hand.

Mathematik ist allgemein und mein Computerprogramm nicht auf statistische Modelle beschränkt. Es löst Binomgleichungssysteme unabhängig von ihrer konkreten Anwendung und kann vielfältig eingesetzt werden. Eine weitere Herausforderung wartete in einem Forschungsartikel von Prof. Sturmfels, Prof. Steve Evans von der UC Berkeley, und der Doktorandin Caroline Uhler. Sie zeigten, dass bestimmte Zufallsbewegungen auch durch Binomgleichungen beschrieben werden. Für deren Analyse verwendeten die Autoren zunächst die allgemeine Software für Polynomgleichungen, was die Anzahl der zugänglichen Beispiele stark beschränkte. Am Ende ihres Artikels formulieren sie eine Vermutung über die Gleichungssysteme der Zufallsbewegung. Sie besagt, dass die Gleichungen keine redundante Information enthalten und daher die einfachste mögliche Beschreibung der Zufallsbewegung sind. Redundante Information steckt in zu hohen Potenzen der Variablen in den Gleichungen. Das Quadrat einer Variablen ist null, genau dann wenn die Variable selbst null ist. Man sollte also die zweite Beschreibung verwenden, da sie einfacher ist. Die Vermutung besagt, dass Redundanzen dieser Art bei den Zufallsbewegungen nicht auftauchen. Mein Programm konnte nach einigen Tagen Rechenzeit die Lösungsmenge eines größeren Beispiels bestimmen. Darüber hinaus fand es auch eine minimale Beschreibung. Da sie nicht mit der Originalbeschreibung übereinstimmt, muss redundante Information vorliegen. Die Situation ist also leider nicht so einfach, wie erhofft. Für die Zukunft ergab sich die neue Fragestellung, Gleichungssysteme ohne Redundanzen zu finden.

Das Problem, unmögliche und unwahrscheinliche Ergebnisse voneinander zu unterscheiden, hat mich auf die Spur der Binomgleichungssysteme gebracht. Um es zu lösen, mussten sowohl Theorie als auch Praxis weiterentwickelt werden. Dadurch wurden wieder neue Ansätze auch für andere Probleme greifbar. Dieses Wechselspiel zwischen Theorie und Anwendung macht die Schönheit der Mathematik aus. Meinen

Werkzeugkasten für statistische Modelle werde ich als nächstes in einem Forschungsaufenthalt am Isaac Newton Institut in Cambridge im Rahmen des Programms zur statistischen Versuchsplanung nutzen.